CHEP 2016 Conference, San Francisco, October 8-14, 2016

Monday 10 October 2016 - Friday 14 October 2016
San Francisco Marriott Marquis

Book of Abstracts
## Contents

- Experiment Management System for the SND Detector 0
- Reconstruction software of the silicon tracker of DAMPE mission 2
- HEPData - a repository for high energy physics data exploration 3
- Reconstruction of Micropattern Detector Signals using Convolutional Neural Networks 4
- Federated data storage system prototype for LHC experiments and data intensive science 4
- BelleII@home: Integrate volunteer computing resources into DIRAC in a secure way 7
- Reconstruction and calibration of MRPC endcap TOF of BESII 8
- RootJS: Node.js Bindings for ROOT 6 9
- C++ Software Quality in the ATLAS experiment: Tools and Experience 10
- An automated meta-monitoring mobile application and frontend interface for the WLCG computing model 11
- Experience of Google’s latest Deep Learning library, TensorFlow, with Docker in a WLCG cluster 12
- Flexible online monitoring for high-energy physics with Pyrame 13
- Simulation of orientational coherent effects via Geant4 14
- Detector control system for the AFP detector in ATLAS experiment at CERN 15
- The InfiniBand based Event Builder implementation for the LHCb upgrade 16
- JavaScript ROOT v4 17
- The evolution of monitoring system: the INFN-CNAF case study 18
- Statistical and Data Analysis Package in SWIFT 19
- Analysis Tools in Geant4 10.2 20
- Online & Offline Storage and Processing for the upcoming European XFEL Experiments 21
- Future approach to tier-0 extension 22
- Internal security consulting, reviews and penetration testing at CERN 23
Web technology detection - for asset inventory and vulnerability management 24
Integrating Containers in the CERN Private Cloud 25
CERN Computing in Commercial Clouds 26
Real-time complex event processing for cloud resources 27
Benchmarking cloud resources 28
RapidIO as a multi-purpose interconnect 29
Evolution of CERN Print Services 30
Windows Terminal Server Orchestration 31
Update on CERN Search based on SharePoint 2013 33
First experience with the new .CERN top level domain 34
Flash is Dead. Finally. 35
Access to WLCG resources: The X509-free pilot 36
Deploying FTS with Docker Swarm and Openstack Magnum 37
DNS Load Balancing in the CERN Cloud 38
Customization of the general fitting tool genfit2 in PandaRoot 39
A world-wide databridge supported by a commercial cloud provider 40
DPM Evolution: a Disk Operations Management Engine for DPM 41
Making the most of cloud storage - a toolkit for exploitation by WLCG experiments. 42
EOS Cross Tier Federation 43
EOS Developments 44
Using HEP Computing Tools, Grid and Supercomputers for Genome Sequencing Studies 45
Support Vector Machines in HEP 46
Production Management System for AMS Remote Computing Centers 48
Evolution of Monitoring System for AMS Science Operation Center 49
Storage Strategy of AMS Science Data at Science Operation Center at CERN 50
AMS Data Production Facilities at Science Operation Center at CERN 51
Context-aware distributed cloud computing using CloudScheduler 52
XrootdFS, a posix file system for XrootD 53
Research and application of OpenStack in Chinese Spallation Neutron Source Computing environment 54
Next Generation high performance, multi-dimensional scalable data transfer 55
Globally Distributed Software Defined Storage 56
An efficient, modular and simple tape archiving solution for LHC Run-3 57
Data Center Environmental Sensor for safeguarding the CERN Data Archive 58
PaaS for web applications with OpenShift Origin 59
40-Gbps Data-Acquisition System for NectarCAM, a camera for the Medium Size Telescopes of the Cherenkov Telescope Array 60
Model-independent partial wave analysis using a massively-parallel fitting framework 61
CERN openlab Researched Technologies That Might Become Game Changers in Software Development 62
CERN openlab Knowledge Transfer and Innovation Projects 64
CERN data services for LHC computing 66
CERN AFS Replacement project 67
From Physics to industry: EOS outside HEP 68
Big Data Analytics for the Future Circular Collider Reliability and Availability Studies 69
CERNBox: the data hub for data analysis 70
Tape SCSI monitoring and encryption at CERN 71
Wi-Fi service enhancement at CERN 72
A graphical performance analysis and exploration tool for Linux-perf 73
The role of dedicated computing centers in the age of cloud computing 74
Next Generation Monitoring 75
The merit of data processing application elasticity 76
Identifying memory allocation patterns in HEP software 77
Finding unused memory allocations with FOM-tools 78
SDN-NGenIA A Software Defined Next Generation integrated Architecture for HEP and Data Intensive Science 79
A Comparison of Deep Learning Architectures with GPU Acceleration and Their Applications 80
CERN’s Ceph infrastructure: OpenStack, NFS, CVMFS, CASTOR, and more! 81
LHCb Kalman Filter cross architectures studies 82
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object-based storage integration within the ATLAS DDM system</td>
<td>92</td>
</tr>
<tr>
<td>Experiences with the new ATLAS Distributed Data Management System</td>
<td>93</td>
</tr>
<tr>
<td>Rucio Auditor - Consistency in the ATLAS Distributed Data Management System</td>
<td>93</td>
</tr>
<tr>
<td>How To Review 4 Million Lines of ATLAS Code</td>
<td>94</td>
</tr>
<tr>
<td>Frozen-shower simulation of electromagnetic showers in the ATLAS forward calorimeters</td>
<td>95</td>
</tr>
<tr>
<td>ATLAS World-cloud and networking in PanDA</td>
<td>95</td>
</tr>
<tr>
<td>Data intensive ATLAS workflows in the Cloud</td>
<td>96</td>
</tr>
<tr>
<td>Volunteer Computing Experience with ATLAS@Home</td>
<td>97</td>
</tr>
<tr>
<td>Assessment of Geant4 Maintainability with respect to Software Engineering References</td>
<td>97</td>
</tr>
<tr>
<td>Migrating the Belle II Collaborative Services and Tools</td>
<td>98</td>
</tr>
<tr>
<td>ROOT and new programming paradigms</td>
<td>99</td>
</tr>
<tr>
<td>Status and Evolution of ROOT</td>
<td>100</td>
</tr>
<tr>
<td>Computing Performance of GeantV Physics Models</td>
<td>101</td>
</tr>
<tr>
<td>Software Aspects of the Geant4 Validation Repository</td>
<td>102</td>
</tr>
<tr>
<td>MPEXS: A CUDA MonteCarlo of the simulation of electromagnetic interactions</td>
<td>103</td>
</tr>
<tr>
<td>Multi-threaded Geant4 on Intel Many Integrated Core architectures</td>
<td>103</td>
</tr>
<tr>
<td>CEPHFS: a new generation storage platform for Australian high energy physics</td>
<td>104</td>
</tr>
<tr>
<td>Evaluation of lightweight site setups within WLCG infrastructure</td>
<td>105</td>
</tr>
<tr>
<td>The Cherenkov Telescope Array production system for Monte Carlo simulations and analysis</td>
<td>106</td>
</tr>
<tr>
<td>Processing and Quality Monitoring for the ATLAS Tile Hadronic Calorimeter data</td>
<td>107</td>
</tr>
<tr>
<td>Data acquisition and processing in the ATLAS Tile Calorimeter Phase-II Upgrade Demonstrator</td>
<td>108</td>
</tr>
<tr>
<td>Validation of Electromagnetic Physics Models for Parallel Computing Architectures in the GeantV project</td>
<td>109</td>
</tr>
<tr>
<td>Stochastic optimisation of GeantV code by use of genetic algorithms</td>
<td>109</td>
</tr>
<tr>
<td>GeantV phase 2: developing the particle transport library</td>
<td>110</td>
</tr>
<tr>
<td>New Directions in the CernVM File System</td>
<td>111</td>
</tr>
<tr>
<td>Toward a CERN Virtual Visit Service</td>
<td>111</td>
</tr>
<tr>
<td>Performance and evolution of the DAQ system of the CMS experiment for Run-2</td>
<td>112</td>
</tr>
<tr>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Upgrading and Expanding Lustre Storage for use with the WLCG</td>
<td>113</td>
</tr>
<tr>
<td>XRootD Popularity on Hadoop Clusters</td>
<td>114</td>
</tr>
<tr>
<td>The Vacuum Platform</td>
<td>114</td>
</tr>
<tr>
<td>The Machine/Job Features mechanism</td>
<td>115</td>
</tr>
<tr>
<td>Integration of grid and local batch system resources at DESY</td>
<td>116</td>
</tr>
<tr>
<td>Web Proxy Auto Discovery for WLCG</td>
<td>116</td>
</tr>
<tr>
<td>Visualization of historical data for the ATLAS detector controls</td>
<td>117</td>
</tr>
<tr>
<td>The detector read-out in ALICE during Run 3 and 4</td>
<td>118</td>
</tr>
<tr>
<td>Grid Access with Federated Identities</td>
<td>119</td>
</tr>
<tr>
<td>Recent progress of Geant4 electromagnetic physics for LHC and other applications</td>
<td>119</td>
</tr>
<tr>
<td>Scaling Up a CMS Tier-3 Site with Campus Resources and a 100 Gb/s Network Connection: What Could Go Wrong?</td>
<td>120</td>
</tr>
<tr>
<td>Facilitating the deployment and exploitation of HEP Phenomenology codes using INDIGO-Datacloud tools</td>
<td>121</td>
</tr>
<tr>
<td>C2MON: a modern open-source platform for data acquisition, monitoring and control</td>
<td>122</td>
</tr>
<tr>
<td>ARC CE cache as a solution for lightweight Grid sites in ATLAS</td>
<td>123</td>
</tr>
<tr>
<td>Exploiting Opportunistic Resources for ATLAS with ARC CE and the Event Service</td>
<td>123</td>
</tr>
<tr>
<td>Evolution of user analysis on the Grid in ATLAS</td>
<td>124</td>
</tr>
<tr>
<td>AthenaMT: Upgrading the ATLAS Software Framework for the Many-Core World with Multi-Threading</td>
<td>125</td>
</tr>
<tr>
<td>An Oracle-based Event Index for ATLAS</td>
<td>125</td>
</tr>
<tr>
<td>Collecting conditions usage metadata to optimize current and future ATLAS software and processing</td>
<td>126</td>
</tr>
<tr>
<td>Integration of the Titan supercomputer at OLCF with the ATLAS Production System</td>
<td>127</td>
</tr>
<tr>
<td>First use of LHC Run 3 Conditions Database infrastructure for auxiliary data files in ATLAS</td>
<td>128</td>
</tr>
<tr>
<td>Multi-threaded ATLAS Simulation on Intel Knights Landing Processors</td>
<td>128</td>
</tr>
<tr>
<td>The Fast Simulation Chain for ATLAS</td>
<td>129</td>
</tr>
<tr>
<td>Integration of the Chinese HPC Grid in ATLAS Distributed Computing</td>
<td>130</td>
</tr>
<tr>
<td>How to keep the Grid full and working with ATLAS production and physics jobs</td>
<td>130</td>
</tr>
<tr>
<td>A study of data representations in Hadoop to optimize data storage and search performance of the ATLAS EventIndex</td>
<td>131</td>
</tr>
</tbody>
</table>
AGIS: Integration of new technologies used in ATLAS Distributed Computing 201

A Roadmap to Continuous Integration for ATLAS software development 202

ATLAS Metadata Interface (AMI), a generic metadata framework 204

Deploying the ATLAS Metadata Interface (AMI) on the cloud with Jenkins 205

ATLAS Fast Physics Monitoring: TADA 206

A new mechanism to persistify the detector geometry of ATLAS and serving it through an experiment-agnostic REST API 207

ATLAS Distributed Computing experience and performance during the LHC Run-2 208

Benefits and performance of ATLAS approaches to utilizing opportunistic resources 209

DIRAC universal pilots 211

The ATLAS Computing Agora : a resource web site for citizen science projects 212

HiggsHunters - a citizen science project for ATLAS 213

Big Data Analytics Tools as Applied to ATLAS Event Data 215

Event visualisation in ATLAS: current software technologies / future prospects and trends 216

DIRAC in Large Particle Physics Experiments 217

ATLAS Data Preparation in Run 2 218

Caching Servers for ATLAS 219

An Analysis of Reproducibility and Non-Determinism in HEP Software and ROOT Data 220

New developments in Geant4 version 10 series 221

Application of econometric and ecology analysis methods in physics software 222

Opportunistic Computing with Lobster: Lessons Learned From Scaling up to 25k Non-Dedicated Cores 224

Mixing HTC and HPC Workloads With HTCondor and Slurm 225

Dark matter research cluster based on computational science 226

JUNO performance evaluation and optimization on virtual platform 227

Opticks : GPU Optical Photon Simulation for Particle Physics with NVIDIA OptiX 228

First results from a combined analysis of CERN computing infrastructure metrics 229

Online Data Compression in the ALICE O2 facility 230

Hadoop and friends - first experience at CERN with a new platform for high throughput analysis steps 231
OCCAM: a flexible, multi-purpose and extendable HPC cluster 232

Automatised data quality monitoring of the LHCb Vertex Locator 233

Monitoring performance of a highly distributed and complex computing infrastructure in LHCb 234

Using OSG Computing Resources with (iLC)DIRAC 235

Indico 2.0 - the whole Iceberg 236

Tracking Machine Learning Challenge 237

First experiences with a parallel architecture testbed in the LHCb trigger system 238

Evolution through revolution - A 2-year Agile experience with the Indico Team 240

GEANT4-based full simulation of the PADME experiment at the DAFNE BTF 241

Development and test of a DRS4-based DAQ system for the PADME experiment at the DAFNE BTF 242

Evaluation of clustering algorithms at the < 1 GeV energy scale for the electromagnetic calorimeter of the PADME experiment 243

LHCbDIRAC as Apache Mesos microservices 244

STAR Online Meta Data Collection Framework: Integration with the Pre-existing Controls Infrastructure 245

Provenance-aware optimization of workload for distributed data production 246

The Cloud Area Padovana: from pilot to production 247

Evolution of the Building Management System in the INFN CNAF Tier-1 datacenter facility 248

The LHCb software and computing upgrade for Run3: opportunities and challenges 250

An educational distributed Cosmic Ray detector network based on ArduSiPM using micro-controllers as data acquisition node NTP protocol as time distribution and IoT technology for data aggregation. 251

A Security Monitoring Framework For Container Based HEP Computing Infrastructures 252

MCBooster: a library for fast Monte Carlo generation of phase-space decays in massively parallel platforms. 253

Towards redundant services in dCache 255

Numerical Optimization for Fast Track Finding Based on the Artificial Retina Algorithm 256

SWAN: a Service for Web-Based Data Analysis in the Cloud 257

Expressing Parallelism in ROOT 258
Towards automation of data quality system for CERN CMS experiment 259
The New ROOT Interface: Jupyter Notebooks 260
Muon trigger for mobile phones 261
Flexible trigger menu implementation on the Global Trigger for the CMS Level-1 trigger upgrade 262
Developing and optimizing applications for the Hadoop environment 263
New operator assistance features in the CMS Run Control System 264
The ZEUS data preservation project 265
Experiences with ALICE Mesos Infrastructure 266
Online computing architecture for the CBM experiment at FAIR 267
An Job Accounting tool for IHEP Computing 268
Building a Regional Computing Grid for the University of California at 100gbps 269
Extreme I/O on HPC for HE using the Burst Buffer at NERSC 270
Review of Terabit/sec SDN demonstrations at Supercomputing 2015 and plans for SC16 271
Data Transfer Nodes and Demonstration of 100G - 400G Wide Area Throughput Using the Caltech SDN Testbed 272
Amplitude analysis of four-body decays using a massively-parallel fitting framework 273
Enabling research network connectivity to clouds with a virtual router 273
An OSG-based distributed campus computing infrastructure 276
The OSG Open Facility: An on-ramp for opportunistic scientific computing 277
A Software Toolkit to Study Systematic Uncertainties of the Physics Models of the Geant4 Simulation Package 278
Integration of Openstack Cloud Resources in BES III Computing Cluster 280
dCache on steroids - delegated storage solutions 281
AMS-02 Monte Carlo Simulation in Science Operation Center at Southeast University 282
Particle Physics Analysis with the CERN Open Data Portal 283
Design and Implementation of Monte Carlo Production Monitoring Tool for AMS Experiment 284
Computational Steering of GEM Simulations 285
STAR Reconstruction Improvements for Tracking with the Heavy Flavor Tracker 286
Representing Misalignments of the STAR Geometry Model using AgML 287
Design and Implementation of Elastic Computing Resource Management System with HTCondor on Openstack 288
OSiRIS: A Distributed Ceph Deployment Using Software Defined Networking for Multi-Institutional Research 289
A dashboard for the Italian Computing in ALICE 290
Scaling the PuNDIT Project for Wide Area Deployments 291
The Simulation Library of the Belle II Experiment 292
Using Shifter to Bring Containerized CVMFS to HPC 293
Tracks pattern recognition for the SHiP Spectrometer Tracker 294
GRID Storage Optimization in Transparent and User-Friendly Way for LHCb datasets 295
The LHCb Grid Simulation 296
Pythonization API for Cppyy 297
Scalable Global Grid catalogue for LHC Run3 and beyond 298
The CMS Data Acquisition - Architectures for the Phase-2 Upgrade 299
Extraction of the luminosity at PANDA 300
A Multipurpose Computing Center with Distributed Resources 301
Consolidating WLCG topology and configuration in the Computing Resource Information Catalogue 302
The Instant Glidein: A generic approach for the late-binding jobs to various resource types 303
IPv6 Security 304
LHCb data and software dependencies in the long-term future preservation 305
Software Quality Control at Belle II 306
High Level Interface to Conditions Data at Belle II 307
Status of the Calibration and Alignment Framework at the Belle II Experiment 308
Status of the ComPWA Framework 310
Acceleration of Cherenkov angle reconstruction with the new Intel Xeon/FPGA compute platform for the particle identification in the LHCb Upgrade 312
ALICE HLT Cluster operation during ALICE Run 2 313
Experiment Software and Projects on the Web with VISPA 315
Design and Execution of make-like distributed Analyses based on Spotify's Pipelining Package luigi 316
CBM First-level Event Selector Input Interface Development 317
Engaging Industry for Innovation in the LHC run3-4 R&D program: CERN openlab 318
Workflow Management for Complex HEP Analyses 320
New Machine Learning Developments in ROOT 321
The FCC software - how to keep SW experiment independent 322
Consolidation of Docker use in HTC and its evolution to HPC at INFN-Pisa 323
Towards more common build tools - experience with using spack in HEP 324
Low latency network and distributed storage for next generation HPC systems: the ExaNest project. 325
Vibration monitoring system for the RACF data center at BNL 326
Evolution of the Ceph Based Storage Systems at the RACF 327
Dynamo - The dynamic data management system for the distributed CMS computing system 328
Using 3D Engineering Models in a GEANT Simulation 329
A lightweight task submission and management infrastructure 330
Enabling volunteer computing for the BESIII Simulation system 331
Microservices for systematic profiling and monitoring of the refactoring process at the LHCb experiment 333
Validation of Physics Models of Geant4 using data from CMS experiment 334
LHCb trigger streams optimization 337
Detector Simulations with DD4hep 339
Storage Quality-of-Service in Cloud-based Scientific Environments: A Standardization Approach 340
Conversion of Belle’s data to Belle II’s data format 341
Conditions Database for the Belle II Experiment 342
Effective HTCondor-based monitoring system for CMS 343
Neuromorphic Computing for High Energy Physics 344
Highlights of the Belle II Computing 345
A comparison of different database technologies for the CMS AsyncStageOut transfer database 346
The Trigger and Data Acquisition System for the KM3NeT-Italy neutrino telescope 348
A programming framework for data streaming on the Xeon Phi 349
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benchmarking worker nodes using LHCb simulation productions and comparing with HEP: Spec06</td>
<td>350</td>
</tr>
<tr>
<td>ALICE HLT Run2 performance overview</td>
<td>351</td>
</tr>
<tr>
<td>Support for Online Calibration in the ALICE HLT Framework</td>
<td>352</td>
</tr>
<tr>
<td>Latest Developments in Delphes 3</td>
<td>353</td>
</tr>
<tr>
<td>Exploring private and commercial clouds for BESIII</td>
<td>354</td>
</tr>
<tr>
<td>Experience with Beignet OpenCL on Low Power Intel SoCs</td>
<td>356</td>
</tr>
<tr>
<td>Optimizing the resource usage in Cloud based environments: the Synergy approach</td>
<td>357</td>
</tr>
<tr>
<td>Gaudi Evolution for Future Challenges</td>
<td>358</td>
</tr>
<tr>
<td>Grid Site Availability Evaluation and Monitoring at CMS</td>
<td>359</td>
</tr>
<tr>
<td>“Big Data” in HEP: A comprehensive use case study</td>
<td>360</td>
</tr>
<tr>
<td>CMS Full Simulation Status</td>
<td>361</td>
</tr>
<tr>
<td>Design and Deployment of a Elastic Network Test-bed in IHEP Data Center based on SDN</td>
<td>362</td>
</tr>
<tr>
<td>CMS readiness for multi-core workload scheduling</td>
<td>363</td>
</tr>
<tr>
<td>Dynfarm: A Dynamic Site Extension</td>
<td>364</td>
</tr>
<tr>
<td>Application of StoRM+Lustre storage system in IHEP’s distributed computing</td>
<td>365</td>
</tr>
<tr>
<td>Accelerated tracking using GPUs at CMS High Level Trigger for Run 3</td>
<td>366</td>
</tr>
<tr>
<td>Improved Cloud resource allocation: how INDIGO-Datacloud is overcoming the current limitations in Cloud schedulers</td>
<td>367</td>
</tr>
<tr>
<td>Netfinder, a real-time network topology discovery tool</td>
<td>368</td>
</tr>
<tr>
<td>Exploiting analytics techniques in CMS computing monitoring</td>
<td>369</td>
</tr>
<tr>
<td>Experience on HTCondor batch system for HEP and other research fields at KISTI-GSDC</td>
<td>370</td>
</tr>
<tr>
<td>Experience in using commercial clouds in CMS</td>
<td>371</td>
</tr>
<tr>
<td>Benchmarking High Performance Computing Architectures with CMS’ Skeleton Framework</td>
<td>372</td>
</tr>
<tr>
<td>CDSLabs: Towards the Next Generation CERN Institutional Repository</td>
<td>373</td>
</tr>
<tr>
<td>A FairShare Scheduling Service for OpenNebula</td>
<td>374</td>
</tr>
<tr>
<td>A new VMDIRAC based approach for an Elastic Cloud</td>
<td>375</td>
</tr>
<tr>
<td>HNType : a diverse trace and migration mechanism in the block based Hierarchical Storage System named HazelNut</td>
<td>376</td>
</tr>
</tbody>
</table>
INDIGO-Datacloud: a Cloud-based Platform as a Service oriented to scientific computing for the exploitation of heterogeneous resources 377

Data Flow Simulation of the ALICE Computing System with OMNET++ 378

LHCb migration from Subversion to Git 379

LHCb Dockerized Build Environment 380

A New Data Access Mechanism for HDFS 381

GridPP Dirac: Supporting non-LHC VOs on LHC centric resources 383

Unified data access to e-Infrastructure, Cloud and personal storage within INDIGO-DataCloud 384

Stealth Cloud: How not to waste CPU during grid to cloud transitions 386

GPU-accelerated track reconstruction in the ALICE High Level Trigger 387

Improvements of the ALICE HLT data transport framework for LHC Run 2 389

The INDIGO-DataCloud Authentication and Authorisation Infrastructure 392

Accelerating Navigation in the VecGeom Geometry Modeller 393

Fast GPU Nearest Neighbors search algorithms for the CMS experiment at LHC 394

Realtime processing of LOFAR data for the detection of particles with the Moon 395

Software and Experience with Managing Workflows for the Computing Operation of the CMS Experiment 396

The HNSciCloud project 397

Resources monitoring and automatic management system for multi-VO distributed computing system 398

Technical challenges of HNSciCloud 399

On-demand provisioning of HEP compute resources on cloud sites and shared HPC centers 400

How Pre-Commercial Procurement can fund innovative cloud services 401

Migration Routes to Virtual Technologies 402

Stability and scalability of the CMS Global Pool: Pushing HTCondor and glideinWMS to new limits 403

Track Finding in CMS for the Level-1 Trigger at the HL-LHC 404

Integration of the SCD Cloud and Aquilon 406

Scheduling beams at CERN, the new AD central timing 407

Connecting restricted, high-availability, or low-latency resources to a seamless Global Pool for CMS 408
Storageless and caching Tier-2 models in the UK context 409

The high-rate data challenge: computing for the CBM experiment 410

Index files for Belle II - very small skim containers 411

Dynamic resource provisioning of the CMS online cluster using a cloud overlay 412

Machine Learning with TensorFlow as an alternative to TMVA 413

Elastic extension of a local analysis facility on external Clouds for the LHC experiments 414

A precision device needs precise simulation: Software description of the CBM Silicon Tracking System 415

The LHCb Starterkit 416

One network metric datastore to track them all: The OSG Network Service 417

Finding the needle in the haystack: a charmonium trigger for the CBM experiment 419

STAR Data Production at NERSC/Cori, an adaptable Docker container approach for HPC 420

Extending the farm on external sites: the INFN Tier-1 experience 421

Monitoring the LHCb data quality system 422

Achieving Cost/Performance Balance Ratio Using Tiered Storage Caching Techniques: A Case Study with CephFS 424

Development of stable Grid service at the next generation system of KEKCC 425

Opportunistic resource usage at Argonne Facility for CMS 426

Last development of the Long Term Data Preservation project for CDF at INFN CNAF 427

CMS Software Performance Measurements "in the wild" 428

Opportunistic data locality for end user data analysis 430

TOSCA-based orchestration of complex clusters at the IaaS level 431

CMS Connect 432

A lightweight access control solution for Openstack 434

Primary Vertex finding in the RHIC high precision measurement era - enhancement and experience in STAR 435

Parallel Monte Carlo search for Hough transform 436

Parallel metric trees for multi-billion body simulations 437

Application of state quantization-based methods in HEP particle transport simulation 438

CMS Event Processing Multi-core Efficiency Status 439
Python object-oriented framework for consuming, manipulating and releasing non-event data for the CMS alignment and calibration. 440

grid-control - The Swiss Army knife of job submission tools 441

Cloud Environment Automation: from infrastructure deployment to application monitoring 442

A multi-group and preemptive computing resource scheduling system based on HTCondor 444

Readout and trigger for AFP detector at ATLAS experiment 445

Data Resilience in dCache Storage System 446

Real-time analysis with the LHCb trigger in Run-II 447

Parallel Algorithms for Online Trackfinding at PANDA Using GPUs 448

A browser-based event display for the CMS Experiment at the LHC using WebGL 449

Integrated monitoring of the ATLAS online computing farm 450

A Web-based application for the collection, management and release of Alignment and Calibration configurations used in data processing at the Compact Muon Solenoid experiment 451

Real time analysis with the upgraded LHCb trigger in Run-III 452

OpenFabrics Interface and NanoMsg 453

Master of Puppets 454

Design and evaluation of Hybrid storage system in HEP Environment 455

Functional tests of a prototype for the CMS-ATLAS common non-event data handling framework 457

Implementation of the ATLAS trigger within the ATLAS MultiThreaded Software Framework AthenaMT 458

Monitoring of the infrastructure and services used to handle and automatically produce Alignment and Calibration conditions at CMS 459

Developing the WLCG Tier-1 center GridKa as topical center in a multidisciplinary research environment 461

Triggering on leptons and photons on ATLAS 462

MonALISA, An Agent-Based Monitoring and Control System for the LHC Experiments: Status and Outlook 463

The ATLAS Run 2 Trigger: Design, Menu, Performance and Operational Aspects 464

Development, Validation and Integration of the ATLAS Trigger System Software in Run 2 465
Invenio digital library technology for open research data repositories 466

The Resource Manager of the ATLAS Trigger and Data Acquisition system 467

The Payload Inspector: a tool for the visualization of Calibration and Alignment constants stored in the CMS condition database 468

Trigger Menu-aware Monitoring for the ATLAS experiment 469

Everware toolkit - supporting reproducible science and challenge driven education 470

Deployment of IPv6 only CPU resources at WLCG sites 471

Geographically distributed Batch System as a Service: the INDIGO-DataCloud approach exploiting HTCondor 472

DD4hep based event reconstruction 473

The Czech National Grid Infrastructure 474

Deployment of 464XLAT (RFC6877) alongside IPv6-only CPU resources at WLCG sites 475

ATLAS Trigger and Data Acquisition Upgrades for High Luminosity LHC 476

The design of a fast Level 1 track trigger for the ATLAS High Luminosity Upgrade 477

The new inter process communication middle-ware for the ATLAS Trigger and Data Acquisition system 478

System upgrade of the KEK central computing system 480

FELIX: the new detector readout system for the ATLAS experiment 481

Multi-Threaded Algorithms for General purpose Graphics Processor Units in the ATLAS High Level Trigger 482

The ATLAS Data Acquisition System LHC Run 2 483

Investigations of Future Computing Platforms for High Energy Physics 484

The ATLAS Level-1 Topological Trigger Performance in Run 2 485

A web-based solution to visualize operational monitoring data in the Trigger and Data Acquisition system of the ATLAS experiment at the LHC 486

Frameworks to monitor and predict resource usage in the ATLAS High Level Trigger 487

High-Throughput Network Communication with NetIO 488

The HEP Software and Computing Knowledge Base 490

Abstracting application deployment on Cloud infrastructures 491

Expanding the user base beyond HEP for the Ganga distributed analysis user interface 492

Efficient monitoring of CRAB3 jobs at CMS 493

Use of DAGMan in CRAB3 to improve the splitting of CMS user jobs 494
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monitoring of the data processing and simulated production at CMS with a web-based service: the Production Monitoring Platform (pMp)</td>
<td>336</td>
</tr>
<tr>
<td>Managing the CMS Data and Monte Carlo Processing during LHC Run 2</td>
<td>337</td>
</tr>
<tr>
<td>Configuration validation in the art event-processing framework</td>
<td>338</td>
</tr>
<tr>
<td>Using Docker in HEP</td>
<td>338</td>
</tr>
<tr>
<td>Research on an universal Openstack upgrade solution</td>
<td>339</td>
</tr>
<tr>
<td>The end-to-end network performance troubleshooting cookbook</td>
<td>340</td>
</tr>
<tr>
<td>Accessing Data Federations with CVMFS</td>
<td>341</td>
</tr>
<tr>
<td>Exploring Compression Techniques for ROOT IO</td>
<td>341</td>
</tr>
<tr>
<td>Contributing opportunistic resources to the grid with HTCondor-CE-Bosco</td>
<td>342</td>
</tr>
<tr>
<td>GRACC: New Generation of the OSG Accounting</td>
<td>343</td>
</tr>
<tr>
<td>Latest generation interconnect technologies in APEnet+ networking infrastructure</td>
<td>343</td>
</tr>
<tr>
<td>Evolution, design, management and support for the CMS Online computing cluster</td>
<td>345</td>
</tr>
<tr>
<td>HEP Track Finding with the Micron Automata Processor and Comparison with an FPGA-based Solution</td>
<td>346</td>
</tr>
<tr>
<td>IceProd 2 usage experience</td>
<td>347</td>
</tr>
<tr>
<td>Applications of modern Erasure Coded Big Data paradigms to WLCG data resilience</td>
<td>347</td>
</tr>
<tr>
<td>Pyglidein - a simple HTCondor glidein service</td>
<td>348</td>
</tr>
<tr>
<td>Plancton: an opportunistic distributed computing project based on Docker containers</td>
<td>349</td>
</tr>
<tr>
<td>Interconnected Private Clouds for Universities and Researchers</td>
<td>350</td>
</tr>
<tr>
<td>SciDAC-Data, A Project to Enabling Data Driven Modeling of Exascale Computing</td>
<td>351</td>
</tr>
<tr>
<td>GEMPix detector as beam monitor at CNAO Hadrontherapy Center: Geant4 simulation and measurements</td>
<td>352</td>
</tr>
<tr>
<td>Analyzing how we do Analysis and consume data, Results from the SciDAC-Data Project</td>
<td>353</td>
</tr>
<tr>
<td>Adjusting the fairshare policy to prevent computing power loss</td>
<td>354</td>
</tr>
<tr>
<td>Design of the ProtoDUNE experiment data management infrastructure</td>
<td>354</td>
</tr>
<tr>
<td>Alignment of the CMS Tracker: Latest Results from LHC Run-II</td>
<td>355</td>
</tr>
<tr>
<td>The Application of SNiPER to the JUNO Simulation</td>
<td>356</td>
</tr>
<tr>
<td>YARR - A PCIe based readout concept for current and future ATLAS Pixel Modules</td>
<td>357</td>
</tr>
<tr>
<td>Web Based Monitoring project at CMS experiment</td>
<td>357</td>
</tr>
<tr>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Developing the Traceability Model to meet the Requirements of an Evolving Distributed Computing Infrastructure</td>
<td>522</td>
</tr>
<tr>
<td>Offline Software for the CMS Level-1 Trigger</td>
<td>523</td>
</tr>
<tr>
<td>Evaluation of the pre-production implementation of the Intel Omni-Path interconnect technology</td>
<td>524</td>
</tr>
<tr>
<td>Challenges in Scaling NLO Generators to Leadership Computers</td>
<td>525</td>
</tr>
<tr>
<td>The HEP Cloud Facility: elastic computing for High Energy Physics</td>
<td>526</td>
</tr>
<tr>
<td>Approximating Likelihood Ratios with Calibrated Classifiers</td>
<td>527</td>
</tr>
<tr>
<td>dCache, managed Cloud Storage</td>
<td>528</td>
</tr>
<tr>
<td>Evaluation of the LHCb Online run bookkeeping and data consistency system</td>
<td>529</td>
</tr>
<tr>
<td>jade: An End-To-End Data Transfer and Catalog Tool</td>
<td>530</td>
</tr>
<tr>
<td>Using container orchestration to improve service management at the RAL Tier 1</td>
<td>531</td>
</tr>
<tr>
<td>Use of a hardware token for Grid authentication by the MICE data distribution framework</td>
<td>532</td>
</tr>
<tr>
<td>Evaluation of ZFS as an efficient WLCG storage backend</td>
<td>533</td>
</tr>
<tr>
<td>dCache, towards Federated Identities and Anonymized Delegation</td>
<td>534</td>
</tr>
<tr>
<td>A Tool to Convert CAD Models for Importation into Geant4</td>
<td>535</td>
</tr>
<tr>
<td>A build system for multiple package development utilizing Spack</td>
<td>536</td>
</tr>
<tr>
<td>NaNet: a Configurable Network Interface Card for Trigger and DAQ Systems</td>
<td>537</td>
</tr>
<tr>
<td>Mesos exploitation within INDIGO-DataCloud for multi-site long-running services, computational workloads and advanced PaaS features</td>
<td>538</td>
</tr>
<tr>
<td>LArSoft: a toolkit for simulation, reconstruction and analysis of liquid argon TPC experiments</td>
<td>539</td>
</tr>
<tr>
<td>Data Management and Database Framework for the MICE Experiment</td>
<td>540</td>
</tr>
<tr>
<td>Taking HEP data management outside of HEP</td>
<td>543</td>
</tr>
<tr>
<td>Virtual Machine Provisioning, Code Management and Data Movement Design for the Fermilab HEPCloud Facility</td>
<td>545</td>
</tr>
<tr>
<td>The LArIAT Experiment’s Data Acquisition and Trigger System</td>
<td>546</td>
</tr>
<tr>
<td>artdaq: DAQ Software Development Made Simple</td>
<td>548</td>
</tr>
<tr>
<td>Integrating Prediction, Provenance, and Optimization into High Energy Workflows</td>
<td>549</td>
</tr>
<tr>
<td>Continuous and fast calibration of the CMS experiment: design of the automated workflows and operational experience</td>
<td>550</td>
</tr>
<tr>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Advances in Grid Computing for the FabrIc for Frontier Experiments Project at Fermilab</td>
<td>379</td>
</tr>
<tr>
<td>Using computing models from particle physics to investigate dose-toxicity correlations in cancer radiotherapy</td>
<td>380</td>
</tr>
<tr>
<td>Optical follow-up of gravitational wave triggers with DECam</td>
<td>381</td>
</tr>
<tr>
<td>Deep-Learning Analysis Pipelines on Raw HEP Data from the Daya Bay Neutrino Experiment at NERSC</td>
<td>382</td>
</tr>
<tr>
<td>An artificial retina processor for track reconstruction at the full LHC crossing rate</td>
<td>383</td>
</tr>
<tr>
<td>The deployment of a large scale object store at the RAL Tier 1</td>
<td>384</td>
</tr>
<tr>
<td>Astronomy data delivery and processing services with CADC and CANFAR</td>
<td>385</td>
</tr>
<tr>
<td>Tracking in the Muon g-2 Experiment</td>
<td>385</td>
</tr>
<tr>
<td>Progress in Analysis Preservation with a Focus on Reinterpretation</td>
<td>386</td>
</tr>
<tr>
<td>Integrating HPC, Cloud, and Containers for Data-Intensive Computing Platforms</td>
<td>387</td>
</tr>
<tr>
<td>Integrating Visualization Applications, such as ParaView, into HEP Software Frameworks for In-situ Event Displays</td>
<td>387</td>
</tr>
<tr>
<td>Neutrino Identification With A Convolutional Neural Network in the NOvA Detectors</td>
<td>388</td>
</tr>
<tr>
<td>CMS use of allocation based HPC resources</td>
<td>389</td>
</tr>
<tr>
<td>Evolution of CMS Software</td>
<td>390</td>
</tr>
<tr>
<td>Simulating storage part of application with Simgrid</td>
<td>390</td>
</tr>
<tr>
<td>Machine learning and parallelism in the reconstruction of LHCb and its upgrade</td>
<td>391</td>
</tr>
<tr>
<td>Novel real-time alignment and calibration of the LHCb Detector in Run2</td>
<td>392</td>
</tr>
<tr>
<td>Large-scale distributed usage of NLO and multi-leg Monte Carlo event generators on the grid and at Argonne Leadership Computing Facility</td>
<td>392</td>
</tr>
<tr>
<td>Performance studies of GooFit on GPUs versus RooFit on CPUs while estimating the statistical significance of a new physical signal</td>
<td>393</td>
</tr>
<tr>
<td>HTTP as a Data Access Protocol: Trials with XrootD in CMS’ AAA Project</td>
<td>394</td>
</tr>
<tr>
<td>The CMS Data Analysis School experience</td>
<td>394</td>
</tr>
<tr>
<td>SWATCH: Common software for controlling and monitoring the upgraded CMS Level-1 trigger</td>
<td>395</td>
</tr>
<tr>
<td>Efficient Access to Massive Amounts of Tape-Resident Data</td>
<td>396</td>
</tr>
<tr>
<td>Fast emulation of track reconstruction in CMS simulation</td>
<td>397</td>
</tr>
<tr>
<td>Divergence techniques in high energy physics</td>
<td>397</td>
</tr>
<tr>
<td>Impact of tracker layout on track reconstruction in pp collisions with high pileup at CMS.</td>
<td>378 - 398</td>
</tr>
<tr>
<td>LStore: Logistical Storage</td>
<td>399</td>
</tr>
<tr>
<td>A SDN based approach for the ATLAS data acquisition network</td>
<td>399</td>
</tr>
<tr>
<td>Identification of High-Momentum Top Quarks, Higgs Bosons, and W and Z Bosons Using Boosted Event Shapes</td>
<td>400</td>
</tr>
<tr>
<td>Experiments Toward a Modern Analysis Environment: Using TMVA and other tools in a functional world with continuous integration for analysis</td>
<td>401</td>
</tr>
<tr>
<td>The Detector Final State pattern: Using the Web Ontology Language to describe a Physics Analysis</td>
<td>402</td>
</tr>
<tr>
<td>The Performance of AMS Offline Software on IBM Blue Gene/Q Architecture</td>
<td>402</td>
</tr>
<tr>
<td>EMMA: A New Paradigm in Configurable Software</td>
<td>403</td>
</tr>
<tr>
<td>Welcome</td>
<td>404</td>
</tr>
<tr>
<td>DOE HEP Viewpoint</td>
<td>404</td>
</tr>
<tr>
<td>Future of Compute</td>
<td>404</td>
</tr>
<tr>
<td>The US Exascale Project</td>
<td>404</td>
</tr>
<tr>
<td>Diversity and Bias: Creating Inclusive Workplaces, See Bias, Block Bias</td>
<td>404</td>
</tr>
<tr>
<td>Diversity and Bias: Panel discussion</td>
<td>404</td>
</tr>
<tr>
<td>TDAQ - Tracking in the Trigger, FPGAs and other approaches</td>
<td>405</td>
</tr>
<tr>
<td>CHEP 2018, Sofia</td>
<td>405</td>
</tr>
<tr>
<td>Closing</td>
<td>405</td>
</tr>
<tr>
<td>Machine Learning in Science and Beyond</td>
<td>405</td>
</tr>
<tr>
<td>Opportunities to Collaborate in Data Science + Data Engineering</td>
<td>405</td>
</tr>
<tr>
<td>Storage - Sony Everspan</td>
<td>405</td>
</tr>
<tr>
<td>Storage - The Future</td>
<td>406</td>
</tr>
<tr>
<td>Non-HEP: Data Management for LSST</td>
<td>406</td>
</tr>
<tr>
<td>Non-HEP: Computing Challenges at Light Sources</td>
<td>406</td>
</tr>
<tr>
<td>test</td>
<td>406</td>
</tr>
<tr>
<td>Computing at the Electron-Ion Collider</td>
<td>406</td>
</tr>
<tr>
<td>Networking - View from ESnet</td>
<td>407</td>
</tr>
<tr>
<td>Networking - View from HEP</td>
<td>407</td>
</tr>
</tbody>
</table>
Simulation - Key Issues for the Coming Decade 613
Future of Distributed Data and Workload Management 614
Software - HEP Software Foundation Community White Paper Status 615
Software Challenges and Commonalities - Panel 616
Dell 617
Clouds and Competitors 618
Computing for HENP Experiments and Programs - Blurring Online and Offline 619
Computing for HENP Experiments and Programs - Data Streaming Trends 620
Quantum Computing 621
Supermicro 622
Track 1: Online Computing - Highlights 623
Track 2: Offline Computing - Highlights 624
Track 3: Distributed Computing - Highlights 625
Track 4: Data Handling - Highlights 626
Track 5: Software Development - Highlights 627
Track 6: Infrastructures - Highlights 628
Track 7: Middleware, Monitoring and Accounting - Highlights 629
Track 8: Security, Policy and Outreach - Highlights 630
Track 9: Future Directions - Highlights 631
How far can COTS HPC go? 632
Poster Competition Winner 633
Experiment Management System for the SND Detector

Author: Konstantin Pugachev

Co-author: Aleksandr Korol

1 Budker Institute of Nuclear Physics (RU)

Corresponding Authors: sekrasoft@gmail.com, aleksandr.korol@cern.ch

We present a new experiment management system for the SND detector at the VEPP-2000 collider (Novosibirsk). Substantially, it includes as important part operator access to experimental databases (configuration, conditions and metadata).

The system is designed in client-server architecture. A user interacts with it via web-interface. The server side includes several logical layers: user interface templates, template variables description and initialization, implementation details like database interaction. The templates are believed to have a simple enough structure to be used not only IT professionals but also by physicists.

Experiment configuration, conditions and metadata are stored in a database being managed by DBMS MySQL, ones being composed as records having hierarchical structure. To implement the server side NodeJS, a modern JavaScript framework, has been chosen. A new template engine is designed. The important feature of our engine is asynchronous computations hiding. The engine provides heterogeneous synchronous-style expressions (including synchronous or asynchronous values or functions calls). This helps template creators to focus on values to get but not on callbacks to handle.

A part of the system is put into production. It includes templates dealing with showing and editing first level trigger configuration and equipment configuration and also showing experiment metadata and experiment conditions data index.

Secondary Keyword (Optional):
Monitoring

Primary Keyword (Mandatory):
Databases

Tertiary Keyword (Optional):

Track 2: Offline Computing / 2

Reconstruction software of the silicon tracker of DAMPE mission

Author: Andrii Tykhonov

Co-author: Xin Wu

1 Université de Genève (CH)

Corresponding Authors: andrii.tykhonov@cern.ch, xin.wu@cern.ch

DAMPE is a powerful space telescope launched in December 2015, able to detect electrons and photons in a wide range of energy (5 GeV to 10 TeV) and with unprecedented energy resolution. Silicon tracker is a crucial component of detector, able to determine the direction of detected particles and trace the origin of incoming gamma rays. This contribution covers the reconstruction software of the tracker, comprising the geometry convertor, track reconstruction and detector alignment algorithms. The convertor is an in-house, standalone system that converts the CAD drawings of the
detector and implements the detector geometry in the GDML (Geometry Description Markup Language) format. Next, the particle track finding algorithm is described. Since the DAMPE tracker identifies independently the particle trajectory in two orthogonal projections, there is an inherent ambiguity in combining the two measurements. Therefore, the 3D track reconstruction becomes a computationally intensive task and the number of possible combinations increases quadratically with the number of particle tracks. To alleviate the problem, a special technique is developed, which reconstructs track fragments independently in two projections and combine the final result using a 3D Kalman fit of pre-selected points. Finally, the detector alignment algorithm allows to align the detector geometry based on real data with precision better than the resolution of tracker. The algorithm optimises a set of around four thousand parameters (offsets and rotations of detecting elements) in an iterative procedure, based on the minimisation of the global likelihood fit of reconstructed tracks. Since the algorithm is agnostic of the detector premises, it could be used for similar optimisation problems with minor modifications by other projects. This contribution will give an insight into the developed algorithms and the results obtained during the first years of operational experience on ground and on orbit.

Tertiary Keyword (Optional):
Collaborative tools

Secondary Keyword (Optional):
Algorithms

Primary Keyword (Mandatory):
Reconstruction

Track 8: Security, Policy and Outreach / 3

HEPData - a repository for high energy physics data exploration

Author: Eamonn James Maguire¹
Co-authors: Frank Martin Krauss ²; Graeme Watt ³; Jan Andrzej Stypka ⁴; Lukas Alexander Heinrich ⁵; Michael Whalley ⁶; Salvatore Mele ¹

¹ CERN
² University of Durham (GB)
³ Durham University
⁴ AGH University of Science and Technology (PL)
⁵ New York University (US)

Corresponding Authors: m.r.whalley@durham.ac.uk, graeme.watt@durham.ac.uk, jan.stypka@cern.ch, salvatore.mele@cern.ch, eamonn.james.maguire@cern.ch, lukas.heinrich@cern.ch

The Durham High Energy Physics Database (HEPData) has been built up over the past four decades as a unique open-access repository for scattering data from experimental particle physics. It is comprised of data points from plots and tables underlying over eight thousand publications, some of which are from the Large Hadron Collider (LHC) at CERN.

HEPData has been rewritten from the ground up in the Python programming language and is now based on the Invenio 3 framework. The software is open source with the current site available at http://hepdata.net with: 1) a more stream-lined submission system; 2) advanced submission reviewing functionalities; 3) powerful full repository search; 4) an interactive data plotting library; 5) an attractive, easy to use interface; and 6) a new data driven visual exploration tool.

Here we will report on our efforts to bring findable, accessible, interoperable, and reusable (FAIR) principles to high energy physics.
Our presentation will cover the background of HEPData, limitations of the current tool, and why we created the new system using Invenio 3. We will present our system by considering four important aspects of the work: 1) the submission process; 2) making the data discoverable; 3) making data first class citable objects; and 4) making data interoperable and reusable.

Tertiary Keyword (Optional):
Visualization

Primary Keyword (Mandatory):
Databases

Secondary Keyword (Optional):
Preservation of analysis and data

Track 1: Online Computing / 4

Reconstruction of Micropattern Detector Signals using Convolutional Neural Networks

Author: Lucie Flekova ¹
Co-authors: Andreas Christian Dudder ²; Matthias Schott ²

¹ Technical University of Darmstadt
² Johannes-Gutenberg-Universitaet Mainz (DE)

Corresponding Authors: andreas.christian.dudder@cern.ch, flekova@ukp.informatik.tu-darmstadt.de, matthias.schott@cern.ch

Micropattern gaseous detector (MPGD) technologies, such as GEMs or MicroMegas, are particularly suitable for precision tracking and triggering in high rate environments. Given their relatively low production costs, MPGDs are an exemplary candidate for the next generation of particle detectors. Having acknowledged these advantages, both the ATLAS and CMS collaborations at the LHC are exploiting these new technologies for their detector upgrade programs in the coming years. When MPGDs are utilized for triggering purposes, the measured signals need to be precisely reconstructed within less than 200 ns, which can be achieved by the usage of FPGAs.

In this work, we present for a novel approach to identify reconstructed signals, their timing and the corresponding spatial position on the detector. In particular, we study the effect of noise and dead readout strips on the reconstruction performance. Our approach leverages the potential of convolutional neural networks (CNNs), which have been recently manifesting an outstanding performance in a range of modeling tasks. The proposed neural network architecture of our CNN is designed simply enough, so that it can be modeled directly by an FPGA and thus provide precise information on reconstructed signals already in trigger level.

Primary Keyword (Mandatory):
Artificial intelligence/Machine learning

Secondary Keyword (Optional):
Reconstruction

Tertiary Keyword (Optional):

Posters B / Break / 6
Federated data storage system prototype for LHC experiments and data intensive science

Author: Alexei Klimentov

Co-authors: Andrey Kirianov; Andrey Zarochentsev; Dimitrii Krasnopevtsev; Eygene Ryabinkin

1 Brookhaven National Laboratory (US)
2 B.P. Konstantinov Petersburg Nuclear Physics Institute - PNPI (RU)
3 St. Petersburg State University (RU)
4 National Research Nuclear University MEPhI (RU)
5 National Research Centre Kurchatov Institute (RU)

Corresponding Authors: alexei.klimentov@cern.ch, dimitriy.krasnopevtsev@cern.ch, rea@grid.kiae.ru, andrey.zarochencev@cern.ch, andrey.kiryanov@cern.ch

Rapid increase of data volume from the experiments running at the Large Hadron Collider (LHC) prompted national physics groups to evaluate new data handling and processing solutions. Russian grid sites and universities’ clusters scattered over a large area aim at the task of uniting their resources for future productive work, at the same time giving an opportunity to support large physics collaborations.

In our project we address the fundamental problem of designing a computing architecture to integrate distributed storage resources for LHC experiments and other data-intensive science applications and to provide access to data from heterogeneous computing facilities. Studies include development and implementation of federated data storage prototype for Worldwide LHC Computing Grid (WLCG) centers of different levels and University clusters within one National Cloud. The prototype is based on computing resources located in Moscow, Dubna, St.-Petersburg, Gatchina and Geneva.

This project intends to implement a federated distributed storage for all kind of operations such as read/write/transfer and access via WAN from Grid centers, university clusters, supercomputers, academic and commercial clouds. The efficiency and performance of the system are demonstrated using synthetic and experiment-specific tests including real data processing and analysis workflows from ATLAS and ALICE experiments, as well as compute-intensive bioinformatics applications (PALEOMIX) running on supercomputer.

We present topology and architecture of the designed system, report performance and statistics for different access patterns and show how federated data storage can be used efficiently by physicists and biologists. We also describe how sharing data on a widely distributed storage system can lead to a new computing model and reformations of computing style, for instance how bioinformatics program running on supercomputer can read/write data from federated storage.

Primary Keyword (Mandatory):
Distributed data handling

Secondary Keyword (Optional):
Storage systems

Tertiary Keyword (Optional):

Posters A / Break / 7

BelleII@home: Integrate volunteer computing resources into DIRAC in a secure way

Author: Wenjing Wu

Co-authors: Hideki Miyake; I Ueda; Phillip Urquijo; Takanori Hara; Wenyao Kan

1 Computer Center, IHEP, CAS
2 KEK
The exploitation of volunteer computing resources has become a popular practice in the HEP computing community as the huge amount of potential computing power it provides. In the recent HEP experiments, the grid middleware has been used to organize the services and the resources, however it relies heavily on the X.509 authentication, which is contradictory to the untrusted feature of volunteer computing resources, therefore one big challenge to utilize the volunteer computing resources is how to integrate them into the grid middleware in a secure way. The DIRAC interware which is commonly used as the major component of the grid computing infrastructure for several HEP experiments proposes an even bigger challenge to this paradox as its pilot is more closely coupled with operations requiring the X.509 authentication compared to the implementations of pilot in its peer grid interware. The Belle II experiment is a B-factory experiment at KEK, and it uses DIRAC for its distributed computing. In the project of BelleII@home, in order to integrate the volunteer computing resources into the Belle II distributed computing platform in a secure way, we adopted a new approach which detaches the payload running from the Belle II DIRAC pilot which is a customized pilot pulling and processing jobs from the Belle II distributed computing platform, so that the payload can run on volunteer computers without requiring any X.509 authentication. In this approach we developed a gateway service running on a trusted server which handles all the operations requiring the X.509 authentication. So far, we have developed and deployed the prototype of BelleII@home, and tested its full workflow which proves the feasibility of this approach. This approach can also be applied on HPC systems whose work nodes do not have outbound connectivity to interact with the DIRAC system in general.

Reconstruction and calibration of MRPC endcap TOF of BESIII

Author: Shengsen Sun

Institute of High Energy Physics, Chinese Academy of Sciences

Corresponding Author: sunss77@hotmail.com

The endcap time of flight(TOF) detector of the BESIII experiment at the BEPCII was upgraded based on multigap resistive plate chamber technology. During 2015-2016 data taking the TOF system has achieved a total time resolution of 65ps for electrons in Bhabha events. Details of reconstruction system and calibration procedures, detector alignment and performance with data will be described.
Reconstruction

Secondary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Track 5: Software Development / 9

RootJS: Node.js Bindings for ROOT 6

Authors: Christoph Haas¹; Christoph Wolff³; Jonas Schwabe¹; Marek Szuba¹; Maximilian Früh¹; Sachin Rajgopal¹; Theo Belfart¹

¹ KIT - Karlsruhe Institute of Technology (DE)
² KIT - Karlsruhe Institute of Technology (DE)

Corresponding Authors: marek.szuba@cern.ch, uwdyu@student.kit.edu, utdtq@student.kit.edu, tbeffart@gmail.com, maxi.frueh@gmail.com, uieis@student.kit.edu, usdwj@student.kit.edu

We present rootJS, an interface making it possible to seamlessly integrate ROOT 6 into applications written for Node.js, the JavaScript runtime platform increasingly commonly used to create high-performance Web applications. ROOT features can be called both directly from Node.js code and by JIT-compiling C++ macros. All rootJS methods are invoked asynchronously and support callback functions, allowing non-blocking operation of Node.js applications using them. Last but not least, our bindings have been designed to platform-independent and should therefore work on all systems supporting both ROOT 6 and Node.js.

Thanks to rootJS it is now possible to create ROOT-aware Web applications taking full advantage of the high performance and extensive capabilities of Node.js. Examples include platforms for the quality assurance of acquired, reconstructed or simulated data, book-keeping and e-log systems, and even Web browser-based data visualisation and analysis.

Tertiary Keyword (Optional):
Network systems and solutions

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

Secondary Keyword (Optional):
Software development process and tools

Posters A / Break / 10

C++ Software Quality in the ATLAS experiment: Tools and Experience

Author: Shaun Roe¹

Co-authors: Emel Obreshkov ²; Graeme Stewart ³; Peter Sherwood ¹; Rolf Seuster ⁵; Scott Snyder ⁶; Stefan Kluth ⁷; Stewart Martin-Haugh ⁸

¹ CERN
² University of Texas at Arlington (US)
In this paper we explain how the C++ code quality is managed in ATLAS using a range of tools from compile-time through to run time testing and reflect on the substantial progress made in the last two years largely through the use of static analysis tools such as Coverity®, an industry-standard tool which enables quality comparison with general open source C++ code. Other available code analysis tools are also discussed, as is the role of unit testing with an example of how the googlemock framework can be applied to our codebase.

Primary Keyword (Mandatory):
Software development process and tools

Tertiary Keyword (Optional): 

Secondary Keyword (Optional): 

Posters A / Break / 11

An automated meta-monitoring mobile application and frontend interface for the WLCG computing model

Author: Gen Kawamura

Co-authors: Arnulf Quadt; Gerhard Ferdinand Rzehorz

Efficient administration of computing centres requires advanced tools for the monitoring and frontend interface of their infrastructure. The large-scale distributed grid systems, like the Worldwide LHC Computing Grid (WLCG) and ATLAS computing, offer many existing web pages and information sources indicating the status of the services, systems, requests and user jobs at grid sites. These monitoring tasks are crucial especially in the management of each WLCG site and federated systems across 130 worldwide sites, and accesses from above 1,000 active users. A meta-monitoring mobile application which automatically collects the information from such monitoring floods could give every administrator a sophisticated and flexible interface for the production-level infrastructure. We offer such a solution; the MadFace mobile application. It is a HappyFace compatible mobile application which has a user-friendly interface. MadFace is an evolution of a model of the HappyFace meta-monitoring system which has demonstrated monitoring several WLCG sites. We present the key concepts of MadFace, including its browser crawler, image processor, Bayesian analyst, mobile viewer and a model of how to manage a complex infrastructure like the grid. For the design and technology of MadFace it employs specifically many recent frameworks such as Ionic, Cordova, AngularJS, Node.js, Jetpack Manager and various Bayesian analysis packages. We also show an actual use of the prototype application. MadFace becomes very feasible in a meta-monitoring platform which automatically investigates the states and troubles from different sources and provides access of the administration roles for non-experts.

Tertiary Keyword (Optional):
Computing models
Experience of Google’s latest Deep Learning library, TensorFlow, with Docker in a WLCG cluster

Author: Gen Kawamura
Co-authors: Arnulf Quadt; Gerhard Ferdinand Rzehorz; Joshua Wyatt Smith

1 Georg-August-Universitaet Goettingen (DE)

Corresponding Authors: joshua.wyatt.smith@cern.ch, arnulf.quadt@cern.ch, gen.kawamura@cern.ch, gerhard.ferdinand.rzehorz@cern.ch

The researchers at the Google Brain team released their second generation Deep Learning library, TensorFlow, as an open-source package under the Apache 2.0 license in November, 2015. Google has already deployed the first generation library using DistBelief in various systems such as Google Search, advertising systems, speech recognition systems, Google Images, Google Maps, Street View, Google Translate and many other recent products. In addition, many researchers in high energy physics have recently started to understand and use Deep Learning algorithms in their own research and analysis. We conceive a first use-case scenario of the TensorFlow library to create the Deep Learning models from high-dimensional inputs like physics analysis data and such environments in a large-scale WLCG computing cluster. TensorFlow carries out computations using a dataflow model and graph structure onto a wide variety of different hardware platforms and systems, such as many CPU architectures, GPUs and smartphone platforms. Having a single library that can distribute the computations to create a model to the various platforms and systems would significantly simplify the use of Deep Learning algorithms in high energy physics. Docker presents a solution in which we can merge the application libraries and the Linux kernel into a production-level WLCG computing cluster. We therefore employ the Docker container environments for TensorFlow and present the first use in our grid system.

Flexible online monitoring for high-energy physics with Pyrame

Author: Miguel Rubio-Roy
Co-authors: Floris Thiant; Frederic Bruno Magniette
Data quality monitoring (DQM) in high-energy physics (HEP) experiments is essential and widely implemented in most large experiments. It provides important real-time information during the commissioning and production phases that allows the early identification of potential issues and eases their resolution.

Existing and performant solutions for online monitoring exist for large experiments such as CMS [1], ALICE [2] and others. For small setups, proprietary solutions such as LabView provide tools to implement it but fail on scaling to high-throughput experiments.

The present work reports on the new online monitoring capabilities of the software Pyrame [3], an open-source framework designed for HEP applications. Pyrame provides an easy-to-deploy solution for command, control and data-acquisition of particle detectors and related test-benches.

Pyrame’s new online monitoring architecture is based on the distribution of the data treatment operations among any module of the system, with multiple input and output (I/O) streams. A common problem in such systems is that different treatment speeds can lead to uncontrolled data loss. To avoid such situation we implement a mechanism that breaks the real-time constraint, at any treatment level, by buffering data and distributing it at the speed of the consumers (potentially leading to subsampling in the worst case). In addition to the distributed data treatment capabilities, Pyrame includes a performance-oriented module dedicated to real-time data acquisition, capable of handling and storing data at 4Gb/s for further treatment.

This approach allows us to use the same code for online and offline data treatment. In the latter case, subsampling is simply forbidden. Legacy ROOT, R, or Python/Panda offline data treatment codes can be adapted to be fed by a file in offline mode or by one of the available online data sources.

The distribution of the data treatment chain over any module allows us to use a modular design. Data filtering operations can be processed first, providing cut or tagged data with a first level of grouping (block level). The resulting data can then be used in concentrating modules, such as multiplexers (e.g.: from multiple detection layers). Finally, the output can be used in higher level treatment processes like online event-builders/reconstruction modules.

References:


**Primary Keyword (Mandatory):**
Data processing workflows and frameworks/pipelines

**Secondary Keyword (Optional):**
DAQ

**Tertiary Keyword (Optional):**

---

**Posters B / Break / 14**

**Simulation of orientational coherent effects via Geant4**

**Author:** Enrico Bagli\(^1\)
Beam manipulation of high- and very-high-energy particle beams is a hot topic in accelerator physics. Coherent effects of ultra-relativistic particles in bent crystals allow the steering of particle trajectories thanks to the strong electrical field generated between atomic planes. Recently, a collimation experiment with bent crystals was carried out at the CERN-LHC [1], paving the way to the usage of such technology in current and future accelerators. Geant4 [2] is a widely used object-oriented tool-kit for the Monte Carlo simulation of the interaction of particles with matter in high-energy physics. Moreover, its areas of application include also nuclear and accelerator physics, as well as studies in medical and space science. We present the Geant4 model for the simulation of orientational effects in straight and bent crystals for high energy charged particles [3]. The model allows the manipulation of particle trajectories by means of straight and bent crystals and the scaling of the cross sections of hadronic and electromagnetic processes for channeled particles. Based on such a model, the extension of the Geant4 toolkit has been developed. The code and the model have been validated by comparison with published experimental data regarding the deflection efficiency via channeling and the variation of the rate of inelastic nuclear interactions.


Primary Keyword (Mandatory):
Simulation

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Posters B / Break / 15

Detector control system for the AFP detector in ATLAS experiment at CERN

Author: Elzbieta Banas

Co-authors: Davide Caforio 2; Jolanta Barbara Olszowska 1; Luis Filipe Oleiro Seabra 3; Petr Sicho 4; Sabina Czekierda 1; Zbyszek Hajduk 1

1 Polish Academy of Sciences (PL)
2 Czech Technical University (CZ)
3 LIP Laboratorio de Instrumentacao e Fisica Experimental de Part
4 Acad. of Sciences of the Czech Rep. (CZ)
Corresponding Authors: petr.sicho@cern.ch, sabina.czekierda@cern.ch, zbigniew.hajduk@cern.ch, jolanta.olszowska@cern.ch, elzbieta.banas@cern.ch, davide.caforio@cern.ch, luis.seabra@cern.ch

The ATLAS Forward Proton (AFP) detector upgrade project consists of two forward detectors located at 205 m and 217 m on each side of the ATLAS experiment. The aim is to measure momenta and angles of diffractively scattered protons. In 2016 two detector stations on one side of the ATLAS interaction point have been installed and are being commissioned.

The detector infrastructure and necessary services were installed and are being supervised by the Detector Control System (DCS), which is responsible for the coherent and safe operation of the detector.

Based on radiation conditions in the tunnel, easy access and maintainability during the collider activity, it was decided to locate only the necessary hardware close to detector stations. The second stage of the low voltage powering system, based on the radiation hard voltage regulators giving accurate voltage levels individually for each sensor, and optical converter module are placed at 212m, in between of both stations, while the vortex coolers are close to each station.

All other equipment is located in the ATLAS underground counting room (USA15), at about 330 m distance to the stations.

A large variety of used equipment represents a considerable challenge for the AFP DCS design. Industrial Supervisory Control and Data Acquisition (SCADA) product Siemens WinCCOA together with the CERN Joint Control Project (JCOP) framework and standard industrial and custom developed server applications and protocols are used for reading, processing, monitoring and archiving of detector parameters. Graphical user interfaces allow for overall detector operation and visualization of the detector status. Parameters, important for detector safety, are used for alert generation and interlock mechanisms.

The actual status of the AFP DCS and first experience gained during commissioning and tests of the detector is described in this contribution.

Tertiary Keyword (Optional):
Visualization

Primary Keyword (Mandatory):
Control systems

Secondary Keyword (Optional):
Monitoring

Track 1: Online Computing / 16

The InfiniBand based Event Builder implementation for the LHCb upgrade

Authors: Antonio Falabella¹; Balazs Voneki²; Francesco Giacomini³; Matteo Manzali¹; Niko Neufeld²; Sebastien Valat²; Umberto Marconi⁵

¹ Universita e INFN, Bologna (IT)
² CERN
³ INFN CNAF
⁴ Universita di Ferrara & INFN (IT)
⁵ Universita e INFN (IT)

Corresponding Authors: niko.neufeld@cern.ch, francesco.giacomini@cern.ch, balazs.voneki@cern.ch, sebastien.jean.valat@cern.ch, umberto.marconi@bo.infn.it, antonio.falabella@cnaf.infn.it, matteo.manzali@cern.ch

The LHCb experiment will undergo a major upgrade during the second long shutdown (2018 - 2019). The upgrade will concern both the detector and the Data Acquisition (DAQ) system, to be rebuilt in order to optimally exploit the foreseen higher event rate. The Event Builder (EB) is the key component of the DAQ system which gathers data from the sub-detectors and build up the whole event.
The EB network has to manage an incoming data flux of 32 Tbps running at 40 MHz, with a cardinality of about 500 nodes. In this contribution we present the EB implementation based on the InfiniBand (IB) network technology. The EB software relies on IB verbs, which offer user space API to employ the Remote Direct Memory Access (RDMA) capabilities provided by IB the network devices. We will present the performance of the EB on different High Performance Computing (HPC) clusters.

**Primary Keyword (Mandatory):**
DAQ

**Secondary Keyword (Optional):**

**Tertiary Keyword (Optional):**

---

**Posters B / Break / 17**

**JavaScript ROOT v4**

**Authors:** Bertrand Bellenot¹; Sergey Linev²

1 *CERN*
2 *GSI DARMSTADT*

**Corresponding Authors:** s.linev@gsi.de, bertrand.bellenot@cern.ch

JavaScript ROOT (JSROOT) aims to provide ROOT-like graphics in web browsers. JSROOT supports reading of binary and JSON ROOT files, and drawing of ROOT classes like histograms (TH1/TH2/TH3), graphs (TGraph), functions (TF1) and many others. JSROOT implements a user interface for THttpServer-based applications.

With the version 4 of JSROOT, many improvements and new features are provided:

- the binary file I/O performance has been improved by a factor of 5;
- the drawing performance for one and two dimensional histograms has been improved by a factor of 10;
- new tooltips providing better and precise information about displayed objects;
- drawing of TGeo-based geometries has been implemented;
- many new classes and new drawing options are supported;
- new generic object inspector, for any object stored in ROOT files.

JSROOT provide an intuitive interface for browsing ROOT files and displaying objects within different layouts like grids or tabs. At the same time flexible and simple JSROOT API can be used to construct custom HTML pages and display any supported ROOT classes inside them.

JSROOT, with documentation and examples, can be found on https://root.cern.ch/js/ website. The developer repository is https://github.com/linev/jsroot/. JSROOT can also be obtained via the bower package manager and easily integrated into Node.js based applications.

**Primary Keyword (Mandatory):**
Visualization

**Secondary Keyword (Optional):**
Monitoring

**Tertiary Keyword (Optional):**
The evolution of monitoring system: the INFN-CNAF case study

Authors: DIEGO MICHELOTTO¹; Stefano Bovina¹

¹ INFN - National Institute for Nuclear Physics

Corresponding Authors: stefano.bovina@cnaf.infn.it, diego.michelotto@cnaf.infn.it

Over the past two years, the operations at INFN-CNAF have undergone significant changes. The adoption of configuration management tools, such as Puppet and the constant increase of dynamic and cloud infrastructures, have led us to investigate a new monitoring approach. Our aim is the centralization of the monitoring service at CNAF through a scalable and highly configurable monitoring infrastructure.

The selection of tools has been made taking into account the following requirements given by our users: adaptability to dynamic infrastructures, ease of configuration and maintenance, capability to provide more flexibility, compatibility with existing monitoring system, re-usability and ease of access to information and data.

We are going to describe our monitoring infrastructure composed of the following components: Sensu as monitoring router, InfluxDB as time series database to store data gathered from sensors and Grafana as a tool to create dashboards and to visualize time series metrics.

Primary Keyword (Mandatory): Monitoring

Secondary Keyword (Optional): Cloud technologies

Tertiary Keyword (Optional):

Statistical and Data Analysis Package in SWIFT

Author: Claude Andre Pruneau¹

¹ Wayne State University (US)

Corresponding Author: claude.andre.pruneau@cern.ch

SWIFT is a compiled object-oriented language similar in spirit to C++ but with the coding simplicity of a scripting language. Built with the LLVM compiler framework used within Xcode 6 and later versions, SWIFT features interoperability with C, Objective-C, and C++ code, truly comprehensive debugging and documentation features, and a host of language features that make for rapid and effective development of robust and easily debuggable software and APPs. As of version 2.2, SWIFT is now open source and made available under APACHE license 2.0 for development on APPLE platforms and LINUX.

I present the design and implementation of a math and statistics package which features single and multi-dimensional functions, including most common functions used in math such as Bessel functions, Laguerre and Legendre polynomials, etc. The package also features classes for vectors, matrices and related linear algebra, a limited set of physics tools including rotations, Lorentz vectors, etc, multi-dimensional histograms, fast and robust moments calculation, calculation of correlation functions, statistical tests, maximum likelihood and least square fits, and extensible random number generation tools, as well as basic plotting capabilities. The code is developed based on a relatively...
small number of classes implementing a basic set of protocols. Given SWIFT’s interoperability with other languages, the presented package should be easy to integrate within existing computing environments, such as ROOT. I developed the presented package in two months during the summer 2015.

**Primary Keyword (Mandatory):**
Analysis tools and techniques

**Secondary Keyword (Optional):**
Algorithms

**Tertiary Keyword (Optional):**

Posters A / Break / 20

**Analysis Tools in Geant4 10.2**

**Authors:** Guy Barrand\(^1\); Ivana Hrivnacova\(^2\)

\(^1\) Laboratoire de l’Accélérateur Linéaire (LAL), Université Paris-Sud, CNRS-IN2P3, France
\(^2\) Institut de Physique Nucléaire (IPNO), Université Paris-Sud, CNRS-IN2P3, France

**Corresponding Authors:** barrand@lal.in2p3.fr, ivana@ipno.in2p3.fr

A new analysis category based on g4tools was added in Geant4 release 9.5 with the aim of providing users with a lightweight analysis tool available as part of the Geant4 installation without the need to link to an external analysis package. It has progressively replaced the usage of external tools based on AIDA (Abstract Interfaces for Data Analysis) in all Geant4 examples. Frequent questions in the Geant4 users forum show its increasing popularity in the Geant4 users community.

The analysis category consists of the analysis manager classes and the g4tools package.

g4tools, originally part of the inlib and exlib packages, provides a very light and easy to install set of C++ classes that can be used to perform analysis in a Geant4 batch program. It allows to create and manipulate histograms, profiles and ntuples, write them in several supported file formats (ROOT, CSV, AIDA XML, and HBOOK) and, when needed, also read them back from the files. Since the last Geant4 release, 10.2, it has been enhanced with functions for batch plotting and MPI messaging.

Analysis manager classes provide a uniform interface to the g4tools objects and also hide the differences between the classes for different supported output formats. They take care of higher level management of the g4tools objects, handle allocation and removal of the objects in memory and provide the methods to access them via indexes. In addition, various features specific to Geant4 applications are implemented in the analysis classes following users requests, such as for example the g4tools objects activation, support for Geant4 units or a rich set of Geant4 user interface commands.

In this presentation, we will give a brief overview of the category, then report on new developments since our CHEP 2013 contribution and on upcoming new features.

**Tertiary Keyword (Optional):**
Software development process and tools

**Primary Keyword (Mandatory):**
Analysis tools and techniques

**Secondary Keyword (Optional):**
Simulation
Online & Offline Storage and Processing for the upcoming European XFEL Experiments

Authors: Janusz Malka 1; Martin Gasthuber 1; Stefan Dietrich 1

Co-authors: Birgit Lewendel 2; Djelloul Boukhelef 3; Janusz Szuba 3; Krzysztof Wrona 4; Manuela Kuhn 1; Uwe Ensslin 1; Volker Guelzow 2

1 DESY
2 Deutsches Elektronen-Synchrotron (DE)
3 XFEL
4 European XFEL

Corresponding Authors: stefan.dietrich@desy.de, jmalka@mail.desy.de, janusz.szuba@xfel.eu, birgit.lewendel@desy.de, volker.guelzow@desy.de, krzysztof.wrona@xfel.eu, manuela.kuhn@desy.de, uwe.ensslin@desy.de, djelloul.boukhelef@xfel.eu, martin.gasthuber@cern.ch

For the upcoming experiments at the European XFEL light source facility, a new online and offline data processing and storage infrastructure is currently being built and verified. Based on the experience of the system being developed for the Petra III light source at DESY, presented at the last CHEP conference, we further develop the system to cope with the much higher volumes and rates (>50GB/sec) together with a more complex data analysis and infrastructure conditions (i.e. long range InfiniBand connections). This work will be carried out in collaboration of DESY/IT, European XFEL and technology support from IBM/Research.

This presentation will shortly wrap up the experience of ~1 year runtime of the PetraIII system, continue with a short description of the challenges for the European XFEL experiments and the main section, showing the proposed system for online and offline with initial result from real implementation (HW & SW). This will cover the selected cluster filesystem GPFS including Quality of Service (QOS), extensive use of Flash subsystems and other new and unique features this architecture will benefit from.

Tertiary Keyword (Optional): High performance computing

Primary Keyword (Mandatory): Storage systems

Secondary Keyword (Optional): Computing facilities

Future approach to tier-0 extension

Authors: Ben Jones 1; Gavin McCance 1

Co-author: Domenico Giordano 1

1 CERN
The current tier-0 processing at CERN is done on two managed sites, the CERN computer centre and the Wigner computer centre. With the proliferation of public cloud resources at increasingly competitive prices, we have been investigating how to transparently increase our compute capacity to include these providers. The approach taken has been to integrate these resources using our existing deployment and computer management tools and to provide them in a way that exposes them to users as part of the same site. The paper will describe the architecture, the toolset and the current production experiences of this model.

Tertiary Keyword (Optional): Virtualization

Primary Keyword (Mandatory): Computing facilities

Secondary Keyword (Optional): Cloud technologies

The CERN Computer Security Team is assisting teams and individuals at CERN who want to address security concerns related to their computing endeavours. For projects in the early stages, we help incorporate security in system architecture and design. For software that is already implemented, we do penetration testing. For particularly sensitive components, we perform code reviews. Finally, for everyone undertaking threat modelling or risk assessment, we provide input and expertise. After several years of these internal security consulting efforts, it seems a good moment to analyse experiences, recognise patterns and draw some conclusions. Additionally, it’s worth mentioning two offspring activities that emerged in the last year or so: White Hat training, and the IT Consulting service.

Primary Keyword (Mandatory): Security and policies

Secondary Keyword (Optional): Software development process and tools

Tertiary Keyword (Optional):
In order to patch web servers and web application in a timely manner, we first need to know which software packages are used, and where. But, a typical web stack is composed of multiple layers, including the operating system, web server, application server, programming platform and libraries, database server, web framework, content management system etc. as well as client-side tools. Keeping track of all the technologies used, especially in a heterogeneous computing environment as found in research labs and academia, is particularly difficult. WAD, a tool developed at CERN based on a browser plugin called Wappalyzer, makes it possible to automate this task by detecting technologies behind a given URL. It allows for establishing and maintaining an inventory of web assets, and consequently greatly improves the coverage of any vulnerability management activities.

Primary Keyword (Mandatory):
Security and policies

Secondary Keyword (Optional):
Network systems and solutions

Tertiary Keyword (Optional):

Track 7: Middleware, Monitoring and Accounting / 25

Integrating Containers in the CERN Private Cloud

Author: Ricardo Brito Da Rocha

Co-authors: Bertrand Noel; Mathieu Velten

Corresponding Authors: ricardo.rocha@cern.ch, bertrand.noel@cern.ch, mathieu.velten@cern.ch

Containers remain a hot topic in computing, with new use cases and tools appearing every day. Basic functionality such as spawning containers seems to have settled, but topics like volume support or networking are still evolving. Solutions like Docker Swarm, Kubernetes or Mesos provide similar functionality but target different use cases, exposing distinct interfaces and APIs.

The CERN private cloud is made of thousands of nodes and users, with many different use cases. A single solution for container deployment would not cover every one of them, and supporting multiple solutions involves repeating the same process multiple times for integration with authentication services, storage services or networking.

In this presentation we will describe OpenStack Magnum as the solution to offer container management in the CERN cloud. We will cover its main functionality and some advanced use cases using Docker Swarm and Kubernetes, highlighting some relevant differences between the two. We will describe the most common use cases in HEP and how we integrated popular services like CVMFS or AFS in the most transparent way possible, along with some limitations found. Finally we will look into ongoing work on advanced scheduling for both Swarm and Kubernetes, support for running batch like workloads and integration of container networking technologies with the CERN infrastructure.

Tertiary Keyword (Optional):
Virtualization
CERN Computing in Commercial Clouds

Author: Domenico Giordano
Co-authors: Cristovao Cordeiro; Laurence Field

With the imminent upgrades to the LHC and the consequent increase of the amount and complexity of data collected by the experiments, CERN’s computing infrastructures will be facing a large and challenging demand of computing resources. Within this scope, the adoption of cloud computing at CERN has been evaluated and has opened the doors for procuring external cloud services from providers, which can supply the computing services needed to extend the current CERN infrastructure.

Over the past two years the CERN procurement initiatives and partnership agreements have led to several cloud computing activities between the CERN IT department and firms like ATOS, Microsoft Azure, T-Systems, Deutsche Börse Cloud Exchange and IBM SoftLayer. As of summer 2016 more than 10 Million core-hours of computing resources will have been delivered by commercial cloud providers to the 4 LHC experiments to run their production workloads, from simulation to full chain processing.

In this paper we describe the experience gained in procuring and exploiting commercial cloud resources for the computing needs of the LHC experiments. The mechanisms used for provisioning, monitoring, accounting, alarming and benchmarking will be discussed, as well as the feedback received from the LHC collaborations in terms of managing experiment workflows within a multi-cloud environment.

Real-time complex event processing for cloud resources

Author: Cristovao Cordeiro
Co-authors: Domenico Giordano; Laurence Field; Luca Magnoni; Martin Adam

1 CERN
2 Acad. of Sciences of the Czech Rep. (CZ)
The ongoing integration of clouds into the WLCG raises the need for a detailed health and performance monitoring of the virtual resources in order to prevent problems of degraded service and interruptions due to undetected failures. When working in scale, the existing monitoring diversity can lead to a metric overflow whereby the operators need to manually collect and correlate data from several monitoring tools and frameworks, resulting in tens of different metrics to be interpreted and analysed per virtual machine, constantly.

In this paper we present an ESPER based standalone application which is able to process complex monitoring events coming from various sources and automatically interpret data in order to issue alarms upon the resources’ statuses, without interfering with the actual resources and data sources. We will describe how this application has been used with both commercial and non-commercial cloud activities, allowing the operators to quickly be alarmed and react upon VMs and clusters running with a low CPU load and low network traffic, among other anomalies, resulting then in either the recycling of the misbehaving VMs or fixes on the submission of the LHC experiments workflows. Finally we’ll also present the pattern analysis mechanisms being used as well as the surrounding Elastic and REST API interfaces where the alarms are collected and served to users.

Primary Keyword (Mandatory):
Artificial intelligence/Machine learning

Secondary Keyword (Optional):
Monitoring

Tertiary Keyword (Optional):

---

**Posters A / Break / 28**

**Benchmarking cloud resources**

*Authors: Alessandro De Salvo¹; Alessandro Di Girolamo²; Arne Wiebalck²; Cristovao Cordeiro²; Domenico Giordano²; Laurence Field²; Manfred Alef³*

¹ Universita e INFN, Roma I (IT)
² CERN
³ Karlsruhe Institute of Technology (KIT)

**Corresponding Authors:** laurence.field@cern.ch, arne.wiebalck@cern.ch, alessandro.di.girolamo@cern.ch, manfred.alef@kit.edu, domenico.giordano@cern.ch, cristovao.cordeiro@cern.ch, alessandro.de.salvo@cern.ch

Performance measurements and monitoring are essential for the efficient use of computing resources. In a commercial cloud environment an exhaustive resource profiling has additional benefits due to the intrinsic variability of the virtualised environment. In this context resource profiling via synthetic benchmarking quickly allows to identify issues and mitigate them. Ultimately it provides information about the actual delivered performance of invoiced resources.

In the context of its commercial cloud initiatives, CERN has acquired extensive experience in benchmarking commercial cloud resources, including Amazon, Microsoft Azure, IBM, ATOS, T-Systems, the Deutsche Boerse Cloud Exchange. The CERN cloud procurement process has greatly profited of the benchmark measurements to assess the compliance of the bids with the requested technical specifications. During the cloud production activities, the job performance has been compared with the benchmark measurements.

In this report we will discuss the experience acquired and the results collected using several benchmark metrics. Those benchmarks span from generic open-source benchmarks (encoding algorithm
and kernel compilers) to experiment specific benchmarks (ATLAS KitValidation) and synthetic benchmarks (Whetstone and random number generators). The workflow put in place to collect and analyse performance metrics will be also described.

Primary Keyword (Mandatory):
Cloud technologies

Secondary Keyword (Optional):
Accounting and information

Tertiary Keyword (Optional):

Track 6: Infrastructures / 29

RapidIO as a multi-purpose interconnect

Author: Simaolhoda Baymani

1 CERN

Corresponding Author: sima.baymani@cern.ch

RapidIO (http://rapidio.org/) technology is a packet-switched high-performance fabric, which has been under active development since 1997. Originally meant to be a front side bus, it developed into a system level interconnect which is today used in all 4G/LTE base stations world wide. RapidIO is often used in embedded systems that require high reliability, low latency and scalability in a heterogeneous environment - features that are highly interesting for several use cases, such as data analytics and data acquisition networks.

We will present the results of evaluating RapidIO in a Data Analytics environment, from setup to benchmark. Specifically, we will share the experience of running ROOT and Hadoop on top of RapidIO.

To demonstrate the multi-purpose characteristics of RapidIO, we will also present the results of investigating RapidIO as a technology for high-speed Data Acquisition networks using a generic multi-protocol event-building emulation tool.

In addition we will present lessons learned from implementing native ports of CERN applications to RapidIO.

Primary Keyword (Mandatory):
High performance computing

Secondary Keyword (Optional):
DAQ

Tertiary Keyword (Optional):

Posters A / Break / 30

Evolution of CERN Print Services

Authors: Natalie Kane; Vincent Nicolas Bippus

Co-author: Tim Smith
CERN Print Services include over 1000 printers and multi-function devices as well as a centralised print shop. Every year, some 12 million pages are printed. We will present the recent evolution of CERN print services, both from the technical perspective (automated web-based configuration of printers, Mail2Print) and the service management perspective.

Primary Keyword (Mandatory): Collaborative tools

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Windows Terminal Server Orchestration

Author: Sebastian Bukowiec

Co-author: Tim Smith

Windows Terminal Servers provide application gateways for various parts of the CERN accelerator complex, used by hundreds of CERN users every day. The combination of new tools such as Puppet, HAProxy and Microsoft System Center suite enable automation of provisioning workflows to provide a terminal server infrastructure that can scale up and down in an automated manner. The orchestration does not only reduce the time and effort necessary to deploy new instances, but also facilitates operations such as patching, analysis and recreation of compromised nodes as well as catering for workload peaks.

Primary Keyword (Mandatory): Computing facilities

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Update on CERN Search based on SharePoint 2013

Authors: Andreas Wagner; Eduardo Alvarez Fernandez; Ismael Posada Trobo; Sergio Fernandez Casado

Co-authors: Alexandre Lossent; Bruno Silva De Sousa

1 CERN IT-CDA

2 CERN IT-CDA
CERN's enterprise Search solution "CERN Search" provides a central search solution for users and CERN service providers. A total of about 20 million public and protected documents from a wide range of document collections is indexed, including Indico, TWiki, Drupal, SharePoint, JACOW, E-group archives, EDMS, and CERN Web pages.

In spring 2015, CERN Search was migrated to a new infrastructure based on SharePoint 2013. In the context of this upgrade, the document pre-processing and indexing process was redesigned and generalised. The new data feeding framework allows to profit from new functionality and to facilitate the long term maintenance of the system.

Primary Keyword (Mandatory): Collaborative tools

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

First experience with the new .CERN top level domain

Authors: Andreas Wagner¹; Eduardo Alvarez Fernandez²; Manuel Malo de Molina¹; Marek Salwerowicz¹
Co-authors: Bruno Silva De Sousa ¹; Tim Smith ¹

¹ CERN IT-CDA

Corresponding Authors: andreas.wagner@cern.ch, bruno.sousa@cern.ch, manuel.malo.de.molina.martinez@cern.ch, tim.smith@cern.ch, eduardo.alvarez.fernandez@cern.ch, marek.salwerowicz@cern.ch

In October 2015, CERN’s core website has been moved to a new address, [http://home.cern](http://home.cern), marking the launch of the brand new top-level domain .cern. In combination with a formal governance and registration policy, the IT infrastructure needed to be extended to accommodate the hosting of Web sites in this new top level domain. We will present the technical implementation in the framework of the CERN Web Services that allows to provide virtual hosting and a reverse proxy solution and includes the provisioning of SSL server certificates for secure communications.

Primary Keyword (Mandatory): Collaborative tools

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

Flash is Dead. Finally.

Author: Marek Domaracky ¹

Co-authors: Magdalena Zofia Peksa ²; Natalia Karina Juszka ²; Thomas Baron ¹

¹ CERN
It’s been for almost 10 years that CERN has been providing live webcast of events using Adobe Flash technology. This year is finally the year that flash died at CERN! At CERN we closely follow the broadcast industry and are always trying to provide our users with the same experience as they have on other commercial streaming services. With Flash being slowly phased out on most of the streaming platforms, we moved as well from Flash to HTTP streaming. All our live streams are delivered via the HTTP Live Streaming (HLS) protocol, which is supported in all modern browsers on desktops and mobile devices. Thanks to HTML5 and the THEOPlayer we are able to deliver the same experience as we did with Adobe Flash based players. Our users can still enjoy video of the speaker synchronised with video of the presentation, so they have the same experience as sitting in the auditoria.

For On Demand Video, to reach our users on any device, we improved the process of publishing recorded lectures with the new release of the CERN Lecture Archiving system - Micala. We improved the lecture viewer, which gives our users the best possible experience of watching recorded lectures, with both video of the speaker and slides in high resolution. For the mobile devices we improved quality and usability to watch any video from CDS even on low bandwidth conditions. We introduced DVR functionality for all our live webcasts. Users that arrived late on the webcast website, now have a possibility to go back to the beginning of the webcast or if they missed something they can seek back to watch it again. With DVR functionality we are able to provide recording right after the webcast is finished.

With 19 CERN rooms capable of webcast and recording, about 300 live webcasts and 1200 lectures recorded every year, we needed a tool for our operators to start webcast and recording easily. We developed a Central Encoding Interface, from which our operators see all the events for a given day and with one click can start webcasting and/or recording. With this new interface we manage to almost eliminate issues where operators forget to start the webcast and with an automatic stop, we now support webcasts and recording which finish out of standard working hours without additional expenses.

Primary Keyword (Mandatory):
Collaborative tools

Secondary Keyword (Optional):
Outreach

Tertiary Keyword (Optional):

Track 8: Security, Policy and Outreach / 36

Access to WLCG resources: The X509-free pilot

Authors: Andrea Manzi\(^1\); Andrey Kirianov\(^2\); Hannah Short\(^1\); Henri Johannes Mikkonen\(^3\); Oliver Keeble\(^1\); Paolo Tedesco\(^1\); Romain Wartel\(^1\); Vincenzo De Notaris\(^1\)

\(^1\) CERN
\(^2\) B.P. Konstantinov Petersburg Nuclear Physics Institute - PNPI
\(^3\) Helsinki Institute of Physics (FI)

Corresponding Authors: oliver.keeble@cern.ch, paolo.tedesco@cern.ch, romain.wartel@cern.ch, andrea.manzi@cern.ch, andrey.kiryanov@cern.ch, vincenzo.de.notaris@cern.ch, henri.mikkonen@cern.ch, hannah.short@cern.ch

Access to WLCG resources is authenticated using an X509 and PKI infrastructure. Even though HEP users have always been exposed to certificates directly, the development of modern Web Applications by the LHC experiments calls for simplified authentication processes keeping the underlying software unmodified.

In this work we will show an integrated Web-oriented solution (code name Kipper) with the goal...
of providing access to WLCG resources using the user's home organisation's credentials, without
the need for user-acquired X.509 certificates. In particular, we focus on identity providers within
eduGAIN, which interconnects research and education organisations worldwide, and enables the
trustworthy exchange of identity-related information.
eduGAIN has been integrated at CERN in the SSO infrastructure so that users can authenticate with-
out the need of a CERN account.
This solution achieves "X.509-free" access to Grid resources with the help of two services: STS and
an online CA. The STS (Security Token Service) allows credential translation from the SAML2 for-
mat used by Identity Federations to the VOMS-enabled X.509 used by most of the Grid. The IOTA
(Identifier-Only Trust Assurance) CA is responsible for the automatic issuing of short-lived X.509
certificates.
The IOTA CA deployed at CERN has been accepted by EUGridPMA as the CERN LCG IOTA CA,
included in the IGTF trust anchor distribution and installed by the sites in WLCG.
We will also describe the first example of Kipper allowing eduGAIN access to WLCG, the WebFTS in-
terface to the FTS3 data transfer engine, enabled by integration of multiple services: WebFTS, CERN
SSO, CERN LCG IOTA CA, STS, and VOMS

Primary Keyword (Mandatory):
Security and policies

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

Posters A / Break / 37

Deploying FTS with Docker Swarm and Openstack Magnum

Authors: Alejandro Alvarez Ayllon\(^1\); Andrea Manzi\(^1\); Maria Arsuaga Rios\(^1\); Ricardo Brito Da Rocha\(^1\)

\(^1\) CERN

Corresponding Authors: andrea.manzi@cern.ch, maria.arsuaga.rios@cern.ch, ricardo.rocha@cern.ch, alejandro.alvarez.ayllon@cern.ch

The deployment of Openstack Magnum at CERN has given the possibility to manage container or-
chestration engines such as Docker and Kubernetes as first class resources in Openstack.
In this poster we will show the work done to exploit a docker Swarm cluster deployed via Magnum
to setup a docker infrastructure running FTS (the WLCG file transfer service). FTS has been chosen
as one of the pilots to validate the Magnum and docker integration with the rest of the CERN infras-
tructure tools. The FTS service has an architecture that is suitable for the exploitation of containers:
the functionality now offered by a VM cluster can be decomposed in dedicated containers and sepa-
rately scaled according to the load and user interactions.
The pilot is under evaluation with a view to a docker-based FTS production deployment.

Primary Keyword (Mandatory):
Virtualization

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Posters A / Break / 38

DNS Load Balancing in the CERN Cloud
Author: Ignacio Reguero
Co-author: Lorena Lobato Pardavila

1 CERN
2 Universidad de Oviedo (ES)

Corresponding Authors: ignacio.reguero@cern.ch, lorena.lobato@cern.ch

Load Balancing is one of the technologies enabling deployment of large scale applications on cloud resources. At CERN we have developed a DNS Load Balancer as a cost-effective way to do it for applications accepting DNS timing dynamics and not requiring memory. We serve 378 load balanced aliases with two small VMs acting as master and slave. These aliases are based on ‘delegated’ DNS zones the we manage with DYN-DNS based on a load metric collected with SNMP from the alias members.

In the last years we have done several improvements to the software, for instance support for IPV6 AAAA records, parallelization of the SNMP requests, as well as reimplementing the client in python allowing for multiple aliases with differentiated state on the same machine, support for Roger state and other new features.

The configuration of the Load Balancer is built with a Puppet type that gets the alias members dynamically from PuppetDB and consumes the alias definitions from a REST service.

We have produced a self-service GUI for the management of the LB aliases based on the REST service above implementing a form of Load Balancing as a Service (LBaaS). Both the GUI and REST API have authorisation based in hostgroups. All this is implemented with Open Software without much CERN specific code.

Tertiary Keyword (Optional):
Network systems and solutions

Primary Keyword (Mandatory):
Distributed data handling

Secondary Keyword (Optional):
Cloud technologies

Posters A / Break / 39

Customization of the general fitting tool genfit2 in PandaRoot

Author: Elisabetta Prencipe
Co-authors: Stefano Spataro; Tobias Stockmanns

1 Forschungszentrum Juelich
2 University of Turin

Corresponding Authors: spataro@to.infn.it, e.prencipe@fz-juelich.de, t.stockmanns@fz-juelich.de

PANDA is a planned experiment at FAIR (Darmstadt, Germany) with a cooled antiproton beam in a range [1.5; 15] GeV/c, allowing a wide physics program in nuclear and particle physics. It is the only experiment worldwide, which combines a solenoid field (B=2T) and a dipole field (B=2Tm) in an experiment with a fixed target topology, in that energy regime. The tracking system of PANDA involves the presence of a high performance silicon vertex detector, a GEM detector, a Straw- Tubes central tracker, a forward tracking system, and a luminosity monitor. The offline tracking algorithm
is developed within the PandaRoot framework, which is a part of the FairRoot project. The tool here presented is based on algorithms containing the Kalman Filter equations and a deterministic annealing filter. This general fitting tool (genfit2) offers to users also a Runge-Kutta track representation, and interfaces with Millepede II (useful for alignment) and RAVE (vertex finder). It is independent on the detector geometry and B field map, and written in C++ o.o. modular code. Several fitting algorithms are available with genfit2, with user-adjustable parameters; therefore the tool is of friendly usage. A check on the fit convergence is done by genfit2 as well. The Kalman-Filter-based algorithms have a wide range of applications; among those in particle physics they can perform extrapolations of track parameters and covariance matrices. The impact of genfit2 on physics simulations performed for the PANDA experiment is shown, with the PandaRoot framework: significant improvement is reported for those channels where a good low momentum tracking is required (pT < 400 MeV/c).

Primary Keyword (Mandatory):
Reconstruction

Secondary Keyword (Optional):
Algorithms

Tertiary Keyword (Optional):

Posters A / Break / 40

A world-wide databridge supported by a commercial cloud provider

Authors: Fabrizio Furano\(^1\); Kwong Tat Cheung\(^2\); Laurence Field\(^1\)

\(^1\) CERN
\(^2\) University of West England

Corresponding Authors: kwong.cheung@live.uwe.ac.uk, laurence.field@cern.ch, fabrizio.furano@cern.ch

Volunteer computing has the potential to provide significant additional computing capacity for the LHC experiments. One of the challenges with exploiting volunteer computing is to support a global community of volunteers that provides heterogeneous resources. However, HEP applications require more data input and output than the CPU intensive applications that are typically used by other volunteer computing projects. While the so-called "databridge" has already been successfully proposed as a method to span the untrusted and trusted domains of volunteer computing and Grid computing respective, globally transferring data between potentially poor-performing public networks at home and CERN can be fragile and lead to wasted resources usage. The expectation is that by placing closer to the volunteers a storage endpoint that is part of a wider, flexible geographical databridge deployment, the transfer success rate and the overall performance can be improved. This contribution investigates the provision of a globally distributed databridge implemented upon a commercial cloud provider.

Tertiary Keyword (Optional):
Storage systems

Primary Keyword (Mandatory):
Distributed data handling

Secondary Keyword (Optional):
DPM Evolution: a Disk Operations Management Engine for DPM

Authors: Andrea Manzi¹; Fabrizio Furano¹; Georgios Bitzes¹; Oliver Keeble¹

¹ CERN

Corresponding Authors: fabrizio.furano@cern.ch, georgios.bitzes@cern.ch, andrea.manzi@cern.ch, oliver.keeble@cern.ch

The DPM (Disk Pool Manager) project is the most widely deployed solution for storage of large data repositories on Grid sites, and is completing the most important upgrade in its history, with the aim of bringing important new features, performance and easier long term maintainability. Work has been done to make the so-called “legacy stack” optional, and substitute it with an advanced implementation that is based on the fastCGI and RESTful technologies. Beside the obvious gain in making optional several legacy components that are difficult to maintain, this step brings important features together with performance enhancements. Among the most important features we can cite the simplification of the configuration, the possibility of working in a totally SRM-free mode, the implementation of quotas, free/used space on directories, and the implementation of volatile pools that can pull files from external sources, which can be used to deploy data caches. Moreover, the communication with the new core, called DOME (Disk Operations Management Engine) now happens through secure HTTPS channels through an extensively documented, industry-compliant protocol.

For this leap, referred to with the codename “DPM Evolution”, the help of the DPM collaboration has been very important in the beta testing phases, and here we report about the technical choices and the first site experiences.

Primary Keyword (Mandatory):
Storage systems

Secondary Keyword (Optional):
Computing middleware

Tertiary Keyword (Optional):

Making the most of cloud storage - a toolkit for exploitation by WLCG experiments.

Author: Oliver Keeble¹

¹ CERN

Corresponding Authors: georgios.bitzes@cern.ch, alejandro.alvarez.ayllon@cern.ch, oliver.keeble@cern.ch, fabrizio.furano@cern.ch, maria.arsuaga.rios@cern.ch, andrea.manzi@cern.ch
Understanding how cloud storage can be effectively used, either standalone or in support of its associated compute, is now an important consideration for WLCG.

We report on a suite of extensions to familiar tools targeted at enabling the integration of cloud object stores into traditional grid infrastructures and workflows. Notable updates include support for a number of object store flavours in FTS3, Davix and gfal2, including mitigations for lack of vector reads; the extension of Dynafed to operate as a bridge between grid and cloud domains; protocol translation in FTS3; the implementation of extensions to DPM (also implemented by the dCache project) to allow 3rd party transfers over HTTP.

The result is a toolkit which facilitates data movement and access between grid and cloud infrastructures, broadening the range of workflows suitable for cloud. We report on deployment scenarios and prototype experience, explaining how, for example, an Amazon S3 or Azure allocation can be exploited by grid workflows.

Tertiary Keyword (Optional):
Storage systems

Primary Keyword (Mandatory):
Cloud technologies

Secondary Keyword (Optional):
Object stores

Posters A / Break / 43

EOS Cross Tier Federation

Authors: Andreas Joachim Peters¹; Geoffray Michel Adde¹

¹ CERN

Corresponding Authors: andreas.joachim.peters@cern.ch, geoffray.adde@cern.ch

Within the WLCG project EOS is evaluated as a platform to demonstrate efficient deployment of geographically distributed storage. Aim of distributed storage deployments is to reduce the number of individual end-points for LHC experiments (>100 today) and to minimize the required effort for small storage sites. The split of meta-data and data component in EOS allows to operate one regional high-available meta data service (MGM) and to deploy the easier to operate file storage component (FST) in geographically distributed sites. EOS has built-in support for geolocation-aware access scheduling, file placement policies and replication workflows.

This contribution will introduce the various concepts and discuss demonstrator deployments for several LHC experiments.

Primary Keyword (Mandatory):
Storage systems

Secondary Keyword (Optional):
Distributed data handling

Tertiary Keyword (Optional):

Track 4: Data Handling / 44
EOS Developments

Authors: Andreas Joachim Peters\textsuperscript{1}; Elvin Alin Sindrilaru\textsuperscript{1}; Geoffray Michel Adde\textsuperscript{1}; Paul Hermann Lensing\textsuperscript{None}

\textsuperscript{1} CERN

Corresponding Authors: andreas.joachim.peters@cern.ch, geoffray.adde@cern.ch, paul.lensing@cern.ch, elvin.alin.sindrilaru@cern.ch

CERN has been developing and operating EOS as a disk storage solution successfully for 5 years. The CERN deployment provides 135 PB and stores 1.2 billion replicas distributed over two computer centres. Deployment includes four LHC instances, a shared instance for smaller experiments and since last year an instance for individual user data as well. The user instance represents the backbone of the CERNBOX service for file sharing. New use cases like synchronisation and sharing, the planned migration to reduce AFS usage at CERN and the continuous growth has brought EOS to new challenges.

Recent developments include the integration and evaluation of various technologies to do the transition from a single active in-memory namespace to a scale-out implementation distributed over many meta-data servers. The new architecture aims to separate the data from the application logic and user interface code, thus providing flexibility and scalability to the namespace component.

Another important goal is to provide EOS as a CERN-wide mounted filesystem with strong authentication making it a single storage repository accessible via various services and front-ends (\texttt{/eos initiative}). This required new developments in the security infrastructure of the EOS Fuse implementation. Furthermore, there were a series of improvements targeting the end-user experience like tighter consistency and latency optimisations.

In collaboration with SEAGATE as openlab partner, EOS has a complete integration of OpenKinetic object drive cluster as a high-throughput, high-availability, low-cost storage solution.

This contribution will discuss these three main development projects and present new performance metrics.

Primary Keyword (Mandatory):

Storage systems

Secondary Keyword (Optional):

Object stores

Tertiary Keyword (Optional):

Track 3: Distributed Computing / 45

Using HEP Computing Tools, Grid and Supercomputers for Genome Sequencing Studies

Authors: Alexei Klimentov\textsuperscript{1}; Alexey Poyda\textsuperscript{2}; Ruslan Mashinistov\textsuperscript{2}

Co-authors: Alexander Novikov\textsuperscript{2}; Anton Teslyuk\textsuperscript{2}; Artem Nedoluzhko\textsuperscript{2}; Eugene Ryabinkin\textsuperscript{2}; Fedor Shargo\textsuperscript{2}; Ivan Tertychnyy\textsuperscript{2}; Kaushik De\textsuperscript{3}; Tadashi Maeno\textsuperscript{1}; Torre Wenaus\textsuperscript{1}

\textsuperscript{1} Brookhaven National Laboratory (US)
\textsuperscript{2} National Research Centre Kurchatov Institute (RU)
\textsuperscript{3} University of Texas at Arlington (US)

Corresponding Authors: wenaus@gmail.com, tmaeno@bnl.gov, poyda@wdbc.ru, ivan.tertychnyy@cern.ch, kaushik@uta.edu, fedosic@gmail.com, ruslan.mashinistov@cern.ch, alexei.klimentov@cern.ch, novikov@wdbc.ru, nedoluzhko@gmail.com, rea@grid.kiae.ru, anthony.teslyuk@gmail.com

PanDA - Production and Distributed Analysis Workload Management System has been developed to address ATLAS experiment at LHC data processing and analysis challenges. Recently PanDA has been extended to run HEP scientific applications on Leadership Class Facilities and supercomputers.
The success of the projects to use PanDA beyond HEP and Grid has drawn attention from other compute intensive sciences such as bioinformatics.

Modern biology uses complex algorithms and sophisticated software, which is impossible to run without access to significant computing resources. Recent advances of Next Generation Genome Sequencing (NGS) technology led to increasing streams of sequencing data that need to be processed, analysed and made available for bioinformaticians worldwide. Analysis of ancient genomes sequencing data using popular software pipeline PALEOMIX can take a month even running it on the powerful computer resource. PALEOMIX include typical set of software used to process NGS data including adapter trimming, read filtering, sequence alignment, genotyping and phylogenetic or metagenomic analysis. Sophisticated computing software WMS and efficient usage of the supercomputers can greatly enhance this process.

In this paper we will describe the adaptation the PALEOMIX pipeline to run it on a distributed computing environment powered by PanDA. We used PanDA to manage computational tasks on a multi-node parallel supercomputer. To run pipeline we split input files into chunks which are run separately on different nodes as separate inputs for PALEOMIX and finally merge output file, it is very similar to what it done by ATLAS to process and to simulate data. We dramatically decreased the total walltime because of jobs (re)submission automation and brokering within PanDA, what was earlier demonstrated for the ATLAS applications on the Grid. Using software tools developed initially for HEP and Grid can reduce payload execution time for Mammoths DNA samples from weeks to days.

Tertiary Keyword (Optional):
High performance computing

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

Secondary Keyword (Optional):
Experience/plans from outside experimental HEP/NP

Support Vector Machines in HEP

Author: Adrian Bevan

Co-authors: Agni Bethani; Jonathan Hays; Thomas James Stevenson

1 University of London (GB)
2 Centre National de la Recherche Scientifique (FR)

Corresponding Authors: thomas.james.stevenson@cern.ch, j.hays@qmul.ac.uk, agni.bethani@cern.ch, a.j.bevan@qmul.ac.uk

We review the concept of support vector machines before proceeding to discuss examples of their use in a number of scenarios. Using the Toolkit for Multivariate Analysis (TMVA) implementation we discuss examples relevant to HEP including background suppression for H->tau+tau- at the LHC. The use of several different kernel functions and performance benchmarking is discussed.

Tertiary Keyword (Optional):
Artificial intelligence/Machine learning

Primary Keyword (Mandatory):
Algorithms

Secondary Keyword (Optional):
Production Management System for AMS Remote Computing Centers

Author: Baosong Shan

Co-authors: Alexander Egorov; Alexandre Eline; Oleg Demakov; Renli Shi; Vitaly Choutko

1 Beihang University (CN)
2 Massachusetts Inst. of Technology (US)
3 Southeast University (CN)

The Alpha Magnetic Spectrometer (AMS) on board of the International Space Station (ISS) requires a large amount of computing power for data production and Monte Carlo simulation. A large fraction of the computing resource has been contributed by the computing centers among the AMS collaboration. AMS has 12 "remote" computing centers outside of Science Operation Center at CERN, with different hardware and software configurations.

This paper presents a production management system for remote computing sites, to automate the processes including job acquiring, submitting, monitoring, transferring and accounting. The system is designed to be modularized, light-weighted, and easy-to-be-deployed. It is based on Deterministic Finite Automaton, and implemented by script languages, Python and Perl, and the built-in Sqlite3 database on Linux operating systems. Different batch management systems (LSF, PBS, Condor ...), file system storage (GPFS, Lustre, EOS ...), and transferring protocols (GRIDFTP, XROOTD ...) are supported. In addition, the recent experience of the integration of the system with Open Science Grid is also described.

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Evolution of Monitoring System for AMS Science Operation Center

Author: Baosong Shan

Co-authors: Alexander Egorov; Alexandre Eline; Oleg Demakov; Renli Shi; Vitaly Choutko

1 Beihang University (CN)
2 Massachusetts Inst. of Technology (US)
3 Southeast University (CN)
Corresponding Authors: alexander.egorov@cern.ch, alexandre.eline@cern.ch, baosong.shan@cern.ch, oleg.demakov@cern.ch, renli.shi@cern.ch, vitaly.choutko@cern.ch

This paper introduces the evolution of the monitoring system of the Alpha Magnetic Spectrometer (AMS) Science Operation Center (SOC) at CERN.

The AMS SOC monitoring system includes several independent tools: Network Monitor to poll the health metrics of AMS local computing farm, Production Monitor to show the production status, Frame Monitor to record the flight data arriving status, and SOC monitor to check the production latency.

Currently CERN has adopted Metrics as the main monitoring platform, and we are working to integrate our monitoring tools to this platform to provide dashboard like monitoring pages which will show the overall status of SOC as well as more detailed information. A diagnostic tool based on set of expandable rules and capable to automatically locate the possible issues and provide suggestions for the fixes is also being designed.

Primary Keyword (Mandatory):
Monitoring

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

---

Storage Strategy of AMS Science Data at Science Operation Center at CERN

Authors: Baosong Shan\(^1\); Vitaly Choutko\(^2\)

Co-authors: Alexander Egorov\(^2\); Alexandre Eline\(^2\); Oleg Demakov\(^2\); Renli Shi\(^3\)

\(^1\) Beihang University (CN)  
\(^2\) Massachusetts Inst. of Technology (US)  
\(^3\) Southeast University (CN)

Corresponding Authors: alexandre.eline@cern.ch, vitaly.choutko@cern.ch, baosong.shan@cern.ch, alexander.egorov@cern.ch, renli.shi@cern.ch, oleg.demakov@cern.ch

This paper introduces the storage strategy and tools of the science data of the Alpha Magnetic Spectrometer (AMS) at Science Operation Center (SOC) at CERN.

The AMS science data includes flight data, reconstructed data and simulation data, as well as the metadata of them. The data volume is 1070 TB per year of operation, and currently reached 5086 TB in total. We have two storage levels: active/live data which is ready for analysis, and backups on CASTOR. Active/live data is stored on AMS local storage, and after the introduction of CERN EOS system, also on EOS. A system is designed to automate the data moving and data backup.

The data validation, the metadata design, and the ways to preserve the consistency between the data and the metadata are also presented.

Primary Keyword (Mandatory):
Storage systems

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
AMSI Data Production Facilities at Science Operation Center at CERN

Authors: Alexander Egorov 1; Vitaly Choutko 1

Co-authors: Alexandre Eline 1; Baosong Shan 2; Oleg Demakov 1; Renli Shi 3

1 Massachusetts Inst. of Technology (US)
2 Beihang University (CN)
3 Southeast University (CN)

Corresponding Authors: baosong.shan@cern.ch, alexandre.eline@cern.ch, oleg.demakov@cern.ch, vitaly.choutko@cern.ch, alexander.egorov@cern.ch, renli.shi@cern.ch

The AMS data production uses different programming modules for job submission, execution and management, as well as for validation of produced data. The modules communicate with each other using CORBA interface. The main module is the AMS production server, a scalable distributed service which links all modules together starting from job submission request and ending with writing data to disk storage. Each running instance of the server has the capacity to manage around 1000 multithreaded jobs capable to serve up to 64K CPUs. Monitoring and management tools with enhanced GUI are also described.

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Context-aware distributed cloud computing using CloudScheduler

Author: Randall Sobie 1

Co-authors: Colin Roy Leavett-Brown 1; Darryl Ring 1; Kevin Casteels 1; Matthew Murray 1; Michael Paterson 1; Rolf Seuster 1; Ryan Taylor 1

1 University of Victoria

Corresponding Authors: rsobie@uvic.ca, colin.roy.leavett-brown@cern.ch, asteels@uvic.ca, rptaylor@uvic.ca, ringda@uvic.ca, matthewmurrayy@gmail.com, mhp@uvic.ca, rolf.seuster@cern.ch

The distributed cloud using the CloudScheduler VM provisioning service is one of the longest running systems for HEP workloads. It has run millions of jobs for ATLAS and Belle II over the past few years using private and commercial clouds around the world. Our goal is to scale the distributed cloud to the 10,000-core level, with the ability to run any type of application (low I/O, high I/O and high memory) on any cloud. To achieve this goal, we have been implementing changes that utilize context-aware computing designs that are currently employed in the mobile communication industry. Context-awareness makes use of real-time and archived data to respond to user or system requirements. In our distributed cloud, we have many opportunistic clouds with no local HEP services, software or storage repositories. A context-aware design significantly benefits the reliability and performance of our system by locating the nearest or optimal location of the required services. We describe how we are collecting and managing contextual information from our workload management systems, the clouds, the virtual machine and our services. This information is used not
only to monitor the system but also to carry out automated corrective actions. We are incrementally adding new alerting and response services to our distributed cloud. This will enable us to scale the number of clouds and virtual machines. Further, a context-aware design will enable us to run analysis or high I/O application on opportunistic clouds. We envisage an open-source HTTP data federation (for example, the Dynafed system at CERN) as a service that would provide us access to existing storage elements used by the HEP experiments.

Primary Keyword (Mandatory):
Cloud technologies

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Posters B / Break / 53

**XrootdFS, a posix file system for XrootD**

**Author:** Wei Yang¹

**Co-author:** Andrew Bohdan Hanushevsky ²

¹ SLAC National Accelerator Laboratory (US)

**Corresponding Authors:** yangw@slac.stanford.edu, andrew.bohdan.hanushevsky@cern.ch

When we first introduced XRootD storage system to the LHC, we needed a filesystem interface so that XRootD system could function as a Grid Storage Element. The result was XRootDfs, a FUSE based mountable posix filesystem. It glues all the data servers in a XRootD storage system together and presents it as a single, posix compliant, multi-user networked filesystem. XRootD’s unique redirection mechanism requires special handling of IO operations and metadata operations in the XRootDfs. This includes a throttling mechanism to gracefully handle extreme metadata operations; handling of returned results from all data servers in a consistent way; hiding delays of metadata operations, including storage media latency; enhancing the performance of concurrent IO by multiple applications; and using an advanced security plugin to ensure secure data access in a multi-user environment. Over the last several years XRootDfs have been adopted by many XRootD sites for data management as well as data access by applications that were not specifically designed to use the native XRootD interface. Many of the technical methods mentioned above can also be used to glue together other types (i.e. non-XRootD) data servers to provide seamless data access.

Primary Keyword (Mandatory):
Storage systems

Secondary Keyword (Optional):
Distributed data handling

Tertiary Keyword (Optional):

Posters B / Break / 54

**Research and application of OpenStack in Chinese Spallation Neutron Source Computing environment**

**Author:** Yakang Li¹
Cloud computing can make IT resources configuration flexible and reduce the hardware cost. It also can provide computing service according to the real need. We are applying this computing mode to the Chinese Spallation Neutron Source (CSNS) computing environment. So from the research and practice aspects, firstly, the application status of cloud computing science in High Energy Physics Experiments and the special requirements of CSNS are introduced in this paper. Secondly, the design and practice of cloud computing platform based on OpenStack are mainly demonstrated from the aspects of cloud computing system framework, elastic distribution of resources and so on. Thirdly, some improvements to OpenStack we made are discussed further. Finally, some future prospects of CSNS cloud computing environment are summarized in the ending of this paper.

Primary Keyword (Mandatory):
Cloud technologies

Secondary Keyword (Optional):
Computing facilities

Tertiary Keyword (Optional):

Track 4: Data Handling / 55

Next Generation high performance, multi-dimensional scalable data transfer

Authors: Chin Fang\(^1\); Roger Cottrell\(^2\)

Co-authors: Andrew Bohdan Hanushevsky\(^1\); Wei Yang\(^3\); Wilko Kroeger\(^2\)

\(^1\) Zettar Inc.
\(^2\) SLAC National Accelerator Laboratory
\(^3\) SLAC National Accelerator Laboratory (US)

Corresponding Authors: cottrell@slac.stanford.edu, yangw@slac.stanford.edu, wilko@slac.stanford.edu, andrew.bohdan.hanushevsky@cern.ch, fangchin@zettar.com

The exponentially increasing need for high speed data transfer is driven by big data, cloud computing together with the needs of data intensive science, High Performance Computing (HPC), defense, the oil and gas industry etc. We report on the Zettar ZX software that has been developed since 2013 to meet these growing needs by providing high performance data transfer and encryption in a scalable, balanced, easy to deploy and use way while minimizing power and space utilization. Working with the Zettar development team we have deployed the ZX software on existing Data Transfer Nodes (DTNs) at SLAC and NERSC and transferred data between a Lustre cluster at SLAC and a GPFS cluster at NERSC. ZX supports multi-dimensional scaling in terms of: computation power, network interfaces, and analysis and customization of the storage interface configuration resulting in optimized use of IOPS. It also pays close attention to successfully and efficiently transferring lots of small files and avoids the latency issues of external services. We have verified that the data rates are comparable with the data rates achieved with those of a commonly used high performance data transfer application (bbcp). In collaboration with several commercial vendors, a Proof of Concept (PoC) has also been put together using off the shelf components to test the ZX scalability and ability to balance services using multiple cores and network interfaces to provide much higher throughput. The PoC consists of two 4-node clusters with the flash storage per node aggregated with a parallel file system, plus 10Gbps, 40Gbps and 100Gbps network interfaces. Each cluster plus the two associated switches occupies 4 rack units and draws less than 3.6KW. Using the PoC, between clusters we have achieved 155Gbps memory-to-memory over a 16x10Gbps link aggregated channel (LAG) and 70Gbps file-to-file with encryption over a 5000 mile 100Gbps link.
Globally Distributed Software Defined Storage

**Author:** Andrey Shevel

**Co-authors:** Arsen Kairkanov; Mikhail Kompaniets; Oleg Sadov; Sergey Khoruzhnikov; Vladimír Grudinin

1 Petersburg Nuclear Physics Institute - PNPI, ITMO University
2 ITMO University
3 St. Petersburg State University (RU)

**Corresponding Authors:** oleg.sadov@niuitmo.ru, grudinin@niuitmo.ru, mikhail.kompaniets@cern.ch, anrsen.kairkanov@gmail.com, xse@vuztc.ru, andrey.shevel@cern.ch

The volume of the coming data in HEP is growing. Also growing volume of the data to be hold long time. Actually large volume of data - big data - is distributed around the planet. In other words now there is situation where the data storage does integrate storage resources from many data centers located far from each other. That means the methods, approaches how to organize, manage the globally distributed data storage are required.

For personal needs the distributed storage has several examples like owncloud.org, pydio.com, seafile.com, sparkleshare.org. For enterprise level there are a number of distributed storage systems SWIFT (part of Openstack), CEPH and the like which are mostly object storage.

When distributed storage integrate several data center resources the organization of data links becomes very important issue especially if several parallel data links between data centers are used. The situation on data centers and in data links might be changed each half a day or more often. All that means each part of distributed data storage has to be able to rearrange usage of data links and storage servers in each data center. In addition for each customer of distributed storage the different requirements have to be satisfied.

Mentioned topics and more are planned to present in the proposal under discussion.

An efficient, modular and simple tape archiving solution for LHC Run-3
Authors: Daniele Francesco Kruse\textsuperscript{1}; Eric Cano\textsuperscript{1}; Steven Murray\textsuperscript{1}

Co-authors: German Cancio \textsuperscript{1}; Julien Leduc \textsuperscript{1}; Viktor Kotliar \textsuperscript{2}; Vladimir Bahyl \textsuperscript{1}

\textsuperscript{1} CERN
\textsuperscript{2} Institute for High Energy Physics (RU)

Corresponding Authors: german.cancio.melia@cern.ch, eric.cano@cern.ch, daniele.francesco.kruse@cern.ch, vladimir.bahyl@cern.ch, julien.leduc@cern.ch, steven.murray@cern.ch, victor.kotlyar@cern.ch

The IT Storage group at CERN develops the software responsible for archiving to tape the custodial copy of the physics data generated by the LHC experiments. Physics run 3 will start in 2021 and will introduce two major challenges for which the tape archive software must be evolved. Firstly the software will need to make more efficient use of tape drives in order to sustain the predicted data rate of 100 petabytes per year as opposed to the current 40 petabytes per year of Run-2. Secondly the software will need to be seamlessly integrated with EOS, which has become the de facto disk storage system provided by the IT Storage group for physics data.

The tape storage software for LHC physics run 3 is code named CTA (the CERN Tape Archive). This paper describes how CTA will introduce a pre-emptive drive scheduler to use tape drives more efficiently, will encapsulate all tape software into a single module that will sit behind one or more EOS systems, and will be simpler by dropping support for obsolete backwards compatibility.

Primary Keyword (Mandatory):
Storage systems

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

---

Data Center Environmental Sensor for safeguarding the CERN Data Archive

Author: Julien Leduc\textsuperscript{1}

Co-authors: Daniele Francesco Kruse \textsuperscript{1}; Eric Cano \textsuperscript{1}; German Cancio Melia \textsuperscript{1}; Quentin Genoud \textsuperscript{2}; Steven Murray \textsuperscript{1}; Vlado Bahyl \textsuperscript{1}

\textsuperscript{1} CERN
\textsuperscript{2} GEII Université de Savoie

Corresponding Authors: daniele.francesco.kruse@cern.ch, vladimir.bahyl@cern.ch, steven.murray@cern.ch, quentin.genoud@free.fr, eric.cano@cern.ch, german.cancio.melia@cern.ch, julien.leduc@cern.ch

CERN has been archiving data on tapes in its Computer Center for decades and its archive system is now holding more than 135 PB of HEP data in its premises on high density tapes.

For the last 20 years, tape areal bit density has been doubling every 30 months, closely following HEP data growth trends. During this period, bits on the tape magnetic substrate have been shrinking exponentially; today’s bits are now smaller than most airborne dust particles or even bacteria. Therefore tape media is now more sensitive to contamination from airborne dust particles that can land on the rollers, reels or heads.

These can cause scratches on the tape media as it is being mounted or wound on the tape drive resulting in the loss of significant amounts of data.

To mitigate this threat, CERN has prototyped and built custom environmental sensors that are hosted in the production tape libraries, sampling the same airflow as the surrounding drives. This paper
will expose the problems and challenges we are facing and the solutions we developed in production to better monitor CERN Computer Center environment in tape libraries and to limit the impact of airborne particles on the LHC data.

Tertiary Keyword (Optional):
Monitoring

Primary Keyword (Mandatory):
Storage systems

Secondary Keyword (Optional):
Preservation of analysis and data

Posters B / Break / 59

PaaS for web applications with OpenShift Origin

Authors: Alberto Rodriguez Peon; Alexandre Lossent

1 CERN

Corresponding Authors: alexandre.lossent@cern.ch, alberto.rodriguez.peon@cern.ch

The CERN Web Frameworks team has deployed OpenShift Origin to facilitate deployment of web applications and improve resource efficiency. OpenShift leverages Docker containers and Kubernetes orchestration to provide a Platform-as-a-service solution oriented for web applications. We will review use cases and how OpenShift was integrated with other services such as source control, web site management and authentication services.

Primary Keyword (Mandatory):
Cloud technologies

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Track 1: Online Computing / 60

40-Gbps Data-Acquisition System for NectarCAM, a camera for the Medium Size Telescopes of the Cherenkov Telescope Array

Authors: Dirk Hoffmann; Julien Houles

1 Aix Marseille Université, CNRS/IN2P3, CPPM UMR 7346, 13288, Marseille, France

Corresponding Authors: dirk.hoffmann@cern.ch, houles@cppm.in2p3.fr

The Cherenkov Telescope Array (CTA) will be the next generation ground-based gamma-ray observatory. It will be made up of approximately 100 telescopes of three different sizes, from 4 to 23 meters in diameter. The previously presented prototype of a high speed data acquisition (DAQ) system for CTA (CHEP 2012) has become concrete within the NectarCAM project, one of the most challenging camera projects with very demanding needs for bandwidth of data handling.
We designed a Linux-PC system able to concentrate and process without packet loss the 40 Gb/s average data rate coming from the 265 Front End Boards (FEB) through Gigabit Ethernet links, and to reduce data to fit the two ten-Gigabit Ethernet downstream links by external trigger decisions as well as custom tailored compression algorithms. Within the given constraints, we implemented de-randomisation of the event fragments received as relatively small UDP packets emitted by the FEB, using off-the-shelf equipment as required by the project and for an operation period of at least 30 years.

We tested out-of-the-box interfaces and used original techniques to cope with these requirements, and set up a test bench with hundreds of synchronous Gigabit links in order to validate and tune the acquisition chain including downstream data logging based on zeroMQ and Google Protocol-Buffers.

Model-independent partial wave analysis using a massively-parallel fitting framework

**Author:** Liang Sun

**Co-authors:** Alberto Correa Dos Reis; Michael David Sokoloff; Rafael Aoude

1 Wuhan University (CN)
2 CBPF - Centro Brasileiro de Pesquisas Fisicas (CBPF)
3 University of Cincinnati (US)
4 CBPF - Centro Brasileiro de Pesquisas Fisicas (BR)

**Corresponding Authors:** michael.david.sokoloff@cern.ch, rafael.tourinho.jadallah.aoude@cern.ch, liang.sun@cern.ch, alberto@cbpf.br

Based on GooFit, a GPU-friendly framework for doing maximum-likelihood fits, we have developed a tool for extracting model-independent S-wave amplitudes from three-body decays such as $D^+ \rightarrow h^+\cdot h^++h^-$. A full amplitude analysis is done where the magnitudes and phases of the S-wave amplitudes (or alternatively, the real and imaginary components), are anchored at a finite number of $m^2(h^\cdot h^+)$, and a cubic spline is used to interpolate between these points. The amplitudes for P-wave and D-wave resonant states are modeled as spin-dependent Breit-Wigners. GooFit uses the Thrust library to launch all kernels, with a CUDA back-end for nVidia GPUs and an OpenMP back-end for compute nodes with conventional CPUs. Performance on a variety of these platforms is compared. Execution time on systems with GPUs is a few hundred times faster than running the same algorithm on a single CPU.
CERN openlab Researched Technologies That Might Become Game Changers in Software Development

Author: Fons Rademakers

Corresponding Author: fons.rademakers@cern.ch

CERN openlab is a unique public-private partnership between CERN and leading IT companies and research institutes. Several of the CERN openlab projects investigate technologies that have the potential to become game changers in HEP software development (like Intel Xeon-FPGA, Intel 3DX-point memory, Micron Automata Processor, etc.). In this presentation I will highlight a number of these technologies in detail and describe in what way they might change current software development techniques and practices.

Primary Keyword (Mandatory): Software development process and tools

Secondary Keyword (Optional): Algorithms

Tertiary Keyword (Optional):

CERN openlab Knowledge Transfer and Innovation Projects

Author: Fons Rademakers

Corresponding Author: fons.rademakers@cern.ch

CERN openlab is a unique public-private partnership between CERN and leading IT companies and research institutes. Having learned a lot from the close collaboration with industry in many different projects we now are using this experience to transfer some of our knowledge to other scientific fields, specifically in the areas of code optimization for the simulations of biological dynamics and the advanced usage of ROOT for the storage and processing of genomics data. In this presentation I will give an overview of the knowledge transfer projects we are currently engaged in. How they are relevant and beneficial for all parties involved, the interesting technologies that are being developed and what the potential and exciting results will be.

Tertiary Keyword (Optional): Outreach

Primary Keyword (Mandatory): Experience/plans from outside experimental HEP/NP
CERN data services for LHC computing

Author: Xavier Espinal Curull¹

¹ CERN

Corresponding Author: xavier.espinal@cern.ch

Dependability, resilience, adaptability, and efficiency. Growing requirements require tailoring storage services and novel solutions. Unprecedented volumes of data coming from the detectors need to be quickly available in a highly scalable way for large-scale processing and data distribution while in parallel they are routed to tape for long-term archival. These activities are critical for the success of HEP experiments. Nowadays we operate at high incoming throughput (14GB/s during 2015 LHC Pb-Pb run) and with concurrent complex production work-loads. In parallel our systems provide the platform for the continuous user and experiment driven work-loads for large-scale data analysis, including end-user access and sharing. The storage services at CERN cover the needs of our community: EOS and CASTOR as a large-scale storage; CERNBox for end-user access and sharing; Ceph as data back-end for the CERN OpenStack infrastructure, NFS services and S3 functionality; AFS for legacy distributed-file-system services. In this paper we will summarise the experience in supporting LHC experiments and the transition of our infrastructure from static monolithic systems to flexible components providing a more coherent environment with pluggable protocols, tunable QoS, sharing capabilities and fine grained ACLs management while continuing to guarantee the dependable and robust services.

CERN AFS Replacement project

Author: Jan Iven¹

¹ CERN

Co-authors: Alberto Pace ¹; Massimo Lamanna ¹

Corresponding Authors: massimo.lamanna@cern.ch, jan.iven@cern.ch, alberto.pace@cern.ch

OpenAFS is the legacy solution for a variety of use cases at CERN, most notably home-directory services. OpenAFS has been used as the primary shared file-system for Linux (and other) clients for more than 20 years, but despite an excellent track record the project’s age and architectural limitations are becoming more evident. We are now working to offer an alternative solution based on existing CERN storage services. The new solution will offer evolved functionality while reducing risk factors compared to the present status, and is expected to eventually benefit from operational synergies. In this paper we will present CERN’s usage and an analysis of our technical choices: we
will focus on the alternatives chosen for the various use cases (among them EOS, CERNBox, CASTOR); on implementing the migration process over the coming years; and the challenges expected to come up during the migration.

Primary Keyword (Mandatory):
Storage systems

Secondary Keyword (Optional):
Preservation of analysis and data

Tertiary Keyword (Optional):

Track 3: Distributed Computing / 68

From Physics to industry: EOS outside HEP

Author: Xavier Espinal Curull

1 CERN

Corresponding Author: xavier.espinal@cern.ch

In the competitive ‘market’ for large-scale storage solutions, EOS has been showing its excellence in the multi-Petabyte high-concurrency regime. It has also shown a disruptive potential in powering the CERNBox service in providing sync&share capabilities and in supporting innovative analysis environments along the storage of LHC data. EOS has also generated interest as generic storage solution ranging from university systems to very large installations for non-HEP applications. While preserving EOS as an open software solution for our community, we teamed up with the Comtrade company (within the CERN OpenLab framework) to productise this HEP contribution to ease its adoption by interested parties, notably outside HEP. In this paper we will deliver a status report of this collaboration and of EOS adoption by other institutes.

Primary Keyword (Mandatory):
Experience/plans from outside experimental HEP/NP

Secondary Keyword (Optional):
Outreach

Tertiary Keyword (Optional):

Track 5: Software Development / 69

Big Data Analytics for the Future Circular Collider Reliability and Availability Studies

Authors: Andrea Apollonio1; Antonio Romero Marin1; Arto Niemi1; Elena Rogova2; Johannes Gutleber1; Jussi-Pekka Penttinen1; Manuel Martin Marquez2; Peter Sollander1; Volodimir Begy4

1 CERN
2 Technische Universität Delft (NL)
3 Ramentor Oy
4 University of Vienna (AT)
European Strategy for Particle Physics update 2013, the study explores different designs of circular colliders for the post-LHC era. Reaching unprecedented energies and luminosities require to understand system reliability behaviour from the concept phase onwards and to design for availability and sustainable operation. The study explores industrial approaches to model and simulate the reliability and availability the entire particle collider complex. Estimates are based on an in-depth study of the CERN injector chain and LHC collider and are carried out as a cooperative effort with the HL-LHC project. The work so far has revealed that a major challenge is obtaining accelerator monitoring and operation data with sufficient quality, to automate the data quality annotation and calculation of reliability distribution functions for systems, subsystems and components where needed. A flexible data management and analytics environment that permits integrating the heterogenous data sources, the domain-specific data quality management algorithms and the reliability modelling and simulation suite is a key enabler to complete this accelerator operation study. This paper describes the Big Data infrastructure and analytics ecosystem that has been put in operation at CERN, serving as the foundation on which reliability and availability analysis and simulations can be built. This contribution focuses on data infrastructure and data management aspects and gives practical data analytics examples.

Tertiary Keyword (Optional):
Visualization

Primary Keyword (Mandatory):
Analysis tools and techniques

Secondary Keyword (Optional):
Control systems

CERNBox: the data hub for data analysis

Author: Jakub Moscicki\textsuperscript{1}

\textsuperscript{1} CERN

Corresponding Author: jakub.moscicki@cern.ch

A new approach to providing scientific computing services is currently investigated at CERN. It combines solid existing components and services (EOS Storage, CERNBox Cloud Sync&Share layer, ROOT Analysis Framework) with rising new technologies (Jupyter Notebooks) to create a unique environment for Interactive Data Science, Scientific Computing and Education Applications. EOS is the main disk storage system handling LHC data in the 100PB range. CERNBox offers a convenient sync&share layer and it is available everywhere: web, desktop and mobile. The Jupyter Notebook is a web application that allows users to create and share documents that contain live code, equations, visualizations and explanatory text. ROOT is a modular scientific software framework which provides the functionality to deal with big data processing, statistical analysis, visualisation and storage.

The system will be integrated in all major work-flows for scientific computing and with existing scientific data repositories at CERN. File access will be provided using a range of access protocols and tools: physics data analysis applications access CERNBox via xrootd protocol; Jupyter Notebooks interact with the storage via file-system interfaces provided by EOS fuse mounts; Grid jobs use webdav access authenticated with Grid certificates whereas batch jobs may use local Krb5 credentials for authentication. We report on early experience with this technology and applicable use-cases, also in a broader scientific and research context.
Tertiary Keyword (Optional):
Outreach

Primary Keyword (Mandatory):
Cloud technologies

Secondary Keyword (Optional):
Storage systems

Posters B / Break / 71

Tape SCSI monitoring and encryption at CERN

Author: Stefanos Laskaridis¹
Co-authors: Daniele Francesco Kruse ²; Eric Cano ²; German Cancio Melia ²; Julien Leduc ²; Steven Murray ²; Vlado Bahyl ²

¹ Aristotle Univ. of Thessaloniki (GR)
² CERN

Corresponding Authors: stefanos.laskaridis@cern.ch, julien.leduc@cern.ch, vladimir.bahyl@cern.ch, steven.murray@cern.ch, eric.cano@cern.ch, german.cancio.melia@cern.ch, daniele.francesco.kruse@cern.ch

CERN currently manages the largest data archive in the HEP domain; over 135PB of custodial data is archived across 7 enterprise tape libraries containing more than 20,000 tapes and using over 80 tape drives. Archival storage at this scale requires a leading edge monitoring infrastructure that acquires live and lifelong metrics from the hardware in order to assess and proactively identify potential drive and media level issues. In addition, protecting the privacy of sensitive archival data is becoming increasingly important and with it the need for a scalable, compute-efficient and cost-effective solution for data encryption.

In this paper we first describe the implementation of acquiring tape medium and drive related metrics reported by the SCSI interface and its integration with our monitoring system. We then address the incorporation of tape drive real-time encryption with dedicated drive hardware into the CASTOR hierarchical mass storage system.

Tertiary Keyword (Optional):
Security and policies

Primary Keyword (Mandatory):
Storage systems

Secondary Keyword (Optional):
Monitoring

Track 6: Infrastructures / 72

Wi-Fi service enhancement at CERN

Author: Vincent Ducret¹
Co-authors: Adam Wojciech Sosnowski ¹; Benito Gonzalez ¹
Over the last few years, the number of mobile devices connected to the CERN internal network has increased from a handful in 2006 to more than 10,000 in 2015. Wireless access is no longer a "nice to have" or just for conference and meeting rooms, now support for mobility is expected by most, if not all, of the CERN community. In this context, a full renewal of the CERN Wi-Fi network has been launched in order to provide a state-of-the-art Campus-wide Wi-Fi Infrastructure. Which technologies can provide an end-user experience comparable, for most applications, to a wired connection? Which solution can cover more than 200 office buildings, which represent a total surface of more than 400,000 m², while keeping a single, simple, flexible and open management platform? The presentation will focus on the studies and tests performed at CERN to address these issues, as well as some feedback about the global project organisation.

**Primary Keyword (Mandatory):**
Network systems and solutions

**Secondary Keyword (Optional):**
Collaborative tools

**Tertiary Keyword (Optional):**

---

**A graphical performance analysis and exploration tool for Linux-perf**

**Authors:** Aram Santogidis¹; Omar Awile¹

¹ CERN

Corresponding Authors: omar.awile@cern.ch, aram.santogidis@cern.ch

Application performance is often assessed using the Performance Monitoring Unit (PMU) capabilities present in modern processors. One popular tool that can read the PMU’s performance counters is the Linux-perf tool. pmu-tools is a toolkit built around Linux-perf that provides a more powerful interface to the different PMU events and give a more abstracted view of the events. Unfortunately pmu-tools report results only in text form or simple static graphs, limiting their usability.

We report on our efforts of developing a web-based front-end for pmu-tools allowing the application developer to more easily visualize, analyse and interpret performance monitoring results. Our contribution should boost programmer productivity and encourage continuous monitoring of the application’s performance. Furthermore, we discuss our tool’s capability to quickly construct and test new performance metrics for characterizing application performance. This will allow the user to experiment with new high level metrics that reflect the performance requirements of his application more accurately.

**Primary Keyword (Mandatory):**
Software development process and tools

**Secondary Keyword (Optional):**
Computing facilities

**Tertiary Keyword (Optional):**
Track 6: Infrastructures / 74

The role of dedicated computing centers in the age of cloud computing

Authors: Alexandr Zaytsev1; Christopher Hollowell2; Costin Caramarcu1; Tony Wong2; William Strecker-Kellogg3

1 Brookhaven National Laboratory (US)
2 Brookhaven National Laboratory
3 Brookhaven National Lab

Corresponding Authors:alezayt@bnl.gov, hollowec@bnl.gov, willsk@bnl.gov, tony@bnl.gov, costin.caramarcu@cern.ch

Brookhaven National Laboratory (BNL) anticipates significant growth in scientific programs with large computing and data storage needs in the near future and has recently re-organized support for scientific computing to meet these needs.

A key component is the enhanced role of the RHIC-ATLAS Computing Facility (RACF) in support of high-throughput and high-performance computing (HTC and HPC) at BNL.

This presentation discusses the evolving role of the RACF at BNL, in light of its growing portfolio of responsibilities and its increasing integration with cloud (academic and for-profit) computing activities.

We also discuss BNL’s plan to build a new computing center to support the new responsibilities of the RACF and present a summary of the cost benefit analysis done, including the types of computing activities that benefit most from a local data center vs. cloud computing. This analysis is partly based on an updated cost comparison of Amazon EC2 computing services and the RACF, which was originally conducted in 2012.

Tertiary Keyword (Optional):
Cloud technologies

Primary Keyword (Mandatory):
Computing facilities

Secondary Keyword (Optional):
High performance computing

Posters B / Break / 75

Next Generation Monitoring

Authors: John Bland1; Robert Fay2

Co-author: Stephen Jones3

1 University of Liverpool
2 University of Liverpool (GB)
3 Liverpool University

Corresponding Authors: j bland@hep.ph.liv.ac.uk, sjones@hep.ph.liv.ac.uk, robert.fay@cern.ch

Monitoring of IT infrastructure and services is essential to maximize availability and minimize disruption, by detecting failures and developing issues to allow rapid intervention.
The HEP group at Liverpool have been working on a project to modernize local monitoring infrastructure (previously provided using Nagios and ganglia) with the goal of increasing coverage, improving visualization capabilities, and streamlining configuration and maintenance. Here we discuss some of the tools evaluated, the different approaches they take, and how they can be combined to complement each other to form a comprehensive monitoring infrastructure. An overview of the resulting system and progress on implementation to date will be presented, which is currently as follows:

The system is configured with Puppet. Basic system checks are configured in Puppet using Hiera, and managed by Sensu. Centralised logging is managed with Elasticsearch, together with Logstash and Filebeat. Kibana provides an interface for interactive analysis, including visualization and dashboards. Metric collection is also configured in Puppet, with ganglia, Sensu, riemann.io, and collectd amongst the tools being considered. Metrics are sent to Graphite, with Grafana providing a visualization and dashboard tool. Additional checks on the collated logs and on metric trends are also configured in Puppet and managed by Sensu.

The Uchiwa dashboard for Sensu provides a web interface for viewing infrastructure status. Alert capabilities are provided via external handlers. Liverpool are developing a custom handler to provide an easily configurable, extensible and maintainable alert facility.

Primary Keyword (Mandatory):
Monitoring

Secondary Keyword (Optional):
Computing facilities

Tertiary Keyword (Optional):

Posters B / Break / 76

The merit of data processing application elasticity

Author: Vardan Gyurjyan

Co-authors: Bryan Muffit 2; Carl Timmer 3; David Abbott 1; Ed Jastrzembski 2; Graham Heyes 2; Ricardo Oyarzun 4; Sebastian Mancilla 5; William Gu 2

1 Jefferson Lab
2 JLAB
3 TJNAF
4 Universidad Tecnica Federico Santa Maria, Chile
5 Departamento de Fisica-Univ. Tecnica Federico Santa Maria (UTFSM)

Corresponding Authors: abbottd@jlab.org, timmer@jlab.org, royarzun@gmail.com, jastrzem@jlab.org, gurjyan@jlab.org, sebastian.mancilla@cern.ch, heyes@jlab.org, jgu@jlab.org, muffin@jlab.org

IO optimizations along with the vertical and horizontal elasticity of an application are essential to achieve data processing performance linear scalability. However to deploy these three critical concepts in a unified software environment presents a challenge and as a result most of the existing data processing frameworks rely on external solutions to address them. For example in a multicore environment we run multiple copies of an application to attain “synthetic” vertical scalability. We rely on complex batch processing systems (with tons of overhead) to imitate so-called horizontal scaling. IO optimizations are not addressed most of the time, because entire effort is spent to perform data processing algorithmic optimizations. Note that IO and algorithmic optimizations by nature are very different and are difficult to address them simultaneously in a tightly coupled software environment. In this paper we present CLAS12 reconstruction and analyses (CLARA) framework based data processing application design experiences and results. CLARA is a real-time data stream-processing framework, that implements service-oriented architecture (SOA) in a flow based programming (FBP) paradigm. The choice of a paradigm with conjunction...
of a publish-subscribe message-passing middleware (MPM) allows integrating above-mentioned critical requirements in a unified software framework. CLARA presents an environment for developing agile, elastic, multilingual data processing applications, presenting solutions, capable of processing large volumes of distributed data interactively.

Tertiary Keyword (Optional):
Distributed data handling

Secondary Keyword (Optional):
Parallelization

Primary Keyword (Mandatory):
Software development process and tools

Track 5: Software Development / 77

Identifying memory allocation patterns in HEP software

Authors: Nathalie Rauschmayr\textsuperscript{1}; Sami Kama\textsuperscript{2}

\textsuperscript{1} CERN
\textsuperscript{2} Southern Methodist University (US)

Corresponding Authors: nathalie.rauschmayr@cern.ch, sami.kama@cern.ch

HEP applications perform an excessive amount of allocations/deallocations within short time intervals which results in memory churn, poor locality and performance degradation. These issues are already known for a decade, but due to the complexity of software frameworks and the large amount of allocations (which are in the order of billions for a single job), up until recently no efficient mechanism has been available to correlate these issues with source code lines. However, with the advent of the Big Data era, many tools and platforms are available nowadays in order to do memory profiling at large scale. Therefore, a prototype program has been developed to track and identify each single de-/allocation. The CERN IT Hadoop cluster is used to compute memory key metrics, like locality, variation, lifetime and density of allocations. The prototype further provides a web based visualization backend that allows the user to explore the results generated on the Hadoop cluster. Plotting these metrics for each single allocation over time gives new insight into application’s memory handling. For instance, it shows which algorithms cause which kind of memory allocation patterns, which function flow causes how many shortlived objects, what are the most commonly allocated sizes etc. The paper will give an insight into the prototype and will show profiling examples for LHC reconstruction, digitization and simulation jobs.

Primary Keyword (Mandatory):
Software development process and tools

Secondary Keyword (Optional):
Analysis tools and techniques

Tertiary Keyword (Optional):

Posters B / Break / 78

Finding unused memory allocations with FOM-tools
Authors: Nathalie Rauschmayr\textsuperscript{1}; Sami Kama\textsuperscript{2}

\textsuperscript{1} CERN
\textsuperscript{2} Southern Methodist University (US)

Corresponding Authors: nathalie.rauschmayr@cern.ch, sami.kama@cern.ch

Memory has become a critical parameter for many HEP applications and as a consequence some experiments had already to move from single- to multicore jobs. However in the case of LHC experiment software, benchmark studies have shown that many applications are able to run with a much lower memory footprint than what is actually allocated. In certain cases even half of the allocated memory being swapped out does not result in any runtime penalty. As a consequence many allocated objects are kept much longer in memory than needed and remain therefore unused. In order to identify and quantify such unused (obsolete) memory, FOM-tools has been developed. The paper presents the functionalities of the tool and shows concrete examples on how FOM-tools helped to remove unused memory allocations in HEP software.

Primary Keyword (Mandatory):
Software development process and tools

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Posters A / Break / 79

SDN-NGenIA A Software Defined Next Generation integrated Architecture for HEP and Data Intensive Science

Author: Harvey Newman\textsuperscript{1}

Co-author: Dorian Kcira\textsuperscript{2}

\textsuperscript{1} Caltech
\textsuperscript{2} California Institute of Technology (US)

Corresponding Authors: harvey.newman@cern.ch, dorian.kcira@cern.ch

The SDN Next Generation Integrated Architecture (SDN-NGeNIA) program addresses some of the key challenges facing the present and next generations of science programs in HEP, astrophysics, and other fields whose potential discoveries depend on their ability to distribute, process and analyze globally distributed petascale to exascale datasets. The SDN-NGenIA system under development by the Caltech and partner HEP and network teams is focused on the coordinated use of network, computing and storage infrastructures, through a set of developments that build on the experience gained in recently completed and previous projects that use dynamic circuits with bandwidth guarantees to support major network flows, as demonstrated across LHCONE and in large scale demonstrations over the last three years, and recently integrated with CMS' PhEDEx and ASO data management applications. The SDN-NGenIA development cycle is designed to progress from the scale required at LHC Run2 (0.3 to 1 exabyte under management and 100 Gbps networks) to the 50-100 Exabyte datasets and 0.8-1.2 terabit/sec networks required by the HL LHC and programs such as the SKA and the Joint Genomics Institute within the next decade. Elements of the system include (1) Software Defined Network (SDN) controllers and adaptive methods that flexibly allocate bandwidth and load balance multiple large flows over diverse paths spanning multi-domain networks, (2) high throughput transfer methods (FDT, RDMA) and data storage and transfer nodes (DTNs) designed to support smooth flows of 100 Gbps and up, (3) pervasive agent-based real-time monitoring services (in the MonALISA framework) that support coordinated operations among the SDN controllers, and help triggering re-allocation and load-balancing operations where needed, (4) SDN transfer optimization services...
developed by the teams in the context of OpenDaylight, and (5) machine learning coupled to prototype system modeling, to identify the key variables and optimize the overall throughput of the system, and (6) a “consistent operations” paradigm that limits the flows of the major science programs to a level compatible with the capacity of the campus, regional and wide area networks, and with other network usage.

In addition to general program goal of supporting the network needs of the LHC and other science programs with similar needs, a recent focus is the use of the Leadership HPC facility at Argonne National Lab (ALCF) for data intensive applications. This includes installation of state of the art DTNs at the site edge, specific SDN-NGenIA applications and the development of prototypical services aimed at securely transporting, processing and returning data “chunks” on an appropriate scale between the ALCF, and LHC Tier1 and Tier2 sites: from tens of terabytes now, to hundreds of petabytes using 400G links by 2019, and a petabyte at a time using terabit/sec links when the first exaflop HPC systems are installed circa 2023.

**Tertiary Keyword (Optional):**
Data processing workflows and frameworks/pipelines

**Primary Keyword (Mandatory):**
Network systems and solutions

**Secondary Keyword (Optional):**
Distributed data handling

**Track 5: Software Development / 80**

**A Comparison of Deep Learning Architectures with GPU Acceleration and Their Applications**

**Author:** Jin Huang¹

**Co-authors:** Daicui Zhou ²; Jianlin Zhu ³

¹ Wuhan Textile University (CN)
² Central China Normal University CCNU (CN)
³ South-central University For Nationalities (CN)

**Corresponding Authors:** derick0320@gmail.com, daicui.zhou@cern.ch, jianlin.zhu@cern.ch

The goal of the comparison is to summarize the state-of-the-art techniques of deep learning which is boosted with modern GPUs. Deep learning, which is also known as deep structured learning or hierarchical learning, is a branch of machine learning based on a set of algorithms that attempt to model high-level abstractions in data by using multiple processing layers composed of multiple non-linear transformations. Deep learning is part of a broader family of machine learning methods based on learning representations of data. The representations are inspired by advances in neuroscience and are loosely based on interpretation of information processing and communication patterns in a nervous system, such as neural coding which attempts to define a relationship between various stimuli and associated neuronal responses in the brain. In this paper, a brief history of deep learning research is discussed first. Then, different deep learning models such as deep neural networks, convolutional deep neural networks, deep belief networks and recurrent neural networks are analyzed to summarize major work reported in the deep learning literature. Then we discuss the general deep learning system architecture including hardware layer and software middleware. In this architecture GPU subsystem is widely used to accelerate computation and its architecture is especially discussed. To show the performance of the deep learning system with GPU acceleration, we choose various deep learning models, compare their performances with/without GPU and list various acceleration rates. Various deep learning models have been applied to fields like computer vision, automatic speech recognition, natural language processing, audio recognition and bioinformatics. Selected applications are reviewed to show state-of-the-art results on various tasks. Finally, future directions of deep learning are discussed.
Track 4: Data Handling / 81

**CERN’s Ceph infrastructure: OpenStack, NFS, CVMFS, CASTOR, and more!**

**Author:** Dan van der Ster

**Co-author:** Xavier Espinal Curull

1 CERN

**Corresponding Authors:** daniel.vanderster@cern.ch, xavier.espinal@cern.ch

This work will present the status of Ceph-related operations and development within the CERN IT Storage Group: we summarise significant production experience at the petabyte scale as well as strategic developments to integrate with our core storage services. As our primary back-end for OpenStack Cinder and Glance, Ceph has provided reliable storage to thousands of VMs for more than 3 years; this functionality is used by the full range of IT services and experiment applications.

Ceph at the LHC scale (above 10’s of PB) has required novel contributions both in the development and operational side. For this reason, we have performed scale testing in cooperation with the core Ceph team. This work has been incorporated into the latest Ceph releases and enables Ceph to operate with at least 7,200 OSDs (totaling 30 PB in our tests). CASTOR has been evolved with the possibility to use a Ceph cluster as extensible high-performance data pool. The main advantages of this solution are the drastic reduction of the operational load and the possibility to deliver high single-stream performances to efficiently drive the CASTOR tape infrastructure. Ceph is currently our laboratory to explore S3 usage in HEP and to evolve other infrastructure services.

In this paper, we will highlight our Ceph-based services, the NFS Filer and CVMFS, both of which use virtual machines and Ceph block devices at their core. We will then discuss the experience in running Ceph at LHC scale (most notably early results with Ceph-CASTOR).

Tertiary Keyword (Optional):
Cloud technologies

Primary Keyword (Mandatory):
Storage systems

Secondary Keyword (Optional):
Object stores
LHCb Kalman Filter cross architectures studies

Author: Daniel Hugo Campora Perez
Co-author: Cedric Potterat

1 Universidad de Sevilla (ES)
2 Univ. Federal do Rio de Janeiro (BR)

Corresponding Authors: daniel.hugo.camopora.perez@cern.ch, cedric.potterat@cern.ch

The 2020 upgrade of the LHCb detector will vastly increase the rate of collisions the Online system needs to process in software, in order to filter events in real time. 30 million collisions per second will pass through a selection chain, where each step is executed conditional to its prior acceptance.

The Kalman Filter is a fit applied to all reconstructed tracks which, due to its time characteristics and early execution in the selection chain, consumes 40% of the whole reconstruction time in the current detector software trigger. This fact makes it a critical item as the LHCb trigger evolves into a full software trigger in the Upgrade.

We present acceleration studies for the Kalman Filter process, and optimize its execution for a variety of architectures, including x86_64 and Power8 architectures, and accelerators such as the Intel Xeon Phi and NVIDIA GPUs. We compare inter-architecture results, factoring in data moving operations and power consumption.

Primary Keyword (Mandatory): Algorithms
Secondary Keyword (Optional): Reconstruction
Tertiary Keyword (Optional):

Track 7: Middleware, Monitoring and Accounting / 83

Unified Monitoring Architecture for IT and Grid Services

Author: Alberto Aimar

Co-authors: Borja Garrido Bear; Daniel Zolnai; Edward Karavakis; Hassen Riahi; Javier Rodriguez Martinez; Luca Magnoni; Maria-Varvara Georgiou; Pablo Saiz; Pedro Andrade; Rocio Rama Ballesteros; Sergey Belov

1 CERN
2 Universidad de Oviedo (ES)
3 Budapest University of Technology and Economics (HU)
4 Athens University of Economics and Business (GR)
5 Joint Inst. for Nuclear Research (RU)

Corresponding Authors: pablo.saiz@cern.ch, pedro.andrade@cern.ch, hassen.riahi@cern.ch, alberto.aimar@cern.ch, sergey.belov@cern.ch, georgiou.mpv@cern.ch, javier.rodriguez.martinez@cern.ch, borja.garrido.bear@cern.ch, luca.magnoni@cern.ch, daniel.zolnai@cern.ch, rocio.rama.ballesteros@cern.ch, edward.karavakis@cern.ch

For over a decade, LHC experiments have been relying on advanced and specialized WLCG dashboards for monitoring, visualizing and reporting the status and progress of the job execution, data management transfers and sites availability across the WLCG distributed grid resources.
In the recent years, in order to cope with the increase of volume and variety of the grid resources, the WLCG monitoring had started to evolve towards data analytics technologies such as ElasticSearch, Hadoop and Spark. Therefore, at the end of 2015, it was agreed to merge these WLCG monitoring services, resources and technologies with the internal CERN IT data centres monitoring services also based on the same solutions.

The overall mandate was to migrate, in concertation with representatives of the users of the LHC experiments, the WLCG monitoring to the same technologies used for the IT monitoring. It started by merging the two small IT and WLCG monitoring teams, in order to join forces to review, rethink and optimize the IT and WLCG monitoring and dashboards within a single common architecture, using the same technologies and workflows used by the CERN IT monitoring services.

This work, in early 2016, resulted in the definition and the development of a Unified Monitoring Architecture aiming at satisfying the requirements to collect, transport, store, search, process and visualize both IT and WLCG monitoring data. The newly-developed architecture, relying on state-of-the-art open source technologies and on open data formats, will provide solutions for visualization and reporting that can be extended or modified directly by the users according to their needs and their role. For instance it will be possible to create new dashboards for the shifters and new reports for the managers, or implement additional notifications and new data aggregations directly by the service managers, with the help of the monitoring support team but without any specific modification or development in the monitoring service.

This contribution provides an overview of the Unified Monitoring Architecture, currently based on technologies such as Flume, ElasticSearch, Hadoop, Spark, Kibana and Zeppelin, with insight and details on the lessons learned, and explaining the work done to monitor both the CERN IT data centres and the WLCG job, data transfers and sites and services.

Primary Keyword (Mandatory):
Monitoring

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

Track 1: Online Computing / 84

Development of DAQ Software for CULTASK Experiment

Author: Soohyung Lee

Corresponding Author: soohyunglee@ibs.re.kr

Axion is a candidate of dark matter and is believed to be a breakthrough of strong CP problem in QCD. CULTASK (CAPP Ultra-Low Temperature Axion Search in Korea) experiment is an axion search experiment which is being performed at Center for Axion and Precision Physics Research (CAPP), Institute for Basic Science (IBS) in Korea. Based on Sikivie’s method, CULTASK uses a resonant cavity to discover axion signal. To get higher axion conversion power and better signal sensitivity, quality factor of the cavity, system temperature, cavity volume, and external magnetic field are important. To maximize those factors, the experiment is in research and development stage to date.

As a part of the research and development, CULDAQ, the DAQ software for CULTASK, is being developed. It controls related equipments such as network analyzer, signal analyzer, Piezo actuator, and so on through various interfaces like GPIB and USB. It also acquires data from those devices, reprocess it to convenient format, and store them to storage. The lowest layer of the software is written in C++, and higher level is written in Python, so the run sequence can be defined in runtime.
Online monitoring is also a feature of it. Online variables are all stored and shared in a database, and the monitoring and controlling are available by using an usual web browser. For this user interface part, PHP, HTML5, and jquery are employed. It is implemented as Model-View-Controller (MVC) scheme to make the framework much flexible and responsive.

In this presentation, the details of CULDAQ will be introduced, and real running experiences in engineering runs are shared as well.


The new ATLAS Fast Calorimeter Simulation

Author: Jana Schaarschmidt

Weizmann Institute of Science (IL)

Corresponding Author: jana.schaarschmidt@cern.ch

Many physics and performance studies with the ATLAS detector at the Large Hadron Collider require very large samples of simulated events, and producing these using the full GEANT4 detector simulation is highly CPU intensive. Often, a very detailed detector simulation is not needed, and in these cases fast simulation tools can be used to reduce the calorimeter simulation time by a few orders of magnitude.

The new ATLAS Fast Calorimeter Simulation (FastCaloSim) is an improved parametrisation compared to the one used in the LHC Run-1. It provides a simulation of the particle energy response at the calorimeter read-out cell level, taking into account the detailed particle shower shapes and the correlations between the energy depositions in the various calorimeter layers. It is interfaced to the standard ATLAS digitization and reconstruction software, and can be tuned to data more easily than with GEANT4. The new FastCaloSim incorporates developments in geometry and physics lists of the last five years and benefits from knowledge acquired with the Run-1 data. It makes use of statistical techniques such as principal component analysis, and a neural network parametrisation to optimise the amount of information to store in the ATLAS simulation infrastructure. It is planned to use this new FastCaloSim parameterization to simulate several billion events in the upcoming LHC runs. In this talk, we will describe this new FastCaloSim parametrisation.

Primary Keyword (Mandatory):
Simulation
ATLAS computing on Swiss Cloud SWITCHengines

Author: Sigve Haug
Co-author: Francesco Giovanni Sciacca

1 Universitaet Bern (CH)

Corresponding Authors: sigve.haug@lhep.unibe.ch, gsciacca@mail.cern.ch

Consolidation towards more computing at flat budgets beyond what pure chip technology can offer, is a requirement for the full scientific exploitation of the future data from the Large Hadron Collider. One consolidation measure is to exploit cloud infrastructures whenever they are financially competitive. We report on the technical solutions and the performance used and achieved running ATLAS production on SWITCHengines. SWITCHengines is the new cloud infrastructure offered to Swiss academia by the National Research and Education Network SWITCH. While solutions and performances are general, financial considerations and policies, which we also report on, are country specific.

Global EOS: exploring the 300-ms-latency region

Author: Luca Mascetti

1 CERN

Corresponding Author: luca.mascetti@cern.ch

EOS, the CERN open-source distributed disk storage system, provides the high-performance storage solution for HEP analysis and the back-end for various work-flows. Recently EOS became the back-end of CERNBox, the cloud synchronisation service for CERN users.

EOS can be used to take advantage of wide-area distributed installations: for the last few years CERN EOS uses a common deployment across two computer centres (Geneva-Meyrin and Budapest-Wigner) about 1,000 km apart (20-ms latency) with about 200 PB of disk (JBOD). In late 2015, the CERN-IT Storage group and AARNET (Australia) set-up a challenging R&D project: a single EOS instance between CERN and AARNET with more than 300 ms latency (16,500 km apart).
This paper will report about the success in deploy and run a distributed storage system between Europe (Geneva, Budapest), Australia (Melbourne) and later in Asia (ASGC Taipei), allowing different type of data placement and data access across these four sites.

**Primary Keyword (Mandatory):**
Storage systems

**Tertiary Keyword (Optional):**

**Secondary Keyword (Optional):**

**Track 6: Infrastructures / 88**

**ATLAS and LHC computing on CRAY**

**Author:** Sigve Haug¹

**Co-author:** Francesco Giovanni Sciaccia ³

¹ Universitaet Bern (CH)

**Corresponding Authors:** sigve.haug@lhep.unibe.ch, gsciaccia@mail.cern.ch

Access and exploitation of large scale computing resources, such as those offered by general purpose HPC centres, is one import measure for ATLAS and the other Large Hadron Collider experiments in order to meet the challenge posed by the full exploitation of the future data within the constraints of flat budgets.

We report on the effort moving the Swiss WLCG T2 computing, serving ATLAS, CMS and LHCb, from a dedicated cluster to large CRAY systems at the Swiss national supercomputing centre, CSCS. These systems do not only offer very efficient hardware, cooling and highly competent operators, but also have large backfill potentials due to size and multidisciplinary usage, and potential gains due to economy of scale.

Technical solutions, performance, expected return and future plans are discussed.

**Tertiary Keyword (Optional):**
Computing facilities

**Primary Keyword (Mandatory):**
High performance computing

**Secondary Keyword (Optional):**
Computing middleware

**Posters A / Break / 89**

**Analysis of empty ATLAS pilot jobs and subsequent resource usage on grid sites**

**Authors:** Alessandra Forti¹; Peter Love²

¹ University of Manchester (GB)
Lancaster University (GB)

Corresponding Authors: peter.love@cern.ch, alessandra.forti@cern.ch

The pilot model employed by the ATLAS production system has been in use for many years. The model has proven to be a success, with many advantages over push models. However one of the negative side-effects of using a pilot model is the presence of ‘empty pilots’ running on sites, consuming a small amount of walltime and not running a useful payload job. The impact on a site can be significant, with previous studies showing a total 0.5% walltime usage with no benefit to either the site or to ATLAS. Another impact is the number of empty pilots being processed by a site’s Compute Element and batch system, which can be 5% of the total number of pilots being handled.

In this paper we review the latest statistics using both ATLAS and site data and highlight edge cases where the number of empty pilots dominate. We also study the effect of tuning the pilot factories to reduce the number of empty pilots.

Primary Keyword (Mandatory):
Computing middleware

Secondary Keyword (Optional):
Computing facilities

Tertiary Keyword (Optional):

Track 9: Future Directions / 90

The ATLAS computing challenge for HL-LHC

Author: Simone Campana

CERN

Corresponding Author: simone.campana@cern.ch

The ATLAS experiment successfully commissioned a software and computing infrastructure to support the physics program during LHC Run 2. The next phases of the accelerator upgrade will present new challenges in the offline area. In particular, at High Luminosity LHC (also known as Run 4) the data taking conditions will be very demanding in terms of computing resources: between 5 and 10 KHz of event rate from the HLT to be reconstructed (and possibly further reprocessed) with an average pile-up of up to 200 events per collision and an equivalent number of simulated samples to be produced. The same parameters for the current run are lower by up to an order of magnitude.

While processing and storage resources would need to scale accordingly, the funding situation allows one at best to consider a flat budget over the next few years for offline computing needs. In this paper we present a study quantifying the challenge in terms of computing resources for HL-LHC and present ideas about the possible evolution of the ATLAS computing model, the distributed computing tools, and the offline software to cope with such a challenge.

Primary Keyword (Mandatory):
Computing models
Networks in ATLAS

Networks have played a critical role in high-energy physics (HEP), enabling us to access and effectively utilize globally distributed resources to meet the needs of our physicists.

Because of their importance in enabling our grid computing infrastructure many physicists have taken leading roles in research and education (R&E) networking, participating in, and even convening, network related meetings and research programs with the broader networking community worldwide. This has led to HEP benefiting from excellent global networking capabilities for little to no direct cost. However, as other science domains ramp-up their need for similar networking it becomes less clear that this situation will continue unchanged.

What this means for ATLAS in particular needs to be understood. ATLAS has evolved its computing model since the LHC started based upon its experience with using globally distributed resources. The most significant theme of those changes has been increased reliance upon, and use of, its networks.

We will report on a number of networking initiatives in ATLAS including the integration of network awareness into PANDA, the changes in our DDM system to allow remote access to data and participation in the global perfSONAR network monitoring efforts of WLCG.

We will also discuss new efforts underway that are exploring the inclusion and use of software defined networks (SDN) and how ATLAS might benefit from: Orchestration and optimization of distributed data access and data movement. Better control of workloads, end to end. Enabling prioritization of time-critical vs normal tasks Improvements in the efficiency of resource usage

Tertiary Keyword (Optional):
Distributed data handling

Primary Keyword (Mandatory):
Network systems and solutions

Secondary Keyword (Optional):
Distributed workload management

Track 4: Data Handling / 92

Production Experience with the ATLAS Event Service
The ATLAS Event Service (ES) has been designed and implemented for efficient running of ATLAS production workflows on a variety of computing platforms, ranging from conventional Grid sites to opportunistic, often short-lived resources, such as spot market commercial clouds, supercomputers and volunteer computing. The Event Service architecture allows real time delivery of fine grained workloads to running payload applications which process dispatched events or event ranges and immediately stream the outputs to highly scalable Object Stores. Thanks to its agile and flexible architecture the ES is currently being used by grid sites for assigning low priority workloads to otherwise idle computing resources; similarly harvesting HPC resources in an efficient back-fill mode; and massively scaling out to the 50-100k concurrent core level on the Amazon spot market to efficiently utilize those transient resources for peak production needs. Platform ports in development include ATLAS@Home (BOINC) and the Goggle Compute Engine, and a growing number of HPC platforms.

After briefly reviewing the concept and the architecture of the Event Service, we will report the status and experience gained in ES commissioning and production operations on various computing platforms, and our plans for extending ES application beyond Geant4 simulation to other workflows, such as reconstruction and data analysis.

Tertiary Keyword (Optional):
Distributed data handling

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

Secondary Keyword (Optional):
Distributed workload management
shift at the start of LHC Run2 in 2015. The main goal was to rely on a person with a good overview of the ADC activities to ease the ADC experts’ workload. The CRC shifter keeps track of ADC tasks related to their fields of expertise and responsibility. At the same time, the shifter maintains a global view of the day-to-day operations of the ADC system. During Run1, this task was accomplished by the ADC Manager on Duty (AMOD), a position that was removed during the shutdown period due to the reduced number and availability of ADC experts foreseen for Run2. The CRC position was proposed to cover some of the AMOD’s former functions, while allowing more people involved in computing to participate. In this way, CRC shifters help train future ADC experts. The CRC shifters coordinate daily ADC shift operations, including tracking open issues, reporting, and representing ADC in relevant meetings. The CRC also facilitates communication between the ADC experts team and the other ADC shifters. These include the Distributed Analysis Support Team (DAST), which is the first point of contact for addressing all distributed analysis questions, and the ATLAS Distributed Computing Shifters (ADCoS), which check and report problems in central services, Tier-0 export, data transfers and production tasks. Finally, the CRC looks at the level of ADC activities on a weekly or monthly timescale to ensure that ADC resources are used efficiently.

Tertiary Keyword (Optional):
Distributed data handling

Primary Keyword (Mandatory):
Monitoring

Secondary Keyword (Optional):
Distributed workload management

Consolidation of Cloud Computing in ATLAS

Author: Ryan Taylor

1 University of Victoria (CA)

Corresponding Author: rptaylor@uvic.ca

Throughout the first year of LHC Run 2, ATLAS Cloud Computing has undergone a period of consolidation, characterized by building upon previously established systems, with the aim of reducing operational effort, improving robustness, and reaching higher scale. This paper describes the current state of ATLAS Cloud Computing. Cloud activities are converging on a common contextualization approach for virtual machines, and cloud resources are sharing monitoring and service discovery components.

We describe the integration of Vac resources, streamlined usage of the High Level Trigger cloud for simulation and reconstruction, extreme scaling on Amazon EC2, and procurement of commercial cloud capacity in Europe. Building on the previously established monitoring infrastructure, we have deployed a real-time monitoring and alerting platform which coalesces data from multiple sources, provides flexible visualization via customizable dashboards, and issues alerts and carries out corrective actions in response to problems. Finally, a versatile analytics platform for data mining of log files is being used to analyze benchmark data and diagnose and gain insight on job errors.
In preparation for the XENON1T Dark Matter data acquisition, we have prototyped and implemented a new computing model. The XENON signal and data processing software is developed fully in Python 3, and makes extensive use of generic scientific data analysis libraries, such as the SciPy stack. A certain tension between modern “Big Data” solutions and existing HEP frameworks is typically experienced in smaller particle physics experiments. ROOT is still the “standard” data format in our field, defined by large experiments (ATLAS, CMS). To ease the transition, our computing model caters to both analysis paradigms, leaving the choice of using ROOT-specific C++ libraries, or alternatively, Python and its data analytics tools, as a front-end choice of developing physics algorithms. We present our path on harmonizing these two ecosystems, which allowed us to use off-the-shelf software libraries (e.g., NumPy, SciPy, scikit-learn, matplotlib) and lower the cost of development and maintenance.

To analyse the data, our software allows researchers to easily create “mini-trees”; small, tabular ROOT structures for Python analysis, which can be read directly into pandas DataFrame structures. One of our goals was making ROOT available as a cross-platform binary for an easy installation from the Anaconda Cloud (without going through the “dependency hell”). In addition to helping us discover dark matter interactions, lowering this barrier helps shift the particle physics toward non-domain-specific code.
Track 6: Infrastructures / 97

Advancing data management and analysis in different scientific disciplines

Authors: Andre Giesler\textsuperscript{None}, Christopher Jung\textsuperscript{None}, Fabian Rigoll\textsuperscript{1}, Jörg Meyer\textsuperscript{1}, Marcus Hardt\textsuperscript{2}, Martin Gasthuber\textsuperscript{3}, Rainer Stotzka\textsuperscript{1}, Sören Fleischer\textsuperscript{4}

Co-authors: Achim Streit\textsuperscript{5}, Max Fischer\textsuperscript{5}

\textsuperscript{1} Karlsruhe Institute of Technology
\textsuperscript{2} Karlsruhe Institute of Technology
\textsuperscript{3} DESY
\textsuperscript{4} GSI - Helmholtzzentrum für Schwerionenforschung
\textsuperscript{5} KIT - Karlsruhe Institute of Technology (DE)

Corresponding Authors: a.giesler@fz-juelich.de, s.fleischer@gsi.de, hardt@kit.edu, fabian.rigoll@kit.edu, rainer.stotzka@kit.edu, achim.streit@cern.ch, joerg.meyer2@kit.edu, christopher.jung@kit.edu, max.fischer@cern.ch

Over the past several years, rapid growth of data has affected many fields of science. This has often resulted in the need for overhauling or exchanging the tools and approaches in the disciplines’ data life cycles, allowing the application of new data analysis methods and facilitating improved data sharing.

The project Large-Scale Data Management and Analysis (LSDMA) of the German Helmholtz Association has been addressing both specific and generic requirements in its data life cycle successfully since 2012. Its data scientists work together with researchers from the fields such as climatology, energy and neuroscience to improve the community-specific data life cycles, in several cases even all stages of the data life cycle, i.e. from data acquisition to data archival. LSDMA scientists also study methods and tools that are of importance to many communities, e.g. data repositories and authentication and authorization Infrastructure.

In this presentation, we will discuss selected highlights of LSDMA’s research and development activities. Specifically, we will address how the results have advanced the user communities in their data-driven research. We will conclude with the lessons we have learned in the past few years.

Primary Keyword (Mandatory):
Experience/plans from outside experimental HEP/NP

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Track 7: Middleware, Monitoring and Accounting / 98

Interfacing HTCondor-CE with OpenStack

Author: Jose Caballero Bejar\textsuperscript{1}

Co-authors: Brian Paul Bockelman\textsuperscript{2}; John Hover\textsuperscript{3}

\textsuperscript{1} Brookhaven National Laboratory (US)
\textsuperscript{2} University of Nebraska (US)
\textsuperscript{3} Brookhaven National Laboratory (BNL)-Unknown-Unknown
Corresponding Authors: jose.caballero@cern.ch, brian.bockelman@cern.ch, john.hover@cern.ch

Over the past few years, Grid Computing technologies have reached a high level of maturity. One key aspect of this success has been the development and adoption of newer Compute Elements to interface the external Grid users with local batch systems. These new Compute Elements allow for better handling of jobs requirements and a more precise management of diverse local resources.

However, despite this level of maturity, the Grid Computing world is lacking diversity in local execution platforms. As Grid Computing technologies have historically been driven by the needs of the High Energy Physics community, most resource providers run the platform (operating system version and architecture) that best suits the needs of their particular users.

In parallel, the development of virtualization and cloud technologies has accelerated recently, making available a variety of solutions, both commercial and academic, proprietary and open source. Virtualization facilitates performing computational tasks on platforms not available at most computing sites.

This work attempts to join the technologies, allowing users to interact with computing sites through one of the standard Computing Elements, HTCondor-CE, but running their jobs within VMs on a local cloud platform, OpenStack, when needed.

The system will re-route, in a transparent way, end user jobs into dynamically-launched VM worker nodes when they have requirements that cannot be satisfied by the static local batch system nodes. Also, once the automated mechanisms are in place, it becomes straightforward to allow an end user to invoke a custom Virtual Machine at the site. This will allow cloud resources to be used without requiring the user to establish a separate account. Both scenarios are described in this work.

Tertiary Keyword (Optional):
Computing facilities

Primary Keyword (Mandatory):
Virtualization

Secondary Keyword (Optional):
Computing middleware

Track 5: Software Development / 100

PODIO - Applying plain-old-data for defining physics data models

Authors: Benedikt Hegner¹; Frank-Dieter Gaede²; Pere Mato Vila¹

¹ CERN
² Deutsches Elektronen-Synchrotron (DE)

Corresponding Authors: benedikt.hegner@cern.ch, pere.mato@cern.ch, frank-dieter.gaede@cern.ch

PODIO is a C++ library that supports the automatic creation and efficient handling of HEP event data, developed as a new EDM toolkit for future particle physics experiments in the context of the AIDA2020 EU programme. Event data models (EDMs) are at the core of every HEP experiment’s software framework, essential for providing a communication channel between different algorithms in the data processing chain as well as for efficient I/O. Experience from LHC and the Linear collider community shows that existing solutions partly suffer from overly complex data models with deep object-hierarchies or unfavourable I/O performance. The PODIO project was created in order to address these problems.
PODIO is based on the idea of employing plain-old-data (POD) data structures wherever possible, while avoiding deep object-hierarchies and virtual inheritance. At the same time it provides the necessary high-level interface towards the developer physicist, such as the support for inter-object relations, and automatic memory-management, as well as a ROOT-assisted Python interface. To simplify the creation of efficient data models, PODIO employs code generation from a simple yams-based markup language. In addition, it was developed with concurrency in mind in order to support the usage of modern CPU features, for example giving basic support for vectorisation techniques. This contribution presents the PODIO design, first experience in the context of the Future Circular Collider (FCC) and Liner Collider (LC) software projects, as well as performance figures when using ROOT as storage backend.

Primary Keyword (Mandatory):
Data model

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Posters A / Break / 102

Distributed Metadata Management of Mass Storage System in High Energy of Physics

Author: Qiulan Huang

Co-authors: Jingyan Shi; Yaodong Cheng

1 Chinese Academy of Sciences (CN)
2 IHEP

Corresponding Authors: qiulan.huang@cern.ch, chyd@ihep.ac.cn, jingyan.shi@cern.ch

The new generation of high energy physics (HEP) experiments have been producing gigantic data. How to store and access those data with high performance have been challenging the availability, scalability, and I/O performance of the underlying massive storage system. At the same time, a series of researches focusing on big data have been more and more active, and the research about metadata management is one of them. Metadata management is quite important to overall system performance in large-scale distributed storage systems, especially in the big data era. Metadata performance would produce a big effect on the scalability, availability and high performance of the massive storage system. In order to manage metadata effectively, so that data can be allocated and accessed efficiently, we design and implement a dynamic and scalable distributed metadata management system to HEP mass storage system.

In this contribution, the open source file system Gluster is reviewed and the architecture of the distributed metadata management system are introduced. Particularly, we discuss the key technologies of the distributed metadata management system and the way to optimize the metadata performance of Gluster file system by modifying the DHT(Distributed Hash Table) layer. We propose a new algorithm named Adaptive Directory Sub-tree Partition(ADSP) for metadata distribution. ADSP divides the filesystem namespace into sub-trees with directory granularity. Sub-trees will be stored on storage devices in flat structure, whose locality information and file attributes are recorded as extended attributes. The placement of sub-tree is adjusted adaptively according to the load of metadata cluster so that the load balance could be improved and metadata cluster could be extended dynamically. ADSP is an improved sub-tree partition algorithm with low computational complexity, also easy to be implemented. Experiments show that ADSP achieves higher metadata performance and scalability compared to Gluster and Lustre. The performance evaluation demonstrates the performance of metadata of Gluster file system is greatly improved. We also propose a new algorithm called Distributed Unified Layout(DULA) to improve dynamic scalability and efficiency of data positioning. A system with DULA could provide uniform data distribution and efficient data positioning. DULA is an improved consistent hashing algorithm which is able to locate data in O(1) without the help of...
routing information. Experiments prove that the better uniform data distribution and efficient data access can be achieved by DULA. This work is validated in YBJ experiment. In addition, three evaluation criteria of hash algorithm in massive storage system are presented. And a comparative analysis of legacy hash algorithms has been carried out in both theory and software simulation according to those criteria. The results of that provide the theoretical basis for the choice of hash algorithm of DULA.

Primary Keyword (Mandatory):
Storage systems

Secondary Keyword (Optional):
Distributed data handling

Tertiary Keyword (Optional):

Track 7: Middleware, Monitoring and Accounting / 103

SCEAPI: A Unified Restful Web APIs for High-Performance Computing

Author: Rongqiang Cao

Co-authors: Haili Xiao; Shasha Lu; Yining Zhao

1 Computer Network Information Center, Chinese Academy of Sciences

Corresponding Authors: haili@sccas.cn, caorq@sccas.cn, zhaoyin@sccas.cn,lusha721@sccas.cn

The development of scientific computing is increasingly moving to web and mobile applications. All these clients need high-quality implementations of accessing heterogeneous computing resources provided by clusters, grid computing or cloud computing. We present a web service called SCEAPI and describe how it can abstract away many details and complexities involved in the use of scientific computing and provide essential RESTful Web APIs including authentication, data transferring and job managements of creating, monitoring and scheduling jobs. Then we discuss how to build our computing environment that integrates computing resources from 15 HPC centers all over China, and how to add and encapsulate new applications into this computing environment so as to provide a unified way of using these applications in different high-performance clusters based on SCEAPI. Finally, use cases are given to show how SCEAPI works, including examples of installing the ATLAS Monte Carlo Simulation application and processing jobs submitted by the ARC Computing Element (ARC-CE) from CERN.

Primary Keyword (Mandatory):
Computing middleware

Secondary Keyword (Optional):
High performance computing

Tertiary Keyword (Optional):

Track 1: Online Computing / 104

Message Queues for Online Reconstruction on the Example of the PANDA Experiment

Author: Tobias Stockmanns

1
One of the large challenges of future particle physics experiments is the trend to run without a first level hardware trigger. The typical data rates exceed easily hundreds of GBytes/s, which is way too much to be stored permanently for an offline analysis. Therefore a strong data reduction has to be done by selection only those data, which is physically interesting. This implies that all detector data is read out and has to be processed with the same rate as it is produced. Several different hardware approaches from FPGAs, GPUs to multicore CPUs and mixtures of these systems are under study. Common to all of them is the need to process the data in massive parallel systems.

One very convenient way to realize parallel systems on CPUs is the usage of message queue based multiprocessing. One package that allow development of such application is the FairMQ module in the FairRoot simulation framework developed at GSI. FairRoot is used by several different experiments at and outside the GSI including the PANDA experiment. FairMQ is an abstract layer for message queue base application, it has two implementations: ZeroMQ and NanoMSG. For the PANDA experiment, FairMQ is under test in two different ways. On one hand side to online process test beam data of prototypes of sub-detectors of PANDA and, in a more generalized way, on time-based simulated data of the complete detector system. In the presentation results from both tests will be shown.

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

Secondary Keyword (Optional):
Reconstruction

Tertiary Keyword (Optional):

---

**XROOT development update - support for metalinks and extreme copy**

*Author:* Michal Kamil Simon

*Co-author:* Andrew Hanushevsky

1. **CERN**

2. **STANFORD LINEAR ACCELERATOR CENTER**

**Corresponding Authors:** michal.simon@cern.ch, abh@stanford.edu

XRootD is a distributed, scalable system for low-latency file access. It is the primary data access framework for the high-energy physics community. One of the latest developments in the project has been to incorporate metalink and segmented file transfer technologies.

We report on the implementation of the metalink metadata format support within XRootD client. This includes both the CLI and the API semantics. Moreover, we give an overview of the employed segmented file transfer mechanism that exploits metalink-based data sources. Its aim is to provide multisource file transmission (BitTorrent-like), which results in increased transfer rates.

This contribution summarizes these two development projects and presents the outcomes.

Tertiary Keyword (Optional):
Algorithms

Primary Keyword (Mandatory):
Storage systems

Secondary Keyword (Optional):
The DAQ system for the AEgIS experiment

Author: Francesco Prelz

Corresponding Author: francesco.prelz@mi.infn.it

In the sociology of small- to mid-sized (O(100) collaborators) experiments the issue of data collection and storage is sometimes felt as a residual problem for which well-established solutions are known. Still, the DAQ system can be one of the few forces that drive towards the integration of otherwise loosely coupled detector systems. As such it may be hard to complete with off-the-shelf components only.

LabVIEW and ROOT are the (only) two software systems that were assumed to be familiar enough to all collaborators of the AEgIS (AD6) experiment at CERN: working out of the GXML representation of LabVIEW Data types, a semantically equivalent representation as ROOT TTrees was developed for permanent storage and analysis. All data in the experiment is cast into this common format and can be produced and consumed on both systems and transferred over TCP and/or multicast over UDP for immediate sharing over the experiment LAN.

We describe the setup that has been able to cater to all run data logging and long term monitoring needs of the AEgIS experiment so far.

The Muon Ionization Cooling Experiment Analysis User Software

Author: Adam Dobbs

Co-author: Durga Rajaram

Corresponding Authors: a.dobbs07@imperial.ac.uk, durga@fnal.gov

The Muon Ionization Cooling Experiment (MICE) is a proof-of-principle experiment designed to demonstrate muon ionization cooling for the first time. MICE is currently on Step IV of its data taking programme, where transverse emittance reduction will be demonstrated. The MICE Analysis User Software (MAUS) is the reconstruction, simulation and analysis framework for the MICE experiment. MAUS is used for both offline data analysis and fast online data reconstruction and visualisation to serve MICE data taking.
This paper provides an introduction to the MAUS framework, describing the central Python and C++ based framework, code management procedure, the current performance for detector reconstruction and results from real data analysis of the recent MICE Step IV data. The ongoing development goals will also be described, including introducing multithreaded processing for the online detector reconstruction.

Tertiary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Primary Keyword (Mandatory):
Reconstruction

Secondary Keyword (Optional):
Analysis tools and techniques

The future of academic computing security

Author: Romain Wartel

1 CERN

Corresponding Author: roomain.wartel@cern.ch

This presentation offers an overview of the current security landscape - the threats, tools, techniques and procedures followed by attackers. These attackers range from cybercriminals aiming to make a profit, to nation-states searching for valuable information. Threat vectors have evolved in recent years; focus has shifted significantly, from targeting computer services directly, to aiming for the people managing the computational, financial and strategical resources instead. The academic community is at a crucial time and must proactively manage the resulting risks. Today, high quality threat intelligence is paramount, as it is the key means of responding and providing defendable computing services. Efforts are necessary, not only to obtain actionable intelligence, but also to process it, match it with traceability information such as network traffic and service logs, and to manage the findings appropriately. In order to achieve this, the community needs to take a three-fold approach: exploit its well-established international collaboration network; participate in vetted trust groups; further liaise with the private sector and law enforcement.

Tertiary Keyword (Optional):
Network systems and solutions

Primary Keyword (Mandatory):
Security and policies

Secondary Keyword (Optional):
Computing models

Optimizing ROOT’s Performance Using C++ Modules

Author: Vasil Georgiev Vasilev

1
ROOT version 6 comes with a C++ compliant interpreter cling. Cling needs to know everything about the code in libraries to be able to interact with them. This translates into increased memory usage with respect to previous versions of ROOT.

During the runtime automatic library loading process, ROOT6 re-parses a set of header files, which describe the library; and enters “recursive” parsing. The former has a noticeable effect on CPU and memory performance, whereas the latter is fragile and can introduce correctness issues. An elegant solution to the shortcoming is to feed the necessary information only when required and in a non-recursive way.

The LLVM community has started working on a powerful tool for reducing build times and peak memory usage of the clang compiler called “C++ Modules”. The feature matured and it is on its way to the C++ standard. C++ Modules are a flexible concept, which can be employed to match CMS and other experiments’ requirement for ROOT: to optimize both runtime memory usage and performance.

The implementation of the missing concepts in cling and its underlying LLVM libraries and adopting the changes in ROOT is a complex endeavor. I describe the scope of the work and I present a few techniques used to lower ROOT’s runtime memory footprint. I discuss the status of the C++ Modules in the context of ROOT and show some preliminary performance results.

Primary Keyword (Mandatory):
Analysis tools and techniques

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

Enabling Federated Access for HEP

Author: Hannah Short 1
Co-author: Romain Wartel 1

1 CERN

HEP has long been considered an exemplary field in Federated Computing; the benefit of this technology has been recognised by the thousands of researchers who have used the grid for nearly 15 years. Whilst the infrastructure is mature and highly successful, Federated Identity Management (FIM) is one area in which the HEP community should continue to evolve.

The ability for a researcher to use HEP resources with their existing account reflects the structure of a research team – team members continue to represent their own home organisation whilst collaborating. Through eduGAIN, the inter-federation service, an extensive suite of research services and a pool of international users have been unlocked within the scientific community. Establishing
adequate trust between these federation participants, as well as the relevant technologies, is the key to enable the effective adoption of FIM.

What is the current landscape of FIM for HEP? How do we see this landscape in the future? How do we get there? We will be addressing these questions in the context of CERN and the WLCG.

Tertiary Keyword (Optional):
Network systems and solutions

Primary Keyword (Mandatory):
Collaborative tools

Secondary Keyword (Optional):
Security and policies

Track 2: Offline Computing / 112

Full and Fast Simulation Framework for the Future Circular Collider Studies

Author: Anna Zaborowska

Warsaw University of Technology (PL)

Corresponding Author: anna.zaborowska@cern.ch

Software for the next generation of experiments at the Future Circular Collider (FCC), should by design efficiently exploit the available computing resources and therefore support of parallel execution is a particular requirement. The simulation package of the FCC Common Software Framework (FCCSW) makes use of the Gaudi parallel data processing framework and external packages commonly used in HEP simulation, including the Geant4 simulation toolkit, the DD4HEP geometry toolkit and the Delphes framework for simulating detector response.

Using Geant4 for Full simulation implies taking into account all physics processes for transporting the particles through detector material and this is highly CPU-intensive. At the early stage of detector design and for some physics studies such accuracy is not needed. Therefore, the overall response of the detector may be simulated in a parametric way. Geant4 provides the tools to define a parametrisation, which for the tracking detectors is performed by smearing the particle space-momentum coordinates and for calorimeters by reproducing the particle showers.

The parametrisation may come from either external sources, or from the Full simulation (being detector-dependent but also more accurate). The tracker resolutions may be derived from measurements of the existing detectors or from the external tools, for instance tkLayout, used in the CMS tracker performance studies. Regarding the calorimeters, the longitudinal and radial shower profiles can be parametrised using the GFlash library. The Geant4 Fast simulation can be applied to any type of particle in any region of the detector. The possibility to run both Full and Fast simulation in Geant4 creates a chance for an interplay, performing the CPU-consuming Full simulation only for the regions and particles of interest.

FCCSW also incorporates the Delphes framework for Fast simulation studies in a multipurpose detector. Phenomenological studies may be performed in an idealised geometry model, simulating the overall response of the detector. Having Delphes inside FCCSW allows users to create the analysis tools that may be used for Full simulation studies as well.

This presentation will show the status of the simulation package of the FCC common software framework.

Primary Keyword (Mandatory):
Simulation

Tertiary Keyword (Optional):
FPGA based data processing in the ALICE High-Level Trigger in LHC Run 2

Author: Heiko Engel
Co-author: Torsten Alt

1 Johann-Wolfgang-Goethe Univ. (DE)
2 FIAS - Frankfurt Institute for Advanced Studies

Corresponding Authors: hengel@cern.ch, torsten.alt@cern.ch

ALICE (A Large Ion Collider Experiment) is a detector system optimized for the study of heavy ion collision detector at the CERN LHC. The ALICE High Level Trigger (HLT) is a computing cluster dedicated to the online reconstruction, analysis and compression of experimental data. The High-Level Trigger receives detector data via serial optical links into custom PCI-Express based FPGA readout cards installed in the cluster machines. The readout cards optionally process the data on a per-link level already inside the FPGA and provide it to the host machines via Direct Memory Access (DMA). The HLT data transport framework collects the data from all machines and performs reconstruction, analysis and compression with CPUs and GPUs as a distributed application across the full cluster.

FPGA based data processing is enabled for the biggest detector of ALICE, the Time Projection Chamber (TPC). TPC raw data is processed in the FPGA with a hardware cluster finding algorithm that is faster than a software implementation and saves a significant amount of CPU resources in the HLT cluster. It also provides some data reduction while introducing only a marginal additional latency into the readout path. This algorithm is an essential part of the HLT already since LHC Run 1 for both proton and heavy ion runs. It was ported to the new HLT readout hardware for Run 2, was improved for higher link rates and adjusted to the recently upgraded TPC Readout Control Unit (RCU2). A flexible firmware implementation allows both the old and the new TPC data format and link rates to be handled transparently. Extended protocol and data error detection, error handling and the enhanced RCU2 data ordering scheme provide an improved physics performance of the cluster finder.

This contribution describes the integration of the FPGA based readout and processing into the HLT framework as well as the FPGA based TPC cluster finding and its adoption to the changed readout hardware during Run 2.
Events visualisation in ALICE - current status and strategy for Run 3

Author: Jeremi Niedziela¹
Co-author: Barthelemy Von Haller ²

¹ Warsaw University of Technology (PL)
² CERN

Corresponding Authors: jeremi.niedziela@cern.ch, barthelemy.von.haller@cern.ch

Events visualisation in ALICE - current status and strategy for Run 3

Jeremi Niedziela for the ALICE Collaboration

A Large Ion Collider Experiment (ALICE) is one of the four big experiments running at the Large Hadron Collider (LHC), which focuses on the study of the Quark-Gluon Plasma (QGP) being produced in heavy-ion collisions.

The ALICE Event Visualisation Environment (AliEVE) is a tool allowing to draw an interactive 3D model of the detector’s geometry and a graphical representation of the data. In addition, together with the online reconstruction module, it provides important quality monitoring of the recorded data. As a consequence it is in permanent use in the ALICE Run Control Center.

Total Event Display (TEV) is another visualisation tool recently developed by the CERN Media Lab and that can be used for all LHC experiments. It can be easily deployed on any platform, including web and mobile platforms. Integration with ALICE has already progressed well and will continue.

The current status of both solutions will be presented, as well as their numerous usages in ALICE. Finally an outlook of the strategy for Run 3 will be given showing how both AliEVE and TEV will be adapted to fit the ALICE O2 project.

Kalman filter tracking on parallel architectures

Authors: Avi Yagil¹; Daniel Sherman Riley²; Frank Wuerthwein¹; Giuseppe Cerati¹; Kevin McDermott²; Matevz Tadel¹; Matthieu Lefebvre³; Peter Elmer³; Peter Wittich²; Slava Krutelyov²; Steven R Lantz²
Limits on power dissipation have pushed CPUs to grow in parallel processing capabilities rather than clock rate, leading to the rise of “manycore” or GPU-like processors. In order to achieve the best performance, applications must be able to take full advantage of vector units across multiple cores, or some analogous arrangement on an accelerator card. Such parallel performance is becoming a critical requirement for methods to reconstruct the tracks of charged particles at the Large Hadron Collider and, in the future, at the High Luminosity LHC. This is because the steady increase in luminosity is causing an exponential growth in the overall event reconstruction time, and tracking is by far the most demanding task for both online and offline processing. Many past and present collider experiments adopted Kalman filter-based algorithms for tracking because of their robustness and their excellent physics performance, especially for solid state detectors where material interactions play a significant role. We report on the progress of our studies towards a Kalman filter track reconstruction algorithm with optimal performance on manycore architectures. The combinatorial structure of these algorithms is not immediately compatible with an efficient SIMD (or SIMT) implementation; the challenge for us is to recast the existing software so it can readily generate hundreds of shared-memory threads that exploit the underlying instruction set of modern processors. We show how the data and associated tasks can be organized in a way that is conducive to both multi-threading and vectorization. We demonstrate very good performance on Intel Xeon and Xeon Phi architectures, as well as promising first results on NVIDIA GPUs. We discuss the current limitations and the plan to achieve full scalability and efficiency in collision data processing.

Tertiary Keyword (Optional):
Algorithms

Secondary Keyword (Optional):
Parallelization

Primary Keyword (Mandatory):
Reconstruction

Performance of the CMS Event Builder

Author: Remi Mommsen

Co-authors: Andre Georg Holzer; Attila Racz; Ben Craigs; Christian Delldicque; Christoph Paus; Christoph Schwick; Cristian Contescu; Dainius Simelevicius; Dominique Gigi; Emilio Meschi; Frank Glege; Frans Meijers; Georgiana Lavinia Darlea; Guillelmo Gomez Ceballos Retuerto; Hannes Sakulin; James Gordon Branson; Jean-Marc Olivier Andre; Jeroen Hegeman; Jonathan Fulcher; Lorenzo Masetti; Luciano Orsini; Marc Dobson; Marco Pieri; Nicolas Doualot; Olivier Chaze; Petr Zejdl; Philipp Maximilian Brummer; Raul Jimenez Estupinan; Samim Erhan; Sergio Cittolin; Srecko Morovic; Thomas Reis; Ulf Behrens; Vivian O’Dell; Zeynep Demiragli

1 Fermi National Accelerator Lab. (US)
2 Univ. of California San Diego (US)
3 CERN
4 University of Northumbria (GB)
5 Massachusetts Inst. of Technology (US)
6 Vilnius University (LT)
The data acquisition system (DAQ) of the CMS experiment at the CERN Large Hadron Collider (LHC) assembles events at a rate of 100 kHz. It transports event data at an aggregate throughput of ~100 GB/s to the high-level trigger (HLT) farm. The CMS DAQ system has been completely rebuilt during the first long shutdown of the LHC in 2013/14. The new DAQ architecture is based on state-of-the-art network technologies for the event building. For the data concentration, 10/40 Gb/s Ethernet technologies are used together with a reduced TCP/IP protocol implemented in FPGA for a reliable transport between custom electronics and commercial computing hardware. A 56 Gb/s Infiniband FDR CLOS network has been chosen for the event builder. We report on the performance of the event builder system and the steps taken to exploit the full potential of the network technologies.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Network systems and solutions

Primary Keyword (Mandatory):

DAQ

An interactive and comprehensive working environment for high-energy physics software with Python and jupyter notebooks

Authors: Nils Braun¹; Thomas Hauth¹

Co-authors: Christian Pulvermacher ¹, Martin Ritter ²

¹ KIT
² LMU / Cluster Universe

Corresponding Authors: nils.braun@kit.edu, christian.pulvermacher@kit.edu, thomas.hauth@kit.edu, ritter@mpp.mpg.de

Today’s analyses for high energy physics experiments involve processing a large amount of data with highly specialized algorithms. The contemporary workflow from recorded data to final results is based on the execution of small scripts - often written in Python or ROOT macros which call complex compiled algorithms in the background - to perform fitting procedures and generate plots. During recent years interactive programming environments, such as jupyter, became popular. Jupyter allows to develop Python-based applications, so-called notebooks, which bundle code, documentation and results, e.g. plots. The advantages over classical script-based approaches is the feature to recompute only parts of the analysis code, which allows for fast and iterative development,
and a web-based user frontend, which can be hosted centrally and only requires a browser on the user side.

In our novel approach, Python and jupyter are tightly integrated into the Belle II Analysis Software Framework 2 (basf2), currently being developed for the Belle II experiment in Japan. This allows to develop code in jupyter notebooks for every aspect of the event simulation, reconstruction and analysis chain. Basf2 is based on software modules implemented in C++11 which have Python bindings created with Boost Python and PyROOT. These interactive notebooks can be hosted as a centralized web service via jupyterhub with docker and used by all scientists of the Belle II Collaboration. Because of its generality and encapsulation, the setup can easily be scaled to large installations.

This contribution will describe the technical implementations of the jupyter integration into basf2. The required code is generic enough to be easily applicable to other high energy physics frameworks and even to software from other research domains. The talk presents a full example of a jupyter-based analysis and some notebooks already successfully used in outreach and educational purposes.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Analysis tools and techniques

Primary Keyword (Mandatory):

Data processing workflows and frameworks/pipelines

Track 1: Online Computing / 118

SND DAQ system evolution

Author: Alexander Bogdanchikov¹

¹ Budker Institute of Nuclear Physics (RU)

Corresponding Author: alexander.bogdanchikov@cern.ch

The SND detector takes data at the e+e- collider VEPP-2000 in Novosibirsk. We present here recent upgrades of the SND DAQ system which are mainly aimed to handle the enhanced events rate load after the collider modernization. To maintain acceptable events selection quality the electronics throughput and computational power should be increased. These goals are achieved with the new fast (Flash ADC) digitizing electronics and distributed data taking. The data flow for the most congested detector subsystems is distributed and processed separately. We describe the new distributed SND DAQ software architecture, its computational and network infrastructure.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):

DAQ
Using ALFA for high throughput, distributed data transmission in ALICE O2 system

Author: Adam Tadeusz Wegrynek

Warsaw University of Technology (PL)

Corresponding Author: adam.wegrzynek@cern.ch

ALICE (A Large Ion Collider Experiment) is the heavy-ion detector designed to study the physics of strongly interacting matter and the quark-gluon plasma at the CERN LHC (Large Hadron Collider).

ALICE has been successfully collecting physics data of Run 2 since spring 2015. In parallel, preparations for a major upgrade of the computing system, called O2 (Online-Offline) and scheduled for the Long Shutdown 2 in 2019-2020, are being made. One of the major requirements is the capacity to transport data between so-called FLPs (First Level Processors), equipped with readout cards, and the EPNs (Event Processing Nodes), performing data aggregation, frame building and partial reconstruction. It is foreseen to have 268 FLPs dispatching data to 1500 EPNs with an average output of 20 Gb/s each. In overall, the O2 processing system will operate at terabits per second of throughput while handling millions of concurrent connections.

The ALFA framework will standardize and handle software related tasks such as readout, data transport, frame building, calibration, online reconstruction and more in the upgraded computing system.

ALFA supports two data transport libraries: ZeroMQ and nanomsg. This paper discusses the efficiency of ALFA in terms of high throughput data transport. The tests were performed using multiple FLPs, each of them pushing data to multiple EPNs. The transfer was done using push-pull communication pattern with multipart message support enable or disabled. The test setup was optimized for the benchmarks to get the most performant results for each hardware configuration. The paper presents the measurement process and final results – data throughput combined with computing resources usage as a function of block size, and in some cases as a function of time.

The high number of nodes and connections in the final set up may cause race conditions that can lead to uneven load balancing and poor scalability. The performed tests allow to validate whether the traffic is distributed evenly over all receivers. It also measures the behavior of the network in saturation and evaluates scalability from a 1-to-1 an N-to-N solution.

Design of the data quality control system for the ALICE O2

Authors: Barthelemy Von Haller; Jacek Tomasz Otwinowski; Patryk Lesiak

CERN
Corresponding Authors: barthelemy.von.haller@cern.ch, patryk.lesiakk@gmail.com, jacek.tomasz.otwinowski@cern.ch

ALICE (A Large Ion Collider Experiment) is the heavy-ion detector designed to study the physics of strongly interacting matter and the quark-gluon plasma at the CERN Large Hadron Collider (LHC). A major upgrade of the experiment is planned for 2020. In order to cope with a data rate 100 times higher and with the continuous readout of the Time Projection Chamber (TPC), it is necessary to upgrade the Online and Offline Computing to a new common system called O2.

The online Data Quality Monitoring (DQM) and the offline Quality Assurance (QA) are critical aspects of the data acquisition and reconstruction software chains. The former intends to provide shifters with precise and complete information to quickly identify and overcome problems while the latter aims at providing good quality data for physics analyses. DQM and QA typically involve the gathering of data, its distributed analysis by user-defined algorithms, the merging of the resulting objects and their visualization.

This paper discusses the architecture and the design of the data Quality Control (QC) system that regroups the DQM and QA in O2. In addition it presents the main design requirements and early results of a working prototype. A special focus is put on the merging of monitoring objects generated by the QC tasks. The merging is a crucial and challenging step of the O2 system, not only for QC but also for the calibration. Various scenarios and implementations have been made and large-scale tests carried out. This document presents the final results of this extensive work on merging.

We conclude with the plan of work for the coming years that will bring the QC to production by 2019.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Monitoring

Primary Keyword (Mandatory):
DAQ

Posters A / Break / 121

A performance study of WebDav access to storages within the Belle II collaboration.

Author: Silvio Pardi

Corresponding Author: spardi@na.infn.it

The use of Webdav protocol to access at large storage areas is becoming popular in the High Energy Physics community. All the main Grid and Cloud storage solutions provide such kind of interface, in this scenario, tuning the storage systems and performance evaluation became crucial aspects to promote the adoption of these protocols within the Belle II community.

In this work, we present the results of a large-scale test activity, made with the goal to evaluate performances and reliability of the WebDAV protocol, and study a possible adoption for the user analysis, in integration or in alternative of the most used protocols.

More specifically, we considered a pilot infrastructure composed by a set of storage elements configured with the webdav interface, hosted at the Belle II sites. The performance tests include also a comparison with xrootd, popular in the HEP community.

As reference tests, we used a set of analysis jobs running in the Belle II software framework, accessing the input data with the ROOT I/O library, in order to simulate as much as possible a realistic
user activity.
The final analysis shows the possibility to achieve promising performances with webdav on different storage systems, and gives an interesting feedback, for Belle II community and for other high energy physics experiments.

Tertiary Keyword (Optional):
Distributed data handling

Secondary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Primary Keyword (Mandatory):
Storage systems

Posters A / Break / 122

A lightweight federation of the Belle II storages through dynafed

Author: Silvio Pardi¹

¹ INFN

Corresponding Author: spardi@na.infn.it

Belle II experiment can take advantage from Data federation technologies to simplify access to distributed datasets and file replicas. The increasing adoption of http and webdav protocol by sites, enable to create lightweight solutions to give an aggregate view of the distributed storage.

In this work, we make a study on the possible usage of the software Dynafed developed by CERN for the creation of an on-the-fly data federation.

We created a first dynafed server, hosted in the datacentre in Napoli, and connected with about the 50% of the production storages of Belle II. Then we aggregated all the file systems under a unique http path. We implemented as well an additional view, in order to browse the single storage file system.

On this infrastructure, we performed a stress test in order to evaluate the impact of federation overall performances, the service resilience, and to study the capability of redirect clients properly to the file replica in case of fault, temporary unavailability of a server.

The results show a good potentiality of the service and suggest additional investigation for additional setup.

Tertiary Keyword (Optional):
Storage systems

Secondary Keyword (Optional):
Distributed data handling

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines
Integration of Oracle and Hadoop: hybrid databases affordable at scale

Authors: Luca Canali\textsuperscript{1}; Prasanth Kothuri\textsuperscript{1}; Zbigniew Baranowski\textsuperscript{2}

\textsuperscript{1} CERN

Corresponding Authors: prasanth.kothuri@cern.ch, luca.canali@cern.ch, zbigniew.baranowski@cern.ch

This work reports on the activities of integrating Oracle and Hadoop technologies for CERN database services and in particular in the development of solutions for offloading data and queries from Oracle databases into Hadoop-based systems. This is of interest to increase the scalability and reduce cost for some our largest Oracle databases. These concepts have been applied, among others, to build offline copies of controls and logging databases, which allow reports to be run without affecting critical production and also reduces the storage cost. Other use cases include making data stored in Hadoop/Hive available from Oracle SQL, which opens the possibility for building applications that integrate data from both sources.

Secondary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Primary Keyword (Mandatory):
Databases

Tertiary Keyword (Optional):

Track 3: Distributed Computing / 124

The ATLAS Production System Evolution. New Data Processing and Analysis Paradigm for the LHC Run2 and High-Luminosity

Author: Misha Borodin\textsuperscript{1}

\textsuperscript{1} University of Iowa (US)

Corresponding Author: mikhail.borodin@cern.ch

The second generation of the ATLAS production system called ProdSys2 is a distributed workload manager that runs daily hundreds of thousands of jobs, from dozens of different ATLAS specific workflows, across more than hundred heterogeneous sites. It achieves high utilization by combining dynamic job definition based on many criteria, such as input and output size, memory requirements and CPU consumption, with manageable scheduling policies and by supporting different kind of computational resources, such as GRID, clouds, supercomputers and volunteering computers. The system dynamically assigns a group of jobs (task) to a group of geographically distributed computing resources. Dynamic assignment and resources utilization is one of the major features of the system, it didn’t exist in the earliest versions of the production system where Grid resources topology has been predefined using national or/and geographical pattern.

Production System has a sophisticated job fault-recovery mechanism, which efficiently allows to run a multi-Terabyte tasks without human intervention. We have implemented train model and open-ended production which allows to submit tasks automatically as soon as new set of data is available and to chain physics groups data processing and analysis with central production run by the experiment.
ProdSys2 simplifies life to ATLAS scientists by offering a flexible web user interface, which implements a user-friendly environment for main ATLAS workflows, e.g., simple way of combining different data flows, and a real-time monitoring optimised to present a huge amount of information.

We present an overview of the ATLAS Production System and its major components features and architecture: task definition, web user interface and monitoring. We describe the important design decisions and lessons learned from an operational experience during the first years of LHC Run2. We also report the performance of the designed system and how various workflows such as data (re)processing, Monte-Carlo and physics group production, users analysis are scheduled and executed within one production system on heterogeneous computing resources.

Tertiary Keyword (Optional):
High performance computing

Secondary Keyword (Optional):
Distributed data handling

Primary Keyword (Mandatory):
Distributed workload management

Posters B / Break / 125

Large scale software building with CMake in ATLAS

Author: Johannes Elmsheuser

1 Brookhaven National Laboratory (US)

Corresponding Author: johannes.elmsheuser@cern.ch

The offline software of the ATLAS experiment at the LHC (Large Hadron Collider) serves as the platform for detector data reconstruction, simulation and analysis. It is also used in the detector trigger system to select LHC collision events during data taking. ATLAS offline software consists of several million lines of C++ and Python code organized in a modular design of more than 2000 specialized packages. Because of different workflows many stable numbered releases are in parallel production use. To accommodate specific workflow requests, software patches with modified libraries are distributed on top of existing software releases on a daily basis. The different ATLAS software applications require a flexible build system that strongly supports unit and integration tests. Within the last year this build system was migrated to CMake. A CMake configuration has been developed that allows one to easily set up and build the mentioned software packages. This also makes it possible to develop and test new and modified packages on top of existing releases. The system also allows one to detect and execute partial rebuilds of the release based on single package changes. The build system makes use of CPack for building RPM packages out of the software releases, and CTest for running unit and integration tests. We report on the migration and integration of the ATLAS...
software to CMake and show working examples of this large scale project in production.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):

Software development process and tools

---

Track 7: Middleware, Monitoring and Accounting / 127

Monitoring of Computing Resource Use of Active Software Releases at ATLAS

**Author:** Antonio Limosani

1 University of Sydney (AU)

**Corresponding Author:** antonio.limosani@cern.ch

The LHC is the world’s most powerful particle accelerator, colliding protons at centre of mass energy of 13 TeV. As the energy and frequency of collisions has grown in the search for new physics, so too has demand for computing resources needed for event reconstruction. We will report on the evolution of resource usage in terms of CPU and RAM in key ATLAS offline reconstruction workflows at the Tier0 at CERN and on the WLCG. Monitoring of workflows is achieved using the ATLAS PerfMon package, which is the standard ATLAS performance monitoring system running inside Athena jobs. Systematic daily monitoring has recently been expanded to include all workflows beginning at Monte Carlo generation through to end user physics analysis, beyond that of event reconstruction. Moreover, the move to a multiprocessor mode in production jobs has facilitated the use of tools, such as "MemoryMonitor", to measure the memory shared across processors in jobs. Resource consumption is broken down into software domains and displayed in plots generated using Python visualization libraries and collected into preformatted auto-generated Web pages, which allow ATLAS' developer community to track the performance of their algorithms. This information is however preferentially filtered to domain leaders and developers through the use of JIRA and via reports given at ATLAS software meetings. Finally, we take a glimpse of the future by reporting on the expected CPU and RAM usage in benchmark workflows associated with the High Luminosity LHC and anticipate the ways performance monitoring will evolve to understand and benchmark future workflows.

---

Secondary Keyword (Optional):

Data processing workflows and frameworks/pipelines

Primary Keyword (Mandatory):

Monitoring

Tertiary Keyword (Optional):
Development and performance of track reconstruction algorithms at the energy frontier with the ATLAS detector

Author: Heather Gray
Co-author: Louis-Guillaume Gagnon

1 CERN
2 Université de Montréal (CA)

Corresponding Authors: heather.gray@cern.ch, louis.guillaume.gagnon@cern.ch

ATLAS track reconstruction code is continuously evolving to match the demands from the increasing instantaneous luminosity of LHC, as well as the increased centre-of-mass energy. With the increase in energy, events with dense environments, e.g. the cores of jets or boosted tau leptons, become much more abundant. These environments are characterised by charged particle separations on the order of ATLAS inner detector sensor dimensions and are created by the decay of boosted objects. Significant upgrades were made to the track reconstruction code to cope with the expected conditions during LHC Run 2. In particular, new algorithms targeting dense environments were developed. These changes lead to a substantial reduction of reconstruction time while at the same time improving physics performance. The employed methods are presented and the prospects for future applications are discussed. In addition, physics performance studies are shown, e.g. a measurement of the fraction of lost tracks in jets with high transverse momentum.

Primary Vertex Reconstruction with the ATLAS experiment

Author: Heather Gray
Co-author: Kathryn Grimm

1 CERN
2 Lancaster University (GB)

Corresponding Authors: heather.gray@cern.ch, kathryn.grimm@cern.ch

Efficient and precise reconstruction of the primary vertex in an LHC collision is essential in both the reconstruction of the full kinematic properties of a hard-scatter event and of soft interactions as a measure of the amount of pile-up. The reconstruction of primary vertices in the busy, high pile-up environment of Run-2 of the LHC is a challenging task. New methods have been developed by the ATLAS experiment to reconstruct vertices in such environments. Advances in vertex seeding include methods taken from medical imaging, which allow for reconstruction of multiple vertices with small spatial separation. The adoption of this new seeding algorithm within the ATLAS adaptive vertex finding and fitting procedure will be discussed, and the first results of the new techniques from Run-2 data will be presented. Additionally, data-driven methods to
evaluate vertex resolution will be presented with special focus on correct
methods to evaluate the effect of the beam spot constraint; results from
these methods in Run-2 data will be presented.

Secondary Keyword (Optional):
Algorithms

Primary Keyword (Mandatory):
Reconstruction

Tertiary Keyword (Optional):

Track 4: Data Handling / 131

Using machine learning algorithms to forecast network and system load metrics for ATLAS Distributed Computing

Author: Mario Lassnig¹

¹ CERN

Corresponding Author: mario.lassnig@cern.ch

The increasing volume of physics data is posing a critical challenge to the ATLAS experiment. In anticipation of high luminosity physics, automation of everyday data management tasks has become necessary. Previously many of these tasks required human decision-making and operation. Recent advances in hardware and software have made it possible to entrust more complicated duties to automated systems using models trained by machine learning algorithms. In this contribution we show results from three ongoing automation efforts. First, we describe our framework for Machine Learning as a Service. This service is built atop the ATLAS Open Analytics Platform and can automatically extract and aggregate data, train models with various machine learning algorithms, and eventually score the resulting models and parameters. Second, we use these models to forecast metrics relevant for network-aware job scheduling and data brokering. We show the characteristics of the data and evaluate the forecasting accuracy of our models. Third, we describe the automation of data management operations tasks. The service is able to classify and cluster run-time metrics based on operational needs. The operator is notified upon a significant event, and potential resolutions are proposed. The framework learns the decisions of the operator through reinforcement algorithms over time, yielding better classification of events and proposals for notification or automated resolution.

Tertiary Keyword (Optional):
Network systems and solutions

Secondary Keyword (Optional):
Distributed data handling

Primary Keyword (Mandatory):
Artificial intelligence/Machine learning
Automatic rebalancing of data in ATLAS distributed data management

Author: Martin Barisits

CERN

Corresponding Author: martin.barisits@cern.ch

The ATLAS Distributed Data Management system stores more than 180PB of physics data across more than 130 sites globally. Rucio, the new data management system of the ATLAS collaboration, has now been successfully operated for over a year. However, with the forthcoming resumption of data taking for Run 2 and its expected workload and utilization, more automated and advanced methods of managing the data are needed. In this article we present an extension to the data management system, which is in charge of detecting and forecasting data imbalances as well as storage elements reaching and surpassing their capacity limit. The system automatically and dynamically rebalances the data to other storage elements, while respecting and guaranteeing data distribution policies and ensuring the availability of the data. This concept not only lowers the operational burden, as these cumbersome procedures had previously to be done manually, but it also enables the system to use its distributed resources more efficiently, which not only affects the data management system itself, but in consequence also the workload management and production systems. This contribution describes the concept and architecture behind those components and shows the benefits made by the system.

Track 2: Offline Computing

ATLAS Simulation using Real Data: Embedding and Overlay

Author: Andrew Haas

New York University

Corresponding Author: drandyhaas@gmail.com

or some physics processes studied with the ATLAS detector, a more accurate simulation in some respects can be achieved by including real data into simulated events, with substantial potential improvements in the CPU, disk space, and memory usage of the standard simulation configuration, at the cost of significant database and networking challenges. Real proton-proton background events can be overlaid (at the detector digitization output stage) on a simulated hard-scatter process, to account for pileup background (from nearby bunch crossings), cavern background, and
detector noise. A similar method is used to account for the large underlying event from heavy ion collisions, rather than directly simulating the full collision. Embedding replaces the muons found in Z→mumu decays in data with simulated taus at the same 4-momenta, thus preserving the underlying event and pileup from the original data event. In all these cases, care must be taken to exactly match detector conditions (beamspot, magnetic fields, alignments, dead sensors, etc.) between the real data event and the simulation.

We will discuss the current status of these overlay and embedding techniques within ATLAS software and computing.

**Primary Keyword (Mandatory):** Simulation

**Secondary Keyword (Optional):**

**Tertiary Keyword (Optional):**

---

**Posters B / Break / 135**

**Metadata for fine-grained processing at ATLAS**

**Author:** Jack Cranshaw

1 Argonne National Laboratory (US)

**Corresponding Author:** cranshaw@anl.gov

High energy physics experiments are implementing highly parallel solutions for event processing on resources that support concurrency at multiple levels. These range from the inherent large-scale parallelism of HPC resources to the multiprocessing and multithreading needed for effective use of multi-core and GPU-augmented nodes.

Such modes of processing, and the efficient opportunistic use of transiently-available resources, lead to finer-grained processing of event data. Previously metadata systems were tailored to jobs that were atomic and processed large, well-defined units of data.

The new environment requires a more fine-grained approach to metadata handling, especially with regard to bookkeeping. For opportunistic resources metadata propagation needs to work even if individual jobs are not finalized.

This contribution describes ATLAS solutions to this problem in the context of the multiprocessing framework currently in use for LHC Run 2, development underway for the ATLAS multithreaded framework (AthenaMT) and the ATLAS EventService.

**Secondary Keyword (Optional):**

Data model

**Primary Keyword (Mandatory):**

Data processing workflows and frameworks/pipelines

**Tertiary Keyword (Optional):**
The ATLAS EventIndex General Dataflow and Monitoring Infrastructure

Author: Alvaro Fernandez Casani

Co-authors: Andrea Favareto, Carlos Garcia Montoro, Dario Barberis, Fedor Prokoshin, Javier Sanchez, Jose Salt, Julius Hrivnac, Rainer Toebbicke, Ruijun Yuan, Santiago Gonzalez De La Hoz

1 Instituto de Física Corpuscular (ES)
2 Università degli Studi e INFN Genova
3 Università e INFN Genova (IT)
4 Federico Santa Maria Technical University (CL)
5 Université de Paris-Sud 11 (FR)
6 CERN
7 Laboratoire de l’Accélérateur Linéaire (FR)

Corresponding Authors: alvaro.fernandez.casani@cern.ch, javier.sanchez@cern.ch, andrea.favareto@cern.ch, ruijun.yuan@cern.ch, jose.salt@cern.ch, fedor.prokoshin@cern.ch, rainer.toebbicke@cern.ch, dario.barberis@cern.ch, carlos.garcia.montoro@cern.ch, santiago.gonzalezdelahoz@cern.ch, julius.hrivnac@cern.ch

The ATLAS EventIndex has been running in production since mid-2015, reliably collecting information worldwide about all produced events and storing them in a central Hadoop infrastructure at CERN. A subset of this information is copied to an Oracle relational database for fast access.

The system design and its optimization is serving event picking from requests of a few events up to scales of tens of thousand of events, and in addition, data consistency checks are performed for large production campaigns. Detecting duplicate events with a scope of physics collections has recently arisen as an important use case.

This paper describes the general architecture of the project and the data flow and operation issues, which are addressed by recent developments to improve the throughput of the overall system. In this direction, the data collection system is reducing the usage of the messaging infrastructure to overcome the performance shortcomings detected during production peaks; an object storage approach is instead used to convey the event index information, and messages to signal their location and status. Recent changes in the Producer/Consumer architecture are also presented in detail, as well as the monitoring infrastructure.

Secondary Keyword (Optional):
Object stores

Primary Keyword (Mandatory):
Distributed data handling

Tertiary Keyword (Optional):
The reconstruction of charged particles trajectories is a crucial task for most particle physics experiments. The high instantaneous luminosity achieved at the LHC leads to a high number of proton-proton collisions per bunch crossing, which has put the track reconstruction software of the LHC experiments through a thorough test. Preserving track reconstruction performance under increasingly difficult experimental conditions, while keeping the usage of computational resources at a reasonable level, is an inherent problem for many HEP experiments. Exploiting concurrent algorithms and using multivariate techniques for track identification are the primary strategies to achieve that goal.

Starting from current ATLAS software, the ACTS project aims to encapsulate track reconstruction into a generic package, which can be built against the Gaudi(Hive) framework. It provides a set of high-level algorithms and data structures for performing track reconstruction tasks as well as fast track simulation. The software is developed with special emphasis on thread-safety to support parallel execution of the code and data structures are optimized for vectorization to speed up linear algebra operations. The implementation is agnostic to the details of the detection technologies and magnetic field configuration which makes it applicable to many different experiments.

Tertiary Keyword (Optional):
Parallelization

Secondary Keyword (Optional):
Algorithms

Primary Keyword (Mandatory):
Reconstruction

Posters A / Break / 138

ATLAS software stack on ARM64

Author: Joshua Wyatt Smith

Co-authors: Arnulf Quadt; Graeme Stewart; Rolf Seuster

1 Georg-August-Universitaet Goettingen (DE)
2 University of Glasgow (GB)
3 University of Victoria (CA)

Corresponding Authors: joshua.wyatt.smith@cern.ch, arnulf.quadt@cern.ch, rolf.seuster@cern.ch, graeme.andrew.stewart@cern.ch

The ATLAS experiment explores new hardware and software platforms that, in the future, may be more suited to its data intensive workloads. One such alternative hardware platform is the ARM architecture, which is designed to be extremely power efficient and is found in most smartphones and tablets.

CERN openlab recently installed a small cluster of ARM 64-bit evaluation prototype servers. Each server is based on a single-socket ARM 64-bit system on a chip, with 32 Cortex-A57 cores. In total, each server has 128 GB RAM connected with four fast memory channels. This paper reports on the port of the ATLAS software stack onto these new prototype ARM64 servers. This included building the "external" packages that the ATLAS software relies on. Patches were needed to introduce this new architecture into the build as well as patches that correct for platform specific code that caused failures on non-x86 architectures. These patches were applied such that porting to further platforms will need no or only very little adjustments. A few additional modifications were needed to account for the different operating system, Ubuntu instead of Scientific Linux 6 / CentOS7. Selected results from the validation of the physics outputs on these ARM 64-bit servers will be reported. CPU, memory and IO intensive benchmarks using ATLAS specific environment
and infrastructure have been performed, with a particular emphasis on the performance vs. energy consumption.

**Tertiary Keyword (Optional):**
High performance computing

**Secondary Keyword (Optional):**
Processor architectures

**Primary Keyword (Mandatory):**
Software development process and tools

---

**Track 2: Offline Computing / 139**

**Modernizing the ATLAS Simulation Infrastructure**

**Author:** Zachary Louis Marshall¹  
**Co-author:** Andrea Di Simone ²

¹ Lawrence Berkeley National Lab. (US)  
² Albert-Ludwigs-Universitaet Freiburg (DE)

**Corresponding Authors:** zach.marshall@cern.ch, andrea.di.simone@cern.ch

The ATLAS Simulation infrastructure has been used to produce upwards of 50 billion proton-proton collision events for analyses ranging from detailed Standard Model measurements to searches for exotic new phenomena. In the last several years, the infrastructure has been heavily revised to allow intuitive multithreading and significantly improved maintainability. Such a massive update of a legacy code base requires careful choices about what pieces of code to completely rewrite and what to wrap or revise. The initialization of the complex geometry was generalized to allow new tools and geometry description languages, popular in some detector groups. The addition of multithreading requires Geant4 MT and GaudiHive, two frameworks with fundamentally different approaches to multithreading, to work together. It also required enforcing thread safety throughout a large code base, which required the redesign of several aspects of the simulation, including "truth," the record of particle interactions with the detector during the simulation. These advances were possible thanks to close interactions with the Geant4 developers.

**Secondary Keyword (Optional):**
Software development process and tools

**Primary Keyword (Mandatory):**
Simulation

**Tertiary Keyword (Optional):**
PanDA for ATLAS distributed computing in the next decade

Author: Tadashi Maeno¹

¹ Brookhaven National Laboratory (US)

Corresponding Author: tmaeno@bnl.gov

The Production and Distributed Analysis (PanDA) system has been developed to meet ATLAS production and analysis requirements for a data-driven workload management system capable of operating at the Large Hadron Collider (LHC) data processing scale. Heterogeneous resources used by the ATLAS experiment are distributed worldwide at hundreds of sites, thousands of physicists analyse the data remotely, the volume of processed data is beyond the exabyte scale, dozens of scientific applications are supported, while data processing requires more than a few billion hours of computing usage per year. PanDA performed very well over the last decade including the LHC Run 1 data taking period. However, it was decided to upgrade the whole system concurrently with the LHC’s first long shutdown in order to cope with rapidly changing computing infrastructure.

After two years of reengineering efforts, PanDA has embedded capabilities for fully dynamic and flexible workload management. The static batch job paradigm was discarded in favor of a more automated and scalable model. Workloads are dynamically tailored for optimal usage of resources, with the brokerage taking network traffic and forecasts into account. Computing resources are partitioned based on dynamic knowledge of their status and characteristics. The pilot has been re-factored around a plugin structure for easier development and deployment. Bookkeeping is handled with both coarse and fine granularities for efficient utilization of pledged or opportunistic resources. Leveraging direct remote data access and federated storage relaxes the geographical coupling between processing and data. An in-house security mechanism authenticates the pilot and data management services in off-grid environments such as volunteer computing and private local clusters.

The PanDA monitor has been extensively optimized for performance and extended with analytics to provide aggregated summaries of the system as well as drill-down to operational details. There are as well many other challenges planned or recently implemented, and adoption by non-LHC experiments such as bioinformatics groups successfully running Paleomix (microbial genome and metagenomes) payload on supercomputers. In this talk we will focus on the new and planned features that are most important to the next decade of distributed computing workload management.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):

Distributed workload management

Track 3: Distributed Computing / 141

Memory handling in the ATLAS submission system from job definition to sites limits

Author: Alessandra Forti¹

¹ University of Manchester (GB)
The ATLAS workload management system is a pilot system based on a late binding philosophy that avoided for many years to pass fine grained job requirements to the batch system. In particular for memory most of the requirements were set to request 4GB vmem as defined in the EGI portal VO card, i.e. 2GB RAM + 2GB swap. However in the past few years several changes have happened in the operating system kernel and in the applications that make such a definition of memory to use for requesting slots obsolete and ATLAS has introduced the new PRODSYS2 workload management which has a more flexible system to evaluate the memory requirements and to submit to appropriate queues. The work stemmed in particular from the introduction of 64bit multicore workloads and the increased memory requirements of some of the single core applications. This paper describes the overall review and changes of memory handling starting from the definition of tasks, the way tasks memory requirements are set using scout jobs and the new memory tool produced in the process to do that, how the factories set these values and finally how the jobs are treated by the sites through the CEs, batch systems and ultimately the kernel.

Tertiary Keyword (Optional):
Computing facilities

Secondary Keyword (Optional):
Computing middleware

Primary Keyword (Mandatory):
Distributed workload management

---

Evolution and experience with the ATLAS Simulation at Point1 Project

Author: Tomas Kouba¹

¹ Acad. of Sciences of the Czech Rep. (CZ)

Corresponding Author: tomas.kouba@cern.ch

The Simulation at Point1 project is successfully running traditional ATLAS simulation jobs on the trigger and data acquisition high level trigger resources. The pool of the available resources changes dynamically and quickly, therefore we need to be very effective in exploiting the available computing cycles.

We will present our experience with using the Event Service that provides the event-level granularity for computations. We will show the design decisions and overhead time related to the usage of the Event Service. The improved utilization of the resources will also be presented with the recent development in the monitoring and the automatic alerting, as well as in the deployment and GUI.

Tertiary Keyword (Optional):
Simulation

Secondary Keyword (Optional):
C3PO - A Dynamic Data Placement Agent for ATLAS Distributed Data Management

Author: Vincent Garonne
Co-author: Thomas Beermann

1 University of Oslo (NO)
2 CERN

Corresponding Authors: vincent.garonne@cern.ch, thomas.beermann@cern.ch

This contribution introduces a new dynamic data placement agent for the ATLAS distributed data management system. This agent is designed to pre-place potentially popular data to make it more widely available. It uses data from a variety of sources. Those include input datasets and sites workload information from the ATLAS workload management system, network metrics from different sources like FTS and PerfSonar, historical popularity data collected through a tracer mechanism and more. With this data it decides if, when and where to place new replicas that then can be used by the WMS to distribute the workload more evenly over available computing resources and then ultimately reduce job waiting times. The new replicas are created with a short lifetime that gets extended, when the data is accessed and therefore the system behaves like a big cache.

This paper gives an overview of the architecture and the final implementation of this new agent. The paper also includes an evaluation of different placement algorithms by comparing the transfer times and the new replica usage.

Primary Keyword (Mandatory):
Distributed data handling

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

Rucio WebUI - The Web Interface for the ATLAS Distributed Data Management

Author: Vincent Garonne
Co-author: Thomas Beermann

1 University of Oslo (NO)
With the current distributed data management system for ATLAS, called Rucio, all user interactions, e.g. the Rucio command line tools or the ATLAS workload management system, communicate with Rucio through the same REST-API. This common interface makes it possible to interact with Rucio using a lot of different programming languages, including Javascript. Using common web application frameworks like JQuery and web.py, a web application for Rucio was built. The main component is R2D2 - the Rucio Rule Definition Droid - which gives the users a simple way to manage their data on the grid. They can search for particular datasets and get details about its metadata and available replicas and easily create rules to create new replicas and delete them if not needed anymore. On the other hand, it is possible for site admins to restrict transfers to their site by setting quotas and manually approve transfers. Besides, R2D2 additional features include transfer backlog monitoring for shifters, group space monitoring for group admins, a bad file replica summary and more.

This paper describes the general architecture of this web application and will detail all the important parts of it.

**Secondary Keyword (Optional):**
Monitoring

**Primary Keyword (Mandatory):**
Distributed data handling

**Tertiary Keyword (Optional):**

---

In this paper, we'll talk about our experiences with different data storage technologies within the ATLAS Distributed Data Management system, and in particular about object-based storage. Object-based storage differs in many points from traditional file system storage and offers a highly scalable, simple and most common storage solution for the cloud. First, we describe the needed changes in the Rucio software to integrate this technology, then we present for which use cases we have evaluated them. Finally, we conclude by reporting the results, performances and the potential future by exploiting more of their specificities and features, like metadata support.

**Tertiary Keyword (Optional):**
Object stores

**Secondary Keyword (Optional):**
Experiences with the new ATLAS Distributed Data Management System

Author: Vincent Garonne

University of Oslo (NO)

Corresponding Author: vincent.garonne@cern.ch

The ATLAS Distributed Data Management (DDM) system has evolved drastically in the last two years with the Rucio software fully replacing the previous system before the start of LHC Run-2. The ATLAS DDM system manages now more than 200 petabytes spread on 130 storage sites and can handle file transfer rates of up to 30Hz. In this talk, we discuss our experience acquired in developing, commissioning, running and maintaining such a large system. First, we describe the general architecture of the system, our integration with external services like the WLCG File Transfer Service and the evolution of the system over its first year of production. Then, we show the performance of the system, describe the integration of new technologies such as object stores, and outline future developments which mainly focus on performance and automation. Finally we discuss the long term evolution of ATLAS data management.

Rucio Auditor - Consistency in the ATLAS Distributed Data Management System

Author: Vincent Garonne

University of Oslo (NO)

Corresponding Author: vincent.garonne@cern.ch

One of the biggest challenge with Large scale data management system is to ensure the consistency between the global file catalog and what is physically on all storage elements.
To tackle this issue, the Rucio software which is used by the ATLAS Distributed Data Management system has been extended to automatically handle lost or unregistered files (aka Dark Data). This system automatically detects these inconsistencies and take actions like recovery or deletion of unneeded files in a central manner. In this talk, we will present this system, explain the internals and give some results.

Secondary Keyword (Optional):
Storage systems

Primary Keyword (Mandatory):
Distributed data handling

Tertiary Keyword (Optional):

Track 5: Software Development / 149

How To Review 4 Million Lines of ATLAS Code

Author: Graeme Stewart
Co-author: Walter Lampl

1 University of Glasgow (GB)
2 University of Arizona (US)

Corresponding Authors: graeme.andrew.stewart@cern.ch, walter.lampl@cern.ch

As the ATLAS Experiment prepares to move to a multi-threaded framework (AthenaMT) for Run3, we are faced with the problem of how to migrate 4 million lines of C++ source code. This code has been written over the past 15 years and has often been adapted, re-written or extended to the changing requirements and circumstances of LHC data taking. The code was developed by different authors, many of whom are no longer active, and under the deep assumption that processing ATLAS data would be done in a serial fashion.

In order to understand the scale of the problem faced by the ATLAS software community, and to plan appropriately the significant efforts posed by the new AthenaMT framework, ATLAS embarked on a wide ranging review of our offline code, covering all areas of activity: event generation, simulation, trigger, reconstruction. We discuss the difficulties in even logistically organising such reviews in an already busy community, how to examine areas in sufficient depth to learn key areas in need of upgrade, yet also to finish the reviews in a timely fashion.

We show how the reviews were organised and how the ouptuts were captured in a way that the sub-system communities could then tackle the problems uncovered on a realistic timeline. Further, we discuss how the review influenced overall planning for the ATLAS Run3 use of AthenaMT and report on how progress is being made towards realistic framework prototypes.

Tertiary Keyword (Optional):
Software development process and tools

Secondary Keyword (Optional):

Tertiary Keyword (Optional):
Frozen-shower simulation of electromagnetic showers in the ATLAS forward calorimeters

Author: Alexander Glazov
Co-author: Ksenia Gasnikova

1 Deutsches Elektronen-Synchrotron (DE)

Corresponding Authors: glazov@mail.desy.de, ksenia.gasnikova@cern.ch

Accurate simulation of calorimeter response for high energy electromagnetic particles is essential for the LHC experiments. Detailed simulation of the electromagnetic showers using Geant4 is however very CPU intensive and various fast simulation methods were proposed instead. The frozen shower simulation substitutes the full propagation of the showers for energies below 1 GeV by showers taken from a pre-simulated library. The method is used for production of the main ATLAS Monte Carlo samples, greatly improving the production time. The frozen showers describe shower shapes, sampling fraction, sampling and noise-related fluctuations very well, while description of the constant term, related to calorimeter non-uniformity, requires a careful choice of the shower library binning. A new method is proposed to tune the binning variables, using multivariate techniques. The method is tested and optimized for the description of the ATLAS forward calorimeter.

Secondary Keyword (Optional):
Artificial intelligence/Machine learning

Primary Keyword (Mandatory):
Simulation

Tertiary Keyword (Optional):

ATLAS World-cloud and networking in PanDA

Author: Fernando Harald Barreiro Megino

1 University of Texas at Arlington

Corresponding Author: fernando.harald.barreiro.megino@cern.ch

The ATLAS computing model was originally designed as static clouds (usually national or geographical groupings of sites) around the Tier 1 centers, which confined tasks and most of the data traffic. Since those early days, the sites’
network bandwidth has increased at O(1000) and the difference in functionalities between Tier 1s and Tier 2s has reduced. After years of manual, intermediate solutions, we have now ramped up to full usage of World-cloud, the latest step in the PanDA Workload Management System to increase resource utilization on the ATLAS Grid, for all workflows (MC production, data (re)processing, etc.).

We have based the development on two new site concepts. Nuclei sites are the Tier 1s and large Tier 2s, where tasks will be assigned and the output aggregated, and satellites are the sites that will execute the jobs and send the output to their nucleus. Nuclei and satellite sites are dynamically paired by PanDA for each task based on the input data availability, capability matching, site load and network connectivity. This contribution will introduce the conceptual changes for World-cloud, the development necessary in PanDA, an insight into the network model and the first half year of operational experience.

Secondary Keyword (Optional):
Computing models

Primary Keyword (Mandatory):
Distributed workload management

Tertiary Keyword (Optional):

Posters A / Break / 152

Data intensive ATLAS workflows in the Cloud

Author: Gerhard Ferdinand Rzehorz

Co-authors: Arnulf Quadt; Gen Kawamura; Oliver Keeble

1 Georg-August-Universitaet Goettingen (DE)
2 CERN

Corresponding Author: gerhard.ferdinand.rzehorz@cern.ch

This contribution reports on the feasibility of executing data intensive workflows on Cloud infrastructures. In order to assess this, the metric ETC = Events/Time/Cost is formed, which quantifies the different workflow and infrastructure configurations that are tested against each other. In these tests ATLAS reconstruction jobs are run, examining the effects of overcommitting (more parallel processes running than CPU cores available), scheduling (staggered execution) and scaling (number of cores). The desirability of commissioning storage in the cloud is evaluated, in conjunction with a simple analytical model of the system, and correlated with questions about the network bandwidth, caches and what kind of storage to utilise. In the end a cost/benefit evaluation of different infrastructure configurations and workflows is undertaken, with the goal to find the maximum of the ETC value.

Tertiary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Secondary Keyword (Optional):
Distributed data handling

Primary Keyword (Mandatory):
Volunteer Computing Experience with ATLAS@Home

Author: David Cameron

Corresponding Author: david.cameron@cern.ch

ATLAS@Home is a volunteer computing project which allows the public to contribute to computing for the ATLAS experiment through their home or office computers. The project has grown continuously since its creation in mid-2014 and now counts almost 100,000 volunteers. The combined volunteers’ resources make up a sizable fraction of overall resources for ATLAS simulation. This paper takes stock of the experience gained so far and describes the next steps in the evolution of the project. These improvements include running natively on Linux to ease the deployment on, for example, university clusters, using multiple cores inside one job to reduce the memory requirements and running different types of workload, such as event generation. In addition to technical details, the success of ATLAS@Home as an outreach tool is evaluated.

Secondary Keyword (Optional):
Outreach

Primary Keyword (Mandatory):
Distributed workload management

Assessment of Geant4 Maintainability with respect to Software Engineering

Authors: Elisabetta Ronchieri; Maria Grazia Pia

Corresponding Authors: maria.grazia.pia@cern.ch, elisabetta.ronchieri@cnaf.infn.it

Maintainability is a critical issue for large scale, widely used software systems, characterized by a long life cycle. It is of paramount importance for a software toolkit, such as Geant4, which is a key instrument for research and industrial applications in many fields, not limited to high energy physics.
Maintainability is related to a number of objective metrics associated with pertinent characteristics of the software. We present an extensive set of these metrics, gathered over recent Geant4 versions with multi-threaded execution capability: they include estimates of the software size, complexity and object-oriented design features.

The collected metrics have been analyzed with various statistical methods to assess the status of Geant4 code with respect to reference values established in software engineering literature, which represent thresholds for risk. The assessment has been detailed to a fine grained level across Geant4 packages to identify potential problematic areas effectively, also taking into account specific peculiarities of different simulation domains.

The evaluation of the metrics suggests preventive actions to facilitate the maintainability of the toolkit over an extended life cycle.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Simulation

Primary Keyword (Mandatory):
Software development process and tools

Posters B / Break / 155

Migrating the Belle II Collaborative Services and Tools

Author: Andreas Gellrich

Co-authors: Bastian Vennemann; Carsten Niebuhr; Oliver Frost; Peter van der Reest; Takanori Hara; Thomas Kuhr

1 DESY
2 KEK
3 LMU

Corresponding Authors: takanori.hara@kek.jp, bastian.vennemann@desy.de, niebuhr@mail.desy.de, andreas.gellrich@desy.de, oliver.frost@desy.de, peter.van.der.reest@desy.de, thomas.kuhr@cern.ch

Collaborative services and tools are essential for any (HEP) experiment. They help to integrate global virtual communities by allowing to share and exchange relevant information among members by way of web-based services. Typical examples are public and internal web pages, wikis, mailing list services, issue tracking system, services for meeting organization and document and authorship management as well as build services and code repositories.

In order to achieve stable and reliable services, the Belle II collaboration decided after considering different options to migrate the current set of services into the existing IT infrastructure at DESY. The DESY IT infrastructure is built to be reliable and highly available and has been providing sustainable services for many scientific groups such as CFEL and EXFEL as well as for HEP. It includes fail-over mechanisms, back-up and archiving options for all services. As for all computer centers, security is a major issue at DESY which is thoroughly considered for all services.
In order to make optimal use of the existing services and support structures at DESY, implementation switches are necessary for some services. Hence not all services can simply be copied but will have to be adapted or even restructured. In preparation for the final migration small groups of experts went through the details of all services to identify problems and to develop alternative solutions.

In the future the Belle II collaboration will be visible in web under the domain 'belle2.org' which is owned and hosted by DESY. It will also be necessary to equip all Belle II members with credentials to access DESY services. For this purpose a certificate and VO membership based portal is set up which allows for authentication and authorization of users.

After the approval by the Belle II bodies the migration process started in spring. It must be finished before the KEK summer shutdown in July.

In the contribution to CHEP2016, at a time when the migration process must have finished, we will describe various aspects migration process and the services. Furthermore we plan to share experiences, reveal details, and give useful hints for similar approaches.

Tertiary Keyword (Optional):
Software development process and tools

Secondary Keyword (Optional):
Computing facilities

Primary Keyword (Mandatory):
Collaborative tools

---

**Track 5: Software Development / 156**

**ROOT and new programming paradigms**

**Author:** Philippe Canal

**Co-author:** Axel Naumann

---

1 *Fermi National Accelerator Lab. (US)*

2 *CERN*

**Corresponding Authors:** philippe.canal@cern.ch, axel.naumann@cern.ch

ROOT is one of the core software tool for physicists. For more than a decade it has a central position in the physicists’ analysis code and the experiments’ frameworks thanks in parts to its stability and simplicity of use. This allowed software development for analysis and frameworks to use ROOT as a “common language” for HEP, across virtually all experiments.

Software development in general and in HEP frameworks in particular has become increasingly complex over the years. From straightforward code fitting in a single FORTRAN source file, HEP software has grown over the years to span millions of lines of code spread amongst many, more or so collaborating, packages and libraries. To add to the complexity, in an effort to better exploit current and upcoming hardware, this code is being adapted to move from purely scalar and serial algorithm to complex multithread, multi-tasked and/or vectorized versions.
The C++ language itself and the software development community’s understanding of the best way to leverage its strength has evolved significantly. One of the best examples of this being the "C++ Core Guidelines" which purports to get a “smaller, simpler and safer language” out of C++. At the same time new tools and techniques are being developed to facilitate proving and testing the correctness of software programs, as exemplified by the C++ Guideline Support Library, but those require the tool to be able to understand the semantic of the interfaces. Design patterns and interface tricks that were appropriate in the early days of C++ are often no longer the best choices for API design. ROOT is at the heart of virtually all physics analysis and most HEP frameworks and as such needs to lead the way and help demonstrate and facilitate the application of those modern paradigms.

This presentation will review what theses lessons are and how they can be applied to an evolution of the ROOT C++ interfaces, striking a balance between conserving familiarity with the legacy interfaces (to facilitate both the transition of existing code and the learning of the new interfaces) and significantly improving the expressiveness, clarity, (re)usability, thread friendliness, and robustness of the user code.

Tertiary Keyword (Optional):
High performance computing

Secondary Keyword (Optional):
Software development process and tools

Primary Keyword (Mandatory):
Analysis tools and techniques

Track 5: Software Development / 157

Status and Evolution of ROOT

Author: Axel Naumann

1 CERN

Corresponding Author: axel.naumann@cern.ch

With ROOT 6 in production in most experiments, ROOT has changed gear during the past year: the development focus on the interpreter has been redirected into other areas.

This presentation will summarize the developments that have happened in all areas of ROOT, for instance concurrency mechanisms, the serialization of C++11 types, new graphics palettes, new “glue” packages for multivariate analyses, and the state of the Jupyter and JavaScript interfaces and language bindings.

It will lay out the short term plans for ROOT 5 and ROOT 6 and try to forecast the future evolution of ROOT, for instance with respect to more robust interfaces and a fundamental change in the graphics and GUI system.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):
Analysis tools and techniques
Computing Performance of GeantV Physics Models

Author: Soon Yung Jun

Co-authors: Abhijit Bhattacharyya ; Ajit Kumar Mohanty ; Alberto Ribon ; Ananya Ananya ; Andrei Gheata ; Calebe De Paula Bianchini ; Farah Hariri ; Federico Carminati ; Guilherme Amadio ; Guilherme Lima ; Harphool Kumawat ; Ilias Goulas ; John Apostolakis ; Krzysztof Genser ; Laurent Duhem ; Marilena Bandieramonte ; Mihaela Gheata ; Mihaly Novak ; Oksana Shadura ; Philippe Canal ; Raman Sehgal ; Rene Brun ; Rogerio Iope ; Sandro Christian Wenzel ; Sofia Vallecorsa ; Tatiana Nikitina ; Victor Daniel Elvira ; Witold Pokorski ; Yang Zhang

1 Fermi National Accelerator Lab. (US)
2 Bhabha Atomic Research Centre (IN)
3 CERN
4 IIT- Indian Institute of Technology (IN)
5 UNESP - Universidade Estadual Paulista (BR)
6 FermiLab (US)
7 Fermi National Accelerator Laboratory (US)
8 ISS - Institute of Space Science (RO)
9 National Technical Univ. of Ukraine "Kyiv Polytechnic Institute
10 Gangneung-Wonju National University (KR)
11 KIT - Karlsruhe Institute of Technology (DE)

Corresponding Authors: guilherme.amadio@cern.ch, philippe.canal@cern.ch, alberto.ribon@cern.ch, ilias.goulas@cern.ch, raman.sehgal@cern.ch, rogerio.iope@cern.ch, calebe.bianchini@cern.ch, mihaly.novak@cern.ch, genser@fnal.gov, harphool.kumawat@cern.ch, marilena.bandieramonte@cern.ch, guilherme.lima@cern.ch, andrei.gheata@cern.ch, rene.brun@cern.ch, oksana.shadura@cern.ch, farah.hariri@cern.ch, sofia.vallecorsa@cern.ch, soon.yung.jun@cern.ch, laurent.duhem@cern.ch, ajit.kumar.mohanty@cern.ch, yang.zhang@cern.ch, ananya.ananya@cern.ch, abhijit.bhattacharyya@cern.ch, witold.pokorski@cern.ch, federico.carminati@cern.ch, daniel.elvira@cern.ch, tatiana.nikitina@cern.ch, mihaela.gheata@cern.ch, john.apostolakis@cern.ch, sandro.wenzel@cern.ch

The recent progress in parallel hardware architectures with deeper vector pipelines or many-cores technologies brings opportunities for HEP experiments to take advantage of SIMD and SIMT computing models. Launched in 2013, the GeantV project studies performance gains in propagating multiple particles in parallel, improving instruction throughput and data locality in HEP event simulation. One of challenges in developing highly parallel and efficient detector simulation is the minimization of the number of conditional branches or thread divergence during the particle transportation process. Due to the complexity of geometry description and physics algorithms of a typical HEP application, performance analysis is indispensable in identifying factors limiting parallel execution. In this report, we will present design considerations and computing performance of GeantV physics models on coprocessors (Intel Xeon Phi and NVidia GPUS) as well as on mainstream CPUs. As the characteristics of these platforms are very different, it is essential to collect profiling data with a variety of tools and to analyze hardware specific metrics and their derivatives to be able to evaluate and tune the performance. We will also show how the performance of parallelized physics models factorizes from the rest of GeantV event simulation.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

High performance computing
Software Aspects of the Geant4 Validation Repository

Authors: Alberto Ribon\textsuperscript{1}; Andrea Dotti\textsuperscript{2}; Dima Konstantinov\textsuperscript{3}; Federico Carminati\textsuperscript{1}; Gunter Folger\textsuperscript{1}; Hans-Joachim Wenzel\textsuperscript{1}; Julia Yarba\textsuperscript{4}; Krzysztof Genser\textsuperscript{5}; Victor Daniel Elvira\textsuperscript{4}; Witold Pokorski\textsuperscript{1}

\textsuperscript{1} CERN  \\
\textsuperscript{2} SLAC National Accelerator Laboratory (US)  \\
\textsuperscript{3} IHEP  \\
\textsuperscript{4} Fermi National Accelerator Lab. (US)  \\
\textsuperscript{5} Fermi National Accelerator Laboratory (US)

Corresponding Authors: dkonst@mail.cern.ch, andrea.dotti@cern.ch, witold.pokorski@cern.ch, wenzel@fnal.gov, genser@fnal.gov, yarba_j@fnal.gov, daniel.elvira@cern.ch, gunter.folger@cern.ch, federico.carminati@cern.ch, alberto.ribon@cern.ch

Geant4 is a toolkit for the simulation of the passage of particles through matter. Its areas of application include high energy, nuclear and accelerator physics as well as studies in medical and space science.

The Geant4 collaboration regularly performs validation and regression tests through its development cycle. A validation test compares results obtained with a specific Geant4 version with data obtained by various experiments. On the other hand, a regression test compares results of two or more versions of Geant4 for any observable. To make validation materials easily available to both collaborators and the user community in general, most of the validation data are stored in one central repository. The availability of this data should help experimenters to find answers to questions like:

- What data is used to validate Geant4 physics?
- How well does Geant4 describe the data of interest to the experiment?
- Which model provided by Geant4 best describes the data of interest for the experiment?
- What are the benefits of switching to the latest improved version of Geant4?

Having easy access to this data might also help in estimating the systematic uncertainties stemming from the simulation of physical processes like the response of a detector or predicting the flux of neutrinos produced when a target is illuminated with protons.

The repository consists of a relational database that stores experimental data and Geant4 test results, accessed through a Java API and a web application, which allows search, selection and graphical display of the data. In this presentation we will describe these components and the technology choices we made.

Future plans include providing a web API, expanding the number of experimental data sets and providing quantitative statistical tests. We also want to stress that the project is not specific to Geant4 and it can be used with other Monte Carlo (e.g. GENIE) tools as well.
MPEXS: A CUDA MonteCarlo of the simulation of electromagnetic interactions

Authors: Akinori Kimura\textsuperscript{None}; Andrea Dotti\textsuperscript{1}; Koichi Murakami\textsuperscript{None}; Makoto Asai\textsuperscript{1}; Margot Gerritsen\textsuperscript{2}; Nick Henderson\textsuperscript{2}; Shogo Okada\textsuperscript{3}; Takashi Sasaki\textsuperscript{4}

\textsuperscript{1} SLAC National Accelerator Laboratory (US)
\textsuperscript{2} Institute for Computational & Mathematical Engineering, Stanford University
\textsuperscript{3} KEK
\textsuperscript{4} High Energy Accelerator Research Organization (JP)

Corresponding Authors: takashi.sasaki@kek.jp, andrea.dotti@cern.ch, nwh@stanford.edu, asai@slac.stanford.edu, shogo@post.kek.jp, margot.gerritsen@stanford.edu, akimura@ashitech.ac.jp, koichi.murakami@kek.jp

As more detailed and complex simulations are required in different application domains, there is much interest in adapting the code for parallel and multi-core architectures. Parallelism can be achieved by tracking many particles at the same time. This work presents MPEXS, a CUDA implementation of the core Geant4 algorithm used for the simulation of electromagnetic interactions (electron, positron and gamma primaries) in voxelised geometries.

A detailed analysis of the code has been performed to maximise the performances on GPU hardware, we have reduced thread divergence, perform coalesced memory reads and writes and introduced atomic reduction of dose calculation. The voxelised geometry allows for a simplified navigation that further reduces thread divergence.

These optimizations allow for very encouraging results and experiments with simulations in MPEXS demonstrate speedups of more than 200 times over Geant4.

A rigorous physics validation has shown that the calculations obtained with MPEXS are equivalent to the Geant4 predictions.

The primary interest of MPEXS code is to simulate benchmarks for radiation-therapy calculations, however the general lessons learn from this code can be applied to more general applications. As an example we have extended MPEXS to treat the simulation of production and transport of radicals (the so called Geant4-DNA extension). We are currently studying its extension to other particles species (neutrons and low-energy hadrons) that could be of interest of domains outside the radiation therapy one.

Tertiary Keyword (Optional):
Parallelization

Secondary Keyword (Optional):
Simulation

Primary Keyword (Mandatory):
High performance computing

Multi-threaded Geant4 on Intel Many Integrated Core architectures

Authors: Andrea Dotti\textsuperscript{1}; Makoto Asai\textsuperscript{1}; Steven Andrew Farrell\textsuperscript{2}
In the midst of the multi- and many-core era, the computing models employed by HEP experiments are evolving to embrace the trends of new hardware technologies. As the computing needs of present and future HEP experiments - particularly those at the Large Hadron Collider - grow, adoption of many-core architectures and highly-parallel programming models is essential to prevent degradation in scientific capability.

Simulation of particle interactions is typically a major consumer of CPU resources in HEP experiments. The recent release of a highly performant multi-threaded version of Geant4 opens the door for experiments to fully take advantage of highly-parallel technologies.

The Many Integrated Core (MIC) architecture of Intel, known as the Xeon Phi family of products, provide a platform for highly-parallel applications. Their large number of cores and Linux-based environment make them an attractive compromise between conventional CPUs and general-purpose GPUs. Xeon Phi processors will be appearing in next-generation supercomputers such as Cori Phase 2 at NERSC.

To prepare for using these next-generation supercomputers, a Geant4 application has been developed to test and study HEP particle simulations on the MIC Intel architectures (HepExpMT).

This application serves as a demonstrator of the feasibility and computing-opportunity of utilizing this advanced architecture with a complex detector geometry.

We have measured the performances of the application on the first generation of Xeon Phi coprocessors (code name Knights Corner, KNC). In this work we extend the scalability measurements to the second generation of Xeon Phi architectures (code name Knights Landing, KNL) in preparation of further testing on Cori Phase 2 supercomputer at NERSC.

Tertiary Keyword (Optional):
Parallelization

Secondary Keyword (Optional):
Simulation

Primary Keyword (Mandatory):
High performance computing

Track 4: Data Handling / 162

CEPHFS: a new generation storage platform for Australian high energy physics

Authors: Goncalo Borges1; Lucien Philip Boland2; Sean Crosby2

1 University of Sydney (AU)
2 University of Melbourne (AU)

Corresponding Authors: sean.christopher.crosby@cern.ch, goncalo@physics.usyd.edu.au, lucien.boland@cern.ch

CEPH is a cutting edge, open source, self-healing distributed data storage technology which is exciting both the enterprise and academic worlds. CEPH delivers an object storage layer (RADOS), block
storage layer, and file system storage in a single unified system. CEPH object and block storage implementations are widely used in a broad spectrum of enterprise contexts, from dynamic provision of bare block storage to object storage backends of virtual machines images in cloud platforms. The High Energy Particle Physics community has also recognized its potential by deploying CEPH object storage clusters both at the Tier-0 (CERN) and in some Tier-1s, and by developing support for the GRIDFTP and XROOTD (a bespoke HEP) transfer and access protocols. However, the CEPH filesystem (CEPHFS) has not been subject to the same level of interest. CEPHFS layers a distributed POSIX file system over CEPH’s RADOS using a cluster of metadata servers dynamically partitioning responsibility for the file system namespace and distributing the metadata workload based on client accesses. It is the less mature CEPH product and has been waiting to be tagged as a production-like product for a long time.

In this paper we present a CEPHFS use case implementation at the Center of Excellence for Particle Physics at the TeraScale (CoEPP). CoEPP operates the Australia Tier-2 for ATLAS and joins experimental and theoretical researchers from the Universities of Adelaide, Melbourne, Sydney and Monash. CEPHFS is used to provide a unique object storage system, deployed on commodity hardware and without single points of failure, used by Australian HEP researchers in the different CoEPP locations to store, process and share data, independent of their geographical location. CEPHFS is also working in combination with a SRM and XROOTD implementation, integrated in ATLAS Data Management operations, and used by HEP researchers for XROOTD or/and POSIX-like access to ATLAS Tier-2 user areas. We will provide details on the architecture, its implementation and tuning, and report performance I/O metrics as experienced by different clients deployed over WAN. We will also explain our plan to collaborate with Red Hat Inc. on extending our current model so that the metadata cluster distribution becomes multi-site aware, such that regions of the namespace can be tied or migrated to metadata servers in different data centers.

In its current status, CoEPP’s CEPHFS has already been in operation for almost a year (at the time of the conference). It has proven to be a service that follows the best industry standards at a significantly lower cost and fundamental to promote data sharing and collaboration between Australian HEP researchers.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Distributed data handling

Primary Keyword (Mandatory):

Storage systems

Track 6: Infrastructures / 165

Evaluation of lightweight site setups within WLCG infrastructure

Authors: Alessandra Forti¹; Maarten Litmaath²

Co-authors: Alessandro Di Girolamo ²; Andrea Sciaba ²; Andrea Valassi ²; Jose Flix Molina ³; Julia Andreeva ²; Maria Alandes Pradillo ²; Maria Dimou ²

¹ University of Manchester (GB)
² CERN
³ Centro de Investigaciones Energ. Medioambientales y Tecn. - (ES)

Corresponding Authors: maria.alandes.pradillo@cern.ch, jose.flix.molina@cern.ch, alessandro.di.girolamo@cern.ch, andrea.sciaba@cern.ch, maria.dimou@cern.ch, andrea.valassi@cern.ch, alessandra.forti@cern.ch, julia.andreeva@cern.ch, maarten.litmaath@cern.ch
The Worldwide LHC Computing Grid (WLCG) infrastructure allows the use of resources from more than 150 sites. Until recently the setup of the resources and the middleware at a site were typically dictated by the partner grid project (EGI, OSG, NorduGrid) to which the site is affiliated. Since a few years, however, changes in hardware, software, funding and experiment computing requirements have increasingly affected the way resources are shared and supported. At the WLCG level this implies a need for more flexible and lightweight methods of resource provisioning.

In the WLCG cost optimisation survey presented at CHEP 2015 the concept of lightweight sites was introduced, viz. sites essentially providing only computing resources and aggregating around core sites that provide also storage. The efficient use of lightweight sites requires a fundamental reorganisation not only in the way jobs run, but also in the topology of the infrastructure and the consolidation or elimination of some established site services.

This contribution gives an overview of the solutions being investigated through “demonstrators” of a variety of lightweight site setups, either already in use or planned to be tested in experiment frameworks.

**Secondary Keyword (Optional):**
Computing middleware

**Primary Keyword (Mandatory):**
Computing facilities

**Tertiary Keyword (Optional):**

---

**Track 3: Distributed Computing / 166**

**The Cherenkov Telescope Array production system for Monte Carlo simulations and analysis**

**Author:** Luisa Arrabito

**Co-authors:** Andreas Haupt; Andrei Tsaregorodtsev; Federico Stagni; Gernot Maier; Johann bregeon; Konrad Bernloehr; Nadine Neyroud; Tarek Hassan; abelardo moralejo

---

The Cherenkov Telescope Array (CTA) – an array of many tens of Imaging Atmospheric Cherenkov Telescopes deployed on an unprecedented scale – is the next-generation instrument in the field of very high energy gamma-ray astronomy. An average data stream of about 0.9 GB/s for about 1300 hours of observation per year is expected, therefore resulting in 4 PB of raw data per year and a total of 27 PB/year, including archive and data processing. The start of CTA operation is foreseen in 2018 and it will last about 30 years. The installation of the first telescopes in the two pre-selected locations (Paranal ESO, Chile and La Palma, Spain) will start in 2017. In order to select the best site
candidate to host CTA telescopes (in the North and in the South hemispheres), massive Monte Carlo simulations have been performed since 2012. Once the two sites have been selected, we have started new Monte Carlo simulations to determine the optimal array layout with respect to the obtained sensitivity. Taking into account that CTA may be finally composed by 7 different telescope types coming in 3 different sizes, many different combinations of telescope position and multiplicity as a function of the telescope type have been proposed. This last Monte Carlo campaign represented a huge computational effort, since several hundreds of telescope positions have been simulated, while for future instrument response function simulations, only the operating telescopes will be considered. In particular, during the last 8 months, about 1.4 PB of MC data have been produced and processed with different analysis chains, with a corresponding overall CPU consumption of about 125x10^6 HS06 hours. In these proceedings, we describe the employed computing model, based on the use of grid resources, as well as the production system setup, which relies on the DIRAC (Distributed Infrastructure with Remote Agent Control) framework. Finally, we present the envisaged evolutions of the CTA production system for the off-line data processing during CTA operations and the instrument response function simulations.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Distributed workload management

Primary Keyword (Mandatory):
Computing models

Posters A / Break / 167

Processing and Quality Monitoring for the ATLAS Tile Hadronic Calorimeter data

Author: Blake Oliver Burghgrave

1 Northern Illinois University (US)

Corresponding Author: blake.burghgrave@cern.ch

We present an overview of Data Processing and Data Quality (DQ) Monitoring for the ATLAS Tile Hadronic Calorimeter. Calibration runs are monitored from a data quality perspective and used as a cross-check for physics runs. Data quality in physics runs is monitored extensively and continuously. Any problems are reported and immediately investigated. The DQ efficiency achieved was 99.6% in 2012 and 100% in 2015, after the detector maintenance in 2013-2014.

Changes to detector status or calibrations are entered into the conditions database during a brief calibration loop between when a run ends and bulk processing begins. Bulk processed data is reviewed and certified for the ATLAS Good Run List if no problem is detected. Experts maintain the tools used by DQ shifters and the calibration teams during normal operation, and prepare new conditions for data reprocessing and MC production campaigns. Conditions data are stored in 3 databases: Online DB, Offline DB for data and a special DB for Monte Carlo. Database updates can be performed through a custom-made web interface.

Tertiary Keyword (Optional):
Reconstruction
Data acquisition and processing in the ATLAS Tile Calorimeter Phase-II Upgrade Demonstrator

Author: Alberto Valero Biot

Corresponding Author: alberto.valero@cern.ch

The LHC has planned a series of upgrades culminating in the High Luminosity LHC (HL-LHC) which will have an average luminosity 5-7 times larger than the nominal Run-2 value. The ATLAS Tile Calorimeter (TileCal) will undergo an upgrade to accommodate to the HL-LHC parameters. The TileCal read-out electronics will be redesigned, introducing a new read-out strategy.

The photomultiplier signals will be digitized and transferred to the TileCal PreProcessors (TilePPr) located off-detector for every bunch crossing, requiring a data bandwidth of 80 Tbps. The TilePPr will provide preprocessed information to the first level of trigger and in parallel will store the samples in pipeline memories. The data for the events selected by the trigger system will be transferred to the ATLAS global Data AcQuisition (DAQ) system for further processing.

A demonstrator drawer has been built to evaluate the new proposed readout architecture and prototypes of all the components. In the demonstrator, the detector data received in the TilePPr are stored in pipeline buffers and, upon the reception of an external trigger signal, the data events are processed, packed and read out in parallel through the legacy ROD system, the new Front-End Link eXchange (FELIX) system and an ethernet connection for monitoring purposes.

The data are processed in the Digital Signal Processors of the RODs and transmitted to the ATLAS DAQ system where the data are reconstructed using the ATLAS standard software framework. The data read out through FELIX and the monitoring ethernet connection use a new custom data-format and they are processed using special software packages.

This contribution will describe in detail the data processing and the hardware, firmware and software components of the TileCal demonstrator readout system. In addition, the system integration tests and results from the two test-beam periods planned for 2016 will be presented.
Validation of Electromagnetic Physics Models for Parallel Computing Architectures in the GeantV project

Author: Marilena Bandieramonte

1 CERN

Corresponding Author: marilena.bandieramonte@cern.ch

High-energy particle physics (HEP) has advanced greatly over recent years and current plans for the future foresee even more ambitious targets and challenges that have to be coped with. Amongst the many computer technology R&D areas, simulation of particle detectors stands out as the most time consuming part of HEP computing. An intensive R&D and programming effort is required to exploit the new opportunities offered by technological developments in order to support the scientific progress and the corresponding increasing demand of computing power necessary for future experimental HEP programs. The GeantV project aims at narrowing the gap between the performance of the existing HEP detector simulation software and the ideal performance achievable, exploiting the latest advances in computer technology. The project has developed a particle detector simulation prototype capable of transporting in parallel particles in complex geometries profiting by instruction level parallelism (SIMD and SIMT) and task-level parallelism (multithreading), following both the multi-core and the many-core opportunities. We present preliminary validation results concerning the electromagnetic (EM) physics models developed for parallel computing architectures within the GeantV project. In order to exploit the potential of vectorization and accelerators and to make the physics model effectively parallelizable, alternative sampling techniques have been implemented and tested. Some of these techniques introduce intervals and discrete tables. We identify artefacts that are introduced by different discrete sampling techniques and determine the energy range in which these methods provide acceptable approximation. We introduce a set of automated statistical analysis in order to verify the vectorized models by checking their consistency with the corresponding Geant4 models and to validate them against experimental data. The validation presented here is part of a larger effort, involving Cern, Fermilab and SLAC, for a common development of new physics validation framework designed for various particle physics detector simulation software and is focused on the extension for GeantV.

Track 2: Offline Computing / 169

Tertiary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Secondary Keyword (Optional):
Reconstruction

Primary Keyword (Mandatory):
DAQ

High performance computing

Simulation

Track 2: Offline Computing / 170
Stochastic optimisation of GeantV code by use of genetic algorithms

Author: Oksana Shadura

1 National Technical Univ. of Ukraine "Kyiv Polytechnic Institute

Corresponding Author: oksana.shadura@cern.ch

GeantV simulation is a complex system based on the interaction of different modules needed for detector simulation, which include transportation (heuristically managed mechanism of sets of predefined navigators), scheduling policies, physics models (cross-sections and reaction final states) and a geometrical modeler library with geometry algorithms. The GeantV project is recasting the simulation framework to get maximum benefit from SIMD/MIMD computational architecture and highly massive parallel systems. This involves finding the appropriate balance of several aspects influencing computational performance (floating-point performance, usage of off-chip memory bandwidth, specification of cache hierarchy, and etc.) and a large number of program parameters that have to be optimized to achieve the best speedup of simulation. This optimisation task can be treated as a "black-box" optimization problem, which requires searching the optimum set of parameters using only point-wise function evaluations. The goal of this study is to provide a mechanism for optimizing complex systems (high energy physics particle transport simulations) with the help of genetic algorithms and evolution strategies as a tuning process for massive coarse-grain parallel simulations. One of the described approaches is based on introduction of specific multivariate analysis operator that could be used in case of resource expensive or time consuming evaluations of fitness functions, in order to speed-up the convergence of the "black-box" optimization problem.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Artificial intelligence/Machine learning

Primary Keyword (Mandatory):
Simulation

GeantV phase 2: developing the particle transport library

Author: Andrei Gheata

1 CERN

Corresponding Author: andrei.gheata@cern.ch

After an initial R&D stage of prototyping portable performance for particle transport simulation, the GeantV project reaches a new phase where the different components such as kernel libraries, scheduling, geometry and physics are rapidly developing. The increase in complexity is accelerating by the multiplication of demonstrator examples and tested platforms, while trying to maintain a balance between code stability and new developments. While some of the development efforts start being available for the HEP community such as the geometry and vector core libraries, GeantV is passing to the stage of demonstrator in order to validate and extend its previous performance achievements on a variety of HEP detector setups. A strategy for adding native support for fast simulation was foreseen for both framework and user-defined parametrisations. This will allow integrating naturally fast simulation within the GeantV parallel workflow, without the need to run any additional programs.

We will present the current status of the project, its most recent results and benchmarks, giving a perspective on the future usage of the software.
New Directions in the CernVM File System

Author: Jakob Blomer

Co-authors: Gerardo Ganis; Predrag Buncic; Rene Meusel

Corresponding Authors: rene.meusel@cern.ch, pbuncic@cern.ch, jakob.blomer@cern.ch, gerardo.ganis@cern.ch

The CernVM File System today is commonly used to host and distribute application software stacks. In addition to this core task, recent developments expand the scope of the file system into two new areas. Firstly, CernVM-FS emerges as a good match for container engines to distribute the container image contents. Compared to native container image distribution (e.g. through the “Docker registry”), CernVM-FS massively reduces the network traffic for image distribution. This has been shown, for instance, by a prototype integration of CernVM-FS into Mesos developed by Mesosphere, Inc. We present possible paths for a smooth integration in Docker and necessary changes to the CernVM-FS server to support the typical container image work flow and lifecycle.

Secondly, CernVM-FS recently raised interest as an option for the distribution of experiment data file catalogs. While very powerful tools are in use for accessing data files in a distributed and scalable manner, finding the file names is typically done by a central, experiment specific SQL database. A name space on CernVM-FS can particularly benefit from an existing, scalable infrastructure, from the POSIX interface and the end-to-end content verification. For this use case, we outline necessary modifications to the CernVM-FS server in order to provide a generic, distributed namespace that supports billions of files and thousands of writes per second.

Toward a CERN Virtual Visit Service

Author: Steven Goldfarb
The installation of Virtual Visit services by the LHC collaborations began shortly after the first high energy collisions were provided by the CERN accelerator in 2010. The experiments: ATLAS, CMS, LHCb, and ALICE have all joined in this popular and effective method to bring the excitement of scientific exploration and discovery into classrooms and other public venues around the world. Their programmes, which use a combination of video conference, webcast, and video recording to communicate with remote audiences have already reached tens of thousands of viewers, and the demand only continues to grow. Other venues, such as the CERN Control Centre, are also considering similar permanent installations.

We present a summary of the development of the various systems in use around CERN today, including the technology deployed and a variety of use cases. We then lay down the arguments for the creation of a CERN-wide service that would support these programmes in a more coherent and effective manner. Potential services include a central booking system and operational management similar to what is currently provided for the common CERN video conference facilities. Key technological choices would provide additional functionality that could support communication and outreach programmes based on popular tools including (but not limited to) Skype, Google Hangouts, and Periscope. Successful implementation of the project, which relies on close partnership between the experiments, CERN IT CDA, and CERN IR ECO, has the potential to reach an even larger, global audience, more effectively than ever before.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):
Outreach

Track 1: Online Computing / 174

Performance and evolution of the DAQ system of the CMS experiment for Run-2

Authors: Frans Meijers1; Remi Mommsen2

1 CERN
2 Fermi National Accelerator Lab. (US)

Corresponding Authors: frans.meijers@cern.ch, remigius.mommsen@cern.ch

The data acquisition system (DAQ) of the CMS experiment at the CERN Large Hadron Collider assembles events at a rate of 100 kHz, transporting event data at an aggregate throughput of 100 GByte/s to the high-level trigger (HLT) farm. The HLT farm selects and classifies interesting events for storage and offline analysis at a rate of around 1 kHz.
The DAQ system has been redesigned during the accelerator shutdown (LS1) in 2013/14. In order to handle higher LHC luminosities and event pileup, a number of sub-detectors are upgraded, increasing the number of readout channels and replacing the off-detector readout electronics with a μTCA implementation. The new DAQ system support the read-out of of the off-detector electronics with point to point links of both the legacy systems, as well as the new μTCA based systems with a fibre based implementation up to 10 Gbps and reliable protocol.

The new DAQ architecture takes advantage of the latest developments in the computing industry. For data concentration, 10/40 Gbit Ethernet technologies are used, as well as an implementation of a reduced TCP/IP in FPGA for a reliable transport between DAQ custom electronics and commercial computing hardware. A 56 Gbps Infiniband FDR CLOS network has been chosen for the event builder with a throughput of ~4 Tbps. The HLT processing is entirely file-based. This allows the DAQ and HLT systems to be independent, and to use the same framework for the HLT as for the offline processing. The fully built events are sent to the HLT with 1/10/40 Gbit Ethernet via network file systems. A hierarchical collection of HLT accepted events and monitoring meta-data are stored in to a global file system. The monitoring of the HLT farm is done with the Elasticsearch analytics tool. This paper presents the requirements, implementation, and performance of the system. Experience is reported on the operation for the LHC pp runs as well as at the heavy ion Pb-Pb runs. The evolution of the DAQ system will be presented including the expansion to accommodate new detectors.

Primary Keyword (Mandatory):
DAQ

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

Posters B / Break / 175

Upgrading and Expanding Lustre Storage for use with the WLCG

Author: Daniel Traynor

Co-author: Terry Froy

1 Queen Mary University of London

Corresponding Authors: t.froy@qmul.ac.uk, d.traynor@qmul.ac.uk

The Queen Mary University of London grid site’s Lustre file system has recently undergone a major upgrade from version 1.8 to the most recent 2.8 release, and the capacity increased to over 3 PB. Lustre is an open source, POSIX compatible, clustered file system presented to the Grid using the StoRM Storage Resource Manager. The motivation and benefits of upgrading including hardware and software choices are discussed. The testing, performance tuning and data migration procedure are outlined as is the source code modifications needed for StoRM compatibility. Benchmarks and real world performance are presented and future plans discussed.

Tertiary Keyword (Optional):
High performance computing

Secondary Keyword (Optional):
Distributed data handling

Primary Keyword (Mandatory):
Storage systems
XRootD Popularity on Hadoop Clusters

Author: Marco Meoni

Co-authors: Domenico Giordano; Luca Menichetti; Nicolo Magini; Tommaso Boccali

1 Universita di Pisa & INFN (IT)
2 CERN
3 Fermi National Accelerator Lab. (US)

The CMS experiment has implemented a computing model where distributed monitoring infrastructures are collecting any kind of data and metadata about the performance of the computing operations. This data can be probed further by harnessing Big Data analytics approaches and discovering patterns and correlations that can improve the throughput and the efficiency of the computing model.

CMS has already begun to store a large set of operational data - user activities, job submissions, resources, file transfers, site efficiencies, software releases, network traffic, machine logs - in a Hadoop cluster. This offers the ability to run fast arbitrary query on the data and test several computing MapReduce-based frameworks.

In this work we analyze the XrootD logs collected in Hadoop through Gled and Flume and we benchmark their aggregation at the level of dataset for monitoring purpose of popularity queries, thus proving how dashboard and monitoring systems can benefit from Hadoop parallelism. Processing time on existing Oracle DBMS of XrootD time-series logs does not scale linearly with data volume. Conversely, Big Data architectures do and make it very effective re-processing any user-defined time interval. The entire set of existing Oracle queries is replicated in the Hadoop data store and result validation is performed accordingly.

These results constitute the set of features on top of which a mining platform is designed to predict the popularity of a new dataset, the best location for replicas or the proper amount of CPU and storage in future timeframes. Learning techniques applied to Big Data architectures are extensively explored to study the correlations between aggregated data and seek for patterns in the CMS computing ecosystem. Examples of this kind are primarily represented by operational information like file access statistics or dataset attributes, which are organized in samples suitable for feeding several classifiers.

Secondary Keyword (Optional):
Databases

Primary Keyword (Mandatory):
Analysis tools and techniques

The Vacuum Platform

Author: Andrew McNab

1 University of Manchester
This paper describes GridPP’s Vacuum Platform for managing virtual machines (VMs), which has been used to run production workloads for WLCG, other HEP experiments, and some astronomy projects. The platform provides a uniform interface between VMs and the sites they run at, whether the site is organised as an Infrastructure-as-a-Service cloud system such as OpenStack with a push model, or an Infrastructure-as-a-Client system such as Vac with a pull model. The paper describes our experience in using this platform, in developing and operating VM lifecycle managers Vac and Vcycle, and in interacting with VMs provided by LHCb, ATLAS, CMS, and the GridPP DIRAC service to run production workloads.

The Machine/Job Features mechanism

Author: Andrew McNab

Co-authors: Igor Sfiligoi; Jan Justinus Keijser; Manfred Alef; Stefan Roiser; Tony Cass; Ulrich Schwickerath

1 University of Manchester
2 University of California San Diego
3 Karlsruhe Institute of Technology (KIT)
4 CERN

Within the HEPiX virtualization group and the WLCG MJF Task Force, a mechanism has been developed which provides access to detailed information about the current host and the current job to the job itself. This allows user payloads to access meta information, independent of the current batch system or virtual machine model. The information can be accessed either locally via the filesystem on a worker node, or remotely via HTTP(S) from a webserver. This paper describes the final version of the specification from 2016 which was published as an HEP Software Foundation technical note, and the design of the implementations of this version for batch and virtual machine platforms. We discuss early experiences with these implementations and how they can be exploited by experiment frameworks.
Track 6: Infrastructures / 179

Integration of grid and local batch system resources at DESY

Authors: Andreas Gellrich\(^1\); Birgit Lewendel\(^2\); Christoph Beyer\(^3\); Thomas Finnern\(^1\); Thomas Hartmann\(^5\); Yves Kemp\(^2\)

\(^1\) DESY  
\(^2\) Deutsches Elektronen-Synchrotron (DE)  
\(^3\) DESY HH

Corresponding Authors: birgit.lewendel@desy.de, thomas.finnern@desy.de, christoph.beyer@desy.de, thomas.hartmann@cern.ch, yves.kemp@cern.ch, andreas.gellrich@desy.de

We present the consolidated batch system at DESY. As one of the largest resource centres DESY has to support differing work flows by HEP experiments in WLCG or Belle II as well as local users. By abandoning specific worker node setups in favour of generic flat nodes with middleware resources provided via CVMFS, we gain flexibility to subsume different use cases in a homogeneous environment.

Grid jobs and the local batch system are managed in a HTCondor based setup, accepting pilot, user and containerized jobs. The unified setup allows dynamic re-assignement of resources between the different use cases. Furthermore, overspill to external cloud resources is investigated as response to peak demands.

Monitoring is implemented on global batch system metrics as well as on a per job level utilizing corresponding cgroup information.

Primary Keyword (Mandatory): Computing facilities

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

Track 3: Distributed Computing / 180

Web Proxy Auto Discovery for WLCG

Author: Dave Dykstra\(^1\)

Co-authors: Alastair Dewhurst \(^2\); Alessandro De Salvo \(^3\); Barry Jay Blumenfeld \(^4\); Jakob Blomer \(^5\); Vassil Verguilov \(^6\)

\(^1\) Fermi National Accelerator Lab. (US)  
\(^2\) STFC - Rutherford Appleton Lab. (GB)  
\(^3\) Universita e INFN, Roma I (IT)  
\(^4\) Johns Hopkins University (US)  
\(^5\) CERN  
\(^6\) Bulgarian Academy of Sciences (BG)

Corresponding Authors: vassil.verguilov@cern.ch, bjb@jhu.edu, jakob.blomer@cern.ch, alessandro.de.salvo@cern.ch, alastair.dewhurst@cern.ch, dwd@fnal.gov

All four of the LHC experiments depend on web proxies (that is, squids) at each grid site in order to support software distribution by the CernVM FileSystem (CVMFS). CMS and ATLAS also use web proxies for conditions data distributed through the Frontier Distributed Database caching system. ATLAS & CMS each have their own methods for their grid jobs to find out which web proxy to use for Frontier at each site, and CVMFS has a third method. Those diverse methods limit usability
and flexibility, particularly for opportunistic use cases. This paper describes a new Worldwide LHC Computing Grid (WLCG) system for discovering the addresses of web proxies that is based on an internet standard called Web Proxy Auto Discovery (WPAD). WPAD is in turn based on another standard called Proxy Auto Configuration (PAC) files. Both the Frontier and CVMFS clients support this standard. The input into the WLCG system comes from squids registered by sites in the Grid Configuration Database (GOCDB) and the OSG Information Management (OIM) system, combined with some exceptions manually configured by people from ATLAS and CMS who participate in WLCG squid monitoring. Central WPAD servers at CERN respond to http requests from grid nodes all over the world with a PAC file describing how grid jobs can find their web proxies, based on IP addresses matched in a database that contains the IP address ranges registered to organizations. Large grid sites are encouraged to supply their own WPAD web servers for more flexibility, to avoid being affected by short term long distance network outages, and to offload the WPAD servers at CERN. The CERN WPAD servers additionally support requests from jobs running at non-grid sites (particularly for LHC@Home) which it directs to the nearest publicly accessible web proxy servers. The responses to those requests are based on a separate database that maps IP addresses to longitude and latitude.

Secondary Keyword (Optional):
Distributed data handling

Primary Keyword (Mandatory):
Distributed workload management

Tertiary Keyword (Optional):

Posters B / Break / 181

Visualization of historical data for the ATLAS detector controls

Author: Julian Piotr Maciejewski

Co-authors: Charilaos Tsarouchas; Stefan Schlenker

1 CERN
2 EOS commerce

Corresponding Authors: julian.maciejewski@cern.ch, tsarouch@gmail.com, stefan.schlenker@cern.ch

The ATLAS experiment is one of four detectors located on the Large Hardon Collider (LHC) based at CERN. Its detector control system (DCS) stores the slow control data acquired within the back-end of distributed WinCC OA applications. The data can be retrieved for future analysis, debugging and detector development in an Oracle relational database.

The ATLAS DCS Data Viewer (DDV) is a client-server application providing access to the historical data outside the experiment network. The server builds optimized SQL queries, retrieves the data from the database and serves it to the clients via HTTP connections. The server also implements protection methods to prevent malicious use of the database.

The client is an AJAX-type web application based on the Google Web Toolkit (GWT) which gives users the possibility to access the data with ease. The DCS metadata can be selected using a column-tree navigation or a search engine supporting regular expressions. The data is visualised by a selection of output modules such as a java script value-over time plot or a lazy loading table widget. Additional plugins give the users the possibility to retrieve data in ROOT format or as ASCII file. Control system alarms can be visualized in a dedicated table. Python mock-up scripts can be generated by the client, allowing the user to query the pythonic DDV server directly, such that the users can embed the scripts into more complex analysis programs. Users can store searches and output configurations as XML on the server to share with others by URL or embed in HTML.
In a recent major release of DDV, the code was migrated to use the Vaadin Java Web framework for the user interface implementation, greatly improving aspects such as browsers and platform independence. The update has helped with reducing development and maintenance timescales. In addition, the tool now supports and visualizes metadata evolution which allows users to access the data consistently over decades. It is able to trace changes of hardware mappings or changes resulting from back-end software migration/restructuring. Furthermore, users can use DDV to review database insert rates e.g. to spot elements causing excessive database storage consumption. The client now also provides each user a usage history which is stored on the server allowing quick access to previously used configurations. Finally, the application has been generalised to be compatible with any other WinCC OA based RDB archive which allowed it to be set up for other control systems of the CERN accelerator infrastructure without any additional development.

Tertiary Keyword (Optional):

Secondary Keyword (Optional): Databases

Primary Keyword (Mandatory): Control systems

Track 1: Online Computing / 182

The detector read-out in ALICE during Run 3 and 4

Author: Filippo Costa¹

¹ CERN

Corresponding Author: filippo.costa@cern.ch

ALICE, the general purpose, heavy ion collision detector at the CERN LHC is designed to study the physics of strongly interacting matter using proton-proton, nucleus-nucleus and proton-nucleus collisions at high energies. The ALICE experiment will be upgraded during the Long Shutdown 2 in order to exploit the full scientific potential of the future LHC. The requirements will then be significantly different from what they were during the original design of the experiment and will require major changes to the detector read-out.

The main physics topics addressed by the ALICE upgrade are characterized by rare processes with a very small signal-to-background ratio, requiring very large statistics of fully reconstructed events. In order to keep up with the 50 kHz interaction rate, the upgraded detectors will be read out continuously. However, triggered read-out will be used by some detectors and for commissioning and some calibration runs.

The total data volume collected from the detectors will increase significantly reaching a sustained data throughput of up to 3 TB/s with the zero-suppression of the TPC data performed after the data transfer to the detector read-out system. A flexible mechanism of bandwidth throttling will allow the system to gracefully degrade the effective rate of recorded interactions in case of saturation of the computing system.

This paper includes a summary of these updated requirements and presents a refined design of the detector read-out and of the interface with the detectors and the online systems. It also elaborates on the system behaviour in continuous and triggered readout and defines ways to throttle the data read-out in both cases.

Tertiary Keyword (Optional):

Data processing workflows and frameworks/pipelines

Secondary Keyword (Optional): Trigger
Grid Access with Federated Identities

Authors: Dave Dykstra¹; Jeny Teheran¹; Mine Altunay¹

Co-authors: Dennis Box¹; Kenneth Richard Herner²; Neha Sharma¹; Tanya Levshina¹

¹ Fermi National Accelerator Laboratory
² Fermi National Accelerator Laboratory (US)

Corresponding Authors: neha@fnal.gov, maltunay@fnal.gov, dbox@fnal.gov, tlevshin@fnal.gov, jteheran@fnal.gov, dwd@fnal.gov, kenneth.richard.herner@cern.ch

It is well known that submitting jobs to the grid and transferring the resulting data are not trivial tasks, especially when users are required to manage their own X.509 certificates. Asking users to manage their own certificates means that they need to keep the certificates secure, remember to renew them periodically, frequently create proxy certificates, and make them available to long-running grid jobs. We have made those tasks easier by creating and managing certificates for users. In order to do this we have written a new general purpose open source tool called 'cigetcert' that takes advantage of the existing InCommon federated identity infrastructure and the InCommon X.509 certificate creation service, CILogon. The tool uses the SAML Enhanced Client or Proxy (ECP) profile protocol which was designed for non-web browser environments, so it fits well with traditional command line-based grid access. The tool authenticates with the local institution’s Identity Provider (IdP) using either Kerberos or the institutional username/password, retrieves a user certificate from CILogon Basic CA, stores a relatively short-lived proxy certificate on the local disk, and stores a longer-lived proxy certificate in a MyProxy server. The local disk proxy certificate is then available to submit jobs, and the grid job submission system reads the proxy certificate out of the MyProxy server and uses that to authorize data transfers for long-lived grid jobs. This paper describes the motivation, design, implementation, and deployment of this system that provides grid access with federated identities.

Secondary Keyword (Optional):
Distributed workload management

Primary Keyword (Mandatory):
Security and policies

Tertiary Keyword (Optional):

Recent progress of Geant4 electromagnetic physics for LHC and other applications
In this work we report on recent progress of the Geant4 electromagnetic (EM) physics sub-packages. A number of new interfaces and models recently introduced are already used in LHC applications and may be useful for any type of simulation.

To improve usability, a new set of User Interface (UI) commands and corresponding C++ interfaces have been added for easier configuration of EM physics. In particular, photo-absorption ionisation model may be enabled per detector region using corresponding UI command. Also low-energy limit for charged particle tracking may be selected, because this low-energy limit is now a new EM parameter making easier user control.

Significant developments were carried out for the modeling of single and multiple scattering of charged particles. Corrections to scattering of positrons and to sampling of displacement have been recently added to the Geant4 default Urban model. The Goudsmit-Saunderson (GS) model was reviewed and re-written. The advantage of this model is that it is fully theory based. This new variant demonstrates both equivalent physics performance for simulation of thin target experiments compared to the Urban model in its best accuracy configuration together with good CPU performances. For testing purposes we provide a configuration of electron scattering based on the GS model instead of the Urban model. In addition, another fully theory-based model for single scattering of electrons with Mott correction has been introduced. This model is an important tool to study performance of tracking devices and cross validation of multiple scattering models.

In this report, we will also present developments of EM model in view of the simulation for the new FCC facility. The simulation of EM processes is important for optimization of FCC interaction region and for study of various conceptions of FCC detectors. This requires an extension of validity of EM models for energies higher than the ones used for LHC experiments. The current results and limitations will be discussed.

Important developments were also recently carried out in low-energy EM models, which may be of interest to various application domains. In particular, a possibility to simulate full Auger cascades and a new version of polarized Compton scattering were added. Regarding the very-low-energy regime, new cross sections models for an accurate tracking of electrons in liquid water were also implemented.

These developments are included in the recent Geant4 10.2 release and in the new development version 10.3beta.

Tertiary Keyword (Optional):
Simulation

Secondary Keyword (Optional):
Parallelization

Primary Keyword (Mandatory):
Simulation

Scaling Up a CMS Tier-3 Site with Campus Resources and a 100 Gb/s Network Connection: What Could Go Wrong?
Authors: Anna Elizabeth Woodard¹; Matthias Wolf²

Co-authors: Benjamin Tovar ²; Douglas Thain ²; Kenyi Paolo Hurtado Anampa ¹; Kevin Patrick Lannon ¹; Mike Hildreth ¹; Patrick Donnelly ²; Paul Brenner ²

¹ University of Notre Dame (US)
² University of Notre Dame

Corresponding Authors: anna.elizabeth.woodard@cern.ch, pdonnel3@nd.edu, m+cms@sushinara.net, btovar@nd.edu, paul.r.brenner@nd.edu, klannon@nd.edu, dthain@cse.nd.edu, kenyi.paolo.hurtado.anampa@cern.ch, mikeh@omega.hep.nd.edu

The University of Notre Dame (ND) CMS group operates a modest-sized Tier-3 site suitable for local, final-stage analysis of CMS data. However, through the ND Center for Research Computing (CRC), Notre Dame researchers have opportunistic access to roughly 25k CPU cores of computing and a 100 Gb/s WAN network link. To understand the limits of what might be possible in this scenario, we undertook to use these resources for a wide range of CMS computing tasks from user analysis through large-scale Monte Carlo production (including both detector simulation and data reconstruction.) We will discuss the challenges inherent in effectively utilizing CRC resources for these tasks and the solutions deployed to overcome those challenges. We will also discuss current performance and potential for future refinements as well as interactions with the broader CMS computing infrastructure.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):
Computing facilities

Posters B / Break / 186

Facilitating the deployment and exploitation of HEP Phenomenology codes using INDIGO-Datacloud tools

Authors: Alvaro Lopez Garcia¹; Emanuele Angelo Bagnaschi²; Isabel Campos Plasencia³; Pablo Orviz Fernandez¹; Sven Heinemeyer⁴

¹ CSIC
² DESY Hamburg
³ Consejo Superior de Investigaciones Cientificas (CSIC) (ES)
⁴ CSIC (Santander, ES)

Corresponding Authors: aloga@ifca.unican.es, isabel.campos@csic.es, pablo.orviz.fernandez@cern.ch, sven.heinemeyer@cern.ch, emanuele.bagnaschi@desy.de

The MasterCode collaboration (http://cern.ch/mastercode) is concerned with the investigation of supersymmetric models that go beyond the current status of the Standard Model of particle physics. It involves teams from CERN, DESY, Fermilab, SLAC, CSIC, INFN, NIKHEF, Imperial College London, King’s College London, the Universities of Amsterdam, Antwerpen, Bristol, Minnesota and ETH Zurich.

Within the MasterCode collaboration, state-of-the-art HEP Phenomenology codes are consistently combined to provide the most precise prediction for supersymmetric models to be confronted with experimental data.

Generally speaking, for the type of software developed in HEP Phenomenology, there is a lack of tools to enable the easy development and deployment of applications. Phenomenology applications
have many dependences in terms of libraries and compilers that makes it difficult to deploy on traditional batch clusters due to system software version conflicts and related issues.

In this work we propose a framework based on the developments of the project INDIGO-Datacloud to fill this gap. In particular such developments allow us to easily build, modify, distribute and run Mastercode in containerized form over multiple Cloud infrastructures.

Other advanced capabilities imply running Mastercode through dynamically instantiated batch systems provisioned by general computing infrastructures. In particular, making possible the automatic handling of parametric runs (parameter sweeping).

Such an advanced computing framework has the potential of speeding up the phases of development and deployment of complex scientific software used in our research, with the corresponding impact in the results.

The development of Mastercode is supported by the ERC Advanced Grant 267352, "Exploring the Terauniverse with the LHC, Astrophysics and Cosmology" led by Prof. John Ellis.

INDIGO-Datacloud receives funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement RIA 653549.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Cloud technologies

Primary Keyword (Mandatory):
Software development process and tools

Track 7: Middleware, Monitoring and Accounting / 187

C2MON: a modern open-source platform for data acquisition, monitoring and control

Author: Justin Lewis Salmon

Co-authors: Anna Suwalska; Matthias Braeger

1 CERN

Corresponding Authors: justin.lewis.salmon@cern.ch, matthias.braeger@cern.ch, anna.suwalska@cern.ch

The CERN Control and Monitoring Platform (C2MON) is a modular, clusterable framework designed to meet a wide range of monitoring, control, acquisition, scalability and availability requirements. It is based on modern Java technologies and has support for several industry-standard communication protocols. C2MON has been reliably utilised for several years as the basis of multiple monitoring systems at CERN, including the Technical Infrastructure Monitoring (TIM) service and the DIAgnostics and MONitoring (DIAMON) service. The central Technical Infrastructure alarm service for the accelerator complex (LASER) is in the final migration phase. Furthermore, three more services at CERN are currently being prototyped with C2MON.

Until now, usage of C2MON has been limited to internal CERN projects. However, C2MON is trusted and mature enough to be made publically available. Aiming to build a user community, encourage collaboration with external institutes and create industry partnerships, the C2MON platform will be distributed as an open-source package under the LGPLv3 licence within the context of the knowledge transfer initiative at CERN.
This paper gives an overview of the C2MON platform focusing on its ease of use, integration with modern technologies, and its other features such as standards-based web support and flexible archiving techniques. The challenges faced when preparing an in-house platform for general release to external users are also described.

**Tertiary Keyword (Optional):**
DAQ

**Secondary Keyword (Optional):**
Outreach

**Primary Keyword (Mandatory):**
Monitoring

---

**Posters A / Break / 188**

**ARC CE cache as a solution for lightweight Grid sites in ATLAS**

**Author:** David Cameron¹

¹ University of Oslo (NO)

**Corresponding Author:** david.cameron@cern.ch

Many Grid sites have the need to reduce operational manpower, and running a storage element consumes a large amount of effort. In addition, setting up a new Grid site including a storage element involves a steep learning curve and large investment of time. For these reasons so-called storage-less sites are becoming more popular as a way to provide Grid computing resources with less operational overhead. ARC CE is a widely-used and mature Grid middleware which was designed from the start to be used on sites with no persistent storage element. Instead, it maintains a local self-managing cache of data which retains popular data for future jobs. As the cache is simply an area on a local posix shared filesystem with no external-facing service, it requires no extra maintenance. The cache can be scaled up as required by increasing the size of the filesystem or adding new filesystems. This paper describes how ARC CE and its cache are an ideal solution for lightweight Grid sites in the ATLAS experiment, and the integration of the ARC CE cache and the ATLAS data management system.

**Tertiary Keyword (Optional):**

**Secondary Keyword (Optional):**
Computing facilities

**Primary Keyword (Mandatory):**
Distributed data handling

---

**Track 3: Distributed Computing / 189**
Exploiting Opportunistic Resources for ATLAS with ARC CE and the Event Service

Author: David Cameron

1 University of Oslo (NO)

Corresponding Author: david.cameron@cern.ch

With ever-greater computing needs and fixed budgets, big scientific experiments are turning to opportunistic resources as a means to add much-needed extra computing power. These resources can be very different in design from the resources that comprise the Grid computing of most experiments, therefore exploiting these resources requires a change in strategy for the experiment. The resources may be highly restrictive in what can be run or in connections to the outside world, or tolerate opportunistic usage only on condition that tasks may be terminated without warning. The ARC CE with its non-intrusive architecture is designed to integrate resources such as High Performance Computing (HPC) systems into a computing Grid. The ATLAS experiment developed the Event Service primarily to address the issue of jobs that can be terminated at any point when opportunistic resources are needed by someone else.

This paper describes the integration of these two systems in order to exploit opportunistic resources for ATLAS in a restrictive environment. In addition to the technical details, results from deployment of this solution in the SuperMUC HPC in Munich are shown.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

High performance computing

Primary Keyword (Mandatory):

Distributed workload management

Evolution of user analysis on the Grid in ATLAS

Author: Federica Legger

1 Ludwig-Maximilians-Univ. Muench (DE)

Corresponding Author: federica.legger@physik.uni-muenchen.de

More than one thousand physicists analyse data collected by the ATLAS experiment at the Large Hadron Collider (LHC) at CERN through 150 computing facilities around the world. Efficient distributed analysis requires optimal resource usage and the interplay of several factors: robust grid and software infrastructures, and system capability to adapt to different workloads. The continuous automatic validation of grid sites and the user support provided by a dedicated team of expert shifters have been proven to provide a solid distributed analysis system for ATLAS users.

Based on the experience from the first run of the LHC, substantial improvements to the ATLAS computing system have been made to optimize both production and analysis workflows. These include the re-design of the production and data management systems, a new analysis data format and event model, and the development of common reduction and analysis frameworks. The impact
of such changes on the distributed analysis system is evaluated. More than 100 million user jobs in the period 2015-2016 are analysed for the first time with analytics tools such as Elastic Search. Typical user workflows and their associated metrics are studied and the improvement in the usage of distributed resources due to the common analysis data format and the reduction framework is assessed. Measurements of user job performance and typical requirements are also shown.

Tertiary Keyword (Optional):
Distributed data handling

Secondary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Primary Keyword (Mandatory):
Computing models

Track 2: Offline Computing / 191

AthenaMT: Upgrading the ATLAS Software Framework for the Many-Core World with Multi-Threading

Author: Charles Leggett¹

¹ Lawrence Berkeley National Lab. (US)

Corresponding Author: charles.g.leggett@gmail.com

ATLAS’s current software framework, Gaudi/Athena, has been very successful for the experiment in LHC Runs 1 and 2. However, its single threaded design has been recognised for some time to be increasingly problematic as CPUs have increased core counts and decreased available memory per core. Even the multi-process version of Athena, AthenaMP, will not scale to the range of architectures we expect to use beyond Run2.

After concluding a rigorous requirements phase, where many design components were examined in detail, ATLAS has begun the migration to a new data-flow driven, multi-threaded framework, which enables the simultaneous processing of singleton, thread unsafe legacy Algorithms, cloned Algorithms that execute concurrently in their own threads with different Event contexts, and fully re-entrant, thread safe Algorithms.

In this paper we will report on the process of modifying the framework to safely process multiple concurrent events in different threads, which entails significant changes in the underlying handling of features such as event and time dependent data, asynchronous callbacks, metadata, integration with the Online High Level Trigger for partial processing in certain regions of interest, concurrent I/O, as well as ensuring thread safety of core services. We will also report on the migration of user code to the new framework, including that of upgrading select Algorithms to be fully re-entrant.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Parallelizarion

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines
An Oracle-based Event Index for ATLAS

Author: Elizabeth Gallas

Co-authors: Andrea Formica; Andrei Dumitru; Gancho Dimitrov; Luca Canali; Petya Tsvetanova Petrova; Zbigniew Baranowski

1 University of Oxford (GB)
2 CEA/IRFU, Centre d’etude de Saclay Gif-sur-Yvette (FR)
3 CERN
4 University of Texas at Arlington (US)

Corresponding Authors: elizabeth.gallas@physics.ox.ac.uk, luca.canali@cern.ch, petya.petrova@cern.ch, gancho.dimitrov@cern.ch, andrei.dumitru@cern.ch, andrea.formica@cern.ch, zbigniew.baranowski@cern.ch

The ATLAS EventIndex System has amassed a set of key quantities for a large number of ATLAS events into a Hadoop based infrastructure for the purpose of providing the experiment with a number of event-wise services. Collecting this data in one place provides the opportunity to investigate various storage formats and technologies and assess which best serve the various use cases as well as consider what other benefits alternative storage systems provide.

In this presentation we describe how the data are imported into an Oracle RDBMS, the services we have built based on this architecture, and our experience with it. We’ve indexed about 26 billion real data events thus far and have designed the system to accommodate future data which has expected rates of 5 and 20 billion events per year. We have found this system offers outstanding performance for some fundamental use cases. In addition, profiting from the co-location of this data with other complementary metadata in ATLAS, the system has been easily extended to perform essential assessments of data integrity and completeness and to identify event duplication, including at what step in processing the duplication occurred.

Secondary Keyword (Optional):
Storage systems

Primary Keyword (Mandatory):
Databases

Tertiary Keyword (Optional):

Collecting conditions usage metadata to optimize current and future ATLAS software and processing

Co-authors: Andrea Formica; Dario Barberis; Elizabeth Gallas; Grigori Rybkin; Lorenzo Rinaldi; Monica Verducci; Susumu Oda

1 CEA/IRFU, Centre d’etude de Saclay Gif-sur-Yvette (FR)
2 Università e INFN Genova (IT)
3 University of Oxford (GB)
4 Laboratoire de l’Accélérateur Lineaire (FR)
5 Universita e INFN, Bologna (IT)
6 Universita e INFN, Roma I (IT)
7 Kyushu University (JP)
Conditions data (for example: alignment, calibration, data quality) are used extensively in the processing of real and simulated data in ATLAS. The volume and variety of the conditions data needed by different types of processing are quite diverse, so optimizing its access requires a careful understanding of conditions usage patterns. These patterns can be quantified by mining representative log files from each type of processing and gathering detailed information about conditions usage for that type of processing into a central repository.

In this presentation, we describe the systems developed to collect this conditions usage metadata per job type and describe a few specific (but very different) ways in which it has been used. For example, it can be used to cull specific conditions data into a much more compact package to be used by jobs doing similar types of processing; these customized collections can then be shipped with jobs to be executed on isolated worker nodes (such as HPC farms) that have no network access to conditions. Another usage is in the design of future ATLAS software: to provide Run 3 software developers essential information about the nature of current conditions accessed by software. This helps to optimize internal handling of conditions data to minimize its memory footprint while facilitating access to this data by the sub-processes that need it.

**Tertiary Keyword (Optional):**
Software development process and tools

**Secondary Keyword (Optional):**
Distributed data handling

**Primary Keyword (Mandatory):**
Databases

---

**Integration of the Titan supercomputer at OLCF with the ATLAS Production System**

**Author:** Sergey Panitkin

1 *Brookhaven National Laboratory (US)*

**Corresponding Author:** panitkin@bnl.gov

The PanDA (Production and Distributed Analysis) workload management system was developed to meet the scale and complexity of distributed computing for the ATLAS experiment. PanDA managed resources are distributed worldwide, on hundreds of computing sites, with thousands of physicists accessing hundreds of Petabytes of data and the rate of data processing already exceeds Exabyte per year.

While PanDA currently uses more than 200,000 cores at well over 100 Grid sites, future LHC data taking runs will require more resources than Grid computing can possibly provide. Additional computing and storage resources are required. Therefore ATLAS is engaged in an ambitious program to expand the current computing model to include additional resources such as the opportunistic use of supercomputers.

In this talk we will describe a project aimed at integration of ATLAS Production System with Titan supercomputer at Oak Ridge Leadership Computing Facility (OLCF). Current approach utilizes modified PanDA Pilot framework for job submission to Titan’s batch queues and local data management, with lightweight MPI wrappers to run single node workloads in parallel on Titan’s multi-core worker nodes. It provides for running of standard ATLAS production jobs on unused resources (backfill) on Titan.

The system already allowed ATLAS to collect on Titan millions of core-hours per month, execute hundreds of thousands jobs, while simultaneously improving Titan’s utilization efficiency.
We will discuss details of the implementation, current experience with running the system, as well as future plans aimed at improvements in scalability and efficiency.

Secondary Keyword (Optional):
Distributed workload management

Primary Keyword (Mandatory):
High performance computing

First use of LHC Run 3 Conditions Database infrastructure for auxiliary data files in ATLAS

Co-authors: Andrea Formica ¹; Elizabeth Gallas ²; Grigori Rybkin ³; Lorenzo Rinaldi ⁴; Ludovica Aperio Bella ⁵; Shaun Roe ⁶; Will Buttinger ⁷

¹ CEA/IRFU,Centre d'étude de Saclay Gif-sur-Yvette (FR)
² University of Oxford (GB)
³ Laboratoire de l’Accélérateur Lineaire (FR)
⁴ Universita e INFN, Bologna (IT)
⁵ University of Birmingham (GB)
⁶ CERN
⁷ Brookhaven National Laboratory (US)

Corresponding Authors: elizabeth.gallas@physics.ox.ac.uk, andrea.formica@cern.ch, shaun.roe@cern.ch, grigori.rybkine@cern.ch, lorenzo.rinaldi@cern.ch, will@cern.ch, ludovica.aperio.bella@cern.ch

Processing of the large amount of data produced by the ATLAS experiment requires fast and reliable access to what we call Auxiliary Data Files (ADF). These files, produced by Combined Performance, Trigger and Physics groups, contain conditions, calibrations, and other derived data used by the ATLAS software. In ATLAS this data has, thus far for historical reasons, been collected and accessed outside the ATLAS Conditions Database infrastructure and related software. For this reason, along with the fact that ADF data is effectively read by the software as binary objects, makes this class of data ideal for testing the proposed Run 3 Conditions data infrastructure now in development. This paper will describe this implementation as well as describe the lessons learned in exploring and refining the new infrastructure with the potential for deployment during Run 2.

Tertiary Keyword (Optional):
Distributed data handling

Secondary Keyword (Optional):
Storage systems

Primary Keyword (Mandatory):
Databases

Track 2: Offline Computing / 196
Multi-threaded ATLAS Simulation on Intel Knights Landing Processors

Author: Steven Andrew Farrell

1 Lawrence Berkeley National Lab. (US)

Corresponding Author: steven.farrell@cern.ch

The Knights Landing (KNL) release of the Intel Many Integrated Core (MIC) Xeon Phi line of processors is a potential game changer for HEP computing. With 72 cores and deep vector registers, the KNL cards promise significant performance benefits for highly-parallel, compute-heavy applications. Cori, the newest supercomputer at the National Energy Research Scientific Computing Center (NERSC), will be delivered to its users in two phases with the first phase online now and the second phase expected in mid-2016. Cori Phase 2 will be based on the KNL architecture and will contain over 9000 compute nodes with 96GB DDR4 memory. ATLAS simulation with the multithreaded Athena Framework (AthenaMT) is a great use-case for the KNL architecture and supercomputers like Cori. Simulation jobs have a high ratio of CPU computation to disk I/O and have been shown to scale well in multi-threading and across many nodes. In this presentation we will give an overview of the ATLAS simulation application with details on its multi-threaded design. Then, we will present a performance analysis of the application on KNL devices and compare it to a similar study done on a traditional x86 platform (NERSC’s Edison) to demonstrate the capabilities of the architecture and evaluate the benefits of utilizing KNL platforms like cori for ATLAS production.

Tertiary Keyword (Optional):
Processor architectures

Secondary Keyword (Optional):
Parallelizarion

Primary Keyword (Mandatory):
Simulation

The Fast Simulation Chain for ATLAS

Author: Zachary Louis Marshall

1 Lawrence Berkeley National Lab. (US)

Corresponding Author: zach.marshall@cern.ch

In order to generate the huge number of Monte Carlo events that will be required by the ATLAS experiment over the next several runs, a very fast simulation is critical. Fast detector simulation alone, however, is insufficient: with very high numbers of simultaneous proton-proton collisions expected in Run 3 and beyond, the digitization (detector response emulation) and event reconstruction time quickly become comparable to the time required for detector simulation. The ATLAS Fast Chain simulation has been developed to solve this problem. Modules are implemented for fast simulation, fast digitization, and fast track reconstruction. The application is sufficiently fast – several orders of magnitude faster than the standard simulation – that the simultaneous proton-proton collisions can be generated during the simulation job, so Pythia8 also runs concurrently with the rest of the algorithms. The Fast Chain has been built to be extremely modular and flexible, so that each sample can be custom-tailored to match the resource and modeling accuracy needs of an analysis. It is ideally suited for analysis templating, systematic uncertainty evaluation, signal parameter space scans, simulation with alternative detector configurations (e.g. upgrade), among other applications.

Tertiary Keyword (Optional):
Integration of the Chinese HPC Grid in ATLAS Distributed Computing

Author: Andrej Filipcic

Jozef Stefan Institute (SI)

Corresponding Author: andrej.filipcic@ijs.si

Fifteen Chinese High Performance Computing sites, many of them on the TOP500 list of most powerful supercomputers, are integrated into a common infrastructure providing coherent access to a user through an interface based on a RESTful interface called SCEAPI. These resources have been integrated into the ATLAS Grid production system using a bridge between ATLAS and SCEAPI which translates the authorization and job submission protocols between the two environments. The ARC Computing Element (ARC CE) forms the bridge using an extended batch system interface to allow job submission to SCEAPI. The ARC CE was setup at the Institute for High Energy Physics, Beijing, in order to be as close as possible to the SCEAPI front-end interface at the Computing Network Information Center, also in Beijing. This paper describes the technical details of the integration between ARC CE and SCEAPI and presents results so far with two supercomputer centers, Tianhe-IA and ERA. These two centers have been the pilots for ATLAS Monte Carlo Simulation in SCEAPI and have been providing CPU power since fall 2015.

How to keep the Grid full and working with ATLAS production and physics jobs

Author: Andreu Pacheco Pages

Institut de Física d'Altes Energies - Barcelona (ES)

Corresponding Author: pacheco@ifae.es

The ATLAS production system has provided the infrastructure to process of tens of thousand of events during LHC Run1 and the first year of the LHC Run2 using grid, clouds and high performance computing. We address in this contribution all the strategies and improvements added to the production system to optimize its performance to get the maximum efficiency of available resources from operational perspective and focussing in detail in the recent developments.
A study of data representations in Hadoop to optimize data storage and search performance of the ATLAS EventIndex

Authors: Dario Barberis\(^1\); Julius Hrivnac\(^2\); Luca Canali\(^3\); Rainer Toebbicke\(^3\); Zbigniew Baranowski\(^3\)

\(^1\) Università e INFN Genova (IT)
\(^2\) Université de Paris-Sud 11 (FR)
\(^3\) CERN

Corresponding Authors: dario.barberis@cern.ch, rainer.toebbicke@cern.ch, julius.hrivnac@cern.ch, luca.canali@cern.ch, zbigniew.baranowski@cern.ch

This paper reports on the activities aimed at improving the architecture and performance of the ATLAS EventIndex implementation in Hadoop. The EventIndex contains tens of billions event records, each of which consisting of ~100 bytes, all having the same probability to be searched or counted. Data formats represent one important area for optimizing the performance and storage footprint of applications based on Hadoop. This work reports on the production usage and on tests using several data formats including Map Files, Apache Parquet, Avro, and various compression algorithms. The query engine plays also a critical role in the architecture. This paper reports on the use of HBase for the EventIndex, focussing on the optimizations performed in production and on the scalability tests. Additional engines that have been tested include Cloudera Impala, in particular for its SQL interface, and the optimizations for data warehouse workloads and reports.

AGIS: Integration of new technologies used in ATLAS Distributed Computing

Author: Alexey Anisenkov\(^1\)

\(^1\) Budker Institute of Nuclear Physics (RU)

Corresponding Author: alexey.anisenkov@cern.ch
The variety of the ATLAS Distributed Computing infrastructure requires a central information system to define the topology of computing resources and to store the different parameters and configuration data which are needed by the various ATLAS software components.

The ATLAS Grid Information System (AGIS) is the system designed to integrate configuration and status information about resources, services and topology of the computing infrastructure used by ATLAS Distributed Computing applications and services. Being an intermediate middleware system between clients and external information sources (like central BDII, GOCDB, MyOSG), AGIS defines the relations between experiment specific used resources and physical distributed computing capabilities.

Being in production during LHC Run1 AGIS became the central information system for Distributed Computing in ATLAS and it is continuously evolving to fulfill new user requests, enable enhanced operations and follow the extension of ATLAS Computing model. ATLAS Computing model and data structures used by Distributed Computing applications and services are continuously evolving and trend to fit newer requirements from ADC community. In this note, we describe the evolution and the recent developments of AGIS functionalities, related to integration of new technologies recently become widely used in ATLAS Computing like flexible computing utilization of opportunistic Cloud and HPC resources, ObjectStore services integration for Distributed Data Management (Rucio) and ATLAS workload management (PanDA) systems, unified storage protocols declaration required for PandDA Pilot site movers and others.

The improvements of information model and general updates are also shown, in particular we explain how other Collaborations outside ATLAS could benefit the system as a computing resources information catalogue. AGIS is evolving toward a common information system not coupled to a specific experiment.

Tertiary Keyword (Optional):
Databases

Secondary Keyword (Optional):
Accounting and information

Primary Keyword (Mandatory):
Computing middleware

---

**Posters A / Break / 202**

**A Roadmap to Continuous Integration for ATLAS software development**

**Author:** Alexander Undrus

**Co-author:** Attila Krasznahorkay

**1 Brookhaven National Laboratory (US)**

**2 CERN**

**Corresponding Authors:** undrus@bnl.gov, attila.krasznahorkay@cern.ch

The ATLAS software infrastructure facilitates efforts of more than 1000 developers working on the code base of 2200 packages with 4 million C++ and 1.4 million python lines. The ATLAS offline code management system is the powerful, flexible framework for processing new package versions requests, probing code changes in the Nightly Build System, migration to new platforms and compilers, deployment of production releases for worldwide access and supporting physicists with tools and interfaces for efficient software use. It maintains multi-stream, parallel development
environment with about 70 multi-platform branches of nightly releases and provides vast opportunities for testing new packages, for verifying patches to existing software and for migrating to new platforms and compilers. The system evolution is currently aimed at the adoption of modern continuous integration (CI) practices focused on building nightly releases early and often, with rigorous unit and integration testing. This presentation describes the CI incorporation program for the ATLAS software infrastructure. It brings modern open source tools such as Jenkins and CTest into the ATLAS Nightly System, rationalizes hardware resource allocation and administrative operations, provides improved feedback and means to fix broken builds promptly for developers. Once adopted, ATLAS CI practices will improve and accelerate innovation cycles and result in increased confidence in new software deployments. The presentation reports the status of Jenkins integration with the ATLAS Nightly System as well as short and long term plans for the incorporation of CI practices.

**Primary Keyword (Mandatory):**
Software development process and tools

**Secondary Keyword (Optional):**

**Tertiary Keyword (Optional):**

---

**Posters A / Break / 204**

**ATLAS Metadata Interface (AMI), a generic metadata framework**

**Author:** Jerome Henri Fulachier

1. Centre National de la Recherche Scientifique (FR)

**Corresponding Author:** jerome.fulachier@lpsc.in2p3.fr

The ATLAS Metadata Interface (AMI) is a mature application of more than 15 years of existence. Mainly used by the ATLAS experiment at CERN, it consists of a very generic tool ecosystem for metadata aggregation and cataloguing. We briefly describe the architecture, the main services and the benefits of using AMI in big collaborations, especially for high energy physics. We focus on the recent improvements, for instance: the lightweight clients (Python, Javascript, C++), the new smart task server system and the Web 2.0 AMI framework for simplifying the development of metadata-oriented web interfaces.

**Tertiary Keyword (Optional):**
Cloud technologies

**Secondary Keyword (Optional):**
Computing middleware

**Primary Keyword (Mandatory):**
Databases

---

**Posters A / Break / 205**
Deploying the ATLAS Metadata Interface (AMI) on the cloud with Jenkins

Author: Fabian Lambert

1 Centre National de la Recherche Scientifique (FR)

Corresponding Author: fabian.lambert@lpsc.in2p3.fr

The ATLAS Metadata Interface (AMI) is a mature application of more than 15 years of existence. Mainly used by the ATLAS experiment at CERN, it consists of a very generic tool ecosystem for metadata aggregation and cataloguing. AMI is used by the ATLAS production system, therefore the service must guarantee a high level of availability. We describe our monitoring system and the Jenkins-based strategy used to dynamically test and deploy cloud OpenStack nodes on demand. Moreover, we describe how to switch to a distant replica in case of downtime.

Tertiary Keyword (Optional):
Computing middleware

Secondary Keyword (Optional):
Databases

Primary Keyword (Mandatory):
Cloud technologies

Posters A / Break / 206

ATLAS Fast Physics Monitoring: TADA

Author: Gabriele Sabato

1 Nikhef National institute for subatomic physics (NL)

Corresponding Author: gabriele.sabato@cern.ch

The ATLAS Experiment at the LHC is recording data from proton-proton collisions with 13 TeV center-of-mass energy since spring 2015. The ATLAS collaboration has set up, updated and optimized a fast physics monitoring framework (TADA) to automatically perform a broad range of validation and to scan for signatures of new physics in the rapidly growing data. TADA is designed to provide fast feedback in two or three days after the data are available. The system can monitor a huge range of physics channels, offline data quality and physics performance. TADA output is available in a, constantly updated, website accessible by the whole collaboration. Hints of potentially interesting physics signals obtained this way are followed up by the physics groups. The poster will report about the technical aspects of TADA: the software structure to obtain the input TAG files, the framework workflow and structure, the webpage and its implementation.

Tertiary Keyword (Optional):
Analysis tools and techniques

Secondary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Primary Keyword (Mandatory):
Monitoring
A new mechanism to persistify the detector geometry of ATLAS and serving it through an experiment-agnostic REST API

Author: Riccardo Maria Bianchi

Co-authors: Ilija Vukotic; Joseph Boudreau

1 University of Pittsburgh (US)
2 University of Chicago (US)
3 University of Pittsburgh

Corresponding Authors: riccardo.maria.bianchi@cern.ch, boudreau@pitt.edu, ilija.vukotic@cern.ch

The complex geometry of the whole detector of the ATLAS experiment at LHC is currently stored only in custom online databases, from which it is built on-the-fly on request. Accessing the online geometry guarantees accessing the latest version of the detector description, but requires the setup of the full ATLAS software framework “Athena”, which provides the online services and the tools to retrieve the data from the database. This operation is cumbersome and slows down the applications that need to access the geometry. Moreover, all applications that need to access the detector geometry need to be built and run on the same platform as the ATLAS framework, preventing the usage of the actual detector geometry in stand-alone applications.

Here we propose a new mechanism to persistify and serve the geometry of HEP experiments. The new mechanism is composed by a new file format and a REST API. The new file format allows to store the whole detector description locally in a flat file, and it is especially optimized to describe large complex detectors with the minimum file size, making use of shared instances and storing compressed representations of geometry transformations. On the other side, the dedicated REST API is meant to serve the geometry in standard formats like JSON, to let users and applications download specific partial geometry information.

With this new geometry persistification a new generation of applications could be developed, which can use the actual detector geometry while being platform-independent and experiment-agnostic.

Tertiary Keyword (Optional):
Visualization

Secondary Keyword (Optional):
Software development process and tools

Primary Keyword (Mandatory):
Data model

ATLAS Distributed Computing experience and performance during the LHC Run-2

Author: Andrej Filipcic
ATLAS Distributed Computing during LHC Run-1 was challenged by steadily increasing computing, storage and network requirements. In addition, the complexity of processing task workflows and their associated data management requirements led to a new paradigm in the ATLAS computing model for Run-2, accompanied by extensive evolution and redesign of the workflow and data management systems. The new systems were put into production at the end of 2014, and gained robustness and maturity during 2015 data taking. ProdSys2, the new request and task interface; JEDI, the dynamic job execution engine developed as an extension to PanDA; and Rucio, the new data management system, form the core of the Run-2 ATLAS distributed computing engine.

One of the big changes for Run-2 was the adoption of the Derivation Framework, which moves the chaotic CPU and data intensive part of the user analysis into the centrally organized train production, delivering derived AOD datasets to user groups for final analysis. The effectiveness of the new model was demonstrated through the delivery of analysis datasets to users just one week after data taking, by completing the calibration loop, Tier-0 processing and train production steps promptly. The great flexibility of the new system also makes it possible to execute part of the Tier-0 processing on the grid when Tier-0 resources experience a backlog during high data-taking periods.

The introduction of the data lifetime model, where each dataset is assigned a finite lifetime (with extensions possible for frequently accessed data), was made possible by Rucio. Thanks to this the storage crises experienced in Run-1 have not reappeared during Run-2. In addition, the distinction between Tier-1 and Tier-2 disk storage, now largely artificial given the quality of Tier-2 resources and their networking, has been removed through the introduction of dynamic ATLAS clouds that group the storage endpoint nucleus and its close-by execution satellite sites. All stable ATLAS sites are now able to store unique or primary copies of the datasets.

ATLAS Distributed Computing is further evolving to speed up request processing by introducing network awareness, using machine learning and optimization of the latencies during the execution of the full chain of tasks. The Event Service, a new workflow and job execution engine, is designed around check-pointing at the level of event processing to use opportunistic resources more efficiently.
Benefits and performance of ATLAS approaches to utilizing opportunistic resources

Author: Andrej Filipcic¹

¹ Jozef Stefan Institute (SI)

Corresponding Author: andrej.filipcic@ijs.si

ATLAS has been extensively exploring possibilities of using computing resources extending beyond conventional grid sites in the WLCG fabric to deliver as many computing cycles as possible and thereby enhance the significance of the Monte-Carlo samples to deliver better physics results.

The difficulties of using such opportunistic resources come from architectural differences such as unavailability of grid services, the absence of network connectivity on worker nodes or inability to use standard authorization protocols. Nevertheless, ATLAS has been extremely successful in running production payloads on a variety of sites, thanks largely to the job execution workflow design in which the job assignment, input data provisioning and execution steps are clearly separated and can be offloaded to custom services. To transparently include the opportunistic sites in the ATLAS central production system, several models with supporting services have been developed to mimic the functionality of a full WLCG site. Some are extending Computing Element services to manage job submission to non-standard local resource management systems, some are incorporating pilot functionality on edge services managing the batch systems, while the others emulate a grid site inside a fully virtualized cloud environment.

The exploitation of opportunistic resources was at an early stage throughout 2015, at the level of 10% of the total ATLAS computing power, but in the next few years it is expected to deliver much more. In addition, demonstrating the ability to use an opportunistic resource can lead to securing ATLAS allocations on the facility, hence the importance of this work goes beyond merely the initial CPU cycles gained.

In this presentation, we give an overview and compare the performance, development effort, flexibility and robustness of the various approaches. Full descriptions of each of those models are given in other contributions to this conference.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

Track 7: Middleware, Monitoring and Accounting / 211

DIRAC universal pilots

Author: Federico Stagni¹

Co-authors: Andrew McNab ²; Cinzia Luzzi ¹; Wojciech Jan Krzemien ³

¹ CERN
² University of Manchester
³ National Centre for Nuclear Research (PL)

Corresponding Authors: andrew.mcnab@cern.ch, wojciech.krzemien@cern.ch, federico.stagni@cern.ch, cinzia.luzzi@cern.ch
In the last few years, new types of computing models, such as IAAS (Infrastructure as a Service) and IAAC (Infrastructure as a Client), gained popularity. New resources may come as part of pledged resources, while others are in the form of opportunistic ones. Most but not all of these new infrastructures are based on virtualization techniques. In addition, some of them, present opportunities for multi-processor computing slots to the users. Virtual Organizations are therefore facing heterogeneity of the available resources and the use of an Interware software like DIRAC to provide the transparent, uniform interface has become essential. The transparent access to the underlying resources is realized by implementing the pilot model. DIRAC’s newest generation of generic pilots (the so-called Pilots 2.0) are the “pilots for all the skies”, and have been successfully released in production more than a year ago. They use a plugin mechanism that makes them easily adaptable. Pilots 2.0 have been used for fetching and running jobs on every type of resource, being it a Worker Node (WN) behind a CREAM/ARC/HTCondor/DIRAC Computing element, a Virtual Machine running on IaaS infrastructures like Vac or BOINC, on IaaS cloud resources managed by Vcycle, the LHCb High Level Trigger farm nodes, and any type of opportunistic computing resource. Make a machine a “Pilot Machine”, and all diversities between them will disappear. This contribution describes how pilots are made suitable for different resources, and the recent steps taken towards a fully unified framework, including monitoring. Also, the cases of multi-processor computing slots either on real or virtual machines, with the whole node or a partition of it, will be discussed.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Distributed workload management

Primary Keyword (Mandatory):

Computing middleware

Posters B / Break / 212

The ATLAS Computing Agora : a resource web site for citizen science projects

Author: Claire Adam Bourdarios

Co-author: Riccardo Maria Bianchi

1 Laboratoire de l’Accelerateur Lineaire (FR)

2 University of Pittsburgh (US)

Corresponding Authors: claire.bourdarios@cern.ch, riccardo.maria.bianchi@cern.ch

The ATLAS collaboration has recently setup a number of citizen science projects which have a strong IT component and could not have been envisaged without the growth of general public computing resources and network connectivity: event simulation through volunteer computing, algorithms improvement via Machine Learning challenges, event display analysis on citizen science platforms, use of open data, etc.

Most of the interactions with volunteers are handled through message boards, but specific outreach material was also developed, giving an enhanced visibility to the ATLAS software and computing techniques, challenges and community.

In this talk the Atlas Computing Agora (ACA) web platform will be presented as well as some of the specific material developed for some of the projects. The considerable interest triggered in the public and the lessons learned over two years will be summarized.

Title:
The ATLAS Computing Agora : a resource web site for citizen science projects
Abstract:
The ATLAS collaboration has recently set up a number of citizen science projects which have a strong IT component and could not have been envisaged without the growth of general public computing resources and network connectivity: event simulation through volunteer computing, algorithms improvement via Machine Learning challenges, event display analysis on citizen science platforms, use of open data, etc.

Most of the interactions with volunteers are handled through message boards, but specific outreach material was also developed, giving an enhanced visibility to the ATLAS software and computing techniques, challenges and community.

In this talk the Atlas Computing Agora (ACA) web platform will be presented as well as some of the specific material developed for some of the projects. The considerable interest triggered in the public and the lessons learned over two years will be summarized.

Primary Keyword (Mandatory):
Outreach

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

Track 8: Security, Policy and Outreach / 213

HiggsHunters - a citizen science project for ATLAS

Author: Alan Barr¹
Co-author: Andrew Haas ²

¹ University of Oxford (GB)
² New York University

Corresponding Authors: alan.barr@cern.ch, drandyhaas@gmail.com

Since the launch of HiggsHunters.org in November 2014, citizen science volunteers have classified more than a million points of interest in images from the ATLAS experiment at the LHC. Volunteers have been looking for displaced vertices and unusual features in images recorded during LHC Run-1. We discuss the design of the project, its impact on the public, and the surprising results of how the human volunteers performed relative to the computer algorithms in identifying displaced secondary vertices.

Primary Keyword (Mandatory):
Outreach

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

Track 5: Software Development / 215

Big Data Analytics Tools as Applied to ATLAS Event Data

Author: Ilija Vukotic¹
Big Data technologies have proven to be very useful for storage, processing and visualization of derived metrics associated with ATLAS distributed computing (ADC) services. Log file data and database records, and metadata from a diversity of systems have been aggregated and indexed to create an analytics platform for ATLAS ADC operations analysis. Dashboards, wide area data access cost metrics, user analysis patterns, and resource utilization efficiency charts are produced flexibly through queries against a powerful analytics cluster. Here we explore whether these techniques and analytics ecosystem can be applied to add new modes of open, quick, and pervasive access to ATLAS event data so as to simplify access and broaden the reach of ATLAS public data to new communities of users. An ability to efficiently store, filter, search and deliver ATLAS data at the event and/or sub-event level in a widely supported format would enable or significantly simplify usage of big data, statistical and machine learning tools like Spark, Jupyter, R, SciPy, Caffe, TensorFlow, etc.. Machine learning challenges such as the Higgs Boson Machine Learning Challenge, the Tracking challenge, Event viewers (VP1, ATLANTIS, ATLASrift), and still to be developed educational and outreach tools would be able to access the data through a simple REST API. In this preliminary investigation we focus on derived xAOD data sets. These are much smaller than the primary xAODs having containers, variables, and events of interest to a particular analysis. Being encouraged with the performance of Elasticsearch for the ADC analytics platform, we developed an algorithm for indexing derived xAOD event data. We have made an appropriate document mapping and have imported a full set of standard model W/Z datasets. We compare the disk space efficiency of this approach to that of standard ROOT files, the performance in simple cut flow type of data analysis, and will present preliminary results on its scaling characteristics with different numbers of clients, query complexity, and size of the data retrieved.

Tertiary Keyword (Optional):
Analysis tools and techniques

Secondary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Primary Keyword (Mandatory):
Data model

Track 5: Software Development / 216

Event visualisation in ATLAS: current software technologies / future prospects and trends

Author: Riccardo Maria Bianchi
At the beginning, HEP experiments made use of photographic images both to record and store experimental data and to illustrate their findings. Then the experiments evolved and needed to find ways to visualize their data. With the availability of computer graphics, software packages to display event data and the detector geometry started to be developed. Here a brief history of event displays is presented, with an overview of the different event display tools used today in HEP experiments in general, and in the LHC experiments in particular.

Then the case of the ATLAS experiment is considered in more detail and two widely used event display packages are presented, Atlantis and VP1, focusing on the software technologies they employ, as well as their strengths, differences and their usage in the experiment: from physics analysis to detector development, and from online monitoring to outreach and communication.

Future development plans and improvements in the ATLAS event display packages will also be discussed, as well as an outlook on interesting technologies for future event display tools for HEP: from web-based interactive visualizations to the usage of game engines.
The DIRAC project is developing interware to build and operate distributed computing systems. It provides a development framework and a rich set of services for both Workload and Data Management tasks of large scientific communities. A number of High Energy Physics and Astrophysics collaborations have adopted DIRAC as the base for their computing models. DIRAC was initially developed for the LHCb experiment at LHC, CERN. Later, the Belle II, BES III and CTA experiments as well as the linear collider detector collaborations started using DIRAC for their computing systems.

Some of the experiments built their DIRAC-based systems from scratch, others migrated from previous solutions, ad hoc or based on different middlewares. Adaptation of DIRAC for a particular experiment was enabled through the creation of extensions to meet their specific requirements. Each experiment has a heterogeneous set of computing and storage resources at their disposal that were aggregated through DIRAC into a coherent pool. Users from different experiments can interact with the system in different ways depending on their specific tasks, expertise level and previous experience using command line tools, python APIs or Web Portals. In this contribution we will summarize the experience of using DIRAC in particle physics collaborations. The problems of migration to DIRAC from previous systems and their solutions will be presented. Overview of specific DIRAC extensions will be given. We hope that this review will be useful for the experiments considering an update or a new from-scratch design of their production computing models.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Distributed workload management

Primary Keyword (Mandatory):
Computing middleware

Track 2: Offline Computing / 218

ATLAS Data Preparation in Run 2

Author: Paul James Laycock¹

¹ University of Liverpool (GB)

Corresponding Author: paul.james.laycock@cern.ch

In this presentation, the data preparation workflows for Run 2 are presented. Online data quality uses a new hybrid software release that incorporates the latest offline data quality monitoring software for the online environment. This is used to provide fast feedback in the control room during a data acquisition (DAQ) run, via a histogram-based monitoring framework as well as the online Event Display. Data are sent to several streams for offline processing at the dedicated Tier-0 computing facility, including dedicated calibration streams and an “express” physics stream containing approximately 2% of the main physics stream. This express stream is processed as data arrives, allowing a first look at the offline data quality within hours of a run end.

A prompt calibration loop starts once an ATLAS DAQ run ends, nominally defining a 48 hour period in which calibrations and alignments can be derived using the dedicated calibration and express streams. The bulk processing of the main physics stream starts on expiry of the prompt
calibration loop, normally providing the primary physics format after a further 24 hours. Physics data quality is assessed using the same monitoring packages, allowing data exclusion down to a granularity of one luminosity block or 1 minute. Meanwhile, the primary reconstruction output is passed to the ATLAS data reduction framework, providing data to users typically within 5 days of the end of a DAQ run, and on the same time scale as the data quality good run list.

Tertiary Keyword (Optional):
Computing facilities

Secondary Keyword (Optional):
Monitoring

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

Track 4: Data Handling / 219

Caching Servers for ATLAS

Author: Robert William Gardner Jr

1 University of Chicago (US)

Corresponding Author: robert.w.gardner@cern.ch

As many Tier 3 and some Tier 2 centers look toward streamlining operations, they are considering autonomously managed storage elements as part of the solution. These storage elements are essentially file caching servers. They can operate as whole file or data block level caches. Several implementations exist. In this paper we explore using XRootD caching servers that can operate in either mode. They can also operate autonomously (i.e. demand driven), be centrally managed (i.e. a Rucio managed cache), or operate in both modes. We explore the pros and cons of various configurations as well as practical requirements for caching to be effective. While we focus on XRootD caches, the analysis should apply to other kinds of caches as well.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Storage systems

Primary Keyword (Mandatory):
Distributed data handling

Track 8: Security, Policy and Outreach / 220

An Analysis of Reproducibility and Non-Determinism in HEP Software and ROOT Data

Author: Douglas Thain

Co-authors: Charles Zheng; Peter Ivie
Reproducibility is an essential component of the scientific process. It is often necessary to check whether multiple runs of the same software produce the same result. This may be done to validate whether a new machine produces correct results on old software, whether new software produces correct results on an old machine, or to compare the equality of two different approaches to the same objective.

Unfortunately, many technical issues in computing make it surprisingly hard to get the same result twice. Non-determinism in both codes and data can arise unexpectedly from the use of concurrency, random number sources, real-time clocks, I/O operations, and other sources. Some of these issues are merely cosmetic, like displaying the current time and machine for diagnostic purposes. Some of these issues are much more subtle, such as race conditions between threads changing the order of records in a file. As a result, one cannot simply compare objects at the binary level: different bits might reflect fundamentally different algorithms, or the might reflect accidents of the runtime environment.

In this talk, we will present an analysis of non-determinism and reproducibility issues in standard HEP software and the ROOT data format. We will present an evaluation of simulation and analysis codes to reveal sources of non-determinism, so that this property can be eliminated (where possible) and understood (where necessary). Where non-determinism is unavoidable, we present tool (root-diff) for the detailed comparison of output files to evaluate differences between structured data, so as to distinguish between files that differ in fundamental data, as opposed to structure, ordering, or timestamps. We conclude with recommendations for best practices in software development, program analysis, and operations.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Accounting and information

Primary Keyword (Mandatory):
Preservation of analysis and data

Track 2: Offline Computing / 221

New developments in Geant4 version 10 series

Authors: Makoto Asai\textsuperscript{1}; Marc Verderi\textsuperscript{2}

\textsuperscript{1} SLAC National Accelerator Laboratory (US)
\textsuperscript{2} Centre National de la Recherche Scientifique (FR)

Corresponding Authors: asai@slac.stanford.edu, marc.verderi@gmail.com

The Geant4 Collaboration released a new generation of the Geant4 simulation toolkit (version 10) in December 2013 and reported its new features at CHEP 2015. Since then, the Collaboration continues to improve its physics and computing performance and usability. This presentation will survey the major improvements made since version 10.0. On the physics side, it includes fully revised multiple scattering models, new Auger atomic de-excitation cascade simulation, significant improvements in string models, and an extension of the low-energy neutron model to protons and light ions. Extensions and improvements of the unified solid library provide more functionality and better computing.
performance, while a review of the navigation algorithm improved code stability. The continued ef- 
fort to reduce memory consumption per thread allows for massive parallelism of large applications 
in the multithreaded mode. Toolkit usability was improved with an evolved real-time visualization in 
multithreaded mode. Prospects for short- and long-term development will also be discussed.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Parallelizarion

Primary Keyword (Mandatory):
Simulation

Track 5: Software Development / 222

Application of econometric and ecology analysis methods in physics software

Authors: Chan Hyeong Kim\textsuperscript{1}; Elisabetta Ronchieri\textsuperscript{None}; Gabriela Hoff\textsuperscript{2}; Maria Grazia Pia\textsuperscript{3}; Mincheol Han\textsuperscript{1}; Sung Hun Kim\textsuperscript{1}

\textsuperscript{1} Hanyang University
\textsuperscript{2} CAPES
\textsuperscript{3} Universita e INFN Genova (IT)

Corresponding Authors: mchan@hanyang.ac.kr, chkim@hanyang.ac.kr, ghoff.gesic@gmail.com, elisabetta.ronchieri@cnaf.infn.it, ksh4249@hanyang.ac.kr, maria.grazia.pia@cern.ch

Some data analysis methods typically used in econometric studies and in ecology have been eval-
uated and applied in physics software environments. They concern the evolution of observables 
through objective identification of change points and trends, and measurements of inequality, di-
versity and evenness across a data set. Within each one of these analysis areas, several statistical 
tests and measures have been examined, often comparing multiple implementations of the same 
algorithm available in R or developed by us.

The presentation will introduce the analysis methods and the details of their statistical formula-
tion, and will review their relation with information theory concepts, such as Shannon entropy. It 
will report the results of their use in two real-life scenarios, which pertain to diverse application 
domains: the validation of simulation models and the quantification of software quality. It will dis-
cuss the lessons learned, highlighting the capabilities and shortcomings identified in this pioneering 
study.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Experience/plans from outside experimental HEP/NP

Primary Keyword (Mandatory):
Analysis tools and techniques

Track 3: Distributed Computing / 224
Opportunistic Computing with Lobster: Lessons Learned From Scaling up to 25k Non-Dedicated Cores

Authors: Anna Elizabeth Woodard\textsuperscript{1}; Matthias Wolf\textsuperscript{6}

Co-authors: Anna Yannakopoulos\textsuperscript{2}; Benjamin Tovar\textsuperscript{3}; Douglas Thain\textsuperscript{3}; Kenyi Paolo Hurtado Anampa\textsuperscript{1}; Kevin Patrick Lannon\textsuperscript{1}; Mike Hildreth\textsuperscript{1}; Patrick Donnelly\textsuperscript{3}; Paul Brenner\textsuperscript{3}; Wenzhao Li\textsuperscript{3}

\textsuperscript{1} University of Notre Dame (US)  
\textsuperscript{2} Notre Dame / Florida State University  
\textsuperscript{3} University of Notre Dame

Corresponding Authors: dthain@cse.nd.edu, paul.r.brenner@nd.edu, pdonnel5@nd.edu, btovar@nd.edu, m+cms@sushinara.net, anna.elizabeth.woodard@cern.ch, mikeh@omega.hep.nd.edu, klannon@nd.edu, kenyi.paolo.hurtado.anampa@cern.ch, wenzhao.li.247@nd.edu, anna.yannakopoulos@gmail.com

We previously described Lobster, a workflow management tool for exploiting volatile opportunistic computing resources for computation in HEP. We will discuss the various challenges that have been encountered while scaling up the simultaneous CPU core utilization and the software improvements required to overcome these challenges.

Categories: Workflows can now be divided into categories based on their required system resources. This allows the batch queueing system to optimize assignment of tasks to nodes with the appropriate capabilities. Within each category, limits can be specified for the number of running jobs to regulate the utilization of communication bandwidth. System resource specifications for a task category can now be modified while a project is running, avoiding the need to restart the project if resource requirements differ from the initial estimates. Lobster now implements time limits on each task category to voluntarily terminate tasks. This allows partially completed work to be recovered.

Workflow dependency specification: One workflow often requires data from other workflows as input. Rather than waiting for earlier workflows to be completed before beginning later ones, Lobster now allows dependent tasks to begin as soon as sufficient input data has accumulated.

Resource monitoring: Lobster utilizes a new capability in Work Queue to monitor the system resources each task requires in order to identify bottlenecks and optimally assign tasks.

The capability of the Lobster opportunistic workflow management system for HEP computation has been significantly increased. We have demonstrated efficient utilization of 25K non-dedicated cores and achieved a data input rate of 9 Gb/s and an output rate of 400 GB/h. This has required new capabilities in task categorization, workflow dependency specification, and resource monitoring.

Primary Keyword (Mandatory): Distributed workload management

Secondary Keyword (Optional): 

Tertiary Keyword (Optional): 

Track 6: Infrastructures / 225

Mixing HTC and HPC Workloads With HTCondor and Slurm

Author: Christopher Hollowell\textsuperscript{1}

Co-authors: Alexandr Zaytsev\textsuperscript{2}; Antonio Wong\textsuperscript{2}; Costin Caramarcu\textsuperscript{2}; Joshua Barnett\textsuperscript{1}; William Strecker-Kellogg\textsuperscript{3}

\textsuperscript{1} Brookhaven National Laboratory
Traditionally, the RHIC/ATLAS Computing Facility (RACF) at Brookhaven National Laboratory has only maintained High Throughput Computing (HTC) resources for our HEP/NP user community. We've been using HTCondor as our batch system for many years, as this software is particularly well suited for managing HTC processor farm resources. Recently, the RACF has also begun to design/administrate some High Performance Computing (HPC) systems for a multidisciplinary user community at BNL. In this presentation, we'll discuss our experiences using HTCondor and Slurm in an HPC context, and our facility’s attempts to allow our HTC and HPC processing farms/clusters to make opportunistic use of each other’s computing resources.

**Primary Keyword (Mandatory):**
Computing facilities

**Secondary Keyword (Optional):**

**Tertiary Keyword (Optional):**

---

**Dark matter research cluster based on computational science**

**Author:** Kihyeon Cho

**Corresponding Author:** cho@kisti.re.kr

Let me introduce the convergence research cluster for dark matter which is supported by National Research Council of Science and Technology in Korea. The goal is to build research cluster of nationwide institutes from accelerator-based physics to astrophysics based on computational science using infrastructures at KISTI (Korea Institute of Science Technology Information) and KASI (Korea Astronomy and Space Science Institute). The key of the computational science, simulation will be discussed.

**Secondary Keyword (Optional):**
Outreach

**Primary Keyword (Mandatory):**
Simulation

**Tertiary Keyword (Optional):**

---

**JUNO performance evaluation and optimization on virtual platform**

**Authors:** Ce Yang; Xiaomei Zhang; Ying Lv; ZhenTai Ma

**Co-authors:** Tian Yan; Xianghu Zhao

1. ZheJiang University
JUNO (Jiangmen Underground Neutrino Observatory) is a multi-purpose neutrino experiment designed to measure the neutrino mass hierarchy and mixing parameters. JUNO is estimated to be in operation in 2019 with 2PB/year raw data rate. The IHEP computing center plans to build up virtualization infrastructure to manage computing resources in the coming years and JUNO is selected to be one of the first experiments to run on virtual platform. Before migrating, performance evaluation and optimization for JUNO software on virtual platform is necessary. With benchmark tools and current JUNO offline software, the paper will present the design of a complete set of tests to find out the best choices of virtualization infrastructures including hardware, hypervisor, memory and size of VMs, etc. To facilitate testing procedures, automatic tools have been developed. The findings during tests and the suggestions to future improvements of JUNO software will also be described in the paper. In the optimization part, we will describe the factors affecting performance and the ways we manage to improve JUNO simulation and reconstruction processes in virtual platform by 10% ~20% in multi-VM cases. Besides detailed tests in single machine, we also do the scale tests to find out performance behaviors in real application scenarios.

Primary Keyword (Mandatory):
Virtualization

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

Track 2: Offline Computing / 228

Opticks: GPU Optical Photon Simulation for Particle Physics with NVIDIA OptiX

Author: simon blyth

Corresponding Author: simon.c.blyth@gmail.com

Opticks is an open source project that integrates the NVIDIA OptiX GPU ray tracing engine with Geant4 toolkit based simulations. Massive parallelism brings drastic performance improvements with optical photon simulation speedup expected to exceed 1000 times Geant4 when using workstation GPUs. Optical photon simulation time becomes effectively zero compared to the rest of the simulation.

Optical photons from scintillation and Cherenkov processes are allocated, generated and propagated entirely on the GPU, minimizing transfer overheads and allowing CPU memory usage to be restricted to optical photons that hit photomultiplier tubes or other photon detectors. Collecting hits into standard Geant4 hit collections then allows the rest of the simulation chain to proceed unmodified.

Optical physics processes of scattering, absorption, reemission and boundary processes are implemented as CUDA OptiX programs based on the Geant4 implementations. Wavelength dependent material and surface properties as well as inverse cumulative distribution functions for reemission are interleaved into GPU textures providing fast interpolated property lookup or wavelength generation.
Geometry is provided to OptiX in the form of CUDA programs that return bounding boxes for each primitive and single ray geometry intersection results. Some critical parts of the geometry such as photomultiplier tubes have been implemented analytically with the remainder being tesselated. OptiX handles the creation and application of a choice of acceleration structures such as boundary volume hierarchies and the transparent use of multiple GPUs. OptiX interoperation with OpenGL and CUDA Thrust has enabled unprecedented visualisations of photon propagations to be developed using OpenGL geometry shaders to provide interactive time scrubbing and CUDA Thrust photon indexing to provide interactive history selection. Validation and performance results are shown for the photomultiplier based Daya Bay and JUNO Neutrino detectors.

Tertiary Keyword (Optional):
Visualization

Secondary Keyword (Optional):
Parallelizarion

Primary Keyword (Mandatory):
Simulation

Track 5: Software Development / 229

First results from a combined analysis of CERN computing infrastructure metrics

Author: Dirk Duellmann

Co-authors: Christian Nieke; Oliver Keeble

1 CERN

2 Brunswick Technical University (DE)

Corresponding Authors: dirk.duellmann@cern.ch, christian.nieke@cern.ch, oliver.keeble@cern.ch

The IT Analysis Working Group (AWG) has been formed at CERN across individual computing units and the experiments to attempt a cross cutting analysis of computing infrastructure and application metrics. In this presentation we will describe the first results obtained using medium/long term data (1 months - 1 year) correlating box level metrics, job level metrics from LSF and HTCondor, I/O metrics from the physics analysis disk pools (EOS) and networking and application level metrics from the experiment dashboards. We will cover in particular the measurement of hardware performance and prediction of job durations, the latency sensitivity of different job types and a search for bottlenecks with the production job mix in the current infrastructure. The presentation will conclude with the proposal of a small set of metrics to simplify drawing conclusions also in the more constrained environment of public cloud deployments.

Secondary Keyword (Optional):
Monitoring

Primary Keyword (Mandatory):
Analysis tools and techniques

Tertiary Keyword (Optional):
Online Data Compression in the ALICE O2 facility

Author: Matthias Richter

1 University of Oslo (NO)

Corresponding Author: matthias.richter@cern.ch

The ALICE Collaboration and the ALICE O2 project have carried out detailed studies for a new online computing facility planned to be deployed for Run 3 of the Large Hadron Collider (LHC) at CERN. Some of the main aspects of the data handling concept are partial reconstruction of raw data organized in so called time frames, and based on that information reduction of the data rate without significant loss in the physics information.

A production solution for data compression is running for the ALICE Time Projection Chamber (TPC) in the ALICE High Level Trigger online system since 2011. The solution is based on reconstruction of space points from raw data. These so called clusters are the input for reconstruction of particle trajectories by the tracking algorithm. Clusters are stored instead of raw data after a transformation of required parameters into an optimized format and subsequent lossless data compression techniques. With this approach, a reduction of 4.4 has been achieved on average.

For Run 3, a significantly higher reduction is required. Several options are under study for cluster data to be stored. As the first group of options, alternative lossless techniques like e.g. arithmetic coding have been investigated.

Furthermore, theoretical studies had shown a significant potential of compressed data formats for clusters relative to the particle trajectory they belong to. In the present scheme, cluster parameters are stored in uncalibrated detector format while the track as reference for residual calculation is described in Cartesian space. This results into higher entropy of the parameter residuals and smaller data reduction. The track reconstruction scheme of the O2 system will allow for storing calibrated clusters. The distribution of residuals has a smaller entropy and is better suited for data compression. A further contribution is expected from adaptive precision for individual cluster parameters based on reconstructed particle trajectories.

As one major difference in the mode of operation, the increase in the flux of particles leads to larger accumulation of space charge in the detector volume and significant distortions of cluster positions relative to the physical particle trajectory. The influence of the space charge distortions to the data compression is under study.

Though data compression is being studied for the TPC as premier use case, concept and code development is kept open to be applied to other detectors as well.

In this contribution we report on general concepts of data compression in ALICE O2 and recent results for all different options under study.

Tertiary Keyword (Optional): DAO

Secondary Keyword (Optional): Reconstruction

Primary Keyword (Mandatory): Data processing workflows and frameworks/pipelines
Hadoop and friends - first experience at CERN with a new platform for high throughput analysis steps

Author: Dirk Duellmann

Co-authors: Kacper Surdy; Luca Menichetti; Prasanth Kothuri; Rainer Toebbicke; Vineet Menon

1 CERN
2 Bhabha Atomic Research Centre (IN)

Corresponding Authors: kacper.surdy@cern.ch, luca.menichetti@cern.ch, rainer.toebbicke@cern.ch, dirk.duellmann@cern.ch, vineet.menon@cern.ch, prasanth.kothuri@cern.ch

The statistical analysis of infrastructure metrics comes with several specific challenges, including the fairly large volume of unstructured metrics from a large set of independent data sources. Hadoop and Spark provide an ideal environment in particular for the first steps of skimming rapidly through hundreds of TB of low relevance data to find and extract the much smaller data volume that is relevant for statistical analysis and modelling.

This presentation will describe the new Hadoop service at CERN and the use of several of its components for high throughput data aggregation and ad-hoc pattern searches. We will describe the hardware setup used, the service structure with a small set of decoupled clusters and the first experience with co-hosting different applications and performing software upgrades. We will further detail the common infrastructure used for data extraction and preparation from continuous monitoring and database input sources.

OCCAM: a flexible, multi-purpose and extendable HPC cluster

Authors: Marco Aldinucci; Paolo Pasteris; Sergio Rabellino; Stefano Bagnasco; Stefano Lusso

1 Computer Science Department, University of Torino
2 Universita e INFN Torino (IT)
3 Istituto Nazionale di Fisica Nucleare, Torino

Corresponding Authors: rabellino@di.unito.it, aldinuc@di.unito.it, pasteris@di.unito.it, stefano.lusso@to.infn.it

Obtaining CPU cycles on an HPC cluster is nowadays relatively simple and sometimes even cheap for academic institutions. However, in most of the cases providers of HPC services would not allow changes on the configuration, implementation of special features or a lower-level control on the computing infrastructure and networks, for example for testing new computing patterns or conducting research on HPC itself. The variety of use cases proposed by several departments of the University of Torino, including ones from solid-state chemistry, high-energy physics, computer science, big data analytics, computational biology, genomics and many others, called for different and sometimes conflicting configurations; furthermore, several R&D activities in the field of scientific computing, with topics ranging from GPU acceleration to Cloud Computing technologies, needed a platform to be carried out on.
The Open Computing Cluster for Advanced data Manipulation (OCCAM) is a multi-purpose flexible HPC cluster designed and operated by a collaboration between the University of Torino and the Torino branch of the Istituto Nazionale di Fisica Nucleare. It is aimed at providing a flexible, reconfigurable and extendable infrastructure to cater to a wide range of different scientific computing needs, as well as a platform for R&D activities on computational technologies themselves. Extending it with novel architecture CPU, accelerator or hybrid microarchitecture (such as forthcoming Intel Xeon Phi Knights Landing) will be as simple as plugging a node in a rack.

The initial system counts slightly more than 1100 cpu cores and includes different types of computing nodes (standard dual-socket nodes, large quad-sockets nodes with 768 GB RAM, and multi-GPU nodes) and two separate disk storage subsystems: a smaller high-performance scratch area, based on the Lustre file system, intended for direct computational I/O and a larger one, of the order of 1PB, to archive near-line data for archival purposes. All the components of the system are interconnected through a 10Gb/s Ethernet layer with one-level topology and an InfiniBand FDR 56Gbps layer in fat-tree topology.

A system of this kind, heterogeneous and reconfigurable by design, poses a number of challenges related to the frequency at which heterogeneous hardware resources might change their availability and shareability status, which in turn affect methods and means to allocate, manage, optimize, bill, monitor VMs, virtual farms, jobs, interactive bare-metal sessions, etc.

This poster describes some of the use cases that prompted the design construction of the HPC cluster, its architecture and a first characterization of its performance by some synthetic benchmark tools and a few realistic use-case tests.

Secondary Keyword (Optional):
High performance computing

Primary Keyword (Mandatory):
Computing facilities

Tertiary Keyword (Optional):

Posters A / Break / 233

Automatised data quality monitoring of the LHCb Vertex Locator

Author: Marco Gersabeck
Co-authors: Alex Pearce; Andrew Christopher Crocombe; Lennaert Bel; Tomasz Szumlak

1 University of Manchester (GB)
2 University of Warwick (GB)
3 Nikhef National institute for subatomic physics (NL)
4 AGH University of Science and Technology (PL)

Corresponding Authors: l.bel@cern.ch, marco.gersabeck@cern.ch, andrew.christopher.crocombe@cern.ch, alex.pearce@cern.ch, szumlak@agh.edu.pl

The LHCb Vertex Locator (VELO) is a silicon strip semiconductor detector operating at just 8mm distance to the LHC beams. Its 172,000 strips are read at a frequency of 1 MHz and processed by off-detector FPGAs followed by a PC cluster that reduces the event rate to about 10 kHz. During the second run of the LHC, which lasts from 2015 until 2018, the detector performance will undergo continued change due to radiation damage effects. This necessitates a detailed monitoring of the data quality to avoid adverse effects on the physics analysis performance.

The VELO monitoring infrastructure has been re-designed compared to the first run of the LHC when it was based on manual checks. The new system is based around an automatic analysis framework,
which monitors the performance of new data as well as long-term trends and flags issues whenever they arise.

An unbiased subset of the detector data are processed about once per hour by monitoring algorithms. The new analysis framework then analyses the plots that are produced by these algorithms. One of its tasks is to perform custom comparisons between the newly processed data and that from reference runs. A single figure of merit for the current VELO data quality is computed from a tree-like structure, where the value of each node is computed using the values of its child branches. The comparisons and the combination of their output is configurable through steering files and is applied dynamically. Configurable thresholds determine when the data quality is considered insufficient and an alarm is raised. The most-likely scenario in which this analysis would identify an issue is the parameters of the readout electronics no longer being optimal and requiring retuning.

The data of the plots are reduced further, e.g. by evaluating averages, and these quantities are input to long-term trending. This is used to detect slow variation of quantities, which are not detectable by the comparison of two nearby runs. Such gradual change is what is expected due to radiation damage effects. It is essential to detect these changes early such that measures can be taken, e.g. adjustments of the operating voltage, to prevent any impact on the quality of high-level quantities and thus on physics analyses.

The plots as well as the analysis results and trends are made available through graphical user interfaces (GUIs). One is available to run locally on the LHCb computing cluster, the other provides a web interface for remote data quality assessment. The latter operates a server-side queuing system for worker nodes that retrieve the data and pass it on the client for displaying. Both GUIs are dynamically configured by a single configuration that determines the choice and arrangement of plots and trends and ensures a common look-and-feel. The infrastructure underpinning the web GUI is used as well for other monitoring applications of the LHCb experiment.

**Secondary Keyword (Optional):**
Visualization

**Primary Keyword (Mandatory):**
Monitoring

**Tertiary Keyword (Optional):**

---

**Track 7: Middleware, Monitoring and Accounting / 234**

**Monitoring performance of a highly distributed and complex computing infrastructure in LHCb**

**Author:** Zoltan Mathe

**Co-authors:** Christophe Haen; Federico Stagni

**CERN**

**Corresponding Authors:** zoltan.mathe@cern.ch, federico.stagni@cern.ch, christophe.denis.haen@cern.ch

In order to ensure an optimal performance of the LHCb Distributed Computing, based on LHCb-DIRAC, it is necessary to be able to inspect the behavior over time of many components: firstly the agents and services on which the infrastructure is built, but also all the computing tasks and data transfers that are managed by this infrastructure. This consists of recording and then analyzing time series of a large number of observables, for which the usage of SQL relational databases is far from optimal. Therefore within DIRAC we have been studying novel possibilities based on NoSQL databases (ElasticSearch, OpenTSDB and InfluxDB) as a result of this study we developed a new monitoring system based on ElasticSearch. It has been deployed on the LHCb Distributed Computing infrastructure for which it collects data from all the components (agents, services, jobs) and allows creating reports through Kibana and a web user interface, which is based on the DIRAC web framework.
In this paper we describe this new implementation of the DIRAC monitoring system. We give details on the ElasticSearch implementation within the DIRAC general framework, as well as an overview of the advantages of the pipeline aggregation used for creating a dynamic bucketing of the time series. We present the advantages of using the ElasticSearch DSL high-level library for creating and running queries. Finally we shall present the performances of that system.

Secondary Keyword (Optional):
Databases

Primary Keyword (Mandatory):
Monitoring

Tertiary Keyword (Optional):

Posters B / Break / 235

Using OSG Computing Resources with (iLC)DIRAC

Authors: Andre Sailer\textsuperscript{1}; Marko Petric\textsuperscript{1}

\textsuperscript{1} CERN

Corresponding Authors: marko.petric@cern.ch, andre.philippe.sailer@cern.ch

CPU cycles for small experiments and projects can be scarce, thus making use of all available resources, whether dedicated or opportunistic, is mandatory. While enabling uniform access to the LCG computing elements (ARC, CREAM), the DIRAC grid interware was not able to use OSG computing elements (GlobusCE, HTCondor-CE) without dedicated support at the grid site through so called 'SiteDirectors', which directly submits to the local batch system. Which in turn requires additional dedicated effort for small experiments on the grid site. Adding interfaces to the OSG CEs through the respective grid middleware is therefore allowing accessing them within the DIRAC software without additional site-specific infrastructure. Thus enabling greater use of opportunistic resources for experiments and projects without dedicated clusters or an established computing infrastructure with the DIRAC software.

To send jobs to HTCondor-CE and legacy Globus computing elements inside DIRAC the required wrapper modules were developed. Not only is the usage these types of computing elements now completely transparent for all DIRAC instances, which makes DIRAC a flexible solution for OSG based virtual organisations, but also allows LCG Grid Sites to move to the HTCondor-CE software, without shutting DIRAC based VOs out of their site.

In this presentation we detail how we interfaced the DIRAC system to the HTCondor-CE and Globus computing elements and explain the encountered obstacles and solutions or workarounds developed, and how the linear collider uses resources in the OSG.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Computing middleware

Primary Keyword (Mandatory):
Computing models
Indico 2.0 - the whole Iceberg

Author: Adrian Mönnich

Co-authors: Alejandro Avilés; Florian Vessaz; Ilias Trichopoulou; Michal Kolodziejski; Pedro Ferreira

1 CERN
2 Ecole Polytechnique Federale de Lausanne (CH)
3 Lodz Technical University (PL)

Corresponding Authors: michal.kolodziejski@cern.ch, ilias.trichopoulou@cern.ch, alejandro.aviles@cern.ch, flor- rian.vessaz@cern.ch, adrian.moennich@cern.ch, pedro.ferreira@cern.ch

The last two years have been atypical to the Indico community, as the development team undertook an extensive rewrite of the application and deployed no less than 9 major releases of the system. Users at CERN have had the opportunity to experience the results of this ambitious endeavour. They have only seen, however, the “tip of the iceberg”.

Indico 2.0 employs a completely new stack, leveraging open source packages in order to provide a web application that is not only more feature-rich but, more importantly, builds on a solid foundation of modern technologies and patterns. But this milestone represents not only a complete change in technology - it is also an important step in terms of user experience and usability that opens the way to many potential improvements in the years to come.

In this article, we will describe the technology and all the different dimensions in which Indico 2.0 constitutes an evolution vis-à-vis its predecessor and what it can provide to users and server administrators alike. We will go over all major system features and explain what has changed, the reasoning behind the most significant modifications and the new possibilities that they pave the way for.

Primary Keyword (Mandatory):
Collaborative tools

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

Tracking Machine Learning Challenge

Author: David Rousseau

Co-authors: Andreas Salzburger; Cecile Germain; Davide Costanzo; Isabelle Guyon; Jean-Roch Vlimant; Markus Elsing; Michael Aaron Kagan; Paolo Calafiura; Rebecca Carney; Riccardo Cenci; Steven Andrew Farrell; Tobias Golling; Tony Tong; Vincenzo Innocente

1 LAL-Orsay, FR
2 CERN
3 Universite Paris Sud
4 University of Sheffield (GB)
5 California Institute of Technology (US)
6 SLAC National Accelerator Laboratory (US)
7 Lawrence Berkeley National Lab. (US)
The instantaneous luminosity of the LHC is expected to increase at HL-LHC so that the amount of pile-up can reach a level of 200 interaction per bunch crossing, almost a factor of 10 w.r.t the luminosity reached at the end of run 1. In addition, the experiments plan a 10-fold increase of the readout rate. This will be a challenge for the ATLAS and CMS experiments, in particular for the tracking, which will be performed with a new all Silicon tracker in both experiments. In terms of software, the increased combinatorial complexity will have to be dealt with within flat budget at best. Preliminary studies show that the CPU time to reconstruct the events explodes with the increased pileup level. The increase is dominated by the increase of the CPU time of the tracking, itself dominated by the increase of the CPU time of the pattern recognition stage. In addition to traditional CPU optimisation and better use of parallelism, exploration of completely new approaches to pattern recognition has been started.

To reach out to Computer Science specialists, a Tracking Machine Learning challenge (trackML) has been set up, building on the experience of the successful Higgs Machine Learning challenge in 2014 (see talk by Glen Cowan at CHEP 2015). It associates ATLAS and CMS physicists with Computer Scientists. A few relevant points:

- A dataset consisting of a simulation of a typical full Silicon LHC experiments has been created, listing for each event the measured 3D points, and the list of 3D points associated to a true track. The data set is large to allow the training of data hungry Machine Learning methods: the orders of magnitude are: one million event, 10 billion tracks, 1 terabyte.
- The participants to the challenge should find the tracks in an additional test dataset, meaning building the list of 3D points belonging to each track (deriving the track parameters is not the topic of the challenge)
- A figure of merit has been defined which combines the CPU time, the efficiency and the fake rate (with an emphasis on CPU time)
- The challenge platforms allow measuring the figure of merit and to rate the different algorithms submitted.

The emphasis is to expose innovative approaches, rather than hyper-optimising known approaches. Machine Learning specialists have shown a deep interest to participate to the challenge, with new approaches like Convolutional Neural Network, Deep Neural Net, Monte Carlo Tree Search and others.

Tertiary Keyword (Optional): Outreach
Secondary Keyword (Optional): Artificial intelligence/Machine learning
Primary Keyword (Mandatory): Algorithms
First experiences with a parallel architecture testbed in the LHCb trigger system

Authors: Alessio Gianelle¹; Donatella Lucchesi¹; Marco Corvo²; Silvia Amerio¹; Stefano Gallorini¹

¹ Universita e INFN, Padova (IT)  
² Universita di Ferrara & INFN (IT)

In view of Run3 (2020) the LHCb experiment is planning a major upgrade to fully readout events at 40 MHz collision rate. This in order to highly increase the statistic of the collected samples and go further in precision beyond Run2. An unprecedented amount of data will be produced, which will be fully reconstructed real-time to perform fast selection and categorization of interesting events. The collaboration has decided to go for a fully software trigger which will have a total time budget of 13 ms to take a decision. This calls for faster hardware and software.

In this talk we will present our efforts on the application of new technologies, such as GPU cards, to LHCb trigger system. During Run2 a node equipped with a GPU has been inserted in LHCb online monitoring system; during normal data taking, a subset of real events is sent to the node and processed in parallel by GPU-based and CPU-based track reconstruction algorithms. This gives us the unique opportunity to test the new hardware and the new algorithms in a realistic environment.

We will present the setup of the testbed, the algorithms developed for parallel architectures and discuss the performance compared to the current LHCb track reconstruction algorithms.

Evolution through revolution - A 2-year Agile experience with the Indico Team

Author: Pedro Ferreira¹

Co-authors: Adrian Mönnich¹, Alejandro Avilés¹, Florian Vessaz², Ilias Trichopoulos¹, Michal Kołodziejski³

¹ CERN  
² Ecole Polytechnique Federale de Lausanne (CH)  
³ Lodz Technical University (PL)

Corresponding Authors: michal.kolodziejski@cern.ch, ilias.trichopoulos@cern.ch, florian.vessaz@cern.ch, alejandro.aviles@cern.ch, pedro.ferreira@cern.ch, adrian.moennich@cern.ch

Over the last two years, a small team of developers worked on an extensive rewrite of the Indico application based on a new technology stack. The result, Indico 2.0, leverages open source packages in order to provide a web application that is not only more feature-rich but, more importantly, builds on a solid foundation of modern technologies and patterns.

Indico 2.0 has the peculiarity of looking like an evolution (in terms of user experience and design), while constituting a de facto revolution. An extensive amount of code (75%) was rewritten, not to mention a complete change of database and some of the most basic components of the system.
In this article, we will explain the process by which, over a period of approximately two years, we have managed to deliver and deploy a completely new version of an application that is used on a daily basis by the CERN community and HEP at large, in a gradual way, with no major periods of unavailability and with virtually no impact in performance and stability. We will focus particularly on how such an endeavor would not have been possible without the use of Agile Methodologies of software development. We will provide examples of practices and tools that we have adopted and display the evolution of development habits in the team over the period in question, as well as their impact in code quality and maintainability.

Secondary Keyword (Optional):
Collaborative tools

Primary Keyword (Mandatory):
Software development process and tools

Tertiary Keyword (Optional):

Track 2: Offline Computing / 241

GEANT4-based full simulation of the PADME experiment at the DAFNE BTF

Authors: Emanuele Leonardi\textsuperscript{1}; Emanuele Leonardi\textsuperscript{2}

Co-authors: Mauro Raggi \textsuperscript{3}; Venelin Kozhuharov \textsuperscript{4}

\textsuperscript{1} INFN Roma
\textsuperscript{2} Universita e INFN, Roma I (IT)
\textsuperscript{3} LNF INFN
\textsuperscript{4} University of Sofia (BG)

Corresponding Authors: mauro.raggi@cern.ch, emanuele.leonardi@roma1.infn.it, venelin.kozhuharov@cern.ch, emanuele.leonardi@cern.ch

The long standing problem of reconciling the cosmological evidence of the existence of dark matter with the lack of any clear experimental observation of it, has recently revived the idea that the new particles are not directly connected with the Standard Model gauge fields, but only through mediator fields or "portals", connecting our world with new "secluded" or "hidden" sectors. One of the simplest models just adds an additional U(1) symmetry, with its corresponding vector boson $A'$. At the end of 2015 INFN has formally approved a new experiment, PADME (Positron Annihilation into Dark Matter Experiment), to search for invisible decays of the $A'$ at the DAFNE BTF in Frascati. The experiment is designed to detect dark photons produced in positron on fixed target annihilations $(e^+e^- \rightarrow \gamma A')$ decaying to dark matter by measuring the final state missing mass. The collaboration aims to complete the design and construction of the experiment by the end of 2017 and to collect $\sim 10^{13}$ positrons on target by the end of 2018, thus allowing to reach the $\epsilon \sim 10^{-3}$ sensitivity up to a dark photon mass of $\sim 24$ MeV/c\textsuperscript{2}.

The experiment will be composed by a thin active diamond target where the positron beam from the DAFNE Linac will impinge to produce $e^+e^-$ annihilation events. The surviving beam will be deflected with a $\text{cuiO}(0.5$ Tesla) magnet, on loan from the CERN PS, while the photons produced in the annihilation will be measured by a calorimeter composed of BGO crystals recovered from the L3 experiment at LEP. To reject the background from bremsstrahlung gamma production, a set of segmented plastic scintillator vetoes will be used to detect positrons exiting the target with an energy below that of the beam, while a fast small angle calorimeter will be used to reject the $e^+e^- \rightarrow \gamma\gamma(\gamma)$ background.
To optimize the experimental layout in terms of signal acceptance and background rejection, the full layout of the experiment has been modeled with the GEANT4 simulation package. In this talk we will describe the details of the simulation and report on the results obtained with the software.

Primary Keyword (Mandatory):
Simulation

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

Posters A / Break / 242

Development and test of a DRS4-based DAQ system for the PADME experiment at the DAFNE BTF

Author: Emanuele Leonardi

Co-authors: Mauro Raggi; Paolo Valente

1 INFN Roma
2 LNF INFN
3 Università e INFN, Roma I (IT)

Corresponding Authors: paolo.valente@roma1.infn.it, emanuele.leonardi@roma1.infn.it, mauro.raggi@cern.ch

The long standing problem of reconciling the cosmological evidence of the existence of dark matter with the lack of any clear experimental observation of it, has recently revived the idea that the new particles are not directly connected with the Standard Model gauge fields, but only through mediator fields or "portals", connecting our world with new "secluded" or "hidden" sectors. One of the simplest models just adds an additional U(1) symmetry, with its corresponding vector boson $A'$.

At the end of 2015 INFN has formally approved a new experiment, PADME (Positron Annihilation into Dark Matter Experiment), to search for invisible decays of the $A'$ at the DAFNE BTF in Frascati. The experiment is designed to detect dark photons produced in positron on fixed target annihilations ($e^+ e^- \rightarrow \gamma A'$) decaying to dark matter by measuring the final state missing mass.

The collaboration aims to complete the design and construction of the experiment by the end of 2017 and to collect $\sim 10^{13}$ positrons on target by the end of 2018, thus allowing to reach the $\epsilon \sim 10^{-3}$ sensitivity up to a dark photon mass of $\sim 24$ MeV/c$^2$.

The experiment will be composed by a thin active diamond target where the positron beam from the DAFNE Linac will impinge to produce $e^+ e^-$ annihilation events. The surviving beam will be deflected with a $\mathcal{O}(0.5$ Tesla) magnet, on loan from the CERN PS, while the photons produced in the annihilation will be measured by a calorimeter composed of BGO crystals recovered from the L3 experiment at LEP. To reject the background from bremsstrahlung gamma production, a set of segmented plastic scintillator vetoes will be used to detect positrons exiting the target with an energy below that of the beam, while a fast small angle calorimeter will be used to reject the $e^+ e^- \rightarrow \gamma \gamma (\gamma)$ background.

The DAQ system of the PADME experiment will handle a total of $\mathcal{O}(1000)$ channels, with an expected DAQ rate, defined by the DAFNE Linac cycle, of 50 Hz. To satisfy these requirements, we plan to acquire all the channels using the CAEN V1742 board, a 32 channels 5 GS/s digitizer based on the DRS4 ( Domino Ring Sampler) chip.

Three such boards have successfully been used during the 2015 and 2016 tests at the DAFNE Beam Test Facility (BTF), where a complete DAQ system, prototypal to the one which will be used for the final experiment, has been set up. The DAQ system includes a centralized Run Control unit, which
interacts with a distributed set of software modules handling the V1742 boards readout, and an Event Builder, which collects the data from the boards and creates the final raw events. In this talk we will report on the details of the DAQ system, with specific reference to our experience with the V1742 board.

**Tertiary Keyword (Optional):**

**Secondary Keyword (Optional):**

**Primary Keyword (Mandatory):**

**DAQ**

---

**Posters A / Break / 243**

**Evaluation of clustering algorithms at the \(< 1\) GeV energy scale for the electromagnetic calorimeter of the PADME experiment**

**Author:** Emanuele Leonardi\(^1\)

**Co-authors:** Gabriele Piperno \(^2\); Mauro Raggi \(^3\)

\(^1\) INFN Roma  
\(^2\) INFN LNF  
\(^3\) LNF INFN

**Corresponding Authors:** emanuele.leonardi@roma1.infn.it, mauro.raggi@cern.ch, gabriele.piperno@lnf.infn.it

The long standing problem of reconciling the cosmological evidence of the existence of dark matter with the lack of any clear experimental observation of it, has recently revived the idea that the new particles are not directly connected with the Standard Model gauge fields, but only through mediator fields or "portals", connecting our world with new "secluded" or "hidden" sectors. One of the simplest models just adds an additional U(1) symmetry, with its corresponding vector boson \(A'\).

At the end of 2015 INFN has formally approved a new experiment, PADME (Positron Annihilation into Dark Matter Experiment), to search for invisible decays of the \(A'\) at the DAFNE BTF in Frascati. The experiment is designed to detect dark photons produced in positron on fixed target annihilations \((e^+e^- \to \gamma A')\) decaying to dark matter by measuring the final state missing mass.

The collaboration aims to complete the design and construction of the experiment by the end of 2017 and to collect \(\sim 10^{13}\) positrons on target by the end of 2018, thus allowing to reach the \(\epsilon \sim 10^{-3}\) sensitivity up to a dark photon mass of \(\sim 24\) MeV/c\(^2\).

One of the key roles of the experiment will be played by the electromagnetic calorimeter, which will be used to measure the properties of the final state recoil \(\gamma\), as the final error on the measurement of the \(A'\) mass will directly depend on its energy, time and angular resolutions. The calorimeter will be built using 616 2x2x22 cm\(^3\) BGO crystals recovered from the electromagnetic calorimeter end-caps of the L3 experiment at LEP. The crystals will be oriented with the long axis parallel to the beam direction and will be disposed in a roughly circular shape with a central hole to avoid the pile up due to the large number of low angle bremsstrahlung photons.

The total energy and position of the electromagnetic shower generated by a photon impacting on the calorimeter can be reconstructed by collecting the energy deposits in the set of crystals intercepted by the shower. This set of crystals is not known a priori and must be reconstructed with an ad hoc clustering algorithm.
In PADME we tested two different clustering algorithms: one based on the definition of a squared set of crystals centered on a local energy maximum, and one based on a modified version of the “island” algorithm in use by the CMS collaboration, where a cluster starts from a local energy maximum and is expanded by including available neighboring crystals.

In this talk we will describe the implementations of the two algorithms and report on the energy and spatial resolution obtained with them at the PADME energy scale (< 1 GeV), both with a GEANT4 based simulation and with an existing 5x5 matrix of BGO crystals tested at the DAFNE Beam Test Facility (BTF).

Secondary Keyword (Optional): Reconstruction

Primary Keyword (Mandatory): Algorithms

Posters B / Break / 244

LHCbDIRAC as Apache Mesos microservices

Authors: Ben Couturier¹; Christophe Haen¹

¹ CERN

Corresponding Authors: christophe.denis.haen@cern.ch, ben.couturier@cern.ch

The LHCb experiment relies on LHCbDIRAC, an extension of DIRAC, to drive its offline computing. This middleware provides a development framework and a complete set of components for building distributed computing systems. These components are currently installed and ran on virtual machines (VM) or bare metal hardware. Due to the increased load of work, high availability is becoming more and more important for the LHCbDIRAC services, and the current installation model is showing its limitations.

Apache Mesos is a cluster manager which aims at abstracting heterogeneous physical resources on which various tasks can be distributed thanks to so called “framework”. The Marathon framework is suitable for long running tasks such as the DIRAC services, while the Chronos framework meets the needs of cron-like tasks like the DIRAC agents. A combination of the service discovery tool Consul together with HAProxy allows to expose the running containers to the outside world while hiding their dynamic placements.

Such an architecture would bring a greater flexibility in the deployment of LHCbDirac services, allowing for easier deployment maintenance and scaling of services on demand (e.g LHCbDirac relies on 138 services and 116 agents). Higher reliability would also be easier, as clustering is part of the toolset, which allows constraints on the location of the services.

This paper describes the investigations carried out to package the LHCbDIRAC and DIRAC components into Docker containers and orchestrate them using the previously described set of tools.

Tertiary Keyword (Optional): Cloud technologies

Secondary Keyword (Optional):
Computing middleware

Primary Keyword (Mandatory):
Computing facilities

Track 1: Online Computing / 245

STAR Online Meta Data Collection Framework: Integration with the Pre-existing Controls Infrastructure

Authors: Dmitry Arkhipkin; Jerome LAURET

1 Brookhaven National Laboratory

Corresponding Authors: arkhipkin@bnl.gov, jlauret@bnl.gov

One of the STAR experiment’s modular Messaging Interface and Reliable Architecture framework (MIRA) integration goals is to provide seamless and automatic connections with the existing control systems. After an initial proof of concept and operation of the MIRA system as a parallel data collection system for online use and real-time monitoring, the STAR Software and Computing group is now working on the integration of Experimental Physics and Industrial Control System (EPICS) with MIRA’s interfaces. This integration goals are to allow functional interoperability and, later on, to replace the existing/legacy Detector Control System components at the service level.

In this report, we describe the evolutionary integration process on the example of EPICS Alarm Handler conversion. We review the complete upgrade procedure starting with the integration of EPICS-originated alarm signals propagation into MIRA, followed by the replacement of the existing operator interface based on Motif Editor and Display Manager (MEDM) with modern portable web-based Alarm Handler interface. To achieve this aim, we have built an EPICS-to-MQTT bridging service, and recreated the functionality of the original Alarm Handler using low-latency web messaging technologies. The integration of EPICS alarm handling into our messaging framework allowed STAR to improve the DCS alarm awareness of existing STAR DAQ and RTS services, which use MIRA as a primary source of experiment control information.

Secondary Keyword (Optional):
Monitoring

Primary Keyword (Mandatory):
Control systems

Tertiary Keyword (Optional):

Track 3: Distributed Computing / 246

Provenance-aware optimization of workload for distributed data production

Authors: Dzmitry Makatun; Hana Rudová; Jerome LAURET; Michal Sumbera

1 Acad. of Sciences of the Czech Rep. (CZ)
2 Faculty of Informatics, Masaryk University, Brno, Czech Republic
3 Brookhaven National Laboratory
4 Nuclear Physics Institute, Acad. of Sciences of the Czech Rep. (CZ)
Distributed data processing in High Energy and Nuclear Physics (HENP) is a prominent example of big data analysis. Having petabytes of data being processed at tens of computational sites with thousands of CPUs, standard job scheduling approaches either do not address well the problem complexity or are dedicated to one specific aspect of the problem only (CPU, network or storage). As a result, the general orchestration of the system is left to the production managers and requires reconsideration each time new resources are added or withdrawn. In previous research we have developed a new job scheduling approach dedicated to distributed data production – an essential part of data processing in HENP (pre-processing in big data terminology). In our approach the load balancing across sites is provided by forwarding data in peer-to-peer manner, but guided by a centrally created (and periodically updated) plan, aiming to achieve global optimality. The planner considers network and CPU performance as well as available storage space at each site and plans data movements between them in order to maximize an overall processing throughput. In this work we extend our approach by distributed data production where multiple input data sources are initially available. Multi-source or provenance is common in user analysis scenario whereas the produced data may be immediately copied to several destinations. The initial input data set would hence be already partially replicated to multiple locations and the task of the scheduler is to maximize overall computational throughput considering possible data movements and CPU allocation. In particular, the planner should decide if it makes sense to transfer files to other sites or if they should be processed at the site where they are available. Reasoning about multiple data replicas allows to broaden the planner applicability beyond the scope of data production towards user analysis in HENP and other big data processing applications. In this contribution, we discuss the load balancing with multiple data sources, present recent improvements made to our planner and provide results of simulations which demonstrate the advantage against standard scheduling policies for the new use case. The studies have shown that our approach can provide a significant gain in overall computational performance in a wide scope of simulations considering realistic size of computational Grid, background network traffic and various input data distribution. The approach is scalable and adjusts itself to resource outage/addition/reconfiguration which becomes even more important with growing usage of cloud resources. The reasonable complexity of the underlying algorithm meets requirements for online planning of computational networks as large as one of the currently largest HENP experiments.

Tertiary Keyword (Optional):
Reconstruction

Secondary Keyword (Optional):
Distributed data handling

Primary Keyword (Mandatory):
Distributed workload management

Track 3: Distributed Computing / 247

The Cloud Area Padovana: from pilot to production

Authors: Alberto Crescente¹; Alvise Dorigo¹; Federica Fanzago¹; Fulvia Costa²; Lisa Zangrando¹; Marco Verlato¹; Massimo Sgaravatto¹; Paolo Andreetto³; Sergio Fantinel³; Sergio Traldi³

¹ Universita e INFN, Padova (IT)
² INFN Padova
³ INFN Laboratori Nazionali di Legnaro

Corresponding Authors: sergio.fantinel@lnl.infn.it, massimo.sgaravatto@pd.infn.it, sergio.traldi@pd.infn.it, fulvia.costa@pd.infn.it, marco.verlato@pd.infn.it, alvise.dorigo@pd.infn.it, paolo.andreetto@pd.infn.it, federica.fanzago@cern.ch, alberto.crescente@cern.ch, lisa.zangrando@pd.infn.it
The Cloud Area Padovana has been running for almost two years. This is an OpenStack-based scientific cloud, spread across two different sites: the INFN Padova Unit and the INFN Legnaro National Labs.

The hardware resources have been scaled horizontally and vertically, by upgrading some hypervisors and by adding new ones: currently it provides about 1100 cores.

Some in-house developments were also integrated in the OpenStack dashboard, such as a tool for user and project registrations with direct support for the INFN-AAI Identity Provider as a new option for the user authentication.

In collaboration with the EU-funded Indigo DataCloud project, the integration with Docker-based containers have been experimented and will be available in production soon.

This computing facility now satisfies the computational and storage demands of more than 70 users afferent to about 20 research projects.

We present here the architecture of this Cloud infrastructure, the tools and procedures used to operate it. We also focus on the lessons learnt in these two years, describing the problems that were found and the corrective actions that had to be applied. We also discuss about the chosen strategy for upgrades, which combines the need to promptly integrate the OpenStack new developments, the demand to reduce the downtimes of the infrastructure, and the need to limit the effort requested for such updates.

We also discuss how this Cloud infrastructure is being used. In particular we focus on two big physics experiments which are intensively exploiting this computing facility: CMS and SPES.

CMS deployed on the cloud a complex computational infrastructure, composed of several user interfaces for job submission in the Grid environment/local batch queues or for interactive processes; this is fully integrated with the local Tier-2 facility. To avoid a static allocation of the resources, an elastic cluster, based on cernVM, has been configured: it allows to automatically create and delete virtual machines according to the user needs.

SPES, using a client-server system called TraceWin, exploits INFN’s virtual resources performing a very large number of simulations on about a thousand nodes elastically managed.

Secondary Keyword (Optional):
Computing facilities

Primary Keyword (Mandatory):
Cloud technologies

Tertiary Keyword (Optional):

Track 6: Infrastructures / 248

Evolution of the Building Management System in the INFN CNAF Tier-1 datacenter facility

Authors: Andrea Mazza¹; Luca dell’Agnello²; Massimo Donatelli²; Michele Onofri¹; Pier Paolo Ricci¹

¹ INFN CNAF
² INFN-CNAF

Corresponding Authors: massimo.donatelli@cnaf.infn.it, michele.onofri@cnaf.infn.it, andrea.mazza@cnaf.infn.it, pierpaolo.ricci@cnaf.infn.it

The INFN CNAF Tier-1 computing center is composed by 2 different main rooms containing IT resources and 4 additional locations that hosts the necessary technology infrastructures providing the electrical power and refrigeration to the facility. The power supply and continuity are ensured by a dedicated room with three 15,000 to 400 V transformers in a separate part of the principal building and 2 redundant 1.4MW diesel rotary uninterruptible power supplies. The cooling is provided by six free cooling chillers of 320 kW each with a N+2 redundancy configuration. Clearly, considering the complex physical distribution of the technical plants, a detailed Building Management System (BMS) was designed and implemented as part of the original project in order to monitor and collect...
all the necessary information and for providing alarms in case of malfunctions or major failures. After almost 10 years of service, a revision of the BMS system was somewhat necessary. In addition, the increasing cost of electrical power is nowadays a strong motivation for improving the energy efficiency of the infrastructure. Therefore the exact calculation of the power usage effectiveness (PUE) metric has become one of the most important factors when aiming for the optimization of a modern data center. For these reasons, an evolution of the BMS system was designed using the Schneider StruxureWare infrastructure hardware and software products. This solution demonstrates to be a natural and flexible development of the previous TAC Vista software with advantages in the ease of use and the possibility to customize the data collection and the graphical interfaces display. Moreover, the addition of protocols like open standard Web services gives the possibility to communicate with the BMS from custom user application and permits the exchange of data and information through the Web between different third-party systems. Specific Web services SOAP requests has been implemented in our Tier-1 monitoring system in order to collect historical trends of power demands and calculate the partial PUE (pPUE) of specific area of the infrastructure. This would help in the identification of "spots" that may need optimization in the power usage. The StruxureWare system maintains compatibility with standard protocols like Modbus as well as native LonWorks, making possible reusing the existing network between physical locations as well as a considerable number of programmable controller and I/O modules that interact with the facility. The high increase in detailed statistical information of power consumption and the HVAC (heat, ventilation and air conditioning) parameters could prove to be a very valuable strategic choice for improving the overall PUE. This will bring remarkable benefits for the overall management costs, despite the limits in the actual location of the facility, and it will help the process of building a more energy efficient data center that embraces the concept of green IT.

Tertiary Keyword (Optional):
Monitoring

Secondary Keyword (Optional):
Computing facilities

Primary Keyword (Mandatory):
Control systems

Track 9: Future Directions / 250

The LHCb software and computing upgrade for Run3: opportunities and challenges

Authors: Concezio Bozzi¹; Stefan Roiser²

¹ CERN and INFN Ferrara
² CERN

Corresponding Authors: stefan.roiser@cern.ch, bozzi@fe.infn.it

The LHCb detector will be upgraded for the LHC Run 3 and will be readout at 40 MHz, with major implications on the software-only trigger and offline computing. If the current computing model is kept, the data storage capacity and computing power required to process data at this rate, and to generate and reconstruct equivalent samples of simulated events, will exceed the current capacity by a couple of orders of magnitude. A redesign of the software framework, including scheduling, the event model, the detector description and the conditions database, is needed to fully exploit the computing power of new architectures. Data processing and the analysis model will also change towards an early streaming of different data types, in order to limit storage resources, with further implications for the data analysis workflows. Fast simulation will allow to obtain a reasonable parameterization of the detector response in considerably less computing time. Finally, the upgrade of LHCb will be a good opportunity to review and implement changes in the domains of software design, test and review, and analysis workflow and preservation. In this contribution, activities and recent results in all the above areas are presented.
An educational distributed Cosmic Ray detector network based on ArduSiPM using microcontrollers as data acquisition node NTP protocol as time distribution and IoT technology for data aggregation.

Author: Valerio Bocci

1 Universita e INFN, Roma I (IT)

Corresponding Author: valerio.bocci@cern.ch

The advent of microcontrollers with enough CPU power and with analog and digital peripherals give the possibility to design a complete acquisition system in one chip. The existence of an world wide data infrastructure as internet allows to think at distributed network of detectors capable to elaborate and send data or respond to settings commands.

The internet infrastructure allow us to do things unthinkable a few years ago, like to distribute the absolute time with tens of milliseconds precision to simple devices far apart from a few meters to thousands of kilometers and to create a Crowdsourcing experiment platform using simple detectors. The terms of IoT (Internet of Things) define a set of data communication protocols and the capability of single embedded electronics objects to communicate using the internet.

The MQTT (Message Queue Telemetry Transport) is one of the main protocol used in IoT device for data transmission over TCP/IP, the client version can run easily in nowadays microcontrollers, the MQTT broker (the server version) can run also in credit card-sized single-board computers as well in big server.

The ArduSiPM (1) is an easy hand-held battery operated data acquisition system based with an Arduino board, which is used to detect cosmic rays and nuclear radiation.

The ArduSiPM uses an Arduino DUE (an open Software/Hardware board based on an ARM Cortex-M3 microcontroller) as processor board and a piggyback custom designed board (Shield), these are controlled by custom developed software and interface. The Shield contains different electronics features both to monitor, to set and to acquire the SiPM signals using the microcontroller board. The SiPM photon counting detector can be coupled to a cheap plastic scintillator to realize a cosmic ray detector (mainly muon particles). An ArduSiPM channel give informations about rate of events, arrival time and number of photons produced by muons, it contains all the feature from controls to data acquisition typical of High Energy Physics channel at a cost affordable for single user or school.

The ArduSiPM send data over serial protocol instead of use ethernet interface to the network. In the market comes up SoCs (System on Chip) like the Espressif ESP8266 a low-cost Wi-Fi chip with full TCP/IP stack and a 32-bit RISC CPU running at 80 MHz.ESP8266 can be used to manage MQTT packets and to retrieve using Network Time Protocol (NTP) the absolute time from the network, with a precision of tens milliseconds. The network time can be used from a cloud of ArduSiPMs to detect offline coincidence events linked to Ultra High Energy Cosmic Ray to realize an educational distributed detector with a centralized MQTT broker that concentrate data.

(1) The ArduSiPM a compact trasportable Software/Hardware Data Acquisition system for SiPM detector.
A Security Monitoring Framework For Container Based HEP Computing Infrastructures

Author: Andres Gomez Ramirez
Co-authors: Camilo Lara; Udo Wolfgang Kebschull

1 Johann-Wolfgang-Goethe Univ. (DE)

Corresponding Authors: andres.gomez@cern.ch, camilo.ernesto.lara.martinez@cern.ch, uk@rz.uni-frankfurt.de

Distributed computing infrastructures require automatic tools to strengthen, monitor and analyze the security behavior of computing devices. These tools should inspect monitoring data such as resource usage, log entries, traces and even processes' system calls. They also should detect anomalies that could indicate the presence of a cyber-attack. Besides, they should react to attacks without administrator intervention, depending on custom configuration parameters. We describe the development of a novel framework that implements these requirements for HEP systems. It is based on Linux container technologies. A previously unexplored deployment of Kubernetes on top of Mesos as Grid site container based batch system, and Heapster as a monitoring solution are being utilized. We show how we achieve a fully virtualized environment that improves the security by isolating services and jobs without an appreciable performance impact. We also describe an novel benchmark dataset for Machine Learning based Intrusion Prevention and Detection Systems on Grid computing. This dataset is built upon resource consumption, logs, and system call data collected from jobs running in a test site that has been developed for the ALICE Grid at CERN as a described framework’s proof of concept. Further, we will use this dataset to develop a Machine Learning module that will be integrated with the framework, performing the autonomous Intrusion Detection task.
MCBooster: a library for fast Monte Carlo generation of phase-space decays in massively parallel platforms.

Author: Antonio Augusto Alves Junior
Co-author: Michael David Sokoloff

1 University of Cincinnati (US)

Corresponding Authors: antonio.augusto.alves.junior@cern.ch, michael.david.sokoloff@cern.ch

MCBooster is a header-only, C++11-compliant library for the generation of large samples of phase-space Monte Carlo events on massively parallel platforms. It was released on GitHub in the spring of 2016. The library core algorithms implement the Raubold-Lynch method; they are able to generate the full kinematics of decays with up to nine particles in the final state. The library supports the generation of sequential decays as well as the parallel evaluation of arbitrary functions over the generated events.

The output of MCBooster completely accords with popular and well-tested software packages such as GENBOD (W515 from CERNLIB) and TGenPhaseSpace from the ROOT framework. MCBooster is developed on top of the Thrust library and runs on Linux systems. It deploys transparently on NVidia CUDA-enabled GPUs as well as multicore CPUs.

This contribution summarizes the main features of MCBooster. A basic description of the user interface and some examples of applications are provided, along with measurements of performance in a variety of environments.

Tertiary Keyword (Optional):
Analysi tools and techniques

Secondary Keyword (Optional):
High performance computing

Primary Keyword (Mandatory):
Event generators

Towards redundant services in dCache

Authors: Albert Rossi; Anupam Ashish; Christian Bernardt; Dmitry Litvinsev; Gerd Behrmann; Jürgen Starek; Marina Sahakyan; Patrick Fuhrmann; Paul Millar; Tigran Mkrtchyan

1 Fermilab
2 DESY
3 Deutsches Elektronen-Synchrotron (DE)
4 FNAL
5 NEIC

Corresponding Authors: litvinse@fnal.gov, tigran.mkrtchyan@desy.de, arossi@fnal.gov, christian.bernardt@desy.de, anupam.ashish@desy.de, juergen.starek@desy.de, patrick.fuhrmann@desy.de, marina.sahakyan@desy.de, paul.millar@desy.de, behrmann@gmail.com

As a robust and scalable storage system, dCache has always allowed the number of storage nodes and user accessible endpoints to be scaled horizontally, providing several levels of fault tolerance and high throughput. Core management services like the POSIX name space and central load balancing...
components however are merely vertically scalable. This greatly limits the scalability of the core services as well as provides single points of failures. Such single points of failures are not just a concern for fault tolerance, but also prevent zero downtime rolling upgrades.

For large sites, redundant and horizontally scalable services translate to higher uptime, easier upgrades, and higher maximum request rates. In an effort to move towards redundant services in dCache, we are reporting on attacking this problem at three levels. At the lowest level dCache needs a service to locate the various dCache nodes. In the past a simple non-redundant UDP service was used for this, but in the latest release this functionality has been ported to Apache ZooKeeper. ZooKeeper is originally part of Hadoop and is a redundant, persistent, hierarchical directory service with strong ordering guarantees. In particular, the strong ordering guarantees make ZooKeeper perfect for coordinating higher level services. On top of the location service, dCache uses a common message passing system to communicate between various services. In the past this relied on a simple star topology, with all messages going through a central broker. This broker forms a single point of failure and is possibly a bottleneck under extreme load conditions. This problem is addressed with a multi-rooted multi-path topology consistent of a set of core brokers forming a fully connected mesh and all other services connecting to all brokers. Finally, each of the central services is made scalable and redundant. For some services this is trivial, as they maintain minimal internal state. For others, the ability of Apache ZooKeeper to act as a coordination service is central. In some cases a leader election procedure ensures that various background tasks are only executed on a single node. In other cases shared state can be stored in the ZooKeeper, e.g. to ensure that a file is only staged from tape once. Further changes to the internal message routing logic will allow load balancing over multiple instances of a service.

The first two steps outlined above will have been deployed in production by the time this paper has been published. Redundancy and scalability of higher level services is currently only available for the trivial services, while other services will be extended over the following releases.

Primary Keyword (Mandatory):
Storage systems

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

Track 1: Online Computing / 256

Numerical Optimization for Fast Track Finding Based on the Artificial Retina Algorithm

Author: Maxim Borisyak

Co-authors: Andrey Ustyuzhanin ¹; Denis Derkach ¹; Mikhail Belous ²

¹ National Research University Higher School of Economics (HSE) (RU); Yandex School of Data Analysis (RU)
² Yandex School of Data Analysis (RU)

Corresponding Authors: maxim.borisyak@cern.ch, denis.derkach@cern.ch, mikhail.belous@cern.ch, andrey.ustyuzhanin@cern.ch

High-energy physics experiments rely on reconstruction of the trajectories of particles produced at the interaction point. This is a challenging task, especially in the high track multiplicity environment generated by p-p collisions at the LHC energies. A typical event includes hundreds of signal examples (interesting decays) and a significant amount of noise (uninteresting examples).

This work describes a modification of Artificial Retina Algorithm for fast track finding: numerical optimization methods were adopted for fast local track search. This approach allows for considerably reduction of the total computational time per event. Test results on simplified simulated model of LHCb VELO (VErtex LOcator) detector are presented. Also this approach is well-suited for
implementation for paralleled computations as GPU which looks very attractive in the context of upcoming detector upgrade.

**Tertiary Keyword (Optional):**
Algorithms

**Secondary Keyword (Optional):**
Reconstruction

**Primary Keyword (Mandatory):**
Artificial intelligence/Machine learning

---

**Track 6: Infrastructures / 257**

**SWAN: a Service for Web-Based Data Analysis in the Cloud**

**Authors:** Danilo Piparo\(^1\); Enric Tejedor Saavedra\(^1\)

**Co-authors:** Jakub Moscicki\(^1\); Luca Mascetti\(^2\); Massimo Lamanna\(^1\); Pere Mato Vila\(^1\)

\(^1\) CERN

**Corresponding Authors:** luca.mascetti@cern.ch, jakub.moscicki@cern.ch, pere.mato@cern.ch, enric.tejedor.saavedra@cern.ch, massimo.lamanna@cern.ch, danilo.piparo@cern.ch

SWAN is a novel service to perform interactive data analysis in the cloud. SWAN allows users to write and run their data analyses with only a web browser, leveraging the widely-adopted Jupyter notebook interface. The user code, executions and data live entirely in the cloud. SWAN makes it easier to produce and share results and scientific code, access scientific software, produce tutorials and demonstrations as well as preserve analyses. Furthermore, it is also a powerful tool for non-scientific data analytics.

The SWAN backend combines state-of-the-art software technologies, like Docker containers, with a set of existing IT services such as user authentication, virtual computing infrastructure, mass storage, file synchronisation and sharing, specialised clusters and batch systems. In this contribution, the architecture of the service and its integration with the aforementioned CERN services is described. SWAN acts as a “federator of services” and the reasons why this feature boosts the existing CERN IT infrastructure are reviewed.

Furthermore, the main characteristics of SWAN are compared to similar products offered by commercial and free providers. Use-cases extracted from workflows at CERN are outlined. Finally, the experience and feedback acquired during the first months of its operation are discussed.

**Tertiary Keyword (Optional):**
Data processing workflows and frameworks/pipelines

**Secondary Keyword (Optional):**
Analysis tools and techniques

**Primary Keyword (Mandatory):**
Cloud technologies

---

**Track 5: Software Development / 258**
Expressing Parallelism in ROOT

Authors: Danilo Piparo¹; Enric Tejedor Saavedra¹

Co-authors: Enrico Guiraud ²; Gerardo Ganis ¹; Lorenzo Moneta ³; Pere Mato Vila ¹; Philippe Canal ³; Xavier Valls Pla ⁴

¹ CERN
² Università degli Studi e INFN Milano (IT)
³ Fermi National Accelerator Lab. (US)
⁴ University Jaume I (ES)

Corresponding Authors: pere.mato@cern.ch, danilo.piparo@cern.ch, philippe.canal@cern.ch, lorenzo.moneta@cern.ch, gerardo.ganis@cern.ch, enric.tejedor.saavedra@cern.ch, enrico.guiraud@cern.ch, xavier.valls.pla@cern.ch

The need for processing the ever-increasing amount of data generated by the LHC experiments in a more efficient way has motivated ROOT to further develop its support for parallelism. Such support is being tackled both for shared-memory and distributed-memory environments.

The incarnations of the aforementioned parallelism are multi-threading, multi-processing and cluster-wide executions. In the area of multi-threading, we discuss the new implicit parallelism and related interfaces, as well as the new building blocks to safely operate with ROOT objects in a multi-threaded environment. Regarding multi-processing, we review the new MultiProc framework, comparing it with similar tools (e.g. multiprocessing module in Python). Finally, as an alternative to PROOF for cluster-wide executions, we introduce the efforts on integrating ROOT with state-of-the-art distributed data processing technologies like Spark, both in terms of programming model and runtime design (with EOS as one of the main components).

For all the levels of parallelism, we discuss, based on real-life examples and measurements, how our proposals can increase the productivity of scientists.

Tertiary Keyword (Optional):
High performance computing

Secondary Keyword (Optional):
Analysis tools and techniques

Primary Keyword (Mandatory):
Parallelization

Towards automation of data quality system for CERN CMS experiment

Author: Maxim Borisyak¹

Co-authors: Andrey Ustyuzhanin ¹; Dmitry Smolyakov ²; Jean-Roch Vlimant ³; Maria Stenina ⁴; Maurizio Pierini ⁵

¹ National Research University Higher School of Economics (HSE) (RU); Yandex School of Data Analysis (RU)
² Yandex School of Data Analysis (RU)
³ California Institute of Technology (US)
⁴ Yandex (RU)
⁵ CERN
Daily operation of a large scale experimental setup is a challenging task both in terms of maintenance and monitoring. In this work we describe an approach for automated Data Quality system. Based on the Machine Learning methods it can be trained online on manually-labeled data by human experts. Trained model can assist data quality managers filtering obvious cases (both good and bad) and asking for further estimation only of fraction of poorly-recognizable datasets.

The system is trained on CERN open data portal data published by CMS experiment. We demonstrate that our system is able to save at least 20% of person power without increase in pollution (false positive) and loss (false negative) rates. In addition, for data not labeled automatically system provides its estimates and hints for a possible source of anomalies which leads to overall improvement of data quality estimations speed and higher purity of collected data.

Track 5: Software Development / 260

The New ROOT Interface: Jupyter Notebooks

Authors: Danilo Piparo\textsuperscript{1}; Enric Tejedor Saavedra\textsuperscript{2}

Co-author: Pere Mato Vila \textsuperscript{1}

\textsuperscript{1} CERN

Corresponding Authors: danilo.piparo@cern.ch, pere.mato@cern.ch, enric.tejedor.saavedra@cern.ch

Notebooks represent an exciting new approach that will considerably facilitate collaborative physics analysis.

They are a modern and widely-adopted tool to express computational narratives comprising, among other elements, rich text, code and data visualisations. Several notebook flavours exist, although one of them has been particularly successful: the Jupyter open source project.

In this contribution we demonstrate how the ROOT framework is integrated with the Jupyter technology, reviewing features such as an unprecedented integration of Python and C++ languages and interactive data visualisation with JavaScript ROOT. In this context, we show the potential of the complete interoperability of ROOT with other analysis ecosystems such as SciPy.

We discuss through examples and use-cases how the notebook approach boosts the productivity of physicists, engineers and non-coding lab scientists. Opportunities in the field of outreach, education and open-data initiatives are also reviewed.
Muon trigger for mobile phones

Author: Maxim Borisyak

Co-authors: Andrey Ustyuzhanin\textsuperscript{1}; Chase Owen Shimmin\textsuperscript{2}; Michael J Mulhearn\textsuperscript{3}; Mikhail Usvyatsov\textsuperscript{4}

\textsuperscript{1} National Research University Higher School of Economics (HSE) (RU); Yandex School of Data Analysis (RU)
\textsuperscript{2} University of California Irvine (US)
\textsuperscript{3} University of California Davis (US)
\textsuperscript{4} Moscow Institute of Physics and Technology (RU); Skolkovo Institute of Science and Technology (RU)

The CRAYFIS experiment proposes usage of private mobile phones as a ground detector for Ultra High Energy Cosmic Rays. Interacting with Earth’s atmosphere they produce extensive particle showers which can be detected by cameras on mobile phones. A typical shower contains minimally-ionizing particles such as muons. As they interact with CMOS detector they leave low-energy tracks that sometimes are hard to distinguish from random detector noise. Triggers that rely on the presence of very bright pixels within an image frame are not efficient in this case.

We present a muon trigger based on Convolutional Neural Networks which play role of trigger sequence and are evaluated in a ‘lazy’ manner: response of the successive layer is computed only if activation of the current layer satisfy continuation criterion. Usage of the neural networks allows to increase sensitivity considerably. Also this modification allows for execution of the trigger under limited computational power constraint, e.g. on mobile phones.

Secondary Keyword (Optional):
Trigger

Primary Keyword (Mandatory):
Artificial intelligence/Machine learning

Tertiary Keyword (Optional):

Flexible trigger menu implementation on the Global Trigger for the CMS Level-1 trigger upgrade

Author: Takashi Matsushita\textsuperscript{1}

\textsuperscript{1} Austrian Academy of Sciences (AT)

Corresponding Author: takashi.matsushita@cern.ch

In 2016 the Large Hadron Collider (LHC) will continue to explore the physics at the high-energy frontier. The integrated luminosity is expected to be about $25 \text{fb}^{-1}$ in 2016 with the estimated peak
luminosity of around $1.1 \times 10^{34} \text{ cm}^{-2} \text{ s}^{-1}$ and the peak mean pile-up of about 30. The CMS experiment will upgrade its hardware-based Level-1 trigger system to keep its performance for new physics searches and precision measurements with the data collected in the higher luminosities.

The Global Trigger is the final step of the CMS Level-1 Trigger and implements a trigger menu, a set of selection requirements applied to the final list of objects from calorimeter and muon triggers, for reducing the 40 MHz collision rate to 100 kHz. The Global Trigger is being upgraded with the state-of-the-art FPGA processors on Advanced Mezzanine Cards with optical links running at 10 GHz in a MicroTCA crate. The upgraded system will benefit from increased processing resources, enabling more algorithms at a time than previously possible as well as allowing CMS to be more flexible in how it handles the available trigger bandwidth. Algorithms for a trigger menu, including topological requirements on multi-objects, can be realised on the Global Trigger using the newly developed trigger menu specification grammar. Analysis-like trigger algorithms can be represented in an intuitive manner then the algorithms are translated to corresponding VHDL code blocks to build a firmware. The grammar can be extended in future as the needs arise. The experience of implementing trigger menus on the upgraded Global Trigger system will be presented.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):
Trigger

Track 5: Software Development / 263

Developing and optimizing applications for the Hadoop environment

Author: Prasanth Kothuri

Co-authors: Daniel Lanza Garcia; Joeri Hermans; Kacper Surdy; Zbigniew Baranowski

1 CERN
2 Ministere des affaires etrangeres et europeennes (FR)
3 Universiteit Maastricht (NL)

Corresponding Authors: kacper.surdy@cern.ch, zbigniew.baranowski@cern.ch, prasanth.kothuri@cern.ch, daniel.lanza@cern.ch, joeri.hermans@cern.ch

This contribution is about sharing our recent experiences of building Hadoop based application. Hadoop ecosystem now offers myriad of tools which can overwhelm new users, yet there are successful ways these tools can be leveraged to solve problems. We look at factors to consider when using Hadoop to model and store data, best practices for moving data in and out of the system and common processing patterns, at each stage relating with the real world experience gained while developing such application. We share many of the design choices, tools developed and how to profile a distributed application which can be applied for other scenarios as well. In conclusion, the goal of the presentation is to provide guidance to architect Hadoop based application and share some of the reusable components developed in this process.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Software development process and tools

Primary Keyword (Mandatory):
New operator assistance features in the CMS Run Control System

Author: Hannes Sakulin

Co-authors: Andre Georg Holzner 1; Attila Racz 1; Benjamin G. Craigs 1; Christian Deldicque 1; Christoph Paus 2; Christoph Schwick 1; Cristian Contescu 4; Dainius Simelevicius 5; Dominique Gigi 1; Emilio Meschi 2; Frank Glege 1; Frans Meijers 1; Georgiana Lavinia Darlea 1; Guillermo Gomez Ceballos Retuerto 1; James Gordon Branson 2; Jean-Marc Olivier Andre 4; Jeroen Hegeman 1; Jonathan Fulcher 1; Lorenzo Masetti 1; Luciano Orsini 1; Maciej Szymon Gladki 1; Marc Dobson 1; Marco Pieri 2; Nicolas Doualot 4; Olivier Chaze 1; Petr Zejdl 4; Philipp Brummer 1; Raul Jimenez-Estupinan 1; Remi Mommersen 4; Samim Erhan 2; Sergio Cittolin 2; Srecko Morovic 1; Thomas Reis 1; Ulf Behrens 7; Vivian O’Dell 8; Zeynep Demiragli 3

1 CERN
2 Univ. of California San Diego (US)
3 Massachusetts Inst. of Technology (US)
4 Fermi National Accelerator Lab. (US)
5 Vilnius University (LT)
6 Univ. of California Los Angeles (US)
7 Deutsches Elektronen-Synchrotron (DE)
8 Fermi National Accelerator Laboratory (FNAL)

Corresponding Authors: luciano.orsini@cern.ch, cristian.contescu@cern.ch, guillelmo.gomez.ceballos@cern.ch, christoph.schwick@cern.ch, hannes.sakulin@cern.ch, dainius.simelevicius@cern.ch, nicolas.doualot@cern.ch, marc.dobson@cern.ch, zeynep.demiragli@cern.ch, srecko.morovic@cern.ch, sergio.cittolin@cern.ch, frank.glege@cern.ch, christoph.steck@cern.ch, thomas.reis@cern.ch, lorenzo.masetti@cern.ch, petr.zejdl@cern.ch, branson@ucsd.edu, paus@mit.edu, odell@fnal.gov, dominique.gigi@cern.ch, benjamin.gordon.craigs@gmail.com, olivier.chaze@cern.ch, marco.pieri@cern.ch, christian.deldicque@cern.ch, jeroen.hegeman@cern.ch, remigius.mommersen@cern.ch, lavinia.darlea@cern.ch, philipp.maximilian.brummer@cern.ch, maciej.szymon.gladki@cern.ch, samim.erhan@cern.ch, ulf.behrens@cern.ch, jm.andre@cern.ch, emilio.meschi@cern.ch, andre.georg.holzner@cern.ch, frans.meijers@cern.ch, jonathan.fulcher@cern.ch, raul.jimenez.estupinan@cern.ch

The Run Control System of the Compact Muon Solenoid (CMS) experiment at CERN is a distributed Java web application running on Apache Tomcat servers. During Run-1 of the LHC, many operational procedures have been automated. When detector high voltages are ramped up or down or upon certain beam mode changes of the LHC, the DAQ system is automatically partially reconfigured with new parameters. Certain types of errors such as errors caused by single-event upsets may trigger an automatic recovery procedure. Furthermore, the top-level control node continuously performs cross-checks to detect sub-system actions becoming necessary because of changes in configuration keys, changes in the set of included front-end drivers or because of potential clock instabilities. The operator is guided to perform the necessary actions through graphical indicators displayed next to the relevant command buttons in the user interface. Through these indicators, consistent configuration of CMS is insured. However, manually following the indicators can still be inefficient at times. A new assistant to the operator has therefore been developed that can automatically perform all the necessary actions in a streamlined order. If additional problems arise, the new assistant tries to automatically recover from these. With the new assistant, a run can be started from any state of the subsystems with a single click. An ongoing run may be recovered with a single click, once the appropriate recovery action has been selected. We review the automation features of the CMS run control system and discuss the new assistant in detail including first operational experience.
The ZEUS data preservation project

Author: Andrii Verbytskyi

1 Max-Planck-Institut fuer Physik (Werner-Heisenberg-Institut) (D)

Corresponding Author: andrii.verbytskyi@cern.ch

The ZEUS data preservation (ZEUS DP) project assures continued access to the analysis software, experimental data and related documentation. The ZEUS DP project supports the possibility to derive valuable scientific results from the ZEUS data in the future. The implementation of the data preservation is discussed in the context of contemporary data analyses and of planning of future experiments and their corresponding data preservation efforts. The data are made available using standard protocols and authentication techniques. In addition to preserved analysis facilities a virtualization option is made available providing the ZEUS software and a validated environment bundled with a virtual machine.

Experiences with ALICE Mesos Infrastructure

Authors: Dario Berzano; Giulio Eulisse

Co-author: Kevin Napoli

1 CERN

2 University of Malta (MT)

Corresponding Authors: giulio.eulisse@cern.ch, kevin.napoli@cern.ch, dario.berzano@cern.ch
Apache Mesos is a resource management system for large data centres, initially developed by UC Berkeley, and now maintained under the Apache Foundation umbrella. It is widely used in the industry by companies like Apple, Twitter, and AirBnB and it’s known to scale to 10’000s of nodes. Together with other tools of its ecosystem, like Mesosphere Marathon or Chronos, it provides an end-to-end solution for datacenter operations and a unified way to exploit large distributed systems.

We present the experience of the ALICE Experiment Offline & Computing in deploying and using in production the Apache Mesos ecosystem for a variety of tasks on a small 500 cores cluster, using hybrid OpenStack and bare metal resources.

We will initially introduce the architecture of our setup and its operation, we will then describe the tasks which are performed by it, including release building and QA, release validation, and simple MonteCarlo production.

We will show how we developed Mesos enabled components (a.k.a. Mesos Frameworks) to carry out ALICE specific needs. In particular we will illustrate our effort to integrate Workqueue, a lightweight batch processing engine developed by University of Notre Dame, which ALICE uses to run release validation.

Finally we will give an outlook on how to use Mesos as resource manager for DDS, a software deployment system developed by GSI which will be the foundation of the system deployment for ALICE next generation Online-Offline (O2).

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Computing middleware

Primary Keyword (Mandatory):
Cloud technologies

Track 1: Online Computing / 267

Online computing architecture for the CBM experiment at FAIR

Author: Jan de Cuveland

Co-author: Volker Lindenstruth

1 Johann-Wolfgang-Goethe Univ. (DE)

Corresponding Authors: voli@compeng.de, jan.de.cuveland@cern.ch

The Compressed Baryonic Matter (CBM) experiment is currently under construction at the upcoming FAIR accelerator facility in Darmstadt, Germany. Searching for rare probes, the experiment requires complex online event selection criteria at a high event rate.

To achieve this, all event selection is performed in a large online processing farm of several hundred nodes, the "First-level Event Selector" (FLES). This compute farm will consist primarily of standard PC components including GPGPUs and many-core architectures. The data rate at the input to this compute farm is expected to exceed 1 TByte/s of time-stamped signal messages from the detectors. The distributed input interface will be realized using custom FPGA-based PCIe add-on cards, which preprocess and index the incoming data streams.

At event rates of up to 10 MHz, data from several events overlaps. Thus, there is no a priori assignment of data messages to events. Instead, event recognition is performed in combination with track
reconstruction. Employing a new container data format to decontextualize the information from specific time intervals, data segments can be distributed on the farm and processed independently. This allows to optimize the event reconstruction and analysis code without additional networking overhead and aids parallel computation in the online analysis task chain.

Time slice building, the continuous process of collecting the data of a time interval simultaneously from all detectors, places a high load on the network and requires careful scheduling and management. Using InfiniBand FDR hardware, this process has been demonstrated at rates of up to 6 GByte/s per node in a prototype system.

The design of the event selector system is optimized for modern computer architectures. This includes minimizing copy operations of data in memory, using DMA/RDMA wherever possible, reducing data interdependencies, and employing large memory buffers to limit the critical network transaction rate. A fault-tolerant control system will ensure high availability of the event selector.

This presentation will give an overview of the online event selection architecture of the upcoming CBM experiment and discuss the premises and benefits of the design. The presented material includes latest results from performance studies on different prototype systems.

Tertiary Keyword (Optional):
DAQ

Secondary Keyword (Optional):
High performance computing

Primary Keyword (Mandatory):
Trigger

---

**An Job Accounting tool for IHEP Computing**

**Author:** Jingyan Shi

**Corresponding Author:** shijy@ihep.ac.cn

The computing services running at computing center of IHEP support some HEP experiments and bio-medicine study. It provides 120,000 cpu cores including 3 local cluster and a Tier 2 grid site. A private cloud with 1000 cpu cores has been established to fit the experiment peak requirement. Besides, the computing center has several remote clusters as its distributing computing sub-site. Torque and HTCondor are two schedulers to manage clusters and there are more than 500,000 jobs running at computing cetner of IHEP each day.

We design and develop a Job Accounting tool to collect all the job information from clusters, clouds and remote sub-sites. The tool gives a fine-grained statistics for both users and system managers in time. Both jobs status and cpu cores utility can be accounted and showed in any time period. Since the amount of the jobs information grows fast day by day, MySQL cluster with optimization is chosen as the job database to provide quick query. As a serial of standard APIs are defined to collect job information, response job info query,it is easy to provide accounting service to new cluster. A web portal is developed as the user interface to accept on-line job query and to show the statistics in HTML5 graph.

Tertiary Keyword (Optional):
Building a Regional Computing Grid for the University of California at 100gbps

Author: Jeffrey Michael Dost

Co-authors: Alja Mrak Tadel 1; Edgar Fajardo Hernandez 1; Frank Wurthwein 2; Matevz Tadel 1; Terrence Martin 2; Tyson Jones 3

1 Univ. of California San Diego (US)  
2 UCSD  
3 Monash

Corresponding Authors: matevz.tadel@cern.ch, alja.mrak.tadel@cern.ch, tmartin@ucsd.edu, fkw@ucsd.edu, jdost@ucsd.edu, tjon14@student.monash.edu, edgar.mauricio.fajardo.hernandez@cern.ch

The Pacific Research Platform is an initiative to interconnect Science DMZs between campuses across the West Coast of the United States over a 100 gbps network. The LHC @ UC is a proof of concept pilot project that focuses on interconnecting 6 University of California campuses. It is spearheaded by computing specialists from the UCSD Tier 2 Center in collaboration with the San Diego Supercomputer Center. A machine has been shipped to each campus extending the concept of the Data Transfer Node to a "cluster in a box" that is fully integrated into the local compute, storage, and networking infrastructure. The node contains a full HTCondor batch system, and also an XRootD proxy cache. User jobs routed to the DTN can run on 40 additional slots provided by the machine, and can also flock to a common GlideinWMS pilot pool, which sends jobs out to any of the participating UCs, as well as to Comet, the new supercomputer at SDSC. In addition, a common XRootD federation has been created to interconnect the UCs and give the ability to arbitrarily export data from the home university, to make it available wherever the jobs run. The UC level federation also statically redirects to either the ATLAS FAX or CMS AAA federation respectively to make globally published datasets available, depending on end user VO membership credentials. XRootD read operations from the federation transfer through the nearest DTN proxy cache located at the site where the jobs run. This reduces wide area network overhead for subsequent accesses, and improves overall read performance. Details on the technical implementation, challenges faced and overcome in setting up the infrastructure, and an analysis of usage patterns and system scalability will be presented.

Tertiary Keyword (Optional):
Distributed data handling

Secondary Keyword (Optional):
Computing middleware

Primary Keyword (Mandatory):
Computing facilities
Extreme I/O on HPC for HEP using the Burst Buffer at NERSC

Author: Wahid Bhimji
Co-authors: Chris Daley; Deborah Bard; Jeff Porter; Kaylan Burleigh; Lisa Gerhardt; Markus Fasel; Peter Nugent; Steven Andrew Farrell; Vakho Tsulaia

1 Lawrence Berkeley National Lab. (US)
2 LBNL

Corresponding Authors: wahid.bhimji@cern.ch, penugent@lbl.gov, markus.fasel@cern.ch, steven.farrell@cern.ch, vakhtang.tsulaia@cern.ch, kburleigh@lbl.gov, rjporter@lbl.gov, djbard@lbl.gov, csdaley@lbl.gov, lgerhardt@lbl.gov

In recent years there has been increasing use of HPC facilities for HEP experiments. This has initially focussed on less I/O intensive workloads such as generator-level or detector simulation. We now demonstrate the efficient running of I/O-heavy ‘analysis’ workloads for the ATLAS and ALICE collaborations on HPC facilities at NERSC, as well as astronomical image analysis for DESI.

To do this we exploit a new 900 TB NVRAM-based storage system recently installed at NERSC, termed a ‘Burst Buffer’. This is a novel approach to HPC storage that builds on-demand filesystems on all-SSD hardware that is placed on the high-speed network of the new Cori supercomputer. The system provides over 900 GB/s bandwidth and 12.5 million I/O operations per second.

We describe the hardware and software involved in this system, and give an overview of its capabilities and use-cases beyond the HEP community before focussing in detail on how the ATLAS, ALICE and astronomical workflows were adapted to work on this system. To achieve this, we have also made use of other novel techniques, such as use of docker-like container technology, and tuning of the I/O layer experiment software.

We describe these modifications and the resulting performance results, including comparisons to other approaches and filesystems. We provide detailed performance studies and results, demonstrating that we can meet the challenging I/O requirements of HEP experiments and scale to tens of thousands of cores accessing a single storage system.

Tertiary Keyword (Optional):
Secondary Keyword (Optional):
Storage systems
Primary Keyword (Mandatory):
High performance computing

Review of Terabit/sec SDN demonstrations at Supercomputing 2015 and plans for SC16

Author: Azher Mughal
Co-author: Dorian Kcira

1 California Institute of Technology (US)

Corresponding Authors: azher.mughal@cern.ch, dorian.kcira@cern.ch
The HEP prototypical systems at the Supercomputing conferences each year have served to illustrate the ongoing state of the art developments in high throughput, software-defined networked systems important for future data operations at the LHC and for other data intensive programs. The Supercomputing 2015 SDN demonstration revolved around an OpenFlow ring connecting 7 different booths and the WAN connections. Some of the WAN connections were built using the Open Grid Forum’s Network Service Interface (NSI) and then stitched together using a custom SDN application developed at Caltech. This helped create an intelligent network design, where large scientific data flows traverse various paths provisioned dynamically with guaranteed bandwidth, with the path selection based on either the shortest or fastest routes available, or through other conditions. An interesting aspect of the demonstrations at SC15 is that all the local and remote network switches were controlled by a single SDN controller in the Caltech booth on the show floor. The SDN controller used at SC 15 was built on top of the OpenDaylight (Lithium) software framework. The software library was written in Python and has been made publicly available at pypi and github: pypi.python.org/pypi/python-odl/.

At SC 16 we plan to further improve and extend the SDN network design, we plan to enhance the SDN controller by introducing a number of higher level services, including the Application-Layer Traffic Optimization (ALTO) software and its path computation engine (PCE) in the OpenDaylight controller framework. In addition, we will use OpenvSwitch at the network edges and incorporate its rate-limiting features in the SDN data transfer plane. The CMS data transfer applications PhEDEx and ASO will be used as high level services to oversee large data transactions. The scale of storage to storage operations will be scaled up further relative to past demonstrations, working at the leading edge of NVMe storage and switching fabric technologies.

Secondary Keyword (Optional):
Network systems and solutions

Primary Keyword (Mandatory):
High performance computing

Tertiary Keyword (Optional):

Posters A / Break / 272

Data Transfer Nodes and Demonstration of 100G - 400G Wide Area Throughput Using the Caltech SDN Testbed

Author: Azher Mughal

Co-author: Harvey Newman

1 California Institute of Technology (US)

Corresponding Authors: azher.mughal@cern.ch, harvey.newman@cern.ch

We review and demonstrate the design of efficient data transfer nodes (DTNs), from the perspectives of the highest throughput over both local and wide area networks, as well as the highest performance per unit cost. A careful system-level design is required for the hardware, firmware, OS and software components. Furthermore, additional tuning of these components, and the identification and elimination of any remaining bottlenecks, is needed once the system is assembled and commissioned, in order to obtain optimal performance. For high throughput data transfers, specialized software is used to overcome the traditional limits in performance caused by the OS, file system, file structures used, etc. Concretely, we will discuss and present the latest results using Fast Data Transfer (FDT), developed by Caltech, and RFTP developed by Stonybrook together with BNL.

We will present and discuss the design choices for three generations of Caltech DTNs. Their transfer capabilities range from 40Gbps to 400Gbps. Disk throughput is still the biggest challenge in the current generation of available hardware. However, new NVMe drives combined with RDMA
and a new NVME network fabric are expected to improve the overall data-transfer throughput and simultaneously reduce the CPU load on the end nodes.

Secondary Keyword (Optional):
Network systems and solutions

Primary Keyword (Mandatory):
High performance computing

Tertiary Keyword (Optional):

Posters A / Break / 273

**Amplitude analysis of four-body decays using a massively-parallel fitting framework**

Authors: Christoph Hasse\(^1\), Michael David Sokoloff\(^2\)

Co-authors: Antonio Augusto Alves Junior \(^2\); Johannes Albrecht \(^1\); Jonas Rademacker \(^3\); Philippe D’Argent \(^4\); Timothy David Evans \(^5\)

\(^1\) Technische Universitaet Dortmund (DE)
\(^2\) University of Cincinnati (US)
\(^3\) University of Bristol (GB)
\(^4\) Ruprecht-Karls-Universitaet Heidelberg (DE)
\(^5\) University of Oxford (GB)

Corresponding Authors: jonas.rademacker@bristol.ac.uk, johannes.albrecht@cern.ch, p.dargent@cern.ch, antonio.augusto.alves.junior@cern.ch, timothy.david.evans@cern.ch, christoph.hasse@cern.ch, michael.david.sokoloff@cern.ch

GooFit, a GPU-friendly framework for doing maximum-likelihood fits, has been extended in functionality to do a full amplitude analysis of scalar mesons decaying into four final states via various combinations of intermediate resonances. Recurring resonances in different amplitudes are recognized and only calculated once, to save memory and execution time. As an example, this tool can be used to study the amplitude structure of the decay \(D^0 \rightarrow K^- \pi^+ \pi^+ \pi^-\) as well as a time-dependent amplitude analysis of \(D^0 \rightarrow K^+ \pi^+ \pi^- \pi^-\) to determine particle-antiparticle oscillation and CP violation parameters. GooFit uses the Thrust library to launch all kernels, with a CUDA back-end for nVidia GPUs and an OpenMP back-end for compute nodes with conventional CPUs. Performance of the algorithm will be compared between a variety of supported platforms.

Tertiary Keyword (Optional):
High performance computing

Secondary Keyword (Optional):
Parallelizarion

Primary Keyword (Mandatory):
Analysis tools and techniques

Posters A / Break / 275

**Enabling research network connectivity to clouds with a virtual router**
The use of opportunistic cloud resources by HEP experiments has significantly increased over the past few years. Clouds that are owned or managed by the HEP community are connected to the LHCONE network or the research network with global access to HEP computing resources. Private clouds, such as those supported by non-HEP research funds are generally connected to the international research network; however, commercial clouds are either not connected to the research network or only connect to research sites within their national boundaries. Since research network connectivity is a requirement for HEP applications, we need to find a solution that provides a high-speed connection. We are studying a solution with a virtual router that will address the use case when a commercial cloud has research network connectivity in a limited region. In this situation, we host a virtual router in our HEP site and require that all traffic from the commercial site transit through the virtual router. Although this may increase the network path and also the load on the HEP site, it is a workable solution that would enable the use of the remote cloud for low I/O applications. We are exploring some simple open-source solutions but expect that an SDN solution will be required to meet the bandwidth requirements. In this paper, we present the results of our studies and how it will benefit our use of private and public clouds for HEP computing.

Tertiary Keyword (Optional):
Virtualization

Secondary Keyword (Optional):
Cloud technologies

Primary Keyword (Mandatory):
Network systems and solutions

An OSG-based distributed campus computing infrastructure

We describe the development and deployment of a distributed campus computing infrastructure consisting of a single job submission portal linked to multiple local campus resources, as well the wider computational fabric of the Open Science Grid (OSG). Campus resources consist of existing OSG-enabled clusters and clusters with no previous interface to the OSG. Users accessing the single submission portal then seamlessly submit jobs to either resource type using the standard OSG toolkit of HTCondor for job submission and scheduling and the CERN Virtual Machine File System (CVMFS)
for software distribution. The usage of the Bosco job submission manager in HTCondor allows for submission to the campus HPC clusters without any access level beyond that of a regular user. The use of Condor flocking also allows user jobs to land at over a hundred clusters throughout the US that constitute the OSG Open Facility. We present the prototype of this facility, which enabled the Alpha Magnetic Spectrometer (AMS) experiment to utilize over 9 million computational hours in 6 weeks. We also present plans, including the usage of the new BoscoCE software stack to allow jobs submitted from outside the campus to land on any of the connected resources.

Secondary Keyword (Optional):
Distributed workload management

Primary Keyword (Mandatory):
Computing facilities

Tertiary Keyword (Optional):

Track 6: Infrastructures / 277

The OSG Open Facility: An on-ramp for opportunistic scientific computing

Authors: Bo Jayatilaka¹; Chander Sehgal²; Frank Wuerthwein³; Mats Rynge⁴; Robert William Gardner Jr⁵; Tanya Levshina²

¹ Fermi National Accelerator Lab. (US)
² Fermilab
³ Univ. of California San Diego (US)
⁴ RENCI UNC Chapel Hill
⁵ University of Chicago (US)

The Open Science Grid (OSG) is a large, robust computing grid that started primarily as a collection of sites associated with large HEP experiments such as ATLAS, CDF, CMS, and DZero, but has evolved in recent years to a much larger user and resource platform. In addition to meeting the US LHC community’s computational needs, the OSG continues to be one of the largest providers of distributed high-throughput computing (DHTC) to researchers from a wide variety of disciplines via the OSG Open Facility. The Open Facility consists of OSG resources that are available opportunistically to users other than resource owners and their collaborators. In the past two years, the Open Facility has doubled its annual throughput to over 200 million wall hours. More than half of these resources are used by over 100 individual researchers from over 60 institutions in fields such as biology, medicine, math, economics, and many others. Over 10% of these individual users utilized in excess of 1 million computational hours each in the past year. The largest source of these cycles is temporary unused capacity at institutions affiliated with US LHC computational sites. An increasing fraction, however, comes from university HPC clusters and large national infrastructure supercomputers offering unused capacity. Such expansions have allowed the OSG to provide ample computational resources to both individual researchers and small groups as well as sizeable international science collaborations such as LIGO, AMS, IceCube, and sPHENIX. Opening up access to the Fermilab FabrIc for Frontier Experiments (FIFE) project has also allowed experiments such as mu2e and NOvA to make substantial use of Open Facility resources, the former with over 40 million wall hours in a year. We present how this expansion was accomplished as well as future plans for keeping the OSG Open Facility at the forefront of enabling scientific research by way of DHTC.

Tertiary Keyword (Optional):
High performance computing

Primary Keyword (Mandatory):
Computing facilities

Track 2: Offline Computing / 278

A Software Toolkit to Study Systematic Uncertainties of the Physics Models of the Geant4 Simulation Package

Authors: Dennis Herbert Wright¹; Gabriel Perdue²; Hans-Joachim Wenzel²; Julia Yarba²; Krzysztof Genser²; Michael Kelsey³; Robert William Hatcher²

¹ SLAC National Accelerator Laboratory
² Fermi National Accelerator Laboratory (US)
³ SLAC National Accelerator Laboratory (US)

Corresponding Authors: julia.yarba@cern.ch, genser@fnal.gov, perdue@fnal.gov, wenzel@fnal.gov, rhatcher@fnal.gov, dennis.herbert.wright@cern.ch, kelsey@slac.stanford.edu

Particle physics experiments make heavy use of the Geant4 simulation package to model interactions between subatomic particles and bulk matter. Geant4 itself employs a set of carefully validated physics models that span a wide range of interaction energies. They rely on measured cross-sections and phenomenological models with the physically motivated parameters that are tuned to cover many application domains. The aggregated sum of these components is what experiments use to study their apparatus. This raises a critical question of what uncertainties are associated with a particular tune of one or another Geant4 physics model, or a group of models, involved in modeling and optimization of a detector design.

In response to multiple requests from the simulation community, the Geant4 Collaboration has started an effort to address the challenge.

We have designed and implemented a comprehensive, modular, user-friendly software toolkit that allows modifications of parameters of one or several Geant4 physics models involved in the simulation studies, and to perform collective analysis of multiple variants of the resulting physics observables of interest, in order to estimate an uncertainty on a measurement due to the simulation model choices.

Based on modern event-processing infrastructure software, the toolkit offers a variety of attractive features, e.g. flexible run-time configurable workflow, comprehensive bookkeeping, easy to expand collection of analytical components.

Design, implementation technology, and key functionalities of the toolkit will be presented and highlighted with selected results.

Keywords: Geant4 model parameters perturbation, systematic uncertainty in detector simulation

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):
Simulation

Posters B / Break / 280
Integration of Openstack Cloud Resources in BES III Computing Cluster

Author: Haibo Li

Co-authors: Jingyan Shi; Yaodong Cheng; Qiulan Huang

1 Institute of High Energy Physics Chinese Academy of Science
2 IHEP
3 Institute of High Energy Physics, Beijing

Corresponding Authors: lihaibo@ihep.ac.cn, chyd@ihep.ac.cn, shijy@ihep.ac.cn, huangql@ihep.ac.cn

The development and new discoveries of a new generation of high-energy physics cannot be separated from the mass data processing and analysis. The BESIII experiments studies physics in the tau-charm energy region from 2GeV to 4.6 GeV, at the Institute of High Energy Physics (IHEP) in Beijing, China, which is a typical data-intensive computing requiring mass storage and efficient computing resources. With the rapid growth of experimental data, the data processing system encounters many problems such as low resource utilization, migration complex and so on, which makes it urgent to enhance the data analysis system ability. Cloud computing which uses virtualization technology provides many advantages to solve these problems in a cost-effective way.

However, offline software design, resource allocation and job scheduling of BESIII experiment are all based on physical machine. To make use of the cloud computing resources such as Opensatck, the integration of Openstack and existing computing cluster is a key issue. In this contribution we present an on-going work that aims to integrate openstack cloud resources into BESIII computing cluster for providing distributed compute clouds for the BES III physics experiment and supplying seamless combination. In particular, we discuss our design of the cloud scheduler which used to integrate OpenStack cloud resources into existing TORQUE and Maui cluster. In aspect of job scheduling, we adopt the thought of pull mode as described follows: when torque asking for launching new VMs in the cloud, the job agent residing in the vm will pull the suitable job to local machine and the job will be sent back after completed. It is transparent to users and they can continue to submit jobs using qsub command without knowing anything about cloud. Lastly, we report on our development work of adaptive job scheduling strategy to improve resource utilization and job processing efficiency.

Tertiary Keyword (Optional):
Secondary Keyword (Optional): Computing models
Primary Keyword (Mandatory): Virtualization

Track 4: Data Handling / 281

dCache on steroids - delegated storage solutions

Authors: Gerd Behrmann; Patrick Fuhrmann; Paul Millar; Tigran Mkrtchyan

1 NDGF
2 DESY

Corresponding Authors: behrmann@ndgf.org, paul.millar@desy.de, tigran.mkrtchyan@desy.de, patrick.fuhrmann@desy.de

For over a decade, dCache.ORG has provided robust software that is used at more than 80 Universities and research institutes around the world, allowing these sites to provide reliable storage services for
the WLCG experiments and many other scientific communities. The flexible architecture of dCache allows running it in a wide variety of configurations and platforms - from all-in-one Raspberry-Pi up to hundreds of nodes in multi-petabyte infrastructures.

Due to lack of managed storage at the time, dCache implemented data placement, replication and data integrity directly. Today, many alternatives are available: S3, GlusterFS, CEPH and others. While such systems position themselves as scalable storage systems, they can not be used by many scientific communities out of the box. The absence of specific authentication and authorization mechanisms, the use of product specific protocols and the lack of namespace are some of reasons that prevent wide-scale adoption of these alternatives.

Most of these limitations are already solved by dCache. By delegating low level storage management functionality to the above mentioned new systems and providing the missing layer through dCache, we provide a system which combines the benefits of both worlds - industry standard storage building blocks with the access protocols and authentication required by scientific communities.

In this presentation, we focus on CEPH, a popular software for clustered storage that supports file, block and object interfaces. CEPH is often used in modern computing centres, for example as a backend to OpenStack services. We will show prototypes of dCache running with a CEPH backend and discuss the benefits and limitations of such an approach. We will also outline the roadmap for supporting ‘delegated storage’ within the dCache releases.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Object stores

Primary Keyword (Mandatory):

Storage systems

Track 6: Infrastructures / 282

AMS-02 Monte Carlo Simulation in Science Operation Center at Southeast University

Authors: Aibo Song\textsuperscript{1}; Alexander Egorov\textsuperscript{2}; Alexandre Eline\textsuperscript{2}; Fang Dong\textsuperscript{1}; Feiqiao Huang\textsuperscript{3}; Jinghui Zhang\textsuperscript{3}; Jiyuan Shi\textsuperscript{1}; Junzhou Luo\textsuperscript{3}; Renli Shi\textsuperscript{3}; Runqun Xiong\textsuperscript{3}; Vitaly Choutko\textsuperscript{2}

\textsuperscript{1} Southeast University (CN)
\textsuperscript{2} Massachusetts Inst. of Technology (US)
\textsuperscript{3} Southeast University (CN)

Corresponding Authors: feiqiao.huang@cern.ch, vitaly.choutko@cern.ch, rxiong@seu.edu.cn, absong@seu.edu.cn, fdong@seu.edu.cn, renli.shi@cern.ch, jiyuanshi@seu.edu.cn, alexander.egorov@cern.ch, alexandre.eline@cern.ch, zhang_jinghui@seu.edu.cn, jluo@seu.edu.cn

Abstract: Southeast University Science Operation Center (SEUSOC) is one of the computing centers of the Alpha Magnetic Spectrometer (AMS-02) experiment. It provides 2000 CPU cores for AMS scientific computing and a dedicated 1Gbps Long Fat Network (LFN) for AMS data transmission between SEU and CERN. In this paper, the workflows of SEUSOC Monte Carlo (MC) production are discussed in detail, including the process of the MC job request and execution, the data transmission strategy, the MC database and the MC production monitoring tool. Moreover, to speed up the data transmission in LFN between SEU and CERN, an optimized transmission strategy in TCP layer and application layer is further introduced.

Tertiary Keyword (Optional):

Simulation
Particle Physics Analysis with the CERN Open Data Portal

Authors: Aleksei Rogozhnikov\(^1\); Andrey Ustyuzhanin\(^1\); Chris Parkes\(^2\); Denis Derkach\(^1\); Malin Litwinski\(^2\); Marco Gersabeck\(^2\); Silvia Amerio\(^3\); Sunje Dallmeier-Tiessen\(^4\); Tim Head\(^5\); george gilliver\(^6\)

\(^1\) Yandex School of Data Analysis (RU)
\(^2\) University of Manchester (GB)
\(^3\) Universita e INFN, Padova (IT)
\(^4\) CERN
\(^5\) Ecole Polytechnique Federale de Lausanne (CH)
\(^6\) The University of Manchester (GB)

Corresponding Authors: sunje.dallmeier-tiessen@cern.ch, chris.parkes@cern.ch, marco.gersabeck@cern.ch, silvia.amerio@cern.ch, denis.derkach@cern.ch, alex.rogozhnikov@cern.ch, malin.litwinski@student.manchester.ac.uk, george_gilliver@hotmail.co.uk, tim.head@cern.ch, andrey.ustyuzhanin@cern.ch

A framework for performing a simplified particle physics data analysis has been created. The project analyses a pre-selected sample from the full 2011 LHCb data. The analysis aims to measure matter antimatter asymmetries. It broadly follows the steps in a significant LHCb publication where large CP violation effects are observed in charged B meson three-body decays to charged pions and kaons. The project is a first-of-its-kind analysis on the CERN open portal as its students are guided through elements of a full particle physics analysis but use a simplified interface. The analysis has multiple stages culminating in the observation of matter anti-matter differences between Dalitz plots of the B\(^+\) and B\(^-\) meson decay. The project uses the open source Jupyter Notebook project, the Docker open platform for distributed applications, and can be hosted through the open source Everware platform. The target audience includes advanced high school students, undergraduate societies and enthusiastic scientifically literate members of the general public. The public use of this data set has been approved by the LHCb collaboration. The project plans to launch for the public in summer 2016 through the CERN Open Data Portal. The project development has been supported by final year undergraduates at the University of Manchester, Yandex school of data analysis and the CERN Open Data team.

Posters A / Break / 284
Design and Implementation of Monte Carlo Production Monitoring Tool for AMS Experiment

Authors: Alexander Egorov¹; Alexandre Eline¹; Baosong Shan²; Fang Dong³; Feiqiao Huang³; Jinghui Zhang³; Junzhou Luo³; Oleg Demakov¹; Renli Shi³; Runqun Xiong³; Vitaly Choutko¹

¹ Massachusetts Inst. of Technology (US)
² Beihang University (CN)
³ Southeast University (CN)

Corresponding Authors: alexandre.eline@cern.ch, feiqiao.huang@cern.ch, vitaly.choutko@cern.ch, oleg.demakov@cern.ch, zhang_jinghui@seu.edu.cn, baosong.shan@cern.ch, alexander.egorov@cern.ch, fdong@seu.edu.cn, renli.shi@cern.ch, rxiong@seu.edu.cn, jluo@seu.edu.cn

Abstract: Monte Carlo (MC) simulation production plays an important part in physics analysis of the Alpha Magnetic Spectrometer (AMS-02) experiment. To facilitate the metadata retrieving for data analysis needs among the millions of database records, we developed a monitoring tool to analyze and visualize the production status and progress. In this paper, we discuss the workflow of the monitoring tool and present its features and technical details.

Tertiary Keyword (Optional): Databases

Secondary Keyword (Optional): Monitoring

Primary Keyword (Mandatory): Software development process and tools

Posters A / Break / 285

Computational Steering of GEM Simulations

Author: Ali Sheharyar¹

Co-author: Othmane Bouhali ¹

¹ Texas A & M University (US)

Corresponding Authors: ali.sheharyar@gmail.com, othmane.bouhali@qatar.tamu.edu

One of the primary objectives of the research on GEMs at CERN is the testing and simulation of prototypes, manufacturing of large-scale GEM detectors and installation into CMS detector sections at the outer layer, where only highly energetic muons particles are detected. When a muon particle traverses a GEM detector, it ionizes the gas molecules generating a freely moving electron that starts ionizing the gas molecules and produces the secondary electrons. These secondary electrons also ionize the gas and subsequently form an avalanche of electrons under the influence of the applied drift field. The simulations of this physical phenomenon especially those with complex scenarios such as those involving high detector voltages or gas with larger gains are computationally intensive and may take several days or even weeks to complete.

These long-running simulations usually run on the high-performance supercomputers in batch mode. If the results lead to unexpected behavior, then the simulation might be rerun with different parameters. However, the simulations (or jobs) have to wait in a queue until they get a chance to run again because the supercomputer is a shared resource that maintains a queue of all other users programs as well and executes them as time and priorities permit. It results in inefficient utilization of computing resource and increases the turnaround time for the scientific experiment.
To overcome this issue, the monitoring of the behavior of a simulation, while it is running (or live), is essential. One method of monitoring is to write the data, produced by the simulation, periodically to the disk. But, the disk being inherently slow can become a bottleneck and can affect the performance. Another approach is to use the computational steering technique, in which simulation is coupled with a visualization system to enable the exploration of “live” data as it is produced by the simulation.

In this work, we employ the computational steering method by coupling the GEM simulations with a visualization package named VisIt. A user can connect to the running simulation with the VisIt client over the network, and can visualize the “live” data to monitor the simulation behavior. Also, the simulation can be restarted immediately on the fly with different parameters without requiring resubmitting the job on the supercomputer.

Secondary Keyword (Optional):
High performance computing

Primary Keyword (Mandatory):
Visualization

Tertiary Keyword (Optional):

Track 2: Offline Computing / 286

STAR Reconstruction Improvements for Tracking with the Heavy Flavor Tracker

Authors: Dmitri Smirnov\textsuperscript{1}; Gene Van Buren\textsuperscript{2}; Jason Webb\textsuperscript{3}; Jerome LAURET\textsuperscript{2}; Victor Perevoztchikov\textsuperscript{2}

\textsuperscript{1} BNL
\textsuperscript{2} Brookhaven National Laboratory
\textsuperscript{3} Brookhaven National Lab

Corresponding Authors: perev@bnl.gov, jwebb@bnl.gov, dsmirnov@bnl.gov, jlauret@bnl.gov, gene@bnl.gov

The reconstruction and identification of charmed hadron decays provides an important tool for the study of heavy quark behavior in the Quark Gluon Plasma. Such measurements require high resolution to topologically identify decay daughters at vertices displaced <100 microns from the primary collision vertex, placing stringent demands on track reconstruction software. To enable these measurements at RHIC, the STAR experiment has designed and employed the Heavy Flavor Tracker (HFT). It is composed of silicon-based tracking detectors, providing four layers of high-precision position measurements which are used in combination with hits from the Time Projection Chamber (TPC) to reconstruct track candidates.

The STAR integrated tracking software (Sti) has delivered a decade of world-class physics. It was designed to leverage the discrete azimuthal symmetry of the detector and its simple radial ordering of components, permitting a flat representation of the detector geometry in terms of concentric cylinders and planes, and an approximate track propagation code. These design choices reflected a careful balancing of competing priorities, trading precision for speed in track reconstruction.

To simplify the task of integrating new detectors, tools were developed to automatically generate the Sti geometry model, tying both reconstruction and simulation to the single source AgML geometry model. The increased precision and complexity of the HFT detector required a careful reassessment of this single geometry path and implementation choices. In this paper we will discuss the tools developed to optimize track reconstruction with the HFT, our lesson learn with tracking with high precision detectors and the tradeoffs between precision, speed and ease of use which were required.

Tertiary Keyword (Optional):
Representing Misalignments of the STAR Geometry Model using AgML

Authors: Dmitri Smirnov\(^1\); Gene Van Buren\(^2\); Jason Webb\(^3\); Jerome LAURET\(^2\); Victor Perevoztchikov\(^2\)

\(^1\)BNL
\(^2\)Brookhaven National Laboratory
\(^3\)Brookhaven National Lab

Corresponding Authors: jwebb@bnl.gov, perev@bnl.gov, dsmirnov@bnl.gov, gene@bnl.gov, jlauret@bnl.gov

The STAR Heavy Flavor Tracker (HFT) was designed to provide high-precision tracking for the identification of charmed hadron decays in heavy ion collisions at RHIC. It consists of three independently mounted subsystems, providing four precision measurements along the track trajectory, with the goal of pointing decay daughters back to vertices displaced by <100 microns from the primary event vertex. The ultimate efficiency and resolution of the physics analysis will be driven by the quality of the simulation and reconstruction of events in heavy ion collisions. In particular, it is important that the geometry model properly accounts for the relative misalignments of the HFT subsystems, along with the alignment of the HFT relative to STAR’s primary tracking detector, the Time Projection Chamber (TPC).

The Abstract Geometry Modeling Language (AgML) provides the single description of the STAR geometry, generating both our simulation (GEANT 3) and reconstruction geometries (ROOT). AgML implements an ideal detector model. Misalignments are stored separately in database tables, and have historically been applied at the hit level – detector hits, whether simulated or real, are moved from their ideal position described in AgML to their misaligned position according to the database. This scheme has worked well as hit errors have been negligible compared with the size of sensitive volumes. The precision and complexity of the HFT detector require us to apply misalignments to the detector volumes themselves. In this paper we summarize the extension of the AgML language and support libraries to enable the static misalignment of our reconstruction and simulation geometries, discussing the design goals, limitations and path to full misalignment support in ROOT/VMC based simulation.
Design and Implementation of Elastic Computing Resource Management System with HTCondor on Openstack

Author: zhen jing cheng

Co-authors: Gang CHEN; Yaodong Cheng; qiulan huang

1 IHEP
2 INSTITUTE OF HIGH ENERGY PHYSICS
3 Institute of High Energy Physics, Beijing

Corresponding Authors: chyd@ihep.ac.cn, gang.chen@ihep.ac.cn, chengzj@ihep.ac.cn, huangql@ihep.ac.cn

As a new approach to manage resource, virtualization technology is more and more widely applied in high-energy physics field. A virtual computing cluster based on Openstack was built at IHEP, and with HTCondor as the job queue management system. An accounting system which can record the resource usages of different experiment groups in details was also developed. There are two types of the virtual computing cluster, static and dynamic. In traditional static cluster, fixed number of virtual machines are pre-allocated to the job queue of different experiments. But it cannot meet peak requirements of different experiments gradually. To solve this problem, we designed and implemented an elastic computing resource management with HTCondor on Openstack. This system performs unified management of virtual computing nodes on the basis of job queue in HTCondor. It is consisted of four loosely-coupled components, including job status monitoring, computing node management, load balance system and the daemon. Job status monitoring system communicates with HTCondor to get the current status of each job queue and each computing node of one specific experiment. Computing node management component communicates with Openstack to launch or destroy virtual machines. After a VM is created, it will be added to the resource pool of corresponding experiment group. Then the job will run at the virtual machine. After the job finishes, the virtual machine will be shutdown. When the VM shutdown in Openstack, it will be removed from the resource pool. Meanwhile, the computing node management system provides an interface to query virtual resources usage. Load balance system provides an interface to get the information of available virtual resources for each experiment. The daemon component asks load balance system to decide how much available virtual resources. It also communicates with job status monitoring system to get the number of queued jobs. Finally, it calls computing node management system to launch or destroy a few of virtual computing nodes.

This paper will present several use cases of LHAASO and JUNO experiments. The results show virtual computing resource dynamic expanded or shrunk while computing requirements change. Additionally, CPU utilization ratio of computing resource is significantly increased compared with traditional resource management. The system also has good performance when there are multiple condor schedulers and multiple job queues. And it is stable and easy to maintain as well.

Secondary Keyword (Optional):
Virtualization

Primary Keyword (Mandatory):
Cloud technologies

Tertiary Keyword (Optional):

Track 4: Data Handling / 289

OSiRIS: A Distributed Ceph Deployment Using Software Defined Networking for Multi-Institutional Research

Authors: Martin Swany; Shawn Mc Kee

Co-authors: Benjeman Jay Meekhof; Ezra Kissel

1 Helsinki Institute of Physics (FI)
We will report on the first year of the OSiRIS project (NSF Award #1541335, UM, IU, MSU and WSU) which is targeting the creation of a distributed Ceph storage infrastructure coupled together with software-defined networking to provide high-performance access for well-connected locations on any participating campus. The project’s goal is to provide a single scalable, distributed storage infrastructure that allows researchers at each campus to read, write, manage and share data directly from their own computing locations. The NSF CC*DNI DIBBs program which funded OSiRIS is seeking solutions to the challenges of multi-institutional collaborations involving large amounts of data and we are exploring the creative use of Ceph and networking to address those challenges.

While OSiRIS will eventually be serving a broad range of science domains, its first adopter will be ATLAS, via the ATLAS Great Lakes Tier-2 (AGLT2), jointly located at the University of Michigan and Michigan State University. Part of our presentation will cover how ATLAS is using the OSiRIS infrastructure and our experiences integrating our first user community. The presentation will also review the motivations for and goals of the project, cover the technical details of the OSiRIS infrastructure, the challenges in providing such an infrastructure, and the technical choices made to address those challenges. We will conclude with our plans for the remaining 4 years of the project and our vision for what we hope to deliver by the project’s end.

Tertiary Keyword (Optional):
Network systems and solutions

Secondary Keyword (Optional):
Object stores

Primary Keyword (Mandatory):
Storage systems

A dashboard for the Italian Computing in ALICE

Author: Domenico Elia

Co-authors: Antonio Franco, Giacinto Donvito, Gioacchino Vino, Stefano Bagnasco, Stefano Piano, Vincenzo Spinoso

1 INFN Bari
2 Universita e INFN, Bari (IT)
3 INFN-Bari
4 I.N.F.N. TORINO
5 INFN (IT)
6 Universita e INFN (IT)

Corresponding Authors: bagnasco@to.infn.it, antonio.franco@cern.ch, stefano.piano@cern.ch, vincenzo.spinoso@ba.infn.it, giacinto.donvito@ba.infn.it, domenico.elia@cern.ch, gioacchino.vino@cern.ch

The ALICE experiment at CERN was designed to study the properties of the strongly-interacting hot and dense matter created in heavy-ion collisions at the LHC energies. The computing model of the experiment currently relies on the hierarchical Tier-based structure, with a top-level Grid site at CERN (Tier-0, also extended to Wigner) and several globally distributed datacenters at national and regional level (Tier-1 and Tier-2 sites). The Italian computing infrastructure is mainly composed by
a Tier-1 site at CNAF (Bologna) and four Tier-2 sites (at Bari, Catania, Padova-Legnaro and Torino), with the addition of two small WLCG centers in Cagliari and Trieste. Globally it contributes by about 15% to the overall ALICE computing resources.

Actually the management of a Tier-2 site is based on a few complementary monitoring tools, each looking at the ALICE activity from a different point of view: for instance, MonALISA is used to extract information from the experiment side, the Local Batch System allows to store statistical data on the overall site activity and the Local Monitoring System provides the status of the computing machines. This typical schema makes somewhat difficult to figure out at a glance the status of the ALICE activity in the site and to compare information extracted from different sources for debugging purposes.

In this contribution, a monitoring system able to gather information from all the available sources to improve the management of an ALICE Tier-2 site will be presented. A centralized site dashboard based on specific tools selected to meet tight technical requirements, like the capability to manage a huge amount of data in a fast way and through an interactive and customizable Graphical User Interface, has been developed. The current version, running in the Bari Tier-2 site since more than one year, relies on an open source time-series database (InfluxDB): a dataset of about 20 M values is currently stored in 400 MB with on-the-fly aggregation allowing to return downsampled series with a factor of 10 gain in the retrieval time. A dashboard builder for visualizing time-series metrics (Grafana) has been identified as best suited option, while dedicated code has been written to implement the gathering phase. Details of the dashboard performance as observed along the last year will be also provided.

The system is currently being exported to all the other sites in order to allow a next step where a unique centralized dashboard for the ALICE computing in Italy will be implemented. Prospects of such an Italian dashboard and further developments on this side will be discussed. They also include the design of a more general monitoring system for distributed datacenters able to provide active support to site administrators in detecting critical events as well as in improving problem solving and debugging procedures.

Tertiary Keyword (Optional):
Cloud technologies

Secondary Keyword (Optional):
Accounting and information

Primary Keyword (Mandatory):
Monitoring

Track 6: Infrastructures / 291

Scaling the PuNDIT Project for Wide Area Deployments

Authors: Constantine Dovrolis\(^1\); Shawn Mc Kee\(^2\)

Co-authors: Danny Lee \(^1\); Gabriele Carcassi \(^3\); Jorge Batista \(^4\)

\(^1\) Georgia Tech
\(^2\) University of Michigan (US)
\(^3\) University of Michigan
\(^4\) U

Corresponding Authors: dannylee@gatech.edu, carcassi@umich.edu, batistaj@umich.edu, dovrolis@cc.gatech.edu, shawn.mckee@cern.ch

In today’s world of distributed scientific collaborations, there are many challenges to providing reliable inter-domain network infrastructure. Network operators use a combination of active monitoring and trouble tickets to detect problems, but these are often ineffective at identifying issues that impact wide-area network users. Additionally, these approaches do not scale to wide area inter-domain networks due to unavailability of data from all the domains along typical network
paths. The Pythia Network Diagnostic InfrasTructure (PuNDIT) project aims to create a scalable infrastructure for automating the detection and localization of problems across these networks.

The project goal is to gather and analyze metrics from existing perfSONAR monitoring infrastructures to identify the signatures of possible problems, locate affected network links, and report them to the user in an intuitive fashion. Simply put, PuNDIT seeks to convert complex network metrics into easily understood diagnoses in an automated manner.

At CHEP 2016, we plan to present our findings from deploying a first version of PuNDIT in one or more communities that are already using perfSONAR. We will report on the project progress to-date in working with the OSG and various WLCG communities, describe the current implementation architecture and demonstrate the various user interfaces it supports. We will also show examples of how PuNDIT is being used and where we see the project going in the future.

Tertiary Keyword (Optional):
Monitoring

Secondary Keyword (Optional):
Computing facilities

Primary Keyword (Mandatory):
Network systems and solutions

---

**The Simulation Library of the Belle II Experiment**

**Author:** Doris Kim

**Co-author:** Martin Ritter

1 Soongsil University

2 LMU Munich

**Corresponding Authors:** dorisykim@yahoo.com, ritter@mpp.mpg.de

SuperKEKB, a next generation B factory, has finished being constructed in Japan as an upgrade of the KEKB e+e- collider. Currently it is running with the BEAST II detector, whose purpose is to understand the interaction and background events at the beam collision region in preparation for the 2018 launch of the Belle II detector. Overall SuperKEKB is expected to deliver a rich data set for the Belle II experiment, which will be 50 times larger than the previous Belle sample. Both the triggered physics event rate and the background event rate will be at least 10 times that of the previous experiment, which creates a challenging data taking environment. The software system of the Belle II experiment has been designed to execute this demanding task. A full detector simulation library, which is a part of the Belle II software system, has been created based on Geant4 and tested thoroughly. Recently the library was updated to Geant4 version 10.1. The library is behaving as expected and is utilized actively in producing Monte Carlo data sets for diverse physics and background situations. In this talk we explain the detailed structure of the simulation library and its interfaces to other packages such as generators, geometry, and background event simulation.

**Primary Keyword (Mandatory):**
Simulation

**Secondary Keyword (Optional):**

**Tertiary Keyword (Optional):**
Track 6: Infrastructures / 293

Using Shifter to Bring Containerized CVMFS to HPC

Author: Lisa Gerhardt

Co-authors: Douglas Jacobsen 2; Jeff Porter 3; Shane Canon 4; Vakho Tsulaia 3; Wahid Bhimji 3

1 LBNL
2 Lawrence Berkeley National Laboratory
3 Lawrence Berkeley National Lab. (US)
4 NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Corresponding Authors: wahid.bhimji@cern.ch, rjporter@lbl.gov, canon@nersc.gov, vakhtang.tsulaia@cern.ch, lgerhardt@lbl.gov, dmj@lbl.gov

Bringing HEP computing to HPC can be difficult. Software stacks are often very complicated with numerous dependencies that are difficult to get installed on an HPC system. To address this issue, amongst others, NERSC has created Shifter, a framework that delivers Docker-like functionality to HPC. It works by extracting images from native formats (such as a Docker image) and converting them to a common format that is optimally tuned for the HPC environment. We have used Shifter to deliver the CVMFS software stack for ALICE, ATLAS, and CMS on the Edison and Cori supercomputers at NERSC. As well as enabling the distribution of TBs of software to HPC, this approach also offers performance advantages. We show that software startup times are significantly reduced (up a factor of 4 relative to the Lustre file system in the ATLAS case) as well as scaling with minimal variation to 1000s of nodes. We will discuss how this was accomplished as well as future outlook for Shifter and HEP at NERSC.

Tertiary Keyword (Optional): Virtualization

Secondary Keyword (Optional): High performance computing

Primary Keyword (Mandatory): HPC

Posters B / Break / 294

Tracks pattern recognition for the SHiP Spectrometer Tracker

Author: Mikhail Hushchyn

Co-authors: Andrey Ustyuzhanin 1; Eric Van Herwijnen 2; Oleg Alenkin 3

1 Yandex School of Data Analysis (RU)
2 CERN
3 Moscow Institute of Physics and Technology (RU)

Corresponding Authors: eric.van.herwijnen@cern.ch, mikhail.hushchyn@cern.ch, alenkin.oleg@phystech.edu, andrey.ustyuhanin@cern.ch

The SHiP is a new fixed-target experiment at the CERN SPS accelerator. The goal of the experiment is searching for hidden particles predicted by the models of Hidden Sectors. The purposes of the SHiP Spectrometer Tracker is to reconstruct the tracks of charged particles from the decay of neutral New Physics objects with high efficiency, while rejecting background events. The problem is to develop a
method of the tracks pattern recognition based on the SHiP Spectrometer Tracker design. Baseline algorithms gives efficiency of 95%.

In this study we compare different tracks pattern recognition methods adapted to the experiment geometry. We present how widely-used algorithms such as RANSAC regression and Hough transformation can be effectively used in the experiment. We compare their performances and show their advantages and limitations.

In addition, in this study we develop new linear regression algorithm which effectively uses the SHiP Spectrometer Tracker strawtubes structure and geometry. We describe properties of the regression and how it helps to reach the high tracks efficiency.

Moreover, we demonstrate how the tracks pattern recognition can be solved in term of classification problem. We show how classifiers helps to search tracks with high efficiency. We compare this approach with others and describe its advantages.

All the methods presented demonstrate track efficiency statistically better than baseline algorithm.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Algorithms

Primary Keyword (Mandatory):
Reconstruction

Track 4: Data Handling / 295

GRID Storage Optimization in Transparent and User-Friendly Way for LHCb datasets

Author: Mikhail Hushchyn
Co-authors: Andrey Ustyuzhanin 1; Christophe Haen 2; Philippe Charpentier 2

1 Yandex School of Data Analysis (RU)
2 CERN

Corresponding Authors: christophe.denis.haen@cern.ch, andrey.ustyuzhanin@cern.ch, mikhail.hushchyn@cern.ch, philippe.charpentier@cern.ch

The LHCb collaboration is one of the four major experiments at the Large Hadron Collider at CERN. Petabytes of data are generated by the detectors and Monte-Carlo simulations. The LHCb Grid interware LHCBDIRAC is used to make data available to all collaboration members around the world. The data is replicated to the Grid sites in different locations. However, disk storage on the Grid is limited and does not allow to keep replicas of each file at all sites. Thus, it is essential to determine the optimal number of replicas in order to achieve a good Grid performance.

In this study, we present an approach of data replication and distribution strategy based on data popularity prediction different from that previous presented at CHEP2015[1]. Each file can be described by the following features: age, reuse time interval, access frequency, type, size and some others parameters. Based on these features and access history, the probability that the file will be accessed in the long-term future can be predicted using machine learning algorithms. In addition, time series analysis and access history can be used to forecast the number of accesses to the file in the short-term period in the future. We describe a metrics that combines these predictions. This metrics helps to determine for which files the number of replicas can be increased or decreased depending on how much disk space is available or how much space needs to be freed. Moreover, the metrics indicates...
when all replicas of the file can be removed from disk storage. The proposed approach is being tested in LHCb production. In the study, we show the results of the simulation studies and results of the tests in the production. We demonstrate that the method outperforms our previous study, while it requires a minimal number of parameters and gives more easily interpretable predictions.

Reference:


Posters B / Break / 296

The LHCb Grid Simulation

Author: Andrey Ustyuzhanin

Co-authors: Alexander Baranov 1; Kenenbek Arzymatov 2; Mikhail Hushchyn 1

1 Yandex School of Data Analysis (RU)
2 Moscow Institute of Physics and Technology (RU)

Corresponding Authors: kenenbek@gmail.com, a.baranov@cern.ch, andrey.ustyuzhanin@cern.ch, mikhail.hushchyn@cern.ch

The LHCb Grid access if based on the LHCbDirac system. It provides access to data and computational resources to researchers with different geographical locations. The Grid has a hierarchical topology with multiple sites distributed over the world. The sites differ from each other by their number of CPUs, amount of disk storage and connection bandwidth. These parameters are essential for the Grid work. Moreover, job scheduling and data distribution strategy have a great impact on the grid performance. However, it is hard to choose an appropriate algorithm and strategies as they need a lot of time to be tested on the real grid.

In this study, we describe the LHCb Grid simulator. The simulator reproduces the LHCb Grid structure with its sites and their number of CPUs, amount of disk storage and bandwidth connection. We demonstrate how well the simulator reproduces job scheduling and network anomalies, consider methods for their detection and resolution. In addition, we compare different algorithms for job scheduling and different data distribution strategies.

Primary Keyword (Mandatory):
Simulation

Secondary Keyword (Optional):

Tertiary Keyword (Optional):
Pythonization API for Cppy

Authors: Toby St Clere Smithe\(^1\); Wim Lavrijsen\(^2\)

\(^1\) Warwick University
\(^2\) Lawrence Berkeley National Lab. (US)

Corresponding Authors: wim.lavrijsen@cern.ch, mail@tsmithe.net

Cppyy provides fully automatic Python/C++ language bindings and so doing covers a vast number of use cases. Use of conventions and known common patterns in C++ (such as smart pointers, STL iterators, etc.) allow us to make these C++ constructs more "pythonistic." We call these treatments "pythonizations", as the strictly bound C++ code is turned into bound code that has a Python "feel." However, there are always a few corner cases that can be improved with manual intervention. Historically, this was done with helpers or wrapper code on the C++ or Python side.

In this paper, we present the new pythonization API that standardizes these manual tasks, covering the common use cases and in so doing improving scalability and interoperability. This API has been provided for both CPython and PyPy. We describe the fundamental abstractions that it covers, how it can be used to resolve conflicts across packages, and its performance.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):
Analysis tools and techniques

Scalable Global Grid catalogue for LHC Run3 and beyond

Author: Miguel Martinez Pedreira\(^1\)

Co-author: Costin Grigoras \(^1\)

\(^1\) CERN

Corresponding Authors: miguel.martinez.pedreira@cern.ch, costin.grigoras@cern.ch

The AliEn file catalogue is a global unique namespace providing mapping between a UNIX-like logical name structure and the corresponding physical files distributed over 80 storage elements worldwide. Powerful search tools and hierarchical metadata information are integral part of the system and are used by the Grid jobs as well as local users to store and access all files on the Grid storage elements.

The catalogue is in production since 2005 and over the past 11 years has grown to more than 2 billion logical file names. The back-end is a set of distributed relational databases, ensuring smooth growth and fast access. Due to the anticipated fast future growth, we are looking for ways to enhance the performance and scalability by simplifying the catalogue schema while keeping the functionality intact. We investigated different back-end solutions, such as distributed key value stores, as replacement for the relational database. This contribution covers the architectural changes in the system, together with the technology evaluation, benchmark results and conclusions.
The CMS Data Acquisition - Architectures for the Phase-2 Upgrade

Author: Emilio Meschi

Co-authors: Andre Georg Holzner; Ben Craigs; Christian Deldicque; Christoph Paus; Christoph Schwick; Cristian Contescu; Dainius Simelevicius; Dominique Gigi; Frank Glege; Frans Meijers; Georgiana Lavinia Darlea; Guillermo Gomez Ceballos Retuerto; Hannes Sakulin; James Gordon Branson; Jean-Marc Olivier Andre; Jeroen Hegeman; Jonathan Fulcher; Lorenzo Masetti; Luciano Orsini; Marc Dobson; Marco Pieri; Nicolas Doualot; Olivier Chaze; Petr Zejdl; Philipp Maximilian Brummer; Raul Jimenez Estupinan; Remi Mommsen; Samim Erhan; Sergio Cittolin; Srecko Morovic; Thomas Reis; Ulf Behrens; Vivian O’Dell; Zeynep Demiragli

1 CERN
2 Univ. of California San Diego (US)
3 CERN, Geneva, Switzerland
4 Massachusetts Inst. of Technology (US)
5 Fermi National Accelerator Lab. (US)
6 Univ. of California Los Angeles (US)
7 Deutsches Elektronen-Synchrotron (DE)
8 Fermi National Accelerator Laboratory (FNAL)

Corresponding Authors: paus@mit.edu, petr.zejdl@cern.ch, jeroen.hegeman@cern.ch, philipp.maximilian.brummer@cern.ch, raul.jimenez.estupinan@cern.ch, guillelmo.gomez.ceballos@cern.ch, zeynep.demiragli@cern.ch, luciano.orsini@cern.ch, lorenzo.masetti@cern.ch, frans.meijers@cern.ch, ulf.behrens@cern.ch, jean-marc.olivier.andre@cern.ch, emilio.meschi@cern.ch, christian.deldicque@cern.ch, jonathan.fulcher@cern.ch, cristian.contescu@cern.ch, branson@ucsd.edu, benjamin.gordon.craigs@cern.ch, marco.pieri@cern.ch, andre.georg.holzner@cern.ch, dainius.simelevicius@cern.ch, remigius.mommsen@cern.ch, hannes.sakulin@cern.ch, odell@fnal.gov, thomas.reis@cern.ch, nicolas.doualot@cern.ch, dominique.gigi@cern.ch, samim.erhan@cern.ch, christoph.schwick@cern.ch, frank.glege@cern.ch, olivier.chaze@cern.ch, srecko.morovic@cern.ch, lavinia.darlea@cern.ch, marc.dobson@cern.ch, attila.racz@cern.ch, sergio.cittolin@cern.ch

In Long Shutdown 3 the CMS Detector will undergo a major upgrade to prepare for the second phase of the LHC physics program, starting around 2026. The HL-LHC upgrade will bring instantaneous luminosity up to 5x10^{34} cm^{-2} s^{-1} (levelled), at the price of extreme pileup of 200 interactions per crossing. A new silicon tracker with trigger capabilities and extended coverage, and new high granularity forward calorimetry will enhance the CMS acceptance and selection power. This will enable precision measurements of the Higgs boson properties, as well as extend the discovery reach for physics beyond the standard model, while coping with conditions dictated by the HL-LHC parameters.

Following the tradition, the CMS Data Acquisition System will continue to feature two trigger levels. The detector will be read out at an unprecedented data rate of up to 50 Tb/s read at a Level-1 rate of 750 kHz from some 50k high-speed optical detector links, for an average expected event size of 5MB. Complete events will be analysed by a software trigger (HLT) running on standard processing nodes, and selected events will be stored permanently at a rate of up to 10 kHz for offline processing and analysis.
In this paper we discuss the baseline design of the DAQ and HLT systems for the Run 4, taking into account the projected evolution of high speed network fabrics for event building and distribution, and the anticipated performance of many-core CPU and their memory and I/O architectures.

Assuming a modest improvement of the processing power of 12.5% per year for the standard Intel architecture CPU and the affordability, by 2026, of 100-200 Gb/s links, and scaling the current HLT CPU needs for increased event size, pileup, and rate, the CMS DAQ will require about:

- 800 100 Gb/s links connecting 400 servers for event building
- 4500 processing servers, or a total of about 11 MHS for HLT

Implications on hardware and infrastructure requirements for the DAQ “data center” are analysed. Emerging technologies for data reduction, in particular of CPU-FPGA hybrid systems, but also alternative CPU architectures, are considered. These technologies may in the future help containing the TCO of the system, while improving the energy performance and reducing the cooling requirements.

Novel possible approaches to event building and online processing are also examined, which are inspired by trending developments in other areas of computing dealing with large masses of data.

We conclude by discussing the opportunities offered by reading out and processing parts of the detector, wherever the front-end electronics allows, at the machine clock rate (40 MHz). While the full detector is being read out and processed at the Level-1 rate, a second, parallel DAQ system would run as an “opportunistic experiment” processing tracker trigger and calorimeter data at 40 MHz. This idea presents interesting challenges and its physics potential should be studied.

Tertiary Keyword (Optional):
High performance computing

Secondary Keyword (Optional):
Trigger

Primary Keyword (Mandatory):
DAQ

Posters A / Break / 300

Extraction of the luminosity at PANDA

Author: Stefan Pflueger

Co-authors: Christof Motzko; Florian Feldbauer; Heinrich Leithoff; Miriam Fritsch; Stephan Maldaner

1 Helmholtz Institute Mainz
2 Helmholtz-Institut Mainz
3 Johannes-Gutenberg-Universitaet Mainz (DE)

Corresponding Authors: s.pflueger@gsi.de, spfluege@googlemail.com, leithoff@kph.uni-mainz.de, motzko@kph.uni-mainz.de, florian@ep1.rub.de, miriam@slac.stanford.edu, stephan.maldaner@cern.ch

The high precision experiment PANDA is specifically designed to shed new light on the structure and properties of hadrons. PANDA is a fixed target antiproton proton experiment and will be part of Facility for Antiproton and Ion Research (FAIR) in Darmstadt, Germany. When measuring the total cross sections or determining the properties of intermediate states very precisely e.g. via the energy scan method, the precise determination of the luminosity is mandatory.
For this purpose, the PANDA luminosity detector will measure the 2D angular distribution of the elastically scattered antiproton trajectories. For the determination of the luminosity the parametrization of the differential cross section in dependence on the scattering angle is fitted to the measured angular distribution. The fit function is highly complex as it is not only able to correct for the detection efficiency and resolution, but also the antiproton beam shift, spotsize, tilt and divergence. As most of these parameters are extracted from the fit, this method is extremely powerful as it delivers also beam properties.

A sophisticated software package was developed to perform these extensive calculations, which is capable of extracting the luminosity with an accuracy in the permille level. The systematic uncertainties of the determination of the time-integrated luminosity are dominated by the elastic scattering model uncertainty and background contributions.

This talk will cover the complete luminosity determination procedure.

**Secondary Keyword (Optional)**
High performance computing

**Primary Keyword (Mandatory)**
Analysis tools and techniques

**Tertiary Keyword (Optional)**

---

**A Multipurpose Computing Center with Distributed Resources**

**Author:** Jiri Chudoba

**Co-authors:** Alexandr Mikula; Dagmar Adamova; Jan Svec; Martin Adam; Petr Vokac; Tomas Koubota; Václav Říkal; Jana Uhlirova

1 Acad. of Sciences of the Czech Rep. (CZ)
2 Czech Technical University (CZ)
3 Institute of Physics of the CAS

**Corresponding Authors:** alexandr.mikula@cern.ch, jiri.chudoba@cern.ch, dagmar.adamova@cern.ch, martin.adam@cern.ch, jan.svec@cern.ch, rikal@fzu.cz, uhlirova@fzu.cz, petr.vokac@cern.ch, koubat@fzu.cz

The Computing Center of the Institute of Physics (CC IoP) of the Czech Academy of Sciences serves a broad spectrum of users with various computing needs. It runs WLCG Tier-2 center for the ALICE and the ATLAS experiments; the same group of services is used by astroparticle physics projects the Pierre Auger Observatory (PAO) and the Cherenkov Telescope Array (CTA). OSG stack is installed for the NOvA experiment. Other groups of users use directly local batch system. Storage capacity is distributed to several locations. DPM servers used by the ATLAS and the PAO are all in the same server room, but several xrootd servers for the ALICE experiment are operated in the Nuclear Physics Institute in Rez, about 10 km away. The storage capacity for the ATLAS and the PAO is extended by resources of the CESNET - the Czech National Grid Initiative representative. Those resources are in Plzen and Jihlava, more than 100 km away from the CC IoP. Both distant sites use a hierarchical storage solution based on disks and tapes. They installed one common dCache instance, which is published in the CC IoP BDII. ATLAS users can use these resources using the standard ATLAS tools in the same way as the local storage without noticing this geographical distribution.

Computing clusters LUNA and EXMAG dedicated to users mostly from the Solid State Physics departments offer resources for parallel computing. They are part of the Czech NGI infrastructure MetaCentrum with distributed batch system based on torque with a custom scheduler. Clusters are installed remotely by the MetaCentrum team and a local contact helps only when needed. Users from IoP have exclusive access only to a part of these two clusters and take advantage of higher priorities on the rest (1500 cores in total), which can also be used by any user of the MetaCentrum.
IoP researchers can also use distant resources located in several towns of the Czech Republic with a capacity of more than 12000 cores in total.

This contribution will describe installation and maintenance procedures, transition from cfengine to puppet, monitoring infrastructure based on tools like nagios, munin, ganglia and organization of the user support via Request Tracker. We will share our experience with log file processing using ELK stack. The network infrastructure description and its load will also be given.

Tertiary Keyword (Optional):
Distributed data handling

Secondary Keyword (Optional):
Monitoring

Primary Keyword (Mandatory):
Computing facilities

Track 7: Middleware, Monitoring and Accounting / 302

Consolidating WLCG topology and configuration in the Computing Resource Information Catalogue

Authors: Alessandro Di Girolamo\(^1\); Alexey Anisenkov\(^2\); Maria Alandes Pradillo\(^1\)

Co-authors: Alessandra Forti\(^3\); Andrea Sciaba\(^1\); Andrea Valassi\(^1\); Edward Karavakis\(^1\); Giuseppe Bagliesi\(^4\); Jose Flix Molina\(^1\); Julia Andreeva\(^1\); Maarten Litmaath\(^1\); Maria Dimou\(^1\); Simone Campana\(^1\); Stefano Belforte\(^6\); Stephan Lammel\(^7\)

1 CERN
2 Budker Institute of Nuclear Physics (RU)
3 University of Manchester (GB)
4 Universita di Pisa & INFN (IT)
5 Centro de Investigaciones Energ. Medioambientales y Tecn. - (ES
6 Universita e INFN, Trieste (IT)
7 Fermi National Accelerator Lab. (US)

Corresponding Authors: julia.andreeva@cern.ch, lammel@fnal.gov, simone.campana@cern.ch, giuseppe.bagliesi@cern.ch, jose.flix.molina@cern.ch, andrea.valassi@cern.ch, stefano.belforte@cern.ch, alessandro.digirolamo@cern.ch, maria.alandes.pradillo@cern.ch, maria.dimou@cern.ch, andrea.sciaba@cern.ch, maarten.litmaath@cern.ch, alessandra.forti@cern.ch, alexey.anisenkov@cern.ch, edward.karavakis@cern.ch

The Worldwide LHC Computing Grid infrastructure links about 200 participating computing centers affiliated with several partner projects. It is built by integrating heterogeneous computer and storage resources in diverse data centers all over the world and provides CPU and storage capacity to the LHC experiments to perform data processing and physics analysis. In order to be used by the experiments, these distributed resources should be well described, which implies easy service discovery and detailed description of service configuration. Currently this information is scattered over multiple generic information sources like GOCDB, OIM, BDII and experiment-specific information systems. Such a model does not allow to validate topology and configuration information easily. Moreover, information in various sources is not always consistent. Finally, the evolution of computing technologies introduces new challenges. Experiments are more and more relying on opportunistic resources, which by their nature are more dynamic and should also be well described in the WLCG information system.

This contribution describes the new WLCG configuration service CRIC (Computing Resource Information Catalog) which collects information from various information providers, performs validation
and provides a consistent set of UIs and APIs to the LHC VOs for service discovery and usage configuration. The main requirements for CRIC are simplicity, agility and robustness. CRIC should be able to be quickly adapted to new types of computing resources, new information sources, and allow for new data structures to be implemented easily following the evolution of the computing models and operations of the experiments.

The implementation of CRIC was inspired by the successful experience with the ATLAS Grid Information System (AGIS). The first prototype was put in place in a short time thanks to the fact that the substantial part of AGIS code was re-used though some re-factoring required in order to perform clean decoupling in two parts:

- A core which describes all physical endpoints and provides a single entry point for WLCG service discovery.
- Experiment-specific extensions (optional), implemented as plugins. They describe how the physical resources are used by the experiments and contain additional attributes and configuration which are required by the experiments for operations and organization of their data and workflows.

CRIC not only provides a current view of the WLCG infrastructure, but also keeps track of performed changes and audit information. Its administration interface allows authorized users to make changes. Authentication and authorization are subject to experiment policies in terms of data access and update privileges.

Secondary Keyword (Optional):
Accounting and information

Primary Keyword (Mandatory):
Computing middleware

Tertiary Keyword (Optional):

Posters B / Break / 303

The Instant Glidein: A generic approach for the late-binding jobs to various resource types

Author: Laurence Field
Co-author: Iain Bradford Steers

1 CERN

Corresponding Authors: laurence.field@cern.ch, iain.steers@cern.ch

High-throughput computing requires resources to be allocated so that jobs can be run. In a highly distributed environment that may be comprised of multiple levels of queueing, it may not be certain where, what and when jobs will run. It is therefore desirable to first acquire the resource before assigning it a job. This late-binding approach has been implemented in resources managed by batch systems using the pilot jobs paradigm, with the HTCondor glidein being a reference implementation. For resources that are managed by other methods such as the IaaS alternative, other approaches for late-binding may be required. This contribution describes one such approach, the instant glidein, as a generic method for implementing late-binding for many resource types.

Tertiary Keyword (Optional):
Distributed workload management

Secondary Keyword (Optional):
IPv6 Security

Authors: Dave Kelsey\(^1\), Francesco Prelz\(^2\)
Co-author: Romain Wartel \(^3\)

\(^1\) STFC - Rutherford Appleton Lab. (GB)
\(^2\) Università degli Studi e INFN Milano (IT)
\(^3\) CERN

Corresponding Authors: romain.wartel@cern.ch, david.kelsey@stfc.ac.uk, francesco.prelz@mi.infn.it

IPv4 network addresses are running out and the deployment of IPv6 networking in many places is now well underway. Following the work of the HEPiX IPv6 Working Group, a growing number of sites in the Worldwide Large Hadron Collider Computing Grid (WLCG) have deployed dual-stack IPv6/IPv4 services. The aim of this is to support the use of IPv6-only clients, i.e. worker nodes, virtual machines or containers.

The IPv6 networking protocols while they do contain features aimed at improving security also bring new challenges for operational IT security. We have spent many decades understanding and fixing security problems and concerns in the IPv4 world. Many WLCG IT support teams have only just started to consider IPv6 security and they are far from ready to follow best practice, the guidance for which is not easy to find. The lack of maturity of IPv6 implementations together with the increased complexity of the protocol standards and the fact that the new protocol stack allows for pretty much the same attack vectors as IPv4, raise many new issues for operational security teams.

The HEPiX IPv6 Working Group is producing guidance on best practices in this area. This paper will consider some of the security concerns for WLCG in an IPv6 world and present the HEPiX IPv6 working group guidance both for the system administrators who manage IT services on the WLCG distributed infrastructure and also for their related security and networking teams.
The Large Hadron Collider beauty (LHCb) experiment at CERN specializes in investigating the slight differences between matter and antimatter by studying the decays of beauty or bottom (B) and charm (D) hadrons. The detector has been recording data from proton-proton collisions since 2010. Data preservation (DP) project at the LHCb insures preservation of the experimental and simulated (Monte Carlo) data, the scientific software and the documentation. The project will assist in analysing the old data in the future, which is important for replicating the old analysis, looking for the signals predicted by a new theory and improving the current measurements.

The LHCb data are processed with the software and hardware that have been changing over time. Information about the data and the software have been logged in the internal databases and web portals. A current goal of the DP team is to collect and structure these information into a singular, solid database that can be used immediately for scientific purposes and for the long-term future preservation. The database is being implemented as a graph in Neo4j 2.3 and the supporting components are done in Py2Neo. It contains complete details of the official production of LHCb real, experimental data from 2010 to 2015. The data is represented as nodes with collision type, run dates and ID, that is related to the data-taking year, energy of the beam, reconstruction, applications etc. The model applies to the both simulated and the real data.

Data taken at different points in time is compatible with different software versions. The LHCb software stack is built on Gaudi framework, which can run various packages and applications depending on the activity. The applications require other components and these dependencies are captured in the database. The database will recommend the user which software to run in order to analyse specific data. Interface to the database is provided as a web page with a search engine, which allows the users to query and explore the database. The information in the database are worthwhile in the current research; and from DP point of view, they are crucial for replicating an old analyses in the future.

Once the database is fully completed and tested, we will use the graph to identify legacy software and critical components, for example those used in official collaboration productions. These components should be specially treated for the long-term preservation.

Software Quality Control at Belle II

Authors: Christian Pulvermacher\(^1\); Martin Ritter\(^2\); Michal Kristof\(^3\); Thomas Hauth\(^4\); Thomas Kuhr\(^5\); Timothy Gebhard\(^1\)

1 KIT
2 LMU / Cluster Universe
3 Charles University
4 KIT - Karlsruhe Institute of Technology (DE)
5 LMU
Over the last seven years the software stack of the next generation B factory experiment Belle II has grown to over 400,000 lines of C++ and python code, counting only the part included in offline software releases. There are several thousand commits to the central repository by about 100 individual developers per year. To keep a coherent software stack of high quality such that it can be sustained and used efficiently for data acquisition, simulation, reconstruction, and analysis over the lifetime of the Belle II experiment is a challenge.

A set of tools is employed to monitor the quality of the software and provide fast feedback to the developers. They are integrated in a machinery that is controlled by a buildbot master and automates the quality checks. The tools include different compilers, cppcheck, the clang static analyzer, valgrind memcheck, doxygen, a geometry overlap checker, a check for missing or extra library links, unit tests, steering file level tests, a sophisticated high-level validation suite, and an issue tracker. The technological development infrastructure is complemented by organizational means to coordinate the development.

The talk will describe the software development process and tools at Belle II and assess its successes and limitations.

Primary Keyword (Mandatory):
Software development process and tools

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

High Level Interface to Conditions Data at Belle II

Authors: Martin Ritter¹; Thomas Kuhr²; marko staric

¹ LMU / Cluster Universe
² LMU

Corresponding Authors: thomas.kuhr@cern.ch, marko.staric@ijs.si, ritter@mpp.mpg.de

The Belle II experiment at the SuperKEKB e+e- accelerator is preparing for taking first collision data next year. For the success of the experiment it is essential to have information about varying conditions available in the simulation, reconstruction, and analysis code.

The interface to the conditions data in the client code was designed to make the life for developers as easy as possible. Two classes, one for single objects and one for arrays of objects, provide a type-safe access. Their interface resembles that of the classes for the access to event-level data with which the developers are already familiar. Changes of the referred conditions objects are usually transparent to the client code, but they can be checked for and functions or methods can be registered that are called back whenever a conditions data object is updated. Relations between objects in arrays can be established by a templated class that looks like a pointer and can use any method return value as key to identify the referred object. The framework behind the interface fetches objects from the back-end database only when needed and caches them while they are valid. It can transparently handle validity ranges that are shorter than a run which is the finest granularity for the validity of payloads in the database. Besides an access to the central database the framework supports local conditions data storage which can be used as fallback solution or to overwrite values in the central database with custom ones.
The talk will present the design of the conditions database interface in the Belle II software, show examples of its application, and report about usage experiences in large-scale Monte Carlo productions and calibration exercises.

Secondary Keyword (Optional):
Databases

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

Tertiary Keyword (Optional):

Track 1: Online Computing / 308

Status of the Calibration and Alignment Framework at the Belle II Experiment

Author: David Dossett

Co-authors: Martin Ritter; Martin Sevior; Sergey Yaschenko; Tadeas Bilka; Thomas Kuhr

1 University of Melbourne
2 LMU / Cluster Universe
3 University of Melbourne (AU)
4 DESY

Corresponding Authors: thomas.kuhr@cern.ch, sergey.yaschenko@desy.de, ritter@mpp.mpg.de, bilka@ipnp.troja.mff.cuni.cz, david.dossett@unimelb.edu.au, martines@unimelb.edu.au

The SuperKEKB $e^+ e^-$ collider has now completed its first turns. The planned running luminosity is 40 times higher than its previous record during the KEKB operation. The Belle II detector placed at the interaction point will acquire a data sample 50 times larger than its predecessor. The monetary and time costs associated with storing and processing this quantity of data mean that it is crucial for the detector components at Belle II to be calibrated quickly and accurately. A fast and accurate calibration allows the trigger to increase the efficiency of event selection, and gives users analysis-quality reconstruction promptly. A flexible framework for fast production of calibration constants is being developed in the Belle II Analysis Software Framework (basf2). Detector experts only need to create two components from C++ base classes. The first collects data from Belle II datasets and passes it to the second stage, which uses this much smaller set of data to run calibration algorithms to produce calibration constants. A Python framework coordinates the input files, order of processing, upload to the conditions database, and monitoring of the output. Splitting the operation into collection and algorithm processing stages allows the framework to optionally parallelize the collection stage in a distributed environment. Additionally, moving the workflow logic to a separate Python framework allows fast development and easier integration with DIRAC; The grid middleware system used at Belle II. The current status of this calibration and alignment framework will be presented.

Tertiary Keyword (Optional):
Databases

Secondary Keyword (Optional):
Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

Posters A / Break / 310

Status of the ComPWA Framework

Author: Mathias Michel

Co-authors: Frank Nerling; Klaus Goetzen; Klaus Peters; Miriam Fritsch; Peter Weidenkaff; Ralf Kliemt; Stefan Pflueger; Wolfgang Gradl

1 Helmholtz Institute Mainz
2 Johannes-Gutenberg-Universitaet Mainz (DE)
3 GSI Darmstadt
4 Institut fuer Experimentalphysik I
5 GSI
6 University of Mainz

Corresponding Authors: k.goetzen@gsi.de, k.peters@gsi.de, matmichel84@googlemail.com, frank.nerling@cern.ch, weidenka@kph.uni-mainz.de, r.kliemt@gsi.de, miriam@slac.stanford.edu, gradl@kph.uni-mainz.de, spfluege@googlemail.com

A large part of the programs of hadron physics experiments deal with the search for new conventional and exotic hadronic states like e.g. hybrids and glueballs. In a majority of analyses a Partial Wave Analysis (PWA) is needed to identify possible exotic states and to classify known states. Of special interest is the comparison or combination of data from multiple experiments. Therefore, a new, agile, and efficient PWA framework ComPWA is being developed. It is modularized to provide easy extension with models and formalisms as well as fitting of multiple datasets, even from different experiments. It provides various modules for fitness estimations and interfaces to the optimization routines from the Minuit2 and the Geneva libraries are currently implemented. The modularity allows complex fit methods like e.g. the model-independent extraction of partial waves. The software aims on the analysis of data from today’s experiments as well as on data from future experiments like e.g. Panda@Fair. Currently ComPWA is used for a model-independent extraction of scalar resonances in radiative $J/\psi$ decays and a D-meson Dalitz plot analysis with data from the BESIII experiment. An update on the status of the ComPWA framework is presented and an overview of the first analyses is given.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):
Analysis tools and techniques

Track 1: Online Computing / 312

Acceleration of Cherenkov angle reconstruction with the new Intel Xeon/FPGA compute platform for the particle identification in the LHCb Upgrade.

Author: Christian Faerber

1
The LHCb experiment at the LHC will upgrade its detector by 2018/2019 to a 'triggerless' readout scheme, where all the readout electronics and several sub-detector parts will be replaced. The new readout electronics will be able to readout the detector at 40MHz. This increases the data bandwidth from the detector down to the event filter farm to 40TBit/s, which also has to be processed to select the interesting proton-proton collision for later storage. The architecture of such a computing farm, which can process this amount of data as efficiently as possible, is a challenging task and several compute accelerator technologies are being considered for use inside the new event filter farm.

In the high performance computing sector more and more FPGA compute accelerators are used to improve the compute performance and reduce the power consumption (e.g. in the Microsoft Catapult project and Bing search engine). Also for the LHCb upgrade the usage of an experimental FPGA accelerated computing platform in the event building or in the event filter farm (trigger) is being considered and therefore tested. This platform from Intel hosts a general CPU and a high performance FPGA linked via a high speed link which is for this platform a QPI link. On the FPGA an accelerator is implemented. The used system is a two socket platform from Intel with a Xeon CPU and an FPGA. The FPGA has cache-coherent memory access to the main memory of the server and can collaborate with the CPU.

As a first step, a computing intensive algorithm to reconstruct Cherenkov angles for the LHCb RICH particle identification was successfully ported to the Intel Xeon/FPGA platform and accelerated by a factor of 35. Also other FPGA accelerators, GPUs, and High Energy Physics trigger algorithms were tested for performance and power consumption. The results show that the Intel Xeon/FPGA platforms, which are built in general for high performance computing, are also very interesting for the High Energy Physics community. Furthermore, the new Intel Xeon/FPGA with Arria10 will be tested.
ALICE (A Large Ion Collider Experiment) is one of the four major detectors located at the LHC at CERN, focusing on the study of heavy-ion collisions. The ALICE High Level Trigger (HLT) is a compute cluster which reconstructs the events and compresses the data in real-time. The data compression by the HLT is a vital part of data taking especially during the heavy-ion runs in order to be able to store the data which implies that reliability of the whole cluster is an important matter.

To guarantee a consistent state among all compute nodes of the HLT cluster we have automatized the operation as much as possible. For automatic deployment of the nodes we use Foreman with locally mirrored repositories and for configuration management of the nodes we use Puppet. Important parameters like temperatures of the nodes are monitored with Zabbix.

During periods without beam the HLT cluster is used for tests and as one of the WLCG Grid sites to compute offline jobs in order maximize the usage of our cluster. To prevent interference with normal HLT operations we introduced a separation via virtual LANs between the normal HLT operation and the grid jobs running inside virtual machines.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

High performance computing

Primary Keyword (Mandatory):

Computing facilities

Posters B / Break / 315

Experiment Software and Projects on the Web with VISPA

Authors: Benjamin Fischer\textsuperscript{1}; Christian Glaser\textsuperscript{1}; Christoph Welling\textsuperscript{1}; Fabian-Andree Heidemann\textsuperscript{1}; Gero Müller\textsuperscript{1}; Marcel Rieger\textsuperscript{1}; Martin Erdmann\textsuperscript{1}; Martin Urban\textsuperscript{1}; Ralf Florian von Cube\textsuperscript{1}; Robert Fischer\textsuperscript{1}; Thorben Quast\textsuperscript{1}

\textsuperscript{1} Rheinisch-Westfälische Tech. Hoch. (DE)

Corresponding Authors: glaser@physik.rwth-aachen.de, martin.urban@physik.rwth-aachen.de, marcel.rieger@cern.ch, florian.von.cube@rwth-aachen.de, thorben.quast@rwth-aachen.de, gmueller@physik.rwth-aachen.de, robert.fischer@cern.ch, benjamin.fischer@cern.ch, fabian-andree.heidemann@cern.ch, martin.erdmann@cern.ch, christoph.welling@rwth-aachen.de

The Visual Physics Analysis (VISPA) project defines a toolbox for accessing software via the web. It is based on latest web technologies and provides a powerful extension mechanism that enables to interface a wide range of applications. Beyond basic applications such as a code editor, a file browser, or a terminal, it meets the demands of sophisticated experiment-specific use cases that focus on physics data analyses and typically require a high degree of interactivity. As an example, we developed a data inspector that is capable of browsing interactively through event content of several data formats, e.g., „MiniAOD” which is utilized by the CMS collaboration. The VISPA extension mechanism can also be used to embed external web-based applications that benefit from dynamic allocation of user-defined computing resources via SSH. For example, by wrapping the „JSROOT” project, ROOT files located on any remote machine can be inspected directly through a VISPA server instance.

We introduced domains that combine groups of users and role-based permissions. Thereby, tailored projects are enabled, e.g. for teaching where access to student’s homework is restricted to a team of tutors, or for experiment-specific data that may only be accessible for members of the collaboration. We present the extension mechanism including corresponding applications and give an outlook onto the new permission system.

Tertiary Keyword (Optional):
Design and Execution of make-like distributed Analyses based on Spotify’s Pipelining Package luigi

Authors: Benjamin Fischer\textsuperscript{1}; Marcel Rieger\textsuperscript{1}; Martin Erdmann\textsuperscript{1}; Robert Fischer\textsuperscript{1}

\textsuperscript{1}Rheinisch-Westfaelische Tech. Hoch. (DE)

Corresponding Authors: marcel.rieger@cern.ch, benjamin.fischer@cern.ch, robert.fischer@cern.ch, martin.erdmann@cern.ch

In particle physics, workflow management systems are primarily used as tailored solutions in dedicated areas such as Monte Carlo production. However, physicists performing data analyses are usually required to steer their individual workflows manually which is time-consuming and often leads to undocumented relations between particular workloads.

We present a generic analysis design pattern that copes with the sophisticated demands of end-to-end HEP analyses and provides a make-like execution environment. It is based on the open-source pipelining package \textit{luigi} which was developed at Spotify and enables the definition of arbitrary workloads, so-called Tasks, and the dependencies between them in a lightweight and scalable structure. Further features are multi-user support, automated dependency resolution and error handling, central scheduling, and status visualization in the web.

In addition to already built-in features for remote jobs and file systems like \textit{Hadoop} and \textit{HDFS}, we added support for \textit{WLCG} infrastructure such as \textit{LSF} and \textit{CREAM} job submission, as well as remote file access through the Grid File Access Library (\textit{GFAL2}). Furthermore, we implemented automated resubmission functionality, software sandboxing, and a command line interface with auto-completion for a convenient working environment.

For the implementation of a $ttH$ cross section measurement with CMS, we created a generic Python interface that provides programmatic access to all external information such as datasets, physics processes, statistical models, and additional files and values. In summary, the setup enables the execution of the entire analysis in a parallelized and distributed fashion with a single command.
CBM is a heavy-ion experiment at the future FAIR facility in Darmstadt, Germany. Featuring self-triggered front-end electronics and free-streaming read-out event selection will exclusively be done by the First Level Event Selector (FLES). Designed as an HPC cluster, its task is an online analysis and selection of the physics data at a total input data rate exceeding 1 TByte/s. To allow efficient event selection, the FLES performs timeslice building, which combines the data from all given input links to self-contained, overlapping processing intervals and distributes them to compute nodes. Partitioning the input data streams into specialized containers allows this task to be performed very efficiently.

The FLES Input Interface defines the linkage between the FEE and the FLES data transport framework. A custom FPGA PCIe board, the FLES Interface Board (FLIB), is used to receive data via optical links and transfer them via DMA to the host’s memory. The current prototype of the FLIB features a Kintex-7 FPGA and provides up to eight 10 GBit/s optical links. A custom FPGA design has been developed for this board. DMA transfers and data structures are optimized for subsequent timeslice building. Index tables generated by the FPGA enable fast random access to the written data containers. In addition the DMA target buffers can directly serve as InfiniBand RDMA source buffers without copying the data. The usage of POSIX shared memory for these buffers allows data access from multiple processes. An accompanying HDL module has been developed to integrate the FLES link into the front-end FPGA designs. It implements the front-end logic interface as well as the link protocol.

Prototypes of all Input Interface components have been implemented and integrated into the FLES test framework. This allows the implementation and evaluation of the foreseen CBM read-out chain. The full chain from FEE up to timeslices has been tested successfully. Setups with 16 and more free-streaming input links will be used for upcoming beam tests. An overview of the FLES Input Interface as well as latest results from performance and stability measurements will be presented.

**DAQ**
LHC Run3 and Run4 represent an unprecedented challenge for HEP computing in terms of both data volume and complexity. New approaches are needed for how data is collected and filtered, processed, moved, stored and analyzed if these challenges are to be met with a realistic budget. To develop innovative techniques we are fostering relationships with industry leaders. CERN openlab is a unique resource for public-private partnership between CERN and leading Information Communication and Technology (ICT) companies. Its mission is to accelerate the development of cutting-edge solutions to be used by the worldwide HEP community. In 2015, CERN openlab started its phase V with a strong focus on tackling the upcoming LHC challenges. Several R&D programs are ongoing in the areas of data acquisition, networks and connectivity, data storage architectures, computing provisioning, computing platforms and code optimisation and data analytics. In this presentation I will give an overview of the different and innovative technologies that are being explored by CERN openlab V and discusses the long-term strategies that are pursued by the LHC communities with the help of industry in closing the technological gap in processing and storage needs expected in Run3 and Run4.

Tertiary Keyword (Optional):
DAQ

Secondary Keyword (Optional):
Artificial intelligence/Machine learning

Primary Keyword (Mandatory):
Parallelization

---

**Workflow Management for Complex HEP Analyses**

**Author:** Robert Fischer

**Co-authors:** Marcel Rieger; Martin Erdmann; Ralf Florian von Cube

**Corresponding Authors:** martin.erdmann@cern.ch, robert.fischer@cern.ch, marcel.rieger@cern.ch, florian.von.cube@rwth-aachen.de

We present the novel Analysis Workflow Management (AWM) that provides users with the tools and competences of professional large scale workflow systems. The approach presents a paradigm shift from executing parts of the analysis to defining the analysis.

Within AWM an analysis consists of steps. For example, a step defines to run a certain executable for multiple files of an input data collection. Each call to the executable for one of those input files can be submitted to the desired run location, which could be the local computer or a remote batch system. An integrated software manager enables automated user installation of dependencies in the working directory at the run location. Each execution of a step item creates one report for bookkeeping purposes containing error codes and output data or file references. Required files, e.g. created by previous steps, are retrieved automatically. Since data storage and run locations are exchangeable from the step’s perspective, computing resources can be used opportunistically. A visualization of the workflow as a graph of the steps in the web browser provides a high-level view on the analysis. The workflow system is developed and tested alongside a ttbb cross section measurement where, for instance, the event selection is represented by one step and a Bayesian statistical inference is performed by another.
The clear interface and dependencies between steps enables a make-like execution of the whole analysis.

Tertiary Keyword (Optional):
Analysi tools and techniques

Secondary Keyword (Optional):
Distributed workload management

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

Track 5: Software Development / 321

New Machine Learning Developments in ROOT

Authors: Lorenzo Moneta¹; Sergei Gleyzer²
Co-author: Omar Andres Zapata Mesa ³

¹ CERN
² University of Florida (US)
³ University of Antioquia & Metropolitan Institute of Technology

Corresponding Authors: sergei.gleyzer@cern.ch, lorenzo.moneta@cern.ch, andresete.chaos@gmail.com

ROOT provides advanced statistical methods needed by the LHC experiments to analyze their data. These include machine learning tools for classification, regression and clustering. TMVA, a toolkit for multi-variate analysis in ROOT, provides these machine learning methods.

We will present new developments in TMVA, including parallelisation, deep-learning neural networks, new features and additional interfaces to external machine learning packages.
We will show the new modular design of the new version of TMVA, cross-validation and hyper parameter tuning capabilities, feature engineering and deep learning.
We will further describe new parallelisation features including multi-threading, multi-processing, cluster parallelisation and present GPU support for intensive machine learning applications, such as deep learning.

Tertiary Keyword (Optional):
Analysi tools and techniques

Secondary Keyword (Optional):
Algorithms

Primary Keyword (Mandatory):
Artificial intelligence/Machine learning

Track 2: Offline Computing / 322

The FCC software - how to keep SW experiment independent
The Future Circular Collider (FCC) software effort is supporting the different experiment design studies for the three future collider options, hadron-hadron, electron-electron or electron-hadron. The software framework used by data processing applications has to be independent of the detector layout and the collider configuration. The project starts from the premise of using existing software packages that are experiment independent and of leveraging other packages, such as the LHCb simulation framework or the ATLAS tracking software, that can be easily modified to factor out any experiment dependency. At the same time, new components are being developed with a view to allowing usage outside of the FCC software project; for example, the data analysis front-end, which is written in Python, is decoupled from the main software stack and is only dependent on the FCC event data model. The event data model itself is generated from configuration files, allowing customisation, and enables parallelisation by supporting a corresponding data layout. A concise overview of the FCC software project will be presented and developments that can be of use to the HEP community highlighted, including the experiment-independent event data model library, the integrated simulation framework that supports Fast and Full simulation and the Tracking Software package.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Analysis tools and techniques

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

---

Clouds and Virtualization are typically used in computing centers to satisfy diverse needs: different operating systems, software releases or fast servers/services delivery. On the other hand solutions relying on Linux kernel capabilities such as Docker are well suited for applications isolation and...
software developing. In our previous work (Docker experience at INFN-Pisa Grid Data Center*) we discussed the possibility to move an HTC environment such as a CMS Tier2 into a Docker approach. During last year we have consolidated the use of Docker for the HTC part of our computing center. Our computing resources are leveraging a two levels infrastructure. The bare metal servers are operated by a Linux operating system compliant with the hardware (HCA, CPU, etc) and software (typically filesystems) requirements. This is the first layer and defines the system administrator’s domain. The second layer takes care of users’ needs and is administered with Docker. This approach improves the isolation of the user domain from the system administrator domain. It also increases the standardization of systems thus reducing the time needed to put into production.

Given the success with the HTC environment we decided to extend this approach to the HPC part of the INFN-Pisa computing center. Up to now about 25% of our HPC resources are using Docker for the users’ domain. We also decided to simplify the bare metal servers management for this we have started to evaluate the integration of the Docker approach with cluster management tools such as Bright Cluster Manager, one of the HPC market leaders of such tools. In this work we will describe the evolution of our infrastructure from HTC to HPC and the integration of Docker and Bright. Given that the use of Docker in our computer center has become more and more important, it was necessary to develop certain ancillary services starting from the image management. For this purpose we have installed an image repository service based on Portus. This service has been integrated with the INFN AAI. Also this aspects are discussed in this work.


**Posters B / Break / 324**

**Towards more common build tools - experience with using spack in HEP**

**Authors:** Benedikt Hegner¹; Brett Viren²; Elizabeth Sexton-Kennedy³; James Amundson⁴

¹ CERN
² Brookhaven National Laboratory
³ Fermi National Accelerator Lab (US)
⁴ Fermi National Accelerator Lab (US)

**Corresponding Authors:** amundson@fnal.gov, benedikt.hegner@cern.ch, sexton@fnal.gov, bv@bnl.gov

Software development in high energy physics follows the open-source software (OSS) approach and relies heavily on software being developed outside the field. Creating a consistent and working stack out of 100s of external, interdependent packages on a variety of platforms is a non-trivial task. Within HEP, multiple technical solutions exist to configure and build those stacks (so-called build tools). Furthermore, quite often software has to be ported to new platforms and operating systems and subsequently patches to the individual externals need to be created. This is a manual and time consuming task, requiring a very special kind of expert knowledge. None of this work is experiment-specific. For this reason, the HEP Software Foundation (HSF) packaging working group evaluated various HEP and non-HEP tools and identified the HPC tool “spack” as
a very promising candidate for a common experiment-independent build tool. This contribution summarizes the build tool evaluations, presents the first experience with using spack in HEP, the required extensions to it, and discusses its potential for HEP-wide adoption.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):
Software development process and tools

Track 6: Infrastructures / 325

Low latency network and distributed storage for next generation HPC systems: the ExaNeSt project.

Author: Piero Vicini

1 Universita e INFN, Roma I (IT)

Corresponding Author: piero.vicini@cern.ch

With processor architecture evolution, the HPC market has undergone a paradigm shift. The adoption of low-cost, Linux-based clusters extended HPC’s reach from its roots in modeling and simulation of complex physical systems to a broad range of industries, from biotechnology, cloud computing, computer analytics and big data challenges to manufacturing sectors. In this perspective, the near future HPC systems will be composed of millions of low-power-consumption computing cores, tightly interconnected by a low latency high performance network, equipped with a new distributed storage architecture, densely packaged but cooled by an appropriate technology.

In the road towards Exascale-class system, several additional challenges wait for a solution; the storage and interconnect subsystems, as well as a dense packaging technology, are three of them.

The ExaNeSt project, started on December 2015 and funded in EU H2020 research framework (call H2020-FETHPC-2014, n. 671553), is a European initiative aiming to develop the system-level interconnect, the NVM (Non-Volatile Memory) storage and the cooling infrastructure for an ARM-based Exascale-class supercomputers. The ExaNeSt Consortium combines industrial and academic research expertise, especially in the areas of system cooling and packaging, storage, interconnects, and the HPC applications that drive all of the above.

ExaNeSt will develop an in-node storage architecture, leveraging on low cost, low-power consumption NVM devices. The storage distributed sub-system will be accessed by a unified low latency interconnect enabling scalability of storage size and I/O bandwidth with the compute capacity. The unified, low latency, RDMA enhanced network will be designed and validated using a network test-bed based on FPGA and passive copper and/or active optical channels allowing the exploration of different interconnection topologies (from low radix n-dimensional torus mesh to higher radix DragonFly topology), routing functions minimizing data traffic congestion and network support to system resiliency.

ExaNeSt also addresses packaging and advanced liquid cooling, which are of strategic importance for the design of realistic systems, and aims at an optimal integration, dense, scalable, and power efficient. In an early stage of the project an ExaNeSt system prototype, characterized by 1000+ ARM cores, will be available acting as platform demonstrator and hardware emulator.

A set of relevant ambitious applications, including HPC codes for astrophysics, nuclear physics, neural network simulation and big data, will support the co-design of the ExaNeSt system providing specifications during design phase and application benchmarks for the prototype platform.

In this talk a general overview of project motivations and objectives will be discussed and the preliminary developments status will be reported.

Secondary Keyword (Optional):
Vibration monitoring system for the RACF data center at BNL

Author: Alexandr Zaytsev

Co-authors: Christopher Hollowell; Costin Caramaruc; Tony Wong; William Strecker-Kellogg

1 Brookhaven National Laboratory (US)
2 Brookhaven National Laboratory
3 Brookhaven National Lab

Corresponding Authors: willsk@bnl.gov, caramarc@bnl.gov, tony@bnl.gov, alezayt@bnl.gov, hollowec@bnl.gov

RHIC & ATLAS Computing Facility (RACF) at BNL is a 15000 sq. ft. facility hosting the IT equipment of the BNL ATLAS WLCG Tier-1 site, offline farms for the STAR and PHENIX experiments operating at the Relativistic Heavy Ion Collider (RHIC), BNL Cloud installations, various Open Science Grid (OSG) resources, and many other physics research oriented IT installations of a smaller scale. The facility originated in 1990 and grew steadily up to the present configuration with 4 physically isolated IT areas with a maximum rack capacity of about 1000 racks and a total peak power consumption of 1.5 MW, of which about 400 racks plus 9 large robotic tape frames are currently deployed.

These IT areas are provided with a raised floor and a distributed group of chilled-water cooled CRAC units deployed both on the false floor (20 Liebert CRAC units are distributed across the area) and in the basement of the data center building (two large units constructed as a part of the original data center building back to late 1960s). Currently the RACF data center has about 50 PB of storage deployed on top approximately 20k spinning HDDs and 70 PB of data stored on 60k tapes loaded into the robotic silos provided with 180 tape drives, that are potentially sensitive to external sources of vibration. An excessive vibration level could potentially endanger the normal operation of IT equipment, cause the equipment shutdown and even reduce the expected lifetime of the HDDs, unless the source of vibration is detected and eliminated quickly. In our environment the CRAC units deployed on the false floor are the cause of such problems in most of the cases, but sometimes similar issues can be a result of mechanical interference between the equipment deployed in the adjacent racks. Normally the mechanical problems related to the CRAC units are caught within 12-24 hours by performing regular inspections of the area by the RACF data center personnel, yet the need was realized in 2015-2016 for a dedicated and fully automated system that would provide the means of early detection of unwanted vibration sources and gatherer of historical data of the energy spectrum evolution for the known (constantly) present sources, such as nominally operating CRAC units and data storage equipment.

This contribution gives a summary of the initial design of the vibration monitoring system for the RACF data center and the related equipment evaluations performed in 2016Q1-2, as well as the results of the first equipment deployment of this monitoring system (based on high sensitivity MEMS technology triaxial accelerometers with DC response measurement capability) in one of the IT areas of the RACF data center.
Evolution of the Ceph Based Storage Systems at the RACF

Authors: Alexandr Zaytsev¹; Hironori Ito²
Co-authors: Tejas Rao ²; Tony Wong ²; Xin Zhao ¹

¹ Brookhaven National Laboratory (US)
² Brookhaven National Laboratory

Corresponding Authors: raot@bnl.gov, tony@bnl.gov, xzhao@bnl.gov, alezayt@bnl.gov, hito@bnl.gov

Ceph based storage solutions and especially object storage systems based on it are now well recognized and widely used across the HEP/NP community. Both object storage and block storage layers of Ceph are now supporting production ready services for HEP/NP experiments at many research organizations across the globe, including CERN and Brookhaven National Laboratory (BNL), and even the Ceph file system (CephFS) storage layer is now used for that purpose at the RHIC and ATLAS Computing Facility (RACF) at BNL for more than a year. This contribution gives a detailed status report and the foreseen evolution path for the 1 PB scale (by usable capacity, taking into account the internal data redundancy overhead) Ceph based storage system provided with Amazon S3 complaint RADOS gateways, OpenStack Swift to Ceph RADOS API interfaces, and dCache/xRootD over CephFS gateways that is operated in RACF since 2013. The system is currently consisting of two Ceph clusters deployed on top of a heterogeneous set of RAJD arrays altogether containing more than 3.8k 7.2krpm HDDs (one cluster with iSCSI / 10 GbE storage interconnect and another one - with 4 Gb/s Fibre Channel storage interconnect) each provided with an independent IPoIB / 4X FDR Infiniband based fabrics for handling the internal storage traffic. The plans are being made to further increase the scale of this installation up to 5.0k 7.2krpm HDDs and 2 PB of usable capacity before the end of 2016. We also report the performance and stability characteristics observed with our Ceph based storage systems over the last 3 years, and lessons learnt from this experience. The prospects of tighter integration of the Ceph based storage systems with the BNL ATLAS dCache storage infrastructure and the work being done to achieve it are discussed as well.

Track 4: Data Handling

Dynamo - The dynamic data management system for the distributed CMS computing system

Author: Yutaro Iiyama¹
Co-authors: Christoph Paus ¹; Maxim Goncharov ¹
The upgraded Dynamic Data Management framework, Dynamo, is designed to manage the majority of the CMS data in an automated fashion. At the moment all CMS Tier-1 and Tier-2 data centers host about 50 PB of official CMS production data which are all managed by this system. There are presently two main pools that Dynamo manages: the Analysis pool for user analysis data, and the Production pool which is used by the production systems to run (re)-reconstruction and produce Monte Carlo simulation and organize dedicated data transfer tests and tape retrieval. The first goal of the Dynamic Data Management system, to facilitate the management of the data distribution, had already been accomplished shortly after its first deployment in 2014. The second goal of optimizing the accessibility of data for the physics analyses has made major progress in the last year. Apart from the historic data popularity we are now also using the information from analysis jobs queued in the global queue to optimize the data replication for faster analysis job processing. This paper describes the architecture of all relevant components and details the experience of the upgraded system and running it over the last half year.

Posters B / Break / 329

Using 3D Engineering Models in a GEANT Simulation

Author: Leah Welty-Rieger¹

¹ Fermilab

The storage ring for the Muon g-2 experiment is composed of twelve custom vacuum chambers designed to interface with tracking and calorimeter detectors. The irregular shape and complexity of the chamber design made implementing these chambers in a GEANT simulation with native solids difficult. Instead, we have developed a solution that uses the CADMesh libraries to convert STL files from 3D engineering models into tessellated solids. This method reduces design time, improves accuracy and allows for quick updates to the experimental geometry. Details about development and implementation will be discussed.
A lightweight task submission and management infrastructure

Authors: Bing Suo¹; Tian Yan²; Xianghu Zhao³; Xiaomei Zhang⁴
Co-authors: Yao Zhang⁵; ziyan deng⁶

¹ Shandong University
² Institution of High Energy Physics, Chinese Academy of Science
³ NanJing University
⁴ Chinese Academy of Sciences (CN)
⁵ Institute of High Energy Physics and
⁶ Institute of High Energy Physics, Beijing

Corresponding Authors: zhangyao@ihep.ac.cn, suob@ihep.ac.cn, yant@ihep.ac.cn, dengzy@ihep.ac.cn, zhaoxh@ihep.ac.cn, xiaomei.zhang@cern.ch

In the near future, many new experiments (JUNO, LHAASO, CEPC, etc) with challenging data volume are coming into operations or are planned in IHEP, China. The Jiangmen Underground Neutrino Observatory (JUNO) is a multipurpose neutrino experiment to be operational in 2019. The Large High Altitude Air Shower Observatory (LHAASO) is oriented to the study and observation of cosmic rays, which is going to collect data in 2019. The Circular Electron Positron Collider (CEPC) is planned to be a Higgs factory and upgraded to a proton-proton collider in second phase. The DIRAC-based distributed computing system has been enabled to support multi experiments. Development of task submission and management system is the first step for new experiments to have a try or use distributed computing resources in their early stages. In the paper we will present the design and development of a common framework to ease the process of building experiment-specific task submission and management system. Advanced object-oriented programming technology has been used to make infrastructure easy to extend for new experiments. The framework covers the functions including user interface, task creation and submission, run-time workflow control, task monitor and management, dataset management. YAML description language has been used to define tasks, which can be easily interpreted to get configurations from users. The run-time workflow control adopts the concept of DIRAC workflow and allows applications easily to define their several steps in one job and report status separately. Common modules including splitter to split tasks, backend to heterogeneous resources, job factory to generate the related parameters and files for submission have been provided. The monitoring service with web portal has been provided to monitor status for tasks and the related jobs. The dataset management module has been designed to communicate with Dirac File Catalog to implement query and register of dataset. At last the paper will show two experiments JUNO and CEPC how to use this infrastructure to build up their own task submission and management system and complete their first scale try on distributed computing resources.

Secondary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Primary Keyword (Mandatory):
Distributed workload management

Tertiary Keyword (Optional):

Enabling volunteer computing for the BESIII Simulation system

Author: wenxiao kan
Co-authors: Andrei Tsaregorodtsev ¹; Gang CHEN²; Wenjing Wu³

¹ CPPM, Aix-Marseille Université, CNRS/IN2P3, Marseille, France
Traditional cluster computing resources can only partly meet the demand for massive data processing in the High Energy Physics (HEP) experiments, and volunteer computing remains a potential resource for this domain. It collects idle CPU time of desktop computers. Desktop Grid is the infrastructure to aggregate multiple volunteer computers to be included into a larger scale heterogeneous computing environment. Unlike traditional BOINC-based volunteer computing, in Desktop Grids we have to solve the problem of cross platform application deployment and heterogeneous resource integration. This is achieved through the virtualization technology. This way, for example, porting high energy physics simulation on desktop PCs has been implemented at CERN for the LHC experiments.

BES III is a large high energy physics experiment at IHEP, Beijing. In this contribution, we define a six-layer Desktop Grid architecture and introduce the related technologies. Based on that, we construct a volunteer computing system for the BES III simulation. It is integrated with the DIRAC workload management system which in turn aggregates also other grid, cloud and cluster resources. Besides that, the system also integrated PBS system for job submission. We use the CernVM as the basic VM image. However, to meet diverse requirements of VM images, we use a remote VM image service by incorporating the Stratuslab marketplace. The user application deployment is done using the DIRAC Pilot Framework. The paper presents the results of the system performance testing. First, based on the tests of different configuration attributes of virtual machine, we found that memory was the key factor, which effects the job efficiency. The optimal memory requirements are defined. Then, we analyze traces of the BES III simulation job execution on a virtual machine. We present the analysis of the influence of performance of the hard disk throughput and the network bandwidth on the job operation. Finally, the Desktop Grid system architecture based on BOINC for BES III simulation is proposed. The test results show that the system can meet the requirements of the BES III computing and can be also offered to other high energy physics experiments.

**Secondary Keyword (Optional):**
Cloud technologies

**Primary Keyword (Mandatory):**
Virtualization

**Tertiary Keyword (Optional):**

---

**Microservices for systematic profiling and monitoring of the refactoring process at the LHCb experiment**

**Author:** Sasha Mazurov

**Co-author:** Ben Couturier

1 University of Birmingham (GB)

2 CERN

**Corresponding Authors:** ben.couturier@cern.ch, alexander.mazurov@cern.ch

Any time you modify an implementation within a program, change compiler version or operating system, you should also do regression testing. You can do regression testing by rerunning existing tests against the changes to determine whether this breaks anything that worked prior to the change and by writing new tests where necessary. At LHCb we have a huge codebase which is maintained by many people and can be run within different setups. Such situation leads to the crucial necessity
to guide refactoring with a central profiling system that helps to run tests and find the impact of changes.

In our work we present a software architecture and tools for running a profiling system. This system is responsible for systematically running regression tests, collecting and comparing results of these tests, so changes between different setups can be observed and reported. The main feature of our solution is that it is based on a microservices architecture. Microservices break a large project into loosely coupled modules, which communicate with each other through simple APIs. Such modular architectural style helps us to avoid general pitfalls of monolithic architectures such as hard to understand and maintaining of a large codebase and ineffective scalability. Our solution also allows to escape a complexity of microservices deployment process by using software containers and services management tools. Containers and service managers let us quickly deploy linked modules in development, production or in any other environments. Most of the developed modules are generic which means that the proposed architecture and tools can be used not only in LHCb but adopted for other experiments and companies.

Tertiary Keyword (Optional):
Visualization

Secondary Keyword (Optional):
Monitoring

Primary Keyword (Mandatory):
Software development process and tools

Track 2: Offline Computing / 334

Validation of Physics Models of Geant4 using data from CMS experiment

Author: Sunanda Banerjee

Co-authors: Serguei Bityukov; Vladimir Ivantchenko

1 Fermi National Accelerator Lab. (US)
2 ITEP
3 CERN

Corresponding Authors: bityukov@mail.cern.ch, vladimir.ivantchenko@cern.ch, sunanda.banerjee@cern.ch

CMS has tuned its simulation program and chosen a specific physics model of Geant4 by comparing the simulation results with dedicated test beam experiments. CMS continues to validate the physics models inside Geant4 using the test beam data as well as collision data. Several physics lists (collection of physics models) inside the most recent version of Geant4 provide good agreement of the energy response, resolution of pions and protons. The validation results from these studies will be presented. Shower shapes of electrons and photons evaluate the goodness of the description of electromagnetic physics in Geant4 while response of isolated charged particles are used to examine the predictions of hadronic models within Geant4. Use of Geant4 to explain rare anomalous hits in the calorimeter will also be discussed.

Primary Keyword (Mandatory):
Simulation

Secondary Keyword (Optional):

Tertiary Keyword (Optional):
LHCb trigger streams optimization

Authors: Alexander Panin¹; Andrey Ustyuzhanin¹; Artem Redkin²; Denis Derkach¹; Ilya Trofimov²; Mika Anton Vesterinen³; Nikita Kazeev¹; Radoslav Neychev¹

¹ Yandex School of Data Analysis (RU)
² Yandex Data Factory
³ Ruprecht-Karls-Universitaet Heidelberg (DE)

Corresponding Authors: kazeevn@yandex-team.ru, artemredkin@yandex-team.ru, trofim@yandex-team.ru, andrey.ustyuzhanin@cern.ch, denis.derkach@cern.ch, radoslav.neychev@cern.ch, mika.vesterinen@cern.ch, alexander.panin@cern.ch

The LHCb experiment stores around $10^{11}$ collision events per year. A typical physics analysis deals with a final sample of up to $10^7$ events. Event preselection algorithms (lines) are used for data reduction. They are run centrally and check whether an event is useful for a particular physical analysis. The lines are grouped into streams. An event is copied to all the streams its lines belong, possibly duplicating it. Due to the storage format allowing only sequential access, analysis jobs read every event and discard the ones they don’t need.

This scheme efficiency heavily depends on the streams composition. By putting similar lines together and balancing the streams sizes it’s possible to reduce the overhead. There are additional constraints that some lines are meant to be used together so they must go to one stream. The total number of streams is also limited by the file management infrastructure.

We developed a method for finding an optimal streams composition. It can be used for different cost functions, has the number of streams as an input parameter and accommodates the grouping constraint. It has been implemented using Theano and the results are being incorporated into the streaming of the LHCb Turbo output with the projected analysis jobs IO time decrease of 20-50%.

¹ Theano: A Python framework for fast computation of mathematical expressions, The Theano Development Team
² Separate file streams https://gitlab.cern.ch/hschrein/Hlt2StreamStudy, Henry Schreiner et. al
³ The LHCb Turbo Stream, Sean Benson et al., CHEP-2015

Tertiary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Secondary Keyword (Optional):
Distributed workload management

Primary Keyword (Mandatory):
Distributed data handling

Detector Simulations with DD4hep

Authors: Andre Sailer¹; Frank-Dieter Gaede²; Marko Petric¹; Markus Frank¹; Nikiforos Nikiforou¹; Shaojun Lu³

¹ CERN
² Deutsches Elektronen-Synchrotron (DE)
Detector design studies, test beam analyses, or other small particle physics experiments require the simulation of more and more detector geometries and event types, while lacking the resources to build full scale Geant4 applications from scratch. Therefore an easy-to-use yet flexible and powerful simulation program that solves this common problem but can also be adapted to specific requirements is needed.

The groups supporting studies of the linear collider detector concepts ILD, SiD and CLICdp as well as detector development collaborations CALICE and FCal have chosen to use the DD4hep geometry framework and its DDG4 pathway to Geant4 for this purpose. DD4hep with DDG4 offers a powerful tool to create arbitrary detector geometries and gives access to all Geant4 action stages.

The DDG4 plugins suite includes the handling of a wide variety of input formats; access to the Geant4 particle gun or general particles source; the handling of Monte Carlo truth information – e.g., linking hits and the primary particle that caused them – indispensable for performance and efficiency studies. An extendable array of segmentations and sensitive detector allows the simulation of a wide variety of detector technologies.

In this presentation we will show how our DD4hep based simulation program allows one to perform complex Geant4 detector simulations without compiling a single line of additional code by providing a palette of sub-detector components that can be combined and configured via compact XML files, and steering the simulation either completely via the command line or via simple python steering files interpreted by a python executable. We will also show how additional plugins and extensions can be created to increase the functionality.

Secondary Keyword (Optional):

Primary Keyword (Mandatory):
Simulation

Track 4: Data Handling / 340

Storage Quality-of-Service in Cloud-based Scientific Environments: A Standardization Approach

Authors: Benjamin Ertl¹; Maciej Brzezniak²; Marcus Hardt³; Patrick Fuhrmann⁴; Paul Millar⁵

Co-authors: Andrea Ceccanti; Giacinto Donvito⁴; Vladimir Sapunenko⁶

¹ Karlruhe Institute of Technology
² PSNC Poznan Poland
³ Karlsruhe Institute of Technology
⁴ DESY
⁵ INFN-Bari
⁶ INFN-CNAF (IT)
When preparing the Data Management Plan for larger scientific endeavours, PI’s have to balance between the most appropriate qualities of storage space along the line of the planned data lifecycle, its price and the available funding. Storage properties can be the media type, implicitly determining access latency and durability of stored data, the number and locality of replicas, as well as available access protocols or authentication mechanisms. Negotiations between the scientific community and the responsible infrastructures generally happen upfront, where the amount of storage space, media types, like: disk, tape and SSD and the foreseeable data lifecycles are negotiated.

With the introduction of cloud management platforms, both in computing and storage, resources can be brokered to achieve the best price per unit of a given quality. However, in order to allow the platform orchestrators to programatically negotiate the most appropriate resources, a standard vocabulary for different properties of resources and a commonly agreed protocol to communicate those, has to be available. In order to agree on a basic vocabulary for storage space properties, the storage infrastructure group in INDIGO-DataCloud together with INDIGO-associated and external scientific groups, created a working group under the umbrella of the "Research Data Alliance (RDA)". As communication protocol, to query and negotiate storage qualities, the "Cloud Data Management Interface (CDMI)" has been selected. Necessary extensions to CDMI are defined in regular meetings between INDIGO and the "Storage Network Industry Association (SNIA)". Furthermore, INDIGO is contributing to the SNIA CDMI reference implementation as the basis for interfacing the various storage systems in INDIGO to the agreed protocol and to provide an official OpenSource skeleton for systems not being maintained by INDIGO partners.

In a first step, INDIGO will equip its supported storage systems, like dCache, StoRM, IBM GPFS and HPSS and possibly public cloud systems, with the developed interface to enable the INDIGO platform layer to programatically auto-detect the available storage properties and select the most appropriate endpoints based on its own policies.

In a second step INDIGO will provide means to change the quality of storage, mainly to support data life cycle but as well to make data available for on low latency media for demanding HPC application before the requesting jobs are launched, which maps to the ‘bring online’ command in current HEP frameworks.

Our presentation will elaborate on the planned common agreements between the involved scientific communities and the supporting infrastructures, the available software stack, the integration into the general INDIGO framework and our plans for the remaining time of the INDIGO funding period.

Secondary Keyword (Optional):
Virtualization

Primary Keyword (Mandatory):
Storage systems

Tertiary Keyword (Optional):

Posters A / Break / 341

Conversion of Belle’s data to Belle II’s data format

Author: Anze Zupanc

Co-authors: Christian Pulvermacher ; Hulya Atmacan ; Martin Ritter ; Matic Lubej ; Ryosuke ITOH ; Thomas Keck ; pcs4327@gmail.com Park

1 Jozef Stefan Institute
2 KIT
The Belle II experiment is the upgrade of the highly successful Belle experiment located at the KEKB asymmetric-energy e+e- collider at KEK in Tsukuba, Japan. The Belle experiment collected e+e- collision data at or near the centre-of-mass energies corresponding to $\Upsilon(nS)$ ($n \leq 5$) resonances between 1999 and 2010 with the total integrated luminosity of 1 ab$^{-1}$. The data collected by the Belle experiment is still being actively analyzed and is producing more than twenty physics results per year.

Belle II is a next generation B factory experiment that will collect 50 times more data than its predecessor Belle. The higher luminosity at the SuperKEKB accelerator leads to higher background and requires a major upgrade of the detector. The simulation, reconstruction and the analysis software itself has also been upgraded substantially and most parts of the software were newly written taking into account the experience from Belle, other experiments, and advances in technology in order to be able to store, manage and analyse much larger data samples. Newly developed reconstruction algorithms and the need to keep the per-event disk-space requirements low resulted also in a new data format for the Belle II experiment.

The Belle II physics analysis software is organised in a modular way and integrated within the Belle II software framework (basf2). A set of physics analysis modules that perform simple and well-defined tasks and are common to almost all physics analyses are provided. The physics modules do not communicate with each other directly but only through the data access protocols that are part of the basf2. The physics modules are written in C++, Python or a combination of both. Typically, a user performing a physics analysis only needs to provide a job configuration file with the analysis’ specific sequence of physics modules. This approach offers beginners and experts alike the ability to quickly and easily produce physics analysis specific data. Newly developed high-level-analysis tools for the Belle II experiment, such as for example full event tagging, show significantly better performance compared with the tools developed and used at the Belle experiment based on studies performed on simulated data.

This talk will present the Belle to Belle II data format converter that converts simulated and real data collected by the Belle experiment to the data format of the Belle II experiment. The Belle data conversion allows for the testing, validation and to some extent for the calibration of the Belle II physics analysis software and in particular of the high-level-analysis tools well before the first data are collected by the Belle II experiment ensuring faster physics output. On the other hand, the ability to run Belle II tools with better performance over Belle data enables significant improvements to be made on new or existing measurements performed using the existing Belle data.

**Tertiary Keyword (Optional):**

**Secondary Keyword (Optional):**

Preservation of analysis and data

**Primary Keyword (Mandatory):**

Data model

---

**Track 2: Offline Computing / 342**

**Conditions Database for the Belle II Experiment**

**Author:** Lynn Wood

**Co-authors:** Eric Stephan; Malachi Schram; Todd Elsethagen
The Belle II experiment at KEK is preparing for first collisions in 2017. Processing the large amounts of data that will be produced will require conditions data to be readily available to systems worldwide in a fast and efficient manner that is straightforward for both the user and maintainer.

The Belle II conditions database was designed with a straightforward goal: make it as easily maintainable as possible. To this end, HEP-specific software tools were avoided as much as possible and industry standard tools used instead. HTTP REST services were selected as the application interface, which provide a high-level interface to users through the use of standard libraries such as curl. The application interface itself is written in Java and runs in an embedded Payara-Micro Java EE application server. Scalability at the application interface is provided by use of Hazelcast, an open source In-Memory Data Grid (IMDG) providing distributed in-memory computing and supporting the creation and clustering of new application interface instances as demand increases. The IMDG provides fast and efficient access to conditions data, and is persisted or backed by OpenStack’s Trove. Trove manages MySQL databases in a multi-master configuration used to store and replicate data in support of the application cluster.

This talk will present the design of the conditions database environment at Belle II and its use as well as go into detail about the actual implementation of the conditions database, its capabilities, and its performance.

Tertiary Keyword (Optional): Distributed data handling

Secondary Keyword (Optional): Computing middleware

Primary Keyword (Mandatory): Databases

---

**Effective HTCondor-based monitoring system for CMS**

**Authors:** Brian Paul Bockelman¹; Justas Balcas²

**Co-authors:** Antonio Perez-Calero Yzquierdo ³; David Alexander Mason ⁴; Farrukh Aftab Khan ⁵; Jadir Marra Da Silva ⁶; James Letts ⁷; Jean-Roch Vlimant ²; Jose Hernandez ⁸; Marco Mascheroni ⁴

---

¹ University of Nebraska (US)
² California Institute of Technology (US)
³ Centro de Investigaciones Energ. Medioambientales y Tecn. - (ES)
⁴ Fermi National Accelerator Lab. (US)
⁵ National Centre for Physics (PK)
⁶ UNESP - Universidade Estadual Paulista (BR)
⁷ Univ. of California San Diego (US)
⁸ CIEMAT

**Corresponding Authors:** antonio.perez.calero.yzquierdo@cern.ch, marco.mascheroni@cern.ch, dmason@fnal.gov, jadir.marra.da.silva@cern.ch, letts@physics.ucsd.edu, farrukh.aftab.khan@cern.ch, brian.bockelman@cern.ch, chema.hernandez.calama@gmail.com, justas.balca@cern.ch, jean-roch.vlimant@cern.ch
The CMS experiment at LHC relies on HTCondor and glideinWMS as its primary batch and pilot-based Grid provisioning systems. Given the scale of the global queue in CMS, the operators found it increasingly difficult to monitor the pool to find problems and fix them. The operators had to rely on several different web pages, with several different levels of information, and sifting tirelessly through log files in order to monitor the pool completely. Therefore, coming up with a suitable monitoring system was one of the crucial items before the beginning of the LHC Run 2 to ensure early detection of issues and to give a good overview of the whole pool. Our new monitoring page utilizes the HTCondor ClassAd information to provide a complete picture of the whole submission infrastructure in CMS. The monitoring page includes useful information from HTCondor schedulers, central managers, the glideinWMS frontend, and factories. It also incorporates information about users and tasks making it easy for operators to provide support and debug issues.

Primary Keyword (Mandatory):
Monitoring

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

Track 2: Offline Computing / 344

Neuromorphic Computing for High Energy Physics

Authors: David Donofrio¹; Jesse Livezey¹; Kristofer Bouchard¹; Mauricio Garcia-Sciveres²; Paolo Calafiura²; Rebecca Carney³

¹ LBNL  
² Lawrence Berkeley National Lab. (US)  
³ Stockholm University (SE)

Corresponding Authors: mgarcia-sciveres@lbl.gov, jlivezey@lbl.gov, pcalafiura@lbl.gov, ddonofrio@lbl.gov, rebecca.carney@cern.ch, kebouchard@lbl.gov

With the advent of a post-Moore’s law field of computation, novel architectures continue to emerge. HEP experiments, with their ever-increasing computing requirements, are exploring new methods of computation and data handling. With composite multi-million connection neuromorphic chips like IBM’s TrueNorth, neural engineering has now become a feasible technology in this novel computing paradigm.

In this talk we will investigate TrueNorth’s role in High Energy Physics - evaluating its potential in tracking, trigger, and dynamical systems. Our interdisciplinary group relates experiences and challenges in adapting neuromorphic technology for dynamical algorithms such as a Kalman filter, with HEP datasets including high pile-up tracking data.

With this novel approach to data processing comes specific challenges such as the effect of approximate computation on precise predictions and classifications; we will present our experience with these constraints in track reconstruction. It is not only the algorithms that are affected, in this talk we will also explore how the realization of neural networks and Kalman filters affects its performance: be it an implementation as a neuromorphic simulation or an in-Silicon custom chip.

Tertiary Keyword (Optional):
Trigger

Secondary Keyword (Optional):
Artificial intelligence/Machine learning

Primary Keyword (Mandatory):
Reconstruction
Highlights of the Belle II Computing

Author: Takanori Hara

Co-authors: Fabrizio Bianchi; Hideki Miyake; I Ueda; Kiyoshi Hayasaka; Malachi Schram; Martin Sevior; Silvio Pardi; Vikas Bansal; Yuji Kato

1 Universita degli Studi di Torino
2 KEK
3 KEK IPNS
4 Nagoya Univ.
5 University of Melbourne (AU)
6 INFN
7 Pacific Northwest National Laboratory

Corresponding Authors: hideki.miyake@kek.jp, martines@unimelb.edu.au, hayasaka@hepl.phys.nagoya-u.ac.jp, vikas.bansal@pnnl.gov, ueda@post.kek.jp, malachi.schram@pnnl.gov, kato@hepl.phys.nagoya-u.ac.jp, bianchi@to.infn.it, takanori.hara@kek.jp, spardi@na.infn.it

The Belle II is the next-generation flavor factory experiment at the SuperKEKB accelerator in Tsukuba, Japan. The first physics run will take place in 2017, then we plan to increase the luminosity gradually. We will reach the world’s highest luminosity \( L = 8 \times 10^{35} \text{ cm}^{-2} \text{s}^{-1} \) after roughly five years operation and finally collect \( \sim 25 \) Petabyte of raw data per year. Such a huge amount of data allows us to explore the new physics possibilities through a large variety of analyses in quark sectors as well as tau physics and to deepen understanding of nature.

The Belle II computing system is expected to manage the process of massive raw data, production of copious simulation and many concurrent user analysis jobs. The required resource estimation for the Belle II computing system shows a similar evolution profile of the resource pledges in LHC experiments. For the Belle II is a worldwide collaboration of about 700 scientists working in 23 countries and region, we adopted a distributed computing model with DIRAC as a workload and data management system.

In 2015, we performed the successful large-scale MC production campaign with the highest CPU resources for the longest period ever. The biggest difference from the past campaigns is the first practical version of the production system we introduced and tested. We also raked up computational resources such as grid, commercial and academic clouds, and the local batch computing clusters, as much as possible from inside and outside the collaboration. In this March, the commissioning of the SuperKEKB accelerator has started and the Trans-Pacific network was upgraded. Then the full replacement of the KEK central computing system is planned in this summer.

We will present the highlights of the recent achievements, output from the latest MC production campaign and current status of the Belle II computing system in this report.

Secondary Keyword (Optional):
Computing middleware

Primary Keyword (Mandatory):
Computing models

Tertiary Keyword (Optional):

A comparison of different database technologies for the CMS Async-StageOut transfer database
AsyncStageOut (ASO) is the component of the CMS distributed data analysis system (CRAB3) that manages users’ transfers in a centrally controlled way using the File Transfer System (FTS3) at CERN. It addresses a major weakness of the previous, decentralized model, namely that the transfer of the user’s output data to a single remote site was part of the job execution, resulting in inefficient use of job slots and an unacceptable failure rate.

Currently ASO manages up to 600k files of various sizes per day from more than 500 users per month, spread over more than 100 site and uses a NoSQL database (CouchDB) as internal bookkeeping and as way to communicate with other CRAB3 components. Since ASO/CRAB3 were put in production in 2014, the number of transfers constantly increased up to a point where the pressure to the central CouchDB instance became critical, creating new challenges for the system scalability, performance, and monitoring. This forced a re-engineering of the ASO application to increase its scalability and lowering its operational effort.

In this contribution we present a comparison of the performance of the current NoSQL implementation and a new SQL implementation, and of how their different strength and features influenced the design choices and operational experience. We also discuss other architectural changes introduced in the system to handle the increasing load and latency in delivering the output to the user.

Primary Keyword (Mandatory): Databases

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

Track 1: Online Computing / 348

The Trigger and Data Acquisition System for the KM3NeT-Italy neutrino telescope

Authors: Annarita Margiotta¹; Carmelo Pellegrino²; Francesco Giacomini³; Matteo Favaro⁴; Matteo Manzali⁵; Tommaso Chiarusi⁶

¹ Universita e INFN, Bologna (IT)
² INFN
³ INFN CNAF
⁴ INFN - National Institute for Nuclear Physics
The INFN’s project KM3NeT-Italy, supported with Italian PON (National Operative Programs) fundings, has designed a distributed Cherenkov neutrino telescope for collecting photons emitted along the path of the charged particles produced in neutrino interactions. The detector consists of 8 vertical structures, called towers, instrumented with a total number of 672 Optical Modules (OMs) and its deployment is ongoing 3500 meters deep in the Ionian Sea, at about 80 km from the Sicilian coast. In this contribution the Trigger and Data Acquisition System (TriDAS) developed for the KM3NeT-Italy detector is presented. The “all data to shore” approach is adopted to reduce the complexity of the submarine detector: at the shore station the TriDAS collects, processes and filters all the data coming from the towers, storing triggered events to a permanent storage for subsequent analysis. Due to the large optical background in the sea from 40K decays and bioluminescence, the throughput from the sea can range up to 30 Gbps. This puts strong constraints on the performances of the TriDAS processes and the related network infrastructure.

**Secondary Keyword (Optional):**
Trigger

**Primary Keyword (Mandatory):**
DAQ

**Tertiary Keyword (Optional):**

---

**Track 5: Software Development / 349**

**A programming framework for data streaming on the Xeon Phi**

**Author:** Sylvain Chapeland

**1 CERN**

**Corresponding Author:** sylvain.chapeland@cern.ch

ALICE (A Large Ion Collider Experiment) is a heavy-ion detector studying the physics of strongly interacting matter and the quark-gluon plasma at the CERN LHC (Large Hadron Collider). After the second long shut-down of the LHC, the ALICE detector will be upgraded to cope with an interaction rate of 50 kHz in Pb-Pb collisions, producing in the online computing system (O2) a sustained throughput of 3 TB/s. This data will be processed on the fly so that the stream to permanent storage does not exceed 80GB/s peak, the raw data being discarded.

In the context of assessing different computing platforms for the O2 system, we have developed a framework for the Intel Xeon Phi processors (MIC).

It provides the components to build a processing pipeline streaming the data from the PC memory to a pool of permanent threads running on the MIC, and back to the host after processing. It is based on explicit offloading mechanisms (data transfer, asynchronous tasks) and basic building blocks (FIFOs, memory pools, C++11 threads). The user only needs to implement the processing method to be run on the MIC.

We present in this paper the architecture, implementation, and performance of this system.

**Tertiary Keyword (Optional):**
Parallelization
In order to estimate the capabilities of a Computing slot with limited processing time, it is necessary to know with a rather good precision its "power". This allows for example pilot job to match a task for which the required CPU work is known, or to define the number of events to be processed knowing the CPU work per event. Otherwise one always has the risk that the task is aborted because it exceeds the CPU capabilities of the resource. It also allows a better accounting of the consumed resources.

The traditional way the CPU power is estimated in WLCG since 2007 is using the HEP-Spec06 benchmark (HS06) suite that was verified at the time to scale properly with a set of typical HEP applications. However the hardware architecture of processors has evolved, all WLCG experiments moved to using 64-bit applications and use different compilation flags from those advertised for running HS06. It is therefore interesting to check the scaling of HS06 with the HEP applications.

To this purpose, we have been using CPU intensive massive simulation productions from the LHCb experiment and compared their event throughput to the HS06 rating of the worker nodes. We also compared it with a much faster benchmark script that is used by the DIRAC framework used by LHCb for evaluating at run time the performance of the worker nodes.

This contribution reports on the finding of these comparisons: the main observation is that the scaling with HS06 is no longer fulfilled, while the fast benchmarks have a better scaling but are less precise. One can also clearly see that some hardware or software features when enabled on the worker nodes may enhance their performance beyond expectation from either benchmark, depending on external factors.
ALICE HLT Run2 performance overview

M. Krzewicki for the ALICE collaboration

The ALICE High Level Trigger (HLT) is an online reconstruction and data compression system used in the ALICE experiment at CERN. Unique among the LHC experiments, it extensively uses modern coprocessor technologies like general purpose graphic processing units (GPGPU) and field programmable gate arrays (FPGA) in the data flow.

Real-time data compression is performed using a cluster finder algorithm implemented on FPGA boards and subsequent optimisation and Huffman encoding stages. These data, instead of raw clusters, are used in storage and the subsequent offline processing. For Run 2 and beyond, the compression scheme is being extended to provide higher compression ratios.

Track finding is performed using a cellular automaton and a Kalman filter algorithm on GPGPU hardware, where CUDA, OpenCL and OpenMP (for CPU support) technologies can be used interchangeably.

In the context of the upgrade of the readout system the HLT framework was optimised to fully handle the increased data and event rates due to the time projection chamber (TPC) readout upgrade and the increased LHC luminosity.

Online calibration of the TPC using HLT’s online tracking capabilities was deployed. To this end, offline calibration code was adapted to run both online and offline and the HLT framework was extended accordingly. The performance of this schema is important to Run 3 related developments.

Online calibration can, next to being an important exercise for Run 3, reduce the computing workload during the offline calibration and reconstruction cycle already in Run 2.

A new multi-part messaging approach was developed forming at the same time a test bed for the new data flow model of the O2 system, where further development of this concept is ongoing.

Utilising the online reconstruction of many detectors, a new asynchronous monitoring scheme was developed to allow real-time monitoring of the physics performance of the ALICE detector, again making use of the new messaging scheme for both internal and external communication.

The spare compute resource, the development cluster consisting of older HLT infrastructure is run as a tier-2 GRID site using an Openstack-based setup, contributing as many resources as feasible depending on the data taking conditions. In periods of inactivity during shutdowns, both the production and development clusters contribute significant computing power to the ALICE GRID.

**Tertiary Keyword (Optional):**
Distributed data handling

**Secondary Keyword (Optional):**
DAQ

**Primary Keyword (Mandatory):**
Trigger

---

**Track 1: Online Computing / 352**

**Support for Online Calibration in the ALICE HLT Framework**

**Author:** Mikolaj Krzewicki

1. Johann-Wolfgang-Goethe Univ. (DE)
Corresponding Author: mikolaj.krzewicki@cern.ch

Support for Online Calibration in the ALICE HLT Framework

Mikolaj Krzewicki, for the ALICE collaboration

ALICE (A Large Heavy Ion Experiment) is one of the four major experiments at the Large Hadron Collider (LHC) at CERN. The High Level Trigger (HLT) is an online compute farm, which reconstructs events measured by the ALICE detector in real-time. The HLT uses a custom online data-transport framework to distribute the data and the workload among the compute nodes. ALICE employs subdetectors sensitive to environmental conditions such as pressure and temperature, e.g. the Time Projection Chamber (TPC). A precise reconstruction of particle trajectories requires the calibration of these detectors. Performing the calibration in real time in the HLT improves the online reconstructions and renders certain offline calibration steps obsolete, speeding up offline physics analysis. For LHC Run 3, starting in 2020 when data reduction will rely on reconstructed data, online calibration becomes a necessity. In order to run the calibration online, the HLT now supports the processing of tasks that typically run offline. These tasks run massively parallel on all HLT compute nodes, their output is gathered and merged periodically. The calibration results are both stored offline for later use and fed back into the HLT chain via a feedback loop in order to apply calibration information to the track reconstruction. Online calibration and feedback loop are subject to certain time constraints in order to provide up-to-date calibration information and they must not interfere with ALICE data taking. Our approach to run these tasks in asynchronous processes enables us to separate them from normal data taking in a way that makes it failure resilient. We performed a first test of online TPC drift time calibration under real conditions during the heavy-ion run in December 2015. We present an analysis and conclusions of this first test, new improvements and developments based on this, as well as our current scheme to commission this for production use.

Tertiary Keyword (Optional):
Reconstruction

Secondary Keyword (Optional):
Distributed data handling

Primary Keyword (Mandatory):
DAQ

Track 2: Offline Computing / 353

Latest Developments in Delphes 3

Author: Michele Selvaggi

1 Universite Catholique de Louvain (UCL) (BE)

Corresponding Author: michele.selvaggi@cern.ch

A status of recent developments of the DELPHES C++ fast detector simulation framework will be given. New detector cards for the LHCb detector and prototypes for future e+ e- (ILC, FCC-ee) and p-p colliders at 100 TeV (FCC-hh) have been designed. The particle-flow algorithm has been optimised for high multiplicity environments such as high luminosity and boosted regimes. In addition, several new features such as photon conversions/bremsstrahlung and vertex reconstruction including timing information have been included. State-of-the-art pile-up treatment and jet filtering/boosted techniques (such as PUPPI, SoftKiller, SoftDrop, Trimming, N-subjettiness, etc.) have been added. Finally, Delphes has been fully interfaced with the Pythia8 event generator allowing for a complete event generation/detector simulation sequence within the framework.

Tertiary Keyword (Optional):
Exploring private and commercial clouds for BESIII

Authors: Andrei Tsaregorodtsev\textsuperscript{1}; Ricardo Graciani Diaz\textsuperscript{2}; Tian Yan\textsuperscript{3}; Victor Manuel Fernandez Albor\textsuperscript{4}; Victor Mendez Munoz\textsuperscript{4}; Xianghu Zhao\textsuperscript{6}; Xiaomei Zhang\textsuperscript{7}; ZhenTai Ma\textsuperscript{8}

\textsuperscript{1} CPPM, Aix-Marseille Université, CNRS/IN2P3, Marseille, France
\textsuperscript{2} University of Barcelona (ES)
\textsuperscript{3} Institution of High Energy Physics, Chinese Academy of Science
\textsuperscript{4} Universidade de Santiago de Compostela (ES)
\textsuperscript{5} Universitat de Barcelona
\textsuperscript{6} NanJing University
\textsuperscript{7} Chinese Academy of Sciences (CN)
\textsuperscript{8} Institute of High Energy Physics

Corresponding Authors: victormanuel.fernandez@usc.es, graciani@ecm.ub.edu, vmendez@pic.es, yant@ihep.ac.cn, mazt@ihep.ac.cn, atsareg@in2p3.fr, zhaoxh@ihep.ac.cn, xiaomei.zhang@cern.ch

The BESIII experiment located in Beijing is an electron-positron collision experiment to study Tau-Charm physics. Now in its middle age BESIII has aggregated more than 1PB raw data and the distributed computing system has been built up based on DIRAC and put into productions since 2012 to deal with peak demands. Nowadays cloud becomes popular ways to provide resources among BESIII collaborations and VMDIRAC is the first method we adopted to integrate cloud resources, which is an extension in DIRAC of implementing elastic cloud resource scheduling. Instead of submitting pilot jobs, VMDIRAC starts VMs equipped with Job agents through cloud managers according to the demands of DIRAC task queue. The paper firstly will present how we adapt and add the extensions to VMDIRAC to fit into BESIII cases. We also try to share the experiences of using VMDIRAC to integrate heterogeneous cloud resources including OpenStack, OpenNebula and Amazon. The cloudinit has been adopted as the standard way to do contextualization. Also we will describe the performance and price comparisons between private and public clouds in order to give suggestions to BESIII collaborations on resource plans.

In the second part, with the experience of using VMDIRAC, we try to present the design and implementation of a new way of integrating cloud. In this method a CE-like frontend system for cloud has been introduced to start VMs, and accept and assign pilot jobs to the cloud. Instead of changing DIRAC original architecture of pilot-based workload management, the system can keep a uniform architecture to manage cloud same as other resources. In this way, the life cycle of pilots can be well tracked and accounted in DIRAC. At last the paper also will try to compare it with VMDIRAC and figure out the best user cases for two ways.
Experience with Beignet OpenCL on Low Power Intel SoCs

Author: Felice Pantaleo
Co-author: Vincenzo Innocente

1 CERN - Universität Hamburg
2 CERN

Corresponding Authors: felice.pantaleo@cern.ch, vincenzo.innocente@cern.ch

The High Luminosity LHC (HL-LHC) is a project to increase the luminosity of the Large Hadron Collider to $5 \times 10^{34}$ cm$^{-2}$ s$^{-1}$. The CMS experiment is planning a major upgrade in order to cope with an expected average number of overlapping collisions per bunch crossing of 140. The dataset sizes will increase by several orders of magnitude and so will be the request for larger computing infrastructure. The complete exploitation of a machine capability is desirable, if not a requirement, that should anticipate a request for new hardware resources to the funding agencies. Furthermore, energy consumption for computing is becoming more and more an important voice into European data center’s budget. The exploitation of Intel integrated accelerators like graphics processors which are part of our machines, will allow us to achieve much higher energy efficiency and higher performance. This presentation will focus on our first-hand experience in evaluating power efficiency when running part of the CMS track seeding code using open source OpenCL, Beignet, on both the integrated GPU and CPU of a very low power Intel SoC compared to an Intel i7 package.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Processor architectures

Primary Keyword (Mandatory):
High performance computing

Optimizing the resource usage in Cloud based environments: the Synergy approach

Author: Lisa Zangrando
Co-authors: Federica Fanzago; Marco Verlato; Massimo Sgaravatto; Vincent Llorens

1 Universita e INFN, Padova (IT)
2 Unknown
3 CNRS

Corresponding Authors: lisa.zangrando@pd.infn.it, fanzago@pd.infn.it, vincent.llorens@cc.in2p3.fr, marco.verlato@pd.infn.it, massimo.sgaravatto@pd.infn.it

Managing resource allocation in a Cloud based data center serving multiple virtual organizations is a challenging issue. In fact, while batch systems are able to allocate resources to different user groups according to specific shares imposed by the data center administrators, without a static partitioning of such resources, this is not so straightforward in the most common Cloud frameworks, e.g. OpenStack. In the current OpenStack implementation, it is only possible to grant fixed quotas to the different user groups and these resources cannot be exceeded by one group even if there are unused resources...
allocated to other groups. Moreover in the existing OpenStack implementation, when there are no resources available, new requests are simply rejected: it is then up to the user to later re-issue the request. The recently started EU-funded INDIGO-DataCloud project is addressing this issue through ‘Synergy’, a new advanced scheduling service targeted for OpenStack.

In the past we solved the same problem by adopting the batch systems and today, with Synergy, we adopt the same approach by implementing in OpenStack the advanced scheduling logic based on the SLURM fair-share algorithm. This model for resource provisioning ensures that resources are distributed among users following precise shares defined by the administrator. The shares are average values to be guaranteed in a given time window by using an algorithm that takes into account the past usage of such resources.

We present the architecture of Synergy, the status of its implementation, some preliminary results and the foreseen evolution of the service.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Computing middleware

Primary Keyword (Mandatory):
Cloud technologies

Posters A / Break / 358

Gaudi Evolution for Future Challenges

Authors: Benedikt Hegner; Charles Leggett; Marco Clemencic

1 CERN
2 Lawrence Berkeley National Lab. (US)

Corresponding Authors: marco.clemencic@cern.ch, benedikt.hegner@cern.ch, charles.g.leggett@gmail.com

The LHCb Software Framework Gaudi was initially designed and developed almost twenty years ago, when computing was very different from today. It has also been used by a variety of other experiments, including ATLAS, Daya Bay, GLAST, HARP, LZ, and MINERVA. Although it has been always actively developed all these years, stability and backward compatibility have been favoured, reducing the possibilities of adopting new techniques, like multithreaded processing. R&D efforts like GaudiHive have however shown its potential to cope with the new challenges.

In view of the LHC second Long Shutdown approaching and to prepare for the computing challenges for the Upgrade of the collider and the detectors, now is a perfect moment to review the design of Gaudi and plan future developments of the project. To do this LHCb, ATLAS and the Future Circular Collider community joined efforts to bring Gaudi forward and prepare it for the upcoming needs of the experiments.

We present here how Gaudi will evolve in the next years and the long term development plans.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines
Grid Site Availability Evaluation and Monitoring at CMS

Authors: Andrea Sciaba¹; Gaston Lyons Pacini²; Giuseppe Bagliesi³; Rokas Maciulaitis⁴; Stephan Lammel²

¹ CERN
² Fermi National Accelerator Lab. (US)
³ Universita di Pisa & INFN (IT)
⁴ Vilnius University (LT)

Corresponding Authors: giuseppe.bagliesi@cern.ch, lammel@fnal.gov, rokas.ma@gmail.com, andrea.sciaba@cern.ch, gastonlp@fnal.gov

Grid Site Availability Evaluation and Monitoring at CMS

The Compact Muon Solenoid (CMS) experiment at the Large Hadron Collider (LHC) uses distributed grid computing to store, process, and analyze the vast quantity of scientific data recorded every year.

The computing resources are grouped into sites and organized in a tiered structure. A tier consists of sites in various countries around the world. Each site provides computing and storage to the CMS computing grid. In total about 125 sites contribute with resources from hundred to well over tentousand computing cores and storage from tens of TBytes to tens of PBytes.

In such a large computing setup scheduled and unscheduled outages occur continually and are not allowed to significantly impact data handling, processing, and analysis. Unscheduled capacity and performance reductions need to be detected promptly and corrected. CMS developed a sophisticated site evaluation and monitoring system for Run 1 of the LHC based on tools of the Worldwide LHC Computing Grid (WLCG). Sites are supplementing their computing with cloud resources while others focus on increased use of opportunistic resources. For Run 2 of the LHC the site evaluation and monitoring system is being overhauled to enable faster detection/reaction to failures and a more dynamic handling of computing resources. Enhancements to better distinguish site from central service issues and to make evaluations more transparent and informative to site support staff are planned.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Computing facilities

Primary Keyword (Mandatory):
Computing middleware

Track 5: Software Development / 360

“Big Data” in HEP: A comprehensive use case study

Author: Oliver Gutsche¹

Co-authors: Alexey Svyatkovskiy ²; Bo Jayatilaka ¹; Cristina Ana Mantilla Suarez ¹; James Kowalkowski ³; Jim Pivarski ²; Matteo Cremonesi ¹; Saba Sehrish ⁴

¹ Fermi National Accelerator Lab. (US)
² Princeton University
³ Fermi National Accelerator Laboratory (FNAL)
Corresponding Authors: pivarski@fnal.gov, oliver.gutsche@cern.ch, jim.kowalkowski@cern.ch, alexeys@princeton.edu, ssehrish@fnal.gov, bo.jayatilaka@cern.ch, matteo.cremonesi@cern.ch

Experimental Particle Physics has been at the forefront of analyzing the world’s largest datasets for decades. The HEP community was the first to develop suitable software and computing tools for this task. In recent times, new toolkits and systems collectively called “Big Data” technologies have emerged to support the analysis of Petabyte and Exabyte datasets in industry. While the principles of data analysis in HEP have not changed (skimming and slimming experiment-specific data formats), these new technologies use different approaches and promise a fresh look at analysis of very large datasets and could potentially reduce the time-to-physics with increased interactivity.

In this talk, we present an active LHC Run 2 analysis, searching for dark matter with the CMS detector, as a testbed for “Big Data” technologies. We directly compare the traditional NTuple-based analysis with an equivalent analysis using Apache Spark on the Hadoop ecosystem and beyond. In both cases, we start the analysis with the official experiment data formats and produce publication physics plots. We will discuss advantages and disadvantages of each approach and give an outlook on further studies needed.

Tertiary Keyword (Optional):
Software development process and tools

Secondary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Primary Keyword (Mandatory):
Analysis tools and techniques

Track 2: Offline Computing / 361

CMS Full Simulation Status

Author: Vladimir Ivantchenko¹

Co-authors: David Lange ²; Ianna Osborne ³; Mike Hildreth ⁴; Sunanda Banerjee ³

¹ CERN
² Princeton University (US)
³ Fermi National Accelerator Lab. (US)
⁴ Department of Physics-College of Science-University of Notre Da

Corresponding Authors: vladimir.ivantchenko@cern.ch, sunanda.banerjee@cern.ch, ianna.osborne@cern.ch, mikh@undhphp.hep.nd.edu, david.lange@cern.ch

We report current status of the CMS full simulation. For run-II CMS is using Geant4 10.0p02 built in sequential mode. About 8 billion events are produced in 2015. In 2016 any extra production will be done using the same production version. For the development Geant4 10.0p03 with CMS private patches built in multi-threaded mode were established. We plan to use newest Geant4 10.2 for 2017 production. In this work we will present CPU and memory performance of CMS full simulation for various configurations and Geant4 versions, will also discuss technical aspects of the migration to Geant4 10.2. CMS plan to install a new pixel detector for 2017. This allows to perform a revision of the geometry of other sub-detectors and add necessary fixes. For 2016 the digitization of CMS is capable to work in the multi-threaded mode. For simulation of pile up events a method of premixing of QCD events in one file has been established. Performance of CMS digitization will be also discussed in this report.

Tertiary Keyword (Optional):
Design and Deployment of a Elastic Network Test-bed in IHEP Data Center based on SDN

Author: SHAN ZENG
Co-author: Fazhi QI

Corresponding Authors: zengshan@ihep.ac.cn, qfz@ihep.ac.cn

High energy physics experiments produce huge amounts of raw data, while because of the sharing characteristics of the network resources, there is no guarantee of the available bandwidth for each experiment which may cause link competition problems. On the other side, with the development of cloud computing technologies, IHEP have established a cloud platform based on OpenStack which can ensure the flexibility of the computing and storage resources, and more and more computing applications have been moved to this platform, however, under the traditional network architecture, network capability become the bottleneck of restricting the flexible application of cloud computing. This report introduces the SDN implemention in IHEP to solve the above problems, we built a dedicated and elastic network platform based on the data center SDN technologies and network virtualization technologies. Firstly, the elastic network architecture design of the cloud data center based on SDN will be introduced, then in order to provide a high performance network environment in the architecture, a distributed SDN controller based on OpenDaylight is proposed and will be introduced in detail, moreover, the network scheduling algorithm and efficient routing protocol in our environment will also be researched and discussed. In the end, the test results and future works will be shared and analyzed.

Secondary Keyword (Optional): Virtualization
Primary Keyword (Mandatory): Network systems and solutions
Tertiary Keyword (Optional):
In the present run of the LHC, CMS data reconstruction and simulation algorithms benefit greatly from being executed as multiple threads running on several processor cores. The complexity of the Run-2 events requires parallelization of the code in order to reduce the memory-per-core footprint constraining serial-execution programs, thus optimizing the exploitation of present multi-core processor architectures. The allocation of computing resources for multi-core tasks however becomes a complex problem in itself. The CMS workload submission infrastructure employs multi-slot partitionable pilots, built on HTCondor and GlideinWMS native features, to enable the scheduling of single and multi-core jobs simultaneously. This provides a solution for the scheduling problem in a uniform way across grid sites running a diversity of gateways to compute resources and batch system technologies. This contribution will present this strategy and the tools on which it has been implemented. The experience of managing multi-core resources at the Tier-0 and Tier-1 sites during 2015 will be described, along with the current phase of deployment to Tier-2 sites during 2016. The process of performance monitoring and optimization in order to achieve efficient and flexible use of the resources will also be described.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Computing models

Primary Keyword (Mandatory):
Distributed workload management

Posters A / Break / 364

Dynfarm: A Dynamic Site Extension

Authors: Donato De Girolamo\textsuperscript{1}; Vincenzo Ciaschini\textsuperscript{Note}

\textsuperscript{1} INFN

Corresponding Authors: vincenzo.ciaschini@cnaf.infn.it, donato.degirolamo@cnaf.infn.it

Requests for computing resources from LHC experiments are constantly mounting, and so are their peak usage. Since dimensioning a site to handle the peak usage times is impractical due to constraints on resources that many publicly-owned computing centres have, opportunistic usage of resources from external, even commercial cloud providers is becoming more and more interesting, and is even the subject of upcoming initiative from the EU commission, named HelixNebula.

While extra resources are always a good thing, to fully take advantage of them they must be integrated in the site’s own infrastructure and made available to users as if they were local resources.
At the CNAF INFN Tier-1 we have developed a framework, called dynfarm, capable of taking external resources and, placing minimal and easily satisfied requirements upon them, fully integrate them into a pre-existing infrastructure and treat them as if they were local, fully-owned resources.

In this article we for the first time will give a full, complete description of the framework’s architecture along with all of its capabilities, to describe exactly what is possible with it and what are its requirements.

Tertiary Keyword (Optional):
Computing middleware

Secondary Keyword (Optional):
Cloud technologies

Primary Keyword (Mandatory):
Virtualization

Posters A / Break / 365

Application of StoRM+Lustre storage system in IHEP’s distributed computing

Author: Tian Yan¹

Co-authors: Weidong Li ²; Xianghu Zhao ²; Xiaofei Yan ²; Xiaomei Zhang ²

¹ Institution of High Energy Physics, Chinese Academy of Science
² Institute of High Energy Physics, Chinese Academy of Sciences

Corresponding Authors: yanxf@ihep.ac.cn, yant@ihep.ac.cn, liwd@ihep.ac.cn, zhaoxh@ihep.ac.cn, zhangxm@ihep.ac.cn

The distributed computing system in Institute of High Energy Physics (IHEP), China, is based on DIRAC middleware. It integrates about 2000 CPU cores and 500 TB storage contributed by 16 distributed cites. These sites are of various type, such as cluster, grid, cloud and volunteer computing. This system went into production status in 2012. Now it supports multi-VO and serves three HEP experiments: BESIII, CEPC and JUNO.

Several kinds of storage element (SE) are used in IHEP’s distributed computing system, such as dCache, BeStMan and StoRM. In IHEP site, a dCache SE with 128 TB storage capacity served as central grid storage since 2012. The local Lustre storage at IHEP hosts about 4PB data for the above three experiments. Physics data, such as random trigger data and DST data, were uploaded to this dCache SE manually and transferred to remote SEs. Output data of jobs were uploaded to this SE by job wrapper, and then downloaded to local Lustre storage by end user.

To integrate grid storage and local Lustre storage, a scheme of StoRM+Lustre storage system was deployed and tested since 2014. StoRM is a lightweight, scalable, flexible and SRMv2 compliant storage resource manager for disk based storage system. It works on each POSIX file systems, and can take advantage of high performance storage systems based on cluster file system like Lustre. StoRM support both standard Grid access and direct access on data, and it relies on the underlying file system structure to identify the physical data position, instead of querying any databases. These features help us to integrate the grid storage in distributed computing and high capacity Lustre storage systems in each site.

With such a StoRM+Lustre architecture, in which StoRM plays as a role of frontend to the Grid environment, while Lustre as a backend of local accessible, massive and high-performance storage,
users and jobs will feel a nearly unified storage interface. Both local and remote users/jobs exchange data with Lustre storage essentially, without manually data movement between a grid SE and local storage system. Moreover, this architecture can be used to expose physics data in local Lustre to remote sites, therefore it’s a convenient way of sharing data between geographically distributed Lustre file systems.

A StoRM+Lustre instance has been set up at IHEP site, with 66 TB storage capacity. We performed several tests in the past year to assure its performance and reliability, including extensive data transfer test, massive distributed job I/O test, and large-scale concurrency pressure test. A performance and pressure monitoring system was developed for these tests. The testing result is positive. This instance has already been in production since Jan 2015. It shows good reliability in the past months, and plays an important role in Monte-Carlo production as well as data transferring between IHEP and remote sites.

Tertiary Keyword (Optional):
Storage systems

Secondary Keyword (Optional):
Computing middleware

Primary Keyword (Mandatory):
Distributed data handling

Track 1: Online Computing / 366

Accelerated tracking using GPUs at CMS High Level Trigger for Run 3

Author: Felice Pantaleo

Co-authors: Adriano Di Florio; Andrea Bocci; Dario Menasce; Luigi Moroni; Marco Rovere; Mauro Dinardo; Shashi Dugad; Simone Gennai; Vincenzo Innocente

1 CERN - Universität Hamburg
2 Universita e INFN, Bari (IT)
3 CERN
4 Universita & INFN, Milano-Bicocca (IT)
5 Tata Inst. of Fundamental Research (IN)

Corresponding Authors: mauro.dinardo@cern.ch, marco.rovere@cern.ch, luigi.moroni@cern.ch, shashikant.dugad@cern.ch, vincenzo.innocente@cern.ch, andrea.bocci@cern.ch, adriano.diﬂorio@ba.infn.it, dariolivio.menasce@cern.ch, felice.pantaleo@cern.ch, simone.gennai@cern.ch

In 2019 the Large Hadron Collider will undergo upgrades in order to increase the luminosity by a factor two if compared to today’s nominal luminosity. Currently CMS software parallelization strategy is oriented toward scheduling one event per thread. However tracking timing performance depends from the factorial of the pileup leading the current approach to increase latency. When designing a HEP trigger stage, the average processing time is a main constraint and the one-event-per-thread approach will lead to a smaller than ideal fraction of events for which tracking is run. GPUs are becoming wider, with millions of threads running concurrently, and their width is expected to increase in the following years. A many-threads-per-event approach would scale with the pileup offloading the combinatorics to the number of threads available on the GPU. The aim is to have GPUs running at the CMS High Level Trigger during Run 3 for reconstructing Pixel Tracks directly from RAW data. The main advantages would be: - Avoid recurrent data movements between host and device; - Use parallel-friendly data structures without having to transform data into different (OO) representations; - Increase the throughput density of the HLT (events*s^-1 * liter^-1), hence increasing the input rate; - Attract students and give them a set of skills that is very valuable outside HEP.
Improved Cloud resource allocation: how INDIGO-Datacloud is overcoming the current limitations in Cloud schedulers

Authors: Alvaro Lopez Garcia¹; Lisa Zangrando²; Massimo Sgaravatto²; Sara Vallero³; Sonia Taneja¹; Stefano Bagnasco³; Stefano Dal Pra⁶

Co-authors: Davide Salomoni⁴; Giacinto Donvito⁷

¹ Universidad de Cantabria (ES)  
² Universita e INFN, Padova (IT)  
³ Universita e INFN Torino (IT)  
⁴ Universita e INFN, Bologna (IT)  
⁵ L.N.F.N. TORINO  
⁶ INFN  
⁷ INFN-Bari

Performing efficient resource provisioning is a fundamental aspect for any resource provider. Local Resource Management Systems (LRMS) have been used in data centers for decades in order to obtain the best usage of the resources, providing their fair usage and partitioning for the users. In contrast, current cloud schedulers are normally based on the immediate allocation of resources on a first-come, first-served basis, meaning that a request will fail if there are no resources (e.g. OpenStack) or it will be trivially queued ordered by entry time (e.g. OpenNebula). This approach has been identified by the INDIGO-DataCloud project as being too simplistic for accommodating easily scientific workloads. Moreover, the scheduling strategies are based on a static partitioning of the resources, meaning that existing quotas cannot be exceeded, even if there are idle resources allocated to other projects. This is a consequence of the fact that cloud instances are not associated with a maximum execution time and leads to a situation where the resources are under-utilized. This is a non desirable situation in scientific data centers that struggle to obtain the maximum utilization of their resources. The INDIGO-DataCloud project is addressing the described issues in several different areas. On the one hand, by implementing fair-sharing strategies for OpenStack and OpenNebula through the “Synergy” and “FairShare Scheduler(FSS)” components, guaranteeing that the resources are accessed by the users according to the defined fair-share policies established by the system administrator. On the other hand, by implementing a mechanism to execute interruptible (or spot) instances. This way, higher priority instances (such as interactive nodes) can terminate lower priority instances that can be exploited by the users for fault-tolerant processing tasks. This way, it is possible to maximize the overall usage of an infrastructure (by filling the available resources with interruptible instances), without preventing users from running normal instances. Finally, taking into account that scientific data centers are composed of a number of different infrastructures (HPC, Grid, local batch systems, cloud resources), INDIGO is developing a “partition director” to address the problem of granting to a project the availability of a share quota of the physical resources in a center, while balancing their destination over different interface types, such as cloud and batch. This gives a resource provider the ability of dynamically resizing sub-quotas. This ability can also be transferred to the project, who can drive the resizing by controlling the resource request rate on the different infrastructures. This feature can be complemented by other optimizations implemented by INDIGO, such as the Synergy.
component. In this contribution, we will present the work done in the scheduling area during the first year of the INDIGO project in the outlined areas and the foreseen evolutions.

Secondary Keyword (Optional):
Computing middleware

Primary Keyword (Mandatory):
Cloud technologies

Tertiary Keyword (Optional):

Posters B / Break / 368

Netfinder, a real-time network topology discovery tool

Author: Donato De Girolamo

1 INFN

Corresponding Author: donato.degirolamo@cnaf.infn.it

In a large Data Center, such as a LHC Tier-1, where the structure of the Local Area Network and Cloud Computing Systems varies on a daily basis, network management has become more and more complex.

In order to improve the operational management of the network, this article presents a real-time network topology auto-discovery tool named Netfinder.

The information required for effective network management varies according to the task of the moment: it can be the map of the entire physical network, the maps of the overlay networks at VLAN level, i.e. a different map for each VLAN ID, or a “Routing Map” that shows the IP logical connectivity, such as which devices operate as a router and for which IP networks.

The system can operate as a real-time host localization tool: given an hostname or a MAC or an IP address, it is able to to find and map where it is plugged in at the moment a query is made, specifically on which physical port of which switch, as well as the VLAN-ID and the IP Network it belongs to and the physical device that acts as its default gateway.

These informations are completely auto-discovered leveraging the Simple Network Management Protocol (SNMP), Internet Control Message Protocol (ICMP) and the Address Resolution Protocol (ARP). In particular the discovery algorithm is based on standard SNMP MIB-II and ICMP commands, in order to be exploitable in a multi-vendor environment.

This paper will describe both the software architecture and the algorithm used to achieve rapid network topology discovery and real-time host localization, as well as its web interface that allows system and network admins to make queries and visualize the results.

Tertiary Keyword (Optional):
Visualization

Secondary Keyword (Optional):
Monitoring

Primary Keyword (Mandatory):
Network systems and solutions
Exploiting analytics techniques in CMS computing monitoring

Authors: Daniele Bonacorsi¹; Nicolo Magini²
Co-author: Eric Vaandering ²

¹ University of Bologna
² Fermi National Accelerator Lab. (US)

Corresponding Authors: nicolo.magini@cern.ch, daniele.bonacorsi@bo.infn.it, ewv@fnal.gov

The CMS experiment has collected an enormous volume of metadata about its computing operations in its monitoring systems, describing its experience in operating all of the CMS workflows on all of the Worldwide LHC Computing Grid Tiers. Data mining efforts into all these information have rarely been done, but are of crucial importance for a better understanding of how CMS did successful operations, and to reach an adequate and adaptive modelling of the CMS operations, in order to allow detailed optimizations and eventually a prediction of system behaviours. These data are now streamed into the CERN Hadoop data cluster for further analysis. Specific sets of information (e.g. data on how many replicas of datasets CMS wrote on disks at WLCG Tiers, data on which datasets were primarily requested for analysis, etc) were collected on Hadoop and processed with MapReduce applications profiting of the parallelization on the Hadoop cluster. We present the implementation of new monitoring applications on Hadoop, and discuss the new possibilities in CMS computing monitoring introduced with the ability to quickly process big data sets from multiple sources, looking forward to a predictive modeling of the system.

Primary Keyword (Mandatory):
Monitoring

Experience on HTCondor batch system for HEP and other research fields at KISTI-GSDC

Author: Sang Un Ahn¹
Co-authors: Amol Jaikar ²; Byungyun Kong ³; Ilyeon Yeo ¹; Jin Kim ²; Sangwook Bae ¹

¹ KISTI Korea Institute of Science & Technology Information (KR)
² Korea Institute of Science and Technology Information
³ KISTI Korea Institute of Science & Technology Information (KR)

Corresponding Authors: amol@kisti.re.kr, b.kong@cern.ch, wookie@kisti.re.kr, jkim@kisti.re.kr, ilyeon.yeo@cern.ch, sang.un.ahn@cern.ch

Global Science experimental Data hub Center (GSDC) at Korea Institute of Science and Technology Information (KISTI) located at Daejeon in South Korea is the unique data center in the country which helps with its computing resources fundamental research fields deal with the large-scale of data. For historical reason, it has run Torque batch system while recently it starts running HTCondor for new systems. Having different kinds of batch systems implies inefficiency in terms of resource management and utilization. We conducted a research on resource management with HTCondor for several user scenarios corresponding to the user environments that currently GSDC supports.
A recent research on the resource usage patterns at GSDC is considered in this research to build the possible user scenarios. Checkpointing and Super-Collector model of HTCondor give us more efficient and flexible way to manage resources and Grid Gate provided by HTCondor helps interface with Grid environment. The overview on the essential features of HTCondor exploited in this work will be described and the practical examples for HTCondor cluster configuration in our cases will be presented.

Secondary Keyword (Optional):
Distributed workload management

Primary Keyword (Mandatory):
Computing facilities

Tertiary Keyword (Optional):

Track 3: Distributed Computing / 371

Experience in using commercial clouds in CMS

Authors: Gabriele Garzoglio\textsuperscript{None}; Maria Girone\textsuperscript{1}

Co-authors: Anthony Tiradani \textsuperscript{2}; Brian Paul Bockelman \textsuperscript{3}; Burt Holzman \textsuperscript{4}; Dave Dykstra \textsuperscript{4}; David Alexander Mason \textsuperscript{4}; Dirk Hufnagel \textsuperscript{1}; Eric Vaandering \textsuperscript{4}; Hyung Jin Kim \textsuperscript{2}; Ian Fisk \textsuperscript{4}; Oliver Gutsche \textsuperscript{4}; Steven Timm \textsuperscript{2}

\textsuperscript{1} CERN
\textsuperscript{2} Fermilab
\textsuperscript{3} University of Nebraska (US)
\textsuperscript{4} Fermi National Accelerator Lab. (US)

Corresponding Authors: garzoglio@fnal.gov, dirk.hufnagel@cern.ch, dwd@fnal.gov, ian.fisk@cern.ch, oliver.gutsche@cern.ch, ewv@fnal.gov, hjkim@fnal.gov, burt@fnal.gov, timmsteve@yahoo.com, tiradani@fnal.gov, brian.bockelman@cern.ch, dmason@fnal.gov, maria.girone@cern.ch

Historically high energy physics computing has been performed on large purpose-built computing systems. In the beginning there were single site computing facilities, which evolved into the World-wide LHC Computing Grid (WLCG) used today. The vast majority of the WLCG resources are used for LHC computing and the resources are scheduled to be continuously used throughout the year. In the last several years there has been an explosion in capacity and capability of commercial and academic computing clouds. Cloud resources are highly virtualized and intended to be able to be flexibly deployed for a variety of computing tasks. There is a growing interest amongst the cloud providers to demonstrate the capability to perform large scale scientific computing. In this presentation we will discuss results from the CMS experiment using the Fermilab HEP Cloud Facility, which utilized both local Fermilab resources and Amazon Web Services (AWS). The goal was to work with AWS through a matching grant to demonstrate a sustained scale approximately equal to half of the worldwide processing resources available to CMS. We will discuss the planning and technical challenges involved in organizing the most IO intensive CMS workflows on a large-scale set of virtualized resource provisioned by the Fermilab HEPCloud. We will describe the data handing and data management challenges. Also, we will discuss the economic issues and cost and operational efficiency comparison to our dedicated resources. At the end we will consider the changes in the working model of HEP computing in a domain with the availability of large scale resources scheduled at peak times.

Primary Keyword (Mandatory):
Cloud technologies

Secondary Keyword (Optional):

Tertiary Keyword (Optional):
Benchmarking High Performance Computing Architectures with CMS’ Skeleton Framework

Author: Elizabeth Sexton-Kennedy

Co-authors: Christopher Jones; James Amundson; Patrick Gartung

1 Fermi National Accelerator Lab. (US)
2 FNAL
3 Fermilab (US)

In 2012 CMS evaluated which underlying concurrency technology would be the best to use for its multi-threaded framework. The available technologies were evaluated on the high throughput computing systems dominating the resources in use at that time. A skeleton framework benchmarking suite that emulates the tasks performed within a CMSSW application was used to select Intel’s Thread Building Block library, based on the measured overheads in both memory and CPU on the different technologies benchmarked. In 2016 CMS will get access to high performance computing resources that use new many core architectures; machines such as Cori Phase 1&2, Theta, Mira. Because of this we have revived the 2012 benchmark to test it’s performance and conclusions on these new architectures. This talk will discuss the results of this exercise.

CDSLabs: Towards the Next Generation CERN Institutional Repository

Authors: Esteban Gabancho; Harris Tzovanakis; Ludmila Marian; Sebastian Witowski

1 CERN
2 National Technical Univ. of Athens (GR)

CERN Document Server (CDS) is the CERN Institutional Repository, playing a key role in the storage, dissemination and archival for all research material published at CERN, as well as multimedia and some administrative documents. As the CERN’s document hub, it joins together submission and publication workflows dedicated to the CERN experiments, but also to the video and photo teams, to the administrative groups, as well as outreach groups.

In the past year, Invenio, the underlying software platform for CDS, has been undergoing major changes, transitioning from a digital library system to a digital library framework, and moving to a new software stack (Invenio is now built on top of the Flask web development framework, using
Jinja2 template engine, SQLAlchemy ORM, JSONSchema data model, and Elasticsearch for information retrieval). In order to reflect these changes on CDS, we are launching a parallel service, CDSLabs, with the goal of offering our users a continuous view of the reshaping of CDS, as well as increasing the feedback from the community in the development phase, rather than after release.

The talk will provide a detailed view on the new and improved features of the new generation CERN Document Server, as well as its design and architecture. The talk will then cover how the new system is shaped to be more user driven, and to respond better to different needs from different user communities (Library, Experiments, Video team, Photo team, and others), and what mechanisms have been put in place to synchronise the data on the two parallel systems. As a showcase, the talk will present more in depth the architecture and development of a new workflow for submitting and disseminating CERN videos.

Tertiary Keyword (Optional):
Software development process and tools

Secondary Keyword (Optional):
Outreach

Primary Keyword (Mandatory):
Collaborative tools

---

A FairShare Scheduling Service for OpenNebula

Authors: Sara Vallero¹; Stefano Bagnasco²

¹ INFN Torino

Corresponding Authors: bagnasco@to.infn.it, svallero@to.infn.it

In the ideal limit of infinite resources, multi-tenant applications are able to scale in/out on a Cloud driven only by their functional requirements. A large Public Cloud may be a reasonable approximation of this condition, where tenants are normally charged _a posteriori_ for their resource consumption. On the other hand, small scientific computing centres usually work in a saturated regime and tenants are charged _a priori_ for their computing needs by paying for a fraction of the computing/storage resources constituting the Cloud infrastructure. Within this context, an advanced resource allocation policy is needed in order to optimise the use of the data center. We consider a scenario in which a configurable fraction of the available resources is statically assigned and partitioned among projects according to fixed shares. Additional assets are partitioned dynamically following the effective requests per project; efficient and fair access to such resources must be granted to all projects.

The general topic of advanced resource scheduling is addressed by several components of the EU-funded INDIGO-DataCloud project. In this context, dedicated services for the OpenNebula and OpenStack cloud management systems are addressed separately, because of the different internal architectures of the systems.

In this contribution, we describe the FairShare Scheduler Service (FSS) for OpenNebula (ON). The service satisfies resource requests according to an algorithm which prioritises tasks according to an initial weight and to the historical resource usage of the project, irrespective of the number of tasks she has running on the system. The software was designed to be less intrusive as possible in the ON code. By keeping minimal dependencies on the ON implementation details, we expect our code to be fairly independent on future ON internals changes and developments.

The scheduling service is structured as a self-contained module interacting only with the ON XML-RPC interface. Its core component is the Priority Manager (PM), whose main task is to calculate a set of priorities for queued jobs. The manager interacts with a set of pluggable algorithms to calculate priorities. The PM exposes an XML-RPC interface, independent from the ON core one, and uses an independent Priority Database as data back-end. The second fundamental building block of the FSS module is the scheduler itself. The default ON scheduler is responsible for the matching of pending
requests to the most suited physical resources. The queue of pending jobs is retrieved through an XML-RPC call to the ON core and they are served in a first-in-first-out manner. We keep the original scheduler implementation, but the queue of pending jobs to be processed is the one ordered according to priorities as delivered by the PM.

After a description of the module’s architecture, internal data representation and APIs, we show the results of the tests performed on the first prototype.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Virtualization

Primary Keyword (Mandatory):
Cloud technologies

A new VMDIRAC based approach for an Elastic Cloud

Authors: Antonio Amoroso¹; Fabrizio Bianchi²; Flavio Astorino²; Jacopo Pellegrino¹; Liang Yan³; Marco Destefanis¹; Marco Maggiora¹; Stefano Bagnasco³; Tian Yan⁴; Xianghu Zhao⁵; Xiaomei Zhang⁶

¹ University of Turin and INFN Turin
² University of Turin
³ INFN Turin
⁴ IHEP, CAS
⁵ NanJing University

Corresponding Authors: bagnasco@to.infn.it, zhangxm@ihep.ac.cn, destefan@to.infn.it, antonio.amoroso@to.infn.it, yan@to.infn.it, yant@ihep.ac.cn, zhaoxh@ihep.ac.cn, bianchi@to.infn.it, marco.maggiora@to.infn.it, jacopo.pellegrino@to.infn.it, flavio.astorino@gmail.com

The INFN Section of Turin hosts a middle-size multi-tenant cloud infrastructure optimized for scientific computing.

A new approach exploiting the features of VMDIRAC and aiming to allow for dynamic automatic instantiation and destruction of Virtual Machines from different tenants, in order to maximize the global computing efficiency of the infrastructure, has been designed, implemented and is now being tested.

Making use of the standard EC2 API, the two OpenNebula and OpenStack platforms are addressed by this approach.

Tertiary Keyword (Optional):
Distributed workload management

Secondary Keyword (Optional):
Computing middleware

Primary Keyword (Mandatory):
Cloud technologies
HNType: a diverse trace and migration mechanism in the block based Hierarchical Storage System named HazelNut

Author: Ran Du

Co-authors: Gang Chen; Yaodong Cheng; qiulan huang

1 Computing Center, Institute of High Energy Physics, University of Chinese Academy of Sciences
2 Chinese Academy of Sciences (CN)
3 IHEP
4 Institute of High Energy Physics, Beijing

Corresponding Authors: gang.chen@cern.ch, duran@ihep.ac.cn, huangql@ihep.ac.cn, chyd@ihep.ac.cn

HazelNut is a block based Hierarchical Storage System, in which logical data blocks are migrated among storage tiers to achieve better I/O performance. In order to choose migrated blocks, data block I/O process is traced to collect enough information for migration algorithms. There are many ways to trace I/O process and implement block migration. However, how to choose trace metrics and design block migration algorithm is a big problem for system designers. To address this problem, a diverse trace and migration mechanism HNType is proposed. HNType consists two parts, one is a diverse trace mechanism named HNType-t, and the other is a diverse migration mechanism named HNType-s. HNType-t abstracts four base elements and trace operation interfaces based on VFS design concept, which makes it feasible to customize specific trace metrics and trace operations; HNType-s presents three ways of data migration, each way of migration can use customized migration algorithms according to predefined prototypes. Based on HNType, A series of tests are conducted about how block migration is affected by different trace metrics, and three conclusions are drawn according to the experiment results. First, trace metrics of access times and access sectors are not linearly correlated, as a result, these two metrics bring different migration result; Second, I/O completion time used as trace metrics is able to improve sequential I/O by at least 10%; Third, access times used as metrics have a tendency of more migrations upwards.

Tertiary Keyword (Optional):
Secondary Keyword (Optional):
Primary Keyword (Mandatory):
Storage systems

Track 6: Infrastructures / 377

INDIGO-Datacloud: a Cloud-based Platform as a Service oriented to scientific computing for the exploitation of heterogeneous resources

Authors: Davide Salomoni; Giacinto Donvito; Isabel Campos Plasencia; Jesus Marco; Jorge Oliveira Gomes; Luciano Gaido; Lukasz Dutka; Marcin Plociennik; Marcus Hardt; Patrick Fuhrmann; Roberto Barbera

Co-authors: Alessandro Italiano Italiano; Alvaro Lopez Garcia; Alvise Dorigo; Andrea Ceccanti; Bartosz Kryza; Bas Wegh; Christian Bernardt; Diego MICHELOTTO; Doina Cristina Aiftimiei; Elisabetta Ronchieri; Emidio Giorgio; Enrico Fattibene; Federica Fanzago; Frederic Schaer; Germán Moltó; Giovanni Aloisio; Giuseppe Andronico; Ignacio Blanquer; Jiri Sitera; Joao Pina; Lionel Schwarz; Ludek Matyska; Lisa Zangrando; Luukjan Matyska; Marco Fargetta; Marco Verlato; Marica Antonacci; Mario Jorge Moura; Massimo Sgaravatto; Mathieu Velten; Matthew James Viljoen; Michal Owsiak; Michał Urbaniak; Miguel Caballer; Milos Mulac; Pablo Orviz Fernandez; Paul Millar; Peter Solagna; Pierre-Francois Honore; Ricardo Brito Da Rocha; Ricardo Bruno; Sandro Fiore; Sara Vallero; Sonia Taneja; Stanislaw
Corresponding Authors: paul.millar@desy.de, fanzago@pd.infn.it, sitera@civ.zcu.cz, jpina@lip.pt, marco.verlato@pd.infn.it, svallerio@to.infn.it, bagnasco@to.infin.it, david@lip.pt, ricardo.rocha@cern.ch, gaido@to.infn.it, staszek@man.poznan.pl, bas.wegh@kit.edu, sonia.taneja@cern.ch, giovanni.aloisio@unicam.it, mulac@civ.zcu.cz, marcinp@man.poznan.pl, diego.michielotto@cnaf.infn.it, tim.bell@cern.ch, giacinto.donvito@ba.infn.it, matthew.viljoen@cern.ch, enrico.fattibene@cnaf.infn.it, emidio.giorgio@ct.infn.it, pablo.orviz.fernandez@cern.ch, iblanque@dsic.upv.es, marica.antonacci@ba.infin.it, ludek@ics.muni.cz, roberto.barbera@ct.infn.it, lukasz.dutka@cyfronet.pl, lisa.zangrando@pd.infn.it, riccardo.bruno@ct.infin.it, frederic.schaer@cea.fr, lionel.schwarz@in2p3.fr, massimo.sgaravatto@pd.infn.it, bkryza@agh.edu.pl, andrea.ceccanti@cnaf.infn.it, sylvain.reynaud@in2p3.fr, phfonore@cea.fr, t.szepieniec@cyfronet.krakow.pl, sandro.fiore@unile.it, giuseppe.andronico@ct.infn.it, sstr4@cesnet.cz, gmolto@dsic.upv.es, peter.solagna@egi.eu, aloga@ifca.unican.es, michalu@man.poznan.pl, jorge@lip.pt, cristina.aiftimiei@pd.infn.it,
INDIGO-DataCloud (INDIGO for short, https://www.indigo-datacloud.eu) is a project started in April 2015, funded under the EC Horizon 2020 framework program. It includes 26 European partners located in 11 countries and addresses the challenge of developing open source software, deployable in the form of a data/computing platform, aimed to scientific communities and designed to be deployed on public or private Clouds, integrated with existing resources or e-infrastructures.

In this contribution the architectural foundations of the project will be covered, starting from its motivations, discussing technology gaps that currently prevent effective exploitation of distributed computing or storage resources by many scientific communities. The overall structure and timeline of the project will also be described.

The main components of the INDIGO architecture in the three key areas of IaaS, PaaS and User Interfaces will then be illustrated. The modular INDIGO components, addressing the requirements of both scientific users and cloud/data providers, are based upon or extend established open source solutions such as OpenStack, OpenNebula, Docker containers, Kubernetes, Apache Mesos, HTCondor, OpenID-Connect, OAuth, and leverage both de facto and de jure standards.

Starting from the INDIGO components, we will then describe the key solutions that the project has been working on. These solutions are the real driver and objective of the project and derive directly from use cases presented by its many scientific communities, covering areas such as Physics, Astrophysics, Bioinformatics, Structural and molecular biology, Climate modeling, Geophysics, Cultural heritage and others. In this contribution we will specifically highlight how the INDIGO software can be useful to tackle common use cases in the HEP world. For example, we will describe how topics such as batch computing, interactive analysis, distributed authentication and authorization, workload management and data access / placement can be addressed through the INDIGO software. Integration with existing data centers and with well-known tools used in the HEP world such as FTS, Dynafed, HTCondor, dCache, StoRM, with popular distributed filesystems and with Cloud management frameworks such as OpenStack and OpenNebula as well as support for X.509, OpenID-Connect and SAML will also be discussed, together with deployment strategies. A description of the first results and of the available testbeds and infrastructures where the INDIGO software has been deployed will then be given.

Finally, this contribution will discuss how INDIGO-DataCloud can complement and integrate with other projects and communities and with existing multi-national, multi-community infrastructures such as those provided by EGI, EUDAT and the HelixNebula Science Cloud. The importance of INDIGO for upcoming EC initiatives such as the European Open Science Cloud and the European Data Infrastructure will also be highlighted.

Tertiary Keyword (Optional):
Computing middleware

Secondary Keyword (Optional):
Distributed data handling

Primary Keyword (Mandatory):
Cloud technologies

---

Data Flow Simulation of the ALICE Computing System with OM-NET++
Data Flow Simulation of the ALICE Computing System with OMNET++

Rifki Sadikin, Furqon Hensan Muttaqien, Iosif Legrand, Pierre Vande Vyvre for the ALICE Collaboration

The ALICE computing system will be entirely upgraded for Run 3 to address the major challenge of sampling the full 50 kHz Pb-Pb interaction rate increasing by a factor 100 times the present limit. We present, in this paper, models for data flow from detector read-out hosts to storage elements in the upgraded system. The model consists of read-out hosts, network switches, and processing hosts. We simulate storage, buffer and network behavior in discrete event simulations by using OMNET++, a network simulation tool. The simulation assumes that each read-out or processing host is a regular computer host and the event size produced by read-out hosts is set to follow ALICE upgrade requirements. The data, then, flow through TCP/IP-based networks through processing hosts to storage elements. We study the performance of the system for different values of data transfer rate and different data compression/reduction ratio. We use the simulation to estimate storage requirements and the optimal buffer size for network traffic in the upgraded system. Furthermore, we discuss the implications of simulation results for the design.

Tertiary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Secondary Keyword (Optional):
Simulation

Primary Keyword (Mandatory):
DAQ

LHCb migration from Subversion to Git

Authors: Ben Couturier1; Joel Closier1; Marco Cattaneo1; Marco Clemencic1

1 CERN

Corresponding Authors: joel.closier@cern.ch, marco.clemencic@cern.ch, marco.cattaneo@cern.ch, ben.couturier@cern.ch

Because of user demand and to support new development workflows based on code review and multiple development streams, LHCb decided to port the source code management from Subversion to Git, using the CERN GitLab hosting service.

Although tools exist for this kind of migration, LHCb specificities and development models required careful planning of the migration, development of migration tools, changes to the development model, and redefinition of the release procedures. Moreover we had to support a hybrid situation with some software projects hosted in Git and others still in Subversion, or even branches of one projects hosted in different systems.

We present how we addressed the special LHCb issues, the technical details of migrating large non standard Subversion repositories, and how we managed to smoothly migrate the software projects following the schedule of each project manager.
LHCb Dockerized Build Environment

Authors: Ben Couturier\textsuperscript{1}; Joel Closier\textsuperscript{1}; Marco Clemencic\textsuperscript{1}; Mathieu Belin\textsuperscript{2}

\textsuperscript{1} CERN
\textsuperscript{2} Universite Blaise Pascal (FR)

Corresponding Authors: joel.closier@cern.ch, ben.couturier@cern.ch, mathieu.belin@cern.ch, marco.clemencic@cern.ch

Used as lightweight virtual machines or as enhanced chroot environments, Linux containers, and in particular the Docker abstraction over them, are more and more popular in the virtualization communities.

LHCb Core Software team decided to investigate how to use Docker containers to provide stable and reliable build environments for the different supported platforms, including the obsolete ones which cannot be installed on modern hardware, to be used in integration builds, releases and by any developer.

We present here the techniques and procedures set up to define and maintain the Docker images and how these images can be used to develop on modern Linux distributions for platforms otherwise not accessible.

A New Data Access Mechanism for HDFS

Author: Qiang LI\textsuperscript{1}

Co-authors: Gongxing Sun \textsuperscript{1}; Zhanchen WEI \textsuperscript{1}; Zhenyu SUN \textsuperscript{1}

\textsuperscript{1} INSTITUTE OF HIGH ENERGY PHYSICS
With the era of big data emerging, Hadoop has become de facto standard of big data processing. However, it is still difficult to get High Energy Physics (HEP) applications run efficiently on HDFS platform. There are two reasons to explain. Firstly, Random access to events data is not supported by HDFS platform. Secondly, it is difficult to make HEP applications adequate to Hadoop data processing mode. In order to address this problem, a new read and write mechanism of HDFS is proposed. With this mechanism, data access is done on local filesystem instead of through HDFS streaming interface. For data writing, the first file replica is written to the local DataNode, the rest replicas produced by copy of the first replica stored on other DataNodes. The first replica is written under the Blocks storage directory and calculates data checksum after write completion. For data reading, DataNode Daemon provides the data access interface for local Blocks, and Map tasks can read the file replica directly on local DataNode when running locally. To enable files modified by users, three attributes including permissions, owner and group are imposed on Block objects. Blocks stored on DataNode have the same attributes as the file they belong to. Users can modify Blocks when the Map task running locally, and HDFS is responsible to update the rest replicas later after data access done. To further improve the performance of Hadoop system, two optimization on Hadoop scheduler are conducted. Firstly, a Hadoop task selection strategy is presented based on disk I/O performance. With this strategy, an appropriate Map task is selected according to disk workloads, so that disk balance workload is achieved on DataNodes. Secondly, a complete localization task execution mechanism is implemented for I/O intensive jobs. Test results show that average CPU utilization is improved by 10% with the new task selection strategy, data read and write performance is improved about 10% and 40% separately.

**Tertiary Keyword (Optional):**

**Secondary Keyword (Optional):**

Data processing workflows and frameworks/pipelines

**Primary Keyword (Mandatory):**

Cloud technologies

---

**Track 3: Distributed Computing / 383**

**GridPP Dirac: Supporting non-LHC VOs on LHC centric resources**

**Authors:** Daniela Bauer\(^1\); Simon Fayer\(^1\)

\(^1\) Imperial College Sci., Tech. & Med. (GB)

**Corresponding Authors:** daniela.bauer@cern.ch, simon.fayer@cern.ch

The GridPP project in the UK has a long-standing policy of supporting non-LHC VOs with 10% of the provided resources. Up until recently this had only been taken up by a very limited set of VOs, mainly due to a combination of the (perceived) large overhead of getting started, the limited computing support within non-LHC VOs and the ability to fulfill their computing requirements on local batch farms.

In the past year, increased computing requirements and a general tendency for more centralised computing resources including cloud technologies has lead a number of small VOs to reevaluate their strategy.

In response to this, the GridPP project commissioned a multi-VO DIRAC server to act as a unified interface to all its grid and cloud resources. This was then offered as a service to a number of small VOs. Six VOs, four of which were completely new to the grid and two transitioning from a glite/WMS based model, have so far taken up the offer and have used the (mainly) UK grid/cloud infrastructure to complete a significant amount of work each in the last 6 months.
In this talk we present the varied approaches taken by each VO, the support issues arising from these and how these can be re-used by other new communities in the future.

Tertiary Keyword (Optional):
Computing middleware

Secondary Keyword (Optional):
Experience/plans from outside experimental HEP/NP

Primary Keyword (Mandatory):
Distributed workload management

Track 4: Data Handling / 384

Unified data access to e-Infrastructure, Cloud and personal storage within INDIGO-DataCloud

Authors: Bartosz Kryza¹; Lukasz Dutka²; Patrick Fuhrmann³

Co-authors: Davide Salomoni⁴; Giacinto Donvito⁵; Lukasz Dutka

¹ ACC Cyfronet-AGH
² Cyfronet
³ DESY
⁴ INFN CNAF
⁵ INFN-Bari

Corresponding Authors: davide.salomoni@cnaf.infn.it, lukasz.dutka@cyfronet.pl, patrick.fuhrmann@desy.de, giacinto.donvito@ba.infn.it, bkryza@agh.edu.pl, lukasz.dutka@gmail.com

Nowadays users have a variety of options to get access to storage space, including private resources, commercial Cloud storage services as well as storage provided by e-Infrastructures. Unfortunately, all these services provide completely different interfaces for data management (REST, CDMI, command line) and different protocols for data transfer (FTP, GridFTP, HTTP). The goal of the INDIGO-DataCloud project is to give users a unified interface for managing and accessing storage resources provided by different storage providers and to enable them to treat all that space as a single virtual file system with standard interfaces for accessing and transfer, including CDMI and POSIX. This solution enables users to access and manage their data crossing the typical boundaries of federations, created by incompatible technologies and security domains. INDIGO provides ways for storage providers to create and connect trust domains, and allows users to access data across federations, independently of the actual underlying low-level storage technology or security mechanism. The basis of this solution is the Onedata platform (http://www.onedata.org). Onedata is a globally distributed virtual file system, built around the concept of “Spaces”. Each space can be seen as a virtual folder with an arbitrary directory tree structure. The actual storage space can be distributed among several storage providers around the world. Each provider gives the user support for each space in a fixed amount and the actual capacity of the space is the sum of all declared provisions. Each space can be accessed and managed through a web user interface (Dropbox-like), REST and CDMI interfaces, command line as well as mounted directly through POSIX. This gives users several options, the major of which is the ability to access large data sets on remote machines (e.g. worker nodes or Docker containers in the Cloud) without pre-staging and thus interface with existing filesystems. Moreover, Onedata allows for automatic replication and caching of data across different sites and allows cross-interface access (e.g. S3 via POSIX). Performance results covering selected scenarios will also be presented. Besides Onedata, as a complete monolithic middleware, INDIGO offers a data management toolbox allowing communities to provide their own data handling policy engines and delegating the actual work to dedicated services. The INDIGO portfolio ranges from multi tier storage systems with automated media transition based on access profiles and user policies, like StoRM and dCache, via a
reliable and highly scalable file transfer service (FTS), with adaptive data rate management to DynaFed, a lightweight WebDAV storage federation network. FTS is in production for over a decade and is the workhorse of the Worldwide Large Hadron Collider Computing GRID. The DynaFed network federates WebDAV endpoints and lets them appear as a single overlay filesystem.

**Secondary Keyword (Optional):**
Storage systems

**Primary Keyword (Mandatory):**
Distributed data handling

**Tertiary Keyword (Optional):**

---

**Track 6: Infrastructures / 386**

**Stealth Cloud: How not to waste CPU during grid to cloud transitions**

**Authors:** Daniela Bauer¹; Simon Fayer¹

¹ Imperial College Sci., Tech. & Med. (GB)

**Corresponding Authors:** daniela.bauer@cern.ch, simon.fayer@cern.ch

When first looking at converting a part of our site’s grid infrastructure into a cloud based system in late 2013 we needed to ensure the continued accessibility of all of our resources during a potentially lengthy transition period.

Moving a limited number of nodes to the cloud proved ineffective as users expected a significant number of cloud resources to be available to justify the effort of converting their workflows onto the cloud.

Moving a substantial part of the cluster into the cloud carries an inherent risk, such as the cloud nodes sitting idle while waiting for the VOs to finish their development work and other external factors. To mitigate this, we implemented a system to seamlessly move some of the grid workload across to the cloud such that they could use any idle resources. A requirement for this was that the existing grid jobs must be transparently run in a VM without requiring any adjustments by the job owner. To accomplish this we brought together a number of existing tools, ARC-CE, glideinWMS (a pilot-based WMS developed by CMS) & OpenStack.

This talk will focus on the details of the implementation and show that this is a viable long-term solution to maintain resource usage during long periods of transition.

**Secondary Keyword (Optional):**
Computing middleware

**Primary Keyword (Mandatory):**
Cloud technologies

**Tertiary Keyword (Optional):**

---

**Track 1: Online Computing / 387**

**GPU-accelerated track reconstruction in the ALICE High Level Trigger**
Author: David Rohr

Johann-Wolfgang-Goethe Univ. (DE)

Corresponding Author: drohr@jwdt.org

ALICE (A Large Heavy Ion Experiment) is one of the four major experiments at the Large Hadron Collider (LHC) at CERN. The High Level Trigger (HLT) is an online compute farm which reconstructs events measured by the ALICE detector in real-time.

The most compute-intensive part is the reconstruction of particle trajectories called tracking and the most important detector for tracking is the Time Projection Chamber (TPC). The HLT uses a GPU-accelerated algorithm for TPC tracking that is based on the Cellular Automaton principle and on the Kalman filter.

The GPU tracking has been running in 24/7 operation since 2012 in LHC Run 1 and now in Run 2.

In order to better leverage the potential of the GPUs, and speed up the overall HLT reconstruction, we plan to bring more reconstruction steps (e.g. the tracking for other detectors) onto the GPUs.

Moving more steps onto the GPU, and processing them on the GPU at once, will reduce PCI Express transfers and free up CPU resources.

On top of that, modern GPUs and GPU programming APIs provide new features which are not yet exploited by the TPC tracking.

We present our new developments for GPU reconstruction, both with a focus on the online reconstruction on GPU for O2 in ALICE during LHC Run 3, and also taking into account how the current HLT in Run 2 can profit from these improvements.

Tertiary Keyword (Optional):
Parallelization

Secondary Keyword (Optional):
Reconstruction

Primary Keyword (Mandatory):
High performance computing

Track 1: Online Computing / 389

Improvements of the ALICE HLT data transport framework for LHC Run 2

Author: David Rohr

Johann-Wolfgang-Goethe Univ. (DE)

Corresponding Author: drohr@jwdt.org

The ALICE HLT uses a data transport framework based on the publisher subscriber message principle, which transparently handles the communication between processing components over the network and between processing components on the same node via shared memory with a zero copy approach.

We present an analysis of the performance in terms of maximum achievable data rates and event rates as well as processing capabilities during Run 1 and Run 2.

Based on this analysis, we present new optimization we have developed for ALICE in Run 2. These include support for asynchronous transport via Zero-MQ which enables loops in the reconstruction chain graph and which is used to ship QA histograms to DQM.

We have added asynchronous processing capabilities in order to support long-running tasks besides the event-synchronous reconstruction tasks in normal HLT operation.
These asynchronous components run in an isolated process such that the HLT as a whole is resilient even to fatal errors in these asynchronous components. In this way, we can ensure that new developments cannot break data taking. On top of that, we have tuned the processing chain to cope with the higher event and data rates expected from the new TPC readout electronics (RCU2) and we have improved the configuration procedure and the startup time in order to increase the time where ALICE can take physics data. We present an analysis of the maximum achievable data processing rates taking into account processing capabilities of CPUs and GPUs, buffer sizes, network bandwidth, the incoming links from the detectors, and the outgoing links to data acquisition.

Tertiary Keyword (Optional):
Distributed workload management

Secondary Keyword (Optional):
High performance computing

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

The INDIGO-DataCloud Authentication and Authorisation Infrastructure

Authors: Andrea Ceccanti1; Bas Wegh2; Benjamin Ertl3; Enrico Vianello1; Marco Caberletti1; Marcus Hardt4; Paul Millar5; Slávek Licehammer6

1 INFN-CNAF
2 Karlsruhe Institute of Technology
3 Karlsruhe Institute of Technology
4 Karlsruhe Institute of Technology
5 DESY
6 CESNET

Corresponding Authors: marco.caberletti@cnaf.infn.it, enrico.vianello@cnaf.infn.it, slavek@ics.muni.cz, andrea.ceccanti@cnaf.infn.it, bas.wegh@kit.edu, hardt@kit.edu, paul.millar@desy.de, benjamin.ertl@kit.edu

Contemporary distributed computing infrastructures (DCIs) are not easily and securely accessible by common users. Computing environments are typically hard to integrate due to interoperability problems resulting from the use of different authentication mechanisms, identity negotiation protocols and access control policies. Such limitations have a big impact on the user experience making it hard for user communities to port and run their scientific applications on resources aggregated from multiple providers in different organisational and national domains. INDIGO-DataCloud will provide the services and tools needed to enable a secure composition of resources from multiple providers in support of scientific applications. In order to do so, an AAI architecture has to be defined that satisfies the following requirements:

- Is not bound to a single authentication mechanism, and can leverage federated authentication mechanisms
- Provides a layer where identities coming from different sources can be managed in a uniform way
- Defines how attributes linked to these identities are represented and understood by services
- Defines how controlled delegation of privileges across a chain of services can be implemented
• Defines how consistent authorization across heterogeneous services can be achieved and provides the tools to define, propagate, compose and enforce authorization policies

• Is mainly targeted at HTTP services, but can accommodate also non-HTTP services, leveraging token translation

In this contribution, we will present the work done in the first year of the INDIGO project to address the above challenges. In particular, we will introduce the INDIGO AAI architecture, its main components and their status and demonstrate how authentication, delegation and authorisation flows are implemented across services. More precisely, we will describe:

• The INDIGO Identity and Access Management (IAM) service, a central service responsible for user authentication and authorisation, supporting multiple authentication mechanisms (SAML, X.509, OpenID-connect);

• How identity information is exposed to services leveraging the OpenID-Connect standard;

• How we leverage and extend OAuth2 to support flexible delegation of privileges across services;

• How account linking and identity harmonisation is implemented centrally and at site-level;

• How identity provisioning and de-provisioning is implemented leveraging the System for Cross-Domain Identity Management Standard (SCIM);

• How Non-HTTP services, or services requiring different credentials/authentication flows, are integrated with the INDIGO AAI via the INDIGO Token Translation Service;

• How the INDIGO AAI relates and integrates with existing production infrastructures and with architectures being developed in the context of other projects in support of scientific research (e.g., the AARC and EGI-Engage projects)

Secondary Keyword (Optional): Computing middleware

Primary Keyword (Mandatory): Security and policies

Tertiary Keyword (Optional):

Track 5: Software Development / 393

Accelerating Navigation in the VecGeom Geometry Modeller

Authors: Sandro Christian Wenzel1; Yang Zhang2

1 CERN
2 KIT - Karlsruhe Institute of Technology (DE)

Corresponding Authors: sandro.wenzel@cern.ch, yang.zhang@cern.ch

The VecGeom geometry library is a relatively recent effort aiming to provide a modern and high performance geometry service for particle-detector simulation in hierarchical detector geometries common to HEP experiments.

One of its principal targets is the effective use of vector SIMD hardware instructions to accelerate geometry calculations for single-track as well as multiple-track queries. Previously, excellent performance improvements compared to Geant4/ROOT could be reported for elementary geometry algorithms at the level of single shape queries.

In this contribution, we will focus on the higher level navigation algorithms in VecGeom, which are the most important components as seen from the simulation engines.
In this contribution, we will first report on our R&D effort and developments to implement SIMD enhanced data structures to speed up the well-known “voxelized” navigation algorithms, ubiquitously used for particle tracing in complex detector modules consisting of many daughter parts.

Second, we will discuss complementary new approaches to improve navigation algorithms in HEP. These ideas are based on a systematic exploitation of static properties of the detector layout as well as automatic code generation and specialization of the C++ navigator classes. Such specializations reduce the overhead of generic- or virtual function based algorithms and enhance the effectiveness of the SIMD vector units.

These novel approaches go well beyond the existing solutions available in Geant4 or TGeo/ROOT, achieve a significantly superior performance, and might be of interest for a wide range of simulation backends (Geant-V, VMC, Geant4). We exemplify this with concrete benchmarks for the CMS and ALICE detectors.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Simulation

Primary Keyword (Mandatory):
High performance computing

Track 1: Online Computing / 394

Fast GPU Nearest Neighbors search algorithms for the CMS experiment at LHC

Author: Alessandro Degano\(^1\)

\(^1\) Universita e INFN Torino (IT)

Corresponding Authors: alessandro.degano@cern.ch, felice.pantaleo@cern.ch

The increase in instantaneous luminosity, number of interactions per bunch crossing and detector granularity will pose an interesting challenge for the event reconstruction and the High Level Trigger system in the CMS experiment at the High Luminosity LHC (HL-LHC), as the amount of information to be handled will increase by 2 orders of magnitude. In order to reconstruct the Calorimetric clusters for a given event detected by CMS it is necessary to search for all the “hits” in a given volume inside the Calorimeter. In particular, the forward regions of the Electromagnetic Calorimeter (ECAL) will be substituted by an innovative tracking calorimeter, the High Granularity Calorimeter (HGCAL) equipped with 6.8x10^6 readout channels. Online reconstruction of the large events expected at HL-LHC require the development of novel, highly parallel reduction algorithms. In this work, we present algorithms that, leveraging the computational power of a Graphical Processor Unit (GPU), are able to perform a Nearest-Neighbors search with timing performances compatible with the constraints imposed by the Phase 2 conditions. We will describe the process through which the sequential and parallel algorithms have been refined to achieve the best performance to cope with the given task. In particular, we will motivate the engineering decisions implemented in the highly-parallelized GPU-specific code, and report how the knowledge acquired in its development allowed to improve the benchmarks of the sequential CPU code. The final performance of the Nearest Neighbors search in 3x10^5 points randomly generated following a uniform distribution is 850 ms for the sequential CPU algorithm (on an Intel i7-3770) and 41 ms for the GPU parallel algorithm (on a Nvidia Tesla K40c), resulting in an average speedup of ~20. The results on different hardware testbeds are also presented along with consideration on the power requirement.
Realtime processing of LOFAR data for the detection of particles with the Moon

**Author:** Tobias Winchen

**Co-authors:** Anna Nelles; Antonio Bonardi; Arthur Corstanje; Gia Trinh; Heino Falcke; J.E. Enriquez; Jörg Hörandel; Jörg Paul Rachen; Katie Mulrey; Laura Rossetto; Olaf Scholten; Pim Schellart; Pragati Mitra; Sander ter Veen; Satyendra Thoudam; Stijn Buitink

1 Vrije Universiteit Brussel  
2 University of California Irvine  
3 IMAPP - Radboud University Nijmegen  
4 Radboud University Nijmegen  
5 KVI-CART, University of Groningen  
6 Ru Nijmegen/Nikhef  
7 IMAPP / Radboud University Nijmegen  
8 KVI-CART, Univ. of Groningen  
9 R  
10 Vrije Universiteit Brussels  
11 ASTRON  
12 Radboud University  
13 Vrije Universiteit Brussel (VUB)

Corresponding Authors: a.corstanje@astro.ru.nl, a.bonardi@astro.ru.nl, j.horandel@astro.ru.nl, kmulrey@gmail.com, s.buitink@gmail.com, t.n.g.trinh@rug.nl, p.schellart@astro.ru.nl, veen@astron.nl, anna.nelles@gmail.com, hflacke@astro.ru.nl, lrossetto@astro.ru.nl, j.rachen@astro.ru.nl, pmitra@vub.ac.be, tobias.winchen@rwth-aachen.de, j.enriquez@astro.ru.nl, s.thoudam@astro.ru.nl, scholten@kvi.nl

The low flux of the ultra-high energy cosmic rays (UHECR) at the highest energies provides a challenge to answer the long standing question about their origin and nature. Even lower fluxes of neutrinos with energies above $10^{22}$ eV are predicted in certain Grand-Unifying-Theories (GUTs) and e.g. models for super-heavy dark matter (SHDM). The significant increase in detector volume required to detect these particles can be achieved by searching for the nano-second radio pulses that are emitted when a particle interacts in Earth’s moon with current and future radio telescopes.

In this contribution we present the design of an online analysis and trigger pipeline for the detection of nano-second pulses with the LOFAR radio telescope. The most important steps of the processing pipeline are digital focusing of the antennas towards the Moon, correction of the signal for ionospheric dispersion, and synthesis of the time-domain signal from the polyphased-filtered signal in frequency domain. The implementation of the pipeline on a GPU/CPU cluster will be discussed together with the computing performance of the prototype.

Tertiary Keyword (Optional):
Data processing workflows and frameworks/pipelines
Software and Experience with Managing Workflows for the Computing Operation of the CMS Experiment

Author: Jean-Roch Vlimant

1 California Institute of Technology (US)

Corresponding Author: jean-roch.vlimant@cern.ch

We present a system deployed in the summer of 2015 for the automatic assignment of production and reprocessing workflows for simulation and detector data in the frame of the Computing Operation of the CMS experiment at the CERN LHC. Processing requests involves a number of steps in the daily operation, including transferring input datasets where relevant and monitoring them, assigning work to computing resources available on the CMS grid, and delivering the output to the Physics groups. Automatization is critical above a certain number of requests to be handled, especially in the view of using more efficiently computing resources and reducing latencies. An effort to automatize the necessary steps for production and reprocessing recently started and a new system to handle workflows has been developed. The state-machine system described consists in a set of modules whose key feature is the automatic placement of input datasets, balancing the load across multiple sites. By reducing the operation overhead, these agents enable the utilization of more than double the amount of resources with robust storage system. Additional functionalities were added after months of successful operation to further balance the load on the computing system using remote read and additional resources. This system contributed to reducing the delivery time of datasets, a crucial aspect to the analysis of CMS data. We report on lessons learned from operation towards increased efficiency in using a largely heterogeneous distributed system of computing, storage and network elements.

Secondary Keyword (Optional):
Distributed workload management

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

Tertiary Keyword (Optional):

The HNSciCloud project

Author: Helge Meinhard

1 CERN

Corresponding Author: helge.meinhard@cern.ch
HEP is only one of many sciences with sharply increasing compute requirements that cannot be met by profiting from Moore’s law alone. Commercial clouds potentially allow for realising larger economies of scale. While some small-scale experience requiring dedicated effort has been collected, public cloud resources have not been integrated yet with the standard workflows of science organisations in their private data centres; in addition, European science has not ramped up to significant scale yet. The HELIX NEBULA Science Cloud project, partly funded by the European Commission, addresses these points. Ten organisations under CERN’s leadership, covering particle physics, bioinformatics, photon science and other sciences, have joined to procure public cloud resources as well as dedicated development efforts towards this integration. The contribution will give an overview of the project, explain the findings so far, and provide an outlook into the future.

### Tertiary Keyword (Optional):
Experience/plans from outside experimental HEP/NP

### Secondary Keyword (Optional):
Computing models

### Primary Keyword (Mandatory):
Cloud technologies

---

**Resources monitoring and automatic management system for multi-VO distributed computing system**

**Authors:** Igor Pelevanyuk\(^1\); Jiong Chen\(^2\); Yong Sun\(^3\)

**Co-authors:** Alexey Zhemchugov\(^4\); Tian Yan\(^5\); Xianghu Zhao\(^3\); Xiaomei Zhang\(^6\)

\(^1\) Joint Inst. for Nuclear Research (RU)  
\(^2\) Suzhou University  
\(^3\) Institute of High Energy Physics  
\(^4\) Joint Inst. for Nuclear Research (JINR)  
\(^5\) Institution of High Energy Physics, Chinese Academy of Science  
\(^6\) Chinese Academy of Sciences (CN)

**Corresponding Authors:** chenjiong@ihep.ac.cn, zhaoxh@ihep.ac.cn, yant@ihep.ac.cn, igor.pelevanyuk@cern.ch, suny@suda.edu.cn, xiaomei.zhang@cern.ch, alexei.jemtchougov@cern.ch

Multi-VO supports based on DIRAC have been set up to provide workload and data management for several high energy experiments in IHEP. The distributed computing platform has 19 heterogeneous sites including Cluster, Grid and Cloud. The heterogeneous resources belong to different Virtual Organizations. Due to scale and heterogeneity, it is complicated to monitor and manage these resources manually. Moreover, the experts who have a rich knowledge about the underlying system are precious. For the above reasons, the requirement of an easy-to-use monitoring system to monitor and manage these resources accurately and effectively is proposed. The system should take all the heterogeneous resources into account and be suitable for multi-VO. Adopting the idea from Resource Status System (RSS) of DIRAC, this paper will present the designs and implementation of resources monitoring and automatic management system.

The system is composed of three parts: information collection, status decision and automatic control, and information display. The information collection includes active and passive way of gathering status from different sources and stores them in databases. For passive information, the system got information from third party systems periodically, such as storage occupancies, user job efficiency. For active collecting, periodical testing services have been designed and developed to send standard jobs to all sites to know their availability and status. These tests are well defined and classified for each VO according to their special requirements. The status decision and automatic control is used to evaluate the resources status and take control actions on resources automatically. Policies have
been pre-defined to set rules to judge the status in different situations. Combined with collected information and policies, the decision can be made and the appropriate actions will be automatically taken to send out alarm and give controls. A web portal has been designed to display both monitoring and control information. A summary page gives a quick view of all sites status and the detail information can be obtained by tracking down from the top. Besides the real-time information, the historical information is also recorded and displayed give a global view of resources status for certain period of time. All the implementations are based on DIRAC framework. The information and control including sites, policies, web portal for different VOs can be well defined and distinguished within DIRAC user and group management infrastructure.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):
Monitoring

Track 3: Distributed Computing / 399

Technical challenges of HNSciCloud

Author: Martin Gasthuber¹

¹ DESY

Corresponding Author: martin.gasthuber@cern.ch

The HNSciCloud project (presented in general by another contribution) faces the challenge to accelerate developments performed by the selected commercial providers. In order to guarantee cost-efficient usage of IaaS resources across a wide range of scientific communities, the technical requirements had to be carefully constructed. With respect to current IaaS offerings, data-intensive science is the biggest challenge; other points that need to be addressed concern identity federations, network connectivity and how to match business practices of large IaaS providers with those of public research organisations. This contribution will explain the key points of the technical requirements and present first results of the experience of the procurers with the services in comparison to their 'on-premise' infrastructure.

Tertiary Keyword (Optional):
Experience/plans from outside experimental HEP/NP

Secondary Keyword (Optional):
Computing models

Primary Keyword (Mandatory):
Cloud technologies

Track 3: Distributed Computing / 400

On-demand provisioning of HEP compute resources on cloud sites and shared HPC centers
Authors: Frank Fischer¹; Georg Stefan Fleig¹; Günter Quast²; Günther Erli²; Manuel Giffels¹; Matthias Jochen Schnepf¹; Thomas Hauth³

Co-authors: Javier Arnáez de Pedro ³; Jörg Heese ⁴; Katja Leppert ¹; Rainer Sträter ⁴

¹ KIT - Karlsruhe Institute of Technology (DE)
² KIT - Karlsruhe Institute of Technology (DE)
³ 1&1 Interent SE
⁴ 1&1 Internet SE

Corresponding Authors: gunter.quast@cern.ch, jarnaez@arsys.es, matthias.jochen.schnepf@cern.ch, katja.leppert@1und1.de, rainer.straeter@1und1.de, joerg.heese@1und1.de, georg.fleig@cern.ch, frank.fischer@student.kit.edu, guenther.erli@student.kit.edu, thomas.hauth@kit.edu, manuel.giffels@cern.ch

This contribution reports on solutions, experiences and recent developments with the dynamic, on-demand provisioning of remote computing resources for analysis and simulation workflows. Local resources of a physics institute are extended by private and commercial cloud sites, ranging from the inclusion of desktop clusters over institute clusters to HPC centers.

Rather than relying on dedicated HEP computing centers, it is nowadays more reasonable and flexible to utilize remote computing capacity via virtualization techniques or container concepts.

We report on recent experience from incorporating a remote HPC center (NEMO Cluster, Freiburg University) and resources dynamically requested from a commercial provider (1&1 Internet SE), which have been seamlessly tied together with the ROCED scheduler ¹ such that, from the user perspective, local and remote resources form a uniform, virtual computing cluster with a single point-of-entry. On a local test system, the usage of Docker containers has been explored and shown to be a viable and light-weight alternative to full virtualization solutions in trusted environments.

The Freiburg HPC resources are requested via the standard batch system, allowing HPC and HEP applications to be executed simultaneously, such that regular batch jobs run side by side to virtual machines managed via OpenStack. For the inclusion of the 1&1 commercial resources, a Python API and SDK as well as the possibility to upload images were available. Large scale tests prove the capability to serve the scientific use case in the European 1&1 datacenters.

The described environment at the Institut für Experimentelle Kernphysik (IEKP) at KIT serves the needs of researchers participating in the CMS and Belle II experiments. In total, resources exceeding half a million CPU hours have been provided by remote sites.


Secondary Keyword (Optional):
Virtualization

Primary Keyword (Mandatory):
Cloud technologies

Tertiary Keyword (Optional):

Posters B / Break / 401

How Pre-Commercial Procurement can fund innovative cloud services

Author: Helge Meinhard¹

¹ CERN
The HELIX NEBULA Science Cloud (HNSciCloud) project (presented in general by another contribution) is run by a consortium of ten procurers and two other partners; it is funded partly by the European Commission, has a total volume of 5.5 MEUR and runs from January 2016 to June 2018. By its nature as a pre-commercial procurement (PCP) project, it addresses needs that are not covered by any commercially available solution yet. The contribution will explain the steps, including administrative and legal ones, needed to establish and conduct the project, and will describe the procurers’ experience with the project.

Migration Routes to Virtual Technologies

Author: Stephen Jones

Co-authors: John Bland ; Robert Fay

1 Liverpool University
2 University of Liverpool

Corresponding Authors: sjones@hep.ph.liv.ac.uk, fay@hep.ph.liv.ac.uk, jbland@hep.ph.liv.ac.uk

Traditional T2 grid sites still process large amounts of data flowing from the LHC and elsewhere. More flexible technologies, such as virtualisation and containerisation, are rapidly changing the landscape, but the right migration paths to these sunlit uplands are not well defined yet. We report on the innovations and pressures that are driving these changes and we discuss their pros and cons. We specifically examine a recently developed migration route to virtual technology that is currently available to sites called VAC. We installed and tested VAC on a production class cluster and we ran it with a set of VOs for a period of months. We report our test findings and conclude that VAC is suitable for large scale deployment.

Track 3: Distributed Computing
Stability and scalability of the CMS Global Pool: Pushing HTCondor and glideinWMS to new limits

Author: James Letts

Co-authors: Anthony Tiradani; Antonio Perez-Calero Yzquierdo; Brian Paul Bockelman; David Alexander Mason; Dirk Hufnagel; Farrukh Aftab Khan; Jadir Marra Da Silva; Justas Balcas; Kenyi Paolo Hurtado Anampa; Krista Larson; Marco Mascheroni; Vassil Verguilov

1 Univ. of California San Diego (US)
2 Fermilab
3 Centro de Investigaciones Energ. Medioambientales y Tecn. (ES)
4 University of Nebraska (US)
5 Fermi National Accelerator Lab. (US)
6 National Centre for Physics (PK)
7 UNESP - Universidade Estadual Paulista (BR)
8 California Institute of Technology (US)
9 University of Notre Dame (US)
10 Bulgarian Academy of Sciences (BG)

Corresponding Authors: tiradani@fnal.gov, marco.mascheroni@cern.ch, brian.bockelman@cern.ch, farrukh.aftab.khan@cern.ch, dirk.hufnagel@cern.ch, kenyi.paolo.hurtado.anampa@cern.ch, jadir.marra.da.silva@cern.ch, vassil.verguilov@cern.ch, dmason@fnal.gov, krista.larson@cern.ch, justas.balcas@cern.ch, letts@physics.ucsd.edu, antonio.perez.calero.yzquierdo@cern.ch

The CMS Global Pool, based on HTCondor and glideinWMS, is the main computing resource provisioning system for all CMS workflows, including analysis, Monte Carlo production, and detector data reprocessing activities. Total resources at Tier-1 and Tier-2 sites pledged to CMS exceed 100,000 CPU cores, and another 50,000-100,000 CPU cores are available opportunistically, pushing the needs of the Global Pool to higher scales each year. These resources are becoming more diverse in their accessibility and configuration over time. Furthermore, the challenge of stably running at higher and higher scales while introducing new modes of operation such as multi-core pilots, as well as the chaotic nature of physics analysis workflows, place huge strains on the submission infrastructure. This paper details some of the most important challenges to scalability and stability that the Global Pool has faced since the beginning of the LHC Run II and how they were overcome.

Tertiary Keyword (Optional):
Distributed workload management

Secondary Keyword (Optional):
Computing middleware

Primary Keyword (Mandatory):
Computing facilities

Track 1: Online Computing / 404

Track Finding in CMS for the Level-1 Trigger at the HL-LHC

Authors: Anders Ryd; Fabrizio Palla; Ian Tomalin; Kristian Hahn; Mark Pesaresi; Peter Wittich; Tiehui Ted Liu

1 Cornell University (US)
2 Universita di Pisa & INFN (IT)
3 STFC - Rutherford Appleton Lab. (GB)
The High Luminosity LHC (HL-LHC) will deliver luminosities of up to 5x10^{34} cm^{-2}/s, with an average of about 140-200 overlapping proton-proton collisions per bunch crossing. These extreme pileup conditions can significantly degrade the ability of trigger systems to cope with the resulting event rates. A key component of the HL-LHC upgrade of the CMS experiment is a Level-1 (L1) track finding system that will identify tracks with transverse momentum above 3 GeV within \sim 5 \mu s. Output tracks will be merged with information from other sub-detectors in the downstream L1 trigger to improve the identification and resolution of physics objects. The CMS collaboration is exploring several designs for a L1 tracking system that can confront the challenging latency, occupancy and bandwidth requirements associated with L1 tracking. This presentation will review the three state-of-the-art L1 tracking architectures proposed for the CMS HL-LHC upgrade. Two of these architectures ("Tracklet" and "TMT") are fully FPGA-based, while a third ("AM+FPGA") employs a combination of FPAGs and ASICs. The FPGA-based approaches employ a road-search algorithm ("Tracklet") or a Hough transform ("TMT"), while the AM+FPGA approach uses content-addressable memories for pattern recognition. Each approach aims to perform the demanding data distribution, pattern recognition, track reconstruction tasks required of L1 tracking in real-time.

**Primary Keyword (Mandatory):** Trigger

**Secondary Keyword (Optional):**

**Tertiary Keyword (Optional):**

---

**Posters A / Break / 406**

**Integration of the SCD Cloud and Aquilon**

**Author:** Alexander Dibbo

1 **STFC RAL**

**Corresponding Author:** alexander.dibbo@stfc.ac.uk

The Scientific Computing Department of the STFC runs a cloud service for internal users and various user communities. The SCD Cloud is configured using a Configuration Management System called Aquilon. Many of the virtual machine images are also created/configured using Aquilon. These are not unusual however our Integrations also allow Aquilon to be altered by the Cloud. For instance creation or destruction of a Virtual Machine can affect its configuration in Aquilon.

**Tertiary Keyword (Optional):** Virtualization

**Secondary Keyword (Optional):** Computing facilities

**Primary Keyword (Mandatory):** Cloud technologies
Scheduling beams at CERN, the new AD central timing

Author: Andrzej Dworak

Co-author: Jean-Claude Bau

A central timing (CT) is a dedicated system responsible for driving an accelerator behaviour. It allows operation teams to interactively select and schedule cycles. While executing a scheduled cycle a CT sends out events which (a) provide precise synchronization and (b) information what to do - to all equipment operating an accelerator. The events are also used to synchronize accelerators between each other, which allows passing of a beam between them.

At CERN there are currently ten important accelerators. Each of them is different and has some unique functionalities. To support the variety and not to constrain operation teams there are three major types of the CT systems. The one which has been developed most recently handles the Antiproton Decelerator (AD). Uniqueness of the AD machine comes from the fact that it works with antimatter and instead of accelerating particles it decelerates them. As a result, an AD cycle differs from other machines and required development of a new CT.

In this paper we describe the differences and systems which has been developed to support unique AD requirements. In particular, a new AD CT is presented and functionality it offers to operation teams who program the machine. We present also the central timing extensions planned to support a new decelerator ELENA, which will be connected to AD to further slow down the beam. We show that with these extensions the new central timing becomes a very generic system. Generic to a point where it is valid to ask a question if it could be used as a common solution for all CERN accelerators.

Connecting restricted, high-availability, or low-latency resources to a seamless Global Pool for CMS

Author: James Letts

Co-authors: Anthony Tiradani; Antonio Perez-Calero Yzquierdo; Bo Jayatilaka; Brian Paul Bockelman; David Alexander Mason; Dirk Hufnagel; Farrukh Aftab Khan; Jadir Marra Da Silva; Justas Balcas; Kenyi Paolo Hurtado Anampa; Krista Larson; Marco Mascheroni; Stefan Piperov; Vassil Verguilov

1 Univ. of California San Diego (US)
2 Fermilab
3 Centro de Investigaciones Energ. Medioambientales y Tecn. - (ES)
4 Fermi National Accelerator Lab. (US)
The connection of diverse and sometimes non-Grid enabled resource types to the CMS Global Pool, which is based on HTCondor and glideinWMS, has been a major goal of CMS. These resources range in type from a high-availability, low latency facility at CERN for urgent calibration studies, called the CAF, to a local user facility at the Fermilab LPC, allocation-based computing resources at NERSC and SDSC, opportunistic resources provided through the Open Science Grid, commercial clouds, and others, as well as access to opportunistic cycles on the CMS High Level Trigger farm. In addition, we have provided the capability to give priority to local users of beyond WLCG pledged resources at CMS sites. Many of the solutions employed to bring these diverse resource types into the Global Pool have common elements, while some are very specific to a particular project. This paper details some of the strategies and solutions used to access these resources through the Global Pool in a seamless manner.

Tertiary Keyword (Optional): Distributed workload management

Secondary Keyword (Optional): Computing middleware

Primary Keyword (Mandatory): Computing facilities

Operational and other pressures have lead to WLCG experiments moving increasingly to a stratified model for Tier-2 resources, where “fat” Tier-2s (“T2Ds”) and “thin” Tier-2s (“T2Cs”) provide different levels of service. In the UK, this distinction is also encouraged by the terms of the current GridPP5 funding model. In
anticipation of this, testing has been performed on the implications, and potential implementation, of such a distinction in our resources.

In particular, this presentation presents the results of testing of storage T2Cs, where the “thin” nature is expressed by the site having either no local data storage, or only a thin caching layer; data is streamed or copied from a “nearby” T2D when needed by jobs.

In OSG, this model has been adopted successfully for CMS AAA sites; but the network topology and capacity in the USA is significantly different to that in the UK (and much of Europe).

We present the result of several operational tests: the in-production University College London (UCL) site, which runs ATLAS workloads using storage at the Queen Mary University of London (QMUL) site; the Oxford site, which has had scaling tests performed against T2Ds in various locations in the UK (to test network effects); and the Durham site, which has been testing the specific ATLAS caching solution of “Rucio Cache” integration with ARC’s caching layer.

We provide suggestions and future implementation models from this data, along with the results of CMS Tier-3 use of the Glasgow and QMUL sites.

Tertiary Keyword (Optional):
Storage systems

Secondary Keyword (Optional):
Distributed data handling

Primary Keyword (Mandatory):
Data model

Track 9: Future Directions / 410

The high-rate data challenge: computing for the CBM experiment

Author: Volker Friese

1 GSI - Helmholtzzentrum fur Schwerionenforschung GmbH (DE)

Corresponding Author: volker.friese@cern.ch

The Compressed Baryonic Matter experiment (CBM) is a next-generation heavy-ion experiment to be operated at the FAIR facility, currently under construction in Darmstadt, Germany. A key feature of CBM are very high interaction rates, exceeding those of contemporary nuclear collision experiments by several orders of magnitude. Such interaction rates forbid a conventional, hardware-triggered readout; instead, experiment data will be freely streaming from self-triggered frontend electronics.

In order to reduce the huge raw data volume to a recordable rate, data will be selected exclusively on CPU, which necessitates partial event reconstruction in real-time. Consequently, the traditional segregation of online and offline software vanishes: an integrated on- and offline data processing concept is called for. In this paper, we will report on concepts and developments for computing for CBM as well as on the status of preparations for its first physics run.

Tertiary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Secondary Keyword (Optional):
High performance computing

Primary Keyword (Mandatory):
Computing models
Index files for Belle II - very small skim containers

Author: Martin Sevior

Co-authors: Chia-Ling Hsu; Hideki Miyake; Ikuo Ueda; Takanori Hara; Thomas Kuhr; Tristan Bloomfield

1 University of Melbourne
2 High Energy Accelerator Research Organization, Tsukuba, Japan
3 LMU, Munich, Germany

Corresponding Authors: chialingh@student.unimelb.edu.au, thomas.kuhr@cern.ch, t.bloomfield@student.unimelb.edu.au, i.ueda@cern.ch, takanori.hara@kek.jp, hideki.miyake@gmail.com, martines@unimelb.edu.au

The Belle II experiment will generate very large data samples. In order to reduce the time for data analyses, loose selection criteria will be used to create files rich in samples of particular interest for a specific data analysis (data skims). Even so, many of the resultant skims will be very large, particularly for highly inclusive analyses. The Belle II collaboration is investigating the use of “index-files” where instead of the skim recording the actual physics data, we record pointers to events of interest in the complete data sample. This reduces the skim files by 2 orders of magnitude. While this approach was successfully employed by the Belle experiment where the majority of the data analysis was performed on a single center cluster, index files are significantly more challenging in a distributed computing system such as the Belle II computing grid.

The approach we use is for each skim file to record metadata identifying the original parent file, as well as the location of the event within the parent file. Since we employ the ROOT file container system it is straightforward to read just the identified events of interest. For remote file access, we employ the XROOTD protocol to both select identified events and transfer them to the worker nodes used for data analysis.

This presentation will describe the details of the implementation of index-files within the Belle II grid. We will also describe the results of tests where the analyses were performed locally and situations where data was transferred internationally, in particular, from Europe to Australia. The configuration of the XROOTD services for these tests has a very large impact on the performance of the system and will be reported in the presentation.

Tertiary Keyword (Optional):
Distributed workload management

Secondary Keyword (Optional):
Computing middleware

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines
During the past years an increasing number of CMS computing resources are offered as clouds, bringing the flexibility of having virtualised compute resources and centralised management of the Virtual Machines (VMs). CMS has adapted its job submission infrastructure from a traditional Grid site to operation using a cloud service and meanwhile can run all types of offline workflows. The cloud service provided by the online cluster for the Data Acquisition (DAQ) and High Level Trigger (HLT) of the experiment was one of the first facilities to commission and deploy this submission infrastructure. The CMS HLT is a considerable compute resource. It consists currently of approximately 1000 dual socket PC server nodes with a total of ~25 k cores, corresponding to ~500 kHEPSpec06. This compares to a total Tier-0 / Tier-1 CMS resources request of 292 / 461 kHEPSpec06. The HLT has no local mass disk storage and is currently connected to the CERN IT datacenter via a dedicated 160 Gbps network connection.

One of the main requirements for the online cloud facility is the parasitic use of HLT which shall never interfere with its primary function as part of the data acquisition system. Hence a design has been chosen where an Openstack infrastructure is overlaid over the HLT hardware resources. This overlay also abstracts the different hardware and networks that the cluster is composed of. The online cloud is meanwhile a well established facility to substantially augment the CMS computing resources when the HLT is not needed for data acquisition, such as during technical stop periods of the LHC. In this static mode of operation, this facility acts as any other Tier-0 or Tier-1 facility. During high workload periods it provided up to ~40% of the combined Tier-0/Tier-1 capacity, including workflows with demanding I/O requirements. Data needed by the running jobs was read from the remote EOS disk system at CERN and data produced was written back out to EOS. The achieved throughput from the remote EOS came close to the installed bandwidth of the 4x40 Gbps long range links.

The next step is to extend the usage of the online cloud to the opportunistic usage of the periods between LHC fills. These periods are a-priori unscheduled and of undetermined length, typically at least 5 hours and once or more a day. This mode of operation of a dynamic usage of the cloud infrastructure requires a fast turn-around for the starting and stopping of the VMs. A more advanced mode of operation where the VMs are hibernated and jobs are not killed is also being explored. Finally, one could envisage to ramp up VMs while the load on the HLT reduces towards the end of the fill. We will discuss the optimisation of the cloud infrastructure for the dynamic operation and the design and implementation of the mechanism in the DAQ system to gracefully switch from DAQ mode to providing cloud resources based on LHC state or server load.

Secondary Keyword (Optional):

DAQ

Primary Keyword (Mandatory):

Cloud technologies

Tertiary Keyword (Optional):

Machine Learning with TensorFlow as an alternative to TMVA

Author: Martin Sevior

Co-authors: Aidan Dang; Anton Hawthorne

1 University of Melbourne
The Toolkit for Multivariate Analysis (TMVA) is a component of the ROOT data analysis framework and is widely used for classification problems. For example, TMVA might be used for the binary classification problem of distinguishing signal from background events.

The classification methods included in TMVA are standard, well-known machine learning techniques which can be implemented in other languages and hardware architectures. The recently released open source package “TensorFlow” from Google, offers the opportunity to test an alternative implementation. In particular, TensorFlow has the capability to transparently interface GPU acceleration for machine learning with the potential for orders of magnitude increase in performance. Furthermore, TensorFlow enables the construction of sophisticated artificial neural networks capable of “Deep Learning”. We have investigated the use of TensorFlow for general purpose machine learning applications in Particle Physics by interfacing it to the root data format and implementing the TMVA API within the TensorFlow framework. In addition, we have investigated recurrent neural networks (RNN) using TensorFlow.

The presentation will report the performance of TensorFlow compared to TMVA for both general purpose CPUs and a high-performance GPU cluster. We will also report on the effectiveness of RNN for particle physics applications.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):
Analysis tools and techniques
ultimately the data availability for physics analysis. In order to cope with the production peaks, the LHC experiments are exploring the opportunity to access Cloud resources provided by external partners or commercial providers.

In this work we present the proof of concept of the elastic extension of a local analysis facility, specifically the Bologna Tier-3 Grid site, for the LHC experiments hosted at the site, on an external OpenStack infrastructure. We focus on the “Cloud Bursting” of the Grid site using DynFarm, a newly designed tool that allows the dynamic registration of new worker nodes to LSF. In this approach, the dynamically added worker nodes instantiated on the OpenStack infrastructure are transparently accessed by the LHC Grid tools and at the same time they serve as an extension of the farm for the local usage.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Virtualization

Primary Keyword (Mandatory):
Cloud technologies

Track 2: Offline Computing / 415

A precision device needs precise simulation: Software description of the CBM Silicon Tracking System

Author: Hanna Malygina
Co-author: Volker Friese

GSI Darmstadt

Precise modelling of detectors in simulations is the key to the understanding of their performance, which, in turn, is a prerequisite for the proper design choice and, later, for the achievement of valid physics results. In this report, we describe the implementation of the Silicon Tracking System (STS), the main tracking device of the CBM experiment, in the CBM software environment. The STS makes use of double-sided silicon micro-strip sensors with double metal layers. We present a description of transport and detector response simulation, including all relevant physical effects like charge creation and drift, charge collection, cross-talk and digitization. Of particular importance and novelty is the description of the time behaviour of the detector, since its readout will not be externally triggered but continuous. We also cover some aspects of local reconstruction, which in the CBM case has to be performed in real-time and thus requires high-speed algorithms.

Tertiary Keyword (Optional):
Algorithms

Secondary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Primary Keyword (Mandatory):
Simulation
The LHCb Starterkit

Author: Albert Puig Navarro

Co-author: Marco Clemencic

1 Ecole Polytechnique Federale de Lausanne (CH)

2 CERN

Corresponding Authors: albert.puig@epfl.ch, marco.clemencic@cern.ch

The vast majority of high-energy physicists use and produce software every day. Software skills are usually acquired “on the go” and dedicated training courses are rare. The LHCb Starterkit is a new training format for getting LHCb collaborators started in effectively using software to perform their research. The course focuses on teaching basic skills for research computing. Unlike traditional tutorials we focus on starting with basics, performing all the material live, with a high degree of interactivity, giving priority to understanding the tools as opposed to handing out recipes that work “as if by magic”. The LHCb Starterkit was started by two young members of the collaboration inspired by the principles of Software Carpentry (http://software-carpentry.org), and the material is created in a collaborative fashion using the tools we teach. Three successful entry-level workshops, as well as an advance one, have taken place since the start of the initiative in 2015, and were taught largely by PhD students to other PhD students.

Tertiary Keyword (Optional):

Analysis tools and techniques

Secondary Keyword (Optional):

Software development process and tools

Primary Keyword (Mandatory):

Collaborative tools
The Open Science Grid (OSG) relies upon the network as a critical part of the distributed infrastructures it enables. In 2012 OSG added a new focus area in networking with a goal of becoming the primary source of network information for its members and collaborators. This includes gathering, organizing and providing network metrics to guarantee effective network usage and prompt detection and resolution of any network issues, including connection failures, congestion and traffic routing.

In September of 2015 this service was deployed into the OSG production environment. We will report on the creation, implementation, testing and deployment of the OSG Networking Service. Starting from organizing the deployment of perfSONAR toolkits within OSG and its partners, to the challenges of orchestrating regular testing between sites, to reliably gathering the resulting network metrics and making them available for users, virtual organizations and higher level services all aspects of implementation will be reviewed. In particular, several higher level services were developed to bring the OSG network service to its full potential. These include a web-based mesh configuration system, which allows central scheduling and management all the network tests performed by the instances, a set of probes to continually gather metrics from the remote instances and publish it to different sources, a central network datastore (Esmond), which provides interfaces to access the network monitoring information in close to real time and historically (up to a year) giving the state of the tests and the perfSONAR infrastructure monitoring, ensuring the current perfSONAR instances are correctly configured and operating as intended.

We will also describe the challenges we encountered in ongoing operations for the network service and how we have evolved our procedures to address those challenges. Finally we will describe our plans for future extensions and improvements to the service.

Tertiary Keyword (Optional):
Monitoring

Secondary Keyword (Optional):
Network systems and solutions

Primary Keyword (Mandatory):
Collaborative tools

---

Finding the needle in the haystack: a charmonium trigger for the CBM experiment

**Author:** Timur Ablyazimov¹

**Co-authors:** Vikas Singhal ²; Volker Friese ³

1 J

2 Department of Atomic Energy (IN)

3 GSI - Helmholtzzentrum fur Schwerionenforschung GmbH (DE)

**Corresponding Authors:** abl@jinr.ru, vikas.singhal@cern.ch, volker.friese@cern.ch

Charmonium is one of the most interesting, yet most challenging observables for the CBM experiment. CBM will try to measure charmonium in the di-muon decay channel in heavy-ion collisions close to or even below the kinematic threshold for elementary interactions. The expected signal yield is consequently extremely low - less than one in a million collisions. CBM as a high-rate experiment shall be able to cope with this, provided a suitable software trigger can be implemented for online data selection. Since the latter will per performed exclusively on CPU, the performance of the algorithm is crucial for the maximal allowed interaction rate - and thus the sensitivity - and/or for the size of the CBM
online cluster FLES (First-Level Event Selector).
In this report we discuss the CBM charmonium trigger, its implementation on the FLES, and its
performance.

Tertiary Keyword (Optional):
Reconstruction

Secondary Keyword (Optional):
Algorithms

Primary Keyword (Mandatory):
Trigger

Posters B / Break / 420

STAR Data Production at NERSC/Cori, an adaptable Docker container approach for HPC

Authors: Jeff Porter¹; Jerome LAURET²; Mustafa Mustafa³

¹ Lawrence Berkeley National Lab. (US)
² Brookhaven National Laboratory
³ Lawrence Berkeley National Laboratory

Corresponding Authors: mmustafa@lbl.gov, jlauret@bnl.gov, rjporter@lbl.gov

The expected growth in HPC capacity over the next decade makes such resources attractive for meeting future computing needs of HEP/NP experiments, especially as their cost is becoming comparable to traditional clusters. However, HPC facilities rely on features like specialized operating systems and hardware to enhance performance that make them difficult to be used without significant changes to production workflows. Containerized software environment running on HPC systems may very well be an ideal scalable solution to leverage those resources and a promising candidate to replace the outgrown traditional solutions employed at different computing centers.

In this talk we report on the first test of STAR real-data production utilizing Docker containers at the Cori-I supercomputer at NERSC. Our test dataset was taken by the STAR experiment at RHIC in 2014 and is estimated to require ~30M CPU hours for full production. To ensure validity and reproducibility, STAR data production is restricted to a vetted computing environment defined by system architecture, Linux OS, compiler and external libraries versions. Furthermore, each data production task requires certain STAR software tag and database timestamp. In short, STAR’s data production workflow represents a typical embarrassingly parallel HEP/NP computing task. Thus, it is an ideal candidate to test the suitability of running containerized software, normalized to run on a shared HPC systems instead of its traditional dedicated off-the-shelf clusters. This direction, if successful, could very well address current and future experiments computing needs. We will report on the different opportunities and challenges of running in such an environment. We will also present the modifications needed to the workflow in order to optimize Cori resource utilization and streamline the process in this and future productions as well as performance metrics.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
High performance computing

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines
Extending the farm on external sites: the INFN Tier-1 experience

Authors: Andrea Chierici\textsuperscript{1}; Daniele Cesini\textsuperscript{2}; Luca Dell’Agnello\textsuperscript{None}; Stefano Dal Pra\textsuperscript{3}; Stefano Zani\textsuperscript{None}; Tommaso Boccali\textsuperscript{4}; Vladimir Sapunenko\textsuperscript{5}

\begin{itemize}
\item \textsuperscript{1} INFN-CNAF
\item \textsuperscript{2} Universita e INFN, Bologna (IT)
\item \textsuperscript{3} INFN
\item \textsuperscript{4} Universita di Pisa & INFN (IT)
\item \textsuperscript{5} INFN-CNAF (IT)
\end{itemize}

Corresponding Authors: andrea.chierici@cern.ch, stefano.dalpra@cnaf.infn.it, vladimir.sapunenko@cern.ch, stefano.zani@cnaf.infn.it, tommaso.boccali@cern.ch, daniele.cesini@cnaf.infn.it

The Tier-1 at CNAF is the main INFN computing facility offering computing and storage resources to more than 30 different scientific collaborations including the 4 experiments at the LHC. It is also foreseen a huge increase in computing needs in the following years mainly driven by the experiments at the LHC (especially starting with the run 3 from 2021) but also by other upcoming experiments such as CTA.

While we are considering the upgrade of the infrastructure of our data center, we are also evaluating the possibility of using CPU resources available in other data centers or even leased from commercial cloud providers.

Hence, at INFN Tier-1, besides participating to the EU project HNSciCloud, we have also pledged a small amount of computing resources (\textasciitilde2000 cores) located at the Bari ReCaS for the WLCG experiments for 2016 and we are testing the use of resources provided by a commercial cloud provider. While the Bari ReCaS data center is directly connected to the GARR network with the obvious advantage of a low latency and high bandwidth connection, in the case of the commercial provider we rely only on the General Purpose Network.

In this paper we describe the setup phase and the first results of these installations started in the last quarter of 2015, focusing on the issues that we have had to cope with and discussing the measured results in terms of efficiency.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Cloud technologies

Primary Keyword (Mandatory):
Computing facilities

Monitoring the LHCb data quality system

Authors: Alexander Baranov\textsuperscript{1}; Alexander Panin\textsuperscript{1}; Andrey Ustyuzhanin\textsuperscript{1}; Denis Derkach\textsuperscript{1}; Flavio Archilli\textsuperscript{2}; Marco Adinolfi\textsuperscript{2}; Wander Baldini\textsuperscript{1}

\begin{itemize}
\item \textsuperscript{1} Yandex School of Data Analysis (RU)
\item \textsuperscript{2} Nikhef National institute for subatomic physics (NL)
\end{itemize}
Monitoring the quality of the data, DQM, is crucial in a high-energy physics experiment to ensure the correct functioning of the apparatus during the data taking. DQM at LHCb is carried out in two phases. The first one is performed on-site, in real time, using unprocessed data directly from the LHCb detector, while the second, also performed on-site, requires the reconstruction of the data selected by the LHCb trigger system and occurs with some delay.

For the Run II data taking the LHCb collaboration has re-engineered the DQM protocols and the DQM graphical interface, moving the latter to a web-based monitoring system, called Monet, thus allowing researchers to perform the second phase off-site. In order to support the operator’s task, Monet is also equipped with an automated, fully configurable, alarm system, thus allowing its use not only for DQM purposes, but also to track and assess the quality of LHCb software and simulation.

Posters A / Break / 424

Achieving Cost/Performance Balance Ratio Using Tiered Storage Caching Techniques: A Case Study with CephFS

Authors: Jerome LAURET\(^1\); Michael Poat\(^1\)

\(^1\) Brookhaven National Laboratory

Corresponding Authors: mpoat@bnl.gov, jlauret@bnl.gov

As demand for widely accessible storage capacity increases and usage is on the rise, steady IO performance is desired but tends to suffer within multi-user environments. Typical deployments use standard hard drives as the cost per/GB is quite low. On the other hand, HDD based solutions for storage are not known to scale well with process concurrency and soon enough, high rate of IOPs create a “random access” pattern killing performance. Though not all SSDs are alike, SSDs are an established technology often used to address this exact “random access” problem. Whilst the cost per/GB of SSDs has decreased since inception, their costs are still significantly more than standard HDDs. A possible approach could be the use of a mixture of both HDDs and SSDs coupled with a caching mechanism between the two types of drives. With such approach, the most performant drive technology can be exposed to the application while the lower performing drives (in IOPs performance metric) used for storage capacity. Furthermore, least used files could be transparently migrated to the least performing storage in the background. With this agile concept, both low cost and performance may very well be achieved. Flashcache, dm-cache, and bcache represents a non-exhaustive list of low-level disk caching techniques that are designed to create such tiered storage infrastructure.

In this contribution, we will first discuss the IO performance of many different SSD drives (tested in a comparable and standalone manner). We will then be discussing the performance and integrity of at least three low-level disk caching techniques (Flashcache, dm-cache, and bcache) including
individual policies, procedures, and IO performance. Furthermore, the STAR online computing infrastructure currently hosts a POSIX-compliant Ceph distributed storage cluster - while caching is not a native feature of CephFS (but only exists in the Ceph Object store), we will show how one can implement a caching mechanism profiling from an implementation at a lower level. As our illustration, we will present our CephFS setup, IO performance tests, and overall experience from such configuration. We hope this work to service the community’s interest for using disk-caching mechanisms with applicable uses such as distributed storage systems and seeking an overall IO performance gain.

Tertiary Keyword (Optional):
Object stores

Secondary Keyword (Optional):
Distributed data handling

Primary Keyword (Mandatory):
Storage systems

Track 3: Distributed Computing / 425

Development of stable Grid service at the next generation system of KEKCC

Author: Tomoaki Nakamura

Co-authors: Go Iwai; Hiroyuki Matsunaga; Koichi Murakami; Soh Suzuki; Takashi Sasaki; Wataru Takase

1 High Energy Accelerator Research Organization (JP)
2 KEK

Corresponding Authors: takashi.sasaki@kek.jp, hiroyuki.matsunaga@kek.jp, soh.suzuki@kek.jp, koichi.murakami@kek.jp, wataru.takase@kek.jp, go.iwai@kek.jp, tomoaki.nakamura@cern.ch

A lot of experiments in the field of accelerator based science are actively running at High Energy Accelerator Research Organization (KEK) by using SuperKEKB and J-PARC accelerator in Japan. In these days at KEK, the computing demand from the various experiments for the data processing, analysis and MC simulation is monotonically increasing. It is not only for the case with high-energy experiments, the computing requirement from the hadron and neutrino experiments and some projects of astro-particle physics is also rapidly increasing due to the very high precision measurement. Under this situation, several projects, Belle II, T2K, ILC and KAGRA experiments supported by KEK are going to utilize Grid computing infrastructure as the main computing resource. The Grid system and services in KEK, which is already in production, are upgraded for the further stable operation at the same time of whole scale hardware replacement of KEK Central Computer (KEKCC). The next generation system of KEKCC start the operation from the beginning of September 2016. The basic Grid services e.g. BDII, VOMS, LFC, CREAM computing element and StorM storage element are made by the more robust hardware configuration. Since the raw data transfer is one of the most important tasks for the KEKCC, two redundant GridFTP servers are adapted to the StorM service instances with 40Gbps network bandwidth on the LHCONE routing. These are dedicated to the Belle II raw data transfer to the other sites apart from the servers for the data transfer usage of the other VOs. Additionally, we prepare the redundant configuration for the database oriented services like LFC and AMGA by using LifeKeeper. The LFC servers are made by two read/write servers and one read-only server for the Belle II experiment, and all of them have individual database for the purpose of load balancing. The FTS3 service is newly deployed as a service for the Belle II data distribution. The service of CVMFS startum-0 is started for the Belle II software repository, and startum-1 service is prepared for the other VOs. In this way, there are lot of upgrade for the real production service of Grid infrastructure at KEK Computing Research Center. In this presentation, we would like to
introduce the detailed configuration of the hardware for Grid instance, and several mechanism to construct the robust Grid system in the next generation system of KEKCC.

**Secondary Keyword (Optional):**
Computing facilities

**Primary Keyword (Mandatory):**
Computing middleware

**Tertiary Keyword (Optional):**

---

**Track 3: Distributed Computing / 426**

**Opportunistic resource usage at Argonne Facility for CMS**

**Authors:** Jean-Roch Vlimant\(^1\); Justas Balcas\(^1\)

**Co-author:** Josh Bendavid \(^1\)

\(^1\) California Institute of Technology (US)

**Corresponding Authors:** jean-roch.vlimant@cern.ch, justas.balcas@cern.ch, joshbendavid@gmail.com

With the increased load and pressure on required computing power brought by the higher luminosity in LHC during Run2, there is a need to utilize opportunistic resources not currently dedicated to the Compact Muon Solenoid (CMS) collaboration. Furthermore, these additional resources might be needed on demand. The Caltech group together with the Argonne Leadership Computing Facility (ALCF) are collaborating to demonstrate the feasibility of using resources from one of the fastest supercomputers in the world, Mira (10 petaflops IBM Blue Gene/Q system). CMS uses the HTCondor/glideinWMS job submission infrastructure for all its batch processing. On the other hand, Mira only supports MPI applications using Cobalt submission, which is not yet available through HTCondor. Majority of computing facilities utilized by CMS experiment are powered by x86_64 processors while Mira is Blue Gene/Q based (PowerPC Architecture). The CMS Monte-Carlo and Data production makes use of a bulk of pledge resource and other opportunistic resource. For efficient use, Mira’s resource has to be transparently integrated into the CMS production infrastructure. We address the challenges posed by submitting MPI applications through CMS infrastructure to Argonne PowerPC (Mira) supercomputer. We will describe the design and implementation of the computing and networking systems for running CMS Production jobs with first operational prototype for running on PowerPC. We also demonstrate the state of the art high networking from the LHC Grid to ALCF requirement of CMS data intensive computation.

**Secondary Keyword (Optional):**
Data processing workflows and frameworks/pipelines

**Primary Keyword (Mandatory):**
High performance computing

**Tertiary Keyword (Optional):**

---

**Posters B / Break / 427**

**Last development of the Long Term Data Preservation project for CDF at INFN CNAF**

**Authors:** Enrico Fattibene \(^1\); Luca dell’Agnello \(^2\); Pier Paolo Ricci \(^1\); Silvia Amerio \(^2\); Michele Pezzi \(^5\)

Page 286
The long term preservation and sharing of scientific data is becoming nowadays an integral part of any new scientific project. In High Energy Physics experiments (HEP) this is particularly challenging, given the large amount of data to be preserved and the fact that each experiment has its own specific computing model. In the case of HEP experiments that have already concluded the data taking phase, additional difficulties are the preservation of software versions that are not supported anymore and the protection of the knowledge about data and analysis framework.

The INFN Tier-1, located at CNAF, is one of the reference sites for data storage and computing in the LHC community but it also offers resources to many other HEP and non-HEP collaborations. In particular the CDF experiment has used the INFN Tier-1 resources for many years and, after the end of data taking in 2011, it faced the challenge to both preserve the large amount of data produced during several years and the ability to access and reuse the whole amount of it in the future. According to this task the CDF Italian collaboration, together with the INFN CNAF and Fermilab (FNAL), has implemented a long term data preservation project. The tasks of the collaboration comprises the copy of all CDF raw data and user level ntuples (about 4 PB) at the INFN CNAF site and the setup of a dedicated framework which allows to access and analyze the data in the long term future. The full sample of CDF data was successfully copied from FNAL to the INFN CNAF tape library backend and a new method for data access has been set up. Moreover a system for doing regular integrity check of data has been developed: it ensures that all the data are accessible and in case of problems it can automatically retrieve an identical copy of the file from FNAL. In addition to this data access and integrity system, a data analysis framework has been implemented in order to run the complete CDF RUN-1 dataset now stored in old Exabyte tapes.

As an integral part of the project, detailed documentation for users and administrations, has been produced in order to analyse data and maintain the whole system.

In this paper we will illustrate the different aspects of the project: from the difficulties and the technical solutions adopted to copy, store and maintain CDF data to the analysis framework and documentation web pages. We will also discuss the learned lesson from the CDF case which can be used when designing new data preservation projects for other experiments.

Tertiary Keyword (Optional):
Storage systems

Secondary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Primary Keyword (Mandatory):
Preservation of analysis and data

Posters A / Break / 428

CMS Software Performance Measurements "in the wild"

Author: Frank Wuerthwein

Co-authors: Aidan Macdonald; Benjamin Stieger; Brian Paul Bockelman; Kenneth Bloom

1 Univ. of California San Diego (US)
2 UCSD
Corresponding Authors: stiegerb@cern.ch, brian.bockelman@cern.ch, aidan.plenert.macdonald@gmail.com, frank.wuerthwein@cern.ch, kenbloom@unl.edu

CMS deployed a prototype infrastructure based on Elastic Search that stores all classAds from the global pool. This includes detailed information on IO, CPU, datasets, etc. etc. for all analysis as well as production jobs. We will present initial results from analyzing this wealth of data, describe lessons learned, and plans for the future to derive operational benefits from analyzing this data on a routine basis.

Tertiary Keyword (Optional):
Monitoring

Secondary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Primary Keyword (Mandatory):
Accounting and information

Track 3: Distributed Computing / 430

Opportunistic data locality for end user data analysis

Author: Max Fischer

Co-authors: Christoph Heidecker; Eileen Kuhn; Gunter Quast; Manuel Giffels; Marcus Schmitt; Matthias Jochen Schnepf

1 KIT - Karlsruhe Institute of Technology (DE)

Corresponding Authors: manuel.giffels@cern.ch, christoph.heidecker@cern.ch, max.fischer@cern.ch, eileen.kuhn@cern.ch, matthias.jochen.schnepf@cern.ch, gunter.quast@cern.ch, marcus.schmitt@cern.ch

With the LHC Run2, end user analyses are increasingly challenging for both users and resource providers. On the one hand, boosted data rates and more complex analyses favor and require larger data volumes to be processed. On the other hand, efficient analyses and resource provisioning require fast turnaround cycles. This puts the scalability of analysis infrastructures to new limits. Existing approaches to this problem, such as data locality based processing, are difficult to adapt to HEP workflows.

For the first data taking period of Run2, the KIT CMS group has deployed a prototype enabling data locality via coordinated caching. The underlying middleware successfully solves key issues of data locality for HEP:

- Caching joins local high performance devices with large background storage.
- Data selection based on user workflows only allocates critical data to optimize throughput.
- Finally, transparent integration into the batch system and operating system reduces compatibility issues for user software.

While the prototype has sped up user analyses by several factors, the scope has been limited so far. Our prototype is deployed only on static, local processing resources accessing file servers under our own administration.
Thus, recent developments focus on opportunistic infrastructure to prove the viability of our approach.

On the one hand, we focus on volatile resources, i.e. cloud computing.
The nature of caching lends itself nicely to this setup.
Yet, the lack of static infrastructure complicates distributed services, while delocalization makes locality optimizations more complicated.

On the other hand, we explore providing caching as a service. Instead of creating an entire analysis environment, we provide a thin platform integrated into caching and resource provisioning services.
Using docker, we merge this high performance data analysis platform with user analysis environments on demand. This allows using modern operating systems, drivers, and other performance critical components, while satisfying arbitrary user dependencies at the same time.

Tertiary Keyword (Optional):
Virtualization

Secondary Keyword (Optional):
Computing middleware

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

Track 6: Infrastructures / 431

TOSCA-based orchestration of complex clusters at the IaaS level

Authors: Germán Moltó¹; Mathieu Velten²; Miguel Caballer³; Ricardo Brito Da Rocha⁴

Co-author: Giacinto Donvito ⁴

¹ Universidad Politécnica de Valencia
² CERN
³ UPV
⁴ INFN-Bari

Corresponding Authors: gmolto@dsic.upv.es, giacinto.donvito@ba.infn.it, ricardo.rocha@cern.ch, micafer1@upv.es, mathieu.velten@cern.ch

The INDIGO-DataCloud project’s ultimate goal is to provide a sustainable European software infrastructure for science, spanning multiple computer centers and existing public clouds. The participating sites form a set of heterogeneous infrastructures, some running OpenNebula, some running OpenStack. There was the need to find a common denominator for the deployment of both the required PaaS services and the end user applications. CloudFormation or Heat were technically viable options, but tied to specific implementations. The TOSCA Simple Profile in YAML v1.0 specification, on the other hand, is on its way to becoming a standard to describe the topology of Cloud applications, with growing support in different communities.

In the context of INDIGO-DataCloud, TOSCA Templates are used at the IaaS level to describe complex clusters of applications and services and to provide a way to express their automatic configuration via Ansible recipes.

Within OpenStack, the TOSCA interface is implemented in the Heat orchestrator service, while in OpenNebula it is implemented using the Infrastructure Manager (IM), an open-source tool to deploy virtual infrastructures on multiple Clouds.

In a IaaS context both Heat and IM are very useful to ease portable provisioning of resources and deployment of services and applications on dynamically instantiated clusters. One of the advantages of using TOSCA and the IM/Heat approach is that the INDIGO PaaS layer can easily exploit it across different IaaS implementations, increasing the portability of the cluster definitions, and
implementing the provisioning of the required services across multiple IaaS infrastructures through the INDIGO orchestrator.

In this contribution we will outline our enhancements for the TOSCA support in both OpenStack and OpenNebula, done together in close collaboration with industry partners such as IBM. These developments, emerging from the INDIGO requirements, have been contributed upstream to the relevant tools, as they have been considered of general interest. Moreover, we will showcase how it is possible to deploy an elastic cluster managed by a batch system like SLURM or Torque, where nodes are dynamically added and removed from the cluster to adapt to the workload.

**Primary Keyword (Mandatory):**
Cloud technologies

**Secondary Keyword (Optional):**

**Tertiary Keyword (Optional):**

---

**Posters A / Break / 432**

**CMS Connect**

**Author:** Kenyi Paolo Hurtado Anampa

**Co-authors:** Anthony Tiradani; Antonio Perez-Calero Yzquierdo; Bo Jayatilaka; Brian Paul Bockelman; David Alexander Mason; Farrukh Aftab Khan; Jadir Marra Da Silva; James Letts; Justas Balcaс; Kevin Patrick Lannon; Krista Larson; Marco Mascheroni; Robert William Gardner Jr.

---

**Abstract:**

The CMS experiment collects and analyzes large amounts of data coming from high energy particle collisions produced by the Large Hadron Collider (LHC) at CERN. This involves a huge amount of real and simulated data processing that needs to be handled in batch-oriented platforms. The CMS Global Pool of computing resources provide +100K dedicated CPU cores and another 50K to 100K CPU cores from opportunistic resources for these kind of tasks and even though production and event processing analysis workflows are already managed by existing tools, there is still a lack of support to submit final stage condor-like analysis jobs familiar to Tier-3 or local Computing Facilities users into these distributed resources in an integrated (with other CMS services) and friendly way.

CMS Connect is a set of computing tools and services designed to augment existing services in the CMS Physics community focusing on these kind of condor analysis jobs. It is based on the CI-Connect platform developed by the Open Science Grid and uses the CMS GlideInWMS infrastructure to transparently plug CMS global grid resources into a virtual pool accessed via a single submission machine. This paper describes the specific developments and deployment of CMS Connect beyond the CI-Connect platform in order to integrate the service with CMS specific needs, including specific Site submission, accounting of jobs and automated reporting to standard CMS monitoring resources in an effortless way to their users.
A lightweight access control solution for Openstack

Author: 1

Co-authors: Haibo li 2; Tao Cui 3; Yaodong Cheng 1

1 IHEP
2 Institute of High Energy Physics Chinese Academy of Science
3 IHEP(Institute of High Energy Physics, CAS,China)

Corresponding Authors: lihaibo@ihep.ac.cn, cuit@ihep.ac.cn, huqb@ihep.ac.cn, chyd@ihep.ac.cn

OpenStack is an open source cloud computing project that is enjoying wide popularity. More and more organizations and enterprises deploy it to provide their private cloud services. However, most organizations and enterprises cannot achieve unified user management access control to the cloud service, since the authentication and authorization systems of Cloud providers are generic and they cannot be easily adapted to the requirements of each individual organization or enterprise.

In this paper we present the design of a lightweight access control solution that overcomes this problem. Our solution access control is offered as a service by a third trusted party, the Access Control Provider. Access control as a service enhances end-user privacy, eliminates the need for developing complex adaptation protocols, and offers user flexibility to switch among the Cloud service and another different services.

We have implemented and incorporated our solution in the popular open-source Cloud stack OpenStack. Moreover, we have designed and implemented a web application that enables the incorporation of our solution into the UMT of IHEP based on Auth2.0. The UMT of IHEP as a tool which is used to manage IHEP user, record user’s information, account, password and so on. Moreover, UMT provide the unified authentication service.

In our access control solution for Openstack, we create an information table to record all Openstack accounts and their passwords which will be queried when these accounts were authenticated by the third trusted party. As the new registered UMT user login the Cloud service for the first time, our system will create the user’s resources automatically by Openstack API, and record the user information into the information table immediately. Moreover, we still keep Openstack original login web page, so administrators and some special users can access Openstack and do some background management. We have applied the solution to IHEPCloud, an IaaS cloud platform at IHEP. Except UMT, it is easy to expand other third-party authentication tools, for example CERN account management system, google, sina, or tecent.
Primary Vertex finding in the RHIC high precision measurement era - enhancement and experience in STAR

Authors: Dmitri Smirnov\textsuperscript{1}; Gene Van Buren\textsuperscript{2}; Jason Webb\textsuperscript{3}; Jerome LAURET\textsuperscript{2}; Victor Perevoztchikov\textsuperscript{2}

\textsuperscript{1} BNL
\textsuperscript{2} Brookhaven National Laboratory
\textsuperscript{3} Brookhaven National Lab

Corresponding Authors: gene@bnl.gov, perev@bnl.gov, jwebb@bnl.gov, dsmirnov@bnl.gov, jlauret@bnl.gov

Since 2014, the STAR experiment has been exploiting data collected by the Heavy Flavor Tracker (HFT), a group of high precision silicon-based detectors installed to enhance track reconstruction and pointing resolution of the existing Time Projection Chamber (TPC). The significant improvement in the primary vertex resolution resulting from this upgrade prompted us to revisit the variety of vertex reconstruction algorithms currently employed by the experiment. In this contributions we share the experience gained in our search for a unified vertex finder (VF) for STAR and improvements made to the algorithms along the way.

The Relativistic Heavy Ion Collider (RHIC) is capable of providing collisions of particle beams made from a wide range of possible nuclei, from protons up to uranium. Historically, STAR utilized different VFs for heavy ion and proton-proton programs to cope with the distinctive nature of these two types of particle interactions. We investigate the possibility of having a universally acceptable vertex reconstruction algorithm that could equally satisfy all ongoing physics analyses. To achieve this goal we establish a common strategy, reshape generic interfaces over the years, and develop tools to evaluate the performance of diversified implementations in an unbiased way. In particular, we introduce independent measurements of the beamline position into the primary vertex fitting routine common to all VFs. This additional constraint on the vertex position is aimed to strengthen the identification of secondary vertices from short-lived particles decaying near the beamline. Finally, we discuss the vertex ranking schemes used in STAR to mitigate the effect of pile-up events contaminating the identification of triggered event vertices at high instantaneous luminosities. The pile-up hits are inevitable due to the inherently slow readout of the TPC and MAPS-based HFT detectors, therefore the systematic approach established for the VF comparison will be of a great help in the future exploration of various ranking schemes.

Tertiary Keyword (Optional):
Analysis tools and techniques

Secondary Keyword (Optional):
Algorithms

Primary Keyword (Mandatory):
Reconstruction

Parallel Monte Carlo search for Hough transform

Author: Raul Cardoso Lopes\textsuperscript{1}

Co-authors: Ivan Reid \textsuperscript{2}; Peter Hobson \textsuperscript{1}; Virginia Franqueira \textsuperscript{3}

\textsuperscript{1} Brunel University (GB)
\textsuperscript{2} Brunel University London (GB)
We investigate the combination of a Monte Carlo Tree Search, hierarchical space decomposition, Hough Transform techniques and parallel computing to the problem of line detection and shape recognition in general.

Paul Hough introduced in 1962 a method for detecting lines in binary images. Extended in the 1970s to the detection of space forms, what came to be known as the Hough Transform (HT) has been proposed, for example, in the context of track fitting in the LHC ATLAS\footnote{N. Amram, Hough Transform Track Reconstruction in the Cathode Strip Chambers in ATLAS, PhD Thesis, Tel Aviv University, 2008 CERN-THESIS-2008-062} and CMS \cite{2} projects. The HT transfers the problem of line detection, for example, into one of optimization of the peak in a vote counting process for cells which contain the possible points of candidate lines. The detection algorithm can be computationally expensive both in the demands made upon the processor and on memory. Proposals to improve its CPU performance have included the use of Monte Carlo algorithms and parallel computing. However, the detection algorithm can be expensive both in CPU and memory demands. Variations of the HT found in literature have a complexity that is least at cubic in the number of points. In addition, background noise can reduce the HT effectiveness, and statistical techniques or the use of the Radon transform instead have been proposed.

We present results for the practical evaluation of variations of the Hough Transform for line detection and discuss implementations on multi-GPU and multicore architectures.

References


Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Reconstruction

Primary Keyword (Mandatory):

Algorithms

---

Parallel metric trees for multi-billion body simulations

Author: Raul Cardoso Lopes\footnote{Brunel University (GB)}

Co-author: Peter Hobson \footnote{Brunel University (GB)}

\footnote{Brunel University (GB)}

Corresponding Authors: raul.cardoso.lopes@cern.ch, peter.hobson@brunel.ac.uk
This work combines metric and parallel computing on both multi-GPU and distributed memory architectures when applied to multi-million or even billion bodies simulations.

Metric trees are data structures for indexing multidimensional sets of points in arbitrary metric spaces. First proposed by Jeffrey K. Uhlmann [1], as a structure to efficiently solve neighbourhood queries, they have been considered, for example, by Sergey Brin for indexing very large databases.

We propose a parallel algorithm for the construction of metric trees that preserves the theoretical work bound of order $n \log(n)$, for indexing a set of $n$ points.

We discuss possible applications of the parallel algorithms obtained in the context of probabilistic Hough Transform applications for line detection and multi billion body simulations.

N-body simulations are of particular interest for beam dynamics simulations in the context of particle accelerator design. We use a parallel metric tree and a variation of a parallel Fast Multipole Method and evaluate its efficiency in a multi-billion points simulation on three different architectures: a multi-GPU cluster; a 256 core Infiniband, distributed memory cluster; and a multi-core architecture. Of particular interest is the evaluation of effects of locality on communication and performance overall.


Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Parallelizarion

Primary Keyword (Mandatory):
Algorithms

Posters A / Break / 438

Application of state quantization-based methods in HEP particle transport simulation

Author: Lucio Santi

Co-authors: Krzysztof Genser; Nicolás Ponieman; Rodrigo Castro; Soon Yung Jun; Victor Daniel Elvira

1 Universidad de Buenos Aires
2 Fermi National Accelerator Laboratory (US)

Corresponding Authors: lukius@gmail.com, genser@fnal.gov, rcastro@dc.uba.ar, nicoponieman@gmail.com, daniel.elvira@cern.ch, soon.yung.jun@cern.ch
Simulation of particle-matter interactions in complex geometries is one of the main tasks in high energy physics (HEP) research. Geant4 is the most commonly used tool to accomplish it.

An essential aspect of the task is an accurate and efficient handling of particle transport and crossing volume boundaries within a predefined (3D) geometry.

At the core of the Geant4 simulation toolkit, numerical integration solvers approximate the solution of the underlying differential equations that characterize the trajectories of particles in an electromagnetic field with a prescribed accuracy.

A common feature of Geant4 integration algorithms is their discrete-time nature, where the physics state calculations are essentially performed by slicing time into (possibly adaptive) steps.

In contrast, a different class of numerical methods replace time discretization by state variable quantization, resulting in algorithms of an asynchronous, discrete-event nature. The Quantized State Systems (QSS) family of methods is a canonical example of this category. One salient feature of QSS is a simpler, lightweight detection and handling of discontinuities based on explicit root-finding of polynomial functions.

In this work we present a performance comparison between a QSS-based standalone solver and combinations of standard fixed step 4th order Runge-Kutta (RK4) and adaptive step RK4/5 methods in the context of Geant4.

Our results show that QSS performance scales significantly better in situations with increasing number of volume crossings. Finally, we shall present the status of our work in progress related to embedding QSS within the Geant4 framework itself.

Secondary Keyword (Optional):

Algorithms

Primary Keyword (Mandatory):

Simulation

Tertiary Keyword (Optional):

Track 2: Offline Computing / 439

CMS Event Processing Multi-core Efficiency Status

Author: Christopher Jones

1 Fermi National Accelerator Lab. (US)

Corresponding Author: cdj@fnal.gov

In 2015, CMS was the first LHC experiment to begin using a multi-threaded framework for doing event processing. This new framework utilizes Intel’s Thread Building Block library to manage concurrency via a task based processing model. During the 2015 LHC run period, CMS only ran reconstruction jobs using multiple threads because only those jobs were sufficiently thread efficient. Recent work now allows simulation and digitization to be thread efficient. In addition, during 2015 the multi-threaded framework could run events in parallel but could only use one thread per event.
Work done in 2016 now allows multiple threads to be used while processing one event. In this presentation we will show how these recent changes have improved CMS’s overall threading and memory efficiency and we will discuss work to be done to further increase those efficiencies.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Parallelization

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

Posters B / Break / 440

Python object-oriented framework for consuming, manipulating and releasing non-event data for the CMS alignment and calibration.

Author: Joshua Heneage Dawes

1 University of Manchester (GB)

Corresponding Author: joshua.heneage.dawes@cern.ch

The Alignment, Calibrations and Databases group at the CMS Experiment delivers Alignment and Calibration Conditions Data to a large set of workflows which process recorded event data and produce simulated events. The current infrastructure for releasing and consuming Conditions Data was designed in the two years of the first LHC long shutdown to respond to use cases from the preceding data-taking period. During the second run of the LHC, new use cases were defined.

For the consumption of Conditions Metadata, no common interface existed for the detector experts to use in Python-based custom scripts, resulting in many different querying and transaction management patterns. A new metadata consumption framework has been built to address such use cases: a simple object-oriented tool that detector experts can use to read and write Conditions Metadata when using Oracle and SQLite databases, that provides a homogeneous method of querying across all services.

The tool provides mechanisms for segmenting large sets of conditions while releasing them to the production database, allows for uniform error reporting to the client-side from the server-side and optimizes the data transfer to the server. The architecture of the new service has been developed exploiting many of the features made available by the metadata consumption framework to implement the required improvements.

This paper presents the details of the design and implementation of the new metadata consumption and data upload framework, as well as analyses of the new upload service’s performance as the server-side state varies.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Primary Keyword (Mandatory):
Databases
grid-control - The Swiss Army knife of job submission tools

Author: Fred Stober
Co-author: Max Fischer

1 Hamburg University (DE)
2 KIT - Karlsruhe Institute of Technology (DE)

Corresponding Authors: max.fischer@cern.ch, fred.stober@cern.ch

grid-control is an open source job submission tool that supports common HEP workflows. Since 2007 it has been used by a number of HEP analyses to process tasks which routinely reach the order of tens of thousands of jobs.

The tool is very easy to deploy, either from its repository or the python package index (pypi). The project aims at being lightweight and portable. It can run in virtually any environment with access to some submission infrastructure. To achieve this, it avoids any external dependencies. Only a python interpreter (CPython or pypy) supporting any language version from 2.3 upwards, including 3.x, is required.

The program supports job submission to a wide range of local batch systems and grid middleware. For small tasks and tests, it is also possible to run jobs on machines without any batch system.

grid-control is built around a powerful plugin and configuration system, that allows to easily define the workflow that is to be processed.

A particularly useful feature for HEP applications is the job parameter system built into grid-control. It provides a convenient way to define the parameter space that a task is based on. This parameter space can be build up from any number of variables and data sources, which can have complex dependencies on each other. The parameter system is able to handle changes to the data source as well as to other parameters. It can transparently adapt the job submission to the new parameter space at runtime.

grid-control provides several plugins to retrieve datasets from different sources. This ranges from simple file listings, to databases, directory contents, other grid-control tasks and more. These datasets may contain a list of URLs, an optional number of work units (eg. events or file size), arbitrary meta-data (eg. lumi section information or generator parameters) and locality information.

All datasets are processed through a configurable pipeline of dataset filters, partition plugins and partition filters. Several methods to split datasets into partitions are supported. These partition plugins can take the number of files, size of the work units, metadata or combinations thereof into account. Dataset changes on the file level (additions or removals) as well as on the work unit level (expanding or shrinking files), are propagated through the processing pipeline and transparently trigger adjustments to the processed parameter space.

For HEP workflows this allows to run jobs on an ever expanding dataset with a single processing task that regularly queries some datasource and spawns new jobs.

While the core functionality is completely experiment independent, a dedicated plugin is available to simplify running tasks using the CMS experiment software (version 1.x-8.x) and to use the CMS Dataset Bookkeeping Service, CMS Data Aggregation System or PhEDEx as data source.
Cloud Environment Automation: from infrastructure deployment to application monitoring

Authors: Alessandro Costantini; Doina Cristina Aiftimiei

Co-authors: Alessandro Italiano Italiano; Davide Salomoni; Diego MICHELOTTO; Giovanni Zizzi; Marco Saletta; Maria Vistoli; Matteo Panella; Matteo Pergolesi; Riccardo Bucchi; Sergio Traldi

Open City Platform (OCP) is an industrial research project funded by the Italian Ministry of University and Research, started in 2014. It intends to research, develop and test new technological solutions open, interoperable and usable on-demand in the field of Cloud Computing, along with new sustainable organizational models for the public administration, to innovate, with scientific results, with new standards and technological solutions, the provision of services by the Local Public Administration (PAL) and Regional to citizens, businesses and other public administrations.

After the development and integration of the different components of the OCP platform that includes IaaS, PaaS and SaaS services, we had to cope with an increasing number of requests to deploy new testbeds at the PALs interested to deploy the OCP platform starting from the IaaS/OpenStack level. In this contribution we present the OCP solution for the automatization and standardization of installation and configuration procedures. With respect to already existing similar tools like Fuel or Staypuft, that were also tested, the solution adopted allows a very flexible initial customization of the services so that it can easily adapt to the different hardware resources and even virtualization techniques that the experimenting PALs make available.

The solution proposed is leveraging two of the most popular open source automation tools, namely Foreman and Puppet, making use as much as possible of the official OpenStack Puppet modules, as well as of other community supported Puppet modules for services like MySQL/Percona, CEPH. We concentrated on the integration of these modules by developing a new Puppet module, iaas-ha, based on different roles (like controller, compute, storage, monitoring nodes) and profiles (like nova, neutron, zabbix, etc), and making it configurable through the Foreman interface. With our solution we tried to address the different requirements and realities that we met during the collaboration with PAs, including, among others: the configuration of multiple external networks; full management of the configuration of the network layer giving the ability to merge or split the configuration of the various OpenStack networks (management, data, public and external); use of CEPH both as block and object storage backend - configuring the RADOSGW to use CEPH RADOS library in order to expose Swift APIs; fine grained variable configuration through the use of the Foreman GUI allowing site-admins to specify the values of all service specific parameters.

We will also present the first outcome of the work done in order to integrate the Cloud Formation as a Service in our automation tool for the installation and configuration of the OCP PaaS layer. Finally, we will discuss planned future work on the integration of a monitoring information service able to
collect information about resource availability in different infrastructures for IaaS, PaaS and SaaS components.

Tertiary Keyword (Optional):
Experience/plans from outside experimental HEP/NP

Secondary Keyword (Optional):
Virtualization

Primary Keyword (Mandatory):
Cloud technologies

Posters A / Break / 444

A multi-group and preemptive computing resource scheduling system based on HTCondor

Author: xiaowei jiang

Co-authors: Jiaheng Zou; Jingyan Shi; Yaodong Cheng

Corresponding Authors: chyd@ihep.ac.cn, jiangxw@ihep.ac.cn, shijy@ihep.ac.cn, zoujh@ihep.ac.cn

Virtual machines have many features — flexibility, easy controlling and customized system environments. More and more organizations and enterprises begin to deploy virtualization technology and cloud computing to construct their distributed system. Cloud computing is widely used in high energy physics field. In this presentation, we introduce an integration of virtual machines with HTCondor, which support resource management of multiple groups and preemptive scheduling policy. The system makes resources management more flexible and more efficient. Firstly, computing resources belong to different experiments, and each experiment has one or more user groups. All users of a same experiment have the access permission to all the resources owned by that experiment. Therefore, we have two types of groups, resource group and user group. In order to manage the mapping of user group and resource group, we design a permission controlling component to ensure jobs are delivered to suitable resource groups. Secondly, for elastically adjusting the resource scale of a resource group, it is necessary to schedule resources in the way of scheduling jobs. So we design a resource scheduler that focusing on virtual resources. The resource scheduler maintains a resource queue and matches an appropriate amount of virtual machines from the requested resource-group. Thirdly, in some conditions, one case that the resource may be occupied by a resource-group for a long time, it needs to be preempted. This presentation adds the preemptive feature to the resource scheduler based on the group priority. Higher priority leads to lower preemption probability, and lower priority leads to higher preemption probability. Virtual resources can be smoothly preempted, and running jobs are held and re-matched later. The feature is based on HTCondor, storing the held job, releasing the job to idle status and waiting for a secondary matching. We built a distributed virtual computing system based on HTCondor and Openstack. This presentation also shows some use cases of the JUNO and LHAASO experiments. The result shows that multi-group and preemptive resource scheduling perform well. Besides, the permission controlling component are not only used in virtual cluster but also in the local cluster, and the amount of experiments which it supports are expanding.

Tertiary Keyword (Optional):
High performance computing

Secondary Keyword (Optional):
Distributed workload management
Readout and trigger for AFP detector at ATLAS experiment

Author: Catrin Bernius
Co-author: Krzysztof Marian Korcyl

1 New York University (US)
2 Polish Academy of Sciences (PL)

Corresponding Authors: catrin.bernius@cern.ch, krzysztof.korcyl@cern.ch

AFP, the ATLAS Forward Proton detector upgrade project consists of two forward detectors at 205 m and 217 m on each side of the ATLAS experiment at the LHC. The new detectors aim to measure momenta and angles of diffractively scattered protons. In 2016 two detector stations on one side of the ATLAS interaction point have been installed and are being commissioned.

The front-end electronics consists currently of eight tracking modules based on the ATLAS 3D pixel sensors with the FEI4 readout chip. The chips are read via serial lines at 160 Mbps. The transmission line consists of 8 m of electrical twisted pair cable to an optical converter and 200 m of optical fiber. The DAQ system uses a FPGA board based on a Xilinx Artix chip, HSIO-2, and a mezzanine card that plugs into this board. The mezzanine card contains a RCE data processing module based on a Xilinx Zynq chip.

The software for calibration and monitoring of the AFP detectors runs on the ARM processor of the Zynq under Linux. The RCE communicates with the ATLAS Run Control software via the standard ATLAS TDAQ software. The AFP trigger signal is generated from the OR of all pixels of each frontend chip where the signals from individual planes can be logically combined. The resulting trigger signal in the form a NIM pulse is transmitted over a 260 m long air core coaxial cable where it is fed into the ATLAS LVL1 trigger system.

In this contribution we give an technical overview of the AFP detector together and the commissioning steps that have been taken. Furthermore first performance results are presented.
Co-authors: Anupam Ashish ²; Christian Bernardt ³; Dmitry Litvinsev ⁴; Gerd Behrmann ⁵; Jurgen Starek ²; Marina Sahakyan ²; Olufemi Segun Adeyemi ²; Patrick Fuhrmann ²; Paul Millar; Tigran Mkrtchyan ²

1 Fermi National Accelerator Laboratory
2 DESY
3 Deutsches Elektronen-Synchrotron (DE)
4 FNAL
5 NEIC

Corresponding Authors: patrick.fuhrmann@desy.de, christian.bernardt@desy.de, litvinse@fnal.gov, olufemi.segun.adeyemi@desy.de, behrmann@gmail.com, tigran.mkrtchyan@desy.de, arossi@fnal.gov, marina.sahakyan@desy.de, anupam.ashish@desy.de, paul.millar@desy.de, juergen.starek@desy.de

dCache is a distributed multi-tiered data storage system widely used by High Energy Physics and other scientific communities. It natively supports a variety of storage media including spinning disk, SSD and tape devices. Data migration between different media tiers is handled manually or automatically based on policies. In order to provide different levels of quality of service concerning performance, data availability and data durability, dCache manages multiple copies of data on different storage devices. In dCache, this feature is called data resilience.

In this paper we discuss the design and implementation of the Resilience Service in dCache. The service was conceived to meet the requirements of flexibility, fine-grained definition of resilience constraints, ease of configuration and integration with existing dCache services. We will also detail several optimizations that were applied to improve concurrency, consistency and fairness, along with the rich set of diagnostic and control commands available through the dCache admin interface. A test procedure and results will be covered as well.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Storage systems

Primary Keyword (Mandatory):
Storage systems

Track 1: Online Computing / 447

Real-time analysis with the LHCb trigger in Run-II

Author: Sascha Stahl¹

Co-authors: Conor Fitzpatrick ²; Gerhard Raven ³; Mika Anton Vesterinen ⁴

1 CERN
2 Ecole Polytechnique Federale de Lausanne (CH)
3 Nikhef National institute for subatomic physics (NL)
4 Ruprecht-Karls-Universitaet Heidelberg (DE)
The LHCb software trigger underwent a paradigm shift before the start of Run-II. From being a system to select events for later offline reconstruction, it can now perform the event analysis in real-time, and subsequently decide which part of the event information is stored for later analysis.

The new strategy is only possible due to a major upgrade during the LHC long shutdown I (2012-2015). The CPU farm was increased by almost a factor of two and the software trigger was split into two stages. The first stage performs a partial event reconstruction and inclusive selections to reduce the 1 MHz input rate from the hardware trigger to an output rate of 150 kHz. The output is buffered on hard disks distributed across the trigger farm. This allows for an asynchronous execution of the second stage where the CPU farm can be exploited also in between fills, and, as an integral part of the new strategy, the real-time alignment and calibration of sub-detectors before further processing. The second stage performs a full event reconstruction which is identical to the configuration used offline. LHCb is the first high energy collider experiment to do this. Hence, event selections are based on the best quality information and physics analyses can be performed directly in and on the output of the trigger. This concept, called the “LHCb Turbo stream”, where reduced event information is saved, increases the possible output rate while keeping the storage footprint small.

In 2017, around half of the 400 trigger selections send their output to the Turbo stream and, for the first time, the Turbo stream no longer keeps the raw sub-detector data banks that would be needed for a repeated offline event reconstruction. This allows up to a factor of 10 decrease in the size of the events, and thus an equivalent factor higher rate of signals that can be exploited in physics analyses. Additionally, the event format has been made more flexible, which has allowed more used of the turbo stream in more physics analyses. We review the status of this real time analysis and discuss our plans for its evolution during Run-II towards the upgraded LHCb experiment that will begin operation in Run-III.

Secondary Keyword (Optional):
Reconstruction

Primary Keyword (Mandatory):
Trigger

Tertiary Keyword (Optional):

Parallel Algorithms for Online Trackfinding at PANDA Using GPUs

Author: Ludovico Bianchi
Co-authors: Andreas Herten ; James Ritman ; Tobias Stockmanns

1 Forschungszentrum Jülich GmbH

Corresponding Authors: lbianchi@fz-juelich.de, j.ritman@fz-juelich.de, t.stockmanns@fz-juelich.de, a.herten@fz-juelich.de

The PANDA experiment, one of the four scientific pillars of the FAIR facility currently in construction in Darmstadt, Germany, is a next-generation particle detector that will study collisions of antiprotons with beam momenta of 1.5–15 GeV/c on a fixed proton target. Because of the broad physics scope and the similar signature of signal and background events in the energy region of interest, PANDA’s strategy for data acquisition is to continuously record data from the whole detector, and use this global information to perform online event reconstruction and filtering. A real-time rejection factor of up to 1000 must be achieved to match the incoming data rate for offline storage, making all components of the data processing system computationally very challenging.
Online particle track identification and reconstruction is an essential step, since track information is used as input in all following phases. Online tracking algorithms must ensure a delicate balance between high tracking efficiency and quality, and minimal computational footprint. For this reason, a massively parallel solution with multiple Graphic Processing Units (GPUs) is under investigation. The Locus Circle Hough algorithm is currently being developed by our group. Based on the Circle Hough algorithm (doi:10.1088/1742-6596/664/8/082006), it uses the information from the geometric properties of primary track candidates in the Hough space to extend the hit-level parallelism to later phases of the algorithm. Two different strategies, based on curve rasterization and analytical intersections, are considered. The poster will present the core concepts of the Locus Circle Hough algorithm, details of its implementation on GPUs, and results of testing its physics and computational performance.

Secondary Keyword (Optional):
Parallelization

Primary Keyword (Mandatory):
Algorithms

Track 5: Software Development / 449

A browser-based event display for the CMS Experiment at the LHC using WebGL

Author: Thomas Mc Cauley

1 University of Notre Dame (US)

Corresponding Author: thomas.mccauley@cern.ch

Modern web browsers are powerful and sophisticated applications that support an ever-wider range of uses. One such use is rendering high-quality, GPU-accelerated, interactive 2D and 3D graphics in an HTML canvas. This can be done via WebGL, a JavaScript API based on OpenGL ES. Applications delivered via the browser have several distinct benefits for the developer and user. For example, they can be implemented using well-known and well-developed technologies, while distribution and use via a browser allows for rapid prototyping and deployment and ease of installation. In addition, delivery of applications via the browser allows for easy use on mobile, touch-enabled devices such as phones and tablets.

iSpy WebGL is an application for visualization of events detected and reconstructed by the CMS Experiment at the Large Hadron Collider at CERN. The first event display developed for an LHC experiment to use WebGL, iSpy WebGL is a client-side application written in JavaScript, HTML, and CSS and uses the WebGL API three.js. iSpy WebGL is used for monitoring of CMS detector performance, for production of images and animations of CMS collisions events for the public, as a virtual reality application using Google Cardboard, and as a tool available for public education and outreach such as in the CERN Open Data Portal and the CMS masterclasses. We describe here its design, development, and usage as well as future plans.

Tertiary Keyword (Optional):
Collaborative tools

Secondary Keyword (Optional):
Outreach

Primary Keyword (Mandatory):
Integrated monitoring of the ATLAS online computing farm

Author: Diana Scannicchio¹

Co-authors: Christopher Jon Lee ²; Costin Gament ³; Daniel Fazio ⁴; Franco Brasolin ⁵; Matthew Shaun Twomey ⁶; Sergio Ballestrero ⁷

¹ University of California Irvine (US)
² University of Cape Town (ZA)
³ University Politehnica of Bucharest (RO)
⁴ CERN
⁵ Sezione di Bologna (INFN)-Universita e INFN
⁶ University of Washington (US)
⁷ A.D.A.M. Applications of Detectors and accelerators to Medicine

Corresponding Authors: daniel.fazio@cern.ch, diana.scannicchio@cern.ch, chris.lee@cern.ch

The online farm of the ATLAS experiment at the LHC, consisting of nearly 4000 PCs with various characteristics, provides configuration and control of the detector and performs the collection, processing, selection and conveyance of event data from the front-end electronics to mass storage. The status and health of every host must be constantly monitored to ensure the correct and reliable operation of the whole online system. This is the first line of defense, which should not only promptly provide alerts in case of failure but, whenever possible, warn of impending issues. The monitoring system should be able to check up to 100000 health parameters and provide alerts on a selected subset.

In this paper we present the implementation and validation of our new monitoring and alerting system based on Icinga 2 and Ganglia. We describe how the load distribution and high availability features of Icinga 2 allowed us to have a centralised but scalable system, with a configuration model that allows full flexibility while still guaranteeing complete farm coverage. Finally, we cover the integration of Icinga 2 with Ganglia and other data sources, such as SNMP for system information and IPMI for hardware health.

Primary Keyword (Mandatory):
Monitoring

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

A Web-based application for the collection, management and release of Alignment and Calibration configurations used in data processing at the Compact Muon Solenoid experiment

Authors: Audrius Mecionis¹; Salvatore Di Guida²

Co-authors: Andreas Pfeiffer ³; Giacomo Govi ⁴; Gianluca Cerminara ⁵; Giovanni Franzoni ⁵; Marco Musich ⁵
The Compact Muon Solenoid (CMS) experiment makes a vast use of alignment and calibration measurements in several data processing workflows: in the High Level Trigger, in the processing of the recorded collisions and in the production of simulated events for data analysis and studies of detector upgrades. A complete alignment and calibration scenario is factored in approximately three-hundred records, which are updated independently and can have a time-dependent content, to reflect the evolution of the detector and data taking conditions. Given the complexity of the CMS condition scenarios and the large number (50) of experts who actively measure and release calibration data, in 2015 a novel web-based service has been developed to structure and streamline their management: the cmsDbBrowser. cmsDbBrowser provides an intuitive and easily accessible entry point for the navigation of existing conditions by any CMS member, for the bookkeeping of record updates and for the actual composition of complete calibration scenarios. This paper describes the design, choice of technologies and the first year of usage in production of the cmsDbBrowser.

Secondary Keyword (Optional):
Collaborative tools

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

Tertiary Keyword (Optional):

Posters B / Break / 452

Real time analysis with the upgraded LHCb trigger in Run-III

Author: Mika Anton Vesterinen¹
Co-authors: Conor Fitzpatrick ²; Sascha Stahl ³; Tomasz Szumlak ⁴

¹ Ruprecht-Karls-Universitaet Heidelberg (DE)
² Ecole Polytechnique Federale de Lausanne (CH)
³ CERN
⁴ AGH University of Science and Technology (PL)

The current LHCb trigger system consists of a hardware level, which reduces the LHC bunch-crossing rate of 40 MHz to 1 MHz, a rate at which the entire detector is read out. A second level, implemented in a farm of around 20k parallel processing CPUs, the event rate is reduced to around 12.5 kHz. The LHCb experiment plans a major upgrade of the detector and DAQ system in the LHC long shutdown II (2018-2019). In this upgrade, a purely software based trigger system is being developed and it will have to process the full 30 MHz of bunch crossings with inelastic collisions. LHCb will also receive a factor of 5 increase in the instantaneous luminosity, which further contributes to the challenge of reconstructing and selecting events in real time with the CPU farm. We discuss the plans and progress towards achieving efficient reconstruction and selection with a 30 MHz throughput. Another challenge is to exploit the increased signal rate that results from removing the 1 MHz readout bottleneck, combined with the higher instantaneous luminosity. Many charm hadron signals can be recorded at up to 50 times higher rate. LHCb is implementing a new paradigm in the form of real
time data analysis, in which abundant signals are recorded in a reduced event format that can be fed directly to the physics analyses. These data do not need any further offline event reconstruction, which allows a larger fraction of the grid computing resources to be devoted to Monte Carlo productions. We discuss how this real time analysis model is absolutely critical to the LHCb upgrade, and how it will evolve during Run-II.

Secondary Keyword (Optional):
DAQ

Primary Keyword (Mandatory):
Trigger

Tertiary Keyword (Optional):

Posters B / Break / 453

OpenFabrics Interface and NanoMsg

Author: Ioannis Charalampidis¹
Co-authors: Adam Tadeusz Wegrzynek ²; Lazaros Lazaridis ¹

¹ CERN
² Warsaw University of Technology (PL)

Corresponding Authors: ioannis.charalampidis@cern.ch, adam.wegrzynek@cern.ch, lazaros.lazaridis@cern.ch

With the demand for more computing power and the widespread use of parallel and distributed computing, applications are looking for message-based transport solutions for fast, stateless communication. There are many solutions already available, with competing performances, but with varying APIs, making it difficult to support all of them. Trying to find a solution to this problem we decided to explore the combination of two complementing libraries: nanomsg and libfabric. The first one implements various scalability protocols (such as pair, publish-subscribe or push-pull) while the latter introduces the OpenFabrics Interfaces, a fabric communication service to high performance applications. This way, we managed to expose access to high-performance hardware, such as PSM, usNIC and VERBS with a simple API, close to unix sockets, with a minimal overhead.

In this paper we are going to present the technical details of our solution, a series of benchmarks to compare its performance to other solutions and our future plans.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
High performance computing

Primary Keyword (Mandatory):
Network systems and solutions

Posters B / Break / 454

Master of Puppets

Author: Christopher Jon Lee¹
Co-authors: Costin Gament 2; Daniel Fazio 3; Diana Scannicchio 4; Franco Brasolin 5; Matthew Shaun Twomey 6; Sergio Ballestrero 7

1 University of Cape Town (ZA)
2 University Politehnica of Bucharest (RO)
3 CERN
4 University of California Irvine (US)
5 Sezione di Bologna (INFN)-Universita e INFN
6 University of Washington (US)
7 A.D.A.M. Applications of Detectors and accelerators to Medicine

Corresponding Author: chris.lee@cern.ch

Within the ATLAS detector, the Trigger and Data Acquisition system is responsible for the online processing of data streamed from the detector during collisions at the Large Hadron Collider at CERN. The online farm is comprised of ˜4000 servers processing the data read out from ˜100 million detector channels through multiple trigger levels. Configuring of these servers is not an easy task, especially since the detector itself is made up of multiple different sub-detectors, each with their own particular requirements.

The previous method of configuring these servers, using Quattor and a hierarchical scripts system was cumbersome and restrictive. A better, unified system was therefore required to simplify the tasks of the TDAQ Systems Administrators, for both the local and net booted systems, and to be able to fulfil the requirements of TDAQ, Detector Control Systems and the sub-detectors groups.

Various configuration management systems were evaluated, though in the end, Puppet was chosen as the application of choice and was the first such implementation at CERN. In this paper we describe the newly implemented system, detailing the redesign, the configuration and the use of the Puppet manifests to ensure a sane state of the entire farm.

Primary Keyword (Mandatory):
Computing facilities

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

Posters A / Break / 455

Design and evaluation of Hybrid storage system in HEP Environment

Author: Qi Xu 1

Co-authors: Gang Chen 1; Yaodong Cheng 1

1 Institute of High Energy Physics, Chinese Academy of Sciences

Corresponding Authors: xuq@ihep.ac.cn, chyd@ihep.ac.cn, gang.chen@ihep.ac.cn

Abstract: Nowadays, the High Energy Physics experiments produce a large amount of data. These data is stored in massive storage system, which need to balance the cost, performance and manageability. HEP is a typical data-intensive application, and process a lot of data to achieve scientific discoveries. A hybrid storage system including SSD (Solid-state Drive) and HDD (Hard Disk Drive) layers is designed to accelerate data analysis and reduce the cost. The performance of accessing files is one of decisive factors for the HEP computing system. The Hybrid storage system can provide a caching mechanism for the server which will improve the performance dramatically. The system combines the advantages of SSD and HDD. It works on virtual block device and the logic block is made up of SSD and HDD. In this way, system gets excellent I/O performance and large capacity with low cost. This paper describes the Hybrid storage system in detail. Firstly, this paper analyzes
the advantages of Hybrid Storage System in High Energy Physics, summarizes the characteristics of data access mode, evaluates the performance of different Read/Write mode, then proposes a new deployment model of Hybrid Storage System in High Energy Physics, which is proved to have higher I/O performance. The paper also gives detailed evaluation methods and the evaluations are about SSD/HDD ratio, the size of the log block, the size of experiment files and so on. In all evaluations, sequential read, sequential write, random read and random write are all tested to get the comprehensive results. The results show the Hybrid Storage System has good performance in some fields such as accessing big files in HEP. Based on the analysis, I proposed an optimization algorithm taking into account a variety of factors including SSD/HDD ratio, file size, performance, price and so on. The Hybrid storage system can achieve better I/O performance with lower price in High Energy Physics, and also be applied in other fields which have a large amount of data.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):
Storage systems

Track 2: Offline Computing / 457

Functional tests of a prototype for the CMS-ATLAS common non-event data handling framework

Author: Giacomo Govi

Co-authors: Andrea Formica; Andreas Pfeiffer; Giovanni Franzoni; Marco Musich; Roland Sipos

1 Fermi National Accelerator Lab. (US)
2 CEA/IRFU, Centre d'étude de Saclay Gif-sur-Yvette (FR)
3 CERN
4 Université Catholique de Louvain (UCL) (BE)
5 Eotvos Lorand University (HU)

Corresponding Authors: giovanni.franzoni@cern.ch, aformic@cern.ch, marco.musich@cern.ch, andreas.pfeiffer@cern.ch, giacomo.govi@cern.ch, roland.sipos@cern.ch

Since the 2014 the ATLAS and CMS experiments share a common vision for the Condition Database infrastructure required to handle the non-event data for the forthcoming LHC runs. The large commonality in the use cases allows to agree on a common overall design solution meeting the requirements of both experiments. A first prototype implementing these solutions has been completed in 2015 and was made available to both experiments.

The prototype is based on a web service implementing a REST api with a set of functions for the management of conditions data. The choice to use a REST api in the architecture has two main advantages: - the Conditions data are exchanged in a neutral format (JSON or XML), allowing to be processed by different languages and/or technologies in different frameworks. - the client is agnostic with respect to the underlying technology used for the persistency (allowing standard RDBMS and NoSQL back-ends).

The implementation of this prototype server uses standard technologies available in Java for server based applications. This choice has the benefit of easing the integration with the existing java-based applications in use by both experiments, notably the Frontier service in the distributed computing environment.

In this contribution, we describe the testing of this prototype performed within the CMS computing infrastructure, with the aim of validating the support of the main use cases and of suggesting
future improvements. Since the data-model reflected in this prototype is very close to the layout of
the current CMS Condition Database, the tests could be performed directly with the existing CMS
condition data.

The strategy for the integration of the prototype into the experiments’ frameworks consists in re-
placing the innermost software layer handling the Conditions with a plugin. This plugin is capable
of accessing the web service and of decoding the retrieved data into the appropriate object structures
used in the CMS offline software. This strategy has been applied to run a test suite on the specific
physics data samples, used at CMS for the software release validation.

Secondary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Primary Keyword (Mandatory):
Databases

Tertiary Keyword (Optional):

Track 1: Online Computing / 458

Implementation of the ATLAS trigger within the ATLAS MultiThreaded Software Framework AthenaMT

Author: Catrin Bernius¹

Co-author: Benjamin Michael Wynne ²

¹ New York University (US)
² University of Edinburgh (GB)

Corresponding Authors: catrin.bernius@cern.ch, b.m.wynne@ed.ac.uk

We present an implementation of the ATLAS High Level Trigger that provides parallel execution of
trigger algorithms within the ATLAS multithreaded software framework, AthenaMT. This develop-
ment will enable the ATLAS High Level Trigger to meet future challenges due to the evolution of
computing hardware and upgrades of the Large Hadron Collider, LHC, and ATLAS Detector. Dur-
ing the LHC datataking period starting in 2021, luminosity will reach up to three times the original
design value. Luminosity will increase further, to up to 7.5 times the design value, in 2026 following
LHC and ATLAS upgrades. This includes an upgrade of the ATLAS trigger architecture that will re-
sult in an increase in the High Level Trigger input rate by a factor of 4 to 10 compared to the current
maximum rate of 300 kHz.

The current ATLAS multiprocess framework, AthenaMP, manages a number of processes that pro-
cess events independently, executing algorithms sequentially in each process. AthenaMT will pro-
vide a fully multithreaded environment that will enable concurrent execution of algorithms also
within an event. This has the potential to significantly reduce the memory footprint on future manycore
devices. An additional benefit of the High Level Trigger implementation within the AthenaMT is
that it facilitates the integration of offline code into the High Level Trigger. The trigger must retain
high rejection in the face of increasing numbers of pileup collisions. This will be achieved by greater
use of offline algorithms that are designed to maximize the discrimination of signal from background.
Therefore a unification of the High Level Trigger and offline reconstruction software environment
is required. This has been achieved while at the same time retaining important High Level Trigger-
specific optimisations that minimize the computation performed to reach a trigger decision. Such
optimizations include early event rejection and reconstruction within restricted geometrical regions.
We report on a High Level Trigger prototype in which the need for High Level Trigger-specific com-
ponents has been reduced to a minimum. Promising results have been obtained with a prototype
that includes the key elements of trigger functionality including regional reconstruction and early
event rejection. We report on the first experience of migrating trigger selections to this new frame-
work and present the next steps towards a full implementation of the ATLAS trigger within this
framework.
The Compact Muon Solenoid (CMS) experiment makes a vast use of alignment and calibration measurements in several crucial workflows: in the event selection at the High Level Trigger (HLT), in the processing of the recorded collisions and in the production of simulated events. A suite of services addresses the key requirements for the handling of the alignment and calibration conditions such as: recording the status of the experiment and of the ongoing data taking, accepting conditions data updates provided by the detector experts, aggregating and navigating the calibration scenarios, and distributing conditions for consumption by the collaborators. Since a large fraction of such services is critical for the data taking and event filtering in the HLT, a comprehensive monitoring and alarm generating system had to be developed. Such monitoring system has been developed based on the open source industry standard for monitoring and alerting services (Nagios) to monitor: the database back-end, the hosting nodes and key heart-beat functionalities for all the services involved. This paper describes the design, implementation and operational experience with the monitoring system developed and deployed at CMS in 2016.

Primary Keyword (Mandatory):
Monitoring

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

Developing the WLCG Tier-1 center GridKa as topical center in a multidisciplinary research environment

Authors: Andreas Heiss¹; Andreas Petzold¹

¹ Eotvos Lorand University (HU)
The WLCG Tier-1 center GridKa is developed and operated by the Steinbuch Centre for Computing (SCC) at the Karlsruhe Institute of Technology (KIT). It was the origin of further Big Data research activities and infrastructures at SCC, e.g. the Large Scale Data Facility (LSDF), providing petabyte scale data storage for various non-HEP research communities.

Several ideas and plans exist to address the increasing demand for large scale data storage and data management services and computing resources by more and more science communities within Germany and Europe, e.g. the Helmholtz Data Federation (HDF) or the European Open Science Cloud (EOSC).

In the era of LHC run 3 and Belle-II, high energy physics will produce even more data than in the past years, requiring improvements of the computing and data infrastructures as well as computing models.

We present our plans to further develop GridKa as topical center within the multidisciplinary research environments at KIT, in Germany and Europe and in the light of increasing requirements and advanced computing models to provide the best possible services to high energy physics.

**Primary Keyword (Mandatory):**
Computing facilities

**Secondary Keyword (Optional):**

**Tertiary Keyword (Optional):**

---

**Posters B / Break / 462**

**Triggering on leptons and photons on ATLAS**

**Author:** Masahiro Tanaka¹

**Co-author:** Catrin Bernius ²

¹ Tokyo Institute of Technology (JP)
² New York University (US)

**Corresponding Authors:** catrin.bernius@cern.ch, masahiro.tanaka@cern.ch

Electron, muon and photon triggers covering transverse energies from a few GeV to several TeV are essential for signal selection in a wide variety of ATLAS physics analyses to study Standard Model processes and to search for new phenomena. Final states including leptons and photons had, for example, an important role in the discovery and measurement of the Higgs particle. Dedicated triggers are also used to collect data for calibration, efficiency and fake rate measurements. The trigger system of the ATLAS experiment at the LHC is divided in a hardware-based Level 1 and a software based high level trigger, both of which were upgraded during the long shutdown of the LHC in preparation for data taking in 2015. The increasing luminosity and more challenging pile-up conditions as well as the planned higher center-of-mass energy demanded the optimisation of the trigger selections at each level, to control the rates and keep efficiencies high. To control the rate, new hardware selections are implemented at the Level 1. To improve the performance multivariate analysis techniques are introduced for the electron selections.
The evolution of the ATLAS electron, muon and photon triggers and their performance will be presented, including new results from the 2015 LHC Run 2 operation.

**Primary Keyword (Mandatory):**
Trigger

**Tertiary Keyword (Optional):**
Reconstruction

**Secondary Keyword (Optional):**
Algorithms

**MonALISA, An Agent-Based Monitoring and Control System for the LHC Experiments: Status and Outlook**

**Author:** Dorian Kcira

**Co-authors:** Harvey Newman; Iosif-Charles Legrand

1 California Institute of Technology (US)

**Corresponding Authors:** iosif.legrand@cern.ch, harvey.newman@cern.ch, dorian.kcira@cern.ch

MonALISA, which stands for Monitoring Agents using a Large Integrated Services Architecture, has been developed over the last fourteen years by Caltech and its partners with the support of the CMS software and computing program. The framework is based on Dynamic Distributed Service Architecture and is able to provide complete monitoring, control and global optimization services for complex systems.

The MonALISA system is designed as an ensemble of autonomous multi-threaded, self-describing agent-based subsystems which are registered as dynamic services, and are able to collaborate and cooperate in performing a wide range of information gathering and processing tasks. These agents can analyze and process the information, in a distributed way, to provide optimization decisions in large scale distributed applications. An agent-based architecture provides the ability to invest the system with increasing degrees of intelligence, to reduce complexity and make global systems manageable in real time. The scalability of the system derives from the use of multithreaded execution engine to host a variety of loosely coupled self-describing dynamic services or agents and the ability of each service to register itself and then to be discovered and used by any other services, or clients that require such information. The system is designed to easily integrate existing monitoring tools and procedures and to provide this information in a dynamic, customized, self describing way to any other services or clients.

A report of the present status of development in MonALISA as well as outlook on future developments will be given.

**Secondary Keyword (Optional):**
Distributed workload management

**Primary Keyword (Mandatory):** Monitoring
The ATLAS Run 2 Trigger: Design, Menu, Performance and Operational Aspects

Author: Catrin Bernius
Co-author: Tim Martin

1 New York University (US)
2 University of Warwick (GB)

Corresponding Authors: catrin.bernius@cern.ch, tim.martin@cern.ch

The LHC, at design capacity, has a bunch-crossing rate of 40 MHz whereas the ATLAS experiment at the LHC has an average recording rate of about 1000 Hz. To reduce the rate of events but still maintain a high efficiency of selecting rare events such as physics signals beyond the Standard Model, a two-level trigger system is used in ATLAS. Events are selected based on physics signatures such as presence of energetic leptons, photons, jets or large missing energy. Despite the limited time available for processing collision events, the trigger system is able to exploit topological information, as well as using multi-variate methods. In total, the ATLAS trigger system consists of thousands of different individual triggers. The ATLAS trigger menu specifies which triggers are used during data taking and how much rate a given trigger is allocated. This menu reflects not only the physics goals of the collaboration but also takes the instantaneous luminosity of the LHC, the design limits of the ATLAS detector and the offline processing Tier0 farm into consideration.

We describe the criteria for designing the ATLAS trigger menu used for the LHC Run 2 period. Furthermore, we discuss the different phases of the deployment of the trigger menu for data-taking: validation, decision on allocated rates for different triggers (ahead of running, or during data-taking in case of sudden rate changes), and monitoring during data-taking itself. Additionally the performance of the high-level trigger algorithms used to identify leptons, hadrons and global event quantities which are crucial for event selection and relevant to wide range of physics analyses is presented at hand of a few examples.

Tertiary Keyword (Optional):
Reconstruction

Secondary Keyword (Optional):
Algorithms

Primary Keyword (Mandatory):
Trigger

Posters A / Break / 465

Development, Validation and Integration of the ATLAS Trigger System Software in Run 2

Author: Catrin Bernius
Co-author: Simon George

1 New York University (US)
2 Royal Holloway, University of London

Corresponding Authors: catrin.bernius@cern.ch, s.george@rhul.ac.uk
The trigger system of the ATLAS detector at the LHC is a combination of hardware, firmware and software, associated to various sub-detectors that must seamlessly cooperate in order to select 1 collision of interest out of every 40,000 delivered by the LHC every millisecond. This talk will discuss the challenges, workflow and organization of the ongoing trigger software development, validation and deployment. This development, from the top level integration and configuration to the individual components responsible for each sub system, is done to ensure that the most up to date algorithms are used to optimize the performance of the experiment. This optimization hinges on the reliability and predictability of the software performance, which is why validation is of the utmost importance. The software adheres to a hierarchical release structure, with newly validated releases propagating upwards. Integration tests are carried out on a daily basis to ensure that the releases deployed to the online trigger farm during data taking run as desired. Releases at all levels are validated by fully reconstructing the data from the raw files of a benchmark run, mimicking the reconstruction that occurs during normal data taking. This exercise is computationally demanding and thus runs on the ATLAS high performance computing grid with high priority. Performance metrics ranging from low level memory and CPU requirements, to shapes and efficiencies of high level physics quantities are visualized and validated by a range of experts. This is a multifaceted critical task that ties together many aspects of the experimental effort that directly influences the overall performance of the ATLAS experiment.

Tertiary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Secondary Keyword (Optional):
Reconstruction

Primary Keyword (Mandatory):
Trigger

Track 8: Security, Policy and Outreach / 466

Invenio digital library technology for open research data repositories

Authors: Jiri Kuncar¹; Lars Holm Nielsen¹; Tibor Simko¹

¹ CERN

Corresponding Authors: jiri.kuncar@cern.ch, tibor.simko@cern.ch, lars.holm.nielsen@cern.ch

We present the new Invenio 3 digital library framework and demonstrate its application in the field of open research data repositories. We notably look at how the Invenio technology has been applied in two research data services: (1) the CERN Open Data portal that provides access to the approved open datasets and software of the ALICE, ATLAS, CMS and LHCb collaborations; (2) the Zenodo service that offers an open research data archiving solution to world-wide scientific communities in any research discipline.

Invenio digital library framework is composed of more than sixty independently developed packages on top of the Flask web development environment. The packages share a set of common patterns and communicate together via well-established APIs. The packages come with extensive test suite and example applications and use Travis continuous integration practices to ensure quality. The packages are often developed by independent teams with special focus on topical use cases (e.g. library circulation, multimedia, research data). The separation of packages in the Invenio 3 ecosystem enables their independent
development, maintenance and rapid release cycle. This also allows the prospective digital repository managers who are interested in deploying an Invenio solution at their institutions to cherry-pick the individual modules of interest with the aim of building a customised digital repository solution targeting their particular needs and use cases.

We discuss the application of the Invenio package ecosystem in the research data repository problem domain. We present how a researcher can easily archive their data files as well as their analysis software code or their Jupyter notebooks via GitHub <-> Zenodo integration. The archived data and software is minted with persistent identifiers to ensure their further citeability. We present how the JSON Schema technology is used to define the data model describing all the data managed by the repository. The conformance to versioned JSON schemas ensure the coherence of metadata structure across the managed assets. The data is further indexed using Elasticsearch for the information retrieval needs. We describe the role of the CERN EOS system used as the underlying data storage via a Pythonic XRootD based protocol. Finally, we discuss the role of virtual environments (CernVM) and container-based solutions (Docker) used with the aim of reproducing the archived research data and analyses software even many years after their publication.

**Tertiary Keyword (Optional):**
Experience/plans from outside experimental HEP/NP

**Secondary Keyword (Optional):**
Preservation of analysis and data

**Primary Keyword (Mandatory):**
Collaborative tools

---

**Posters B / Break / 467**

**The Resource Manager of the ATLAS Trigger and Data Acquisition system**

**Author:** Catrin Bernius

**Co-author:** William Panduro Vazquez

1. New York University (US)
2. Royal Holloway, University of London

**Corresponding Authors:** catrin.bernius@cern.ch, j.panduro.vazquez@cern.ch

The Resource Manager is one of the core components of the Data Acquisition system of the ATLAS experiment at the LHC. The Resource Manager marshals the right for applications to access resources which may exist in multiple but limited copies, in order to avoid conflicts due to program faults or operator errors.

The access to resources is managed in a manner similar to what a lock manager would do in other software systems. All the available resources and their association to software processes are described in the Data Acquisition configuration database. The Resource Manager is queried about the availability of resources every time an application needs to be started.

The Resource Manager’s design is based on a client-server model, hence it consists of two components: the Resource Manager “server” application and the “client” shared library. The Resource Manager server implements all the needed functionalities, while the Resource Manager client library provides remote access to the “server” (i.e., to allocate and free resources, to query about the status
During the LHC’s Long Shutdown period, the Resource Manager’s requirements have been reviewed at the light of the experience gained during the LHC’s Run I. As a consequence, the Resource Manager has undergone a full re-design and re-implementation cycle with the result of a reduction of the code base by 40% with respect to the previous implementation.

This contribution will focus on the way the design and the implementation of the Resource Manager could leverage the new features available in the C++11 standard, and how the introduction of external libraries (like Boot multi-container) led to a more maintainable system. Additionally, particular attention will be given to the technical solutions adopted to ensure the Resource Manager could effort the typical requests rates of the Data Acquisition system which is about 30000 requests in a time window of few seconds coming from O(1000) clients.

Secondary Keyword (Optional):
Control systems

Primary Keyword (Mandatory):
DAQ

Tertiary Keyword (Optional):

Posters B / Break / 468

The Payload Inspector: a tool for the visualization of Calibration and Alignment constants stored in the CMS condition database

Author: Audrius Mecionis
Co-authors: Andreas Pfeiffer, Giacomo Govi, Giovanni Franzoni, Marco Musich, Salvatore Di Guida

1 Vilnius University (LT)
2 CERN
3 Fermi National Accelerator Lab. (US)
4 Universite Catholique de Louvain (UCL) (BE)
5 Universita degli Studi Guglielmo Marconi (IT)

Corresponding Authors: giacomo.govi@cern.ch, andreas.pfeiffer@cern.ch, marco.musich@cern.ch, giovanni.franzoni@cern.ch, salvatore.di.guida@cern.ch, audrius.mecionis@cern.ch

The Compact Muon Solenoid (CMS) experiment makes a vast use of alignment and calibration measurements in several data processing workflows. Such measurements are produced either by automated workflows or by analysis tasks carried out by experts in charge. Very frequently, experts want to inspect and exchange with others in CMS the time evolution of a given calibration, or want to monitor the values produced by one of the automated procedures. To address and simplify these operations, a Payload Inspector platform has been introduced as a web-based service to present historical plots and maps of calibrations directly retrieved by the CMS production condition database. The Payload Inspector has been designed to allow for great flexibility in the drawing capabilities while keeping the visualization layer agnostic to the internal structure of the specific calibration. This is achieved through a multi-layered approach: the drawing layer is implemented as a plugin in the CMS offline software which consumes the calibrations while a web-service deals with the visualization aspects. This paper reports the design and development of the Payload Inspector platform, the choice of technologies (python, Flask and bokeh) and reports the operational experience after one year of use.

Primary Keyword (Mandatory):
Visualization
Trigger Menu-aware Monitoring for the ATLAS experiment

Author: Catrin Bernius¹
Co-author: Xanthe Hoad ²

¹ New York University (US)
² University of Edinburgh (GB)

Corresponding Authors: catrin.bernius@cern.ch, xanthe.hoad@cern.ch

Changes in the trigger menu, the online algorithmic event-selection of the ATLAS experiment at the LHC in response to luminosity and detector changes are followed by adjustments in their monitoring system. This is done to ensure that the collected data is useful, and can be properly reconstructed at Tier-0, the first level of the computing grid. During Run 1, ATLAS deployed monitoring updates with the installation of new software releases at Tier-0. This created unnecessary overhead for developers and operators, and unavoidably led to different releases for the data-taking and the monitoring setup.

We present a “trigger menu-aware” monitoring system designed for the ATLAS Run 2 data-taking. The new monitoring system aims to simplify the ATLAS operational workflows, and allows for easy and flexible monitoring configuration changes at the Tier-0 site via an Oracle DB interface. We present the design and the implementation of the menu-aware monitoring, along with lessons from the operational experience of the new system with the 2016 collision data.

Tertiary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Secondary Keyword (Optional):
Trigger

Primary Keyword (Mandatory):
Monitoring

Track 5: Software Development / 470

Everware toolkit - supporting reproducible science and challenge driven education

Authors: Alexander Tiunov¹; Andrey Ustyuzhanin²; Igor Babuschkin³; Tim Head⁴

¹ Higher School of Economics
² Yandex School of Data Analysis (RU)
³ University of Manchester (GB)
⁴ Ecole Polytechnique Federale de Lausanne (CH)
Reproducibility is a fundamental piece of the scientific method and increasingly complex problems demand ever wider collaboration between scientists. To make research fully reproducible and accessible to collaborators a researcher has to take care of several aspects: research protocol description, data access, preservation of the execution environment, workflow pipeline, and analysis script preservation.

Version control systems like git help with the workflow and analysis scripts part. Virtualization techniques like containers or virtual machines help with sharing execution environments. Jupyter notebooks are a powerful tool to capture the computational narrative of a data analysis project.

We present project Everware that seamlessly integrates github/gitlab, Docker and Jupyter helping with a) sharing results of real research and b) boosts education activities. With the help of everware one can not only share the final artifacts of research but all the depth of the research process. This been shown to be extremely helpful during organization of several data analysis hackathons and machine learning schools. Using everware participants could start from an existing solution instead of starting from scratch. They could start contributing immediately.

Everware allows its users to make use of their own computational resources to run the workflows they are interested in, which enables Everware to scale to large numbers of users.

Everware is supported by the Mozilla science lab and Yandex. It is being evaluated as an option for analysis preservation at LHCb. It is an open-source project that welcomes contributions of all kinds at: https://github.com/everware/everware.

Tertiary Keyword (Optional):
Outreach

Secondary Keyword (Optional):
Preservation of analysis and data

Primary Keyword (Mandatory):
Analysis tools and techniques

Track 6: Infrastructures / 471

Deployment of IPv6 only CPU resources at WLCG sites

Authors: Alastair Dewhurst¹; Andrea Sciaba²; Costin Grigoras²; Dave Kelsey¹; Raja Nandakumar¹

Co-authors: Bruno Heinrich Hoeft; Daniel Peter Traynor³; Duncan Rand¹; Edoardo Martelli³; Fernando Lopez Munoz⁵; Francesco Prelz⁶; Jiri Chudoba⁷; Kars Ohrenberg⁸; Kashif Hafeez¹; Marian Babik⁹; Ramiro Voicu⁹; Simon Fayer⁴; Terry Froy¹⁰; Thomas Finnern⁸; Tiju Idiculla¹¹; Ulf Bobson Severin Tigerstedt¹²

¹ STFC - Rutherford Appleton Lab. (GB)
² CERN
³ University of London (GB)
⁴ Imperial College Sci., Tech. & Med. (GB)
⁵ Port d’Informació Científica (ES)
⁶ Università degli Studi e INFN Milano (IT)
⁷ Acad. of Sciences of the Czech Rep. (CZ)
⁸ DESY
⁹ California Institute of Technology (US)
¹⁰ Queen Mary University of London
¹¹ STFC - Rutherford Appleton Laboratory
¹² STFC - Rutherford Appleton Laboratory
The fraction of internet traffic carried over IPv6 continues to grow rapidly. IPv6 support from network hardware vendors and carriers is pervasive and becoming mature. A network infrastructure upgrade often offers sites an excellent window of opportunity to configure and enable IPv6.

There is a significant overhead when setting up and maintaining dual stack machines, so where possible sites would like to upgrade their services directly to IPv6 only. In doing so, they are also expediting the transition process towards its desired completion. While the LHC experiments accept there is a need to move to IPv6, it is currently not directly affecting their work. Sites are unwilling to upgrade if they will be unable to run LHC experiment workflows. This has resulted in a very slow uptake of IPv6 from WLCG sites.

For several years the HEPiX IPv6 Working Group has been testing a range of WLCG services to ensure they are IPv6 compliant. Several sites are now running many of their services as dual stack. The working group, driven by the requirements of the LHC VOs to be able to use IPv6-only opportunistic resources, continues to encourage wider deployment of dual-stack services to make the use of such IPv6-only clients viable.

This paper will present the HEPiX plan and progress so far to allow sites to deploy IPv6 only CPU resources. This will include making experiment central services dual stack as well as a number of storage services. The monitoring, accounting and information services that are used by jobs also needs to be upgraded. Finally the VO testing that has taken place on hosts connected via IPv6 only will be reported.
One of the challenges a scientific computing center has to face is to keep delivering a computational framework well consolidated within the community (i.e. the batch farm), while complying to modern computing paradigms. The aim is to ease system administration at all levels (from hardware to applications) and to provide a smooth end-user experience.

HTCondor is a LRMS widely used in the scientific community and it’s Cloud aware (i.e. it does not resent from a volatile environment where resources might dynamically change). Apache Mesos is a tool that allows to abstract computing resources away from the physical or virtual hosts and to deal with the entire computing infrastructure as a single pool of resources to be shared among services. Within the INDIGO-DataCloud project, we adopted two different approaches to implement a PaaS-level, on-demand Batch Farm Service based on HTCondor and Mesos.

In the first approach, the various HTCondor daemons are packaged inside pre-configured Docker images and deployed as Long Running Service (LRS) through Marathon, profiting from its health checks and failover capabilities.

In the second approach, we have implemented an HTCondor framework for Mesos, that can be used by itself or as a component in the more complex INDIGO PaaS system in conjunction with an orchestration layer like i.e. Marathon. The new framework consists of a scheduler to implement HTCondor policies on the resource offers provided by the Mesos master and a dedicated executor to launch tasks on the slave nodes. The benefits of an ad-hoc framework are first of all a fine-grained level of control on the tasks the application is responsible for. Moreover, it is possible to implement the preferred authorization rules and roles for multi-tenancy and to define application-specific scaling rules. Application isolation and packetization are achieved with the Docker Containerizer module of Mesos.

The most difficult aspects of both approaches concern networking and storage. For instance, the usage of the shared port mode within HTCondor has been evaluated in order to avoid dynamically assigned ports; container-to-container communication and isolation have been addressed exploring solutions based on overlay networks (including e.g. the Calico Project implementation).

Finally, we have studied the possibility to deploy an HTCondor cluster that spans over different sites, also exploiting the Condor Connection Brokering (CCB) component that allows communication across a private network boundary or firewall, as in case of multi-site deployments. Concerning the storage aspects, where factors such as scalability, performance and reliability have to be taken into account, we have explored the usage of CVMFS (using Parrot) and the integration with the INDIGO Data Services (Onedata, Dynafed, FTS).

In this contribution, we are going to describe and motivate our implementative choices and to show the results of the first tests performed.
The DD4hep detector description tool-kit offers a flexible and easy to use solution for the consistent and complete description of particle physics detectors in one single system. The sub-component DDRec provides a dedicated interface to the detector geometry as needed for event reconstruction. With DDRec there is no need to define an additional, separate reconstruction geometry as is often done in HEP, but one can transparently extend the existing detailed simulation model to be also used for the reconstruction.

Based on the extension mechanism of DD4hep, DDRec allows one to attach user defined data structures to detector elements at all levels of the geometry hierarchy. These data structures define a high level view onto the detectors describing their physical properties, such as measurement layers, point resolutions and cell sizes.

For the purpose of charged particles track reconstruction dedicated surface objects can be attached to every volume in the detector geometry. These surfaces provide the measurement directions, local-to-global coordinate transforms and material properties. The material properties, essential for the correct treatment of multiple scattering and energy loss effects, are automatically averaged from the detailed geometry model along the normal of the surface.

Additionally a generic interface allows the user to query material properties at any given point or between any two points in the detector’s world volume.

In this talk we will present DDRec and how it is used together with the generic tracking toolkit aidaTT and the the particle flow package PandoraPFA for full event reconstruction of the ILC detector concepts ILD and SiD and of CLICdp.

This flexible tool chain is also well suited for other future accelerator projects such as FCC and CEPC.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Data processing workflows and frameworks/pipelines

Primary Keyword (Mandatory):

Reconstruction

The Czech National Grid Infrastructure

Authors: Jiri Chudoba¹; Miroslav Ruda¹

¹ CESNET

Corresponding Authors: jiri.chudoba@cesnet.cz, ruda@ics.muni.cz, jiri.chudoba@cern.ch

The Czech National Grid Infrastructure is operated by MetaCentrum, a CESNET department responsible for coordinating and managing activities related to distributed computing. CESNET as the Czech National Research and Education Network (NREN) provides many e-infrastructure services, which are used by 94% of the scientific and research community in the Czech Republic. Computing and storage resources owned by different organizations are connected by fast enough network to provide transparent access to all resources. We describe in more detail the computing infrastructure, which is based on several different technologies and covers grid, cloud and map-reduce environment. While the largest part of CPUs is still accessible via distributed torque servers, providing environment for long batch jobs, part of infrastructure is available via standard EGI tools in EGI, subset of
NGI resources is provided into EGI FedCloud environment with cloud interface and there is also Hadoop cluster provided by the same e-infrastructure. A broad spectrum of computing servers is offered; users can choose from standard 2 CPU servers to large SMP machines with up to 6 TB of RAM or servers with GPU cards. Different groups have different priorities on various resources, resource owners can even have an exclusive access. The software is distributed via AFS. Storage servers offering up to tens of terabytes of disk space to individual users are connected via NFS4 on top of GPFS and access to long term HSM storage with peta-byte capacity is also provided. Overview of available resources and recent statistics of usage will be given.

Secondary Keyword (Optional):
Distributed workload management

Primary Keyword (Mandatory):
Computing facilities

Tertiary Keyword (Optional):

Posters A / Break / 475

Deployment of 464XLAT (RFC6877) alongside IPv6-only CPU resources at WLCG sites

Author: Terry Froy
Co-authors: Christopher John Walker, Daniel Peter Traynor

1 Queen Mary, University of London
2 University of London (GB)

Corresponding Authors: t.froy@qmul.ac.uk, christopher.john.walker@cern.ch, d.traynor@qmul.ac.uk

With many parts of the world having run out of IPv4 address space and the Internet Engineering Task Force (IETF) deprecating IPv4 the use of and migration to IPv6 is becoming a pressing issue. A significant amount of effort has already been expended by the HEPix IPv6 Working Group (http://hepix-ipv6.web.cern.ch/) on testing dual-stacked hosts and IPv6-only CPU resources. The Queen Mary grid site has been at the forefront of adopting IPv6 throughout its cluster and it use within the WLCG. A process to migrate world accessible grid services, such as CREAM, Storm and ARGUS, to be accessible via dual stack IPv4/IPv6 is presented. However, dual stack adds complexity and administrative overhead to sites that may already be starved of resource. This has resulted in a very slow uptake of IPv6 from WLCG sites. 464XLAT (RFC6877) is intended for IPv6 single-stack environments that require the ability to communicate with IPv4-only endpoints, similar to the way IPv4 Network Address Translation (NAT) allows private IPv4 addresses to route communications to public IPv4 addresses around the world. This paper will present a deployment strategy for 464XLAT, operational experiences of using 464XLAT in production at a WLCG site and important information to consider prior to deploying 464XLAT.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):
Network systems and solutions

Track 1: Online Computing / 476
ATLAS Trigger and Data Acquisition Upgrades for High Luminosity LHC

Author: Jose Guillermo Panduro Vazquez
Co-author: Simon George

Corresponding Authors: j.panduro.vazquez@cern.ch, s.george@rhul.ac.uk

The ATLAS experiment at CERN is planning a second phase of upgrades to prepare for the "High Luminosity LHC", a 4th major run due to start in 2026. In order to deliver an order of magnitude more data than previous runs, 14 TeV protons will collide with an instantaneous luminosity of $7.5 \times 10^{34} \text{ cm}^{-2}\text{s}^{-1}$, resulting in much higher pileup and data rates than the current experiment was designed to handle. While this extreme scenario is essential to realise the physics programme, it is a huge challenge for the detector, trigger, data acquisition and computing. The detector upgrades themselves also present new requirements and opportunities for the trigger and data acquisition system.

Initial upgrade designs for the trigger and data acquisition system are shown, including the real time low latency hardware trigger, hardware-based tracking, the high throughput data acquisition system and the commodity hardware and software-based data handling and event filtering. The motivation, overall architecture and expected performance are explained. Some details of the key components are given. Open issues and plans are discussed.

Secondary Keyword (Optional):
DAQ

Primary Keyword (Mandatory):
Trigger

Tertiary Keyword (Optional):

Track 1: Online Computing / 477

The design of a fast Level 1 track trigger for the ATLAS High Luminosity Upgrade

Author: Jose Guillermo Panduro Vazquez
Co-author: Benedict Allbrooke

Corresponding Authors: j.panduro.vazquez@cern.ch, benedict.allbrooke@cern.ch

The ATLAS experiment at the high-luminosity LHC will face a five-fold increase in the number of interactions per collision relative to the ongoing Run 2. This will require a proportional improvement in rejection power at the earliest levels of the detector trigger system, while preserving good signal efficiency.

One critical aspect of this improvement will be the implementation of precise track reconstruction, through which sharper turn-on curves, b-tagging and tau-tagging techniques can in principle be implemented. The challenge of such a project comes in the development of a fast, precise custom electronic device integrated in the hardware-based first trigger level of the experiment, with repercussions propagating as far as the detector read-out philosophy.
This talk will discuss the projected performance of the system in terms of tracking, timing and physics.

Tertiary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Secondary Keyword (Optional):
Reconstruction

Primary Keyword (Mandatory):
Trigger

Posters B / Break / 478

The new inter process communication middle-ware for the ATLAS Trigger and Data Acquisition system.

Author: Jose Guillermo Panduro Vazquez

1 Royal Holloway, University of London

Corresponding Author: j.panduro.vazquez@cern.ch

The ATLAS Trigger & Data Acquisition project was started almost twenty years ago with the aim of providing a scalable distributed data collection system for the experiment. While the software dealing with physics dataflow was implemented by directly using low level communication protocols, like TCP and UDP, the control and monitoring infrastructure services for the system were implemented on top of the CORBA communication middle-ware. CORBA provides a high-level object oriented abstraction for inter-process communication, hiding communication complexity from the developers. This approach speeds up and simplifies development of communication services but incurs some extra cost in terms of performance and resource overhead.

The ATLAS experience of using CORBA for control and monitoring data exchange in the distributed trigger and data acquisition system has been very successful, mostly due to the outstanding quality of the CORBA brokers which were used in the project: omniORB for C++ and JacORB for Java. However, due to a number of shortcomings and technical issues the CORBA standard has been gradually loosing its initial popularity in the last decade and the availability of long term support for the open source implementations of CORBA is becoming uncertain. Taking into account the time scale of the ATLAS experiment, which goes beyond the next two decades, the trigger and data acquisition infrastructure team reviewed the requirements for inter-process communication middle-ware and performed a survey of the communication software market in order to access modern technologies developed in recent years. Based on the result of that survey several technologies have been evaluated, estimating the long term benefits and drawbacks of using them as a possible replacement for CORBA during the next long LHC shutdown which is scheduled for two years from now. The evaluation was recently concluded with the recommendation to use a communication library called ZeroMQ in place of CORBA.

This presentation will discuss the methodology and results of the evaluation as well as the plans for organizing the migration from CORBA to ZeroMQ.

Tertiary Keyword (Optional):
Control systems
The KEK central computer system (KEKCC) supports various activities in KEK, such as the Belle / Belle II, J-PARC experiments, etc. The system is now under replacement and will be put into production in September 2016. The computing resources, CPU and storage, in the next system are much enhanced as recent increase of computing resource demand. We will have 10,000 CPU cores, 13 PB disk storage, and 70 PB maximum capacity of tape system.

Grid computing can help distribute large amount of data in geographically dispersed sites and share data in an efficient way for world wide collaborations. But the data centers of host institutes of large HEP experiments have to take into serious consideration for managing huge amount of data. For example, the Belle II experiment expects that several hundred PB storage has to be stored in the KEK site even if Grid computing is taken as an analysis model. The challenge is not only for storage capacity. I/O scalability, usability and power efficiency and so on should be considered for the system design of storage system. Our storage system is designed to meet requirements for managing 100 PB-order data. We introduce IBM Elastic Storage Server (ESS) and DDN SFA12K as storage hardware, and adopts GPFS parallel file system to realize high I/O performance. The GPFS file system can take several tiers for recalling data from local SSD cache in computing nodes to HDD and tape storage. We take full advantage of this hierarchical storage management in the next system. Actually we have long history of using HPSS system as HSM of tape system. Since the current system, we introduced GHI (GPFS HPSS Interface) as the layer of disk and tape system, which enables high I/O performance and good usability as GPFS disk file system for tape data.

In this talk, we mainly focus on the design and performance of our new storage system. In addition, issues on workload management, system monitoring, data migration and so on are described. Our knowledge, experience and challenges can be usefully shared among HEP data centers as a data-intensive computing facility for the next generation of HEP experiments.
FELIX: the new detector readout system for the ATLAS experiment

Author: Jose Guillermo Panduro Vazquez
Co-author: Soo Ryu

1 Royal Holloway, University of London
2 Argonne National Laboratory (US)

After the Phase-I upgrade and onward, the Front-End Link eXchange (FELIX) system will be the interface between the data handling system and the detector front-end electronics and trigger electronics at the ATLAS experiment. FELIX will function as a router between custom serial links and a commodity switch network which will use standard technologies (Ethernet or Infiniband) to communicate with data collecting and processing components. The system architecture of FELIX will be described and the results of the demonstrator program currently in progress will be presented.

Tertiary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Secondary Keyword (Optional):
Network systems and solutions

Primary Keyword (Mandatory):
DAQ

Multi-Threaded Algorithms for General purpose Graphics Processor Units in the ATLAS High Level Trigger

Author: Jose Guillermo Panduro Vazquez
Co-author: Patricia Conde Muino

1 Royal Holloway, University of London
2 LIP Laboratorio de Instrumentacao e Fisica Experimental de Part

General purpose Graphics Processor Units (GPGPU) are being evaluated for possible future inclusion in an upgraded ATLAS High Level Trigger farm. We have developed a demonstrator including GPGPU implementations of Inner Detector and Muon tracking and Calorimeter clustering within the ATLAS software framework. ATLAS is a general purpose particle physics experiment located on the LHC collider at CERN. The ATLAS Trigger system consists of two levels, with level 1 implemented in hardware and the High Level Trigger implemented in software running on a farm of commodity CPU.

The High Level Trigger reduces the trigger rate from the 100 kHz level 1 acceptance rate to 1 kHz for recording, requiring an average per-event processing time of ~250 ms for this task. The selection in the high level trigger is based on reconstructing tracks in the Inner Detector and Muon Spectrometer and clusters of energy deposited in the Calorimeter. Performing this reconstruction within the
available farm resources presents a significant challenge that will increase significantly with future LHC upgrades. During the LHC data-taking period starting in 2021, luminosity will reach up to three times the original design value. Luminosity will increase further to 7.5 times the design value in 2026 following LHC and ATLAS upgrades. Corresponding improvements in the speed of the reconstruction code will be needed to provide the required trigger selection power within affordable computing resources.

Key factors determining the potential benefit of including GPGPU as part of the HLT processor farm are the relative speed of the CPU and GPU algorithm implementations, the relative execution times of the GPU algorithms and serial code remaining on the CPU, the number of GPU required and the relative financial cost of the selected GPU. We give a brief overview of the algorithms implemented and present new measurements that compare the performance of various configurations exploiting different GPU cards.

Tertiary Keyword (Optional):
Algorithms

Secondary Keyword (Optional):
Parallelizarion

Primary Keyword (Mandatory):
Trigger

Posters B / Break / 483

The ATLAS Data Acquisition System LHC Run 2

Author: Jose Guillermo Panduro Vazquez

1 Royal Holloway, University of London

Corresponding Author: j.panduro.vazquez@cern.ch

The LHC has been providing pp collisions with record luminosity and energy since the start of Run 2 in 2015. In the ATLAS experiment the Trigger and Data Acquisition system has been upgraded to deal with the increased event rates. The dataflow element of the system is distributed across hardware and software and is responsible for buffering and transporting event data from the Readout system to the High Level Trigger and on to event storage. The dataflow system has been reshaped in order benefit from technological progress and to maximize the flexibility and efficiency of the data selection process.

The updated dataflow system is radically different from the previous implementation both in terms of architecture and performance. The previous two level software filtering architecture, known as L2 and the Event Filter, have been merged with the Event Builder function into a single process, performing incremental data collection and analysis. This design has many advantages, among which are: radical simplification of the architecture, flexible and automatically balanced distribution of computing resources, and the sharing of code and services on nodes. In addition, logical farm slicing, with each slice managed by a dedicated supervisor, has been dropped in favour of global management by a single farm master operating at 100 kHz. This farm master has also since been integrated with a new software based region of interest builder, replacing the previous VMEbus bases system.

The Data Collection network, that connects the HLT processing nodes to the Readout and the storage systems has evolved to provide network connectivity as required by the new dataflow architecture. The old Data Collection and Back-End networks have been merged into a single Ethernet network and the Readout PCs have been directly connected to the network cores. The aggregate throughput and port density have been increased by an order of magnitude and the introduction of Multi-Chassis Trunking significantly enhanced fault tolerance and redundancy. The Readout system itself has been
completely refitted with new higher performance, lower footprint server machines housing a new custom front-end interface card.

This presentation will cover overall design of the system, along with performance results from the start up phase of LHC Run 2.

Tertiary Keyword (Optional):
Distributed data handling

Secondary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Primary Keyword (Mandatory):
DAQ

Track 5: Software Development / 484

Investigations of Future Computing Platforms for High Energy Physics

Authors: Aritz Brosa Iartza\textsuperscript{1}; Daniel Hugo Campora Perez\textsuperscript{2}; David Abdurachmanov\textsuperscript{3}; Felice Pantaleo\textsuperscript{4}; Liviu Valsan\textsuperscript{5}; Marco Guerri\textsuperscript{5}; Niko Neufeld\textsuperscript{6}; Omar Awile\textsuperscript{5}; Pawel Szostek\textsuperscript{5}; Peter Elmer\textsuperscript{6}; Robert Knight\textsuperscript{6}

\textsuperscript{1} Universidad de Oviedo (ES) \\
\textsuperscript{2} Universidad de Sevilla (ES) \\
\textsuperscript{3} Fermi National Accelerator Lab. (US) \\
\textsuperscript{4} CERN - Universität Hamburg \\
\textsuperscript{5} CERN \\
\textsuperscript{6} Princeton University (US)

Corresponding Authors: aritz.brosa.iartza@cern.ch, niko.neufeld@cern.ch, daniel.hugo.campora.perez@cern.ch, david.abdurachmanov@cern.ch, pawel.szostek@cern.ch, knight@princeton.edu, felice.pantaleo@cern.ch, peter.elmer@cern.ch, liviu.valsan@cern.ch, marco.guerri@cern.ch, omar.awile@cern.ch

Around the year 2000, the convergence on Linux and commodity x86_64 processors provided a homogeneous scientific computing platform which enabled the construction of the Worldwide LHC Computing Grid (WLCG) for LHC data processing. In the last decade the size and density of computing infrastructure has grown significantly. Consequently, power availability and dissipation have become important limiting factors for modern data centres. The on-chip power density limitations, which have brought us into the multicore era, are driving the computing market towards a greater variety of solutions. This, in turn, necessitates a broader look at the future computing infrastructure.

Given the planned High Luminosity LHC and detector upgrades, changes are required to the computing models and infrastructure to enable data processing at increased rates through the early 2030s. Understanding how to maximize throughput for minimum initial cost and power consumption will be a critical aspect for computing in HEP in the coming years.

We present results from our work to compare performance and energy efficiency for different general purpose architectures, such as traditional x86_64 processors, ARMv8, PowerPC 64-bit, as well as specialized parallel architectures, including Xeon Phi and GPUs. In our tests we use a variety of HEP-related benchmarks, including HEP-SPEC 2006, GEANT4 ParFullCMS and the realistic production codes from the LHC experiments. Finally we conclude on the suitability of the architectures under test for the future computing needs of HEP data centres.
The ATLAS Level-1 Topological Trigger Performance in Run 2

Author: Jose Guillermo Panduro Vazquez
Co-author: Imma Riu

1 Royal Holloway, University of London
2 IFAE Barcelona (ES)

Corresponding Authors: j.panduro.vazquez@cern.ch, imma.riu@cern.ch

The LHC will collide protons in the ATLAS detector with increasing luminosity through 2016, placing stringent operational and physical requirements to the ATLAS trigger system in order to reduce the 40 MHz collision rate to a manageable event storage rate of about 1 kHz, while not rejecting interesting physics events. The Level-1 trigger is the first rate-reducing step in the ATLAS trigger system with an output rate of up to 100 kHz and decision latency smaller than 2.5 μs. It consists of a calorimeter trigger, muon trigger and a central trigger processor.

During the LHC shutdown after the Run 1 finished in 2013, the Level-1 trigger system was upgraded at hardware, firmware and software levels. In particular, a new electronics sub-system was introduced in the real-time data processing path: the Topological Processor System (L1Topo). It consists of a single AdvancedCTA shelf equipped with two Level-1 topological processor blades. They receive real-time information from the Level-1 calorimeter and muon triggers, which is processed by four individual state-of-the-art FPGAs. It needs to deal with a large input bandwidth of up to 6 Tb/s, optical connectivity and low processing latency on the real-time data path. The L1Topo firmware includes measurements of angles between jets and/or leptons and determination of kinematic variables based on lists of selected or sorted trigger objects. All these complex calculations are executed in hardware within 200 ns. Over one hundred VHDL algorithms are producing trigger outputs that are incorporated into the logic of the central trigger processor, responsible of generating the Level-1 acceptance signal. The detailed additional information provided by L1Topo, will improve the ATLAS physics reach in a harsher collisions environment.

The system has been installed and commissioning started during 2015 and continued during 2016. As part of the firmware commissioning, the physics output from individual algorithms needs to be simulated and compared with the hardware response. An overview of the design, commissioning process and early impact on physics results of the new L1Topo system will be illustrated.
A web-based solution to visualize operational monitoring data in the Trigger and Data Acquisition system of the ATLAS experiment at the LHC

Author: Jose Guillermo Panduro Vazquez
Co-author: Jorn Schumacher

1 Royal Holloway, University of London
2 University of Paderborn (DE)

Corresponding Authors: j.panduro.vazquez@cern.ch, joern.schumacher@cern.ch

The Trigger and Data Acquisition system of the ATLAS detector at the Large Hadron Collider at CERN is composed of a large number of distributed hardware and software components (about 3000 machines and more than 25000 applications) which, in a coordinated manner, provide the data-taking functionality of the overall system.

During data taking runs, a huge flow of operational data is produced in order to constantly monitor the system and allow proper detection of anomalies or misbehaviors. In the ATLAS trigger and data acquisition system, operational data are archived and made available to applications by the P-Beast (Persistent Back-End for the Atlas Information System of TDAQ) service, implementing a custom time-series database.

The possibility to efficiently visualize both real-time and historical operational data is a great asset facilitating both online identification of problems and post-mortem analysis. This paper will present a web-based solution developed to achieve such a goal: the solution leverages the flexibility of the P-Beast archiver to retrieve data, and exploits the versatility of the Grafana dashboard builder to offer a very rich user experience. Additionally, particular attention will be given to the way some technical challenges (like the efficient visualization of a huge amount of data and the integration of the P-Beast data source in Grafana) have been faced and solved.

Frameworks to monitor and predict resource usage in the ATLAS High Level Trigger

Author: Jose Guillermo Panduro Vazquez
Co-author: Tim Martin

1 Royal Holloway, University of London
2 University of Warwick (GB)
The ATLAS High Level Trigger Farm consists of around 30,000 CPU cores which filter events at up to 100 kHz input rate. A costing framework is built into the high level trigger, this enables detailed monitoring of the system and allows for data-driven predictions to be made utilising specialist datasets. This talk will present an overview of how ATLAS collects in-situ monitoring data on both CPU usage and dataflow over the data-acquisition network during the trigger execution, and how these data are processed to yield both low level monitoring of individual selection-algorithms and high level data on the overall performance of the farm. For development and prediction purposes, ATLAS uses a special 'Enhanced Bias' event selection. This mechanism will be explained along with how is used to profile expected resource usage and output event-rate of new physics selections, before they are executed on the actual high level trigger farm.

Tertiary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Secondary Keyword (Optional):
DAQ

Primary Keyword (Mandatory):
Trigger

Track 6: Infrastructures / 488

High-Throughput Network Communication with NetIO

Author: Jose Guillermo Panduro Vazquez

Co-author: Jorn Schumacher

1 Royal Holloway, University of London
2 University of Paderborn (DE)

NetIO is a simple, novel asynchronous message service that can operate on Ethernet, Infiniband and similar network fabrics. In our presentation we describe HPC network technologies like Infiniband, TrueScale or OmniPath provide low-latency and high-throughput communication between hosts, which makes them attractive options for data-acquisition systems in large-scale high-energy physics experiments. Like HPC networks, data acquisition networks are local and include a well specified number of systems. Unfortunately traditional network communication APIs for HPC clusters like MPI or PGAS exclusively target the HPC community and are not well suited for data acquisition applications. It is possible to build distributed data acquisition applications using low-level system APIs like Infiniband Verbs, but it requires non negligible effort and expert knowledge.

On the other hand, message services like 0MQ have gained popularity in the HEP community. Such APIs facilitate the building of distributed applications with a high-level approach and provide good performance. Unfortunately their usage usually limits developers to TCP/IP-based networks. While it is possible to operate a TCP/IP stack on top of Infiniband and OmniPath, this approach may not be very efficient compared to direct use of native APIs.

NetIO is a simple, novel asynchronous message service that can operate on Ethernet, Infiniband and similar network fabrics. In our presentation we describe
the design and implementation of NetIO, evaluate its use in comparison to other approaches and show performance studies.

NetIO supports different high-level programming models and typical workloads of HEP applications. The ATLAS front end link exchange project successfully uses NetIO as its central communication platform.

The NetIO architecture consists of two layers:

- The outer layer provides users with a choice of several socket types for different message-based communication patterns. At the moment NetIO features a low-latency point-to-point send/receive socket pair, a high-throughput point-to-point send/receive socket pair, and a high-throughput publish/subscribe socket pair.
- The inner layer is pluggable and provides a basic send/receive socket pair to the upper layer to provide a consistent, uniform API across different network technologies.

There are currently two working backends for NetIO:

- The Ethernet backend is based on TCP/IP and POSIX sockets.
- The Infiniband backend relies on libfabric with the Verbs provider from the OpenFabrics Interfaces Working Group.

The libfabric package also supports other fabric technologies like jWarp, Cisco usNic, Cray GNI, Mellanox MXM and others. Via PSM and PSM2 it also natively supports Intel TrueScale and Intel OmniPath. Since libfabric is already used for the Infiniband backend, we do not foresee major challenges for porting NetIO to OmniPath, and a native OmniPath backend is currently under development.

Tertiary Keyword (Optional):
Computing middleware

Secondary Keyword (Optional):
Distributed data handling

Primary Keyword (Mandatory):
Network systems and solutions

The HEP Software and Computing Knowledge Base

Author: Torre Wenaus

1 Brookhaven National Laboratory (US)

Corresponding Author: wenaus@gmail.com

HEP software today is a rich and diverse domain in itself and exists within the mushrooming world of open source software. As HEP software developers and users we can be more productive and effective if our work and our choices are informed by a good knowledge of what others in our community have created or found useful. The HEP Software and Computing Knowledge Base, hep-software.org, was created to facilitate this by serving as a collection point and information exchange on software projects and products, services, training, computing facilities, and relating them to the
projects, experiments, organizations and science domains that offer them or use them. It was created as a contribution to the HEP Software Foundation, for which a HEP S&C knowledge base was a much requested early deliverable. This contribution will motivate and describe the system, what it offers, its content and contributions both existing and needed, and its implementation (node.js based web service and javascript client app) which has emphasized ease of use for both users and contributors.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Software development process and tools

Primary Keyword (Mandatory):
Collaborative tools

Abstracting application deployment on Cloud infrastructures

Authors: Davide Salomoni¹; Doina Cristina Aiftimiei²; Enrico Fattibene³; Matteo Panella¹; Riccardo Gargana⁴

¹ Universita e INFN, Bologna (IT)
² INFN - CNAF, IFIN-HH
³ INFN - National Institute for Nuclear Physics
⁴ INFN

Corresponding Authors: davide.salomoni@cnaf.infn.it, matteo.panella@cnaf.infn.it, riccardo.gargana@lnf.infn.it, enrico.fattibene@cnaf.infn.it, cristina.aiftimiei@cnaf.infn.it

Deploying a complex application on a Cloud-based infrastructure can be a challenging task. Among other things, the complexity can derive from software components the application relies on, from requirements coming from the use cases (i.e. high availability of the components, autoscaling, disaster recovery), from the skills of the users that have to run the application.

Using an orchestration service allows to hide the complex deployment of the application components, the order of each of them in the instantiation process and the relationships among them. In order to further simplify the application deployment to users not familiar with Cloud infrastructures and above layers, it can be worthwhile to provide an abstraction layer on top of the orchestration one.

In this contribution we present an approach for Cloud-based deployment of applications and its implementation in the framework of several projects, such as "CHAOS: a cloud of controls, a project funded by MIUR (Italian Ministry of Research and Education) to create a Cloud-based deployment of a control system and data acquisition framework, "INDIGO-DataCloud", an EC H2020 project targeting among other things high-level deployment of applications on hybrid Clouds, and "Open City Platform", an Italian project aiming to provide open Cloud solutions for Italian Public Administrations.

Through orchestration services, we prototyped a dynamic, on-demand, scalable platform of software components, based on OpenStack infrastructures. A set of Heat templates developed ad-hoc allow to automatically deploy all the application components, minimize the faulty situations and guarantee the same configuration every time they run. The automatic orchestration is an example of Platform as a Service, that can be instantiated both via command-line or OpenStack dashboard, presuming a certain level of knowledge of OpenStack usage.

On top of the orchestration services we developed a prototype of a web interface exploiting the Heat APIs, that can be related to a specific application provided that ad-hoc Heat templates are available. The user can start an instance of the application without having knowledge about the underlying Cloud infrastructure and services. Moreover, the platform instance can be customized by choosing parameters related to the application such as the size of a File System or the number of instances of a NoSQL DB cluster. As soon as the desired platform is running, the web interface offers the
possibility to scale some infrastructure components. By providing this abstraction layer, users have a simplified access to Cloud resources and data center administrators can limit the degrees of freedom granted to the users.

In this contribution we describe the solution design and implementation, based on the application requirements, the details of the development of both the Heat templates and of the web interface, together with possible exploitation strategies of this work in Cloud data centers.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):
Cloud technologies

Track 3: Distributed Computing / 492

Expanding the user base beyond HEP for the Ganga distributed analysis user interface

Authors: Robert Andrew Currie¹; Ulrik Egede¹

Co-authors: Alexander John Richards ¹; Mark Slater ²; Matt Williams ³

¹ Imperial College Sci., Tech. & Med. (GB)
² Birmingham University
³ University of Birmingham (GB)

Corresponding Authors: mws@hep.ph.bham.ac.uk, rcurrie@cern.ch, matthew.peter.williams@gmail.com, alexander.richards@cern.ch, u.egede@imperial.ac.uk

This talk will present the result of recent developments to support new users from the Large Scale Survey Telescope (LSST) group on the GridPP DIRAC instance. I will describe a workflow used for galaxy shape identification analyses whilst highlighting specific challenges as well as the solutions currently being explored. The result of this work allows this community to make best use of available computing resources.

The LSST workflow is CPU limited, producing a large amount of highly distributed output which is managed and collected in an automated way for the user. We have made use of the Ganga distributed analysis user interface to manage physics-driven workflows with large numbers of user generated jobs.

I will also present the successes of working with a new user community to take advantage of HEP related computing resources as this community migrates to make use of a more distributed computing environment.

Tertiary Keyword (Optional):
Preservation of analysis and data

Secondary Keyword (Optional):
Outreach

Primary Keyword (Mandatory):
Experience/plans from outside experimental HEP/NP
Efficient monitoring of CRAB3 jobs at CMS

Authors: Emilis Antanas Rupeika\(^1\); Marco Mascheroni\(^2\)

Co-authors: Diego Ciangottini \(^3\); Eric Vaandering \(^2\); Jadir Marra Da Silva \(^4\); Jose Hernandez \(^5\); Justas Balcas \(^6\); Stefano Belforte \(^7\)

\(^1\) Vilnius University (LT)  
\(^2\) Fermi National Accelerator Lab. (US)  
\(^3\) Universita e INFN, Perugia (IT)  
\(^4\) UNESP - Universidade Estadual Paulista (BR)  
\(^5\) CIEMAT  
\(^6\) California Institute of Technology (US)  
\(^7\) Universita e INFN, Trieste (IT)

Corresponding Authors: emilis.antanas.rupeika@cern.ch, chema.hernandez.calama@gmail.com, marco.mascheroni@cern.ch, stefano.belforte@cern.ch, jadir.marra.da.silva@cern.ch, diego.ciangottini@cern.ch, ewv@fnal.gov, justas.balcas@cern.ch

CRAB3 is a tool used by more than 500 users all over the world for distributed Grid analysis of CMS data. Users can submit sets of Grid jobs with similar requirements (tasks) with a single user request. CRAB3 uses a client-server architecture, where a lightweight client, a server, and ancillary services work together and are maintained by CMS operators at CERN.

As with most complex software, good monitoring tools are crucial for efficient use and long-term maintainability. This work gives an overview of the monitoring tools developed to ensure the CRAB3 server and infrastructure are functional, help operators debug user problems, and minimize overhead and operating cost.

CRABMonitor is a dedicated javascript-based monitoring page dedicated to monitoring system operation. It gathers the results from multiple CRAB3 APIs exposed via a REST interface. It is used by CRAB3 operators to monitor the details of submitted jobs, task status, configuration and parameters, user code, and log files. Links to relevant data are provided when a problem must be investigated. The software is largely javascript developed in-house to maximize flexibility and maintainability, although jQuery is also utilized.

In addition to CRABMonitor, a range of Kibana "dashboards" which have been developed to provide real-time monitoring of system operations will also be presented.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Monitoring

Primary Keyword (Mandatory):

Distributed workload management

Track 3: Distributed Computing / 494

Use of DAGMan in CRAB3 to improve the splitting of CMS user jobs

Authors: Marco Mascheroni\(^2\); Matthias Wolf\(^2\)
CRAB3 is a workload management tool used by more than 500 CMS physicists every month to analyze data acquired by the Compact Muon Solenoid (CMS) detector at the CERN Large Hadron Collider (LHC). CRAB3 allows users to analyze a large collection of input files (datasets), splitting the input into multiple Grid jobs depending on parameters provided by users.

The process of manually specifying exactly how a large project is divided into jobs is tedious and often results in sub-optimal splitting due to its dependence on the performance of the user code and the content of the input dataset. This introduces two types of problems; jobs that are too big will have excessive runtimes and will not distribute the work across all of the available nodes. However, splitting the project into a large number of very small jobs is also inefficient, as each job creates additional overhead which increases load on scheduling infrastructure resources.

In this work we present a new feature called “automatic splitting” which removes the need for users to manually specify job splitting parameters. We discuss how HTCondor DAGMan can be used to build dynamic Directed Acyclic Graphs (DAGs) on the fly to optimize the performance of large CMS analysis jobs on the Grid.

We use DAGMan to dynamically generate interconnected DAGs that estimate the time per event of the user code, then run a set of jobs of preconfigured runtime to analyze the dataset. If some jobs have terminated before completion, the unfinished portions are assembled into smaller jobs and resubmitted to the worker nodes.
Corresponding Authors: jacob.walker@cern.ch, giovanni.franzoni@cern.ch, phat.srimonobhas@gmail.com, antanas.norkus@cern.ch, adrian.pol@cern.ch

Physics analysis at the Compact Muon Solenoid (CMS) requires both a vast production of simulated events and an extensive processing of the data collected by the experiment.

Since the end of the LHC runI in 2012, CMS has produced over 20 Billion simulated events, from 75 thousand processing requests organised in one hundred different campaigns, which emulate different configurations of collision events, CMS detector and LHC running conditions. In the same time span, sixteen data processing campaigns have taken place to reconstruct different portions of the runI and runII data with ever improving algorithms and calibrations.

The scale and complexity of the events simulation and processing and the requirement that multiple campaigns must proceed in parallel, demand that a comprehensive, frequently updated and easily accessible monitoring be made available to the CMS collaboration.

Such monitoring must serve both the analysts, who want to know which and when datasets will become available, and the central teams in charge of submitting, prioritizing and running the requests across the distributed computing infrastructure of CMS.

The Production Monitoring Platform (pMp) web-based service, has been developed in 2015 to address those needs. It aggregates information from multiple services used to define, organize and run the processing requests; pMp updates hourly a dedicated Elastic database, and provides multiple configurable views to assess the status of single datasets as well as entire production campaigns.

This contribution will cover the pMp development, the evolution of its functionalities and one and half year of operational experience.

Secondary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Primary Keyword (Mandatory):
Monitoring

Tertiary Keyword (Optional):

Track 3: Distributed Computing / 496

Managing the CMS Data and Monte Carlo Processing during LHC Run 2

Author: Christoph Wissing¹

¹ Deutsches Elektronen-Synchrotron (DE)

Corresponding Author: christoph.wissing@desy.de

The CMS Computing and Offline group has put in a number of enhancements into the main software packages and tools used for centrally managed processing and data transfers in order to cope with the challenges expected during the LHC Run 2. In the presentation we will highlight these improvements that allow CMS to deal with the increased trigger output rate and the increased collision pileup in the context of the evolution in computing technology. The overall system aims for higher usage efficiency through increased automation and enhanced operational flexibility in terms of data transfers (dynamic) and workflow handling. The tight coupling of workflow classes to types of sites has been drastically reduced. Reliable and high-performing networking between most of the computing sites and the successful deployment of a data-federation allow the execution of workflows using remote data access. Another step towards flexibility has been the introduction of one large global HTCondor Pool for all types of processing workflows and analysis jobs implementing the 'late binding' principle. Besides classical grid resources also some opportunistic resources as well
Configuration validation in the art event-processing framework

Author: Kyle Knoepfel

1 Fermi National Accelerator Laboratory

Corresponding Author: knoepfel@fnal.gov

One of the difficulties experimenters encounter when using a modular event-processing framework is determining the appropriate configuration for the workflow they intend to execute. A typical solution is to provide documentation external to the C++ code source that explains how a given component of the workflow is to be configured. This solution is fragile, because the documentation and the code will tend to diverge. A better solution is to implement a configuration-checking system that is embedded into the C++ code source itself. With modern C++ techniques, it is possible to cleanly (and concisely) implement a configuration-validation system that self-documents an allowed configuration and validates and provides access to a user-provided configuration. I will be presenting such a system as implemented in the art framework. The techniques used, however, can be applied to any system that represents a user-provided configuration as a C++ object.

Using Docker in HEP

Author: Marc Paterno

Co-authors: Adam Lyon; Jim Kowalkowski

1 Fermilab
Docker is a container technology that provides a way to "wrap up a piece of software in a complete filesystem that contains everything it needs to run". We have experimented with Docker to investigate its utility in three broad realms: (1) allowing existing complex software to run in very different environments from that in which the software was built (such as Cori, NERSC’s newest supercomputer), (2) as a means of delivering the same development environment to multiple operating systems (including laptops), and allowing the use of tools from both the host and container system to their best advantage, and (3) as a way of encapsulating entire software suites (in particular, a popular cosmology-based MCMC parameter estimation system), allowing them to be supported for use on multiple operating systems without additional effort. We report on the strengths and weaknesses of Docker for the HEP community, and show results (including performance) from our experiments.

1 "What is Docker?", https://www.docker.com/what-docker.

Posters B / Break / 499

Research on an universal Openstack upgrade solution

Author: Tao Cui

Co-author: Yaodong Cheng

1 IHEP(Institute of High Energy Physics, CAS,China)
2 IHEP

Corresponding Authors: cuit@ihep.ac.cn, chyd@ihep.ac.cn

IhepCloud is a multi-user virtualization platform which based on Openstack icehouse and deployed at Nov. 2014. The platform provides multiple types virtual machine, such as test VM, UI and WN, is a part of local computing system. There are 21 physical machines and 120 users on this platform and about 300 virtual machines running on it. Upgrading IhepCloud from Icehouse to Kilo is difficult, because there is big change from Icehouse to Kilo. Kilo requests SL7 and has difference database structure. Thus we must reinstall OS and adjust auxiliary components such NETDB we developed. The upgrading must ensure the integrity and consistency of user information, quota, security group, operation strategy, network topology and their virtual machines. Finally, the upgrading should consider improvement of the new platform and make it better. The traditional approach is to rebuild the platform based on new Openstack version and users re-launched new VMs and migrates data. Original virtual machines environment will be lost.

We discuss a universal solution which can upgrade multi-user Openstack platform. This is a migration solution from old version platform to new one. Platform information such security group, operation strategy, network configuration and user quota and so on will re-configure on new platform. Some operations, such as virtual machine user changes, need to directly modify the database. Virtual machines from old platform will be migrated to new platform with a short downtime. The
solution avoids complex upgrading steps based on old platform and is helpful to re-deploy a new platform with more advanced technologies and even new network topology. This solution is suitable for large-span Openstack version migration. In this paper, we will present the detail how to migrate from Icehouse to Kilo on IhepCloud.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):
Cloud technologies

Posters B / Break / 500

The end-to-end network performance troubleshooting cookbook

Author: Vincenzo Capone

Co-author: Richard Hughes-Jones

Corresponding Authors: richard.hughes-jones@geant.org, vincenzo.capone@geant.org

The growth in size and geographical distribution of scientific collaborations, while enabling researcher to achieve always higher and bolder results, also poses new technological challenges, one of these being the additional efforts to analyse and troubleshoot network flows that travel for thousands of miles, traversing a number of different network domains. While the day-to-day multi-domain monitoring and fault detection and handling procedures are firmly established and agreed on by the network operators in the R&E community, a cleverer end-to-end traffic analysis and troubleshooting is still something users are in need of, since the network providers not always have specific tools in place aimed to deal with this category of problems.

The well-known perfSONAR framework makes available to the users several testing instruments able to investigate a number of transmission parameters, like latency, jitter and end-to-end throughput measurement. Notwithstanding its high effectiveness in testing the path between two networks, a proper end-to-end monitoring between two servers in production is beyond the reach of perfSONAR.

Indeed, a single (either software or hardware) testing tool will not be capable of grasping the complete end-to-end performance analysis including all the pieces that take part in the data transfer between two end points. When a data movement happens, what we have is a long series of interactions between several components and domains, starting with a storage device (that could be further divided into even smaller parts: hard disk, controller, FC network, etc.), through a CPU, a network interface card, a switch (more likely, a whole LAN), a firewall, a router, then a network provider, then one or more backbone networks, and then the reverse path when approaching the other end of the data transmission. Not to mention the many software elements involved, including the science-specific applications, the authentication and authorization tools, the operating system sub-components, and so on.

It’s then clear that what is needed to face this challenge is a set of techniques and good practices that leverage on different tools, able to interact with the different layers and interaction domains, which collectively form the end-to-end application data transfer.

What we will present is a structured and systematic approach to a complete and effective network performance analysis, which can be carried out by any networking or system manager with proper access rights to the local infrastructure. The talk will explain the different domains on which the analysis needs to be performed, identifying the most appropriate tools to use and parameters to measure, that collectively will likely lead to find out where, along the path, the problem lies.

Tertiary Keyword (Optional):
Track 4: Data Handling / 501

Accessing Data Federations with CVMFS

Author: Derek John Weitzel

Co-authors: Brian Paul Bockelman ¹; Dave Dykstra ²; Jakob Blomer ³; Rene Meusel ³

¹ University of Nebraska (US)  
² Fermi National Accelerator Lab. (US)  
³ CERN

Corresponding Authors: rene.meusel@cern.ch, jakob.blomer@cern.ch, dwd@fnal.gov, brian.bockelman@cern.ch, derek.weitzel@cern.ch

Data federations have become an increasingly common tool for large collaborations such as CMS and Atlas to efficiently distribute large data files. Unfortunately, these typically come with weak namespace semantics and a non-POSIX API. On the other hand, CVMFS has provided a POSIX-compliant read-only interface for use cases with a small working set size (such as software distribution). The metadata required for the CVMFS POSIX interface distributed through a caching hierarchy, allowing it to scale to the level of about a hundred thousand hosts. In this paper, we will describe our contributions to CVMFS that merges the data scalability of XRootD-based data federations (such as AAA) with metadata scalability and POSIX interface of CVMFS. We modified CVMFS so it can serve unmodified files without copying them to the repository server. CVMFS 2.2.0 is also able to redirect requests for data files to servers outside of the CVMFS content distribution network. Finally, we added the ability to manage authorization and authentication using security credentials such as X509 proxy certificates. We combined these modifications with the OSG’s StashCache regional XRootD caching infrastructure to create a cached data distribution network. We will show performance metrics accessing the data federation through CVMFS compared to direct data federation access. Additionally, we will discuss the improved user experience of providing access to a data federation through a POSIX filesystem.

Tertiary Keyword (Optional):
Storage systems

Secondary Keyword (Optional):
Computing middleware

Primary Keyword (Mandatory):
Distributed data handling

Track 5: Software Development / 502

Exploring Compression Techniques for ROOT IO

Author: Brian Paul Bockelman
Co-author: Zhang Zhe 2

1 University of Nebraska (US)
2 University of Nebraska-Lincoln

Corresponding Authors: brian.bockelman@cern.ch, zhan0915@huskers.unl.edu

ROOT provides an extremely flexible format used throughout the HEP community. The number of use cases – from an archival data format to end-stage analysis – has required a number of tradeoffs to be exposed to the user. For example, a high "compression level" in the traditional DEFLATE algorithm will result in a smaller file (saving disk space) at the cost of slower decompression (costing CPU time when read). If not done correctly, at the scale of a LHC experiment, poor design choices can result in terabytes of wasted space.

We explore and attempt to quantify some of these tradeoffs. Specifically, we explore: the use of alternate compression algorithms to optimize for read performance; an alternate method of compression individual events to allow efficient random access; and a new approach to whole-file compression. Quantitative results are given, as well as guidance on how to make compression decisions for different use cases.

Track 7: Middleware, Monitoring and Accounting / 503

Contributing opportunistic resources to the grid with HTCondor-CE-Bosco

Author: Derek John Weitzel 1

Co-author: Brian Paul Bockelman 1

1 University of Nebraska (US)

Corresponding Authors: derek.weitzel@cern.ch, brian.bockelman@cern.ch

The HTCondor-CE is the primary Compute Element (CE) software for the Open Science Grid. While it offers many advantages for large sites, for smaller, WLCG Tier-3 sites or opportunistic clusters, it can be a difficult task to install and configure the HTCondor-CE. Installing a CE typically involves understanding several pieces of software, installing hundreds of packages on a dedicated node, updating several configuration files, and implementing grid authentication mechanisms. On the other hand, accessing remote clusters from personal computers has been dramatically improved with Bosco: site admins only need to setup SSH public key authentication and appropriate accounts on a login host. In this paper, we take a new approach with the HTCondor-CE-Bosco, a CE which combines the flexibility and reliability of the HTCondor-CE with the easy-to-install Bosco. The administrators of the opportunistic resource are not required to install any software: only SSH access and a user account are required from the host site. The OSG can then run the grid-specific portions from a central location. This provides a new, more centralized, model for running grid services, which complements the traditional distributed model. We will show the architecture of a HTCondor-CE-Bosco enabled site, as well as feedback from multiple sites that have deployed it.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Throughout the last decade the Open Science Grid (OSG) has been fielding requests from user communities, resource owners, and funding agencies to provide information about utilization of OSG resources. Requested data include traditional “accounting” - core-hours utilized - as well as user’s certificate Distinguished Name, their affiliations, and field of science. The OSG accounting service, Gratia, developed in 2006, is able to provide this information and much more. However, with the rapid expansion and transformation of the OSG resources and access to them, we are faced with several challenges in adapting and maintaining the current accounting service. The newest changes include, but are not limited to, acceptance of users from numerous university campuses, whose jobs are flocking to OSG resources, expansion into new types of resources (public and private clouds, allocation-based HPC resources, and GPU farms), migration to pilot-based systems, and migration to multicore environments. In order to have a scalable, sustainable and expandable accounting service for the next few years, we are embarking on the development of the next-generation OSG accounting service, GRACC, that will be based on open-source technology and will be compatible with the existing system. It will consist of swappable, independent components, such as Logstash, Elasticsearch, Grafana, and RabbitMQ, that communicate through a data exchange. GRACC will continue to interface EGI and XSEDE accounting services and provide information in accordance with existing agreements. We will present the current architecture and working prototype.
Latest generation interconnect technologies in APEnet+ networking infrastructure

Authors: Alessandro Lonardo¹; Andrea Biagioni²; Davide Rossetti¹; Elena Pastorelli²; Francesca Lo Cicero²; Francesco Simula³; Laura Tosoratto³; Michele Martinelli³; Ottorino Frezza³; Pier Stanislas Paolucci³; Piero Vicini³; Roberto Ammendola⁴

¹ Università e INFN, Roma I (IT)  
² INFN  
³ INFN Rome Section  
⁴ Università e INFN Roma Tor Vergata (IT)

Corresponding Authors: andrea.biagioni@roma1.infn.it, alessandro.lonardo@cern.ch, elena.pastorelli@roma1.infn.it, michele.martinelli@roma1.infn.it, ottorino.frezza@roma1.infn.it, francesco.simula@roma1.infn.it, davide.rossetti@cern.ch, piero.vicini@roma1.infn.it, francesca.locicero@roma1.infn.it, pier.paolucci@roma1.infn.it, roberto.ammendola@roma2.infn.it, laura.tosoratto@roma1.infn.it

Hybrid systems are emerging as an efficient solution in the HPC arena, with an abundance of approaches for integration of accelerators into the system (i.e. GPU, FPGA). In this context, one of the most important features is the chance of being able to address the accelerators, whether they be local or off-node, on an equal footing. Correct balancing and high performance in how the network supports this kind of transfers in internode traffic become critical factors in global system efficiency.

The goal of the APEnet project is the design and development of a point-to-point, low-latency and high-throughput interconnect adapter, to be employed in High Performance Computing clusters with a 3D toroidal network mesh.

In this paper we present the status of the 3rd generation design of the board (V5) built around the 28nm Altera StratixV FPGA; it features a PCIe Gen3 x8 interface and enhanced embedded transceivers with a maximum capability of 50.0Gbps. The network architecture is designed according to the Remote DMA paradigm. APEnet implements the NVIDIA GPUDirect RDMA and V2 (“peer-to-peer”) protocols to directly access GPU memory, overcoming the bottleneck represented by transfers between GPU/host memory.

The APEnet+ V5 prototype is built upon the StratixV Dev Kit with the addition of a proprietary, third party IP core implementing multi-DMA engines. Support for zero-copy communication is assured by the possibility of DMA-accessing either host and GPU memory, offloading the CPU from the chore of data copying. Current implementation shows an upper limit for the memory read bandwidth of 4.8GB/s. Here we describe the memory write process hardware optimization relying on the use of two independent DMA engines and an improved TLB and characterization of software enhancements aimed at exploiting the hardware capabilities to the most, e.g. using CUDA 7.5 features and the driver migration to user-space, this latter allowing us to either better pinpoint software-induced overhead - compared to a kernel-space only driver implementation - and to lower the perceived latency to the application.

The APEnet+ V5 prototype offers three APElink high performance data transmission channels. The X channel was implemented by bonding 4 lanes of the QSFP connector available on the board; the Y and Z channels were implemented onto the HSMC interface. In this paper we describe the Transmission Control Logic that manages the data flow by encapsulating packets into a light, low-level, “word-stuffing” proprietary protocol able to detect transmission errors via CRC. The current implementation of the APElink TCL is able to sustain the link bandwidth of about 5GB/s at an operating frequency of ~312MHz. Measures of performance (latency and bandwidth) on host-to-host and GPU-to-GPU between two servers will be provided.

Finally, as regards future developments, we discuss work undertaken towards compliance with next generation FPGAs with hard IP processors on board.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Network systems and solutions

Primary Keyword (Mandatory):
High performance computing

Posters A / Break / 506

Evolution, design, management and support for the CMS Online computing cluster

Author: Marc Dobson

Co-authors: Andre Georg Holzner 2; Attila Racz 1; Benjamin Gordon Craigs 1; Christian Deldicque 1; Christoph Paus 1; Christoph Schwick 1; Cristian Contescu 1; Dainius Simelevicius 1; Dominique Gigi 1; Emilio Meschi 1; Frank Glege 1; Frans Meijers 1; Georgiana Lavinia Darlea 3; Guillelmo Gomez Ceballos Retuerto 3; Hannes Sakulin 1; James Gordon Branson 1; Jean-Marc Olivier Andre 1; Jeroen Hegeman 1; Jonathan Fulcher 1; Lorenzo Masetti 1; Luciano Orsini 1; Marco Pieri 1; Nicolas Doualot 1; Olivier Chaze 1; Petr Zejdl 1; Philipp Maximilian Brummer 1; Raul Jimenez Estupinan 1; Remigius K. Mommsen 1; Samim Erhan 1; Sergio Cittolin 1; Srecko Morovic 1; Thomas Reis 1; Ulf Behrens 3; Vivian O’Dell 7; Zeynep Demiragli 3

1 CERN
2 Univ. of California San Diego (US)
3 Massachusetts Inst. of Technology (US)
4 Fermi National Accelerator Lab. (US)
5 Univ. of California Los Angeles (US)
6 Deutsches Elektronen-Synchrotron (DE)
7 Fermi National Accelerator Laboratory (FNAL)

Corresponding Authors: hannes.sakulin@cern.ch, attila.racz@cern.ch, christian.deldicque@cern.ch, dominique.gigi@cern.ch, jonathan.fulcher@cern.ch, branson@ucsd.edu, guillelmo.gomez.ceballos@cern.ch, marco.pieri@cern.ch, srecko.morovic@cern.ch, frank.glege@cern.ch, luciano.orsini@cern.ch, samim.erhan@cern.ch, nicolas.doualot@cern.ch, remigius.mommsen@cern.ch, benjamin.gordon.craigs@cern.ch, raul.jimenez.estupinan@cern.ch, andre.georg.holzner@cern.ch, dainius.simelevicius@cern.ch, philipp.maximilian.brummer@cern.ch, jean-marc.olivier.andre@cern.ch, Lorenzo.masetti@cern.ch, emilio.meschi@cern.ch, olivier.chaze@cern.ch, paus@mit.edu, frans.meijers@cern.ch, cristian.contescu@cern.ch, sergio.cittolin@cern.ch, odell@fnal.gov, ulf.behrens@cern.ch, lavinia.darlea@cern.ch, thomas.reis@cern.ch, marc.dobson@cern.ch, jeroen.hegeman@cern.ch, petr.zejdl@cern.ch, zeynep.demiragli@cern.ch, christoph.schwick@cern.ch

After two years of maintenance and upgrade, the Large Hadron Collider (LHC) has started its second four year run. In the mean time, the CMS experiment at the LHC has also undergone two years of maintenance and upgrade, especially in the field of the Data Acquisition and online computing cluster, where the system was largely redesigned and replaced. Various aspects of the supporting computing system will be addressed here.

The increasing processing power and the use of high end networking technologies (10/40Gb/s Ethernet and 56Gb/s Infiniband) has reduced the number of DAQ event building nodes, since the performance of the individual nodes has increased by an order of magnitude since the start of LHC. The pressure on using the systems in an optimal way has increased accordingly, thereby also increasing the importance of proper configuration and careful monitoring to catch any deviation from standard behaviour. The upgraded monitoring system based on Ganglia and Icinga2 will be presented with the different mechanisms used to monitor and troubleshoot the crucial elements of the system.

The evolution of the various sub-detector applications, the data acquisition and high level trigger, following their upgraded hardware and designs over the upgrade and running periods, require a performant and flexible management and configuration infrastructure. The puppet based configuration and management system put in place for this phase, will be presented, showing it’s flexibility to support a large heterogeneous system, as well as, it’s ability to do bulk installations from scratch or rapid installations of CMS software cluster wide. A number of custom tools have been developed to
support the update of rpm based installations by the end users, a feature not typically supported in a datacenter environment. The performance of the system will also be presented with insights into its scaling with the increasing farm size over this data taking run.

Such a large and complex system requires redundant, flexible core infrastructure services to support them. Details will be given on how a flexible and highly available infrastructure has been put in place, leveraging various high availability technologies, from network redundancy, through virtualisation, to high availability services with Pacemaker/Corosync.

To conclude, a roundup of the different tools and solutions used in the CMS cluster administration will be given, pulling all the above into a coherent, performant and scalable system.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

DAQ

Primary Keyword (Mandatory):
Computing facilities

Track 1: Online Computing / 507

HEP Track Finding with the Micron Automata Processor and Comparison with an FPGA-based Solution

Author: Michael Wang

Co-authors: Christopher Green; Ted Zmuda; Tiehui Ted Liu; gustavo cancelo

Fermi National Accelerator Lab. (US)
Department of Physics
Fermilab
fermilab

Corresponding Authors: zmuda@fnal.gov, mwang@fnal.gov, cancelo@fnal.gov, greenc@fnal.gov, thliu@fnal.gov, jcfree@fnal.gov

Moore’s Law has defied our expectations and remained relevant in the semiconductor industry in the past 50 years, but many believe it is only a matter of time before an insurmountable technical barrier brings about its eventual demise. Many in the computing industry are now developing post-Moore’s Law processing solutions based on new and novel architectures. An example is the Micron Automata Processor (AP) which uses a non-von Neumann architecture based on the hardware realization of Non-deterministic Finite Automata. Although it is a dedicated pattern search engine designed primarily for text-based searches in the Internet search industry, the AP is also suitable for pattern recognition applications in HEP such as track finding. We describe our work in demonstrating a proof-of-concept for the suitability of the AP in HEP track finding applications. We compare our AP-based approach with a similar one based on Content-Addressable Memories on an FPGA. Pros and cons of each approach are considered and compared based on processing performance and ease of implementation.

Tertiary Keyword (Optional):
Processor architectures

Secondary Keyword (Optional):
Reconstruction
Primary Keyword (Mandatory):
Trigger

Track 7: Middleware, Monitoring and Accounting / 508

IceProd 2 usage experience

Authors: David Delventhal¹; David Schultz¹
Co-author: Juan Carlos Diaz Velez ¹

¹ University of Wisconsin-Madison

Corresponding Authors: david.schultz@icecube.wisc.edu, juancarlos@icecube.wisc.edu, david.delventhal@icecube.wisc.edu

IceProd is a data processing and management framework developed by the IceCube Neutrino Observatory for processing of Monte Carlo simulations, detector data, and analysis levels. It runs as a separate layer on top of grid and batch systems. This is accomplished by a set of daemons which process job workflow, maintaining configuration and status information on the job before, during, and after processing. IceProd can also manage complex workflow DAGs across distributed computing grids in order to optimize usage of resources.

IceProd has recently been rewritten to increase its scaling capabilities, handle user analysis workflows together with simulation production, and facilitate the integration with 3rd party scheduling tools. IceProd 2, the second generation of IceProd, has been running in production for several months now. We share our experience setting up the system and things we’ve learned along the way.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):
Computing middleware

Posters A / Break / 509

Applications of modern Erasure Coded Big Data paradigms to WLCG data resilience

Author: Samuel Cadellin Skipsey ¹
Co-author: Marcus Ebert ¹

¹ University of Edinburgh (GB)

Corresponding Authors: samuel.cadellin.skipsey@cern.ch, marcus.ebert@ed.ac.uk

Previous research has shown that it is relatively easy to apply a simple shim to conventional WLCG storage interfaces, in order to add Erasure coded distributed resilience to data.

One issue with simple EC models is that, while they can recover from losses without needing additional full copies of data, recovery often involves reading the all of the distributed chunks of the file (and their parity chunks). This causes efficiency losses, especially when the chunks are widely
distributed on a global level. Facebook, and others, have developed "Locally Repairable Codes" which avoid this issue, by adding additional parity chunks summing over subsets of the total chunk distribution, or by entangling the parity of two stripes to provide additional local information.

Applying these approaches to data distribution on WLCG storage resources, we provide a modified encoding tool, based on our previous approach, to generate LRC encoded files, and distribute them appropriately. We also discuss the potential application to the natural chunking of WLCG style data, with reference to single-event data access models and WAN data placement. In particular, we consider the advantages of mechanisms to distribute load across potentially contended "fat" Tier-2 storage nodes which may need to serve "thin" Tier-2 and Tier-3 resources in their geographical region.

Tertiary Keyword (Optional): Object stores
Secondary Keyword (Optional): Storage systems
Primary Keyword (Mandatory): Distributed data handling

Posters B / Break / 510

Pyglidein - a simple HTCondor glidein service

Author: David Schultz
Co-authors: Benedikt Riedel, Gonzalo Merino

1 University of Wisconsin-Madison
2 University of Alberta
3 IceCube

Corresponding Authors: benedikt.riedel@icecube.wisc.edu, david.schultz@icecube.wisc.edu, gonzalo.merino@icecube.wisc.edu

A major challenge for data production at the IceCube Neutrino Observatory presents itself in connecting a large set of small clusters together to form a larger computing grid. Most of these clusters do not provide a Grid interface. Using a local account on each submit machine, HTCondor glideins can be submitted to virtually any type of scheduler. The glideins then connect back to a main HTCondor pool, where jobs can run normally with no special syntax. To respond to dynamic load, a simple server advertises the number of idle jobs in the queue and the resources they request. The submit script can query this server to optimize glideins to what is needed, or not submit if there is no demand. Configuring HTCondor dynamic slots in the glideins allows us to efficiently handle varying memory requirements as well as whole-node jobs.

One step of the IceCube simulation chain, photon propagation in the ice, heavily relies on GPUs for faster execution. Therefore, one important requirement for any workload management system in IceCube is that it can handle GPU resources. Within the pyglidein system, we have successfully configured HTCondor glideins to use any GPU allocated to it, with jobs using the standard HTCondor GPU syntax to request and use a GPU. This mechanism allows us to seamlessly integrate our local GPU cluster with remote non-Grid GPU clusters, including specially allocated resources at XSEDE Supercomputers.

Tertiary Keyword (Optional):
Secondary Keyword (Optional): 

Page 348
Computing middleware

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

Posters B / Break / 511

Plancton: an opportunistic distributed computing project based on Docker containers

Author: Matteo Concas¹

Co-authors: Dario Berzano ²; Massimo Masera ¹; Maximiliano Puccio ³; Sara Vallero ¹; Stefano Bagnasco ³; Stefano Lusso ⁴

¹ Universita e INFN Torino (IT)
² CERN
³ I.N.F.N. TORINO
⁴ INFN-TO

Corresponding Authors: dario.berzano@cern.ch, massimo.masera@cern.ch, bagnasco@to.infn.it, stefano.lusso@to.infn.it, svallero@to.infn.it, maximiliano.puccio@cern.ch, matteo.concas@cern.ch

The computing power of most modern commodity computers is far from being fully exploited by standard usage patterns.

The work we present describes the development and setup of a virtual computing cluster based on Docker containers used as worker nodes. The facility is based on Plancton¹: a lightweight fire-and-forget background service that spawns and controls a local pool of Docker containers on a host with free resources by constantly monitoring its CPU utilisation. Plancton is designed to release the resources allocated opportunistically whenever another demanding task is run by the host user, according to configurable thresholds: this is attained by killing a number of running containers.

The resources comprising the facility are a collection of heterogeneous non-dedicated Linux hosts ideally inside the same local network, with no guaranteed network bandwidth, made available by members of a collaboration or institute. The user has agreed to donate its spare CPU cycles and remains the administrator of the involved host. Since the system is based on Docker containers performance isolation and security are guaranteed through sandboxing.

Using a thin virtualization layer such as Docker has the effect of having containers that are started almost instantly upon request. We will show how fast startup and disposal of containers finally enables us to implement the formation of the opportunistic cluster in a headless fashion, where our containers are mere pilots.

As an example we are running pilot HTCondor containers automatically joining a given cluster and terminating right after executing a job or in a short while if no new job is available. Software is provided through CVMFS on Parrot, making the execution environment suitable for HEP jobs.

Finally we will show how the uncomplicated approach of Plancton to containers deployment makes it suitable for setting up dedicated computing facilities too, provided that the underlying use case is sufficiently simple.

¹ https://github.com/mconcas/plancton

Tertiary Keyword (Optional):
Distributed workload management

Secondary Keyword (Optional):
Interconnected Private Clouds for Universities and Researchers

Author: Christian Schmitz
Co-authors: Jakub Moscicki; Peter Szegedi

1 ownCloud Inc
2 CERN

Corresponding Authors: szegedi@terena.org, cschmitz@owncloud.com, jakub.moscicki@cern.ch

1. Statement
OpenCloudMesh has a very simple goal: to be an open and vendor agnostic standard for private cloud interoperability.
To address the YetAnotherDataSilo problem, a working group under the umbrella of the GÉANT Association is has been created with the goal of ensuring neutrality and a clear context for this project.
All leading partners of the OpenCloudMesh project - GÉANT, CERN and ownCloud Inc. - are fully committed to the open API design principle. This means that - from day one - the OCM sharing API should be discussed, designed and developed as a vendor neutral protocol to be adopted by any on-premise sync&share product vendor or service provider. We acknowledge the fact that the piloting of the first working interface prototype was carried out in an ownCloud environment and has been in production since 2015 with a design size of 500,000 users, called "Sciebo", interconnecting dozens of private research clouds in Germany. Pydo adopting the standard in March 2016 underlines that this did and will not affect the adoption of the open API in any other vendor or service/domain provider.

2. OpenCloudMesh talk at CHEP2016 titled “Interconnected Private Clouds for Universities and Researchers”
The content of the presentation is an overview over the project currently managed by GEANT (Peter Szegedi), CERN (Dr. Jakub Moscicki) and ownCloud (Christian Schmitz), outlining overall concepts, past, present achievements and future milestones with a clear call to participation.
The presentation will summarize the problem originally scoped, then making a shift to the past success milestones (ex. demonstrated simultaneous interoperability between CERN, AARnet, Sciebo, UniVienna and, at the time of the talk expected, interoperability between different clouds running different software vendors) and then shift to future milestones, moonshot challenges and a call to participation.

3. OpenCloudMesh Moonshot scope
The problems, concepts and solution approaches to solving this are absolutely cutting edge as of 2016, hence they offer both practical and research challenges. Science and research in its open and peer reviewed nature has become a truly globalized project with OCM having the potential to be the "usability fabric" of a network of private clouds acting as one global research cloud.

4. Links
Project Wiki
https://wiki.geant.org/display/OCM/Open+Cloud+Mesh
SciDAC-Data, A Project to Enabling Data Driven Modeling of Exascale Computing

Authors: Andrew Norman¹; Leonidas Aliaga Soplin²

Co-authors: Adam Lyon ¹; Aristeidis Tsaris ¹; Leonidas Aliaga Soplin ¹; Misbah Mubarek ³; Pengfei Ding ⁴; Robert Ross ³

¹ Fermilab
² College of William and Mary (US)
³ Argonne
⁴ The University of Manchester

Corresponding Authors: mmubarak@anl.gov, dingpf@fnal.gov, atsaris@fnal.gov, lyon@fnal.gov, rross@mcs.anl.gov, anorman@fnal.gov, laliagasoplin@email.wm.edu, leonidas.aliaga.soplin@cern.ch

The SciDAC-Data project is a DOE funded initiative to analyze and exploit two decades of information and analytics that have been collected, by the Fermilab Data Center, on the organization, movement, and consumption of High Energy Physics data. The project is designed to analyze the analysis patterns and data organization that have been used by the CDF, DØ, NOØA, Minos, Minerva and other experiments, to develop realistic models of HEP analysis workflows and data processing. The SciDAC-Data projects aims to provide both realistic input vectors and corresponding output data which can be used to optimize and validate simulations of HEP analysis in different high performance computing (HPC) environments. These simulations are designed to address questions of data handling, cache optimization and workflow structures that are the prerequisites for modern HEP analysis chains to be mapped and optimized to run on the next generation of leadership class exascale computing facilities.

We will address the use of the SciDAC-Data distributions acquired from over 5.6 million analysis workflows and corresponding to over 410,000 HEP datasets, as the input to detailed queuing simulations that model the expected data consumption and caching behaviors of the work running in HPC environments. In particular we describe in detail how the SAM data handling system in combination with the dCache/Enstore based data archive facilities have been analyzed to develop the radically different models of the analysis of collider data and that of neutrino datasets. We present how the data is being used for model output validation and tuning of these simulations. The paper will address the next stages of the SciDAC-Data project which will extend this work to more detailed modeling and optimization of the models for use in real HPC environments.
Purpose
The aim of this work consists in the full simulation and measurements of a GEMPix (Gas Electron Multiplier) detector for a possible application as monitor for beam verification at CNAO Center (National Center for Oncological Hadrontherapy). A triple GEMPix detector read by 4 Timepix chips could provide a beam monitoring, dose verification and quality checks with good resolution and optimal radiation background control with respect to the existing devices. The Monte Carlo Geant4 10.01 patch 03 toolkit was used to simulate the complete CNAO extraction beamline (beam delivered with active scanning modality). The simulation allowed the characterization of the GEM detector response carbon ions with respect to reference detectors (EBT3 radiochromic films).

Methods
The GEMPix detector is fully simulated: an homogeneous electric field was implemented to take into account the drift of secondary particle in gas gap. An ArCO2CF4 gas mixture was simulated to reproduce the GEMPix response. The complete measurement setup was simulated with the GEMPix placed in a water phantom and irradiated with carbon ions at different energies. Important beam parameters such as the transverse FWHM were compared with experimental measurements at CNAO Center. A triple GEM detector prototype, with a 55 μm pitch pixelated ASIC for the readout, was tested at CNAO in Pavia for a detailed characterization and measurements of energy deposition inside the water phantom. The energy deposition was measured at different positions in depth allowing a 3D reconstruction of the beam inside the phantom.

Results
The simulation results are very encouraging since they reproduce in a few percent the experimental data set. All the measurements are carried out in a stable setup at CNAO Center and are acquired in...
several experimental sessions with different parameters settings. Experimental measurements are still ongoing.

**Conclusions**

Even further validations must be done, the good results so far obtained by this work point out and confirm that GEMPix detector is suitable to be used as beam monitor in hadrontherapy.

**Secondary Keyword (Optional):**

Analysis tools and techniques

**Primary Keyword (Mandatory):**

Simulation

**Tertiary Keyword (Optional):**

Track 7: Middleware, Monitoring and Accounting / 515

**Analyzing how we do Analysis and consume data, Results from the SciDAC-Data Project.**

**Author:** Andrew Norman

**Co-authors:** Adam Lyon; Aristeidis Tsaris; Leonidas Aliaga Soplin; Misbah Mubarak; Pengfei Ding; Robert Ross

1 Fermilab
2 Argonne

**Corresponding Authors:** mmubarak@anl.gov, rross@mcs.anl.gov, dingpf@fnal.gov, anorman@fnal.gov, laliaga-soplin@email.wm.edu, lyon@fnal.gov, atsaris@fnal.gov

One of the principle goals of the Dept. of Energy funded SciDAC-Data project is to analyze the more than 410,000 high energy physics “datasets” that have been collected, generated and defined over the past two decades by experiments using the Fermilab storage facilities. These datasets have been used as the input to over 5.6 million recorded analysis projects, for which detailed analytics have been gathered. The analytics and meta information regarding these for these datasets and analysis projects are being combined with knowledge of their part of the HEP analysis chains for major experiments to understand how modern computing and data delivery is being used.

We present the first results of this project, which examine in detail how the CDF, DØ and NOØA experiments have organized, classified and consumed petascale datasets to produce their physics results. The results include the analysis of the correlations in dataset/file overlap, data usage patterns, data popularity, dataset dependency and temporary dataset consumption. The results provide critical insight into how workflows and data delivery schemes can be combined with different caching strategies to more efficiently perform the work required to mine these large HEP data volumes and to understand the physics analysis requirements for the next generation of HEP computing facilities.

In particular we present detailed analysis of the NOØA data organization and consumption model corresponding to their first and second oscillation results (2014-2016) and the first look at the analysis of the Tevatron Run II experiments. We present statistical distributions for the characterization of these data and data driven models describing their consumption.

**Tertiary Keyword (Optional):**

Data processing workflows and frameworks/pipelines

**Secondary Keyword (Optional):**

Computing facilities
Adjusting the fairshare policy to prevent computing power loss

Author: Stefano Dal Pra

1 INFN

Corresponding Author: stefano.dalpra@cnaf.infn.it

On a typical WLCG site providing batch access to computing resources according to a fairshare policy, the idle timelapse after a job ends and before a new one begins on a given slot is negligible if compared to the duration of typical jobs. The overall amount of these intervals over a time window increases with the size of the cluster and the inverse of job duration and can be considered equivalent to an average number of unavailable slots over that time window. This value has been investigated for the Tier-1 at CNAF, and observed to occasionally grow and reach up to more than the 10% of the about 20,000 available computing slots. Analysis reveals that this happens when a sustained rate of short jobs is submitted to the cluster and dispatched by the batch system. Because of how the default fairshare policy works, it increases the dynamic priority of those users mostly submitting short jobs, since they are not accumulating runtime, and will dispatch more of their jobs at the next round, thus worsening the situation until the submission flow ends. To address this problem the default behaviour of the fairshare have been altered by adding a correcting term to the default formula for the dynamic priority. The LSF batch system, currently adopted at CNAF, provides a way to define its value by invoking a C function, which returns it for each user in the cluster. The correcting term works by rounding up to a minimum defined runtime the most recently done jobs. Doing so, each short job looks almost like a regular one and the dynamic priority value equilibrates to a proper value. The net effect is a reduction of the dispatching rate of short jobs and, consequently, the average number of available slots greatly improves. Furthermore, a potential starvation problem, actually observed at least once is also prevented. After describing short jobs and reporting about their impact on the cluster, possible workarounds are discussed and the selected solution is motivated. Details on the most critical aspects of the implementation are explained and the observed results are presented.
The Deep Underground Neutrino Experiment (DUNE) will employ a uniquely large (40kt) Liquid Argon Time Projection chamber as the main component of its Far Detector. In order to validate this design and characterize the detector performance an ambitious experimental program (called "protoDUNE") has been created which includes a beam test of a large-scale DUNE prototype at CERN. The amount of data to be collected in this test is substantial and on par with the LHC experiments in LHC Run 1. The protoDUNE experiment will require careful design of the DAQ and data handling systems, as well as mechanisms to distribute data to a number of the DUNE distributed computing sites. We present our approach to solving these problems by leveraging the expertise and components created at Fermilab, in a broader context of integration with the systems at other National Laboratories in the US as well as at CERN and other European sites.

Alignement of the CMS Tracker: Latest Results from LHC Run-II

The all-silicon design of the tracking system of the CMS experiment provides excellent resolution for charged tracks and an efficient tagging of jets. As the CMS tracker, and in particular its pixel detector, underwent repairs and experienced changed conditions with the start of the LHC Run-II in 2015, the position and orientation of each of the 15148 silicon strip and 1440 silicon pixel modules needed to be determined with a precision of several micrometers. The alignment also needs to be quickly recalculated each time the state of the CMS magnet is changed between 0T and 3.8T. Latest Run-II results of the CMS tracker alignment and resolution performance are presented that were obtained using several million reconstructed tracks from collisions and cosmic rays data of 2015 and 2016. The geometries and the resulting performance of physics observables are finally carefully validated with data-driven methods.
The Application of SNiPER to the JUNO Simulation

Authors: Jiaheng Zou¹; Tao Lin¹
Co-authors: Guofu Cao¹; Weidong Li¹; Xiao Fang²; Xingtao Huang³; Zhengyun You⁴; Ziyan Deng¹

¹ IHEP
² Sichuan University
³ Shandong University
⁴ Sun Yat-Sen University

Corresponding Authors: huangxt@sdu.edu.cn, youzhy5@mail.sysu.edu.cn, weidong.li@ihep.ac.cn, dengzy@ihep.ac.cn, fangx@ihep.ac.cn, caogf@ihep.ac.cn, zoujh@ihep.ac.cn, lintao@ihep.ac.cn, lintao51@gmail.com

The JUNO (Jiangmen Underground Neutrino Observatory) is a multipurpose neutrino experiment which is mainly designed to determine neutrino mass hierarchy and precisely measure oscillation parameters. As one of the most important systems, the JUNO offline software is being developed using the SNiPER software. In this presentation, we focus on the requirements of JUNO simulation and present the working solution based on the SNiPER.

The JUNO simulation framework is in charge of managing event data, detector geometries and materials, physics processes, simulation truth information etc. It glues physics generator, detector simulation and electronics simulation modules together to achieve a full simulation chain. In the implementation of the framework, many attractive characteristics of the SNiPER have been used, such as dynamic loading, flexible flow control, multiple event management and Python binding. Furthermore, additional efforts have been made to make both detector and electronics simulation flexible enough to accommodate and optimize different detector designs.

For the Geant4-based detector simulation, each subdetector component is implemented as a SNiPER tool which is a dynamically loadable and configurable plugin. So it is possible to select the detector configuration in runtime. The framework provides the event loop to drive the detector simulation and interacts with the Geant4 which is implemented as a passive service. All levels of user actions are wrapped into different customizable tools, so that user functions can be easily extended by just adding new tools. The electronics simulation has been implemented by following an event driven scheme. The SNiPER task component is used to simulate data processing steps in the electronics modules. The electronics and trigger are synchronized by triggered events containing possible physics signals.

Now the JUNO simulation software has been released and is being used by the JUNO collaboration to do detector design optimization, event reconstruction algorithm development and physics sensitivity studies. The concurrent computing using GPU and phi-coprocessor is being studied in order to speed up the simulation of light propagation in the large liquid scintillator detector.

Secondary Keyword (Optional):
Simulation
YARR - A PCIe based readout concept for current and future ATLAS Pixel Modules

Author: Timon Heim

Co-authors: Ian Hinchliffe; Mauricio Garcia-Sciveres

1 Lawrence Berkeley National Lab. (US)

Corresponding Authors: timon.heim@cern.ch, i_hinchliffe@lbl.gov, mgarcia-sciveres@lbl.gov

The Yet Another Rapid Readout (YARR) system is a DAQ system designed for the readout of current generation ATLAS Pixel FE-I4 and next generation ATLAS ITk chips. It utilises a commercial-of-the-shelf PCIe FPGA card as a reconfigurable I/O interface, which acts as a simple gateway to pipe all data from the pixel chips via the high speed PCIe connection into the host systems memory. Relying on modern CPU architectures which enables the usage of parallelised processing threads and commercial high speed interfaces in everyday computers, it is possible to perform all processing on a software level in the host CPU. Although FPGAs are very powerful at parallel signal processing their firmware is hard to maintain and constrained by their connected hardware, software on the other hand is very portable and upgraded frequently with new features coming at no cost. A DAQ concept which does not rely on the underlying hardware for acceleration also eases the transition from prototyping in the laboratory to the full scale implementation in the experiment. The overall concept and data flow will be outlined, as well as the challenges and possible bottlenecks which can be encountered when moving the processing from hardware to software.
We present the Web-Based Monitoring project of the CMS experiment at the LHC at CERN. With the growth in size and complexity of High Energy Physics experiments and the accompanying increase in the number of collaborators spread across the globe, the importance of broadly accessible monitoring has grown. The same can be said about the increasing relevance of operation and reporting web tools used by run coordinators, sub-system experts and managers. CMS Web-Based Monitoring has played a crucial role providing that for the CMS experiment through the commissioning phase and the LHC Run I data taking period in 2010-2012. It has adapted to many CMS changes and new requirements during the Long Shutdown 1 and even now, during the ongoing LHC Run II. We have developed a suite of web tools to present data to the users from many underlying heterogeneous sources, from real time messaging systems to relational databases. The tools combine, correlate and visualize information in both graphical and tabular formats of interest to the experimentalist, with data such as beam conditions, luminosity, trigger rates, DAQ, detector conditions, operational efficiency and more, allowing for flexibility on the user side. In addition, we provide data aggregation, not only at display level but also at database level. An upgrade of the Web Based Monitoring project is being planned, implying major changes, and that is also discussed here.

Tertiary Keyword (Optional):
Trigger

Secondary Keyword (Optional):
DAQ

Primary Keyword (Mandatory):
Monitoring

---

Developing the Traceability Model to meet the Requirements of an Evolving Distributed Computing Infrastructure

Author: Ian Collier

Co-authors: Andrew David Lahiff; Dave Kelsey; Romain Wartel; Vincent Brillault

1 STFC - Rutherford Appleton Lab. (GB)
2 CERN

Corresponding Authors: vincent.brillault@cern.ch, roomain.wartel@cern.ch, ian.peter.collier@cern.ch, andrew.david.lahiff@cern.ch, david.kelsey@stfc.ac.uk

The growing use of private and public clouds, and volunteer computing are driving significant changes in the way large parts of the distributed computing for our communities are carried out. Traditionally HEP workloads within WLCG were almost exclusively run via grid computing at sites where site administrators are responsible for and have full sight of the execution environment. The experiment virtual organisations (VOs) are increasingly taking more control of those execution environments. In addition, the development of container and control group technologies offer new possibilities for isolating processes and workloads.

The absolute requirement for detailed information allowing incident response teams to answer the basic questions of who, did what, when and where, remains. But we can no longer rely on central logging from within the execution environment at resource providers sites to provide all of that information. Certainly, in the case of commercial public cloud providers that information is unlikely to be accessible at all. Shifting the focus to the externally observable behaviour of processes (including virtual machines and containers) and looking to the VO workflow management systems for user identification would be one approach to ensuring the required traceability.

The newly created WLCG Traceability & Isolation Working Group is investigating the feasibility of the technologies involved and developing technical requirements both for gathering traceability...
information from VO workflow management and technologies for separating processes and isolating processes and so protecting users and their data from one another.

We discuss the technical requirements as well as the policy issues raised and make some initial proposals for solutions to these problems.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Distributed workload management

Primary Keyword (Mandatory):
Security and policies

---

Offline Software for the CMS Level-1 Trigger

Author: Michael J Mulhearn¹

¹ University of California Davis (US)

Corresponding Author: mulhearn@physics.ucdavis.edu

The offline software for the CMS Level-1 trigger provides a reliable bitwise emulation of the high-speed custom FPGA-based hardware at the foundation of the CMS data acquisition system. The staged upgrade of the trigger system requires flexible software that accurately reproduces the system at each stage using recorded running conditions. The high intensity of the upgraded LHC necessitates new advanced algorithms which reconstruct physics objects in real time. We'll discuss the status and performance of the upgraded trigger software.

Tertiary Keyword (Optional):
Simulation

Secondary Keyword (Optional):
DAQ

Primary Keyword (Mandatory):
Trigger

---

Evaluation of the pre-production implementation of the Intel Omni-Path interconnect technology

Author: Alexandr Zaytsev¹

Co-authors: Christopher Hollowell ²; Costin Caramaru ²; Tony Wong ²; William Strecker-Kellogg ³

¹ Brookhaven National Laboratory (US)
² Brookhaven National Laboratory
³
This contribution gives a report on the remote evaluation of the pre-production Intel Omni-Path (OPA) interconnect hardware and software performed by RHIC & ATLAS Computing Facility (RACF) at BNL in Dec 2015 - Feb 2016 time period using a 32 node “Diamond” cluster with a single Omni-Path Host Fabric Interface (HFI) installed on each and a single 48-port Omni-Path switch with the non-blocking fabric (capable of carrying up to 9.4 Tbps of the aggregate traffic if all ports are involved) provided by Intel. The main purpose of the tests was to assess the basic features and functionality of the control and diagnostic tools available for the pre-production version of the pre-production version of the Intel Omni-Path low latency interconnect technology, as well as the Omni-Path interconnect performance in a realistic environment of a multi-node HPC cluster running RedHat Enterprise Linux 7 x86_64 OS. The interconnect performance metering was performed using the low level fabric layer and MPI communication layer benchmarking tools available in the OpenFabrics Enterprise Distribution (OFED), Intel Fabric Suite and OpenMPI v1.10.0 distributions with pre-production support of the Intel OPA interconnect technology built with both GCC v4.9.2 and Intel Compiler v15.0.2 versions and provided in the existing test cluster setup. A subset of the tests were performed with benchmarking tools built with GCC and Intel Compiler, with and without explicit mapping of a test processes to the physical CPU cores on the compute nodes in order to determine whether these changes result in a statistically significant difference in performance observed. Despite the limited scale of the test cluster used, the test environment provided was sufficient to carry out a large variety of RDMA, native and Intel OpenMPI, and IP over Omni-Path performance measurements and functionality tests. In addition to presenting the results of the performance benchmarks we also discuss the prospects for the future used of the Intel Omni-Path technology as a future interconnect solutions for both the HPC and HTC scientific workloads.

**Tertiary Keyword (Optional):**

**Secondary Keyword (Optional):**

High performance computing

**Primary Keyword (Mandatory):**

Network systems and solutions

---

**Track 5: Software Development / 525**

**Challenges in Scaling NLO Generators to Leadership Computers**

**Author:** Taylor Childers

**Co-authors:** Doug Benjamin; Thomas Le Compte; Thomas Uram

1 Argonne National Laboratory (US)

2 Duke University (US)

3 Argonne Leadership Computing Facility

**Corresponding Authors:** tomuram@gmail.com, lecompte@anl.gov, douglas.benjamin@cern.ch, john.taylor.childers@cern.ch

Exascale computing resources are roughly a decade away and will be capable of 100 times more computing than current supercomputers. In the last year, Energy Frontier experiments crossed a milestone of 100 million core-hours used at the Argonne Leadership Computing Facility, Oak Ridge Leadership Computing Facility, and NERSC. The Fortran-based leading-order parton generator called Alpgen was successfully scaled to millions of threads to achieve this level of usage on Mira. Sherpa and MadGraph are next-to-leading order generators used heavily by LHC experiments for simulation. Integration times for high-multiplicity or rare NLO processes can take a week or more on standard Grid machines, even using all 16-cores. We will describe our work to scale these generators to...
millions of threads on leadership-class machines to reduce run times to less than a day. This work allows the experiments to leverage large-scale parallel supercomputers for event generation today, freeing tens of millions of grid hours for other work, and paving the way for future applications (simulation, reconstruction) on these and future supercomputers.

**Tertiary Keyword (Optional):**
Event generators

**Secondary Keyword (Optional):**
Parallelization

**Primary Keyword (Mandatory):**
High performance computing

---

**Track 3: Distributed Computing / 526**

**The HEP Cloud Facility: elastic computing for High Energy Physics**

**Authors:** Anthony Tiradani¹; Burt Holzman²; Gabriele Garzoglio³ and Robert Kennedy⁴; Steven Timm⁵; Stuart Fuess⁶

¹ Fermilab  
² Fermi National Accelerator Lab. (US)

**Corresponding Authors:** kennedy@fnal.gov, burt@fnal.gov, tiradani@fnal.gov, garzoglio@fnal.gov, timmsteve@yahoo.com, fuess@fnal.gov

The need for computing in the HEP community follows cycles of peaks and valleys mainly driven by conference dates, accelerator shutdown, holiday schedules, and other factors. Because of this, the classical method of provisioning these resources at providing facilities has drawbacks such as potential overprovisioning. As the appetite for computing increases, however, so does the need to maximize cost efficiency by developing a model for dynamically provisioning resources only when needed.

To address this issue, the HEP Cloud project was launched by the Fermilab Scientific Computing Division in June 2015. Its goal is to develop a facility that provides a common interface to a variety of resources, including local clusters, grids, high performance computers, and community and commercial Clouds. Initially targeted experiments include CMS and NOvA, as well as other Fermilab stakeholders.

In its first phase, the project has demonstrated the use of the “elastic” provisioning model offered by commercial clouds, such as Amazon Web Services. In this model, resources are rented and provisioned automatically over the Internet upon request. In January 2016, the project demonstrated the ability to increase the total amount of global CMS resources by 58,000 cores from 150,000 cores - a 25 percent increase - in preparation for the Rencontres de Moriond. In March 2016, the NOvA experiment has also demonstrated resource burst capabilities with an additional 7,300 cores, achieving a scale almost four times as large as the local allocated resources and utilizing the local AWS s3 storage to optimize data handling operations and costs. NOvA was using the same familiar services used for local computations, such as data handling and job submission, in preparation for the Neutrino 2016 conference. In both cases, the cost was contained by the use of the Amazon Spot Instance Market and the Decision Engine, a HEP Cloud component that aims at minimizing cost and job interruption.

This paper describes the Fermilab HEP Cloud Facility and the challenges overcome for the CMS and NOvA communities.
Approximating Likelihood Ratios with Calibrated Classifiers

Authors: Gilles Louppe¹; Juan Guillermo Pavez Sepulveda²; Kyle Stuart Cranmer¹

¹ New York University (US)  
² Federico Santa Maria Technical University (CL)

Corresponding Authors: juan.pavezs@alumnos.usm.cl, kyle.cranmer@cern.ch, g.louppe@cern.ch

Likelihood ratio tests are a well established technique for statistical inference in HEP. Because of the complicated detector response, we usually cannot evaluate the likelihood function directly. Instead, we usually build templates based on (Monte Carlo) samples from a simulator (or generative model). However, this approach doesn’t scale well to high dimensional observations.

We describe a technique that is an generalization of HEP usage of machine learning for discrete signal vs. background classification to a situation with continuous parameters. We use the simulator to describe the complex processes that tie parameters θ of an underlying theory and measurement apparatus to the observations. we show that likelihood ratios are invariant under a specific class of dimensionality reduction maps \( \mathbb{R}^p \rightarrow \mathbb{R} \). As a direct consequence, we show that discriminative classifiers can be used to approximate the generalized likelihood ratio statistic when only a generative model for the data is available.

We have implemented the method in 'carl', a python based package that extends the interfaces of scikit-learn and also takes advantage of theano for the implementation of simple generative models.

The method has many applications as it extends our traditional likelihood-based measurements to high-dimensional inputs. The likelihood ratio can also be used for multivariate reweighting and as a goodness of fit statistic. Experimental results on artificial problems with known exact likelihoods illustrate the potential of the proposed method. We also have preliminary results applying the method to Higgs Effective Field Theory.
dCache, managed Cloud Storage

Authors: Albert Rossi¹; Anupam Ashish²; Christian Bernardt³; Dmitry Litvintsev¹; Femi Adeyemi²; Gerd Behrmann⁵; Juergen Starek²; Marina Sahakyan²; Patrick Fuhrmann²; Paul Millar⁵; Tigran Mkrtchyan²

¹ Fermilab  
² DESY  
³ Deutsches Elektronen-Synchrotron (DE)  
⁴ FNAL  
⁵ NDGF

Corresponding Authors: litvinse@fnal.gov, anupam.ashish@desy.de, juergen.starek@desy.de, behrmann@ndgf.org, paul.millar@desy.de, marina.sahakyan@desy.de, tigran.mkrtchyan@desy.de, christian.bernardt@desy.de, olufemi.adeyemi@desy.de, patrick.fuhrmann@desy.de, arossi@fnal.gov, patrick.fuhrmann@cern.ch

For the previous decade, high performance, high capacity Open Source storage systems have been designed and implemented, accommodating the demanding needs of the LHC experiments. However, with the general move away from the concept of local computer centers, supporting their associated communities, towards large infrastructures, providing Cloud-like solutions to a large variety of different scientific groups, storage systems needed to adjust their capabilities in many areas, as there are federated identities, non authenticated delegation to portals or platforms, modern sharing and user defined Quality of Storage.

This presentation will give an overview on how dCache is keeping up with modern Cloud storage requirements by partnering with EU projects, which provide the necessary contact to a large set of Scientific Communities.

Regarding authentication, there is no strict relationship anymore between the individual scientist, the scientific community and the infrastructure, providing resources. Federated identity systems like SAML or “OpenID Connect” are growing into the method-of-choice for new scientific groups and are even sneaking their way into HEP.

Therefor, under the umbrella of the INDIGO-DataCloud project, dCache is implementing those authentication mechanisms in addition to the already established ones, like username/password, Kerberos and X509 Certificates.

To simplify the use of dCache as back-end of scientific portals, dCache is experimenting with new anonymous delegation methods, like “Macaroons”, which the dCache team would like to introduce in order to start a discussion, targeting their broader acceptance in portals and at the level of service providers.

As the separation between managing scientific mass data and scientific semi-private data, like publications, is no longer strict, large data management systems are supposed to provide a simple interface to easily share data among individuals or groups. While some systems are offering that feature through web portals only, dCache will show that this can be provided uniquely for all protocols the system supports, including NFS and GridFTP.

Furthermore, in modern storage infrastructures, storage media, and consequently the quality and price of the request storage space are no longer negotiated with the responsible system administrators but dynamically selected by the end user or by automated computing platforms. The same is true for data migration between different qualities of storage. To accommodate this conceptual change, dCache is exposing it’s entire data management interface through a RESTful service and a graphical user interface. The implemented mechanisms are following the recommendation of the corresponding working groups in RDA and SNIA and are agreed-upon with the INDIGO-DataCloud project to be compatible with similar functionalities of other INDIGO provided storage systems.

Tertiary Keyword (Optional): Distributed data handling

Secondary Keyword (Optional): Storage systems

Primary Keyword (Mandatory):
Evolution of the LHCb Online run bookkeeping and data consistency system

Author: Rainer Schwemmer

Co-author: Niko Neufeld

1 CERN

Corresponding Authors: niko.neufeld@cern.ch, rainer.schwemmer@cern.ch

Since its original commissioning in 2008, the LHCb data acquisition system has seen several fundamental architectural changes. The original design had a single, continuous stream of data in mind, going from the read-out boards through a software trigger straight to a small set of parallelly written files. Over the years the enormous increase in available storage capacity has made it possible to reduce the amount of real-time computing at the experiment site and move to a more Offline like processing of data. A reduced, software based trigger is pre-processing the detector data which is then stored in a pool of storage elements in our computing farm. This data is then further processed at a later time, when computing resources are available.

Today, the LHCb Online System is keeping track of several thousand files and hundreds of runs being processed concurrently. The additional storage and parallel processing made it necessary to scale the run and file bookkeeping system far beyond its original design specifications. The additional complexity of the data flow also called for improved sanity checks and post processing, before the data can be shipped to the grid for analysis.

In this paper we are going to show the evolution of our scaled up system, with particular focus on handling several runs in parallel, output file merging for easier offline processing, data integrity checking and assurance that events are only sent offline once.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Data processing workflows and frameworks/pipelines

Primary Keyword (Mandatory):
DAQ

jade: An End-To-End Data Transfer and Catalog Tool

Author: PATRICK MEADE

1 University of Wisconsin-Madison

Corresponding Author: patrick.meade@icecube.wisc.edu

The IceCube Neutrino Observatory is a cubic kilometer neutrino telescope located at the Geographic South Pole. IceCube collects 1 TB of data every day. An online filtering farm processes this data

Page 364
in real time and selects 10% to be sent via satellite to the main data center at the University of Wisconsin-Madison. IceCube has two year-round on-site operators. New operators are hired every year, due to the hard conditions of wintering at the South Pole. These operators are tasked with the daily operations of running a complex detector in serious isolation conditions. One of the systems they operate is the data archiving and transfer system. Due to these challenging operational conditions, the data archive and transfer system must above all be simple and robust. It must also share the limited resource of satellite bandwidth, and collect and preserve useful metadata. The original data archive and transfer software for IceCube was written in 2005. After running in production for several years, the decision was taken to fully rewrite it, in order to address a number of structural drawbacks. The new data archive and transfer software (JADE2) has been in production for several months providing improved performance and resiliency. One of the main goals for JADE2 is to provide a unified system that handles the IceCube data end-to-end: from collection at the South Pole, all the way to long-term archive and preservation in dedicated repositories at the North. In this contribution, we describe our experiences and lessons learned from developing and operating the data archive and transfer software for a particle physics experiment in extreme operational conditions like IceCube.

**Tertiary Keyword (Optional):**

**Secondary Keyword (Optional):**

Software development process and tools

**Primary Keyword (Mandatory):**

Data processing workflows and frameworks/pipelines

---

**Track 6: Infrastructures / 531**

**Using container orchestration to improve service management at the RAL Tier 1**

**Author:** Andrew David Lahiff

**Co-author:** Ian Collier

1 STFC - Rutherford Appleton Lab. (GB)

**Corresponding Authors:** andrew.david.lahiff@cern.ch, ian.peter.collier@cern.ch

At the RAL Tier-1 we have been deploying production services on both bare metal and a variety of virtualisation platforms for many years. Despite the significant simplification of configuration and deployment of services due to the use of a configuration management system, maintaining services still requires a lot of effort. Also, the current approach of running services on static machines results in a lack of fault tolerance, which lowers availability and increases the amount of manual interventions required. In the current climate more and more non-LHC communities are becoming important, with the potential for the need to run additional instances of existing services as well as new services, but at the same time comes the likelihood that staff effort is more likely to decrease rather than increase. It is therefore important that we are able to reduce the amount of effort required to maintain services whilst ideally improving availability, in addition to being able to maximise the utilisation of resources and become more adaptive to changing conditions.

These problems are not unique to RAL, and from looking at what is happening in the wider world it is clear that container orchestration has the possibility to provide a solution to many of these issues. Therefore last year we began investigating the migration of services to an Apache Mesos cluster running on bare metal. In this model the concept of individual machines is abstracted away and services are run on the cluster in Docker containers, managed by a scheduler. This means that any host or application failures, as well as procedures such as rolling starts or upgrades, can be handled automatically and no longer require any human intervention. Similarly, the number of instances of
applications can be scaled automatically in response to changes in load. On top of this it also gives us the important benefit of being able to run a wide range of services on a single set of resources without involving virtualisation.

In this presentation we will describe the Mesos infrastructure that has been deployed at RAL, including how we deal with service discovery, the challenge of monitoring, logging and alerting in a dynamic environment and how it integrates with our existing traditional infrastructure. We will report on our experiences in migrating both stateless and stateful applications, the security issues surrounding running services in containers, and finally discuss some aspects of our internal process for making Mesos a platform for running production services.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):
Cloud technologies

Use of a hardware token for Grid authentication by the MICE data distribution framework

Author: J.J. Nebrensky

Co-author: Janusz Martyniak

1 Brunel University
2 Imperial College

Corresponding Authors: janusz.martyniak@imperial.ac.uk, j.nebrensky@brunel.ac.uk

The international Muon Ionization Cooling Experiment (MICE) is designed to demonstrate the principle of muon ionisation cooling for the first time, for application to a future Neutrino Factory or Muon Collider. The experiment is currently under construction at the ISIS synchrotron at the Rutherford Appleton Laboratory, UK. As presently envisaged, the programme is divided into three Steps: characterisation of the muon beams (complete), characterisation of the Cooling Channel and Absorbers (data-taking restarting in 2016), and demonstration of Ionisation Cooling (2018).

Data distribution and archiving, batch reprocessing, and simulation are all carried out using the EGI Grid infrastructure, in particular the facilities provided by GridPP in the UK. To prevent interference - especially accidental data deletion - these activities are separated by different VOMS roles.

Data acquisition, in particular, can involve 24/7 operation for a number of weeks and so for moving the data out of the MICE Local Control Room at the experiment a valid, VOMS-enabled, Grid proxy must be made available continuously over that time. Long-lifetime proxies and password-less certificates raise security concerns regarding their exposure, whereas requiring a particular certificate owner to log in and renew the proxy manually at half-day intervals for weeks on end is operationally unsustainable. The MyProxy service still requires maintaining a valid local proxy, to talk to the MyProxy server, and also requires that a long-lifetime proxy be held at a remote site.

The MICE “Data Mover” agent, responsible for transferring the raw data from the experiment DAQ to tape and initial replication on the Grid whence it can be read with other credentials, is now using a robot certificate stored on a hardware token (Feitian ePass2003) from which a cron job generates a "plain" proxy (using the scripts distributed by NIKHEF) to which the VOMS extensions are added in a separate transaction. A valid short-lifetime proxy is thus continuously available to the Data Mover process.
The Feitian ePass2003 was chosen because it was both significantly cheaper and easier to actually purchase than the token commonly referred to in the community at that time; however there was no software support for the hardware. This paper will detail the software packages, process and commands used to deploy the token into production. A similar arrangement (with a different certificate) is to be put in place for MICE’ Offline Reconstruction data distribution.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Distributed data handling

Primary Keyword (Mandatory):
Security and policies

Track 4: Data Handling / 533

Evaluation of ZFS as an efficient WLCG storage backend

Author: Marcus Ebert
Co-author: Andrew John Washbrook

1 University of Edinburgh (GB)

Corresponding Authors: marcus.ebert@ed.ac.uk, andrew.washbrook@cern.ch

ZFS is a combination of file system, logical volume manager, and software raid system developed by SUN Microsystems for the Solaris OS. ZFS simplifies the administration of disk storage and on Solaris it has been well regarded for its high performance, reliability, and stability for many years. It is used successfully for enterprise storage administration around the globe, but so far on such systems ZFS was mainly used to provide storage, like for users home directories, through NFS and similar network related protocols. Since ZFS became available in a stable version on Linux recently, here we will present the usage and benefits of ZFS as backend for WLCG storage servers based on Linux and its advantages over current WLCG storage practices using hardware raid systems.

We tested ZFS in comparison to hardware raid configurations on WLCG DPM storage servers used to provide data storage to the LHC experiments. Tests investigated the performance as well as reliability and behavior in different failure scenarios, such as simulating failures of single disks and whole storage devices. The test results comparing ZFS to other file systems based on a hardware raid vdev will be presented, as well as recommendations for a ZFS based storage setup for a WLCG data storage server based on our test results. Among others, we tested the performance under different vdev and redundancy configurations, behaviour in failure situations, and redundancy rebuild behaviour. We will also report on the importance of ZFS’ own unique features and their benefits for WLCG storage. For example, initial tests using ZFS’ built in compression on sample data containing ROOT files indicated a reduction in space of 4% without any negative impact on the performance. We will report on space reduction and how the compression performance scales to 1PB of LHC experiment data. Scaled to the whole LHC experiments’ data amount, that could provide a significant amount of additional storage at no extra costs to the sites. Since more sites provide data storage also to other non-LHC experiments, be able to use compression could be of even greater benefit to the overall disk capacity provided by a site.

After very promising first results on using ZFS on Linux at one of the NGI UK distributed Tier-2 ScotGrid sites together with the much easier administration and better reliability compared to hardware raid systems, we switched the whole storage on this site to ZFS and will report about the longer term experience of using it, too.

All ZFS tests are based on a Linux system (SL6) with the latest stable ZFS-on-Linux version instead of using a traditional Solaris based system. To make the test results transferable to other WLCG
sites, typical storage servers were used as client machines managing 36 disks of different capacity, used before in hardware raid configurations based on typical hardware raid controllers.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):
Storage systems

Track 8: Security, Policy and Outreach / 534

dCache, towards Federated Identities and Anonymized Delegation

Authors: Albert Rossi¹; Anupam Ashish²; Christian Bernardt³; Dmitry Litvintsev⁴; Gerd Behrmann¹; Juergen Starek¹; Marina Sahakyan¹; Olufemi Segun Adeyemi⁵; Patrick Fuhrmann¹; Paul Millar¹; Tigran Mkrtchyan²

¹ Fermi National Accelerator Laboratory (US)
² Deutsches Elektronen-Synchrotron (DE)
³ FNAL
⁴ NDGF
⁵ Deutsches Elektronen-Synchrotron
⁶ Deutsches Elektronen-Synchrotron (DE)

Corresponding Authors: patrick.fuhrmann@desy.de, arossi@fnal.gov, olufemi.segun.adeyemi@desy.de, marina.sahakyan@desy.de, tigran.mkrtchyan@desy.de, litvinse@fnal.gov, juergen.starek@desy.de, christian.bernardt@desy.de, behrmann@ndgf.org, anupam.ashish@desy.de, paul.millar@desy.de

For over a decade, X509 Proxy Certificates are used in High Energy Physics (HEP) to authenticate users and guarantee their membership in Virtual Organizations, on which subsequent authorization, e.g. for data access, is based upon. Although the established infrastructure worked well and provided sufficient security, the implementation of procedures and the underlying software is often seen as a burden, especially by smaller communities trying to adopt existing HEP software stacks. In addition, it is more efficient to guarantee the identity of a scientist at his home institute, since the necessary identity validation has already been performed. Scientists also depend on service portals for data access and processing, on their behalf. As a result, it is imperative for the infrastructure providers to support delegation of access to these portals for their end-users without compromising data security and identity privacy.

The growing usage of distributed services for similar data sharing and processing have led to the development of novel solutions like OpenID Connect, SAML etc. OpenID Connect is a mechanism for establishing the identity of an end-user based on authentication performed by a trusted third-party identity provider, which thereof can be used by infrastructures to delegate the identity verification and establishment to the trusted entity. After a successful authentication, the portal is in possession of an authenticated token, which can be further used to operate on infrastructure services on behalf of the scientist. Furthermore, these authenticated tokens can be exchanged for more flexible authorized credentials, like Macaroons. Macaroons are bearer tokens and can be used by services to ascertain whether a request is originating from an authorized portal. They are cryptographically verifiable entities and can be embedded with caveats to attenuate their scope before delegation.

In this presentation, we describe how OpenID Connect is integrated with dCache and how it can be used by a service portal to obtain a token for an end-user, based on authentication performed with a trusted third-party identity-provider. We also propose how this token can be exchanged for a Macaroon by an end-user and we show how dCache can be enabled to accept requests bearing delegated Macaroons.
A Tool to Convert CAD Models for Importation into Geant4

Author: Carl Vuosalo

Co-authors: Duncan Carlsmith; Kimberly Palladino; Sridhara Dasu

1 University of Wisconsin-Madison (US)
2 Wisconsin
3 University of Wisconsin-Madison

Corresponding Authors: duncan@hep.wisc.edu, kpalladino@wisc.edu, carl.vuosalo@cern.ch, dasu@hep.wisc.edu

The engineering design of a particle detector is usually performed in a Computer Aided Design (CAD) program, and simulation of the detector’s performance can be done with a Geant4-based program. However, transferring the detector design from the CAD program to Geant4 can be laborious and error-prone.

SW2GDML is a tool that reads a design in the popular SolidWorks CAD program and outputs Geometry Description Markup Language (GDML), used by Geant4 for importing and exporting detector geometries. SW2GDML utilizes the SolidWorks Application Programming Interface for direct access to the design and then converts the geometric shapes described in SolidWorks into standard GDML solids.

Other methods for outputting CAD designs are available, such as the STEP and STL formats, and tools exist to convert these formats into GDML. However, these conversion methods produce very large and unwieldy designs composed of tessellated solids that can reduce Geant4 performance. In contrast, SW2GDML produces compact, human-readable GDML that employs standard geometric shapes rather than tessellated solids.

This talk will describe the development and current capabilities of SW2GDML and plans for its enhancement. The aim of this tool is to automate importation of detector engineering models into Geant4-based simulation programs to support rapid, iterative cycles of detector design, simulation, and optimization.
A build system for multiple package development utilizing Spack

Authors: James Amundson\textsuperscript{1}; Patrick Gartung\textsuperscript{2}

\textsuperscript{1} Fermilab
\textsuperscript{2} Fermilab (US)

Corresponding Authors: amundson@fnal.gov, gartung@fnal.gov

The software suite required to support a modern high energy physics experiment is typically made up of many experiment-specific packages in addition to a large set of external packages. The developer-level build system has to deal with external package discovery, versioning, build variants, user environments, etc. We find that various systems for handling these requirements divide the problem in different ways, making simple substitution of one set of build tools for another impossible. Recently, there has been a growing interest in the HEP community in using Spack,\url{https://github.com/llnl/spack}, to handle various aspects of the external package portion of the build problem. We describe a new build system that utilizes Spack for external dependencies and emphasizes common open source software solutions for the rest of the build process.

Track 1: Online Computing / 537

NaNet: a Configurable Network Interface Card for Trigger and DAQ Systems

Authors: Alessandro Lonardo\textsuperscript{1}; Andrea Biagioni\textsuperscript{2}; Francesca Lo Cicero\textsuperscript{2}; Gianluca Lamanna\textsuperscript{1}; Luca Pontisso\textsuperscript{4}; Michele Martinelli\textsuperscript{1}; Ottorino Frezza\textsuperscript{2}; Paolo Cretaro\textsuperscript{2}; Piero Vicini\textsuperscript{1}; Roberto Ammendola\textsuperscript{6}

Co-authors: Davide Rossetti \textsuperscript{1}; Elena Pastorelli \textsuperscript{2}; Ilaria Neri \textsuperscript{1}; Marco Sozzi \textsuperscript{4}; Massimiliano Fiorini \textsuperscript{7}; Pier Stanislao Paolucci \textsuperscript{4}; Riccardo Fantechi \textsuperscript{2}; Roberto Piandani \textsuperscript{4}; Simula Francesco \textsuperscript{2}; Stefano Di Lorenzo \textsuperscript{4}

\textsuperscript{1} Universita e INFN, Roma I (IT)
\textsuperscript{2} INFN Roma
\textsuperscript{3} Istituto Nazionale Fisica Nucleare Frascati (IT)
\textsuperscript{4} Universita di Pisa & INFN (IT)
\textsuperscript{5} INFN
\textsuperscript{6} Universita e INFN Roma Tor Vergata (IT)
\textsuperscript{7} Universita di Ferrara & INFN (IT)

Corresponding Authors: marco.sozzi@cern.ch, alessandro.lonardo@cern.ch, massimiliano.fiorini@cern.ch, elena.pastorelli@roma1.infn.it, davide.rosetti@cern.ch, michele.martinelli@roma1.infn.it, luca.pontisso@cern.ch, ilaria.neri@cern.ch, francesco.simula@roma1.infn.it, pier.piolucci@roma1.infn.it, roberto.piandani@cern.ch, piero.vicini@cern.ch, paolo.cretaro@roma1.infn.it, gianluca.lamanna@cern.ch, stefano.di.lorenzo@cern.ch, roberto.ammendola@roma2.infn.it, andrea.biagoni@cern.ch, ottorino.frezza@roma1.infn.it, francesca.locicero@roma1.infn.it, riccardo.fantechi@cern.ch

In order to face the LHC luminosity increase planned for the next years, new high-throughput network mechanisms interfacing the detectors readout to the software trigger computing nodes are
being developed in several CERN experiments. Adopting many-core computing architectures such as Graphics Processing Units (GPUs) or the Many Integrated Core (MIC) would allow to reduce drastically the size of the server farm, or to develop completely new algorithms with improved trigger selectivity. NaNet project goal is the design and implementation of PCI Express (PCIe) Network Interface Cards (NICs) featuring low-latency, real-time data transport towards CPUs and nVIDIA GPU accelerators.

Being an FPGA-based NIC, NaNet natively supports a number of different link technologies allowing for its straightforward integration in diverse experimental setups. One of the key features of the design is the capability of managing the network protocol stack in hardware, thus avoiding OS jitter effects and guaranteeing a deterministic behaviour of the communication latency. Furthermore, NaNet integrates a processing stage which is able to reorganize data coming from detectors on the fly in order to improve the efficiency of applications running on the host node. On a per experiment basis different solutions can be implemented, e.g. data compression/decompression and reformatting or merging of event fragments. NaNet accomplishes zero-copy networking by means of a hardware implemented memory copy engine that follows the RDMA paradigm for both CPU and GPU, supporting the nVIDIA GPUDirect RDMA protocol. The RDMA engine is assisted by a proprietary Translation Look-aside Buffer based on Content Addressable Memory performing virtual-to-physical memory address translations. Finally, thanks to its PCIe interface NaNet can be configured either as Gen2 or Gen3 x8 PCIe endpoint.

On the software side, a Linux kernel device driver offers its services to an application level library, which provides the user with a series of functions to: open/close the device; register and de-register circular lists of receiving buffers (CLOPs) in CPU and/or GPU memory; manage software events generated when a receiving CLOP buffer is full (or when a configurable timeout is reached) and received data are ready to be consumed. A configuration of the NaNet design, featuring four 10GbE channels for the I/O and a PCIe x8 Gen3 host interface, has successfully been integrated in the CERN NA62 experiment to interface the readout of the RICH detector to a GPU accelerated server performing multi-ring pattern reconstruction. Results will be then sent to the central L0 processor, where the trigger decision is made taking into account information from other detectors, within the overall time budget of 1 ms. We will describe two multi-rings pattern recognition algorithms we developed specifically to exploit the many-core parallelism of GPUs and discuss the results we obtained during the NA62 2016 data taking.

Tertiary Keyword (Optional):
Parallelization

Secondary Keyword (Optional):
Trigger

Primary Keyword (Mandatory):
Network systems and solutions

Posters B / Break / 538

Mesos exploitation within INDIGO-DataCloud for multi-site long-running services, computational workloads and advanced PaaS features

Authors: Davide Salomoni1; Doina Cristina Aiftimiei2; Germán Moltó3; Giacinto Donvito4; Ignacio Blanquer5; Marica Antonacci6; Tomasz Szepieniec7

1 INFN CNAF
2 INFN - CNAF, IFIN-HH
3 Universidad Politécnica de Valencia
4 INFN-Bari
5 UPV
The European project INDIGO-DataCloud aims at developing an advanced computing and data platform. It provides advanced PaaS functionalities to orchestrate the deployment of Long-Running Services (LRS) and the execution of jobs (workloads) across multiple sites through a federated AAI architecture.

The multi-level and multi-site orchestration and scheduling capabilities of the INDIGO PaaS layer are presented in this contribution, highlighting the benefits introduced by the project to already available infrastructures and data centers.

User application/service deployment requests are expressed using TOSCA, an OASIS standard to specify the topology of services provisioned in IT infrastructures; the TOSCA template describing the application/service deployment is processed by the INDIGO Orchestrator, implementing a complex workflow aimed at fulfilling a user request using information about the health status and capabilities of underlying IaaS and their resource availability, QoS/SLA constraints, the status of the data files and storage resources needed by the service/application. This process allows to achieve the best allocation of the resources among multiple IaaS sites.

On top of the enhanced Cloud Management Frameworks scheduling capabilities developed by the project and described in other contributions, a two-level scheduling exploiting Apache Mesos has been implemented, where the Orchestrator is able to coordinate the deployment of applications/services on top of one or more Mesos clusters.

Mesos allows to share cluster resources (CPU, RAM) across different distributed applications (frameworks) organizing the cluster architecture in two sets of nodes: masters coordinating the work, and slaves executing it.

INDIGO uses and improves two already available Mesos frameworks: Marathon, which allows to deploy and manage LRS, and Chronos, which allows to execute jobs. Important features that are currently missing in Mesos and that are being added by INDIGO include: the elasticity of a Mesos cluster so that it can automatically shrink or expand depending on the tasks queue, the automatic scaling of the user services running on top of the Mesos cluster and a strong authentication mechanism based on OpenID-Connect. Docker containers are widely used in order to simplify the installation and configuration of both services and applications. A further application of this architecture and of these enhancements addresses one of the objectives of the INDIGO project, namely to provide a flexible Batch Systems as a Service, i.e. the possibility to request and deploy a virtual cluster on-demand for submitting batch jobs. To this purpose, the INDIGO team is implementing the integration of HTCondor with Mesos and Docker, as described in detail in another contribution.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Distributed workload management

Primary Keyword (Mandatory):

Cloud technologies

Track 2: Offline Computing / 539

LArSoft: a toolkit for simulation, reconstruction and analysis of liquid argon TPC experiments

Author: Erica Snider

Co-authors: Christopher Jones; Gianluca Petrillo
LArSoft is an integrated, experiment-agnostic set of software tools for liquid argon (LAr) neutrino experiments to perform simulation, reconstruction and analysis within Fermilab art framework. Along with common algorithms, the toolkit provides generic interfaces and extensibility that accommodate the needs of detectors of very different size and configuration. To date, LArSoft has been successfully used by ArgoNeuT, MicroBooNE, LArIAT, SBND, DUNE, and other prototype single phase LAr TPCs. Work is in progress to include support for Dual Phase argon TPCs such as some of the DUNE prototypes.

The LArSoft suite provides a wide selection of algorithms for event generation, simulation of LAr TPCs including optical readouts, and facilities for signal processing and the simulation of “auxiliary” detectors (e.g., scintillators). Additionally, it offers algorithms for the full range of reconstruction from hits on single channels, up to track trajectories, identification of electromagnetic cascades, estimation of particle momentum and energy, and particle identification. LArSoft presents data structures describing common physics objects and concepts, that constitute a protocol connecting the algorithms. It also includes the visualisation of generated and reconstructed objects, which helps with algorithm development.

LArSoft content is contributed by the adopting experiments. New experiments must provide a description of their detector geometry and specific code for the treatment of the TPC electronic signals. With that, the experiments gain instant access to the full set of algorithms within the suite. The improvements which they achieve can be pushed back into LArSoft, allowing the rest of the community to take advantage of them. The sharing of algorithms enabled by LArSoft has been a major advantage for small experiments that have little effort to devote to the creation of equivalent software infrastructure and tools.

LArSoft is also a collaboration of experiments, Fermilab and associated software projects which cooperate in setting requirements, priorities and schedules. A core project team provides support for infrastructure, architecture, software, documentation and coordination, with oversight and input from the collaborating experiments and projects.

In this talk, we outline the general architecture of the software and the interaction with external libraries and experiment specific code. We also describe the dynamics of LArSoft development between the contributing experiments, the projects supporting the software infrastructure LArSoft relies on, and the LArSoft support core project.

Tertiary Keyword (Optional):
Reconstruction

Secondary Keyword (Optional):
Algorithms

Primary Keyword (Mandatory):
Software development process and tools
Data Management and Database Framework for the MICE Experiment

Author: Janusz Martyniak
Co-authors: Durga Rajaram ¹; Henry Nebrensky ²

¹ IIT, Chicago
² Brunel University

Corresponding Authors: j.nebrensky@brunel.ac.uk, janusz.martyniak@imperial.ac.uk, durga@fnal.gov

The international Muon Ionization Cooling Experiment (MICE) currently operating at the Rutherford Appleton Laboratory in the UK, is designed to demonstrate the principle of muon ionization cooling for application to a future Neutrino Factory or Muon Collider. We present the status of the framework for the movement and curation of both raw and reconstructed data. We also review the implementation of a robust database system that has been designed for MICE.

A raw data-mover has been designed to safely upload data files onto permanent tape storage as soon as they have been written out. The process has been automated, and checks have been built in to ensure the integrity of data at every stage of the transfer. The data processing framework has been recently redesigned in order to provide fast turnaround of reconstructed data for analysis. The automated reconstruction is performed on a dedicated machine in the MICE control room and any reprocessing is done at Tier-2 GRID sites. In conjunction with this redesign, a new reconstructed-data-mover has been designed and implemented.

The processing of data, whether raw or Monte Carlo, requires accurate knowledge of the experimental conditions. MICE has several complex elements ranging from beamline magnets to particle identification detectors to superconducting magnets. A Configuration Database which contains information about the experimental conditions (magnet currents, absorber material, detector calibrations, etc.) at any given time has been developed to ensure accurate and reproducible simulation and reconstruction. A fully replicated, hot-standby database system has been implemented with a firewall-protected read-write master running in the control room, and a read-only slave running at a different location. The actual database is hidden from end users by a Web Service layer which provides platform and programming language-independent access to the data.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Databases

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

Taking HEP data management outside of HEP

Author: Robert Illingworth
Co-authors: Bo Jayatilaka ²; Ian Fisk ²

¹ Fermilab
² Fermi National Accelerator Lab. (US)
Corresponding Authors: ian.fisk@cern.ch, illingwo@fnal.gov, bo.jayatilaka@cern.ch

High Energy Physics experiments have long had to deal with huge amounts of data. Other fields of study are now being faced with comparable volumes of experimental data and have similar requirements to organize access by a distributed community of researchers. Fermilab is partnering with the Simons Foundation Autism Research Initiative (SFARI) to adapt Fermilab’s custom HEP data management system (SAM) to catalog genome data. SFARI has petabyte scale datasets stored in the Fermilab Active Archive Facility and needs to catalog the data, organizing it according to metadata for processing and analysis by a diverse community of researchers. The SAM system is used for data management by multiple HEP experiments at Fermilab and is flexible enough to provide the basis for handling other types of data. This presentation describes both the similarities and the differences in requirements and the challenges in adapting an existing system to a new field.

Secondary Keyword (Optional):
Distributed data handling

Primary Keyword (Mandatory):
Experience/plans from outside experimental HEP/NP

Tertiary Keyword (Optional):

---

Virtual Machine Provisioning, Code Management and Data Movement Design for the Fermilab HEPCloud Facility

Author: Steven Timm

Co-authors: Anthony Tiradani 1; Burt Holzman 2; Davide Grassano 1; Gabriele Garzoglio; Hao Wu 3; Hyun Woo Kim 1; R. Glenn Cooper 1; Rahul Krishnamurthy 1; Robert Kennedy 1; Shangping Ren 3; Shivakumar Vinayagam 3; Stuart Fuess 1

1 Fermilab
2 Fermi National Accelerator Lab. (US)
3 Illinois Institute of Technology

Corresponding Authors: hyunwoo@fnal.gov, davidecrivella@yahoo.com, burt@fnal.gov, timmsteve@yahoo.com, tiradami@fnal.gov, v.shivakumar15@gmail.com, fueess@fnal.gov, kennedy@fnal.gov, rahul096@gmail.com, gcooper@fnal.gov, garzoglio@fnal.gov, hwu28@hawk.iit.edu, ren@iit.edu

The Fermilab HEPCloud Facility Project has as its goal to extend the current Fermilab facility interface to provide transparent access to disparate resources including commercial and community clouds, grid federations, and HPC centers. This facility enables experiments to perform the full spectrum of computing tasks, including data-intensive simulation and reconstruction. We have evaluated the use of the commercial cloud to provide elasticity to respond to peaks of demand without over-provisioning local resources. Full scale data-intensive workflows have been successfully completed on Amazon Web Services for two High Energy Physics Experiments, CMS and NOvA, at the scale of 58000 simultaneous cores. This paper describes the significant improvements that were made to the virtual image provisioning system, code caching system, and data movement system to accomplish this work. The virtual image provisioning and contextualization service was extended to multiple AWS regions, and to support experiment-specific data configurations. A prototype Decision Engine was written to determine the optimal availability zone and instance type to run on, minimizing cost and job interruptions. We have deployed a scalable on-demand caching service to deliver code and database information to jobs running on the commercial cloud. It uses the frontier-squid server and CERN VM File System (CVMFS) clients on EC2 instances and utilizes various services provided by AWS to build the infrastructure (stack). We discuss the architecture and load testing benchmarks on the squid servers. We also describe various approaches that were evaluated to transport experimental data to and from the cloud, and the optimal solutions that were used for the bulk of the data transport. Finally we summarize lessons learned from this scale test, and our future plans to expand and improve the Fermilab HEP Cloud Facility.
The LArIAT Experiment’s Data Acquisition and Trigger System

Author: William Badgett

 correponding Author: william.badgett@cern.ch

The LArIAT Liquid Argon Time Projection Chamber (TPC) in a Test Beam experiment explores the interaction of charged particles such as pions, kaons, electrons, muons and protons within the active liquid argon volume of the TPC detector. The LArIAT experiment started data collection at the Fermilab Test Beam Facility (FTBF) in April 2015 and continues to run in 2016. LArIAT provides important particle identification and cross section measurements for future liquid argon detector such as DUNE. The LArIAT detector consists of a 480 wire TPC, integrated PMT and SiPM light collection systems, downstream muon catcher blocks, four upstream multi-wire tracking chambers, time of flight scintillators, particle ID Cerenkov detectors and cosmic rag paddles. Each disparate detector element has independent timing, data acquisition and trigger systems with the significant challenge of integrating all of them into one seamless whole, operating reliably and flexibly within a tightly constrained budget, all while handling the unique test beam timing cycles. We will describe the integrated data acquisition solutions, the central and unusually flexibly trigger mechanism and the challenge of correlating event data across asynchronous subsystems, all of which must be nimble in the fast changing particle test beam world.

Track 1: Online Computing / 548

artdaq: DAQ Software Development Made Simple

Author: John Freeman

co-authors: Eric Flumerfelt, Gennadiy Lukhanin, Kurt Biery, Ronald Rechenmacher, Wesley Ketchum
For a few years now, the artdaq data acquisition software toolkit has provided numerous experiments with ready-to-use components which allow for rapid development and deployment of DAQ systems. Developed within the Fermilab Scientific Computing Division, artdaq provides data transfer, event building, run control, and event analysis functionality. This latter feature includes built-in support for the art event analysis framework, allowing experiments to run art modules for real-time filtering, compression, disk writing and online monitoring; as art, also developed at Fermilab, is also used for offline analysis, a major advantage of artdaq is that it allows developers to easily switch between developing online and offline software.

artdaq continues to be improved. Support for an alternate mode of running whereby data from some subdetector components are only streamed if requested has been added; this option will reduce unnecessary DAQ throughput. Real-time reporting of DAQ metrics has been implemented, along with the flexibility to choose the format through which experiments receive the reports; these formats include the Ganglia, Graphite and syslog software packages, along with flat ASCII files. Additionally, work has been performed investigating more flexible modes of online monitoring, including the capability of being able to run multiple online monitoring processes on different hosts, each running its own set of art modules. Finally, a web-based GUI interface through which users can configure details of their DAQ system has been implemented, increasing the ease of use of the system.

Already successfully deployed on the LArIAT, DarkSide-50, DUNE 35ton and Mu2e experiments, artdaq will be employed for SBND and is a strong candidate for use on ICARUS and protoDUNE. With each experiment comes new ideas for how artdaq can be made more flexible and powerful; the above improvements will be described, along with potential ideas for the future.
Motivated by the complex workflows within Belle II, we propose an approach for efficient execution of workflows on distributed resources that integrates provenance, performance modeling, and optimization-based scheduling. The key components of this framework include modeling and simulation methods to quantitatively predict workflow component behavior; optimized decision making such as choosing an optimal subset of resources to meet demand, assignment of tasks to resources, and placement of data to minimize data movement; prototypical testbeds for workflow execution on distributed resources; and provenance methods for collecting appropriate performance data.

The Belle II experiments deal with massive amounts of data. Designed to probe the interactions of the fundamental constituents of our universe, the Belle II experiments will generate 25 petabytes of raw data per year. During the course of the experiments, the necessary storage is expected to reach over 350 petabytes. Data is generated by the Belle II detector, Monte Carlo simulations, and user analysis. The detector’s experimental data is processed and re-processed through a complex set of operations, which are followed by analysis in a collaborative manner. Users, data, storage and compute resources are geographically distributed across the world creating a complex data intensive workflow.

Belle II workflows necessitate decision making at several levels. We therefore present a hierarchical framework for data driven decision making. Given an estimated demand for compute and storage resources for a period of time, the first (top) level of decision making involves identifying an optimal (sub)set of resources that can meet the predicted demand. We use the analogy of unit commitment problem in electric power grids to solve this problem. Once a cost-efficient set of resources are chosen, the next step is to assign individual tasks from the workflow to specific resources. For Belle II, we consider the situation of Monte Carlo campaigns that involves a set of independent tasks (bag-of-tasks) that need to be assigned on distributed resources.

In order to support accurate and efficient scheduling, predictive performance modeling is employed to rapidly quantify expected task performance across available hardware platforms. The goals of this performance modeling work are to gain insight into the relationship between workload parameters, system characteristics, and performance metrics of interest (e.g., task throughput or scheduling latency); to characterize observed performance; and to guide future and runtime optimizations (including task/module scheduling). Of particular interest, these quantitative and predictive models provide the cost estimates to the higher-level task scheduling algorithms, allowing the scheduler to make informed decisions concerning the optimal resources to utilize for task execution.

Tertiary Keyword (Optional): Distributed workload management
Secondary Keyword (Optional): Data processing workflows and frameworks/pipelines
Primary Keyword (Mandatory): High performance computing

Track 1: Online Computing / 550

Continuous and fast calibration of the CMS experiment: design of the automated workflows and operational experience

Authors: Gianluca Cerminara¹; Piotr Karol Oramus²
Co-authors: Andreas Pfeiffer ¹; Giacomo Govi ³; Giovanni Franzoni ¹; Marco Musich ⁴; Salvatore Di Guida ⁵
The exploitation of the full physics potential of the LHC experiments requires fast and efficient processing of the largest possible dataset with the most refined understanding of the detector conditions. To face this challenge, the CMS collaboration has set up an infrastructure for the continuous unattended computation of the alignment and calibration constants, allowing for a refined knowledge of the most time-critical parameters already a few hours after the data have been saved to disk. This is the prompt calibration framework which, since the beginning of the LHC RunI, enables the analysis and the High Level Trigger of the experiment to consume the most up-to-date conditions optimizing the performance of the physics objects. In RunII this setup has been further expanded to include even more complex calibration algorithms requiring higher statistics to reach the needed precision. This imposed the introduction of a new paradigm in the creation of the calibration datasets for unattended workflows and opened the door to a further step in performance.

The presentation reviews the design of these automated calibration workflows, the operational experience in RunII and the monitoring infrastructure developed to ensure the reliability of the service.

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

Track 3: Distributed Computing / 551

Advances in Grid Computing for the FabrIc for Frontier Experiments Project at Fermilab

Authors: Anna Mazzacane¹; Dave Dykstra²; Gabriele Garzoglio³; Jeny Teheran³; Joseph Boyd¹; Kenneth Richard Herner⁴; Kevin Retzke¹; Marc Mengel¹; Margaret Votava¹; Michael Kirby³; Mine Altunay³; Neha Sharma¹; Tanya Levshina³

¹ FNAL
² Fermi National Accelerator Lab. (US)
³ Fermilab
⁴ Fermi National Accelerator Laboratory (US)
⁵ Fermi National Accelerator Laboratory

Corresponding Authors: kretzke@fnal.gov, votava@fnal.gov, dwdi@fnal.gov, mengel@fnal.gov, maultunay@fnal.gov, garzoglio@fnal.gov, jteheran@fnal.gov, kirby@fnal.gov, boyd@fnal.gov, mazzacan@fnal.gov, neha@fnal.gov, tlevshin@fnal.gov, kenneth.richard.herner@cern.ch

The FabrIc for Frontier Experiments (FIFE) project is a major initiative within the Fermilab Scientific Computing Division charged with leading the computing model for Fermilab experiments. Work within the FIFE project creates close collaboration between experimenters and computing professionals to serve high-energy physics experiments of differing size, scope, and physics area. The FIFE project has worked to develop common tools for job submission, certificate management, software
and reference data distribution through CVMFS repositories, robust data transfer, job monitoring, and databases for project tracking. Since the project’s inception the experiments under the FIFE umbrella have significantly matured, and present an increasingly complex list of requirements to service providers. To meet these requirements, the FIFE project has been involved in transitioning the Fermilab General Purpose Grid cluster to support a partitionable slot model, expanding the resources available to experiments via the Open Science Grid, assisting with commissioning dedicated high-throughput computing resources for individual experiments, supporting the efforts of the HEP Cloud projects to provision a variety of back end resources, including public clouds and high performance computers, and developing rapid onboarding procedures for new experiments and collaborations. The larger demands also require enhanced job monitoring tools, which the project has developed using such tools as ElasticSearch and Grafana. FIFE has also closely worked with the Fermilab Scientific Computing Division’s Offline Production Operations Support Group (OPOS) in helping experiments manage their large-scale production workflows. This group in turn requires a structured service to facilitate smooth management of experiment requests, which FIFE provides in the form of the Production Operations Management Service (POMS). POMS is designed to track and manage requests from the FIFE experiments to run particular workflows, and support troubleshooting and triage in case of problems. Recently we have started to work on a new certificate management infrastructure called Distributed Computing Access with Federated Identities (DCAFI) that will eliminate our dependence on a specific third-party Certificate Authority service and better accommodate FIFE collaborators without a Fermilab Kerberos account. DCAFI integrates the existing InCommon federated identity infrastructure, CILogon Basic CA, and a MyProxy service using a new general purpose open source tool. We will discuss the general FIFE onboarding strategy, progress in expanding FIFE experiments’ presence on the Open Science Grid, new tools for job monitoring, the POMS service, and the DCAFI project. We will also discuss lessons learned from collaborating with the OPOS effort and how they can be applied to improve efficiency in current and future experiment’s computational work.

**Tertiary Keyword (Optional):**
Distributed workload management

**Secondary Keyword (Optional):**
Computing middleware

**Primary Keyword (Mandatory):**
Computing models

---

**Track 5: Software Development / 552**

**Using computing models from particle physics to investigate dose-toxicity correlations in cancer radiotherapy**

**Author:** Karl Harrison

**Co-authors:** Andrew Sultana 1; Andy Parker 1; Hannah Pullen 1; Lin Yeap 1; Marina Romanchikova 2; Michael Sutcliffe 1; Patrick Elwood 1; Simon Thomas 2

1 University of Cambridge
2 Cambridge University Hospitals NHS Foundation Trust

**Corresponding Authors:** mr588@cam.ac.uk, ads73@cam.ac.uk, simon.thomas@addenbrookes.nhs.uk, ypl1991@gmail.com, hp346@cam.ac.uk, parker@hep.phy.cam.ac.uk, mpfs@eng.cam.ac.uk, pje40@cam.ac.uk, karl.harrison@cern.ch

Radiotherapy is planned with the aim of delivering a lethal dose of radiation to a tumour, while keeping doses to nearby healthy organs at an acceptable level. Organ movements and shape changes, over a course of treatment typically lasting four to eight weeks, can result in actual doses being different from planned. The UK-based VoxTox project aims to compute actual doses, at the level of millimetre-scale volume elements (voxels), and to correlate with short- and long-term side effects
(toxicity). The initial focuses are prostate cancer, and cancers of the head and neck. Results may suggest improved treatment strategies, personalised to individual patients.

The VoxTox studies require analysis of anonymised patient data. Production tasks include: calculations of actual dose, based on material distributions shown in computed-tomography (CT) scans recorded at treatment time to guide patient positioning; pattern recognition to locate organs of interest in these scans; mapping of toxicity data to standard scoring systems. User tasks include: understanding differences between planned and actual dose; evaluating the pattern recognition; searching for correlations between actual dose and toxicity scores. To provide for the range of production and user tasks, an analysis system has been developed that uses computing models and software tools from particle physics.

The VoxTox software framework is implemented in Python, but is inspired by the Gaudi C++ software framework of ATLAS and LHCb. Like Gaudi, it maintains a distinction between data objects, which are processed, and algorithm objects, which perform processing. It also provides services to simplify common operations. Applications are built as ordered sets of algorithm objects, which may be passed configuration parameters at run time. Analysis algorithms make use of ROOT. An application using Geant4 to simulate CT guidance scans is under development.

Drawing again from ATLAS and LHCb, VoxTox computing jobs are created and managed within Ganga. This allows transparent switching between different processing platforms, provides cross-platform job monitoring, performs job splitting and output merging, and maintains a record of job definitions. For VoxTox, Ganga has been extended through the addition of components with built-in knowledge of the software framework and of patient data. Jobs can be split based on either patients or guidance scans per sub-job.

This presentation details use of computing models and software tools from particle physics to develop the data-analysis system for the VoxTox project, investigating dose-toxicity correlations in cancer radiotherapy. Experience of performing large-scale data processing on an HTCondor cluster is summarised, and example results are shown.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):
Experience/plans from outside experimental HEP/NP

Track 1: Online Computing / 553

Optical follow-up of gravitational wave triggers with DECam

Authors: Alex Drlica-Wagner 1; Brian Yanny 1; Dillon Brout 2; Edo Berger 2; Eric Neilsen 1; Flavia Sobreira 1; James Annis 1; Kenneth Richard Herner 4; Marcelle Soares-Santos 1; Masao Sako 5; Philip Cowperthwaite 7; Richard Kessler 8; William Wester 9

Co-authors: Alistair Walker 10; Armin Rest 11; Ben Farr 8; Dan Scolnic 3; David Finley 1; Elizabeth Buckley-Geer 1; Hsin-Yu Chen 8; Huan Lin 1; John Marriner 1; Josh Frieman 1; Juan Garcia-Bellido 12; Robert Gruendl 13; Ryan Foley 14; Thomas Diehl 1; Zoheyr Doctor 8; brenna flaugher 1; daniel holz 8

1 Fermilab
2 University of Pennsylvania
3 CFA Harvard University
4 FERMI NATIONAL ACCELERATOR LABORATORY
5 Laboratorio Interinstitucional de e-Astronomia
6 Fermi National Accelerator Laboratory (US)
Gravitational wave (GW) events can have several possible progenitors, including binary black hole mergers, cosmic string cusps, core-collapse supernovae, black hole-neutron star mergers, and neutron star-neutron star mergers. The latter three are expected to produce an electromagnetic signature that would be detectable by optical and infrared telescopes. To that end, the LIGO-Virgo Collaboration (LVC) has agreements with a number of partners to send an alert following a possible GW event detection so that the partners can begin to search for an electromagnetic counterpart. One such partner is the Dark Energy Survey (DES), which makes use of the Dark Energy Camera (DECam), situated on the 4m Blanco Telescope at the Cerro Tololo Inter-American Observatory in Chile. DECam is an ideal instrument for performing optical followup of GW triggers in the southern sky. The DES-GW followup program compares new search images to template images of the same region of sky taken in the past, and selects new candidate objects not present in previous images for further analysis.

Due to the short decay timescale of the expected EM counterparts and the need to quickly eliminate survey areas with no counterpart candidates, it is critical to complete the initial analysis of each night’s images within 24 hours. The computational challenges in achieving this goal include maintaining robust I/O pipelines in the processing, being able to quickly acquire template images of new sky regions outside of the typical DES observing regions, and being able to rapidly provision additional batch computing resources with little advance notice. We will discuss the search area determination, imaging pipeline, general data transfer strategy, and methods to quickly increase the available amount of batch computing through opportunistic use of the Open Science Grid, NERSC, and commercial clouds. We will conclude with results from the first season of observations from September 2015 to January 2016.

Tertiary Keyword (Optional):
Computing models

Secondary Keyword (Optional):
Distributed workload management

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

---

**Track 5: Software Development / 554**

**Deep-Learning Analysis Pipelines on Raw HEP Data from the Daya Bay Neutrino Experiment at NERSC**

**Author:** Evan Racah

---
The use of up-to-date machine learning methods, including deep neural networks, running directly on raw data has significant potential in High Energy Physics for revealing patterns in detector signals and as a result improving reconstruction and the sensitivity of the final physics analyses. In this work, we describe a machine-learning analysis pipeline developed and operating at the National Energy Research Scientific Computing Center (NERSC), processing data from the Daya Bay Neutrino Experiment. We apply convolutional neural networks to raw data from Daya Bay in an unsupervised mode where no input physics knowledge or training labels are used.

Tertiary Keyword (Optional):
High performance computing

Secondary Keyword (Optional):
Analysis tools and techniques

Primary Keyword (Mandatory):
Artificial intelligence/Machine learning

Track 1: Online Computing / 555

An artificial retina processor for track reconstruction at the full LHC crossing rate

Authors: Alessio Piucci1; Daniele Ninci2; Diego Tonelli3; Franco Bedeschi1; Franco Spinella1; Giovanni Punzi1; John Walsh1; Luciano Frances Ristori2; Michael J. Morello4; Pietro Marino5; Riccardo Cenci1; Simone Stracka1

1 Universita di Pisa & INFN (IT)
2 INFN Pisa
3 INFN Trieste
4 Pisa University and INFN
5 Fermi National Accelerator Lab. (US)
6 SNS and INFN-Pisa (IT)
7 SNS & INFN-Pisa (IT)
8 SNS and INFN-Pisa, Italy

The goal of the "INFN-RETINA" R&D project is to develop and implement a parallel computational methodology that allows to reconstruct events with an extremely high number (>100) of charged-particle tracks in pixel and silicon strip detectors at 40 MHz, thus matching the requirements for processing LHC events at the full crossing frequency.
Our approach relies on a massively parallel pattern-recognition algorithm, dubbed "artificial retina", inspired by studies of the processing of visual images by the brain as it happens in nature. Preliminary studies on simulation already showed that high-quality tracking in large detectors is possible with sub-microsecond latencies when this algorithm is implemented in modern, high-speed, high-bandwidth FPGA devices, opening a possibility of making track reconstruction happen transparently as part of the detector readout.

In order to demonstrate that a track-processing system based on the retina algorithm is feasible, we built a sizable prototype of a tracking processor with 3,000 patterns, based on already existing readout boards equipped with Altera Stratix III FPGAs. The detailed geometry and charged-particle activity of a large tracking detector currently in operation are used to assess its performances.

All the processing steps, like dispatching hit data and finding local maxima in the track parameters space, have been successfully implemented in the board at the nominal clock frequency (160 MHz). We test the whole processing chain providing hit sequences as input, and correct parameters for reconstructed tracks were received on the output. Hits are processed at a 1.8-MHz event rate, using boards that had originally been designed for a 1-MHz readout-only functionality.

We report on the test results with such a prototype, and on the scalability prospects to larger detector systems and to higher event rates.

Tertiary Keyword (Optional):
Parallelization

Secondary Keyword (Optional):
Reconstruction

Primary Keyword (Mandatory):
Trigger

The deployment of a large scale object store at the RAL Tier 1

Author: Alastair Dewhurst

Co-authors: Bruno Canning; George Ryall; George Vasilakakos; Ian Johnson; James Adams

Since 2014, the RAL Tier 1 has been working on deploying a Ceph backed object store. The aim is to replace Castor for disk storage. This new service must be scalable to meet the data demands of the LHC to 2020 and beyond. As well as offering access protocols the LHC experiments currently use, it must also provide industry standard access protocols. In order to keep costs down the service must use erasure coding rather than replication to ensure data reliability. This paper will present details of the storage service setup, which has been named Echo, as well as the experience gained from running and upgrading it.

In October 2015 a pre-production service offering the S3 and Swift APIs was launched. This paper will present details of the setup as well as the testing that has been done. This includes the use of S3
as a backend for the CVMFS Stratum 1s, for writing ATLAS log files and for testing FTS transfers. Additionally throughput testing from local experiments based at RAL will be discussed.

While there is certainly interest from the LHC experiments regarding the S3 and Swift APIs, they are still currently dependant on the XrootD and GridFTP protocols. The RAL Tier 1 has therefore also been developing an XrootD and GridFTP plugin for Ceph. Both plugins are built on top of the same libraries that write striped data into Ceph and therefore data written by one protocol will be accessible by the other. In the long term we hope the LHC experiments will migrate to industry standard protocols, therefore these plugins will only provide the features needed by the LHC VOs. This paper will report on the development and testing of these plugins.

**Tertiary Keyword (Optional):**
Distributed data handling

**Secondary Keyword (Optional):**
Computing facilities

**Primary Keyword (Mandatory):**
Object stores

---

**Track 4: Data Handling / 557**

**Astronomy data delivery and processing services with CADC and CANFAR**

**Author:** Sebastien Fabbro

1 NRC Herzberg

**Corresponding Author:** sebfabbro@gmail.com

The Canadian Advanced Network For Astronomical Research (CANFAR) is a digital infrastructure that has been operational for the last six years.

The platform allows astronomers to store, collaborate, distribute and analyze large astronomical datasets. We have implemented multi-site storage and in collaboration with an HEP group at University of Victoria, multi-cloud processing. CANFAR is deeply integrated with the Canadian Astronomy Data Centre (CADC), one of the first public astronomy data delivery service initiated 30 years ago that is now expanding its services beyond data delivery. Individual astronomers, official telescope archives, and large astronomy survey collaborations are the current CANFAR users.

This talk will describe some CANFAR use cases, the internal infrastructure, the lessons learned and future directions.

**Secondary Keyword (Optional):**
Computing facilities

**Primary Keyword (Mandatory):**
Experience/plans from outside experimental HEP/NP

**Tertiary Keyword (Optional):**
Track 2: Offline Computing / 558

Tracking in the Muon g-2 Experiment

Author: Tammy Walton

Corresponding Author: twalton.physics@gmail.com

The Muon g-2 experiment will measure the precession rate of positive charged muons subjected to an external magnetic field in a storage ring. To prevent interference in the magnetic field, both the calorimeter and tracker detectors are situated along the ring and measure the muon’s properties via the decay positron. The influence of the magnetic field and oscillation motions of the muon beam result to parts per million corrections in the muon precession rate. The tracker detectors are designed to precisely measure the profile of the muon beam. The Muon g-2 software uses Fermilab support framework, "art", to manage the data handling and organization of the tracking algorithms. A sophisticated tracking infrastructure is needed to execute efficient and multiple algorithms performing hit pattern recognition, track fitting, and track extrapolation of the decay positrons. In addition, the framework handles the linkage and coordination of data between the tracker and calorimeter detectors. The tracking software takes advantage of all available resources to reconstruct high quality tracks for understanding the beam profile of the muons in the g-2 experiment.

Tertiary Keyword (Optional):
Software development process and tools

Secondary Keyword (Optional):
Data model

Primary Keyword (Mandatory):
Reconstruction

Track 8: Security, Policy and Outreach / 559

Progress in Analysis Preservation with a Focus on Reinterpretation

Authors: Kyle Stuart Cranmer¹; Lukas Alexander Heinrich¹

¹ New York University (US)

Corresponding Authors: lukas.heinrich@cern.ch, kyle.cranmer@cern.ch

LHC data analyses consist of workflows that utilize a diverse set of software tools to produce physics results. The different set of tools range from large software frameworks like Gaudi/Athena to single-purpose scripts written by the analysis teams. The analysis steps that lead to a particular physics result are often not reproducible without significant assistance from the original authors. This severely limits the ability to re-execute the original analysis or to re-use its analysis procedures in new contexts: for instance, reinterpreting the results of a search in the context of a new physics model. We will describe two related packages that have been developed to archive analysis code and the corresponding analysis workflow, which enables both re-execution and re-use.

Following the data model of the W3C PROV standard, we express analysis workflows as a collection of activities (individual data processing steps) that generate entities (data products, such as collections of selected events). An activity is captured as a parametrized executable program in conjunction with its required execution environment. Among various technologies, Docker has been explored most extensively due to its versatility and support both in academic and industry environments. Input parameters are provided in form of JSON objects, while output data is written to storage that is...
separate from the execution environment and shared among all activities of the workflow. Further, each activity publishes JSON data, in order to allow for semantic access to its outputs.

The workflow itself is modeled as a directed acyclic graph (DAG) with nodes representing activities and directed edges denoting dependencies between two activities. Frequently, the complete graph structure and activity parameters are not known until execution time. Therefore, the workflow graph is iteratively built during run-time by a sequence of graph extensions, that collectively represent a workflow template. These extensions, also referred to as stages, schedule new nodes and edges as soon as the required information is available. As the dependency structure of the activities is fully captured in the DAG, mutually independent activities can be computed in parallel, distributed across multiple computing resources, e.g. using container orchestration tools such as Docker Swarm or Kubernetes.

Both activities and workflow stages descriptions are defined using an extensible JSON schema that allows for composition and re-use of both individual activities as well as partial analysis workflows across separate analyses. Finally, it enables us to store and richly query, inspect and present workflow information in the context of analysis archives such as the CERN Analysis Preservation Portal.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Data processing workflows and frameworks/pipelines

Primary Keyword (Mandatory):

Preservation of analysis and data

Posters A / Break / 560

Integrating HPC, Cloud, and Containers for Data-Intensive Computing Platforms

Author: Daniel Murphy-Olson

1 Argonne National Laboratory

Corresponding Author: dolson@mcs.anl.gov

Argonne provides a broad portfolio of computing resources to researchers. Since 2011 we have been providing a cloud computing resource to researchers, primarily using Openstack. Over the last year we’ve been working to better support containers in the context of HPC. Several of our operating environments now leverage a combination of the three technologies which provides infrastructure tailored to the needs of the specific workload. This paper summarizes our experiences integrating HPC, Cloud, and Container environments.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):

Cloud technologies
Integrating Visualization Applications, such as ParaView, into HEP Software Frameworks for In-situ Event Displays

Author: Adam Lyon
Co-authors: Christopher Jones; Jim Kowalkowski

1 Fermilab
2 Fermi National Accelerator Lab. (US)

Corresponding Authors: jbk@fnal.gov, lyon@fnal.gov, cdj@fnal.gov

ParaView is a high performance visualization application not widely used in HEP. It is a long standing open source project led by Kitware[2] and involves several DOE and DOD laboratories and has been adopted by many DOE supercomputing centers and other sites. ParaView is unique in speed and efficiency by using state-of-the-art techniques developed by the academic visualization community which are not found in applications written by the HEP community. In-situ visualization of events, where event details are visualized during processing/analysis, is a common task for experiment software frameworks. Kitware supplies Catalyst[3], a library that enables scientific software to serve visualization objects to client ParaView viewers yielding a real-time event display. Connecting ParaView to the Fermilab art framework and the LArSoft toolkit (a toolkit for reconstruction and analysis with Liquid Argon TPC neutrino detectors), will be described and the capabilities it brings discussed. We will further discuss introducing a visualization “protocol” and generalizing this capability to other visualization systems.


Secondary Keyword (Optional): Analysis tools and techniques
Primary Keyword (Mandatory): Visualization
Tertiary Keyword (Optional):

Neutrino Identification With A Convolutional Neural Network in the NOvA Detectors

Author: Adam Aurisano
Co-authors: Alexander Himmel; Alexander Radovic; Dominick Rocco; Evan Niner; Fernanda Psihas

1 University of Cincinnati
2 Fermilab
3 College of William and Mary
4 University of Minnesota
5 Indiana University
The observation of neutrino oscillation provides evidence of physics beyond the standard model, and the precise measurement of those oscillations remains an important goal for the field of particle physics. Using two finely segmented liquid scintillator detectors located 14 mrad off-axis from the NuMI muon-neutrino beam, NOvA is in a prime position to contribute to precision measurements of the neutrino mass splitting, mass hierarchy, and CP violation.

A key part of that precise measurement is the accurate characterization of neutrino interactions in our detector. This presentation will describe a convolutional neural network based approach to neutrino interaction type identification in the NOvA detectors. The Convolutional Visual Network (CVN) algorithm is an innovative and powerful new approach to event identification which uses the technology of convolutional neural networks, developed in the computer vision community, to identify events in the detector without requiring detailed reconstruction. This approach has produced a 40% improvement in electron-neutrino efficiency without a loss in purity as compared to selectors previously used by NOvA. We will discuss the core concept of convolutional neural networks, modern innovations in convolutional neural network architecture related to the nascent field of deep learning, and the performance of our own novel network architecture in event selection for the NOvA oscillation analyses. This talk will also discuss the architecture and performance of two new variants of CVN. One variant classifies constituent particles of an interaction rather than the neutrino origin which will allow for detailed investigations into event topology and the separation of hadronic and lepton energy depositions in the interaction. The other variant focuses on classifying interactions in the Near Detector to improve cross-section analyses as well as making it possible to search for anomalous tau-neutrino appearance at short baselines.

Tertiary Keyword (Optional):

Algorithms

Secondary Keyword (Optional):

Reconstruction

Primary Keyword (Mandatory):

Analysis tools and techniques

---

**CMS use of allocation based HPC resources**

**Authors:** Bo Jayatilaka1; Dirk Hufnagel1; Elizabeth Sexton-Kennedy1; Stefan Piperov2

1 *Fermi National Accelerator Lab. (US)*
2 *Brown University (US)*

**Corresponding Authors:** dirk.hufnagel@cern.ch, sexton@fnal.gov, piperov@fnal.gov, bo.jayatilaka@cern.ch

The higher energy and luminosity from the LHC in Run2 has put increased pressure on CMS computing resources. Extrapolating to even higher luminosities (and thus higher event complexities and trigger rates) in Run3 and beyond, it becomes clear the current model of CMS computing alone will not scale accordingly. High Performance Computing (HPC) facilities, widely used in scientific computing outside of HEP, present a (at least so far) largely untapped computing resource for CMS. Even just being able use a small fraction of HPC facilities computing resources could significantly increase the overall computing available to CMS. Here we describe the CMS strategy for integrating HPC resources into CMS computing, the unique challenges these facilities present, and how we plan on overcoming these challenges. We also present the current status of ongoing CMS efforts at HPC sites such as NERSC (Cori cluster), SDSC (Comet cluster) and TACC (Stampede cluster).
Evolution of CMS Software

Author: David Lange

1 Princeton University (US)

Corresponding Author: david.lange@cern.ch

The algorithms and infrastructure of the CMS offline software are under continuous change in order to adapt to a changing accelerator, detector and computing environment. In this presentation, we discuss the most important technical aspects of this evolution, the corresponding gains in performance and capability, and the prospects for continued software improvement in the face of challenges posed by the high-luminosity LHC program. Developers in CMS are now able to support the significant detector changes that have occurred during Run 2, while at the same time improving software performance to keep up with increasing event complexity and detector effects due to increased LHC luminosity. We will describe the methods used to achieve and monitor this flexibility in configuration. Finally, the CMS software stack continues to evolve towards modern compilers and techniques, while at the same time addressing short comings across our suite of reconstruction and simulation algorithms. We will discuss our achievements and their impact on the CMS physics program during Run 2 and looking forward to the future.

Simulating storage part of application with Simgrid

Author: cong wang

1 CC-IHEP

Corresponding Author: congfairy2536@gmail.com

Simulation has been used for decades in various areas of computing science, such as network protocol design, microprocessor design. By comparison, current practice in storage simulation is in its infancy. So we are trying to fulfill a simulator with Simgrid to simulate the storage part of application. Cluefs is a lightweight utility to collect data on the I/O events induced by an application when interacting with a file system. Simgrid is a counter-intuitive design approach. It is accepted wisdom
that it must be highly specialized to a target domain and it can be both accurate, fast and scalable. Users can use cluefs to get I/O action sequence of the application, and input all the sequence to the simulator, the simulator will simulate I/O action. The simulator has two mode, sleep mode to simulate all the action according to the time in the trace file, it will produce the log including exact executing information and will produce another trace file which is used for visualization, the visualization part can visualize all the I/O file action sequence of each process to make it clear how the application works in storage part. The second mode is action mode, it simulates all the I/O actions with a platform configuration file. In this file, users can change the storage structure and storage parameters such as Reading bandwidth of disk. And then users can get the executing time of different platform file to make sure if it is worth to change some configuration of your real storage system. So the simulator is useful when you want to know how the storage part of your application works and when you want to know how your application will work with different disks and different structure of storage, you can simulate it with the simulator and you will gain a lot.

Tertiary Keyword (Optional):
Visualization

Secondary Keyword (Optional):
Storage systems

Primary Keyword (Mandatory):
Simulation

Track 2: Offline Computing / 567

Machine learning and parallelism in the reconstruction of LHCb and its upgrade

Author: Lucia Grillo

1 Universita & INFN, Milano-Bicocca (IT)

Corresponding Authors: lucia.grillo@cern.ch, marian.stahl@cern.ch

The LHCb detector at the LHC is a general purpose detector in the forward region with a focus on reconstructing decays of c- and b-hadrons. For Run II of the LHC, a new trigger strategy with a real-time reconstruction, alignment and calibration was developed and employed. This was made possible by implementing an offline-like track reconstruction in the high level trigger. However, the ever increasing need for a higher throughput and the move to parallelism in the CPU architectures in the last years necessitated the use of vectorization techniques to achieve the desired speed and a more extensive use of machine learning to veto bad events early on. This document discusses selected improvements in computationally expensive parts of the track reconstruction, like the Kalman filter, as well as an improved approach to eliminate fake tracks using fast machine learning techniques. In the last part, a short overview of the track reconstruction challenges for the upgrade of LHCb, is given: Running a fully software-based trigger, a large gain in speed in the reconstruction has to be achieved to cope with the 40MHz bunch-crossing rate. Two possible approaches for techniques exploiting massive parallelization are discussed.

Secondary Keyword (Optional):
Parallelizarion

Primary Keyword (Mandatory):
Reconstruction

Tertiary Keyword (Optional):
Track 1: Online Computing / 568

Novel real-time alignment and calibration of the LHCb Detector in Run2

Author: Maurizio Martinelli¹

Co-author: Lucia Grillo ²

¹ Ecole Polytechnique Federale de Lausanne (CH)
² Universita & INFN, Milano-Bicocca (IT)

Corresponding Authors: lucia.grillo@cern.ch, maurizio.martinelli@cern.ch

LHCb has introduced a novel real-time detector alignment and calibration strategy for LHC Run 2. Data collected at the start of the fill are processed in a few minutes and used to update the alignment parameters, while the calibration constants are evaluated for each run. This procedure improves the quality of the online reconstruction. For example, the vertex locator is retracted and reinserted for stable beam conditions in each fill to be centred on the primary vertex position in the transverse plane. Consequently its position changes on a fill-by-fill basis. Critically, this new real-time alignment and calibration procedure allows identical constants to be used in the online and offline reconstruction, thus improving the correlation between triggered and offline selected events. This offers the opportunity to optimise the event selection in the trigger by applying stronger constraints. The required computing time constraints are met thanks to a new dedicated framework using the multi-core farm infrastructure for the trigger. The motivation for a real-time alignment and calibration of the LHCb detector is discussed from both the operational and physics performance points of view. Specific challenges of this novel configuration are discussed, as well as the working procedures of the framework and its performance.

Secondary Keyword (Optional):
Trigger

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

Tertiary Keyword (Optional):

Posters B / Break / 569

Large-scale distributed usage of NLO and multi-leg Monte Carlo event generators on the grid and at Argonne Leadership Computing Facility

Author: Josh Bendavid¹

¹ California Institute of Technology (US)

Corresponding Author: joshbendavid@gmail.com

The increases in both luminosity and center of mass energy of the LHC in Run 2 impose more stringent requirements on the accuracy of the Monte Carlo simulation. An important element in this is the inclusion of matrix elements with high parton multiplicity and NLO accuracy, with the corresponding increase in computing requirements for the matrix element generation step posing a significant challenge. We discuss the large-scale distributed usage of such generators in CMS Monte Carlo production, using both traditional grid resources, as well as the Argonne Leadership Computing Facility (ALCF), including associated challenges in software integration, effective parallelization, and efficient handling of output data.
Performance studies of GooFit on GPUs versus RooFit on CPUs while estimating the statistical significance of a new physical signal

Authors: Adriano Di Florio\(^1\); Alexis Pompili\(^1\)

\(^1\) Universita e INFN, Bari (IT)

Corresponding Authors: alexis.pompili@cern.ch, adriano.diflorio@ba.infn.it

Graphical Processing Units (GPUs) represent one of the most sophisticated and versatile parallel computing architectures available that are nowadays entering the High Energy Physics field. GooFit is an open source tool interfacing ROOT/RooFit to the CUDA platform on nVidia GPUs (it also supports OpeMP). Specifically it acts as an interface between the MINUIT minimization algorithm and a parallel processor which allows a Probability Density Function (PDF) to be evaluated in parallel. In order to test the computing capabilities of GPUs with respect to traditional CPU cores, a high-statistics pseudo-experiment technique has been implemented both in ROOT/RooFit and GooFit frameworks with the purpose of estimating the local statistical significance of the structure observed by CMS close to the kinematical threshold of the J/\psi phi invariant mass in the B\(^+\) to J/\psi phi K\(^+\) decay. The optimized GooFit application running on GPUs provides striking speed-up performances with respect to the RooFit application parallelised on multiple CPU workers through the PROOF-Lite tool. The described technique has been extended to situations when, dealing with an unexpected signal, a global significance must be estimated. The LEE is taken into account by means of a scanning technique in order to consider - within the same background-only fluctuation and everywhere in the relevant mass spectrum - any fluctuating peaking behavior with respect to the background model. The execution time of the fitting procedure for each MC toy considerably increases, thus the RooFit-based approach is not only time-expensive but gets unreliable and the use of GooFit as a reliable tool is mandatory to carry out this p-value estimation method.
HTTP as a Data Access Protocol: Trials with XrootD in CMS’ AAA Project

Author: Justas Balcas

Co-authors: Brian Paul Bockelman; Dorian Kcira; Harvey Newman; Jean-Roch Vlimant; Wayne Hendricks

1 California Institute of Technology (US)
2 University of Nebraska (US)

Corresponding Authors: jean-roch.vlimant@cern.ch, brian.bockelman@cern.ch, twh@caltech.edu, harvey.newman@cern.ch, dorian.kcira@cern.ch, justas.balcas@cern.ch

The main goal of the project is to demonstrate the ability of using HTTP data federations in a manner analogous to today’s AAA infrastructure used from the CMS experiment. An initial testbed at Caltech has been built and changes in the CMS software (CMSSW) are being implemented in order to improve HTTP support. A set of machines is already set up at the Caltech Tier2 in order to improve the support infrastructure for data federations at CMS. As a first step, we are building systems that produce and ingest network data transfers up to 80 Gbps. In collaboration with AAA, HTTP support is enabled at the US redirector and the Caltech testbed. A plugin for CMSSW is being developed for HTTP access based on the libdavix software. It will replace the present fork/exec or curl for HTTP access. In addition, extensions to the Xrootd HTTP implementation are being developed to add functionality to it, such as client-based monitoring identifiers. In the future, patches will be developed to better integrate HTTP-over-Xrootd with the OSG distribution. First results of the transfer tests using HTTP will be presented together with details about the setup.

Secondary Keyword (Optional):
Network systems and solutions

Primary Keyword (Mandatory):
Distributed data handling

Tertiary Keyword (Optional):

The CMS Data Analysis School experience

Author: Nicola De Filippis

1 Politecnico e INFN Bari (IT)

Corresponding Author: nicola.de.filippis@cern.ch

An overview of the CMS Data analysis school (CMSDAS) model and experience is provided. The CMSDAS is the official school that CMS organize every year in US, in Europe and in Asia to train students, Ph.D and young post-docs for the physics analysis. It consists of two days of short exercises about physics objects reconstruction and identification and 2.5 days of long exercises about physics analysis with real data taken LHC. More that 1000 physicists were trained since the first school in 2010. This effort has proven to be a key for the new and young physicists to jump start and contribute to the physics goals of CMS by looking for new physics with the collision data. A description of the
scopes and the organization of the school is provided and results in terms of statistical surveys are presented. Plans for the future are also exposed.

Secondary Keyword (Optional):
Outreach

Primary Keyword (Mandatory):
Analysis tools and techniques

Tertiary Keyword (Optional):

Posters B / Break / 573

SWATCH: Common software for controlling and monitoring the upgraded CMS Level-1 trigger

Author: Alessandro Thea

Co-authors: Carlos Ghabrous Larrea; Christos Lazaridis; Giuseppe Codispoti; Glenn Ronny I. Dirkx; Joschka Lingemann; Karol Bunkowski; Lukasz Kreczko; Simone Bologna; Tom Williams

1 STFC - Rutherford Appleton Lab. (GB)
2 CERN
3 University of Wisconsin-Madison (US)
4 Universita e INFN, Bologna (IT)
5 KU Leuven (BE)
6 University of Warsaw (PL)
7 University of Bristol (GB)
8 Universita & INFN, Milano-Bicocca (IT)

Corresponding Authors: lukasz.kreczko@cern.ch, giuseppe.codispoti@cern.ch, joschka.lingemann@cern.ch, carlos.ghabrous@cern.ch, alessandro.thea@cern.ch, glenn.dirkx@cern.ch, karol.bunkowski@cern.ch, simone.bologna@cern.ch, t.williams@cern.ch, christos.lazaridis@cern.ch

The Large Hadron Collider at CERN restarted in 2015 with a higher centre-of-mass energy of 13 TeV. The instantaneous luminosity is expected to increase significantly in the coming years. An upgraded Level-1 trigger system is being deployed in the CMS experiment in order to maintain the same efficiencies for searches and precision measurements as those achieved in the previous run. This system must be controlled and monitored coherently through software, with high operational efficiency.

The legacy system is composed of approximately 4000 data processor boards, of several custom application-specific designs. These boards are organised into several subsystems; each subsystem receives data from different detector systems (calorimeters, barrel/endcap muon detectors), or with differing granularity. These boards have been controlled and monitored by a medium-sized distributed system of over 40 computers and 200 processes. Only a small fraction of the control and monitoring software was common between the different subsystems; the configuration data was stored in a database, with a different schema for each subsystem. This large proportion of subsystem-specific software resulted in high long-term maintenance costs, and a high risk of losing critical knowledge through the turnover of software developers in the Level-1 trigger project.
The upgraded system is composed of a set of general purpose boards, that follow the MicroTCA specification, and transmit data over optical links, resulting in a more homogeneous system. This system will contain the order of 100 boards connected by 3000 optical links, which must be controlled and monitored coherently. The associated software is based on generic C++ classes corresponding to the firmware blocks that are shared across different cards, regardless of the role that the card plays in the system. A common database schema will also be used to describe the hardware composition and configuration data. Whilst providing a generic description of the upgrade hardware, its monitoring data, and control interface, this software framework (SWATCH) must also have the flexibility to allow each subsystem to specify different configuration sequences and monitoring data depending on its role. By increasing the proportion of common software, the upgrade system’s software will require less manpower for development and maintenance. By defining a generic hardware description of significantly finer granularity, the SWATCH framework will be able to provide a more uniform graphical interface across the different subsystems compared with the legacy system, simplifying the training of the shift crew, on-call experts, and other operation personnel.

We present here, the design of the control software for the upgrade Level-1 Trigger, and experience from using this software to commission the upgraded system.

Tertiary Keyword (Optional):
Control systems

Secondary Keyword (Optional):
DAQ

Primary Keyword (Mandatory):
Trigger

---

Efficient Access to Massive Amounts of Tape-Resident Data

Author: David Yu¹

Co-author: Jerome LAURET ²

¹ Brookhaven National Laboratory (US)
² Brookhaven National Laboratory

Corresponding Authors: jlauret@bnl.gov, david.yu@cern.ch

Randomly restoring files from tapes degrades the read performance primarily due to frequent tape mounts. The high latency and time-consuming tape mount and dismount is a major issue when accessing massive amounts of data from tape storage. BNL’s mass storage system currently holds more than 80 PB of data on tapes, managed by HPSS. To restore files from HPSS, we make use of a scheduler software, called ERADAT. This scheduler system was originally based on code from Oak Ridge National Lab, developed in the early 2000s. After some major modifications and enhancements, ERADAT now provides advanced HPSS resource management, priority queuing, resource sharing,
web-browser visibility of real-time staging activities and advanced real-time statistics and graphs. ERADAT is also integrated with ACSLS and HPSS for near real-time mount statistics and resource control in HPSS. ERADAT is also the interface between HPSS and other applications such as the locally developed Data Carousel providing fair resource-sharing policies and related capabilities. ERADAT has demonstrated great performance at BNL and other scientific organizations.

Secondary Keyword (Optional):
Storage systems

Primary Keyword (Mandatory):
Software development process and tools

Tertiary Keyword (Optional):

Posters A / Break / 576

Fast emulation of track reconstruction in CMS simulation

Author: Matthias Komm

1 Universite Catholique de Louvain (UCL) (BE)

Corresponding Author: matthias.komm@cern.ch

Simulated samples of various physics processes are a key ingredient within analyses to unlock the physics behind LHC collision data. Samples with more and more statistics are required to keep up with the increasing amounts of recorded data. During sample generation, significant computing time is spent on the reconstruction of charged particle tracks from energy deposits which additionally scales with the pileup conditions. In CMS, the Fast Simulation package is developed providing a fast alternative to the standard simulation and reconstruction work flow. It employs various techniques to emulate track reconstruction effects in particle collision events amongst others. Several analysis groups in CMS are utilizing the package, in particular those requiring many samples to scan the parameter space of physics models (e.g. SUSY) or for the purpose of estimating systematic uncertainties. The strategies for and recent developments in this emulation are presented which features a novel, flexible implementation of tracking emulation while retaining a sufficient, tuneable accuracy.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Reconstruction

Primary Keyword (Mandatory):
Simulation

Posters A / Break / 577

Divergence techniques in high energy physics

Author: Petr Bouř

1
Binary decision trees are a widely used tool for supervised classification of high-dimensional data, for example among particle physicists. We present our proposal of the supervised binary divergence decision tree with nested separation method based on kernel density estimation. A key insight we provide is the clustering driven only by a few selected physical variables. The proper selection consists of the variables achieving the maximal divergence measure between two different subclasses of data. Further we apply our method to Monte Carlo data set from the particle accelerator Tevatron at the D0 experiment in Fermilab. We also introduce the modification of statistical tests applicable to weighted data sets in order to test homogeneity of the Monte Carlo simulation and real data.

Tertiary Keyword (Optional):
Artificial intelligence/Machine learning

Secondary Keyword (Optional):
Analysis tools and techniques

Primary Keyword (Mandatory):
Algorithms

Impact of tracker layout on track reconstruction in pp collisions with high pileup at CMS.

Author: Slava Krutelyov
Co-authors: Avi Yagil; Frank Wuerthwein; Giuseppe Cerati; Matevz Tadel

High luminosity operations of the LHC are expected to deliver proton-proton collisions to experiments with average number of pp interactions reaching 200 every bunch crossing. Reconstruction of charged particle tracks in this environment is computationally challenging.

At CMS, charged particle tracking in the outer silicon tracker detector is among the largest contributors to the overall CPU budget of the tracking. Expected costs of the tracking detector upgrades are comparable to that of the computing costs associated with the track reconstruction. We explore potential gains that could be achieved for tracking computational costs for a range of realistic changes in the tracker layout.

A layout with grouped layers placed at shorter distance than the traditional equidistant layer separation shows potential benefits in several aspects:
- increase in locality of track reconstruction up to track segments measured on these layer groups and reduction of combinatorial backgrounds.
- A simplified detector layout emulation based on CMS upgrade tracker detector simulation is used to quantify dependence of tracking computational performance for different detector layouts.
LStore: Logistical Storage

Author: Andrew Malone Melo

1 Vanderbilt University (US)

Corresponding Author: andrew.malone.melo@cern.ch

LStore was developed to satisfy the ever-growing need for cost-effective, fault-tolerant, distributed storage. By using erasure coding for fault-tolerance, LStore has an order of magnitude lower probability of data loss than traditional 3-replica storage while incurring 1/2 the storage overhead. LStore was integrated with the Data Logistics Toolkit (DLT) to introduce LStore to a wider audience. We describe our experiences with the CMS experiment’s multi-petabyte installation capable of reaching sustained transfer rates of hundreds of gigabits per second.

Tertiary Keyword (Optional): Storage systems

A SDN based approach for the ATLAS data acquisition network

Author: William Panduro Vazquez

Co-author: Jorn Schumacher

1 Royal Holloway, University of London
2 University of Paderborn (DE)

Corresponding Authors: j.panduro.vazquez@cern.ch, joern.schumacher@cern.ch

ATLAS is a high energy physics experiment in the Large Hadron Collider located at CERN. During the so called Long Shutdown 2 period scheduled for late 2018, ATLAS will undergo several modifications and upgrades on its data acquisition system in
order to cope with the higher luminosity requirements. As part of these activities, a new read-out chain will be built for the New Small Wheel muon detector and the one of the Liquid Argon calorimeter will be upgraded. The subdetector specific electronic boards will be replaced with new commodity-server-based systems and instead of the custom SLINK-based communication, the new system will make use of a yet to be chosen commercial network technology. The new network will be used as a data acquisition network and at the same time it is intended to allow communication for the control, calibration and monitoring of the subdetectors. Therefore several types of traffic with different bandwidth requirements and different criticality will be competing for the same underlying hardware. One possible way to address this problem is using a SDN based solution.

SDN stands for Software Defined Networking and it is an innovative approach to network management. Instead of the classic network protocols used to build a network topology and to create traffic forwarding rules, SDN allows a centralized controller application to programatically build the topology and create the rules that are loaded into the network devices. The controller can react very fast to new conditions and new rules can be installed on the fly. A typical use case is a network topology change due to a device failure which is handled promptly by the SDN controller. Dynamically assigning bandwidth to different traffic types based on different criteria is also possible. On the other hand, several difficulties can be anticipated such as the connectivity to the controller when the network is booted and the scalability of the number of rules as the network grows.

This work summarizes the evaluation of the SDN technology in the context of the research carried out for the ATLAS data acquisition system upgrade. The benefits and drawbacks of the new approach will be discussed and a deployment proposal will be made.

Secondary Keyword (Optional):
DAQ

Primary Keyword (Mandatory):
Network systems and solutions

Tertiary Keyword (Optional):

Posters A / Break / 582

Identification of High-Momentum Top Quarks, Higgs Bosons, and W and Z Bosons Using Boosted Event Shapes
Author: Ramya Bhaskar

1 University of California, Davis

Corresponding Author: rbhaskar@ucdavis.edu

At the Large Hadron Collider, numerous physics processes expected within the standard model and theories beyond it give rise to very high momentum particles decaying to multihadronic final states. Development of algorithms for efficient identification of such "boosted" particles while rejecting the background from multihadron jets from light quarks and gluons can greatly aid in the sensitivity of measurements and new particle searches. Here we present a new method for identifying boosted high-mass particles by reconstruction of jets and event shapes in Lorentz-boosted reference frames. Variables calculated in these frames for multihadronic jets can then be used as input to a large artificial neural network to discriminate their origin.

Tertiary Keyword (Optional):
Artificial intelligence/Machine learning

Secondary Keyword (Optional):
Algorithms

Primary Keyword (Mandatory):
Reconstruction

Posters B / Break / 583

Experiments Toward a Modern Analysis Environment: Using TMVA and other tools in a functional world with continuous integration for analysis

Author: Gordon Watts

1 University of Washington (US)

Corresponding Author: gwatts@uw.edu

A modern high energy physics analysis code is complex. As it has for decades, it must handle high speed data I/O, corrections to physics objects applied at the last minute, and multi-pass scans to calculate corrections. An analysis has to accommodate multi-100 GB dataset sizes, multi-variate signal/background separation techniques, larger collaborative teams, and reproducibility and data preservation requirements. The result is often a series of scripts and separate programs stitched together by hand or automated by small driver programs scattered around an analysis team’s working directory and disks. Worse, the code is often much harder to read and understand because most of it is dealing with these requirements, not with the physics. This paper describes a framework that is built around the functional and declarative features of the C# language and its Language Integrated Query (LINQ) extensions to declare an analysis. The framework uses language tools to convert the analysis into C++ and runs ROOT or PROOF as a backend to determine the results. This gives the analyzer the full power of an object-oriented programming language to put together the analysis and at the same time the speed of C++ for the analysis loop. A fluent interface has been created for TMVA to fit into this framework, and can be used as a model for incorporating other complex long-running processes into similar frameworks. A by-product of the design is the ability to cache results between runs, dramatically reducing the cost of adding one-more-plot. This lends the analysis to running on a continuous integration server after every check-in (Jenkins). To aid to data preservation a backend that accesses GRID datasets by name and transforms has been added as well. This paper will describe this framework in general terms along with the significant improvements described above.
The Detector Final State pattern: Using the Web Ontology Language to describe a Physics Analysis

Author: Gordon Watts

Co-author: David Carral

Corresponding Authors: carral.2@wright.edu, gwatts@uw.edu

The Data and Software Preservation for Open Science (DASPOS) collaboration has developed an ontology for describing particle physics analyses. The ontology, a series of data triples, is designed to describe dataset, selection cuts, and measured quantities for an analysis. The ontology specification, written in the Web Ontology Language (OWL), is designed to be interpreted by many pre-existing tools, including search engines, and to apply to both theory and experiment published papers. This paper gives an introduction to OWL and this branch of library science from a particle physicist’s point of view, specifics of the Detector Final State Pattern, and how it is designed to be used in the field of particle physics primarily to archive and search analyses. Also included is a description of a SPARK end-point for meta-data powered search. A general introduction to DASPOS and how its other work fits in with this topic will also be described.

The Performance of AMS Offline Software on IBM Blue Gene/Q Architecture

Author: Vitaly Choutko

Co-authors: Alexander Egorov; Baosong Shan

Corresponding Authors: vitaly.choutko@cern.ch, baosong.shan@cern.ch, alexander.egorov@cern.ch

1 Massachusetts Inst. of Technology (US)

2 Beihang University (CN)
The Alpha Magnetic Spectrometer (AMS) on board of the International Space Station (ISS) requires a large amount of computing power for data production and Monte Carlo simulation. Recently the AMS Offline software was ported to IBM Blue Gene/Q architecture. The supporting software/libraries which have been successfully ported include: ROOT 5.34, GEANT4.10, CERNLIB, and AMS offline data reconstruction and simulation software. The operating system on IBM Blue Gene/Q computing nodes is Compute Node Kernel (CNK) on which there is no Linux shell and only a limited set of system calls are supported. A system shell wrapper class is implemented to support the existing shell command calls in the code. To support standalone simulation jobs running on computing nodes, an MPI emulator class is designed to initialize, get arguments, read data cards, start simulation threads, and finally finalize with coordination across all the jobs within the submission to achieve the best use of CPU time. The AMS offline software as well as the supporting software/libraries have been built and tested in JUQUEEN computing center in Juelich, Germany. The performance of the AMS offline software on IBM Blue Gene/Q architecture is presented.

Tertiary Keyword (Optional):
Parallelization

Secondary Keyword (Optional):
High performance computing

Primary Keyword (Mandatory):
Software development process and tools

---

EMMA: A New Paradigm in Configurable Software

Author: Jerzy Nogiec

1 FERMI NATIONAL ACCELERATOR LABORATORY

Corresponding Author: nogiec@fnal.gov

EMMA is a framework designed to build a family of configurable systems, with emphasis on extensibility and flexibility. It is based on a loosely coupled, event driven architecture. The architecture relies on asynchronous communicating components as a basis for decomposition of the system.

EMMA is embracing a fine-grained, component-based architecture, which produces a network of communicating processes organized into components. Components are identified based on the cohesion criterion—the drive to keep related code and functionality grouped together. Each component is independent and can evolve internally as long as its functionality boundaries and external interface remain unchanged. Components are short lived, created for the duration of running a particular system, with the same or different set of components (configuration) for different applications. A component has a set of properties that can be initialized to required values and dynamically modified to alter the behavior of the particular component at run-time.

The system is in nature decentralized with each component accessing its own DAQ hardware or data stores and components communicating over a simple and efficient middleware implemented as a software bus.

The functionality of each application is orchestrated by scripts. Several different scripts can execute on a given set of components providing different behaviors, with each script parameterized to provide an easy way to tailor the application’s behavior.

The EMMA framework has been built upon the premise of composing test systems from independent components. It opens up opportunities for reuse of components and their functionality and composing them together in many different ways. It provides the developer with a lightweight alternative
to microservices, while sharing their various advantages, including composable, loose coupling, encapsulation, and reuse.

Tertiary Keyword (Optional):

Secondary Keyword (Optional):
Software development process and tools

Primary Keyword (Mandatory):
Data processing workflows and frameworks/pipelines

Plenary Session 1 / 592

Welcome

Plenary Session 1 / 593

DOE HEP Viewpoint

Corresponding Author: jim.siegrist@science.doe.gov

Plenary Session 1 / 594

Future of Compute

Corresponding Author: mark.seager@intel.com

Plenary Session 2 / 595

The US Exascale Project

Corresponding Author: messina@anl.gov

Plenary Session 2 / 596

Diversity and Bias: Creating Inclusive Workplaces, See Bias, Block Bias

Corresponding Author: csimard@stanford.edu
Plenary Session 2 / 597

Diversity and Bias: Panel discussion

Plenary Session 8 / 599

TDAQ - Tracking in the Trigger, FPGAs and other approaches

Corresponding Author: brian.winer@cern.ch

Plenary Session 9 / 601

CHEP 2018, Sofia

Corresponding Authors: peter.hristov@cern.ch, petya.petrova@cern.ch

Plenary Session 9 / 602

Closing

Corresponding Author: richard.mount@slac.stanford.edu

Plenary Session 7 / 603

Machine Learning in Science and Beyond

Corresponding Author: pfaldbi@ics.uci.edu

Plenary Session 7 / 604

Opportunities to Collaborate in Data Science + Data Engineering

Corresponding Author: kcopic@gmail.com

Plenary Session 8 / 605

Storage - Sony Everspan

Corresponding Author: horst.schellong@am.sony.com
Storage - The Future

Corresponding Author: per@yttibrium.com

Non-HEP: Data Management for LSST

Corresponding Author: becla@slac.stanford.edu

Non-HEP: Computing Challenges at Light Sources

Corresponding Author: dyparkinson@lbl.gov

Computing at the Electron-Ion Collider

Author: markus diefenthaler¹

¹ Thomas Jefferson National Laboratory

Corresponding Author: mdiefent@jlab.org

The Electron-Ion Collider (EIC) is envisioned as the next-generation U.S. facility to study quarks and gluons in strongly interacting matter. Developing the physics program for the EIC, and designing the detectors needed to realize it, requires a plethora of software tools and multifaceted analysis efforts. Many of these tools have yet to be developed or need to be expanded and tuned for the physics reach of the EIC. Currently, various groups use disparate sets of software tools to achieve the same or similar analysis tasks such as Monte Carlo event generation, detector simulations, track reconstruction, event visualization, and data storage to name a few examples. With a long-range goal of the successful execution of the EIC scientific program in mind, it is clear that early investment in the development of well-defined interfaces for communicating, sharing, and collaborating, will facilitate a timely completion of not just...
the planning and design of an EIC but ultimate delivery the physics capable with an EIC. In this presentation, we give an outline of forward-looking global objectives that we think will help sustain a software community for more than a decade. We then identify the high-priority projects for immediate development and also those, which will ensure an open-source development environment for the future.

Primary Keyword (Mandatory):
Computing models

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

Plenary Session 3 / 611

Networking - View from ESnet

Corresponding Author: imonga@es.net

Plenary Session 3 / 612

Networking - View from HEP

Corresponding Author: shawn.mckee@cern.ch

Plenary Session 5 / 613

Simulation - Key Issues for the Coming Decade

Corresponding Author: federico.carminati@cern.ch

Plenary Session 1 / 614

Future of Distributed Data and Workload Management

Corresponding Author: ifisk@simonsfoundation.org

Plenary Session 4 / 615

Software - HEP Software Foundation Community White Paper Status

Corresponding Author: peter.elmer@cern.ch
Software Challenges and Commonalities - Panel

Corresponding Authors: david.britton@cern.ch, graeme.andrew.stewart@cern.ch, wahid.bhimji@cern.ch, frank.wuerthwein@cern.ch, ian.bird@cern.ch, sexton@fnal.gov

Plenary Session 3 / 617

Dell

Clouds and Competitors

Corresponding Authors: karanbhatia@google.com, tim.bell@cern.ch, tanewill@microsoft.com, rso-bie@uvic.ca

Plenary Session 6 / 619

Computing for HENP Experiments and Programs - Blurring Online and Offline

Corresponding Author: peter.hristov@cern.ch

Plenary Session 6 / 620

Computing for HENP Experiments and Programs - Data Streaming Trends

Corresponding Author: mcromaz@lbl.gov

Plenary Session 5 / 621

Quantum Computing

Corresponding Author: jmartinis@google.com

Track 6: Infrastructures / 622

Supermicro
Plenary Session 8 / 623

Track 1: Online Computing - Highlights
Authors: Frank Winklmeier\textsuperscript{1}; Gene Van Buren\textsuperscript{2}

\textsuperscript{1} University of Oregon (US)
\textsuperscript{2} Brookhaven National Laboratory

Corresponding Authors: frank.winklmeier@cern.ch, gene@bnl.gov

Track 2: Offline Computing - Highlights
Authors: David Lange\textsuperscript{1}; Deborah BARD\textsuperscript{2}; John Marshall\textsuperscript{3}

\textsuperscript{1} Princeton University (US)
\textsuperscript{2} LBL
\textsuperscript{3} University of Cambridge

Corresponding Authors: david.lange@cern.ch, marshall@hep.phy.cam.ac.uk, djbard@lbl.gov

Track 3: Distributed Computing - Highlights
Authors: Latchezar Betev\textsuperscript{1}; Tanya Levshina\textsuperscript{None}; Weidong Li\textsuperscript{2}

\textsuperscript{1} CERN
\textsuperscript{2} Chinese Academy of Sciences (CN)

Corresponding Authors: tlevshin@fnal.gov, latchezar.betev@cern.ch, weidong.li@cern.ch
Plenary Session 8 / 626

Track 4: Data Handling - Highlights

Authors: Elizabeth Gallas¹; Maria Girone²; Patrick Fuhrmann³; Wahid Bhimji⁴

¹ University of Oxford (GB)
² CERN
³ DESY
⁴ Lawrence Berkeley National Lab. (US)

Corresponding Authors: elizabeth.gallas@physics.ox.ac.uk, maria.girone@cern.ch, wahid.bhimji@cern.ch, patrick.fuhrmann@desy.de

Tertiary Keyword (Optional):

Primary Keyword (Mandatory):

Secondary Keyword (Optional):

Plenary Session 9 / 627

Track 5: Software Development - Highlights

Authors: Alberto Aimar¹; Concetta Cartaro²; Florian Uhlig³

¹ CERN
² SLAC
³ GSI - Helmholtzzentrum fur Schwerionenforschung GmbH (DE)

Corresponding Authors: cartaro@slac.stanford.edu, f.uhlig@gsi.de, alberto.aimar@cern.ch

Primary Keyword (Mandatory):

Secondary Keyword (Optional):

Tertiary Keyword (Optional):

Plenary Session 9 / 628

Track 6: Infrastructures - Highlights

Authors: Catherine Biscarat¹; Francesco Prezë²; Margaret Votava³; Olof Barring⁴

¹ LPSC Grenoble, IN2P3/CNRS
² Università degli Studi e INFN Milano (IT)
³ Fermi National Accelerator Laboratory
⁴ CERN
Corresponding Authors: olof.barring@cern.ch, votava@fnal.gov, catherine.biscarat@cern.ch, francesco.prelz@mi.infn.it

Tertiary Keyword (Optional):

Primary Keyword (Mandatory):

Secondary Keyword (Optional):

Plenary Session 8 / 629

Track 7: Middleware, Monitoring and Accounting - Highlights

Authors: Farida Fassi\textsuperscript{1}; Jeremy Coles\textsuperscript{2}; Randy Sobie\textsuperscript{3}

\textsuperscript{1} Universite Mohammed V (MA)
\textsuperscript{2} University of Cambridge (GB)
\textsuperscript{3} University of Victoria (CA)

Corresponding Authors: jeremy.coles@cern.ch, farida.fassi@cern.ch, rsobie@uvic.ca

Tertiary Keyword (Optional):

Primary Keyword (Mandatory):

Secondary Keyword (Optional):

Plenary Session 9 / 630

Track 8: Security, Policy and Outreach - Highlights

Authors: David Michael South\textsuperscript{1}; Hannah Short\textsuperscript{2}; Torre Wenaus\textsuperscript{3}

\textsuperscript{1} Deutsches Elektronen-Synchrotron (DE)
\textsuperscript{2} CERN
\textsuperscript{3} Brookhaven National Laboratory (US)

Corresponding Authors: hannah.short@cern.ch, david.south@cern.ch, wenaus@gmail.com

Tertiary Keyword (Optional):

Primary Keyword (Mandatory):

Secondary Keyword (Optional):

Plenary Session 9 / 631

Track 9: Future Directions - Highlights

Authors: Jan Fridolf Strube\textsuperscript{None}; Oxana Smirnova\textsuperscript{1}; Takanori Hara\textsuperscript{None}
How far can COTS HPC go?

Author: Leng Tau

COTS HPC has evolved for two decades to become an undeniable mainstream computing solution. It represents a major shift away from yesterday’s proprietary, vector-based processors and architectures to modern supercomputing clusters built on open industry standard hardware. This shift enabled the Industry with a cost-effective path to high-performance, scalable and flexible supercomputers (from very simple to extremely complex clusters) to accelerate research & engineering.

Today the landscape is undergoing a new change. New technologies have evolved once again including accelerators, higher speed interconnect, faster memory and wider IO architecture and could break the economies of scale engine driving the COTS model. Even with these remarkable improvements, path to Exascale is a difficult challenge with power consumption an additional glaring constraint. How do we approach a new era in supercomputing with the COTS model? How far can COTS HPC go?

Tertiary Keyword (Optional):

Secondary Keyword (Optional):

Primary Keyword (Mandatory):
Computing facilities