21st International Conference on Computing in High Energy and Nuclear Physics (CHEP2015)

Monday 13 April 2015 - Friday 17 April 2015

OIST

21st International Conference on Computing in High Energy and Nuclear Physics **CHEP2015** Okinawa Japan: April 13 - 17, 2015

Book of Abstracts

Contents

The Front-End Electronics and the Data Acquisition System for a Kinetic Inductance Detector 1 1

Mean PB to Failure – Initial results from a long-term study of disk storage patterns at the RACF 2 1

The Effect of NUMA Tunings on CPU Performance 3 2

The LHCb turbo stream 4 2

Designing a future Conditions Database based on LHC experience 5 3

How much higher can HTCondor fly 6 3

Running and Testing T2 Grid Services with Puppet at GRIF-IRFU 7 4

Modular and scalable RESTful API to sustain STAR collaboration's record keeping 8 . . . 4

Development and test of the DAQ system for a Micromegas prototype installed into the ATLAS experiment 9 5

Pyrame, a rapid-prototyping framework for online systems 10 6

Data Preservation at the Fermilab Tevatron 11 7

LHCb topological trigger reoptimization 12 7

Progress in Multi-Disciplinary Data Life Cycle Management 13 8

Intrusion Detection in Grid computing by Intelligent Analysis of Jobs Behavior –The LHC ALICE Case 14 8

A Case Study in Preserving a High Energy Physics Application 16 9

Performance evaluation of the ATLAS IBL Calibration 17 10

Automation of Large-scale Computer Cluster Monitoring Information Analysis 18 10

Extending DIRAC File Management with Erasure-Coding for efficient storage. 20 11

Data Acquisition for the New Muon g-2 Experiment at Fermilab 21 12

Enabling Object Storage via shims for Grid Middleware 22 12

Current Status of the Ceph Based Storage Systems at the RACF 23 12

Operational Experience Running Hadoop XRootD Fallback 24 13

HistFitter: a flexible framework for statistical data analysis 25 13

Scaling Agile Infrastructure to people 26 14

Optimizing the transport layer of the ALFA framework for the Intel Xeon Phi co-processor 27 15

The Careful Puppet Master: Reducing risk and fortifying acceptance testing with Jenkins CI 28 15

The ATLAS Data Flow system for the second LHC run 29 16

The ATLAS Trigger System: Ready for Run-2 30 17

Operation of the upgraded ATLAS Level-1 Central Trigger System 31 17

ATLAS Jet Trigger Performance in LHC Run I and Initial Run II Updates 32 18

The ATLAS Trigger Core Configuration and Execution System in Light of the ATLAS Upgrade for LHC Run 2 33 19

Upgrade of the ATLAS Level-1 Trigger with event topology information 34 20

Upgrade of the ATLAS Control and Configuration Software for Run 2 35 20

A Validation System for the Complex Event Processing Directives of the ATLAS Shifter Assistant Tool 36 21

ATLAS TDAQ System Administration: evolution and re-design 37 22

Triggering events with GPUs at ATLAS 39 23

gFEX, the ATLAS Calorimeter Level 1 Real Time Processor 40 23

FELIX: a High-throughput network approach for interfacing to front end electronics for ATLAS upgrades 41 24

Evolution of Database Replication Technologies for WLCG 42 25

Scale Out Databases for CERN Use Cases 43 26

Visualization of dCache accounting information with state-of-the-art Data Analysis Tools. 45 26

dCache: Moving to open source development model and tools. 46 27

Computer security: surviving and operating services despite highly skilled and well-funded organised crime groups 47 27

Enabling identity federation for the research and academic communities 48 27

Real-time flavour tagging selection in ATLAS 49 28

Using an Intel Galileo board as a Hardware Token Server 50 28

Distributed Root analysis on an ARM cluster 51 29

Online-Analysis of Hits in the Belle-II Pixeldetector for Separation of Slow Pions from Background 52 . . . 29

The performance and development of the Inner Detector Trigger at ATLAS for LHC Run 2 53 . . . 30

Monte Carlo Production Management at CMS 54 . . . 30

Performance of the ATLAS Muon Trigger in Run I and Upgrades for Run II 55 . . . 31

ALDIRAC, a commercial extension to DIRAC 56 . . . 31

Multi-threaded Object Streaming 57 . . . 32

Online data handling and storage at the CMS experiment 58 . . . 32

Experiences and challenges running CERN's high-capacity tape archive 59 . . . 33

Automated workflows for critical time-dependent calibrations at the CMS experiment. 60 34

Indico - the road to 2.0 61 . . . 34

Vidyo@CERN: A Service Update 62 . . . 35

Offering Global Collaboration Services beyond CERN and HEP 63 . . . 35

The new CERN tape software - getting ready for total performance 64 . . . 36

Big Data Analytics as a Service Infrastructure: Challenges, Desired Properties and Solutions 65 . . . 37

Kalman Filter Tracking on Parallel Architectures 66 . . . 37

Data Science for Improving CERN's Accelerator Complex Control Systems 67 . . . 38

The new ALICE DQM client: a web access to ROOT-based objects 68 . . . 38

COMPUTING STRATEGY OF THE AMS-02 EXPERIMENT 69 . . . 39

A New Event Builder for CMS Run II 70 . . . 39

High-Speed Mobile Communications in Hostile Environments 71 . . . 40

Towards a 21st Century Telephone Exchange at CERN 72 . . . 40

The ALICE Glance Shift Accounting Management System 73 . . . 41

The ALICE Glance Membership Management System 74 . . . 42

Architectures and methodologies for future deployment of multi-site Zettabyte-Exascale data handling platforms 75 . . . 42

NoSQL technologies for the CMS Conditions Database 76 . . . 43

The Application of DAQ-Middleware to the J-PARC E16 Experiment 77 . . . 43

Software for implementing trigger algorithms on the upgraded CMS Global Trigger System 79 . . . 46

Scaling the CERN OpenStack cloud 80 47

Enterprise Social Networking Systems for HEP community 81 47

Towards a production volunteer computing infrastructure for HEP 82 48

Energy Reconstruction using Artificial Neural Networks and different analytic methods in a Highly Granularity Semi-Digital Hadronic Calorimeter. 83 48

Progress in Geant4 electromagnetic physics modeling and validation 84 49

The CERN Lync system : voice-over-ip and much more 85 49

Benchmarking and accounting for the (private) cloud 86 50

Fast TPC online tracking on GPUs and asynchronous data-processing in the ALICE HLT to enable online calibration 87 50

Experience of public procurement of Open Compute servers 88 51

LHCb Build and Deployment Infrastructure for RUNII 89 52

The GÉANT network: addressing current and future needs for the High Energy Physics community. 91 52

ATLAS computing on the HPC Piz Daint machine 92 53

SWATCH: common control SW for the uTCA-based upgraded CMS L1 Trigger 93 53

FPGAs go wireless 94 54

Accelerating Scientific Analysis with SciDB 95 54

A history-based estimation for LHCb job requirements 96 55

Background elimination using the SNIP method for Bragg reflections from a protein crystal measured by a time-of-flight single-crystal neutron diffractometer 97 55

ATLAS Monte Carlo production Run-1 experience and readiness for Run-2 challenges 99 56

Scaling up ATLAS production system for the LHC Run 2 and beyond: project ProdSys2 100 56

Evaluating the power efficiency and performance of multi-core platforms using HEP workloads 101 57

Implementation and use of an highly available and innovative IaaS solution: the Cloud Area Padovana 102 57

SModelS: a framework to confront theories beyond the standard model with experimental data 103 58

CMS@Home: Enabling Volunteer Computing Usage for CMS 104 59

LHCOPN and LHCONE: Status and Future Evolution 105 59

Investigating machine learning to classify events in CMS 107 59

The LHCb Data Aquisition and High Level Trigger Processing Architecture 108 60

Monitoring WLCG with lambda-architecture: a new scalable data store and analytics platform for monitoring at petabyte scale. 109 60

Extending software repository hosting to code review and testing 110 61

Monitoring the Delivery of Virtualized Resources to the LHC Experiments 111 61

Jobs masonry with elastic Grid Jobs 112 62

Pilots 2.0: DIRAC pilots for all the skies 113 62

FTS3 on the Web 114 63

Studies of Big Data meta-data segmentation between relational and non-relational databases 115 64

The Data Quality Monitoring Software for the CMS experiment at the LHC 116 64

The Bayesian analysis toolkit: version 1.0 and beyond 117 65

Open access to high-level data and analysis tools in the CMS experiment at the LHC 118 65

The CMS Tier-0 goes Cloud and Grid for LHC Run 2 119 66

Using the CMS Threaded Framework In A Production Environment 120 66

Developments and applications of DAQ framework DABC v2 121 67

CMS Experience with a World-Wide Data Federation 122 67

Enabling opportunistic resources for CMS Computing Operations 123 68

Exploiting Volatile Opportunistic Computing Resources as a CMS User 124 68

Multimedia Content in the CERN Document Server 125 69

CMS Full Simulation for Run-II 126 69

PROOF Analysis Framework (PAF) 127 70

Pooling the resources of the CMS Tier-1 sites 128 70

DDG4 A Simulation Framework using the DD4hep Detector Description Toolkit 129 . . . 71

The CMS Condition Database system 130 72

HEP cloud production using the CloudScheduler/HTCondor Architecture 131 72

The Library Event Matching classifier for ν_e events in NOvA 132 73

ATLAS Fast Tracker Simulation Challenges 133 73

Distributed analysis in ATLAS 137 74

ATLAS Distributed Computing in LHC Run2 138 75

The ATLAS fast chain MC production project 139 75

Fine grained event processing on HPCs with the ATLAS Yoda system 140 76

Data Preservation in ATLAS 141 76

Analysis Preservation in ATLAS 142 77

The Future of PanDA in ATLAS Distributed Computing 144 77

Dynamic Resource Allocation with the ARC Control Tower 145 78

Evolution of Cloud Computing in ATLAS 146 78

Preparing ATLAS Reconstruction for LHC Run 2 147 79

Status and future evolution of the ATLAS offline software 148 79

Improvements of LHC data analysis techniques at italian WLCG sites. Case-study of the transfer of this technology to other research areas 149 79

The ATLAS Higgs Machine Learning Challenge 150 80

Requirements for a Next Generation Framework: ATLAS Experience 151 81

Integration of PanDA workload management system with Titan supercomputer at OLCF. 152 81

Bringing ATLAS production to HPC resources - A use case with the Hydra supercomputer of the Max Planck Society 153 82

Spanish ATLAS Tier-2 facing up to Run-2 period of LHC 154 83

gLExec integration with the ATLAS PanDA workload management system 155 83

The importance of having an appropriate relational data segmentation in ATLAS 156 . . 84

New data access with HTTP/WebDAV in the ATLAS experiment 157 84

Improved ATLAS HammerCloud Monitoring for local Site Administration 159 85

The ATLAS ARC ssh back-end to HPC 161 85

Exploring JavaScript and ROOT technologies to create Web-based ATLAS analysis and monitoring tools 162 86

ATLAS strategy for primary vertex reconstruction during run-II of the LHC 163 86

A New Petabyte-scale Data Derivation Framework for ATLAS 164 87

Running ATLAS workloads within massively parallel distributed applications using Athena Multi-Process framework (AthenaMP) 165 87

Development of a Next Generation Concurrent Framework for the ATLAS Experiment 166 88

An object-oriented approach to generating highly configurable Web interfaces for the ATLAS experiment 167 88

AGIS: Evolution of Distributed Computing information system for ATLAS 168 89

Design, Results, Evolution and Status of the ATLAS simulation in Point1 project. 169 . . 90

ATLAS@Home: Harnessing Volunteer Computing for HEP 170 90

ATLAS I/O Performance Optimization in As-Deployed Environments 171 91

ATLAS Metadata Infrastructure Evolution for Run 2 and Beyond 172 92

Alignment of the ATLAS Inner Detector Upgraded for the LHC Run II 173 92

ATLAS Tracking Detector Upgrade studies using the fast simulation engine (FATRAS) 174 93

Event-Driven Messaging for Offline Data Quality Monitoring at ATLAS 176 93

Dual-use tools and systematics-aware analysis workflows in the ATLAS Run-II analysis model 177 93

PROOF-based analysis on ATLAS Italian Tiers with Prodsys2 and Rucio 179 94

The migration of the ATLAS Metadata Interface (AMI) to Web 2.0 180 95

Evolution of the Architecture of the ATLAS Metadata Interface (AMI) 181 95

Implementation of the ATLAS Run 2 event data model 182 95

The ATLAS Event Service: A new approach to event processing 183 96

Integration of Russian Tier-1 Grid Center with High Performance Computers at NRC-KI for LHC experiments and beyond HENP 184 97

Recent developments and upgrade to the Geant4 visualization Qt driver 185 97

Overview of Different Exciting Technologies Being Researched in CERN openlab V 186 . 98

Using S3 cloud storage with ROOT and CvmFS 187 98

Protocol benchmarking for HEP data access using HTTP and Xrootd 188 99

Seamless access to HTTP/WebDAV distributed storage: the LHCb storage federation case study and prototype 189 99

Docker experience at INFN-Pisa Grid Data Center 190 100

Virtual Circuits in PhEDEx, an update from the ANSE project 191 100

Bandwidth-sharing in LHCONE, an analysis of the problem 192 101

Dimensions of Data Management: A taxonomy of data-transfer solutions in ATLAS and CMS 193 102

Understanding the CMS Tier-2 network traffic during Run-1 194 103

Lightweight user and production job monitoring system for GlideinWMS 196 103

Comprehensive Monitoring for Heterogeneous Geographically Distributed Storage. 197 . 104

Implementation of an Upward-Going Muon Trigger for Indirect Dark Matter Searches with the NOvA Far Detector 198 104

100G Deployment@(DE-KIT) 199 105

Recent Evolution of the Offline Computing Model of the NOvA Experiment 200 105

Software framework testing at the Intensity Frontier 201 106

Matrix Element Method for High Performance Computing platforms 202 107

Evolution of ATLAS conditions data and its management for LHC run-2 203 107

The ATLAS Software Installation System v2: a highly available system to install and validate Grid and Cloud sites via Panda 204 108

The ATLAS Data Management system - Rucio: commissioning, migration and operational experiences 205 108

Monitoring and controlling ATLAS data management: The Rucio web user interface 206 . 109

Resource control in ATLAS distributed data management: Rucio Accounting and Quotas 207 110

The ATLAS EventIndex: architecture, design choices, deployment and first operation experience. 208 110

Optimisation of the ATLAS Track Reconstruction Software for Run-2 209 111

A JEE Restful service to access Conditions Data in ATLAS 210 111

Applying deep neural networks to HEP job statistics 211 112

BESIII physical offline data analysis on virtualization platform 212 112

The NOvA Simulation Chain 213 113

Data Handling with SAM and ART at the NOvA Experiment 214 114

The Performance of the H.E.S.S. Target of Opportunity Alert System 215 114

Accessing commercial cloud resources within the European Helix Nebula cloud marketplace 216 115

A scalable monitoring for the CMS Filter Farm based on elasticsearch 217 115

File-based data flow in the CMS Filter Farm 218 116

THE DAQ NEEDLE IN THE BIG-DATA HAYSTACK 219 117

Integration of the EventIndex with other ATLAS systems 220 118

QuerySpaces on Hadoop for the ATLAS EventIndex 221 118

Distributed Data Collection for the ATLAS EventIndex. 222 119

A study on dynamic data placement for the ATLAS Distributed Data Management system 223 119

Scalable and fail-safe deployment of the ATLAS Distributed Data Management system Rucio 224 120

AsyncStageOut: Distributed user data management for CMS Analysis 225 120

Cloud Federation - the new way to build distributed clouds 226 121

A prototype Infrastructure for Cloud-based distributed services in High Availability over WAN 227 122

Data preservation for the HERA Experiments @ DESY using dCache technology 228 . . . 122

Dynamic Data Management for the Distributed CMS Computing System 229 123

The diverse use of clouds by CMS 230 124

Quantitative transfer monitoring for FTS3 232 124

Software Development at Belle II 233 125

dCache, evolution by tackling new challenges. 234 125

DEAP-3600 Data Acquisition System 235 126

Evolution of the T2K-ND280 Computing Model 236 126

Integrating Network Awareness in ATLAS Distributed Computing Using the ANSE Project 237 127

The evolution of CERN EDMS 238 127

BESIII production with distributed computing 239 128

Setup of a resilient FTS3 service at GridKa 240 128

HappyFace as a monitoring tool for the ATLAS experiment 241 129

Geant4 VMC 3.0 242 130

ROOT/RooFIT optimizations for Intel Xeon Phi's - first results 243 130

A new Self-Adaptive disPatching System for local cluster 244 131

Optimising Costs in WLCG Operations 246 132

Performance benchmark of LHCb code on state-of-the-art x86 architectures 247 132

Open access for ALICE analysis based on virtualisation technology 248 133

Analysis of Public CMS Data on the VISPA Internet Platform 249 133

ATLAS user analysis on private cloud resources at GoeGrid 250 134

A first look at 100 Gbps LAN technologies, with an emphasis on future DAQ applications 251 134

File Access Optimization with the Lustre Filesystem at Florida CMS T2 252 135

Free cooling on the Mediterranean shore: Energy efficiency upgrades at PIC 253 135

Unlocking data: federated identity with LSDMA and dCache 254 136

New adventures in storage: cloud storage and CDMI 255 136

The VISPA Internet Platform for Scientific Research, Outreach and Education 256 137

Design and development of a Virtual Computing Platform Integrated with Batch Job System 257 138

New developments in the FairRoot framework 258 138

Deep Storage for Big Scientific Data 259 139

The Electronics, Online Trigger System and Data Acquisition System of the J-PARC E16 Experiment 260 139

Tier-1 in Kurchatov Institute: status before Run-2 and HPC integration 261 141

Physics Analysis Software Framework for Belle II 262 141

ARC Control Tower: A flexible generic distributed job management framework 263 . . . 142

Improvements in the CMS Computing System for Run2 264 142

Optimizing CMS build infrastructure via Apache Mesos 265 143

Development of GEM trigger electronics for the J-PARC E16 experiment 266 143

Optimization of the LHCb track reconstruction 267 145

SIMD studies in the LHCb reconstruction software 268 145

LHCb experience with running jobs in virtual machines 269 146

Analysis of CERN Computing Infrastructure and Monitoring Data 270 146

Managing virtual machines with Vac and Vcycle 271 147

Integrating Puppet and Gitolite to provide a novel solution for scalable system management at the MPPMU Tier2 centre 272 147

A Model for Forecasting Data Centre Infrastructure Costs 273 148

The Database Driven ATLAS Trigger Configuration System 275 148

CMS reconstruction improvements for the tracking in large pile-up events 276 149

Event Reconstruction Techniques in NOvA 277 150

The DII-HEP OpenStack based CMS Data Analysis for secure cloud resources 278 151

Dynamic provisioning of local and remote compute resources with OpenStack 279 151

A Comparison of the Overheads Associated with WLCG Federation Technologies 280 . . 152

Scheduling multicore workload on shared multipurpose clusters 281 153

RAL Tier-1 evolution as global CernVM-FS service provider 282 153

Prototype of a production system for CTA with DIRAC 283 154

dCache, enabling tape systems to handle small files efficiently. 284 155

dCache, Sync-and-Share for Big Data 285 155

THttpServer class in ROOT 286 156

Ceph-based storage services for Run2 and beyond 287 157

JSROOT version 3 –JavaScript library for ROOT 288 157

Using the glideinWMS System as a Common Resource Provisioning Layer in CMS 289 . . 158

Integration of DD4hep in the Linear Collider Software Framework 290 158

Named Data Networking in Climate Research and HEP Applications 291 159

USER AND GROUP STORAGE MANAGEMENT AT THE CMS CERN T2 CENTRE 292 . . 161

Software Management for the NOvA Experiment 293 161

Utilizing cloud computing resources for BelleII 294 162

Updating the LISE^{++} software and future upgrade plans 295 163

EOS as the present and future solution for data storage at CERN 296 163

Integrating CEPH in EOS 297 164

Archiving tools for EOS 298 165

The production deployment of IPv6 on WLCG 300 165

LHCb EventIndex 301 166

Disk storage management for LHCb based on Data Popularity estimator 303 167

Glint: VM image distribution in a multi-cloud environment 304 167

CernVM WebAPI - Controling Virtual Machines from the Web 305 168

Cellular Automaton based Track Finding for the Central Drift Chamber of Belle II 306 . . 168

SDN implementations at IHEP 307 169

BESIII Physics Data Storing and Processing on HBase and MapReduce 308 169

The SNiPER offline software framework for non-collider physics experiments 309 170

Accelerating Debugging In A Highly Distributed Environment 310 171

The data acquisition system of the XMASS experiment 311 171

Renovation of HEPnet-J for near-future experiments 312 171

Directory Search Performance Optimization of AMGA for the Belle II Experiment 313 . . 172

Monitoring system for the Belle II distributed computing 314 173

Investigation of High-Level Synthesis tools' applicability to data acquisition systems design based on the CMS ECAL Data Concentrator Card example 315 173

HPC in a HEP lab: lessons learned from setting up cost-effective HPC clusters 316 174

Development of tracker alignment software for the J-PARC E16 experiment 317 174

An integrated solution for remote data access 318 175

Continuous Readout Simulation with FairRoot on the Example of the PANDA Experiment 319 176

The Simulation Library of the Belle II Software System 320 176

Fast event generation on graphics processing unit (GPU) and its integration into the MadGraph system. 321 177

The DIRAC Web Portal 2.0 322 178

Disk storage at CERN 323 178

Federating LHCb datasets using the Dirac File Catalog 324 179

Data Management System of the DIRAC Project 325 179

Monitoring tools of COMPASS experiment at CERN 326 180

Cernbox + EOS: End-user Storage for Science 327 181

Evaluation of NoSQL databases for DIRAC monitoring and beyond 328 181

Belle II production system 329 182

Migration experiences of the LHCb Online cluster to Puppet and Icinga2 330 182

Tile-in-ONE: A web platform which integrates Tile Calorimeter data quality and calibration assessment 331 183

Multicore job scheduling in the Worldwide LHC Computing Grid 333 183

The GridPP DIRAC project - DIRAC for non-LHC communities 334 184

Exploiting CMS data popularity to model the evolution of data management for Run-2 and beyond 335 185

Job monitoring on DIRAC for Belle II distributed computing 337 186

An Research and Development for Evolving Architecture for the Beyond Standard Model 338 186

Public Outreach at RAL: Engaging the Next Generation of Scientists and Engineers 339 . 187

Testing WAN access to storage over IPv4 and IPv6 using multiple transfer protocols 340 . 187

Local storage federation through XRootD architecture for interactive distributed analysis 341 188

Belle II public and private clouds management in VMDIRAC system. 342 188

Base ROOT reference guide on Doxygen 343 189

Experimental quantification of Geant4 PhysicsList recommendations: methods and results 344 189

CMS data distributed analysis with CRAB3 345 190

The GridPP DIRAC project - Implementation of a multi-VO DIRAC service 346 191

Geant4 simulation for a study of a possible use of carbon ion pencil beams for the treatment of ocular melanomas with the active scanning system at CNAO. 347 191

Testable physics by design 348 193

Possibilities for Named Data Networking in HEP 350 194

Dynamic partitioning as a way to exploit new computing paradigms: the cloud usecase. 351 194

Data-analysis scheme and infrastructure at the X-ray free electron laser facility, SACLA 352 195

Efficient time frame building for online data reconstruction in ALICE experiment 353 . . 197

Scientometrics of Monte Carlo simulation: lessons learned and how HEP can profit from them 355 197

Evaluation of containers as a virtualisation alternative for HEP workloads 356 198

Deployment and usage of perfSonar Networking tools for non-HEP communities 358 . . 199

Filesize distribution of WLCG data at the Rutherford Appleton Laboratory Tier1 359 . . . 199

A Grid-based Batch Reconstruction Framework for MICE 360 200

Storage Interface Usage at a Large, Multi-Experiment Teir1 362 200

Online tracking with GPUs at PANDA 363 201

Analysis Traceability and Provenance for HEP 364 202

Distributed Computing for Pierre Auger Observatory 365 203

Judith: A Software Package for Synchronised Analysis of Test-beam Data 366 204

fwk: a go-based concurrent control framework 367 204

docker & HEP: containerization of applications for development, distribution and preservation. 368 205

Evaluation of NoSQL database MongoDB for HEP analyses 369 206

Challenge and Future of Job Distribution at a Multi-VO Grid Site 370 206

Pushing HTCondor and glideinWMS to 200K+ Jobs in a Global Pool for CMS before LHC Run 2 371 207

A New Pileup Mixing Framework for CMS 372 208

Status and Roadmap of CernVM 373 208

GridPP –preparing for Run-2 and the wider context 374 209

The CMS High Level Trigger 375 209

Large-Scale Merging of Histograms using Distributed In-Memory Computing 376 209

Performance of muon-based triggers at the CMS High Level Trigger 377 210

Electrons and photons at High Level Trigger in CMS for Run II 378 210

Performance of Tracking, b-tagging and Jet/MET reconstruction at the CMS High Level Trigger 379 211

Implementing a Domain Specific Language to configure and run LHCb Continuous Integration builds 380 211

ROOT6: a quest for performance 381 212

HLT configuration management system 382 212

CMS High Level Trigger Timing Measurements 383 213

Using DD4Hep through Gaudi for new experiments and LHCb 384 213

Improved interface for the LHCb Continuous Integration System 385 214

Clusteralive 386 214

Managing competing elastic Grid and Cloud scientific computing applications using OpenNebula 387 215

Development of site-oriented Analytics for Grid computing centres 388 215

Integrated Monitoring-as-a-service for Scientific Computing Cloud applications using the Elasticsearch ecosystem 389 216

A flexible and modular data format ROOT-based implementation for HEP 390 217

pawgo: an interactive analysis workstation 391 217

Acceleration of ensemble machine learning methods using many-core devices 392 218

SkyGrid - where cloud meets grid computing 393 219

Interpolation between multi-dimensional histograms using a new non-linear moment morphing method 394 220

Towards Reproducible Experiment Platform 395 220

CMS Detector Description for Run II and Beyond 396 221

Towards generic volunteer computing platform 397 221

A Generic Framework for Rapid Development of OPC UA Servers 399 222

Configuration Management and Infrastructure Monitoring using CFEngine and Icinga for real-time heterogeneous data taking environment 400 222

Recent advancements in user-job management with Ganga 401 223

POSIX and Object Distributed Storage system –performance Comparison studies and real-life usage in an experimental data taking context leveraging OpenStack/Ceph. 402 . . 224

The Geant4 physics validation repository 403 225

Open Data and Data Analysis Preservation Services for LHC Experiments 405 226

Statistical analysis of virtualization performance in high energy physics software 406 . . 227

Integrating network and transfer metrics to optimize transfer efficiency and experiment workflows 407 228

Identifying and Localizing Network Problems using the PuNDIT Project 408 228

Evolution of CMS workload management towards multicore job support 409 229

Monitoring data transfer latency in CMS computing operations 410 230

ROOT 6 and beyond: TObject, C++14 and many cores. 411 230

Data-driven estimation of neutral pileup particle multiplicity in high-luminosity hadron collider environments 412 231

GPU Accelerated Event-by-event Reweighting for a T2K Neutrino Oscillation Analysis 413 231

Alignment and calibration of Belle II tracking detectors 414 232

Geant4 Computing Performance Benchmarking and Monitoring 415 232

Optimisation of the usage of LHC and local computing resources in a multidisciplinary physics department hosting a WLCG Tier-2 centre 416 233

CosmoSIS: a system for MC parameter estimation 417 233

Exploring Two Approaches for an End-to-End Scientific Analysis Workflow 418 234

Integration of XRootD into the cloud infrastructure for ALICE data analysis 419 235

Deep Integration: Python in the Cling World 420 235

Breaking the Silos: The art Documentation Suite 421 236

ALFA: The new ALICE-FAIR software framework 422 236

A Comparative Analysis of Event Processing Frameworks used in HEP 423 237

Online/Offline reconstruction of trigger-less readout in the R3B experiment at FAIR 425 . 237

MiniAOD: A new analysis data format for CMS 426 238

Evolution of the Open Science Grid Application Software Installation Service (OASIS) 427 238

Detector Simulation On Modern Coprocessors 428 239

Simulation and Reconstruction Upgrades for the CMS experiment 429 240

Tracker software for Phase-II CMS 430 240

Use of containerisation as an alternative to full virtualisation in grid environments. 431 . 240

The evolving grid paradigm and code ”tuning” for modern architectures- are the two mutually exclusive? 432 241

ConDB - Conditions Database for Particle Physics Experiments 434 241

ATLAS Public Web Pages: Online Management of HEP External Communication Content 435 242

High Speed Fault Tolerant Secure Communication for Muon Chamber using FPGA based GBTx Emulator 436 242

Redundant Web Services Infrastructure for Data Intensive and Interactive Applications 437 243

Maintaining Traceability in an Evolving Distributed Computing Environment 438 244

A design study for the upgraded ALICE O2 computing facility 439 244

Subtlenoise: reducing cognitive load when monitoring distributed computing operations 440 245

Real-time alignment and calibration of the LHCb Detector in Run2 441 246

The OSG Open Facility: A Sharing Ecosystem Using Harvested Opportunistic Resources 442 246

Engineering the CernVM-FileSystem as a High Bandwidth Distributed Filesystem for Auxiliary Physics Data 443 247

Advances in Distributed High Throughput Computing for the Fabric for Frontier Experiments Project at Fermilab 444 248

slic : A full-featured Geant4 simulation program 445 248

org.lcsim: A Java-based tracking toolkit 446 249

Interoperating Cloud-based Virtual Farms 447 250

Cloud services for the Fermilab scientific stakeholders 448 250

Integrating grid and cloud resources at the RAL Tier-1 449 251

Implementation of the vacuum model using HTCondor 450 251

MAUS: The MICE Analysis and User Software 451 251

Experience with batch systems and clouds sharing the same physical resources 452 . . . 252

A Data Summary File Structure and Analysis Tools for Neutrino Oscillation Analysis at the NOvA Experiment 453 252

The Linear Collider Software Framework 454 253

Efficient provisioning for multicore applications with LSF 455 253

Migrating to 100GE WAN Infrastructure at Fermilab 456 254

Multicore-Aware Data Transfer Middleware (MDTM) 457 255

Recent Developments in the Infrastructure and Use of artdaq 458 256

Application-Oriented Network Traffic Analysis based on GPUs 459 256

A virtual validation cluster for ALICE software releases based on CernVM 460 257

Lightweight scheduling of elastic analysis containers in a competitive cloud environment: a Docked Analysis Facility for ALICE 461 258

Archiving Scientific Data outside of the traditional High Energy Physics Domain, using the National Archive Facility at Fermilab 462 258

Replacing the Engines without Stopping The Train; How A Production Data Handling System was Re-engineered and Replaced without anyone Noticing. 463 259

Large Scale Management of Physicist's Personal Analysis Data 464 259

Large Scale Monte Carlo Simulation of neutrino interactions using the Open Science Grid and Commercial Clouds 465 260

Improvement of AMGA Python Client Library for the Belle II Experiment 466 261

Dark Energy Survey Computing on FermiGrid 467 261

The Belle II analysis on Grid 468 262

The GENFIT Library for Track Fitting and its Performance in Belle II 469 262

A quantative evaluation of different methods for instantiating private cloud virtual machines 470 263

Rate-Equation -based model for describing the resonance laser ionization process 472 . . 263

How do particle physicists learn the programming concepts they need? 473 264

Using R in ROOT with the ROOT-R package 474 265

New developments in the ROOT function and fitting classes 475 265

Clad - Automatic Differentiation Using Cling in ROOT 476 265

High performance data analysis via coordinated caches 477 266

IgProf profiler support for power efficient computing 478 267

Multi-VO Support in IHEP's Distributed Computing Environment 479 267

VecGeom: A new vectorised geometry library for particle-detector simulation 480 268

A multi-port 10GbE PCIe NIC featuring UDP offload and GPUDirect capabilities. 481 . . 269

Status report of the migration of the CERN Document Server to the invenio-next package 482 . . . 270

Hardware and Software Design of FPGA-based PCIe Gen3 interface for APENet+ network interconnect system 483 . . . 270

Effective administration through gaining the portability of the LHC Computing Grid sites 484 . . . 271

First statistical analysis of Geant4 quality software metrics 485 . . . 272

Background decomposition of the GERDA data with BAT 487 . . . 273

Evaluation of 'OpenCL for FPGA' for Data Acquisition and Acceleration in High Energy Physics applications 488 . . . 273

A Virtual Geant4 Environment 489 . . . 274

Active job monitoring in pilots 490 . . . 274

Getting prepared for the LHC Run2: the PIC Tier-1 case 491 . . . 275

Network technology research on IhepCloud platform 492 . . . 276

Future Computing Platforms for Science in a Power Constrained Era 493 . . . 277

Storage solutions for a production-level cloud infrastructure 494 . . . 277

A Grid and cloud computing farm exploiting HTCondor 495 . . . 278

IceProd2: A Next Generation Data Analysis Framework for the IceCube Neutrino Observatory 496 . . . 279

The Belle II Conditions Database 497 . . . 280

Architecture of a new data taking and analysis infrastructure and services for the next generation detectors of Petra3 at DESY 498 . . . 280

Designing Computing System Architecture and Models for the HL-LHC era 499 . . . 281

Building a Tier-3 Based on ARMv8 64-bit Server-on-Chip for the WLCG 500 . . . 282

Building a bridge between cloud storage and GPFS 501 . . . 282

The ALICE High Level Trigger, status and plans 502 . . . 283

Geant4 Version 10 Series 503 . . . 283

Processing of data from innovative parabolic strip telescope. 504 . . . 284

Pilot run of the new DAQ of the COMPASS experiment 506 . . . 285

Event Building Process for Time streamed data 507 . . . 286

The CMS BRIL Data Acquisition system 508 . . . 287

Discovering matter-antimatter asymmetries with GPUs 509 . . . 288

The Heavy Photon Search Experiment Software Environment 510 288

Use of jumbo frames for data transfer over the WAN 511 289

The Pattern Recognition software for the PANDA experiment 512 289

The NOvA DAQ Monitor System 513 290

Performance of the NOvA Data Driven Triggering System with the full 14 kT Far Detector 515 291

Integration of the Super Nova Early Warning System with the NOvA Trigger 516 292

Commissioning HTCondor-CE for the Open Science Grid 519 292

Timing in the NOvA detectors with atomic clock based time transfers between Fermilab, the Soudan mine and the Nova Far detector. 520 292

Development of New Data Acquisition System at Super-Kamiokande for Nearby Supernova Bursts 522 293

MAD –Monitoring ALICE Dataflow 523 294

Experience in running relational databases on clustered storage 524 294

DataBase on Demand : insight how to build your own DBaaS 525 295

How the Monte Carlo production of a wide variety of different samples if centrally handled in the LHCb experiment. 526 295

Agile Research - Strengthening Reproducibility in Collaborative Data Analysis Projects 527 296

Performance of the NOvA Data Acquisition System with the full 14 kT Far Detector 528 . 296

The NOvA Data Acquisition Error Handling System 529 297

Monitoring cloud-based infrastructures 530 298

The GeantV project: preparing the future of simulation 531 298

WLCG Monitoring, consolidation and further evolution 532 299

Sharing lattice QCD data over a widely distributed file system 534 299

Data Integrity for Silent Data Corruption in Gfarm File System 535 300

Simulation of LHC events on a million threads 536 301

Achieving production-level use of HEP software at the Argonne Leadership Computing Facility 537 302

4-Dimensional Event Building in the First-Level Event Selection of the CBM Experiment. 538 302

Monitoring Evolution at CERN 539 303

Performance of the CMS High Level Trigger 545 303

ATLAS High-Level Trigger algorithms for Run-2 data-taking 546 304

Welcome Address 547 304

Welcome Address 548 304

Future plan of KEK 549 305

Computing at the Belle-II experiment 550 305

Evolution of Computing and Software at LHC: from Run 2 to HL-LHC 551 305

Intel Corporation 552 305

Dell Inc. 553 305

Computing in Intensity Frontier Accelerator Experiments 554 305

KAGRA and the Global Network of Gravitational Wave Detectors -Construction Status and Prospects for GW Astronomy with Data Sharing Era- 555 305

Challenges of Developing and Maintaining HEP ”Community” Software 556 305

IBM Corporation 557 306

Lenovo Corporation 558 306

Diversity in Computing Technologies: Grid, Cloud, HPC ... and Strategies for Dynamic Resource Allocation 559 306

Distributed Data Management and Distributed File Systems 560 306

Computing at FAIR 561 306

Expanding OpenStack community in academic fields 562 306

Computer Security for HEP 563 306

EMC Corporation 564 307

Future of DAQ Frameworks and Approaches, and Their Evolution toward ‘Internet of Things’ 565 307

The Changing Face of Networks and Implications for Future HEP Computing Models 566 307

Experiences on File Systems: Which is the best FS for you? 567 307

Track5 Summary 568 307

Track7 Summary 569 307

Track8 Summary 570 308

Track1 Summary 571 308

Track2 Summary 572 308

Track3 Summary 573 308

Track4 Summary 574 308

Track6 Summary 575 308

Lightning Talks 576 308

Invitation to CHEP2016 577 309

Closing Address 578 309

HEP Computing: A Tradition of Scientific Leadership 579 309

Introduction to CHEP parallel program summaries 580 309

Track 1 Session / 1

The Front-End Electronics and the Data Acquisition System for a Kinetic Inductance Detector

Author: paolo branchini[1]

Co-authors: Antonio Budano [1]; Dedalo Marchetti [1]; Luciano Capasso [1]

[1] *INFN Roma Tre*

Corresponding Author: paolo.branchini@roma3.infn.it

The Data Acquisition System (DAQ) and the Front-End electronics for an array of Kinetic Inductance Detectors (KIDs) are described. KIDs are superconductive detectors, in which electrons are organized in Cooper pairs. Any incident radiation could break such pairs generating quasi-particles, whose effect is increasing the inductance of the detector. Electrically, any KID is equivalent to a parallel RLC resonant circuit. An array of N KIDs is composed of N pixels, each one resonating on its own frequency. A feed line passes close to each KID and delivers a unique Stimulus signal containing all the resonant frequencies. If one of the KIDs was hit by some radiation, its inductance would change and the corresponding sine component in the readout signal would have its intensity reduced and its phase shifted. The DAQ system we developed is a hardware/software co-design, based on state machines and a Microprocessor embedded into an FPGA. A commercial DAC/ADC board is used to interface the FPGA to the analog environment of the array of KIDs. The DAQ system generates a Stimulus signal for an array of up to 128 KIDs, by creating and adding up 128 sinusoids parted by one MHz. The Stimulus is in the form of a Look-Up Table and it is provided to the DAC device. The analog signal generated is up-mixed with a 3 GHz carrier wave and it then travels through the KIDs array. The read-out signal from the detector is down-mixed with respect to the 3 GHz sine wave and it is read back by the ADC device. The microprocessor stores the read out data via a PCI bus into an external disk. It also elaborates the Fast Fourier Transform of the acquired read out signal: this allows to extrapolate which KID interacted and the energy of the impinging radiation. Simulations and tests have been performed successfully and experimental results are presented.

Track 3 Session / 2

Mean PB to Failure – Initial results from a long-term study of disk storage patterns at the RACF

Author: Tony Wong[1]

Co-authors: Alexandr Zaytsev [2]; Christopher Hollowell [1]; Costin Caramarcu [2]; Tejas Rao [1]; William Strecker-Kellogg [3]

[1] *Brookhaven National Laboratory*

[2] *Brookhaven National Laboratory (US)*

[3] *Brookhaven National Lab*

Corresponding Author: hollowec@bnl.gov

The RACF (RHIC-ATLAS Computing Facility) has operated a large, multi-purpose dedicated computing facility since the mid-1990's, serving a worldwide, geographically diverse scientific community that is a major contributor to various HEPN projects. A central component of the RACF is the Linux-based worker node cluster that is used for both computing and data storage purposes. It currently has nearly 50,000 computing cores and over 23 PB of storage capacity distributed over 12,000+ (non-SSD) disk drives.

The majority of the 12,000+ disk drives provides a cost-effective solution for dCache/xRootd-managed storage, and a key concern is the reliability of this solution over the lifetime of the hardware, particularly as the number of disk drives and the storage capacity of individual drives grow. We report

initial results of a long-term study to measure lifetime PB read/written to disk drives in the worker node cluster. We discuss the historical disk drive mortality rate, disk drive manufacturers' published MPBTF (Mean PB to Failure) data and how they are correlated to our results.

The results helps the RACF understand the productivity and reliability of its storage solutions and has implications for other highly-available storage systems (NFS, GPFS, CVMFS, etc) with large I/O requirements.

Track 8 Session / 3

The Effect of NUMA Tunings on CPU Performance

Author: Christopher Hollowell[1]

Co-authors: Alexandr Zaytsev [2]; Costin Caramarcu [2]; Tony Wong [1]; William Strecker-Kellogg [3]

[1] *Brookhaven National Laboratory*

[2] *Brookhaven National Laboratory (US)*

[3] *Brookhaven National Lab*

Corresponding Author: hollowec@bnl.gov

Non-uniform memory access (NUMA) is a memory architecture for symmetric multiprocessing (SMP) systems where each processor is directly connected to separate memory. Indirect access to other CPU's (remote) RAM is still possible, but such requests are slower as they must also pass through that memory's controlling CPU. In concert with a NUMA-aware operating system, the NUMA hardware architecture can help eliminate the memory performance reductions generally seen in SMP systems when multiple processors simultaneously attempt to access memory.

The x86 CPU architecture has supported NUMA for a number of years. Modern operating systems such as Linux support NUMA-aware scheduling, where the OS attempts to schedule a process to the CPU directly attached to the majority of its RAM. In Linux, it is possible to further manually tune the NUMA subsystem using the "numactl" utility. With the release of Red Hat Enterprise Linux (RHEL) 6.3, the "numad" daemon became available in this distribution. This daemon monitors a system's NUMA topology and utilization, and automatically makes adjustments to optimize locality.

As the number of cores in x86 servers continues to grow, efficient NUMA mappings of processes to CPUs/memory will become increasingly important. This presentation gives a brief overview of NUMA, and discusses the effects of manual tunings and numad on the performance of the HEPSPEC06 benchmark.

Track 1 Session / 4

The LHCb turbo stream

Authors: Mika Anton Vesterinen[1]; Sean Benson[2]

Co-authors: J Michael Williams [3]; Vladimir Gligorov [2]

[1] *Ruprecht-Karls-Universitaet Heidelberg (DE)*

[2] *CERN*

[3] *Massachusetts Inst. of Technology (US)*

Corresponding Author: sean.benson@cern.ch

The LHCb experiment will record an unprecedented dataset of beauty and charm hadron decays during Run II of the LHC, set to take place between 2015 and 2018. A key computing challenge is to store and process this data, which limits the maximum output rate of the LHCb trigger. So far, LHCb has written out a few kHz of events containing the full raw sub-detector data, which are passed through a full offline event reconstruction before being considered for physics analysis. Charm physics in particular is limited by trigger output rate constraints. A new streaming strategy includes the possibility to perform the physics analysis with candidates reconstructed in the trigger, thus bypassing the offline reconstruction. In the "turbo stream" the trigger will write out a compact summary of "physics" objects containing all information necessary for analyses, and this will allow an increased output rate and thus higher average efficiencies and smaller selection biases. This idea will be commissioned and developed during 2015 with a selection of physics analyses. It is anticipated that the turbo stream will be adopted by an increasing number of analyses during the remainder of LHC Run-II (2015-2018) and ultimately in Run-III (starting in 2020) with the upgraded LHCb detector.

Track 3 Session / 5

Designing a future Conditions Database based on LHC experience

Authors: Andrea Formica[1]; Elizabeth Gallas[2]

Co-authors: Andreas Pfeiffer [3]; Dario Barberis [4]; Giacomo Govi [5]; Giovanna Lehmann Miotto [3]

[1] *CEA/IRFU,Centre d'etude de Saclay Gif-sur-Yvette (FR)*

[2] *University of Oxford (GB)*

[3] *CERN*

[4] *Università e INFN Genova (IT)*

[5] *Fermi National Accelerator Lab. (US)*

Corresponding Author: andrea.formica@cern.ch

The ATLAS and CMS Conditions Database infrastructures have served each of the respective experiments well through LHC Run 1, providing efficient access to a wide variety of conditions information needed in online data taking and offline processing and analysis. During the long shutdown between Run 1 and Run 2, we have taken various measures to improve our systems for Run 2. In some cases, a drastic change was not possible because of the relatively short time scale to prepare for Run 2. In this process, and in the process of comparing to the systems used by other experiments, we realized that for Run 3, we should consider more fundamental changes and possibilities.

We seek changes which would streamline conditions data management, improve monitoring tools, better integrate the use of metadata, incorporate analytics to better understand conditions usage, as well as investigate fundamental changes in the storage technology, which might be more efficient while minimizing maintenance of the data as well as simplify the access to it. This contribution will describe architectures we are evaluating and testing which we think will address the problematic areas while providing improved services.

6

How much higher can HTCondor fly

Authors: Edgar Fajardo Hernandez[1]; Jeffrey Michael Dost[1]

Co-authors: Anthony Tiradani [2]; Brian Paul Bockelman [3]; Burt Holzman [4]; David Alexander Mason [4]; Jaime Frey [5]; James Letts [1]; Todd Tannenbaum [6]

[1] *Univ. of California San Diego (US)*

[2] *Fermilab*

[3] *University of Nebraska (US)*

[4] *Fermi National Accelerator Lab. (US)*

[5] *University of Wisconsin*

[6] *Univ of Wisconsin-Madison, Wisconsin, USA*

Corresponding Author: edgar.mauricio.fajardo.hernandez@cern.ch

The HTCondor batch system is heavily used in the HEP community as the batch system for several WLCG resources. Moreover it is the backbone of the GlideInWMS, the main pilot system used by CMS. To prepare for LHC Run 2, we are probing the scalability limits of new versions and configurations of HTCondor with the goal of reaching at least 200,000 simultaneous running jobs in a single pool.

A sleeper pool of this size was achieved without a major impact in real jobs by using only 10,000 real slots distributed at several WLCG sites. We will report on how this was made and the impact it has in future scalability tests, not only of HTCondor but of any web faced service.

High configurability is one of the main capabilities of HTCondor. In addition to the test conditions and the testbed topology, we include the suggested configuration options used to obtain the scaling results.Finally, we will list the features present in newer versions of HTCondor that allow for sustained operations at scales well beyond what was previously possible.

7

Running and Testing T2 Grid Services with Puppet at GRIF-IRFU

Author: Sophie Catherine Ferry[1]

Co-authors: Frederic Schaer [2]; Pierrick Micout [3]

[1] *CEA/IRFU,Centre d'etude de Saclay Gif-sur-Yvette (FR)*

[2] *CEA*

[3] *Unknown*

Corresponding Author: sophie.catherine.ferry@cern.ch

GRIF is a distributed Tiers2 center, made of 6 different centers in the Paris region (France), and serving many VOs. The sub-sites are connected with 10Gbits/s private network and share tools for central management.
One of the sub-sites, GRIF-IRFU held and maintained in the CEA-Saclay center, moved a year ago, to a configuration management using Puppet.
Thanks to the versatility of Puppet/Foreman automation, the GRIF-IRFU site maintains usual grid services, with among them: a Cream-CE with a Torque+Maui (running a batch with more than 4000 jobs slots), a DPM storage of more than 2PB, a Nagios monitoring essentially based on check_mk, as well as centralized services for the French NGI, like the accounting, or the argus central banning system.
We report on the actual functionalities of Puppet and present the last tests and evolutions including a monitoring with Graphite, a HT-condor multi core batch accessed with an ARC-CE and a CEPH storage file system.

8

Modular and scalable RESTful API to sustain STAR collaboration's record keeping

Authors: Dmitry Arkhipkin[1]; Jerome LAURET[2]; Prashanth Shanmuganathan[3]

[1] *Brookhaven National Laboratory*

[2] *BROOKHAVEN NATIONAL LABORATORY*

[3] *Kent State University, USA*

Corresponding Author: sprashan@kent.edu

STAR collaboration's record system is a collection of heterogeneous and sparse information associated to each members and institutions. In its original incarnation, only flat information were stored revealing many restrictions such as the lack of historical change information, the inability to keep track of members leaving and re-joining or the ability to easily extend the saved information as new requirements appear.

In mid-2013, a new project was launched covering for an extensive set of revisited requirements. The requirements led us to a design based on a RESTful API, back-end storage engine relying on key/value pair data representation model coupled with a tiered architecture design. This design was motivated by the fact that unifying many STAR tools, relying on the same business logic and storage engine, was a key and central feature for the maintainability of records. This central service API would leave no ambiguities and provide easy service integration between STAR tools.

The new design stores the changes in records dynamically and allows tracking the changes chronologically. The storage engine is extensible as new field of information emerges (member specific or general) without affecting the presentation or the business logic layers. The new record system features a convenient administrative interface, fuzzy algorithms for data entry and search and provides basic statistics and graphs. Finally, this modular approach is supplemented with access control, allowing sensitive information and administrative operations away from public users.

In this contribution, we will review the requirements, present our design and its benefits as well as illustrate the power of the approach using practical examples for record keeping in a large collaboration like STAR.

9

Development and test of the DAQ system for a Micromegas prototype installed into the ATLAS experiment

Authors: Andre Zibell[1]; Michele Bianco[2]; Ourania Sidiropoulou[3]; Philipp Jonathan Loesel[4]; Sorin Martoiu[5]

[1] *Bayerische Julius Max. Universitaet Wuerzburg (DE)*

[2] *CERN*

[3] *Aristotle Univ. of Thessaloniki (GR)*

[4] *Ludwig-Maximilians-Univ. Muenchen (DE)*

[5] *IFIN-HH Bucharest (RO)*

Corresponding Author: andre.zibell@cern.ch

Ourania Sidiropoulou
on behalf of the ATLAS Muon Collaboration

A Micromegas (MM) quadruplet prototype with an active area of {0.5 m^2} that adopts the general design foreseen for the upgrade of the innermost forward muon tracking systems (Small Wheels) of the ATLAS detector in 2018-2019, has been built at CERN and is going to be tested in the ATLAS cavern environment during the LHC RUN-II period 2015-2017.

The integration of this prototype detector into the ATLAS data acquisition system using custom ATCA equipment is presented. An ATLAS compatible ReadOutDriver (ROD) based on the Scalable Readout System (SRS), the Scalable Readout Unit (SRU), will be used in order to transmit the data after generating valid event fragments to the high-level Read Out System (ROS). The SRU will be synchronized with the LHC bunch crossing clock (40.08 MHz) and will receive the Level-1 trigger signals from the Central Trigger Processor (CTP) through the TTCrx receiver ASIC. The configuration of the system will be driven directly from the ATLAS Run Control System. By using the ATLAS TDAQ Software, a dedicated Micromegas segment has been implemented, in order to include the detector inside the main ATLAS DAQ partition. A full set of tests, on the hardware and software aspects, is presented.

Track 1 Session / 10

Pyrame, a rapid-prototyping framework for online systems

Authors: Frederic Bruno Magniette[1]; Miguel Rubio-Roy[2]

[1] *Ecole Polytechnique (FR)*

[2] *CNRS*

Corresponding Author: frederic.bruno.magniette@cern.ch

High-energy physics experiments produce huge amounts of data that need to be processed and stored for further analysis and eventually treated in real time for triggering and monitoring purposes. In addition, more and more often these requirements are also being found on other fields such as on-line video processing, proteomics and astronomical facilities.

The complexity of such experiments usually involves long development times on which simple test-bench prototypes evolve over time and require increasingly performant software solutions for both data acquisition and control systems. Flexibility and wide application range are, therefore, important conditions in order to keep a single software solution that can evolve over the duration of the development period.

Existing solutions such as LabView provide proprietary solutions for control applications and small scale data acquisition, but fail on scaling to high-throughput experiments, add an important performance penalty and become difficult to maintain when the complexity of the software flow increases. Other lower-level solutions such as LabWindows/C, Matlab or TANGO allow to build more complex systems but still fail on providing an integrated and close to turn-key solution for the specific needs of prototypes evolving from bench-based tests to distributed environments with high data-throughput needs.

The present work reports on the software Pyrame, an open-source framework designed with high-energy physics applications in mind and providing a light-weight, distributed, stable, performant and easy-to-deploy solution. Pyrame is a data-acquisition chain, a data-exchange network protocol and a wide set of drivers allowing the control of hardware components. Data-acquisition throughput is on the order of 1.7 Gb/s for UDP/IP and Pyrame's protocol overhead is about 1.6 ms per command/response using the stock tools.

Pyrame provides basic blocks (modules) for hardware control or data acquisition. It contains an important number of modules to manage commonly used hardware: high and low voltage power supplies (Agilent, CAEN, Hameg, Keithley...), pattern generators (Agilent), digital oscilloscopes (LeCroy), motion controllers (Newport) and bus adapters (GPIB, USB, RS-232, Ethernet, UDP, TCP···). Generic modules give a unified access to machines from different generations and native or emulated functions depending on the possibilities of the machines.

Pyrame also provides service modules, including: a multimedia and multistream high-performance acquisition chain; a variable module allowing to share data between all the hardware modules; and a configuration management module allowing to save or load an image of the global configuration of the system at any time.

These blocks can be assembled together through Pyrame's software bus and protocol to quickly obtain complete systems for test benches. The framework is very flexible and provides a wide range of options, allowing the system to evolve as fast as the prototype.

In order to ease the development of extensions, Pyrame uses open standards: TCP and XML. They are implemented on most platforms and in most languages making the development of new modules fast and easy. The use of TCP/IP eases the distribution of code on multi-computer setups, but also its use on embedded platforms.

Multiple test-beams at DESY have validated the stability of the system. In particular, our tests have involved data acquisition sessions of hundreds of hours from up to 10 simultaneous Si-based particle detectors distributed over multiple computers. Very long calibration procedures involving hundreds of thousands of configuration operations have also been performed. On another optical bench, simultaneous control and data acquisition of oscilloscopes, motorized stages and power supplies confirmed the same level of stability for long acquisition sessions.

Track 5 Session / 11

Data Preservation at the Fermilab Tevatron

Authors: Bodhitha Jayatilaka[1]; Joe Boyd[1]; Ken Herner[1]; Willis Sakumoto[2]

Co-author: Robert Roser [1]

[1] *Fermilab*

[2] *University of Rochester (US)*

Corresponding Author: bo.jayatilaka@cern.ch

The Fermilab Tevatron collider's data-taking run ended in September 2011, yielding a dataset with rich scientific potential. The CDF and D0 experiments each have nearly 9 PB of collider and simulated data stored on tape. A large computing infrastructure consisting of tape storage, disk cache, and distributed grid computing for physics analysis with the Tevatron data is present at Fermilab.

The Fermilab Run II data preservation project intends to keep this analysis capability sustained through the year 2020 or beyond. To achieve this, we are implementing a system that utilizes virtualization, automated validation, and migration to new standards in both software and data storage technology as well as leveraging resources available from currently-running experiments at Fermilab. These efforts will provide useful lessons in ensuring long-term data access for numerous experiments throughout high-energy physics, and provide a roadmap for high-quality scientific output for years to come. We will present the status, benefits, and challenges of data preservation efforts within the CDF and D0 collaborations at Fermilab.

Track 1 Session / 12

LHCb topological trigger reoptimization

Author: Tatiana Likhomanenko[1]

Co-authors: Andrey Ustyuzhanin [2]; J Michael Williams [3]; Philip Ilten [3]

[1] *National Research Centre Kurchatov Institute (RU)*

[2] *ITEP Institute for Theoretical and Experimental Physics (RU)*

[3] *Massachusetts Inst. of Technology (US)*

Corresponding Author: tatiana.likhomanenko@cern.ch

The main b-physics trigger algorithm used by the LHCb experiment is the so-called topological trigger. The topological trigger selects vertices which are a) detached from the primary proton-proton collision and b) compatible with coming from the decay of a b-hadron. In the LHC Run 1, this trigger utilized a custom boosted decision tree algorithm, selected an almost 100% pure sample of b-hadrons with a typical efficiency of 60-70%, and its output was used in about 60% of LHCb papers. This talk presents studies carried out to optimize the topological trigger for LHC Run 2. In particular, we have carried out a detailed comparison of various machine learning classifier algorithms, e.g., AdaBoost, MatrixNet and neural networks. The topological trigger algorithm is designed to select all "interesting" decays of b-hadrons, but cannot be trained on every such decay. Studies have therefore been performed to determine how to optimize the performance of the classification algorithm on decays not used in the training. These include cascading, ensembling and blending techniques. Furthermore, novel boosting techniques have been implemented that will help reduce systematic uncertainties in Run 2 measurements. We demonstrate that the reoptimized topological trigger is expected to significantly improve on the Run 1 performance for a wide range of b-hadron decays.

13

Progress in Multi-Disciplinary Data Life Cycle Management

Authors: André Giesler[1]; Christopher Jung[2]; Fabian Rigoll[3]; Jörg Meyer[3]; Kilian Schwarz[4]; Marcus Hardt[5]; Martin Gasthuber[6]; Rainer Stotzka[3]

Co-author: Achim Streit [3]

[1] *FZ Jülich*

[2] *KIT - Karlsruhe Institute of Technology (DE)*

[3] *KIT*

[4] *GSI - Helmholtzzentrum fur Schwerionenforschung GmbH (DE)*

[5] *Kalrsruhe Institute of Technology*

[6] *Deutsches Elektronen-Synchrotron (DE)*

Corresponding Author: christopher.jung@kit.edu

Modern science is most often driven by data. Improvements in state-of-the-art technologies and methods in many scientific disciplines lead not only to increasing data rates, but also to the need to improve or even completely overhaul their data life cycle management.

Communities usually face two kinds of challenges: generic ones like federated authorization and authentication infrastructures and data preservation, and ones that are specific to their community and their respective data life cycle. In practice, the specific requirements often hinder the use of generic tools and methods.

The German Helmholtz Association project "Large-Scale Data Management and Analysis" (LSDMA) addresses both challenges: its five Data Life Cycle Labs (DLCLs) closely collaborate with communities in joint research and development to optimize the communities'data life cycle management, while its Data Services Integration Team (DSIT) provides generic data tools and services.

We present most recent developments and results from the DLCLs covering communities ranging from heavy ion physics and photon science to high-throughput microscopy, and from DSIT.

Track 4 Session / 14

Intrusion Detection in Grid computing by Intelligent Analysis of Jobs Behavior –The LHC ALICE Case

Author: Andres Gomez Ramirez[1]

Co-authors: Camilo Ernesto Lara Martinez [1]; Udo Wolfgang Kebschull [1]

[1] *Johann-Wolfgang-Goethe Univ. (DE)*

Corresponding Author: andres.gomez@cern.ch

Grid infrastructures allow users flexible on-demand usage of computing resources using an Internet connection. A remarkable example of a Grid in High Energy Physics (HEP) research is used by the ALICE experiment at European Organization for Nuclear Research CERN. Physicists can submit jobs used to process the huge amount of particle collision data produced by the Large Hadron Collider (LHC) at CERN. Grids allow the submission of user developed jobs (code and data). They also have interfaces to Internet, storage systems, experiment infrastructure and other networks. This creates an important security challenge. Even when Grid system administrators perform a careful security assessment of sites, worker nodes, storage elements and central services, an attacker could still take advantage of unknown vulnerabilities (zero day). This attacker could enter and escalate her access privileges to misuse the computational resources for unauthorized or even criminal purposes. She could even manipulate the data of the experiments.

Accordingly, Grid systems require an automatic tool to monitor and analyze the behavior of user jobs. This tool should analyze data generated by jobs such as log entries, traces, system calls, to determine if they run in a desired behavior or are performing any kind of attack on the system. The tool should react to the attack by sending alerts, logging information about relevant events and performing automatic defensive actions (for example stopping a suspicious process). This piece of software could
be classified as Grid Intrusion Detection Systems (Grid-IDS). Traditional IDS allow detection of attacks by fixed if-then rules based on signatures. It compares the input data with known predefined conditions from previous events. This strategy fails when a new type of intrusion is used, even with a slightly difference. Artificial intelligence algorithms have been suggested as a method to improve Intrusion Detection Systems. By the usage of a Machine Learning approach it is possible to train the IDS on generalizing among attacks even when they are completely new. Intelligent IDS can also analyze the huge amount of data generated in Grid logs and process traces to determine a misbehaving scenario (data mining). This Grid IDS has to be adapted to highly distributed scenarios, when collaboration among geographically separate sites is necessary and reliability on central services is not always an option.

Currently there is no framework that allows us to fulfill all the above requirements. We will design and build such framework. This framework should allow the monitoring and analysis of grid job behavior to detect attack attempts, even if new techniques or zero day vulnerabilities are utilized. This framework should also perform required countermeasures for its protection. In a first step, we plan to analyze the
behavior of the usual job execution in the ALICE experiment Grid. We will determine the most important metrics to characterize a "bad"behavior (an attack). Later we will collect data from the Grid logs using these metrics and will use this data to train a machine learning algorithm. The algorithm will allow us classification of jobs as in desired or undesired state depending on the data produced in their execution. We plan to implement the proposed framework as a software prototype that will be tested as a component of the ALICE Grid middelware.

Keywords – grid computing, distributed computing, distributed System security, artificial intelligence,
data mining, Intrusion Detection Systems.

16

A Case Study in Preserving a High Energy Physics Application

Authors: Douglas Thain[1]; Haiyan Meng[2]

Co-authors: Anna Elizabeth Woodard [3]; Matthias Wolf [3]; Michael Hildreth [1]; Peter Ivie [1]

[1] *University of Notre Dame*

[2] *U*

[3] *University of Notre Dame (US)*

Corresponding Authors: dthain@nd.edu, hmeng@nd.edu, mhildret@nd.edu

The reproducibility of scientific results increasingly depends upon the preservation of computational artifacts. Although preserving a computation to be used later sounds trivial, it is surprisingly difficult due to the complexity of existing software and systems. Implicit dependencies, networked resources, and shifting compatibility all conspire to break applications that appear to work well. Tools are needed which can automatically identify both local and remote dependencies, so that they can be captured and preserved.

To investigate these issues, we present a case study of preserving a CMS application using Parrot. We analyze the application and attempt several methods at extracting its dependencies for the purposes of preservation. We demonstrate a fine-grained dependency management toolkit which can observe both the local filesystem and remote network dependencies, using the system call tracing capabilities of Parrot. We observe that even a simple application depends upon 22,068 files and directories totalling 21 GB of data and software drawn from 8 different sources including cvmfs, hdfs, afs, git, http, cvs, PanFS and local root filesystem.

Once the dependencies are observed, a portable execution package can be generated. This package is not tied to any particular technology and can be re-run using Parrot, Docker, a chroot Jail, or as a virtual machine image, depending on the technology available at the execution site. We will report on the performance and completeness of re-execution using both public and private clouds and offer some guidance for future work in application preservation.

17

Performance evaluation of the ATLAS IBL Calibration

Author: Moritz Kretz[1]

[1] *Ruprecht-Karls-Universitaet Heidelberg (DE)*

Corresponding Author: moritz.kretz@cern.ch

With the installation of the Insertable B-Layer in 2014 the Pixel Detector of the ATLAS experiment has been extended by about 12 million pixels. Scanning and tuning procedures have been implemented by employing newly designed read-out hardware which is now able to support the full detector bandwidth even for calibration. The hardware is supported by an embedded software stack running on the read-out boards. Compute-intensive operations that are necessary during the calibration process are performed on a small cluster of commodity servers, which are included in the ATLAS TDAQ framework. In order to exploit the potential of this heterogenious data processing architecture to significantly speed-up detector calibration, timing and performance of the various software components is critical.

We present a timing model including the hardware and software components and procedures that are involved in the calibration process. By instrumenting these components at critical locations we are then able to accurately evaluate the timing of a calibration scan, analyze the performance, and eventually locate and eliminate bottlenecks in the chain.

18

Automation of Large-scale Computer Cluster Monitoring Information Analysis

Author: Erekle Magradze[1]

Co-authors: Arnulf Quadt [1]; Gen Kawamura [1]; Haykuhi Musheghyan [1]; Jordi Nadal Serrano [1]

[1] *Georg-August-Universitaet Goettingen (DE)*

Corresponding Author: erekle.magradze@cern.ch

High-throughput computing platforms consist of complex infrastructure and provide a number of services apt to failures. To mitigate the impact of failures on the quality of the provided services, a constant monitoring and in time reaction is required, which is impossible without automation of the system administration
processes. This paper introduces a way of automation of the process of monitoring information analysis to provide long and short term predictions of the service response time (SRT) of the mass storage and the batch systems and to identify the status of a service at a given time. The approach for the SRT predictions is based on Adaptive Neuro Fuzzy Inference System (ANFIS) while for a proper service status
identification the K-means clustering algorithm was employed. An evaluation of the approaches is performed on real monitoring data from the WLCG Tier 2 center GoeGrid. Ten fold cross validation results demonstrate high efficiency of both approaches in comparison to known methods.

20

Extending DIRAC File Management with Erasure-Coding for efficient storage.

Author: Samuel Cadellin Skipsey[None]

Co-authors: Dan Protopopescu [1]; David Britton [1]; David Crooks [1]; Gareth Douglas Roy [1]; Janusz Martyniak [2]; Paulin Todev [3]

[1] *University of Glasgow (GB)*

[2] *Imperial College London*

[3] *University of Glasgow*

Corresponding Author: samuel.cadellin.skipsey@cern.ch

The state of the art in Grid style data management is to achieve increased resilience of data via multiple complete replicas of data files across multiple storage endpoints. While this is effective, it is not the most space-efficient approach to resilience, especially when the reliability of individual storage endpoints is sufficiently high that only a few will be inactive at any point in time.

We extended the Dirac File Catalogue and file management interface to allow the placement of *erasure-coded* files: each file distributed as N identically-sized chunks of data striped across a vector of storage endpoints, encoded such that any M chunks can be lost and the original file can be reconstructed.

The tools developed are transparent to the user, and, as well as allowing up and downloading of data to Grid storage, also provide the possibility of parallelizing access across all of the distributed chunks at once, improving data transfer and IO performance.
We expect this approach to be of most interest to smaller VOs, who have tighter bounds on the storage available to them, but larger VOs may be interested as their total data increases during Run 2.

With this in mind, we tested the applicability of our tools to the NA62 analysis model, which already expects data to be distributed across multiple Tier-2 sites.

We provide an analysis of the costs and benefits of the approach, along with future development and implementation plans.

Track 1 Session / 21

Data Acquisition for the New Muon g-2 Experiment at Fermilab

Author: Wesley Gohn[1]

[1] *U*

Corresponding Author: gohn@pa.uky.edu

A new measurement of the anomalous magnetic moment of the muon, $a_\mu \equiv (g-2)/2$, will be performed at the Fermi National Accelerator Laboratory. The most recent measurement, performed at Brookhaven National Laboratory and completed in 2001, shows a 3.3-3.6 standard deviation discrepancy with the standard model value of g-2. The new measurement will accumulate 20 times those statistics, measuring g-2 to 0.14 ppm and improving the uncertainty by a factor of 4 over that of the previous measurement.

The data acquisition system for this experiment must have the ability to create deadtime-free records from 700 μs muon spills at a raw data rate 18 GB per second. Data will be collected using 1296 channels of μTCA-based 800 MSPS, 12 bit waveform digitizers and processed in a layered array of networked commodity processors with 24 GPUs working in parallel to perform a fast recording of the muon decays during the spill. The system will be controlled using the MIDAS data acquisition software package. The described data acquisition system is currently being constructed, and will be fully operational before the start of the experiment in 2016.

Track 3 Session / 22

Enabling Object Storage via shims for Grid Middleware

Author: Samuel Cadellin Skipsey[None]

Co-authors: Alastair Dewhurst [1]; David Britton [2]; David Crooks [2]; Gareth Douglas Roy [2]; Wahid Bhimji [3]

[1] *STFC - Rutherford Appleton Lab. (GB)*

[2] *University of Glasgow (GB)*

[3] *University of Edinburgh (GB)*

Corresponding Author: samuel.cadellin.skipsey@cern.ch

The *Object Store* model has quickly become the de-facto basis of most commercially successful mass storage infrastructure, backing so-called "Cloud" storage such as Amazon S3, but also underlying the implementation of most parallel distributed storage systems.
Many of the assumptions in object store design are similar, but not identical, to concepts in the design of Grid Storage Elements, although the requirement for "POSIX-like" filesystem structures on top of SEs makes the disjunction seem larger.
As modern object stores provide many features that most Grid SEs do not (block level striping, parallel access, automatic file repair, etc), it is of interest to see how easily we can provide interfaces to typical object stores via plugins and shims for Grid tools, and how well experiments can adapt their data models to them.

We present the experience of developing plugins for the currently-popular *ceph* parallel distributed filesystem, for GFAL2 and also other Grid file access and data management tools.

Additionally, we present evaluation of, and first-deployment experiences with, (for example) Xrootd-Ceph interfaces for direct object-store access, for ATLAS at RAL.

Track 3 Session / 23

Current Status of the Ceph Based Storage Systems at the RACF

Authors: Alexandr Zaytsev[1]; Hironori Ito[1]

Co-authors: Christopher Hollowell [1]; Tejas Rao [1]; Tony Wong [1]

[1] *Brookhaven National Laboratory (US)*

Corresponding Author: hito@bnl.gov

Ceph based storage solutions are becoming increasingly popular within the HEP/NP community over the last few years. With the current status of the Ceph project, both its object storage and block storage layers are production ready on a large scale, and even the Ceph file system (CephFS) storage layer is rapidly getting to that state as well. This contribution contains a thorough review of various functionality, performance, and stability tests performed with all three (object storage, block storage and file system) levels of Ceph by using the resources of the RHIC and ATLAS Computing Facility (RACF) at Brookhaven National Laboratory (BNL) in 2012-2014 on various hardware platforms (including HP Moonshot) and with different networking solutions based on 10/40 Gbps Ethernet and IPoIB/4X FDR Infiniband interconnect technology. We also report the current status of a 1 PB scale (by usable capacity, taking into account the internal data replication factor 3x) Ceph based object storage system provided with Amazon S3 complaint RADOS Gateway interfaces that was built for the RACF in 2013, as well as the first results on its performance and the experience obtained while operating it in production over the last 8 months.

Track 3 Session / 24

Operational Experience Running Hadoop XRootD Fallback

Author: Jeffrey Michael Dost[1]

Co-authors: Alja Mrak Tadel [1]; Frank Wuerthwein [1]; Matevz Tadel [1]

[1] *Univ. of California San Diego (US)*

Corresponding Author: jdost@ucsd.edu

In April of 2014, the UCSD T2 Center deployed hdfs-xrootd-fallback, a UCSD-developed software system that interfaces Hadoop with XRootD to increase reliability of the Hadoop file system. The hdfs-xrootd-fallback system allows a site to depend less on local file replication and more on global replication provided by the XRootD federation to ensure data redundancy. Deploying the software has allowed us to reduce Hadoop replication on a significant subset of files in our cluster, freeing hundreds of terabytes in our local storage, and to recover HDFS blocks lost due to storage degradation. An overview of the architecture of the hdfs-xrootd-fallback system will be presented, as well as details of our experience operating the service over the past year.

Track 2 Session / 25

HistFitter: a flexible framework for statistical data analysis

Authors: Aleksej Koutsman[1]; David Cote[2]; Geert Jan Besjes[3]; Jeanette Miriam Lorenz[4]; Max Baak[5]

[1] *TRIUMF (CA)*

[2] *University of Texas at Arlington (US)*

[3] *Radboud University Nijmegen (NL)*

[4] *Ludwig-Maximilians-Univ. Muenchen (DE)*

[5] *CERN*

Corresponding Author: geert-jan.besjes@cern.ch

We present a software framework for statistical data analysis, called *HistFitter*, that has been used extensively in the ATLAS Collaboration to analyze data of proton-proton collisions produced by the Large Hadron Collider at CERN. Most notably, HistFitter has become a de-facto standard in searches for supersymmetric particles since 2012, with some usage for Exotic and Higgs boson physics. HistFitter coherently combines several statistics tools in a programmable and flexible framework that is capable of bookkeeping hundreds of data models under study using thousands of generated input histograms. The key innovations of HistFitter are to weave the concepts of control, validation and signal regions into its very fabric, and to treat them with rigorous methods, while providing multiple tools to visualize and interpret the results through a simple configuration interface, as will become clear throughout this presentation.

HistFitter interfaces with the statistics tools
HistFactory and RooStats to construct parametric models and to perform statistical
tests of the data, and extends these tools in four key areas:

1. Programmable framework: HistFitter puts tools from several sources together in a coherent and programmable
framework, capable of performing a complete statistical analysis of pre-formatted input data samples.

2. Bookkeeping: HistFitter can perform statistical tests and scan over parameter values of hundreds of
signal hypotheses in an organized way from a single user-defined configuration file.

3. Analysis strategy: HistFitter uses built-in the concepts of control, signal and validation regions which are used
to constrain, extrapolate and validate data-model predictions across analysis regions.
HistFitter also introduces a rigorous treatment of validation regions that is new in high-energy physics.

4. Presentation and interpretation: The HistFitter framework keeps track of data models before and after
fits to the data, and includes a collection of methods to determine the statistical
significance of all tested hypotheses and to produce tables and plots expressing these
results with publication-quality style.

26

Scaling Agile Infrastructure to people

Author: Ben Jones[1]

Co-authors: Gavin McCance [1]; Nacho Barrientos Arias [1]; Steve Traylen [1]

[1] *CERN*

Corresponding Author: ben.dylan.jones@cern.ch

When CERN migrated its infrastructure away from home-grown fabric management tools to emerging industry-standard open-source solutions, the immediate technical challenges and motivation were clear. The move to a multi-site Cloud Computing model meant that the toolchains that were growing around this ecosystem would be a good choice, the challenge was to leverage them.

The use of open-source tools brings challenges other than merely how to deploy them. Home-grown software, for all the deficiencies identified at the outset of the project, has the benefit of growing with the organization. This presentation will examine what challenges there were in adapting open-source tools to the needs of the organization, particularly in the areas of multi-group development and security.

Additionally, the increase in scale of the plant required changes to how Change Management was organized and managed. Continuous Integration techniques are used in order to manage the rate of change across multiple groups, and the tools and workflow for this will be examined.

27

Optimizing the transport layer of the ALFA framework for the Intel Xeon Phi co-processor

Author: Aram Santogidis[1]

[1] *CERN*

Corresponding Author: aram.santogidis@cern.ch

ALFA is the common framework of the next generation software for ALICE and FAIR high energy physics experiments. It supports both offline and online processing which includes ALICE DAQ/HLT/Offline and the FairRoot project. The framework is designed based on a data-flow model with message-oriented middleware (MOM) serving as a transport layer. By using multiple data-flows concurrently it facilitates parallel processing which maps naturally to emerging multi-core and many-core computing architectures.

With the introduction of the Intel Xeon Phi co-processor in the industry it is interesting to investigate whether it can be used by ALFA to increase the processing efficiency. The co-processor can be used in three main computing modes. These are offload in which portions of code are accelerated on the device, native where the full program is executed on the device and symmetric where complete tasks are executed on the device and the host processor. For acceleration via offloading there are many competitive platforms such as GPUs and FPGAs. Although Xeon Phi can be used as an accelerator, it is particularly interesting to investigate the possibility of utilizing the co-processor in the *symmetric mode* of operation. Since it is x86_64 compatible it is possible to port complete *task processes* to the device and take advantage of the manycore architecture. It is also worth noting that the next generation of the Xeon Phi codenamed Knights Landing (KNL) will be manufactured in a socket variant as well. The research of using the co-processor as an independent node as opposed to just an offload accelerator can serve as a preliminary study for the future KNL servers.

The software components ported to the device will be connected with the rest of the system via the transport layer therefore there is strong motivation to optimize it for Xeon Phi. The metrics against which it is optimized are **throughput**, **latency** and **energy** consumption, with throughput being the primary target. The two core MOM technologies of choice for ALFA are ZeroMQ and NanoMSG. The out-of-the box versions of these libraries use primarily TCP as the transport protocol which is known to provide limited performance on Xeon Phi in terms of data transfer throughput. In this effort these libraries are extended with support for SCIF, the Xeon Phi native transport protocol over PCIe and additionally with the Co-processor Communication Link (CCL), an RDMA technology used for efficient internode communication. By introducing these extensions we will demonstrate improvements in the data transfer performance by collecting performance monitoring results both in isolation with micro-benchmarks and integrated in the ALFA framework with the respective transport layer benchmark. A successful completion of the optimizations of the transport layer will improve the performance of NanoMSG and ZeroMQ on Xeon Phi. This achievement will potentially make this architecture a viable choice for certain use cases for ALFA which will further enrich its heterogeneous computing capabilities.

Track 4 Session / 28

The Careful Puppet Master: Reducing risk and fortifying acceptance testing with Jenkins CI

Authors: Gabriel Richman[1]; Jason Alexander Smith[2]

Co-authors: James Pryor [2]; John Steven De Stefano Jr [2]; Tejas Rao [2]; Tony Wong [2]; William Strecker-Kellogg [2]

[1] *Oberlin College*

[2] *Brookhaven National Laboratory*

Corresponding Author: smithj4@bnl.gov

Using centralized configuration management, including automation tools such as Puppet, can greatly increase provisioning speed and efficiency when configuring new systems or making changes to existing systems, reduce duplication of work, and improve automated processes. However, centralized management also brings with it a level of inherent risk: a single change in just one file can quickly be pushed out to thousands of computers and, if that change is not properly and thoroughly tested and contains an error, could result in catastrophic damage to many services, potentially bringing an entire computer facility offline.

Change management procedures can – and should – be formalized in order to prevent such accidents. However, like the configuration management process itself, if such procedures are not automated, they can be difficult to enforce strictly. Therefore, to reduce the risk of merging potentially harmful changes into our production Puppet environment, we have created an automated testing system, which includes the Jenkins CI tool, to manage our Puppet testing process. This system includes the proposed changes and runs Puppet on a pool of dozens of RHEV VMs that replicate most of our important production services. This paper describes our automated test system and how it hooks into our production approval process for automatic acceptance testing. All pending changes that have been pushed to production must pass this validation process before they can be approved and merged into production.

Track 1 Session / 29

The ATLAS Data Flow system for the second LHC run

Author: Reiner Hauser[1]

[1] *Michigan State University (US)*

Corresponding Author: reiner.hauser@cern.ch

After its first shutdown, LHC will provide pp collisions with increased luminosity and energy. In the ATLAS experiment the Trigger and Data Acquisition (TDAQ) system has been upgraded to deal with the increased event rates. The Data Flow (DF) element of the TDAQ is a distributed hardware and software system responsible for buffering and transporting event data from the Readout system to the High Level Trigger (HLT) and to the event storage. The DF has been reshaped in order to profit from the technological progress and to maximize the flexibility and efficiency of the data selection process.

The updated DF is radically different from the previous implementation both in terms of architecture and expected performance. The pre-existing two level software filtering, known as L2 and the Event Filter, and the Event Building are now merged into a single process, performing incremental data collection and analysis. This design has many advantages, among which are: the radical simplification of the architecture, the flexible and automatically balanced distribution of the computing resources, 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.

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 Data Flow 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.

We will discuss the design choices, the strategies employed to minimize the data-collection latency, the results of scaling tests done during the commissioning phase and the operational performance after the first months of data taking.

Track 1 Session / 30

The ATLAS Trigger System: Ready for Run-2

Author: Yu Higuchi[1]

[1] *High Energy Accelerator Research Organization (JP)*

Corresponding Author: yu.nakahama@cern.ch

The ATLAS trigger has been used very successfully for the online event selection during the first run of the LHC between 2009-2013 at a centre-of-mass energy between 900 GeV and 8 TeV. The trigger system consists of a hardware Level-1 (L1) and a software based high-level trigger (HLT) that reduces the event rate from the design bunch-crossing rate of 40 MHz to an average recording rate of a few hundred Hz. During the next data-taking period starting in early 2015 (Run-2) the LHC will operate at a centre-of-mass energy of about 13 TeV resulting in roughly five times higher trigger rates.

We will review the upgrades to the ATLAS Trigger system that have been implemented during the shutdown and that will allow us to cope with these increased trigger rates while maintaining or even improving our efficiency to select relevant physics processes. This includes changes to the L1 calorimeter trigger, the introduction of a new L1 topological trigger module, improvements in the L1 muon system and the merging of the previously two-level HLT system into a single event filter farm. At hand of a few examples, we will show the impressive performance improvements in the HLT trigger algorithms used to identify leptons, hadrons and global event quantities like missing transverse energy. And finally, we will summarize the commissioning status of the Trigger system in view of the imminent restart of data-taking.

Track 1 Session / 31

Operation of the upgraded ATLAS Level-1 Central Trigger System

Author: Julian Glatzer[1]

[1] *CERN*

Corresponding Author: julian.glatzer@cern.ch

The ATLAS Level-1 Central Trigger (L1CT) system is a central part of ATLAS data-taking and is configured, controlled, and monitored by a software framework with emphasis on reliability and flexibility. The hardware has undergone a major upgrade for Run 2 of the LHC, in order to cope with the expected increase of instantaneous luminosity of a factor of 2 with respect to Run 1. It offers more flexibility in the trigger decisions due to the double amount of trigger inputs and usable trigger channels. It also provides an interface to the new topological trigger system. Operationally - particularly useful for commissioning, calibration and test runs - it allows concurrent, independent triggering of up to 3 different sub-detector combinations.

In this contribution, we give an overview of the fully operational software framework of the L1CT system with particular emphasis on the configuration, controls and monitoring aspects. The software framework allows the L1CT system to be configured consistently with the ATLAS experiment and the LHC machine, upstream and downstream trigger processors, and the data acquisition. Trigger and dead-time rates are monitored coherently at all stages of processing and are logged by the online computing system for physics analysis, data quality assurance and operation debugging. In addition, the synchronisation of trigger inputs is watched based on bunch-by-bunch trigger information. Several software tools allow to efficiently display the relevant information in the control room in a way useful for shifters and experts. The design of the framework aims at reliability, flexiblity, and robustness of the system and takes into account the operational experience gained during Run 1. We present the overall performance during cosmic-ray data taking with the full ATLAS detector and the experience with first beams in 2015.

32

ATLAS Jet Trigger Performance in LHC Run I and Initial Run II Updates

Author: Shima Shimizu[1]

[1] *Kobe University (JP)*

Corresponding Author: shima.shimizu@cern.ch

The immense rate of proton-proton collisions at the Large Hadron Collider (LHC) must be reduced from the nominal bunch-crossing rate of 40 MHz to approximately 1 kHz before the data can be written on disk offline. The ATLAS Trigger System performs real-time selection of these events in order to achieve this reduction. Dedicated selection of events containing jets is uniquely challenging at a hadron collider where nearly every event contains significant hadronic energy.

Following the very successful first LHC run from 2010 to 2012, the ATLAS Trigger was much improved, including a new hardware topological module and a restructured High Level Trigger system, merging two previous software-based processing levels. This allowed the optimization of resources and a much greater re-use of the precise but costly offline software base. After summarising the overall performance of the jet trigger during the first LHC run, the software design choices and use of the topological module will be reviewed. The expected performance of jet trigger for the second LHC run, to start

in 2015, will be described together with the available commissioning measurements from early data taking.

33

The ATLAS Trigger Core Configuration and Execution System in Light of the ATLAS Upgrade for LHC Run 2

Author: Lukas Alexander Heinrich[1]

[1] *New York University (US)*

Corresponding Author: lukas.heinrich@cern.ch

During the 2013/14 shutdown of the Large Hadron Collider (LHC) the ATLAS first level trigger (L1T) and the data acquisition system (DAQ) were substantially upgraded to cope with the increase in luminosity and collision multiplicity, expected to be delivered by the LHC in 2015.

To name a few, the L1T was extended on the calorimeter side (L1Calo) to better cope with pile-up and apply better-tuned isolation criteria on electron, photon, and jet candidates. The central trigger (CT) was widened to analyze twice as many inputs, provide more trigger lines, and serve multiple sub-detectors in parallel during calibration periods. A new FPGA-based trigger, capable of analyzing event topologies at 40 MHz, was added to provide further input to forming the level 1 trigger decision (L1Topo). On the DAQ side the dataflow was completely remodeled, merging the two previously existing stages of the software-based high level trigger into one.

Partially because of these changes, partially because of the new trigger paradigm to have more full event analysis, the high level trigger (HLT) execution framework and the trigger configuration system had to be upgraded, tools and data content had to be adapted to the new ATLAS analysis model. In this paper we describe this work:

The algorithm execution framework was changed to seamlessly work within the merged HLT, the data access providers were adapted to the new dataflow. The event building, at which point all data are retrieved from the readout system, can now dynamically change with progressing event feature extraction, allowing a more flexible adjustment to dataflow constraints. The cost monitoring framework which analyzes data access and CPU consumption, even prior to data taking, was extended to work within the merged system, several other improvements followed.

The HLT execution was moved to a memory-saving multi-process application, in which many event processors are forked after the system configuration. They thus share common data such as geometry and conditions information, leading to a dramatic reduction in the overall memory consumption. Compared to Run 1 many more event processors can run on each machine.

Upon request from the ATLAS physics groups a new kind of data stream was implement, in which only a small subset of the reconstructed trigger objects and no detector data is stored. Since the trigger reconstruction in Run 2 almost compares to the offline in resolution, these data make search analyses that require high statistics feasible.

As a consequence of these changes, the new ATLAS data model, and the new dual environment analysis approach, the tools that are provided for trigger aware analysis had to be completely restructured. In particular the reduction and specialization of data content in derived data sets was posing a challenge for the trigger, a new trigger data slimming was invented.

The database driven trigger configuration system, which describes the physics implemented at L1 and HLT, needs to reflect all changes in the L1 and HLT system. It now incorporates the configuration for the new L1Topo trigger, has extended the configuration capabilities of the L1Calo and CT and the describes are merged HLT. A new system for automatically adjust the trigger prescale factors to the dropping luminosity during a run was devised and implemented.

We also present measurements of the trigger execution on first data with the new ATLAS trigger algorithms and selection.

Track 1 Session / 34

Upgrade of the ATLAS Level-1 Trigger with event topology information

Author: Eduard Ebron Simioni[1]

Co-authors: Adam Kaluza [1]; Alexander Vogel [1]; Andreas Dominik Reiss [1]; Bruno Bauss [1]; Christian Kahra [1]; Jan Schaffer [1]; Katharina Bianca Jakobi [1]; Manuel Sebastian Simon [1]; Markus Zinser [1]; Reinold Degele [1]; Sebastian Artz [1]; Stefan Tapprogge [1]; Uli Schaefer [1]; Volker Buescher [1]

[1] *Johannes-Gutenberg-Universitaet Mainz (DE)*

Corresponding Author: eduard.simioni@cern.ch

The Large Hadron Collider (LHC) in 2015 will collide proton beams with increased luminosity from 10^{34} up to 3×10^{34} cm^{-2} s^{-1}. ATLAS is an LHC experiment designed to measure decay properties of highly energetic particles produced in these proton-collisions. The high luminosity places stringent physical and operational requirements on the ATLAS Trigger in order to reduce the 40 MHz collision rate to an event storage rate of 1 kHz, thereby retaining events with valuable physics content. The hardware-based first ATLAS trigger level (Level-1) has an output rate of 100 kHz and decision latency of less than 2.5 µs. It is composed of the Calorimeter Trigger (L1Calo), the Muon Trigger (L1Muon) and the Central Trigger Processor. In 2014, there will be a new trigger system has been added: the Topological Processor System (L1Topo system).

The L1Topo system consists of a single AdvancedTCA shelf equipped with three L1Topo processor blades. It processes detailed information from L1Calo and L1Muon in individual state-of-the-art FPGA processors to derive desitions based on the topology of each collision event. Such topologies are the angles between jets and/or leptons or global event variables based on lists of pre-selected/-sorted objects. The system is designed to receive and process up to 6 Tb/s of real time data. The talk is about the relevant upgrades of the Level-1 trigger with focus on the topological processor design and commissioning.

35

Upgrade of the ATLAS Control and Configuration Software for Run 2

Author: Giuseppe Avolio[1]

Co-authors: Alejandro Santos [2]; Alina Corso Radu [3]; Andrei Kazarov [4]; Andrew James Lankford [3]; Diana Scannicchio [3]; Florian Laurent [5]; Gabriel Anders [1]; Giovanna Lehmann Miotto [1]; Gokhan Unel [3]; Igor Aleksandrov [6]; Igor Soloviev [3]; Jose Seixas [7]; Julio De Castro Vargas Fernandes [7]; Luca Magnoni [1]; Lykourgos Papaevgeniou [8]; Maria Teresa Dova [2]; Mihai Caprini [9]; Nikolai Klopov [10]; Serguei Oleshko [4]; Vladislav Kotov [11]; Yoshiji Yasu [12]; Yury Ryabov [4]

[1] *CERN*

[2] *Universidad Nacional de La Plata (AR)*

[3] *University of California Irvine (US)*

[4] *B.P. Konstantinov Petersburg Nuclear Physics Institute - PNPI (*

[5] *Ecole Polytechnique Federale de Lausanne (CH)*

[6] *Joint Institute for Nuclear Research (JINR)*

[7] *Univ. Federal do Rio de Janeiro (BR)*

[8] *University of Athens (GR)*

[9] *IFIN-HH Bucharest (RO)*

[10] *Petersburg Nuclear Physics Institute (PNPI)*

[11] *Joint Inst. for Nuclear Research (RU)*

[12] *High Energy Accelerator Research Organization (JP)*

Corresponding Author: giuseppe.avolio@cern.ch

The ATLAS data acquisition (DAQ) system is controlled and configured via a software infrastructure that takes care of coherently orchestrating the data taking. While the overall architecture, established at the end of the 90's, has proven to be solid and flexible, many software components have undergone a complete redesign or re-implementation in 2013/2014 in order to fold-in the additional requirements that appeared in the course of LHC's Run 1, to profit from new technologies and to re-factorise and cleanup software.
This paper describes the approach that was taken to plan, organise and carry out this software upgrade project. It highlights the main technical choices that have guided the overall work, describes the major achievements and outlines how the control and configuration software may be further improved or re-shaped in the future.

Track 4 Session / 36

A Validation System for the Complex Event Processing Directives of the ATLAS Shifter Assistant Tool

Author: Giuseppe Avolio[1]

Co-authors: Alejandro Santos [2]; Andrei Kazarov [3]; Gabriel Anders [1]; Giovanna Lehmann Miotto [1]; Igor Soloviev [4]

[1] *CERN*

[2] *Universidad Nacional de La Plata (AR)*

[3] *B.P. Konstantinov Petersburg Nuclear Physics Institute - PNPI (*

[4] *University of California Irvine (US)*

Corresponding Author: giuseppe.avolio@cern.ch

Complex Event Processing (CEP) is a methodology that combines data from different sources in order to identify events or patterns that need particular attention. It has gained a lot of momentum in the computing world in the past few years and is used in ATLAS to continuously monitor the behaviour

of the data acquisition system, to trigger corrective actions and to guide the experiment's operators. This technology is very powerful, if experts regularly insert and update their knowledge about the system's behaviour into the CEP engine. Nevertheless, writing or modifying CEP directives is not trivial since the used programming paradigm is quite different with respect to what developers are normally familiar with. In order to help experts verify that the directives work as expected, we have thus developed a complete testing and validation environment.
This system consists of three main parts: the first is the persistent storage of all relevant data streams that are produced during data taking, the second is a playback tool that allows to re-inject data of specific data taking sessions from the past into the CEP engine and the third is a reporting tool that shows the output that the directives loaded into the engine would have produced in the live system. In this paper we describe the design, implementation and performance of this validation system, highlight its strengths and short-comings and indicate how such a system could be re-used in similar projects.

Track 1 Session / 37

ATLAS TDAQ System Administration: evolution and re-design

Author: Christopher Jon Lee[1]

Co-authors: Aleksandr Korol [2]; Alexander Bogdanchikov [2]; Artem Voronkov [2]; Cristian Contescu [3]; Daniel Fazio [4]; Diana Scannicchio [5]; Franco Brasolin [6]; Matthew Shaun Twomey [7]; Sergei Dubrov [2]; Sergio Ballestrero [1]

[1] *University of Johannesburg (ZA)*

[2] *Budker Institute of Nuclear Physics (RU)*

[3] *Polytechnic University of Bucharest (RO)*

[4] *CERN*

[5] *University of California Irvine (US)*

[6] *Universita e INFN (IT)*

[7] *University of Washington (US)*

Corresponding Author: chris.lee@cern.ch

The ATLAS Trigger and Data Acquisition (TDAQ) system is responsible for the online processing of live data, streaming from the ATLAS experiment at the Large Hadron Collider (LHC) at CERN. The online farm is composed of ~3000 servers, processing the data readout from ~100 million detector channels through multiple trigger levels.
During the two years of the first Long Shutdown (LS1) there has been a tremendous amount of work done by the ATLAS TDAQ System Administrators, implementing numerous new software applications, upgrading the OS and the hardware, changing some design philosophies and exploiting the High Level Trigger farm with different purposes.
During the data taking only critical security updates are applied and broken hardware is replaced to ensure a stable operational environment. The LS1 provided an excellent opportunity to look into new technologies and applications that would help to improve and streamline the daily tasks of not only the System Administrators, but also of the scientists who will be working during the upcoming data taking period (Run-II).
The OS version has been upgraded to SLC6; for the largest part of the farm, which is composed by netbooted nodes, this required a completely new design of the netbooting system. In parallel, the migration to Puppet of the Configuration Management systems has been completed for both netbooted and localbooted hosts; the Post-Boot Scripts system and Quattor have been consequently dismissed. Various new ATCA-based readout systems, with specific network requirements, have also been integrated into the overall system.
Virtual Machine (VM) usage has been investigated and tested and many of our core servers are now running on VMs. This provides us with the functionality of rapidly replacing them in case of failures and increasing the number of servers when needed.
Virtualization has also been used to adapt the HLT farm as a batch system, which has been used for running Monte Carlo production jobs that are mostly CPU and not I/O bound.
In Run-II this feature could be exploited during the downtimes of the LHC.
A new Satellite Control Room (SCR) has also been commissioned and in the ATLAS Control Room

(ACR) the PC-over-IP network connections have been upgraded to a fully redundant network. The migration to SLC6 has also had an impact on the Control Room Desktop (CRD), the in house KDE-based desktop environment designed to enforce access policies while fulfilling the needs of the people working in the ACR and the SCR.
Finally, monitoring the health and the status of ˜3000 machines in the experimental area is obviously of the utmost importance, so the obsolete Nagios v2 has been replaced with Icinga, complemented by Ganglia for performance data.
This paper serves for reporting "What", "Why" and "How" we did in order to improve and produce a system capable of performing for the next 3 years of ATLAS data taking.

Track 8 Session / 39

Triggering events with GPUs at ATLAS

Author: Sami Kama[1]

Co-authors: Ademar Tavares Delgado [2]; Andrea Messina [3]; Antonio Sidoti [4]; Dmitry Emeliyanov [5]; John Baines [5]; Jose Soares Augusto [6]; Lorenzo Rinaldi [7]; Lourenco Vaz Gil Lopes [2]; Matteo Bauce [8]; Matteo Negrini [7]; Patricia Conde Muino [2]; Ricardo Jose Morais Silva Goncalo [2]; Salvatore Tupputi [9]; Tomasz Bold [10]

[1] *Southern Methodist University (US)*

[2] *LIP Laboratorio de Instrumentacao e Fisica Experimental de Part*

[3] *Universita e INFN, Roma I*

[4] *INFN Bologna (IT)*

[5] *STFC - Rutherford Appleton Lab. (GB)*

[6] *LIP/FCUL*

[7] *Universita e INFN (IT)*

[8] *Universita e INFN, Roma I (IT)*

[9] *Universita e INFN, Bologna (IT)*

[10] *AGH Univ. of Science and Technology, Krakow*

Corresponding Author: sami.kama@cern.ch

The growing size and complexity of events produced at the high luminosities expected in 2015 at the Large Hadron Collider demands much more computing power for the online event selection and for the offline data reconstruction than in the previous data taking period. In recent years, the explosive performance growth of low-cost, massively parallel processors like Graphical Processing Units (GPUs) - both in computing power and in low energy consumption - make GPUs extremely attractive for solving the challenging computing tasks of high energy physics experiment like ATLAS.

After the optimisation of the computing intensive algorithms and their adaptation to GPUs, thus exploiting the paradigm of massively parallel computing, a small scale prototype of the full reconstruction chain of the ATLAS High Level Trigger under inclusion of the GPU-based algorithms has been implemented. We discuss the integration procedure of this prototype, the achieved performance and the prospects for the future.

Track 1 Session / 40

gFEX, the ATLAS Calorimeter Level 1 Real Time Processor

Author: Helio Takai[1]

Co-authors: Hucheng Chen [1]; Michael Begel [1]

[1] *Brookhaven National Laboratory (US)*

Corresponding Author: takai@bnl.gov

The global feature extractor (gFEX) is a component of the Level-1 Calorimeter trigger Phase-I upgrade for the ATLAS experiment. It is intended to identify patterns of energy associated with the hadronic decays of high momentum Higgs, W, & Z bosons, topquarks, and exotic particles in real time at the LHC crossing rate. The single processor board will be implemented as a fast reconfigurable processor based on four largeFPGAs. The board will receive coarse-granularity information from all the ATLAS calorimeters on 264 optical fibers with the data transferred at the 40 MHz LHC clock frequency. The gFEX will be controlled by a single system-on-chip processor, ZYNQ, that will be used to configure FPGAs, monitor board health, and interface to externalsignals. Although the board is being designed specifically for the ATLAS experiment, itis sufficiently generic that it could be used for fast data processing at other HEP or NP experiments. We will present the design of the gFEX board and discuss how it is beingimplemented.

Track 1 Session / 41

FELIX: a High-throughput network approach for interfacing to front end electronics for ATLAS upgrades

Author: Jorn Schumacher[1]

Co-authors: Alexander Roich [2]; Andrea Borga [3]; Benedetto Gorini [4]; David Francis [4]; Francesco Lanni [5]; Frans Philip Schreuder [3]; Gary Drake [6]; Giovanna Lehmann Miotto [4]; Henk Boterenbrood [3]; Hucheng Chen [5]; Jinlong Zhang [7]; John Thomas Anderson [8]; Joseph Vermeulen [3]; Julia Narevicius [2]; Kai Chen [9]; Lorne Levinson [2]; Wainer Vandelli [4]

[1] *University of Paderborn (DE)*

[2] *Weizmann Institute of Science (IL)*

[3] *NIKHEF (NL)*

[4] *CERN*

[5] *Brookhaven National Laboratory (US)*

[6] *Argonne National Laboratory (US)*

[7] *Shandong University/Brookhaven National Laboratory*

[8] *A*

[9] *BNL*

Corresponding Author: joern.schumacher@cern.ch

The ATLAS experiment at CERN is planning the full deployment of a new, unified link technology for connecting detector front-end electronics on the timescale of the LHC Run 4 (2025). It is estimated that roughly 8000 Gigabit Transceiver links (GBT), with transfer rates probably up to 9.6 Gbps, will replace existing links used for readout, detector control and distribution of timing and trigger information. In particular the links used for readout are often detector-specific. Already in Run 3 this technology will be deployed in conjunction with new muon detectors, additional muon first-level triggering electronics and new on-detector and off-detector liquid argon calorimeter electronics to be used for first level triggering. A total of roughly 2000 GBT or GBT-like links (for connecting to off-detector trigger electronics) will be needed. A new class of devices will need to be developed to interface many GBT links to the rest of the trigger, data-acquisition and detector control systems. In this paper we present FELIX, the Front End LInk eXchange. The purpose of FELIX is to route data from and to multiple GBT links via a high-performance general purpose network capable of a total throughput that could be of O(20 Tbps).

FELIX is designed to be used as a multi-purpose routing device; data can be forwarded to or from multiple destinations connected to the network, e.g. data-acquisition, monitoring, detector control, calibration, or configuration systems. The system will be capable of handling special bi-directional low-latency channels, such as needed for transferring Timing, Trigger and Control, TTC, signals via GBT links or for GBT-to-GBT Direct-Output-Low-Latency connections to first level trigger electronics. The software layer will integrate support for data sampling, quality-of-service prioritization, load balancing based on a round-robin scheme and automatic failover.

FELIX is the central element of a new readout architecture that replaces the legacy point-to-point links with a switched network. The new architecture will be more dynamic, flexible and adaptable. By replacing point-to-point links with a switched network, the number of single points-of-failure can be reduced and a more robust and fault-tolerant system can be built, also implementing traffic equalization schemes.

In preparation for the deployment in Run 3 and to support the ongoing detector and trigger electronics development, a FELIX technology demonstrator is planned to be available early 2015, using commercial-off-the-shelf server PC technology in combination with a commercial FPGA-based PCIe Gen3 I/O card interfacing to up to 24 GBT links and with TTC connectivity provided by an FMC-based mezzanine card. Dedicated firmware for the Xilinx Virtex-7 FPGA installed on the I/O card alongside an interrupt-driven Linux kernel driver and user-space software will provide the required functionality. On the network side, the FELIX demonstrator connects to an Ethernet-based network. Extensions to other high-performance network technologies, such as Infiniband are possible and will be discussed in the paper.

In this paper we introduce a new approach to interfacing to on-detector and trigger electronics on the basis of GBT link or GBT-like link technology for ATLAS in Run 3 and Run 4, highlighting the innovative elements and advantages. We will then derive the functional and performance requirements on FELIX and present the design and implementation of the FELIX demonstrator. Furthermore, we will show throughput performance results as well as networking and data processing benchmarks. We intend to show that our design is a viable solution for a multi-purpose routing device in the anticipated ATLAS architecture of LHC Run 4.

Track 3 Session / 42

Evolution of Database Replication Technologies for WLCG

Author: Zbigniew Baranowski[1]

Co-authors: Gancho Dimitrov [1]; Lorena Lobato Pardavila [1]; Luca Canali [1]; Marcin Blaszczyk [1]

[1] *CERN*

Corresponding Author: zbigniew.baranowski@cern.ch

During LHC run 1 ATLAS and LHCb databases have been using Oracle Streams replication technology for their use cases of data movement between online and offline Oracle databases. Moreover ATLAS has been using Streams to replicate conditions data from CERN to selected Tier 1s. GoldenGate is a new technology introduced by Oracle to replace and improve on Streams, by providing better performance, robustness, manageability and support. At the same time Oracle Active Data Guard is an another technology that fulfils several use cases in this area, by allowing the creation of read-only copies of production databases, currently in use by CMS, ALICE and more recently by ATLAS for controls data. In this paper we will provide a short review of when GoldenGate is preferable to Streams/Active Data Guard. Further we will report the details of the migration of CERN and Tier 1s database replication setups/configurations from Streams to GoldenGate. In particular we will describe the architecture of GoldenGate replication services, our performance tests for remote replication done in collaboration with ATLAS and Tier 1 database administrators and the return of experience from running GoldenGate in production. We will also summarize our tests for future work on how GoldenGate can be used to reduce the downtime for Oracle version upgrades and on using GoldenGate for heterogeneous database replication.

43

Scale Out Databases for CERN Use Cases

Author: Zbigniew Baranowski[1]

Co-authors: Daniel Lanza Garcia [2]; Luca Canali [1]; Maciej Grzybek [3]

[1] *CERN*

[2] *Univ. Extremadura, Cen. Uni. Merida (ES)*

[3] *Warsaw University of Technology (PL)*

Corresponding Author: zbigniew.baranowski@cern.ch

Data generation rates are expected to grow very fast for some database workloads going into LHC run 2 and beyond. In particular this is expected for data coming from controls, logging and monitoring systems. Storing, administering and accessing big data sets in a relational database system is in certain cases very demanding on the technology and therefore on costs. Notably one of the critical parts in the architecture of Oracle database clusters is the use of shared storage. Therefore there is a high interest in the CERN database community to look for alternative solutions for storing and querying big data volumes with fast and scalable data access time. Scale out database engines are an emerging and rapidly developing area. Recently a technical solution that has attracted attention is Cloudera Impala with columnar storage provided by Parquet on top of Hadoop Distributed File System. This solution has the additional benefit of offering SQL as the main data access interface which makes it easy to integrate with existing client application. In this paper we will describe the architecture of database systems based on Impala Hadoop clusters and we will discuss the results of our tests, including tests of data loading and integration with existing data sources, notably Oracle databases. We will report on query performance tests done with various data sets of interest at CERN, notably the accelerator log database.

Track 4 Session / 45

Visualization of dCache accounting information with state-of-the-art Data Analysis Tools.

Author: Tigran Mkrtchyan[1]

Co-authors: Birgit Lewendel [1]; Lusine Yakovleva [1]; Patrick Fuhrmann [1]

[1] *DESY*

Corresponding Author: kofemann@gmail.com

Over the previous years, storage providers in scientific infrastructures were facing a significant change in the usage profile of their resources. While in the past, a small number of experiment frameworks were accessing those resources in a coherent manner, now, a large amount of small groups or even individuals request access in a completely chaotic way. Moreover, scientific laboratories have been recently forced to provide detailed accounting information for their communities and individuals. Another consequence of the chaotic access profiles is the difficulty, for often rather small operating teams, to detect malfunctions in extremely complex storage systems, composed of a large variety of different hardware components. Although information about usage and possible malfunction is available in the corresponding log and billing files, the sheer amount of collected meta data makes it extremely difficult to be handled or interpreted. Simply the dCache production instances at DESY are producing Gigabytes of meta data per day. To cope with those pressing issues, DESY has been evaluating and put into production a Big Data processing tool, enabling our operation team to analyze log and billing information by providing a configurable and easy to interpret visualization of that data.
This presentation will demonstrate how DESY built a real-time monitoring system, visualizing dCache billing files and providing an intuitive and simple to operate Web interface, using ElasticSearch, Logstash and Kibana.

46

dCache: Moving to open source development model and tools.

Author: Tigran Mkrtchyan Mkrtchyan[1]

Co-authors: Gerd Behrmann [2]; Patrick Fuhrmann [3]; Paul Millar [4]

[1] *Deutsches Elektronen-Synchrotron DESY*

[2] *NDGF*

[3] *DESY*

[4] *Deutsches Elektronen-Synchrotron (DE)*

Corresponding Author: kofemann@gmail.com

For over a decade, dCache.ORG has provided software which is used at more than 80 sites around the world, providing reliable services for WLCG experiments and others. This can be achieved only with a well established process from white board, where ideas are created, through packages, installed on the production systems. Since early 2013 we have moved to git as out source code management system as well as to GitHub for our source code hosting service.

This presentation will show how techniques and tools are used by dCache developers to manage code and releases, how to handle external contributions and still keep the quality and reliability of the provided packages.

Track 5 Session / 47

Computer security: surviving and operating services despite highly skilled and well-funded organised crime groups

Author: Romain Wartel[1]

[1] *CERN*

Corresponding Author: romain.wartel@cern.ch

This presentation gives an overview of the current computer security landscape. It describes the main vectors of compromises in the academic community including lessons learnt, reveals inner mechanisms of the underground economy to expose how our computing resources are exploited by organised crime groups, and gives recommendations how to better protect our computing infrastructures. By showing how these attacks are both sophisticated and profitable, the presentation concludes that an important mean to adopt an appropriate response is to build a tight international collaboration and trusted information sharing mechanisms within the community.

48

Enabling identity federation for the research and academic communities

Author: Romain Wartel[1]

[1] *CERN*

Corresponding Author: romain.wartel@cern.ch

Federated identity management (FIM) is an arrangement made among multiple organisations that lets subscribers use the same identification data, e.g. account names & credentials, to obtain access to the secured resources and computing services of all other organisations in the group. Specifically in the various research communities there is an increased interest in a common approach as there is obviously a large potential for synergies.
Several research communities, infrastructures and as well as federations have worked together in this area. This presentation will give an overview of this ongoing effort, including a common vision for FIM, a set of requirements and a number of recommendations for ensuring secure and interoperable operations across multiple administrative domains. In particular, the current SCI / Sir-T-Fi work will be presented, providing details on a common trust framework proposing minimal requirements for participants, as well as a common security incident response policy for federations.

49

Real-time flavour tagging selection in ATLAS

Author: Claudia Bertella[1]

[1] *Johannes-Gutenberg-Universitaet Mainz (DE)*

Corresponding Author: claudia.bertella@cern.ch

In high-energy physics experiments, online selection is crucial to identify the few interesting collisions from the large data volume processed. In the overall ATLAS trigger strategy, b-jet triggers are designed to identify heavy-flavor content in real-time and, in particular, provide the only option to efficiently record events with fully hadronic final states containing b-jets. In doing so, two different, but related, challenges are faced. The physics goal is to optimise as far as possible the rejection of light jets from QCD processes, while retaining a high efficiency on selecting jets from beauty, while maintaining affordable trigger rates without raising jet energy thresholds. This maps into a challenging computing task, as charged tracks and their corresponding vertexes must be reconstructed and analysed for each jet above the desired threshold, regardless of the increasingly harsh pile-up conditions. The performance of b-jet triggers during the LHC Run 1 data-taking campaigns is presented, together with an overview of the new online b-tagging strategy and algorithms, designed to face the above mentioned challenges, which will be adopted during Run 2.

50

Using an Intel Galileo board as a Hardware Token Server

Author: Jan Justinus Keijser[1]

[1] *NIKHEF*

Corresponding Author: janjust@nikhef.nl

The Intel Galileo Arduino board is a low cost, low power 32bit Pentium-class computer. It is normally used for embedded devices but it can also run a full-blown version of Linux.

Grid security can be greatly enhanced using hardware token for two-factor authentication. Two-factor autentication is based on the idea that in order to obtain access you need both something you know (i.e. a password) and something you possess (i.e. a hardware token). Pilot job security is improved by storing the pilot job grid certificate/private key pair on such a hardware token. In the Netherlands we use SafeNet eToken PRO's for generating and storing private keys for robot services. Typically, robot hardware tokens are inserted into server-class systems inside a data center for maximum security.

This research was done to address the power consumption issue of such server-class systems. An Intel Galileo board typically draws 5 W of power, which is less than even the most energy efficient Atom-based server. As part of this research we built a 32bit version of CentOS 7, tuned specifically to run on the Galileo board. On top of this we installed and configured the hardware token drivers and the 'myproxy' software needed to upload robot grid proxies to a Globus MyProxy or Dirac Proxy store.

Next to the poster a live demonstration of the Galileo+eToken setup is shown.

51

Distributed Root analysis on an ARM cluster

Author: Marcus Hardt[1]

Co-authors: Bernd Rische [2]; Christopher Jung [3]; Diana Gudu [2]; Matthias Schnepf [2]

[1] *Kalrsruhe Institute of Technology*

[2] *KIT*

[3] *KIT - Karlsruhe Institute of Technology (DE)*

Corresponding Author: christopher.jung@kit.edu

An ARM cluster, CEPH, ROOT and the energy balance

The total cost of ownershipt (TCO) of todays computer centres are increasingly driven the power consumption of computing equipment. The question arises if Intel based CPUs are still the best choice for analysis tasks. Furthermore, data-driven computing models are emerging.

This contribution compares performance, TCO, power and energy consumption of an ARM based CEPH cluster with off the shelf computers and an average GridKa compute node. The ARM cluster was constructed from 16 CubieBoards that comprise a dual Core ARM-A7 CPU at 1GHz as well as a SATA controller.

As analysis use case of this comparison we chose a simple high-energy di-muon events simulated for LHC.

The goal is to assess which configuration is better suited for future analysis hardware.

Track 8 Session / 52

Online-Analysis of Hits in the Belle-II Pixeldetector for Separation of Slow Pions from Background

Author: Steffen Baehr[1]

Co-authors: Christian Pulvermacher [1]; Juergen Becker [1]; Martin Heck [1]; Michael Feindt [1]; Oliver Sander [1]

[1] *Karlsruhe Institute of Technology*

Corresponding Author: steffen.baehr@kit.edu

The impending Upgrade of the Belle experiment is expected to increase the generated data set by a factor of 50.
This means that for the planned pixeldetector, which is the closest to the interaction point, the data rates are going to increase to over 20 GB/s.

Combined with data generated by the other detectors, this rate is too big to be efficiently send out to offline processing.
This is makes the employment of online data reduction schemes necessary, in which background is detected and rejected in order to reduce the data rates.
In this paper an approach for efficient online data reduction for the planned pixeldetector of Belle-II is presented.
The approach centers around the usage of an algorithm, the NeuroBayes, that is based on multivariate analysis, allowing the identification of signal and background by analysing clusters of hits in the pixeldetector on FPGAs.
The algorithm is leveraging the fact that hits of signal particles can have very different characteristics, compared to background, when passing through the pixeldetector.
The applicability and advantages in performance are shown through the D* decay.
In Belle-II these decays produce pions with such a small transversal momentum, that they barley escape the pixel detector itself.
In a common approach like extrapolation of tracks from outer detectors to RoIs, these pions are simply lost, since they do not reach all necessary layers of a detector.
Meanwhile usage of the cluster analysis succeeds in separating those pions from background, allowing to retain the data.
For that characteristics of corresponding hits, like the total amount of charge deposited in the pixels, are used for separation.
The capability for effective data reduction is underlined by a background reduction of at least 90% and signal efficiency of 95 %, for slow pions.
An implementation of the algorithm for usage on Virtex-6 FPGAs, that are used at the pixeldetector, was performed.
It is shown that the resulting implementation succeeds in replicating the efficiency of the algorithm, implemented in software, while throughputs that suffice hard realtime constraints, set by the read out system of Belle-II, are achieved and efficient use of the resources present on the FPGA is made.

53

The performance and development of the Inner Detector Trigger at ATLAS for LHC Run 2

Author: Stewart Martin-Haugh[1]

[1] *STFC - Rutherford Appleton Lab. (GB)*

Corresponding Author: stewart.martin-haugh@cern.ch

A description of the design and performance of the newly reimplemented tracking algorithms for the ATLAS trigger for LHC Run 2, to commence in spring 2015, is provided. The ATLAS High Level Trigger (HLT) has been restructured to run as a more flexible single stage process, rather than the two separate Level 2 and Event Filter stages used during Run 1. To make optimal use of this new scenario, a new tracking strategy has been implemented for Run 2 for the HLT. This new strategy will use a Fast Track Finder (FTF) algorithm to directly seed the subsequent Precision Tracking, and will result in improved track parameter resolution and significantly faster execution times than achieved during Run 1 but with no significant reduction in efficiency. The performance and timing of the algorithms for numerous physics signatures in the trigger are presented. The profiling infrastructure, constructed to provide prompt feedback from the optimisation, is described, including the methods used to monitor the relative perfor- mance improvements as the code evolves. The online deployment and commissioning, together with the first measurements with the Run 2 data are also discussed.

54

Monte Carlo Production Management at CMS

Author: Giovanni Franzoni[1]

Co-authors: Adrian Pol [1]; Antanas Norkus [2]; Gaelle Boudoul [3]; Jean-Roch Vlimant [4]; Phat Srimanobhas [5]

[1] *CERN*

[2] *Vilnius University (LT)*

[3] *Universite Claude Bernard-Lyon I (FR)*

[4] *California Institute of Technology (US)*

[5] *Chulalongkorn University (TH)*

Corresponding Author: giovanni.franzoni@cern.ch

The analysis of the LHC data at the Compact Muon Solenoid (CMS) experiment requires the production of a large number of simulated events. During the runI of LHC (2010-2012), CMS has produced over 12 Billion simulated events, organized in approximately sixty different campaigns each emulating specific detector conditions and LHC running conditions (pile up).

In order to aggregate the information needed for the configuration and prioritization of the events production, assure the book-keeping and of all the processing requests placed by the physics analysis groups, and to interface with the CMS production infrastructure, the web-based service 'Monte Carlo Management' (McM) has been developed and put in production in 2012.

McM is based on recent server infrastructure technology (CherryPy + java) and relies on a CouchDB database back-end.

This contribution will cover the one and half year of operational experience managing samples of simulated events for CMS, the evolution of its functionalities and the extension of its capability to monitor the status and advancement of the events production.

55

Performance of the ATLAS Muon Trigger in Run I and Upgrades for Run II

Author: Dai Kobayashi[1]

[1] *Tokyo Institute of Technology (JP)*

Corresponding Author: dai.kobayashi@cern.ch

The ATLAS experiment at the Large Hadron Collider (LHC) has taken data at a centre-of-mass energy between 900 GeV and 8 TeV during Run I (2009-2013). The LHC delivered an integrated luminosity of about 20 fb^{-1} in 2012, which required dedicated strategies to guard the highest possible physics output while reducing effectively the event rate. The Muon High Level Trigger has successfully adapted to the changing environment of a low luminosity in 2010 to the luminosities encountered in 2012. The selection strategy has been optimized for the various physics analyses involving muons in the final state. We will present the excellent performance achieved during Run I.

In preparation for the next data taking period (Run II) several hardware and software upgrades to the ATLAS Muon Trigger have been performed to deal with the increased trigger rate expected at higher center of mass energy and increased instantaneous luminosity. We will highlight the development of novel algorithms that have been developed to maintain a highly efficient event selection while reducing the processing time by a factor of three. In addition, the two stages of the high level trigger that was deployed in Run I will be merged for Run II. We will discuss novel approaches that are being developed to further improve the trigger algorithms for Run II and beyond.

56

ALDIRAC, a commercial extension to DIRAC

Author: Stephane Guillaume Poss[1]

Co-authors: Andrei Tsaregorodtsev [2]; Ricardo Graciani Diaz [3]

[1] *Alpes Lasers SA*

[2] *CPPM, Aix-Marseille Université, CNRS/IN2P3, Marseille, France*

[3] *University of Barcelona (ES)*

Corresponding Author: stephane.poss@alpeslasers.ch

We provide a report on ALDIRAC, the first DIRAC extension for a commercial application. DIRAC is a complete distributed computing solution, initially implemented for the LHCb experiment but now used by a wider community. The ALDIRAC extension is designed for the Alpes Lasers SA company in Neuchatel, Switzerland, to perform the simulation of the properties of Quantum Cascade Lasers on a Cloud system, namely Amazon EC2.

In this report, we will demonstrate that DIRAC is well suited to be used as a commercial solution. We will put an emphasis on the software developments required to use it in such a context. In particular, the resources used will be detailed. Moreover, as the Intellectual Property is an essential aspect of the business, a special treatment of the simulation software installation was implemented. Additional developments were necessary: due to the limited in-house computing resources, in particular network bandwidth, a system was designed to automatically deploy a complete DIRAC server on Amazon EC2 based on external signals. The machines provided by Amazon EC2 give the ability to quickly scale up and down the capabilities of the service. We will show the challenges faced and the solutions provided to allow agility with controlled costs. Finally the chosen data model will be presented: it is based on a postgresql database, where the every simulation result is stored individually. A set of meta data is used to select interesting simulation results.

57

Multi-threaded Object Streaming

Author: Andreas Pfeiffer[1]

Co-authors: Giacomo Govi [2]; Miguel Ojeda Sandonis [1]

[1] *CERN*

[2] *Fermi National Accelerator Lab. (US)*

Corresponding Author: andreas.pfeiffer@cern.ch

The CMS experiment at CERN's Large Hadron Collider in Geneva redesigned the code handling the conditions data during the last years, aiming to increase performance and enhance maintainability. The new design includes a move to serialise all payloads before storing them into the database, allowing the handling of the payloads in external tools independent of a given software release. In this talk we present the results of performance studies done using the serialisation package from the Boost suite as well as serialisation done with the ROOT (v5) tools. Furthermore, as the Boost tools allow parallel (de-)serialisation, we show the performance gains achieved with parallel threads when de-serialising a realistic set of conditions in CMS. Without specific optimisations an overall speed up of a factor of 3-4 was achieved using multi-threaded loading and de-serialisation of our conditions.

Track 1 Session / 58

Online data handling and storage at the CMS experiment

Authors: Anastasios Andronidis[1]; Andre Georg Holzner[None]; Andrea Petrucci[2]; Andrew Kevin Forrest[3]; Attila Racz[2]; Aymeric Arnaud Dupont[2]; Benjamin Stieger[2]; Carlos Nunez Barranco Fernandez[2]; Christian Deldicque[2]; Christoph Paus[4]; Christoph Schwick[2]; Christopher Colin Wakefield[5]; Dominique Gigi[2]; Emilio Meschi[2]; Frank Glege[2]; Frans Meijers[6]; Georgiana Lavinia Darlea[4]; Guillelmo Gomez Ceballos Retuerto[4]; Hannes Sakulin[2]; James Gordon Branson[7]; Jan Veverka[4]; Jean-Marc Olivier Andre[8]; Jeroen Hegeman[2]; Konstanty Sumorok[4]; Lorenzo Masetti[2]; Luciano Orsini[2]; Marc Dobson[2]; Marco Pieri[7]; Olivier Chaze[2]; Petr Zejdl[2]; Remi Mommsen[8]; Samim Erhan[9]; Sergio Cittolin[7]; Srecko Morovic[2]; Tomasz Adrian Bawej[10]; Ulf Behrens[11]; Vivian O'Dell[12]

[1] *Aristotle Univ. of Thessaloniki (GR)*

[2] *CERN*

[3] *University of Kent (GB)*

[4] *Massachusetts Inst. of Technology (US)*

[5] *Staffordshire University (GB)*

[6] *C*

[7] *Univ. of California San Diego (US)*

[8] *Fermi National Accelerator Lab. (US)*

[9] *Univ. of California Los Angeles (US)*

[10] *University of Wisconsin (US)*

[11] *Deutsches Elektronen-Synchrotron (DE)*

[12] *Fermi National Accelerator Laboratory (FNAL)*

Corresponding Author: lavinia.darlea@cern.ch

The CMS experiment at CERN is one of the two general-purpose detectors on the Large Hadron Collider (LHC) in the Geneva area, Switzerland. Its infrastructure has undergone massive upgrades during 2013 and 2014, which lead to major changes in the philosophy of its DAQ (Data AcQuisition) system. One of the major components of this system is the Storage Manager, which is responsible for buffering the online data generated at the level of the readout units (RUs), which crosses the computing farm undergoing different processing and filtering stages all the way to the builder units (BUs). The Storage Manager at CMS is made up of three components: the distributed file system, the Merger service and the Transfer System. In the production DAQ system there will be around 50 BUs which will concurrently write their output data at an expected aggregated rate of 2 GB/s. A merger service has been put in place to aggregate this data. Counting the merger reading of the data provided by the BUs, its writing back of the merged data and the final reading for transfer to Tier0, an estimated bandwidth of 7GB/s in concurrent r/w mode is needed. Moreover, the Storage Manager has to be able to cope with being disconnected from Tier0 for 1 week, so an estimated of 250TB of total usable disk space is required. A unified name space file system (Lustre) has been chosen in order to cope with these requirements. Three different implementations of the merger service are proposed, each one providing different advantages: robustness, easy debugging, bandwidth requirements reduction. The merger is also providing the bookkeeping needed for establishing the handshake protocol between the Transfer System and the Tier0 facility at CERN, which ensures data consistency and integrity throughout the transfer process. Even if the handshake protocol itself is identical to the one from the Run1, some adjustments needed to be performed in the Transfer System in order to accommodate the new structures providing the required meta-information related to the acquired data. In addition to the nominal transfer of data to Tier0, the Transfer System needs to intelligently distribute the data, for a number of data streams need to be (also) stored locally in the CMS network for various consumers to process on site. In this article we present the various technological and implementation choices of the three components of the Storage Manager.

Track 3 Session / 59

Experiences and challenges running CERN's high-capacity tape archive

Authors: German Cancio Melia[1]; Vlado Bahyl[1]

Co-author: Eric Cano [1]

[1] *CERN*

Corresponding Author: eric.cano@cern.ch

CERN's tape-based archive system has collected over 70 Petabytes of data during the first run of the LHC. The Long Shutdown is being used for migrating the complete 100 Petabytes data archive to higher-density tape media. During LHC Run 2, the archive will have to cope with yearly growth rates of up to 40-50 Petabytes. In this contribution, we will describe the scalable architecture for coping with the storage and long-term archival of such massive data amounts, as well as the procedures and tools developed for the proactive and efficient operation of the tape infrastructure. This will include also the features developed for automated problem detection, identification and notification. We will also review the challenges resulting and mechanisms devised for measuring and enhancing availability and reliability, as well as ensuring the long-term integrity and bit-level preservation of the complete data repository. Finally, we will present an outlook in terms of the future performance and capacity requirements growth and how they match the expected tape technology evolution.

60

Automated workflows for critical time-dependent calibrations at the CMS experiment.

Author: Gianluca Cerminara[1]

[1] *CERN*

Corresponding Author: gianluca.cerminara@cern.ch

Fast and efficient methods for the calibration and the alignment of the detector are a key asset to exploit the physics potential of the Compact Muon Solenoid (CMS) detector and to ensure timely preparation of results for conferences and publications.

To achieve this goal, the CMS experiment has set up a powerful framework. This includes automated workflows in the context of a prompt calibration concept, which allows for a quick turnaround of the calibration process following as fast as possible any change in running conditions.

The presentation will review the design and operational experience of these workflows and the related monitoring system during the LHC RunI and focus on the development, deployment and commissioning in preparation of RunII.

Track 6 Session / 61

Indico - the road to 2.0

Author: Pedro Ferreira[1]

[1] *CERN*

Corresponding Author: pedro.ferreira@cern.ch

Indico has come a long way since it was first used to organize CHEP 2004.
More than ten years of development have brought new features and projects, widening the application's feature set and enabling event organizers to work even more efficiently. While this has boosted the tool's usage and facilitated its adoption by a remarkable 300,000 events (at CERN only), it has also generated a whole new range of challenges, which have been the target of the team's attention for the last 2 years. One of them was that of scalability and the maintainability of the current database solution (ZODB).

After careful consideration, the decision was taken to move away from ZODB to PostgreSQL, a relational and widely-adopted solution that will permit the development of a more ambitious feature set as well as improved performance and scalability. A change of this type is by no means trivial in nature and requires the refactoring of most backend code as well as the full rewrite of significant portions of it. We are taking this opportunity to modernize Indico, by employing standard web modules, technologies and concepts that not only make development and maintenance easier but also constitute an upgrade to Indico's stack. The first results are already visible since August 2014, with the full migration of the Room Booking module to the new paradigm.

In this paper we explain what has been done so far in the context of this ambitious migration, what have been the main findings and challenges, as well as the main technologies and concepts that will constitute the foundation of the resultant Indico 2.0.

Track 6 Session / 62

Vidyo@CERN: A Service Update

Author: Joao Correia Fernandes[1]

[1] *CERN*

Corresponding Author: joao.fernandes@cern.ch

We will present an overview of the current real-time video service offering for the LHC, in particular the operation of the CERN Vidyo service will be described in terms of consolidated performance and scale: The service is an increasingly critical part of the daily activity of the LHC collaborations, topping recently more than 50 million minutes of communication in one year, with peaks of up to 852 simultaneous connections. We will elaborate on the improvement of some front-end key features such as the integration with CERN Indico, or the enhancements of the Unified Client and also on new ones, released or in the pipeline, such as a new WebRTC client and CERN SSO/Federated SSO integration. An overview of future infrastructure improvements, such as virtualization techniques of Vidyo routers and geo-location mechanisms for load-balancing and optimum user distribution across the service infrastructure will also be discussed. The work done by CERN to improve the monitoring of its Vidyo network will also be presented and demoed.
As a last point, we will touch the roadmap and strategy established by CERN and Vidyo with a clear objective of optimizing the service both on the end client and backend infrastructure to make it truly universal, to serve Global Science; that includes the introduction of the multi-tenant concept to serve different communities, as the follow up of CERN's decision to offer the Vidyo service currently operated for the LHC, to other Sciences, Institutions and Virtual Organizations beyond HEP that might express interest for it.

Track 6 Session / 63

Offering Global Collaboration Services beyond CERN and HEP

Author: Thomas Baron[1]

Co-authors: Joao Correia Fernandes [1]; Pedro Ferreira [1]

[1] *CERN*

Corresponding Author: joao.fernandes@cern.ch

The CERN IT department has built over the years a performant and integrated ecosystem of collaboration tools, from videoconference and webcast services to event management software. These services have been designed and evolved in very close collaboration with the various communities

surrounding the laboratory and have been massively adopted by CERN users. To cope with this very heavy usage, global infrastructures have been deployed which take full advantage of CERN's international and global nature.
If these services and tools are instrumental in enabling the worldwide collaboration which generates major HEP breakthroughs, they would certainly also benefit other sectors of science in which globalisation has already taken place. Some of these services are driven by commercial software (Vidyo or Wowza for example), some others have been developed internally and have already been made available to the world as Open Source Software in line with CERN's spirit and mission. Indico for example is now installed in 100+ institutes worldwide.
But providing the software is often not enough and institutes, collaborations and project teams do not always possess the expertise, or human or material resources that are needed to set up and maintain such services. Regional and national institutions have to answer needs which are growingly global and often contradict their operational capabilities or organisational mandate and so are looking at existing worldwide service offers such as CERN's.
We believe that the accumulated experience obtained through the operation of a large scale worldwide collaboration service combined with CERN's global network and its recently-deployed Agile Infrastructure would allow the Organisation to set up and operate collaborative services, such as Indico and Vidyo, at a much larger scale and on behalf of worldwide research and education institutions and thus answer these pressing demands while optimizing resources at a global level. Such services would be built over a robust and massively scalable Indico server to which the concept of communities would be added, and which would then serve as a hub for accessing other collaboration services such as Vidyo, on the same simple and successful model currently in place for CERN users. This talk will describe this vision, its benefits and the steps which have already been taken to make it come to life.

64

The new CERN tape software - getting ready for total performance

Authors: Eric Cano[1]; Steven Murray[1]

Co-authors: Daniele Francesco Kruse [1]; David Come [2]; Viktor Kotliar [3]

[1] *CERN*

[2] *CERN/ISAE-Supaéro*

[3] *Institute for High Energy Physics (RU)*

Corresponding Author: eric.cano@cern.ch

CASTOR (the CERN Advanced STORage system) is used to store the custodial copy of all of the physics data collected from the CERN experiments, both past and present. CASTOR is a hierarchical storage management system that has a disk-based front-end and a tape-based back-end. The software responsible for controlling the tape back-end has been redesigned and redeveloped over the last year and shall be put into production at the beginning of 2015. This paper summarises the motives behind the redesign, describes in detail the redevelopment work and concludes with the short and long-term benefits.

Modern tape drives achieve 250 MB/s and speeds up to 1GB/s are on the roadmaps. To achieve this performance and to drop support for obsolete requirements, a new tape server software has been designed from the ground up, with a fully pipelined architecture. Disk and tape transfers, instruction fetches and result reporting are all running in independent threads communicating through queues to prevent external latencies from impacting tape drive performance.

The software has been developed in C++, facilitating modularity, unit testing, and faster development turnaround to face future challenges. The main developments included a novel methodology for SCSI development, polymorphic drive classes and factories to minimize extra coding, a tape file layer reimplementing the existing file format, polymorphic disk file access methods for easy extension to new protocols, thread-safe FIFO containers, easy to use threading building blocks and a memory management system for buffering data between disk and tape transfers.

Development was backed by a continuous integration system that included unit testing. Thanks to the use of mock objects, such as simulated tape drives, the unit tests covered complex use cases comparable with full production sessions. The unit tests were also validated by memory leak and race condition detectors.

Thanks to strict adherence to unit testing methodologies, the project went very smoothly. The development started on the 31st May 2013. The first read-only test version was delivered to tape operators mid-August 2014 for running against production stagers and the read-write version followed in mid-September. The project took the equivalent 2 FTEs for 5 people involved.

Starting from this solid base, the CERN tape infrastructure will evolve to address future needs such as long term data preservation, striped file systems access for higher throughput and session preemption to improve the granularity of drive allocations and to ensure for the full utilization of all drives at all times.

65

Big Data Analytics as a Service Infrastructure: Challenges, Desired Properties and Solutions

Author: Manuel Martin Marquez[1]

[1] *CERN*

Corresponding Author: manuel.martin.marquez@cern.ch

CERN's accelerator complex is an extreme data generator, every second an important amount of comprehensively heterogeneous data coming from control equipment and monitoring agents is persisted and needs to be analysed. Over the decades, CERN's researching and engineering teams have applied different approaches, techniques and technologies. This situation has minimized the necessary collaboration and more relevantly the cross data analytics over different domains. These two factors are essential to unlock hidden insights and correlations between the underlying processes, which enable better and more efficient daily-based accelerators operations and more informed decisions.
The proposed Big Data Analytics as a Service Infrastructure aims to: (1) Integrate the existing developments. (2) Centralize and standardize the complex data analytics needs for the wide CERN research and engineering community. (3) Deliver real time and batch data analytics capabilities and (4) provide transparent access and extraction-transformation-load, ETL, mechanisms to the different and critical-mission existing data repositories.
This paper reflects the desired properties resulting from the analysis on CERN's data analytics requirements; the main challenges: technological, collaborative and educational; and finally potential solutions and lessons learned.

Track 2 Session / 66

Kalman Filter Tracking on Parallel Architectures

Authors: Avi Yagil[1]; Daniel Sherman Riley[2]; Frank Wuerthwein[3]; Giuseppe Cerati[1]; Ian Macneill[1]; Kevin Mcdermott[2]; Matevz Tadel[1]; Peter Elmer[4]; Peter Wittich[2]; Steven Lantz[5]

[1] *Univ. of California San Diego (US)*

[2] *Cornell University (US)*

[3] *UCSD*

[4] *Princeton University (US)*

[5] *Cornell University*

Corresponding Author: giuseppe.cerati@cern.ch

Power density constraints are limiting the performance improvements of modern CPUs. To address this we have seen the introduction of lower-power, multi-core processors, but the future will be even more exciting. In order to stay within the power density limits but still obtain Moore's Law performance/price gains, it will be necessary to parallelize algorithms to exploit larger numbers of lightweight cores and specialized functions like large vector units. Example technologies today include Intel's Xeon Phi and GPGPUs.

Track finding and fitting is one of the most computationally challenging problems for event reconstruction in particle physics. At the High Luminosity LHC, for example, this will be by far the dominant problem. The need for greater parallelism has driven investigations of very different track finding techniques including Cellular Automata or returning to Hough Transform. The most common track finding techniques in use today are however those based on the Kalman Filter. Significant experience has been accumulated with these techniques on real tracking detector systems, both in the trigger and offline. They are known to provide high physics performance, are robust and are exactly those being used today for the design of the tracking system for HL-LHC.

Our previous investigations showed that, using optimized data structures, track fitting with Kalman Filter can achieve large speedup both with Intel Xeon and Xeon Phi. We report here our further progress towards an end-to-end track reconstruction algorithm fully exploiting vectorization and parallelization techniques in a realistic simulation setup.

Track 1 Session / 67

Data Science for Improving CERN's Accelerator Complex Control Systems

Author: Manuel Martin Marquez[1]

[1] *CERN*

Corresponding Author: manuel.martin.marquez@cern.ch

Data science is about unlocking valuable insights and obtaining deep knowledge out of the data. Its application enables more efficient daily-based operations and more intelligent decision-making processes. CERN has been very successful on developing custom data-driven control and monitoring systems. Several millions of control devices: sensors, front-end equipment, etc., make up these critical-mission services and have lead to a significant investment in terms of data persistency. Exploiting CERN's historical investment on data and evolve the controls and monitoring infrastructures to intelligent, predictive and proactive control and monitoring systems has become an essential task to overcome some the most important challenges to be faced in the coming years. CERN's Engineering, Beams and Information Technology departments have documented this fact. This paper describes how applied data science: classification, clustering, time series forecasting and information discovery techniques, allows us to improve the efficiency of CERN's control and monitoring systems by performing efficient and intelligent predictive maintenance on control equipment, detecting potential anomalies and spotting hidden root causes. Future developments and applications are also discussed.

68

The new ALICE DQM client: a web access to ROOT-based objects

Author: Barthelemy Von Haller[1]

Co-authors: Adam Wegrzynek [2]; Adriana Telesca [1]; Charles Delort [3]; Csaba Soos [1]; Ervin Denes [4]; Filippo Costa [1]; Franco Carena [1]; Giuseppe Simonetti [5]; Jeremi Niedziela [2]; Pierre Vande Vyvre [1]; Roberto Divia [1]; Sylvain Chapeland [1]; Ulrich Fuchs [1]; Vasco Chibante Barroso [1]; Wisla Carena [1]

[1] *CERN*

[2] *Warsaw University of Technology (PL)*

[3] *Ministere des affaires etrangeres et europeennes (FR)*

[4] *Hungarian Academy of Sciences (HU)*

[5] *Ludwig-Maximilians-Univ. Muenchen (DE)*

Corresponding Author: barthelemy.von.haller@cern.ch

A Large Ion Collider Experiment (ALICE) 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). The online Data Quality Monitoring (DQM) plays an essential role in the experiment operation by providing shifters with immediate feedback on the data being recorded in order to quickly identify and overcome problems.

An immediate access to the DQM results is needed not only by shifters in the control room but also by detector experts worldwide.
As a consequence, a new web application has been developed to dynamically display and manipulate the ROOT-based objects produced by the DQM system in a flexible and user friendly interface.

This paper describes the architecture and design of the tool, its main features and the technologies that were used, both on the server and the client side. In particular, we detail how we took advantage of the most recent ROOT JavaScript I/O and web server library to give interactive access to ROOT objects stored in a database. We describe as well the use of modern web techniques and packages such as AJAX, DHTMLX and Jquery, which has been instrumental in the successful implementation of a reactive and efficient application.

We finally present the performance of this application under normal and peak load and how code quality was ensured. We conclude with a roadmap for future technical and functional developments.

Track 5 Session / 69

COMPUTING STRATEGY OF THE AMS-02 EXPERIMENT

Authors: Baosong Shan[1]; vitali choutko[2]

Co-authors: Alexander Egorov [2]; Alexandre Eline [2]

[1] *Beihang University (CN)*

[2] *Massachusetts Inst. of Technology (US)*

Corresponding Author: baosong.shan@cern.ch

The Alpha Magnetic Spectrometer (AMS) is a high energy physics experiment installed and operating on board of the International Space Station (ISS) from May 2011 and expected to last through Year 2024 and beyond. The computing strategy of the AMS experiment is discussed in the paper, including software design, data processing and modelling details, simulation of the detector performance and overall computing organization.

Track 1 Session / 70

A New Event Builder for CMS Run II

Author: Remi Mommsen[1]

Co-authors: Andre Georg Holzner [2]; Andrea Petrucci [3]; Andrew Kevin Forrest [4]; Andronidis Anastasios [3]; Attila Racz [3]; Aymeric Arnaud Dupont [3]; Benjamin Stieger [3]; Carlos Nunez Barranco Fernandez [3]; Christian Deldicque [3]; Christoph Paus [5]; Christoph Schwick [3]; Christopher Colin Wakefield [6]; Dominique Gigi [3]; Emilio Meschi [3]; Frank Glege [3]; Frans Meijers [3]; Georgiana Lavinia Darlea [5]; Guillelmo Gomez Ceballos Retuerto [5]; Hannes Sakulin [3]; James Gordon Branson [2]; Jan Veverka [5]; Jean-Marc Olivier Andre [7]; Jeroen Hegeman [3]; Konstanty Sumorok [5]; Lorenzo Masetti [3]; Luciano Orsini [3]; Marc Dobson [3]; Marco Pieri [2]; Olivier Chaze [3]; Petr Zejdl [3]; Samim Erhan [8]; Sergio Cittolin [2]; Srecko Morovic [3]; Tomasz Adrian Bawej [9]; Ulf Behrens [10]; Vivian O'Dell [11]

[1] *Fermi National Accelerator Lab. (US)*

[2] *Univ. of California San Diego (US)*

[3] *CERN*

[4] *University of Kent (GB)*

[5] *Massachusetts Inst. of Technology (US)*

[6] *Staffordshire University (GB)*

[7] *FNAL*

[8] *Univ. of California Los Angeles (US)*

[9] *University of Wisconsin (US)*

[10] *Deutsches Elektronen-Synchrotron (DE)*

[11] *Fermi National Accelerator Laboratory (FNAL)*

Corresponding Author: remigius.mommsen@cern.ch

The data acquisition system (DAQ) of the CMS experiment at the CERN Large Hadron Collider (LHC) assembles events at a rate of 100 kHz, transporting event data at an aggregate throughput of 100 GB/s to the high-level trigger (HLT) farm. The DAQ system has been redesigned during the LHC shutdown 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 with a throughput of 4 Tb/s. This paper will discuss the software design, protocols and optimizations for exploiting the hardware capabilities. We will present performance measurements from small-scale prototypes and from the full-scale production system.

Track 6 Session / 71

High-Speed Mobile Communications in Hostile Environments

Author: Stefano Agosta[1]

Co-authors: Frederic Chapron [1]; Rodrigo Sierra [1]

[1] *CERN*

Corresponding Author: stefano.agosta@cern.ch

With the inexorable increase in the use of mobile devices, for both general communications and mission-critical applications, wireless connectivity is required anytime and anywhere. This requirement is addressed in office buildings through the use of Wi-Fi technology but Wi-Fi is ill adapted for use in large experiment halls and complex underground environments such as the LHC tunnel and experimental caverns. CERN is instead investigating the use of 4G/LTE technology to address issues such as radiation tolerance and distance constraints. This presentation will describe these studies, presenting results on the level of data throughput that can be achieved and discussing issues such as the provision of a consistent user experience as devices migrate between Wi-Fi and 4G/LTE.

72

Towards a 21st Century Telephone Exchange at CERN

Author: Francisco Valentin Vinagrero[1]

Co-authors: Frederic Chapron [1]; Rodrigo Sierra [1]; Tony Cass [1]

[1] *CERN*

Corresponding Author: tony.cass@cern.ch

The advent of mobile telephony and VoIP has significantly impacted the traditional telephone exchange industry—to such an extent that private branch exchanges are likely to disappear completely in the near future. For large organisations, such as CERN, it is important to be able to smooth this transition by implementing new voice platforms that can protect past investments and the flexibility needed to securely interconnect emerging VoIP solutions and forthcoming developments such as VoLTE. We present the results of ongoing studies and tests at CERN of the latest technologies in this area.

73

The ALICE Glance Shift Accounting Management System

Author: Heron Henrique Martins Silva[1]

Co-authors: Adriana Telesca [2]; Carmen Maidantchik [1]; Federico Ronchetti [3]

[1] *Univ. Federal do Rio de Janeiro (BR)*

[2] *CERN*

[3] *Istituto Nazionale Fisica Nucleare (IT)*

Corresponding Author: adriana.telesca@cern.ch

ALICE (A Large Ion Collider Experiment) is an experiment at the CERN LHC (Large Hadron Collider) studying the physics of strongly interacting matter and the quark-gluon plasma.

The experiment operation requires a 24 hours per day and 7 days a week "shift"crew at the experimental site, composed by the ALICE collaboration members. Shift duties are calculated for each institute according to their correlated members. In order to ensure the full coverage of the experiment operation as well as its good quality, the ALICE Shift Accounting Management System (SAMS) is used to manage the shift bookings as well as the needed training.

ALICE SAMS is the result of a joint effort between the Federal University of Rio de Janeiro (UFRJ) and the ALICE Collaboration. The Glance technology, developed by the UFRJ and the ATLAS experiment, sits at the basis of the system as an intermediate layer isolating the particularities of the databases.

In this paper, we describe the ALICE SAMS development process and functionalities. The database has been modelled according to the collaboration needs and is fully integrated with the ALICE Collaboration repository in order to access members information and respectively roles and activities. Run, period and training coordinators can manage their subsystem operation and ensure an efficient personnel management. Members of the ALICE collaboration can book shifts and on call according to pre-defined rights.

ALICE SAMS features a user's profile containing all the statistics and user contact information as well as the Institutes profile. Both the user and institute profiles are public (within the scope of the collaboration) and show the done over due credit balance in real time. A shift calendar allows the run Coordinator to plan data taking periods in terms of which subsystems shifts are enabled or disabled and on call responsibles and slots. An overview display presents the shift crew attending

the control room and allows the run coordination team to confirm the presence of both regular and trainees shift personnel, necessary for credit accounting.

74

The ALICE Glance Membership Management System

Authors: Heron Henrique Martins Silva[1]; Stefan Boettger[2]

Co-authors: Adriana Telesca [3]; Carmen Maidantchik [1]; Igor Abreu Da Silva [1]

[1] *Univ. Federal do Rio de Janeiro (BR)*

[2] *Johann-Wolfgang-Goethe Univ. (DE)*

[3] *CERN*

Corresponding Author: adriana.telesca@cern.ch

ALICE (A Large Ion Collider Experiment) is an experiment at the CERN LHC (Large Hadron Collider) studying the physics of strongly interacting matter and the quark-gluon plasma. The experiment collaboration counts more than 1500 members from 148 institutes in 39 countries.

During the experiment start up in 2008 and the following years of data taking the information about members was manually maintaned in a static database called the ALICE Collaboration Database (ACDB). The increased size and complexity of the collaboration in terms of contracts diversity and institutions responsibility in physics and experiment operation made clear the need of using a dynamic and flexible system.

In this paper, we introduce the ALICE Glance Membership system, which is the result of a fruitful joint effort between UFRJ (Federal University of Rio de Janeiro) and the ALICE Collaboration at CERN. The Glance tecnology, developed by the UFRJ and the ATLAS experiment, sits at the basis of the system as an intermediate layer isolating the particularities of the databases.

The developed Web system manages members data and their employments contracts, activities and appointments in the collaboration. It handles institutes and funding agencies information, agreements and representatives, which is essential to give the right access to ALICE members, create mailing lists, electronic groups and automatic notifications as well as generating the list of publication authors. The ALICE Glance Membership interacts with external software such as the CERN Foundation database and the Greybook. The system supports the decentralization of functions, allowing Institute leaders to directly manage their team and members to access their own information, enabling a dynamic collaboration network through the Web.

Track 3 Session / 75

Architectures and methodologies for future deployment of multi-site Zettabyte-Exascale data handling platforms

Author: Manuel Delfino Reznicek[1]

Co-authors: Andreas Heiss [2]; Andrew Sansum [3]; Benoit Delaunay [4]; Brian Matthews [3]; Christian Neissner [5]; David Corney [3]; German Cancio Melia [6]; Giovanni Lamanna [7]; Ian Bird [6]; Ian Peter Collier [3]; Jamie Shiers [6]; Jean-Yves Nief [4]; Jose Flix Molina [8]; Luca dell'Agnello [9]; Marcello Maggi [9]; Maria Del Carmen Porto Fernandez [8]; Markus Schulz [6]; Martin Gasthuber [10]; Patrick Fuhrmann [11]; Pierre-Etienne Macchi [4]; Tommaso Boccali [9]; Vanessa Acin Portella [5]; Volker Guelzow [11]

[1] *Universitat Autònoma de Barcelona (ES)*

[2] *KIT*

[3] *RAL*

[4] *IN2P3*
[5] *IFAE*
[6] *CERN*
[7] *CNRS*
[8] *CIEMAT*
[9] *INFN*
[10] *Deutsches Elektronen-Synchrotron (DE)*
[11] *DESY*

Corresponding Author: delfino@pic.es

Several scientific fields, including Astrophysics, Astroparticle Physics, Cosmology, Nuclear and Particle Physics, and Research with Photons, are estimating that by the 2020 decade they will require data handling systems with data volumes approaching the Zettabyte distributed amongst as many as 10¹⁸ individually addressable data objects (Zettabyte-Exascale systems). It may be convenient or necessary to deploy such systems using multiple physical sites. This paper describes the findings of a working group composed of experts from several large European scientific data centres on architectures and methodologies that should be studied by building proof-of-concept systems, in order to prepare the way for building reliable and economic Zettabyte-Exascale systems. Key ideas emerging from the study are: the introduction of a global Storage Virtualization Layer which is logically separated from the individual storage sites; the need for maximal simplification and automation in the deployment of the physical sites; the need to present the user with an integrated view of their custom metadata and technical metadata (such as the last time an object was accessed, etc.); the need to apply modern efficient techniques to handle the large metadata volumes (e.g. Petabytes) that will be involved; and the challenges generated by the very large rate of technical metadata updates. It also addresses the challenges associated with the need to preserve scientific data for many decades. The paper is presented in the spirit of sharing the findings with both the user communities and data centre experts, in order to receive feedback and generate interest in starting prototyping work on the Zettabyte-Exascale challenges.

Track 3 Session / 76

NoSQL technologies for the CMS Conditions Database

Author: Roland Sipos[1]

[1] *Eotvos Lorand University (HU)*

Corresponding Author: roland.sipos@cern.ch

With the restart of the LHC in 2015, the growth of the CMS Conditions dataset will continue, therefore the need of consistent and highly available access to the Conditions makes a great cause to revisit different aspects of the current data storage solutions.

We present a study of alternative data storage backends for the Conditions Databases, by evaluating some of the most popular NoSQL databases to support a key-value representation of the CMS Conditions. An important detail about the Conditions that the payloads are stored as BLOBs, and they can reach sizes that may require special treatment (splitting) in these NoSQL databases. As big binary objects may be a bottleneck in several database systems, and also to give an accurate baseline, a testing framework extension was implemented to measure the characteristics of the handling of arbitrary binary data in these databases. Based on the evaluation, prototypes of a document store, using a column-oriented and plain key-value store, are deployed. An adaption layer to access the backends in the CMS Offline software was developed to provide transparent support for these NoSQL databases in the CMS context. Additional data modelling approaches and considerations in the software layer, deployment and automatization of the databases are also covered in the research. In this paper we present the results of the evaluation as well as a performance comparison of the prototypes studied.

Track 1 Session / 77

The Application of DAQ-Middleware to the J-PARC E16 Experiment

Author: Eitaro Hamada[1]

Co-authors: Daisuke Kawama [2]; Hiroshi Sendai [1]; Kyoichiro Ozawa [1]; Manobu Tanaka [1]; Masahiro Ikeno [1]; Satoshi Yokkaichi [2]; Tomonori Takahashi [3]; Wataru Nakai [4]; Yuhei Morino [1]; Yuki Obara [4]

[1] *High Energy Accelerator Research Organization (KEK)*

[2] *RIKEN Nishina Center*

[3] *Research Center for Nuclear Physics, Osaka University*

[4] *University of Tokyo*

Corresponding Author: ehamada@post.kek.jp

1. Introduction

We developed a DAQ system of the J-PARC E16 Experiment by using the DAQ-Middleware. We evaluated the DAQ system and confirmed that the DAQ system can be applied to the experiment. The DAQ system receives an average 660MB/spill of data (2-seconds spill per 6 seconds cycle). In order to receive such a large quantity of data, we need a network-distributed system. DAQ-Middleware is a software framework of a network-distributed DAQ system. Therefore, the framework is a useful tool for the DAQ system development. In our talk, we are going to talk about useful features of DAQ-Middleware, an architecture and DAQ performance of the J-PARC E16 Experiment DAQ system.

2. The J-PARC E16 Experiment

The aim of the J-PARC E16 Experiment is to measure mass spectra of vector mesons in nucleus using electron pair decays with a huge statistics. For such purpose, a high intensity proton beam is used and the interaction rate of the experiment becomes 1×10^{10}. To cope with such high intensity and the rate, we plan to use the Gas Electron Multiplier (GEM) Tracker and strip read outs. Figure1 shows the estimation of data transfer to DAQ PCs.

Indico rendering error

```
Could not include image: [404] Error fetching image
```

3. DAQ-Middleware

DAQ-Middleware is a software framework of a network-distributed DAQ system. The framework consists of some software components called DAQ-Components. The framework provides the basic functionalities, such as communication between DAQ-Components, transferring data, starting and stopping the DAQ system. DAQ-Components can be set on separate computers.

4. DAQ system for the E16 Experiment

The DAQ system performs following functions.

- store all data on storage devices
- build and monitor event data from a part of all events

Figure2 shows architecture of an entire DAQ system.

Indico rendering error

```
Could not include image: [404] Error fetching image
```

The DAQ system consists of two stages. 1st stage PCs read and store all data. The 2nd stage PC builds and monitors event data from a part of all events.

Figure3 shows DAQ-Component architecture on one 1st stage PC.

Indico rendering error

Could not include image: [404] Error fetching image

Blue boxes of the figure represent DAQ-Component. Gatherer reads data from one read-out module. Merger receives data from multiple Gatherers and sends the data to Dispatcher. Dispatcher sends data to Logger and Filter. Logger stores data on storage devices. Filter sends data, which meet specific conditions, to a next DAQ-Component on the 2nd stage PC. 1st stage consists of multiple PCs which have these DAQ-Components.

Figure4 shows DAQ-Component architecture on one 2nd stage PC.

Indico rendering error

Could not include image: [404] Error fetching image

Blue boxes of the figure represent DAQ-Component. Merger receives data from multiple Filters on the 1st stage PC and sends the data to Eventbuilder. This Merger is the same as one of the 1st stage PC. Eventbuilder builds event data from received data. Monitor analysis and monitor the event data.

5. Evaluation for the DAQ System

Because we did not have enough PCs, we evaluated 1st stage and 2nd stage separately.

Indico rendering error

Could not include image: [404] Error fetching image

5.1 1st stage evaluation

Indico rendering error

Could not include image: [404] Error fetching image

We measured maximum throughput of one 1st stage PC. Figure6 shows the environment of the evaluation. Because we did not have read-out modules, we used the emulators instead of those. Data format of the emulators was the same as that of the read-out modules. We installed DAQ-Components of the 1st stage to an evaluation PC. Figure5 shows the specification of the evaluation PC. We prepared the 2nd stage PC which is installed Skeltonsink. Skeltonsink received data and is used only for 1st stage evaluation.
Each emulator sent test data to the evaluation PC at a maximum rate. As shown in Table 1, one event data size is 45KB per one read-out module and event rate is 2000Hz.

Indico rendering error

Could not include image: [404] Error fetching image

Figure7 shows the evaluation result. The points mean measured value, line means ideal value. When the number of emulators was up to 7, measured value matched ideal value. The result shows the evaluation PC can process up to 7 emulators, when read-out modules send data at a maximum rate. One evaluation PC can process around 600MB/s of data at a maximum. During this evaluation, we observed data loss size. There was no data loss when the number of emulators was up to 7.

5.2 2nd stage evaluation

Indico rendering error

Could not include image: [404] Error fetching image

We evaluated 2nd stage. Figure8 shows the environment of the evaluation. We installed DAQ-Components of the 2nd stage on the evaluation PC. Figure5 shows specification of the evaluation PC.
We assumed that 50 read-out modules transferred 45kB/event of data and event rate was 10 Hz after passing Filter. In this case, the 2nd PC received 22MB/s of data. We confirmed that the evaluation PC was able to process 22MB/s of data losing no data. This result shows the evaluation PC has an enough capability.

5. Conclusion

The DAQ system of the J-PARC E16 Experiment consists of two stages. One evaluation PC on 1st stage can process 600MB/s of data at a maximum. Average total data transfer Rate to DAQ PCs is 330MB/s. Therefore, 1st stage can be developed by one PC or a few PCs. 2nd stage can be developed by one PC. Therefore, we were able to confirm the DAQ system can be applied to the experiment by a few PCs.

79

Software for implementing trigger algorithms on the upgraded CMS Global Trigger System

Author: Takashi Matsushita[1]

[1] *Austrian Academy of Sciences (AT)*

Corresponding Author: takashi.matsushita@cern.ch

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 to meet the physics objectives. The conditions for trigger object selection, with possible topological requirements on multi-object triggers, are combined by simple combinatorial logic (AND-OR-NOT) to form the algorithms. The most basic algorithms consist of applying E_T or p_T threshold to single objects. The present Global Trigger is comprised of several VME modules with FPGAs.

When the LHC resumes its operation in 2015, the collision-energy will be increased from 8 TeV to 13 TeV, with the luminosity expected to go up from 0.75 x 10^{34} $cm^{-2}s^{-1}$ to 2 x 10^{34} $cm^{-2}s^{-1}$. These operating environments will provide new challenges for the CMS trigger system. The CMS level-1 trigger system will be upgraded to improve its performance for selecting interesting physics events and to operate within the predefined data-acquisition rate.

Together with the upgrade of other level-1 trigger systems, the Global Trigger will be re-implemented on modern FPGAs on an Advanced Mezzanine Card in MicroTCA crate. The upgraded system will benefit from the ability to process complex algorithms with DSP slices and increased processing resources with optical links running at 10 Gbit/s, enabling more algorithms at a time than previously possible and allowing CMS to be more flexible in how it handles the trigger bandwidth. CMS also will be able to match different objects, e.g. muons with jets, with higher resolution and efficiency and be able to calculate more sophisticated quantities such as the mass of a pair of objects. In 2015, CMS plans to keep the present triggers running and commission the new ones simultaneously. Detailed comparisons will be performed between both to test everything from technical implementation to whether the new triggers perform better, before the older system will be turned off.

A software for handling trigger menu implementation on the present system is strongly coupled with the current hardware design. In order to handle the increased complexity of the trigger menu implemented on the upgraded Global Trigger, a set of new software has been developed. The software allows a physicist to define a menu with analysis-like triggers using intuitive user interface. The

menu is then realised on FPGAs with further software processing, instantiating predefined firmware blocks. The menu plays a central role in trigger selection and is shared by the second level trigger, known as the High Level Trigger as well as a trigger emulating software in an offline software environment. The menu information is stored in either an XML file or in a database for sharing information with other systems.

The design and implementation of the software for preparing a menu for the upgraded CMS Global Trigger system will be presented.

Track 7 Session / 80

Scaling the CERN OpenStack cloud

Author: Tim Bell[1]

Co-authors: Arne Wiebalck [1]; Belmiro Moreira [1]; Jan Van Eldik ; Jose Castro Leon [1]; Luis Fernandez Alvarez [1]; Stefano Zilli [1]

[1] *CERN*

Corresponding Author: stefano.zilli@cern.ch

CERN has been running a production OpenStack cloud since July 2013 to support physics computing and infrastructure services for the site. This is expected to reach over 100,000 cores by the end of 2015.

This talk will cover the different use cases for this service and experiences with this deployment in areas such as user management, deployment, metering and configuration of thousands of servers across two data centres.

Ongoing research, such as bare metal management, containers, software defined networking and orchestration will also be covered.

Track 6 Session / 81

Enterprise Social Networking Systems for HEP community

Author: Bruno Silva De Sousa[1]

Co-authors: Andreas Wagner [1]; Emmanuel Ormancey [1]; Pawel Grzywaczewski [1]

[1] *CERN*

Corresponding Author: bruno.sousa@cern.ch

The emergence of social media platforms in the consumer space unlocked new ways of interaction between individuals on the web. People develop now their social networks and relations based on common interests and activities with the choice to opt-in or opt-out on content of their interest. This kind of platforms have also an important place to fill inside large organizations and enterprises where communication and collaborators interaction are keys for development. They are enablers to unlock hidden opportunities.

Enterprise Social Networking Systems (ESN) add value to an organization by encouraging information sharing, capturing knowledge, enabling action and empowering people. For such they propose microblogging, profiles, social networking, suggestion systems and discussion forums. Microblogging introduces a lightweight and informal communication channel that primes information sharing, promotes discussion and gives an equal opportunity to everyone to reach a broad audience. Profiles and social networking are the perfect tool to seek for skills and can act as a catalyzer for

new projects and opportunities. Suggestion systems and discussion forums are the perfect tools to identify evolving trends and allow new communities to born.

CERN is currently rolling out an ESN which aims to unify and provide a single point of access to the multitude of information sources in the organization. It also implements social features that can be added on top of existing communication channels. This talk will review the experiences with this deployment, the risks and benefits and outlook for the future as a social community tool for HEP.

Track 7 Session / 82

Towards a production volunteer computing infrastructure for HEP

Authors: Miguel Marquina[1]; Nils Hoimyr[1]

Co-authors: Alvaro Gonzalez Alvarez [1]; Laurence Field [1]; Pete Jones [1]; Tomi Juhani Asp [2]

[1] *CERN*

[2] *University of Jyvaskyla (FI)*

Corresponding Author: miguel.marquina@cern.ch

Using virtualisation with CernVM has emerged as a de-facto standard among HEP experiments; it allows for running of HEP analysis and simulation programs in cloud environments. Following the integration of virtualisation with BOINC and CernVM(link is external), first pioneered for simulation of event generation in the Theory group at CERN, the LHC experiments ATLAS, CMS and LHCb have all adopted volunteer computing as part of their strategy to benefit from opportunistic computing resources. This presentation will describe the current technology for volunteer computing and the evolution of the BOINC service at CERN from project support for LHC@home(link sends e-mail) towards a general service. It will also provide some recommendations for teams and experiments wishing to benefit from volunteer computing resources.

Track 2 Session / 83

Energy Reconstruction using Artificial Neural Networks and different analytic methods in a Highly Granularity Semi-Digital Hadronic Calorimeter.

Author: Sameh Mannai[1]

Co-authors: Eduardo Cortina Gil [2]; Imad Laktineh [3]

[1] *Universite Catholique de Louvain (UCL) (BE)*

[2] *Universite catholique de Louvain*

[3] *Universite Claude Bernard-Lyon I (FR)*

Corresponding Author: sameh.mannai@cern.ch

The Semi-Digital Hadronic CALorimeter(SDHCAL) using Glass Resistive Plate Chambers (GRPCs) is
one of the two hadronic calorimeter options proposed by the ILD (International Large Detector) project for the future (ILC) International Linear Collider experiments.

It is a sampling calorimeter with 48 layers. Each layer has a size of 1 m² and
finely segmented into cells of 1 cm² ensuring a high granularity which is required for

the application of the Particle Flow Algorithm (PFA) in order to improve the jet energy resolution which is the corner stone of ILC experiments.

The electronic of SDHCAL provide 2-bit readout. It is equiped with power pulisng mode reducing the power consumption and thus heating related problems.

The performance of the SDHCAL technological prototype was tested successfully in beam tests at CERN during 2012. The next beam test will take place
at CERN in December 2014 with new improvements in hardware developments. Results of this test beam will be shown.

One of the main points to be discussed concerns the energy reconstruction in SDHCAL.

Based on Monte Carlo Simulation of the SDHCAL prototype with Geant4, we will show different analytic energy reconstruction methods. We will present the single particle energy resolution and the linearity of the detecor response to hadrons obtained with these methods.
In particular, we will highlight a new approch based on the Artificial Neural Networks used in the energy reconstruction and giving promising results compared to the classical analytic methods. Results will be presented for both simulation and real data in the aim to compare them.
In the same context, we will discuss the application of The Artifical Neural Network to purify the beam test data from contamination.
Results of particle separation obtained with the Artificial Neural Network will be shown and compared to classical event selection methods.

84

Progress in Geant4 electromagnetic physics modeling and validation

Author: Vladimir Ivantchenko[1]

Co-authors: Alexandre Bagulya [2]; John Apostolakis [1]; Makoto Asai [3]

[1] *CERN*

[2] *P.N. Lebedev Institute of Physics (FIAN)*

[3] *SLAC National Accelerator Laboratory (US)*

Corresponding Author: asai@slac.stanford.edu

The Geant4 electromagnetic (EM) physics sub-packages are key components of any simulation; in particular, the simulation of LHC experiments. A small variation of EM physics may affect prediction accuracy and CPU performance of large scale Monte Carlo simulations for HEP, medicine or space science.

In this work we report on recent improvements of the EM models and on new validations of EM physics of Geant4. Improvements were made in models of the photoelectric effect, Compton scattering, gamma conversion to electron and muon pairs, fluctuations of energy loss, multiple scattering, synchrotron radiation, and high energy positron annihilation. The results of these developments are included in the new Geant4 version 10.1 and in patches to previous versions 9.6 and 10.0 that are planned to be used for production for run-2 at LHC.

The Geant4 validation suite for EM physics has been extended and new validation results will be shown in this work. In particular, the effect of gamma-nuclear interactions on EM shower shape at LHC energies will be discussed.

85

The CERN Lync system : voice-over-ip and much more

Author: Pawel Grzywaczewski[1]

Co-author: Bruno Silva De Sousa [1]

[1] *CERN*

Corresponding Author: bruno.sousa@cern.ch

While travelling, we expect to have access to Internet, or being able to check a mailbox. But until recently, it was difficult to maintain voice conversations while outside of your working place. For some cases we can use mobile phones but the roaming charges are high when abroad. At CERN we have deployed Lync, a Voice-over-IP system that fills this gap. Once a CERN user has requested a migration of a landline phone number to the Lync Phone system, people can place calls from range of supported clients: Lync IP phones, Lync application on Windows/MAC, or Lync application on a smartphones (iPhone, Android, Windows Phone). This reduces costs and allows people to place phone calls using their CERN number from any place around the world.

This talk will explain the new functionalities such as instant messaging, presence detection, desktop sharing and integration with the CERN phone system, the technical architecture and the experiences with deploying the Lync solution at CERN. Lync Federation with other organisations and bridging to Skype and Google Talk will also be described.

Track 7 Session / 86

Benchmarking and accounting for the (private) cloud

Author: Ulrich Schwickerath[1]

Co-authors: Janos Daniel Pek [1]; Jerome Belleman [1]

[1] *CERN*

Corresponding Author: ulrich.schwickerath@cern.ch

As part of CERN's Agile Infrastructure project, large parts of the CERN batch farm have been moved to virtual machines running on CERNs private IaaS (link is external) cloud. During this process a large fraction of the resources, which had previously been used as physical batch worker nodes, were converted into hypervisors. Due to the large spread of the per-core performance (rated in HS06) in the farm, caused by its heterogenious nature, it is necessary to have a good knowledge of the performance of the virtual machines. This information is used both for scheduling and accounting. While in the previous setup worker nodes were classified and benchmarked based on the purchase order number, for virtual batch worker nodes this is no longer possible; the information is now either hidden or hard to retrieve. Therefore we developed a new scheme to classify worker nodes in terms of their performance. The new scheme is flexible enough to be usable both for virtual and physical machines in the batch farm. It should be possible to apply it as well to public clouds and more dynamic future batch farms with worker nodes coming and going at a high rate.

The used scheme, experiences and lessons learned will be presented. Possible extensions as well as application to a more general case, for example in the context of accounting within WLCG, will be covered.

Track 8 Session / 87

Fast TPC online tracking on GPUs and asynchronous data-processing in the ALICE HLT to enable online calibration

Author: David Michael Rohr[1]

Co-authors: Mikolaj Krzewicki [1]; Sergey Gorbunov [1]; Volker Lindenstruth [1]

[1] *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, which is today the most powerful particle accelerator worldwide. The High Level Trigger (HLT) is an online compute farm of about 200 nodes, 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 several subdetectors that are sensitive to calibration, e.g. the TPC (Time Projection Chamber). For a precise reconstruction, the HLT has to perform the calibration online. Online-calibration can make certain offline calibration steps obsolete and can thus speed up offline analysis. In ALICE Run 3 starting in 2020, online calibration becomes a necessity. The main detector used for track reconstruction is the TPC. Reconstructing the trajectories in the TPC is the most compute-intense step during event reconstruction. Therefore, a fast tracking implementation is of great importance. Reconstructed TPC tracks build the basis for the calibration making a fast online-tracking mandatory.

We present several components developed for the ALICE High Level Trigger to perform fast event reconstruction and to provide features required for online calibration.

As first topic, we present our TPC tracker, which employs GPUs to speed up the processing, and which bases on a Cellular Automaton and on the Kalman filter. Our TPC tracking algorithm has been successfully used in 2011 and 2012 in the lead-lead and the proton-lead runs. We have improved it to leverage features of newer GPUs and we have ported it to support OpenCL, CUDA, and CPUs with a single common source code. This makes us vendor independent.

As second topic, we present framework extensions, which are required for online calibration. The extensions, however, are generic and can be used for other purposes as well. We have extended the framework to allow asynchronous compute chains, which are required for long-running tasks e.g. for online calibration. And we describe our method to feed in custom data sources in the data flow. This can be external parameters like environmental temperature required for calibration and this can also be used to feed back calibration results in the processing chain.

Overall, the work presented in this contribution makes the ALICE HLT ready for online reconstruction and calibration for the LHC Run 2 starting in 2015.

Track 6 Session / 88

Experience of public procurement of Open Compute servers

Author: Olof Barring[1]

Co-authors: Alain Gentit [1]; Alexandru Grigore [1]; Anthony Grossir [1]; Benoit Clement [1]; Eric Bonfillou [1]; Liviu Valsan [1]; Marco Guerri [1]; Vincent Dore [1]

[1] *CERN*

Corresponding Author: olof.barring@cern.ch

The Open Compute Project, OCP (http://www.opencompute.org/ (link is external)), was launched by Facebook in 2011 with the objective of building efficient computing infrastructures at lowest possible cost. The technologies are released as open hardware, with the goal to develop servers and data centers following the model traditionally associated with open source software projects. In 2013 CERN acquired a few OCP servers in order to compare performance and power consumption with standard hardware. The conclusions were that there are sufficient savings to motivate an attempt to procure a large scale installation. One objective is to evaluate if the OCP market is sufficiently mature and broad to meet the constrains of a public procurement. This talk will summarize this

procurement, which started in September 2014 and involved the Request for information (RFI) to qualify bidders and Request for Tender (RFT).

89

LHCb Build and Deployment Infrastructure for RUNII

Authors: Ben Couturier[1]; Marco Clemencic[1]

[1] *CERN*

Corresponding Author: ben.couturier@cern.ch

After the successful RUN I of the LHC, the LHCb Core software team has taken advantage of the long shutdown to consolidate and improve its build and deployment infrastructure. Several of the related projects have already been presented like the build system using Jenkins, as well as the LHCb Performance and regression testing infrastructure. Some components are completely new, like the Software Configuration Database (using the Graph DB Neo4j), or the new packaging installation using RPM packages. Furthermore all those parts are integrated to allow easier and quicker releases of the LHCb Software stack, therefore reducing the risk of operational errors. Integration and Regression tests are also now easier to implement, allowing to improve further the software checks.

In this poster we describe the various components of the infrastructure and how they fit together to finally allow nearly automated deployment.

Track 6 Session / 91

The GÉANT network: addressing current and future needs for the High Energy Physics community.

Authors: Mian Usman[1]; Vincenzo Capone[1]

[1] *DANTE*

Corresponding Author: vincenzo.capone@dante.net

The GÉANT infrastructure is the backbone that serves the scientific communities in Europe for their data movement needs and their access to international research and education networks. Using the extensive fibre footprint and infrastructure in Europe the GÉANT network delivers a portfolio of services aimed to best fit the specific needs of the users, including Authentication and Authorization Infrastructure, end-to-end performance monitoring, advanced network services (dynamic circuits, L2-L3VPN, MD-VPN).

This talk will outline the factors that help the GÉANT network to respond to the needs of the High Energy Physics community, both in Europe and worldwide.

The Pan-European network provides the connectivity between 40 European national research and education networks. In addition, GÉANT also connects the European NRENs to the R&E networks in other world region and has reach to over 110 NREN worldwide, making GÉANT the best connected Research and Education network, with its multiple intercontinental links to different continents e.g. North and South America, Africa and Asia-Pacific.
The High Energy Physics computational needs have always had (and will keep having) a leading role among the scientific user groups of the GÉANT network: the LHCONE overlay network has been built, in collaboration with the other big world REN, specifically to address the peculiar needs of the LHC data movement. Recently, as a result of a series of coordinated efforts, the LHCONE network has been expanded to the Asia-Pacific area, and is going to include some of the main regional R&E

network in the area.
The LHC community is not the only one that is actively using a distributed computing model (hence the need for a high-performance network); new communities are arising, as BELLE II. GÉANT is deeply involved also with the BELLE II Experiment, to provide full support to their distributed computing model, along with a Perfsonar-based network monitoring system. GÉANT has also coordinated the setup of the network infrastructure to perform the BELLE II Trans-Atlantic Data Challenge, and has been active on helping the BELLE II community to sort out their end-to-end performance issues.

In this talk we will provide information about the current GÉANT network architecture and of the international connectivity, along with the upcoming upgrades and the planned and foreseeable improvements. We will also describe the implementation of the solutions provided to support the LHC and BELLE II experiments.

92

ATLAS computing on the HPC Piz Daint machine

Author: Sigve Haug[1]

Co-authors: Andrej Filipcic [2]; Michi Hostettler [1]; Rodney Walker [3]

[1] *Universitaet Bern (CH)*

[2] *Jozef Stefan Institute (SI)*

[3] *Ludwig-Maximilians-Univ. Muenchen (DE)*

Corresponding Author: michi.hostettler@cern.ch

The Piz Daint Cray XC30 HPC system at CSCS, the Swiss National Supercomputing centre, is in 2014 the highest ranked European system on TOP500, also featuring GPU accelerators. Event generation and detector simulation for the ATLAS experiment has been enabled for this machine. We report on the technical solutions, performance, HPC policy challenges and possible future opportunities for HEP on extreme HPC systems. In particular a custom made integration to the ATLAS job submission system has been developed via the Advanced Resource Connector (ARC) middleware. Further, some GPU acceleration of the GEANT4 detector simulations were implemented to justify the allocation request for this machine.

93

SWATCH: common control SW for the uTCA-based upgraded CMS L1 Trigger

Authors: Alessandro Thea[1]; Carlos Ghabrous Larrea[2]; Christos Lazaridis[2]; Ivan Amos Cali[3]; Jim Brooke[4]; Karol Bunkowski[5]

[1] *STFC - Rutherford Appleton Lab. (GB)*

[2] *University of Wisconsin (US)*

[3] *Massachusetts Inst. of Technology (US)*

[4] *University of Bristol (GB)*

[5] *University of Warsaw (PL)*

Corresponding Author: carlos.ghabrous@cern.ch

The CMS (Compact Muon Solenoid) L1 (Level-1) Trigger electronics are composed of a large number of different cards based on the VMEBus standard. The majority of the system is being replaced to adapt the trigger to the higher collision rates the LHC will deliver after the LS1 (Long Shutdown 1),

the first phase on the CMS upgrade program. As a consequence, the software that controls, monitors and tests the hardware will need to be re-written.

The upgraded trigger will consist of a set of general purpose boards of similar technology that follow the uTCA specification, thus resulting in a more homogeneous system. A great effort has been made to identify the common firmware blocks and components shared across different cards, regardless of the role they play within the trigger data path. A similar line of work has been followed in order to identify all possible common functionalities in the control software, as well as in the database where the hardware initialisation and configuration data are stored. This will not only increase the homogeneity on the software and database sides, but it will also reduce the manpower needed to accommodate the online SW to the changes on hardware.
Due to the fact that the upgrade will take place in different stages, it has been taken into consideration that these new components had to be integrated in the current SW framework.

This paper presents the design of the control SW and configuration database for the upgraded L1 Trigger.

94

FPGAs go wireless

Author: Alberto Aloisio[1]

Co-author: Raffaele Giordano [1]

[1] *Universita' di Napoli Federico II and INFN*

Corresponding Author: aloisio@na.infn.it

We present a feasibility study of a RF transmitters and modulators
based on parametric softcores fully embedded in a general purpose FPGA fabric,
without using external components.
This architecture aims at providing wireless physical layers
to I-o-T and NFC protocols with programmable hardware.
We show preliminary results with latest generation 7-series XILINX FPGA.

Track 2 Session / 95

Accelerating Scientific Analysis with SciDB

Author: Lisa Gerhardt[1]

Co-authors: Carlos Faham [1]; Yushu Yao [1]

[1] *LBNL*

Corresponding Author: lgerhardt@lbl.gov

SciDB is an open-source analytical database for scalable complex analytics on very large array or multi-structured data from a variety of sources, programmable from Python and R. It runs on HPC, commodity hardware grids, or in a cloud and can manage and analyze terabytes of array-structured data and do complex analytics in-database.
We present an overall description of the SciDB framework and describe its implementation at NERSC at Lawrence Berkeley National Laboratory. A case study using SciDB to analyze data from the LUX dark matter detector is described. LUX is a 370 kg liquid xenon time-projection chamber built to directly detect galactic dark matter in an underground laboratory 1 mile under the Black Hills in South Dakota, USA. In the 2013 initial data run, LUX collected 86 million events and wrote 32 TB of data of which only 160 events are retained for final analysis. The data rate for the new dark

matter run starting in 2014 is expected to exceed 250 TB / year. We describe how SciDB is used to dramatically streamline the data collection and analysis, and discuss future plans for a large parallel SciDB array at NERSC.

Track 4 Session / 96

A history-based estimation for LHCb job requirements

Author: Nathalie Rauschmayr[1]

[1] *CERN*

Corresponding Author: nathalie.rauschmayr@cern.ch

The main goal of a Workload Management System (WMS) is to find and allocate resources for the jobs it is handling. The more and more accurate information the WMS receives about the jobs, the easier it will be to accomplish its task, which will directly translate into a better utilization of resources. Traditionally, the information associated with each job, like expected runtime or memory requirement, is in the best case defined at submission time by the Production Manager or fixed by default to arbitrary conservative values. In the case of LHCb's Workload Management System, no mechanisms are provided that automatize the estimation of job requirements. As a result, in order to be conservative, much more CPU time is normally requested than actually needed. Particularly, in the context of multicore jobs this represents a major problem, since single- and multi-core jobs shall share the same resources. Therefore, in order to allow an optimization of the available resources, an accurate estimation of the necessary resources is required. As the main motivation for going to multicore jobs is the reduction of the overall memory footprint, the memory requirement of the jobs should also be correctly estimated.

A detailed workload analysis of past LHCb jobs will be presented. It includes a study of which job features have a correlation with runtime and memory consumption. Based on these features, a supervised learning algorithm has been developed relying on a history-based prediction. The aim is to learn over time how jobs' runtime and memory evolve due to changes in the experimental conditions and the software versions. It will be shown that this estimation can be notably improved if the experimental conditions are taken into account.

97

Background elimination using the SNIP method for Bragg reflections from a protein crystal measured by a time-of-flight single-crystal neutron diffractometer

Author: Katsuaki Tomoyori[1]

Co-authors: Kazuo Kurihara [1]; Taro Tamada [1]

[1] *Japan Atomic Energy Agency*

Corresponding Author: tomoyori.katsuaki@jaea.go.jp

In neutron protein crystallography, it should be also emphasized that the weak Bragg reflections due to the large unit cells may be buried beneath the strong background caused by the incoherent scattering of hydrogen atoms. Therefore, the background estimation from the source is more reliable to improve the accuracy of Bragg integral intensity. We propose the adoption of Statistics-sensitive Nonlinear Iterative Peak-clipping (SNIP) algorithm for background estimation 1, which can eliminate the background from the source spectrum as well as statistically enhance low peaks. In addition, it was recently reported that the Landau and Vavilov distributions, which are used to describe the energy loss of charged particles traversing a thin absorber were found to be in excellent agreement

with the observed TOF profile 2. I show that it may be beneficial to establish a profile-fitting method with the combined use of the Landau/Vavilov functions to faithfully reproduce the TOF peak shape and a SNIP background evaluation algorithm to eliminate statistical fluctuations.
1 Ryan CG et al (1988) Nucl. Instr. Meth. Phys. Res. Sect. B 34:396-402
2 Tomoyori K et al (2013) Nucl. Instr. Meth. Phys. Res. Sect. A 723:128-135

99

ATLAS Monte Carlo production Run-1 experience and readiness for Run-2 challenges

Authors: Andreu Pacheco Pages[1]; Jose Enrique Garcia Navarro[2]

Co-authors: Alexandre Vaniachine [3]; Claire Gwenlan [4]; James Ferrando [5]; Jiahang Zhong [4]; John Derek Chapman [6]; Sascha Mehlhase [7]; Vakho Tsulaia [8]

[1] *Institut de Física d'Altes Energies - Barcelona (ES)*

[2] *Instituto de Fisica Corpuscular (ES)*

[3] *ATLAS*

[4] *University of Oxford (GB)*

[5] *University of Oxford*

[6] *University of Cambridge (GB)*

[7] *Ludwig-Maximilians-Univ. Muenchen (DE)*

[8] *Lawrence Berkeley National Lab. (US)*

Corresponding Author: pacheco@ifae.es

In this presentation we will review the ATLAS Monte Carlo production setup including the different production steps involved in full and fast detector simulation. A report on the Monte Carlo production campaigns during Run-I, Long Shutdown 1 (LS1) and status of the production for Run-2 will be presented. The presentation will include the details on various performance aspects. Important improvements in the workflow and software will be highlighted.

Besides standard Monte Carlo production for data analyses at 7 and 8 TeV, the production accommodates for various specialised activities. These range from extended Monte Carlo validation, Geant4 validation, pileup simulation using zero bias data and production for various upgrade studies. The challenges of these activities will be discussed.

100

Scaling up ATLAS production system for the LHC Run 2 and beyond: project ProdSys2

Author: Alexandre Vaniachine[1]

Co-authors: Alexei Klimentov [2]; Jose Enrique Garcia Navarro [3]; Kaushik De [4]; Misha Borodin [5]; Tadashi Maeno [2]

[1] *ATLAS*

[2] *Brookhaven National Laboratory (US)*

[3] *Instituto de Fisica Corpuscular (ES)*

[4] *University of Texas at Arlington (US)*

[5] *National Research Nuclear University MEPhI (RU)*

Corresponding Author: alexei.klimentov@cern.ch

The Big Data processing needs of the ATLAS experiment grow continuously, as more data and more use cases emerge. For Big Data processing the ATLAS experiment adopted the data transformation approach, where software applications transform the input data into outputs. In the ATLAS production system, each data transformation is represented by a task, a collection of many jobs, submitted by the ATLAS workload management system (PanDA) and executed on the Grid.

Our experience shows that the rate of tasks submission grows exponentially over the years. To scale up the ATLAS production system for new challenges, we started the ProdSys2 project. PanDA has been upgraded with the Job Execution and Definition Interface (JEDI). Patterns in ATLAS data transformation workflows composed of many tasks, provided a scalable production system framework for template definitions of the many-tasks workflows. These workflows are being implemented in the Database Engine for Tasks (DEfT) that generates individual tasks for processing by JEDI.

We report on the ATLAS experience with many-task workflow patterns in preparation for the LHC Run-2.

Track 8 Session / 101

Evaluating the power efficiency and performance of multi-core platforms using HEP workloads

Author: Pawel Szostek[1]

Co-author: Vincenzo Innocente [1]

[1] *CERN*

Corresponding Author: pawel.szostek@cern.ch

As Moore's Law drives the silicon industry towards higher transistor counts, processor designs are becoming more and more complex. The area of development includes core count, execution ports, vector units, uncore architecture and finally instruction sets. This increasing complexity leads us to a place where access to the shared memory is the major limiting factor, making feeding the cores with data a real challenge. On the other hand, the significant focus on power efficiency paves the way for power-aware computing and less complex architectures to data centers. In this paper we try to examine these trends and present results of our experiments with "Haswell-EP" processor family and highly scalable HEP workloads.

102

Implementation and use of an highly available and innovative IaaS solution: the Cloud Area Padovana

Author: Alvise Dorigo[1]

Co-authors: Alberto Crescente [2]; Doina Cristina Aiftimiei [3]; Eric Frizziero [2]; Federica Fanzago [4]; Fulvia Costa [2]; Lisa Zangrando [1]; Marco Verlato [1]; Massimo Biasotto [5]; Massimo Sgaravatto [1]; Massimo Venaruzzo [1]; Michele Gulmini [1]; Michele Michelotto [1]; Paolo Andreetto [2]; Sergio Fantinel [2]; Sergio Traldi [2]; Stefano Dal Pra [2]; sara bertocco [6]

[1] *Universita e INFN (IT)*

[2] *INFN*

[3] *INFN, on leave of absence from "Horia Hulubei" National Institute for Physics and Nuclear Engineering (IFIN-HH)*

[4] *Unknown*

[5] *INFN Legnaro*

[6] *INFN-PD*

Corresponding Author: stefano.dalpra@cnaf.infn.it

While in the business world the cloud paradigm is typically implemented purchasing resources and services from third party providers (e.g. Amazon), in the scientific environment there's usually the need of on-premises IaaS infrastructures which allow efficient usage of the hardware distributed among (and owned by) different scientific administrative domains. In addition, the requirement of open source adoption has leaded to the choice of products like OpenStack by many organizations.

We describe a use case of the Italian National Institute for Nuclear Physics (INFN) which resulted in the implementation of a unique Cloud service, called "Cloud Area Padovana", which encompasses resources spread in two different sites: the INFN Legnaro National Laboratories and the INFN Padova division.
We describe how this IaaS has been implemented, which technologies have been adopted, how services have been configured in high-availability (HA) mode.

We also discuss how identity and authorization management were implemented, adopting a widely accepted standard architecture based on SAML2 and OpenID: by leveraging the versatility of those standards the integration with authentication federations like IDEM was implemented.

We also discuss some other innovative developments, such as a pluggable scheduler, implemented as extension of the native OpenStack scheduler, which allows to allocate resources according to a fair-share based model and which provides a persistent queuing mechanism for handling user requests that can not be immediately served.

Tools, technologies, procedures used to install, configure, monitor, operate this Cloud service are also discussed.

Finally we present some examples that show how this IaaS infrastructure is being used.

103

SModelS: a framework to confront theories beyond the standard model with experimental data

Authors: Andre Lessa[1]; Michael Traub[2]; Sabine Kraml[3]; Suchita Kulkarni[4]; Ursula Laa[4]; Veronika Magerl[4]; Wolfgang Magerl[None]; Wolfgang Waltenberger[4]

[1] *IFGW - UNICAMP*

[2] *HEPHY Vienna*

[3] *Centre National de la Recherche Scientifique (FR)*

[4] *Austrian Academy of Sciences (AT)*

Corresponding Author: walten@cern.ch

We present a general procedure to decompose Beyond the Standard Model (BSM) collider signatures into Simplified Model Spectrum (SMS) topologies. Our method provides a way to cast BSM predictions for the LHC in a model independent framework, which can be directly confronted with the relevant experimental constraints. Our python implementation currently focusses on supersymmetry searches with missing energy, for which a large
variety of SMS results from ATLAS and CMS are available.
We have collected these results into a database, enabling us to quickly confront any BSM theory with about 200 simplified models results.
Originating in the LHC world we wish to extend beyond hadron colliders, exploiting also the information about BSM physics contained in precision measurements, or dark matter searches.
As show-case examples we will discuss application of our procedure to one specific supersymmetric model, show how the limits constrain the model, and point out regions in parameter space still

unchallenged by the current SMS results.
Given experimental hints at new physics, it is our ultimate goal to build the next standard model in a bottom-up fashion from the experimental SMS results
from several experiments.

Track 7 Session / 104

CMS@Home: Enabling Volunteer Computing Usage for CMS

Author: Laurence Field[1]

Co-authors: Daniele Spiga [1]; Hassen Riahi [1]; Hendrik Borras [2]

[1] *CERN*

[2] *Heidelberg University*

Corresponding Author: laurence.field@cern.ch

Volunteer computing remains an untapped opportunistic resource for the LHC experiments. The use of virtualization in this domain was pioneered by the Test4theory project and enabled the running of high-energy particle physics simulations on home computers. This paper describes the model for CMS to run workloads using a similar volunteer computing platform. It is shown how the original approach is explored to map onto the existing CMS workflow and identifies missing functionality along with the components and changes that are required. The final implementation of the prototype is detailed along with the identification of areas that would benefit from further development.

Track 6 Session / 105

LHCOPN and LHCONE: Status and Future Evolution

Authors: Edoardo Martelli[1]; Stefan Nicolae Stancu[1]

[1] *CERN*

Corresponding Author: stefan.stancu@cern.ch

The LHC Optical Private Network, linking CERN and the Tier1s and the LHC Open Network Environment linking these to the Tier2 community successfully supported the data transfer needs of the LHC community during Run 1 and have evolved to better serve the networking requirements of the new computing models for Run 2. We present here the current status and the key changes, notably the delivered and planned bandwidth increases, the ongoing work to better address the needs of the Asia-Pacific region, developments to improve redundancy provision and progress made for the provision of point-to-point links.

107

Investigating machine learning to classify events in CMS

Author: Maria Girone[1]

[1] *CERN*

Corresponding Author: maria.girone@cern.ch

High energy physics experiments are experiencing a growth in the number of collected and processed events that exceeds the rate of growth in computing resources sustainable by technology improvements at a flat yearly cost. This trend is expected to continue into the foreseeable future, and as the field is not expecting a big increase in support, innovative approaches are needed. In areas of science as diverse as biology and cosmology groups are deploying advanced machine learning techniques to learn to classify events based on rules and trained outcomes. High energy physics has used similar decision techniques at the analysis step, and has tried a few specific demonstrations as challenges. In this presentation we present the early exploratory work in CMS to utilize machine learning techniques for real-time event classification for triggering and analysis. The early performance and next steps will be outlined.

Track 1 Session / 108

The LHCb Data Aquisition and High Level Trigger Processing Architecture

Author: Markus Frank[1]

Co-authors: Beat Jost [1]; Clara Gaspar [1]; Niko Neufeld [1]

[1] *CERN*

Corresponding Author: markus.frank@cern.ch

The LHCb experiment at the LHC accelerator at CERN collects collisions of particle bunches at 40 MHz. After a first level of hardware trigger with output of 1 MHz, the physically interesting collisions are selected by running dedicated trigger algorithms in the High Level Trigger (HLT) computing farm. This farm consists of up to roughly 25000 CPU cores in roughly 1600 physical nodes each equipped with 2 TB of local storage space.
This work describes the LHCb online system with an emphasis on the developments implemented during during the current long shutdown (LS1). We will elaborate the architecture to treble the available CPU power of the HLT farm and the technicalities to determine and verify precise calibration and alignment constants which are fed to the HLT event selection procedure. Precise calibration and alignment constants are determined and verified in a separate data acquisition activity as soon as data from particle collisions are delivered by the LHC collider. We will describe how the constants are fed into a two stage HLT event selection facility using extensively the local disk buffering capabilities on the worker nodes. With the installed disk buffers, the installed CPU can be used during periods of up to ten days without beams. These periods in the past accounted to more than 70 % of the total time.

Track 6 Session / 109

Monitoring WLCG with lambda-architecture: a new scalable data store and analytics platform for monitoring at petabyte scale.

Author: Luca Magnoni[1]

Co-authors: Akram Khan [2]; David Smith [2]; Julia Andreeva [1]; Uthayanath Suthakar [2]

[1] *CERN*

[2] *Brunel University*

Corresponding Author: luca.magnoni@cern.ch

Monitoring the WLCG infrastructure requires to gather and to analyze high volume of heterogeneous data (e.g. data transfers, job monitoring, site tests) coming from different services and experiment-specific frameworks to provide a uniform and flexible interface for scientists and sites. The current architecture, where relational database systems are used to store, to process and to serve monitoring data, has limitations in coping with the foreseen extension of the volume (e.g. higher LHC luminosity) and the variety (e.g. new data-transfer protocols and new resource-types, as cloud-computing) of WLCG monitoring events. This paper presents a new scalable data store and analytics platform designed by the Support for Distributed Computing (SDC) group, at the CERN IT department, which leverages on a stack of technology each one targeting specific aspects on big-scale distributed data-processing (commonly referred as lambda-architecture approach). Results on data processing on Hadoop for WLCG data transfers are presented, showing how the new architecture can easily analyze hundreds of millions of transfer logs in few minutes. Moreover, a comparison on data partitioning, compression and file format (e.g. CSV, AVRO) is presented, with particular attention on how the file structure impacts the overall MapReduce performance. In conclusion, the evolution of the current implementation, which focuses on data store and batch processing, towards a complete lambda-architecture is discussed, with consideration on candidate technology for the serving layer (e.g. ElasticSearch) and a description of a proof of concept implementation, based on Esper, for the real-time part which compensates for batch-processing latency and automates problem detection and failures.

110

Extending software repository hosting to code review and testing

Authors: Alexandre Lossent[1]; Alvaro Gonzalez Alvarez[1]; Borja Aparicio Cotarelo[1]; Terje Andersen[1]

Co-authors: Anna Trzcinska [2]; David Asbury [1]; Helge Meinhard [1]; Nils Hoimyr [1]

[1] *CERN*

[2] *Warsaw University of Technology (PL)*

Corresponding Author: helge.meinhard@cern.ch

We will present how CERN's services around Issue Tracking and Version Control have evolved, and what the plans for the future are. We will describe the services' main design, integration and structure, giving special attention to the new requirements from the community of users in terms of collaboration and integration tools and how we address this challenge when defining new services based on GitLab for collaboration and Code Review and Jenkins for Continuous Integration. These new services complement the existing ones to create a new global "development service".

111

Monitoring the Delivery of Virtualized Resources to the LHC Experiments

Author: Cristovao Jose Domingues Cordeiro[1]

Co-authors: Alessandro Di Girolamo [1]; Daniele Spiga [1]; Domenico Giordano [1]; Laurence Field [1]; Luis Villazon Esteban [2]

[1] *CERN*

[2] *Universidad de Oviedo (ES)*

Corresponding Author: cristovao.cordeiro@cern.ch

The adoption of Cloud technologies by the LHC experiments places the fabric management burden of monitoring virtualized resources upon the VO. In addition to monitoring the status of the virtual machines and triaging the results, it must be understood if the resources actually provided match with any agreements relating to the supply. Monitoring the instantiated virtual machines is therefore a fundamental activity and hence this paper describes how the Ganglia monitoring system can be used within the cloud computing scope of the LHC experiments. Extending upon this, it is then shown how the integral of the time-series monitoring data obtained can be repurposed to provide a consumer-side accounting record, which can then be compared with the concrete agreements that exist between the supplier of the resources and the consumer. From this, it is not clear though how the performance of the resources differ both within and between providers. Hence, the case is made for a benchmarking metric to normalize the results along with some preliminary investigation on obtaining such a metric.

Track 4 Session / 112

Jobs masonry with elastic Grid Jobs

Author: Federico Stagni[1]

Co-author: Philippe Charpentier [1]

[1] *CERN*

Corresponding Author: federico.stagni@cern.ch

The DIRAC workload management system used by LHCb Distributed Computing is based on Computing Resource reservation and late binding (also known as pilot job in the case of batch resources) that allows the serial execution of several jobs obtained from a central task queue. CPU resources can usually be reserved for limited duration only (e.g. batch queue time limit) and in order to optimize their usage, it is important to be able to use them for the whole available time. However traditionally the tasks to be performed by jobs are defined at submission time and therefore it may happen that no job fits in the available time.
In LHCb, this so-called job masonry is optimized by the usage of elastic simulation jobs: unlike data processing jobs that must process all events in the input dataset, simulation jobs offer an interesting degree of freedom as one can define the number of events to be simulated even after submission. This requires however knowing three information: the time available in the reserved resource, its CPU power and the average CPU work required for simulating one event. The decision on the number of events can then be made at the very last moment, just before starting the simulation application. This is what we call elastic jobs.
LHCb simulation jobs are now all elastic, with an upper limit on the number of events per job. When several jobs are needed to complete a simulation request, enough jobs are submitted for simulating the total number of events required, assuming this upper limit for each job. New jobs are then submitted depending on the actual number of events simulated by this first batch of jobs.
We will show in this contribution how elastic jobs allow a better backfilling of computing resources as well as using resources with limited work capacity, such as short batch queues or volunteer computing resources. They also allow easily to shutdown virtual machines on cloud resources when sites require them to shutdown within a grace period.

Track 4 Session / 113

Pilots 2.0: DIRAC pilots for all the skies

Author: Federico Stagni[1]

Co-authors: Andrei Tsaregorodtsev [2]; Andrew McNab [3]; Cinzia Luzzi [1]

[1] *CERN*

[2] *CPPM, Aix-Marseille Université, CNRS/IN2P3, Marseille, France*

[3] *University of Manchester (GB)*

Corresponding Author: federico.stagni@cern.ch

In the last few years, new types of computing infrastructures, such as IAAS (Infrastructure as a Service) and IAAC (Infrastructure as a Client), gained popularity. New resource may come as part of pledged resources, while others are in the form of opportunistic ones. Most of these new infrastructures are based on virtualization techniques, others don't. Meanwhile, some concepts, such as distributed queues, lost appeal, while still supporting a vast amount of resources. Virtual Organizations are therefore facing heterogeneity of the available resources and the use of an Interware software like DIRAC to hide the diversity of underlying resources has become essential.
The DIRAC WMS is based on the concept of pilot jobs that was introduced back in 2004. A pilot is what creates the possibility to run jobs on a worker node. The advantages of the pilot job concept are now well established. The pilots are not only increasing the visible efficiency of the user jobs but also help to manage the heterogeneous computing resources presenting them to the central services in a uniform coherent way. Within DIRAC, we developed a new generation of pilot jobs, that we dubbed Pilots 2.0. Pilots 2.0 are not tied to a specific infrastructure; rather they are generic, fully configurable and extendible pilots. A Pilot 2.0 can be sent, as a script to be run, or it can be fetched from a remote location. A pilot 2.0 can run on every computing resource, e.g.: on CREAM Computing elements, on DIRAC Computing elements, on Virtual Machines as part of the contextualization script, or IAAC resources, provided that these machines are properly configured, hiding all the details of the WN's infrastructure. Pilots 2.0 can be generated server and client side. Pilots 2.0 are the "pilots to fly in all the skies", aiming at easy use of computing power, in whatever form it is presented. Another aim is the unification and simplification of the monitoring infrastructure for all kind of computing resources by using pilots as a network of distributed sensors coordinated by a central resource monitoring system.
Pilots 2.0 have been developed using the command pattern: each command is realizing an atomic function, and can be easily activated and de-activated based on the WN type. VOs using DIRAC can tune pilots 2.0 as they need, and extend or replace each and every pilot command in an easy way. In this paper we describe how Pilots 2.0 work with distributed and heterogeneous resources providing the abstraction necessary to deal with different kind of computing resources.

114

FTS3 on the Web

Author: Alejandro Alvarez Ayllon[1]

Co-authors: Andrea Manzi [1]; Andres Abad Rodriguez [1]; Andrey Kiryanov [2]; Christina Skarpathiotaki [3]; Markus Schulz [1]; Michail Salichos [1]; Michal Kamil Simon [1]; Oliver Keeble [1]

[1] *CERN*

[2] *B.P. Konstantinov Petersburg Nuclear Physics Institute - PNPI (*

[3] *National Technical Univ. of Athens (GR)*

Corresponding Author: alejandro.alvarez.ayllon@cern.ch

FTS3 is the service responsible for the distribution of the LHC data across the WLCG Infrastructure. To facilitate its use outside the traditional grid environment we have provided a web application - known as WebFTS - fully oriented towards final users, and easily usable within a browser.

This web application is completely decoupled from the core service, and interfaces with it via a REST API. Following the principle of maximum reusability, we have recently developed a pluggable OAuth2 provider for the FTS3 REST interface, so any third party can now use FTS3 as a framework for building their own transfer-based applications, effectively outsourcing any concern about delegation, scheduling or underlying protocol details.

Here we will describe how FTS3 and WebFTS fit together using this approach, and how the same interface could be easily reused by external developers to create their own web services relying on

FTS3 as a gateway between the Grid, the Web and the Cloud. We will also provide the outline for some potentially interesting uses of this idea, for example a synchronization mechanism between storage elements.

Track 3 Session / 115

Studies of Big Data meta-data segmentation between relational and non-relational databases

Authors: Maria Grigorieva[1]; Marina Golosova[1]

Co-authors: Alexei Klimentov [2]; Eygene Ryabinkin [1]; Gancho Dimitrov [3]; Maxim Potekhin [2]

[1] *National Research Centre "Kurchatov Institute"*

[2] *Brookhaven National Laboratory*

[3] *CERN*

Corresponding Author: golosova.marina@gmail.com

In recent years the concepts of Big Data became well established in IT-technologies. Most systems (for example Distributed Data Management or Workload Management systems) produce metadata that describes actions performed on jobs, stored data or other entities and its volume takes one to the realms of Big Data on many occasions. This metadata can be used to obtain information about the current system state, the aggregation of data for summary purposes and for statistical and trend analysis of the processes this system drives. The latter requires metadata to be stored for a long period of time. On the example of PanDA (Workload Management System for distributed production and analysis for the ATLAS experiment at the LHC and astro-particle experiments AMS and LSST) it can be seen that the growth rate of the volume of stored information has increased significantly over the last few years: from 500k completed jobs per day in 2011 up to 2 million nowadays.
Database is the central component of the PanDA architecture. Currently RDBMS (Oracle or MySQL) is used as the storage backend. To provide better performance and scalability the data stored in relational storage is partitioned into actual ("live") and archive (historical) parts. But even in this scheme, as the "archived"data volume grows, the underlying software and hardware stack encounters certain limits that negatively affect processing speed and the possibilities of metadata analysis. We had investigated a new class of database technologies commonly referred to as NoSQL databases. We suggest to use NoSQL solution for finalized, reference portion of essentially read-only data to improve performance and scalability. We had developed and implemented a heterogeneous storage which consists of both relational and non-relational databases and provides an API for unified access to stored meta-data.
We present methods of partitioning the data between two database classes, methods for efficient storage of NoSQL backend for archived data, including the analysis of different indexing schemes based on the statistics of the most frequently used queries to the historical data. We also present a comparison between different NoSQL databases to conclude their applicability to our solution. Performance of "archived"data storage in the analytical tasks is shown in the quantitative scalability and performance test results, including testing NoSQL storage against Oracle RDBMS.
This work is conducted as a part of the project to expand PanDA beyond HEP and LHC experiments.

Track 2 Session / 116

The Data Quality Monitoring Software for the CMS experiment at the LHC

Author: Marco Rovere[1]

Co-authors: Daniel Duggan [2]; Federico De Guio [1]

[1] *CERN*
[2] *Rutgers, State Univ. of New Jersey (US)*

Corresponding Author: marco.rovere@cern.ch

The Data Quality Monitoring (DQM) Software is a central tool in the CMS experiment. Its flexibility allows for integration in several key environments: Online, for real-time detector monitoring; Offline, for the final, fine-grained data analysis and certification; Release-Validation, to constantly validate the functionalities and the performance of the reconstruction software; in Monte Carlo productions. Since the end of data taking at a center of mass energy of 8 TeV, the environment in which the DQM lives has undergone fundamental changes. In turn, the DQM system has made significant upgrades in many areas to respond to not only the changes in infrastructure, but also the growing specialized needs of the collaboration with an emphasis on more sophisticated methods for evaluating data quality, as well as advancing the DQM system to provide quality assessments of various Monte Carlo simulations versus data distributions, monitoring changes in physical effects due to modifications of algorithms or framework, and enabling regression modeling for long-term effects for the CMS detector. The central tool to deliver Data Quality information is an interactive web site for browsing data quality histograms (DQMGUI), and its transition to becoming a distributed system will also be presented. In this contribution the usage of the DQM Software in the different environments and its integration in the CMS Reconstruction Software Framework (CMSSW) and in all production workflows are presented, with emphasis on recent developments and improvement in advance of the LHC restart at 13 TeV. The main technical challenges and the adopted solutions to them will be also discussed with emphasis on functionality and long-term robustness.

Track 2 Session / 117

The Bayesian analysis toolkit: version 1.0 and beyond

Author: Frederik Beaujean[1]

Co-authors: Allen Caldwell [2]; Daniel Greenwald [3]; Daniel Kollar [4]; Kevin Alexander Kroeninger [5]; Oliver Schulz [6]; Stefan Kluth [2]

[1] *LMU Munich*
[2] *Max Planck institute for physics*
[3] *Technical University Munich*
[4] *CERN*
[5] *Technical University Dortmund*
[6] *MPI for Physics, Munich*

Corresponding Author: frederik.beaujean@lmu.de

The Bayesian analysis toolkit (BAT)
is a C++ package centered around Markov-chain Monte Carlo sampling. It is used in analyses of various particle-physics experiments such as ATLAS and Gerda. The software has matured over the last few years to a version 1.0. We will summarize the lessons learned and report on the current developments of a complete redesign targeting multicore and multiprocessor architectures and supporting many more sampling algorithms both built-in and user-supplied.

118

Open access to high-level data and analysis tools in the CMS experiment at the LHC

Authors: Elizabeth Sexton-Kennedy[1]; Thomas Mc Cauley[2]

[1] *Fermi National Accelerator Lab. (US)*

[2] *University of Notre Dame (US)*

Corresponding Author: sexton@fnal.gov

The CMS experiment, in recognition of its commitment to data preservation and open access as well as to education and outreach, has made its first public release of high-level data: up to half of the proton-proton collision data at 7 TeV from 2010 in CMS Analysis Object Data format. CMS has prepared, in collaboration with CERN and the other LHC experiments, an open data web portal based on Invenio. The portal provides access to CMS public data as well as to analysis tools and documentation for the public. The tools include an event display and histogram application that run in the browser. In addition a virtual machine is available which contains a CMS software environment along with XRootD access to the data. Within the virtual machine the public can analyse CMS data; example code is provided. We describe the accompanying tools and documentation and discuss the first experience of data use.

Track 5 Session / 119

The CMS Tier-0 goes Cloud and Grid for LHC Run 2

Author: Dirk Hufnagel[1]

Co-authors: Anthony Tiradani [2]; Luis Contreras [1]; Oliver Gutsche [1]

[1] *Fermi National Accelerator Lab. (US)*

[2] *Fermilab*

Corresponding Author: dirk.hufnagel@cern.ch

In 2015, CMS will embark on a new era of collecting LHC collisions at unprecedented rates and complexity. This will put a tremendous stress on our computing systems. Prompt Processing of the raw data by the Tier-0 infrastructure will no longer be constrained to CERN alone due to the significantly increased resource requirements. In LHC Run 2, we will need to operate it as a distributed system utilizing both the CERN Cloud-based Agile Infrastructure and a significant fraction of the CMS Tier-1 Grid resources. In another big change for LHC Run 2, we will process all data using the multi-threaded framework to deal with the increased event complexity and to ensure efficient use of the resources. This contribution will cover the evolution of the Tier-0 infrastructure and present scale testing results and experiences from the first data taking in 2015.

Track 2 Session / 120

Using the CMS Threaded Framework In A Production Environment

Author: Christopher Jones[1]

[1] *Fermi National Accelerator Lab. (US)*

Corresponding Author: cdj@fnal.gov

During 2014, the CMS Offline and Computing Organization completed the necessary changes to use the CMS threaded framework in the full production environment. Running reconstruction workflows using the multi-threaded framework is a crucial element of CMS' 2015 and beyond production plan. We will briefly discuss the design of the CMS Threaded Framework, in particular how the design affects scaling performance. We will then cover the effort involved in getting both the CMSSW application software and the workflow management system ready for using multiple threads for production. Finally, we will present metrics on the performance of the application and workflow

system as well as the difficulties which were uncovered. We will end with CMS' plans for using the threaded framework to do production for LHC Run 2.

Track 1 Session / 121

Developments and applications of DAQ framework DABC v2

Authors: Joern Adamczewski-Musch[1]; Nikolaus Kurz[1]; Sergey Linev[2]

[1] *GSI*

[2] *GSI DARMSTADT*

Corresponding Author: s.linev@gsi.de

The *Data Acquisition Backbone Core* (*DABC*) is a C++ software framework that can implement and run various data acquisition solutions on Linux platforms. In 2013 version 2 of *DABC* has been released with several improvements. These developments have taken into account lots of practical experiences of *DABC v1* with detector test beams and laboratory set-ups since first release in 2009. The plug-in interfaces for user code implementation, and configuration procedures have been simplified. Internally the framework has been enhanced by means of smart references and object cleanup mechanisms. Moreover, for monitoring and control a http web server, or a proprietary command channel access have been provided.

In May and August 2014, *DABC v2* was applied for production data taking of the *HADES* collaboration's pion beam time at *GSI*. It replaced the functionality of the established *HADES* event builder software *hadaq*, receiving frontend data via multiple Gigabit UDP streams, combining them, and storing them to RFIO tape and local disks. *DABC* was transparently integrated to the existing *HADES* DAQ environment, and the EPICS based control system. Additionally, the *HADES* quality monitoring process was fed online with data samples by a *DABC* "streamserver" socket instead of reading intermediate disk files, improving reaction time and stability. A maximum of 16 eventbuider processes on 4 Linux server machines were processing data of 32 subevent inputs. Each of the eventbuilder processes could be configured to run either dabc or hadaq software. In the first beamtime block just one single quality monitoring node was replaced by dabc; finally in August eventbuilding was completely handled by *DABC*.

GSI standard DAQ systems often use PCIe boards *PEXOR*/KINPEX that receive data via optical *gosip* protocol from various front-end hardware. For *DABC v2* corresponding plug-ins and drivers have been developed to handle reading out such systems on Linux platform. Functionalities of the previous DAQ solution with the established multi-platform framework *MBS* have been re-implemented with *DABC v2* and were compared with the same hardware against *MBS*.

122

CMS Experience with a World-Wide Data Federation

Authors: Brian Paul Bockelman[1]; Carl Vuosalo[2]; Daniel Charles Bradley[2]; Federica Fanzago[3]; Frank Wuerthwein[4]; Igor Sfiligoi[4]; Kenneth Bloom[1]; Marian Zvada[1]; Matevz Tadel[4]; Sridhara Dasu[2]; Tommaso Boccali[5]

[1] *University of Nebraska (US)*

[2] *University of Wisconsin (US)*

[3] *Universita e INFN, Padova (IT)*

[4] *Univ. of California San Diego (US)*

[5] *Universita di Pisa & INFN (IT)*

Corresponding Author: marian.zvada@cern.ch

Over the past three years, the CMS Collaboration has developed the "Any Data, Anytime, Anywhere" technology to make use of a global data federation that is based on the XrootD protocol. The federation is now deployed across virtually all Tier-1 and Tier-2 sites in the CMS distributed computing system. This data federation gives workflows greater flexibility for location of execution, which has benefits at all scales of operation, from individual users accessing specific collision event records to organized production processing. In preparation for the coming LHC data run, CMS has been testing a number of applications of the data federation at increasingly large scales. In this presentation, we will discuss the applications in use and the results of the tests.

Track 7 Session / 123

Enabling opportunistic resources for CMS Computing Operations

Author: Dirk Hufnagel[1]

Co-authors: Anthony Tiradani [2]; Burt Holzman [1]; David Alexander Mason [1]; Krista Larson [1]; Marco Mambelli [3]; Oliver Gutsche [1]; Parag Mhashilkar [4]

[1] *Fermi National Accelerator Lab. (US)*

[2] *Fermilab*

[3] *University of Chicago (US)*

[4] *Fermi National Accelerator Laboratory*

Corresponding Author: dirk.hufnagel@cern.ch

With the increased pressure on computing brought by the higher energy and luminosity from the LHC in Run 2, CMS Computing Operations expects to require the ability to utilize "opportunistic" resources —resources not owned by, or a priori configured for CMS —to meet peak demands. In addition to our dedicated resources we look to add computing resources from non CMS grids, cloud resources, and national supercomputing centers. CMS uses the HTCondor/glideinWMS job submission infrastructure for all its batch processing, so such resources will need to be transparently integrated into its glideinWMS pool. Bosco and parrot wrappers are used to enable access and bring the CMS environment into these non CMS resources. Here we describe our strategy to supplement our native capabilities with opportunistic resources and our experience so far using them.

124

Exploiting Volatile Opportunistic Computing Resources as a CMS User

Authors: Anna Elizabeth Woodard[1]; Charles Nicholas Mueller[1]; Matthias Wolf[1]

Co-authors: Ben Tovar [1]; Douglas Thain [2]; Kevin Patrick Lannon [1]; Mike Hildreth [1]; Patrick Donnelly [1]; Paul Brenner [1]

[1] *University of Notre Dame (US)*

[2] *University of Notre Dame*

Corresponding Authors: anna.elizabeth.woodard@cern.ch, m+cms@sushinara.net

Individual scientists in high energy physics experiments like the Compact Muon Solenoid (CMS) experiment at the Large Hadron Collider require extensive use of computing resources for analysis of massive data sets. The majority of this analysis work is done at dedicated grid-enabled CMS

computing facilities. University campuses offer considerable additional computing resources, but these are not specifically configured to run CMS software. Furthermore, in many cases, the machines are available for general usage whenever they are idle, but opportunistic jobs can be terminated at any time, leading to a highly volatile computing environment.

As a joint effort involving computer scientists and CMS physicists at Notre Dame, we have developed an opportunistic workflow management tool, Lobster, to harvest available cycles from university campus computing pools. The Lobster framework consists of a management server, file server, and workers submitted to any available computing resource. Only standard user permissions are required to run the entire suite of processes, making it possible to use this tool with any resource on which the user has permission to run. Lobster makes use of the Work Queue system to perform task management, while the CMS specific software environment is provided via CVMFS and Parrot. Data is handled via Chirp and Hadoop for local data storage and XrootD for access to the CMS wide-area data federation. An extensive set of monitoring and diagnostic tools have been developed to facilitate system optimisation. The tool has been tested in a variety of environments, including within OSG Connect. We have tested it at large-scales using the 20,000-core cluster at Notre Dame, achieving approximately 8000 tasks running simultaneously, sustaining approximately 9 Gbit/s of input data and 340 Mbit/s of output data.

125

Multimedia Content in the CERN Document Server

Author: Jean-Yves Le Meur[1]

Co-authors: Charalampos Tzovanakis [2]; Ludmila Marian [1]

[1] *CERN*

[2] *National Technical Univ. of Athens (GR)*

Corresponding Author: nikos.kasioumis@cern.ch

The talk will focus on the recent developments done by the Multimedia team of the Digital Library Services to better acquire grand-public content captured at CERN and disseminate it to the general public.

In collaboration with the CERN communication unit and the Photo & Video Labs, the team has built new facilities to transfer, disseminate and archive multimedia content on the CERN Document Server Invenio-based platform. These original user-oriented features will be introduced for both the owners of multimedia material and the people willing to search for content.

Explanations on how new technologies have been used behind the scene will follow, together with the details on how these developments have been carried out within the Invenio Open Source Digital Library community. The evolution of the front-end into a modern multimedia service will be explained as well as some on-going research work on automated face & object recognition, with direct image tagging functionalities. The re-architecture of the back engine to address the long-term preservation needs and the advantages of a new rich paradigm supporting images embedded into albums will be described.

Finally, it will be shown the impressive peaks of use observed during major events and the recent increase in the needs for new collections. The example of the acquisition of 130'000 CERN diapos from the years 1960th and 1970th will give to all attendees the opportunity to discover not only the technical aspects of the process but also an overview of this unique content.

Track 2 Session / 126

CMS Full Simulation for Run-II

Author: Vladimir Ivantchenko[1]

Co-authors: David Lange [2]; Mike Hildreth [3]

[1] *CERN*

[2] *Lawrence Livermore Nat. Laboratory (US)*

[3] *University of Notre Dame (US)*

Corresponding Author: david.lange@cern.ch

This presentation will discuss new features of the CMS simulation for Run 2, where we have made considerable improvements during LHC shutdown to deal with the increased event complexity and rate for Run 2. For physics improvements migration from Geant4 9.4p03 to Geant4 10.0p02 has been performed. CPU performance was improved by introduction of the Russian roulette method inside CMS calorimeters, optimization of CMS simulation sub-libraries, and usage of statics build of the simulation executable. As a result of these efforts, CMS simulation was speeded up by about factor two. In this work we provide description of these updates and discuss different software components of CMS simulation.

Geant4 version 10.0 is multi-threaded capable. This allows development of multi-threaded version of CMS simulation in parallel with mainstream sequential production version. For CMS multi-threaded Geant4 additional modules and manager classes were added. Geometry, magnetic field, physics, user actions, and sensitive detectors classes are the same for both sequential and multi-threaded versions of CMS simulation. In this work we report on details of implementation of CMS multi-threaded simulation including CPU and memory performance.

127

PROOF Analysis Framework (PAF)

Author: Isidro Gonzalez Caballero[1]

Co-authors: Ana Rodriguez Marrero [2]; Enol Fernández [3]; Javier Delgado Fernández [4]

[1] *Universidad de Oviedo (ES)*

[2] *Universidad de Cantabria (ES)*

[3] *IFCA*

[4] *Universidad de Oviedo*

Corresponding Author: isidro.gonzalez.caballero@cern.ch

The PROOF Analysis Framework (PAF) has been designed to improve the ability of the physicist to develop software for the final stages of an analysis where typically simple ROOT Trees are used and where the amount of data used is in the order of several terabytes. It hides the technicalities of dealing with PROOF leaving the scientist to concentrate on the analysis. PAF is capable of using available non specific resources on, for example, local batch systems, remote grid sites or clouds through the integration of other toolkit like PROOF Cluster or PoD. While it has been successfully used on LHC Run-1 data for some key analysis, including the H->WW dilepton channel, the higher instantaneous and integrated luminosity together with the increase of the center-of-mass energy foreseen for the LHC Run-2, which will increment the total size of the samples by a factor 6 to 20, will demand PAF to improve its scalability and to reduce the latencies as much as possible. In this paper we address the possible problems of processing such big data volumes with PAF and the solutions implemented to overcome them. We will also show the improvements in order to make PAF more modular and accessible to other communities.

Track 3 Session / 128

Pooling the resources of the CMS Tier-1 sites

Authors: Nicolo Magini[1]; Oliver Gutsche[2]

Co-author: Christoph Wissing [3]

[1] *CERN*

[2] *Fermi National Accelerator Lab. (US)*

[3] *Deutsches Elektronen-Synchrotron (DE)*

Corresponding Author: christoph.wissing@desy.de

The CMS experiment at the LHC relies on 7 Tier-1 centres of the WLCG to perform the majority of its bulk processing activity, and to archive its data. During the first run of the LHC, these two functions were tightly coupled as each Tier-1 was constrained to process only the data archived on its hierarchical storage. This lack of flexibility in the assignment of processing workflows occasionally resulted in uneven resource utilisation and in an increased latency in the delivery of the results to the physics community.

The long shutdown of the LHC in 2013-2014 was an opportunity to revisit this mode of operations, disentangling the processing and archive functionalities of the Tier-1 centres. The storage services at the Tier-1s were redeployed breaking the traditional hierarchical model: each site now provides a large disk storage to host input and output data for processing, and an independent tape storage used exclusively for archiving. Movement of data between the tape and disk endpoints is not automated, but triggered externally through the WLCG transfer management systems.

With this new setup, CMS operations actively controls at any time which data is available on disk for processing and which data should be sent to archive. Thanks to the high-bandwidth connectivity guaranteed by the LHCOPN, input data can be freely transferred between disk endpoints as needed to take advantage of free CPU, turning the Tier-1s into a large pool of shared resources. The output data can be validated before archiving them permanently, and temporary data formats can be produced without wasting valuable tape resources. Finally, the data hosted on disk at Tier-1s can now be made available also for user analysis since there is no risk any longer of triggering chaotic staging from tape.

In this contribution, we describe the technical solutions adopted for the new disk and tape endpoints at the sites, and we report on the commissioning and scale testing of the service. We detail the procedures implemented by CMS computing operations to actively manage data on disk at Tier-1 sites, and we give examples of the benefits brought to CMS workflows by the additional flexibility of the new system.

Track 2 Session / 129

DDG4 A Simulation Framework using the DD4hep Detector Description Toolkit

Author: Markus Frank[1]

Co-authors: Andre Sailer [1]; Frank-Dieter Gaede [2]; Marko Petric [3]; Nikiforos Nikiforou [1]

[1] *CERN*

[2] *Deutsches Elektronen-Synchrotron (DE)*

[3] *Jozef Stefan Institute (SI)*

Corresponding Author: markus.frank@cern.ch

The detector description is an essential component that has to be used to analyse and simulate data resulting from particle collisions in high energy physics experiments. Based on the DD4hep detector description toolkit a flexible and data driven simulation

framework was designed using the Geant4 tool-kit. We present this framework and describe the guiding requirements and the architectural design, which was strongly driven by ease of use. The goal was, given an existing detector description, to simulate the detector response to particle collisions in high energy physics experiments with minimal effort, but not impose restrictions to support enhanced or improved behaviour.
Starting from the ROOT based geometry implementation used by DD4hep an automatic conversion mechanism to Geant4 was developed. The physics response and the mechanism to input particle data from generators was highly formalized and can be instantiated on demand using using known factory patterns. A palette of components to model the detector response is provided by default, but improved or more sophisticated components may easily be added using the factory pattern. Only the final configuration of the instantiated components has to be provided by end-users using either C++ or python scripting or an XML based description.

130

The CMS Condition Database system

Author: Giacomo Govi[1]

Co-authors: Andreas Pfeiffer [2]; Miguel Ojeda Sandonis [2]; Salvatore Di Guida [3]

[1] *Fermi National Accelerator Lab. (US)*

[2] *CERN*

[3] *Universita degli Studi Guglielmo Marconi (IT)*

Corresponding Author: giacomo.govi@cern.ch

The Condition Database plays a key role in the CMS computing infrastructure. The complexity of the detector and the variety of the sub-systems involved are setting tight requirements for handling the Conditions. In the last two years the collaboration has put an effort in the re-design of the Condition Database system, with the aim to improve the scalability and the operability for the data taking starting in 2015. The re-design has focused in simplifying the architecture, using the lessons learned during the operation of the previous data-taking period. In the new system the relational features of the database schema are mainly exploited to handle the metadata (Tag and Interval of Validity), allowing for a limited and controlled set of queries. The bulk condition data (Payloads) are stored as unstructured binary data, allowing the storage in a single table with a common layout for all of the condition data types. In this presentation, we describe the full architecture of the system, including the services implemented for uploading payloads and the tools for browsing the database. Furthermore, the implementation choices for the core software will be discussed.

Track 7 Session / 131

HEP cloud production using the CloudScheduler/HTCondor Architecture

Author: Randy Sobie[1]

Co-authors: Colin Roy Leavett-Brown [1]; Frank Olaf Berghaus [1]; Ian Gable [1]; Michael Paterson [2]; Ron Desmarais [3]; Ryan Taylor [1]

[1] *University of Victoria (CA)*

[2] *U*

[3] *University of Victoria*

Corresponding Author: igable@uvic.ca

The use of distributed IaaS clouds with the CloudScheduler/HTCondor architecture has been in production for HEP and astronomy applications for a number of years. The design has proven to be robust and reliable for batch production using HEP clouds, academic non-HEP (opportunistic) clouds and commercial clouds. Further, the system is seamlessly integrated into the existing WLCG infrastructures for the both ATLAS and BelleII experiments. Peak workloads continue to increase and we have utilized over 3000 cores for HEP applications.

We show that the CloudScheduler/HTCondor architecture has adapted well to the evolving cloud ecosystem. Its design preceded the introduction of OpenStack clouds, however, the integration of these clouds has been straightforward. Key developments over the past year include the use of CVMFS for application software distribution. CVMFS together with the global array of Squid web caches and the use of the Squid-discovery service, Shoal, have significantly simplified the management and distribution of virtual machine images. The recent deployment of micro-CernVM images have reduced the size of the images to a few megabytes and make the images independent of the application software. The introduction of Glint, a distributed VM image management service, has made the distribution of images over multiple OpenStack clouds an automated operation. These new services have greatly simplified management of the application software and operating system, and enable us to respond quickly to urgent security issues such as the ShellShock (bash shell) vulnerability.

HEP experiments are beginning to use multi-core applications and this has resulted in a demand for simultaneously running single-core, multi-core or high-memory jobs. We exploit HTCondor's feature that dynamically partitions the slots within a VM instance. As a result, a new VM instance can run either single or multi-core jobs depending on the demand.

The distributed cloud system has been used primarily for low I/O production jobs, however, the goal is run higher I/O production and analysis application jobs. This will require the use a data federation and we are integrating the CERN UGR data federator for managing data distributed over multiple storage elements.

The use of distributed clouds with the CloudScheduler/HTCondor architecture continues to improve and expand in use and functionality. We review the overall progress, and highlight the new features and future challenges.

Track 2 Session / 132

The Library Event Matching classifier for ν_e events in NOvA

Author: Christopher Backhouse[None]

Co-authors: Dominick Rocco [1]; Ryan Patterson [2]

[1] *urn:Google*
[2] *Caltech*

Corresponding Author: dominick.r.rocco@gmail.com

In this paper we present the Library Event Matching (LEM) classification technique for particle identification. The LEM technique was developed for the NOvA electron neutrino appearance analysis as an alternative but complimentary approach to standard multivariate methods. Traditional multivariate PIDs are based on high-level reconstructed quantities which can obscure or discard important low-level detail in high granularity detectors. LEM, by contrast, uses the full hit by hit information of the event, comparing the hit charges and positions of each physics event to a large "template" library of simulated signal and background events. This is a powerful classification technique for the finely segmented NOvA detectors, but poses computational challenges due to the large Monte Carlo template libraries required for to reach the optimal physics sensitivity. We will present both the LEM classification technique as well as its technical implementation for the NOvA experiment exploiting memory mapping techniques on high memory Linux platforms.

133

ATLAS Fast Tracker Simulation Challenges

Author: Alberto Annovi[1]

Co-authors: Alexandre Vaniachine [2]; Dhiman Chakraborty [3]; Dmitri Golubkov [4]; Graeme Stewart [5]; Guido Volpi [6]; Jose Enrique Garcia Navarro [7]; Lauren Alexandra Tompkins [8]; Misha Borodin [9]; Sami Kama [10]; Sergey Panitkin [11]; Yuri Smirnov [12]

[1] *Istituto Nazionale Fisica Nucleare (IT)*

[2] *ATLAS*

[3] *Northern Illinois University*

[4] *Institute for High Energy Physics (IHEP)-Unknown-Unknown*

[5] *University of Glasgow (GB)*

[6] *Sezione di Pisa (IT)*

[7] *Instituto de Fisica Corpuscular (ES)*

[8] *University of Chicago (US)*

[9] *National Research Nuclear University MEPhI (RU)*

[10] *Southern Methodist University (US)*

[11] *Brookhaven National Laboratory (US)*

[12] *Conseil Europeen Recherche Nucl. (CERN)-Unknown-Unknown*

Corresponding Author: graeme.andrew.stewart@cern.ch

To deal with Big Data flood from the ATLAS detector most events have to be rejected in the trigger system. the trigger rejection is complicated by the presence of a large number of minimum-bias events –the pileup. To limit pileup effects in the high luminosity environment of the LHC Run-2, ATLAS relies on full tracking provided by the Fast TracKer (FTK) implemented with custom electronics.

The FTK data processing pipeline has to be simulated in preparation for LHC upgrades to support electronics design and develop trigger strategies at high luminosity. The simulation of the FTK - a highly parallelized system - has inherent performance bottlenecks on general-purpose CPUs. To take advantage of the Grid Computing power, the FTK simulation is integrated with Monte Carlo simulations at the Production System level above the ATLAS workload management system PanDA.

We report on ATLAS experience with FTK simulations on the Grid and next steps for accommodating the growing requirements for resources during the LHC Run-2.

Track 5 Session / 137

Distributed analysis in ATLAS

Author: Alastair Dewhurst[1]

Co-author: Federica Legger [2]

[1] *STFC - Rutherford Appleton Lab. (GB)*

[2] *Ludwig-Maximilians-Univ. Muenchen (DE)*

Corresponding Author: federica.legger@physik.uni-muenchen.de

The ATLAS experiment accumulated more than 140 PB of data during the first run of the Large Hadron Collider (LHC) at CERN. The analysis of such an amount of data for the distributed physics community is a challenging task. The Distributed Analysis (DA) system of the ATLAS experiment is an established and stable component of the ATLAS distributed computing operations. About half a

million user jobs are daily running on DA resources, submitted by more than 1500 ATLAS physicists. The reliability of the DA system during the first run of the LHC and the following shutdown period has been high thanks to the continuous automatic validation of the distributed analysis sites and the user support provided by a dedicated team of expert shifters. During the LHC shutdown, the ATLAS computing model has undergone several changes to improve the analysis workflows, including the re-design of the production system, a new analysis data format and event model, and the development of common reduction and analysis frameworks. We report on the impact such changes have on the DA infrastructure, describe the new DA components, and include first measurements of DA performances in the first months of the second LHC run, starting early 2015.

Track 5 Session / 138

ATLAS Distributed Computing in LHC Run2

Author: Alessandro Di Girolamo[1]

[1] *CERN*

Corresponding Author: simone.campana@cern.ch

The ATLAS Distributed Computing infrastructure has evolved after the first period of LHC data taking in order to cope with the challenges of the upcoming LHC Run2. An increased data rate and computing demands of the Monte-Carlo simulation, as well as new approaches to ATLAS analysis, dictated a more dynamic workload management system (ProdSys2) and data management system (Rucio), overcoming the boundaries imposed by the design of the old computing model. In particular, the commissioning of new central computing system components was the core part of the migration toward the flexible computing model.
The flexible computing utilization exploring the opportunistic resources such as HPC, cloud, and volunteer computing is embedded in the new computing model, the data access mechanisms have been enhanced with the remote access, and the network topology and performance is deeply integrated into the core of the system. Moreover a new data management strategy, based on defined lifetime for each dataset, has been defined to better manage the lifecycle of the data.
In this note, the overview of the operational experience of the new system and its evolution is presented.

139

The ATLAS fast chain MC production project

Authors: Andreas Salzburger[1]; Roland Jansky[2]

[1] *CERN*

[2] *University of Innsbruck (AT)*

Corresponding Author: andreas.salzburger@cern.ch

During the last years ATLAS has successfully deployed a new integrated simulation framework (ISF) which allows a flexible mixture of full and fast detector simulation techniques within the processing of one event. With the ISF, the simulation execution speed could be increased up to a factor 100, which makes subsequent digitisation and reconstruction processing the dominant contributions to the MC production CPU cost. The slowest components of both digitisation and reconstruction are within the Inner Detector due to the complex signal modelling needed in the emulation of the detector readout and in reconstruction due to the combinatorial nature of the problem to solve, respectively. Alternative fast approaches have been developed for these components: for the silicon based detectors a simpler geometrical clustering approach has been deployed replacing the charge drift emulation in the standard digitisation modules, and achieves a very high accuracy in describing the standard output. For the Inner Detector track reconstruction, a Monte Carlo truth based trajectory

building has been deployed for bypassing the CPU intensive pattern recognition. All components have been, together with the ISF, integrated into a new fast MC production chain, aiming to produce fast MC simulated data with acceptable agreement with fully simulated and reconstructed data at a processing time of seconds per event.

Track 8 Session / 140

Fine grained event processing on HPCs with the ATLAS Yoda system

Author: Paolo Calafiura[1]

Co-authors: Danila Oleynik [2]; Paul Nilsson [3]; Peter Van Gemmeren [4]; Sergey Panitkin [3]; Tadashi Maeno [3]; Torre Wenaus [3]; Vakho Tsulaia [1]; Wen Guan [5]

[1] *Lawrence Berkeley National Lab. (US)*

[2] *Joint Inst. for Nuclear Research (RU)*

[3] *Brookhaven National Laboratory (US)*

[4] *Argonne National Laboratory (US)*

[5] *University of Wisconsin (US)*

Corresponding Author: vakhtang.tsulaia@cern.ch

High performance computing facilities present unique challenges and opportunities for HENP event processing. The massive scale of many HPC systems means that fractionally small utilizations can yield large returns in processing throughput. Parallel applications which can dynamically and efficiently fill any scheduling opportunities the resource presents benefit both the facility (maximal utilization) and the (compute-limited) science. The ATLAS Yoda system provides this capability to HENP-like event processing applications by implementing event-level processing in an MPI-based master-client model that integrates seamlessly with the more broadly scoped ATLAS Event Service. Fine grained, event level work assignments are intelligently dispatched to parallel workers to sustain full utilization on all cores, with outputs streamed off to destination object stores in near real time with similarly fine granularity, such that processing can proceed until termination with full utilization. The system offers the efficiency and scheduling flexibility of preemption without requiring the application actually support or employ checkpointing. We will present the new Yoda system, its motivations, architecture, implementation, and applications in ATLAS data processing at several US HPC centers.

Track 5 Session / 141

Data Preservation in ATLAS

Author: Roger Jones[1]

Co-authors: David Michael South [2]; Kyle Stuart Cranmer [3]

[1] *Lancaster University (GB)*

[2] *Deutsches Elektronen-Synchrotron (DE)*

[3] *New York University (US)*

Corresponding Author: roger.jones@cern.ch

Complementary to parallel open access and analysis preservation initiatives, ATLAS is taking steps to ensure that the data taken by the experiment during run-1 remain accessible and available for future analysis by the collaboration. An evaluation of what is required to achieve this is underway, examining the ATLAS data production chain to establish the effort required and potential problems.

Several alternatives are explored, but the favoured solution is to bring the run 1 data and software in line with the equivalent to that which will be used for run 2. This will result in a coherent ATLAS dataset for the data already taken and that to come in the future.

142

Analysis Preservation in ATLAS

Author: Roger Jones[1]

Co-authors: David Michael South [2]; Kyle Stuart Cranmer [3]

[1] *Univ Manchester/Cockcroft Inst.*

[2] *Deutsches Elektronen-Synchrotron (DE)*

[3] *New York University (US)*

Corresponding Author: lukas.heinrich@cern.ch

Long before data taking ATLAS established a policy that all analyses need to be preserved. In the initial data-taking period, this has been achieved by various tools and techniques. ATLAS is now reviewing the analysis preservation with the aim to bring coherence and robustness to the process and with a clearer view of the level of reproducibility that is reasonably achievable. The secondary aim is to reduce the load on the analysts. Once complete, this will serve for our internal preservation needs but also provide a basis for any subsequent sharing of analysis results with external parties.

Track 4 Session / 144

The Future of PanDA in ATLAS Distributed Computing

Author: Kaushik De[1]

Co-authors: Alexandre Vaniachine [2]; Alexei Klimentov [3]; Artem Petrosyan [4]; Danila Oleynik [4]; Jaroslava Schovancova [3]; Paul Nilsson [3]; Sergey Panitkin [3]; Tadashi Maeno [3]; Torre Wenaus [3]

[1] *University of Texas at Arlington (US)*

[2] *ATLAS*

[3] *Brookhaven National Laboratory (US)*

[4] *Joint Inst. for Nuclear Research (RU)*

Corresponding Author: tmaeno@bnl.gov

Experiments at the Large Hadron Collider (LHC) face unprecedented computing challenges. Heterogeneous resources are distributed worldwide at hundreds of sites, thousands of physicists analyze the data remotely, the volume of processed data is beyond the exabyte scale, while data processing requires more than a few billion hours of computing usage per year. The PanDA (Production and Distributed Analysis) system was developed to meet the scale and complexity of LHC distributed computing for the ATLAS experiment. In the process, the old batch job paradigm of locally managed computing in HEP was discarded in favor of a far more automated, flexible and scalable model. The success of PanDA in ATLAS is leading to widespread adoption and testing by other experiments. PanDA is the first exascale workload management system in HEP, already operating at more than a million computing jobs per day, and processing over an exabyte of data in 2013. There are many new challenges that PanDA will face in the near future, in addition to new challenges of scale, heterogeneity and increasing user base. PanDA will need to handle rapidly changing computing infrastructure, will require factorization of code for easier deployment, will need to incorporate additional information sources including network metrics in decision making, be able to control network circuits, handle dynamically sized workload processing, provide improved visualization, and face many other

challenges. In this talk we will focus on the new features, planned or recently implemented, that are relevant to the next decade of distributed computing workload management using PanDA.

Track 4 Session / 145

Dynamic Resource Allocation with the ARC Control Tower

Author: Andrej Filipcic[1]

Co-authors: David Cameron [2]; Jon Kerr Nilsen [2]

[1] *Jozef Stefan Institute (SI)*

[2] *University of Oslo (NO)*

Corresponding Author: andrej.filipcic@ijs.si

Distributed computing resources available for high-energy physics research are becoming less dedicated to one type of workflow and researchers'workloads are increasingly exploiting modern computing technologies such as parallelism. The current pilot job management model used by many experiments relies on static dedicated resources and cannot easily adapt to these changes. The model used for ATLAS in Nordic countries and some other places enables a flexible job management system based on dynamic resources allocation. Rather than a fixed set of resources managed centrally, the model allows resources to be requested on the fly. The ARC Computing Element (ARC-CE) and ARC Control Tower (aCT) are the key components of the model. The aCT requests jobs from the ATLAS job mangement system (Panda) and submits a fully-formed job description to ARC-CEs. ARC-CE can then dynamically request the required resources from the underlying batch system. In this paper we describe the architecture of the model and the experience of running many millions of ATLAS jobs on it.

Track 7 Session / 146

Evolution of Cloud Computing in ATLAS

Author: Ryan Taylor[1]

Co-authors: Alessandro Di Girolamo [2]; Frank Olaf Berghaus [1]; Ian Gable [1]; Jose Caballero Bejar [3]; Laurence Field [4]; Matt Leblanc [1]; Michael Paterson [5]; Peter Love [6]; Randall Sobie [7]; Ron Desmarais [7]

[1] *University of Victoria (CA)*

[2] *C*

[3] *Brookhaven National Laboratory (US)*

[4] *CERN*

[5] *U*

[6] *Lancaster University (GB)*

[7] *University of Victoria*

Corresponding Author: rptaylor@uvic.ca

The ATLAS experiment has successfully incorporated cloud computing technology and cloud resources into its primarily grid-based model of distributed computing. Cloud R&D activities continue to mature and transition into stable production systems, while ongoing evolutionary changes are still needed to adapt and refine the approaches used, in response to changes in prevailing cloud technology. In addition, completely new developments are needed to handle emerging requirements.

This work will describe the overall evolution of cloud computing in ATLAS. The current status of the VM management systems used for harnessing IAAS resources will be discussed. Monitoring and

accounting systems tailored for clouds are needed to complete the integration of cloud resources within ATLAS' distributed computing framework. We are developing and deploying new solutions to address the challenge of operation in a geographically distributed multi-cloud scenario, including a system for managing VM images across multiple clouds, a system for dynamic location-based discovery of caching proxy servers, and the usage of a data federation to unify the worldwide grid of storage elements into a single namespace and access point. The usage of the experiment's HLT farm for Monte Carlo production, in a specialized cloud environment, will be presented. Finally, we evaluate commercial clouds, and have conducted a study assessing cost vs. benchmark performance.

147

Preparing ATLAS Reconstruction for LHC Run 2

Author: Jovan Mitrevski[1]

[1] *Ludwig-Maximilians-Univ. Muenchen (DE)*

Corresponding Author: jovan.mitrevski@cern.ch

In order to maximize the physics potential of the ATLAS detector during LHC's run 2, the Reconstruction software has been updated. Flat computing budgets required a factor of three improved run time, while the new xAOD data format forced changes in the reconstruction algorithms. Physics performance improvements have been made in the reconstruction of various objects, using improved techniques like particle flow, multivariate discriminants, etc. This talk will present an overview of the improvements that have been made.

Track 2 Session / 148

Status and future evolution of the ATLAS offline software

Author: Vakho Tsulaia[1]

Co-authors: Graeme Stewart [2]; Markus Elsing [3]; Rolf Seuster [4]

[1] *Lawrence Berkeley National Lab. (US)*

[2] *University of Glasgow (GB)*

[3] *CERN*

[4] *TRIUMF (CA)*

Corresponding Author: rolf.seuster@cern.ch

The talk will give a summary of the broad spectrum of software upgrade projects to prepare ATLAS for the challenges of the soon coming LHC Run-2. Those projects include the reduction of the CPU required for reconstruction by a factor 3 compared to 2012, which was required to meet the challenges of the expected increase in pileup and the higher data taking rate of up to 1 kHz. As well, the new Integrated Simulation Framework (ISF) has been put into production. By far the most ambitious project is the implementation of a completely new Analysis Model, based on a new ROOT readable reconstruction format xAOD, a reduction framework based on the train model to centrally produce skimmed data samples and an analysis framework. The Data Challenge 2014 has been a first large scale test of most of the foreseen software upgrades. At the end of the talk an overview will be given of future software projects and plans that will lead up to the coming Long Shutdown 2 as the next major ATLAS software upgrade phase.

149

Improvements of LHC data analysis techniques at italian WLCG sites. Case-study of the transfer of this technology to other research areas

Authors: Alessandro De Salvo[1]; Domenico Elia[2]; Laura Perini[3]; Tommaso Boccali[4]

[1] *Universita e INFN, Roma I (IT)*

[2] *INFN Bari*

[3] *Università degli Studi e INFN Milano (IT)*

[4] *Sezione di Pisa (IT)*

Corresponding Authors: alessandro.de.salvo@cern.ch, domenico.elia@cern.ch, laura.perini@mi.infn.it, tommaso.boccali@cern.ch

In 2012, 14 Italian Institutions participating in all major LHC Experiments won a grant from the ITALIAN MINISTRY OF RESEARCH (MIUR), to optimise Analysis activities and in general the Tier2/Tier3 infrastructure. We report on the activities being researched upon, on the considerable improvement in the ease of access to resources by physicists, also those with no specific computing interests. We focused on items like distributed storage federations, access to batch-like facilities, provisioning of user interfaces on demand and cloud systems. R&D on next-generation databases, distributed analysis interfaces, and new computing architectures was also carried on. The project, ending in the first months of 2016, will produce a white paper with recommendations on best practices for data-analysis support by computing centers.

Track 2 Session / 150

The ATLAS Higgs Machine Learning Challenge

Author: David Rousseau[1]

Co-authors: Balázs Kégl [1]; Cecile Germain-Renaud [2]; Claire Adam Bourdarios [3]; Glen Cowan [4]; Isabelle Guyon [5]

[1] *LAL-Orsay, FR*

[2] *Laboratoire de Recherche en Informatique*

[3] *Laboratoire de l'Accelerateur Lineaire (FR)*

[4] *Royal Holloway, University of London*

[5] *Chalearn*

Corresponding Author: g.cowan@rhul.ac.uk

High Energy Physics has been using Machine Learning techniques (commonly known as Multivariate Analysis) since the 1990s with Artificial Neural Net and more recently with Boosted Decision Trees, Random Forest etc. Meanwhile, Machine Learning has become a full blown field of computer science. With the emergence of Big Data, data scientists are developing new Machine Learning algorithms to extract meaning from large heterogeneous data.

HEP has exciting and difficult problems like the extraction of the Higgs boson signal, and at the same time data scientists have advanced algorithms: the goal of the HiggsML project was to bring the two together by a "challenge": participants from all over the world and any scientific background could compete online to obtain the best Higgs to tau tau signal significance on a set of ATLAS fully simulated Monte Carlo signal and background.

Instead of HEP physicists browsing through machine learning papers and trying to infer which new algorithms might be useful for HEP, then coding and tuning them, the challenge has brought realistic HEP data to the data scientists on the Kaggle platform, which is well known in the Machine Learning community.

The challenge has been organized by the ATLAS collaboration associated to data scientists, in partnership with the Paris Saclay Center for Data Science, CERN and Google.
The challenge ran from May to September 2014, drawing considerable attention. 1785 teams participated, making it the most popular challenge ever on the Kaggle platform. New Machine Learning techniques have been used by the participants with significantly better results than usual HEP tools. This presentation has two parts: the first one describes how a HEP problem was simplified (not too much!) and wrapped up into an online challenge, the second what was learned from the challenge, in terms of new Machine Learning algorithms and techniques which could have an impact on future HEP analysis.

Track 2 Session / 151

Requirements for a Next Generation Framework: ATLAS Experience

Author: Graeme Stewart[1]

Co-authors: Benjamin Michael Wynne [2]; Charles Leggett [3]; David Malon [4]; John Baines [5]; Paolo Calafiura [3]; Sami Kama [6]; Tomasz Bold [7]; Walter Lampl [8]

[1] *University of Glasgow (GB)*

[2] *University of Edinburgh (GB)*

[3] *Lawrence Berkeley National Lab. (US)*

[4] *Argonne National Laboratory (US)*

[5] *STFC - Rutherford Appleton Lab. (GB)*

[6] *Southern Methodist University (US)*

[7] *AGH Univ. of Science and Technology, Krakow*

[8] *University of Arizona (US)*

Corresponding Author: sami.kama@cern.ch

The challenge faced by HEP experiments from the current and expected architectural evolution of CPUs and co-processors is how to successfully exploit concurrency and keep memory consumption within reasonable limits. This is a major change from frameworks which were designed for serial event processing on single core processors in the 2000s. ATLAS has recently considered this problem in some detail through its Future Frameworks Requirements group.

Here we report on the major considerations of the group, which was charged with considering the best strategies to exploit current and anticipated CPU technologies. The group has re-examined the basic architecture of event processing and considered how the building blocks of a framework (algorithms, services, tools and incidents) should evolve. The group has also had to take special care to ensure that the use cases of the ATLAS high level trigger are encompassed, which differ in important ways from offline event processing (for example, 99% of events are rejected, which must be done with minimum resource investment). Finally, the group has considered how best to use the wide variety of concurrency techniques, such as multi-processing, multi-threading, offloaded accelerator technology, parallel I/O, and how to exploit resources such as high performance computing sites.

Track 8 Session / 152

Integration of PanDA workload management system with Titan supercomputer at OLCF.

Author: Sergey Panitkin[1]

Co-authors: Alexandre Vaniachine [2]; Alexei Klimentov [1]; Artem Petrosyan [3]; Danila Oleynik [3]; Jaroslava Schovancova [4]; Kaushik De [5]; Torre Wenaus [1]

[1] *Brookhaven National Laboratory (US)*

[2] *ATLAS*

[3] *Joint Inst. for Nuclear Research (RU)*

[4] *CERN*

[5] *University of Texas at Arlington (US)*

Corresponding Author: panitkin@bnl.gov

The PanDA (Production and Distributed Analysis) workload management system (WMS) was developed to meet the scale and complexity of LHC distributed computing for the ATLAS experiment.
While PanDA currently uses more than 100,000 cores at well over 100 Grid sites with a peak performance of 0.3 petaFLOPS, next LHC data taking run will require more resources than Grid computing can possibly provide.
To alleviate these challenges, ATLAS is engaged in an ambitious program to expand the current computing model to include additional resources such as the opportunistic use of supercomputers.

We will describe a project aimed at integration of PanDA WMS 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 light-weight MPI wrappers to run single threaded workloads in parallel on Titan's multi-core worker nodes. It also gives PanDA new capability to collect, in real time, information about unused worker nodes on Titan, which allows precisely define the size and duration of jobs submitted to Titan according to available free resources.
This capability significantly reduces PanDA job wait time while improving Titan's utilization efficiency.
This implementation was tested with a variety of Monte-Carlo workloads on Titan and is being tested on several other supercomputing platforms.

153

Bringing ATLAS production to HPC resources - A use case with the Hydra supercomputer of the Max Planck Society

Author: Luca Mazzaferro[1]

Co-authors: John Kennedy [2]; Rodney Walker [3]; Stefan Kluth [4]

[1] *Universita e INFN Roma Tor Vergata (IT)*

[2] *LMU Munich*

[3] *Ludwig-Maximilians-Univ. Muenchen (DE)*

[4] *Max-Planck-Institut fuer Physik (Werner-Heisenberg-Institut) (D*

Corresponding Author: luca.mazzaferro@cern.ch

The possible usage of HPC resources by ATLAS is now becoming viable due to the changing nature of these systems and it is also very attractive due to the need for increasing amounts of simulated data.

In recent years the architecture of HPC systems has evolved, moving away from specialized monolithic systems, to a more generic linux type platform. This change means that the deployment of non HPC specific codes has become much easier. The timing of this evolution perfectly suits the needs of ATLAS and opens a new window of opportunity.

The ATLAS experiment at CERN will begin a period of high luminosity data taking in 2015. This high luminosity phase will be accompanied by a need for increasing amounts of simulated data which is expected to exceed the capabilities of the current Grid infrastructure.

ATLAS aims to address this need by opportunistically accessing resources such as cloud and HPC systems. This paper presents the results of a pilot project undertaken by ATLAS and the MPP/RZG to provide access to the HYDRA supercomputer facility. Hydra is the supercomputer of the Max Planck Society, it is a linux based supercomputer with over 80000 cores and 4000 physical nodes located at the RZG near Munich.

This paper describes the work undertaken to integrate Hydra into the ATLAS production system by using the Nordugrid ARC-CE and other standard Grid components. The customization of these components and the strategies for HPC usage are discussed as well as possibilities for future directions.

154

Spanish ATLAS Tier-2 facing up to Run-2 period of LHC

Author: Santiago Gonzalez De La Hoz[1]

Co-authors: Almudena Del Rocio Montiel Gonzalez [2]; Alvaro Fernandez Casani [3]; Andreu Pacheco Pages [4]; Elena Oliver Garcia [5]; Farida Fassi [6]; Javier Sanchez [5]; Jose Del Peso [2]; Jose Salt [3]; Miguel Villaplana Perez [7]; Mohammed KACI [8]; Victor Ruben Lacort Pellicer [5]; Victoria Sanchez Martinez [9]

[1] *IFIC-Valencia*

[2] *Universidad Autonoma de Madrid (ES)*

[3] *Instituto de Fisica Corpuscular (ES)*

[4] *Institut de Física d'Altes Energies - Barcelona (ES)*

[5] *Universidad de Valencia (ES)*

[6] *Universite Mohammed V (MA)*

[7] *Università degli Studi e INFN Milano (IT)*

[8] *IFIC - Valencia - Spain*

[9] *Instituto de Fisica Corpuscular (IFIC) UV-CSIC (ES)*

Corresponding Author: santiago.gonzalezdelahoz@cern.ch

The goal of this work is to describe the way of addressing the main challenges of Run-2 by the Spanish ATLAS Tier-2. The considerable increase of energy and luminosity for the upcoming Run-2 w.r.t. Run-1 has led to a revision of the ATLAS computing model as well as some of the main ATLAS computing tools. The adaptation to these changes will be shown, with the peculiarities that it is a distributed Tier-2 composed of three sites and its members are involved on ATLAS computing tasks with a hub of research, innovation and education.

155

gLExec integration with the ATLAS PanDA workload management system

Authors: Alessandro Di Girolamo[1]; Edward Karavakis[2]; Kaushik De[3]; Maarten Litmaath[2]; Paul Nilsson[4]; Ramon Medrano Llamas[2]; Tadashi Maeno[4]; Torre Wenaus[4]

Co-authors: Fernando Harald Barreiro Megino [2]; Simone Campana [2]

[1] *C*

[2] *CERN*

[3] *University of Texas at Arlington (US)*

[4] *Brookhaven National Laboratory (US)*

Corresponding Author: simone.campana@cern.ch

The ATLAS Experiment at the Large Hadron Collider has collected data during Run 1 and is ready to collect data in Run 2. The ATLAS data are distributed, processed and analysed at more than 130 grid and cloud sites across the world. At any given time, there are more than 150,000 concurrent jobs running and about a million jobs are submitted on a daily basis on behalf of thousands of physicists within the ATLAS collaboration. The Production and Distributed Analysis (PanDA) workload management system has proved to be a key component of ATLAS and plays a crucial role in the success of the large-scale distributed computing as it is the sole system for distributed processing of Grid jobs across the collaboration since October 2007.

ATLAS user jobs are executed on worker nodes by pilots sent to the sites by pilot factories. This pilot architecture has greatly improved job reliability and although it has clear advantages, such as making the working environment homogeneous by hiding any potential heterogeneities, the approach presents security and traceability issues distinct from standard batch jobs for which the submitter is also the payload owner. Jobs initially inherit the identity of the pilot submitter, typically a robot certificate with very limited rights. By default the payload jobs then execute directly under that same identity on a Worker Node. This exposes the pilot environment to the payload, requiring any pilot 'secrets' such as the proxy to be hidden; it constrains the rights and identity of the user job to be identical to the pilot; and it requires sites to take extra measures to achieve user traceability and user job isolation.

To address these security risks, the gLExec tool and framework can be used to let the payloads for each user be executed under a different UNIX user identity that uniquely identifies the ATLAS user.

This presentation describes the recent improvements and evolution of the security model within the ATLAS PanDA system, including improvements in the PanDA pilot, in the PanDA server and their integration with MyProxy, a credential caching system that entitles a person or a service to act in the name of the issuer of the credential. Finally, we will present results from ATLAS user jobs running with gLExec and give an insight into future deployment plans.

156

The importance of having an appropriate relational data segmentation in ATLAS

Author: Gancho Dimitrov[1]

[1] *CERN*

Corresponding Author: dario.barberis@cern.ch

In this paper we describe specific technical solutions put in place in various database applications of the ATLAS experiment at LHC where we make use of several partitioning techniques available in Oracle 11g. With the broadly used range partitioning and its option of automatic interval partitioning we add our own logic in PLSQL procedures and scheduler jobs to sustain data sliding windows in order to enforce various data retention policies. We also make use of the new Oracle 11g reference partitioning in the Nightly Build System to achieve uniform data segmentation. However the most challenging issue was to segment the data of the new ATLAS Distributed Data Management system (Rucio), which resulted in tens of thousands list type partitions and sub-partitions. Partition and sub-partition management, index strategy, statistics gathering and queries execution plan stability are important factors when choosing an appropriate physical model for the application data management. The so-far accumulated knowledge and analysis on the new Oracle 12c version features that could be beneficial will be shared with the audience.

Track 3 Session / 157

New data access with HTTP/WebDAV in the ATLAS experiment

Author: Johannes Elmsheuser[1]

Co-authors: Cedric Serfon [1]; Paul Nilsson [2]; Rodney Walker [1]; Sylvain Blunier [3]; Vincenzo Lavorini [4]

[1] *Ludwig-Maximilians-Univ. Muenchen (DE)*

[2] *Brookhaven National Laboratory (US)*

[3] *Pontificia Univ. Catolica de Chile (CL)*

[4] *Universita della Calabria (IT)*

Corresponding Author: johannes.elmsheuser@physik.uni-muenchen.de

With the exponential growth of LHC (Large Hadron Collider) data in the years 2010-2012, distributed computing has become the established way to analyze collider data. The ATLAS experiment Grid infrastructure includes more than 130 sites worldwide, ranging from large national computing centres to smaller university clusters. So far the storage technologies and access protocols to the clusters that host this tremendous amount of data vary from site to site. HTTP/WebDAV offers the possibility to use a unified industry standard to access the storage. We present the deployment and testing of HTTP/WebDAV for local and remote data access in the ATLAS experiment for the new data management system Rucio and the PanDA workload management system. Deployment and large scale tests have been performed using the Grid testing system HammerCloud and the ROOT HTTP plugin Davix.

159

Improved ATLAS HammerCloud Monitoring for local Site Administration

Author: Michael Boehler[1]

Co-authors: Federica Legger [2]; Francesco Giovanni Sciacca [3]; Friedrich Hoenig [2]; Johannes Elmsheuser [2]; Valentina Mancinelli [4]

[1] *Albert-Ludwigs-Universitaet Freiburg (DE)*

[2] *Ludwig-Maximilians-Univ. Muenchen (DE)*

[3] *Universitaet Bern (CH)*

[4] *Universita e INFN (IT)*

Corresponding Author: michael.boehler@cern.ch

Every day hundreds of tests are run on the Worldwide LHC Computing Grid for the ATLAS, CMS, and LHCb experiments in order to evaluate the performance and reliability of the different computing sites. All this activity is steered, controlled, and monitored by the HammerCloud testing infrastructure. Sites with failing functionality tests are auto-excluded from the ATLAS computing grid, therefore it is essential to provide a detailed and well organized web interface for the local site administrators such that they can easily spot and promptly solve site issues. Additional functionalities have been developed to extract and visualize the most relevant information. The site administrators can now be pointed easily to major site issues which lead to site blacklisting as well as possible minor issues that are usually not conspicuous enough to warrant the blacklisting of a specific site, but can still cause undesired effects such as a non-negligible job failure rate. This contribution summarizes the different developments and optimizations of the HammerCloud web interface and gives an overview of typical use cases.

161

The ATLAS ARC ssh back-end to HPC

Author: Michi Hostettler[1]

Co-authors: Francesco Giovanni Sciacca [1]; Sigve Haug [1]

[1] *Universitaet Bern (CH)*

Corresponding Author: gsciacca@mail.cern.ch

The current distributed computing resources used for simulating and processing collision data collected by the LHC experiments are largely based on dedicated Linux clusters. Job control and software provisioning mechanisms are quite different from the common concept of self-contained HPC applications run by particular users on specific HPC systems. This poster reports on the development and the usage of a ssh back-end to the Advanced Resource Connector (ARC) middleware to enable HPC compliant access and on the corresponding software provisioning mechanisms.

162

Exploring JavaScript and ROOT technologies to create Web-based ATLAS analysis and monitoring tools

Author: Arturo Sanchez Pineda[1]

Co-author: Tetiana Moskalets [2]

[1] *Universita di Napoli Federico II-Universita e INFN*

[2] *C*

Corresponding Author: arturo.sanchez.pineda@cern.ch

We explore the potentialities of current web applications to create online interfaces that allow the visualization, interaction and real physics cut-based analysis and monitoring of processes trough a web browser. The project consists in the initial development of web-based and cloud computing services to allow students and researches to perform fast and very useful cut-based analysis on a browser, reading and using real data and official Monte-Carlo simulations stored in ATLAS computing facilities. Several tools are considered: ROOT, JavaScript and HTML. Our study case is the current cut-based H->ZZ->llqq analysis of the ATLAS experiment. Preliminary but satisfactory results have been obtained online; this presentation describes the tests and plans and future upgrades.

163

ATLAS strategy for primary vertex reconstruction during run-II of the LHC

Author: Kirill Prokofiev[1]

Co-authors: Andrew Mark Wharton [2]; Guennadi Borissov [2]; Kathryn Grimm [2]; Lars Egholm Pedersen [3]; Matthew Scott Rudolph [4]; Olivier Arnaez [5]; Simone Pagan Griso [6]

[1] *The Hong Kong University of Science and Technology (HK)*

[2] *Lancaster University (GB)*

[3] *University of Copenhagen (DK)*

[4] *University of Toronto (CA)*

[5] *CERN*

[6] *Lawrence Berkeley National Lab. (US)*

Corresponding Authors: james.catmore@cern.ch, roger.jones@cern.ch

Based on experience gained from run-I of the LHC, the ATLAS vertex reconstruction group has developed a refined primary vertex reconstruction strategy for run-II. With instantaneous luminosity exceeding 10ˆ34 cm-2 s-1, an average of 40 to 50 pp collisions per bunch crossing are expected. Together with the increase of the center-of-mass collision energy from 8 TeV to 13 TeV, this will create a challenging environment for primary vertex pattern recognition.

This contribution explains the ATLAS strategy for primary vertex reconstruction in high pile-up conditions. The new approach is based on vertex seeding with a medical-imaging algorithm, adaptive reconstruction of vertex positions, and iterative recombination of occasional split vertices. The mathematical foundation and software implementation of the method are described in detail. Monte Carlo-based estimates of vertex reconstruction performance for LHC run-II are presented.

Track 2 Session / 164

A New Petabyte-scale Data Derivation Framework for ATLAS

Author: Paul James Laycock[1]

Co-authors: Eirik Gramstad [2]; Graeme Stewart [3]; Jack Cranshaw [4]; James Catmore [2]; Nurcan Ozturk [5]; Thomas Gillam [6]

[1] *University of Liverpool (GB)*

[2] *University of Oslo (NO)*

[3] *University of Glasgow (GB)*

[4] *Argonne National Laboratory (US)*

[5] *University of Texas at Arlington*

[6] *University of Cambridge (GB)*

Corresponding Author: james.catmore@cern.ch

During the Long shutdown of the LHC, the ATLAS collaboration overhauled its analysis model based on experience gained during Run 1. A significant component of the model is a "Derivation Framework" that takes the Petabyte-scale AOD output from ATLAS reconstruction and produces samples, typically Terabytes in size, targeted at specific analyses. The framework incorporates all of the functionality of the core reconstruction software, while producing outputs that are simply configured. Event selections are specified via strings, including support for logical operations. The output content can be highly optimised to minimise disk requirements, while maintaining the same C++ interface. The framework includes an interface to the late-stage physics analysis tools, ensuring that the final outputs are consistent with tool requirements. Finally, the framework allows several outputs to be produced for the same input, providing the possibility to optimise configurations to computing resources.

Track 2 Session / 165

Running ATLAS workloads within massively parallel distributed applications using Athena Multi-Process framework (AthenaMP)

Author: Vakho Tsulaia[1]

Co-authors: Charles Leggett [1]; Paolo Calafiura [1]; Peter Van Gemmeren [2]; Rolf Seuster [3]

[1] *Lawrence Berkeley National Lab. (US)*

[2] *Argonne National Laboratory (US)*

[3] *TRIUMF (CA)*

Corresponding Author: vakhtang.tsulaia@cern.ch

AthenaMP is a multi-process version of the ATLAS reconstruction and data analysis framework Athena. By leveraging Linux fork and copy-on-write, it allows the sharing of memory pages between event processors running on the same compute node with little to no change in the application code. Originally targeted to optimize the memory footprint of reconstruction jobs, AthenaMP has demonstrated that it can reduce the memory usage of certain confugurations of ATLAS production jobs by a factor of 2. AthenaMP has also evolved to become the parallel event-processing core of the recently developed ATLAS infrastructure for fine-grained event processing (Event Service) which allows to run AthenaMP inside massively parallel distributed applications on hundreds of compute nodes simultaneously. We present the architecture of AthenaMP, various strategies implemented by AthenaMP for scheduling workload to worker processes (for example: Shared Event Queue and Shared Distributor of Event Tokens) and the usage of AthenaMP in the diversity of ATLAS event processing workloads on various computing resources: Grid, opportunistic resources and HPC.

Track 2 Session / 166

Development of a Next Generation Concurrent Framework for the ATLAS Experiment

Author: Charles Leggett[1]

Co-authors: Benjamin Michael Wynne [2]; David Malon [3]; Graeme Stewart [4]; Paolo Calafiura [1]; Walter Lampl [5]

[1] *Lawrence Berkeley National Lab. (US)*

[2] *University of Edinburgh (GB)*

[3] *Argonne National Laboratory (US)*

[4] *University of Glasgow (GB)*

[5] *University of Arizona (US)*

Corresponding Author: charles.g.leggett@gmail.com

The ATLAS experiment has successfully used its Gaudi/Athena software framework for data taking and analysis during the first LHC run, with billions of events successfully processed. However, the design of Gaudi/Athena dates from early 2000 and the software and the physics code has been written using a single threaded, serial design. This programming model has increasing difficulty in exploiting the potential of current CPUs, which offer their best performance only through taking full advantage of multiple cores and wide vector registers. Future CPU evolution will intensify this trend, with core counts increasing and memory per core falling. With current memory consumption for 64 bit ATLAS reconstruction in a high luminosity environment approaching 4GB, it will become impossible to fully occupy all cores in a machine without exhausting available memory. However, since maximising performance per watt will be a key metric, a mechanism must be found to use all cores as efficiently as possible.

In this paper we report on our progress with a practical demonstration of the use of multi-threading in the ATLAS reconstruction software, using the GaudiHive framework. We have expanded support to Calorimeter, Inner Detector, and Tracking code, discussing what changes were necessary in order to allow the serially designed ATLAS code to run, both to the framework and to the tools and algorithms used. We report on both the performance gains, and what general lessons were learned about the code patterns that had been employed in the software and which patterns were identified as particularly problematic for multi-threading. We also present our findings on implementing a hybrid multi-threaded / multi-process framework, to take advantage of the strengths of each type of concurrency, while avoiding some of their corresponding limitations.

Track 4 Session / 167

An object-oriented approach to generating highly configurable Web interfaces for the ATLAS experiment

Author: Bruno Lange Ramos[1]

Co-authors: Breno Vieira Arosa [1]; Carmen Maidantchik [1]; Kathy Pommes [2]; Varlen Pavani Neto [1]

[1] *Univ. Federal do Rio de Janeiro (BR)*

[2] *CERN*

Corresponding Authors: bruno.lange.ramos@cern.ch, bruno.lange@cern.ch

In order to manage a heterogeneous and worldwide collaboration, the ATLAS experiment developed web systems that range from supporting the process of publishing scientific papers to monitoring equipment radiation levels. These systems are vastly supported by Glance, a technology that was set forward in 2004 to create an abstraction layer on top of different databases; it automatically recognizes their modelling and generates web search interfaces. Fence (Front ENd ENgine for glaNCE) assembles classes to build applications by making extensive use of configuration files. It produces templates of the core JSON files on top of which it is possible to create Glance-compliant search interfaces. Once the database, its schemas and tables are defined using Glance, its records can be incorporated into the templates by escaping the returned values with a reference to the column identifier wrapped around double enclosing brackets.
The developer may also expand on available configuration files to create HTML forms and securely interact with the database. A token is issued within each deployed form as a random string of characters which must then be matched whenever it is posted. Additionally, once the user is authenticated through CERN's Shibboleth single sign-on, Fence assigns them roles and permissions as stored in the database. Clearance attributes can then be bound to individual inputs within their own JSON description so that whenever they are submitted, the resulting system verifies whether the user has the necessary permissions to edit them. Input validation is primarily carried out on the server side with PHP but, following progressive enhancement guidelines, verification routines may be additionally entrusted to the client side by enabling specific HTML5 data attributes which are then handed over to the jQuery validation plugin. User monitoring is accomplished by logging URL requests along with any POST data. The documentation is automatically published from the source code using the Doxygen tool and made accessible in a web interface. Fence, therefore, speeds up the implementation of Web software products while minimizing maintenance overhead and facilitating the comprehension of embedded rules and requirements.

168

AGIS: Evolution of Distributed Computing information system for ATLAS

Author: Alexey Anisenkov[1]

Co-author: Alessandro Di Girolamo [2]

[1] *Budker Institute of Nuclear Physics (RU)*

[2] *CERN*

Corresponding Author: alexey.anisenkov@cern.ch

The variety of the ATLAS 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.
We describe the evolution and the recent developments of AGIS functionalities, including new bulk update implementation for user interfaces. The improvements of information model are also shown, in particular the consolidation of computing resources definition for the ATLAS workload management system (PanDA). We will explain how the AGIS flexibility and scalability allow the transparent definition of virtual resources like Cloud and HPC capabilities recently become widely used by ATLAS.

169

Design, Results, Evolution and Status of the ATLAS simulation in Point1 project.

Authors: Alexandre Zaytsev[1]; Franco Brasolin[2]

Co-authors: Alessandro Di Girolamo [3]; Alexey Sedov [4]; Christopher Jon Lee [5]; Cristian Contescu [6]; Daniel Fazio [3]; Diana Scannicchio [7]; Fuqiang Wang [8]; Matthew Shaun Twomey [9]; Mikel Eukeni Pozo Astigarraga [3]; Sergio Ballestrero [5]; Silvia-Maria Fressard-Batraneanu [3]

[1] *Institute for High Energy Physics (RU)*

[2] *Universita e INFN (IT)*

[3] *CERN*

[4] *Universitat Autònoma de Barcelona*

[5] *University of Johannesburg (ZA)*

[6] *Polytechnic University of Bucharest (RO)*

[7] *University of California Irvine (US)*

[8] *Lawrence Berkeley National Lab. (US)*

[9] *University of Washington (US)*

Corresponding Author: franco.brasolin@cern.ch

During the LHC long shutdown period (LS1), that started in 2013, the simulation in Point1 (Sim@P1) project takes advantage in an opportunistic way of the trigger and data acquisition (TDAQ) farm of the ATLAS experiment. The farm provides more than 1500 computer nodes, and they are particularly suitable for running event generation and Monte Carlo production jobs that are mostly CPU and not I/O bound. It is capable of running up to 2500 virtual machines (VM) provided with 8 CPU cores each, for a total of up to 20000 parallel running jobs. This contribution gives a thorough review of the design, the results and the evolution of the Sim@P1 project operating a large scale Openstack based virtualized platform deployed on top of the ATLAS TDAQ farm computing resources. During LS1, Sim@P1 was one of the most productive GRID sites: it delivered more than 50 million CPU-hours and it generated more than 1.7 billion Monte Carlo events to various analysis communities within the ATLAS collaboration.
The particular design aspects are presented: the virtualization platform exploited by the Sim@P1 project permits to avoid interferences with TDAQ operations and, more important, it guarantees the security and the usability of the ATLAS private network. The Cloud infrastructure allows to decouple the needed support on both infrastructural (hardware, virtualization layer) and logical (Grid site support and handling the job lifecycle) levels. In particular in this note we focus on the operational aspects of such a large system for the upcoming LHC Run 2 period: customized, simple, reliable and efficient tools are needed to quickly switch from Sim@P1 to TDAQ mode and vice versa to exploit the TDAQ resources when they are not used for the data acquisition, even for short period. We also describe the evolution of the central Openstack infrastructure as it was upgraded from Folsom to Icehouse release and the scalability issues we have addressed.
The success of the Sim@P1 project is due to the continuous combined efforts of the ATLAS TDAQ SysAdmins and NetAdmins teams, CERN IT and the RHIC & ATLAS Computing Facility (RACF) at BNL.

Track 7 Session / 170

ATLAS@Home: Harnessing Volunteer Computing for HEP

Author: David Cameron[1]

Co-authors: Andrej Filipcic [2]; Claire Adam Bourdarios [3]; Eric Christian Lancon [4]; Wenjing Wu [5]

[1] *University of Oslo (NO)*

[2] *Jozef Stefan Institute (SI)*

[3] *Laboratoire de l'Accelerateur Lineaire (FR)*

[4] *CEA/IRFU,Centre d'etude de Saclay Gif-sur-Yvette (FR)*

[5] *Institute of High Energy Physics,Chinese Academy of Sciences (CN)*

Corresponding Author: david.cameron@cern.ch

A recent common theme among HEP computing is exploitation of opportunistic resources in order to provide the maximum statistics possible for Monte-Carlo simulation. Volunteer computing has been used over the last few years in many other scientific fields and by CERN itself to run simulations of the LHC beams. The ATLAS@Home project was started to allow volunteers to run simulations of collisions in the ATLAS detector. So far many thousands of members of the public have signed up to contribute their spare CPU cycles for ATLAS, and there is potential for volunteer computing to provide a significant fraction of ATLAS computing resources. Here we describe the design of the project, the lessons learned so far and the future plans.

Track 3 Session / 171

ATLAS I/O Performance Optimization in As-Deployed Environments

Author: David Malon[1]

[1] *Argonne National Laboratory (US)*

Corresponding Author: thomas.maier@cern.ch

I/O is a fundamental determinant in the overall performance of physics analysis and other data-intensive scientific computing. It is, further, crucial to effective resource delivery by the facilities and infrastructure that support data-intensive science. To understand I/O performance, clean measurements in controlled environments are essential, but effective optimization requires as well an understanding of the complicated realities of as-deployed environments. These include a spectrum of local and wide-area data delivery and resilience models, heterogeneous storage systems, matches and mismatches between data organization and access patterns, multi-user considerations that may help or hinder individual job performance, and more.

The ATLAS experiment has organized an interdisciplinary working group of I/O, persistence, analysis framework, distributed infrastructure, site deployment, and external experts to understand and improve I/O performance in preparation for Run 2 of the Large Hadron Collider. The adoption of a new analysis data model for Run 2 has afforded the collaboration a unique opportunity to incorporate instrumentation and monitoring from the outset.
This paper describes a program of instrumentation, monitoring, measurement, and data collection both in cleanroom and grid environments, and discusses how such information is propagated and employed.
The paper further explores how these findings inform decision-making on many fronts, including persistent data organization, caching, best practices, framework interactions with underlying service layers, and settings at many levels, from sites to application code. Related developments that increase robustness and resilience in the presence of faults by improving communication between frameworks and underlying infrastructure layers are also discussed.

Track 2 Session / 172

ATLAS Metadata Infrastructure Evolution for Run 2 and Beyond

Author: Jack Cranshaw[1]

Co-authors: David Malon [1]; Peter Van Gemmeren [1]

[1] *Argonne National Laboratory (US)*

Corresponding Author: peter.van.gemmeren@cern.ch

ATLAS developed and employed for Run 1 of the Large Hadron Collider a sophisticated infrastructure for metadata handling in event processing jobs. This infrastructure profits from a rich feature set provided by the ATLAS execution control framework, including standardized interfaces and invocation mechanisms for tools and services, segregation of transient data stores with concomitant object lifetime management, and mechanisms for handling occurrences asynchronous to the control framework's state machine transitions.

This metadata infrastructure is evolving and being extended for Run 2 to allow its use and reuse in downstream physics analyses, analyses that may or may not utilize the ATLAS control framework. At the same time, multiprocessing versions of the control framework and the requirements of future multithreaded frameworks are leading to redesign of components that use an incident-handling approach to asynchrony. The increased use of scatter-gather architectures, both local and distributed, requires further enhancement of metadata infrastructure in order to ensure semantic coherence and robust bookkeeping.

This paper describes the evolution of ATLAS metadata infrastructure for Run 2 and beyond, including the transition to dual-use tools—tools that can operate inside or outside the ATLAS control framework—and the implications thereof. It further examines how the design of this infrastructure is changing to accommodate the requirements of future frameworks and emerging event processing architectures.

173

Alignment of the ATLAS Inner Detector Upgraded for the LHC Run II

Author: Pawel Bruckman De Renstrom[1]

[1] *Polish Academy of Sciences (PL)*

Corresponding Author: javier.jimenez.pena@cern.ch

ATLAS is equipped with a tracking system built using different technologies, silicon planar sensors (pixel and micro-strip) and gaseous drift- tubes, all embedded in a 2T solenoidal magnetic field. For the LHC Run II, the system has been upgraded with the installation of a new pixel layer, the Insertable B-layer (IBL). Offline track alignment of the ATLAS tracking system has to deal with about 700,000 degrees of freedom (DoF) defining its geometrical parameters. The task requires using very large data sets and represents a considerable numerical challenge in terms of both CPU time and precision. The adopted strategy uses a hierarchical approach to alignment, combining local and global least squares techniques.
An outline of the track based alignment approach and its implementation within the ATLAS software will be presented. Special attention will be paid to integration to the alignment framework of the IBL, which plays the key role in precise reconstruction of the collider luminous region, interaction vertices and identification of long-lived heavy flavor states.
Techniques allowing to pinpoint and eliminate tracking systematics due to alignment as well as strategies to deal with time-dependent variations will be briefly covered. Performance from Cosmic Ray commissioning run and status from proton-proton collision in LHC Run II will be discussed.

174

ATLAS Tracking Detector Upgrade studies using the fast simulation engine (FATRAS)

Authors: Andreas Salzburger[1]; Noemi Calace[2]

[1] *CERN*

[2] *Universite de Geneve (CH)*

Corresponding Author: andreas.salzburger@cern.ch

The successful physics program of Run-1 of the LHC with the discovery of the higgs boson in 2012 has put a strong emphasis on design studies for future upgrades of the existing LHC detectors and for future accelerators. Ideas how to cope with instantaneous luminosities way beyond the current specifications of the LHC in future tracking detectors are emerging and need sufficiently accurate Monte Carlo simulation techniques to be evaluated. In ATLAS, testing alternative layouts through the full simulation and reconstruction chain is a work-intensive program, which is probably not worth following to full detail for all test layout configurations. We present a novel technique implemented for ATLAS that allows fast definitions of realistic detector layouts in the ATLAS TrackingGeometry library and run the fast simulation program (FATRAS), followed by a fast digitisation and reconstruction chain. The code is also implemented without dependencies on the ATLAS software in the context of the Future Collider Collaboration (FCC) framework.

Track 4 Session / 176

Event-Driven Messaging for Offline Data Quality Monitoring at ATLAS

Author: Peter Onyisi[1]

Co-authors: Eric Torrence [2]; Iurii Ilchenko [1]; James Frost [3]

[1] *University of Texas (US)*

[2] *University of Oregon (US)*

[3] *University of Oxford (GB)*

Corresponding Author: ponyisi@utexas.edu

During LHC Run 1, the information flow through the offline data quality monitoring in ATLAS relied heavily on chains of processes polling each other's outputs for handshaking purposes. This resulted in a fragile architecture with many possible points of failure and an inability to monitor the overall state of the distributed system. We report on the status of a project undertaken during the long shutdown to replace the ad hoc synchronization methods with a uniform message queue system. This enables the use of standard protocols to connect processes on multiple hosts; reliable transmission of messages between possibly unreliable programs; easy monitoring of the information flow; and the removal of inefficient polling-based communication.

177

Dual-use tools and systematics-aware analysis workflows in the ATLAS Run-II analysis model

Author: Steven Andrew Farrell[1]

Co-authors: Attila Krasznahorkay [2]; David Adams [3]; Eric Christian Lancon [4]; Iacopo Vivarelli [5]; Karsten Koeneke [6]; Markus Elsing [2]; Martin Woudstra [7]; Nils Erik Krumnack [8]; Paolo Calafiura [1]; Paul James Laycock [9]; Pierre-Antoine Delsart [10]; Sara Kristina Strandberg [11]; Wim Lavrijsen [1]; Wouter Verkerke [12]; Xiaowen Lei [13]

[1] *Lawrence Berkeley National Lab. (US)*

[2] *CERN*

[3] *Brookhaven National Laboratory (US)*

[4] *CEA/IRFU,Centre d'etude de Saclay Gif-sur-Yvette (FR)*

[5] *University of Sussex (GB)*

[6] *Albert-Ludwigs-Universitaet Freiburg (DE)*

[7] *University of Manchester (GB)*

[8] *Iowa State University (US)*

[9] *University of Liverpool (GB)*

[10] *Centre National de la Recherche Scientifique (FR)*

[11] *Stockholm University (SE)*

[12] *NIKHEF (NL)*

[13] *University of Arizona (US)*

Corresponding Author: steven.andrew.farrell@cern.ch

The ATLAS analysis model has been overhauled for the upcoming run of data collection in 2015 at 13 TeV. One key component of this upgrade was the Event Data Model (EDM), which now allows for greater flexibility in the choice of analysis software framework and provides powerful new features that can be exploited by analysis software tools. A second key component of the upgrade is the introduction of a dual-use tool technology, which provides abstract interfaces for analysis software tools to run in either the Athena framework or a ROOT-based framework. The tool interfaces, including a new interface for handling systematic uncertainties, have been standardized for the development of improved analysis workflows and consolidation of high-level analysis tools. This presentation will cover the details of the dual-use tool functionality, the systematics interface, and how these features fit into a centrally supported analysis environment.

179

PROOF-based analysis on ATLAS Italian Tiers with Prodsys2 and Rucio

Author: Roberto Di Nardo[1]

Co-authors: Elisabetta Vilucchi [2]; Gerardo Ganis [3]

[1] *Istituto Nazionale Fisica Nucleare - Laboratori Nazionali di Frascati (IT)*

[2] *Laboratori Nazionali di Frascati (LNF) - Istituto Nazionale di F*

[3] *CERN*

Corresponding Author: gerardo.ganis@cern.ch

During the LHC Run-1, Grid resources in ATLAS have been managed by the PanDA and DQ2 systems. In order to meet the needs for the LHC Run-2, Prodsys2 and Rucio are used as the new ATLAS Workload and Data Management systems.

The data are stored under various formats in ROOT files and end-user physicists have the choice to use either the ATHENA framework or directly ROOT. Within the ROOT data analysis framework it is possible to perform analysis of huge sets of ROOT files in parallel with PROOF on clusters of computers (usually organised in analysis facilities) or multi-core machines. In addition, PROOF-on-Demand (PoD) can be used to enable PROOF on top of an existing resource management system.

In this work, we present the first performances obtained enabling PROOF-based analysis in the Italian ATLAS Tier-1/Tier-2 sites within the new ATLAS workload system.

Benchmark tests of data access with the httpd protocol, using also the httpd redirector, will be shown. We also present results on the startup latency tests using the new PROOF functionality of dynamic workers addition, which improves the performance of PoD using Grid resources. These new results will be compared with the expected improvements discussed in a previous work

180

The migration of the ATLAS Metadata Interface (AMI) to Web 2.0

Author: Jerome Odier[1]

Co-authors: Fabian Lambert [1]; Jerome Fulachier [1]; Solveig Albrand [1]

[1] *Centre National de la Recherche Scientifique (FR)*

Corresponding Author: jerome.odier@cern.ch

The ATLAS Metadata Interface (AMI) can be considered to be a mature application because it has existed for at least 10 years. Over the last year, we have been adapting the application to some recently available technologies. The web interface, which previously manipulated XML documents using XSL transformations, has been migrated to Asynchronous Java Script (AJAX). Web development has been considerably simplified by the development of a framework for AMI based on JQuery and Twitter Bootstrap. Finally there has been a major upgrade of the python web service client.

181

Evolution of the Architecture of the ATLAS Metadata Interface (AMI)

Author: Jerome Odier[1]

Co-authors: Fabian Lambert [1]; Jerome Fulachier [1]; Osman AIDEL [2]; Solveig Albrand [1]

[1] *Centre National de la Recherche Scientifique (FR)*

[2] *CNRS / CC-IN2P3*

Corresponding Author: jerome.odier@cern.ch

The ATLAS Metadata Interface (AMI) can be considered to be a mature application because it has existed for at least 10 years. Over the years, the number of users and the number of functions provided for these users has increased.

It has been necessary to adapt the hardware infrastructure in a seamless way so that the Quality of Service remains high.

We will describe the evolution of the application from the initial one, using single server with a MySQL backend database, to the current state, where we use a cluster of Virtual Machines on the French Tier 1 Cloud at Lyon, an ORACLE database backend also at Lyon, with replication to CERN using ORACLE streams behind a back-up server.

Track 2 Session / 182

Implementation of the ATLAS Run 2 event data model

Author: Scott Snyder[1]

Co-authors: Attila Krasznahorkay [2]; Dag Gillberg [2]; Edward Moyse [3]; Karsten Koeneke [4]; Marcin Nowak [1]; Markus Elsing [2]; Peter Van Gemmeren [5]; Till Eifert [2]

[1] *Brookhaven National Laboratory (US)*

[2] *CERN*

[3] *University of Massachusetts (US)*

[4] *Albert-Ludwigs-Universitaet Freiburg (DE)*

[5] *Argonne National Laboratory (US)*

Corresponding Author: scott.snyder@cern.ch

During the 2013-2014 shutdown of the Large Hadron Collider, ATLAS switched to a new event data model for analysis, called the xAOD. A key feature of this model is the separation of the object data from the objects themselves (the auxiliary store'). Rather being stored as member variables of the analysis classes, all object data are stored separately, as vectors of simple values. Thus, the data are stored in astructure of arrays' format, while the user still can access it as an 'array of structures'. This organization allows for on-demand partial reading of objects, the selective removal of object properties, and the addition of arbitrary user-defined properties in a uniform manner. It also improves performance by increasing the locality of memory references in typical analysis code. The resulting data structures can be written to ROOT files with data properties represented as simple ROOT tree branches. This talk will focus on the design and implementation of the auxiliary store and its interaction with ROOT. Results on reconstruction and analysis performance will also be discussed.

Track 4 Session / 183

The ATLAS Event Service: A new approach to event processing

Author: Torre Wenaus[1]

Co-authors: Kaushik De [2]; Paolo Calafiura [3]; Paul Nilsson [1]; Peter Van Gemmeren [4]; Tadashi Maeno [1]; Vakho Tsulaia [3]; Wen Guan [5]

[1] *Brookhaven National Laboratory (US)*

[2] *University of Texas at Arlington (US)*

[3] *Lawrence Berkeley National Lab. (US)*

[4] *Argonne National Laboratory (US)*

[5] *University of Wisconsin (US)*

Corresponding Author: wenaus@gmail.com

The ATLAS Event Service (ES) implements a new fine grained approach to HEP event processing, designed to be agile and efficient in exploiting transient, short-lived resources such as HPC hole-filling, spot market commercial clouds, and volunteer computing. Input and output control and data flows, bookkeeping, monitoring, and data storage are all managed at the event level in an implementation capable of supporting ATLAS-scale distributed processing throughputs (about 4M CPU-hours/day). Input data flows utilize remote data repositories with no data locality or prestaging requirements, minimizing the use of costly storage in favor of strongly leveraging powerful networks. Object stores provide a highly scalable means of remotely storing the quasi-continuous, fine grained outputs that give ES based applications a very light data footprint on a processing resource, and ensure negligible losses should the resource suddenly vanish. We will describe the motivations for the ES system, its unique features and capabilities, its architecture and the highly scalable tools and technologies employed in its implementation, and its applications in ATLAS processing on HPCs, commercial cloud resources, volunteer computing, and grid resources.

184

Integration of Russian Tier-1 Grid Center with High Performance Computers at NRC-KI for LHC experiments and beyond HENP

Authors: Alexei Klimentov[1]; Alexey Poyda[2]; Eygene Ryabinkin[3]; Ruslan Mashinistov[4]

Co-authors: Alexander Belyaev [2]; Alexander Novikov [2]; Alexandra Berezhnaya [5]; Andrey Polyakov [2]; Anton Teslyuk [6]; Daniel Drizhuk [2]; Danila Oleynik [7]; Igor Tkachenko [2]; Ilya Lyalin [2]; Kaushik De [8]; Latchezar Betev [9]; Leonid Yasnopolskiy [2]; Predrag Buncic [9]; Yury Lazin [2]

[1] *Brookhaven National Laboratory (US)*

[2] *NATIONAL RESEARCH CENTRE "KURCHATOV INSTITUTE"*

[3] *National Research Centre Kurchatov Institute (RU); Moscow Institute for Physics and Technology, Applied computational geophysics lab*

[4] *NATIONAL RESEARCH CENTRE "KURCHATOV INSTITUTE"; P.N. Lebedev Institute of Physics (Russian Academy of Sciences)*

[5] *National Research Centre "Kurchatov Institute"*

[6] *RRC Kurchatov Institute*

[7] *University of Texas at Arlington (Texas, US); Joint Institute of Nuclear Research (Dubna, Russia)*

[8] *University of Texas at Arlington (US)*

[9] *CERN*

Corresponding Authors: poyda_aa@nrcki.ru, rea@grid.kiae.ru, ruslan.mashinistov@cern.ch

During LHC Run1 ATLAS and ALICE produced more than 30 Petabytes of data, That rate outstripped any other scientific effort going on, even in data-rich fields such as genomics and climate science. To address an unprecedented multi-petabyte data processing challenge, the experiments are relying on the computational grid infrastructure deployed by the Worldwide LHC Computing Grid (WLCG). LHC experiments preparing for the precision measurements and further discoveries that will be made possible by much higher LHC collision rates from early 2015 (Run2). The need for simulation, data processing and analysis and would overwhelm the expected capacity of WLCG computing facilities unless the range and precision of physics studies were to be curtailed. To meet this challenge the integration of the opportunistic resources into LHC computing model is highly important.

Tier-1 facility at Kurchatov Institute (NRC-KI) in Moscow is a part of WLCG and it will process and store up to 10% of total data obtained from ALICE, ATLAS and LHCb experiments. In addition Kurchatov Institute has supercomputers with peak performance 0.12 PFLOPS. Delegation of even a fraction of super-computing resources to the LHC Computing will notably increase total capacity.

In 2014, we have started a pioneer work to develop a portal combining a Tier-1 and a supercomputer in Kurchatov Institute. This portal is aimed to provide interfaces to run Monte-Carlo simulation at the Tier-1 Grid and supercomputer, using common portal and storage. PanDA (Production and Distributed Analysis) workload management system having great results at the ATLAS was chosen as underlying technology. In the nearest future we are planning to evolve the portal to fullscale data- and task- management system for federative heterogeneous resources.

The portal will be used not only for HEP, but also for other data-intensive sciences like biology with genome sequencing analysis; astrophysics with cosmic rays analysis, antimatter and dark matter search, etc. At this moment we have developed portal architecture, deployed prototype and run the first HENP and biology applications. Also we have deployed the local independent PanDA WMS instance in Kurchatov Institute adapted for the local supercomputer.

We want to present our current accomplishments with running PanDA WMS at NRC-KI supercomputers and use PanDA as a portal independent to the computing facilities infrastructure for High-Energy and Nuclear Physics as well as other data-intensive science applications.

185

Recent developments and upgrade to the Geant4 visualization Qt driver

Author: Laurent Garnier[1]

[1] *LAL-IN2P3-CNRS*

Corresponding Author: garnier@lal.in2p3.fr

Geant4 is a toolkit for the simulation of the passage of particles through matter. This object-oriented toolkit supports a variety of visualisation drivers including OpenGL, OpenInventor, HepRep, DAWN, VRML, RayTracer, gMocren and ASCIITree, with diverse and complementary functionalities.

In 2013, Gean4-MT1 has brought multi-threading to Geant4. The OpenGL suite of visualization drivers has been significantly improved by adding many functionalities, some specially for MT, in particular in the OpenGL Qt driver. Users may browse the scene tree and render the scene in multiple frames. The Qt driver is also able to treat multi-thread output separately and change view parameters in a graphic window. Migrating from old OpenGL (based on Display Lists) to modern OpenGL (based on Vertex Buffer Object) compliant code was not only a major implementation, but it also allowed us to include WebGL 2. Furthermore, there have been efforts to make the user interface more user friendly, but still allow users to customize it by adding menus, buttons, toolbars, viewpoints on scene and many others.

1 : http://http://geant4.web.cern.ch
2 : https://www.khronos.org/webgl/

Track 5 Session / 186

Overview of Different Exciting Technologies Being Researched in CERN openlab V

Author: Fons Rademakers[1]

[1] *CERN*

Corresponding Author: fons.rademakers@cern.ch

CERN openlab is a unique public-private partnership between CERN and leading ICT companies. Its mission is to accelerate the development of cutting-edge solutions to be used by the worldwide HEP community. Since January 2015 openlab phase V has started. To bring the openlab conducted research closer to the experiments, phase V has been changed to a project based structure which allows research being done not only in CERN IT but also directly in the different experiments or even in different research institutes. In this talk I will give an overview of the different exciting technologies being researched by CERN openlab V.

Track 7 Session / 187

Using S3 cloud storage with ROOT and CvmFS

Authors: Dirk Duellmann[1]; Maria Arsuaga Rios[1]; Seppo Heikkila[1]

Co-authors: Ben Couturier [1]; Jakob Blomer [1]; Rene Meusel [1]

[1] *CERN*

Corresponding Author: maria.arsuaga.rios@cern.ch

Amazon S3 is a widely adopted protocol for scalable cloud storage that could also fulfill storage requirements of the high-energy physics community. CERN has been evaluating this option using some key HEP applications such as ROOT and the CernVM filesystem (CvmFS) with S3 back-ends. In this contribution we present our evaluation based on two versions of the Huawei UDS storage system used from a large number of clients executing HEP software applications.
The performance of concurrently storing individual objects is presented alongside with more complex data access patterns as produced by the ROOT data analysis framework. We further report on the S3 integration with recent CvmFS versions and summarize the performance and pre-production experience with CvmFS/S3 for publishing daily releases of the full LHCb experiment software stack.

Track 3 Session / 188

Protocol benchmarking for HEP data access using HTTP and Xrootd

Authors: Adrien Devresse[1]; Fabrizio Furano[1]; Oliver Keeble[1]

[1] *CERN*

Corresponding Author: oliver.keeble@cern.ch

The DPM project offers an excellent opportunity for comparative testing of the HTTP and xroot protocols for data analysis.

- The DPM storage itself is multi-protocol, allowing comparisons to be performed on the same hardware
- The DPM has been instrumented to produce an i/o monitoring stream, familiar from the xrootd project, regardless of the protocol being used for access
- The continuous builds of DPM have been instrumented in 2013 with automated testing procedures that regularly produce, and collect metadata stress test information
- The DPM Collaboration involves a number of active grid sites who have made testing resources available. These sites are continuously exercised by a ROOT analysis benchmark several times per day, and comparisons can be conducted in a variety of environments, also involving access through Wide Area Network.

We present the results of our performance analyses of realistic use cases, and discuss their implications for the use of HTTP as a data and metadata access protocol for HEP.

189

Seamless access to HTTP/WebDAV distributed storage: the LHCb storage federation case study and prototype

Authors: Adrien Devresse[1]; Fabrizio Furano[1]; Stefan Roiser[1]

[1] *CERN*

Corresponding Author: stefan.roiser@cern.ch

In this contribution we describe the activities and the technical aspects that led to the construction of a public prototype for LHCb file access that is built on HTTP and WebDAV, supporting file access for distributed computing data management and data processing activities as well as seamless interactive access via web browsers.

The LHCb replica naming scheme provides characteristics that makes it an excellent candidate for evaluating the kind of interaction that an HTTP deployment can give on top of an existing, already running computing model.
Deploying HTTP access in the context of LHCb gives the possibility of accessing all 19 individual experiment storage areas, with 5 different disk-only and tape storage technologies deployed, using web tools like browsers, curl, wget and the Davix client.

These options allow a wide range of applications frameworks such as ROOT to access data, using techniques that may range from simple, single file transfers to vectored reads. Coordinating the deployment of WebDAV access in the sites has been one of the main efforts, and is still ongoing.

Given the very high level of coherency of the LHCb site namespaces, the quick pace at which sites joined the effort, and the desire to quickly setup an ambitious prototype, we decided to use the already existing storage federation testbed that is run at DESY, by configuring it with the HTTP/WebDAV endpoints of the LHCb storage elements as they joined the exercise.
We obtained a Dynamic Federations instance that is already covering a good part of the LHCb data namespace. The front-end shows an unified repository that is composed by the content of all the sites that it aggregates. The aggregation is performed on the fly, by distributing quick WebDAV queries towards the endpoints and scheduling, aggregating and caching the responses. We will further show how the HTTP protocol access will be incorporated into the LHCb/DIRAC distributed computing tool and can be beneficial for individual end users.

190

Docker experience at INFN-Pisa Grid Data Center

Author: Enrico Mazzoni[1]

Co-authors: Alberto Ciampa [2]; Daniele Bonacorsi [3]; Dario Fabiani [1]; Giuseppe Caruso [1]; Silvia Arezzini [4]; Simone Coscetti [5]; Tommaso Boccali [5]

[1] *INFN-Pisa*

[2] *Universita degli Studi di Pisa-INFN, Sezione di Pisa*

[3] *University of Bologna*

[4] *INFN Italy*

[5] *Sezione di Pisa (IT)*

Corresponding Author: silvia.arezzini@pi.infn.it

A large scale computing center, when not dedicated to a single/few users, has to face the problem of meeting ever changing user needs with respect to operating system version, architecture, availability of attached data volumes and logins. While clouds are a typical answer to these types of questions, they introduce resource problems like higher usage of RAM, difficulty to expose bare metal additional hardware, and are in general a few % slower. The Pisa computing center chose to use CHROOT long ago, in order to meet the needs of more than 50 user groups, with distinct needs. The CHROOTed system were started from a local tar image, preloaded on the host systems. We recently started investigating system distribution and setup via Docker, which does not cause any of the resource problems clouds do. We report here on our experience, and on the ease of its deployment on a 8000+ cores computing centers. Docker has also be used to deploy user specific services, like web portals, caching machines and web services. The data volume distribution is obtained via a 2+ PB GPFS system mounted on the host systems, which is attached to the CHROOTed machines whenever needed.

Track 4 Session / 191

Virtual Circuits in PhEDEx, an update from the ANSE project

Authors: Tony Wildish[1]; Vlad Lapadatescu[2]

Co-authors: Andrew Malone Melo [3]; Artur Jerzy Barczyk [2]; Azher Mughal [2]; Harvey Newman [2]; Iosif-Charles Legrand [2]; Paul Sheldon [3]; Ramiro Voicu [2]

[1] *Princeton*

[2] *California Institute of Technology (US)*

[3] *Vanderbilt University (US)*

Corresponding Author: tony.wildish@cern.ch

The ANSE project has been working with the CMS and ATLAS experiments to bring network awareness into their middleware stacks. For CMS, this means enabling control of virtual network circuits in PhEDEx, the CMS data-transfer management system. PhEDEx orchestrates the transfer of data around the CMS experiment to the tune of 1 PB per week spread over about 70 sites.

The goal of ANSE is to improve the overall working efficiency of the experiments, by enabling more deterministic time to completion for a designated set of data transfers, through the use of end-to-end dynamic virtual circuits with guaranteed bandwidth.

ANSE has enhanced PhEDEx, allowing it to create, use and destroy circuits according to it's own needs. PhEDEx can now decide if a circuit is worth creating based on its current workload and past transfer history, which allows circuits to be created only when they will be useful.

This paper reports on the progress made by ANSE in PhEDEx. We show how PhEDEx is now capable of using virtual circuits as a production-quality service, and describe how the mechanism it uses can be refactored for use in other software domains. We present first results of transfers between CMS sites using this mechanism, and report on the stability and performance of PhEDEx when using virtual circuits.

The ability to use dynamic virtual circuits for prioritised large-scale data transfers over shared global network infrastructures represents an important new capability and opens many possibilities. The experience we have gained with ANSE is being incorporated in an evolving picture of future LHC Computing Models, in which the network is considered as an explicit component.

Finally, we describe the remaining work to be done by ANSE in PhEDEx, and discuss future directions for continued development.

Track 6 Session / 192

Bandwidth-sharing in LHCONE, an analysis of the problem

Author: Tony Wildish[1]

[1] *Princeton University (US)*

Corresponding Author: tony.wildish@cern.ch

The LHC experiments have traditionally regarded the network as an unreliable resource, one which was expected to be a major source of errors and inefficiency at the time their original computing models were derived. Now, however, the network is seen as much more capable and reliable. Data are routinely transferred with high efficiency and low latency to wherever computing or storage resources are available to use or manage them.

Although there was sufficient network bandwidth for the experiments' needs during Run-1, they cannot rely on ever-increasing bandwidth as a solution to their data-transfer needs in the future. Sooner or later they need to consider the network as a finite resource that they interact with to manage their traffic, in much the same way as they manage their use of disk and CPU resources.

There are several possible ways for the experiments to integrate management of the network in their software stacks, such as the use of virtual circuits with hard bandwidth guarantees or soft real-time flow-control, with somewhat less firm guarantees. Abstractly, these can all be considered as the users (the experiments, or groups of users within the experiment) expressing a request for a given bandwidth between two points for a given duration of time. The network fabric then grants some allocation to each user, dependent on the sum of all requests and the sum of available resources, and attempts to ensure the requirements are met (either deterministically or statistically).

An unresolved question at this time is how to convert the users' requests into an allocation. Simply put, how do we decide what fraction of a network's bandwidth to allocate to each user when the sum of requests exceeds the available bandwidth? The usual problems of any resource-scheduling system arise here, namely how to ensure the resource is used efficiently and fairly, while still satisfying the needs of the users.

Simply fixing quotas on network paths for each user is likely to lead to inefficient use of the network. If one user cannot use their quota for some reason, that bandwidth is lost. Likewise, there is no incentive for the user to be efficient within their quota, they have nothing to gain by using less than their allocation.

As with CPU farms, some sort of dynamic allocation is more likely to be useful. A promising approach for sharing bandwidth at LHCONE is the 'Progressive Second-Price auction', where users are given a budget and are required to bid from that budget for the specific resources they want to reserve. The auction allows users to effectively determine among themselves the degree of sharing they are willing to accept based on the priorities of their traffic and their global share, as represented by their total budget. The network then implements those allocations using whatever mix of technologies is appropriate or available.

This paper describes how the Progressive Second-Price auction works and presents a prototype simulation of the auction. Results from the simulation are shown, addressing practical questions such as how are budgets set, what strategy should users use to manage their budget, how and how often should the auction be run, and how do we ensure that the goals of fairness and efficiency are met.

193

Dimensions of Data Management: A taxonomy of data-transfer solutions in ATLAS and CMS

Authors: Alessandro Di Girolamo[1]; Mario Lassnig[1]; Simone Campana[1]; Tony Wildish[2]; Vincent Garonne[1]

[1] *CERN*

[2] *Princeton University (US)*

Corresponding Author: tony.wildish@cern.ch

We present an abstract view of data-transfer architectures in use in ATLAS and CMS. We use this to classify data-transfer tools not in terms of their technology, but in terms of their more basic features, such as the properties of the traffic they handle and the use-cases they serve.

This classification moves the focus from programming interfaces and technologies back into the original problem-space, which enables a number of things:

- we can identify how existing tools can be re-factored according to common and technology-independent guidelines, which can facilitate re-use of code, both within and between experiments.

- we can identify design patterns which work well for certain areas of the architectural phase-space. This can lead to greater robustness through re-use of good design and shallower learning curves for developers moving from project to project.
- we can map new use-cases into the architectural phase-space independently of existing tools. This lets us see how components from existing tools could be used to solve new problems, which can lead to more rapid deployment of new solutions.
- we provide a common language for expressing experiment needs to external providers, such as storage element developers, FTS3 developers, or WLCG and OSG.

In this paper we show how data-transfer tools from both ATLAS and CMS fit into this architecture. We give examples of where this architecture would have improved our tool development in the past. Finally, we present a few sample use-cases for the future, and discuss them in terms of this architecture.

194

Understanding the CMS Tier-2 network traffic during Run-1

Author: Tony Wildish[1]

[1] *Princeton University (US)*

Corresponding Author: tony.wildish@cern.ch

At the beginning of Run-1 CMS was operating it's facilities according to the MONARC model, where data-transfers were strictly hierarchical in nature. Direct transfers between Tier-2 nodes was excluded, being perceived as operationally intensive and risky in an era where the network was expected to be a major source of errors. By the end of Run-1 wide-area networks were more capable and stable than originally anticipated. The original data-placement model was largely superceded, and traffic was allowed between Tier-2 nodes.

Tier-2 to Tier-2 traffic in 2012 already exceeded the amount of Tier-2 to Tier-1 traffic, so it clearly has the potential to become important in the future. Moreover, while Tier-2 to Tier-1 traffic is mostly upload of monte-carlo data, the Tier-2 to Tier-2 traffic represents data moved in direct response to requests from the physics analysis community. As such, problems or delays there are more likely to have a direct impact on the user community.

Tier-2 to Tier-2 traffic may also traverse parts of the WAN that are at the 'edge' of our network, with limited network capacity or reliability compared to, say, the Tier-0 to Tier-1 traffic which goes the over LHCOPN network.

CMS is looking to exploit technologies that allow us to interact with the network fabric so that it can manage our traffic better for us, this we hope to achieve before the end of Run-2. Tier-2 to Tier-2 traffic would be the most interesting use-case for such traffic management, precisely because it is close to the users' analysis and far from the 'core' network infrastructure.

As such, a better understanding of our Tier-2 to Tier-2 traffic is important. Knowing the characteristics of our data-flows can help us place our data more intelligently. Knowing how widely the data moves can help us anticipate the requirements for network capacity, and inform the dynamic data placement algorithms we expect to have in place for Run-2.

This paper presents our analysis of the Tier-2 traffic during Run 1. We examine the geographical and temporal distribution of traffic, the transfer quality, and the correlation between data-movement and physics analysis groups. We conclude with recommendations for improving data-placement in Run-2 and suggestions for monitoring improvements to help us better understand the changes in this behaviour as our system evolves.

196

Lightweight user and production job monitoring system for GlideinWMS

Author: Dmytro Kovalskyi[1]

[1] *Univ. of California San Diego (US)*

Corresponding Author: dmytro.kovalskyi@cern.ch

Glidemon is a lightweight monitoring system for GlideinWMS job submission infrastructure. It allows for basic information aggregation based on ClassAds in HTCondor environment of GlideinWMS. It can easily be adopted for a specific application running on top of GlideinWMS. In CMS it is used for user and production job monitoring managed by CRAB and WMAgent. We will review critical design choices that were made to deliver needed functionality keeping the system lightweight using the traditional relational database management systems.

197

Comprehensive Monitoring for Heterogeneous Geographically Distributed Storage.

Authors: Natalia Ratnikova[1]; Tony Wildish[2]

[1] *Fermilab*

[2] *Princeton University (US)*

Corresponding Author: natasha@fnal.gov

Storage capacity at CMS Tier-1 and Tier-2 sites reached over 100 Petabytes in 2014, and will be substantially increased during Run 2 data taking. The allocation of storage for the individual users analysis data, which is not accounted as a centrally managed storage space, will be increased to up to 40%. For comprehensive tracking and monitoring of the storage utilization across all participating sites, CMS developed a Space Monitoring system, which provides a central view of the geographically dispersed heterogeneous storage systems. The first prototype has been deployed at the pilot sites in summer 2014, and has been substantially reworked since then. In this presentation we discuss the functionality and our experience of system deployment and operation on the full CMS scale.

198

Implementation of an Upward-Going Muon Trigger for Indirect Dark Matter Searches with the NOvA Far Detector

Author: Robert Mina[1]

Co-authors: Andrew Norman [2]; Eric Fries [1]; Iuri Oksuzian [3]; Martin Frank [4]; Robert Group [5]

[1] *U. Virginia*

[2] *Fermilab*

[3] *UVa*

[4] *U*

[5] *University of Virginia*

Corresponding Author: ram2aq@virginia.edu

The NOvA collaboration has constructed a 14,000 ton, fine-grained, low-Z, total absorption tracking calorimeter at an off-axis angle to an upgraded NuMI neutrino beam. This detector, with its excellent granularity and energy resolution, and relatively low-energy neutrino thresholds was designed to observe electron neutrino appearance in a muon neutrino beam but it also has unique capabilities suitable for more exotic efforts. In fact, if an efficient upward-going muon trigger with sufficient cosmic ray background rejection can be demonstrated, NOvA will be capable of a competitive indirect dark matter search for low-mass WIMPs. The cosmic ray muon rate at the NOvA far detector is about 100 kHz and provides the primary challenge for triggering and optimizing such a search analysis The status of the NOvA upward-going muon trigger and initial sensitivity estimates for indirect detection of WIMP dark matter will be presented.

Track 6 Session / 199

100G Deployment@(DE-KIT)

Author: Bruno Heinrich Hoeft[1]

Co-author: Andreas Petzold [1]

[1] *KIT - Karlsruhe Institute of Technology (DE)*

Corresponding Author: bruno.hoeft@kit.edu

The Steinbuch Center for Computing (SCC) at Karlsruhe Institute of Technology (KIT) was involved quite early in 100G network technology. In 2010 already a first 100G wide area network testbed over a distance of approx. 450 km was deployed between the national research organizations KIT and FZ-Jülich - initiated by DFN (the German NREN). Only three years later 2013, KIT joined the Caltech SC13 100G "show floor" initiative using the transatlantic ANA-100G link to transfer LHC data from a storage at DE-KIT (GridKa) in Europe to hard disks at the show floor of SC13 in Denver (USA).
The network infrastructure of KIT as well as of the German Tier-1 installation DE-KIT (GridKa) however, is still based on 10Gbps. As highlighted in the contribution "Status and Trends in Networking at LHC Tier1 Facilities" to CHEP 2012, proactive investment is required at the Tier-1 sites. Bandwidth requirements will grow beyond the capacities currently available and the required upgrades are expected to be performed in 2015.
In close cooperation with DFN KIT is driving the upgrade from 10G to 100G. The process is divided into several phases, due to upgrade costs and different requirements in varying parts of the network infrastructure. The first phase will add a 100G interface to combine the interface connecting DE-KIT to LHCONE, where the highest demand for increased bandwidth is currently predicted. LHCONE is a routed virtual private network, connecting several Tier-[123] centers of WLCG and Belle-2. In phase number two, a second 100G interface will provide 100G symmetric interfaces to LHCONE. In phase number three, several of the routing interlinks of the Tier-1 center (DE-KIT) will receive an upgrade to 100G. KIT itself is still based on 10G, yet this will be upgraded in the next phase with two symmetric 100G uplinks. In the last phase, the router interface at KIT will receive a 100G upgrade at the required locations.
The requirements of the different phases as well as the planned topology will be presented. Some of the obstacles we discovered during the deployment will be discussed and solutions or workarounds presented.

Track 5 Session / 200

Recent Evolution of the Offline Computing Model of the NOvA Experiment

Authors: Andrew Norman[1]; Robert Group[2]

Co-author: Alec Habig [3]

[1] *Fermilab*

[2] *University of Virginia*

[3] *Univ. of Minnesota Duluth*

Corresponding Author: ahabig@umn.edu

The NOvA experiment at Fermilab is a long-baseline neutrino experiment designed to study nu-e appearance in a nu-mu beam. Over the last few years there has been intense work to streamline the computing infrastructure in preparation for data, which started to flow in from the far detector in Fall 2013. Major accomplishments for this effort include migration to the use of offsite resources through the use of the Open Science Grid and upgrading the file handling framework from simple disk storage to a tiered system using a comprehensive data management and delivery system to find and access files on either disk, dCache, or tape storage. NOvA has already produced more than 6.5 million files and more than 1 PB of raw data and Monte Carlo simulation generated files which are managed under this model. The current system has demonstrated sustained rates of up to 1 TB per hour of file transfer to permanent storage. NOvA pioneered the use of new tools and this paved the way for their use by other Intensity Frontier experiments at Fermilab. Most importantly, the new framework places the experiment's computing infrastructure on a firm foundation, which is ready to produce the files required for NOvA's first physics results. In this talk we discuss the offline computing model and infrastructure that has been put in place for NOvA and how we have used it to produce the experiment's first neutrino oscillation results.

201

Software framework testing at the Intensity Frontier

Author: Matthew Tamsett[1]

Co-author: Robert Group [2]

[1] *University of Sussex*

[2] *University of Virginia*

Corresponding Author: rcg6p@virginia.edu

The NOvA experiment at Fermilab is a long-baseline neutrino experiment designed to study nu-e appearance in a nu-mu beam. NOvA has already produced more than 1 million Monte Carlo and detector generated files amounting to more than 1 PB in size. This data is divided between a number of parallel streams such as far and near detector beam spills, cosmic ray backgrounds, a number of data-driven triggers and over 20 different Monte Carlo configurations.

Each of these data streams must be processed through the appropriate steps of the rapidly evolving, multi-tiered, interdependent NOvA software framework. In total there are greater than 12 individual software tiers, each of which performs a different function and can be configured differently depending on the input stream.

In order to regularly test and validate that all of these software stages are working correctly NOvA has designed a powerful, modular testing framework that enables detailed validation and benchmarking to be performed in a fast, efficient and accessible way with minimal expert knowledge.

The core of this system is a novel series of python modules which wrap, monitor and handle the underlying C++ software framework and then report the results to a slick front-end web-based interface. This interface utilises modern, cross-platform, visualisation libraries to render the test results in a meaningful way. They are fast and flexible, allowing us to cater, easily, for new tests and datasets.

In total upwards of 14 individual streams are regularly tested amounting to over 70 individual software processes, producing over 25 GB of output files. The rigour enforced through this flexible testing framework enables NOvA to rapidly verify configurations, results and software and thus ensure that data is available for physics analysis in a timely and robust manner.

Track 8 Session / 202

Matrix Element Method for High Performance Computing platforms

Authors: David Chamont[1]; Gilles Grasseau[2]

Co-authors: Florian Beaudette [2]; Lorenzo Bianchini [3]; Luca Mastrolorenzo [2]; Olivier Davignon [2]; Thomas Strebler [2]

[1] *LLR - École polytechnique*

[2] *Ecole Polytechnique (FR)*

[3] *Eidgenoessische Tech. Hochschule Zuerich (CH)*

Corresponding Author: chamont@llr.in2p3.fr

The Matrix Element Method (MEM) is a well known powerful approach in particle physics to extract maximal information of the events arising from the LHC pp collisions. Compared to other methods requiring trainings, the MEM allows direct comparisons between a theory and the observation. Since the phase space has a higher dimensionality to explore, MEM is much more CPU time consuming at the analysis level than classic methods. As a consequence, this method is hardly exploitable with sequential implementation, in particular, when one has to deal with channels with large backgrounds. For the upcoming LHC data-taking, this issue will become even more crucial.

The major challenge of this project is to provide a MEM implementation based on widely used standards such as MPI and OpenCL, which, taking advantage of multi-acccelerators and multi-node architectures, offers a drastic speed-up with low cost technologies.

In this talk, we will present how, in the context of the CMS experiment, we parallelized the MEM with the OpenCL abstract model and the memory distributed model (MPI layer). We will describe the necessary work on the different ingredients: the Parton Distribution Functions library (LHAPDF), the computation of the matrix element (MadGraph) and a small subset of ROOT tools. In addition, we will focus on the main tricky points to have access to an implementation of quality, e.g. uncorrelated random number sequences, minimizing synchronizations, minimizing the number of kernel calls. Finally, we will present the performance obtained on various platforms : CPUs, GPGPU, MIC, multi-GPGPUs or MICs, and cluster of GPU-nodes.

Track 3 Session / 203

Evolution of ATLAS conditions data and its management for LHC run-2

Author: Michael Boehler[1]

Co-authors: Andrea Formica [2]; Elizabeth Gallas [3]; Misha Borodin [4]; Voica Ana Maria Radescu [5]

[1] *Albert-Ludwigs-Universitaet Freiburg (DE)*

[2] *CEA/IRFU,Centre d'etude de Saclay Gif-sur-Yvette (FR)*

[3] *University of Oxford (GB)*

[4] *National Research Nuclear University MEPhI (RU)*

[5] *Deutsches Elektronen-Synchrotron (DE)*

Corresponding Author: michael.boehler@cern.ch

The ATLAS detector consists of several sub-detector systems. Both data taking and Monte Carlo (MC) simulation rely on an accurate description of the detector conditions from every sub system, such as calibration constants, different scenarios of pile-up and noise conditions, size and position of the beam spot, etc. In order to guarantee database availability for critical online applications during

data-taking, two database systems, one for online access and another one for all other database access have been implemented.

The long shutdown period has provided the opportunity to review and improve the run-1 system: revise workflows, include new and innovative monitoring and maintenance tools and implement a new database instance for run-2 conditions data.
The detector conditions are organized by tag identification strings and managed independently from the different sub-detector experts. The individual tags are then collected and associated into a global conditions tag, assuring synchronization of various sub-detector improvements. Furthermore, a new concept was introduced to maintain conditions over all the different data run periods into a single tag, by using Interval of Validity (IOV) dependent detector conditions for the MC database as well. This allows on-flight preservation of past conditions for data and MC and assure their sustainability with software evolution.

This contribution presents an overview of the commissioning of the new database instance, improved tools and workflows, and summarizes the actions taken during the run-2 commissioning phase beginning of 2015.

Track 4 Session / 204

The ATLAS Software Installation System v2: a highly available system to install and validate Grid and Cloud sites via Panda

Author: Alessandro De Salvo[1]

Co-authors: Arturo Sanchez Pineda [2]; Mayuko Kataoka [3]; Yuri Smirnov [4]

[1] *Universita e INFN, Roma I (IT)*

[2] *Universita di Napoli Federico II-Universita e INFN*

[3] *University of Texas at Arlington (US)*

[4] *Conseil Europeen Recherche Nucl. (CERN)-Unknown-Unknown*

Corresponding Author: alessandro.de.salvo@cern.ch

The ATLAS Installation System v2 is the evolution of the original system, used since 2003. The original tool has been completely re-designed in terms of database backend and components, adding support for submission to multiple backends, including the original WMS and the new Panda modules. The database engine has been changed from plain MySQL to Galera/Percona and the table structure has been optimized to allow a full High-Availability (HA) solution over WAN. The
servlets, running on each frontend, have been also decoupled from local settings, to allow an easy scalability of the system, including the possibility of an HA system with multiple sites. The clients can also be run in multiple copies and in different geographical locations, and take care of sending the installation and validation jobs to the target Grid or Cloud sites.
Moreover, the Installation DB is used as source of parameters by the automatic agents running in CVMFS, in order to install the software and distribute it to the sites.
The system is in production for ATLAS since 2013, having as main sites in HA the INFN Roma T2 and CERN AI. The LJSFi2 engine is directly interfacing with Panda for the Job Management, AGIS for the site parameter configurations, and CVMFS for both core components and the installation of the software itself.
LJSFi2 is also able to use other plugins, and is essentially VO-agnostic, so can be directly used and extended to cope with the requirements of any Grid or Cloud enabled VO.
In this work we'll present the architecture, performance, status and possible evolutions to the system for the LHC Run2 and beyond.

Track 4 Session / 205

The ATLAS Data Management system - Rucio: commissioning, migration and operational experiences

Authors: Cedric Serfon[1]; Vincent Garonne[1]

Co-authors: Alessandro Di Girolamo [1]; David Cameron [2]; Mario Lassnig [1]; Martin Barisits [1]; Ralph Vigne [3]; Thomas Beermann [4]; Tomas Kouba [5]; Wen Guan [6]

[1] *CERN*

[2] *University of Oslo (NO)*

[3] *University of Vienna (AT)*

[4] *Bergische Universitaet Wuppertal (DE)*

[5] *Acad. of Sciences of the Czech Rep. (CZ)*

[6] *University of Wisconsin (US)*

Corresponding Author: vincent.garonne@cern.ch

For more than 8 years, the Distributed Data Management (DDM) system of ATLAS called DQ2 has been able to demonstrate very large scale data management capabilities with more than 600M files, 160 petabytes spread worldwide across 130 sites, and accesses from 1,000 active users. However, the system does not scale for LHC run2 and a new DDM system called Rucio has been developed to be DQ2's successor. Rucio is based on different concepts and has new functionalities not provided by DQ2 which make the migration from the old to the new system a big challenge. The main issues are the large amount of data to move between the two systems, the number of users affected by the change, and the fact that the ATLAS Distributing Computing system, on the contrary to the sub-detectors, must stay continuously up and running during the LHC long shutdown to ensure the continuity of analysis and Monte-Carlo production. We will detail here the difficulties of this transition and will present the steps that were realized to ensure a smooth and transparent transition from DQ2 to Rucio. We will also discuss the new features and gains from the Rucio system.

206

Monitoring and controlling ATLAS data management: The Rucio web user interface

Authors: Mario Lassnig[1]; Thomas Beermann[2]

Co-authors: Cedric Serfon [1]; Martin Barisits [1]; Ralph Vigne [3]; Vincent Garonne [1]

[1] *CERN*

[2] *Bergische Universitaet Wuppertal (DE)*

[3] *University of Vienna (AT)*

Corresponding Author: mario.lassnig@cern.ch

The monitoring and controlling interfaces of the previous data management system DQ2 followed the evolutionary requirements and needs of the ATLAS collaboration. The new system, Rucio, has put in place a redesigned web-based interface based upon the lessons learnt from DQ2, and the increased volume of managed information. This interface encompasses both a monitoring and controlling component, and allows easy integration for user-generated views. The interface follows three design principles. First, the collection and storage of data from internal and external systems is asynchronous to reduce latency. This includes the use of technologies like ActiveMQ or Nagios. Second, analysis of the data into information is done massively parallel due to its volume, using a combined approach with an Oracle database and Hadoop MapReduce. Third, sharing of the information does not distinguish between human or programmatic access, making it easy to access selective parts of the information both in constrained frontends like web-browsers as well as remote services. This contribution will detail the reasons for these principles and the design choices taken. Additionally, the implementation, the interactions with external systems, and an evaluation of the system in production, both from a technological and user perspective, conclude this contribution.

Track 4 Session / 207

Resource control in ATLAS distributed data management: Rucio Accounting and Quotas

Author: Martin Barisits[1]

Co-authors: Cedric Serfon [1]; Mario Lassnig [1]; Ralph Vigne [2]; Thomas Beermann [3]; Vincent Garonne [1]

[1] *CERN*

[2] *University of Vienna (AT)*

[3] *Bergische Universitaet Wuppertal (DE)*

Corresponding Author: martin.barisits@cern.ch

The ATLAS Distributed Data Management system stores more than 160PB of physics data across more than 130 sites globally. Rucio, the next-generation data management system of ATLAS has been introduced to cope with the anticipated workload of the coming decade. The previous data management system DQ2 pursued a rather simplistic approach for resource management, but with the increased data volume and more dynamic handling of data workflows required by the experiment, a more elaborate approach to this issue is needed. This document describes how resources, like storage, accounts and replication requests, are accounted in Rucio. Especially the measurement of used logical storage space is fundamentally different in Rucio than it's predecessor DQ2. We introduce a new concept of declaring quota policies (limits) for accounts in Rucio. This new quota concept is based on accounts and RSE (Rucio storage element) expressions, which allows the definition of account limits in a dynamic way. This concept enables the operators of the data management system to establish very specific limits in which users, physics groups and production systems use the distributed data management system while, at the same time, lowering the operational burden. This contribution describes the architecture behind those components, the interfaces to other internal and external components and will show the benefits made by this system.

Track 3 Session / 208

The ATLAS EventIndex: architecture, design choices, deployment and first operation experience.

Author: Dario Barberis[1]

Co-authors: Alvaro Fernandez Casani [2]; Andrea Favareto [3]; Claudia Glasman [4]; David Malon [5]; Elizabeth Gallas [6]; Fedor Prokoshin [7]; JOSE SALT [8]; Jack Cranshaw [5]; Javier Sanchez [9]; Julius Hrivnac [10]; Rainer Toebbicke [11]; Ruijun Yuan [10]; Santiago González de la Hoz [12]

[1] *Università e INFN Genova (IT)*

[2] *Instituto de Fisica Corpuscular (ES)*

[3] *Università degli Studi e INFN Genova*

[4] *Universidad Autonoma de Madrid*

[5] *Argonne National Laboratory (US)*

[6] *University of Oxford (GB)*

[7] *Federico Santa Maria Technical University (CL)*

[8] *IFIC-VALENCIA*

[9] *Universidad de Valencia (ES)*

[10] *Laboratoire de l'Accelerateur Lineaire (FR)*

[11] *CERN*

[12] *IFIC-Valencia*

Corresponding Author: dario.barberis@cern.ch

The EventIndex is the complete catalogue of all ATLAS events, keeping the references to all files that contain a given event in any processing stage. It replaces the TAG database, which had been in use during LHC Run 1. For each event it contains its identifiers, the trigger pattern and the GUIDs of the files containing it. Major use cases are event picking, feeding the Event Service used on some production sites, and technical checks of the completion and consistency of processing campaigns. The system design is highly modular so that its components (data collection system, storage system based on Hadoop, query web service and interfaces to other ATLAS systems) could be developed separately and in parallel during LS1. The EventIndex is in operation for the start of LHC Run 2. This talk describes the high level system architecture, the technical design choices and the deployment process and issues. The performance of the data collection and storage systems, as well as the query services, will be reported.

Track 2 Session / 209

Optimisation of the ATLAS Track Reconstruction Software for Run-2

Author: Andreas Salzburger[1]

[1] *CERN*

Corresponding Author: andreas.salzburger@cern.ch

Track reconstruction is one of the most complex elements of the reconstruction of events recorded by ATLAS from collisions delivered by the LHC. It is the most time consuming reconstruction component in high luminosity environments. After a hugely successful Run-1, the flat budget projections for computing resources for Run-2 of the LHC together with the demands of reconstructing higher pile-up collision data at rates more than double those in Run-1 (an increase from 400 Hz to 1 kHz in trigger output) have put stringent requirements on the track reconstruction software. The ATLAS experiment has performed a two year long software campaign which aimed to reduce the reconstruction rate by a factor of three to meet the resource limitations for Run-2: a major part of the changes to achieve this were improvements to the track reconstruction software and will be presented in this contribution . The CPU processing time of ATLAS track reconstruction was reduced by more than a factor of three during this campaign without any loss of output information of the track reconstruction.
We present the methods used for analysing the tracking software and the code changes and new methods implemented to optimise both algorithmic performance and event data. Although most improvements were obtained without dedicated targetting concurrency strategies, major parts of the ATLAS tracking software were updated to allow for future improvements based on parallelism which will also be discussed.

210

A JEE Restful service to access Conditions Data in ATLAS

Author: Andrea Formica[1]

Co-author: Elizabeth Gallas [2]

[1] *CEA/IRFU,Centre d'etude de Saclay Gif-sur-Yvette (FR)*

[2] *University of Oxford (GB)*

Corresponding Author: andrea.formica@cern.ch

Usage of Conditions Data in ATLAS is extensive for offline reconstruction and analysis (for example: alignment, calibration, data quality). The system is based on the LCG Conditions Database infrastructure, with read and write access via an ad hoc C++ API (COOL), a system which was developed before Run 1 data taking began. The infrastructure dictates that the data is organized into separate schemas (assigned to subsystems/groups storing distinct and independent sets of conditions), making it difficult to access information from several schemas at the same time.

We have thus created PL/SQL functions containing queries to provide content extraction at multi-schema level. The PL/SQL API has been exposed to external clients by means of an intermediate java application server (JBoss), where an application delivering access to the DB via RESTful services has been deployed. The services allow navigation over multiple schema content, via simple URLs. The queried data can be retrieved either in XML or JSON formats, via simple clients (curl, wget or Web browsers).

211

Applying deep neural networks to HEP job statistics

Author: Lu Wang[None]

Corresponding Author: wanglu@ihep.ac.cn

The cluster of CC-IHEP is a middle sized computing system providing 10 thousands CPU cores, 3 PB disk storage, and 40 GB /s IO throughput. Its 1000+ users come from serials of HEP experiments including ATLAS, BESIII, CMS, DYB, JUNO, YBJ etc. In such a system, job statistics is necessary to find performance bottlenecks, locate software pitfalls, identify suspicious behaviors and make resource provisions, especially for new experiments whose computing modeling are still developing and refining.
To fulfill this requirement, we have developed and deployed a job statistics system which consists of an instrumenting agent, a central database, a data summarizer and a visualizer on the IHEP cluster. In the first half of 2014, the system has collected 1 million valid job records from BESIII experiment. Each job record includes static information from batch system, average efficiency from process manager and detailed IO parameters from VFS interfaces.To analyze this dataset we find that DNNs (Deep Neural Networks) is a useful technique for data classification and abnormity detection.
This paper demonstrates how we train a job classifier with DNNs. It firstly describes how we label the dataset semi-automatically from about 20% jobs samples which have hints of job type in their job-option-file names. Then some adapted data pre-processing steps will be presented. After that, it will describe the DNNs model which has achieved a precision of 96.6% with 240'000 labeled job samples (Ratio of training set and testing set is 7:3), 6 classes. It will also compare the results with those from a linear model and a MLPs (Multi-Layer Perceptrons) model. Impacts of meta-parameters including learning rate, batch size will be discussed. Examples of how we leverage the classification results to find software problems and detect abnormal job behaviors will be given at the last part of this paper.

Track 7 Session / 212

BESIII physical offline data analysis on virtualization platform

Author: Qiulan Huang[1]

Co-authors: Bowen Kan [2]; haibo li [3]; xiaofeng lei [3]

[1] *Chinese Academy of Sciences (CN)*

[2] *Institute of High Physics Chinese Academy of Sciences*

[3] *Chinese Academy of Sciences*

Corresponding Author: kanbw@ihep.ac.cn

Mass data processing and analysis contribute much to the development and discoveries of a new generation of High Energy Physics. The BESIII experiment of IHEP(Institute of High Energy Physics, Beijing, China) studies particles in the tau-charm energy region ranges from 2 GeV to 4.6 GeV, and requires massive storage and computing resources, which is a typical kind of data intensive application. With the rapid growth of experimental data, the data processing system encounters many problems, such as low resource utilization, complex migration and so on, which makes it urgent to transplant the data analysis system to a virtualization platform. However, offline software design, resource allocation and job scheduling of BESIII experiment are all based on physical machine. To solve those problems, we bring the virtualization technology of Openstack and KVM to BESIII computing system.

In this contribution we present an ongoing work which aims to make BESIII physical analysis work on virtualized resources to achieve higher resource utilization, dynamic resource management and higher job operating efficiency. Particularly, we discuss the architecture of BESIII offline software and the way to optimize the offline software to reduce the performance loss in virtualized environment by creating event index(event metadata) and do event pre-selection based on index, which significantly reduces the IO throughput and event numbers that need to do analysis, and then greatly improves the job processing efficiency. We also report the optimization of KVM from various factors in hardware and kernel including EPT (Extended Page Tables) and CPU affinity. Experimental results show the CPU performance penalty of KVM can be decreased to about 3%. This work is validated through real use cases of production BESIII jobs by working on physical slots and virtualized slots. In addition, the performance comparison between KVM and physical machines in aspect of CPU, disk IO and network IO is also presented. Finally, we describe our development work of adaptive cloud scheduler, which allocates and reclaims VMs dynamically according to the status of TORQUE queue and the size of resource pool to improve resource utilization and job processing efficiency.

Track 2 Session / 213

The NOvA Simulation Chain

Author: Adam Aurisano[1]

Co-authors: Alexandre Sousa [1]; Christopher Backhouse ; James Musser [2]; Nathan Mayer [3]; Raphael Schroeter [4]; Robert William Hatcher [5]; Ryan Patterson [6]

[1] *University of Cincinnati*

[2] *Indiana University*

[3] *Tufts University*

[4] *Harvard*

[5] *FNAL*

[6] *Caltech*

Corresponding Author: aurisano@fnal.gov

The NOvA experiment is a two-detector, long-baseline neutrino experiment operating in the recently upgraded NuMI muon neutrino beam. Simulating neutrino interactions and backgrounds requires many steps including: the simulation of the neutrino beam flux using FLUKA and the FLUGG interface; cosmic ray generation using CRY; neutrino interaction modeling using GENIE; and a simulation of the energy deposited in the detector using GEANT4. To shorten generation time, the modeling of detector-specific aspects, such as photon transport, detector and electronics noise, and readout electronics, employs custom, parameterized simulation applications. We will describe the NOvA simulation chain, and present details on the techniques used in modeling photon transport near the ends of cells, and in developing a novel data-driven noise simulation.

Due to the high intensity of the NuMI beam, the Near Detector samples a high rate of muons originating in the surrounding rock. In addition, due to its location on the surface at Ash River, MN, the Far Detector collects a large rate (~40 kHz) of cosmic muons. We will discuss the methods used in NOvA for overlaying rock muons and cosmic ray muons with simulated neutrino interactions and show how realistically the final simulation reproduces the preliminary NOvA data.

214

Data Handling with SAM and ART at the NOvA Experiment

Author: Adam Aurisano[1]

Co-authors: Andrew Norman [2]; Christopher Backhouse ; Dominick Rocco [3]; Gavin Davies [4]; Jan Zirnstein [3]; Marc Mengel [2]; Nathan Mayer [5]; Robert Illingworth [2]

[1] *University of Cincinnati*

[2] *Fermilab*

[3] *University of Minnesota*

[4] *Iowa State University*

[5] *Tufts University*

Corresponding Author: aurisano@fnal.gov

During operations, NOvA produces between 5,000 and 7,000 raw files per day with peaks in excess of 12,000. These files must be processed in several stages to produce fully calibrated and reconstructed analysis files. In addition, many simulated neutrino interactions must be produced and processed through the same stages as data. To accommodate the large volume of data and Monte Carlo, production must be possible both on the Fermilab grid and on off-site farms, such as the ones accessible through the Open Science Grid.

To handle the challenge of cataloging these files and to facilitate their off-line processing, we have adopted the SAM system developed at Fermilab. SAM indexes files according to metadata, keeps track of each file's physical locations, provides dataset management facilities, and facilitates data transfer to off-site grids.

To integrate SAM with the Fermilab's ART software framework and the NOvA production workflow, we have developed methods to embed metadata into our configuration files, ART files, and standalone ROOT files. A module in the ART framework propagates the embedded information from configuration files into ART files, and from input ART files to output ART files, allowing us to maintain a complete processing history within our files. Embedding metadata in configuration files also allows configuration files indexed in SAM to be used as inputs to Monte Carlo production jobs. Further, SAM keeps track of the input files used to create each output file. Parentage information enables the construction of self-draining datasets which have become the primary production paradigm used at NOvA. We will present an overview of SAM at NOvA and how it has transformed the file production framework used by the experiment.

Track 1 Session / 215

The Performance of the H.E.S.S. Target of Opportunity Alert System

Authors: Arnim Balzer[1]; Matthias Füßling[2]; Petter Hofverberg[3]; Robert Daniel Parsons[3]

Co-author: Collaboration H.E.S.S. [4]

[1] *Universiteit van Amsterdam*

[2] *DESY Zeuthen*

[3] *Max-Planck-Institut für Kernphysik*

[4] *H.E.S.S. Site Namibia*

Corresponding Author: arnim.balzer@uva.nl

The High Energy Stereoscopic System (H.E.S.S.) is an array of five imaging atmospheric Cherenkov telescopes located in the Khomas Highland in Namibia. Very high energy gamma rays are detected using the Imaging Atmospheric Cherenkov Technique. It separates the Cherenkov light emitted by

the background of mostly hadronic air showers from the light emitted by air showers induced by gamma rays. Using the fifth, larger telescope of the array with a huge mirror area of 600 m^2, it was possible to lower the energy threshold down to ≈ 30 GeV. With this unique ability to observe large amounts of gamma rays in the high energy gamma-ray regime (< 100 GeV) by using the large effective area of the fifth telescope at these energies, the H.E.S.S. experiment is ideally suited to observe short time scale transient events like gamma-ray bursts (GRBs). Originally detected by the Vela satellites in 1967, GRBs are among the most energetic processes in the known Universe. Extrapolating the spectrum of long duration GRBs (i.e. a GRB duration of the order of a few seconds or above) measured by current satellite experiments like Fermi, which measured gamma rays up to 95 GeV for GRB 130427A, a detection of these phenomena with the H.E.S.S. array is possible.
This presentation will give an update on the H.E.S.S. Target of Opportunity (ToO) alert system. It is used for an immediate and fully automatic response to a prompt GRB alert received via the Gamma-Ray Coordinates Network (GCN). The key feature of this system is a fast repointing of the whole array to a new observation position. The recent decrease of the response time to a ToO alert of more than 50 % achieved by improvements in hard- & software as well as the overall performance of the system will be discussed.

216

Accessing commercial cloud resources within the European Helix Nebula cloud marketplace

Author: Domenico Giordano[1]

Co-authors: Alessandro De Salvo [2]; Alessandro Di Girolamo [1]; Bob Jones [1]; Cristovao Jose Domingues Cordeiro [1]; Laurence Field [1]

[1] *CERN*

[2] *Universita e INFN, Roma I (IT)*

Corresponding Author: domenico.giordano@cern.ch

Helix Nebula –the Science Cloud Initiative is a public-private-partnership between Europe's leading scientific research organisations (notably CERN, EMBL and ESA) and European IT cloud providers, that aims to establish a cloud-computing platform for data intensive science within Europe.

Over the past two years, Helix Nebula has built a federated cloud framework –the Helix Nebula Marketplace (HNX) - to provision cloud services from a range of commercial cloud providers and public e-infrastructures, such as the EGI and GEANT. HNX delivers efficient and scalable access to cloud resources through an innovative broker technology deployed within the Helix Nebula project. In addition it complies with EU regulations and legislation to provide trusted cloud services.

CERN contributed to this initiative by providing a flagship use case: the integration of Helix Nebula cloud resources within the workload management system of the ATLAS experiment. Aiming to gain experience in managing and monitoring large-scale deployments as well as in benchmarking the cloud resources, a sizable Monte Carlo production was realized using the Helix Nebula platform. Thousands of concurrent virtual machines were run for several months in order to be comparable with the capacity supplied by a typical WLCG Tier-2 site.

This contribution describes the Helix Nebula initiative and summarizes the CERN experience and the lessons learned in deploying experiment applications with large cloud setups involving several commercial providers. Details about the procurement process, the cost analysis and the cloud benchmark are provided.

Track 1 Session / 217

A scalable monitoring for the CMS Filter Farm based on elasticsearch

Authors: Emilio Meschi[1]; Srecko Morovic[1]

Co-authors: Anastasios Andronidis [2]; Andre Georg Holzner [3]; Andrea Petrucci [1]; Attila Racz [1]; Aymeric Arnaud Dupont [1]; Benjamin Stieger [1]; Carlos Nunez Barranco Fernandez [1]; Christian Deldicque [1]; Christoph Paus [4]; Christoph Schwick [1]; Dominique Gigi [1]; Frank Glege [1]; Frans Meijers [1]; Georgiana Lavinia Darlea [4]; Guillelmo Gomez Ceballos Retuerto [4]; Hannes Sakulin [1]; James Gordon Branson [3]; Jan Veverka [4]; Jean-Marc Olivier Andre [5]; Jeroen Hegeman [1]; Konstanty Sumorok [4]; Lorenzo Masetti [1]; Luciano Orsini [1]; Marc Dobson [1]; Marco Pieri [3]; Olivier Chaze [1]; Penelope Amelia Roberts [6]; Petr Zejdl [1]; Remi Mommsen [5]; Salvatore Zaza [7]; Samim Erhan [8]; Sergio Cittolin [3]; Tomasz Adrian Bawej [9]; Ulf Behrens [10]; Vivian O'Dell [11]

[1] *CERN*

[2] *Aristotle Univ. of Thessaloniki (GR)*

[3] *Univ. of California San Diego (US)*

[4] *Massachusetts Inst. of Technology (US)*

[5] *Fermi National Accelerator Lab. (US)*

[6] *Nottingham Trent University (GB)*

[7] *Sezione di Pisa (IT)*

[8] *Univ. of California Los Angeles (US)*

[9] *University of Wisconsin (US)*

[10] *Deutsches Elektronen-Synchrotron (DE)*

[11] *Fermi National Accelerator Laboratory (FNAL)*

Corresponding Author: srecko.morovic@cern.ch

A flexible monitoring system has been designed for the CMS File-based Filter Farm making use of modern data mining and analytics components. All the metadata and monitoring information concerning data flow and execution of the HLT are generated locally in the form of small "documents" using the JSON encoding. These documents are indexed into a hierarchy of elasticsearch (es) clusters along with process and system log information. Elasticsearch is a search server based on Apache Lucene. It provides a distributed, multitenant-capable search and aggregation engine. Since es is schema-free, any new information can be added seamlessly and the unstructured information can be queried in non-predetermined ways. The leaf es clusters consist of the very same nodes that form the Filter Farm thus providing "natural"horizontal scaling. A separate "central" es cluster is used to collect and index aggregated information. The fine-grained information, all the way to individual processes, remains available in the leaf clusters. The central es cluster provides quasi-real-time high-level monitoring information to any kind of client. Historical data can be retrieved to analyse past problems or correlate them with external information. We discuss the design and performance of this system in the context of the CMS DAQ commissioning for LHC Run 2.

Track 1 Session / 218

File-based data flow in the CMS Filter Farm

Author: Emilio Meschi[1]

Co-authors: Anastasios Andronidis [2]; Andre Georg Holzner [3]; Andrea Petrucci [1]; Attila Racz [1]; Aymeric Arnaud Dupont [1]; Benjamin Stieger [1]; Carlos Nunez Barranco Fernandez [1]; Christian Deldicque [1]; Christoph Paus [4]; Christoph Schwick [1]; Dominique Gigi [1]; Frank Glege [1]; Frans Meijers [1]; Georgiana Lavinia Darlea [4]; Guillelmo Gomez Ceballos Retuerto [4]; Hannes Sakulin [1]; James Gordon Branson [3]; Jan Veverka [4]; Jean-Marc Olivier Andre [5]; Jeroen Hegeman [1]; Konstanty Sumorok [4]; Lorenzo Masetti [1]; Luciano Orsini [1]; Marc Dobson [1]; Marco Pieri [3]; Olivier Chaze [1]; Penelope Amelia Roberts [6]; Petr Zejdl [1]; Remi Mommsen [5]; Salvatore Zaza [7]; Samim Erhan [8]; Sergio Cittolin [3]; Srecko Morovic [1]; Tomasz Adrian Bawej [9]; Ulf Behrens [10]; Vivian O'Dell [11]

[1] *CERN*

[2] *Aristotle Univ. of Thessaloniki (GR)*

[3] *Univ. of California San Diego (US)*

[4] *Massachusetts Inst. of Technology (US)*

[5] *Fermi National Accelerator Lab. (US)*

[6] *Nottingham Trent University (GB)*

[7] *Sezione di Pisa (IT)*

[8] *Univ. of California Los Angeles (US)*

[9] *University of Wisconsin (US)*

[10] *Deutsches Elektronen-Synchrotron (DE)*

[11] *Fermi National Accelerator Laboratory (FNAL)*

Corresponding Author: emilio.meschi@cern.ch

During the LHC Long Shutdown 1, the CMS DAQ system underwent a partial redesign to replace obsolete network equipment, use more homogeneous switching technologies, and prepare the ground for future upgrades of the detector front-ends. The software and hardware infrastructure to provide input, execute the High Level Trigger (HLT) algorithms and deal with output data transport and storage has also been redesigned to be completely file-based. This approach provides a complete decoupling between the HLT algorithms and the input and output data flow. All the metadata needed for bookkeeping of the data flow and the HLT process lifetimes are also generated in the form of small "documents"using the JSON encoding, by either services in the flow of the HLT execution (for rates etc.) or watchdog processes. These "files" can remain memory-resident or be written to disk if they are to be used in another part of the system (e.g. for aggregation of output data). We discuss how this redesign improves to the robustness and flexibility of the CMS DAQ and the performance of the system currently being commissioned for the LHC Run 2.

Track 1 Session / 219

THE DAQ NEEDLE IN THE BIG-DATA HAYSTACK

Author: Emilio Meschi[1]

[1] *CERN*

Corresponding Author: emilio.meschi@cern.ch

Technology convergences in the post-LHC era

In the course of the last three decades HEP experiments have had to face the challenge of manipulating larger and larger masses of data from increasingly complex and heterogeneous detectors with hundreds of millions of electronic channels. The traditional approach of low-level data reduction using ad-hoc electronics working on fast analog signals, followed by global readout and digitisation, and a final stage of centralised processing in a more or less monolithic system has reached its limit before the LHC era. LHC experiments have been forced to turn to a distributed approach, leveraging the appearance of high speed switched networks developed for digital telecommunication and the internet. This has led to a generation of experiments where the use of custom electronics, analysing coarser-granularity analog or digital "fast"data, is limited to the first phase of triggering, where predictable latency and real time processing, as well as reliable, low-jitter clock and trigger distribution, are a necessity dictated by the limits of the front end readout buffers. Low speed monitoring (temperatures, pressures, etc.) and controls (thresholds, calibrations, voltages, etc.) have remained decoupled and considered an altogether separate realm in the detector design and operation.

We believe that it is now time for the HEP community to prepare for the next "revolution". Already, the mass of persistent data produced by e.g. the LHC experiments means that multiple pass end-to-end offline processing is becoming increasingly burdensome. Some experiments (e.g. ALICE) are moving towards a single-pass system for data reduction, relying on fast calibration feedback loops for zero suppression and low-level pattern recognition into the online system. The pristine "raw" channel readouts become thus volatile and no longer permanently stored. Others (e.g. LHCb) read out every channel for each beam crossing and delegate the entirety of the data reduction, reconstruction and selection to a fully software system. The latter approach is particularly attractive if low power techniques can be developed to counter the negative effects of the consequent increase in material budget for services and cooling in the active areas of the detector.

Further developments can be envisaged.
On the one hand very large scale integration paired with progress in radiation hard technologies, as well as the appearance of high-bandwidth bidirectional optical links, both on and off-silicon, could make intelligent very-front-end electronics and high-speed low-power readout a possibility already in the next decade, thus lifting strict latency limitations. At the same time, the traditional distinction between readout, trigger and control data channels will become increasingly artificial, paving the way to the possibility of running fully programmable algorithms at on- or near-detector electronics.
On the other hand, boosted by the "big data"industry, massively parallel and distributed analysis of unstructured data has become ubiquitous in commercial applications. Apart for their attractiveness for use in monitoring of both detector parameters and data flow, as well as data analysis, these technologies indicate in our opinion a possible evolutionary path for future DAQ and trigger architectures. In particular, a new trend is emerging from the data mining and analytics world which consists in "bringing the algorithm to the data". For HEP experiments, this might mean to abandon the consolidated paradigm represented by the triad low-level trigger - event building - high level trigger.
How close can we bring our algorithms to the detector ? Can we take advantage of the ideas, software (and hardware) technologies developed for data mining and search engines ? Can we imagine a future detector with extremely deep multi-stage, asynchronous or even virtual pipelines, where data streams from the various detector channels are analysed and indexed in quasi-real-time, and the final selection is operated as a distributed "search for interesting event parts"?
Can we push these ideas even further, removing the inflexible notion of pre-processed datasets and paving the way to completely new forms of selection and analysis, that can be developed, tested and implemented "online"as aggregation and reduction algorithms making use of the full, unstructured information from the experiment and directly returning the high-level physics quantities of interest ?
We investigate the potential impact of these different developments in the design of detector readout, trigger and data acquisition systems in the next decades.

220

Integration of the EventIndex with other ATLAS systems

Author: Elizabeth Gallas[1]

Co-author: Fedor Prokoshin [2]

[1] *University of Oxford (GB)*

[2] *Federico Santa Maria Technical University (CL)*

Corresponding Author: dario.barberis@cern.ch

The ATLAS EventIndex System, developed for use in LHC Run 2, is designed to index every processed event in ATLAS, replacing the TAG System used in Run 1. Its storage infrastructure, based on Hadoop, necessitates revamping how information in this system relates to other ATLAS systems. In addition, the scope of this new application is different from that of the TAG System. It will store fewer derived quantities, but store more indexes since the fundamental mechanisms for retrieving these indexes will be better integrated into all stages of processing, allowing more events from later stages of processing to be indexed than was possible with the previous system.
Connections with other systems are fundamentally critical to assess dataset completeness, identify data duplication, and check data integrity, but also needed to enhance user and system interfaces accessing information in EventIndex. This presentation will give an overview of the ATLAS systems involved, the relevant metadata, and describe the technologies we are deploying to complete these connections.

221

QuerySpaces on Hadoop for the ATLAS EventIndex

Author: Julius Hrivnac[1]

Co-authors: Andrea Favareto [2]; Claudia Glasman [3]; Jack Cranshaw [4]; Rainer Toebbicke [5]; Ruijun Yuan [1]

[1] *Laboratoire de l'Accelerateur Lineaire (FR)*

[2] *Università degli Studi e INFN Genova*

[3] *Universidad Autonoma de Madrid*

[4] *Argonne National Laboratory (US)*

[5] *CERN*

Corresponding Author: julius.hrivnac@cern.ch

The new ATLAS EventIndex catalogue uses a Hadoop cluster to store information on each event processed by ATLAS. Several tools belonging to the Hadoop eco-system are used to organise the data in HDFS, catalogue it internally, and provide the search functionality. This presentation will describe the Hadoop-based implementation of the adaptive query engine serving as the back-end for the ATLAS EventIndex. The QuerySpaces implementation handles both original data and search results providing fast and efficient mechanisms for new user queries using already accumulated knowledge for optimisation. Detailed description and statistics about user requests are collected in HBase tables and HDFS files. Requests are associated to their results and a graph of relations between them is created to be used to find the most efficient way of providing answers to new requests. The environment is completely transparent to users and is accessible over several command-line interfaces, a Web Service and a programming API.

Track 3 Session / 222

Distributed Data Collection for the ATLAS EventIndex.

Author: Javier Sanchez[1]

Co-authors: Alvaro Fernandez Casani [1]; Javier Sanchez [2]; Santiago Gonzalez De La Hoz [3]

[1] *Instituto de Fisica Corpuscular (ES)*

[2] *Universidad de Valencia (ES)*

[3] *IFIC-Valencia*

Corresponding Author: javier.sanchez@cern.ch

The ATLAS EventIndex contains records of all events processed by ATLAS, in all processing stages. These records include the references to the files containing each event (the GUID of the file) and the internal "pointer"to each event in the file. This information is collected by all jobs that run at Tier-0 or on the Grid and process ATLAS events. Each job produces a snippet of information for each permanent output file. This information is packed and transfered to a central broker at CERN using an ActiveMQ messaging system, and then is unpacked, sorted and reformatted in order to be stored and catalogued into a central Hadoop server. This talk describes in detail the Producer/Consumer architecture to convey this information from the running jobs through the messaging system to the Hadoop server.

Track 5 Session / 223

A study on dynamic data placement for the ATLAS Distributed Data Management system

Author: Thomas Beermann[1]

Co-authors: Graeme Stewart [2]; Peter Mattig [1]

[1] *Bergische Universitaet Wuppertal (DE)*

[2] *University of Glasgow (GB)*

Corresponding Author: thomas.beermann@cern.ch

This contribution presents a study on the applicability and usefulness of dynamic data placement methods for data-intensive systems, such as ATLAS distributed data management (DDM). In this system the jobs are sent to the data, therefore having a good distribution of data is significant. Ways of forecasting workload patterns are examined which then are used to redistribute data to achieve a better overall utilisation of computing resources and to reduce waiting time for jobs before they can run on the grid. This method is based on a tracer infrastructure that is able to monitor and store historical data accesses and which is used to create popularity reports. These reports provide detailed summaries about data accesses in the past, including information about the accessed files, the involved users and the sites. From this past data it is possible to then make near-term forecasts for data popularity in the future. This study evaluates simple prediction methods as well as more complex methods like neural networks. Based on the outcome of the predictions a redistribution algorithm deletes unused replicas and adds new replicas for potentially popular datasets. Finally, a grid simulator is used to examine the effects of the redistribution. The simulator replays workload on different data distributions while measuring the job waiting time and site usage. The study examines how the average waiting time is affected by the amount of data that is moved, how it differs for the various forecasting methods and how that compares to the optimal data distribution.

224

Scalable and fail-safe deployment of the ATLAS Distributed Data Management system Rucio

Authors: Mario Lassnig[1]; Ralph Vigne[2]

Co-authors: Cedric Serfon [1]; Martin Barisits [1]; Thomas Beermann [3]; Vincent Garonne [1]

[1] *CERN*

[2] *University of Vienna (AT)*

[3] *Bergische Universitaet Wuppertal (DE)*

Corresponding Author: mario.lassnig@cern.ch

This contribution details the deployment of Rucio, the ATLAS Distributed Data Management system. The main complication is that Rucio interacts with a wide variety of external services, and connects globally distributed data centres under different technological and administrative control, at an unprecedented data volume. It is therefore not possibly to create a duplicate instance of Rucio for testing or integration. Every software upgrade or configuration change is thus potentially disruptive and requires fail-safe software and automatic error recovery. Rucio uses a three-layer scaling and mitigation strategy based on quasi-realtime monitoring. This strategy mainly employs independent stateless services, automatic failover, and service migration. The technologies used for deployment and mitigation include OpenStack, Puppet, Graphite, HAProxy, Apache, and nginx. In this contribution, the reasons and design decisions for the deployment, the actual implementation, and an evaluation of all involved services and components are discussed.

Track 4 Session / 225

AsyncStageOut: Distributed user data management for CMS Analysis

Authors: Diego Ciangottini[1]; Hassen Riahi[2]; Tony Wildish[3]

Co-authors: Andres Jorge Tanasijczuk [4]; Edward Karavakis [2]; Eric Vaandering [5]; Jose Hernandez [6]; Julia Andreeva [2]; Justas Baclas [7]; Marco Mascheroni [8]

[1] *Universita e INFN (IT)*

[2] *CERN*

[3] *Princeton University (US)*

[4] *Univ. of California San Diego (US)*

[5] *Fermi National Accelerator Lab. (US)*

[6] *CIEMAT, Madrid, Spain*

[7] *DiSCC, Vilnius University, Lithuania*

[8] *Universita & INFN, Milano-Bicocca (IT)*

Corresponding Author: tony.wildish@cern.ch

AsyncStageOut (ASO) is a new component of the distributed data analysis system of CMS, CRAB, designed for managing users' data. It addresses a major weakness of the previous model, namely that data movement was part of the job execution resulting in inefficient use of job slots and an unacceptable failure rate at the end of the jobs.

ASO foresees the management of up to 400k files per day of various sizes, spread worldwide across more than 60 sites. It must handle up to 1000 individual users per month, and work with minimal delay. This creates challenging requirements for system scalability, performance and monitoring.

ASO uses FTS to schedule and execute the transfers between the storage elements of the source and destination sites. It has evolved from a limited prototype to a highly adaptable service, which manages and monitors the user file placement and bookkeeping. To ensure system scalability and data monitoring, it employs new technologies such as a NoSQL database and re-uses existing components of PhEDEx and the FTS Dashboard.

We present the asynchronous stage-out strategy and the architecture of the solution we implemented to deal with those issues and challenges. The deployment model for the high availability and scalability of the service is discussed. The performance of the system during the commissioning and the first phase of production are also shown, along with results from simulations designed to explore the limits of scalability.

Track 7 Session / 226

Cloud Federation - the new way to build distributed clouds

Author: Marek Kamil Denis[1]

[1] *CERN*

Corresponding Author: marek.denis@cern.ch

Cloud federation brings an old concept into new technology, allowing for sharing resources between independent cloud installations. Cloud computing starts to play major role in HEP and e-science allowing resources to be obtained on demand. Cloud federation supports sharing between independent organizations and companies coming from the commercial world such as public clouds, bringing new ways of joint collaboration along with security and usability.

From the OpenStack Icehouse release, cloud federation, based on the SAML2 protocol, is now integrated with the standard open source tree and becoming a reality. Examples of this configuration both command line and web based use of providers such as EduGain or connections to public clouds such as Rackspace.

This presentation will demonstrate what functionalities OpenStack brings, what are the main concepts, how it was designed, how federation with the SAML ECP extension allows use of REST and CLIs, and what are our plans to use real hybrid clouds - identity federation, optimal image sharing and networking layer which would allow for creating distributed vLANs.

227

A prototype Infrastructure for Cloud-based distributed services in High Availability over WAN

Authors: Alessandra Doria[1]; Alessandro De Salvo[2]; Silvio Pardi[3]; Vincenzo Capone[4]

Co-authors: Carlo Graziosi [5]; Cristina Bulfon [6]; Enrico Pasqualucci [7]; Giampaolo Carlino [8]; Lorenzo Puccio [9]; Massimo Carboni [9]; Paolo Bolletta [9]

[1] *Univ. + INFN*

[2] *Universita e INFN, Roma I (IT)*

[3] *INFN*

[4] *DANTE*

[5] *INFN Sez. Roma1*

[6] *Dipartim.di Fisica G.Marconi Rome I-Universita di Roma I "La Sap*

[7] *INFN Roma*

[8] *Universita e INFN (IT)*

[9] *GARR - the Italian National Research and Education Network*

Corresponding Authors: alessandro.de.salvo@cern.ch, spardi@na.infn.it

The advancements in technologies on provisioning end-to-end network services over geographical networks, together with the consolidation of Cloud Technologies, allow the creation of innovative scenarios for data centers.

In this work, we present the architecture and performance studies concerning a prototype of distributed Tier2 infrastructure for HEP, instantiated between the two Italian sites of INFN-Roma1 and INFN-Napoli.

The Network Infrastructure is based on a Layer-2 geographical link, provided by the Italian NREN (GARR), directly connecting the two remote LANs of the named sites. By exploiting the possibilities offered by the new distributed file systems, a shared storage area with synchronous copy has been set up. The Computing Infrastructure, based on a OpenStack facility, is using a set of distributed Hypervisors installed in both sites. The Cloud is then using the common distributed storage facility to allow High Availability (HA) and Crash Recovery.
The result is a strongly coupled system, enabling the creation, management and migration of Virtual Machines in both sites.

The main parameter to be taken into account when managing two remote sites with a single framework is the the effect of the latency, due to the distance and the end-to-end service overhead. In order to understand the capabilities and limits of our setup, the impact of latency has been investigated by means of a set of stress tests, including data I/O throughput, metadata access performance evaluation and network occupancy, during the life cycle of a Virtual Machine. A set of resilience tests has been also performed, in order to verify the stability of the system on event of hardware or software fault.

The results of this work show that the reliability and robustness of the chosen architecture are effective enough to build a real system and to provide common services. This prototype can also be extended to multiple sites, by changing the network topology, thus creating a National Network of Cloud-based distributed services in HA over WAN.

Track 3 Session / 228

Data preservation for the HERA Experiments @ DESY using dCache technology

Authors: Dirk Kruecker[1]; Karsten Schwank[1]

Co-authors: Birgit Lewendel [1]; David South [1]; Patrick Fuhrmann [1]

[1] *DESY*

Corresponding Author: karsten.schwank@desy.de

We report on the status of the data preservation project at DESY for the HERA experiments and present the latest design of the storage which is a central element for bit-preservation. The HEP experiments based at the HERA acceleerator at DESY collected large and unique datasets during the period 1992 to 2007. In addition, corresponding Monte Carlo simulation datasets were produced, which are significantly larger by volume and still being added to as the final analyses are completed. As part of the ongoing DPHEP data preservation efforts at DESY, these data sets must be transferred into storage systems that guarantee a reliable long term access. At the same time, given that the experiments are still active, easy access to the data must be guaranteed for the coming years.
The long term storage system is two-fold: an archive part where the data exists on two tape copies and an online part where the full dataset can be kept available and allows easy access to all HERA data. The archive and online parts are physically separate. The demanding aspect of this data is not only the size of about 1PB but also the large number (about 4 million) of files and the broad range of file sizes from a few KB to a few hundred of GB. To achieve a high level of reliability, we use the dCache distributed storage solution and make use of its replication capabilities and tape interfaces. We describe the dCache installation with tape backend that is used as mass storage together with an newly introduced small files service that allows for the automatic creation of tape friendly container files, containing many single (small) files. From the user's point of view, this is done in a fully transparent way in terms of creation and access to the data.

Track 5 Session / 229

Dynamic Data Management for the Distributed CMS Computing System

Author: Christoph Paus[1]

[1] *Massachusetts Inst. of Technology (US)*

Corresponding Author: paus@mit.edu

The Dynamic Data Management (DDM) framework is designed to manage the majority of the CMS data in an automated fashion. At the moment 51 CMS Tier-2 data centers have the ability to host about 20 PB of data. Tier-1 centers will also be included adding substantially more space. The goal of DDM is to facilitate the management of the data distribution and optimize the accessibility of data for the physics analyses. The basic idea is to replicate highly demanded data and remove data replicas that are not often used. There are four distinct parts that work together on achieving this goal: newly created datasets are inserted into the managed data pool, highly popular datasets are replicated across multiple data centers, data centers are cleaned of the least popular data, and temporarily unavailable sites are accounted by creating replicas of missing datasets. We are also developing various metrics that show whether the system behaves as expected. The CMS Popularity Service is the authoritative source of the popularity information – CPU time, number of accesses, number of users/day, etc. of datasets, blocks and files. The information is collected from various resources namely directly from the CMS Software, from the CMS Remote Analysis Builder (CRAB) and from files accessed via xrootd. It is subsequently summarized and exposed using a REST API to other CMS services for further use. This paper describes the architecture of these components and details the experience of commissioning and running the system over the last year.

Track 7 Session / 230

The diverse use of clouds by CMS

Author: David Colling[1]

Co-authors: Adam Huffman [1]; Alison McCrae [2]; Andrew David Lahiff [3]; Anthony Tiradani [4]; Claudio Grandi [5]; Daniela Bauer [1]; Dirk Hufnagel [6]; Massimo Sgaravatto [7]; Olivier Chaze [2]

[1] *Imperial College Sci., Tech. & Med. (GB)*

[2] *CERN*

[3] *STFC - Rutherford Appleton Lab. (GB)*

[4] *Fermilab*

[5] *INFN - Bologna*

[6] *Fermi National Accelerator Lab. (US)*

[7] *Universita e INFN (IT)*

Corresponding Author: d.colling@imperial.ac.uk

The resources CMS is using are increasingly being offered as clouds. In Run 2 of the LHC the majority of CMS CERN resources, both in Meyrin and at the Wigner Computing Centre, will be presented as cloud resources on which CMS will have to build its own infrastructure. This infrastructure will need to run all of the CMS workflows including: Tier 0, production and user analysis. In addition, the CMS High Level Trigger will provide a compute resource comparable in scale to the total offered by the CMS Tier 1 sites, when it is not running as part of the trigger system. During these periods a cloud infrastructure will be overlaid on this resource, making it accessible for general CMS use. Finally, CMS is starting to utilise cloud resources being offered by individual institutes and is gaining experience to facilitate the use of opportunistically available cloud resources.

Track 4 Session / 232

Quantitative transfer monitoring for FTS3

Authors: Hassen Riahi[1]; Michail Salichos[1]; Oliver Keeble[1]

Co-authors: Alejandro Alvarez Ayllon [1]; Alessandro Di Girolamo [1]; Julia Andreeva [1]; Michal Kamil Simon [2]; Nicolo Magini [3]; Stefan Roiser [1]

[1] *CERN*

[2] *A.D.A.M. Applications of Detectors and accelerators to Medicine*

[3] *Fermi National Accelerator Lab. (US)*

Corresponding Author: oliver.keeble@cern.ch

The overall success of LHC data processing depends heavily on stable, reliable and fast data distribution. The Worldwide LHC Computing Grid (WLCG) relies on the File Transfer Service (FTS) as the data movement middleware for moving sets of files from one site to another.

This paper describes the components of FTS3 monitoring infrastructure and how they are built to satisfy the common and particular requirements of the LHC experiments. We show how the system provides a complete and detailed cross-virtual organization (VO) picture of transfers for sites, operators and VOs. This information has proven critical due to the shared nature of the infrastructure, allowing a complete view of all transfers on shared network links between various workflows and VOs using the same FTS transfer manager.

We also report on the performance of the FTS service itself, using data generated by the aforementioned monitoring infrastructure both during the commissioning and the first phase of production. We also explain how this monitoring information and network metrics produced can be used both as

a starting point for troubleshooting data transfer issues, but also as a mechanism to collect information such as transfer efficiency between sites, achieved throughput and its evolution over time, most common errors, etc, and take decision upon them to further optimize transfer workflows.

The service setup is subject to sites policies to control the network resource usage, as well as all the VOs making use of the Grid resources at the site to satisfy their requirements. FTS3 is the new version of FTS and has been deployed in production in August 2014.

233

Software Development at Belle II

Author: Thomas Kuhr[1]

Co-author: Thomas Hauth [2]

[1] *KIT - Karlsruhe Institute of Technology (DE)*

[2] *KIT*

Corresponding Author: thomas.hauth@cern.ch

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. As a consequence also the simulation, reconstruction, and analysis software has to be upgraded substantially and actually most parts are newly written taking into account the experience from Belle and other experiments and advances in technology.

While large parts of the Belle software were written by KEK staff members, the Belle II software development is distributed around the world and relies on contributions from students and staff members. The regional distribution, the different (cultural) backgrounds and skills of developers are a challenge for a project that has to make sure that all parts fit together. Several tools and organizational measures are employed to allow the developers to focus on their main work and at the same time keep them conscious of the overall goal of providing an easy-to-use, robust, and powerful software suite to the Belle II collaboration.

Track 3 Session / 234

dCache, evolution by tackling new challenges.

Authors: Albert Rossi[1]; Christian Bernardt[2]; Dmitry Litvintsev[1]; Gerd Behrmann[3]; Karsten Schwank[4]; Patrick Fuhrmann[4]; Paul Millar[2]; Tigran Mkrtchyan[4]

[1] *FNAL*

[2] *Deutsches Elektronen-Synchrotron (DE)*

[3] *NDGF*

[4] *DESY*

Corresponding Author: patrick.fuhrmann@desy.de

With the great success of the dCache Storage Technology in the framework of the World Wide LHC Computing Grid, an increasing number of non HEP communities were attracted to use dCache for their data management infrastructure. As a natural consequence, the dCache team was presented with new use-cases that stimulated the development of interesting dCache features.

Perhaps the most important group of new features is the enhanced media awareness. One aspect is the optimized migration of data between random access devices and tertiary storage, e.g. tape systems. Transparently for the user, dCache combines small files into containers before being copied

to tape. Another aspect of this media awareness work is dCache's activity in making more efficient use of SSDs to boost high speed writing and chaotic reading: depending on access profile or protocol, data is placed on the most appropriate media types.

A second hot topic, often requested by scientific communities, is Cloud Storage. By marrying the OwnCloud software with dCache, a unique hybrid system becomes available that provides both the simplicity of a sync-n-share service and dCache's many unique data management features. Beyond simple sync-n-share, users also demand control over the quality of service dCache offers. To support this, dCache is implementing the CDMI standard. With CDMI, dCache can present new functionality in a standard fashion; e.g. storing and querying metadata, triggering media migration or treating dCache as an object store.

As the X509 certificate infrastructure proved unpopular outside of HEP sciences, dCache will support alternative methods of authenticating, like SAML and OpenID Connect. Such support allows sites running dCache to join identity federations as part of international collaborations.

The dCache team will describe its ongoing activities and present its future development road map.

235

DEAP-3600 Data Acquisition System

Author: Thomas Lindner[1]

Co-authors: Ben Smith [1]; Fabrice Retiere [1]; Pierre Gorel [2]; Pierre-Andre Amaudruz [1]

[1] *TRIUMF*

[2] *University of Alberta*

Corresponding Author: lindner@triumf.ca

DEAP-3600 is a dark matter experiment located at SNOLAB in Ontario, Canada. The DEAP detector uses 3600kg of liquid argon to search for the interactions of Weakly Interacting Massive Particles (WIMPs), a dark matter candidate. Light from the WIMP interactions is imaged using an array of 255 PMTs. A critical challenge for the DEAP experiment is the large background from 39Argon beta decays which occur at a rate of 3.6kHz. Beta decays are efficiently eliminated by pulse shape discrimination.

The DEAP Data AcQuisition (DAQ) system has been designed to handle the very large 39Ar event rate without compromising the detection of other interaction occurring at a much lower rate. Key features of the DEAP DAQ include: i) factor of 5 to 10 online 39Ar event rejection using a custom trigger board by analyzing 22 waveforms sampled at 45MS/s in an FPGA housed in a custom board, ii) digitization and pulse processing of the PMT signals using commercial CAEN 250MS/s FADCs iii) Data acquisition software designed using MIDAS DAQ toolkit providing online pulse feature extraction and event filtering capabilities. The full DAQ is capable of providing a factor of 50 to 100 data rate reduction by filtering out beta decays during regular data taking and the capability to handling data rates of 200-300MB/s during calibration runs

236

Evolution of the T2K-ND280 Computing Model

Author: Thomas Lindner[1]

[1] *TRIUMF*

Corresponding Author: lindner@triumf.ca

ND280 is the off-axis near detector for the T2K neutrino experiment; it is designed to characterize the unoscillated T2K neutrino beam and measure neutrino cross-sections. We have developed a complicated system for processing and simulating the ND280 data, using computing resources from North America, Europe and Japan.
Recent work has concentrated on unifying our computing framework across these regions. In particular, we have started using two HEP tools: GANGA and DIRAC. GANGA has provided a unified framework for users scripts; using a unified scripting framework makes it easier for a smaller number of people to maintain and run a multiple types of production scripts. DIRAC is a framework providing a suite of tools for HEP computing. In particular, we have been testing DIRAC as a Workload Manager Server (WMS), replacing the original EGI WMS.
We will describe our experiences with these developments of the ND280 computing model.

Track 4 Session / 237

Integrating Network Awareness in ATLAS Distributed Computing Using the ANSE Project

Author: Kaushik De[1]

Co-authors: Artem Petrosyan [2]; Jorge Batista [3]; Shawn Mc Kee [4]

[1] *University of Texas at Arlington (US)*

[2] *Joint Inst. for Nuclear Research (RU)*

[3] *U*

[4] *University of Michigan (US)*

Corresponding Author: alexei.klimentov@cern.ch

A crucial contributor to the success of the massively scaled global computing system that delivers the analysis needs of the LHC experiments is the networking infrastructure upon which the system is built. The experiments have been able to exploit excellent high-bandwidth networking in adapting their computing models for the most efficient utilization of resources.

New advanced networking technologies now becoming available such as software defined networking hold the potential of further leveraging the network to optimize workflows and dataflows, through proactive control of the network fabric on the part of high level applications such as experiment workload management and data management systems. End to end monitoring of networks using perfSONAR combined with data flow performance metrics further allows applications to adapt based on real time conditions.

We will describe efforts underway in ATLAS on integrating network awareness at the application level, particularly in workload management, building upon the ANSE (Advance Network Services for Experiments) project components. We will show how knowledge of network conditions, both historical and current, are used to optimize PanDA and other systems for ATLAS and describe how software control of end-to-end network paths can augment ATLAS's ability to effectively utilize its distributed resources.

238

The evolution of CERN EDMS

Authors: Aleksandra Wardzinska[1]; Stephan Petit[1]

[1] *CERN*

Corresponding Author: aleksandra.wardzinska@cern.ch

Large-scale long-term projects such as the LHC require the ability to store, manage, organize and distribute large amounts of engineering information, covering a wide spectrum of fields. This information is a living material, evolving in time, following various lifecycles. It has to reach the next generations of engineers so they understand how their predecessors designed, crafted, operated and maintained the most complex machines ever built.

This is the role of CERN EDMS. The Engineering and Equipment Data Management Service has served the High Energy Physics Community for over 15 years. It is the CERN's official PLM (Product Lifecycle Management), supporting engineering communities in their collaborations inside and outside the laboratory. EDMS is integrated with the CAD (Computer-aided Design) and CMMS (Computerized Maintenance Management) systems used at CERN providing tools for engineers who work in different domains and who are not PLM specialists.

Over the years, human collaborations and machines grew in size and complexity. So did the EDMS: it is currently home to more than 2 million files and documents, and has over 6 thousand active users. In April 2014 we released a new major version of EDMS, featuring a complete makeover of the web interface, improved responsiveness and enhanced functionality. Following the results of user surveys and building upon feedback received from key users group, we brought what we think is a system that is more attractive and makes it easy to perform complex tasks.

In this paper we will describe the main functions and the architecture of EDMS. We will discuss the available integration options, which enable further evolution and automation of engineering data management. We will also present our plans for the future development of EDMS.

239

BESIII production with distributed computing

Author: Xiaomei Zhang[1]

Co-authors: Tian Yan [1]; Zhao Xianghu [1]

[1] *Institute of High Energy Physics*

Corresponding Author: zhangxm@ihep.ac.cn

Distributed computing is necessary nowadays for high energy physics experiments to organize heterogeneous computing resources all over the world to process enormous amounts of data.
The BESIII experiment in China, which has aggregated about 3 PB of data over the last 5 years, has established its own distributed computing system, based on DIRAC, as a supplement to local clusters, collecting cluster, grid, desktop and cloud resources from collaborating member institutes around the world. The system consists of workload management and data management to deal with the BESIII Monte Carlo production workflow in a distributed environment. Random trigger data are distributed with a dataset-based data transfer system at a speed of 10 TB/day to enable processing of the whole Monte Carlo simulation and reconstruction workflow at remote sites. File and metadata management tools and a job submission frontend have been developed to provide a virtual layer for BESIII physicists to use distributed resources.
The BESIII distributed computing system, which is composed of more than 3000 CPU cores across 10 sites, with about 400TB storage, has been in production since the end of 2012. Three large-scale production tasks have been completed, with more than 150,000 jobs completed successfully. Measures have been taken to cope with lack of grid experience and low manpower at grid sites, and in particular, monitoring has been strengthened to deal with these issues. Moreover, the paper shows our experience of integrating various kinds of private cloud resources in a dynamic way to greatly ease maintenance of sites. Our efforts to extend the platform to more new high energy experiments in China are also discussed.

240

Setup of a resilient FTS3 service at GridKa

Author: Thomas Hartmann[1]

Co-authors: Andreas Petzold [1]; Kamil Wisniewski [2]; Ludmilla Obholz [3]; Preslav Borislavov Konstantinov [4]

[1] *KIT - Karlsruhe Institute of Technology (DE)*

[2] *Karlsruhe Institut for Technology*

[3] *KIT Karlsruhe*

[4] *Bulgarian Academy of Sciences (BG)*

Corresponding Author: thomas.hartmann@cern.ch

The FTS service provides a transfer job scheduler to distribute and replicate waste amounts of data over the heterogeneous WLCG infrastructures. The most recent version FTS3 simplifies and improves the flexibility compared to the channel model of the previous incarnations while reducing the load to the service components. The improvements allow to handle a higher number of transfers with a single FTS3 setup. Covering now continent-wide transfers compared to the previous version handling only transfers related to specific clouds, a resilient system becomes even more necessary with the increased number of depending users.

Having set up a FTS3 services at the German T1 site *GridKa* at *KIT* in Karlsruhe, we present our experiences on the preparations for a high-availability FTS3 service. Trying to avoid single points of failure, we rely on a database cluster as fault tolerant data back-end and the FTS3 service deployed on an own cluster setup to provide a resilient infrastructure for the users. With the database cluster providing a basic resilience for the data back-end, we ensure on the FTS3 service level a consistent and reliable database access through a proxy solution. On each FTS3 node a HAproxy instance is monitoring the health of each database node and distributes database queries over the whole cluster for load balancing during normal operations; in case of a broken database node, the proxy excludes it transparently to the local FTS3 service. The FTS3 service itself consists of a main and a backup instance, which takes over the identity of the main instance, i.e., IP, in case of an error using a CTDB infrastructure offering clients a consistent service.

241

HappyFace as a monitoring tool for the ATLAS experiment

Author: Haykuhi Musheghyan[1]

Co-authors: Arnulf Quadt [1]; Erekle Magradze [1]; Gen Kawamura [1]; Jordi Nadal Serrano [1]

[1] *Georg-August-Universitaet Goettingen (DE)*

Corresponding Author: haykuhi.musheghyan@cern.ch

The importance of monitoring on HEP grid computing systems is growing due to a significant increase in their complexities. Computer scientists and administrators have been studying and building effective ways to gather information on and clarify a status of each local grid infrastructure.
The HappyFace project aims at making the above-mentioned workflow possible. It aggregates, processes and stores the information and the status of different HEP monitoring resources into the common database of HappyFace. The system displays the information and the status through a single interface.
However, this model of HappyFace relied on the monitoring resources which are always under development in the HEP experiments. Consequently, HappyFace needed to have direct access methods to the grid application and grid service layers in the different HEP grid systems. To cope with this issue, we use a reliable HEP software repository, the CernVM File System. We propose a new implementation and an architecture of HappyFace, the so-called grid-enabled HappyFace. It allows its basic framework to connect directly to the grid user applications and the grid collective services, without involving the monitoring resources in the HEP grid systems.
This approach gives HappyFace several advantages: Portability, to provide an independent and generic monitoring system among the HEP grid systems. Functionality, to allow users to perform

various diagnostic tools in the individual HEP grid systems and grid sites. Flexibility, to make HappyFace beneficial and open for the ATLAS computing environments.
Different grid-enabled modules, to view datasets of the ATLAS Distributed Data Management system (DDM), to connect to the Ganga job monitoring system and to check the performance of grid transfers among the grid sites, have been implemented. The new HappyFace system has been successfully integrated and now it displays the information and the status of both the monitoring resources and the direct access to the grid user applications and the grid collective services in the ATLAS computing system.

Track 2 Session / 242

Geant4 VMC 3.0

Author: Ivana Hrivnacova[1]

Co-author: Andrei Gheata [2]

[1] *IPNO, Université Paris-Sud, CNRS/IN2P3*

[2] *CERN*

Corresponding Author: ivana@ipno.in2p3.fr

Virtual Monte Carlo (VMC) provides an abstract interface into Monte Carlo transport codes. A user VMC based application, independent from the specific Monte Carlo codes, can be then run with any of the supported simulation programs. Developed by the ALICE Offline Project and further included in ROOT, the interface and implementations have reached stability during the last decade and have become a foundation for other detector simulation frameworks, the FAIR facility experiments framework being among the first and largest.

Geant4 VMC, which provides the implementation of the VMC interface for Geant4, is in continuous maintenance and development, driven by the evolution of Geant4 on one side and requirements from users on the other side. Besides the implementation of the VMC interface, Geant4 VMC also provides a set of examples that demonstrate the use of VMC to new users and also serve for testing purposes. Since major release 2.0, it includes the G4Root navigator package, which implements an interface that allows one to run a Geant4 simulation using a ROOT geometry.

The release of Geant4 version 10.00 with the integration of multi-threading processing has triggered the development of the next major version of Geant4 VMC (version 3.0), whose release is planned for this year. A beta version, available for user testing since March, has helped its consolidation and improvement. We will review the new capabilities introduced in this major version, in particular the integration of multi-threading into the VMC design, its impact on the Geant4 VMC and G4Root packages, and the introduction of a new package, MTRoot, providing utility functions for ROOT parallel output in independent files with necessary additions for thread-safety. Migration of user applications to multi-threading that preserves the ease of use of VMC will be also discussed. We will also report on the introduction of a new CMake based build system, the migration to ROOT major release 6 and the improvement of the testing suites.

243

ROOT/RooFIT optimizations for Intel Xeon Phi's - first results

Author: Jan Justinus Keijser[1]

[1] *NIKHEF*

Corresponding Author: janjust@nikhef.nl

In the past, grid worker nodes struck a reasonable balance between the number of cores, the amount of available memory, available diskspace and the maximum network bandwidth. This led to an operating model where worker nodes were "carved up" into single core job slots, each of which would execute a HEP workload job. Typical worker nodes would have up to 16 computing cores , with roughly 2 GB RAM and 32 GB scratch space per job slot.
In the last few years the number of threads and cores per processor has risen rapidly. With the advent of multi-core, many-core and GPU computing the number of cores is expected to rise even more. This leads to an imbalance on modern worker nodes, where the "job per core" model is no longer a valid option.
In view of these development the desire, or perhaps even the necessity, to run programs in parallel has also increased. However, the process of transforming the HEP software to run in parallel has been slow and painful.

One of the more recent development in the GPU and many-core computing field has been the introduction of the Intel Many Integrated Core architecture, known now as the Intel Xeon Phi. The Xeon Phi is a PCI-Express adapter with up to 16 GB of RAM on board and with up to 61 cores, each capable of running 4 threads. The Xeon Phi has a theoretical performance of more than 1 TFLOPS.

Initial attempts to port HEP code to the Xeon Phi has not produced the expected results and often a performance *decrease* has been seen, instead of an increase. This has also been reported by non-HEP researchers, and slowly the reasons for this are becoming known. Using the proper optimization techniques it is possible to achieve 1 TFLOPS, but this requires substantial effort in rewriting and optimizing the application. The main bottleneck here is that the Xeon Phi should not be regarded "61 independent cores" but instead it should be seen as a vector instruction machine. By properly vectorizing an application the performance can improve significantly. This also applies when porting an application to GPUs.

In this talk we will focus on porting and optimizing ROOT, and in particular RooFIT to the Intel Xeon Phi. A zero-order port does indeed show a decrease in performance, but by analyzing and optimizing the 'hot spots' in the code the performance can be improved. It is vital to validate the results of the ported application, however: the Xeon Phi version of RooFIT must produce the same results as the single-core version.

Most interesting is the fact that code optimized for the Xeon Phi also performs better on "regular" Xeon processors. As part of the results we will also discuss the performance increase of the optimized code on an Xeon E5 processor. In that way the Xeon Phi porting effort can be seen as a catalyst to transform the HEP code to run in parallel on modern CPUs.

The main goal of this research and of this talk is to provide a direction for optimizing the HEP software for modern and future computing platforms. The main goal of parallelizing RooFIT is not to make it scale to 60 cores or more, but to find a way to restore the imbalance seen in modern worker nodes. If a relatively simple code optimization leads to HEP software that can scale to roughly 10 cores then the "core imbalance" can be restored. Worker nodes could then be carved up into "10 core job slots". This is a different approach from true High Performance Computing centers, who are usually interested in getting an application to scale beyond 32 cores and more.

Track 8 Session / 244

A new Self-Adaptive disPatching System for local cluster

Author: Bowen Kan[1]

Co-author: Jingyan Shi [2]

[1] *Institute of High Physics Chinese Academy of Sciences*

[2] *IHEP*

Corresponding Author: kanbw@ihep.ac.cn

Scheduler is one of the most important components of high performance cluster. This paper introduces a self-adaptive dispatching system (SAPS) based on torque/maui which increases the resources utilization of cluster effectively and guarantees the high reliability of the computing platform. It provides great convenience for users to run various tasks on the computing platform. First of all, the SAPS implements the GPU scheduling with multi-core. This provides the basis for effective integration and utilization of computing resources, improves the ability of the cluster computing greatly. Secondly, SAPS analysis the relationship between the number of jobs queueing and the idle resources left, tune the priority of users'job dynamically. In this way, more resources are provided for jobs running and less resources idle. Thirdly, integrated the on-line error detection with work nodes, the SAPS can excluded error nodes and include the recovered nodes automatically.
In addition, SAPS provides a monitoring management with fine granularity, a comprehensive scheduling accounting module and a scheduling real-time alarm function, and all of those ensure the cluster runs more high-efficiently, and reliably.
Currently, the SAPS has been running stable on IHEP local cluster (more than 10,000 cores and 30,000 jobs every day) and resource utilization has been improved more than 26%, and the SAPS has reduced costs for both administrator and users greatly.

Track 5 Session / 246

Optimising Costs in WLCG Operations

Authors: Alessandra Forti[1]; Andrea Sciaba[2]; Jose Flix Molina[3]; Maria Alandes Pradillo[2]; Maria Dimou[2]

[1] *University of Manchester (GB)*

[2] *CERN*

[3] *Centro de Investigaciones Energ. Medioambientales y Tecn. - (ES*

Corresponding Author: andrea.sciaba@cern.ch

The Worldwide LHC Computing Grid project (WLCG) provides the computing and storage resources required by the LHC collaborations to store, process and analyse the ~50 Petabytes of data annually generated by the LHC. The WLCG operations are coordinated by a distributed team of managers and experts and performed by people at all participating sites and from all the experiments. Several improvements in the WLCG infrastructure have been implemented during the first long LHC shutdown to prepare for the increasing needs of the experiments during Run2 and beyond. However, constraints in funding will affect not only the computing resources but also the available effort for operations. This paper presents the results of a detailed investigation on the allocation of the effort in the different areas of WLCG operations, identifies the most important sources of inefficiency and proposes viable strategies for optimising the operational cost, taking into account the current trends in the evolution of the computing infrastructure and the computing models of the experiments.

Track 8 Session / 247

Performance benchmark of LHCb code on state-of-the-art x86 architectures

Author: Rainer Schwemmer[1]

Co-authors: Daniel Hugo Campora Perez [1]; Niko Neufeld [1]

[1] *CERN*

Corresponding Author: rainer.schwemmer@cern.ch

For Run 2 of the LHC, LHCb is exchanging a significant part of its event filter farm with new compute nodes. For the evaluation of the best performing solution, we have developed a method to convert our high level trigger application into a stand-alone, bootable benchmark image. With additional instrumentation we turned it into a self-optimising benchmark which explores techniques such as late forking, NUMA balancing and optimal number of threads, i.e. it automatically optimises box-level performance. We have run this procedure on a wide range of Haswell-E CPUs and numerous other architectures from both Intel and AMD, including also the latest Intel micro-blade servers. We present results in terms of performance, power consumption, overheads and relative cost.

248

Open access for ALICE analysis based on virtualisation technology

Author: Mihaela Gheata[1]

Co-authors: Predrag Buncic [2]; Yves Schutz [3]

[1] *ISS - Institute of Space Science (RO)*

[2] *CERN*

[3] *IN2P3 (FR) and CERN*

Corresponding Author: mihaela.gheata@cern.ch

Open access is one of the prerequisites for long term data preservation for a HEP experiment. To guarantee the usability of data analysis tools over long periods of time it is crucial that third party users from the scientific community have access to the data and associated software. The ALICE Collaboration has developed a layer of lightweight components built on top of virtualisation technology to hide the complexity and details of the experiment-specific software. Users can perform basic analysis tasks within CernVM, a lightweight generic virtual machine, paired with an ALICE specific contextualisation. Once the virtual machine is launched, a graphics user interface is automatically started without any additional configuration. This interface allows to download the base ALICE analysis software and run a set of ALICE analysis modules. Currently the available tools include fully documented tutorials for ALICE analysis, such as the measurement of strange particle production or the nuclear modification factor in Pb-Pb collisions. The interface can be easily extended to include an arbitrary number of additional analysis modules. We will present the current status of the tools used by ALICE through the CERN open access portal, our first user experience and the plans for future extensions of this system.

249

Analysis of Public CMS Data on the VISPA Internet Platform

Author: Robert Fischer[1]

Co-authors: Benjamin Fischer [1]; Christian Glaser [2]; Christoph Welling [1]; Daniel Van Asseldonk [1]; Fabian Heidemann [1]; Gero Müller [1]; Marcel Rieger [1]; Martin Urban [1]; Thorben Quast [1]

[1] *Rheinisch-Westfaelische Tech. Hoch. (DE)*

[2] *RWTH Aachen*

Corresponding Author: robert.fischer@cern.ch

Within CERN's new open data portal, the CMS collaboration provides a substantial fraction of its recorded data to the public. To explore and analyse the data, computing resources, an analysis framework, and documentation are required as well. While scientists can use C++ and the experiment software CMSSW in virtual machines, a simpler approach is needed, e.g. for university students who are in the process of becoming familiar with data analyses in particle physics. The VISPA Internet platform provides interactive access to CMS public data together with analysis examples. Here, the Python programming language requires only moderate programming skills. Computing resources are offered, such that CMS analyses can be developed and executed from any device with a standard web browser. We report on the concept of the Internet platform, and present experience gained on the worldwide usage of the public CMS data.

250

ATLAS user analysis on private cloud resources at GoeGrid

Authors: Fabian Glaser[1]; Jordi Nadal Serrano[1]

Co-authors: Arnulf Quadt [1]; Jens Grabowski [2]

[1] *Georg-August-Universitaet Goettingen (DE)*

[2] *Georg-August-Universitaet Goettingen*

Corresponding Author: fabian.glaser@cern.ch

User analysis job demands can exceed available computing resources, especially before major conferences. ATLAS physics results might be slowed down due to this lack of resources available. For these reasons, cloud R&D activities are now included in the skeleton of the ATLAS computing model, which has been extended by using resources from commercial and private cloud providers to satisfy the demand. However, most of these activities are focused on Monte-Carlo production jobs, extending the resources at Tier-2. To evaluate the suitability of the cloud-computing model for user analysis jobs, we developed a framework to launch an ATLAS user analysis cluster in a cloud infrastructure on demand and evaluated two solutions. The first solution is totally integrated in the Grid infrastructure by using the same mechanism, which is already in use at Tier-2: A designated PandaQueue is monitored and additional worker nodes are launched in a cloud environment and assigned to a corresponding HTCondor queue according to the demand. Thereby, the use of cloud resources is totally transparent to the user. However, using this approach, submitted user analysis jobs might still suffer from a certain delay introduced by waiting in the queue. Therefore, our second solution offers the possibility to easily deploy a totally private analysis cluster, i.e., batch or PROOF, on private cloud resources belonging to the university. Thereby, the private analysis cluster is connected to the ATLAS data-management system (DDM) to read and write input/output files.

Track 6 Session / 251

A first look at 100 Gbps LAN technologies, with an emphasis on future DAQ applications

Authors: Adam Jedrzej Otto[1]; Daniel Hugo Campora Perez[2]; Flavio Pisani[3]; Niko Neufeld[2]; Rainer Schwemmer[2]

[1] *Ministere des affaires etrangeres et europeennes (FR)*

[2] *CERN*

[3] *Universita e INFN, Roma I (IT)*

Corresponding Author: adam.otto@cern.ch

The LHCb experiment is preparing a major upgrade of both the detector and the data acquisition system. A system capable of transporting up to 50 Tbps of data will be required. This can only be achieved in a manageable way using 100 Gbps links. Such links recently became available also in the servers, while they have been available between switches already for a while.

We present first measurements with such links (InfiniBand EDR, Ethernet 100 GbE) both using standard benchmarks and using a prototype event-building application. We analyse the CPU load effects by using Remote DMA technologies, and we also show comparison with previous tests on 40 G equipment.

252

File Access Optimization with the Lustre Filesystem at Florida CMS T2

Author: Bockjoo Kim[1]

Co-authors: Dimitri Bourilkov [1]; Jorge Luis Rodriguez [2]; Paul Ralph Avery [1]; Yu Fu [1]

[1] *University of Florida (US)*

[2] *Florida International University (US)*

Corresponding Authors: bockjoo@phys.ufl.edu, bourilkov@phys.ufl.edu, jorge.luis.rodriguez@cern.ch, paul.ralph.avery@cern.ch, yfu@phys.ufl.edu

One of the CMS Tier2 centers, the Florida CMS Tier2 center, has been using the Lustre filesystem for its data storage backend system since 2004. Recently, the data access pattern at our site has changed greatly due to various new access methods that include file transfers through the GridFTP servers, read access from the worker nodes, and remote read access through xrootd. In order to optimize the file access performance, we have to consider all the possible access patterns and each pattern needs to be studied separately. In this presentation, we report on our work to optimize file access with the Lustre filesystem at the Florida CMS T2 using an approach based on the analyzing these access patterns.

253

Free cooling on the Mediterranean shore: Energy efficiency upgrades at PIC

Authors: Manuel Delfino Reznicek[1]; Vanessa Acin Portella[2]

Co-authors: Adrià Herbera Luna [3]; Jordi Hernández Sánchez [3]

[1] *Universitat Autònoma de Barcelona (ES)*

[2] *Institut de Física d'Altes Energies (IFAE)*

[3] *Universitat Autònoma de Barcelona*

Corresponding Author: delfino@pic.es

Energy consumption is an increasing concern for data centers. This paper summarizes recent energy efficiency upgrades at the Port d'Informació Científica (PIC) in Barcelona, Spain which have considerably lowered energy consumption. The upgrades were particularly challenging, as they involved modifying the already existing machine room, which is shared by PIC with the general IT services of the Universitat Autònoma de Barcelona (UAB), with all the services in full operation, as well as the introduction of "free cooling"techniques in a location 20 Km from the Mediterranean sea. The upgrades targeted three distinct areas: First, the segregation of hot and cold air zones using and innovative horizontal layout, where hot air is channelled through openings in a false ceiling to a second

story hot air plenum. This segregation allows increasing the cold air inlet temperature according to the latest ASHRAE recommendations. Second, the introduction of an outside air economizer which replaces obsolete CRAH systems with air-to-air heat exchangers. This system, built entirely from industrial components, also incorporates an adiabatic cooling module and enables the "free"removal of over 300 kW of IT heat load during 6000 hours a year. Third, the introduction of UPS systems based on IGBT technology, in order to better match the impedance characteristics of the IT load. In addition, a transversal activity has been done to fully integrate cooling and UPS infrastructure monitoring into PIC's overall IT monitoring framework based on Nagios. This required development of a ModBus/TCPIP gateway server.

254

Unlocking data: federated identity with LSDMA and dCache

Authors: Arsen Hayrapetyan[1]; Christian Bernardt[2]; Dmitry Litvintsev[3]; Gerd Behrmann[4]; Karsten Schwank[None]; Marcus Hardt[5]; Patrick Fuhrmann[6]; Paul Millar[2]; Tigran Mkrtchyan[7]

[1] *ANSL (Yerevan Physics Institute) (AM)*

[2] *Deutsches Elektronen-Synchrotron (DE)*

[3] *FNAL*

[4] *NDGF*

[5] *Kalrsruhe Institute of Technology*

[6] *DESY*

[7] *Unknown*

Corresponding Author: paul.millar@desy.de

X.509, the dominant identity system from grid computing, has proved unpopular for many user communties. More popular alternatives generally assume the user is interacting via their web-browser. Such alternatives allow a user to authenticate with many services with the same credentials (username and password). They also allow users from different organisations form collaborations quickly and simply.

Scientists generally require that their custom analysis software has direct access to the data. Such direct access is not currently supported by alternatives to X.509, as they require the use of a web-browser.

Various approaches to solve this issue are being investigated as part of the Large Scale Data Management and Analysis (LSDMA) project, a German funded national R&D project. These involve dynamic credential translation (creating an X.509 credential) to allow backwards compatibility in addition to direct SAML- and OpenID Connect-based authentication.

We present a summary of the current state of art and the current status of the federated identity work funded by the LSDMA project along with the future road map.

Track 7 Session / 255

New adventures in storage: cloud storage and CDMI

Authors: Christian Bernardt[1]; Dmitry Litvintsev[2]; Gerd Behrmann[3]; Hermann Heßling[4]; Jana Weschenfelder[4]; Karsten Schwank[None]; Patrick Fuhrmann[5]; Paul Millar[1]; Tigran Mkrtchyan[6]

[1] *Deutsches Elektronen-Synchrotron (DE)*

[2] *FNAL*

[3] *NDGF*

[4] *HTW Berlin*

[5] *DESY*

[6] *urn:Google*

Corresponding Author: paul.millar@desy.de

Traditionally storage systems have had well understood responsibilities and behaviour, codified by the POSIX standards. More sophisticated systems (such as dCache) support additional functionality, such as storing data on media with different latencies (SSDs, HDDs, tapes). From a user's perspective, this forms a relatively simple adjunct to POSIX: providing optional quality-of-service values when writing data and optionally requesting data be staged from tape ahead of use.

The CDMI protocol provides a standard mechanism for clients to discover and use many advanced features. Such features include storing and querying metadata, searching for files matching metadata predicates, controlling a file's quality-of-service and retention policies, providing an object store and alternative protocol discovery.

A CDMI enabled storage has the potential for greatly simplifying a more general service as some high-level functionality can be delegated to the storage system. This reduces and may remove the need to run additional services, which makes it easier for sites to support their users.

By implementing the CDMI standard, dCache can expose new features in a standards compliant fashion. Here, various scenarios are presented where CDMI provides an advantage and the road-map for CDMI support in dCache is explored.

Track 5 Session / 256

The VISPA Internet Platform for Scientific Research, Outreach and Education

Author: Martin Urban[1]

Co-authors: Benjamin Fischer [1]; Christian Glaser [1]; Christoph Welling [1]; Daniel Van Asseldonk [1]; Fabian Heidemann [1]; Gero Mueller [1]; Marcel Rieger [1]; Martin Erdmann [1]; Robert Fischer [1]; Thorben Quast [1]

[1] *Rheinisch-Westfaelische Tech. Hoch. (DE)*

Corresponding Author: martin.urban@physik.rwth-aachen.de

VISPA provides a graphical front-end to computing infrastructures giving its users all functionality needed for working conditions comparable to a personal computer. It is a framework that can be extended with custom applications to support individual needs, e.g. graphical interfaces for experiment-specific software. By design, VISPA serves as a multi-purpose platform for many disciplines and experiments as demonstrated in the following different use-cases. A GUI to the analysis framework OFFLINE of the Pierre Auger collaboration, submission and monitoring of computing jobs, university teaching of hundreds of students, and outreach activity, especially in CERN's open data initiative.
Serving heterogeneous user groups and applications gave us lots of experience. This helps us in

maturing the system, i.e. improving the robustness and responsiveness, and the interplay of the components. Among the lessons learned are the choice of a file system, the implementation of web-sockets, efficient load balancing, and the fine-tuning of existing technologies like the RPC over SSH. We present in detail the improved server setup and report on the performance, the user acceptance and the realized applications of the system.

257

Design and development of a Virtual Computing Platform Integrated with Batch Job System

Author: Jingyan Shi[1]

Co-authors: Bowen Kan [1]; Haibo Li [1]; Zhenyu Sun [2]; qiulan huang [3]

[1] *IHEP*

[2] *ihep*

[3] *Institute of High Energy Physics, Beijing*

Corresponding Author: shijy@ihep.ac.cn

Batch system is a common way for a local cluster to schedule jobs running on work nodes. In some cases, some jobs have to stay in queue without suitable work nodes while some job slots have to keep free without suitable jobs running. The reasons for such case might be various. One of the main reasons is that operating system running on the free work nodes is different from the one that jobs in queues wanted. Another reason to block jobs running on free job slots is the complicated job scheduling policies.
Virtual machine has been proved to run some type of jobs with low penalty, and it could provide more flexible Operating system the job wanted.
We designed and developed a virtual computing platform based on openstack and integrated with batch job system. The same submit command and job script could be used without any changes. The jobs submitted to the virtual computing platform would be run on virtual machines with the correct operating system. All of these are transparent to users. The platform is composed of four parts, Batch Info Collector, Job Matcher, Virtual Machine Manager and Virtual Job Manager. After consults with the info from Batch Info Collector and Virtual Machine Manager, Job Matcher would send request to the Virtual Machine Manager to start virtual machines for the jobs submitted to the virtual queue. All the jobs running on virtual machine are monitored and controlled by virtual job manager.

Track 2 Session / 258

New developments in the FairRoot framework

Author: Florian Uhlig[1]

Co-authors: Alexey Rybalchenko [1]; Anar Manafov [1]; Andrey Lebedev [1]; Denis Bertini [1]; Dmytro Kresan [1]; Mohammad Al-Turany [1]; Nicolas Winkler [1]; Radoslaw Karabowicz [1]

[1] *GSI Darmstadt*

Corresponding Author: f.uhlig@gsi.de

The FairRoot framework is the standard framework for simulation, reconstruction and data analysis developed at GSI for the future experiments at the FAIR facility.

The framework delivers base functionality for simulation, i.e.: Infrastructure to easily implement a set of detectors, fields, and event generators. Moreover, the framework decouples the user code (e.g.:

Geometry description, detector response, etc.) completely from the used MC engine. The framework also handles the Input/Output (IO). The output of single detectors (tasks) can be switched on (made persistence) or off (transient) in a simple and flexible way.

For reconstruction and/or data analysis the user code is organized in modular tasks that implement the different states of a state machine. The order in which these tasks are executed is defined via a so-called steering macro. This scheme allows a very flexible handling of the reconstruction and data analysis configurations, it also allow for mixing of simulation and data reconstruction. Reconstruction tasks can run separately after simulation or directly on the fly within the simulation.

The modular design of the framework has allowed a smooth transition to a message queue based system, which makes it possible to parallelize the execution of the tasks without re-designing or re-writing the existing user code. The new design also allows implementing the processes in different programming languages or on different hardware platforms. For the communication between the different processes modern technologies like protocol buffers and Boost serialization are also used.

The framework with a focus on the basic building blocks and the transition to the message queue based system will be presented.

Track 3 Session / 259

Deep Storage for Big Scientific Data

Author: David Yu[1]

Co-authors: Guangwei Che [1]; Ognian Novakov [1]; Tim Chou [1]

[1] *BNL*

Corresponding Author: david.yu@bnl.gov

Brookhaven National Lab (BNL)'s RHIC and Atlas Computing Facility (RACF), is supporting science experiments such as RHIC as its Tier-0 center and the U.S. ATLAS/LHC as a Tier-1 center. Scientific data is still growing exponentially after each upgrade. The RACF currently manages over 50 petabytes of data on robotic tape libraries, and we expect a 50% increase in data next year. Not only do we have to address the issue of efficiently archiving high bandwidth data to our tapes, but we also have to face the problem of randomly restoring files from tapes. In addition, we have to manage tape resource usage and technology migration, which is moving data from low-capacity media to newer, high-capacity tape media, in order to free space within a tape library. BNL's mass storage system is managed by a software called IBM HPSS. To restore files from HPSS, we have developed a file retrieval scheduling software, called TSX. TSX provides dynamic HPSS resource management, schedules jobs efficiently, and enhances visibility of real-time staging activities and advanced error handling to maximize the tape staging performance.

Track 1 Session / 260

The Electronics, Online Trigger System and Data Acquisition System of the J-PARC E16 Experiment

Author: Tomonori Takahashi[1]

Co-authors: Daisuke Kawama [2]; Eitaro Hamada [3]; Hiroshi Sendai [4]; Kyoichiro Ozawa [4]; M. M. Tanaka [4]; Masahiro Ikeno [4]; Satoshi Yokkaichi [2]; Tomohisa Uchida [4]; Wataru Nakai [5]; Yuhei Morino [4]; Yuki Obara [6]

[1] *Research Center for Nuclear Physics, Osaka University*

[2] *RIKEN, Nishina Center*

[3] *High Energy Accelerator Research Organization (KEK)*

[4] *KEK*

[5] *Department of Physics, University of Tokyo / RIKEN, Nishina Center*

[6] *Department of Physics, University of Tokyo*

Corresponding Author: tomonori@riken.jp

1. Introduction

The J-PARC E16 experiment aims to investigate the chiral symmetry restoration in cold nuclear matter and the origin of the hadron mass through the systematic study of the mass modification of vector mesons.
In the experiment,
e^+e^- decay of slowly-moving ϕ mesons in the normal nuclear matter density are intensively studied using several nuclear targets (H, C, Cu and Pb).
The dependence of the modification on the nuclear size and the meson momentum will be measured for the first time.

2. Experiment

The experiment will be performed in 2016 at the high-momentum beam line of the J-PARC hadron experimental facility,
where a 30-GeV proton beam with a high intensity of 1×10^{10} per pulse (2-second spill per 6-second cycle) is delivered to experimental targets.
Since the material budget around the targets is sensitive to the e^+e^- measurement,
thin detector systems are under construction.
The targets are surrounded by GEM Trackers (GTR) with three tracking planes to achieve the good resolution of 100 μm in the high rate environment of 5 kHz/mm^2.
The electrons (positrons) are identified by two types of counters.
One is the Hadron Blind Detector (HBD), which is a threshold type gas Cherenkov detector using GEM,
and the other is the Lead-glass EM calorimeter (LG).

3. Trigger electronics

The first level trigger is decided by the three fold coincidence of ∼620-ch from the GTR, ∼940-ch from the HBD and ∼1000-ch from the LG.
Cathode foils which face to the read out strips of the most outside GTR and pads of the HBD are divided into trigger segments.
A pulse fired on the GEM cathode foils are fed into an amplifier-shaper-discriminator (ASD) ASIC, which has been developed by our group in cooperation with
Open-It1.
The LG signals are discriminated by a commercial fast comparator.
In order to gather the trigger primitives, which are sent from the GTR, HBD and LG in parallel LVDS signals,
a trigger merger board (TRG-MRG) has been developed.
The TRG-MRG produces time stamps of the trigger primitives with a resolution of less than 4 nsec by using a Xilinx Kintex-7 FPGA.
The time stamps are serialized by the FPGA and transmitted to a global trigger decision module via optical fibers at each link rate of 5 Gbps or more.
The global trigger module utilizes a Belle-II Universal Trigger Board 3.
The first level trigger as well as a global clock of ∼125 MHz is distributed by Belle-II FTSW boards via Category-7 LAN cables to the front-end-modules described bellow.

4. Readout electronics

The numbers of readout channels amount to ∼56k, ∼36k and ∼1k for the GTR, HBD and LG, respectively.
In the current design, waveforms from all of the readout channels will be recorded by using analog memory ASICs to obtain timing and charge deposit information and to distinguish pulse pile-up in the high rate environment for the offline analysis.

The waveform from the GTR and HBD are stored with a 25 nsec cycle in APV25s12 chips and then transferred to the Scalable Readout System, which has been developed by the CERN RD51 Collaboration3 (an R&D collaboration for MGPDs).
The LGs are read out by custom made boards, which employ DRS44 chips to record the pulses at 1 GHz.
Those modules digitize the waveforms and perform the zero suppression at online.
The data are collected by the DAQ-Middleware5 using gigabit Ethernet and 10G Ethernet links.
The expected data rate is 660 MB/spill with the event rate of 2k/spill after zero suppression.

5. Summary

This is an overview talk on the electronics and trigger system
for the J-PARC E16 experiment.
Other contributions for the detail of the DAQ software, trigger ASIC,
and so on are also prepared and submitted by coauthors.

References

1 http://openit.kek.jp/ (in Japanese)

2 M. Raymond *et al.*, IEEE NSS Conf. Rec. 2 (2000) 9/113.

3 http://rd51-public.web.cern.ch/RD51-Public/

4 http://www.psi.ch/drs/

5 http://daqmw.kek.jp/ (in Japanese)

261

Tier-1 in Kurchatov Institute: status before Run-2 and HPC integration

Author: Eygene Ryabinkin[1]

Co-authors: Alexandra Berezhnaya [2]; Igor Tkachenko [2]; Ilya Lyalin [2]; Vasily Velikhov [1]; Yury Lazin [2]

[1] *National Research Centre Kurchatov Institute (RU), Moscow Institute for Physics and Technology (RU)*

[2] *National Research Centre Kurchatov Institute (RU)*

Corresponding Author: rea@grid.kiae.ru

We present the status of RRC-KI-T1, new Russian Tier-1 that supports ALICE, ATLAS and LHCb. Our aim is to enter the full production mode just before the beginning of Run-2 and we will talk about our current setup, deployed services and middleware, workflow, achievements and problems on the route of bringing yet another Tier-1 for WLCG.

Another facet of our current activity is making the parts of the processing resources at our HPC facilities to be available for Grid on-demand usage. This requires us to adopt our HPC environment to be able to process Grid jobs, to use the storage at our Tier-1 as the native storage element and to adopt site-level services for interoperability with HPC parts. We also collaborate with Kurchatov Institute BigData laboratory that adopts PanDA workload management system for HPC environments in the context of ATLAS and ALICE workloads, the task that also requires major tweaks of HPC worker nodes to incorporate them into PanDA and WLCG environments. We will present our hardware and software architecture that powers such integration and results of test for its real-world usage.

Track 2 Session / 262

Physics Analysis Software Framework for Belle II

Authors: Anze Zupanc[1]; marko staric[2]

[1] *Jozef Stefan Institute*

[2] *J. Stefan Institute, Ljubljana, Slovenia*

Corresponding Author: marko.staric@ijs.si

We present software framework being developed for physics analyses using the data collected by the Belle II experiment. The analysis workflow is organized in a modular way 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 combination of both. Typically, a user performing a physics analysis only needs to provide a job configuration file with analysis'specific sequence of physics modules that can be then executed on the Grid.

Track 4 Session / 263

ARC Control Tower: A flexible generic distributed job management framework

Authors: Andrej Filipcic[1]; David Cameron[2]; Jon Kerr Nilsen[2]

[1] *Jozef Stefan Institute (SI)*

[2] *University of Oslo (NO)*

Corresponding Author: j.k.nilsen@fys.uio.no

While current grid middlewares are quite advanced in terms of connecting jobs to resources, their client tools are generally quite minimal and features for managing large sets of jobs are left to the user to implement. The ARC Control Tower (aCT) is a very flexible job management framework that can be run on anything from a single user's laptop to a multi-server distributed setup. aCT was originally designed to enable ATLAS jobs to be submitted to the ARC CE. However, with the recent redesign of aCT where the ATLAS specific elements are clearly separated from the ARC job management parts, the control tower can now easily be reused as a flexible generic distributed job manager for other communities. This paper will give a detailed explanation how aCT works as a job management framework and go through the steps needed to create a simple job manager using aCT and show that it can easily manage thousands of jobs.

Track 5 Session / 264

Improvements in the CMS Computing System for Run2

Authors: Ian Fisk[1]; Maria Girone[2]

[1] *Fermi National Accelerator Lab. (US)*

[2] *CERN*

Corresponding Author: ian.fisk@cern.ch

Beginning in 2015 CMS will collected and produce data and simulation adding to 10B new events a year. In order to realize the physics potential of the experiment these events need to be stored,

processed, and delivered to analysis users on a global scale. CMS has 150k processor cores and 80PB of disk storage and there is constant pressure to reduce the resources needed and increase the efficiency of usage. In this presentation we will comprehensively overview the improvements made in the computing system for Run2 by CMS in the areas of data and simulation processing, data distribution, data management and data access. The system has been examined and we will discuss the improvements in the entire data and workflow systems: CMS processing processing and analysis workflow tools, the development and deployment of dynamic data placement infrastructure, and progress toward operating a global data federation. We will describe the concepts and approaches to utilize the variety of CMS CPU resources, ranging from established Grid sites to HPC centers, Cloud resources and CMS' own High Level Trigger farm. We will explain the strategy for improving how effectively the storage is used and the commissioning, validation and challenge activities will be presented.

265

Optimizing CMS build infrastructure via Apache Mesos

Author: Giulio Eulisse[1]

Co-authors: Alessandro Degano [2]; David Abdurachmanov [3]; David Gonzalo Mendez Lopez [4]; Peter Elmer [5]; Shahzad Malik Muzaffar [1]

[1] *Fermi National Accelerator Lab. (US)*

[2] *Universita e INFN (IT)*

[3] *Vilnius University (LT)*

[4] *Universidad de los Andes (CO)*

[5] *Princeton University (US)*

Corresponding Author: giulio.eulisse@cern.ch

The Offline Software of the CMS Experiment at the Large Hadron Collider (LHC) at CERN consists of 6M lines of in-house code, developed over a decade by nearly 1000 physicists, as well as a comparable amount of general use open-source code. A critical ingredient to the success of the construction and early operation of the WLCG was the convergence, around the year 2000, on the use of a homogeneous environment of commodity x86-64 processors and Linux.

Apache Mesos is a cluster manager that provides efficient resource isolation and sharing across distributed applications, or frameworks. It can run Hadoop, Jenkins, Spark, Aurora, and other applications on a dynamically shared pool of nodes.

We present how we migrated our continuos integration system to schedule jobs on an relatively small Apache Mesos enabled cluster and how this resulted in better resource usage, higher peak performance and lower latency thanks to the dynamic scheduling capabilities of Mesos.

266

Development of GEM trigger electronics for the J-PARC E16 experiment

Author: Yuki Obara[1]

Co-authors: Daisuke Kawama [2]; Eitaro Hamada [3]; Hiroshi Sendai [3]; Kyoichiro Ozawa [3]; M. M. Tanaka [3]; Masahiro Ikeno [3]; Satoshi Yokkaichi [2]; Tomohisa Uchida [3]; Tomonori Takahashi [4]; Wataru Nakai [5]; Yuhei Morino [3]

[1] *University of Tokyo*

[2] *RIKEN Nishina Center*

[3] *High Energy Accelerator Research Organization (KEK)*

[4] *Research Center for Nuclear Physics (RCNP)*

[5] *University of Tokyo, RIKEN Nishina Center*

Corresponding Author: yobara@post.kek.jp

The purpose of the J-PARC E16 experiment is to investigate the origin of hadron mass through the chiral symmetry restoration in nuclear matter. In the experiment, we measure mass spectra of vector mesons in nuclei in the e^+e^- decay channel
with high precision and high statistics.

We use a 30 GeV proton beam with high intensity of 10^{10} per spill to achieve high statistics
and targets of 0.1% interaction length to suppress an electron background caused by γ - conversion in the targets.
In the spectrometer of the experiment, GEM Trackers (GTR), which are composed of three layers of tracking planes in a magnetic field, are used to measure momenta of the decay electrons in the high rate environment.
Hadron Blind Detectors (HBD), which are gas
Cerenkov counters using GEM, and Lead Glass Calorimeters (LG) are placed outside the GTR to identify electrons.
The e^+e^- event trigger consists of three-fold coincidence of signals from the most outside GTR, HBD and LG.
The number of channels for GTR, HBD and LG is about 620, 940, and 1000, respectively.
Efficient trigger system selecting events of e^+e^- decays from
huge background events must be constructed in order to reduce the trigger rate to 1-2 kHz that our DAQ can cope with.

We developed new Amplifier-Shaper-Discriminator (ASD) ASICs which can deal with large detector capacitance originating from GEM and have a short shaping time to handle the high rate counting.
The naive idea to extract the trigger signal from these GEM detectors, such as GTR and HBD, is utilizing signals of strips or pads on the anode readout plane which are used for tracking or electron identification.
This idea requires, however, R&D of complex frontend circuits and a large number of channels for the fast signal outputs.
In order to avoid these problems, the trigger signals are picked up from the cathode plane of the induction gap of these GEM chambers, namely, the last GEM foil in the stack.
Considering to use the signals from the GEM foil, it is difficult to cope with large capacitance of the order of nF by using normal preamps.

Requirements of the trigger electronics for GTR are to cope with the large detector capacitance of about 2 nF, a fast shaping time and a good signal-to-noise ratio for the minimum input charge of 10 fC.
The cathode plane of a GEM foil of the most outside GTR, whose size is $300 \times 300\ \mathrm{mm}^2$, has detector capacitance of about 50
nF.
The GEM foil is divided into 24 segments in order to reduce the capacitance to about 2 nF and a counting rate in each segment, and to roughly track charged particles.
The rough tracking has an important role of decreasing a background of electrons which do not come from targets.
In the forward region of the spectrometer, the maximum hit rate of each segment is expected to be 1-2 MHz.
Thus, we set the shaping time to 25 ns corresponding to a pulse width of about 200 ns.
The electric circuit of the ASIC was designed to suppress Equivalent Noise Charge (ENC) under

2×10^4 for the input detector capacitance of 2 nF.
The ASD ASIC chip satisfying the above requirements has been developed by our group in cooperation with Open-It1.
A modified version of the ASIC for GTR is used as the electronics of the GEM trigger for HBD.

A prototype of a preamp board with the ASIC chips was produced.
Size of the board should be enough small to be installed in narrow spaces of GTR modules.
One of functions of the preamp board is to convert discriminated digital outputs of the ASIC to parallel LVDS signals which are sent to a trigger merger board (TRG-MRG).
Also, the board needs to receive slow control signals from the TRG-MRG to activate and control digital functions of the ASIC.

This contribution will report the development status of the GEM trigger system and result of performance test of the ASIC chip.

Reference: 1. http://openit.kek.jp/

Track 2 Session / 267

Optimization of the LHCb track reconstruction

Authors: Barbara Storaci[1]; Michel De Cian[2]

[1] *Universitaet Zuerich (CH)*

[2] *Ruprecht-Karls-Universitaet Heidelberg (DE)*

Corresponding Author: barbara.storaci@cern.ch

The LHCb track reconstruction uses sophisticated pattern recognition algorithms to reconstruct trajectories of charged particles. Their main feature is the use of a Hough-transform like approach to connect track segments from different subdetectors, allowing for having no tracking stations in the magnet of LHCb. While yielding a high efficiency, the track reconstruction is a major contributor to the overall timing budget of the software trigger of LHCb, and will continue to be so in the light of the higher track multiplicity expected from Run II of the LHC.
In view of this fact, key parts of the pattern recognition have been revised and redesigned. We will present the main features which were studied. A staged approach strategy for the track reconstruction in the software trigger was investigated: it allows unifying complementary sets of tracks coming from the different stages of the high level trigger, resulting in a more flexible trigger strategy and a better overlap between online and offline reconstructed tracks. Furthermore the use of parallelism was investigated, using SIMD instructions for time-critical parts of the software or - in a later stage - using GPU-driven track reconstruction.
In addition a new approach to monitoring was implemented, where quantities important for track reconstruction are monitored on a regular basis, using an automated framework for comparing different figures of merit.

Track 8 Session / 268

SIMD studies in the LHCb reconstruction software

Authors: Ben Couturier[1]; Daniel Hugo Campora Perez[1]

[1] *CERN*

Corresponding Author: daniel.hugo.campora.perez@cern.ch

During the data taking process in the LHC at CERN, millions of collisions are recorded every second by the LHCb Detector. The LHCb "Online" computing farm, counting around 15000 cores, is dedicated to the recontruction of the events in real-time, in order to filter those with interesting Physics. The ones kept are later analysed "Offline" in a more precise fashion on the Grid. This imposes very stringent requirements on the Reconstruction Software, which has to be as efficient as possible.

Modern CPUs support so-called "vector-extensions", which extend their Instruction Sets, allowing for concurrent execution across functional units. Several libraries expose the Single Instruction Multiple Data programming paradigm to issue these instructions. The use of vectorisation in our codebase can provide performance boosts, leading ultimately to Physics reconstruction enhancements.

In this paper, we present vectorisation studies of significant reconstruction algorithms. A variety of vectorisation libraries are analysed and compared in terms of design, maintainability and performance. We also present the steps taken to systematically measure the performance of the released software, to ensure the consistency of the run-time of the vectorized software.

Track 7 Session / 269

LHCb experience with running jobs in virtual machines

Author: Andrew McNab[1]

Co-authors: Cinzia Luzzi [2]; Federico Stagni [2]

[1] *University of Manchester (GB)*

[2] *CERN*

Corresponding Author: andrew.mcnab@cern.ch

The LHCb experiment has been running production jobs in virtual machines since 2013 as part of its DIRAC-based infrastructure. We describe the architecture of these virtual machines and the steps taken to replicate the WLCG worker node environment expected by user and production jobs. This relies on the CernVM 3 system for providing root images for virtual machines. We use the cvmfs distributed filesystem to supply the root partition files, the LHCb software stack, and the bootstrapping scripts necessary to configure the virtual machines for us. Using this approach, we have been able to minimise the amount of contextualisation which must be provided by the virtual machine managers. We explain the process by which the virtual machine is able to receive payload jobs submitted to DIRAC by users and production managers, and how this differs from payloads executed within conventional DIRAC pilot jobs on batch queue based sites. We compare our operational experiences of running production on VM based sites managed using OpenStack, Vac, BOINC, and Condor. Finally we describe our requirements for monitoring which are specific to the additional responsibilities for experiments when operating virtual machines which were previously undertaken by the system managers of worker nodes, and how this is facilitated by the new DIRAC Pilot 2.0 architecture.

Track 6 Session / 270

Analysis of CERN Computing Infrastructure and Monitoring Data

Authors: Christian Nieke[1]; Dirk Duellmann[2]

[1] *Brunswick Technical University (DE)*

[2] *CERN*

Corresponding Author: christian.nieke@cern.ch

Optimising a computing infrastructure on the scale of LHC requires a quantitative understanding of a complex network of many different resources and services. For this purpose the CERN IT department and the LHC experiments are collecting a large multitude of logs and performance probes, which are already successfully used for short-term analysis (e.g. operational dashboards) within each group.
The IT analytics working group has been created with the goal to bring data sources from different services and on different abstraction levels together and to implement a suitable infrastructure for mid- to long-term statistical analysis. It further provides a forum for joint optimization across single service boundaries and the exchange of analysis methods and tools.
To simplify access to the collected data, we implemented an automated repository for cleaned and aggregated data sources based on the Hadoop ecosystem. This contribution describes some of the challenges encountered, such as dealing with heterogeneous data formats, selecting an efficient storage format for map reduce and external access, and will describe the repository user interface.
Using this infrastructure we were able to quantitatively analyse the relationship between CPU/wall fraction, latency/throughput constraints of network and disk and the effective job throughput.
In this contribution we will first describe the design of the shared analysis infrastructure and then present a summary of first analysis results from the combined data sources.

Track 7 Session / 271

Managing virtual machines with Vac and Vcycle

Author: Andrew McNab[1]

Co-authors: Ewan Christopher Mac Mahon [2]; Peter Love [3]

[1] *University of Manchester (GB)*

[2] *Particle Physics-University of Oxford-Unknown*

[3] *Lancaster University (GB)*

Corresponding Author: andrew.mcnab@cern.ch

We compare the Vac and Vcycle virtual machine lifecycle managers and our experiences in providing production job execution services for ATLAS, LHCb, and the GridPP VO at sites in the UK and at CERN. In both the Vac and Vcycle systems, the virtual machines are created outside of the experiment's job submission and pilot framework. In the case of Vac, a daemon runs on each physical host which manages a pool of virtual machines on that host, and a peer-to-peer UDP protocol is used to achieve the desired target shares between experiments across the site. In the case of Vcycle, a daemon manages a pool of virtual machines on an Infrastructure As A Service cloud system such as OpenStack, and has within itself enough information to create the types of virtual machines to achieve the desired target shares. Both systems allow unused shares for one experiment to temporarily taken up by other experiements with work to be done. The virtual machine lifecycle is managed with a minimum of information, gathered from the virtual machine creation mechanism (such as libirt or OpenStack) and using the proposed Machine/Job Features API from WLCG. We demonstrate that the same virtualmachine designs can be used to run production jobs on Vac and Vcycle/OpenStack sites for ATLAS, LHCb, and GridPP, and that these technologies allow sites to be operated in a reliable and robust way.

272

Integrating Puppet and Gitolite to provide a novel solution for scalable system management at the MPPMU Tier2 centre

Author: Luca Mazzaferro[1]

Co-authors: Cesare Delle Fratte [1]; John Kennedy [2]; Stefan Kluth [1]

[1] *Max-Planck-Institut fuer Physik (Werner-Heisenberg-Institut) (D*

[2] *Rechenzentrum Garching (RZG) - Max Planck Society*

Corresponding Author: luca.mazzaferro@rzg.mpg.de

In a grid computing infrastructure tasks such as continuous upgrades, services installations and software deployments are part of an admins daily work. In such an environment tools to help with the management, provisioning and monitoring of the deployed systems and services have become crucial.

As experiments such as the LHC increase in scale, the computing infrastructure also becomes larger and more complex. Moreover, today's admins increasingly work within teams that share responsibilities and tasks. Such a scaled up situation requires tools that not only simplify the workload on administrators but also enable them to work seamlessly in teams.

In this paper will be presented our experience from managing the Max Planck Institute Tier2 using Puppet and Gitolite in a cooperative way to help the system administrator in their daily work.

In addition to describing the Puppet-Gitolite system, best practices and customizations will also be shown.

Track 6 Session / 273

A Model for Forecasting Data Centre Infrastructure Costs

Author: Renaud Vernet[1]

[1] *CC-IN2P3 - Centre de Calcul (FR)*

Corresponding Author: renaud.vernet@cern.ch

The computing needs in the HEP community are increasing steadily, but the current funding situation in many countries is tight. As a consequence experiments, data centres, and funding agencies have to rationalize resource usage and expenditures.

CC-IN2P3 (Lyon, France) provides computing resources to many experiments including LHC, and is a major partner for astroparticle projects like LSST, CTA or Euclid. The financial cost to accommodate all these experiments is substantial and has to be planned well in advance for funding and strategic reasons.

In that perspective, leveraging infrastructure expenses, electric power cost and hardware performance observed in our site over the last years, we have built a model that integrates these data and provides estimates of the investments that would be required to cater to the experiments for the mid-term future.

We present how our model is built and the expenditure forecast it produces, taking into account the experiment roadmaps. We also examine the resource growth predicted by our model over the next years assuming a flat-budget scenario.

275

The Database Driven ATLAS Trigger Configuration System

Author: Alex Christopher Martyniuk[1]

Co-authors: Carlos Chavez [2]; Joerg Stelzer [3]; Mark Stockton [4]; Michele Faucci Giannelli [5]; Werner Wiedenmann [6]; William Panduro Vazquez [7]

[1] *University College London*

[2] *University of Liverpool*

[3] *CERN*

[4] *McGill University (CA)*

[5] *Royal Holloway, University of London*

[6] *University of Wisconsin (US)*

[7] *Imperial College London*

Corresponding Author: alex.christopher.martyniuk@cern.ch

This contribution describes the trigger selection configuration system of the ATLAS low- and high-level trigger (HLT) and the upgrades it received in preparation for LHC Run 2.

The ATLAS trigger configuration system is responsible for applying the physics selection parameters for the online data taking at both trigger levels and the proper connection of the trigger lines across those levels. Here the low-level trigger consists of the already existing central trigger (CT) and the new Level-1 Topological trigger (L1Topo), which has been added for Run 2. In detail the tasks of the configuration system during the online data taking are

- Application of the selection criteria, e.g. energy cuts, minimum multiplicities, trigger object correlation, at the three trigger components L1Topo, CT, and HLT
- On-the-fly, e.g. rate-dependent, generation and application of prescale factors to the CT and HLT to adjust the trigger rates to the data taking conditions, such as falling luminosity or rate spikes in the detector readout
- Recording of the complete trigger configuration for any given point in time, for later use by data analysts

The core of the trigger configuration system is an oracle database (TriggerDB), with a dedicated schema to reflect the L1Topo, CT, and HLT configuration needs. A java-based UI serves as the front-end to the TriggerDB for the trigger experts to store new and modify existing trigger configurations. C++-based database reader software exists for the trigger clients to retrieve configurations from the database. Web interfaces exist to display the information to a large group of ATLAS members.

With the vast amount of upgrades of the CT and HLT during the last two years, and the addition of the L1Topo, substantial changes to the database and software were necessary, which will be presented. Technical problems, such as the low-latency distribution of the configuration across the HLT computing farm and the synchronous application to the data will be addressed. Also the propagation of the trigger configuration database from the data-taking side to various ATLAS data reconstruction sites will be discussed, including a short description how the same trigger configuration mechanism is being used for ATLAS Monte Carlo simulation. New features such an automated luminosity tracking and prescale application system, which optimizes the ATLAS data taking efficiency, will also be shown.

Track 2 Session / 276

CMS reconstruction improvements for the tracking in large pile-up events

Author: Giacomo Sguazzoni[1]

Co-author: Marco Rovere [2]

[1] *INFN (IT)*

[2] *CERN*

Corresponding Author: marco.rovere@cern.ch

The CMS tracking code is organized in several levels, known as 'iterative steps', each optimized to reconstruct a class of particle trajectories, as the ones of particles originating from the primary vertex or displaced tracks from particles resulting from secondary vertices. Each iterative step consists of seeding, pattern recognition and fitting by a kalman filter, and a final filtering and cleaning. Each subsequent step works on hits not yet associated to a reconstructed particle trajectory.

The CMS tracking code is continuously evolving to make the reconstruction computing load compatible with the increasing instantaneous luminosity of LHC, resulting in a large number of primary vertices and tracks per bunch crossing.

The major upgrade put in place during the present LHC Long Shutdown will allow the tracking code to comply with the conditions expected during Run2 and the much larger pile-up. In particular, new algorithms that are intrinsically more robust in high occupancy conditions have been developed, iterations have been re-designed (including a new one, dedicated to specific physics objects), code optimizations have been deployed and new software techniques have been used. The speed improvement has been achieved without significant reduction in term of physics performance.

The methods and the results are presented and the prospects for future applications are discussed.

Track 2 Session / 277

Event Reconstruction Techniques in NOvA

Authors: Evan Niner[1]; Kanika Sachdev[2]; Michael Baird[1]

Co-authors: Adam Aurisano [3]; Dominick Rocco [4]

[1] *Indiana University*

[2] *University of Minnesota*

[3] *University of Cincinnati*

[4] *urn:Google*

Corresponding Author: dominick.r.rocco@gmail.com

The NOvA experiment is a long baseline neutrino oscillation experiment utilizing the NuMI beam generated at Fermilab. The experiment will measure the oscillations within a muon neutrino beam in a 300 ton Near Detector located underground at Fermilab and a functionally-identical 14 kiloton Far Detector placed 810 km away. The detectors are liquid scintillator tracking calorimeters with a fine-grained cellular structure that provides a wealth of information for separating the different particle track and shower topologies. Each detector has its own challenges with the Near Detector seeing multiple overlapping neutrino interactions in each event and the Far Detector having a large background of cosmic rays due to being located on the surface. A series of pattern recognition techniques have been developed to go from event records, to spatially and temporally separating individual interactions, to vertexing and tracking, and particle identification. This combination of methods to achieve the full event reconstruction will be presented.

278

The DII-HEP OpenStack based CMS Data Analysis for secure cloud resources

Author: Lirim Osmani[1]

Co-authors: John White White [2]; Miika Komu [3]; Paula Eerola [4]; Salman Toor [2]; Sasu Tarkoma [1]; Tomas Linden [2]

[1] *Department of Computer Science, University of Helsinki*

[2] *Helsinki Institute of Physics (FI)*

[3] *Ericsson Research*

[4] *University of Helsinki (FI)*

Corresponding Author: lirim.osmani@cs.helsinki.fi

The topic of data storage and analysis on Cloud infrastructures has gained importance in recent years. The High Energy Physics community is interested in performing simulations and data analysis on public or private Cloud facilities. Currently the simulations and analysis are performed mostly on a computing and data Grid. The software and experience of operating on a Grid needs to be adapted for running on Cloud facilities. The approach of harnessing Grid and Cloud technologies ensures a steady and seamless transition towards new ways of working.

In order to fulfill this task, a virtual cluster has been constructed that is Cloud-based and Grid-enabled. For the Cloud part we use OpenStack and for the Grid software, that controls the execution of the physics jobs, we use components of the Advanced Resource Connector (ARC). This allows the end-users to submit the jobs with their preferred Grid or Cloud submission system and at the same time provides flexibility to maintain the infrastructure. The analysis software and libraries are installed via the CERN Virtual Machine file system. The cluster is monitored by running Site Availability Monitoring (SAM) jobs as well by using Graphite.

Our solution uses a hybrid approach of combining elements of Cloud and Grid software components. To manage the virtual machines (VMs) dynamically in an elastic fashion, we are using the EMI authorization service (Argus) and the Execution Environment Service (Argus-EES) with an OpenStack plugin that has been developed for Argus-EES. Our newly developed plugin for the Argus-EES can communicate with multiple OpenStack deployments to expand and shrink resources dynamically upon demand.

The Host Identity Protocol (HIP) has been designed for mobile networks and it provides a secure method for IP mobility and multi homing. HIP separates the end-point identifier and locator role for IP address which improves network agility of applications and the underlying virtual machines. Our solution leverages HIP for traffic management. This is useful for secure connections of hybrid Cloud resources.

We describe the state and current experience with a virtualized computing environment for CMS data analysis in the Datacenter Indirection Infrastructure for Secure High Energy Physics (DII-HEP) project.

Track 7 Session / 279

Dynamic provisioning of local and remote compute resources with OpenStack

Authors: Frank Polgart[1]; Thomas Hauth[1]

Co-authors: Günter Quast [1]; Manuel Giffels [2]

[1] *KIT - Karlsruhe Institute of Technology (DE)*

[2] *KIT - Karlsruhe Institute of Technology (DE)*

Corresponding Author: thomas.hauth@cern.ch

Modern high-energy physics experiments rely on the extensive usage of computing resources, both for the reconstruction of measured events as well as for Monte Carlo simulation. The Institut für Experimentelle Kernphysik (EKP) at KIT is participating in both the CMS and Belle experiments with computing and storage resources. In the upcoming years, these requirements are expected to increase due to growing amount of recorded data and the rise in complexity of the simulated events. It is therefore essential to increase the available computing capabilities by tapping into all resource pools.

At the EKP institute, powerful desktop machines are available to users. Due to the multi-core nature of modern CPUs, vast amounts of CPU time are not utilized by common desktop usage patterns.
Other important providers of compute capabilities are classical HPC data centers at Universities or national research centers. Due to the shared nature of these installations, the standardized software stack required by HEP applications cannot be installed.

A viable way to overcome this constraint and offer a standardized software environment in a transparent manner
is the usage of virtualization technologies. The OpenStack project has become a widely adopted solution
to virtualize hardware and offer additional services like storage and virtual machine management.

This contribution will report on the incorporation of the institute's desktop machines into a private OpenStack cloud. The additional compute resources provisioned via the virtual machines have been used for
Monte Carlo simulation and data analysis. Furthermore, a concept to integrate shared, remote HPC centers
into regular HEP job workflows will be presented. In this approach, local and remote resources are be merged to form a unfiorm, virtual compute cluster with a single point-of-entry for the user. Evaluations
of the performance and stability of this setup and operational experiences will be discussed.

280

A Comparison of the Overheads Associated with WLCG Federation Technologies

Author: Shaun de Witt[1]

Co-authors: Paul Millar [2]; Wahid Bhimji [3]

[1] *STFC*

[2] *Deutsches Elektronen-Synchrotron (DE)*

[3] *University of Edinburgh (GB)*

Corresponding Author: shaun.de-witt@stfc.ac.uk

For many years the Storage Resource Manager (SRM) has been the de-facto federation technology used by WLCG. This technology has, along with the rest of the middleware stack, mediated the transfer of many Petabytes of data since the start of data taking. In recent years, other technologies have become more popular as federation technologies because they offer additional functionalities that are not provided by SRM or they are based on more widely adopted standards. The main technologies currently used or being evaluated with WLCG are xrootd and WebDav.
In this work we investigate the overhead associated with each technology for different commonly used storage systems within WLCG. For jobs running at a local site, delays in file access can make a significant contribution to job efficiency and it is this we report on results from both controlled tests

and in 'real life' operations. While it is difficult to compare the overheads across different installations and hardware configurations, a comparison within an installation can help inform experiments to make the best choice of technology at any site and for any use case.

Track 6 Session / 281

Scheduling multicore workload on shared multipurpose clusters

Author: Jeff Templon[1]

Co-authors: Alessandra Forti [2]; Antonio Perez-Calero Yzquierdo [3]; Carlos Acosta Silva [4]; Jose Flix Molina [3]; Ronald Starink [5]

[1] *NIKHEF (NL)*

[2] *University of Manchester*

[3] *Centro de Investigaciones Energ. Medioambientales y Tecn. - (ES*

[4] *Universitat Autònoma de Barcelona (ES)*

[5] *Nikhef*

Corresponding Author: templon@nikhef.nl

With the advent of workloads containing explicit requests for multiple cores in a single grid job, grid sites faced a new set of challenges in workload scheduling. The most common batch schedulers deployed at HEP computing sites do a poor job at multicore scheduling when using only the native capabilities of those schedulers. This talk describes how efficient multicore scheduling was achieved at the three sites represented in the author list, by implementing dynamically-sized multicore partitions via a minimalistic addition to the Torque/Maui batch system already in use at those sites.

The first part of the talk covers the theory related to this particular problem, which is also applicable to e.g. the scheduling of large-memory jobs or data-aware jobs similarly comprising part of a highly heterogenous workload. The system design is also presented, linking it to previous work at Nikhef on grid-cluster scheduling. The second part of the talk presents an evaluation of several months of production operation at the three sites.

282

RAL Tier-1 evolution as global CernVM-FS service provider

Author: Catalin Condurache[1]

Co-authors: Ian Peter Collier [1]; John Kelly [2]

[1] *STFC - Rutherford Appleton Lab. (GB)*

[2] *STFC Rutherford Appleton Laboratory (UK)*

Corresponding Author: catalin.condurache@stfc.ac.uk

The CernVM-FS is firmly established as a method of software distribution for the LHC experiments at the WLCG sites. Use of CernVM-FS outside WLCG has been growing steadily, with increasing number of Virtual Organizations (VOs) both within High Energy Physics (HEP) communities and in

other disciplines (i.e. Space, Natural and Life Sciences) having identified this technology as a more efficient way of maintaining and accessing their software across the Grid.

We give an overview on the work carried out at RAL to establish an infrastructure able to offer the CernVM-FS service to a broad range of HEP and non-HEP organizations and on the role played by the Tier-1 to support the CernVM-FS developers group.

We discuss the CernVM-FS Replica service delivered as part of the WLCG Stratum-1 network and we then focus on the facility provided to setup a complete service for the non-LHC communities. Master Repository (Stratum-0) and Replica/Mirror (Stratum-1) facilities are presented, together with a customized mechanism to upload and maintain the master repositories by VO Software Grid Managers.

The RAL Tier-1 began offering CernVM-FS to small VOs in the UK in 2012. With support from the EGI the RAL Tier-1 has, since September 2013, been leading the CernVM-FS Task Force which has successfully developed an infrastructure, modeled on that deployed by WLCG, providing a resilient, distributed CernVM-FS service to non-LHC VOs across Europe and indeed replicated around the world. The evolution and current status of the worldwide non-LHC Stratum-0 and Stratum-1 network topology where RAL Tier-1 is a key player is discussed as part of the sustained effort made by the EGI CernVM-FS Task Force to promote the use of CernVM-FS and cooperation on cross-replicating repositories for VOs supported by multiple collaborations.

The pros and cons for a possible CernVM-FS service consolidation are explored. Using a single High Available Stratum-1 system for both WLCG and other HEP and non-HEP communities is cost-effective and will optimise the resources, as long as it is kept as a critical service without rising the maintenance costs.

283

Prototype of a production system for CTA with DIRAC

Author: Luisa Arrabito[1]

Co-authors: Andreas Haupt [2]; Andrei Tsaregorodtsev [3]; Federico Stagni [4]; Johan Bregeon [1]; Ricardo Graciani Diaz [5]

[1] *LUPM-IN2P3 (FR)*

[2] *Deutsches Elektronen-Synchrotron (DE) and Linnaeus University (SE)*

[3] *CPPM, Aix-Marseille Université, CNRS/IN2P3, Marseille, France*

[4] *CERN*

[5] *University of Barcelona (ES)*

Corresponding Author: federico.stagni@cern.ch

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. CTA will operate as an open observatory providing data products to the scientific community. An average data stream of about 1 GB/s for about 1000 hours of observation per year, thus producing several PB/year, is expected. Large CPU time is required for data processing as well for massive Monte Carlo simulations needed for detector calibration purposes.
The current CTA computing model is based on a distributed infrastructure for the archive and the data off-line processing. In order to manage the off-line data processing in a distributed environment, CTA has evaluated the DIRAC system, as base framework for the CTA production system. In particular, a production system prototype has been developed, based on the two main DIRAC components, i.e. the Workload Management and Data Management Systems. After two years of successful exploitation of this prototype, for simulations and analysis, we proved that DIRAC provides suitable functionalities needed for the CTA data processing. Based on these results, the CTA development plan aims to achieve an operational production system, based on the DIRAC Workload Management System, to be ready for the start of CTA operation phase in 2017-2018. One more

important challenge consists of the development of a fully automatized execution of the CTA workflows. For this purpose, we have identified a third DIRAC component, the so-called Transformation System, which offers very interesting functionalities to achieve this automatisation. The Transformation System is a 'data-driven' system, allowing to automatically trigger data processing and data management operations according to pre-defined scenarios. In this paper, we present a brief summary of the DIRAC evaluation done so far, as well as the future developments planned for the CTA production system. In particular, we will focus on the developments of CTA automatic workflows, based on the Transformation System. As a result, we also propose some design optimizations of the Transformation System, in order to fully support the most complex workflows, envisaged in the CTA processing.

284

dCache, enabling tape systems to handle small files efficiently.

Authors: Albert Rossi[1]; Christian Bernardt[2]; Dmitry Litvintsev[1]; Gerd Behrmann[3]; Karsten Schwank[4]; Patrick Fuhrmann[4]; Paul Millar[2]; Tigran Mkrtchyan[5]

[1] *FNAL*

[2] *Deutsches Elektronen-Synchrotron (DE)*

[3] *NDGF*

[4] *DESY*

[5] *Deutsches Elektronen-Synchrotron DESY*

Corresponding Author: karsten.schwank@desy.de

Increasingly, sites are using dCache to support communities that have different requirements from WLCG; as an example, DESY facilities and services now support photon sciences and biology groups. This presents new use-cases for dCache. Of particular interest is the chaotic file size distribution with a peak towards small files. This is problematic because tertiary storage systems, and tape in particular, are optimized for storing large files. Direct storing of the users files results in unacceptably poor performance.

As dCache bridges the filesystem view with the underlying storage and manages transitions between media, it is the natural place to solve the poor performance from storing small files on tape. We achieved this by introducing a new service that reconciles user demand against tape behavior.

The service is transparent to the users and packs files into containers (currently zip files) based on configurable policies. These containers are written directly into the same dCache, which then stores them on tape. Both the small files and the containers file then benefit from dCache features, such as caching and load-balancing. No additional storage is necessary and the service itself scales by running multiple instances within the same dCache, sharing the load.

We describe the design, report on DESY's experience of running the service over the past six months, and detail the future plans for making it generally available.

Track 3 Session / 285

dCache, Sync-and-Share for Big Data

Authors: Albert Rossi[1]; Christian Bernardt[2]; Dmitry Litvintsev[1]; Gerd Behrmann[3]; Karsten Schwank[4]; Patrick Fuhrmann[4]; Paul Millar[2]; Peter van der Reest[4]; Quirin Buchholz[4]; Tigran Mkrtchyan[4]; Volker Guelzow[2]

[1] *FNAL*

[2] *Deutsches Elektronen-Synchrotron (DE)*

[3] *NDGF*

[4] *DESY*

Corresponding Author: paul.millar@desy.de

The availability of cheap, easy-to-use sync-and-share cloud services has split the scientific storage world into the traditional big data management systems and the very attractive sync-and-share services. With the former, the location of data is well understood while the latter is mostly operated in the Cloud, resulting in a rather complex legal situation.

Beside legal issues, those two worlds have little overlap in user authentication and access protocols. While traditional storage technologies, popular in HEP, are based on X509, cloud services and sync-n-share software technologies are generally based on user/password authentication or mechanisms like SAML or Open ID Connect. Similarly, data access models offered by both are somewhat different, with sync-n-share services often using proprietary protocols.

As both approaches are very attractive, dCache.org developed a hybrid system, providing the best of both worlds. To avoid reinvent the wheel, dCache.org decided to embed another Open Source project: OwnCloud. This offers the required modern access capabilities but does not support the managed data functionality needed for large capacity data storage.

With this hybrid system, scientist can share files and synchronize their data with laptops or mobile devices as easy as with any other cloud storage service. On top of this, the same data can be accessed via established mechanisms, like GridFTP to serve the Globus Transfer Service or the WLCG FTS3 tool, or the data can be made available to worker nodes or HPC applications via a mounted filesystem. As dCache provides a flexible authentication module, the same user can access its storage via different authentication mechanisms; e.g., X.509 and SAML. Additionally, users can specify the desired quality of service or trigger media transitions as necessary, so tuning data access latency to the planned access profile. Such features are a natural consequence of using dCache.

We will describe the design of the hybrid dCache/OwnCloud system, report on several months of operations experience running it at DESY, and elucidate on the future road-map.

286

THttpServer class in ROOT

Author: Sergey Linev[1]

[1] *GSI DARMSTADT*

Corresponding Author: s.linev@gsi.de

New THttpServer class in ROOT implements http server for arbitrary ROOT-based application. It is based on embeddable Civetweb server and provides direct access to all registered for the server objects. THttpServer also supports FastCGI interface and therefore can be integrated with many standard web servers like Apache.

Main advantage of http server usage in ROOT –one could access objects data from running applications without need to create temporary files. Together with standard binary representation objects can be converted into JSON (JavaScript Object Notation), using new TBufferJSON class. Server also provides access to data members and collections in registered objects.

Generic user interface for THttpServer implemented with HTML/JavaScript and based on newest JSROOT development. With any modern web browser one could list, display and monitor objects available on http server. Different possibilities are provided to integrate dynamic graphical elements into other HTML pages.

THttpServer available in both 5-34 and 6-02 versions of ROOT. Online documentation and examples can be found on ROOT web site and on http://web-docs.gsi.de/~linev/js/

Track 3 Session / 287

Ceph-based storage services for Run2 and beyond

Author: Dan van der Ster[1]

Co-authors: Andreas Joachim Peters [1]; Herve Rousseau [1]; Luca Mascetti [1]; Massimo Lamanna [1]

[1] *CERN*

Corresponding Author: andreas.joachim.peters@cern.ch

In 2013, CERN IT evaluated then deployed a petabyte-scale Ceph cluster to support OpenStack use-cases in production. As of fall 2014, this cluster stores around 300 TB of data comprising more than a thousand VM images and a similar number of block device volumes. With more than a year of smooth operations, we will present our experience and tuning best-practices.

Beyond the cloud storage use-cases, we have been exploring Ceph-based services to satisfy the growing storage requirements during and after Run2. First, we have developed a Ceph back-end for CASTOR, allowing this service to deploy thin disk server nodes which act as gateways to Ceph; this feature marries the strong data archival and cataloging features of CASTOR with the resilient and performant Ceph subsystem for disk. Second, we have developed RADOSFS, a lightweight storage API which builds a POSIX-like filesystem on top of the Ceph object layer. When combined with Xrootd, RADOSFS can offer a scalable object interface compatible with our HEP data processing applications. Lastly the same object layer is being used to build a scalable and inexpensive NFS service for several user communities.

288

JSROOT version 3 –JavaScript library for ROOT

Authors: Bertrand Bellenot[1]; Sergey Linev[2]

[1] *CERN*

[2] *GSI DARMSTADT*

Corresponding Author: s.linev@gsi.de

This is further development of JSRootIO project. Code was mostly rewritten to make it modular; I/O part was clearly separated from the graphics.

Many new interactive features were implemented:

- loading of required functionality on the fly;
- possibility of dynamic update of objects drawings;
- automatic resize of drawings when browser window is resized;
- move/resize of elements like statbox and title;
- context menu for different elements.

On the central JSROOT page one could open ROOT files, browse their content and display different ROOT objects like histograms, graphs or canvases. Produced drawings can be easily integrated into other HTML pages via 'iframe' tag. Simple and flexible JSROOT API provided; it is used to implement graphical interface for THttpServer class.

JSROOT is now part of actual ROOT distribution. On http://web-docs.gsi.de/~linev/js/ one could found latest version together with documentation and examples.

Track 4 Session / 289

Using the glideinWMS System as a Common Resource Provisioning Layer in CMS

Author: Oliver Gutsche[1]

Co-authors: Alison Mc Crea [2]; Brian Paul Bockelman [3]; Christoph Wissing [4]; David Alexander Mason [1]; David Colling [5]; Dirk Hufnagel [1]; Farrukh Aftab Khan [6]; Igor Sfiligoi [7]; James Letts [2]; Justas Balcas [8]; Krista Larson [1]; Marco Mascheroni [9]; Maria Dolores Saiz Santos [2]; Stefan Piperov [10]; Stefano Belforte [11]

[1] *Fermi National Accelerator Lab. (US)*

[2] *Univ. of California San Diego (US)*

[3] *University of Nebraska (US)*

[4] *Deutsches Elektronen-Synchrotron (DE)*

[5] *Imperial College Sci., Tech. & Med. (GB)*

[6] *National Centre for Physics (PK)*

[7] *University of California San Diego*

[8] *Vilnius University (LT)*

[9] *Universita & INFN, Milano-Bicocca (IT)*

[10] *Brown University (US)*

[11] *Universita e INFN (IT)*

Corresponding Author: letts@physics.ucsd.edu

CMS will require access to more than 125k processor cores for the beginning of Run2 in 2015 to carry out its ambitious physics program with more and higher complexity events. During Run1 these resources were predominantly provided by a mix of grid sites and local batch resources. During the long shut down cloud infrastructures, diverse opportunistic resources and HPC supercomputing centers were made available to CMS, which further complicated the operations of the submission infrastructure. In this presentation we will discuss the CMS effort to adopt and deploy the glideinWMS system as a common resource provisioning layer to grid, cloud, local batch, and opportunistic resources and sites. We will address the challenges associated with integrating the various types of resources, the efficiency gains and simplifications associated with using a common resource provisioning layer, and discuss the solutions found. We will finish with an outlook of future plans for how CMS is moving forward on resource provisioning for more heterogenous architectures and services.

290

Integration of DD4hep in the Linear Collider Software Framework

Author: Andre Sailer[1]

Co-authors: Frank-Dieter Gaede [2]; Marko Petric [3]; Markus Frank [1]; Nikiforos Nikiforou [1]; Shaojun Lu [4]

[1] *CERN*

[2] *CERN/DESY*

[3] *Jozef Stefan Institute (SI)*

[4] *DESY*

Corresponding Author: andre.philippe.sailer@cern.ch

The DD4hep detector description toolkit offers a flexible and easy to use solution for the consistent and complete description of particle physics detectors in one single system. It provides software components addressing visualisation, simulation, reconstruction and analysis of high energy physics

data.
The Linear Collider community has adopted DD4hep early on in the development phase and actively participated in the design of the toolkit. The CLICdp and ILD detector working groups have reimplemented their simulation models in DD4hep, thereby ensuring a well defined hierarchy of the model, an accurate material description and a flexible scaling behaviour for the purpose of detector optimisation.
In parallel, an interface for reconstruction offering a high level view on the detector geometry has been developed.
It is based on surfaces assigned to detector components, which provide access to all relevant material properties and allow for fast and efficient navigation during track and calorimeter reconstruction.
In this talk we give a brief introduction to DD4hep, then focus on the description of the new simulation models and the reconstruction interface and its application to track reconstruction and particle flow at the ILC and CLIC.

Track 6 Session / 291

Named Data Networking in Climate Research and HEP Applications

Authors: Artur Jerzy Barczyk[1]; Susmit Shannigrahi[2]

Co-authors: Alex Sim [3]; Christos Papadopoulos [2]; Edmund Yeh [4]; Harvey Newman [1]; Inder Monga [5]; John Wu [3]

[1] *California Institute of Technology (US)*

[2] *Colorado State University*

[3] *LAWRENCE BERKELEY NATIONAL LABORATORY*

[4] *Northeastern University*

[5] *ESNET*

Corresponding Author: christos@cs.colostate.edu

Introduction

The Computing Models of the LHC experiments continue to evolve from the simple hierarchical MONARC model towards more agile models where data is exchanged among many Tier2 and Tier3 sites, relying on both strategic data placement, and an increased use of remote access with caching through CMS's AAA and ATLAS' FAX projects, for example. The challenges presented by expanding needs for CPU, storage and network capacity have pointed the way towards future more agile pervasive models that make best use of highly distributed heterogeneous resources.

In this paper, we explore the use of **Named Data Networking (NDN)** 1, a new Internet architecture focusing on content rather than the location of the data collections. As NDN has shown considerable promise in another data intensive field, Climate Science, we discuss the similarities and differences between the Climate and HEP use cases, along with specific issues HEP faces and will face during LHC Run2 and beyond, which NDN could address.

NDN

NDN, an instance of Information Centric Networking (ICN), is a new Internet architecture which focuses on the content of the data collections themselves, rather than on where the data resides. The end host addresses are replaced with content names, which, similar to URLs, are hierarchical, unique and human readable. Thus, NDN imposes minimal

structure on applications, which can choose their own naming schemes.
The hierarchical structure of NDN names has several advantages:

1. it is an intuitive, common organizational structure (e.g., file systems, URLs, etc.),
2. it is scalable (similar to hierarchical IP addresses), and
3. coupled with longest prefix matching, it allows for data discovery and enumeration.

NDN has a wide range of potential benefits such as in-network content caching with request deduplication to reduce congestion and improve delivery speed, simpler application configuration, and security built into the network at the data level.

The NDN concepts, structure and initial applications have been developed through an NSF Future Internet Architecture project in its second round of funding, involving eight universities.
NDN has attracted significant interest from industry, including Cisco, Intel, Alcatel, Huawei, and Panasonic, and involves many of these companies through an industry consortium.

NDN and Climate Applications

We have successfully begun to test NDN in the climate application domain 2.
To handle the various naming schemes used in climate applications, we have designed and implemented translators that take existing names with arbitrary structure (produced by climate models, or home-grown) and translated them into NDN-compliant names. Depending on the original name structure, the translation can be fairly direct (e.g., data that complies with the "Data Reference Syntax" from the Coupled Model Intercomparison Project), or complex (from home-grown naming schemes that require the analysis of metadata embedded in the dataset or even user feedback in order to construct proper NDN names).

We have deployed a dedicated 6-node testbed for climate applications that reaches locations such as the Atmospheric and Computer Science Departments at Colorado State University, LBNL and NWCS. The testbed is connected via 10G links by ESnet and is composed of high-end machines each with 40 core CPUs, 128GB RAM and 48TB diskspace. The machines cumulatively host over 50TB of climate data and are used for research, experimentation and development of climate applications.

NDN Support for HEP Applications

Several features of NDN can be beneficial to the HEP computing use case.
Data sources publish new content to the network following an agreed upon naming scheme. Data delivery is always performed in a pull mode, driven by the consumer issuing interest packets. Intermediate nodes in the network dynamically cache data based on content popularity, ready to satisfy subsequent interests directly from the cache, thus lowering the load on servers with popular content.
Combining this with the pull-mode results in a multicast-like data delivery, possibly optimizing both the network utilisation as well as server load.

The use of multiple data sources simultaneously, as well as the native use of multiple paths between client and data source, provide for robust failover in case of network segment, node, or end-site failure.

All these are active research areas today. Caching as well as forwarding strategies, naming schemes, multi-sourcing and multi-path forwarding need to be investigated not only from the network but also the application perspective.

HEP experiments using the World-wide LHC Computing Grid (WLCG) have well-developed, hierarchical naming schemes in use, which already fit the NDN

approach well. We take this logical file name structure as a starting point for investigating the benefits of using NDN as the data distribution and access network for HEP data processing. For this, we use the testbed described above. We further target simultaneous optimization of storage and bandwidth resource utilization through dynamic caching using the VIP framework in 3.
For the scalability study, we complement the testbed with the use of a simulation environment with a representative topology including network nodes and end-sites.

Summary

In this paper, we study data access over an NDN testbed developed for Climate research. We study the behaviour using HEP-like data structures based on the CMS naming scheme, showing data publishing, discovery and retrieval in an NDN network. We demonstrate the benefits of caching, speeding up data delivery in multi-job access from a single source, with jobs executing at multiple sites. We also show the results of the simulation studies of remote data access over an NDN network demonstrating the scalability of the system.

References

1. V. Jacobson, et al.; "Networking Named Content", 2009
2. C. Olschanowsky, et al., "Supporting Climate Research using Named Data Networking", LANMAN, 2014
3. E. Yeh, et al.; "VIP: A Framework for Joint Dynamic Forwarding and Caching in Named Data Networks", Proc. ACM Conf. on Information-Centric Networking, 2014

292

USER AND GROUP STORAGE MANAGEMENT AT THE CMS CERN T2 CENTRE

Authors: Andreas Pfeiffer[1]; Gianluca Cerminara[1]; Giovanni Franzoni[1]

[1] *CERN*

Corresponding Author: giovanni.franzoni@cern.ch

A wide range of detector commissioning, calibration and data analysis tasks is carried out by members of the Compact Muon Solenoid (CMS) collaboration using dedicated storage resources available at the CMS CERN Tier-2 centre.

Relying on the functionalities of the EOS disk-only storage technology, the optimal exploitation of the CMS user/group resources has required the introduction of policies for data access management, data protection, cleanup campaigns based on access pattern, and long term tape archival.

The resource management has been organised around the definition of working groups and the delegation to an identified responsible of each group composition.

In this poster contribution we illustrate the user/group storage management, and the development and operational experience at the CMS CERN Tier-2 centre in the 2012-2015 period.

293

Software Management for the NOvA Experiment

Authors: Brian Rebel[1]; Gavin Davies[2]; Jan Zirnstein[3]; Kanika Sachdev[4]

Co-author: Robert Group [5]

[1] *Fermi National Accelerator Laboratory*

[2] *Indian University, Bloomington, Indiana, USA*

[3] *University of Minnesota, Twin Cities, Minnesota, USA*

[4] *University Of Minnesota, Twin Cities, MN, USA*

[5] *University of Virginia*

Corresponding Author: rcg6p@virginia.edu

The NOvA software (NOvASoft) is written in C++ and built on the Fermilab Computing Division's ART framework that uses ROOT analysis software. NOvASoft makes use of more than 50 external software packages, is developed by more than 50 developers and is used by more than 100 physicists from over 30 universities and laboratories in 3 continents. The software builds are handled by Fermilab's custom version of Software Release Tools (SRT), a UNIX based software management system for large, collaborative projects that is used by several experiments at Fermilab. The system provides software version control with SVN configured in a client-server mode and is based on the code originally developed by the BaBar collaboration. In this paper, we present efforts towards distributing the NOvA software via the CERN VMFS distributed file system. We will also describe our recent work to use CMake build system and Jenkins, the open source continuous integration system, for NOvASoft.

Track 7 Session / 294

Utilizing cloud computing resources for BelleII

Author: Randy Sobie[1]

Co-authors: Martin Sevior [2]; Takanori HARA [3]

[1] *University of Victoria (CA)*

[2] *University of Melbourne (AU)*

[3] *KEK*

Corresponding Author: rsobie@uvic.ca

The BelleII experiment is developing a global computing system for the simulation of MC data prior its collecting real collision data in the next few years. The system utilizes the grid middleware used in the WLCG and uses the DIRAC workload manager. We describe how IaaS cloud resources are being integrated into the BelleII production computing system in Australia and Canada. The IaaS resources include HEP as well as opportunistic and commercial clouds.

In Canada, the cloud resources are managed by a DIRAC installation at the University of Victoria, which acts as a slave to the DIRAC instance at KEK. A workload management service running on the DIRAC server at the University of Victoria submits pilot jobs to an HTCondor queue dedicated to the distributed cloud system. The CloudScheduler VM provisioning service boots the VMs based on the HTCondor queue. The distributed cloud uses resources in Europe and North America including Amazon EC2.

Australia provides it's contribution to the Belle II Distributed Computing solution via CPU resources supplied by the NeCTAR Open Stack cloud. The Australian solution employs a Dynamic Torque Batch system with cloud-based worker nodes as a backend to an EGI CREAM-CE. The DIRAC interware sees this a conventional grid cluster with no further configuration or tuning required. The worker nodes employ an SL6 operating system configured via puppet and with the Belle II application software provided via CVMFS. At the time of writing, the Australian system efficiently supplies 350 worker nodes to the Belle II Distributed Computing solution.

All the clouds have been successfully used in BelleII MC production campaigns, producing a substantial fraction of the simulated data samples.

295

Updating the LISE^{++} software and future upgrade plans

Author: Michelle Kuchera[1]

Co-authors: Bradley Sherrill [1]; Daniel Bazin [2]; Oleg Tarasov [2]

[1] *National Superconducting Cyclotron Laboratory, Michigan State University*

[2] *National Superconducting Cyclotron Laboratory*

Corresponding Author: kucheram@nscl.msu.edu

Production of new isotopes is one of the opportunities at the intensity frontier of nuclear physics. The associated science ranges from tests of the Standard Model to exploration of the origin and evolution of the chemical elements in the universe. Leading facilities in this effort are RIBF at RIKEN, TRIUMF in Canada, and ISOLDE at CERN. New large scale facilities under development at the nuclear intensity frontier include FAIR in Europe, FRIB in the United States, and others in countries including China, France, Korea, and Italy.

This talk will describe capabilities and future upgrade plans for the isotope production and simulation software that is used at many of these facilities, namely LISE
$textsuperscript{++}$ 1. For reference of its wide-scale use, the LISE
$textsuperscript{++}$ website had approximately 3000 unique visitors in 2013, with Japan, USA, Germany, France, and China as the top five countries in terms of visitors.

LISE
$textsuperscript{++}$ is software used to predict beam intensity and purity of rare isotope beams produced in-flight by magnetic and electric separators. The primary use of LISE
$textsuperscript{++}$ at most facilities is to predict and identify the composition of Radioactive Nuclear Beams 1. Intensity and purity of a desired beam can be predicted, along with the separator magnet settings. Included in the LISE
$textsuperscript{++}$ package are models of isotope production mechanisms, ion optical transport through magnetic and electric systems, and ion interactions in matter. The suite includes a full set of utilities for simulation of experiments.

The talk will highlight the process and methods of updating the software while retaining the computational integrity of the code. To accommodate the diversity of our users, we extend the software from Windows to a cross platform application. In addition, the C++ standard will be updated from Borland to C++11. The calculations of beam transport and isotope production are becoming more computationally intense with the new large scale facilities. For example, the 90 m long FRIB separator will have around fifty magnetic elements and ten points of beam interactions with matter. In order to perform the calculations in acceptable time, numerical optimization and parallel methods are applied. Computational improvements as well as the process of updating this large code will be discussed.

1 LISE++: Radioactive beam production with in-flight separators. O.B. Tarasov, D. Bazin

Track 3 Session / 296

EOS as the present and future solution for data storage at CERN

Author: Andreas Joachim Peters[1]

Co-author: Dirk Duellmann [1]

[1] *CERN*

Corresponding Author: andreas.joachim.peters@cern.ch

EOS is an open source distributed disk storage system in production since 2011 at CERN. Development focus has been on low-latency analysis use cases for LHC and non-LHC experiments and life-cycle management using JBOD hardware for multi PB storage installations. The EOS design implies a split of hot and cold storage and introduced a change of the traditional HSM functionality based workflows at CERN.

The 2015 deployment brings storage at CERN to a new scale and foresees to breach 100 PB of disk storage in a distributed environment using tens of thousands of (heterogeneous) hard drives. EOS has brought to CERN major improvements compared to past storage solutions by allowing quick changes in the quality of services of storage pools. This allows the data centre to quickly meet the changing performance and reliability requirements of the LHC experiments with minimal data movements and dynamic reconfigurations. For example, the software stack has met the specific needs of the dual computing centre set-up required by CERN and allowed the fast design of new workflows accommodating the separation of long-term tape archive and disk storage required for the LHC Run II.

The talk will give a high-level state of the art overview of EOS with respect to Run II, introduce new tools and use cases and set the new roadmap for the next storage solutions to come.

Track 3 Session / 297

Integrating CEPH in EOS

Author: Andreas Joachim Peters[1]

Co-authors: Dan van der Ster [1]; Joaquim Manuel Pereira Rocha [1]

[1] *CERN*

Corresponding Author: andreas.joachim.peters@cern.ch

The EOS storage software was designed to cover CERN disk-only storage use cases in the medium-term trading scalability against latency. To cover and prepare for long-term requirements the CERN IT data and storage services group (DSS) is actively conducting R&D and open source contributions to experiment with a next generation storage software based on CEPH.

CEPH provides a scale-out object storage system (RADOS) and additionally various optional high-level services like S3 gateway, RADOS block devices and a POSIX compliant file system (CephFS). The acquisition of CEPH by Redhat underlines the promising role of CEPH as the open source storage platform of the future.

CERN IT is running a CEPH service in the context of OpenStack on a moderate scale of 1 PB replicated storage. Building a 100+PB storage system based on CEPH will require software and hardware tuning. It is of capital importance to demonstrate the feasibility and possibly iron out bottlenecks and blocking issues beforehand.

Main idea behind this R&D is to leverage and contribute to existing building blocks in the CEPH storage stack and implement a few CERN specific requirements in a thin customizable storage layer.

The presentation will introduce various open source developments & contributions, their applicability and first performance figures of a next generation storage platform aka EOS Diamond:

- radosFS - a RADOS based client API providing a scale-out meta data catalog and pseudo-hierarchical storage view with optimized(POSIX-light) directory and file access, parallel meta data queries, modification time propagation, file striping …
- Erasure Codes - local reconstruction and reed solomon codes for storage cost reduction and improved reliability

- a multi-protocol CEPH Overlay Service based on the XRootD framework providing data localization, multi-user policies, strong authentication, WebDAV/XRootD access ...
- a read-write Federated Storage Cloud platform based on storage virtualization and infrastructure aware scheduling

Track 3 Session / 298

Archiving tools for EOS

Author: Elvin Alin Sindrilaru[1]

Co-authors: Andreas Joachim Peters [1]; Dirk Duellmann [1]

[1] *CERN*

Corresponding Author: andreas.joachim.peters@cern.ch

Archiving data to tape is a critical operation for any storage system, especially for the EOS system at CERN which holds production data from all major LHC experiments. Each collaboration has an allocated quota it can use at any given time therefore, a mechanism for archiving "stale" data is needed so that storage space is reclaimed for online analysis operations.

The archiving tool that we propose for EOS aims to provide a robust interface for moving data between EOS and the tape storage system while enforcing data integrity and verification.

The archiving infrastructure is written in Python and is fully based on the XRootD framework. All data transfers are done using the third-party copy mechanism which ensures point-to-point communication between the source and destination, thus providing maximum aggregate throughput. Using ZMQ message-passing paradigm and a process-based approach enabled us to archive optimal utilisation of the resources and a stateless architecture which can easily be tuned during operation.

In conclusion, we make a comparative analysis between archiving a data set in a "managed" way using the archiving tool and the old plain copy method, highlighting the speed-up gain, the data integrity checks performed and the behaviour of the system in different failure scenarios. We expect this tool to considerably improve the data movement work-flow at CERN in both directions between disk and tape.

Track 6 Session / 300

The production deployment of IPv6 on WLCG

Author: Dave Kelsey[1]

Co-authors: Alastair Dewhurst [1]; Andrea Sciaba [2]; Bruno Heinrich Hoeft [3]; Christopher John Walker [4]; Costin Grigoras [2]; Duncan Rand [5]; Edoardo Martelli [2]; Ewan MacMahon [6]; Fernando Lopez Munoz [7]; Francesco Prelz [8]; Jerome Bernier [9]; Jiri Chudoba [10]; Kars Ohrenberg [11]; Keith Chadwick [12]; Marek Elias [13]; Raja Nandakumar [1]; Ramiro Voicu [14]; Simon Fayer [5]; Simone Campana [2]; Thomas Finnern [11]; Tiju Idiculla [15]; Tony Wildish [16]; Ulf Tigerstedt [17]

[1] *STFC - Rutherford Appleton Lab. (GB)*

[2] *CERN*

[3] *KIT - Karlsruhe Institute of Technology (DE)*

[4] *University of London (GB)*

[5] *Imperial College Sci., Tech. & Med. (GB)*

[6] *University of Oxford*

[7] *Universitat Autònoma de Barcelona (ES)*

[8] *Università degli Studi e INFN Milano (IT)*

[9] *IN2P3 Computing Center*

[10] *Acad. of Sciences of the Czech Rep. (CZ)*

[11] *DESY*

[12] *Fermilab*

[13] *FZU ASCR*

[14] *California Institute of Technology (US)*

[15] *STFC - Rutherford Appleton Laboratory*

[16] *Princeton University (US)*

[17] *CSC Oy*

Corresponding Author: david.kelsey@stfc.ac.uk

The world is rapidly running out of IPv4 addresses; the number of IPv6 end systems connected to the internet is increasing; WLCG and the LHC experiments may soon have access to worker nodes and/or virtual machines (VMs) possessing only an IPv6 routable address. The HEPiX IPv6 Working Group (http://hepix-ipv6.web.cern.ch/) has been investigating, testing and planning for dual-stack services on WLCG for several years. Following feedback from our working group, many of the storage technologies in use on WLCG have recently been made IPv6-capable. The worldwide HEP computing community now needs to deploy dual-stack IPv6/IPv4 services on WLCG to allow such use of IPv6-only resources.

This paper will present the IPv6 requirements, tests and plans of each of the four LHC experiments together with the tests performed both on the IPv6 test-bed and in targeted use of WLCG production services. This is primarily aimed at IPv6-only worker nodes or VMs accessing several different implementations of a global dual-stack federated storage service. The changes required to the operational infrastructure, including monitoring and security, will be addressed as will the implications for site management. The working group will present its deployment plan for dual-stack storage services, together with other essential central and monitoring services, to start during 2015.

301

LHCb EventIndex

Author: Alexey Artemov[1]

Co-authors: Andrey Ustyuzhanin [2]; Federico Stagni [3]; Marco Cattaneo [3]; Marco Clemencic [3]; Nikita Kazeev [2]; Philippe Charpentier [3]

[1] *Yandex*

[2] *Moscow Institute of Physics and Technology, Moscow*

[3] *CERN*

Corresponding Authors: andrey.ustyuzhanin@cern.ch, kazeevn@yandex-team.ru

The LHCb experiment routinely generates up to 10ˆ10 events per year. Organizing such an amount of data in a convenient manner for interactive analysis is non-trivial. It becomes even more complicated as every event undergoes several versions of reconstructions, and users have to be able to navigate through many different versions of the same event. This paper presents the LHCb EventIndex: an event search system designed for organizing LHCb events. Its primary goal is to allow a fast selection of subsets of events fulfilling a combination of high-level conditions, such as which trigger has fired, or how many muons were found in the event.

The system is designed with scalability and extensibility in mind. We describe the part of the system that deals with Grid scanning, event storage and user interface. Events are stored by modern open-source NoSQL solutions, with row keys being organized so that certain searches can be run with

only a partial table scan. Such embarrassingly parallel organization makes search time upper bound equal to a full scan of a regional server, a constant independent of the total amount of data stored. User interaction is handled by a master-slave python middleware.

We demonstrate the scalability of the system by performance evaluation on the real LHCb datasets. This system is also being evaluated by the LHC ATLAS experiment.

Track 3 Session / 303

Disk storage management for LHCb based on Data Popularity estimator

Author: Mikhail Hushchyn[1]

Co-authors: Andrey Ustyuzhanin [2]; Marco Cattaneo [3]; Philippe Charpentier [3]

[1] *Moscow Institute of Physics and Technology, Moscow*

[2] *ITEP Institute for Theoretical and Experimental Physics (RU)*

[3] *CERN*

Corresponding Author: hushchyn.mikhail@gmail.com

The amount of data produced by the LHCb experiment every year consists of several petabytes. This data is kept on disk and tape storage systems. Disks are much faster than tapes, but are way more expensive and hence disk space is limited. It is impossible to fit the whole data taken during the experiment's lifetime on disk, but fortunately fast access to datasets are no longer needed after the analysis requiring them are over. So it is highly important to identify which datasets should be kept on disk and which ones should be kept as archives on tape. The metrics to be used for deprecating datasets'caching is based on the "popularity"of this dataset, i.e. whether it is likely to be used in the future or not. We discuss here the approach and the studies carried out for optimizing such a Data Popularity estimator.

Input information to the estimator are the dataset usage history and metadata (size, type, configuration etc). The system is designed to select the datasets which may be used in the future and thus should remain on disk. Studies have therefore been performed on how to optimize the usage of dataset information from the past for predicting its future popularity. In particular, we have carried out a detailed comparison of various time series analysis, machine learning classifier, clustering and regression algorithms. We demonstrate that our approach is capable of improving significantly the disk usage efficiency.

304

Glint: VM image distribution in a multi-cloud environment

Author: Randy Sobie[1]

Co-authors: Colin Roy Leavett-Brown [1]; Frank Olaf Berghaus [1]; Ian Gable [1]; Michael Paterson [2]; Ron Desmarais [3]; Ryan Taylor [1]

[1] *University of Victoria (CA)*

[2] *U*

[3] *University of Victoria*

Corresponding Author: rsobie@uvic.ca

The HEP community is increasingly using clouds that are distributed around the world for running its applications. The stringent software criteria of HEP experiments require that we use the identical

(secure) virtual machine (VM) image at all sites with a minimal set of site-specific customizations. Nearly all cloud systems (such as OpenStack) require that the VM image to be instantiated must be stored in the local repository (e.g. Glance in OpenStack). This requirement forces the user to manually distribute images to target clouds. As the number of clouds utilized grows, this procedure becomes increasingly time-consuming and error-prone. To remedy this situation, we designed and constructed an image replication system, called Glint, to mange VM images in multiple OpenStack Glance repositories.

On a single OpenStack cloud, Glance gives the user (with the appropriate credentials) the ability to use pre-configured images or upload customized images using the OpenStack Horizon web interface or command line interface. A user with access to multiple OpenStack clouds (each with separate credentials) will likely find that the set of pre-configured images will be different at each site and the user will need to upload their customized image to each cloud.

Glint is designed to streamline the distribution of application images to multiple clouds. Glint integrates into the OpenStack framework and provides access to its functions via the OpenStack Horizon web interface. It registers with the OpenStack Keystone Identity service during installation and authorizes requests by validating the user's security token with the local Keystone service. Glint gives the user ability to add remote clouds (with their credentials) and migrate images from one to many cloud sites.

We describe the features of Glint and its integration in the OpenStack framework. The initial use of Glint for HEP application is described. Future developments and potential integration into the OpenStack code base are discussed.

Track 7 Session / 305

CernVM WebAPI - Controling Virtual Machines from the Web

Author: Ioannis Charalampidis[1]

Co-authors: Gerardo Ganis [1]; Jakob Blomer [1]; Peter Skands [1]; Predrag Buncic [1]; Rene Meusel [1]

[1] *CERN*

Corresponding Author: ioannis.charalampidis@cern.ch

Lately there is a trend in scientific projects to look for computing resources in the volunteering community. In addition, to reduce the development effort required to port the scientific software stack to all the known platforms, the use of Virtual Machines (VMs) as end-projects is becoming increasingly popular. Unfortunately, the installation and the interfacing with the existing volunteering computing infrastructure (such as BOINC) is left to the end-user, therefore restricting even more the audience to only sufficiently software-capable people.

CernVM WebAPI is a software solution addressing this specific case in a way that opens wide new application opportunities. It offers a very simple API for setting-up, controlling and interfacing with a VM instance in the user's computer, while in the same time offloading the
user from all the burden of downloading, installing and configuring the hypervisor. WebAPI comes with a lightweight javascript library that guides the user through the application installation process. Malicious usage is prohibited by offering a per-domain PKI validation mechanism.

In this contribution we will overview this new technology, discuss its security features and examine some test cases where it is already in use.

Track 2 Session / 306

Cellular Automaton based Track Finding for the Central Drift Chamber of Belle II

Author: Oliver Frost[1]

Co-author: Claus Kleinwort [2]

[1] *DESY*

[2] *Deutsches Elektronen Synchrotron (DESY)*

Corresponding Author: oliver.frost@desy.de

With the upgraded electron-positron-collider facility, SuperKEKB and Belle II, the Japanese high energy research center KEK strives to exceed its own world record luminosity by a factor of 40.
To provide a solid base for the event reconstruction within the central drift chamber in the enhanced luminosity setup, a powerful track finding algorithm coping with the higher beam induced backgrounds is developed at DESY.
Pursing a bottom-up approach, which is less susceptible to the increased number of background hits compared to global event reconstruction techniques such as the Hough transformation and its successors, we present a generalization of the cellular automaton.
While maintaining the high execution speed by circumventing the combinatorial backtracking in the graph of local hit information and extrapolations naturally arising in bottom-up approaches, this so called weighted cellular automaton integrates the adaptiveness of the Hopfield network into the original algorithm.

307

SDN implementations at IHEP

Author: Shan Zeng[1]

Co-authors: Fazhi Qi [1]; Gang Chen [1]; Tao Cui [1]; Zhihui Sun [1]

[1] *IHEP*

Corresponding Author: zengshan@ihep.ac.cn

This Paper describes two research aspects and practices of SDN at IHEP. The first one is the SDN practice for the data transferring across the internet, in which a virtual private network based on SDN is designed and built,and an intelligent network route algorithm is developed and deployed in the SDN controller to make full use of IPv6 resources. Experimental results show that this practice can solve the network bandwidth problem between cooperation members of the high energy physics experiments around China.The second one is about the deployment of SDN in the internal data center of IHEP, which provides the flexible network management in the virtualization environment, but the latency and the throughput of the SDN controller is far from our expectation. In order to improve the performance of the SDN controller, we compare four controllers in our second test bed, including NOX, NOX-MT,Beacon and Maestro, and analyzed the performance both in latency and in throughput. At last, a new SDN controller design through the method of patching and reducing the data path request frequency,which is dedicated to high performance SDN in the internal data center of IHEP is proposed.

Track 2 Session / 308

BESIII Physics Data Storing and Processing on HBase and MapReduce

Author: Xiaofeng LEI[1]

Co-authors: Gongxing SUN [2]; Qiang LI [2]

[1] *INSTITUE OF HIGH ENERGY PHYSICS, University of Chinese Academy of Sciences*

[2] *INSTITUE OF HIGH ENERGY PHYSICS*

Corresponding Author: leixf@ihep.ac.cn

In the past years, we have successfully applied Hadoop to high-energy physics analysis. Although, we have not only improved the efficiency of data analysis, but also reduced the cost of cluster building so far, there are still some spaces to be optimized, like static pre-selection, low-efficient random data reading and I/O bottleneck caused by Fuse which is used to access HDFS. In order to change this situation, this paper presents a new analysis platform for high-energy physics data storing and analyzing.
The data structure is changed from DST tree-like files to HBase according to the features of the data itself and analysis processes, since HBase is more suitable for processing random data reading than DST files and enable HDFS to be accessed directly. A few of optimization measures are taken for the purpose of getting a good performance and as well as a customized protocol is defined for data serializing and desterilizing for the sake of decreasing the storage space in HBase. In order to make full use of locality of data storing in HBase, utilizing a new MapReduce model and a new split policy for HBase regions are proposed in the paper. In addition, we establish a dynamic pluggable easy-to-use tag (event metadata) based pre-selection subsystem. It can assist physicists even to filter out 999‰ uninterested data, if the conditions are set properly. This means that a lot of I/O resources can be saved, the CPU usage can be improved and consuming time for data analysis can be reduced. Finally, several use cases are designed, the test results show that the new platform has an excellent performance with 3.5 times faster with pre-selection and 20% faster without pre-selection, and the new platform is stable and scalable as well.

309

The SNiPER offline software framework for non-collider physics experiments

Authors: Jiaheng Zou[1]; Weidong Li[1]; Xingtao Huang[2]

Co-authors: Guofu Cao [1]; Kun Zhang [1]; Tao Lin [1]; Teng Li [2]; Ziyan Deng [1]

[1] *IHEP*

[2] *SDU*

Corresponding Authors: zoujh@ihep.ac.cn, liwd@ihep.ac.cn, huangxt@sdu.edu.cn

SNiPER (the abbreviation of Software for Non-collider Physics ExpeRiments) has been developed based on common requirements from both cosmic ray and nuclear reactor neutrino experiments. This contribution will introduce the detailed design and implementation of the SNiPER software. Compared to the existing offline software frameworks in the high energy physics domain, the design of SNiPER is more focused on efficiency and flexibility. It contains a compact kernel for software components management, job configuration, event execution control, input/output data flow, etc. Both event data model and in-memory data management can be customized by different applications. In SNiPER, a job is composed of one or more low coupling tasks, and various services and algorithms can be specified in each task. By using this kind of structure, both multiple I/O streams and conditional execution of algorithm subsets can be easily realized. In order to facilitate event correlation analysis (a necessity for non-collider physics experiments), the FIFO buffer is used to store adjacent events and a sophisticated method of memory updating is applied, so that accessing to successive events within the user-defined window according to event timestamp becomes possible. The SNiPER also supports concurrent computing by running multi-threads concurrently with a single job. Currently a nascent product of SNiPER has been released and is being used by the JUNO and LHAASO experiments. The practices show that the software architecture is universal and expandable, and SNiPER can be used by non-collider physics experiments to build their offline data analysis and processing systems.

Track 4 Session / 310

Accelerating Debugging In A Highly Distributed Environment

Author: Andrew Hanushevsky[1]

[1] *STANFORD LINEAR ACCELERATOR CENTER*

Corresponding Author: abh@stanford.edu

As more experiments move to a federated model of data access the environment becomes highly distributed and decentralized. In many cases this may pose obstacles in quickly resolving site issues; especially given vast time-zone differences. Spurred by ATLAS needs, Release 4 of XRootD incorporates a special mode of access to provide remote debugging capabilities. Essentially, XRootD allows a site to grant secure access to specific individuals to view certain XRootD information (e.g. log files, configuration files, etc). In a virtual view all of the information is laid out in a site independent way regardless of how the site configured its system. This allows experts at other locations to assist in resolving issues and alleviates time-zone vagaries. The view is available through XRootd or, optionally, HTTP. This talk provides the motivation for developing the remote debugging facility, why it is essential in highly distributed environments, and what can actually be done with it.

Track 1 Session / 311

The data acquisition system of the XMASS experiment

Author: Katsuki Hiraide[1]

[1] *the University of Tokyo*

Corresponding Author: hiraide@km.icrr.u-tokyo.ac.jp

XMASS is a multi-purpose low-background experiment with a large volume of liquid xenon scintillator at Kamioka in Japan. The first phase of the experiment aiming at direct detection of dark matter was commissioned in 2010 and is currently taking data.

The detector uses ˜830 kg of liquid xenon viewed by 642 photomultiplier tubes (PMTs). Signals from 642 PMTs are amplified and read out by 1 GS/s digitizers (CAEN V1751) as well as ADC/TDC modules. To reduce data size, we implemented an on-board data suppression algorithm in digitizers. The trigger is generated based on the number of hit PMTs within a 200 ns coincidence window. Recently, it is pointed out that the XMASS detector has also a great potential to detect supernova neutrino burst, and therefore several DAQ upgrades and system extensions are proceeding for this purpose.

We will present the overall design and performance of the XMASS data acquisition system and the status of DAQ upgrade for supernova neutrino burst detection.

312

Renovation of HEPnet-J for near-future experiments

Author: Soh Suzuki[1]

Co-authors: Fukuko YUASA [1]; Malachi Schram ; Takanori HARA [1]; Tomoaki Nakamura [2]; Yukio KARITA [1]

[1] *KEK*

[2] *High Energy Accelerator Research Organization (JP)*

Corresponding Author: soh.suzuki@kek.jp

Formerly most of HEP experiments in japan used
the centralized computing model.

Originally HEPnet-J had only one instance which is
connected to Internet, and recently it has many closed
network which connects domestic sites.
At that time, the network connectivity in Japan
was very poor and the main purpose of HEPnet-J was
providing enough connectivity for interactive use
over domestic and international links funded by KEK.
During last 10 years,
the domestic and international connectivity
provided by each university and NREN has been
dramatically improved and it is enough for
manual transfer of typical mDST files.

The rapid growth of data volume
makes it unable to apply same model to new generation experiments.
As the LCG tier structure for LHC has proved that
the distributed computing model over collaboration sites
is really applicable to the huge scale experiment,
the external connectivity for international collaboration sites
should be more faster, and more secure.

For example,
the Belle II experiment in KEK
will have huge data repositories in US and EU.
The expected throughput from KEK to US is 20Gbps,
thus we have to bypass the security device like a firewall
for this data path.
The computing facility in KEK (KEKCC) is not enclosed
in the firewall of KEK, but it has a dedicated firewall that
is managed by the independent policy.
Its throughput is sufficient to handle an ordinary activity
over Grid VO.
But it is not enough for the mass-data transmission for Belle II,
we have to bypass it in the same way of LHCONE.
Now bypass lines for Belle II are prepared and under tasting.

We will report how we renovate HEPnet-J for experiments
driven by international collaboration.

313

Directory Search Performance Optimization of AMGA for the Belle II Experiment

Author: Geun Chul Park[1]

Co-authors: Jae-Hyuck Kwak Kwak [2]; Soonwook Hwang [2]; Taesang Huh [2]

[1] *KiSTi Korea Institute of Science & Technology Information (KR)*

[2] *KISTI*

Corresponding Author: gcpark@kisti.re.kr

AMGA (ARDA Metadata Grid Application) is a grid metadata catalog system that has been developed as a component of the EU FP7 EMI consortium based on the requirements of the HEP (High-Energy

Physics) and the Biomed user communities. Currently, AMGA is exploited to manage the metadata in the gBasf2 framework at the Belle II which is one of the largest particle physics experiments in the world.

In this paper, we present our efforts to optimize the metadata query performance of AMGA to better support the massive MC Campaign of the Belle II experiment. While AMGA has been proved to show very competitive performance for a relatively small amount of data, as the number of directories created and the size of overall metadata increase (e.g. hundreds of thousands of directories) during the MC Campaign, it suffers from severe query processing performance degradations. To address this problem, we modified the query search mechanism and the database scheme of AMGA which results in dramatic improvements of metadata search performance and query response time.

Throughout our comparative performance analysis of metadata search operations, we show that AMGA can be an optimal solution for a metadata catalog in a large-scale scientific experimental framework.

314

Monitoring system for the Belle II distributed computing

Author: Kiyoshi Hayasaka[1]

Co-authors: Hideki Miyake [2]; I Ueda [3]; Takanori HARA [2]; Yuji Kato

[1] *Nagoya Univ.*

[2] *KEK*

[3] *University of Tokyo (JP)*

Corresponding Author: hayasaka@hepl.phys.nagoya-u.ac.jp

Belle II Experiment is the next generation of B factory at SuperKEKB in Japan. A sample of 50 at^{-1} will be collected at the Υ resonances. In addition a large Monte Carlo (MC) sample will be generated to optimize the event selection criteria. The large data samples are managed by a sophisticated distributed computing system. To utilize the computing resources with a high efficiency, a monitoring system is absolutely needed. The effective monitoring system for the central system and each site has been developed based on some experiences for the MC sample production that was performed at the world wide computing centers. Also, in order to detect troubles on each site, a job to collect the environment has been developed and is submitted periodically. The monitoring system for the Belle II distributed computing will be introduced.

315

Investigation of High-Level Synthesis tools' applicability to data acquisition systems design based on the CMS ECAL Data Concentrator Card example

Author: Michal Husejko[1]

Co-authors: John Evans [1]; Jose Carlos Rasteiro Da Silva [2]

[1] *CERN*

[2] *LIP Laboratorio de Instrumentacao e Fisica Experimental de Part*

Corresponding Author: michal.husejko@cern.ch

High-Level Synthesis (HLS) for Field-Programmable Logic Array (FPGA) programming is becoming a practical alternative to well-established VHDL and Verilog languages. This paper describes

a case study in the use of HLS tools to design an FPGA-based data acquisition systems (DAQ). We will present the implementation of the CERN CMS detector ECAL Data Concentrator Card (DCC) functionality in HLS and lessons learned from using HLS design flow.

The DCC functionality and a definition of the initial system-level performance requirements (latency, bandwidth, and throughput) will be presented . We will describe how its packet processing control-centric algorithm was implemented with VHDL and Verilog languages. We will then show how the HLS flow could speed up design-space exploration by providing loose coupling between function's interface design and function's algorithm implementation.

We conclude with results of real-life hardware tests performed with the HLS flow-generated design with a DCC Tester system.

316

HPC in a HEP lab: lessons learned from setting up cost-effective HPC clusters

Author: Michal Husejko[1]

Co-authors: Helge Meinhard [1]; Ioannis Agtzidis [2]; John Evans [1]; Nils Hoimyr [1]; Pierre Baehler [1]; Tadeusz Jakub Dul [3]

[1] *CERN*

[2] *Aristotle Univ. of Thessaloniki (GR)*

[3] *University of Wroclaw (PL)*

Corresponding Author: michal.husejko@cern.ch

In this paper we present our findings gathered during the evaluation and testing of Windows Server High Performance Computing (Windows HPC) in view of potentially using it as a production HPC system for engineering applications. The Windows HPC package, an extension of Microsoft's Windows Server product, provides all essential interfaces, utilities and management functionality for creating, operating and monitoring a Windows based HPC cluster infrastructure. The evaluation and test phase was focused on verifying the functionalities of Windows HPC, its performance, support of commercial tools and the integration with the users' work environment.

We will describe constraints imposed by the way the CERN Computer Centre is operated, licensing for engineering tools and scalability and behaviour of the HPC Engineering applications used at CERN. We will present an initial set of requirements, which were created based on above constraints and requests from the CERN engineering user community.
We will explain how we have configured Windows HPC clusters to provide job scheduling functionalities required to support the CERN engineering user community, quality of service, user and project based priorities, and fair access to limited resources. Finally, we will present several performance tests we carried out to verify Windows HPC performance and scalability.

317

Development of tracker alignment software for the J-PARC E16 experiment

Author: Wataru Nakai[1]

Co-authors: Daisuke Kawama [2]; Hikari Murakami [3]; Kazuya Aoki [4]; Koki Kanno [1]; Kyoichiro Ozawa [5]; Satoshi Yokkaichi [2]; Tomonori Takahashi [6]; Yuhei Morino [5]; Yuki Obara [3]; Yusuke Komatsu [3]

[1] *University of Tokyo / RIKEN*

[2] *RIKEN*

[3] *University of Tokyo*

[4] *High Energy Accelerator Research Organization (KEK)*

[5] *KEK*

[6] *Research Center for Nuclear Physics (RCNP), Osaka University*

Corresponding Author: wnakai@post.kek.jp

The J-PARC E16 experiment will be performed to measure the mass modification of vector mesons in nuclear matter at J-PARC in order to study the origin of hadron mass.
In the experiment, we will measure invariant mass spectra of vector mesons with the electron and positron decay channel.
We will use 30 GeV proton beam with an intensity of 1×10^{10} protons per pulse at High-momentum beam line, which is to be constructed at J-PARC Hadron Facility in early 2016.

Indico rendering error

```
Could not include image: [404] Error fetching image
```

The design of the E16 spectrometer is shown as Figure 1.
It consists of several modules, and one module consists of three Gas Electron Multiplier Trackers (GTRs), a Hadron Blind Detector, which is a gas Cherenkov counter using CsI evaporated GEM, and Lead-Glass calorimeters. The sizes of GTRs are 100 mm $\times$ 100 mm, 200 mm $\times$ 200 mm, and 300 mm $\times$ 300 mm, respectively. A full spectrometer consists of 26 modules. Nuclear targets such as CH_2, C, Cu, and Pb, are placed at the center of the spectrometer. The magnitude of magnetic field is about 1.8 T at the center.

The E16 experiment aims for the mass resolution of 5 MeV/c^2 around ϕ meson peak. In order to achieve this, the position resolution of 100 μm is required.
So we should align GTRs with the accuracy of better than 100 μm.

Detector R&D has been well performed by the J-PARC E16 collaboration.
Our GTR consists of a drift cathode, a triple GEM, and a readout strip board.
We chose a strip pitch of 350 μm to achieve the required position resolution.
By several beam tests, the requirement is shown to be satisfied.

The photogrammetry system using a CCD camera will be conducted when detectors are installed, while the aimed precision will be finally achieved by using charged particle track-based software alignment.
We are developing its algorithm.
We will acquire calibration run data in which no magnetic field will be present and two additional wire targets will be used.
In the algorithm, we move tracking detectors to arrange wire target position and three hit positions of GTRs on a straight line.
We evaluate this algorithm using a Monte Carlo simulation.
In the simulation, tracks are generated from wire targets after misaligning detectors in random directions from designed places.
For example, distances between simulation true and corrected positions of detector centers are estimated.

We will report the R&D progress of the alignment software for the J-PARC E16 experiment.

318

An integrated solution for remote data access

Author: Vladimir Sapunenko[1]

Co-authors: Domenico D'Urso [2]; Luca dell'Agnello [3]; Vincenzo Vagnoni [2]

[1] *INFN-CNAF (IT)*

[2] *Universita e INFN (IT)*

[3] *INFN-CNAF*

Corresponding Author: vladimir.sapunenko@cern.ch

Data management constitutes one of the major challenges that a geographically-distributed data centre has to face, especially when remote data access is involved. We discuss an integrated solution which enables transparent and efficient access to online and nearline data through high latency networks. This is based on the joint use of the General Parallel File System (GPFS) and of the Tivoli Storage Manager (TSM). Both products, developed by IBM, are well known and extensively used in the HEP computing world. Owing to a new feature introduced in GPFS 3.5, so-called Active File Management (AFM), the definition of a single, geographically-distributed namespace, characterised by automated data flow management between different locations, becomes possible. As a practical example, we present the implementation of AFM-based remote data access between two data centres located in Bologna and Rome, demonstrating the validity of the solution for the use case of the AMS experiment.

Track 2 Session / 319

Continuous Readout Simulation with FairRoot on the Example of the PANDA Experiment

Author: Tobias Stockmanns[1]

[1] *FZ Jülich GmbH*

Corresponding Author: t.stockmanns@fz-juelich.de

Future particle physics experiments are searching more and more for rare decays which have similar signatures in the detector as the huge background. For those events usually simple selection criteria do not exist, which makes it impossible to implement a hardware-trigger based on a small subset of detector data.
Therefore all the detector data is read out continuously and processed on-the-fly to achieve a data reduction suitable for permanent storage and detailed analysis.

To cope with these requirements of a triggerless readout also the simulation software has to be adopted to add a continuous data production with pile-up effects and event overlapping in addition to the event wise simulation. This simulated data is of utmost importance to get a realistic detector simulation, to develop event-building algorithms and to determine the hardware requirements for the DAQ system of the experiments.
The possibility to simulate a continuous data stream was integrated into the FairRoot simulation framework. This running mode is called time-based simulation and a lot of effort was taken that one can switch seamlessly between event-based and a time-based simulation mode.

One experiment, which is using this new feature, is the PANDA experiment. It utilizes a quasi-continuous antiproton beam with a mean time between interactions of 50 ns. Because of the unbunched structure of the beam the interaction time follows a Poisson statistics with a high probability of events with short time distances. Depending on the time resolution of the sub-detectors this leads to an overlap of up to 20 events inside a sub-detector. This makes it an ideal test candidate for the time-based simulation.

In this talk the way the time-based simulation was implemented into FairRoot will be presented and examples of time-based simulations done for the PANDA experiment will be shown.

320

The Simulation Library of the Belle II Software System

Author: Leo Piilonen[1]

[1] *Virginia Tech*

Corresponding Author: piilonen@vt.edu

SuperKEKB and Belle II, the next generation B factory and its detector counterpart, are being constructed in Japan, as an upgrade of KEKB and Belle, respectively. The commissioning of the new SuperKEKB collider will be started in 2015. The luminosity of this e+ e− collider will be increased by a factor of 40, which will create a data sample 50 times larger than the previous Belle sample. Both the triggered and the background event rates will be increased by at least a factor of 10 compared to the previous rates, creating a very challenging data taking environment for the Belle II detector. The software system of the Belle II experiment is designed to execute this ambitious plan. A full detector simulation library, which is a part of the Belle II software system, is created based on Geant4. The construction of the library is progressing steadily and it is utilized actively in producing Monte Carlo data sets for pre-commission studies. In this talk, we will explain the detailed structure of the simulation library and the various interfaces to other packages including generators, geometry, and background simulation.

Track 8 Session / 321

Fast event generation on graphics processing unit (GPU) and its integration into the MadGraph system.

Author: Junichi Kanzaki[1]

[1] *KEK*

Corresponding Author: junichi.kanzaki@gmail.com

Fast event generation system of physics processes is developed using graphics processing unit (GPU). The system is based on the Monte Carlo integration and event generation programs, BASES/SPRING, which were originally developed in FORTRAN.
They were rewritten on the CUDA platform provided by NVIDIA in order for the implementation of these programs to GPUs.
Since the Monte Carlo integration algorithm is composed of a lot of independent function calls at multi-dimensional space, highly parallel architecture of GPU is very suitable for the improvement of their performance.
The performance of event generations based on the integrated results can be also easily improved by the event parallelization on GPU.
Parallelized programs show very good performance in process time compared to the existing event generation programs.

For the computation of cross sections of physics processes on GPU the helicity amplitude calculation package in FORTRAN is implanted in the CUDA framework as "HEGET" library, and new phase space generation library and random number generator are developed in order for the better generation efficiency on GPU.
The event generation system on GPU is tested using general Standard Model processes and computed cross sections are consistent with those obtained with the MadGraph system which is widely used in the field of elementary particle physics.
The total process time of the new system on GPU is compared with the equivalent programs on CPU and its improvement factors for various physics processes range from 10 to 100 depending on the complexity of their final states.

In order to achieve better interface for the generation of various physics processes and also to realize wider application of the fast Monte Carlo integration program the system has been integrated into the MadGraph5.

Also the HEGET library for the helicity amplitude computations is integrated into the ALOHA system of the MadGraph5.

322

The DIRAC Web Portal 2.0

Author: Zoltan Mathe[1]

Co-authors: Adrian Casajus Ramo [2]; Federico Stagni [1]; Nikola Lazovski [3]

[1] *CERN*

[2] *University of Barcelona (ES)*

[3] *University of Barcelona*

Corresponding Author: federico.stagni@cern.ch

For many years the DIRAC interware (Distributed Infrastructure with Remote Agent Control) has had a web interface, allowing the users to monitor DIRAC activities and also interact with the system. Since then many new web technologies have emerged, therefore a redesign and a new implementation of the DIRAC Web portal were necessary, taking into account the lessons learnt using the old portal.
These new technologies allowed to build a more compact and more responsive web interface that is robust and that enables users to have more control over the whole system while keeping a simple interface. The framework provides a large set of "applications", each of which can be used for interacting with various parts of the system. Communities can also create their own set of personalised web applications, and can easily extend already existing web applications with a minimal effort. Each user can configure and personalise the view for each application and save it using the DIRAC User Profile service as RESTful state provider, instead of using cookies.
The owner of a view can share it with other users or within a user community. Compatibility between different browsers is assured, as well as with mobile versions.
In this paper, we present the new DIRAC Web framework as well as the LHCb extension of the DIRAC Web portal.

Track 3 Session / 323

Disk storage at CERN

Author: Luca Mascetti[1]

Co-authors: Alessandro Fiorot [1]; Andrea Ieri [1]; Belinda Chan Kwok Cheong [1]; Dan van der Ster [1]; Giuseppe Lo Presti [1]; Herve Rousseau [1]; Hugo Gonzalez Labrador [2]; Jakub Moscicki [1]; Jan Iven [1]; Massimo Lamanna [1]; Sebastien Ponce [1]; Xavier Espinal Curull [1]

[1] *CERN*

[2] *University of Vigo (ES)*

Corresponding Author: luca.mascetti@cern.ch

CERN IT DSS operates the main storage resources for data taking and physics analysis mainly via three system: AFS, CASTOR and EOS. The total usable space available for users is about 100 PB (with relative ratios 1:20:120). EOS deploys disk resources across the two CERN computer centres (Meyrin and Wigner) with a current ratio 60% to 40%. IT DSS is also providing sizable on-demand resources for general IT services most notably OpenStack and NFS clients. This is provided by our Ceph infrastructure and a few of proprietary servers (NetApp) for a total capacity of ˜1 PB.

We will describe our operational experience and recent changes to these systems with special emphasis to the following items:

- Present usages for LHC data taking (new roles of CASTOR and EOS)
- Convergence to commodity hardware (nodes with 200-TB each with optional SSD) shared across all services
- Detailed study of the failure modes in the different services and approaches (RAID, RAIN, ZFS vs XFS, etc…)
- Disaster recovery strategies (across the two CERN computer centres)
- Experience in coupling commodity and home-grown solution (e.g. Ceph disk pools for AFS, CASTOR and NFS)
- Future evolution of these systems in the WLCG realm and beyond

Track 3 Session / 324

Federating LHCb datasets using the Dirac File Catalog

Author: Christophe Haen[1]

Co-authors: Andrei Tsaregorodtsev [2]; Markus Frank [1]; Philippe Charpentier [1]

[1] *CERN*

[2] *CPPM, Aix-Marseille Université, CNRS/IN2P3, Marseille, France*

Corresponding Author: christophe.denis.haen@cern.ch

In the distributed computing model of LHCb the File Catalog (FC) is a central component that keeps track of each file and replica stored on the Grid. It is federating the LHCb data files in a logical namespace used by all LHCb applications. As a replica catalog, it is used for brokering jobs to sites where their input data is meant to be present, but also by jobs for finding alternative replicas if necessary.
The LCG File Catalog (LFC) used originally by LHCb and other experiments is now being retired and needs to be replaced. The DIRAC File Catalog (DFC) was developed within the framework of the DIRAC Project and presented during CHEP 2012. From the technical point of view, the code powering the DFC follows an Aspect oriented programming (AOP): each type of entity that is manipulated by the DFC (Users, Files, Replicas, etc) is treated as a separate 'concern' in the AOP terminology. Hence, the database schema can also be adapted to the needs of a Virtual Organization. LHCb opted for a highly tuned MySQL database, with optimized requests and stored procedures.
This paper will present the improvements brought to the DFC presented at CHEP 2012, its performance with respect to the LFC, as well as the migration procedure used to migrate the LHCb data from the LFC to the DFC. Finally it will show how a combination of the DFC and the LHCb framework Gaudi allow LHCb to build a data federation at low cost.

325

Data Management System of the DIRAC Project

Authors: Andrei Tsaregorodtsev[1]; Christophe Haen[2]; Philippe Charpentier[2]

[1] *CPPM, Aix-Marseille Université, CNRS/IN2P3, Marseille, France*

[2] *CERN*

Corresponding Author: christophe.denis.haen@cern.ch

The DIRAC Interware provides a development framework and a complete set of components for building distributed computing systems. The DIRAC Data Management System (DMS) offers all the necessary tools to ensure data handling operations for small and large user communities. It supports

transparent access to storage resources based on multiple technologies, and is easily expandable. The information on data files and
replicas is kept in a File Catalog of which DIRAC offers a powerful and versatile implementation (DFC). Data movement can be performed using third party services including FTS3. Bulk data operations are resilient with respect to failures due to the use of the Request Management System (RMS) that keeps track of ongoing tasks.
In this contribution we will present an overview of the DIRAC DMS capabilities and its connection with other DIRAC subsystems such as the Transformation System.
The DIRAC DMS is in use by several user communities now. The contribution will present the experience of the LHCb experiment. The experience of other experiements as well as multi-VO DIRAC services with the use of the DIRAC DMS will be described.

326

Monitoring tools of COMPASS experiment at CERN

Author: Jan Tomsa[1]

Co-authors: Josef Novy [1]; Martin Bodlak [2]; Miroslav Virius [1]; Vladimir Jary [1]

[1] *Czech Technical University in Prague (CZ)*

[2] *CERN (CH)*

Corresponding Authors: jan.tomsa@cern.ch, josef.novy@cern.ch

Nowadays, all modern high energy physics experiments are substantially dependent on fast and reliable data acquisition systems that are able to collect large quantities of data supplied by various detectors. To ensure smooth and errorless operation, it is necessary to control and monitor the behavior and state of processes running in the system.

COMPASS is a high energy particle experiment with fixed target located at SPS at CERN laboratory in Geneva, Switzerland. This poster briefly introduces the data acquisition system of COMPASS and is mainly focused on the part that is responsible for the monitoring of the nodes in the whole newly developed data acquisition system of the experiment.

The technical shutdown of CERN during years 2013 and 2014 has been used to upgrade the data acquisition system of the COMPASS experiment. Both new harware, which uses PFGA cards for event building, and new software has been developed and deployed. The monitoring of the system is managed by tools called Message Logger and Message Browser.

The Message Logger is a simple console application that collects informative and error messages. The messages are received via standard switched network and the communication is handled by the DIM library. The DIM library is a communication tool originally designed and developed for the DELPHI experiment at CERN. It provides C++, JAVA, and Python interfaces. The relevant collected messages are then saved into the central COMPASS DAQ MySQL database.

The Message Browser is a tool with a graphical user interface created in Qt framework based upon a MVC (model - view - controller) design pattern. It is used to fetch and display messages previously stored by the Message Logger. It also features so called “online mode” which serves for acquisition of new messages via network at the time of their origin (this works similarly to the Message Logger). This functionality saves time and system resources - the new messages are displayed immediately and the program does not need to poll the database for new messages. Even though the Message Browser is connected to other nodes via the DIM service, it is designed to run independently from the other nodes (it requires only the database to be connected). Therefore it allows quicker response from the operators of the experiment in case of unexpected behavior or crash of the data acquisition system of the experiment. The Message Browser is also equipped with ordering and rich and intuitive filtering options (by all parameters of the messages).

The central data acquisition database stores system configuration and the messages. The Message_log table is expected to be the largest table in the database. Because the Message Browser requires high amount of data from the database, the Message_log table uses the MyISAM storage engine that is

optimized for environments with heavy read operations. For further increase in speed, several indices are created on some of the columns, most importantly on the Stamp column (ordering by the time of the creation of the message is most important).

The poster also contains performance tests of the created monitoring tools.

Track 4 Session / 327

Cernbox + EOS: End-user Storage for Science

Authors: Andreas Joachim Peters[1]; Jakub Moscicki[1]

Co-authors: Elvin Alin Sindrilaru [1]; Hugo Gonzalez Labrador [2]; Luca Mascetti [1]; Massimo Lamanna [1]

[1] *CERN*

[2] *University of Vigo (ES)*

Corresponding Author: luca.mascetti@cern.ch

Cernbox is a cloud synchronisation service for end-users: it allows to sync and share files on all major mobile and desktop platforms (Linux, Windows, MacOSX, Android, iOS) aiming to provide offline availability to any data stored in the CERN EOS infrastructure.

The successful beta phase of the service confirmed the high demand in the community for such easily accessible cloud storage solution. Integration of the Cernbox service with the EOS storage back-end is the next step towards providing sync and share capabilities for scientific and engineering use-cases.

In this report we will present lessons learnt offering the Cernbox service, key technical aspects of Cernbox/EOS integration and new, emerging usage possibilities. The latter include the ongoing integration of sync and share capabilities with the LHC data analysis tools and transfer services.

Track 3 Session / 328

Evaluation of NoSQL databases for DIRAC monitoring and beyond

Author: Zoltan Mathe[1]

Co-authors: Adrian Casajus Ramo [2]; Federico Stagni [1]; Luca Tomassetti [3]

[1] *CERN*

[2] *University of Barcelona (ES)*

[3] *University of Ferrara and INFN*

Corresponding Author: federico.stagni@cern.ch

Nowadays, many database systems are available but they may not be optimized for storing time series data. The DIRAC job monitoring is a typical use case of such time series. So far it was done using a MySQL database, which is not well suited for such an application. Therefore alternatives have been investigated.
Choosing an appropriate database for storing huge amounts of time series is not trivial as one must take into account different aspects such manageability, scalability, extensibility etc.
We compared the performance of Elasticsearch, OpenTSDB that is based on HBase and InfluxDB time series NoSQL databases using the same set of machines and the same data. We also evaluated the effort required for maintaining them.
Using the LHCb Workload Management System, based on DIRAC, as a use case we have setup a new

monitoring system in parallel with the current MySQL system and we publish the same data into the databases under test. We have evaluated Grafana (for OpenTSDB) and Kibana (for ElasticSearch) metrics and graph editors for creating dashboards in order to have clear picture on the usability of each candidate. In this paper we present the result of this study and the performance of the selected technology. We also give an outlook of other potential applications of NoSQL databases with DIRAC project.

Track 4 Session / 329

Belle II production system

Author: Hideki Miyake[1]

Co-authors: Malachi Schram [2]; Radek Ludacka [3]; Rafal Grzymkowski [4]

[1] *KEK*

[2] *Pacific Northwest National Laboratory*

[3] *Charles Univ. in Prague*

[4] *Institute of Nuclear Physics PAN*

Corresponding Author: hideki.miyake@kek.jp

In Belle II experiment a large amount of physics data will be continuously taken and the production rate is equivalent to LHC experiments.
Considerable resources of computing, storage, and network, are necessary to handle not only the taken data but also substantial simulated data.
Therefore Belle II exploits distributed computing system based on DIRAC interware.
DIRAC is a general software framework to provide unified interface among heterogeneous computing resources.
As well as proven DIRAC software stack, Belle II is developing its own extension called BelleDIRAC.
BelleDIRAC gives a transparent user experience of Belle II analysis framework (basf2) on various environments and access to file information managed by LFC and AMGA metadata catalog.
With unifying DIRAC and BelleDIRAC functionalities, Belle II plans to operate automated mass data processing framework named a production system.
Belle II production system covers sizable raw data transfer from experimental site to raw data centers, followed by massive data processing, and smart output deployment to each remote site.
The production system is also utilized for simulated data production and data analysis.
Although development of the production system is still on-going, recently Belle II has prepared prototype version and evaluated it with large scale of simulated data production test.
In this presentation we will report the evaluation of the prototype system and future development plan.

330

Migration experiences of the LHCb Online cluster to Puppet and Icinga2

Authors: Adam Jedrzej Otto[1]; Christophe Haen[2]; Daniel Hugo Campora Perez[2]; Francesco Sborzacchi[3]; Hristo Mohamed[4]; Loic Brarda[2]; Mohamed Chebbi[2]; Niko Neufeld[2]

[1] *Ministere des affaires etrangeres et europeennes (FR)*

[2] *CERN*

[3] *Istituto Nazionale Fisica Nucleare (IT)*

[4] *University of Sofia (BG)*

Corresponding Author: loic.brarda@cern.ch

The LHCb experiment operates a large computing infrastructure with more than 2000 servers, 300 virtual machines and 400 embedded systems. Many of the systems are operated diskless from NFS or iSCSI root-volumes. They are connected by more than 200 switches and routers. We have recently completed the migration of the management of this system from Quattor to puppet and of the original monitoring structure based on Icinga to icinga2. We also have a tight integration of icinga2 in puppet. In this paper we present challenges and experiences during this migration. We discuss and contrast the expected and observed operational improvements.

331

Tile-in-ONE: A web platform which integrates Tile Calorimeter data quality and calibration assessment

Authors: Alexander Solodkov[1]; Andressa Gomes[2]; Carlos Solans Sanchez[3]; Carmen Maidantchik[2]; Fernando Guimaraes Ferreira[2]

Co-authors: Blake Oliver Burghgrave [4]; Iouri Smirnov [4]

[1] *Institute for High Energy Physics (RU)*

[2] *Univ. Federal do Rio de Janeiro (BR)*

[3] *CERN*

[4] *Northern Illinois University (US)*

Corresponding Authors: andressa@cern.ch, carlos.solans@cern.ch

The ATLAS Tile Calorimeter assesses the quality of data in order to ensure its proper operation. A number of tasks are then performed by running several tools and systems,
which were independently developed to meet distinct collaboration's requirements and do not necessarily builds an effective connection among them. Thus, a program is usually implemented without a global perspective of the detector, requiring basic software features. In addition, functionalities may overlap in their objectives and frequently replicate resources retrieval mechanisms.

Tile-in-ONE is a unique platform that assembles various web systems used by the calorimeter community through a single framework and a standard technology. It provides an infrastructure to support the code implementation, avoiding duplication of work while integrating with an overall view of the detector status. Database connectors smooth the process of information access since developers do not need to be aware of where records are placed and how to extract them. Within the environment, a dashboard stands for a particular Tile operation aspect and gets together plug-ins, i.e. software components that add specific features to an existing application.

A server contains the platform core, which represents the basic environment to deal the configuration, manage user settings and load plug-ins at runtime. A web middleware assists users to develop their own plug-ins, perform tests and integrate them into the platform as a whole. Backends are employed to allow that any type of application is interpreted and displayed in a uniform way.

Track 4 Session / 333

Multicore job scheduling in the Worldwide LHC Computing Grid

Authors: Alessandra Forti[1]; Andrej Filipcic[2]; Andrew David Lahiff[3]; Antonio Perez-Calero Yzquierdo[4]; Carlos Acosta Silva[5]; Christopher John Walker[6]; Daniel Peter Traynor[6]; Jeff Templon[7]; Manfred Alef[8]; Miguel Gila[9]; Rodney Walker[10]; Samuel Cadellin Skipsey[None]; Sebastien Gadrat[11]; Stefano Dal Pra[12]; Thomas Hartmann[13]

[1] *University of Manchester (GB)*

[2] *Jozef Stefan Institute (SI)*

[3] *STFC - Rutherford Appleton Lab. (GB)*

[4] *Centro de Investigaciones Energ. Medioambientales y Tecn. - (ES*

[5] *Universitat Autònoma de Barcelona (ES)*

[6] *University of London (GB)*

[7] *NIKHEF (NL)*

[8] *Karlsruhe Institute of Technology (KIT)*

[9] *ETH Zurich*

[10] *Ludwig-Maximilians-Univ. Muenchen (DE)*

[11] *CC-IN2P3 - Centre de Calcul (FR)*

[12] *INFN*

[13] *KIT - Karlsruhe Institute of Technology (DE)*

Corresponding Author: alessandra.forti@cern.ch

After the successful first run of the LHC, data taking will restart in early 2015 with unprecedented experimental conditions leading to increased data volumes and event complexity. In order to process the data generated in such scenario and exploit the multicore architectures of current CPUs, the LHC experiments have developed parallelized software for data reconstruction and simulation. A good fraction of their computing effort is still expected to be executed as single-core tasks. Therefore, jobs with diverse resources requirements will be distributed across the Worldwide LHC Computing Grid (WLCG), making workload scheduling a complex problem in itself.

In response to this challenge, the WLCG Multicore Deployment Task Force has been created with the purpose of coordinating the joint effort from experiments and WLCG sites. The main objective is to ensure the convergence of approaches from the different LHC Virtual Organizations (VOs) to make the best use of the shared resources in order to satisfy their new computing needs and minimize any inefficiency deriving from the scheduling mechanisms. This should also be achieved without imposing unnecessary complexities in the way sites manage their resources.

Job scheduling in the WLCG involves the use of grid-wide workload submission tools by the VOs linked via Computing Element (CE) middleware to the batch system technologies managing local resources at every WLCG site. Each of these elements and their interaction has been analyzed by the Task Force. The various job submission strategies proposed by the LHC VOs have been evaluated, providing feedback for the evolution of their grid-wide submission models and tools. The diverse capabilities of different CE technologies in passing the resource request from the VOs to the sites have been examined. The technical features of the most common batch systems in WLCG sites have been discussed for a better understanding of their multicore job handling capabilities. Participants in the Task Force have also been encouraged to share their system configurations with the purpose of avoiding duplicated efforts among sites operating the same technologies.

This contribution will present the activities and progress of the Task Force related to the aforementioned topics, including experiences from key sites on how to best use different batch system technologies, the evolution of workload submission tools by the experiments and the knowledge gained from scale tests of the different proposed job submission strategies.

Track 4 Session / 334

The GridPP DIRAC project - DIRAC for non-LHC communities

Authors: Adam Huffman[1]; Alexander John Richards[1]; Daniela Bauer[1]; David Colling[1]; Duncan Rand[1]; Janusz Martyniak[2]; Robert Andrew Currie[1]; Simon Fayer[1]

[1] *Imperial College Sci., Tech. & Med. (GB)*

[2] *Imperial College London*

Corresponding Author: janusz.martyniak@imperial.ac.uk

The GridPP consortium in the UK is currently testing a multi-VO DIRAC service aimed at non-LHC VOs. These VOs are typically small (fewer than two hundred members) and generally do not have a dedicated computing support post. The majority of these represent particle physics experiments (e.g. T2K, NA62 and COMET), although the scope of the DIRAC service is not limited to this field. A few VOs have designed bespoke tools around the EMI-WMS & LFC, while others have so far eschewed distributed resources as they perceive the overhead for accessing them to be too high.
The aim of the GridPP DIRAC project is to provide an easily adaptable toolkit for such VOs in order to lower the threshold for access to distributed resources such as grid and cloud computing. As well as hosting a centrally run DIRAC service, we will also publish our changes and additions to the base DIRAC codebase under an open-source license.
We started by taking a survey of the existing VO specific solutions using the feedback to determine the user requirements that were driving these implementations. These details were then used to map the base requirements to available DIRAC features, implementing any additional common functionality that was needed. Once the groundwork was complete, this knowledge was shared with the existing VOs and we worked with them to adapt their grid model to use the new DIRAC service.
The experience gained from this process was then used to recommend sensible approaches to the new VOs and assist them in getting started with distributed computing. We investigated different support models and found that a mailing list was the most accessible option for the target audience, while GGUS is used for tracking service issues.

We report on the current status of this project and show increasing adoption of DIRAC within the non-LHC communities.

Track 5 Session / 335

Exploiting CMS data popularity to model the evolution of data management for Run-2 and beyond

Author: Daniele Bonacorsi[1]

Co-authors: Domenico Giordano [2]; Maria Girone [2]; Neri Matteo [1]; Nicolo Magini [2]; Tommaso Boccali [3]; Tony Wildish [4]; Valentin Y Kuznetsov [5]

[1] *University of Bologna*

[2] *CERN*

[3] *Sezione di Pisa (IT)*

[4] *Princeton University (US)*

[5] *Cornell University (US)*

Corresponding Author: daniele.bonacorsi@bo.infn.it

During the LHC Run-1 data taking, all experiments collected large data volumes from proton-proton and heavy-ion collisions. The collisions data, together with massive volumes of simulated data, were replicated in multiple copies, transferred among various Tier levels, transformed/slimmed in format/content. These data were then accessed (both locally and remotely) by large groups of distributed analysis communities exploiting the WorldWide LHC Computing Grid infrastructure and services. While efficient data placement strategies - together with optimal data redistribution and deletions on demand - have become the core of static versus dynamic data management projects, little effort has so far been invested in understanding the detailed data-access patterns which surfaced

in Run-1. These patterns, if understood, can be used as input to simulation of computing models at the LHC, to optimise existing systems by tuning their behaviour, and to explore next-generation CPU/storage/network co-scheduling solutions. This is of great importance, given that the scale of the computing problem will increase far faster than the resources available to the experiments, for Run-2 and beyond.

Studying data-access patterns involves the validation of the quality of the monitoring data collected on the "popularity" of each dataset, the analysis of the frequency and pattern of accesses to different datasets by analysis end-users, the exploration of different views of the popularity data (by physics activity, by region, by data type), the study of the evolution of Run-1 data exploitation over time, the evaluation of the impact of different data placement and distribution choices on the available network and storage resources and their impact on the computing operations.

This work presents some insights from studies on the popularity data from the CMS experiment. We present the properties of a range of physics analysis activities as seen by the data popularity, and make recommendations for how to tune the initial distribution of data in anticipation of how it will be used in Run-2 and beyond.

337

Job monitoring on DIRAC for Belle II distributed computing

Author: Yuji Kato[None]

Co-authors: Hideki Miyake [1]; I Ueda [2]; Kiyoshi Hayasaka [3]; Takanori Hara [4]

[1] *KEK*

[2] *University of Tokyo (JP)*

[3] *Nagoya Univ.*

[4] *Osaka University*

Corresponding Author: kato@hepl.phys.nagoya-u.ac.jp

The BelleII is an asymmetric energy e^+e^- collider experiment at SuperKEKB in Japan. One of the main goals of BelleII is to search for physics beyond the Standard Model with a data set of about 5×10^{10} $B\bar{B}$ pairs. In order to store such huge amount of data including MC events and analyze it in a timely manner, BelleII adopted distributed computing model with DIRAC (Distributed Infrastructure with Remote Agent Control), which is a framework developed for the LHCb experiment to manage heterogeneous computing environment.
To effectively maximize resources, we must detect and repair errors quickly. Monitoring job processing is crucial. Currently, DIRAC provides information such as status of jobs and log files. However, this is not sufficient to detect problems. By collecting statistical information such as number of successes/failed jobs and analyzing log files, we can understand the status of each site. Furthermore, these processes must be automated to check all sites efficiently. We implement these functions in DIRAC.
In this paper, we describe the design and experience with the monitor in the MC data production campaign.

338

An Research and Development for Evolving Architecture for the Beyond Standard Model

Author: Kihyeon Cho[1]

Co-authors: Huiyoung Ryu [1]; Junghyun Kim [1]

[1] *KISTI*

Corresponding Authors: hyryu@kisti.re.kr, hyun@kisti.re.kr, cho@kisti.re.kr

For the solution of searching for new physics beyond the standard model, we do a research and development of simulation tool kit based on the evolving computing architecture with international collaboration. Using the tools, we study particle physics beyond the standard model. We present the current status of the research and development for these.

339

Public Outreach at RAL: Engaging the Next Generation of Scientists and Engineers

Authors: George Ryall[1]; Ian Peter Collier[2]; James Adams[3]; Sophy Palmer[1]

Co-authors: Greg Corbett [1]; Rob Appleyard [1]

[1] *STFC*

[2] *STFC - Rutherford Appleton Lab. (GB)*

[3] *STFC RAL*

Corresponding Author: george.ryall@stfc.ac.uk

Rutherford Appleton Laboratory (RAL) is part of the UK's Science and Technology Facilities Council (STFC). The Royal Charter that established the STFC requires us to generate public awareness and encourage public engagement and dialogue in relation to the science we undertake. We firmly support this activity as it is important to encourage the next generation of students to consider studying STEM (Science, Technology, Engineering, and Mathematics) subjects, providing the UK with a highly skilled work-force in the future.

To this end, we undertake a variety of outreach activities. We will describe the outreach activities undertaken by RAL, particularly focussing on those of the Scientific Computing Department. These activities include: an Arduino based activity day for 12-14 year olds to celebrate Ada Lovelace day; running a centre as part of the Young Rewired State –encouraging 11-18 year olds to created web applications with open data; sponsoring a team in the Engineering Education Scheme –supporting a small team of 16/17 year olds to solve a real world engineering problem we set them; as well as the more traditional tours of our facilities. We believe these activities could serve as an example for other sites involved in scientific computing around the globe.

340

Testing WAN access to storage over IPv4 and IPv6 using multiple transfer protocols

Authors: Duncan Rand[1]; Steve Lloyd[2]

[1] *Imperial College Sci., Tech. & Med. (GB)*

[2] *University of London (GB)*

Corresponding Author: duncan.rand@cern.ch

In the lead up to Run 2 of the LHC the WLCG grid middleware, storage access protocols and LHC computing models are in a state of flux. The LCG utilities and SRM middleware are being phased out, IPv6 is being rolled out across the WLCG and LHC experiments are making increasing use of

xrootd federated access to storage elements over the WAN. However, both client and server software and WLCG sites vary in their readiness for IPv6.

Sites within GridPP will need to enable IPv6, support it on relevant software and ensure everything is working correctly. Comprehensive monitoring will help and we report on work to extend existing monitoring of WAN transfer capability that GridPP has been doing for a number of years. WAN access from storage elements to worker nodes over both IPv4 and IPv6 is systematically checked. The gsiftp protocol is checked with *lcg-cp* and *gfal-copy* clients, https with *curl* and xrootd with *xrdcp*. Future work will include job submission and data handling tests using the GridPP DIRAC service.

Results can be viewed at http://pprc.qmul.ac.uk/~lloyd/gridpp/nettest_v6.html

341

Local storage federation through XRootD architecture for interactive distributed analysis

Authors: Antonio Franco[1]; Domenico Colella[1]; Domenico Elia[2]; Giacinto Donvito[3]; Giorgia Miniello[1]; Giorgio Maggi[1]

[1] *Universita e INFN (IT)*

[2] *INFN Bari*

[3] *INFN-Bari*

Corresponding Authors: domenico.elia@cern.ch, giorgia.miniello@ba.infn.it

A cloud-based Virtual Analysis Facility (VAF) for the ALICE experiment at the LHC has been developed in Bari. Similar facilities are currently running in other Italian sites with the aim to create a federation of interoperating farms able to provide their computing resources for interactive distributed analysis. The facility consists in a PROOF cluster of virtual machines dynamically deployed by the Openstack cloud infrastructure built in Bari for the PRISMA project.
The use of cloud technology, along with elastic provisioning of computing resources as an alternative to the grid for running data intensive analyses, is the main challenge of the facility. One of the crucial aspects of the user-driven analysis execution is the data access. A local storage facility has the disadvantage that the stored data can be accessed by PROOF only locally, i.e. from within the single VAF. To overcome such a limitation a federated infrastructure, which provides full access to all the data belonging to the federation independently from the site where they are stored, has been set up. The federation architecture exploits both cloud computing (block storage, posix storage and object storage) and XRootD technologies, in order to provide a dynamic, easy-to-use and well performing solution for data handling. It allows the users to store the files and efficiently retrieve the data, since it implements a dynamic distributed cache among many datacenters in Italy connected to one another through the high-bandwidth national network.
In this contribution we will show the technical solution, based on CEPH, GlusterFS or Swift together with XRootD, which has been implemented to build this distributed storage infrastructure. This solution will be able either to store data for long time periods or just cache data locally, so that they can be processed by the PROOF facilities at any given site. Details on the architecture implementation and the achieved performance in the first prototype tests will be also discussed.

The present work is supported by the Istituto Nazionale di Fisica Nucleare (INFN) of Italy and is partially funded under contract 20108T4XTM of Programmi di Ricerca Scientifica di Rilevante Interesse Nazionale (STOA-LHC PRIN, Italy).

342

Belle II public and private clouds management in VMDIRAC system.

Author: Rafal Zbigniew Grzymkowski[1]

Co-author: Takanori HARA [2]

[1] *Polish Academy of Sciences (PL)*

[2] *KEK*

Corresponding Author: rafal.grzymkowski@ifj.edu.pl

The role of cloud computing technology in the distributed computing for HEP experiments grows rapidly. Some experiments (Atlas, BES-III, LHCb,···) already exploits private and public cloud resources for data processing. Future experiments such as Belle II or upgraded LHC experiments will largely rely on the availability of cloud resources and therefore their computing models have to be adjusted to the specific features of cloud environment, in particular to the on-demand computing paradigm.

Belle II experiment at SuperKEKB will start physics run in 2017. Belle II computing requirements are challenging. The data size at the level of hundred PB is expected after several years of operation, around 2020.
The baseline solution selected for distributed processing is the DIRAC system. DIRAC can handle variety of computing resources including Grids, Clouds and independent clusters. Cloud resources can be connected by VMDIRAC module through public interfaces. In particular the mechanism of dynamic activation of new virtual machines with reserved job slots for new tasks in case of an increasing demand for computing resources is introduced.

This work is focused on VMDIRAC interaction with public (Amazon EC2) and private (CC1) cloud. The solution applied by Belle II experiment and the experience from Monte Carlo production campaigns will be presented. Updated computation costs for different use cases will be shown.

343

Base ROOT reference guide on Doxygen

Author: Olivier Couet[1]

[1] *CERN*

Corresponding Author: olivier.couet@cern.ch

The ROOT reference guide is part of the code. Classes description, methods usage, examples etc… are all embedded with the code itself. Doxygen is the reference model allowing to extract the documentation from such self described system. The ROOT documentation requires the development
of specific tools (scripts) in the Doxygen context. The proposed project is these tools writing.

344

Experimental quantification of Geant4 PhysicsList recommendations: methods and results

Authors: Chan Hyeong Kim[1]; Gabriela Hoff[None]; Maria Grazia Pia[2]; Mincheol Han[3]; Paolo Saracco[4]; Sung Hun Kim[3]

[1] *Hanyang University*

[2] *Universita e INFN (IT)*

[3] *H*

[4] *INFN - Genova (Italy)*

Corresponding Authors: maria.grazia.pia@cern.ch, sung.hun.kim@cern.ch

Geant4 recommends a set of PhysicsLists and related classes (Builders, PhysicsConstructors) to its user community to facilitate the use of Geant4 functionality despite its intrinsic physics complexity. Relative limited documentation is available in the literature regarding Geant4 physics configuration tools, especially concerning the quantification of their accuracy, their computational performance and the stability of their results.

We present the results of a comprehensive experimental quantification project concerning recommendations related to multiple scattering, which are implemented in the "constructors" package of Geant4 PhysicsLists. We show how these findings affect the simulation of energy deposition, both regarding the absolute value of the energy deposited in a volume and its spatial pattern, as well as other observables specifically pertinent to the treatment of multiple scattering.

In-depth analysis was carried out to quantify not only the behaviour of the recommended PhysicsConstructor classes of Geant4 physics_list package, but also of individual settings embedded in them, and their contribution to the overall simulation outcome. We report an extensive assessment of the accuracy, of the computational performance and of alternative settings of the physics configuration options subject to evaluation: all these aspects are quantified with respect to a wide collection of experimental test cases with rigorous statistical methods. This analysis is documented over several Geant4 versions to quantify the stability of the results.

The results of this project provide objective guidelines for the improvement of Geant4 physics models, of their recommended use and of experimental applications of Geant4. We discuss how the methodology developed for this assessment could be extended to other Constructors, Builders and complete PhysicsLists released within the Geant4 Toolkit.

Track 4 Session / 345

CMS data distributed analysis with CRAB3

Author: Marco Mascheroni[1]

Co-authors: Andres Jorge Tanasijczuk [2]; Anna Elizabeth Woodard [3]; Brian Paul Bockelman [4]; Diego Ciangottini [5]; Eric Vaandering [6]; Hassen Riahi [7]; Jose Hernandez Calama [8]; Justas Balcas [9]; Matthias Wolf [3]; Mohd Nizam Bin Yusli [10]; Preslav Borislavov Konstantinov [11]; Stefano Belforte [5]

[1] *Universita & INFN, Milano-Bicocca (IT)*

[2] *Univ. of California San Diego (US)*

[3] *University of Notre Dame (US)*

[4] *University of Nebraska (US)*

[5] *Universita e INFN (IT)*

[6] *Fermi National Accelerator Lab. (US)*

[7] *CERN*

[8] *Centro de Investigaciones Energ. Medioambientales y Tecn. - (ES*

[9] *Vilnius University (LT)*

[10] *University of Malaya (MY)*

[11] *Bulgarian Academy of Sciences (BG)*

Corresponding Author: marco.mascheroni@cern.ch

The CMS Remote Analysis Builder (CRAB) provides the service for managing analysis tasks isolating users from the technical details of the distributed Grid infrastructure. Throughout the LHC Run 1,

CRAB has been successfully employed by an average 350 distinct users every week executing about 200,000 jobs per day.

In order to face the new challenges posed by the LHC Run 2, CRAB has been significantly re-factored. The pieces of the new system are a lightweight client, a central server exposing a REST interface accepting user requests, a number of servers dealing with user analysis tasks and submitting jobs to the CMS resource provisioning system, and a central service to asynchronously move user data from the execution node to the desired storage location. The new system improves the robustness, scalability and sustainability of the service.

This contribution will give an overview of the new system, reporting the status, experience and lessons learnt from the commissioning phase and the production rollout for the initial data taking. It will address all aspects of the project from development to operations and support.

Track 4 Session / 346

The GridPP DIRAC project - Implementation of a multi-VO DIRAC service

Authors: Adam Huffman[1]; Alexander John Richards[1]; Daniela Bauer[1]; David Colling[1]; Duncan Rand[1]; Janusz Martyniak[2]; Robert Andrew Currie[1]; Simon Fayer[1]

[1] *Imperial College Sci., Tech. & Med. (GB)*

[2] *Imperial College London*

Corresponding Author: rcurrie@cern.ch

The DIRAC INTERWARE system was originally developed within the LHCb VO as a common interface to access distributed resources, i.e. grids, clouds and local batch systems. It has been used successfully in this context by the LHCb VO for a number of years. In April 2013 the GridPP consortium in the UK decided to offer a DIRAC service to a number of small VOs. The majority of these had been previously accessing grid resources using the EMI-WMS system and some of them were also using the EMI-LFC for their data management. DIRAC provides integrated workload and data management systems, although these can also be used independently if required.
It was decided that the most maintainable way to deploy the DIRAC server was with one instance shared between all supported VOs rather than a separate one for each VO. While this deployment model was not without precedent, this approach greatly deviates from the model for which DIRAC was originally designed. We set-up a test instance and while using this it was found that a number of core features needed to be improved so that they could be managed within the scope of a VO rather than server-wide. The main features that needed this adaptation were the access controls, the automatic configuration of users & resources and the data management functions.
The DIRAC client toolset is fully-featured and contains analogues to most of the lcg client commands. For the VOs which already have wrappers around the lcg tools, just adapting these would be a possibility however this would still leave every VO with their own solution to maintain. New VOs don't want to spend valuable computing effort to develop custom wrappers for their users based on these tools; a better solution is to try and standardise on a set of common tools. This reduces both the maintenance effort for any single user as bugs get fixed for everyone at once while also minimising the number of different configurations the central experts need to be familiar with. The primary tool we have investigated for this role is Ganga, particularly with its DIRAC backend module.
We report on how we resolved these issues to provide a production grade multi-VO DIRAC service, including where appropriate the details of changes applied to the DIRAC and user interface code bases.

347

Geant4 simulation for a study of a possible use of carbon ion pencil beams for the treatment of ocular melanomas with the active scanning system at CNAO.

Authors: Farina Edoardo[1]; Piersimoni Pierluigi[2]; Riccardi Cristina[3]; Rimoldi Adele[3]; Tamborini Aurora[3]

Co-author: Ciocca Mario [4]

[1] *University of Pavia*

[2] *Loma Linda University Health*

[3] *University of Pavia - INFN Section of Pavia*

[4] *CNAO Foundation*

Corresponding Author: aurora.tamborini@unipv.it

Purpose

The aim of this work is a study of a possible use of carbon ion pencil beams (delivered with active scanning modality) for the treatment of ocular melanomas at the National Centre for Oncological Hadrontherapy (CNAO). The promising aspect of carbon ions radiotherapy for the treatment of this disease lies in its superior relative radiobiological effectiveness (RBE).
The Monte Carlo Geant4 toolkit is used to simulate the complete CNAO extraction beamline, with the active and passive components along it.
A human eye modeled detector, including a realistic target tumor volume, is used as target. Cross check with previous studies at CNAO using protons allows comparisons on possible benefits on using such a technique with respect to proton beams. Experimental data on proton and carbon ion beams transverse distributions are used to validate the simulation.

Description

The ocular melanoma is nowadays the most common early intraocular tumor in adulthood, a malignant tumor which tends to grow both inside the bulb, invading and disrupting the intraocular tissues and outside it, infiltrating the sclera and orbital tissues.
Several possible modalities are available to treat the disease (conventional radiotherapy, brachytherapy, protontherapy); thanks to the physical properties of hadrons penetrating matter, protons or heavy ions treatments can improve the visual prognosis, because the energy is delivered to the target tissue, with very little exposure of surrounding healthy tissues.
In the clinical practice of eye protontherapy, the currently worldwide accepted technique for dose delivery is represented by the passive scattering modality.
However, in the present configuration at CNAO the dose delivery system adopted and only available to treat patients, is represented by the full active scanning modality and a commercial general-purpose (not specifically dedicated to the eye's treatment), image-based treatment planning system (TPS) is used.

Methods

The CNAO synchrotron is able to produce carbon ion beams with a FWHM ranging from a minimum of
4 ± 0.04 mm to a maximum of 10 ± 0.1 mm at the standard target center (isocenter), inversely depending on the beam energy. For protons, the FWHM ranges from 7.0 ± 0.7 mm to 22.3 ± 0.2 mm.
The CNAO beam's energy ranges from 120 MeV/u to 400 MeV/u, for carbon ions and from
63 MeV to 250 MeV for protons, with an uncertainty of 0.05%.
The final part of the CNAO transport beamline, with all its elements, is simulated using the Geant4 toolkit. Projectile charged particles, accelerated in the CNAO synchrotron ring, travel in a long extraction vacuum beam pipe (about 6 m), crossing the magnetic field generated by two orthogonal deflecting magnets. The vacuum beam pipe is sealed by a carbon fiber exit window, after which the accelerated beam reaches the treatment room and travels for several centimeters in air. In the treatment room the beam crosses a fixed structure called nozzle, inside of which two beam-monitoring chambers are embedded to measure in real-time the beam fluence and position.
The standard target center (isocenter) is located at a distance of 64 cm from the downstream edge of the nozzle.
As a first step of the simulation, a detailed description of the human eye with its internal components is used to build an eye-detector; a realistic tumor is also included inside, near the optic nerve. The eye is placed in a water box simulating the human brain, protruding from it as well as in a real human head. Each eye's element is made sensitive in Geant4 simulation to evaluate the dose deposition in each of the biological structures of the eye.

Subsequently, an image of an eye from a Computed Tomography (CT) is implemented as a realistic detector, by importing in Geant4 the CT data. The image's information is stored in a series of DICOM (Digital Imaging and Communication in Medicine) files containing the image of a specific CT slice. The DICOM-detector is built by superimposing 10 contiguous CT files.

Results

Before the eye-detector irradiation a validation of the Geant4 simulation with CNAO experimental data is carried out with both carbon ions and protons.
Important beam parameters such as the transverse FWHM and scanned radiation field 's uniformity are tested within the simulation and compared with experimental measurements at CNAO Centre.
The physical processes involved in secondary particles generation by carbon ions and protons in the eye-detector are reproduced to take into account the additional dose to the primary beam given to irradiated eye's tissues.
A study of beam shaping is carried out to produce a uniform 3D dose distribution (shaped on the tumor) by the use of a spread out Bragg peak. The eye-detector is then irradiated through a two dimensional transverse beam scan at different depths. In the use case the eye-detector is rotated of an angle of 40° in the vertical direction, in order to misalign the tumor from healthy tissues in front of it. The treatment uniformity on the tumor in the eye-detector is tested.
For a more quantitative description of the deposited dose in the eye-detector and for the evaluation of the ratio between the dose deposited in the tumor and the other eye components, proton and carbon DVHs (Dose Volume Histograms) are compared. A high statistics simulated sample is used to minimize statistical errors.
In the simulation a new particle generation method is developed in order to reproduce the experimental treatment plan by importing the DICOM RT-PLAN file, which contains all the information on the irradiation geometries and sequences (treatment plan parameters).

Conclusions

Even further validations must be done, the good results so far obtained by this work point out and confirm the possibility of using carbon ions delivered with active scanning beams to treat the ocular melanoma.

Track 4 Session / 348

Testable physics by design

Authors: Chan Hyeong Kim[1]; Chansoo Choi[2]; Gabriela Hoff[None]; Georg Weidenspointner[3]; Han Sung Kim[1]; Maria Grazia Pia[4]; Markus Kuster[5]; Mincheol Han[6]; Paolo Saracco[7]; Steffen Hauf[5]; Sung Hun Kim[6]

[1] *Hanyang University*

[2] *Hanyang University (KR)*

[3] *MPE Garching*

[4] *Universita e INFN (IT)*

[5] *European XFEL GmbH*

[6] *H*

[7] *INFN - Genova (Italy)*

Corresponding Author: maria.grazia.pia@cern.ch

Testable physics by design

The validation of physics calculations requires the capability to thoroughly test them. The difficulty of exposing parts of the software to adequate testing can be the source of incorrect physics functionality, which in turn may generate hard to identify systematic effects in physics observables produced by the experiments.

Starting from real-life examples encountered in the course of an extensive validation effort of Geant4 physics, we show how software design choices may affect the ability to test basic aspects of physics

functionality and to monitor their evolution in the course of software lifecycle. We document cases where inaccurate physics behaviour was hidden by software design that prevented its testability, and illustrate how improved physics functionality could be achieved by improving the transparency of software design.

Exploiting the experience with Geant4 as a playground for investigation, we discuss methods to enhance the testability of existing software systems by means of refactoring techniques, the identification of inflection points and their coverage. We discuss how these techniques differ from, but can be combined with, the traditional practice of monitoring some physics observables with regression testing.

We also present guidelines to introduce testability into the software design since the early phases of the software development, and to preserve it in the course of the product lifecycle. This issue is especially relevant in the context of ongoing R&D on future simulation systems.

Track 6 Session / 350

Possibilities for Named Data Networking in HEP

Authors: David Colling[1]; Duncan Rand[1]; Simon Fayer[1]

[1] *Imperial College Sci., Tech. & Med. (GB)*

Corresponding Author: duncan.rand@cern.ch

Named Data Networks (NDN) are an emerging network technology based around requesting data from a network rather than a specific host. Intermediate routers in the network cache the data. Each data packet must be signed to allow its provenance to be verified. Data blocks are addressed by a unique name which consists of a hierarchical path, a name and attributes. An example of a valid address could be "/ndn/uk/ac/imperial/ph/hep/data/somefile/1".

The provision for in-network caching makes NDN an ideal choice for transferring large data sets where different endpoints are likely to make use of the same data within close succession. The naming of data rather than nodes also means that a data request could potentially be satisfied by multiple geographically disparate sites allowing for failover and load-balancing.

We believe that the delegation of robustness and reliability to the network itself offers significant possibilities for computing in HEP. For example, the LHC experiments currently pre-place data and have more recently started making use of storage namespace federation and caching through the xrootd protocol. NDN offers interesting possibilities to simplify many of the the experiments'data placement systems and to extend the data caching approach. Another advantage is that in the future it is likely that large commercial content delivery providers will be using NDN and will contribute effort into developing and maintaining the technology. The deployment of NDN will be a slow process, however it can run over the existing IP infrastructure allowing for a phased, non-disruptive parallel rollout.

We will discuss results from an HEP NDN testbed, prototype *GFAL 2* and *root* NDN plugins, aspects of packet signing and security and implications for the LHC computing models.

351

Dynamic partitioning as a way to exploit new computing paradigms: the cloud usecase.

Authors: Stefano Dal Pra[1]; Vincenzo Ciaschini[2]

Co-author: Luca dell'Agnello [1]

[1] *INFN*

[2] *Istituto Nazionale Fisica Nucleare (IT)*

Corresponding Author: stefano.dalpra@cnaf.infn.it

The WLCG community and many groups in the HEP community have based their computing strategy on the Grid paradigm, which proved successful and still ensues its goals. However, Grid technology has not spread much over other communities; in the commercial world, the cloud paradigm is the emerging way to provide computing services.

WLCG experiments aim to achieve integration of their existing current computing model with cloud deployments and take advantage of the so-called opportunistic resources (including HPC facilities) which are usually not Grid compliant. One missing feature in the most common cloud frameworks, is the concept of job scheduler, which plays a key role in a traditional computing centre, by enabling a fairshare based access at the resources to the the experiments in a scenario where demand greatly outstrips availability.

At CNAF we have opened started, as a preproduction service, the possibility to access the Tier-1 computing resources as an OpenStack based cloud service. The system, exploiting the dynamic partitioning mechanism already being used to enable Multicore computing, allowed us to avoid a static splitting of the computing resources in the Tier-1 farm, while permitting a share friendly approach.

The hosts in a dynamically partitioned farm may be moved to or from the partition, according to suitable policyes for request and release of computing resources. Nodes being requested in the partition switch their role and become available to play a different one. In the cloud use case hosts may switch from acting as Worker Node in the Batch system farm to cloud compute node member, made available to tenants.

In this paper we describe the dynamic partitioning concept, its implementation and integration with our current batch system, LSF. We then present results for the dynamic cloud usecase.

Track 5 Session / 352

Data-analysis scheme and infrastructure at the X-ray free electron laser facility, SACLA

Author: Takashi SUGIMOTO[1]

Co-authors: Kensuke OKADA [1]; Makina YABASHI [2]; Mitsuhiro YAMAGA [1]; Ryotaro TANAKA [1]; Takaki Hatsui [2]; Takashi KAMESHIMA [1]; Toshinori Abe [1]; Yasumasa JOTI [1]

[1] *Japan Synchrotron Radiation Research Institute*

[2] *RIKEN SPring-8 Center*

Corresponding Author: takashi.sugimoto@spring8.or.jp

An X-ray free electron laser (XFEL) facility, SACLA, is generating ultra-short, high peak brightness, and full-spatial-coherent X-ray pulses 1. The unique characteristics of the X-ray pulses, which have never been obtained with conventional synchrotron orbital radiation, are now opening new opportunities in a wide range of scientific fields such as atom, molecular and optical physics, ultrafast science, material science, and life science. More than 100 experiments have been performed since the first X-ray delivery to experimental users in March 2012. In this paper, we present an overview of SACLA data acquisition (DAQ) and analysis system with a special emphasis on the analysis scheme

and its infrastructure. In the case of serial femotosecond protein crystallography experiments 2, a typical experiment collects diffraction image patterns of order of 10^6, which demands heavy load to the data analysis system. Each pattern is recorded by a 2000×2000 pixel detector that consists of eight Multiport Charge-Coupled Device (MPCCD) sensors 3. The resolution of the single MPCCD sensor is 1024×512 pixel and data depth of each pixel is 16 bits. The DAQ system consists of detector front-ends 4, data-handling servers, hardware-based event-tag distribution system 5, event-synchronized database 6, two cache storages, tape archive system, and physically-segregated two network system 7. The DAQ has data bandwidth of maximum 6 Gbps to support other experiment setups with various detector configuration of up to twelve MPCCD sensors. In addition to the currently operational beamline BL3, BL2 will operate concurrently through a fast-switching operation mode in 2015. To support this operation mode, the cache storage with capacities of 200 TB (250 TB) is assigned to the beamline BL2 (BL3) respectively 8. These capacities correspond to the accumulated data size for one week operation. Experimental data are periodically moved into the tape archive system. The tape archive system has a capacity of 7 PB, and extendable up to 26 PB by installing additional tape cartridges.

The analysis section has two functions: one is run-by-run analysis to monitor the experimental conditions, and the other is off-line analysis. To implement these functions, the analysis system consists of a PC cluster and a supercomputer. The PC cluster is based on x86_64 processors and has a computing power of 14 TFLOPS. A 160 TB storage is connected to the PC cluster via Infiniband QDR network. To pick up raw image data, the PC cluster is connected to the cache storages and the tape archive system via 10 Gigabit Ethernet. The run-by-run analysis is performed using the PC cluster. The results are saved on the storage in HDF5 format [9]. The PC cluster is also used for off-line analysis, using analysis code developed by the scientific community, such as CrystFEL [10] and SITENNO [11]. The supercomputer with 90 TFLOPS SPARC-based processors (Fujitsu FX10) was installed for the data analysis that requires higher computing power. Storage of the supercomputer is 100+500 TB Lustre-based file system (Fujitsu FEFS). Another 1 PB Lustre file system is also connected to both the supercomputer and the PC cluster to interaccess the experimental data from the two systems. Data analysis that requires much higher computing power is foreseen. For these cases, we are developing a joint analysis mode using both the supercomputer and the 10-PFLOPS K computer [12]. The results of the feasibility study on data transfer and quasi-realtime job submission to the K computer will also be discussed.

1 T. Ishikawa et al., “A compact X-ray free-electron laser emitting in the sub-angstrom region”, Nature Photonics 6, 540-544 (2012).

2 M. Sugahara et al., Nature Methods submitted.

3 T. Kameshima et al., “Development of an X-ray pixel detector with multi-port charge-coupled device for X-ray free-electron laser experiments”, Rev. Sci. Instrum. 85, 033110 (2014).

4 A. Kiyomichi, A. Amselem, et al., “Development of Image Data Acquisition System for 2D Detector at SACLA”, Proceedings of ICALEPCS2011, WEPMN028 (2011).

5 T. Abe et al., “DEVELOPMENT OF NEW TAG SUPPLY SYSTEM FOR DAQ FOR SACLA USER EXPERIMENTS”, Proceedings of IPAC2014, TUPRI108 (2013).

6 M. Yamaga et al., “Event-Synchronized Data Acquisition System of 5 Giga-bps Data Rate for User Experiment at the XFEL Facility, SACLA”, Proceedings of ICALEPCS2011, TUCAUST06 (2011).

7 T. Sugimoto et al., “Large-bandwidth Data-acquisition Network for XFEL Facility, SACLA”, Proceedings of ICALEPCS2011, WEBHAUST03 (2011).

8 K. Okada, et al., “UPGRADE OF SACLA DAQ SYSTEM ADAPTS TO MULTI-BEAMLINE OPERATION”, Proceedings of PCaPAC2014, WCO205 (2014).

[9] The HDF Group, http://www.hdfgroup.org/HDF5/.

[10] T. A. White, et al.. “CrystFEL: a software suite for snapshot serial crystallography”. J. Appl. Cryst. 45, pp.335–341 (2012).

[11] Y. Sekiguchi et al., “Data processing software suite SITENNO for coherent X-ray diffraction imaging using X-ray free electron laser SACLA”, Journal of Synchrotron Radiation 21/ 4, 600-612 (2014).

[12] M. Yokokawa, et al., "The K computer: Japanese next-generation supercomputer development project", ISLPED '11 Proceedings of the 17th IEEE/ACM international symposium on Low-power electronics and design, pp.371-372 (2011).

Track 1 Session / 353

Efficient time frame building for online data reconstruction in ALICE experiment

Author: Alexey Rybalchenko[1]

Co-authors: Charalampos Kouzinopoulos [2]; Mohammad Al-Turany [2]; Nicolas Winckler [3]

[1] *GSI - Helmholtzzentrum fur Schwerionenforschung GmbH (DE)*

[2] *CERN*

[3] *GSI*

Corresponding Author: alexey.rybalchenko@cern.ch

After Long Shutdown 2, the upgraded ALICE detector at the LHC will produce more than a terabyte of data per second. The data, constituted from a continuous un-triggered stream data, have to be distributed from about 250 First Level Processor nodes (FLPs) to O(1000) Event Processing Nodes (EPNs). Each FLP receives a small subset of the detector data that is chopped in sub-timeframes. One EPN needs all the fragments from the 250 FLPs to build a full timeframe. An algorithm is being implemented on the FLPs with the aim of optimizing the usage of the network connecting the FLPs and EPNs. The algorithm minimizes contention when several FLPs are sending to the same EPN. An adequate traffic shaping is implemented by delaying the sending time of each FLP by a unique offset. The payloads are stored in a buffer large enough to accommodate the delay provoked by the maximum number of FLPs. As the buffers are queued for sending, the FLPs can operate with the highest efficiency. Using the time information embedded in the data any further FLP synchronization can be avoided. Moreover, "zero-copy"and multipart messages of ZeroMQ are used to create full timeframes on the EPNs without the overhead of copying the payloads. The concept and the performance measurement of the implementation on a computing cluster are presented.

355

Scientometrics of Monte Carlo simulation: lessons learned and how HEP can profit from them

Authors: Maria Grazia Pia[1]; Tullio Basaglia[2]

Co-authors: Paul Dressendorfer [3]; Zane Bell [4]

[1] *Universita e INFN (IT)*

[2] *CERN*

[3] *IEEE*

[4] *ORNL*

Corresponding Author: maria.grazia.pia@cern.ch

An extensive scientometric assessment of the literature is presented, which documents the prominent role achieved by Monte Carlo methods, and simulation in general, in particle physics and related fields (nuclear physics, astrophysics, medical physics etc.). As an example of their pervasiveness, one can remark that currently approximately 50% of the papers published in major, multi-disciplinary nuclear technology journals mention simulation methods. Scientometric data objectively evidence

their use not only in traditional academic research domains, but also in industry, hospitals and other disciplines not directly related to high energy physics.

The analysis of scientometric data identifies some relevant features related to Monte Carlo methods and codes: it quantifies their impact on HEP itself, on scientific communities other than HEP, on industry and on social aspects of life sciences. Both widely used, general purpose Monte Carlo codes and specialized codes that address specific needs or user communities are examined.

The presentation highlights the contribution of HEP to this scientific domain, and discusses how the HEP community could profit from these achievements by improved communication methods addressed to the civil society, to funding agencies and to government bodies.

Track 7 Session / 356

Evaluation of containers as a virtualisation alternative for HEP workloads

Authors: Andrew John Washbrook[1]; Gareth Douglas Roy[2]

Co-authors: David Britton [2]; David Crooks [2]; Gang Qin [2]; Samuel Cadellin Skipsey

[1] *University of Edinburgh (GB)*

[2] *University of Glasgow (GB)*

Corresponding Author: andrew.washbrook@cern.ch

Cloud computing enables ubiquitous, convenient and on-demand access to a shared pool of configurable computing resources that can be rapidly provisioned with minimal management effort. The flexible and scalable nature of the cloud computing model is attractive to both industry and academia. In HEP, the use of the "cloud"has become more prevalent with LHC experiments making use of standard Cloud technologies to take advantages of elastic resources in both private and commercial computing environments.

A key software technology that has eased transition to a cloud environment is the Virtual Machine (VM). VM's can be dynamically provisioned, managed and run a variety of Operating Systems tailored to user requirements. From a resource utilisation perspective however, VM's are a considered a heavyweight solution. Upon instantiation a VM will contain a complete copy of an operating system and all associated services leading to an increase in resource consumption when compared to standard "bare metal" deployment. This level of virtualisation is not required by the majority of workloads processed in HEP and can lead to increases in execution time on workloads that performs intensive I/O file operations such as LHC data analysis.

An alternative solution which is gaining rapid traction within industry is containerisation. Here the Linux Kernel itself can virtualise and isolate a user-land level instance of the operating system in which applications can run. Less resources are needed compared to a VM because only shared system libraries and files needed by the application are virtualised. Additionally, as the virtualisation takes place via namespaces (a mechanism provided by the Linux Kernel giving an isolated view of a global system resource) performance is close to that of the physical hardware with minimal tuning.

In this study the use of containers will be assessed as a mechanism for running LHC experiment application payloads. Using currently available software tools (Docker) deployment strategies will be investigated by the use of a distributed WLCG Tier-2 facility as an example computing site. The relative performance of Containers and VM's when compared with native execution will be detailed using hardware benchmarking tools such as the HEPSPEC suite and Bonnie++. System-level resource profiling will also be performed on a selection of representative LHC workloads in order to determine the relative efficiency of hardware utilisation in each scenario.

As part of this deployment alternative methods of providing resources to WLCG similar to those found in "Cloud"solutions will be investigated. The integration of Containers in existing Cloud platforms (OpenStack) will be explored as well as new and emerging platforms (CoreOS, Kubernetes).

Additionally, the possibility of hosting Grid Services in containers to ease Middleware deployment will also be assessed.

358

Deployment and usage of perfSonar Networking tools for non-HEP communities

Author: Brian Davies[1]

Co-authors: Alex White [2]; Cristina Del Cano Novales [3]

[1] *STFC (RAL) (GB)*

[2] *DLS Ltd*

[3] *Unknown-Unknown-Unknown*

Corresponding Author: brian.davies@stfc.ac.uk

perfSonar is a network monitoring tool set which enables performance of wide area communications to be analysed and eases problem identification across distributed centres. It has been widely used within WLCG since 2012 and has been crucial in identifying network problems and confirming network changes have the desired effect. We report on examples of this within this presentation.

In addition, at Rutherford Appleton Laboratory (RAL), we have been extending the use of perfSonar to communities outside of High Energy Physics (HEP) such as photonics and climatology. We describe how the knowledge base gained from our experience in WLCG has been used to help these non-HEP communities to deploy perfSonar hosts and incorporate these into the current test infrastructure. We demonstrate improvements that have resulted from the monitoring information obtained such as identification of packet loss across the firewall which had not previously been identified by any other tools used.

359

Filesize distribution of WLCG data at the Rutherford Appleton Laboratory Tier1

Author: Brian Davies[1]

Co-author: Shaun de Witt [2]

[1] *STFC (RAL) GB*

[2] *STFC*

Corresponding Author: brian.davies@stfc.ac.uk

The Rutherford Appleton Laboratory (RAL) operates the UK WLCG Tier1 facility on behalf of GridPP. Tier 1's provide persistent archival storage (on tape at RAL) and online storage for fast access data analysis. RAL is one of the few Tier-1s which supports data management for all the major LHC experiments, as well as a number of smaller Virtual Organisations. This allows us to compare usage of the infrastructure across many different experiments and use cases.

In this post, we look at the distribution of file sizes between VOs and across different storage work flows within the same VO. We also report on how this distribution impacts standard operational procedures such as tape usage and decommissioning of storage nodes.

Further, we also show how a hypothetical change on data storage policy of log files would change the profile of filesize distribution of a significant subset of data currently being stored at RAL and how this change would impact operations to the advantage of the user communities.

360

A Grid-based Batch Reconstruction Framework for MICE

Author: Janusz Martyniak[1]

Co-author: Henry Nebrensky [2]

[1] *Imperial College London*

[2] *Brunel University*

Corresponding Author: janusz.martyniak@imperial.ac.uk

The international Muon Ionisation 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 2015-2016), and demonstration of Ionisation Cooling (2017-2018).

The MICE Batch Reconstruction framework reconstructs all MICE data recorded to date for a particular MICE Step, using a specified MAUS version. Our job manager process holds a local SQLite database of MICE runs already reconstructed with a given MAUS version; it repeatedly checks against the MICE Metadata DB and on identifying data runs that have not yet been reconstructed it will submit a job to the Grid to run MAUS against that data run. For this we select from those Tier2 sites that hold a copy of the raw data. The reconstructed data are stored on the SE at the site where the job runs and registered in the LFC. On successful reconstruction, the Grid job registers an FTS transfer request with a separate File Transfer Controller (Web Service based) to copy the reconstructed data to the Castor SE at RAL, where they are stored on tape. This decouples the FTS transfer request and monitoring from the Grid job. After the output has been transferred to the Castor SE at RAL, a corresponding record is added to the MICE Metadata DB by the File Transfer Controller.
The MICE Metadata DB can then co-ordinate the further distribution of the reconstructed data across the Grid, including to HTTP-enabled SE's to allow MICE collaborators to access the data directly from any web browser.

362

Storage Interface Usage at a Large, Multi-Experiment Teir1

Author: Shaun de Witt[1]

Co-author: Matthew Reggler [1]

[1] *STFC*

Corresponding Author: shaun.de-witt@stfc.ac.uk

Within WLCG much has been discussed concerning the possible demise of the Storage Resource Manager (SRM) and replacing it with different technologies such as XrootD and WebDAV. Each of these storage interfaces presents different functionalities and experiments currently make use of all of these at different sites. At the RAL Tier-1 we have been monitoring the usage of both SRM and XrootD by all the major experiments to assess when and if some of the nodes hosting the SRM should be redeployed to support XRootD.
Initial results were presented at ISGC2014 which showed the SRM still handles the majority of requests issued by most experiments (ALICE do not and have never used the SRM), but with an increasing usage of XrootD, particularly by ATLAS. This poster will update these results based on several months of additional data and show that the SRM is still widely used by all the major WLCG VOs and all smaller experiments such as T2K and NA62. We break down this usage according by read/write and by geographic source to show how a large Tier 1 is used globally. We also analyse usage by 'use case'(archival storage, persistent storage and scratch storage) and show how different experiments make use of the SRM.

Track 1 Session / 363

Online tracking with GPUs at PANDA

Authors: Andreas Herten[1]; Ludovico Bianchi[1]

Co-authors: Andrew Adinetz [2]; Dirk Pleiter [3]; James Ritman [4]; Jiri Kraus [5]; Tobias Stockmanns [6]

[1] *Forschungszentrum Jülich*

[2] *Jülich Supercomputing Centre, Forschungszentrum Jülich*

[3] *DESY*

[4] *CERN*

[5] *NVIDIA GmbH, Germany*

[6] *Forschungszentrum Jülich GmbH*

Corresponding Author: l.bianchi@fz-juelich.de

The PANDA experiment is a next generation particle detector planned for operation at the FAIR facility, currently under construction in Darmstadt, Germany. PANDA will detect events generated by colliding an antiproton beam on a fixed proton target, allowing studies in hadron spectroscopy, hypernuclei production, open charm and nucleon structure.

The nature of hadronic collisions means that signal and background events will look very similar, making a conventional approach, based on a hardware trigger signal generated by a subset of the detectors to start the data acquisition, unfeasible.
Instead, data coming from the detector are acquired continuously, and all online selection is performed in real-time. A rejection factor of about 1000 is needed to reduce the data rate for offline storage, making the data acquisition system computationally very challenging.

Adoption of Graphical Processing Units (GPUs) in many computing applications is increasing, due to their cost-effectiveness, performance, and accessible and versatile development using high-level programming paradigms such as CUDA or OpenCL. Applications of GPU within HEP include Monte Carlo production, analysis, low- and high-level trigger.

Online track reconstruction of charged particles plays an essential part in the event reconstruction and selection process. Our activity within the PANDA collaboration is centered on the development and implementation of particle tracking algorithms on GPUs, and on studying the possibility of performing online tracking using a multi-GPU architecture. Three algorithms are currently under development, using information from the PANDA tracking system: a Hough Transform; a Riemann Track Finder; and a Triplet Finder algorithm, a novel approach finely tuned for the PANDA STT detector. The algorithms are implemented on the GPU in the CUDA C language, utilizing low-level optimizations and non-trivial data packaging in order to exploit to the maximum the capabilities of GPUs.

This talk will present details of the implementation of these algorithms, together with first performance results, and solutions for data transfer to and from GPUs based on message queues for a deeper integration of the algorithms with the FairRoot and PandaRoot frameworks, both for online and offline applications.

Track 5 Session / 364

Analysis Traceability and Provenance for HEP

Author: Jetendr Shamdasani[1]

Co-authors: Andrew Branson [1]; Richard Mcclatchey [1]; Zsolt Kovacs [1]

[1] *University of the West of England (GB)*

Corresponding Author: jetendr.shamdasani@cern.ch

In complex data analyses it is increasingly important to capture information about the usage of data sets in addition to their preservation over time in order to ensure reproducibility of results, to verify the work of others and to ensure appropriate conditions data have been used for specific analyses. This so-called provenance data in the computer science world is defined as the history or derivation of a data (or process) artifact. Many scientific workflow based studies are beginning to realise the benefit of capturing the provenance of their data and the activities used to process, transform and carry out studies on that data. This is especially true in scientific disciplines where the collection of data through experiment is costly and/or difficult to reproduce and where that data needs to be preserved over time. With the increase in importance of provenance data many different models have emerged to capture provenance such as PROV or the OPM models. However, these are more for interoperability of provenance information and do not focus on the capture of provenance related data. There is a clear and emerging requirement for systems to handle the provenance of data over extended timescales with a emphasis on preserving the analysis procedures themselves and the environment in which the analyses were conducted alongside the processed data sets.

A provenance data system that has been built in house at CERN since early 2000 is called CRISTAL. CRISTAL was used to capture the provenance resulting from the design and construction of the CMS ECAL detector over the period 2002-2010. The CRISTAL Kernel (V3.0) has now been launched as open source under the LGPL (V3) licencing scheme and it is available for use by the wider communities including teams involved in the offline analysis of physics data, whether at CMS or other experiments. In addition, in the EC funded neuGRID and N4U projects the original developers have been using CRISTAL to capture the provenance of analyses for neuroscientists running complex pipelines of algorithms in the study of biomarkers for the onset of Alzheimer's disease. In this paper this application is presented with a focus on how its approach can be customised for use in the high energy physics data analysis community at large. The main focus of this is a set of analysis tools (persistency, browsing/querying, visualising and analysis tracking services) which together with a generic analysis model backend can be used to capture the information required to support complex analyses.
The Analysis Tools comprise the following interfaces :

- **The Analysis Web Serivce** – Which is for advanced users, these users are able to programmatically create analyses on the fly.
- **The Analysis Command-line Interface** – This is for users that are intermediate/advanced. It allows users to create analysis using a shell like interface.
- **The Analysis Web Portlet Interface** – This is for novice users, it is a visual interface which is portlet based and allows users to browse datasets and pipelines. It also allows users to create and deploy their analyses in a visual manner.
- **The Analysis Core** – These are a core set of objects used by the above interfaces. These objects connect directly to a customised analysis aware CRISTAL instance which is provenance aware.

During the provenance capture phase the Analysis tools are able to capture :

- **who** ran an analysis, this is a user name,
- for **what** purpose, what their analysis is supposed to achieve,
- **when** they ran it this is a timestamp which denotes when it started and when it finished,
- **where** it was run this is GRID and Cloud related information,
- **which** datasets and algorithms were used to create and run their analyses,
- **how** it was executed, this more detailed infrastructure information
- and lastly **why** the analysis was run, this is a justification from the user.

Also in this paper, we present the case for using the Analysis Services developed using CRISTAL as an avenue for long timescale data preservation. The tools are able to store the provenance metadata surrounding the analyses in a human readable form (XML). This is a light-weight and queryable manner of storing provenance as well as analysis results. In CRISTAL everything is recorded and nothing is thrown away. So another user would be able to replicate the experiment at a later date and time. Besides reproducibility of experiments, users can also share their experiments with other users using the provenance related information.

The analysis tools run currently in a GRID and a Cloud infrastructure. As well as collecting analysis provenance information, they are able to provenance of the infrastructure. There is strong novelty in this work which facilitates allows more precise reproducibility of experiments. This information is known as *infrastructure provenance*. It is currently being collected in the course of the N4U project. This infrastructure provenance can also be applied to HEP to aid in the reproducibility of results. For example, if performance is a factor it can be sent to the same compute node.

Concerning the future of the analysis tools there is currently an emerging standard known as PROV which is used for *provenance interoperability*. In the near future, the analysis provenance information that we have collected with be exported to PROV. This work has already begun, we are looking for mapping patterns to aid in our cause. The reasoning for this is so that people can study and use provenance information in a standard and commonly understood format. This will also allow users to publish the provenance generated from their analyses onto the ever growing linked data cloud as well.

Track 5 Session / 365

Distributed Computing for Pierre Auger Observatory

Author: Jiri Chudoba[1]

[1] *Acad. of Sciences of the Czech Rep. (CZ)*

Corresponding Author: jiri.chudoba@cern.ch

Pierre Auger Observatory operates the largest system of detectors for ultra-high energy cosmic ray measurements. Comparison of theoretical models of interactions with recorded data requires thousands of computing cores for Monte Carlo simulations. Since 2007 distributed resources connected via EGI grid are succesfully used. The first and the second versions of production system based on bash scripts and MySQL database were able to submit jobs to all reliable sites supporting Virtual Organization auger. Many years VO auger belongs to top ten of EGI users based on the total used computing time.

Migration of the production system to DIRAC interware started in 2014. Pilot jobs improve efficiency of computing jobs and eliminate problems with small and less reliable sites used for the bulk production. The new system has also possibility to use available resources in clouds. Dirac File Catalog replaced LFC for new files, which are organized in datasets defined via metadata. CVMFS is used for software distribution since 2014. In the presentation we give a comparison of the old and new production system and report the experience from migration to the new system.

366

Judith: A Software Package for Synchronised Analysis of Test-beam Data

Author: Andrej Gorisek[1]

Co-authors: Garrin Mcgoldrick [2]; Matevz Cerv [3]

[1] *Jozef Stefan Institute (SI)*

[2] *University of Toronto (CA)*

[3] *CERN*

Corresponding Authors: andrej.gorisek@cern.ch, garrin.mcgoldrick@cern.ch, matevz.cerv@cern.ch

The Judith software performs pixel detector analysis tasks utilising two different data streams such as those produced by the reference and tested devices typically found in a testbeam. This software addresses and fixes problems arising from the desynchronization of the two simultaneously triggered heterogeneous data streams by detecting missed triggers in either of the streams. The software can perform all tasks required to generate particle tracks using multiple detector planes: it can align the planes, cluster hits and generate tracks from these clusters. This information can then be used to measure the properties of a particle detector with very fine spatial resolution. It was successfully used by the authors at DESY in a KarTel telescope with an ATLAS Diamond Beam Monitor module as a DUT as well as more recently by other groups testing pixel detectors at CERN PS and SPS.

367

fwk: a go-based concurrent control framework

Author: Sebastien Binet[1]

[1] *IN2P3/LAL*

Corresponding Author: sebastien.binet@cern.ch

fwk: a go-based concurrent control framework

Current HEP control frameworks have been designed and written in the early 2000's, when multi-core architectures were not yet pervasive.
As a consequence, an inherently sequential event processing design emerged.
Evolving current frameworks' APIs and data models encouraging global states, non-reentrancy and non-thread-safety to a more concurrent friendly environment, in an adiabatic way, is a major undertaking, even more so when relying on the building blocks provided by C++.

This paper reports on the development of fwk, a framework investigating and leveraging the built-in tools of the Go language to enable concurrency at the event- and sub-event-levels. The concurrency features, code distribution and tooling of Go will be first presented to set the scene and explain why Go -and its ecosystem at large- is a natural fit for a concurrent framework.
The aim for such a framework is to be usable as a big HEP experiment's control framework. fwk is also meant to be a nimble application for analyses which require quick development/deployment cycles but without paying the price of a VM on the runtime performances side.

The paper will then discuss the design decisions applied to the various components of fwk, introduce its current capabilities

(I/O, histogramming, dataflow, ...) and highlight how the design decisions of the Go language adequately paved the way for fwk.
Then, fads, a fast detector simulation toolkit, will be introduced. 'fads' is the port of a single-threaded fast detector simulation toolkit (Delphes) to fwk.
The paper will present benchmarks (CPU, RSS, I/O, scalability) of real Delphes use cases against fads, as well as comparisons of fads performances with regard to other next-generation concurrent frameworks, using synthetic data.
Performance tools and concurrency debugging tools provided with the Go toolchain and used for this comparison will also be presented.

Finally, the paper will present prospects and prototypes to fill the gap in the nascent Go-based HEP ecosystem and landscape - namely: histograms interactive displays and interactive analysis tools.

Track 7 Session / 368

docker & HEP: containerization of applications for development, distribution and preservation.

Author: Sebastien Binet[1]

Co-author: Ben Couturier [2]

[1] *IN2P3/LAL*

[2] *CERN*

Corresponding Author: ben.couturier@cern.ch

docker & HEP: containerization of applications for development, distribution and preservation.

HEP software stacks are not shallow.
Indeed, HEP experiments' software are usually many applications in one (reconstruction, simulation, analysis, ...) and thus require many libraries - developed in-house or by third parties - to be properly compiled and installed.
Moreover, because of resource constraints, experiments' software is usually installed, tested, validated and deployed on a very narrow set of platforms, architectures, toolchains and operating systems. As a consequence, bootstrapping a software environment on a developer machine or deploying the software on production or user machines is usually perceived as tedious and iterative work, especially when one wants the native performances of bare metal.

Docker containers provide an interesting avenue for packaging applications and development environment, relying on the Linux kernel capabilities for process isolation, adding "git"-like capabilities to the filesystem layer and providing (close to) native CPU, memory and I/O performances.

This paper will introduce in more details the modus operandi of Docker containers and then focus on the hepsw/docks containers which provide containerized software stacks for -among others- LHCb. The paper will then discuss various strategies experimented with to package software (e.g. CVMFS, RPMs, from-source) and

applied to optimize provisioning speed and disk usage, leveraging the caching system of Docker.

Finally, the paper will report on benchmarks comparing workloads on native, VM and Docker containers setups.

369

Evaluation of NoSQL database MongoDB for HEP analyses

Author: Jörg Meyer[1]

Co-authors: Achim Streit [2]; Christopher Jung [2]

[1] *Karlsruhe Institute of Technology*

[2] *KIT - Karlsruhe Institute of Technology (DE)*

Corresponding Author: christopher.jung@kit.edu

Most analyses in experimental high-energy physics (HEP) are based on the data analysis framework ROOT. Therefore, simulated as well as measured events are stored in ROOT trees. A typical analysis loops over events in ROOT files and selects relevant events for further processing according to certain selection criteria.
The emergence of NoSQL databases provide a new mean for large scale data storage and analyses. NoSQL databases allow horizontal scaling and can be used as a new approach for HEP analyses.
We present a study that stores simulated top-pair-production events in the document-based NoSQL database MongoDB. We compare analysis steps using MongoDB as input source to the traditional approach using ROOT trees. We demonstrate that many analysis steps beyond the selection of events are supported by the database query language.
The advantages and limitations of the NoSQL approach are discussed in terms of functionality, performance, storage efficiency, and usability.

Track 6 Session / 370

Challenge and Future of Job Distribution at a Multi-VO Grid Site

Author: Andreas Gellrich[1]

Co-author: Birgit Lewendel [2]

[1] *DESY*

[2] *Deutsches Elektronen-Synchrotron (DE)*

Corresponding Author: birgit.lewendel@cern.ch

DESY operates a multi-VO Grid site for 20 HEP and non-HEP collaborations and is one of the world-wide largest Tier-2 sites for ATLAS, CMS, LHCb, and BELLE2. In one common Grid infrastructure computing resources are shared by all VOs according to MoUs and agreements, applying an opportunistic usage model allows to distribute free resources among the VOs. Currently, the Grid site DESY-HH provides roughly 100kHS06 in 10000 job slots, exploiting the queueing system PBS/TORQUE.

As described in former CHEP conferences, resource utilization and job scheduling in a multi-VO environment is a major challenge. On one hand

side all job slots should be occupied, on the other hand jobs with diverging resource usage patterns must be cleverly distributed to the compute nodes in order to guarantee stability and optimal resource usage.

Batch systems such as PBS/TORQUE with the scheduler MAUI only scale up to a few thousand job slots. At DESY-HH an alternative scheduler was developed and brought into operation.

In the preparation for LHC Run 2 as well as the start of BELLE2, in particular the request for the support of multi-core jobs requires appropriate job scheduling strategies which are not available out-of-the-box. Even more, the operation of work load managements system and pilot factories (DIRAC, PANDA) by the big collaborations question the way of how sites provide computing resources in the future. Is cloud computing the future? What about small VOs and individual users who use the standard Grid tools to submit jobs then?

In the contribution to CHEP2015 we will try the review what has been learned by operating a large Grid site for many VOs. We will give a summary of the experiences with the job scheduling at DESY-HH of the last years and we will describe limits of the current system.
A main focus will be put on the discussion of future scenarios, including alternatives to the approach of local resource managements systems (LRMS) which are still widely used.

371

Pushing HTCondor and glideinWMS to 200K+ Jobs in a Global Pool for CMS before LHC Run 2

Author: James Letts[1]

Co-authors: Alison Mc Crea [1]; Brian Paul Bockelman [2]; David Alexander Mason [3]; Farrukh Aftab Khan [4]; Igor Sfiligoi [1]; Justas Balcas [5]; Krista Majewski [3]; Luis Emiro Linares Garcia [6]; Marco Mascheroni [7]; Maria Dolores Saiz Santos [1]; Oliver Gutsche [3]; Stefano Belforte [8]

[1] *Univ. of California San Diego (US)*

[2] *University of Nebraska (US)*

[3] *Fermi National Accelerator Lab. (US)*

[4] *National Centre for Physics (PK)*

[5] *Vilnius University (LT)*

[6] *Universidad de los Andes (CO)*

[7] *Universita & INFN, Milano-Bicocca (IT)*

[8] *Universita e INFN (IT)*

Corresponding Author: letts@physics.ucsd.edu

The CMS experiment at the LHC relies on HTCondor and glideinWMS as its primary batch and pilot-based Grid provisioning system. So far we have been running several independent resource pools, but we are working on unifying them all to reduce the operational load and more effectively share resources between various activities in CMS. The major challenge of this unification activity is scale. The combined pool size is expected to reach 200K job slots, which is significantly bigger than any other multi-user HTCondor based system currently in production. To get there we have studied scaling limitations in our existing pools, the biggest of which tops out at about 70K slots, providing valuable feedback to the development communities, who have responded by delivering improvements which have helped us reach higher and higher scales with more stability. We have also worked on improving the organization and support model for this critical service during Run 2

of the LHC. This contribution will present the results of the scale testing and experiences from the first months of running the Global Pool.

372

A New Pileup Mixing Framework for CMS

Authors: Mike Hildreth[1]; Mike Hildreth[2]

Co-authors: Vladimir Ivantchenko [3]; Vladimir Ivantchenko [3]

[1] *University of Notre Dame (US)*

[2] *Department of Physics-College of Science-University of Notre Da*

[3] *CERN*

Corresponding Author: mikeh@omega.hep.nd.edu

The CMS Simulation uses minimum bias events created by a "standard" event generator (e.g., Pythia) to simulate the additional interactions due to peripheral proton-proton collisions in each bunch crossing at the LHC (also known as pileup). Due to the inherent time constants of the CMS front-end electronics, many bunch crossings before and after the central bunch crossing of interest must be included in the simulation, leading to hundreds of minimum bias events being used for each simulated hard-scatter event.

We report on the performance gains in I/O load and computational speed made possible by a new framework that allows the combination of pileup events in a "pre-mixing" step. This has been made possible by the development of software that allows single-channel information to be combined at the digitization level in each sub-detector, rather than accumulating simulated hits from Geant. The logistics of large-scale production with pre-fabricated pileup distributions is described.

Track 7 Session / 373

Status and Roadmap of CernVM

Authors: Dario Berzano[1]; Gerardo Ganis[1]; Ioannis Charalampidis[1]; Jakob Blomer[1]; Predrag Buncic[1]; Rene Meusel[1]

[1] *CERN*

Corresponding Author: gerardo.ganis@cern.ch

Cloud resources nowadays contribute an essential share of resources for computing in high-energy physics. Such resources can be either provided by private or public IaaS clouds (e.g. OpenStack, Amazon EC2, Google Compute Engine) or by volunteers'computers (e.g. LHC@Home 2.0). In any case, experiments need to prepare a virtual machine image that provides the execution environment for the physics application at hand.
The CernVM virtual machine since version 3 is a minimal and versatile virtual machine image capable of booting different operating systems. The virtual machine image is less than 20 megabyte in size. The actual operating system is delivered on demand by the CernVM File System.
CernVM 3 has matured from a prototype to a production environment. It is used, for instance, to run LHC applications in the cloud, to tune event generators using a network of volunteer computers, and as a container for the historic Scientific Linux 5 and Scientific Linux 4
based software environments in the course of long-term data preservation efforts of the ALICE, CMS, and ALEPH experiments. We present experience and lessons learned from the use of CernVM at scale. We also provide an outlook on the upcoming developments. These developments include adding support for Scientific Linux 7, the use of container virtualization, such as provided by

Docker, and the streamlining of virtual machine contextualization towards the cloud-init industry standard.

Track 6 Session / 374

GridPP –preparing for Run-2 and the wider context

Author: Jeremy Coles[1]

[1] *University of Cambridge (GB)*

Corresponding Author: jeremy.coles@cern.ch

This first section of this paper elaborates upon the operational status and directions within the UK Computing for Particle Physics (GridPP) project as we approach LHC Run-2. It details the pressures that have been gradually reshaping the deployed hardware and middleware environments at GridPP sites –from the increasing adoption of larger multicore nodes to the move towards alternative batch systems and cloud alternatives - as well as changes being driven by funding considerations. The second section focuses on work being done with non-LHC communities, communities that GridPP has supported for many years with varying degrees of engagement, and lays out findings of a survey into where such communities desire more support and their changing requirements. The section also explores some of the early outcomes of adopting a generic DIRAC based job submission and management framework within GridPP. The third and final section of the paper highlights changes being made in GridPP operations in order to meet new challenges that are arising as recent technical advancements (in areas such as cloud/VM provision) reach production readiness.

Track 1 Session / 375

The CMS High Level Trigger

Author: Andrea Bocci[1]

[1] *CERN*

Corresponding Author: andrea.bocci@cern.ch

The CMS experiment has been designed with a 2-level trigger system: the Level 1 Trigger, implemented on custom-designed electronics, and the High Level Trigger (HLT), a streamlined version of the CMS offline reconstruction software running on a computer farm. A software trigger system requires a tradeoff between the complexity of the algorithms running on the available computing power, the sustainable output rate, and the selection efficiency. Here we will present the performance of the main triggers used during the 2012 data taking, ranging from simpler single-object selections to more complex algorithms combining different objects, and applying analysis-level reconstruction and selection. We will discuss the optimisation of the triggers and the specific techniques developed to cope with the increasing LHC pile-up, reducing its impact on the physics performance.

Track 8 Session / 376

Large-Scale Merging of Histograms using Distributed In-Memory Computing

Author: Jakob Blomer[1]

Co-author: Gerardo Ganis [1]

[1] *CERN*

Corresponding Author: jakob.blomer@cern.ch

Most high-energy physics analysis jobs are embarrassingly parallel except for the final merging of the output objects, which are typically histograms. Currently, the merging of output histograms scales badly. The running time for distributed merging depends not only on the overall number of bins but
also on the number partial histogram output files. That means, while the time to analyze data decreases linearly with the number of worker nodes, the time to merge the histograms in fact increases with the number of worker nodes.
On the grid, merging jobs that take a few hours are not unusual. In order to improve the situation, we present a distributed and decentral merging algorithm whose running time is independent of the number of worker nodes. We exploit full bisection bandwidth of local networks and we keep all intermediate results in memory. We present benchmarks from an implementation using the parallel ROOT facility (PROOF) and RAMCloud, a distributed key-value store that keeps all data in DRAM. Our results show that a real-world collection of ten thousand histograms with overall ten million non-zero bins can be merged in less than one minute.

377

Performance of muon-based triggers at the CMS High Level Trigger

Author: Benjamin Radburn-Smith[1]

[1] *Purdue University (US)*

Corresponding Author: benjamin.radburn-smith@cern.ch

The trigger systems of LHC detectors play a fundamental role in defining the physics capabilities of the experiments. A reduction of several orders of magnitude in the rate of collected events, with respect to the proton-proton bunch crossing rate generated by the LHC, is mandatory to cope with the limits imposed by the readout and storage systems limits. An accurate and efficient online selection mechanism is thus required to fulfill the task keeping maximal the acceptance to physics signals. The CMS experiment operates using a two-level trigger system. Firstly a Level-1 Trigger (L1T) system, implemented using custom-designed electronics, is designed to reduce the event rate to a limit compatible to the CMS Data Acquisition (DAQ) capabilities. A High Level Trigger System (HLT) follows, aimed at further reducing the rate of collected events finally stored for analysis purposes. The latter consists of a streamlined version of the CMS offline reconstruction software and operates on a computer farm. It runs algorithms optimized to make a trade-off between computational complexity, rate reduction and high selection efficiency. With the computing power available in 2012 the maximum reconstruction time at HLT was about 200 ms per event, at the nominal L1T rate of 100 kHz. An efficient selection of muons at HLT, as well as an accurate measurement of their properties, such as transverse momentum and isolation, is fundamental for the CMS physics programme. The performance of the muon HLT for single and double muon triggers achieved in Run I will be presented. Results from new developments, aimed at improving the performance of the algorithms for the harsher scenarios of collisions per event (pile-up) and luminosity expected for Run II will also be discussed.

378

Electrons and photons at High Level Trigger in CMS for Run II

Author: Afiq Aizuddin Bin Anuar[1]

[1] *University of Malaya (MY)*

Corresponding Author: afiq.anuar@cern.ch

The CMS experiment has been designed with a 2-level trigger system. The first level is implemented using custom-designed electronics. The second level is the so-called High Level Trigger (HLT), a streamlined version of the CMS offline reconstruction software running on a computer farm. For Run II of the Large Hadron Collider, the increase in center-of-mass energy and luminosity will raise the event rate to a level challenging for the HLT algorithms. New approaches have been studied to keep the HLT output rate manageable while maintaining thresholds low enough to cover physics analyses. The strategy mainly relies on porting online the ingredients that have been successfully applied in the offline reconstruction, thus allowing to move HLT selection closer to offline cuts. Improvements in HLT electron and photon definitions will be presented, focusing in particular on: updated clustering algorithm and the energy calibration procedure, new Particle-Flow-based isolation approach and pileup mitigation techniques, and the electron-dedicated track fitting algorithm based on Gaussian Sum Filter.

379

Performance of Tracking, b-tagging and Jet/MET reconstruction at the CMS High Level Trigger

Author: Mia Tosi[1]

[1] *Universita' degli Studi di Padova e INFN (IT)*

Corresponding Author: mia.tosi@cern.ch

The trigger systems of the LHC detectors play a crucial role in determining the physics capabilities of experiments. In 2015, the center-of-mass energy of proton-proton collisions will reach 13 TeV up to an unprecedented luminosity of 1e34 cm-2s-1. A reduction of several orders of magnitude of the event rate is needed to reach values compatible with detector readout, offline storage and analysis capabilities. The CMS experiment has been designed with a two-level trigger system: the Level-1 Trigger (L1T), implemented on custom-designed electronics, and the High Level Trigger (HLT), a streamlined version of the offline reconstruction software running on a computer farm. A software trigger system requires a trade-off between the complexity of the algorithms, the sustainable output rate, and the selection efficiency. With the computing power available during the 2012 data taking the maximum reconstruction time at HLT was about 200 ms per event, at the nominal L1T rate of 100 kHz. Tracking algorithms are widely used in the HLT in the object reconstruction through particle-flow techniques as well as in the identification of b-jets and lepton isolation. Reconstructed tracks are also used to distinguish the primary vertex, which identifies the hard interaction process, from the pileup ones. This task is particularly important in the LHC environment given the large number of interactions per bunch crossing: on average 25 in 2012, and expected to be around 40 in Run II with a large contribution from out-of-time particles. In order to cope with tougher conditions the tracking and vertexing techniques used in 2012 have been largely improved in terms of timing and efficiency in order to keep the physics reach at the level of Run-I conditions. We will present the performance of these newly developed algorithms, discussing their impact on the b-tagging performances as well as on the jet and met reconstruction.

380

Implementing a Domain Specific Language to configure and run LHCb Continuous Integration builds

Author: Marco Clemencic[1]

Co-author: Ben Couturier [1]

[1] *CERN*

Corresponding Author: marco.clemencic@cern.ch

The new LHCb nightly build system described at CHEP 2013 was limited by the use of JSON files for its configuration. JSON had been chosen as a temporary solution to maintain backward compatibility towards the old XML format by means of a translation function.

Modern languages like Python leverage on meta-programming techniques to enable the development of Domain Specific Languages (DSLs).

In this contribution we will present the advantages of such techniques and how they have been used to implement a DSL that can be used to both describe the configuration of the LHCb Nightly Builds and actually operate them.

Track 4 Session / 381

ROOT6: a quest for performance

Author: Danilo Piparo[1]

[1] *CERN*

Corresponding Author: danilo.piparo@cern.ch

The sixth release cycle of ROOT is characterised by a radical modernisation in the core software technologies the tookit relies on: language standard, interpreter, hardware exploitation mechanisms.
If on the one hand, the change offered the opportunity of consolidating the existing codebase, in presence of such innovations, maintaing the balance between full backward compatibility and software performance was not easy.
In this contribution we review the challenges and the solutions identified and implemented in the area of CPU and memory consumption as well as I/O capabilities in terms of patterns.
Moreover, we present some of the new ROOT components which are offered to the users to improve the performance of third party applications.

382

HLT configuration management system

Author: Vincenzo Daponte[1]

[1] *Universite de Geneve (CH)*

Corresponding Author: vincenzo.daponte@cern.ch

The CMS High Level Trigger (HLT) is implemented running a streamlined version of the CMS offline reconstruction software running on thousands of CPUs. The CMS software is written mostly in C++, using Python as its configuration language through an embedded CPython interpreter. The configuration of each process is made up of hundreds of "modules", organized in "sequences" and "paths". As an example, the HLT configurations used for 2011 data taking comprised over 2200 different modules, organized in more than 400 independent trigger paths. The complexity of the HLT configurations and the large number of configuration produced require the design of a suitable data management system. The present work describes the designed solution to manage the considerable number of configurations developed and to assist the editing of new configurations. The system is required to be remotely accessible and OS-independent as well as easly maintainable easy to use. To meet these requirements a three-layers architecture has been choosen. On top of the "ConfDB"

database a business logic manager has been introduced to handle the database operations, to perform the read and write rights check and to send a configuration to the user in a suitable format for the user interface. The graphical user interface (GUI) provides the features to display, modify and manage the configurations. The GUI design was carried out first by exposing paper sketches to the end-users and based on the their feedbacks a software mockup was implemented. At the end of the development process usability test will be carried out in order to measure the impact that the new GUI has on the development of configurations for the CMS-HLT.

383

CMS High Level Trigger Timing Measurements

Author: Clint Allan Richardson[1]

[1] *Boston University (US)*

Corresponding Author: clint.allan.richardson@cern.ch

The two-level trigger system employed by CMS consists of the Level 1 (L1) Trigger, which is implemented using custom-built electronics, and the High Level Trigger (HLT), a farm of commercial cpus running a streamlined version of the offline CMS reconstruction software. The operational L1 output rate of 100 kHz, together with the number of cpus in the HLT farm, imposes a fundamental constraint on the amount of time available for the HLT to process events. Exceeding this limit impacts the experiment's ability to collect data efficiently. Hence, there is a critical need to characterize the performance of the HLT farm as well as the algorithms run prior to startup in order to ensure optimal data taking. Additional complications arise from the fact that the HLT farm consists of multiple generations of hardware and there can be subtleties in machine performance. We will present our methods of measuring the timing performance of the CMS HLT, including the challenges of making such measurements. Results for the performance of various intel Xeon architectures from 2009-2014 and different data taking scenarios will also be presented.

384

Using DD4Hep through Gaudi for new experiments and LHCb

Authors: Anastasia Karachaliou[1]; Marco Clemencic[2]

[1] *National and Kapodistrian University of Athens (GR)*

[2] *CERN*

Corresponding Author: marco.clemencic@cern.ch

The LHCb Software Framework Gaudi is a C++ software framework for HEP applications used by several experiments.

Although Gaudi is extremely flexible and extensible, its adoption is limited by the lack of certain components that are fundamental for the software framework of an experiment, in particular a detector description framework, whose implementation is delegated to the adopters.

To enable future experiments to quickly adopt Gaudi, we integrated the DD4Hep toolkit in the existing software framework, and, as a proof of concept, we used it with the LHCb software applications, from simulation to reconstruction and analysis.

We will describe how the DD4Hep toolkit can be used by a new experiment, as well as how we can migrate an existing detector description framework to the new toolkit.

385

Improved interface for the LHCb Continuous Integration System

Author: Sofia Qirjazi[1]

Co-authors: Ben Couturier [2]; Marco Clemencic [2]

[1] *National and Kapodistrian University of Athens (GR)*

[2] *CERN*

Corresponding Authors: ben.couturier@cern.ch, marco.clemencic@cern.ch

The purpose of this paper is to describe the steps that led to an improved interface for LHCb's Nightly Builds Dashboard. The goal was to have an efficient application that meets the needs of both the project developers, by providing them with a user friendly interface, as well as those of the computing team supporting the system by providing them with a dashboard allowing for better monitoring of build job themselves. In line with what is already used by LHCb, the web interface is implemented with the Flask Python framework for future maintainability and code clarity. The Database chosen to host the data is the schema-less CouchDB, serving the purpose of flexibility in document form changes. To improve the user experience, we use JavaScript libraries such as JQuery.

386

Clusteralive

Author: Giuseppe Caruso[1]

Co-authors: Alberto Ciampa [1]; Antonino Formuso [1]; Silvia Arezzini [1]

[1] *INFN - Pisa*

Corresponding Author: silvia.arezzini@pi.infn.it

"CLUSTERALIVE"

"Clusteralive" is an integrated system developed in order to monitor and manage few important tasks in our HPC environment.
We have also other management systems, but now, with "Clusteralive"we can know immediately, just seeing our screen, if Clusters are up and running and we are sure that the most important functionality are well instanced.
"Clusteralive" is a php scripts suite able to monitor and perform tasks of automatic management about services, parameters and basic activities related in particular to computing nodes of the HPC clusters but also of the entire computing infrastructure.
At this link: http://farmsmon.pi.infn.it/clusteralive/monitoring.php you have a look at the specific application of "Clusteralive" dedicated to our principal projects.
In particular "Clusteralive" controls the cluster (dedicated to theoretical physics) called 'Zefiro' and funded by the SUMA project (SUper MAssive computing project, link: http://vh2.pi.infn.it/ a special project approved by Italian Research Ministery). Zefiro (2048 cores total: AMD Opteron 6380-2.5GHz) consists of 32 machines each one with 512 GB of RAM and 4 processors (16 cores for processors fo 64 cores total, grouped into 2 jobslot). Nodes are linked via Infiniband QDR connections operated with Mellanox IS5100 switch with 108 ports. The accesses are regulated by the IBM LSF (V.9) scheduler.
"Clusteralive" has also been recently extended to all the resources of the HPC in Pisa used for academic and industrial collaborations (more than 4000 computing cores total).
The monitoring system allows, via web browser, the view of essential informations about the status of each compute nodes and about a specific service and status of the entire HPC infrastructure.
For each computing node are displayed few informations like: the state (used / free / closed for maintenance), users that have assigned resources on it (used or reserved), information about the communication between the machines (ping) via Ethernet and via Infiniband, the percentage of used disk space, the used or reserved resources, the CPU load, the loked processes called zombie

(due to a termination or to a bug in the application), daemons for user's autentication (nscd / nslcd), communication daemons (sshd) and node status daemons (gmond / hsflowd).
For the entire cluster the status of jobs on the specific queues are shown , specifically the running/pending and suspended jobs end in last two cases, it is possible to know reasons
for the specific status. The monitoring system also shows the physically turned off and turned on computing nodes and the nodes with the file system unmounted.
The automatic management system, has been developed for the recovery of some services in case of their malfunction. Specifically, an automatic restart of stopped and blocked services and an automatic closing system for all compute nodes in case of malfunction of the filesystem which supports the entire cluster HPC has been implemented togheter a system of automatic recovery of all the services necessary for communication between computing nodes and user identification.
An additional feature is going to be implemented and will permit the automatic cleaning of processes (zombie) after an unexpected termination or at the end of specific job.

387

Managing competing elastic Grid and Cloud scientific computing applications using OpenNebula

Authors: Dario Berzano[1]; Massimo Masera[2]; Sara Vallero[2]; Stefano Bagnasco[3]; Stefano Lusso[4]

[1] *CERN*

[2] *Universita e INFN (IT)*

[3] *I.N.F.N. TORINO*

[4] *INFN-TO*

Corresponding Author: bagnasco@to.infn.it

Elastic cloud computing applications, i.e. applications that automatically scale according to computing needs, work on the ideal assumption of infinite resources. While large public cloud infrastructures may be a reasonable approximation of this condition, scientific computing centres like WLCG Grid sites usually work in a saturated regime, in which applications compete for scarce resources through queues, priorities and scheduling policies, and keeping a fraction of the computing cores idle to allow for headroom is usually not an option.
In our particular environment one of the applications (a WLCG Tier-2 Grid site) is much larger than all the others, so a possible strategy is to make it shrink and release resources when smaller, higher priority ones require them.
The implementation of said model in our infrastructure, based on the OpenNebula cloud stack, will be described and the very first operational experiences with a small number of strategies for timely allocation and release of resources will be discussed.
Such strategies, which aim to be non-invasive with respect to our most common class of applications, i.e. Grid jobs, include for example the tuning of virtual Worker Node parameters (number of cores and lifetime) to match the statistical distribution of job durations, or the balance between the amount of resources statically pinned to an application and the amount left available for competitive seizing by elastic applications.

388

Development of site-oriented Analytics for Grid computing centres

Authors: Andrew John Washbrook[1]; David Crooks[2]

Co-authors: David Britton [2]; Gang Qin [2]; Gareth Douglas Roy [2]; Samuel Cadellin Skipsey

[1] *University of Edinburgh (GB)*

[2] *University of Glasgow (GB)*

Corresponding Authors: andrew.washbrook@cern.ch, david.crooks@cern.ch

The field of analytics, the process of analysing data to visualise meaningful patterns and trends, has become increasingly important to a wide range of scientific applications as the volume and variety of accessible data available to process (so called Big Data) has significantly increased. There are a number of scalable analytic platforms and services which have risen in prominence (such as Elasticsearch) which enable unstructured data from numerous sources to be gathered, curated and visualised through a single extensible interface. There is ongoing work in the HEP community evaluating these tools, for example in the augmentation of system management at regional computing centres. In this context the provisioning of analytic solutions for computing sites pledging resources to the Worldwide LHC Computing Grid (WLCG) is an area of considerable interest.

Each Grid computing centre generates a wealth of monitoring data from multiple sources as part of their ongoing operations. These include system logging, Grid middleware services, LRMS scheduling information, network and storage utilisation and workload performance. This rich set of data is available for exploitation using analytics tools to enable post-facto diagnostics and a more comprehensive understanding of site systems, extending existing work on site monitoring. A site-oriented analytics portal would allow administrators to more easily leverage their available logging and monitoring data to determine the causes in variations in workload performance that may be unclear from a single data source.

In this study we will explore the components necessary for a WLCG site-oriented analytics platform. A corpus of relevant time-series based monitoring and logging data collected at two UK Grid computing centres (ECDF and Glasgow) will be categorised and stored. We will then explore the use of this data as part of a distributed analytics system. A necessary part of this work will be an examination of the appropriate level of visibility for different categories of site data. Furthermore, we will explore the extent to which machine learning techniques could be harnessed to provide predictive capability in error detection by using curated site data as a continuous training set.

This model is being developed with a particular focus on providing a solution for Grid computing centres rather than attempting to cover all data sources generated by a Virtual Organisation (VO). Such an approach is intended to complement analytic development from larger VOs (such as the LHC experiments) whilst benefiting smaller VOs who may not have the resources available to develop these types of tools. Based on this work we will look towards areas of best practice in developing systems of this nature.

Track 7 Session / 389

Integrated Monitoring-as-a-service for Scientific Computing Cloud applications using the Elasticsearch ecosystem

Authors: Andrea Guarise[1]; Dario Berzano[2]; Massimo Masera[3]; Sara Vallero[3]; Stefano Bagnasco[4]; Stefano Lusso[5]

[1] *Istituto Nazionale Fisica Nucleare (IT)*

[2] *CERN*

[3] *Universita e INFN (IT)*

[4] *I.N.F.N. TORINO*

[5] *INFN-TO*

Corresponding Author: svallero@to.infn.it

The INFN computing centre in Torino hosts a private Cloud, which is managed with the OpenNebula cloud controller. The infrastructure offers IaaS services to different scientific computing applications. The main stakeholders of the facility are a grid Tier-2 site for the ALICE collaboration at LHC, an interactive analysis facility for the same experiment and a separate grid Tier-2 site for the BES-III collaboration, plus an increasing number of other smaller tenants. The dynamic allocation of resources to tenants is partially automated. This feature requires detailed monitoring and accounting of the resource usage.

We set up a monitoring framework to inspect the site activities both in terms of IaaS and applications running on the hosted virtual instances. For this purpose we used the Elasticsearch, Logstash and Kibana stack. The infrastructure relies on an SQL database back-end for data preservation and to ensure flexibility to switch to a different monitoring solution if needed. The heterogeneous accounting information is transferred from the database to the Elasticsearch engine via a custom Logstash plugin. Each use-case is indexed separately in Elasticsearch and we setup a set of Kibana dashboards with pre-defined queries in order to monitor the relevant information in each case.
For the IaaS metering, we developed sensors for the OpenNebula API. The IaaS level information gathered through the API is sent to the MySQL database through an ad-hoc developed RESTful web service. Moreover, we have developed a billing system for our private Cloud, which relies on the RabbitMQ message queue for asynchronous communication to the database and on the ELK stack for its graphical interface.
Concerning the application level, we used the Root plugin TProofMonSenderSQL to collect accounting data from the interactive analysis facility. The BES-III virtual instances used to be monitored with Zabbix, as a proof of concept we also retrieve the information contained in the Zabbix database.
Finally, we have defined a model for monitoring-as-a-service, based on the tools described above, which the Cloud tenants can easily configure to suit their needs.
In this way we have achieved a uniform monitoring interface for both the IaaS and the scientific applications, mostly leveraging off-the-shelf tools.

390

A flexible and modular data format ROOT-based implementation for HEP

Authors: Domenico D'Urso[1]; Matteo Duranti[1]

Co-author: Paolo Zuccon [2]

[1] *Universita e INFN (IT)*

[2] *Massachusetts Inst. of Technology (US)*

Corresponding Author: domenico.d'urso@cern.ch

A flexible and modular data format implementation for HEP applications is presented.
Designed to face HEP data issues, the implementation is based on the CERN ROOT toolkit.
The design is aimed to create a data format as much as possible modular and easily upgradable and extendable.
Event informations are split into different files, that may contain different parts of the event (i.e. different sub-detectors) or different levels of abstraction (i.e. from raw information to high level quantities), in a fully transparent way for the final user.
Fully exploiting the ROOT "Friend" concept and the C++ inheritance, it has been possible to achieve the desired modularity and flexibility: the possibility to upgrade and extend without a full reprocessing, but also the simplicity of a standard ROOT-ple functionality (i.e. TTree::Draw()).
The file splitting in self-consistent TFile friends allows also a much more efficient distribution and upgrade of the data to the regional centers of a typical HEP experiment. It also gives the possibility to download and process only a small section of the event information, for example for subdetector studies or specific analyses, even on the user laptop.

391

pawgo: an interactive analysis workstation

Author: Sebastien Binet[1]

[1] *IN2P3/LAL*

Corresponding Author: sebastien.binet@cern.ch

pawgo: an interactive analysis workstation

Current interactive analysis toolkits usually leverage a Turing-complete general programming language, such as C++ or python, married with some kind of interpreter (*e.g.:* CINT or CLing) and a graphical user interface to present results (ROOT, matplotlib or Chaco.)
An obvious advantage of using a general programming language is that one can tap the full power of that language to build any analysis.
But conversely, users have to comply with the sometimes verbose ways of that language to describe "graphics" entities or express manipulations to be applied to these entities, letting much to be improved on the interactivity and reactivity fronts.

pawgo is a prototype re-investigating the use of a Domain Specific Language (DSL) - whith a nod to the old PAW and PAW++ applications.

This paper will first introduce the overall architecture of pawgo, a set of reusable Go libraries packaged as a portable application, and the tools it leverages (d3.js, WebSockets and RPC to name a few) to enable both remote and local work as well as batch execution mode.
Most importantly, the paper will describe the use of the grammar of graphics to overcome the impedance mismatch between general programming languages and graphics manipulations, thus easily allowing users to create new graphic objects, animations or styles.
Indeed, the grammar of graphics lets users and developers focus on the "what" to display, letting the underlying library deal with the minute details and ceremony of the "how" best to display.

Finally, the paper will present prospects and work ahead for pawgo to graduate from a prototype to a full-fledged analysis workstation.

392

Acceleration of ensemble machine learning methods using many-core devices

Author: Andrew John Washbrook[1]

Co-author: Wyeth Daniel [2]

[1] *University of Edinburgh (GB)*

[2] *Toshiba Medical Visualisation Systems Europe (TMVSE)*

Corresponding Author: andrew.washbrook@cern.ch

Multivariate training and classification methods using machine learning techniques are commonly applied in data analysis at HEP experiments. Despite their success in looking for signatures of new physics beyond the standard model it is known that some of these techniques are computationally bound when input sample size and model complexity are increased. Investigating opportunities for potential performance improvements is therefore of great importance if these techniques are to be used with the much larger data volumes expected from Run 2 operations at the Large Hadron Collider.

It has been previously shown that a large degree of algorithm parallelisation was possible for MLP-based artificial neural networks by the use of many-core devices such as GPUs. Improved scaling was observed in network complexity and qualitative performance gains were attainable through the simultaneous processing of multiple neural networks. Here we investigate how many-core devices

can be used to accelerate ensemble machine learning methods that are gaining traction in HEP data analysis.

We present a case study into the acceleration of decision forests using many-core devices in collaboration with Toshiba Medical Visualisation Systems Europe (TMVSE). TMVSE have developed software to process three-dimensional medical imaging data (such as CT or MRI scans), using automatic detection of anatomical landmarks defined on the skeleton, vasculature and major organs. Landmark detection underpins a semantic understanding of the medical data and thus has many diverse applications, for example, it facilitates rapid navigation to a named organ. TMVSE have applied ensemble machine learning methods such as classification by random decision forests to efficiently compute the bounding boxes of organs in processed image data volumes. It is important that their applications using this algorithm runs efficiently and quickly. After data preparation and optimisation the execution time is on average 4.5 seconds per volume with a sub-second processing time being desirable.

Using representative medical image data as input and pre-trained decision trees we will demonstrate how the decision forest classification method maps onto the GPU data processing model. It was found that a GPU-based version of the classification method resulted in over 130 times speed-up over a single-threaded CPU implementation with further improvements possible. We will outline the main optimisation steps undertaken to maximise GPU performance and detail how this was implemented using device profiling to evaluate thread occupancy and execution efficiency.

As this solution was developed to be context independent we will demonstrate how this work can be applied to a suitably formed HEP dataset to determine potential gains in event throughput and classifier discrimination. We will also explore how the advanced analysis techniques applied to automatic landmark detection in medical data can be applied to HEP dataset to achieve further increases in performance.

Track 7 Session / 393

SkyGrid - where cloud meets grid computing

Authors: Alexander Baranov[1]; Andrey Ustyuzhanin[1]

Co-authors: Dimitris Sarigiannis ; Konstantin Nikitin [2]

[1] *ITEP Institute for Theoretical and Experimental Physics (RU)*

[2] *Yandex*

Corresponding Author: a.baranov@cern.ch

Computational grid (or simply 'grid') infrastructures are powerful but restricted by several aspects: grids are incapable of running user jobs compiled with a non-authentic set of libraries and it is difficult to restructure grids to adapt to peak loads. At the same time if grids are not loaded with user-tasks, owners still have to pay for electricity and hardware maintenance. So a grid is not cheap and small/medium scientific experiments have difficulties working with such a computational model.

To address these inflexibility issues we present SkyGrid - a system that integrates cloud technologies into grid systems. By cloud technologies we mean mainly virtualization that allows both virtualization of user jobs and computational hardware. Virtualization of user jobs is performed by means of Docker - a lightweight Virtual Machine system. Virtualization of hardware is provided by YARN or similar platforms. SkyGrid unties users from the computational cluster architecture and configuration. Also the virtualization approach we propose enables some extreme cases like volunteer computing inside browsers by means of PNaCl technology and running jobs on super-computers. In this paper we present the requirements and architecture of SkyGrid, interfaces available to end-users and interfaces to other systems. Also we provide a description of a real case study of system usage for the SHiP experiment integrating private cloud resources from several universities as well as volunteer computing resources into a single computational infrastructure.

Track 2 Session / 394

Interpolation between multi-dimensional histograms using a new non-linear moment morphing method

Authors: Max Baak[1]; Robert Harrington[2]; Stefan Gadatsch[3]; Wouter Verkerke[3]

[1] *CERN*

[2] *University of Edinburgh*

[3] *NIKHEF (NL)*

Corresponding Author: stefan.gadatsch@cern.ch

In particle physics experiments data analyses generally use Monte Carlo (MC) simulation templates to interpret the observed data. These simulated samples may depend on one or multiple model parameters, such as a shifting mass parameter, and a set of such samples may be required to scan over the various parameter values. Since detailed detector MC simulation can be time-consuming, there is often a need to interpolate between the limited number of available MC simulation templates. Only several interpolation techniques exist for this. For example, the statistical tests widely used in particle physics, e.g. for the discovery of Higgs boson, rely critically on continuous and smooth parametric models that describe the physics processes in the data.

We present a new template morphing technique, moment morphing, for the interpolation between multi-dimensional distribution templates based on one or multiple model parameters. Moment morphing is fast, numerically stable, and is not restricted in the number of input templates, the number of model parameters or the number of input observables. For the first time, statistical tests may include the impact of a non-factorizable response between different model parameters, where varying one model parameter at a time is insufficient to capture the full response function.

395

Towards Reproducible Experiment Platform

Authors: Alexander Baranov[1]; Alexey Artemov[2]; Alexey Rogozhnikov[3]; Andrey Ustyuzhanin[1]; Egor Khairullin[3]; Tatiana Likhomanenko[4]

Co-authors: J Michael Williams [5]; Vladimir Gligorov [6]

[1] *ITEP Institute for Theoretical and Experimental Physics (RU)*

[2] *Yandex School of Data Analysis*

[3] *Yandex School of Data Analysis, Moscow*

[4] *National Research Centre Kurchatov Institute (RU)*

[5] *Massachusetts Inst. of Technology (US)*

[6] *CERN*

Corresponding Author: andrey.ustyuzhanin@cern.ch

Data analysis in fundamental sciences nowadays is essential process that pushes frontiers of our knowledge and leads to new discoveries. At the same time we can see that complexity of those analysis increases exponentially due to a) enormous volumes of datasets being analyzed, b) variety of techniques and algorithms one have to check inside a single analysis, c) distributed nature of research teams that requires special communication media for knowledge and information exchange between individual researchers. There is a lot of resemblance between techniques and problems arising in the areas of industrial information retrieval and particle physics. To address those problems we propose a Reproducible Experiment Platform (REP) - a software infrastructure to support a collaborative ecosystem for computational science. It is a Python-based solution for research teams that allows running computational experiments on big shared datasets, obtaining reproducible and repeatable results, and consistent comparisons of the obtained results. REP supports many data

formats including ROOT, allowing for easy integration with existing HEP software and analyses. We present some key features of REP based on case studies which include trigger optimization and physics analysis studies at the LHCb experiment, as well as an example case of applying the prototype of such a system in Information Retrieval research that led to a performance increase of two orders of magnitude.

Track 2 Session / 396

CMS Detector Description for Run II and Beyond

Author: Ianna Osborne[1]

Co-authors: Christopher Jones [1]; David Lange [2]; Elizabeth Sexton-Kennedy [1]; Gaelle Boudoul [3]; Giulio Eulisse [1]

[1] *Fermi National Accelerator Lab. (US)*

[2] *Lawrence Livermore Nat. Laboratory (US)*

[3] *Universite Claude Bernard-Lyon I (FR)*

Corresponding Author: boudoul@ipnl.in2p3.fr

CMS Detector Description (DD) is an integral part of the CMSSW software multithreaded framework. CMS software has evolved to be more flexible and to take advantage of new techniques, but many of the original concepts remain and are in active use. In this presentation we will discuss the limitations of the Run I DD model and changes implemented for the restart of the LHC program in 2015. Responses to Run II challenges and transition to multithreaded environment are discussed.

The DD is a common source of information for Simulation, Reconstruction, Analysis, and Visualisation, while allowing for different representations as well as specific information for each application. The DD model usage for CMS Magnetic field map description allows seamless access to variable field during the same run. Examples of the integration of DD in the GEANT4 simulation and in the reconstruction applications are provided.

397

Towards generic volunteer computing platform

Authors: Alexander Baranov[1]; Andrey Ustyuzhanin[1]; Konstantin Nikitin[2]

[1] *ITEP Institute for Theoretical and Experimental Physics (RU)*

[2] *Yandex School of Data Analysis, Moscow*

Corresponding Author: a.baranov@cern.ch

Computational power that is distributed amongst user hardware: laptops, PCs and even smartphones is enormous. It is not exceptional that volunteer computational networks provide computational power comparable to the power of the modern supercomputers. The problem is that utilization of those resources is difficult from volunteer (user) point of view as well as from computation provider (client) point of view. Users have to install special software, clients have to convert essential computational algorithms into special format and test it thoroughly.
Dibrocop project makes an attempt to use modern in-browser computation platforms to solve those problems. Modern browsers have special code paths for low overhead calculations, e.g. ASM.js by Firefox and PNaCl by Chromium. In contrast to SETI@Home and other public volunteer computing projects, our approach requires no software installation and starts as soon as you open provider's page in your web browser. It doesn't have any requirements for operating system or device capabilities.

In this paper we are going to share our experience in moving computational tasks for Monte Carlo event simulation and processing from regular GRID resources to ASM.js and PNaCl. Algorithms used for demonstration of Dibrocop are taken from newly developed SHiP experiment at CERN.

399

A Generic Framework for Rapid Development of OPC UA Servers

Author: Piotr Pawel Nikiel[1]

Co-authors: Benjamin Farnham [1]; Stefan Schlenker [1]; Viatcheslav Filimonov [2]

[1] *CERN*

[2] *B.P. Konstantinov Petersburg Nuclear Physics Institute - PNPI (*

Corresponding Authors: benjamin.farnham@cern.ch, piotr.nikiel@cern.ch

This paper describes a new approach for generic design and efficient development of OPC UA servers. Development starts with creation of a design file, in XML format, describing an object-oriented information model of the target system or device. Using this model, the framework generates an executable OPC UA server application, which exposes the per-design OPC UA address space, without the developer writing a single line of code. Furthermore, the framework generates skeleton code into which the developer adds the necessary logic for integration to the target system or device.

This approach allows both developers unfamiliar with the OPC UA standard, and advanced OPC UA developers, to create servers for the systems they are experts in while greatly reducing design and development effort as compared to developments based purely on COTS OPC UA toolkits. Higher level software may further benefit from the explicit OPC UA server model by using the XML design description as the basis for generating client connectivity configuration and server data representation. Moreover, having the XML design description at hand facilitates automatic generation of validation tools.

In this contribution, the concept and implementation of this framework is detailed along with examples of actual production-level usage in the detector control system of the ATLAS experiment at CERN and beyond.

400

Configuration Management and Infrastructure Monitoring using CFEngine and Icinga for real-time heterogeneous data taking environment

Authors: Jerome LAURET[1]; Michael Poat[2]; Wayne Betts[2]

[1] *BROOKHAVEN NATIONAL LABORATORY*

[2] *Brookhaven National Laboratory*

Corresponding Author: mpoat@bnl.gov

The STAR online computing environment is an intensive ever-growing system used for real-time data collection and analysis. Composed of heterogeneous and sometimes custom-tuned machine groups (Data Acquisition or DAQ computing, Trigger group, Sow Control or user-end data quality monitoring resources do not have the same requirements) the computing infrastructure was managed by manual configurations and inconsistently monitored by combinations of known tools

(Ganglia for monitoring for example) and home-made scripts sending Email reports to the (many) administrators of the diverse systems. This situation lead to configuration inconsistencies and an overload of repetitive tasks that needed to be passed from system groups to system groups with no global configuration and view of problems. Worst off, as the need to communicate between systems increased, globally securing the cyber-infrastructure was not possible to achieve and as more resources are moving closer to where the data is generated (due to event filtering and so-called "High Level Trigger farms), an agile policy-driven system ensuring consistency was seek.

STAR has narrowed down its strategy toward deploying a versatile and sustainable solution by leveraging the configuration management tool CFEngine to automate configurations along with the deployment of the Infrastructure monitoring system Icinga providing a dashboard view of the system' s health. In its first incarnation, Icinga 1 can be seen as a fork of the more commonly known tool Nagios monitoring while its version 2 is a core framework replacement and rewrite. Conjointly, CFEngine and Icinga have strengthened automation and sanctioning intricate development of the monitoring system. STAR has over 150 online systems spanning over four major sub-systems, each becoming critical during Runs. With a bird's eye view of each system, keeping track of the infrastructure becomes an ease. Similarly, STAR can now swiftly upgrade and modify the environment to our needs with ease as well as promptly react to cyber-security nodes whether it appears as a global patch for the Shellshock vulnerability or the reconfiguration of Secure shell. Though, as the DAQ resources do not need the same configurations than the user based resources (as one comparative example) but are all required to follow the same baseline, the infrastructure lay out is intricate and our strategy allows each system group (and sometimes machine) to be configured and monitored for its own particulars. Modular configuration is the key to consistency where differentiated plug-ins are distributed and configurations are updated ubiquitously. By creating a sustainable long term monitoring solution, the rate of failure detection has gone up from days to minutes, allowing rapid actions before the issue becomes a dire problem potentially causing loss of precious experimental data.

Alternatives to our configuration management choice of CFEngine such as Chef and Puppet have been proposed and evaluated. We have chosen an open source minimally dependent agile tool that is powerful and simple to use. Our monitoring tool has offered STAR administrators a crisp interface reporting each system's state simultaneously and allowing historical system status lookup.

In this report, we will make a brief reminder and comparison of the diverse configuration management systems available on the market and justify our choice by requirements and functionalities. We will discuss the details and procedures for developing practical uses with configuration management and infrastructure monitoring. Our extensions to the community's plugin have been re-integrated into the main development branch and we will provide examples of extension of Icinga and demonstrate through example the versatility of the framework for adding metric.

401

Recent advancements in user-job management with Ganga

Author: Ulrik Egede[1]

Co-authors: Alexander John Richards [1]; Johannes Elmsheuser [2]; Mark William Slater [3]; Matt Williams [4]; Patrick Haworth Owen [1]; Robert Andrew Currie [1]; Robert Fay [5]; William Lawrence Sutcliffe [1]

[1] *Imperial College Sci., Tech. & Med. (GB)*

[2] *Ludwig-Maximilians-Univ. Muenchen (DE)*

[3] *University of Birmingham (GB)*

[4] *University of Warwick (GB)*

[5] *University of Liverpool*

Corresponding Author: rcurrie@cern.ch

The Ganga project (http://cern.ch/ganga) has long been used by several experimental communities within HEP, most notably Atlas and LHCb. This talk describes the most recent developments in job

submission and management within Ganga with a focus on newly developed tools and features.
Ganga offers a powerful unified interface for submitting complex user-jobs to many different backends, this allows the user to exploit their available computing resources without the need to worry about implementation details.
The release of Ganga 6.1 in 2014/2015 will offer a new unified approach to file management within distributed analyses. One of the benefits of this is that it will allow for a user's data to be transferred directly between worker nodes and various storage solutions without the need for a centrally managed instance of Ganga.
In addition to this, Ganga now also supports a powerful new job queuing system. This system allows for users to perform job manipulation in a more automated way, providing them with more freedom over their job management.
Working closely with recent developments in DIRAC, Ganga will soon support bulk job submission to the grid allowing users to more easily exploit the currently available distributed resources. Alongside this, and with recent GridPP developments at Imperial College, Ganga offers an out of the box solution for experiments looking to work with vanilla DIRAC.
Ganga offers a modular codebase allowing for smaller experiments to easily adopt it with the ability to readily expand and develop additional modules to suit their particular requirements. Offering support for multiple HEP experiments Ganga development follows a "release early, release often" mantra. This allows for the development of new features whilst still focusing on stability and support for the existing user-base. Recent Ganga developments have also added a new Jenkins based unit-testing system to provide more comprehensive testing and debugging.

Track 3 Session / 402

POSIX and Object Distributed Storage system –performance Comparison studies and real-life usage in an experimental data taking context leveraging OpenStack/Ceph.

Authors: Jerome LAURET[1]; Michael Poat[2]; Wayne Betts[2]

[1] *BROOKHAVEN NATIONAL LABORATORY*

[2] *Brookhaven National Laboratory*

Corresponding Author: mpoat@bnl.gov

The STAR online computing environment is an intensive ever-growing system used for first-hand data collection and analysis. As systems become more sophisticated, they result in a more detailed dense collection of data output and inefficient limited storage systems have become an impediment to fast feedback to the online shift crews relying on data processing at near real-time speed. Motivation for a centrally accessible, scalable and redundant storage solution was led from an expansion of data growth and user data processing necessity. However, standard solutions (NAS, SAN ···GPFS) are "expensive"solutions and it became clear a balance of affordability and cost effectiveness was needed. Furthermore, the vast amount of sparse and distributed storage (disk attached to individual nodes) made the aggregation of the storage an attractive path. Yet, as shift crews are often composed of novice members sustaining the daily operations, providing a POSIX compliant interface within a standard namespace was, for this environment, a strong requirement for any retained solution.

The acquisition of reused hardware has offered STAR an opportunity to deploy a storage strategy at minimal to no cost. We have analyzed multiple open source object oriented cloud inspired solutions and a POSIX compliant storage system. Openstack Swift Object Storage and Ceph Object Storage were put to the test with single and parallel I/O tests emulating real world scenario for data processing and workflows. The Ceph file system storage, offering a POSIX compliant file system mounted similar to an NFS share which maintain owner/group permissions and historical lookup, was of particular interest as aligned with our requirements and was retained as our solution.

The Ceph storage system will allow scalability and redundancy without user or system interruption. Initial configuration for user setup is minimal while maintaining security and data integrity across

the entire cluster. A distributed storage system becomes and interconnected web of machines utilizing load balancing and redundancy, however if a subset of machines or all machines go down data loss is unlikely to occur. The expansion of a cloud storage system is trivial and can allow system administrators do add capacity and load balancing in a complete transparent manner.

In this report, we will review the necessary steps and requirements for our system, tools leveraged for performance testing and present comparative IO performance results between Swift and Ceph Object storage approach in similar context as well as a cross comparison of performance between the Object and POSIX compliant Ceph approaches. We will also discuss the benefit of a backbone private network for enhanced performance and fast communication between the distributed storage system components and scalability considerations for Meta-Data access and report on actual user's experience in a data taking environment. Finally, we will also present in great details the essential steps necessary to setup a Ceph storage cluster, including tweaks and hardware changes made along the way –the recipes we will present are not always easy to find in the manual and we hope our presentation will serve well the community's interest for distributed storage solutions.

403

The Geant4 physics validation repository

Author: Hans-Joachim Wenzel[1]

Co-author: Julia Yarba [1]

[1] *Fermi National Accelerator Lab. (US)*

Corresponding Author: wenzel@fnal.gov

We describe the Geant4 physics validation repository and the technology used
to implement it. The Geant4 collaboration regularly performs validation and regression tests where results
obtained with a new Geant4 version are compared to data obtained by various HEP experiments
or the results of previous releases. As the number of regularly performed validation tests
increases and the collection of results grows, storing them and making them available to the
collaborators and users community becomes a challenge. We decided to organize the materials
in one central repository and to make this data easily available via a web application.
The Physics Validation Repository consists of three components:
The first component is a PostgresSQL relational database that stores both simulated and experimental data
in form of images with meta data or raw data points. The meta data describes the test and lists the references
from where the experimental data for comparison was obtained from, as well as other parameters that describe the test
(e.g. beam particle, target material, etc.). The second component is a Java API which is based on the data
access object (DAO) design pattern and provides an abstract interface to the database. Finally a web application
based on Java Platform, Enterprise Edition (Java EE) is deployed on a GlassFish Application server. The web application
allows to interactively select and overlay compatible data, e.g. to compare the results of different releases,
different physics models or target materials in one plot.
The web application also provides security and authentication to grant access to groups of functions and data
that are internal to the Geant4 collaboration, e.g. viewing results from development releases, upload of new tests
and/or modification of selected tests.
We use the PrimeFaces JSF (Java Server Faces) Framework to create interactive, modern looking web interfaces and
the jfreechart java library or the Highcharts JavaScript library for plotting.

Primary Authors:
Wenzel, Hans
Yarba, Julia
On behalf of the Physics Validation Working Group of the Geant4 Collaboration

Track 5 Session / 405

Open Data and Data Analysis Preservation Services for LHC Experiments

Authors: Jake Cowton[1]; Jiri Kuncar[2]; Laura Rueda Garcia[2]; Pamfilos Fokianos[3]; Patricia Sigrid Herterich[4]; Sunje Dallmeier-Tiessen[4]; Tibor Simko[2]; Tim Smith[2]

[1] *University of Northumbria at Newcastle (GB)*

[2] *CERN*

[3] *National and Kapodistrian University of Athens (GR)*

[4] *Humboldt-Universitaet zu Berlin (DE)*

Corresponding Author: tim.smith@cern.ch

In this paper we present newly launched services for open data and for long-term preservation and reuse of high-energy-physics data analyses. We follow the "data continuum" practices through several progressive data analysis phases up to the final publication. The aim is to capture all digital assets and associated knowledge inherent in the data analysis process for subsequent generations, and to make a subset available rapidly to the public.

A data analysis preservation pilot study was launched in order to assess the usual workflow practices in LHC collaborations. Leveraging on synergies between ALICE, ATLAS, CMS and LHCb experiments, the analysed data was followed through various "analysis train" steps, from the initial capture and pre-selection of primary data, through several intermediate selection steps yielding more greatly reduced datasets, up to the final selection of N-tuples used for producing high-level plots appearing in scientific journal publications. Most of the analysis chain is kept strictly restricted within a given collaboration; only the final plots, presentations and papers are usually made public. It is therefore essential to handle access rights and embargo periods as part of the data life cycle.

The study revealed many similarities between collaborations, even though the variety of different practices existing in different groups within the collaborations make it hard to reproduce an analysis at a later time in a uniform way. One recurring problem underlined by the study was to ensure an efficient "knowledge capture" related to user code when the principal author of an analysis (e.g. a PhD student) leaves the collaboration later.

The pilot solution has been prototyped using the Invenio digital library platform which was extended with several data-handling capabilities. The aim was to preserve information about datasets, the underlying OS platform and the user software used to study it. The configuration parameters, the high-level physics information such as physics object selection, and any necessary documentation and discussions are optionally being recorded alongside the process as well. The metadata representation of captured assets uses the MARC bibliographic standard which had to be customised and extended in relation to specific analysis-related fields. The captured digital assets are being minted with Digital Object Identifiers, ensuring later referencability and citability of preserved data and software. Connectors were built in the platform to large-scale data storage systems (CASTOR, EOS, Ceph). In addition, to facilitate information exchange among concerned services, further connectors were built to

the internal information management systems of LHC experiments (e.g. CMS CADI), to the discussion platforms (e.g. TWiki, SharePoint), and to the final publication servers (e.g. CDS, INSPIRE) used in the process. Finally, the platform draws inspiration from the Open Archival Information System (OAIS) recommended practices in order to ensure long-term preservation of captured assets.

The ultimate goal of the analysis preservation platform is to capture enough information about the process in order to facilitate reproduction of an analysis even many years after its initial publication, permitting to extend the impact of preserved analyses through future revalidation and recasting services.

A related "open data" service was launched for the benefit of the general public. The LHC experimental collaborations are committed to make their data open after a certain embargo period. Moreover, the collaborations also release simplified datasets for the general public within the framework of the international particle physics masterclass program. The primary and reduced datasets that the collaborations release for public use are being collected within the CERN Open Data portal service, allowing any physicist or general data scientist to access, explore, and further study the data on their own.

The CERN Open Data portal offers several high-level tools which help to visualise and work with the data, such as an interactive event display permitting to visualise CMS detector events on portal web pages, or a basic histogram plotting interface permitting to create live plots out of CMS reduced datasets. The platform guides high-school teachers and students to online masterclasses to further explore the data and improve their knowledge of particle physics. A considerable part of the CERN Open Data portal was therefore devoted to attractive presentation and ease-of-use of captured data and associated information.

The CERN Open Data platform not only offers datasets and live tools to explore them, but it also preserves the software tools used to analyse the data. It notably offers the download of Virtual Machine images permitting users to start their own working environment in order to further explore the data; for this the platform uses CernVM based images prepared by the collaborations. Moreover, the CERN Open Data platform preserves examples of user analysis code, illustrating how the general public could write their own code to perform further analyses.

406

Statistical analysis of virtualization performance in high energy physics software

Authors: Marcos Seco Miguelez[1]; Victor Manuel Fernandez Albor[1]

Co-authors: Juan Jose Saborido Silva [1]; Ricardo Graciani Diaz [2]; Tomás Fernandez [3]; Victor Mendez Munoz [4]

[1] *Universidade de Santiago de Compostela (ES)*

[2] *University of Barcelona (ES)*

[3] *Santiago de Compostela University*

[4] *Universitat de Barcelona*

Corresponding Author: victormanuel.fernandez@usc.es

Cloud Computing is emerging today as the new approach followed by computing centres, since the flexibility the Cloud provides
is a powerful component to manage their resources. Through the use of virtualization, cloud promise to address with the same
shared set of physical resources a large user base with different needs. However, virtualization may induce significant performance
penalties for the demanding scientific computing workloads. This work presents an evaluation of the usefulness of the current
cloud computing services for scientific applications. The performace of a sample of High Energy Physics (HEP) software running
in a Kernel-based Virtual Machine (KVM) under different set-ups is analysed with the use of a multivariate analysis of variance
(MANOVA). While clouds are still changing, our results indicate that the current cloud services have to take into account the
different setups of Cores and Memory in order to get reasonable performances in HEP-Software.

Track 6 Session / 407

Integrating network and transfer metrics to optimize transfer efficiency and experiment workflows

Authors: Marian Babik[1]; Shawn Mc Kee[2]

Co-authors: Alessandra Forti [3]; Alessandro De Salvo [4]; Andreas Petzold [5]; Bruno Heinrich Hoeft [5]; Christopher John Walker [6]; Costin Grigoras [1]; Duncan Rand [7]; Enrico Mazzoni [8]; Fernando Lopez Munoz [9]; Frederique Chollet [10]; Guenter Duckeck [11]; HSIN YEN CHEN [12]; Ian Gable [13]; Ilija Vukotic [14]; Jason Zurawski [15]; Joel Closier [1]; John Shade [1]; Jorge Alberto Diaz Cruz [16]; Jose Flix Molina [17]; Kaushik De [18]; Michael OConnor [19]; Michail Salichos [1]; Simone Campana [1]; Tony Wildish [20]; Ulf Bobson Severin Tigerstedt [21]; Vincent Garonne [1]; Volodymyr Kotlyar [22]

[1] *CERN*

[2] *University of Michigan (US)*

[3] *University of Manchester (GB)*

[4] *Universita e INFN, Roma I (IT)*

[5] *KIT - Karlsruhe Institute of Technology (DE)*

[6] *University of London (GB)*

[7] *Imperial College Sci., Tech. & Med. (GB)*

[8] *Sezione di Pisa (IT)*

[9] *PIC (ES)*

[10] *Centre National de la Recherche Scientifique (FR)*

[11] *Ludwig-Maximilians-Univ. Muenchen (DE)*

[12] *ASGC*

[13] *University of Victoria (CA)*

[14] *University of Chicago (US)*

[15] *ESnet*

[16] *Fermi National Accelerator Lab. (US)*

[17] *Centro de Investigaciones Energ. Medioambientales y Tecn. - (ES*

[18] *University of Texas at Arlington (US)*

[19] *E*

[20] *Princeton University (US)*

[21] *Helsinki Institute of Physics (FI)*

[22] *Institute for Theoretical Physics, National Science Center Kharkov Institute of Physics and Technology, Nat. Acad. of Sciences of Ukraine (UA)*

Corresponding Author: shawn.mckee@cern.ch

The Worldwide LHC Computing Grid relies on the network as a critical part of its infrastructure and therefore needs to guarantee effective network usage and prompt detection and resolution of any network issues, including connection failures, congestion, traffic routing, etc. The WLCG Network and Transfer Metrics project aims to integrate and combine all network-related monitoring data collected by the WLCG infrastructure. This includes FTS monitoring information, monitoring data from the XRootD federation, as well as results of the perfSONAR tests. The main challenge consists of further integrating and analyzing this information so that it can be turned into actionable insight for optimization of data transfers and workload management systems of the LHC experiments.The presentation will include technical description of the WLCG network monitoring infrastructure as well as results of the analysis of the collected data. It will also highlight how results of this analysis can be used in order to improve efficiency of the WLCG computing activities.

408

Identifying and Localizing Network Problems using the PuNDIT Project

Authors: Constantine Dovrolis[1]; Shawn Mc Kee[2]

Co-authors: Danny Lee [1]; Jorge Batista [3]

[1] *Georgia Institute of Technology*

[2] *University of Michigan (US)*

[3] *U*

Corresponding Author: 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. However, some of these approaches do not scale to wide area inter-domain networks due to unavailability of data. 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 objective is to gather and analyze metrics from monitoring infrastructure to identify the signatures of possible problems and locate affected network links. A primary goal for PuNDIT is to convert complex network metrics into easily understood diagnoses in an automated manner.

PuNDIT is building upon the de-facto standard perfSONAR network measurement infrastructure deployed in Open Science Grid and the Worldwide LHC Computing Grid. The PuNDIT Team is working closely with the perfSONAR developers from ESnet and Internet2 to integrate PuNDIT as part of the perfSONAR Toolkit.

We will report on the project progress to-date in working with the OSG and WLCG communities and describe the current implementation architecture. We will also discuss some initial results, future plans and the project timeline.

Track 4 Session / 409

Evolution of CMS workload management towards multicore job support

Authors: Alan Malta Rodrigues[1]; Alison Mc Crea[2]; Antonio Perez-Calero Yzquierdo[3]; Eric Vaandering[1]; Farrukh Aftab Khan[4]; James Letts[2]; Jose Hernandez Calama[3]; Krista Larson[1]

[1] *Fermi National Accelerator Lab. (US)*

[2] *Univ. of California San Diego (US)*

[3] *Centro de Investigaciones Energ. Medioambientales y Tecn. - (ES*

[4] *National Centre for Physics (PK)*

Corresponding Author: antonio.perez.calero.yzquierdo@cern.ch

The successful exploitation of the multicore processor architectures available at the computing sites is a key element of the LHC distributed computing system in the coming era of the LHC Run 2. High-pileup complex-collision events represent a challenge for the traditional sequential programming in terms of memory and processing time budget. The CMS data production and processing framework has introduced the parallel execution of the reconstruction and simulation algorithms to overcome these limitations.

CMS plans to execute the data reconstruction and simulation as multicore processing yet supporting single-core processing for other tasks difficult to parallelize, such as user analysis. The CMS strategy for job management across the Grid thus aims at integrating single and multicore job scheduling. This is accomplished by scheduling multicore pilots with dynamic partitioning of the allocated resources, capable of running jobs with various core counts within a single pilot.

An extensive test programme has been conducted to enable multicore scheduling with the various local batch systems available at CMS sites. Scale tests have been run to optimize the scheduling

strategy and to ensure the most efficient use of the distributed resources. This contribution will present in detail the evolution of the CMS job management and resource provisioning systems in order to support this hybrid scheduling model, as well as its optimization and deployment, which will enable CMS to transition to a multicore production model by the restart of the LHC.

410

Monitoring data transfer latency in CMS computing operations

Authors: Daniele Bonacorsi[1]; Nicolo Magini[2]; Tony Wildish[3]

Co-authors: Andrea Sartirana [4]; Meric Taze [5]; Tommaso Diotalevi [1]

[1] *University of Bologna*

[2] *CERN*

[3] *Princeton University (US)*

[4] *Ecole Polytechnique (FR)*

[5] *Cukurova University (TR)*

Corresponding Authors: daniele.bonacorsi@bo.infn.it, nicolo.magini@cern.ch, tony.wildish@cern.ch

During the first LHC run, the CMS experiment collected tens of Petabytes of collision and simulated data, which need to be distributed among dozens of computing centres with low latency in order to make efficient use of the resources. While the desired level of throughput has been successfully achieved, it is still common to observe transfer workflows that cannot reach full completion in a timely manner due to a small fraction of stuck files which require operator intervention.

For this reason, in 2012 the CMS transfer management system, PhEDEx, was instrumented with a monitoring system to measure file transfer latencies, and to predict the completion time for the transfer of a data set. The operators can detect abnormal patterns in transfer latencies while the transfer is still in progress, and monitor the long-term performance of the transfer infrastructure to plan the data placement strategy.

Based on the data collected with the latency monitoring system, we present a study on the different factors that contribute to transfer completion time. As case studies, we analyze several typical CMS transfer workflows, such as distribution of collision event data from CERN or upload of simulated event data from the Tier-2 centres to the archival Tier-1 centres. For each workflow, we present the typical patterns of transfer latencies that have been identified with the latency monitor.

We identify the areas in PhEDEx where a development effort can reduce the latency, and we show how we are able to detect stuck transfers which need operator intervention. We propose a set of metrics to alert about stuck subscriptions and prompt for manual intervention, with the aim of improving transfer completion times.

Track 4 Session / 411

ROOT 6 and beyond: TObject, C++14 and many cores.

Author: Philippe Canal[1]

Co-authors: Axel Naumann [2]; Danilo Piparo [2]; Vasil Georgiev Vasilev [2]

[1] *Fermi National Accelerator Lab. (US)*

[2] *CERN*

Corresponding Author: philippe.canal@cern.ch

Following the release of version 6, ROOT has entered a new area of development. It will leverage the industrial strength compiler library shipping in ROOT 6 and its support of the C++11/14 standard, to significantly simplify and harden ROOT's interfaces and to clarify and substantially improve ROOT's support for multi-threaded environments.

This talk will also recap the most important new features and enhancements in ROOT in general, focusing on those allowed by the improved interpreter and better compiler support, including I/O for smart pointers, easier type safe access to the content of TTrees and enhanced multi processor support.

412

Data-driven estimation of neutral pileup particle multiplicity in high-luminosity hadron collider environments

Author: Federico Colecchia[1]

[1] *Brunel University London*

Corresponding Author: federico.colecchia@brunel.ac.uk

The contamination from low-energy strong interactions is a major issue for data analysis at the Large Hadron Collider, particularly with reference to pileup, i.e. to proton-proton collisions from other bunch crossings. With a view to improving on the performance of pileup subtraction in higher-luminosity regimes, particle weighting methods have recently been proposed whereby the weights are used to rescale the particle four-momentum vectors. We describe an algorithm based on a different approach that instead employs the weights to reshape the particle-level kinematic distributions in the data. We have applied the algorithm to the task of estimating the number of neutral pileup particles in different kinematic regions inside individual simulated events. Because of the simplicity and parallelisation potential of this technique, we foresee the possibility of using it in conjunction with existing methods at the reconstruction level in future high-luminosity hadron collider environments.

Track 8 Session / 413

GPU Accelerated Event-by-event Reweighting for a T2K Neutrino Oscillation Analysis

Author: Richard Calland[None]

Corresponding Author: rcalland@hep.ph.liv.ac.uk

The Tokai-to-Kamioka (T2K) experiment is a second generation long baseline neutrino experiment, which uses a near detector to constrain systematic uncertainties for oscillation measurements with its far detector. Event-by-event reweighting of Monte Carlo (MC) events is applied to model systematic effects and construct PDFs describing predicted event distributions. However when analysing simultaneously several data samples from both near and far detectors, the computational overhead can become a limiting factor in an oscillation analysis. Because reweighting each MC event is an independent process, it can be parallelized using graphics processing units (GPUs). For a recent T2K analysis, several bottlenecking calculations were offloaded onto NVIDIA GPUs using CUDA: the calculation of oscillation probabilities with matter effects and the evaluation of non-linear parameter responses with cubic splines. Individually, these methods achieved 40-180x speed-ups in standalone benchmarks. When implemented into the analysis software suite, an improvement of ˜20x was seen to the overall analysis. This talk will discuss the motivation and implementation of GPU reweighting into the T2K oscillation analysis, and prospects for further improvements using GPUs.

414

Alignment and calibration of Belle II tracking detectors

Author: Tadeas Bilka[1]

Co-authors: Claus Kleinwort [2]; Peter Kvasnicka [3]; Sergey Yaschenko [4]; Zdenek Dolezal [3]

[1] *Charles University in Prague*

[2] *Deutsches Elektronen Synchrotron (DESY)*

[3] *Charles University (CZ)*

[4] *DESY*

Corresponding Author: bilka@ipnp.troja.mff.cuni.cz

The Belle II experiment will start taking data in 2017. The SuperKEKB accelerator will deliver a factor 40 higher luminosity in comparison to its predecessor, KEKB, to acquire a 50 times larger data sample of B-$\bar{B}$ events. In order to manage higher occupancy and background, a new silicon vertex detector consisting of two inner layers of DEPFET pixel sensors surrounded by four layers of double-sided strip sensors will be installed. The high target performance of the detector, in particular vertex and momentum resolution, motivated the development of reliable alignment and calibration procedures. Combined with calibration of the upgraded drift chamber tracker, this task requires advanced software design and considerable computing resources. We present track-based alignment and calibration of the vertex and tracking detector using the global approach of the Millepede II tool in combination with an advanced track parametrization by the General Broken Lines (GBL) model. The procedure is implemented within the software framework of the Belle II experiment and GBL has also been integrated into the experiment-independent track-fitting toolkit GENFIT.

415

Geant4 Computing Performance Benchmarking and Monitoring

Authors: Andrea Dotti[1]; Gunter Folger[2]; Jim Kowalkowski[3]; Krzysztof Genser[3]; Marc Paterno[3]; Soon Yung Jun[4]; Victor Daniel Elvira[4]

[1] *SLAC National Accelerator Laboratory (US)*

[2] *CERN*

[3] *Fermilab*

[4] *Fermi National Accelerator Lab. (US)*

Corresponding Author: soon.yung.jun@cern.ch

Performance evaluation and analysis of large-scale computing applications is essential for optimizing the use of resources. As detector simulation is one of the most compute-intensive tasks and Geant4 is the simulation toolkit most widely used in contemporary high energy physics (HEP) experiments, it is important to monitor Geant4 through its development cycle for changes in computing performance and to identify problems and opportunities for code improvements. All development and public releases are being profiled with a set of applications that utilize different input event samples, physics lists, and detector configurations. Results from multiple benchmarking runs are compared to previous public and development reference releases to monitor CPU and memory usage. Observed changes are evaluated and correlated with code modifications. Besides the full summary of call stack data and memory footprint, a detailed call graph analysis is available to Geant4 developers for further analysis. The set of software tools used in the performance evaluation procedure,

both in sequential and multi-threaded modes, include FAST, IgProf and Open|Speedshop. The scalability of the CPU time and memory utilization of multi-threaded applications is evaluated by measuring event throughput and memory usage as a function of the number of threads for selected event samples. We will describe the procedure of Geant4 computing performance profiling and benchmarking and present recent results.

416

Optimisation of the usage of LHC and local computing resources in a multidisciplinary physics department hosting a WLCG Tier-2 centre

Authors: David Rebatto[1]; Francesco Prelz[1]; Franco Leveraro[2]; Leonardo Carminati[3]; Luca Vaccarossa[1]; Miguel Villaplana Perez[1]; Ruggero Turra[1]; Simone Michele Mazza[1]; Stefano Barberis[4]; laura perini[5]

[1] *Università degli Studi e INFN Milano (IT)*

[2] *Università degli Studi e INFN Milano (IT)*

[3] *INFN Sezione di Milano (INFN)*

[4] *Università degli Studi e INFN Milano*

[5] *Milan University and INFN*

Corresponding Author: miguel.villaplana.perez@cern.ch

We present the approach of the University of Milan Physics Department and the local unit of INFN to allow and encourage the sharing among different research areas of computing, storage and networking resources (the largest ones being those composing the Milan WLCG Tier-2 centre and tailored to the needs of the ATLAS experiment).

Computing resources are organised as independent HTCondor pools, with a global master in charge of monitoring them and optimising their usage. The configuration has to provide satisfactory throughput for both serial and parallel (multicore, MPI) jobs. A combination of local, remote and cloud storage options are available. The experience of users from different research areas operating on this shared infrastructure is discussed.

The promising direction of improving scientific computing throughput by federating access to distributed computing and storage also seems to fit very well with the objectives listed in the European Horizon 2020 framework for research and development.

Track 2 Session / 417

CosmoSIS: a system for MC parameter estimation

Authors: Joe Zuntz[1]; Marc Paterno[2]

Co-authors: Alessandro Manzotti [3]; Douglas Rudd [4]; Elise Jennings ; Jim Kowalkowski [2]; Saba Sehrish [1]; Sarah Bridle [5]; scott dodelson [6]

[1] *urn:Google*

[2] *Fermilab*

[3] *The University of Chicago*

[4] *University of Chicago*

[5] *UCL*

[6] *fermilab*

Corresponding Author: manzotti.alessandro@gmail.com

CosmoSIS [http://arxiv.org/abs/1409.3409] is a modular system for cosmological parameter estimation, based on Markov Chain Monte Carlo (MCMC) and related techniques. It provides a series of samplers, which drive the exploration of the parameter space, and a series of modules, which calculate the likelihood of the observed data for a given physical model, determined by the location of a sample in the parameter space. While CosmoSIS ships with a set of modules that calculate quantities of interest to cosmologists, there is nothing about the framework itself, nor in the MCMC technique, that is specific to cosmology. Thus CosmoSIS could be used for parameter estimation problems in other fields, including HEP.

This presentation will describe the features of CosmoSIS and show an example of its use outside of cosmology. It will also discuss how collaborative development strategies differ between two different communities: that of HEP physicists, accustomed to working in large collaborations, and that of cosmologists, who have traditionally not worked in large groups. For example, because there is no collaboration to enforce a language choice, the framework supports programming in multiple languages. Additionally, since scientists in the cosmology community are used to working independently, a system was needed for helping ensure that proper attribution is given to authors of contributed algorithms.

418

Exploring Two Approaches for an End-to-End Scientific Analysis Workflow

Author: Saba Sehrish[1]

Co-authors: Jim Kowalkowski [2]; Marc Paterno [2]

[1] *urn:Google*

[2] *Fermilab*

Corresponding Author: paterno@fnal.gov

The scientific discovery process can be advanced by the integration of independently-developed programs run on disparate computing facilities into coherent workflows usable by scientists who are not experts in computing. For such advancement, we need a system which scientists can use to formulate analysis workflows, to integrate new components to these workflows, and to execute different components on resources that are best suited to run those components. In addition, we need to monitor the status of the workflow as components get scheduled and executed, and to access the intermediate and final output for visual exploration and analysis. Finally, it is important for scientists to be able to share their workflows with collaborators.

We are involved with a project to develop such an analysis framework for the Large Synoptic Survey Telescope (LSST) Dark Energy Science Collaboration (DESC). Following upon the development of several detailed use cases for LSST DESC, we have been working on two approaches for the framework; the first one is based on the use and extension of Galaxy, a web-based portal for biomedical research, and the second one is based on a programming language, Python. There are benefits to each approach as we discovered while implementing one example use case. Both approaches allow scientists to run complicated workflows that involve the use of a

variety of computational resources (including grid resources,
supercomputing resources at NERSC, and local compute nodes) for the
execution of workflows on simulations of LSST images.
Adding a new application in the Python-based workflow description is straight forward,
however,
adding new applications through the Galaxy interface requires expert knowledge of
the Galaxy system and interaction with Galaxy infrastructure.

In this paper, we present a brief description of the two approaches,
describe the kinds of extensions to the Galaxy system we have found necessary
in order to support the wide variety of scientific analysis in the
cosmology community, and discuss how similar efforts might be of benefit
to the HEP community.

419

Integration of XRootD into the cloud infrastructure for ALICE data analysis

Authors: Andrey Zarochentsev[1]; Dmitrii Batkovich[2]; Mikhail Kompaniets[1]; Oksana Shadura[3]; Volodymyr Yurchenko[None]

Co-author: Eygene Ryabinkin [4]

[1] *St. Petersburg State University (RU)*

[2] *St.Petersburg State University*

[3] *National Technical Univ. of Ukraine "Kyiv Polytechnic Institute*

[4] *National Research Centre Kurchatov Institute (RU)*

Corresponding Author: rea@grid.kiae.ru

Cloud technologies allow easy load balancing between different tasks and projects. From the viewpoint of the data analysis in the ALICE experiment, cloud allows to deploy software using Cern Virtual Machine (CernVM) and CernVM File System (CVMFS), to run different (including outdated) versions of software for long term data preservation and to dynamically allocate resources for different computing activities, e.g. grid site, ALICE Analysis Facility (AAF) and possible usage for local projects or other LHC experiments.

We present a cloud solution for Tier-3 sites based on OpenStack and CEPH distributed storage with an integrated XRootD based storage element (SE). One of the key features of the solution is that CEPH has been used as a backend for Cinder Block Storage service for OpenStack and in the same time as a storage backend for XRootD, with redundancy and availability of data preserved by CEPH settings.

For faster and easier OpenStack deployment the PackStack solution, which is based on the Puppet configuration management system, is applied. CEPH installation and configuration chains are structured, converted to Puppet manifests and integrated into Packstack. This solution can be easily deployed, maintained and used even in small groups with limited computing resources and small organisations which usually have lack of IT support.
The proposed infrastructure has been tested on two different clouds (SPbSU \& BITP) and integrates succesfully with the ALICE data analysis model.

The present work is supported in part by Saint-Petersburg State University research grants 11.38.66.2012 and 11.38.197.2014.

420

Deep Integration: Python in the Cling World

Author: Wim Lavrijsen[1]

[1] *Lawrence Berkeley National Lab. (US)*

Corresponding Author: wim.lavrijsen@cern.ch

The language improvements in C++11/14 greatly reduce the amount of boilerplate code required and allow resource ownership to be clarified in interfaces. On top, the Cling C++ interpreter brings a truly interactive experience and real dynamic behavior to the language. Taken together, these developments bring C++ much closer to Python in ability, allowing the combination of PyROOT/cppyy and Cling to integrate the two languages on a new level. This paper describes the current state of the art, including cross-language callbacks, automatic resource management, software transactional memory, automatic template instantiations, and the ability to use Python from Cling.

Track 2 Session / 421

Breaking the Silos: The art Documentation Suite

Author: Robert Kutschke[1]

Co-author: Marc Paterno [2]

[1] *Femilab*

[2] *Fermilab*

Corresponding Author: kutschke@fnal.gov

The art event processing framework is used by almost all new experiments at Fermilab, and by several outside of Fermilab. All use art as an external product in the same sense that the compiler, ROOT, Geant4, CLHEP and boost are external products. The art team has embarked on a campaign to document art and develop training materials for new users. Many new users of art have little or no knowledge of C++, software engineering, build systems or the many external packages used by art or their experiments, such as ROOT, CLHEP, HEPPDT, and boost. To effectively teach art requires that the training materials include appropriate introductions to these topics as they are encountered. Experience has shown that simply refering readers to the existing native documentation does not work; too often a simple idea that they need to understand is described in a context that presumes prerequisites that are unimportant for a beginning user of art.
There is the additional complication that the training materials must be presented in a way that does not presume knowledge of any of the experiments using art. Finally, new users of art arrive at random times throughout the year and the training materials must allow them to start to learn art at any time. This presentation will explain the strategies adopted by the art team to develop a documentation suite that complies with these boundary conditions. It will also show the present status of the documentation suite, including feedback we have received from pilot users.

Track 2 Session / 422

ALFA: The new ALICE-FAIR software framework

Author: Mohammad Al-Turany[1]

Co-authors: Peter Hristov [1]; Pierre Vande Vyvre [1]; Predrag Buncic [1]; Thorsten Sven Kollegger [2]; Volker Lindenstruth [3]

[1] *CERN*

[2] *GSI - Helmholtzzentrum fur Schwerionenforschung GmbH (DE)*

[3] *Johann-Wolfgang-Goethe Univ. (DE)*

Corresponding Author: mohammad.al-turany@cern.ch

The commonalities between the ALICE and FAIR experiments and their computing requirements lead to the development of large parts of a common software framework in an experiment independent way. The FairRoot project has already shown the feasibility of such an approach for the FAIR experiments and extending it beyond FAIR to experiments at other facilities. The ALFA framework is a joint development between ALICE Online-Offline (O2) and FairRoot teams.

ALFA is designed as a flexible, elastic system, which balances reliability and ease of development with performance using multi-processing and multi-threading. A message-based approach has been adopted; such an approach will support the use of the software on different hardware platforms, including heterogeneous systems. Each process in ALFA assumes limited communication and reliance on other processes. Such a design will add horizontal scaling (multiple processes) to "vertical scaling"provided by multiple threads to meet computing and throughput demands.
ALFA does not dictate any application protocols. Potentially, any content-based processor or any source can change the application protocol. The framework supports different serialization standards for data exchange between different hardware and software languages. The concept and design of this new framework as well as the already implemented set of utilities and interfaces will be presented.

Track 5 Session / 423

A Comparative Analysis of Event Processing Frameworks used in HEP

Author: Elizabeth Sexton-Kennedy[1]

[1] *Fermi National Accelerator Lab. (US)*

Corresponding Author: sexton@fnal.gov

Today there are many different experimental event processing frameworks in use by running or about to be running experiments. This talk will compare and contrast the different components of these frameworks and highlight the different solutions chosen by different groups. In the past there have been attempts at shared framework projects for example the collaborations on the BaBar framework (between BaBar, CDF, and CLEO), on the Gaudi framework (between LHCb and ATLAS), on AliROOT/FairROOT (between Alice and GSI/Fair), and in some ways on art (Fermilab based experiments) and CMS'framework. However, for reasons that will be discussed, these collaborations did not result in common frameworks shared among the intended experiments. Though importantly, two of the resulting projects have succeeded in providing frameworks that are shared among many customer experiments: Fermilab's art framework and GSI/Fair's FairROOT. Interestingly, several projects are considering remerging their frameworks after many years apart. I'll report on an investigation and analysis of these realities. With the advent of the need for multi-threaded frameworks and the scarce available manpower, it is important to collaborate in the future; however it is also important to understand why previous attempts at multi-experiment frameworks either worked or didn't work.

Track 1 Session / 425

Online/Offline reconstruction of trigger-less readout in the R3B experiment at FAIR

Author: Dmytro Kresan[1]

Co-authors: Denis Bertini [2]; Florian Uhlig [1]; Mohammad Al-Turany [3]

[1] *GSI - Helmholtzzentrum fur Schwerionenforschung GmbH (DE)*

[2] *GSI Darmstadt*

[3] *CERN*

Corresponding Author: d.kresan@gsi.de

The R3B (Reactions with Rare Radioactive Beams) experiment is one of the planned experiments at the future FAIR facility at GSI Darmstadt. R3B will cover experimental reaction studies with exotic nuclei far off stability, thus enabling a broad physics programs with rare-isotope beams with emphasis on nuclear structure and dynamics. Several different detection subsystems as well as sophisticated DAQ system and data-analysis software are being developed for this purpose.
The data analysis software for R3B is based on FairRoot framework and called R3BRoot. R3BRoot is being used for simulation and detector design studies for the last few years. Recently, it was successfully used directly with the data acquisition and for the analysis of the R3B test beam-time in April 2014. For the future beam times the framework has to deal with the free streaming readout of the detectors. The implementation within R3BRoot to fulfill this trigger-less run mode will be presented as well as the set of tools developed for the online reconstruction and quality assurance of the data during the run.

Track 2 Session / 426

MiniAOD: A new analysis data format for CMS

Authors: Carl Vuosalo[1]; David Lange[2]

[1] *University of Wisconsin (US)*

[2] *Lawrence Livermore Nat. Laboratory (US)*

Corresponding Author: carl.vuosalo@cern.ch

The CMS experiment has developed a new analysis object format (the “mini-AOD”) targeted to be less than 10% of the size of the Run 1 AOD format. The motivation for the Mini-AOD format is to have a small and quickly derived data format from which the majority of CMS analysis users can perform their analysis work. This format is targeted at having sufficient information to serve about 80% of CMS analysis, while dramatically simplifying the disk and I/O resources needed for analysis. This improvement should bring substantial improvements in resource needs and turn-around time for analysis applications. Such large reductions were achieved using a number of techniques, including defining light-weight physics object candidate representations, increasing transverse momentum thresholds for storing physics-object candidates, and reduced numerical precision when it is not required at the analysis level. In this contribution we discuss the critical components of the mini-AOD format, our experience with its deployment and the planned physics analysis model for Run 2 based on the mini-AOD.

427

Evolution of the Open Science Grid Application Software Installation Service (OASIS)

Authors: Brian Paul Bockelman[1]; John Hover[2]; John Steven De Stefano Jr[3]; Jose Caballero Bejar[3]; Robert Quick[4]; Scott Werner Teige[5]; Vince Neal[6]

[1] *University of Nebraska (US)*

[2] *Brookhaven National Laboratory (BNL)-Unknown-Unknown*

[3] *Brookhaven National Laboratory (US)*

[4] *Indiana University*

[5] *Indiana University (US)*

[6] *University of Indiana*

Corresponding Authors: brian.bockelman@cern.ch, jose.caballero@cern.ch

The Open Science Grid Application Software Installation Service (OASIS)
provides an application installation service for Open Science Grid (OSG) virtual organizations (VOs) built on top of the CERN Virtual Machine File System (CVMFS).
This paper provides an overview and progress report of the OASIS service, which has been in production for over 18 months.
OASIS can be used either directly, as a service run by OSG Operations, or as a standalone software product.
In either use case, the VOs' files feed into a network of Stratum One replication servers, along with other partners within a federation.
The OASIS network has the flexibility of providing pure CVMFS repositories, the OSG-run central OASIS repository, or VO-run OASIS repositories.
Depending on their needs, VOs can integrate into OASIS at several different points, leading to a more complex trust model than that of the standard CVMFS service.

This paper describes the evolution of usage and implementation of OASIS over the past 18 months, as well as its operational successes, evolutions and issues.
Topics such as monitoring and auditing of the distributed content, handling of cryptographic keys, and establishing technical and organizational trust relationships are discussed.

Track 8 Session / 428

Detector Simulation On Modern Coprocessors

Authors: Philippe Canal[1]; Soon Yung Jun[1]

Co-authors: Andrei Gheata [2]; Federico Carminati [2]; Georgios Bitzes [3]; Guilherme Lima [4]; Johannes Christof de Fine Licht [5]; John Apostolakis [2]; Laurent Duhem ; Marilena Bandieramonte [6]; Mihaly Novak [2]; Oksana Shadura [7]; Raman Sehgal [8]; Rene Brun [2]; Sandro Christian Wenzel [2]; Victor Daniel Elvira [1]

[1] *Fermi National Accelerator Lab. (US)*

[2] *CERN*

[3] *National and Kapodistrian University of Athens (GR)*

[4] *FermiLab (US)*

[5] *University of Copenhagen (DK)*

[6] *Universita e INFN (IT)*

[7] *National Technical Univ. of Ukraine "Kyiv Polytechnic Institute*

[8] *Bhabha Atomic Research Centre (IN)*

Corresponding Author: philippe.canal@cern.ch

The recent prevalence of hardware architectures of many-core or accelerated processors opens opportunities for concurrent programming models taking advantages of both SIMD and SIMT architectures. The Geant Vector Prototype has been designed both to exploit the vector capability of main stream CPUs and to take advantage of Coprocessors including NVidia's GPU and Intel Xeon Phi. The characteristics of each of those architectures are very different in term of the vectorization depth, parallelization needed to achieve optimal performance or memory access latency and speed. Between each platforms the number of individual tasks to be processed 'at once'for efficient use of the hardware varies sometimes by an order of magnitude.
The granularity of the code executed may also need to be dynamically adjusted.
An additional challenge is to avoid the code duplication often inherent to supporting heterogeneous platforms. We will present the challenges, solutions

and resulting performance of running an end to end detector simulation concurrently on a main stream CPU and a coprocessor and detail the broker implementation bridging the disparity between the two architectures. The impacts of task decomposition, vectorization, efficient sampling techniques and data look-up using track level parallelism will be also evaluated on vector and massively parallel architectures.

Track 2 Session / 429

Simulation and Reconstruction Upgrades for the CMS experiment

Author: David Lange[1]

[1] *Lawrence Livermore Nat. Laboratory (US)*

Corresponding Author: david.lange@cern.ch

Over the past several years, the CMS experiment has made significant changes to its detector simulation and reconstruction applications motivated by the planned program of detector upgrades over the next decade. These upgrades include both completely new tracker and calorimetry systems and changes to essentially all major detector components to meet the requirements of very high pileup operations. We have untaken an ambitious program to implement these changes into the Geant4-based simulation and reconstruction applications of CMS in order to perform physics analyses to both demonstrate the improvements from the detector upgrades and to influence the detector design itself. In this presentation, we will discuss our approach to generalizing much of the CMS codebase to efficiently and effectively work for multiple detector designs. We will describe our computational solutions for very high pileup reconstruction and simulation algorithms, and show results based on our work.

430

Tracker software for Phase-II CMS

Author: Gaelle Boudoul[1]

[1] *Universite Claude Bernard-Lyon I (FR)*

Corresponding Author: boudoul@ipnl.in2p3.fr

The CMS experiment is has multi-faceted detector upgrade program planned over the next decade. The silicon tracker system plans an improved pixel detector for 2017 and proposes an entirely new tracker for the high-luminosity LHC run. In this presentation, we discuss the tools developed and used in the design, simulation and reconstruction of the upgraded tracker including completely new trigger functionalities. We discuss how we adapted the standard CMS tracker software to achieve excellent physics and technical performance results in each of the upgraded tracker scenarios, and our implementation of the track trigger simulation. We discuss our results and will demonstrate how the results of our simulation and reconstruction work have impacted the overall program of the upgraded tracker designs.

431

Use of containerisation as an alternative to full virtualisation in grid environments.

Authors: Robin Eamonn Long[1]; Roger Jones[1]

[1] *Lancaster University (GB)*

Corresponding Author: roger.jones@cern.ch

Virtualisation is a key tool on the grid. It can be used to provide varying work environments or as pat of a cloud infrastructure. Virtualisation itself carries certain overheads that decrease the performance of the system through requiring extra resources to virtualise the software and hardware stack, and CPU-cycles wasted instantiating or destroying virtual machines for each job. With the rise and improvements in containerisation, where only the software stack is kept separate, and no hardware or kernel virtualisation is used, their is scope for speed improvements and efficiency increases over standard virtualisation. We compare containerisation and virtualisation, including a comparison against bare-metal machines as a benchmark.

432

The evolving grid paradigm and code "tuning" for modern architectures- are the two mutually exclusive?

Authors: Robin Eamonn Long[1]; Roger Jones[1]

[1] *Lancaster University (GB)*

Corresponding Author: roger.jones@cern.ch

With the data output from the LHC increasing, many of the LHC experiments have made significant improvements to their code to take more advantage of the underlying CPU architecture and advanced features. With the grid environment changing to heavily include virtualisation and cloud services, we look at whether these two systems
can be compatible, or whether improvements in code are lost through virtualisation. This is done by looking at the efficiency increases achieved in the latest iterations of ATLAS code and testing it within various grid paradigms.

434

ConDB - Conditions Database for Particle Physics Experiments

Author: Igor Mandrichenko[1]

Co-authors: Jonathan Paley [2]; Vladimir Podstavkov [1]

[1] *FNAL*

[2] *Argonne National Laboratory*

Corresponding Author: ivm@fnal.gov

Abstract: Conditions or calibration data are an important part of High Energy experiments. This kind of data is typically organized in terms of intervals of validity that require a special type of database table schema and API structure. At Fermilab we have designed and developed ConDB, a general tool to store, manage and retrieve conditions data organized into validity intervals in a database. ConDB is highly scalable, has been proven to be capable of storing data for more than million channels, has many additional features such as data compression, a direct access API and a scalable high performance web interface. We present the conceptual design, product features, database representation and system architecture of this very powerful tool.

Track 6 Session / 435

ATLAS Public Web Pages: Online Management of HEP External Communication Content

Author: Steven Goldfarb[1]

Co-authors: Abha Eli Phoboo [2]; Claudia Marcelloni De Oliveira [3]; Kate Shaw [4]

[1] *University of Michigan (US)*

[2] *University of Manchester (GB)*

[3] *University of Birmingham (GB)*

[4] *INFN Gruppo Collegato di Udine and ICTP Trieste*

Corresponding Author: steven.goldfarb@cern.ch

The ATLAS Education and Outreach Group is in the process of migrating its public online content to a professionally designed set of web pages built on a Drupal-based content management system. Development of the front-end design passed through several key stages, including audience surveys, stakeholder interviews, usage analytics, and a series of fast design iterations, called sprints. Implementation of the web site involved application of the html design using Drupal templates, refined development iterations, and the overall population of the site with content.

We present the design and development processes and share the lessons learnt along the way, including the results of the data-driven discovery studies. We also demonstrate the advantages of selecting a back-end supported by content management, with a focus on workflow. Finally, we discuss usage of the new public web pages to implement Outreach strategy through implementation of clearly presented themes, consistent audience targeting and messaging, and the enforcement of a well-defined visual identity.

436

High Speed Fault Tolerant Secure Communication for Muon Chamber using FPGA based GBTx Emulator

Authors: Suman Sau[1]; Swagata Mandal[2]

Co-authors: Amlan Chakrabarty [3]; Jogender Saini [2]; Subhasis Chattapadhya [2]

[1] *Calcutta University*

[2] *VECC,Kolkata*

[3] *Cacutta University*

Corresponding Author: sumansau@yahoo.co.in

The Compressed Baryonic Matter (CBM) experiment is a part of the Facility for Antiproton and Ion Research (FAIR) in Darmstadt at the GSI. This experiment will examine heavy-ion collisions in fixed target geometry and will be able to measure hadrons, electrons and muons. Muon Chamber (MUCH) is used to detect low momentum muons in an environment of high particle densities. Basic read out chain of MUCH consists of MUCH-XYTER (Front End Electronics), Gigabit Transceiver (GBTx), Data Processing Board (DPB), First level event selector interface board (FLIB) 1. MUCH- XYTER is a self-triggered Application-Specific Integrated Circuit (ASIC), which is directly connected to the detector and send the receive data to the GBTx through serial peripheral interface (SPI) like E-link. GBTx will be connected to the DPB through 4.8 Gbps optical link. In between DPB and FLIB there will be 10 Gbps optical link and at the end FLIB will be connected to the data acquisition system through Peripheral Component Interconnect Express (PCIe).
As a part of the implementation of read out chain of MUCH in India, we have implemented FPGA emulator of GBTx 2. GBTx is a radiation tolerant ASIC that can be used to implement multipurpose high

speed bidirectional optical links for high energy physics (HEP) experiments. It is developed by CERN. It consists of packet generator, scrambler-descrambler, encoder-decoder, interleaver-deinterleaver, gearbox, and transceiver.Packet generator is used to generate test pattern (either static or dynamic) to test the optical link. Scrambler is used to scramble the data so that clock can be recovered properly in the receiver side. In the encoder block single error correcting (15, 11) Reed-Solomon (RS) encoding is used to mitigate error occurred in the communication channel due to radiation. Interleaver is used to simply interleave the encoded data to make the coded data more robust against burst error. Gearbox is used to divide the 120 bit data frame into three 40 bit wards. These 40 bit wards are transmitted through multi gigabit transceiver (MGT) and reference clock used for MGT is 120 MHz. In the receiver side apart from the predefined block one frame aligner block is also used. Frame aligner is used to detect header in the data frame properly in the receiver side. GBTx will be used in highly irradiated area and more prone to be affected by multi-bit error. To mitigate this effect instead of single bit error correcting (15, 11) RS code we have used two bit error correcting (15, 7) BCH code 3. It will increase the redundancy, which in turn increases the reliability in the coded data. So the coded data will be less prone to be affected by noise due to radiation. Normally in the wired communication between any two fixed points there will be no such security issue. But when multiple stations will be used for long distance communication the question of security will come into play. So to make the data transmitting through optical fiber more secure, we use advanced encryption standard (AES) (a symmetric key cryptography) 4 is used after channel coding. The AES is a specification for the encryption of electronic data established by the U.S. National Institute of Standards and Technology (NIST) in 2001. AES encryption algorithm acts as a block cipher. The data will be always 128 bit block but the key size can be select from 128 bit, 192 bit or 256 bit. Here for testing purpose we used AES- 128 and AES- 256. In this paper our key contributions can be summarized as follows:
• **Implementation of multibit error correcting BCH code to make optical communication more robust against error due to radiation.**
• **Implementation of AES along with GBTx to make the data communication more secure.**
We have implemented GBTx emulator on two Xilinx Kintex-7 boards (KC705). One will act as transmitter and other will act as receiver and they are connected through optical fiber using small form-factor pluggable (SFP) port. We have performed the run-time verification of the system using Xilinx Chipscope Pro Analyzer and also measured the resource utilization, throughput and power utilization of the implemented design.

References:
1. FLES/DAQ summary and Outlook, 24th CBM week, Walter F.J. Müller, FAIR, Darmstadt
2. S. Baron, M. Barros Marin, "GBT-FPGA user guide,"Version 1.00.
3. Error Detection and Correction using BCH code, Hank Wallance, 2001.
4. A.M Deshpande, M.S Deshpande, D.N Kayatanavar, "FPGA implementation of AES encryption and decryption,"International Conference on Control, Automation, Communication and Energy Conservation (INCACEC), 2009.

437

Redundant Web Services Infrastructure for Data Intensive and Interactive Applications

Author: Igor Mandrichenko[1]

Co-author: Vladimir Podstavkov [2]

[1] *Fermilab*

[2] *FNAL*

Corresponding Author: ivm@fnal.gov

RESTful web services are popular solution for distributed data access and information management. Performance, scalability and reliability of such services is critical for the success of data production and analysis in High Energy Physics as well as other areas of science.

At FNAL, we have been successfully using HTTP/REST-based data access architecture to provide access to various types of data, in particular, access to scientific databases. We have built a simple yet versatile infrastructure which allows us to use redundant copies of web application servers to increase service performance, scalability and availability. It is designed to handle both state-less and state-full data access methods using distributed web server.

The redundant infrastructure allows us to add or remove individual application servers at any time without a visible interruption of the service. This infrastructure has been successfully used for several years now with data web services as well as with interactive web applications.

We will present components of our infrastructure and several examples of how it can be used.

438

Maintaining Traceability in an Evolving Distributed Computing Environment

Author: Ian Peter Collier[1]

Co-author: Romain Wartel [2]

[1] *STFC - Rutherford Appleton Lab. (GB)*

[2] *CERN*

Corresponding Author: ian.peter.collier@cern.ch

The management of risk is fundamental to the operation of any distributed computing infrastructure. Identifying the cause of incidents is essential to prevent them from re-occurring. In addition, it is a goal to contain the impact of an incident while keeping services operational. For response to incidents to be acceptable this needs to be commensurate with the scale of the problem.

The minimum level of traceability for distributed computing infrastructure usage is to be able to identify the source of all actions (executables, file transfers, pilot jobs, portal jobs, etc) and the individual who initiated them. In addition, sufficiently fine-grained controls, such as blocking the originating user and monitoring to detect abnormal behaviour, are necessary for keeping services operational. It is essential to be able to understand the cause and to fix any problems before re-enabling access for the user.

The aim is to be able to answer the basic questions who, what, where, and when concerning any incident. This requires retaining all relevant information, including timestamps and the digital identity of the user, sufficient to identify, for each service instance, and for every security event including at least the following: connect, authenticate, authorize (including identity changes) and disconnect.

In traditional grid infrastructures (WLCG, EGI, OSG etc.) best practices and procedures for gathering and maintaining the information required to maintain traceability are well established. In particular, sites collect and store information required to ensure traceability of events at their sites.

With the increased use of virtualisation and private and public clouds for HEP workloads established procedures, which are unable to see 'inside' running virtual machines no longer capture all the information required. Maintaining traceability will at least involve a shift of responsibility from sites to Virtual Organisations (VOs) bringing with it new requirements for their logging infrastructures. VOs indeed need to fulfil a new operational role and become fully active participants in the incident response process.

We present an analysis of the changing requirements to maintain traceability for virtualised and cloud based workflows with particular reference to the work of the WLCG Traceability Working Group.

Track 1 Session / 439

A design study for the upgraded ALICE O2 computing facility

Author: Matthias Richter[1]

[1] *University of Oslo (NO)*

Corresponding Author: matthias.richter@cern.ch

An upgrade of the ALICE detector is currently prepared for the Run 3 period of the Large Hadron Collider (LHC) at CERN starting in 2020. The physics topics under study by ALICE during this period will require the inspection of all collisions at a rate of 50 kHz for minimum bias Pb-Pb and 200
kHz for pp and p-Pb collisions in order to extract physics signals embedded into a large background.

The upgraded ALICE detector will produce more than 1 TByte/s of data. Both collision and data rate impose new challenges onto the detector readout and compute system. Some detectors will not use a triggered readout, which will require a continuous processing of the detector data.

Although various online systems are existing for event based reconstruction, the application of a production system for time-based data processing and reconstruction is a novel case in HEP. The project
will benefit from the experience gained with the current ALICE High Level Trigger online system, which already implements a modular concept combining data transport, algorithms and heterogeneous hardware. Processing of individual events will however have to be replaced by the continuous processing of the data stream segmented according to a time-frame structure.

One challenge is the distribution of data within the compute nodes. Time-correlated data sets are received by the First Level Processors (FLP) and must be coherently transported to and aggregated on
the Event Processing Nodes (EPN). Several approaches for the distribution of data are being studied. Aggregated time-frame data is processed on the EPN with the primary goal to reconstruct particle properties. On-the-fly and short-latency detector calibration is necessary for the reconstruction. The
impact of the calibration strategy to the reconstruction performance is under study. Based on the partially reconstructed data, events corresponding to particular collisions can be assembled from the time-based data. The original raw data are then replaced by these preprocessed data. This transformation together with the application of lossless data compression algorithms will provide a data volume reduction of a factor of 20 before data is passed onto the storage system.

Building on messaging solutions, the design and development of a flexible framework for transparent
data flow, online reconstruction, and data compression has started. The system uses parallel processing
on the level of processes and threads within processes in order to achieve an optimal utilization of CPU
cores and memory. Furthermore, the framework provides the necessary abstraction to run common code on heterogeneous platforms including various hardware accelerator cards.

We present in this contribution the first results of a prototype with estimates for scalability and feasibility for a full scale system.

440

Subtlenoise: reducing cognitive load when monitoring distributed computing operations

Author: Peter Love[1]

[1] *Lancaster University (GB)*

Corresponding Author: peter.love@cern.ch

The operation of distributed computing systems requires comprehensive monitoring to ensure reliability and robustness. There are two components found in most monitoring systems: one being visually rich time-series graphs and another being notification systems for alerting operators under certain pre-defined conditions. In this paper the sonification of monitoring messages is explored using an architecture which fits easily within existing infrastructures based on mature opensource technologies such as ZeroMQ, Logstash, and Supercollider (a synth engine). Message attributes are mapped onto audio attributes based on broad classification of the message (continuous or discrete metrics) but keeping the audio stream subtle in nature. The benefits of audio-rendering are described in the context of distributed computing operations and may provide a less intrusive way to understand the operational health of these systems.

441

Real-time alignment and calibration of the LHCb Detector in Run2

Authors: Barbara Storaci[1]; Giulio Dujany[2]

[1] *Universitaet Zuerich (CH)*

[2] *University of Manchester (GB)*

Corresponding Author: barbara.storaci@cern.ch

Stable, precise spatial alignment and PID calibration are necessary to achieve optimal detector performances.
During Run2, LHCb will have a new real-time detector alignment and calibration to reach equivalent performances in the online and offline reconstruction.
This offers the opportunity to optimise the event selection by applying stronger constraints as well as hadronic particle identification at the trigger level.
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 the operative and physics performance point of view.
Specific challenges of this configuration are discussed, as well as the designed framework and its performance.

Track 5 Session / 442

The OSG Open Facility: A Sharing Ecosystem Using Harvested Opportunistic Resources

Authors: Bodhitha Jayatilaka[1]; Chander Sehgal[1]; Marko Slyz[1]; Mats Rynge[2]; Tanya Levshina[3]

[1] *Fermilab*

[2] *RENCI UNC Chapel Hill*

[3] *FERMILAB*

Corresponding Author: bo.jayatilaka@cern.ch

The Open Science Grid (OSG) ties together individual experiments' computing power, connecting their resources to create a large, robust computing grid; this computing infrastructure started primarily as a collection of sites associated with large HEP experiments such as ATLAS, CDF, CMS, and DZero. OSG has been funded by the Department of Energy Office of Science and National Science Foundation since 2006 to meet the US LHC community's computational needs. In the years since, the OSG has broadened its focus to also address the needs of other US researchers and increased delivery of Distributed High Throughput Computing (DHTC) to users from a wide variety of disciplines via the OSG Open Facility. Presently, the Open Facility delivers about 100 million computing wall hours per year to researchers who are not already associated with the owners of the computing sites; this is primarily accomplished by harvesting and organizing the temporarily unused capacity (i.e. opportunistic cycles) from the sites in the OSG. We present an overview of the infrastructure developed to accomplish this from flocking architecture to harvesting opportunistic resources to providing user support to a diverse set of researchers. We also present our experiences since becoming a high-throughput computing service provider as part of the NSF's XD program. Using these methods, OSG resource providers and scientists share computing hours with researchers in many other fields to enable their science, striving to make sure that these computing resources are used with maximal efficiency. We believe that expanded access to DHTC is an essential tool for scientific innovation and work continues in expanding this service.

Track 3 Session / 443

Engineering the CernVM-FileSystem as a High Bandwidth Distributed Filesystem for Auxiliary Physics Data

Author: Dave Dykstra[1]

Co-authors: Brian Paul Bockelman [2]; Jakob Blomer [3]; Ken Herner [4]; Marko Slyz [4]; Tanya Levshina [4]

[1] *Fermi National Accelerator Lab. (US)*

[2] *University of Nebraska (US)*

[3] *CERN*

[4] *Fermilab*

Corresponding Author: jakob.blomer@cern.ch

Fermilab has several physics experiments including NOvA, MicroBooNE, and the Dark Energy Survey that have computing grid-based applications that need to read from a shared set of data files. We call this type of data Auxiliary data to distinguish it from (a) Event data which tends to be different for every job, and (b) Conditions data which tends to be the same for each job in a batch of jobs. Conditions data also tends to be relatively small per job (100s of Megabytes) where both Event data and Auxiliary data are larger per job (Gigabytes), but unlike Event data, Auxiliary data comes from a limited working set (10s to 100s of Gigabytes) of shared files. Since there is some sharing of the Auxiliary data, it appeared at first that the CernVM-Filesystem (CVMFS) infrastructure built for distributing software to the grid would also be the best way to distribute Auxiliary data. However, because grid jobs tend to be started in large batches running the same code, the software distribution use case of CVMFS has a very high cache hit ratio, so the bandwidth requirements on the per-site http proxy caches (squids) is quite low. As a result those proxy caches have been engineered with relatively low bandwidth, and they are easily overwhelmed by Auxiliary data with its relatively low cache hit ratio. A new approach was needed. We are taking advantage of a CVMFS client feature called "alien cache" to cache data on site-local high-bandwidth data servers that were engineered for storing Event data. This site-shared cache replaces caching CVMFS files on both the worker node local disks and on the site-local squids. We have tested this alien cache with the dCache NFSv4.1 interface, Lustre, and Hadoop-FS-fuse, and found that they all perform well. In addition, we use high-bandwidth data servers at

central sites to perform the CVMFS Stratum 1 function instead of the low-bandwidth web servers deployed for the CVMFS software distribution function. We have tested this using the dCache http interface. As a result, we have an end-to-end high-bandwidth widely distributed caching read-only filesystem, using existing client software already widely deployed to grid worker nodes and existing file servers already widely installed at grid sites. Files are published in a central place and are soon available on demand throughout the grid and cached locally on the site with a convenient POSIX interface. This paper discusses the details of the architecture and reports performance measurements.

Track 4 Session / 444

Advances in Distributed High Throughput Computing for the Fabric for Frontier Experiments Project at Fermilab

Authors: Arthur Kreymer[1]; Dave Dykstra[2]; Dennis Box[3]; Gabriele Garzoglio[4]; Ken Herner[5]; Michael Kirby[6]; Parag Mhashilkar[6]; Tanya Levshina[1]

[1] *FERMILAB*

[2] *Fermi National Accelerator Lab. (US)*

[3] *F*

[4] *FERMI NATIONAL ACCELERATOR LABORATORY*

[5] *SUNY Stony Brook*

[6] *Fermi National Accelerator Laboratory*

Corresponding Author: parag@fnal.gov

The FabrIc for Frontier Experiments (FIFE) program is an ambitious, major-impact initiative within the Fermilab Scientific Computing Division designed to lead the computing model development for Fermilab experiments and external projects. FIFE is a collaborative effort between physicists and computing professionals to provide computing solutions for experiments of varying scale, needs, and infrastructure from all areas of Particle Physics: from Neutrino to Collider to Astro-Particle Physics. The major focus of the FIFE project is a single, unified high throughput computing solution for all experiments by development, deployment, and integration of workload management system (WMS), data management layer, and database access layer working seamlessly at sites in Open Science Grid (OSG), commercial and community cloud providers (e.g. Amazon AWS and FermiCloud), and local computing farms. Two development areas where FIFE has made significant progress are a job submission system and reference data distribution. Jobsub is a redesigned scalable, reliable, and robust tiered jobs submission system that integrates with the GlideinWMS workload management system to run complex scientific workflows in Grids and clouds. Jobsub is responsible for functions such as site selection via GlideinWMS, managing credentials, and handling data transfers, with GlideinWMS responsible for provisioning computing resources. Through the development of Alien Cache for CVMFS, the FIFE program has considerably expanded the capabilities of CVMFS for reference data distribution. In addition to job submission and reference data distribution, the FIFE project has also made significant progress integrating services into experiment computing operations such as a flexible data transfer client and access to opportunistic resources on the Open Science Grid. The progress with current experiments and plans for expansion with additional projects will be discussed. FIFE has taken the leading role in defining the computing model for Fermilab experiments, aided in the design of experiments beyond Fermilab, and will continue to lead the future direction of high throughput computing for future physics experiments world wide.

Track 2 Session / 445

slic : A full-featured Geant4 simulation program

Author: Norman Anthony Graf[1]

Co-author: Jeremy McCormick [2]

[1] *SLAC National Accelerator Laboratory (US)*

[2] *SLAC National Accelerator Laboratory*

Corresponding Author: norman.graf@slac.stanford.edu

As the complexity and resolution of particle detectors increases, the need for detailed simulation of the experimental setup also increases. We have developed efficient and flexible tools for detailed physics and detector response simulations which build on the power of the Geant4 toolkit but free the end user from any C++ coding. Geant4 is the de facto high-energy physics standard for simulating the interaction of particles with fields and materials. However, the end user is required to write their own C++ program, and the learning curve for setting up the detector geometry and defining sensitive elements and readout can be quite daunting, especially for those without previous experience or not associated with large collaborations. We have developed the Geant4-based detector simulation program, slic, which employs generic IO formats as well as a textual detector description. Extending the pure geometric capabilities of GDML, LCDD enables fields, regions, sensitive detector readout elements, etc. to be fully described at runtime using an xml file. We also describe how more complex geometries, such as those from CAD programs, can be seamlessly incorporated into the xml files. We have defined generic "hits"which can be used to model sophisticated tracking and calorimetry readouts, but the native Geant4 scoring functionality can also be used for simpler applications. Although developed within the context of HEP collider detectors, the program is completely flexible and can be used to simulate detectors in many different fields. We present a software toolkit and computing infrastructure which allows physicists to quickly and easily contribute to detector design by modeling detector elements without requiring either C++ coding expertise or experience with Geant4. Examples of its use designing collider detectors for the ILC and CLIC as well as its use in simulating the detector response and physics performance of a fixed-target experiment at the Thomas Jefferson National Laboratory will be presented.

446

org.lcsim: A Java-based tracking toolkit

Author: Norman Anthony Graf[1]

[1] *SLAC National Accelerator Laboratory (US)*

Corresponding Author: norman.graf@slac.stanford.edu

We describe a software toolkit for full event simulation and reconstruction in silicon tracking detectors. It features modular packages providing sophisticated simulations of the response of silicon detectors to the passage of charged particles. Sensor classes allow very detailed descriptions of charge carrier movement in silicon detectors: one can list the collecting, absorbing and reflecting regions, properties of silicon (doping, mobility, diffusion length, etc.), and electric and magnetic fields (including TCAD maps). After the charge carriers are generated and collected, the electronics simulation processes this into digital signals. We have defined an interface to specify how any such simulation should communicate with other parts of the package. Since details of signal processing are very sensor specific, it is anticipated that any sensor option will have its own class handling such processing, but we have implemented a number of readout technologies of interest to HEP detectors, such as CCDs and active pixel devices. Common to all the specific electronics simulation are the addition of electronics noise, propagation of the signal to readout, thresholding, and digitization of the signal. The final output is then a list of electronics channels with their corresponding ADC counts, and optionally the time for the signal, replicating the readout from a real detector. We also provide code for cluster finding, pattern recognition, track-finding and fitting, and analysis. The detector is defined by the same xml input files used for the Geant4 detector response simulation, ensuring that simulation and reconstruction geometries are always commensurate by construction.
We describe an easy-to-use software toolkit used to fully simulate all aspects of silicon trackers, from signal development in pixels and microstrips, through pattern recognition and track fitting, to analysis. Originally developed within the context of collider detector development for the ILC, it is

being used by the Heavy Photon Search Detector, a fixed-target experiment at the Thomas Jefferson National Laboratory. We will describe its use and performance in a running experiment.

Track 7 Session / 447

Interoperating Cloud-based Virtual Farms

Authors: Domenico Elia[1]; Grazia Luparello[2]; Massimo Venaruzzo[2]; Stefano Bagnasco[3]; Stefano Piano[4]

[1] *INFN Bari*

[2] *Universita e INFN (IT)*

[3] *I.N.F.N. TORINO*

[4] *INFN (IT)*

Corresponding Author: bagnasco@to.infn.it

The present work aims at optimizing the use of computing resources available at the grid Italian Tier-2 sites of the ALICE experiment at CERN LHC by making them accessible to interactive distributed analysis, thanks to modern solutions based on cloud computing. The scalability and elasticity of the computing resources via dynamic ("on-demand") provisioning is essentially limited by the size of the computing site, reaching the theoretical optimum only in the asymptotic case of infinite resources. The main challenge of the project is to overcome this limitation by federating different sites through a distributed cloud facility. Storage capacities of the participating sites are seen as a single federated storage area, preventing from the need of mirroring data across them: high data access efficiency is guaranteed by location-aware analysis software and storage interfaces, in a transparent way from an end-user perspective. Moreover, the interactive analysis on the federated cloud reduces the execution time with respect to grid batch jobs. The tests of the investigated solutions for both cloud computing and distributed storage on wide area network will be presented.

Track 7 Session / 448

Cloud services for the Fermilab scientific stakeholders

Authors: Gabriele Garzoglio[1]; Steven Timm[2]

Co-authors: Gerard Bernabeu [1]; Hyunwoo Kim [1]; Joe Boyd [1]; Neha Sharma [1]; Nicholas Peregonow [1]; Parag Mhashilkar [3]; SANDEEP PALUR [4]; Seoyoung Noh [5]

[1] *FERMI NATIONAL ACCELERATOR LABORATORY*

[2] *Fermilab*

[3] *Fermi National Accelerator Laboratory*

[4] *ILLINOIS INSTITUTE OF TECHNOLOGY*

[5] *KISTI Korea Institute of Science & Technology Information (KR)*

Corresponding Author: parag@fnal.gov

As part of the Fermilab/KISTI cooperative research project, Fermilab has successfully run an experimental simulation workflow at scale on a federation of Amazon Web Services (AWS), FermiCloud, and local FermiGrid resources. We used the CernVM-FS (CVMFS) file system to deliver the application software. We established Squid caching servers in AWS as well, using the Shoal system to let each individual virtual machine find the closest squid server. We also developed an automatic virtual machine conversion system so that we could transition virtual machines made on FermiCloud to Amazon Web Services.
We used this system to successfully run a cosmic ray simulation of the NOvA detector at Fermilab, making use of both AWS spot pricing and network bandwidth discounts to minimize the cost. On

FermiCloud we also were able to run the workflow at the scale of 1000 virtual machines, using a private network routable inside of Fermilab. We present the details of the technological improvements that were used to make this successfully.

449

Integrating grid and cloud resources at the RAL Tier-1

Author: Andrew David Lahiff[1]

Co-author: Ian Peter Collier [1]

[1] *STFC - Rutherford Appleton Lab. (GB)*

Corresponding Author: andrew.david.lahiff@cern.ch

Today the primary method by which the LHC and other experiments run computing work at WLCG sites is grid job submission. Jobs are submitted to computing element middleware which in turn submits jobs to a batch system managing the local compute resources. With the increasing interest and usage of cloud technology, a new challenge facing sites which support multiple experiments in recent years is a need to provide both traditional grid as well as cloud interfaces, but without partitioning the underlying resources. When the batch system is busy but the cloud is idle, it should be possible for the unused cloud resources to be included in the batch system. Similarly, when the batch system is idle but the cloud is busy, the unused batch resources should be available for users within the cloud. At the RAL Tier-1 a cloud based on OpenNebula has been under development for some time and will made available to the LHC experiments and others, as well as being used internally by staff for activities such as testing and developement. Here we present our experience unifying the cloud with our production HTCondor batch system in a way that avoids static partitioning, ensures that resources are used efficiency and that allocations are respected.

Track 7 Session / 450

Implementation of the vacuum model using HTCondor

Author: Andrew David Lahiff[1]

[1] *STFC - Rutherford Appleton Lab. (GB)*

Corresponding Author: andrew.david.lahiff@cern.ch

The recently introduced vacuum model offers an alternative to the traditional methods that virtual organisations (VOs) use to run computing tasks at sites, where they either submit jobs using grid middleware or create virtual machines (VMs) using cloud APIs. In the vacuum model VMs are created and contextualized by the site itself, and start the appropriate pilot job framework which fetches real jobs. This allows sites to avoid the effort required for running grid middleware or a cloud. Here we present an implementation of the vacuum model based entirely on HTCondor, making use of HTCondor's ability to manage VMs. Extensive use is made of job hooks, including for preparing fast local storage for use in the VMs, carrying out contextualization, and updating job ClassAds about the status of the VMs and their payloads. VMs for each supported VO are created at regular intervals. If there is no work or there are fatal errors, no additional VMs are created. On the other hand, if there is real work running, further VMs can be created. Since the HTCondor negotiator decides whether to run the VMs or not, fairshares are naturally respected. Normal grid or locally-submitted jobs can run on the same resources and share the same physical worker nodes that are also used as hypervisors for running VMs.

Track 2 Session / 451

MAUS: The MICE Analysis and User Software

Author: Durga Rajaram[1]

[1] *IIT, Chicago*

Corresponding Author: janusz.martyniak@imperial.ac.uk

The Muon Ionization Cooling Experiment (MICE) has developed the MICE Analysis User Software (MAUS) to simulate and analyse experimental data. It serves as the primary codebase for the experiment, providing for offline batch simulation and reconstruction as well as online data quality checks . The software provides both traditional particle physics functionalities such as track reconstruction and particle identification, and accelerator physics functions such as calculating transfer matrices and emittances. The code is structured in a Map-Reduce framework to allow parallelization whether on a personal computer or in the control room. MAUS allows users to develop in either Python or C++ and provides APIs for both. Various software engineering practices from industry are also used to ensure correct and maintainable physics code, which include unit, functional and integration tests, continuous integration and load testing, code reviews, and distributed version control systems. The software framework and the simulation and reconstruction capabilities are described.

452

Experience with batch systems and clouds sharing the same physical resources

Author: Andrew David Lahiff[1]

Co-authors: Frazer Barnsley [1]; George Ryall [2]; Ian Peter Collier [1]

[1] *STFC - Rutherford Appleton Lab. (GB)*

[2] *STFC*

Corresponding Author: andrew.david.lahiff@cern.ch

Today it is becoming increasingly common for WLCG sites to provide both grid and cloud compute resources. In order to avoid the inefficiencies caused by static partitioning of resources it is necessary to integrate grid and cloud resources. There are two options to consider when doing this. The simplest option is to have the cloud manage all the physical hardware and use entirely virtualised worker nodes in the batch system. The downside of this is that everything is virtualised, whereas it might be useful to be able to run jobs in the batch system directly on hardware in order to achieve the best performance. In such a configuration it is essential that batch jobs can't interfere with the virtual machines running on the same node or affect the hypervisors in a negative way, hence containerisation of batch jobs is particularly important. It is also important to consider the implications of having both a batch system and cloud scheduling jobs and virtual machines on the same machines. We present an investigation into the feasibility of having a batch system and cloud sharing the same set of physical resources carried out at the RAL Tier-1.

453

A Data Summary File Structure and Analysis Tools for Neutrino Oscillation Analysis at the NOvA Experiment

Authors: Christopher Backhouse[None]; Dominick Rocco[1]

[1] *urn:Google*

Corresponding Author: dominick.r.rocco@gmail.com

The NuMI Off-axis Neutrino Experiment (NOvA) is designed to study neutrino oscillations in the NuMI beam at Fermilab. Neutrinos at the Main Injector (NuMI) is currently being upgraded to provide 700 kW for NOvA. A 14 kt Far Detector in Ash River, MN and a functionally identical 0.3 kt Near Detector at Fermilab are positioned 810 km apart in the NuMI beam line. The fine granularity of the NOvA detectors provides a detailed representation of particle trajectories. The data volume associated with such granularity, however, poses problems for analyzing data with ease and speed. NOvA has developed a data summary file structure which discards the full event record in favor of higher-level reconstructed information. A general-purpose framework for neutrino oscillation measurements has been developed for analysis of these data summary files. We present the design methodology for this new file format as well as the analysis framework and the role it plays in producing NOvA physics results.

454

The Linear Collider Software Framework

Author: Norman Anthony Graf[1]

Co-authors: Akiya Miyamoto [2]; Andre Sailer [3]; Frank-Dieter Gaede [4]

[1] *SLAC National Accelerator Laboratory (US)*

[2] *High Energy Accelerator Research Organization (JP)*

[3] *CERN*

[4] *Deutsches Elektronen-Synchrotron (DE)*

Corresponding Author: norman.graf@slac.stanford.edu

Detectors at future electron-positron linear colliders such as ILC or CLIC will require unprecedentedly precise tracking, vertexing, and calorimetry in order to meet the ambitious physics goals of the experimental program. The physics performance of different detector geometries and technologies has to be realistically estimated. These assessments require sophisticated and flexible full detector simulation and reconstruction software. At the heart of the linear collider detectors lies the particle flow reconstruction. The goal of unambiguously associating tracks and showers to individual particles requires the combination of precise, lightweight trackers combined with fine-grained "imaging" calorimeters and advanced clustering software. Detailed physics and detector optimization studies are taking place in the CLICdp, ILD, and SiD groups. The similarities between the different detector concepts allow for the use of common software tools. All the concepts share an event data model and persistency format which enables the sharing of files and applications across the concepts. Particle flow clustering, vertexing and flavor tagging is already provided by standalone packages via lightweight interfaces. In the near future the geometry information for all detector layouts will be provided by a common source for the simulation and reconstruction programs, providing further re-use of software between the collaborations. In addition, a track reconstruction package is currently under development. The sharing and development of flexible software tools not only saves precious time and resources, the use of common tools for different detectors also helps to uncover bugs or inefficiencies that would be harder to spot without multiple users. The concept of generic software tools and some of the programs themselves can be beneficial to experiments beyond the linear collider community.

455

Efficient provisioning for multicore applications with LSF

Author: Stefano Dal Pra[1]

[1] *INFN*

Corresponding Author: stefano.dalpra@cnaf.infn.it

Tier-1 sites providing computing power for HEP experiments are usually tightly designed for high throughput performances. This is pursued by reducing the variety of supported usecases and tuning for performances those ones, the most important of which have been that of single-core jobs. Moreover, the usual workload is saturation: each available core in the farm is in use and there are queued jobs waiting for their turn to run. Enabling multicore jobs thus requires dedicating a number of hosts where to run, and waiting for them to free the needed number of cores. This drain-time introduces a loss of computing power driven by the number of unusable empty cores.

As an increasing demand for multicore capable resources have emerged, a Task Force have been constituted in WLCG, with the goal to define a simple and efficient multicore resource provisioning model.

This paper details the work done at the INFN Tier1 to enable multicore support for the LSF batch system, with the intent of reducing to the minimum the average number of unused cores.

The adopted strategy has been that of dedicating to multicore a dynamic set of nodes, whose dimension is mainly driven by the number of pending multicore requests and fairshare priority of the submitting user. The node status transition, from single to multi core et vice versa, is driven by a finite state machine which is implemented in a custom multicore director script, running in the cluster.

After describing and motivating both the implementation and the details specific to the LSF batch system, results about performance are reported. Factors having positive and negative impact on the overall efficiency are discussed and solutions to reduce at most the negative ones are proposed.

Track 6 Session / 456

Migrating to 100GE WAN Infrastructure at Fermilab

Author: Phil Demar[1]

Co-authors: Andrey Bobyshev [2]; Vyto Grigaliunas [1]

[1] *Fermilab*

[2] *FERMILAB*

Corresponding Author: demar@fnal.gov

Fermilab is in the process of upgrading its wide-area network facilities to 100GE technology. One might assume that migration to be relatively straightforward, with forklift upgrades of our existing network perimeter devices to 100GE-capable platforms, and accompanying deployment of 100GE WAN services. However, our migration to 100GE WAN technology has proven to be significantly more complicated than that. Several key factors are driving this complexity:

1 The Laboratory has historically separated its high impact science (LHC) data movement from its general internet traffic. For a decade, this has been implemented with additional physical infrastructure, included a separate border router and multiple 10GE WAN links dedicated to the science data traffic. However, the extreme cost of 100GE routers, coupled with corresponding high cost of

100GE WAN links, has necessitated a consolidation of the current WAN infrastructure into a simpler configuration, one in which science data and general internet traffic must share the same 100GE infrastructure.

2 The Laboratory has also been able to rely on the diversity inherent in its 8x10GE WAN link configuration to provide acceptable resiliency in its WAN services. Replacement of the multiple 10GE links with a single 100GE link has the potential to seriously compromise that diversity. To alleviate that vulnerability, plans for deployment of a second 100GE WAN link in early 2015 are already under way. However, preserving full geographic diversity with limited 100GE equipment has presented additional challenges.

3 With the availability of 100GE WAN infrastructure, as well as 40GE-attached network R&D systems, traffic generated as part of network research activities also potentially becomes much more significant. While this is a positive thing for the Laboratory's network R&D efforts, it adds yet another class of high-impact traffic competing for the new 100GE WAN resources.

This talk will describe how Fermilab is consolidating its WAN infrastructure in order to implement 100GE technology, while still preserving its long term philosophy about keeping high impact science data traffic, its general internet traffic, and it network R&D traffic logically isolated from each other. Current virtualization techniques being deployed will be discussed, as will investigations into use of emerging Software Defined Networking (SDN) and OpenFlow technologies as a longer term solution.

Track 4 Session / 457

Multicore-Aware Data Transfer Middleware (MDTM)

Author: Wenji Wu[1]

Co-authors: Liang Zhang [2]; Phil Demar [2]

[1] *Fermi National Accelerator Laboratory*

[2] *Fermilab*

Corresponding Author: wenji@fnal.gov

Multicore and manycore have become the norm for scientific computing environments. Multicore/manycore platform architectures provide advanced capabilities and features that can be exploited to enhance data movement performance for large-scale distributed computing environments, such as LHC. However, existing data movement tools do not take full advantage of these capabilities and features. The result is inefficiencies in data movement operations, particularly within the wide area scope. These inefficiencies will become more pronounced as networks upgrade to 100GE infrastructure, and host systems correspondingly migrate to 10GE, Nx10GE, and 40GE technologies. To address these inefficiencies and limitations, DOE's Advanced Scientific Computing Research (ASCR) office has funded Fermilab and Brookhaven National Laboratory to collaboratively work on the Multicore-Aware Data Transfer Middleware (MDTM) project. MDTM aims to accelerate data movement toolkits on multicore systems. Essentially, the MDTM project consists of two major components:

- MDTM middleware services to harness multicore parallelism and make intelligent decisions that align CPU, memory, and I/O device operations in a manner which optimizes performance for higher layer applications.
- MDTM-enabled data transfer applications (client or server) that can utilize MDTM middleware to reserve and manage multiple CPU cores, memory, network devices, disk storage as an integrated resource entity, thus achieving higher throughput and enhanced quality of service when compared with existing approaches.

A prototype version of MDTM is currently undergoing testing and evaluation. This talk will describe MDTM's architectural and design principles, how it works in implementation, and initial test results

in comparison to standard GridFTP and BBCP operations. In addition, future directions for the project will be discussed, including the notion of enabling an external resource scheduling capability for MDTM, thus making it a reservable component for application-driven end-to-end path resource reservations.

458

Recent Developments in the Infrastructure and Use of artdaq

Author: Kurt Biery[1]

Co-authors: Christopher Green [2]; Eric Flumerfelt [3]; Gennadiy Lukhanin [1]; Jim Kowalkowski [3]; John Freeman [3]; Marc Paterno [3]; Ronald Rechenmacher [1]

[1] *Fermi National Accelerator Lab. (US)*

[2] *Department of Physics*

[3] *Fermilab*

Corresponding Author: biery@fnal.gov

The artdaq data acquisition software toolkit has been developed within the Fermilab Scientific Computing Division, and it is being used by a growing number of high-energy and cosmology experiments. It currently provides data transfer, event building, run control, and event analysis functionality. The event analysis functionality is provided by the art framework, which has also been developed at Fermilab and which is used for offline reconstruction and simulation by many experiments at Fermilab. The format that is used to store the data is ROOT.

Recent enhancements to the toolkit include a custom vector class to improve the performance of buffer allocation, configurable options for dealing with back-pressure, and improved organization of the software packages that are part of the toolkit. The enhancements that are currently in development include the addition of web-based control and monitoring interfaces, improved monitoring of DAQ system performance, and improved handling of online monitoring data.

Artdaq is successfully being used for the DarkSide-50, LArIAT, LBNE 35 ton prototype, and Mu2e experiments. In addition, a demo system has been developed so that prospective and new users can gain experience in using the toolkit to develop their own DAQ system.

We will describe the functionality that is currently available in the toolkit, the advantages that are gained by its use, our experience in deploying it to experiments, future plans, and opportunities for integrating with commercial off-the-shelf hardware modules.

459

Application-Oriented Network Traffic Analysis based on GPUs

Author: Phil Demar[1]

Co-authors: Liang Zhang [1]; Wenji Wu [2]

[1] *Fermilab*

[2] *Fermi National Accelerator Laboratory*

Corresponding Authors: demar@fnal.gov, wenji@fnal.gov

Software-Defined Networking (SDN) has emerged as a major development direction in network technology. Conceptually, SDN enables customization of forwarding through network infrastructure on

a per-flow basis. With SDN, a high impact LHC data flow could be allocated a "slice"of the network infrastructure. Functionally, the data flow would have a private path through the network infrastructure, potentially with performance guarantees. SDN offers a technology path to meet extreme data movement requirements, like LHC, over highly capacious, general use network infrastructure. Standardized implementations of SDN, specifically OpenFlow, are now becoming available on network routers and switches.

Conventionally, SDN configuration is implemented by users or applications. User-driven SDN configuration is manual, and therefore tends to be static and inherently inefficient. It doesn't scale well to complex traffic patterns, such as the traffic patterns generated by LHC experiments. Application-driven SDN configuration is typically dynamic, representing a more efficient on-demand type of approach for network resources. However, building SDN-awareness into applications is highly complex and extremely challenging. It places ongoing software adaptation and maintenance burdens on the application developers. Finally, there will always be major policy issues or constraints on an application's ability to modify the network.

We believe networks need an application-awareness capability. With such a capability, traffic identified as "special"by the network could be provided with a custom network "slice"without manual intervention or modification to applications. To facilitate this capability, we are currently developing a network traffic analysis tool to identify "special"types of network traffic. Our tool is designed to perform this analysis in real-time. We accomplish this through GPU technology, which is extremely well-suited to parallelize the processing and analysis of the network traffic. Our traffic analysis uses basic traffic characteristics such as packet size, packet spacing, flow duration, and multi-stream flow presence to develop the traffic pattern signatures and templates necessary for traffic classification and/or identification. Our ultimate objective is to create a tool that could facilitate dynamic customization of SDN-capable network infrastructure, based on real-time traffic analysis and detection.

A prototype of our analysis tool is currently in development. This talk will describe the tool's architectural and design principles, how it works in implementation, the challenges faced in high speed (40GE/100GE) network environments, and our initial trial results. In addition, future directions for the project, as well as wider applicability for the underlying technology, will be discussed.

460

A virtual validation cluster for ALICE software releases based on CernVM

Author: Dario Berzano[1]

Co-author: Mikolaj Krzewicki [2]

[1] *CERN*

[2] *Johann-Wolfgang-Goethe Univ. (DE)*

Corresponding Author: dario.berzano@cern.ch

One of the most important steps of software lifecycle is Quality Assurance: this process comprehends both automatic tests and manual reviews, and all of them must pass successfully before the software is approved for production. Some tests, such as source code static analysis, are executed on a single dedicated service: in High Energy Physics, a full simulation and reconstruction chain on a distributed computing environment, backed with a sample "golden"dataset, is also necessary for the quality sign-off. The ALICE experiment uses dedicated and virtualized computing infrastructures for the Release Validation in order not to taint the production environment (i.e. CVMFS and the Grid) with non-validated software and validation jobs: the ALICE Release Validation cluster is a disposable virtual cluster appliance based on CernVM and the Virtual Analysis Facility, capable of deploying on-demand, and with a single command, a dedicated virtual HTCondor cluster with an automatically scalable number of virtual workers on any cloud supporting the standard EC2 interface. Input and output data are externally stored on EOS, and a dedicated CVMFS service is used to provide the software to be validated. We will show how the Release Validation Cluster deployment

and disposal are completely transparent for the Release Manager, who simply triggers the validation from the ALICE build system's web interface. CernVM 3, based entirely on CVMFS, permits to boot any snapshot of the operating system in time: we will show how this allows us to certify each ALICE software release for an exact CernVM snapshot, addressing the problem of Long Term Data Preservation by ensuring a consistent environment for software execution and data reprocessing in the future.

Track 7 Session / 461

Lightweight scheduling of elastic analysis containers in a competitive cloud environment: a Docked Analysis Facility for ALICE

Author: Dario Berzano[1]

Co-authors: Gerardo Ganis [1]; Ioannis Charalampidis [1]; Jakob Blomer [1]; Predrag Buncic [1]; Rene Meusel [1]

[1] *CERN*

Corresponding Author: dario.berzano@cern.ch

During the last years, several Grid computing centers chose virtualization as a better way to manage diverse use cases with self-consistent environments on the same bare infrastructure. The maturity of control interfaces (such as OpenNebula and OpenStack) opened the possibility to easily change the amount of resources assigned to each use case by simply turning on and off virtual machines. Some of those private clouds use, in production, copies of the Virtual Analysis Facility, a fully virtualized and self-contained batch analysis cluster capable of expanding and shrinking automatically upon need: however, resources starvation occurs frequently as expansion has to compete with other virtual machines running long-living batch jobs. Such batch nodes cannot relinquish their resources in a timely fashion: the more jobs they run, the longer it takes to drain them and shut off, and making one-job virtual machines introduces a non-negligible virtualization overhead. By improving several components of the Virtual Analysis Facility we have realized an experimental "Docked"Analysis Facility for ALICE, which leverages containers instead of virtual machines for providing performance and security isolation. We will present the techniques we have used to address practical problems, such as software provisioning through CVMFS, as well as our considerations on the maturity of containers for High Performance Computing. As the abstraction layer is thinner, our Docked Analysis Facilities may feature a more fine-grained sizing, down to single-job node containers: we will show how this approach will positively impact automatic cluster resizing by deploying lightweight pilot containers instead of replacing central queue polls.

Track 3 Session / 462

Archiving Scientific Data outside of the traditional High Energy Physics Domain, using the National Archive Facility at Fermilab

Author: Andrew Norman[1]

Co-authors: Adam Lyon [1]; Marc Mengel [1]; Michael Diesburg [2]; Michael Gheith [1]; Robert Illingworth [1]

[1] *Fermilab*

[2] *F*

Corresponding Author: anorman@fnal.gov

Many experiments in the HEP and Astrophysics communities generate large extremely valuable datasets, which need to be efficiently cataloged and recorded to archival storage. These datasets,

both new and legacy, are often structured in a manner that is not conducive to storage and cataloging with modern data handling systems and large file archive facilities. In this paper we discuss in detail how we have created a robust toolset and simple portal into the Fermilab Archive Facility, which allows for scientific data to be quickly imported, organized and retrieved from the 0.650 Exabyte facility.

In particular we discuss how the data from the Sudbury Neutrino Observatory (SNO) for the COUPP dark matter detector was aggregated, cataloged, archived and re-organized to permit it to be retrieved and analyzed using modern distributed computing resources both at Fermilab and on the Open Science Grid. We pay particular attention to the methods that were employed to "uniquify" the namespaces for the data, derive metadata for the over 460,000 image series taken by the COUP experiment and what was required to map that information into coherent datasets that could be stored and retrieved using the large scale archives systems.

We describe the data transfer and cataloging engines that are used for data importation and how these engines have been setup to import data from the data acquisition systems of ongoing experiments at non-Fermilab remote sites including the Laboratori Nazionali del Gran Sasso and the Ash River Laboratory in Orr, Minnesota. We also describe how large University computing sites around the world are using the system to store and retrieve large volumes of simulation and experiment data for physics analysis.

Track 4 Session / 463

Replacing the Engines without Stopping The Train; How A Production Data Handling System was Re-engineered and Replaced without anyone Noticing.

Author: Marc Mengel[1]

Co-authors: Adam Lyon [1]; Andrew Norman [1]; Michael Diesburg [2]; Michael Gheith [1]; Robert Illingworth [1]; Steve White [1]

[1] *Fermilab*

[2] *F*

Corresponding Author: anorman@fnal.gov

As high energy physics experiments have grown, their operational needs and requirements they place on computing systems change. These changes often require new technical solutions to meet the increased demands and functionalities of the science. How do you affect sweeping change to core infrastructure, without causing major interruptions to the scientific programs?

This paper explores the operational challenges, procedures and techniques that were used to completely replace the core data handling infrastructure for the Fermilab experimental program, while continuing to store and deliver more than 1PB of data per month to the analysis and computing efforts of the experiments. It discusses designs patterns like the Command Pattern, the Façade and Delegation that were employed at different stages of the project and how they worked and didn' t work to hide the underlying changes that were being made. We discuss how parallel production and parasitic integration systems were used to perform "at scale"performance testing and data validation prior to switch overs and how this allowed for more robust test environments which would not have been possible to replicate in traditional test settings. We describe how the experimenters were engaged in the transition process and how they were used to propagate subtle but necessary changes to the client interfaces over a period of 18 months.

Track 5 Session / 464

Large Scale Management of Physicist's Personal Analysis Data

Author: Andrew Norman[1]

Co-authors: Marc Mengel [1]; Matthew Tamsett [2]; Robert Group [3]; Robert Illingworth [1]

[1] *Fermilab*

[2] *University of Sussex*

[3] *University of Virginia*

Corresponding Author: anorman@fnal.gov

The ability of modern HEP experiments to acquire and process unprecedented amounts of data and simulation have led to an explosion in the volume of information that individual scientists deal with on a daily basis. This explosion has resulted in a need for individuals to generate and keep large "personal analysis"data sets which represent the skimmed portions of official data collections pertaining to their specific analysis. These personal analysis and simulation sets represent a significant reduction in size compared to the original data, but they can still be many terabytes or tens of terabytes in size and consist of tens of thousands of files. When this personal data is aggregated across the many physicists in a single analysis group or experiment it can represent data volumes on par with or exceeding the official "production"samples which require special data handling techniques and storage systems to deal with effectively.

In this paper we explore the toolsets, analysis models and changes to the Fermilab computing infrastructure which have been developed and deployed by the NOvA experiment to allow experimenters to effectively manage their personal analysis data and other data that falls outside of the typically centrally managed production chains. In particular we describe the models and tools that are being used to allow NOvA to leverage Fermilab storage resources that are sufficient to meet their analysis needs, without imposing management burdens of specific quotas on users or groups of users, without relying on traditional central disk facilities and without having to constantly police individuals users usage. We discuss the storage mechanisms and the caching algorithms that are being used as well as the toolkits that have been developed to allow the users to easily operate with terascale+ datasets.

465

Large Scale Monte Carlo Simulation of neutrino interactions using the Open Science Grid and Commercial Clouds

Author: Andrew Norman[1]

Co-authors: Eric Flumerfelt [1]; Gavin Davies [2]; Matthew Tamsett [3]; Nathan Mayer [4]; Steven Timm [1]

[1] *Fermilab*

[2] *Iowa State University*

[3] *University of Sussex*

[4] *Tufts University*

Corresponding Author: anorman@fnal.gov

Modern long baseline neutrino experiments like the NOvA experiment at Fermilab, require large scale, compute intensive simulations of their neutrino beam fluxes and backgrounds induced by cosmic rays. The amount of simulation required to keep the systematic uncertainties in the simulation from dominating the final physics results is often 10x to 100x that of the actual detector exposure. For the first physics results from NOvA this has meant the simulation of more than 2 billion cosmic ray events in the far detector and more than 300 million NuMI beam spill simulations. Performing these high statistics levels of simulation have been made possible for NOvA through the use of the Open Science Grid and through large scale runs on commercial clouds like Amazon EC2.

This paper details the challenges in performing large scale simulation in these environments and how the computing infrastructure for the NOvA experiment has been adapted to seamlessly support the routing of different simulation and data processing tasks to these resources. We discuss the optimization of the simulation computing model, data movement and computation to match the commercial computing environment.

466

Improvement of AMGA Python Client Library for the Belle II Experiment

Author: Jae-Hyuck Kwak[1]

Co-authors: Geun Chul Park [1]; Soonwook Hwang [1]; Tae Sang Huh [1]

[1] *KISTI*

Corresponding Author: jhkwak@kisti.re.kr

This paper describes the recent improvement of AMGA python client library for the Belle II Experiment. We were drawn to the action items about library improvement after in-depth discussions with the developer of the Belle II distributed computing group. It includes GSI support, client-side metadata federation support and atomic operation support. Some of the improvements were already applied to AMGA python client library bundled with the Belle II distributed computing software. The recent mass Monte-Carlo production campaign shows AMGA python client library operates in a reliable stability.

467

Dark Energy Survey Computing on FermiGrid

Authors: Brian Yanny[1]; Elizabeth Buckley-Geer[1]; Marko Slyz[1]

[1] *FNAL*

Corresponding Author: mslyz@fnal.gov

The Dark Energy Survey (DES) uses a CCD camera installed in the Blanco telescope in Cerro Tololo, Chile. The goal of the survey is to study the effect known as Dark Energy.

DES uses Fermigrid for nightly processing, for quality assessement of images, and for the detection of type 1A Super Novae. Nighly processing needs to be carried out for each of the 105 nights in a season that DES acquires data, and must be completed before observations begin on the following night. This was seen as feasible on Fermigrid because the requirements for memory and CPU were similar to those of HEP jobs. Fermigrid used some HEP computing techniques – among them, the CernVM File System (CVMFS) for storing software, and disks on the worker nodes for each job's scratch space – that were novel to cosmology experiments accustomed to using HPC machines.

At the same time, we learned of other compute requirements which were not well served by the the existing model, but were still well suited to the basic approach of loosely-coupled, high throughput computing used in HEP. We are working to support workflows with large memory

requirements, workflows that require multi-core cpus, and workflows that cache calibration data on the worker nodes.

DES started running production on Fermigrad in August of 2014. We present how we are addressing some notable problems with the current system, including the variable amount of time it takes for files to be visible on worker nodes after they're first uploaded to CVMFS, low CPU efficiency, and jobs hanging while finishing.

468

The Belle II analysis on Grid

Authors: Chia-Ling Hsu[1]; Martin Sevior[2]

[1] *University of Melbourne*

[2] *University of Melbourne (AU)*

Corresponding Author: chialingh@student.unimelb.edu.au

The basf2 software framework has been developed the Belle II experiment, the next generation B-factory experiment at the KEK Laboratory. Belle II will collect 50 times more data than the previous Belle experiment and has a commensurate increase in computing requirements.
Consequently Belle II has adopted a distributed computing solution to provide the computing resources required of the experiment. The interface is called gbasf2 and is designed to provide an easy transition from an analysis done with basf2 to an analysis performed on the grid. The output of the raw data processing are mDST files, consisting of all relevant information for physics analyses. Another file format called μDST has been proposed which also records the reconstructed particle information.
We present a study of physics analysis with gbasf2 and compare the performance of the analysis methods which employ mDST and μDST files on grid sites and on local resources.

469

The GENFIT Library for Track Fitting and its Performance in Belle II

Authors: Johannes Rauch[1]; Tobias Schlüter[2]

[1] *Technische Universität München*

[2] *LMU München*

Corresponding Author: tobias.schlueter@physik.uni-muenchen.de

GENFIT is an experiment-independent, universal track-fitting package, available under a free software license. It implements a variety of track-fitting algorithms and provides the surrounding functionality needed by particle physics experiments: general handling of detector hits, supplemented with example implementations for various detector types; track extrapolation code; a track-data model allowing for flexible storage and track-level operations such as the chaining of tracks from different subdetectors; residual calculation; interfaces to alignment codes; track and fit visualization, which can also serve didactical purposes. The package has been significantly overhauled and extended in the course of the development of the Belle II software framework, taking into account

lessons learned also in the COMPASS, FOPI and PANDA experiments. GENFIT is particularly suited e.g. for new experiments that want to be able to use track fits of simulated data already at early stages in order to more realistically assess the impact of design choices. We will report on the current status of the project, performance of the code within Belle II, and share some of the lessons learned during the work on the project.

470

A quantative evaluation of different methods for instantiating private cloud virtual machines

Authors: Alexander Dibbo[1]; George Ryall[2]; Ian Peter Collier[3]

Co-authors: Andrew David Lahiff [3]; Frazer Barnsley [2]

[1] *STFC RAL*

[2] *STFC*

[3] *STFC - Rutherford Appleton Lab. (GB)*

Corresponding Author: ian.peter.collier@cern.ch

The RAL Tier-1 has been deploying production virtual machines for several years. As we move to providing a production private cloud, managed using OpenNebula, we have experimented with a range of different ways of deploying virtual machine images on hypervisors. We present a quantative comparison of a variety of virtual machine image and storage combinations, including monolithic Scientific Linux VMs copied and run locally on hy[pervisors, the same images run on Ceph shared storage, lean images with volatile filesystems created and run locally as required. We will also examine the same combinations using the micro-CernVM - where the bulk of the operating system is provided by the CernVM filesystem. We will compare performance, with regards to both instantiation time and computational performance for HEP workloads, and also discuss the virtual machine managememt implications of each combination.

472

Rate-Equation -based model for describing the resonance laser ionization process

Author: Mikael Reponen[1]

Co-author: Volker Sonnenschein [2]

[1] *RIKEN*

[2] *University of Jyväskylä*

Corresponding Author: mikael.reponen@riken.jp

The nucleus perturbs the atomic energy levels of atoms and ions at the ppm level and although this is a small absolute effect it is readily probed and measured by modern laser spectroscopic methods. These methods are particularly suitable for the study of short-lived radionuclides with lifetimes as short as a few milliseconds and production rates often only a few isotopes/isomers per second in the case of resonance laser ionization spectroscopy (RIS)1. This ionization technique, which utilizes stepwise excitation of the unique atomic state of different elements, has been demonstrated to be a powerful tool for probing nuclei far from stability[2, 3].

The analysis of the RIS hyperfine spectra is typically performed with the help of relative intensities calculated using angular momentum coupling. The analysis yields the

nuclear magnetic dipole and electric quadrupole moments and changes in mean-square charge radii in a model-independent manner. However, recent studies have observed that the relative intensities of the measured hyperfine transitions do not necessarily follow the calculated intensities. This has partially been attributed to optical pumping effects which are not addressed by the typical procedures. Due to this new methods, based on modeling the complete ionization process 4, are required to restrict the number of free parameters during analysis.

Here we present a model for resonance ionization process based on rate-equations as a stepping stone towards a density matrix -based model. The software is written in C++ utilizing Boost Odeint, Eigen and Sundials -libraries. In addition, the software facilitates parallel solving capabilities through the OpenMP library.

1 T. Cocolios et al., Phys. Rev. Lett. 103,(2009), 102501

2 R. Ferrer et al., Phys.Lett. B 728, (2014), 191

3 S. Rothe et al., Nat.Commun.4, (2013), 1835.

4 S. Gheysen et al., Phys Rev C 69, (2004), 064310

473

How do particle physicists learn the programming concepts they need?

Authors: Maria Grazia Pia[1]; Peter Steinbach[2]; Stefan Kluth[3]; Thomas Schoerner-Sadenius[4]; Thomas Velz[5]

[1] *Universita e INFN (IT)*

[2] *MPI-CBG*

[3] *Max-Planck-Institut fuer Physik (Werner-Heisenberg-Institut) (D*

[4] *DESY*

[5] *Universitaet Bonn (DE)*

Corresponding Authors: maria.grazia.pia@cern.ch, steinbac@mpi-cbg.de, skluth@mpp.mpg.de, thomas.schoerner@cern.ch, thomas.velz@cern.ch

The ability to judge, use and develop code efficiently and successfully is a key ingredient in modern particle physics. Software design plays a fundamental role in the software development process and is instrumental to many critical aspects in the life-cycle of an experiment: the transparency of software design enables the validation of physics results, contributes to the effective use of human and computational resources, facilitates the evolution and the maintainability of the software over the lifetime of an experiment. However, despite the wide consensus on this appraisal in the HEP community, many concepts and techniques that are essential to working effectively in a HEP experiment are not covered during university studies, and limited opportunities are available to students to learn them after that.

We report the experience of a training program, identified as “Advanced Programming Concepts”, that introduces software concepts, methods and techniques to work effectively on a daily basis in a HEP experiment or other programming intensive fields. The program is targeted at students and young researchers involved in physics analysis and detector development, not only at core software developers of relevant scientific code bases. The presented workshop introduces basic and advanced programming techniques (principles of object-oriented design, design patterns, meta-template programming etc.), as well as elements of the software development process and project management skills. Emphasis is given to methods to work effectively with existing code and to improve it as well as to building a basis for further self-improvement in the field.

This presentation illustrates the principles and methods that shape the “Advanced Computing Concepts” training program, the knowledge base that it conveys, an analysis of the feedback received so far, and the integration of these concepts in the software development process of the experiments

as well as its applicability to a wider audience. It intends to promote a discussion in the software-oriented particle physics community on the responsibility of better preparing our young people for their work in the experiments, and on how the experiments could profit from a wider knowledge of advanced software methods and techniques.

Track 2 Session / 474

Using R in ROOT with the ROOT-R package

Author: Lorenzo Moneta[1]

Co-author: Omar Zapata Mesa [2]

[1] *CERN*

[2] *Columbia*

Corresponding Author: lorenzo.moneta@cern.ch

ROOT is a C++ data analysis framework, providing advanced statistical methods needed by the HEP experiments for analysing their data. R is a free software framework for statistical computing, which complements the functionality of ROOT, by including some of the latest tools developed by statistics and computing research groups. We will present the ROOT-R package, a module in ROOT, which allows to use from the ROOT environment R functions using the low-level R C++ API provided by R. This interface opens the possibility to use in ROOT and with data stored in ROOT objects, the very large set of statistical tools present in R. We will describe how this interface works, by converting ROOT C++ objects in R's objects, which can be passed to the R functions and then by converting the result back in ROOT objects. We will show as well examples how the R tools can be used inside ROOT, in particular by presenting a ROOT plug-in module based on the R optimization package, which can be used to minimise functions in ROOT and also for fitting.

475

New developments in the ROOT function and fitting classes

Author: Lorenzo Moneta[1]

[1] *CERN*

Corresponding Author: lorenzo.moneta@cern.ch

Several improvements have been introduced in the new version 6 of ROOT in the Math work package. We will report on the improvements in the ROOT function classes used for fitting data objects like histograms or trees. These include the usage of a new TFormula class, based on the capabilities Cling, which makes easier for the user to build complex expressions, which can be compiled on-the fly using the JIT provided by Cling. We will present also how the ROOT function classes have been now extended with the capabilities to auto-normalise, making them usable as probability density functions in fitting. We will show how the functions can be added together, to perform an extended likelihood fit of several normalised components, and also how it is possible to perform function convolution using the Fourier transforms, provided in ROOT by the FFTW package. Finally, we will present the changes in the function interfaces needed to use SIMD vectorisation to speed-up the function evaluation in fitting.

476

Clad - Automatic Differentiation Using Cling in ROOT

Authors: Lorenzo Moneta[1]; Vasil Georgiev Vasilev[1]

[1] *CERN*

Corresponding Author: lorenzo.moneta@cern.ch

Differentiation is ubiquitous in high energy physics, for instance for minimization algorithms used for fitting and statistical analysis, detector alignment and calibration, and for theoretical physics. Automatic differentiation (AD) avoids well-known limitations in round-offs and speed, which symbolic and numerical differentiation suffer from, by transforming the source code of functions.
We will present how AD can be used to compute the gradient of multi-variate functions and functor objects. We will explain approaches to implement an AD tool. We will show how LLVM, Clang and Cling (ROOT's C++11 interpreter) simplifies the creation of such a tool. We describe how the tool will be integrated within ROOT to be used by statistical tools such as RooFit and RooStats. We will demonstrate a simple proof-of-concept prototype, called clad, which is able to generate n-th order derivatives of C++ functions and other language constructs. We also demonstrate how clad can offload laborious computations from the CPU using OpenCL.

Track 8 Session / 477

High performance data analysis via coordinated caches

Authors: Christian Metzlaff[1]; Eileen Kuhn[2]; Max Fischer[2]

Co-authors: Christopher Jung [2]; Gunter Quast [2]; Manuel Giffels [2]; Thomas Hauth [3]

[1] *KIT - Karlsruhe Institute of Technology (DE)*

[2] *KIT - Karlsruhe Institute of Technology (DE)*

[3] *CERN*

Corresponding Author: max.fischer@cern.ch

With the second run period of the LHC, high energy physics collaborations will have to face increasing computing infrastructural needs.
Opportunistic resources are expected to absorb many computationally expensive tasks, such as Monte Carlo event simulation.
This leaves dedicated HEP infrastructure with an increased load of analysis tasks that in turn will need to process an increased volume of data.
In addition to storage capacities, a key factor for future computing infrastructure is therefore input bandwidth available per core.

Modern data analysis infrastructure relies on one of two paradigms:
data is kept on dedicated storage and accessed via network or distributed over all compute nodes and accessed locally.
Dedicated storage allows data volume to grow independently of processing capacities, whereas local access allows processing capacities to scale linearly.
However, with the growing data volume and processing requirements, HEP will require both of these features.

For enabling adequate user analyses in the future, the KIT CMS group is merging both paradigms:
High-throughput data is spread over a local disk layer on compute nodes,
while any data is available from an arbitrarily sized background storage.
This concept is implemented as a pool of distributed caches, which are loosely coordinated by a central service.
A Tier 3 prototype cluster is currently being set up for performant user analyses of both local and remote data.

The contribution will discuss the current topology of computing resources available for HEP user analyses.

Based on this, an overview on the KIT CMS analysis cluster design and implementation is presented. Finally, operational experience in terms of performance and reliability is presented.

Track 4 Session / 478

IgProf profiler support for power efficient computing

Authors: Filip Nybäck[1]; Giulio Eulisse[2]

Co-authors: David Abdurachmanov [3]; Jukka Nurminen [1]; Kashif Nizam Khan [4]; Ou Zhonghong [1]; Peter Elmer [5]; Tapio Petteri Niemi [4]

[1] *Aalto University*

[2] *Fermi National Accelerator Lab. (US)*

[3] *Vilnius University (LT)*

[4] *Helsinki Institute of Physics (FI)*

[5] *Princeton University (US)*

Corresponding Author: giulio.eulisse@cern.ch

In recent years the size and scale of scientific computing has grown significantly. Computing facilities have grown to the point where energy availability and costs have become important limiting factors for data-center size and density. At the same time, power density limitations in processors themselves are driving interest in more heterogeneous processor architectures. Optimizing application performance is no longer required merely to obtain results faster, but also to stay within the economic limits and constraints imposed by power hungry datacenters.

IgProf is an open-source, general purpose memory and performance profiler suite. We present the improvements we have made to permit optimizing the power efficiency of an application. This functionality builds on direct measurements of power related quantities using a newly developed IgProf module which exploits novel on-chip power monitoring capabilities (such as RAPL on new Intel processors). We also explore indirect methods which extrapolate from other measured application characteristics and processor behaviors, using CPU performance counters and processor power states.

We demonstrate the use of our tools both on small micro-benchmarks, developed to better understand problems and tuning measurements, and with complex, large-scale C++ applications, derived from millions of lines of code.

Track 4 Session / 479

Multi-VO Support in IHEP's Distributed Computing Environment

Author: Tian Yan[1]

Co-authors: Bing SUO [2]; Xianghu ZHAO [2]; Xiaomei ZHANG [2]

[1] *Institution of High Energy Physics, Chinese Academy of Science*

[2] *IHEP, CAS, China*

Corresponding Author: yant@ihep.ac.cn

For Beijing Spectrometer III (BESIII) experiment located at the Institute of High Energy Physics (IHEP), China, the distributed computing environment (DCE) has been setup and been in production status since 2012. The basic framework or middleware is DIRAC (Distributed Infrastructure with Remote Agent Control) with BES-DIRAC extensions. About 2000 CPU cores and 400 TB storage contributed by BESIII collaboration members are integrated to serve MC simulation and reconstruction jobs, as well as data transfer between sites.

Recently, this distributed computing environment has been extended to support other experiments operated by IHEP, such as Jiangmen Underground Neutrino Observatory (JUNO), Circular Electron Positron Collider (CEPC), Large High Altitude Air Shower Observatory (LHAASO), Hard X-ray Modulation Telescope (HXMT), etc. Each experiment's community naturally forms a virtual organization (VO). They have various heterogeneous computing and storage resources located at geographically distributed universities or institutions worldwide, with larger or smaller size. They wish to integrate those resources and supply to their VO's users. The already established BES-DIRAC DCE is chosen as platform to serve their requirements, with considerable multi-VO support technologies implemented.

The advantage of building DIRAC as a service is obvious in our case. First of all, DIRAC servers need dedicated hardware and expert manpower to maintain, some small VOs are not willing to afford that. Second, many universities and institutions in China are collaboration members of several experiments listed above, they would like to contribute resources to more than one experiment, a single DIRAC setup will be the easiest way to manage these resources. Therefore, Instead of setting up dedicated DCE for each experiment, operating a single setup for all the experiments is an effective way to make maximal use of shared resources and manpower while remaining flexible and open for new experiments to join.

The key point for this single setup is to design and develop a multi-VO support scheme based on BES-DIRAC framework.

The VO Management System (VOMS) plays a central role in multi-VO members and resources information management. Other services will retrieve user proxy information and resource per-VO configurations from VOMS server.

In data management subsystem, DIRAC file catalogs (DFC) with dynamic dataset functionality is extended to handle files and datasets belong to users from different VOs. StoRM SE is employed to manage the storage resources contributed by universities belong to more than two VOs.

In workload management subsystem, the pilot, job matcher and scheduler should take into consideration the VO properties of jobs, priorities of VOs in resources configurations, etc.

In frontend subsystem, a task based request submission and management system is developed. Different VO's users can specify their own workflows. It allows users from different VOs to submit jobs, manage data transferring, get monitoring and accounting information of their requests.

The experience of designing multi-VO support in IHEP's distributed computing environment may be interested and useful for other high energy physics center operating several experiments.

480

VecGeom: A new vectorised geometry library for particle-detector simulation

Author: Sandro Christian Wenzel[1]

Co-authors: Andrei Gheata [1]; Federico Carminati [1]; Gabriele Cosmo [1]; Georgios Bitzes [2]; Guilherme Lima [3]; Johannes Christof de Fine Licht [4]; John Apostolakis [1]; Laurent Duhem ; Marilena Bandieramonte [5]; Mihaly Novak [1]; Oksana Shadura [6]; Philippe Canal [7]; Raman Sehgal [8]; Rene Brun [1]; Soon Yung Jun [7]; Tatiana Nikitina [9]; Victor Daniel Elvira [7]

[1] *CERN*

[2] *National and Kapodistrian University of Athens (GR)*

[3] *FermiLab (US)*

[4] *University of Copenhagen (DK)*

[5] *Universita e INFN (IT)*

[6] *National Technical Univ. of Ukraine "Kyiv Polytechnic Institute*

[7] *Fermi National Accelerator Lab. (US)*

[8] *Bhabha Atomic Research Centre (IN)*

[9] *Gangneung-Wonju National University (KR)*

Corresponding Author: sandro.wenzel@cern.ch

A geometry modeller library is among the most important components of the software simulating the passage of particles in a detector, and many experiment simulations are currently based on the geometry implementations offered by Geant4 or ROOT. Here, we report on our effort to extend, re-engineer and evolve thes libraries in multiple directions in order to make them ready for the future challenges of computing in HEP.
This includes primarily an extended API as well as SIMD-vectorised algorithms able to efficiently handle geometry queries for multiple particles at the same time.
Secondly, we aim for a native support of the geometry module on the GPU or on mixed heterogeneous CPU/GPU platforms.
This effort is one of the essential ingredients of the Geant-Vector project that is focusing on a fine-grained multithreaded and vectorised design, propagating many particles from different events at the same time.

Our presentation will give an overview of the new vectorizsed, generic and templated geometry library "VecGeom" that accomplishes those primary goals and which also improves the overall geometry performance for current simulation frameworks. Beyond discussing new performance numbers for elementary and higher level geometry algorithms, we will show a first global evaluation of the new library on a realistic detector in form of the geometry description from the CMS experiment

481

A multi-port 10GbE PCIe NIC featuring UDP offload and GPUDirect capabilities.

Authors: Alessandro Lonardo[1]; Andrea Biagioni[2]; Davide Rossetti[3]; Elena Pastorelli[4]; Francesca Lo Cicero[4]; Francesco Simula[4]; Laura Tosoratto[2]; Michele Martinelli[4]; Ottorino Frezza[4]; Pier Stanislao Paolucci[4]; Piero Vicini[5]; Roberto Ammendola[2]

Co-authors: Gianluca Lamanna [6]; Luca Pontisso [7]; Marco Sozzi [8]

[1] *Universita e INFN, Roma I (IT)*

[2] *INFN*

[3] *nVIDIA Corp.*

[4] *INFN - Sezione di Roma*

[5] *INFN Rome Section*

[6] *INFN LNF*

[7] *INFN - Sezione di Pisa*

[8] *Sezione di Pisa (IT)*

Corresponding Author: andrea.biagioni@roma1.infn.it

NaNet-10 is a four-ports 10GbE PCIe Network Interface Card designed for low-latency real-time operations with GPU systems.

For this purpose the design includes a UDP offload module, for a fast and deterministic to clock-cyle handling of transport layer protocol, plus a GPUDirect P2P/RDMA engine for low-latency communication with nVIDIA Tesla GPU devices.
A dedicate module (Merger) can optionally process input UDP streams before data are delivered through PCIe DMA to their destination devices, e.g. coalescing
payload data from different streams according to a reconfigurable algorithm.
NaNet-10 is going to be integrated in the NA62 CERN experiment in order to assess the suitability of GPGPU systems as real-time triggers, we will report results and lessons learned in this activity.

Track 6 Session / 482

Status report of the migration of the CERN Document Server to the invenio-next package

Author: Nikos Kasioumis[1]

Co-authors: Esteban Gabancho [1]; Sebastian Witowski [1]

[1] *CERN*

Corresponding Author: esteban.jose.garcia.gabancho@cern.ch

With the expected release of invenio next stable version, CDS is preparing a 'lab' service where users will have the opportunity to experience the powerful features of the new software.

After a short introduction of Invenio next, the talk will explain the mechanisms that have been implemented to allow to run parallel services with the same content exposed from two different designs and the related constraints of this architecture. It will then focus on the main advantages of the new service for end-users.

The modern personalized collection home pages that have been designed after interviewing random users, will be detailed. Both the process to converge to a User Interface answering the major requirements and the result of the development (how a user can easily decide on the more interesting content to be displayed first) will be presented to the audience.

Many other improvements to the current CDS will also be introduced, like the advanced search clusters, the new commenting scheme heavily used by the experiments, the new auto-suggest feature looking in multiple author databases and the move of the CDS service monitoring tool to a dedicated ElasticSearch/Kibana instance.

Finally, the on-going project of creating CERN Authors profile will be over-viewed, from the mandatory disambiguation process that uniquely identifies people to the design of the most appropriate profile pages.

Track 8 Session / 483

Hardware and Software Design of FPGA-based PCIe Gen3 interface for APENet+ network interconnect system

Authors: Alessandro Lonardo[1]; Andrea Biagioni[2]; Davide Rossetti[3]; Elena Pastorelli[4]; Francesca Lo Cicero[4]; Francesco Simula[4]; Laura Tosoratto[2]; Michele Martinelli[4]; Ottorino Frezza[4]; Pier Stanislao Paolucci[4]; Piero Vicini[4]; Roberto Ammendola[2]

[1] *Universita e INFN, Roma I (IT)*

[2] *INFN*

[3] *INFN Rome and NVidia Corp. (USA)*

[4] *INFN Rome*

Corresponding Author: michele.martinelli@roma1.infn.it

The computing nodes of modern hybrid HPC systems are built using the CPU+GPU paradigm. When this class of systems is scaled to large size, the efficiency of the network connecting GPUs mesh and supporting the internode traffic is a critical factor. The adoption of a low latency, high performance dedicated network architecture, exploiting peculiar characteristics of CPU and GPU hardware, allows to guarantee scalability and a good level of sustained performances.

In the attempt to develop a custom interconnection architecture optimized for scientific computing we designed APEnet+, a point-to-point, low-latency and high-performance 3D torus network controller which supports 6 fully bidirectional off-board links.
The first release of APEnet+ (named V4), was a board based on a high end 40nm Altera FPGA that integrates multiple (6) channels at 34Gbps of raw bandwidth per direction and a PCIe Gen2 x8 host interface. APEnet+ board was the first-of-its-kind to implement a Remote Direct Memory Access (RDMA) protocol to directly read/write data from/to Fermi and Kepler NVIDIA GPUs using the Nvidia "peer-to-peer"and "GPUDirect RDMA"protocols, obtaining real zero-copy, low-latency GPU-to-GPU transfers over the network and reducing the performance bottleneck due to the costly copies of data from user to kernel space(and vice-versa).

The last generation of APEnet+ systems (V5), currently under development, is based on state-of-the-art high end FPGA, 28nm Altera Stratix V, offering a number of multi-standard fast transceivers (up to 14.4 Gbps), huge amount of configurable internal resources and hardware IP cores to support main interconnection standard protocols.
APEnet+ V5 implements a PCIe Gen3 x8 interface, the current standard protocol for high end system peripherals, in order to gain performance on the critical CPU/GPU connection and mitigate the effect of the bottleneck represented by GPUs memory access.
Furthermore the FPGA technology advancement, allowed us to integrate in V5, new off-board torus channels characterized by a target speed of 56 Gbps.
Both Linux Device Driver and the low-level libraries, have been redesigned to support the PCIe Gen3 protocol, introducing optimizations and solutions based on hardware/software co-design.

In this paper we present the architecture of APEnet+ V5 and discuss the status of APEnet+ V5 PCIe Gen3 hardware and system software design. Measures of performance in terms of latency and bandwidth, both for the local APEnet+ to CPU-GPU connection (with Kepler class GPU) and host-to-host via torus links, will also be provided.

484

Effective administration through gaining the portability of the LHC Computing Grid sites

Author: Gen Kawamura[1]

[1] *International Cloud Cooperation*

Corresponding Author: gen.kawamura@cern.ch

Grid computing enables deployments of large scale distributed computational infrastructures among different research facilities. It has been recently proposed that the Grid infrastructure be based on cloud computing. Provisioning systems and automated management frameworks using Cobbler, Rocks, Cfengine and Puppet are being successfully applied to many systems. Having implemented these new concepts in Grid sites within the Worldwide LHC Computing Grid (WLCG), they are solving many problems with huge world-wide collaborations and experiments, such as ATLAS and CMS.

Software provisioning like Cobbler and Rocks involves the process of selecting a target machine, a server, loading the different systems (bare operating systems, middlewares, and applications), and customizing individual components and configuring the cluster to make them ready for production.

The provisioning system needs a variety of actions: for example, creating or changing boot images; specifying particular hardware and network parameters; and lastly, starting the machine, its softwares, Grid middlewares and high energy physics applications. Typically, because of the complexity of the systems, system administrators will perform these tasks using various provisioning and management tools. However, there is a lack of generic control tools for their huge collaborations that can adequately integrate and automate these tasks for the large scale Grid sites and local experimental systems in the WLCG environments. Therefore, even though fundamental components between the large Grid production site and the small experimental system are similar, every administrator needs to customize and optimize these different tools and administrative codes when the site needs to be built and merged with WLCG collaborations.

In terms of resource provisioning, system management, meta-scheduling and portability of the Grid clusters, the combination of Grid and cloud resources for deploying the WLCG computing infrastructure introduces new challenges and opportunities. We demonstrate a way of effective administration and collaboration by using typical Grid middlewares and various services: globus, TorquePBS, CREAMCE, dCache, StoRM and gLite middlewares. Multiple virtual Grid clusters have been prepared and the deployments show the effectiveness of a generalized approach gaining the portability of the grid clusters.

In the present paper, we argue for integrating the concept of a localized LHC Computing Grid in a testing and deployment framework with that of the so-called Portable-Grid system. Using virtualization and revision control technologies to share generic system management codes, procedures installing grid systems, required libraries, experimental codes and bare operating systems, the remotely prepared Grid systems can be immediately deployed into virtual images and executed under suitable hypervisors installed on target host platforms and grid clusters. This approach is allowing the instant construction of localized Grid sites in a scenario based on OpenStack cloud computing. Further, it enables a new paradigm applicable also to general experiments in physics and computer science.

Track 4 Session / 485

First statistical analysis of Geant4 quality software metrics

Author: Elisabetta Ronchieri[1]

Co-authors: Francesco Giacomini [2]; Maria Grazia Pia [1]

[1] *INFN*

[2] *INFN CNAF*

Corresponding Author: elisabetta.ronchieri@cnaf.infn.it

Geant4 is a widespread simulation system of "particles through matter" used in several experimental areas from high energy physics and nuclear experiments to medical studies. Some of its applications may involve critical use cases; therefore they would benefit from an objective assessment of the software quality of Geant4. The issue of maintainability is especially relevant for such a widely used, mature software system, which at the present time is the result of 20 years of development.

We performed a quantitative analysis of Geant4 software quality with emphasis on maintainability.

To evaluate the maintainability of Geant4 software, we used existing standards, such as ISO/IEC 9126, that identifies the software characteristics. Furthermore, we exploited a set of product metrics - aggregated in the program size, code distribution, control flow complexity and object-orientation metrics categories - that allows to understand the code state. By using various software metrics tools, we were able to collect a large amount of measurements of software characteristics.

In this paper, we provide a first statistical evaluation of software metrics data related to a set of Geant4 physics packages. The analysis determined what metrics are most effective at identifying risks for the considered Geant4 packages and their correlations. We also evaluated the applicability

of existing quality standards, which may derive from different application environments, to the Geant4 context.

The findings of this pilot study set the ground for further extensions to the whole of Geant4 and to other HEP software systems.

487

Background decomposition of the GERDA data with BAT

Author: Oliver Schulz[1]

[1] *MPI for Physics, Munich*

Corresponding Author: oliver.schulz@tu-dortmund.de

GERDA is an ultra-low background experiment, designed to search for the neutrinoless double beta decay of Ge-76. The main background sources of such an experiment are minute radioactive ontaminations of the experimental setup itself. Gaining a good understanding of the individual contributions to this radioactive background is vital not only for data analysis, but also guides the design of the next stage of the experiment.

The Bayesian Analysis Toolkit (BAT) was used to perform a full background decomposition of the GERDA Phase-I data. The Bayesian approach allowed the implementation of prior knowledge and the ability to handle competing models in a consistent way. It also yields a straightforward uncertainty propagation, taking the correlations between model components into account.

We describe the formulation of the analysis problem and the technical realization. The techniques described here are of general validity and interest, and have proven to be very successful.

Track 8 Session / 488

Evaluation of 'OpenCL for FPGA' for Data Acquisition and Acceleration in High Energy Physics applications

Author: Srikanth Sridharan[1]

Co-author: Niko Neufeld [1]

[1] *CERN*

Corresponding Author: srikanth.sridharan@cern.ch

The proposed upgrade for the Large Hadron Collider LHCb experiment at CERN envisages a system of 500 Data sources each generating data at 100 Gbps, the acquisition and processing of which is a challenge even for state of the art FPGAs. This challenge splits into two, the Data Acquisition (DAQ) part and the Algorithm acceleration part, the later not necessarily immediately following the former.

Looking first at the DAQ part, a Header Generator module was needed to packetize the streaming data coming in from the front-end electronics of the detectors, for easy access and processing by the servers. This necessitates FPGA architectures that not only handle the data generated by the experiment in real-time but also dynamically adapt to potential inadequacies of other components, such as the network and PCs, while ensuring system stability and overall data integrity. Since the

data source has no flow control, this module needs to modify the stream data by dropping datasets in a controlled fashion in the event of receiving a back pressure signal from the downstream modules. Also needed was a front-end source emulator capable of generating the various data patterns, that can act as a test bed to validate the functionality and performance of the Header Generator. Such a system was earlier designed and realized in VHDL. The results from this were presented as a paper, 'Dynamically Adaptive Header Generator and Front-End Source Emulator for a 100 Gbps FPGA based DAQ'presented at the IEEE Real-Time Conference earlier in 2014 (RT2014).

While this process has been traditionally carried out using hardware description languages (HDLs), the possibility exists of using OpenCL to design a DAQ system. This has the potential to simplify development for physicists using the tools, who are more familiar with traditional software as opposed to HDLs, so they can understand the system and make modifications in the future. This is challenging due to fact that the OpenCL language is designed for Parallel Processing and not really targeted at real-time DAQ and there are major challenges in representing the cycle-accurate data acquisition and processing system in OpenCL. However, OpenCL for FPGAs may be applicable from a high level synthesis perspective. Achieving this will enable the movement of the entire FPGA design flow for High Energy Physics applications to OpenCL, rather than just the algorithm acceleration portion that involves parallel processing.

For the algorithm acceleration part, the Hough transformation will be implemented in OpenCL. This is a method to reconstruct lines from points in 2D/3D space and can be used to identify particle tracks from hits in the VELO detector elements. Variations of this algorithm are also used for feature identification on the data from other detectors too.

This work explores the feasibility of implementing Data Acquisition and Processing system on OpenCL and evaluates the performance of this OpenCL implementation with the HDL based implementation. Development is using the Altera OpenCL compiler for FPGA.

This work was is funded under ICE-DIP, a European Industrial Doctorate project in the European Community's 7th Framework programme Marie Curie Actions under grant PITN-GA-2012-316596.

489

A Virtual Geant4 Environment

Author: Go Iwai[1]

[1] *KEK*

Corresponding Author: go.iwai@kek.jp

We describe the development of an environment for Geant4 consisting of the application and data that enables users a faster and easier way to access the Geant4 applications without having to download and build the software locally. The environment is platform neutral and offers the users near-real time performance. The environment consists of data and Geant4 libraries built using the LLVM tools which can then result in bitcode that can be embedded in HTML and accessed via the browser. The bitcode is downloaded to the local machine via the browser and can then be configured by the user.

This approach provides a way of minimising the risk of leaking potentially sensitive data used to construct the Geant4 model and has application in the medical domain for treatment planning.

We describe several applications that have used this approach and describe the performance of applications built with this approach compared to native applications. We also describe potential user communities that could benefit from this approach.

Track 6 Session / 490

Active job monitoring in pilots

Author: Eileen Kuhn[1]

Co-authors: Andreas Petzold [1]; Christopher Jung [1]; Manuel Giffels [1]; Max Fischer [1]

[1] *KIT - Karlsruhe Institute of Technology (DE)*

Corresponding Author: manuel.giffels@cern.ch

Recent developments in high energy physics (HEP) including multi-core jobs and multi-core pilots require data centres to gain a deep understanding of the system to correctly design and upgrade computing clusters. Especially networking is a critical component as the increased usage of data federations relies on WAN connectivity and availability as a fallback to access data. The specific demands of different experiments and communities, but also the need for identification of misbehaving batch jobs requires an active monitoring.

Existing monitoring tools are not capable of measuring fine-grained information at batch job level. This complicates network-specific scheduling and optimisations. In addition, pilots add another layer of abstraction. They behave like batch systems themselves by managing and executing payloads of jobs internally. As the original batch system has no access to internal information about the scheduling process inside the pilots, there is an unpredictable number of jobs being executed. Therefore, the comparability of jobs and pilots cannot be ensured to predict runtime behaviour or network performance. Hence, the identification of the actual payload is of interest.

At the GridKa Tier 1 centre a specific monitoring tool is in use, that allows the monitoring of network traffic information at batch job level. A first analysis using machine learning algorithms showed the relevance of the measured data, but indicated a possible improvement by subdividing pilots into separate jobs.

This contribution will present the current monitoring approach and will discuss recent efforts and importance to identify pilots and their substructures inside the batch system. It will also show how to determine monitoring data of specific jobs from identified pilots. Finally, the approach is evaluated and adapted to the former analysis and the results are presented.

Track 6 Session / 491

Getting prepared for the LHC Run2: the PIC Tier-1 case

Author: Jose Flix Molina[1]

Co-authors: Alexey Sedov [2]; Andreu Pacheco Pages [3]; Antonio Perez-Calero Yzquierdo [1]; Bruno Rodriguez [4]; Carlos Acosta Silva [2]; Elena Planas [5]; Esther Accion Garcia [2]; Fernando Lopez Munoz [2]; Jordi Casals Hernandez [6]; Manuel Delfino Reznicek [2]; Marc Caubet Serrabou [2]; Maria Del Carmen Porto Fernandez [7]; Ricard Cruz [4]; Vanessa Acin Portella [2]

[1] *Centro de Investigaciones Energ. Medioambientales y Tecn. - (ES*

[2] *Universitat Autònoma de Barcelona (ES)*

[3] *Institut de Física d'Altes Energies - Barcelona (ES)*

[4] *UAB/PIC*

[5] *PIC*

[6] *U*

[7] *Unknown*

Corresponding Author: jose.flix.molina@cern.ch

The LHC experiments will collect unprecedented data volumes in the next Physics run, with high pile-up collisions resulting in events which require a more complex processing. The collaborations have been asked to update their Computing Models to optimize the use of the available resources in

order to cope with the Run2 conditions, in the midst of widespread funding restrictions. The changes in computing for Run2 represent significant efforts for the collaboration, as well as significant repercussions on how the WLCG sites are built and operated.

This contribution focuses on how these changes have been implemented and integrated in the Spanish WLCG Tier-1 center at Port d'Informació Científica (PIC), which serves ATLAS, CMS and LHCb experiments. The approach to adapt a multi-VO site to the new requirements while maintaining top reliability levels for all the experiments, is described. The main activities covered in this contribution include setting up dCache disk-only pools together with access via HTTP/XrootD protocols to expose the most demanded data; enabling end user analysis at the center; efficient integration of multi-core job handling in the site batch system and scheduler by means of the dynamic allocation of computing nodes; implementation of dynamic high memory queues; simplification, automation and virtualization of services deployment; and setting up a dCache test environment to assess the storage management readiness against experiment workflows. In addition, innovative free-cooling techniques along with a modulation of computing power versus electricity costs have been implemented, achieving a significant reduction of the electricity costs of our infrastructure.

The work has been done to reduce the operational and maintenance costs of the Spanish Tier-1 center, in agreement with the expectations from WLCG. With the state of the optimizations currently being implemented and the work foreseen during the coming years to further improve the effectiveness of the use of the provided resources, it is expected that the resources deployed in WLCG will approximately double by the end of Run2. All of the implementations done in PIC are flexible enough to rapidly evolve following changing technologies.

492

Network technology research on IhepCloud platform

Author: Tao Cui[1]

Co-author: Yaodong Cheng [2]

[1] *IHEP(Institute of High Energy Physics, CAS,China)*
[2] *IHEP*

Corresponding Authors: cuit@ihep.ac.cn, chyd@ihep.ac.cn

Traditionally, physical computer is used to run high-performance computing jobs. There are many problems such as job interference with each other, operation system crash because of abnormal operation and low computing resource utilization. IhepCloud expects to solve the job isolation, operating system fault isolation and to improve resource utilization by computing resource virtualization. IhepCloud builds virtual infrastructure platform and provides virtual machine to high-performance computing system.
In this report, we introduce our network architecture of high-performance computing system which contained virtual computing resource. We discuss dynamic synchronization between Openstack virtual environment and physic environment such as virtual machine DNS register, virtual machine network monitor, synchronization between virtual machine and some application systems. A service is deployed to collect virtual machine's information such as virtual machine name, IP, MAC and so on. The information will be stored to a network database and be synchronized to each service. These help to transparently integrate virtual system into physical computing environment.
We describe a virtual network topology based on Openstack neutron component in IhepCloud. In this topology, OVS plugin and vlan mode is deployed to achieve L2 interconnection between virtual network and physical network. Each computing node connects to physical switch through trunk links and virtual machine directly access physic network through 802.1Q supported by OVS plugin. Openstack neutron component is only responsible for distributing addresses and supporting 802.1Q. L3 is configured in physical switch. This topology avoids the performance insufficient of pure virtual network and achieves an easy interconnection with physical network. Thus, virtual system integration will not result in large changes in computing environment.
Finally we report operation and maintenance experience in our test-bed. Performance testing about glance and ceph will be given. Testing report shows image copy to each computing node is inefficient and nova with ceph is not ready. We also introduce and analyze vxlan performance testing with physical switch.

Track 8 Session / 493

Future Computing Platforms for Science in a Power Constrained Era

Authors: David Abdurachmanov[1]; Giulio Eulisse[2]; Peter Elmer[3]; Robert Knight[3]

[1] *Vilnius University (LT)*

[2] *Fermi National Accelerator Lab. (US)*

[3] *Princeton University (US)*

Corresponding Author: giulio.eulisse@cern.ch

Power consumption will be a key constraint on the future growth of Distributed High Throughput Computing (DHTC) as used by High Energy Physics (HEP). This makes performance-per-watt a crucial metric for selecting cost-efficient computing solutions. For this paper, we have done a wide survey of current and emerging architectures becoming available on the market including x86-64 variants, ARMv7 32-bit, ARMv8 64-bit, Many-Core and GPU solutions, as well as newer System-on-Chip (SoC) solutions. We compare performance and energy efficiency using an evolving set of standardized HEP-related benchmarks and power measurement techniques we have been developing. We evaluate the potential for use of such computing solutions in the context of DHTC systems, such as the Worldwide LHC Computing Grid (WLCG).

494

Storage solutions for a production-level cloud infrastructure

Authors: Eufemia Tinelli[1]; Fabrizio Ventola[1]; Giacinto Donvito[2]; Marica Antonacci[2]; Roberto Valentini[2]; Stefano Nicotri[2]; Vincenzo Spinoso[3]

[1] *UNIBA*

[2] *INFN-Bari*

[3] *Universita e INFN (IT)*

Corresponding Author: giacinto.donvito@ba.infn.it

At INFN-Bari we have set-up an OpenStack-based cloud infrastructure in the framework of a publicly funded project, PRISMA, aimed at implementing a fully integrated PaaS+IaaS platform to provide services in the field of public administration and scientific data analysis. The IaaS testbed currently consists of 25 compute nodes providing in total almost 600 physical cores, 3 TB of RAM, 400 TB of storage (disks). Connectivity is ensured through 10Gbit/s ethernet connection. Both the back-end (MySQL database and RabbitMq message broker) and the core services (nova, keystone, glance, neutron, etc.) have been configured in high-availability using HA clustering techniques. The full capacity available in the early 2015 will provide 2000 cores and 8 TB of RAM.
In this work we present the storage solutions that we are currently using as backend for our production cloud services.

Storage-as-service is implemented in Openstack by the Block Storage project, Cinder, and the Object Storage project, Swift. Selecting the right software to manage the underlying backend storage for these services is very important, and decisions can depend on many factors, not only merely technical, but also economic: in most cases they result from a trade-off between performance and costs.

Many operators use separate compute and storage hosts. We have decided not to follow this mainstream trend, aiming at the best cost-performance scenario: for us, it makes sense to run compute and storage on the same machines, since we want to be able to dedicate as many of our hosts as possible to running instances. Therefore, each compute node is configured with a significant amount of disk space and a distributed file system (GlusterFS and/or Ceph) ties the disks from each compute

node into a single file-system.
In this case, the reliability and stability of the shared file-system is critical and defines the effort to maintain the compute hosts: tests have been performed to asses the stability of the shared file-systems changing the replica factor. For example, we have tested performance and reliability of GlusterFS in both replica 2 or 3.

We currently use CEPH distributed file system for:
- storage of the running VMs enabling the live migration,
- storage of the virtual images (as primary Glance image store),
- implementation of several of the Cinder backends for block devices.

We have decided to enable Ceph as additional backend of Cinder in order to compare features, reliability and performances of the two solutions.
Our interest in Ceph derives also from the possibility of consolidating the infrastructure overall backend storage into a unified solution. To this aim, we are currently testing Ceph to run Virtual Machines, both using RBD and Ceph-FS protocols, and to implement object storage.

In order to test scalability and performance of the deployed system we used test cases which are derived from the typical pattern of storage utilization.

The testing tools are standard software widely used for this purpose such as: iozone and/or dd for block storage and specific benchmarking tools like Cosbench, swift-bench and ssbench for object storage. Using different tools for testing the file-system and comparing their results with the observation of the real test case, is also a good possibility for verifying the reliability of the benchmarking tools.

Throughput tests have been planned and conducted on the two system configurations in order to understand the performance of both storage solutions and their impacts on applications, aiming at achieving the better SLA and end-users experience.

Implementing our cloud platform, we have also focused on providing transparent access to data using standardized protocols (both de-iure and de-facto standards). In particular, Amazon-compliant S3 and CDMI (Cloud Data Management Interface) interfaces have been installed on top of the Swift Object Storage, in order to promote interoperability at PaaS/SaaS levels also.

Data is important for businesses of all sizes. Therefore, one of the most common user requirement is the possibility to backup data, in order to minimize their loss, stay compliant, and preserve data integrity. Implementing this feature is particularly challenging when the users come from the scientific communities that produce huge quantities of heterogeneous data and/or can have strict constraints.
An interesting feature of the Swift Object Storage is the geographic replica that can be used in order to add a disaster-recovery feature to the set of data and services exposed by our infrastructure.
Also Ceph provides a similar feature: the geo-replication through RADOS gateway.
Therefore, we have installed and configured both a Swift global cluster and a Ceph federated cluster, distributed on three different geographic sites. Results of the performance tests conducted on both clusters are presented along with a description of the parameters tuning that has been performed for optimization. The different replication methods implemented in the two middlewares, Swift and Ceph, are compared in terms of network traffic bandwidth, cpu and memory consumption.

Another important aspect we are taking care of is the QoS (Quality of Service) support, i.e. the capability of providing different levels of storage service optimized wrt the user application profile. This can be achieved defining different tiers of storage and setting parameters like how many I/Os the storage can handle, what limit it should have on latency, what availability levels it should offer and so on.
Our final goal is also to set-up a (semi-)automated system that is able of self-optimising. Therefore we are exploring the cache tiering feature of Ceph, that handles the migration of data between the cache tier and the backing storage tier automatically. Results of these testing activities are shown in this presentation.
In this work we show results achieved in terms of performance and functionalities of INFN-Bari cloud infrastructure while supporting scientific use cases as LHC experiments.

495

A Grid and cloud computing farm exploiting HTCondor

Authors: Alessandro Italiano[1]; Domenico Diacono[1]; Giacinto Donvito[1]; Roberto Valentini[1]; Vincenzo Spinoso[2]

[1] *INFN-Bari*

[2] *Universita e INFN (IT)*

Corresponding Author: alessandro.italiano@ba.infn.it

This work presents the result of several tests that demonstrate the capabilities of HTCondor as batch system for a big computing farm serving both LHC use cases and others scientists.
The HTCondor testbed hosted at INFN-Bari is made of about 300 nodes and 15'000 CPU slots, and meant to sustain about 50'000 job in the queue. The computing farm is used both from Grid users of many VOs (HEP, Bioinformatics, Astrophysics, etc) and from local users, that come from complete different environments: meteorologists, HEP, bioinformatics, humanities, computational chemistry, remote sensing, statistics, etc. These very different use-cases express a wide range of requirements to the batch system.
For instance, the batch system must address both a huge number of very fast jobs and very large MPI based multi-CPU jobs.
The HTCondor configuration tuned in the test aims to allow several scheduling functionalities: Hierarchical FairShare, Quality of Service, priority on users, priority on group, limit on the number of jobs per user/group/queue, job age scheduling, job size scheduling, "consumable resources"scheduling. Also it has been tested the ability to manage different job types: serial job, MPI job, multi-thread jobs, interactive jobs, whole node jobs.
HTCondor has been tested as well on a HPC Cluster with GPU card and Low Latency Infiniband connection, connected to the standard HTC farm at the same time.
Another point of interest is the High Availability configuration for the master node. This possibility is indeed very useful in order to let the cluster working also in case of hardware failure of the server hosting the batch system.
Other interesting features involved in our tests are: capabilities of using ACLs on resources in general; capabilities to deal with the pre-execution and post-execution script, capabilities of Condor_rooster to manage the start-up the worker node as soon as there are job waiting in the queue, and to shutdown them as soon as they are not needed anymore.
A very new item is also instantiating computational resources in a cloud environment. We have developed a plug-in that is able to request the needed resources to our OpenStack local facility, choosing the right image and the flavour that fulfills the user requirements expressed with the ClassAD in the jobs. In this way there is the possibility to support the computational user requests exploiting cloud resources, and this will help the site admin to efficiently and dynamically share resources, among different use-cases (batch jobs, Cloud applications, etc), exploiting the flexibility of a Cloud Computing infrastructure.
Finally, results will be shown in terms of performances achieved in job submission/scheduled/executed, configuration available and behaviour in terms of scheduling features, best practices on how to configure the batch master for High Availability.
The work will also provide the results in terms of stress tests for the configuration of the CREAM Computing Element when exploiting the HTCondor cluster.
Hence, this work summarizes the capabilities of HTCondor when supporting a big computing farm made up of diverse computational resources (Grid, Local HTC Cluster and Cloud Computing).

Track 4 Session / 496

IceProd2: A Next Generation Data Analysis Framework for the IceCube Neutrino Observatory

Author: David Schultz[1]

[1] *University of Wisconsin-Madison*

Corresponding Author: david.schultz@icecube.wisc.edu

We describe the overall structure and new features of the second generation of IceProd, a data processing and management framework. IceProd was developed by the IceCube Neutrino Observatory for processing of Monte Carlo simulations and detector data, and has been a key component of the IceCube offline computing infrastructure since it was first deployed in 2006. It runs fully in user space as a separate layer on top of grid and batch systems. This is accomplished by a set of python daemons which process job workflow and maintain configuration and status information on every job before, during, and after processing. IceProd can also manage complex workflow DAGs across distributed computing grids in order to optimize usage of special resources such as GPUs.

As the second major version of IceProd, substantial improvements have been made to increase security, reliability, scalability, and ease of use. One new goal is to extend usage beyond centralized production to individual users and to facilitate large-scale individual analysis. Currently only a handful of IceCube users have access to more than one computing cluster; this version of IceProd should enable over 200 active IceCube data analyzers to transparently access distributed CPU and GPU resources. The scope of IceProd 2 also extends beyond IceCube-specific applications and can be used as a general grid computing tool.

Track 3 Session / 497

The Belle II Conditions Database

Authors: Lynn Wood[1]; Marko Bracko[2]

[1] *Pacific Northwest National Laboratory, USA*

[2] *Jozef Stefan Institute (SI)*

Corresponding Author: marko.bracko@ijs.si

The Belle II experiment, a next-generation B factory experiment at the KEK laboratory, Tsukuba, Japan, is expected to collect an experimental data sample fifty times larger than its predecessor, the Belle experiment. The data taking and processing rates are expected to be at least one order of magnitude larger as well.

In order to cope with these large data processing rates and huge data samples, the Conditions Database, which stores the time-dependent calibrations, have to be designed carefully for successful and efficient data processing and analysis by the worldwide Belle II Collaboration. The database system needs to offer fast response, has to enable storing experimental information with fine spatial and time granularity, and ensure the scalability and redundancy for robust operation.

The Conditions Database design and implementation details, together with future plans, will be presented in this talk.

Track 3 Session / 498

Architecture of a new data taking and analysis infrastructure and services for the next generation detectors of Petra3 at DESY

Authors: Marco Strutz[1]; Martin Gasthuber[2]; Steven Aplin[1]

Co-authors: Birgit Lewendel [2]; Manuela Kuhn [1]; Stefan Dietrich [1]; Uwe Ensslin [1]; Volker Guelzow [2]

[1] *DESY*

[2] *Deutsches Elektronen-Synchrotron (DE)*

Corresponding Author: martin.gasthuber@cern.ch

Data taking and analysis infrastructures in HEP have evolved during many years to a well known problem domain. In contrast to HEP, third generations synchrotron light sources, existing and upcoming free electron laser are confronted an explosion in data rates which is primarily driven by recent developments in 2D pixel array detectors. The next generation will produce data in the region upwards of 50 Gbytes per second. At synchrotrons, data was traditionally taken away by users following data taking using portable media. This will clearly not scale at all.

We present first experiences of our new architecture and services underlying by results taken from the resumption of data taking in March 2015. Technology choices were undertaking over a period of twelve month. The work involved a close collaboration between central IT, beamline controls, and beamline support staff. In addition a cooperation was established between DESY IT and IBM to include industrial research and development experience and skills.

In technological terms the next generation detectors exceeds current generations in order of magnitudes by data volume, -rate as well as complexity at an exponential growth. We are challenging unpredictable data access demands, computing platform and OS version integration (i.e. Windows), but still requiring acceptable bandwidth for DAQ rate at the final storage system.

Our approach integrates HPC technologies for storage systems and protocols. In particular, our solution uses a single filesystem instance with a multiple protocol access, while operating within a single namespace - ubiquitous NFS & SMB access to same repository. We are targeting a system supporting distributed parity, decent erasure codes - to allow very fast rebuilds and a high availability level for the capacity disk based resources, multiple cluster, asynchronous file replication, high speed networking - Infiniband FDR & 10/40GE Ethernets, storage class memory (SSD & FlashDIMM) to run the high IOPS burst buffering layer within the architecture.

Track 5 Session / 499

Designing Computing System Architecture and Models for the HL-LHC era

Authors: Brian Paul Bockelman[1]; Frank Wuerthwein[2]; Lothar A.T. Bauerdick[3]; Matevz Tadel[2]; Peter Elmer[4]; Stephen Gowdy[3]

[1] *University of Nebraska (US)*

[2] *Univ. of California San Diego (US)*

[3] *Fermi National Accelerator Lab. (US)*

[4] *Princeton University (US)*

Corresponding Author: stephen.gowdy@cern.ch

The global distributed computing system (WLCG) used by the Large Hadron Collider (LHC) is evolving. The treatment of wide-area-networking (WAN) as a scarce resource that needs to be strictly managed is far less necessary that originally foreseen. Static data placement and replication, intended to limit interdependencies among computing centers, is giving way to global data federations building on computing centers whose maturity has increased significantly over the past decade. Different modalities for provisioning resources, including commercial clouds, will coexist with computing centers in our labs and universities. Compute resources may increasingly be shared between HEP and other fields.

By necessity today's computing system is evolving in an adiabatic fashion due to the need to support the next LHC run. In the medium and long term, however, a number of questions arise regarding the appropriate system architecture, for example: What ratio of storage to compute capabilities will be needed? How should storage be deployed geographically to maximally exploit owned, opportunistic and cloud distributed compute capabilities? What is the right mix of placement,

WAN reads, and automated caching to optimize performance? How will the reliability and scalability of the system be impacted by these choices? Can different computing models or computation techniques (map reduce, etc.) be deployed more easily with different system architectures?

In this presentation we report results from our work to simulate future distributed computing system architectures and computing models to answer these questions. When possible we also report our efforts to validate this simulation using today's computing system.

Track 6 Session / 500

Building a Tier-3 Based on ARMv8 64-bit Server-on-Chip for the WLCG

Authors: Brian Paul Bockelman[1]; David Abdurachmanov[2]; Giulio Eulisse[3]; Peter Elmer[4]; Robert Knight[5]

[1] *University of Nebraska (US)*

[2] *Vilnius University (LT)*

[3] *Fermi National Accelerator Lab. (US)*

[4] *Princeton University (US)*

[5] *Princeton University*

Corresponding Author: peter.elmer@cern.ch

Deploying the Worldwide LHC Computing Grid (WLCG) was greatly facilitated by the convergence, around the year 2000, on Linux and commodity x86 processors as a standard scientific computing platform. This homogeneity enabled a relatively simple "build once, run anywhere" model for applications. A number of factors are now driving interest in alternative platforms. Power limitations at the level of individual processors, and in aggregate in computer centers, place greater emphasis on power efficiency issues. The rise of mobile computing, based primarily on ARM processors with a different intellectual property model, has also created interest in ARM as an potential general purpose architecture for the server market.

We report our experience building a demonstrator Grid cluster using systems with ARMv8 64-bit processors, capable of running production-style workflows. We present what we have learned regarding the use of both application and Open Science Grid (OSG) software on ARMv8 64-bit processors, as well as issues related to the need to manage heterogeneous mixes of x86_64 and ARMv8 64-bit processors in a Grid environment. We also report the hardware experience we have gained while building the cluster.

501

Building a bridge between cloud storage and GPFS

Author: Andreas Petzold[1]

Co-author: Jos Van Wezel [1]

[1] *KIT - Karlsruhe Institute of Technology (DE)*

Corresponding Author: andreas.petzold@cern.ch

The possibilities of cloud storage for use in HEP computing have been the topic of many studies and trials. The typical cloud storage values, easily accessible and expandable, relatively cheap and with a light weight interface have become available for local storage as well. Initially as part of larger environments like Open Nebula or OpenStack Swift, vendors now offer value storage with integrated object based interfaces. Although cloud storage itself can be accessed with little effort, the integration with existing storage is done at the application level. However, this approach ties the storage to the application and more importantly, is not usable for all applications because these require regular POSIX semantics. The German Tier 1 GridKa as well as the Large Scale Data Facility (LSDF) at KIT, uses GPFS as the underlying file system for most of its storage. In recent versions GPFS offers a built-in rule and policy engine that allows complex data management operations including migrating data to external storage systems. The policy engine can drive hierarchical storage management (HSM) in which a condition that matches a defined characteristic triggers an external event such as the copy of a file or the movement of files from and to the file system. Interfaced to GPFS, cloud storage has the potential to be used transparently for all applications.

In this paper we present the first results of experiments where data is migrated from and recalled to a GPFS system that serves as storage for an xrootd disk pool. The data moves between the GPFS and local cloud storage and as a result, the xrootd system operates transparently on disk-based virtual storage that is several times larger than the actual file system. The results of the presented solution are compared with the traditional disk to tape HSM in use at GridKa. The experiences with the automated migration will be very helpful for the design of large and cost effective storage systems.

Track 1 Session / 502

The ALICE High Level Trigger, status and plans

Author: Mikolaj Krzewicki[1]

[1] *Johann-Wolfgang-Goethe Univ. (DE)*

Corresponding Author: mikolaj.krzewicki@cern.ch

The ALICE High Level Trigger (HLT) is an online reconstruction, triggering 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. These data, instead of raw clusters, are used in the subsequent processing and storage, resulting in a compression factor of around 4. Track finding is performed using a cellular automaton and a Kalman filter algorithm on GPGPU hardware, where both CUDA and OpenCL technologies can be used interchangeably.
The ALICE upgrade requires further development of online concepts to include detector calibration and stronger data compression. The current HLT farm will be used as a test bed for online calibration and both synchronous and asynchronous processing frameworks already before the upgrade, during Run 2.

Track 2 Session / 503

Geant4 Version 10 Series

Authors: Makoto Asai[1]; Marc Verderi[2]

[1] *SLAC National Accelerator Laboratory (US)*

[2] *Ecole Polytechnique (FR)*

Corresponding Author: asai@slac.stanford.edu

The Geant4 Collaboration released a new generation of the Geant4 simulation toolkit (version 10.0) in December 2013, and continues to improve its physics, computing performance and usability. This presentation will cover the major improvements made since version 10.0. The physics evolutions include improvement of the Fritiof hadronics model, extension of the INCL++ model to higher energy, extension of the data-driven low-energy neutron model approach to proton and light ions, extension of radioactive decay modes to include proton emission and double beta decay, introduction of phonon physics, a first implementation of particle-matter interactions dealing with crystal lattice structure, improvements of multiple-scattering for displacement at volume boundary, and extension of low energy electromagnetic processes in DNA scale. Extension and improvement of the unified solid library will provide more functionality and better computing performance. The continued effort to reduce memory consumption per thread allows for massive parallelism of large applications in the multithreaded mode. Toolkit usability is improved with an evolved real-time visualization in multithreaded mode, a new web-based visualization driver, enhancements in histogramming tool in particular in multithreaded mode, and additional APIs for easier integration with Intel TBB and MPI. In addition, the Collaboration has extended its platform support to the Xeon Phi coprocessors. A redesigned physics validation tool based on modern web technologies will ease users' access to routinely performed validations. We will also discuss our short and long term development perspectives.

504

Processing of data from innovative parabolic strip telescope.

Author: Vladislav Kosejk[1]

Co-authors: Goce Chadzitaskos [1]; Josef Novy [2]

[1] *Czech Technical University Department of Physics*

[2] *Czech Technical University (CZ)*

Corresponding Author: kosejvla@fjfi.cvut.cz

This paper presents innovative telescope design based on usage of parabolic strip as objective.

Isaac Newton was the first one to solve problem of chromatic aberration, which is caused by difference in refractive index in lens. This problem was solved by new kind of telescope with mirror used as objective.
There are many different kind of telescopes.
The most basic one is lens telescope. This kind of telescope uses set of lenses.
The next type is mirror telescope. Objective of such telescope can be concave mirror, spherical parabolic mirror or hyperbolically shaped mirror. Lens speed is depending directly on surface of the mirror.
Both kinds can be combined. Such telescope composes of at least two mirrors and set of lenses. Light is bounced of the primary mirror to the secondary one and then to the lens system. This type has smaller sizes, but also smaller lens speed.

The telescope design presented in this paper have parabolic strip as objective.
Observed objects are projected as line in picture plane. Each line, which size is equal to size of strip, corresponds to sum of intensities of light coming perpendicular to objective from observed object. Series of pictures taken with different rotation and processed by special reconstruction algorithm is

needed to get 2D pictures. The telescope can be also used for fast detection of objects. The rotation and multiple pictures are not need in this mode, just one picture of mirror is needed.

The greatest advantage of the parabolic strip telescope is its simpler and cheaper construction. Just two holders with cut-out for parabolic strip and strip from reflective flexible material are needed to create the objective. Tension in the strip material guarantees stability and precision of the reflective surface shape.
Main idea of reconstruction is based on the same principle as CT (Computer Tomography); reconstruction of 2D image by usage of Radon transformation principle from 1D samples. Four algorithms were implemented and tested.
The most basic one is summation algorithm. It is based on sequential summation of projections (adjusted pictures taken during rotation of the telescope). It algorithm is very fast and very useful for verification of basic functionality of the whole setup.
Multiplicative algorithm is improved version of summation algorithm. It is based on mutual multiplication of picture matrixes. Main advantage of this algorithm is primarily in ability to easily clean dark point from final matrix and thus suppress noise in the reconstructed 2D picture. Disadvantage is complicated weighting of matrixes during multiplication.
The third one is iterative algorithm. This approach is base on multiplication in steps. Pictures are divided into subsets. These subsets are multiplicated and until the reconstructed image is created. Idea of subsets came from CT image processing.
The last algorithm, for processing of pictures from parabolic strip telescope, is Filtered Back Projection method. It is not possible to directly use inverse radon transformation despite the fact that telescope is based on principle of Radon transformation. It is mainly because radon transformation does not have guaranteed analytic solution, thus approximation of this algorithm is used. This approximation is known as Filtered Back Projection and was originally used for reconstruction of the picture from CT. Final picture processed by this method is usually of very good quality and algorithm itself is very well adjustable for particular problem thanks to wide variety of filters.

Construction of whole telescope is relatively simple. It composes of four parts. The base is multi-axis astronomical tripod with automatic Earth movement corrections. The second part is firm holder for servomotor. This servomotor is responsible for rotation of the objective of telescope. Furthermore the CDD camera is attached to the telescope in focal point. Image from objective is projected directly on the CCD chip without any other optical elements. The support elements are made of aluminum and duralaluminum. Whole surface is anodized to black color to minimize light reflection. Mirror holders are made of hardened plastic.

The building of many observatories all around the world is possible thanks to low cost of this design. Observation of large part of the sky would be possible by such system of telescopes with big objective and angular freedom of about 15 degrees. This system is also fit for usage in satellite construction, thanks to the low weight. Best place for such telescope, is an orbit of the Earth, where all the adverse effects of the atmosphere are absent and construction of hundreds meters large objective can be made. Telescope of this magnitude can make observation of distant stars and search for exoplanets more precise and easier. Resolution of the 120 cm large parabolic strip telescope based on preliminary calculation, equal those of parabolic telescope with 150 cm diameter.
This design has yet another very unique usage. Telescope can be used for observation outside of visible spectrum, more precisely wavelength of X-Ray and gama, by exchange of parabolic strip material.

Track 1 Session / 506

Pilot run of the new DAQ of the COMPASS experiment

Author: Josef Novy[1]

Co-authors: Antonin Kveton [1]; Dmytro Levit [2]; Igor Konorov [2]; Jan Tomsa [3]; Martin Bodlak [3]; Miroslav Virius [1]; Stefan Huber [2]; Vladimir Frolov [4]; Vladimir Jary [1]

[1] *Czech Technical University (CZ)*

[2] *Technische Universitaet Muenchen (DE)*

[3] *Charles University (CZ)*

[4] *Joint Inst. for Nuclear Research (RU)*

Corresponding Author: josef.novy@cern.ch

This contribution focuses on the deployment and first results of the new data acquisition system (DAQ) of the COMPASS experiment utilizing FPGA-based event builder. The new DAQ system is developed under name RCCARS (run control, configuration, and readout system).

COMPASS is a high energy physics experiment situated at the SPS particle accelerator at CERN laboratory in Geneva, Switzerland. After two years of preparations and commissioning, the physics data taking started in 2002. The original DAQ consisted of several layers and was based on software event building paradigm. The detector frontend electronics continuously preprocess and digitize data in approximately 300000 channels, the data are readout when trigger signal arrives and are concentrated into 250 custom VME modules. These modules were connected to the event building network using 90 Slinks. The network consisted of two types of servers: readout buffers and event builders. Readout buffers served for data reception and buffering which allowed to distribute the load over the entire SPS accelerator cycle. The collected event fragments were transferred over the switched gigabit Ethernet to the event builders that assembled full events. Full events were written into the local disk space and afterwards send to the central CERN storage facility CASTOR. The system was controlled by adapted ALICE DATE package which implemented run control, event sampling, monitoring, run keeping, and configuration functionality.

Since 2002, number of channels increased from 190000 to approximately 300000, trigger rate increase from 5 kHz to 30 kHz; the average event size remained roughly 35 kB. In order to handle the increased data rates and mainly cope with aging of the system, it has been decided to develop a new DAQ system during technical shutdown of CERN accelerator in 2013-2014. Custom FPGA based data handling cards (DHC) are responsible for building of events in the new system, thus replacing the event building network. The cards have been designed in Compact AMC form factor and they feature 16 high speed serial links, 4GB of DDR3 memory, Gigabit Ethernet connection, and COMPASS Trigger Control System receiver. There are two different versions of firmware: multiplexer and switch. The multiplexer card combines data from 15 incoming Slinks send them into one outgoing, whereas the switch combines data from up to 8 multiplexers and distributes the full events to the readout engine servers equipped by spillbuffer PCI-Express cards that receive the data. DHC cards memory allows to store date of one spill and to distribute the load over the accelerator cycle period. Readout engine servers are also used for monitoring of data consistency and data quality checks.

As the DHC cards perform data flow control and event building, the software serves for configuration, run control, and monitoring. For these purposes, we have developed special software package. The main part of the software is implemented in the C++ language with the Qt framework; JavaScript, PHP, TCL, Python languages are used for support tasks. MySQL database has been selected as storage of system configuration and logs. Communication between processes in the system is implemented using the DIM library. Several types of processes are present in the system. The master is the most important process; it exchanges information and commands between user interface and slave processes. Slaves monitor and configure the DHC cards, process data, and provide interface for data access. User interface can run either in control or in viewer mode. However, only one instance of user interface may be present in the control mode. Message logger collects messages from all processes involved in the RCCARS and stores them in the database. Message browser is a graphical tool which allows to display and to filter these logs. The RCCARS is configured trough the web interface.

The RCCARS has been deployed for the pilot run starting from the September 2014. In the paper, we present performance and stability results of the new DAQ architecture, we compare it with the original system in more details.

507

Event Building Process for Time streamed data

Author: Vikas Singhal[1]

Co-authors: Subhasis Chattopadhyay [1]; Volker Friese [2]

[1] *Department of Atomic Energy (IN)*

[2] *GSI Darmstadt*

Corresponding Author: vikas.singhal@cern.ch

The Compressed Baryonic Matter (CBM) experiment at the Facility for Anti-Proton and Ion Research (FAIR)in Darmstadt, Germany, is going to produce about 1 TByte per second of raw data at an interaction rate of 10 MHz for the measurement of very rare particles. Until now, all the HEP experiments are based on traditional hardware trigger approach; therefore all simulation and reconstruction software frameworks have been working on the basis of event by event data processing. In contrast, CBM is based on a triggerless DAQ concept. This introduces a real challenge for the online computing process, which has to suppress the raw data volume by three orders of magnitude in real-time. In this respect major work has been going in the direction of changing the software framework such that it produces a continuous, time-sorted stream of digitized data packed in fixed time intervals ("time slice") which will be similar to the actual raw data output from the different detectors of the experiment. In this time-sliced data there is no event separation; consequently, the available simulation and reconstruction software cannot be straightforwardly used.

As a first step towards the reconstruction of such free-streaming data, we introduced an event-building process which tags physical events based on the time information of the raw data. The process is based on identifying and analyzing dips in the continuous data stream. It was developed on the example of the muon detecton system. The process measures the raw data rate on a nanosecond level; data between two dips indicate event candidates, which are then further analysed with respect to a number of parameters like minimum and maximum number of hits per event, event duration etc. After event definiton in software, all reconstruction algorithms, working on an event-by-event base, can be used without any modification.

In this paper, we will present and discuss the event selection process and its further development like the implementation on heterogeneous platforms using different parallel computing paradigms like OpenMP, MPI, OpenCL and CUDA, as well as its extension to the other CBM detector systems in order to reach at a global event definition.

508

The CMS BRIL Data Acquisition system

Author: Zhen Xie[1]

Co-authors: Andrea Petrucci [2]; Anne Dabrowski [2]; David Peter Stickland [1]; Luciano Orsini [2]

[1] *Princeton University (US)*

[2] *CERN*

Corresponding Author: zhen.xie@cern.ch

(BRIL) Beam Radiation Instrumentation and Luminosity is a new project within CMS. It consists of several independent sub-detectors for measuring the luminosity, monitoring the beam conditions, and the protection of CMS from serious radiation damage. It is beneficial for the project in the long run to use a single software infrastructure for data acquisition.

Similar to CMS central daq, BRIL daq manages distributed and heterogeneous subsystems.
The difference is that it has no event building and requires less data throughput.
We present the design of the BRIL DAQ system build on the XDAQ and RCMS framework in CMS.

XDAQ is the C++ framework for distributed data acquisition system. Its event-driven architecture enables large numbers of loosely coupled software components and services to exchange information in near real-time.

A BRIL DAQ component is a XDAQ process containing several applications. There are three tiers of BRIL DAQ components: sources, processors and central processors. A source is the hardware readout unit, a processor is responsible for local data aggregation and reduction, and a central processing component aggregates and further reduces data from all subsystems. Different central processors serve different purposes such as data storage, luminosity selection or online data quality monitoring.

XDAQ b2in eventing in publish/subscribe mode is used as data transport mechanism between components. The publish/subscribe model is chosen because it guarantees maximum decoupling of data senders and receivers.

RCMS is the run-control framework in CMS. It is written in Java and with built-in web and database supports. To write a RCMS component means to implement a function manager(FM) with a state machine, state transitions and event callbacks. Some state inputs are visible to the run control and monitoring web page allowing external intervention. State inputs, configurations or other parameters of choice can be stored in the database.

There are about 15 to 20 XDAQ processes running in the BRIL linux cluster. The global configuration and control of them is achieved by a single generic function manager driven by several configuration groups. The division of such groups is decided by the life cycle pattern of the processes. This design is simpler than function manager hierarchy used in the CMS central daq but is sufficient to control BRIDAQ processes which are stateless.

The fact that each XDAQ process containing a HTTP server allows the data acquisition and the control system to collaborate seamlessly via SOAP messaging.

Finally, we present some first experience of using the new BRIL daq system.

509

Discovering matter-antimatter asymmetries with GPUs

Author: Shanzhen Chen[1]

Co-authors: Andrew McNab [1]; Chris Parkes [1]; Jolanta Brodzicka [1]; Marco Gersabeck [1]

[1] *University of Manchester (GB)*

Corresponding Author: andrew.mcnab@cern.ch

The LHCb experiment has recorded the world's largest sample of charmed meson decays. The search for matter-antimatter asymmetries in charm sector requires high precision analysis and thus intensive computing. This contribution will present a powerful method to measure matter-antimatter asymmetries in multi-body decays where GPU systems have been successfully exploited. In this method, local asymmetries in phase-space distributions were explored with an unbinned approach, and the parallelisation of GPU makes this approach feasible for the first time. The performance including on GPUs on the grid will be discussed in detail. With this new method, the world's best sensitivities to particular decay channels have been achieved.

Track 2 Session / 510

The Heavy Photon Search Experiment Software Environment

Author: Norman Anthony Graf[1]

[1] *SLAC National Accelerator Laboratory (US)*

Corresponding Author: norman.graf@slac.stanford.edu

The Heavy Photon Search (HPS) is an experiment at the Thomas Jefferson National Accelerator Facility (JLab) designed to search for a hidden sector photon (A') in fixed target electroproduction. It uses a silicon microstrip tracking and vertexing detector inside a dipole magnet to measure charged particle trajectories and a fast electromagnetic calorimeter just downstream of the magnet to provide a trigger and identify electrons. As the first stage of this project, the HPS Test Run apparatus was constructed and operated in 2012 to demonstrate the experiment's technical feasibility and to confirm that the trigger rates and occupancies were as expected. The full detector is currently being installed and will be commissioned starting in November, 2014. Data taking is expected to commence in the spring of 2015.
The HPS experiment uses both invariant mass and secondary vertex signatures to search for the A'. The overall design of the detector follows from the kinematics of A'production which typically results in a final state particle within a few degrees of the incoming beam. The occupancies of sensors near the beam plane are high, so high-rate detectors, a fast trigger, and excellent time tagging are required to minimize their impact. The trigger comes from a highly-segmented lead-tungstate crystal calorimeter located just downstream of the dipole magnet. The detector was fully simulated using the flexible and performant Geant4-based program slic (abstract 445, this conference). Simulation of the readout and the event reconstruction itself were performed with the Java-based software package org.lcsim (abstract 445, this conference).. The simulation of the detector readout includes full charge deposition, drift and diffusion in the silicon wafers, followed by a detailed simulation of the readout chip and associated electronics. Full accounting of the occupancies was performed by overlaying the expected number of beam backgrounds. Track, cluster and vertex reconstruction for both simulated and real data will be described and preliminary comparisons of the expected and measured detector performance will be presented.
We will begin with an overview of the physics goals of the experiment followed by a short description of the detector design. We will then describe the software tools used to design the detector layout and simulate the expected detector performance. Finally, the event reconstruction chain will be presented.

511

Use of jumbo frames for data transfer over the WAN

Author: Christopher John Walker[1]

Co-author: Daniel Peter Traynor [1]

[1] *University of London (GB)*

Corresponding Authors: christopher.john.walker@cern.ch, d.traynor@qmul.ac.uk

Jumbo frames (with an MTU of 9000 bytes rather than the ethernet standard of 1500) have potential performance advantages for WAN transfers. Whilst many national and international research and education networks support their use, they are not widely supported at end sites. Furthermore, firewalls at some end sites block path MTU discovery leading to potential performance bottlenecks.

QMUL has two data transfer servers, one configured to use jumbo frames, and one one. In this paper, we compare data transfer speeds to a range of sites and assess the impact of jumbo frames, and path MTU blocking on data transfer rates.

512

The Pattern Recognition software for the PANDA experiment

Author: Gianluigi Boca[1]

[1] *University of Pavia and INFN, Italy*

Corresponding Author: boca@pv.infn.it

PANDA is an antiproton-proton experiment that will run at center-of-mass energies from 2.25 to 5.46 GeV at the new facility FAIR in Darmstadt, Germany. In order to achieve the broad range of physics goals of PANDA, a triggerless data acquisition and a high luminosity (20 MHz interaction rate) are necessary. This talk will concentrate on the Pattern Recognition software of the experiment. This software will use the information of all tracking devices in PANDA, namely a microvertex detector (10 ns signal time resolution), straw tube subdetectors (170 ns signal collection time), a GEM detector (few tens of ns time resolution), a fast scintillator tile system (100 ps time resolution), and several planes of gas proportional drift chambers. The absence of a hardware trigger, on one hand, gives maximum flexibility in the selection of events of interest, and on the other hand, poses severe requirements to the pattern recognition code:
1- it has to be very fast to keep up with the 20 MHz interaction rate;
2- it must be very efficient in finding tracks and very selective in rejecting ghost tracks caused by the large signal collection time of the straw systems combined with the 20 MHz interaction rate makes possible the (wrong) assignment of a hit to several physics events.
In this talk the ideas and the various algorithms explored in the Pattern Recognition to tackle those challenges will be shown.

513

The NOvA DAQ Monitor System

Author: Susan Kasahara[1]

[1] *University of Minnesota*

Corresponding Author: schubert@physics.umn.edu

The NOνA (NuMI Off-Axis ν_e Appearance) experiment is a long-baseline neutrino experiment using the NuMI
main injector neutrino beam at Fermilab and is designed to search for ν_μ ($\bar{\nu}_\mu$) to ν_e
($\bar{\nu}_e$) oscillations. The experiment consists of two detectors; both positioned 14 mrad off the beam axis: a
220 ton Near Detector constructed in an underground cavern at Fermilab and a 14 kton Far Detector constructed in Ash
River, MN, 810 km from the beam source. The detectors have similar design, and consist of planes of PVC extrusion
cells containing liquid scintillator and wavelength shifting fibers. The fiber ends are readout by Avalanche
Photodiodes (APDs). The primary task for the Data Acquisition (DAQ) system is to concentrate the data from the large
number of APD channels (340000 channels at the Far Detector, 20000 channels at the Near Detector), buffer this data
long enough to apply an online trigger, and record the selected data.

The health and performance of the DAQ system is monitored with a DAQ Monitor system which monitors 180 custom data concentrator modules, over 200 buffer farm nodes, and 40 manager nodes at two detector sites separated by 810 km. The DAQ Monitor system is based on an open source
third-party product, the Ganglia distributed monitoring system. Ganglia provides much of the functionality needed for
the DAQ Monitor system "out-of-the-box"with the ability to collect standard computing performance metrics such as CPU
usage, memory usage, and network transfer rates. Ganglia also provides the basis for displaying this

information in a web
display and storing this information in a database.

We have augmented the Ganglia system for the specific needs of
the NOνA DAQ Monitor system in the form of a custom metric client application interface which is used for purposes
of constructing and distributing custom metrics by the components which make up the DAQ system. Using the custom
metric client interface, monitored quantities specific to each DAQ component are sent at regular time intervals,
displayed on the Ganglia web display and stored in the Ganglia database. Examples of custom metrics are trigger
rates, data rates and sizes, and data corruption monitored rates. The Ganglia base has also been enhanced with a
server side
application used to read the monitored data from the Ganglia database, compare it to configurable thresholds, and
issue warnings when monitored data falls out-of-range.

The design of the NOνA DAQ Monitor system will be discussed as will experience with its deployment on the NOνA Near and
Far detectors.

Track 1 Session / 515

Performance of the NOvA Data Driven Triggering System with the full 14 kT Far Detector

Author: Andrew Norman[1]

Co-authors: Alec Habig [2]; Jan Zirnstein [3]; Martin Frank [4]; Matthew Tamsett [5]; Robert Group [6]

[1] *Fermilab*

[2] *Univ. of Minnesota Duluth*

[3] *Univeristy of Minnesota*

[4] *U*

[5] *University of Sussex*

[6] *University of Virginia*

Corresponding Author: anorman@fnal.gov

The NOvA experiment uses a continuous, free-running, dead-timeless data acquisition system to collect data from the 14 kT far detector. The DAQ system readouts the more than 344,000 detector channels and assembles the information into an raw unfiltered high bandwidth data stream. The NOvA trigger systems operate in parallel to the readout and asynchronously to the primary DAQ readout/event building chain, where they examine the full (unfiltered) detector data stream and perform complicated high level pattern recognition and reconstruction algorithms to identify rare and unique interaction topologies.

The data driven triggering systems for NOvA are unique in that they examine long contiguous time windows of the high resolution readout data and enable the detector to be sensitive to a wide range of physics interactions from those with fast, nanosecond scale signals up to processes with long delayed coincidences between hits at the ten's of millisecond scale. The trigger system is able to a true 100% live time for the detector, making it sensitive to both beam spill related and off-spill physics.

We present the performance of the trigger system with the full 14 kT NOvA detector during the first year of physics operations. We discuss the real-time and parallel computing techniques that have been used to obtain the demonstrated performance of the trigger framework. We give details

relating to the performance of key triggering algorithms and how they have been used to validate the first observations of neutrinos in the NOvA far detector as well as the challenges of implementing and simulating the performance of these algorithms in the trigger environment.

516

Integration of the Super Nova Early Warning System with the NOvA Trigger

Author: Jan zirnstein[1]

Co-author: Alec Habig [2]

[1] *University of Minnesota*

[2] *Univ. of Minnesota Duluth*

Corresponding Author: ahabig@umn.edu

The NOvA experiment, with a baseline of 810 km, samples Fermilab's upgraded NuMI beam with a Near Detector on-site and a Far Detector (FD) at Ash River, MN, to observe oscillations of muon neutrinos. The 344,064 liquid scintillator-filled cells of the 14 kton FD provide high granularity of a large detector mass and enable us to also study non-accelerator based neutrinos with our Data Driven Trigger framework. This talk/poster will focus on the real time integration of the SNEWS with the NOvA Trigger where we have set up an XML-RPC based messaging system to inject the SNEWS signal directly into our trigger. This presents a departure from the E-Mail based notification mechanism used by SNEWS in the past and allows NOvA more control over propagation and transmission timing. Message propagation time studies as well as DAQ upgrades to accommodate larger data buffers and improved data readout are discussed.

Track 4 Session / 519

Commissioning HTCondor-CE for the Open Science Grid

Author: Brian Paul Bockelman[1]

Co-authors: Brian Lin [2]; Edgar Fajardo Hernandez [3]; Jaime Frey [2]; Marian Zvada [1]; Matyas Selmeci [2]; Tim Cartwright [2]; Todd Tannenbaum [4]

[1] *University of Nebraska (US)*

[2] *University of Wisconsin-Madison*

[3] *Univ. of California San Diego (US)*

[4] *Univ of Wisconsin-Madison, Wisconsin, USA*

Corresponding Author: edgar.mauricio.fajardo.hernandez@cern.ch

The HTCondor-CE is the next-generation gateway software for the Open Science Grid (OSG). This is responsible for providing a network service which authorizes remote users and provides a resource provisioning service (other well-known gatekeepers include Globus GRAM, CREAM, Arc-CE, and Openstack's Nova). Based on the venerable HTCondor software, this new CE is simply a highly-specialized configuration of HTCondor. It was developed and adopted to provide the OSG with a more flexible, scalable, and easier-to-manage gateway software. This software does not exist in a vacuum: to deploy this gateway across the OSG, we had to integrate it with the CE configuration, deploy a corresponding information service, coordinate with sites, and overhaul our documentation.

520

Timing in the NOvA detectors with atomic clock based time transfers between Fermilab, the Soudan mine and the Nova Far detector.

Author: Andrew Norman[1]

Co-authors: Alec Habig [2]; Evan Niner [3]

[1] *Fermilab*

[2] *Univ. of Minnesota Duluth*

[3] *Indiana University*

Corresponding Author: anorman@fnal.gov

The NOvA experiment uses a GPS based timing system to both internally to synchronize the readout of the DAQ components and to establish an absolute "wall clock"reference which can be used to link the Fermilab accelerator complex with the neutrino flux that crosses the NOvA detectors. We describe the methods that were used during the commissioning of the NOvA DAQ and Timing systems to establish the synchronization between the Fermilab beam and the NOvA far detector. We present how high precision atomic clocks were trained and transported between the MINOS and NOvA detectors during a Northern Minnesota blizzard to validate the absolute time offsets of the experiments and make the first observation of beam neutrinos in the NOvA far detector.

Track 1 Session / 522

Development of New Data Acquisition System at Super-Kamiokande for Nearby Supernova Bursts

Author: Asato Orii[1]

Co-authors: Masahiro Ikeno [2]; Tomonobu Tomura [3]; Yoshihisa Obayashi [4]; Yoshinari Hayato [5]

[1] *urn:Facebook*

[2] *High Energy Accelerator Research Organization (KEK)*

[3] *ICRR. University of Tokyo*

[4] *Department of Physics*

[5] *University of Tokyo*

Corresponding Author: asato_ori@yahoo.co.jp

Super-Kamiokande (SK), a 50-kiloton water Cherenkov detector, is one of the most sensitive neutrino detectors. SK is continuously collecting data as the neutrino observatory and can be used also for supernova observations by detecting supernova burst neutrinos.

It is reported that Betelgeuse (640ly) is shrinking 15% in 15 years (C.H.townes et al. 2009) and this may be an indication of the supernova burst. Based on the Livermore model, the simulation study predicts 30MHz neutrino events observed in the SK detector for the neutrino burst from a supernova within a few hundreds of light years.
The current SK data acquisition (DAQ) system can record only about first 20% of the events. To overcome this inefficiency, we developed a new DAQ system that records the number of hit PMTs so that we can store high-rate events and obtain a time profile of the number of neutrinos emitted at the supernova.
This system uses the outputs from the number of hits from existing electronics modules as inputs and it is synchronized to the existing

DAQ system. Therefore the data is easily checked the correlation to that from the existing electronics. The data is transferred to the computers with SiTCP, an implementation of TCP/IP stack in FPGA without CPU. Part of the data are stored in the 4GB DDR2 memory before it transferred and this makes it possible to record detailed time structure of the superonva signal. The design and the production of the new modules were completed and we tested basic functions and the interference with the existing system. The firmware for the module is prepared and now being installed in SK. We will report the development and the status of the operation.

523

MAD –Monitoring ALICE Dataflow

Author: Vasco Chibante Barroso[1]

Co-authors: Adam Wegrzynek [2]; Costin Grigoras [1]; Filippo Costa [1]

[1] *CERN*

[2] *Warsaw University of Technology (PL)*

Corresponding Author: vasco.chibante.barroso@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). Following a successful Run 1, which ended in February 2013, the ALICE data acquisition (DAQ) entered a consolidation phase to prepare for Run 2 which will start in the beginning of 2015. A new software tool has been developed by the data acquisition project to improve the monitoring of the experiment's dataflow, from the data readout in the DAQ farm up to its shipment to CERN's main computer centre. This software, called ALICE MAD (Monitoring ALICE Dataflow), uses the MonALISA framework as core module to gather, process, aggregate and distribute monitoring values from the different processes running in the distributed DAQ farm. Data are not only pulled from the data sources to MAD but can also be pushed by dedicated data collectors or the data source processes. A large set of monitored metrics (from the backpressure status on the readout links to event counters in each of the DAQ nodes and aggregated data rates for the whole data acquisition) is needed to provide a comprehensive view of the DAQ status. MAD also injects alarms in the Orthos alarm system whenever abnormal conditions are detected. The MAD web-based GUI uses WebSockets to provide dynamic and on-time status displays for the ALICE shift crew. Designed as a widget-based system, MAD supports an easy integration of new visualization blocks and also customization of the information displayed to the shift crew based on the ALICE activities.

Track 3 Session / 524

Experience in running relational databases on clustered storage

Author: Ruben Domingo Gaspar Aparicio[1]

Co-author: Miroslav Potocky [1]

[1] *CERN*

Corresponding Author: ruben.gaspar.aparicio@cern.ch

CERN IT-DB group is migrating its storage platform, mainly NetApp NAS's running on 7-mode but also SAN arrays, to a set of NetApp C-mode clusters. The largest one is made of 14 controllers and it will hold a range of critical databases from administration to accelerators control or experiment

control databases. This talk shows our setup: network, monitoring, use of features like transparent movement of file systems, flash pools (SSD + HDD storage pools), snapshots, etc. It will also show how these features are used on our infrastructure to support backup & recovery solutions with different database solutions: Oracle (11g and 12c multi tenancy), MySQL or PostgreSQL. Performance benchmarks and experience collected while running services on this platform will be also shared. It will be also covered the use of the cluster to provide iSCSI (block device) access for OpenStack windows virtual machines.

525

DataBase on Demand : insight how to build your own DBaaS

Author: Ruben Domingo Gaspar Aparicio[1]

Co-authors: David Collados Polidura [1]; Ignacio Coterillo Coz [1]

[1] *CERN*

Corresponding Author: ruben.gaspar.aparicio@cern.ch

Inspired on different database as a service, DBaas, providers, the database group at CERN has developed a platform to allow CERN user community to run a database instance with database administrator privileges providing a full toolkit that allows the instance owner to perform backup/point in time recoveries, monitoring specific database metrics, start/stop of the instance and uploading/downloading specific logging or configuration files. With about 150 instances Oracle (11g and 12c), MySQL and PostgreSQL the platform has been designed and proofed to be flexible to run different RDBMS vendors and to scale up.
Initially running on virtual machines, OracleVM, the instances are represented as objects in the management database toolset, making it independent of its physical representation. Nowadays instances run on physical servers together with virtual machines. A high availability solution has been implemented using Oracle cluster ware.
This talk explains how we have built this platform, different technologies involved, actual user interface, command execution based on a database queue, backups based on snapshots, and possible future evolution (Linux containers, storage replication, OpenStack, Puppet,···).

Track 2 Session / 526

How the Monte Carlo production of a wide variety of different samples if centrally handled in the LHCb experiment.

Authors: Alexey Zhelezov[1]; Ben Couturier[2]; Daniel Patrick O'Hanlon[3]; Federico Stagni[2]; Gloria Corti[2]; Joel Closier[2]; Marco Clemencic[2]; Michal Kreps[3]; Patrick Robbe[4]; Philippe Charpentier[2]; Vladimir Romanovskiy[5]; Zoltan Mathe[2]

[1] *Ruprecht-Karls-Universitaet Heidelberg (DE)*

[2] *CERN*

[3] *University of Warwick (GB)*

[4] *Laboratoire de l'Accelerateur Lineaire (FR)*

[5] *Institute for High Energy Physics (RU)*

Corresponding Author: gloria.corti@cern.ch

In the LHCb experiment all massive processing of data is handled centrally. In the case of simulated data a wide variety of different types of Monte Carlo (MC) events has to be produced, as each physics'analysis needs different sets of signal and background events. In order to cope with this large set of different types of MC events, of the order of several hundreds, a numerical event type identification code has been devised and is used throughout. A dedicated package contains all event

type configurations files, automatically produced from this code, and is released independently from the simulation application. The deployment of the package on the distributed production system is handled centrally via the LHCb distribution tools and newly deployed event types are registered in the Bookkeeping catalogue. MC production requests are submitted via the LHCb production request system where a dedicated customization for MC data is in place. A specific request is made using predefined models centrally prepared to reproduce various data taking periods and selecting the event type from the Bookkeeping catalogue. After formal approval the requests are automatically forwarded to the LHCb Production team that carries them out. As the data are produced in remote sites they are automatically registered to the Bookkeeping catalogue where they can be found in folders specifying the event type, simulation conditions and processing chain. The various elements in the procedure, from writing a file for an event type to retrieving the sample produced, and the conventions established to allow their interplay will be described. The choices made have allowed to automate the MC production and for experts to concentrate on their specific tasks: while the configurations for each event type are prepared and validated by the physicists and simulation software experts, the MC samples are produced transparently on a world-wide distributed system the by LHCb production team.

Track 4 Session / 527

Agile Research - Strengthening Reproducibility in Collaborative Data Analysis Projects

Authors: Andrey Ustyuzhanin[1]; Christian Peter Linn[2]; Sebastian Neubert[2]; Till Moritz Karbach[2]

[1] *ITEP Institute for Theoretical and Experimental Physics (RU)*

[2] *CERN*

Corresponding Author: sebastian.neubert@cern.ch

Reproducibility of results is a fundamental quality of scientific research. However, as data analyses become more and more complex and research is increasingly carried out by larger and larger teams, it becomes a challenge to keep up this standard. The decomposition of complex problems into tasks that can be effectively distributed over a team in a reproducible manner becomes nontrivial.
Overcoming these obstacles requires a shift in both management methodology as well as supporting technology.
The LHCb collaboration is experimenting with different methods and technologies to attack such challenges. In this talk we present a language and thinking framework for laying out data analysis projects. We show how this framework can be supported by specific tools and services that allow teams of researchers to achieve high quality results in a distributed environment. Those methodologies are based on so called agile development approaches that have been adopted very successfully in industry. We show how the approach has been adapted to HEP data analysis projects and report on experiences gathered on a pilot project.

528

Performance of the NOvA Data Acquisition System with the full 14 kT Far Detector

Author: Peter Shanahan[1]

Co-authors: Abbey Waldron [2]; Alec Habig [3]; Andrew Norman [1]; Evan Niner [4]; Holger Meyer [5]; Jonathan Paley [6]; Mark Messier [4]; Mathew Muether [7]; Steve Magill [6]; Susan Kasahara [8]

[1] *Fermilab*

[2] *Oxford/T2K*

[3] *Univ. of Minnesota Duluth*

[4] *Indiana University*

[5] *Wichita State University*

[6] *Argonne National Laboratory*

[7] *Witchita State University*

[8] *University of Minnesota*

Corresponding Author: ahabig@umn.edu

The NOvA experiment studies neutrino oscillations with 2 functionally identical detectors separated by a baseline of 810km. The 14 kT far detector in Ash River, Minnesota, comprises 344,064 channels of liquid scintillator detection cells read out via wavelength-shifting fiber into 32-channel Avalanche Photo Diodes (APD). A custom designed Front End Board (FEB) continuously digitizes and zero-supresses the output signals from each APD. The smaller near detector located at Fermilab has 20,192 channels that are readout in an identical manner. Both detectors are internally synchronized by a GPS-based timing system with maintains a readout-to-readout synchronization to better than 15.6 ns across the full detector. The timing system also provides a universal time base, which is used provide synchronization and correlation between the geographically separated detectors and the Fermilab accelerator complex.

The NOvA Data Acquisition (DAQ) system for the far detector comprises 168 powerPC-based custom computers for reading out and collating data from FEBs, a farm of 196 COTS linux nodes for buffering data for trigger decisions and event building, another 10 for dedicated DAQ functions such as run control, data logging, and DAQ system monitoring. Data is transferred between the detector and the DAQ computing via a multiple bandwidth multilayer networks and fabric routing.

The performance of the recently completed DAQ on the full near and far detectors will be reviewed. The scaling characteristics of the network data flow, event building systems, and DDS based message passing layers will be covered in detail highlighting the computing and operational challenges of bringing the full DAQ and readout system online.

529

The NOvA Data Acquisition Error Handling System

Author: Peter Shanahan[1]

Co-authors: Abbey Waldron [2]; Jonathan Paley [3]; Qiming Lu [4]

[1] *Fermilab*

[2] *S*

[3] *Argonne National Laboratory*

[4] *Fermi National Accelerator Laboratory*

Corresponding Author: shanahan@fnal.gov

The NOvA experiment studies neutrino oscillations with 2 functionally identical detectors separated by a baseline of 810km. The Data Acquisition (DAQ) system for the far detector in Ash River in Minnesota comprises more than 10,000 Front End Boards, and a cluster of 168 custom PPC-based, and 206 COTS x86 linux nodes performing a variety of functions.

An Error Handling system has been developed to facilitate operation of this expansive system, using status messages generated by DAQ applications, a dedicated "rule engine"for analyzing status messages for specific or general patterns, and an Error Handler for taking pre-defined actions based on the outcome of the message analysis.

The performance of the Error Handling system in the context of production data-taking on the near and far NOvA detectors will be presented.

530

Monitoring cloud-based infrastructures

Author: Vincenzo Spinoso[1]

Co-authors: Alessandro Bruno [1]; Giacinto Donvito [1]; Marica Antonacci [1]; Stefano Nicotri [1]

[1] *INFN*

Corresponding Authors: giacinto.donvito@ba.infn.it, vincenzo.spinoso@ba.infn.it

INFN-Bari is involved in PRISMA and RECAS, two national projects aiming respectively at setting up an OpenStack-based cloud infrastructure for the public administration and the scientific data analysis, and upgrading the computing resources to a new T1-sized infrastructure.
As Bari is also a T2 for the CMS and Alice experiments, setting up the cloud resources so that they can be used for high energy physics is one of the main goals: the PaaS+IaaS platform provided by PRISMA will be installed on the RECAS resources at the beginning of 2015, providing about 2000 cores of additional resources to play as regular worker nodes.
It is fundamental to rethink all the monitoring infrastructure, to get a new elastic, scalable and automatic setup.
In this work a new setup for monitoring the cloud resources will be shown: in particular, it allows to know the availability of the underlying IaaS infrastructure and the status of all the IaaS/PaaS services running on the OpenStack tesbed during the whole life of each virtual machine. Also, it is possible to get the history of all the sensors related to the infrastructure, together with graphs. Finally, users are provided with "monitoring as a Service" features: they can instanciate a service they wish together with a machine monitoring the service itself and showing its status to the user. The monitoring infrastructure is based on Zabbix, a powerful and flexible monitoring tool, together with OpenStack itself providing Ceilometer and APIs.
It will be shown also how Zabbix serves the monitoring purposes of the whole remaining farm, basically set up with a classic batch system, cluster file system and grid middleware. All the nodes have sensors monitoring the services they run; also the network infrastructure is monitored against topological loops, high rates of packet collisions, generic unavailability. As the network topology is known in advance, we cross the information coming from the monitoring tools to build a dynamic map of the whole farm: if a machine is moved, the dynamic map moves the server as well after some time.
Thanks to this new design, monitoring definitely helps the system administrator in providing stable services even when dealing with a new big infrastructure and new services.

Track 2 Session / 531

The GeantV project: preparing the future of simulation

Author: Federico Carminati[1]

Co-authors: Andrei Gheata [1]; Guilherme Lima [2]; Johannes Christof de Fine Licht [3]; John Apostolakis [1]; Mihaly Novak [1]; Philippe Canal [4]; Sandro Christian Wenzel [1]; Soon Yung Jun [4]; Victor Daniel Elvira [4]

[1] *CERN*

[2] *FermiLab (US)*

[3] *University of Copenhagen (DK)*

[4] *Fermi National Accelerator Lab. (US)*

Corresponding Author: federico.carminati@cern.ch

Detector simulation is consuming at least half of the HEP computing cycles, and even so, experiments have to take hard decisions on what to simulate, as their needs greatly surpass the availability of computing resources. New experiments still in the design phase such as FCC, CLIC and ILC as well as upgraded versions of the existing LHC detectors will push further the simulation requirements. Since computing resources will not increase at best, it is therefore necessary to sustain the progress of High Energy Physics and to explore innovative ways of speeding up simulation. The GeantV project aims at developing a high performance detector simulation system integrating fast and full simulation that can be ported on different computing architectures, including accelerators. After more than two years of R&D the project has produced a prototype capable of transporting particles in complex geometries exploiting micro-parallelism, SIMD and multithreading. Portability is obtained via C++ template techniques that allow the development of machine-independent computational kernels. A set of tables derived from Geant4 for cross sections and final states provides a realistic shower development and, having been ported into a Geant4 physics list, is also a basis for a performance comparison.
The talk will describe the development of the project and the main R&D results motivating the technical choices of the project. It will review the current results and the major challenges facing the project. We will conclude with an outline of the future roadmaps and major milestones for the project.

532

WLCG Monitoring, consolidation and further evolution

Author: Pablo Saiz[1]

Co-authors: Alberto Aimar [1]; Alessandra Forti [2]; Alessandro Di Girolamo [1]; Andrea Sciaba [1]; David Tuckett [1]; Edward Karavakis [1]; Hector Martin De Los Rios Saiz [3]; Ivan Antoniev Dzhunov [1]; Jacobo Tarragon Cros [1]; Julia Andreeva [1]; Lionel Cons [1]; Luca Magnoni [1]; Maarten Litmaath [1]; Marian Babik [1]; Markus Schulz [1]; Nicolo Magini [1]; Stefan Roiser [1]

[1] *CERN*

[2] *University of Manchester (GB)*

[3] *Universidad Complutense (ES)*

Corresponding Author: luca.magnoni@cern.ch

The WLCG monitoring system solves a challenging task of keeping track of the LHC computing activities on the WLCG infrastructure, ensuring health and performance of the distributed services at more than 160 sites. The current challenge consists of decreasing the effort needed to operate the monitoring service and to satisfy the constantly growing requirements for its scalability and performance. This contribution describes the recent consolidation work aimed to reduce the complexity of the system, and to ensure more effective operations, support and service management. This was done by unifying where possible the implementation of the monitoring components. The contribution also covers further steps like the evaluation of the new technologies for data storage, processing and visualization and migration to a new technology stack

534

Sharing lattice QCD data over a widely distributed file system

Authors: Osamu Tatebe[1]; Tomoteru Yoshie[1]; Toshiyuki Amagasa[1]

Co-authors: Akira Ukawa [2]; Hideo Matsufuru [3]; Hideyuki Jitsumoto [4]; Hiroaki Togawa [5]; Hiroyuki Kamano [5]; Kazumi Fukumura [6]; Ken-Ichi Ishikawa [7]; Kohtaroh Miura [8]; Mitsuhisa Sato [1]; Naoya Ukita [1]; Noriyoshi Ishii [5]; Shinji Takeda [9]; Sinya Aoki [1]; Takeshi Yamazaki [1]; Takumi Doi [10]; Tatsumi Aoyama [8]; Yasumichi Aoki [8]; Yasushi Watanabe [10]; Yoshiaki Mikami [11]; Yukiko Konno [12]

[1] *University of Tsukuba*

[2] *RIKEN Advanced Institute for Computational Science*

[3] *High Energy Accelerator Research Organization (KEK)*

[4] *Tokyo Institute of Technology*

[5] *Osaka University*

[6] *Kyoto University*

[7] *Hiroshima University*

[8] *Nagoya University*

[9] *Kanazawa university*

[10] *RIKEN*

[11] *Hitachi Solutions East Japan, Ltd.*

[12] *Hitachi Solutions East Japan,Ltd.*

Corresponding Author: yoshie@ccs.tsukuba.ac.jp

JLDG is a data-grid for the lattice QCD (LQCD) community in Japan. Several large research groups in Japan have been working on lattice QCD simulations using supercomputers distributed over distant sites. The JLDG provides such collaborations with an efficient method of data management and sharing.

File servers installed on 9 sites are connected to the NII SINET VPN called HEPnet-J/sc and are bound into a single file system with the GFarm, a grid-base file system software. Because the file system looks the same from any sites, users can do analyses (measurement of physical quantities) on a supercomputer on a site, using data generated and stored in the JLDG at a different site.

Since the official start of operation in 2008, the JLDG has been improved in various ways. Among others, the following two have drastically improved usability of the JLDG. 1) Implementation of FUSE mount which enables users to mount the JLDG file system as a unix file system. 2) Development of fast data copy sub-system between the JLDG and the HPCI Shared Storage, where the HPCI is a Japanese national project started in 2011 for constructing High Performance Computing Infrastructure.

We present a brief description of hardware and software of the JLDG, focusing mainly on the above two improvements and report performance and statistics of the JLDG. As of October 2014, 11 research groups (66 users) store their daily research data of 4.0PB including replica and 63 million files in total. Number of publications for works used the JLDG is 105. The large number of publications and recent rapid increase of disk usage convince us that the JLDG has grown up into a useful infrastructure for LQCD community in Japan.

535

Data Integrity for Silent Data Corruption in Gfarm File System

Author: Osamu Tatebe[1]

Co-author: Tomoteru Yoshie [2]

[1] *University of Tsukuba*

[2] *U*

Corresponding Author: tatebe@cs.tsukuba.ac.jp

Files in storage are often corrupted silently without any explicit error. This is typically due to file system software bug, RAID controller firmware bug, and some other reasons. Most critical issue is damaged data is read without any error. Although there are several mechanisms to detect data corruption in different layers such as ECC in disk and memory and TCP checksum, the data may be damaged. To cope with the silent data corruption, the file system level detection is effective. Btrfs and ZFS have a mechanism to detect it by adding checksum in each block. However, data replication is often required to correct the damaged data, which may waste storage capacity in local file system since it is required only for data integrity.

Gfarm file system is a distributed file system that federates storages among several institutions in wide area. Large installations include Japan Lattice Data Grid (JLDG) with 4PB storage capacity in 9 storage sites, and HPCI shared storage with 20PB storage capacity in 3 storage sites. It has file replicas to improve access performance from distant clients, and also to improve fault tolerance. The number and the locations of file replicas are specified by an extended attribute of a directory or a file.

We design the data integrity feature in Gfarm file system by automatically calculating digest like md5 or sha256 when accessing files. The file digest is calculatd at a storage node before writing to a storage when a file is created, and managed in file system metadata. It can detect data corruption even when writing to storage. The digest is also calculated when reading a file. After reading entire data of file, the read system call returns input/output error when the digest mismatches. This ensures corrupted data cannot be read. When creating a file replica, this digest check is also performed. To cope with data corruption during the network transfer from a client, it also supports digest calculation at client side to ensure end-to-end data integrity.

This paper describes a design and implementation of data integrity feature in Gfarm file system. Due to native and required feature of file replicas in Gfarm file system, the data integrity can be supported without any waste storage capacity. To reduce additional overhead for calculating digest, the digest is only calculated when accessing files sequentially. This access pattern is the most typical case in Gfarm file system. For files that is created by random access write, the digest is calculated when creating the file replica. This design enables a minimum additional overhead and enough data integrity support.

Track 8 Session / 536

Simulation of LHC events on a million threads

Authors: Doug Benjamin[1]; Taylor Childers[2]; Thomas Le Compte[2]; Tom Uram[3]

[1] *Duke University (US)*

[2] *Argonne National Laboratory (US)*

[3] *urn:Google*

Corresponding Author: john.taylor.childers@cern.ch

Demand for Grid resources is expected to double during LHC Run II as compared to Run I; the capacity of the grid, however, will not double. The HEP community must consider how to bridge this computing gap. Two approaches to meeting this demand include targeting larger compute resources, and using the available compute resources as efficiently as possible. Argonne's Mira, the fifth fastest supercomputer in the world, can run roughly five times the number of parallel processes that the ATLAS experiment typically uses on the Grid. We have ported Alpgen, a serial x86 code, to run as a parallel application under MPI on the Blue Gene/Q architecture. By analysis of the Alpgen code, we reduced the memory footprint to allow running 64 threads per node, utilizing the four hardware threads available per core on the PowerPC A2 processor. Event generation and unweighting, typically run as independent serial phases, are coupled together in a single job in this scenario, reducing intermediate writes to the filesystem. By these optimizations, we have successfully run LHC proton-proton physics event generation at the scale of a million threads, filling two-thirds of Mira.

Track 4 Session / 537

Achieving production-level use of HEP software at the Argonne Leadership Computing Facility

Authors: Doug Benjamin[1]; Taylor Childers[2]; Thomas Le Compte[2]; Tom Uram[3]

[1] *Duke University (US)*

[2] *Argonne National Laboratory (US)*

[3] *urn:Google*

Corresponding Author: tomuram@gmail.com

HEP's demand for computing resources has grown beyond the capacity of the Grid, and these demands will accelerate with the higher energy and luminosity planned for Run II. Mira, the ten petaflops supercomputer at the Argonne Leadership Computing Facility, is a potentially significant compute resource for HEP research. Through an award of fifty million hours on Mira, we have delivered millions of events to LHC Experiments by establishing the means of marshaling jobs through serial stages on local clusters, and parallel stages on Mira. We are running several HEP applications, including Alpgen, Pythia, Sherpa, and Geant4. Event generators, such as Sherpa, typically have a split workload: a small scale integration phase, and a second, more scalable, event-generation phase. To accommodate this workload on Mira we have developed two Python-based Django applications, Balsam and ARGO. Balsam is a generalized scheduler interface which uses a plugin system for interacting with scheduler software such as Condor, Cobalt, and Torque. ARGO is a workflow manager that submits jobs to instances of Balsam. Through these mechanisms, the serial and parallel tasks within jobs are executed on the appropriate resources. This approach and its integration with the PanDA production system will be discussed.

Track 2 Session / 538

4-Dimensional Event Building in the First-Level Event Selection of the CBM Experiment.

Author: Ivan Kisel[1]

[1] *Johann-Wolfgang-Goethe Univ. (DE)*

Corresponding Author: i.kisel@gsi.de

The future heavy-ion experiment CBM (FAIR/GSI, Darmstadt, Germany) will focus on the measurement of very rare probes at interaction rates up to 10 MHz with data flow of up to 1 TB/s. The beam will provide free stream of beam particles without bunch structure. That requires full online event

reconstruction and selection not only in space, but also in time, so-called 4D event building and selection. This is a task of the First-Level Event Selection (FLES).

The FLES reconstruction and selection package consists of several modules: track finding, track fitting, short-lived particles finding, event building and event selection. Since all detector measurements contain also time information, the event building is done at all stages of the reconstruction process. The input data are distributed within the FLES farm in a form of so-called time-slices, which time length is proportional to a compute power of a processing node. A time-slice is reconstructed in parallel between cores within a CPU, thus minimizing communication between CPUs. After all tracks of the whole time-slice are found and fitted in 4D, they are collected into clusters of tracks originated from common primary vertices, which then are fitted, thus identifying 4D interaction points registered within the time-slice. Secondary tracks are associated with primary vertices according to their estimated production time. After that short-lived particles are found and the full e vent building process is finished. The last stage of the FLES package is a selection of events according to the requested trigger signatures.

We describe in details all stages of the FLES package and present results of tests on many-core computer farms with up to 3000 cores, focusing mainly on parallel implementations of the track finding and the event building stages as the most complicated and time consuming parts of the package. The track finding efficiency remains stable and the processing time grows as a polynomial of second order with respect to the number of events in the time-slice. The first results of J/psi selection are presented and discussed.

Track 6 Session / 539

Monitoring Evolution at CERN

Author: Pedro Andrade[1]

Co-authors: Benjamin Fiorini [1]; Luis Pigueiras [2]; Lukasz Starakiewicz [3]; Miguel Coelho dos Santos [1]; Susie Murphy [1]

[1] *CERN*

[2] *Universidad de Oviedo (ES)*

[3] *Technische Universitaet Muenchen (DE)*

Corresponding Author: pedro.andrade@cern.ch

Over the past two years, the operation of the CERN Data Centres went through significant changes with the introduction of new mechanisms for hardware procurement, new services for cloud infrastructure and configuration management, among other improvements. These changes resulted in an increase of resources being operated in a more dynamic environment. Today, the CERN Data Centres provide over 11000 multi-core processor servers, 130 PB disk servers, 100 PB tape robots, and 150 high performance tape drives.

To cope with these developments, an evolution of the data centre monitoring tools was also required. This modernisation was based on a number of guiding rules: sustain the increase of resources, adapt to the new dynamic nature of the data centres, make monitoring data easier to share, give more flexibility to Service Managers on how they publish and consume monitoring metrics and logs, establish a common repository of monitoring data, optimize the handling of monitoring notifications, and replace the previous toolset by new open source technologies with large adoption and community support.

This talk will explain how these improvements were delivered, present the architecture and technologies of the new monitoring tools, and review the experience of its production deployment.

Track 1 Session / 545

Performance of the CMS High Level Trigger

Author: Andrea Perrotta[1]

[1] *Universita e INFN, Bologna (IT)*

Corresponding Author: andrea.perrotta@cern.ch

The CMS experiment has been designed with a 2-level trigger system. The first level is implemented using custom-designed electronics. The second level is the so-called High Level Trigger (HLT), a streamlined version of the CMS offline reconstruction software running on a computer farm. For Run II of the Large Hadron Collider, the increases in center-of-mass energy and luminosity will raise the event rate to a level challenging for the HLT algorithms. The increase in the number of interactions per bunch crossing: on average 25 in 2012, and expected to be around 40 in Run II will be an additional complication. We will present the performance of the main triggers used during the 2012 run and will also cover new approaches that have been developed since then to cope with the challenges of the new run. This includes improvements in HLT electron and photon reconstruction as well as better performing muon triggers. We will also present the performance of the improved tracking and
vertexing algorithms, discussing their impact on the b-tagging performance as well as on the jet and missing energy reconstruction.

Track 1 Session / 546

ATLAS High-Level Trigger algorithms for Run-2 data-taking

Author: Carlo Schiavi[1]

[1] *Universita e INFN Genova (IT)*

Corresponding Author: carlo.schiavi@cern.ch

Following the successful Run-1 LHC data-taking, the long shutdown gave the opportunity for significant improvements in the ATLAS trigger capabilities, as a result of the introduction of new or improved Level-1 trigger hardware and significant restructuring of the DAQ infrastructure. To make use of these new capabilities, the High-Level trigger (HLT) software has been to a large extent rewritten, introducing in its turn a plethora of new features and improved algorithms for object reconstruction. The HLT algorithms rely heavily on the offline reconstruction algorithms with the aim to have an even larger efficiency for accepted events than in Run-1.

A summary of the HLT algorithms for object reconstruction will be given with the focus on new and improved algorithms together with an outline of the strategy to combine them into a trigger menu. In addition, we will show examples of impressive code speedups and the expected trigger performance for Run-2 data-taking.

547

Welcome Address

Plenary / 548

Welcome Address

Plenary / 549

Future plan of KEK

Plenary / 550

Computing at the Belle-II experiment

Plenary / 551

Evolution of Computing and Software at LHC: from Run 2 to HL-LHC

Corresponding Author: graeme.andrew.stewart@cern.ch

Plenary / 552

Intel Corporation

Plenary / 553

Dell Inc.

Plenary / 554

Computing in Intensity Frontier Accelerator Experiments

Corresponding Author: rcg6p@virginia.edu

Plenary / 555

KAGRA and the Global Network of Gravitational Wave Detectors -Construction Status and Prospects for GW Astronomy with Data Sharing Era-

Plenary / 556

Challenges of Developing and Maintaining HEP "Community" Software

Corresponding Author: amber@slac.stanford.edu

Plenary / 557

IBM Corporation

Plenary / 558

Lenovo Corporation

Plenary / 559

Diversity in Computing Technologies: Grid, Cloud, HPC … and Strategies for Dynamic Resource Allocation

Corresponding Author: oliver.gutsche@cern.ch

Plenary / 560

Distributed Data Management and Distributed File Systems

Corresponding Author: maria.girone@cern.ch

Plenary / 561

Computing at FAIR

Corresponding Author: thorsten.kollegger@cern.ch

Plenary / 562

Expanding OpenStack community in academic fields

Corresponding Author: tom@openstack.org

Plenary / 563

Computer Security for HEP

Corresponding Author: sebastian.lopienski@cern.ch

Plenary / 564

EMC Corporation

Plenary / 565

Future of DAQ Frameworks and Approaches, and Their Evolution toward 'Internet of Things'

Corresponding Author: niko.neufeld@cern.ch

Plenary / 566

The Changing Face of Networks and Implications for Future HEP Computing Models

Corresponding Author: mernst@bnl.gov

Plenary / 567

Experiences on File Systems: Which is the best FS for you?

Corresponding Author: jakob.blomer@cern.ch

Plenary / 568

Track5 Summary

Corresponding Author: tmaeno@bnl.gov

Plenary / 569

Track7 Summary

Corresponding Author: andrew.mcnab@cern.ch

Plenary / 570

Track8 Summary

Corresponding Author: danilo.piparo@cern.ch

Plenary / 571

Track1 Summary

Corresponding Author: andrea.bocci@cern.ch

Plenary / 572

Track2 Summary

Corresponding Author: i.kisel@gsi.de

Plenary / 573

Track3 Summary

Corresponding Author: latchezar.betev@cern.ch

Plenary / 574

Track4 Summary

Corresponding Author: marco.clemencic@cern.ch

Plenary / 575

Track6 Summary

Corresponding Author: jose.flix.molina@cern.ch

Plenary / 576

Lightning Talks

Corresponding Authors: andre.philippe.sailer@cern.ch, andrew.david.lahiff@cern.ch, anna.elizabeth.woodard@cern.ch, aram.santogidis@cern.ch, christophe.denis.haen@cern.ch, dai.kobayashi@cern.ch, erekle.magradze@cern.ch, kato@hepl.phys.nagoya-u.ac.jp

Plenary / **577**

Invitation to CHEP2016

Corresponding Author: richard.mount@slac.stanford.edu

Plenary / **578**

Closing Address

Corresponding Author: sakamoto@icepp.s.u-tokyo.ac.jp

Plenary / **579**

HEP Computing: A Tradition of Scientific Leadership

Plenary / **580**

Introduction to CHEP parallel program summaries

Corresponding Author: daniele.bonacorsi@bo.infn.it