20th International Conference on Computing in High Energy and Nuclear Physics (CHEP2013)

Monday, 14 October 2013 - Friday, 18 October 2013

Amsterdam, Beurs van Berlage

Book of Abstracts

Welcome from the Organisers

The CHEP conference series has come a long way from its first beginnings in Amsterdam in 1985. But the attendee list (about a hundred people) may have come out of a line printer at the time, it did contain many illustriuous names, and many of those first participants are still with us today - and you may still find abstracts from them in this Book. Following the example set by Bob Hertzberger and Walter Hoogland, our International Conference on Computing in High Energy and Nuclear Physics (CHEP) has become a series of successful conferences bringing together the high energy and nuclear physics community, computer scientists and engineers, and information technology experts. CHEP today provides an international forum to exchange the experiences and needs of the community, and to review recent, ongoing, and future activities. After almost 30 years, the 20th CHEP2013 conference is again organized by Nikhef, the Dutch National Institute for Sub-atomic Physics, in collaboration with partners.

The Programme Committee, chaired by Daniele Bonacorsi of INFN Bologna, has done a tremendous job in reviewing, ordering and selecting all the high-quality abstracts submitted. This Book of Abstracts is a reflection of that elaborate and careful process: it contains as of the time of writing the list of accepted contributions for all tracks and all sessions. Those contributions with a designated session attached have been selected for oral presentation. The posters are organised together in a single continuous session - and we hope you can look at many of those as you browse the lunch and refreshment rooms where these will be placed. And if you particularly like a poster, pick up your voting slip and help promote the poster to a plenary lightning talk on Friday!

Once contributions have been scheduled in the time table of the Conference, this book of abstracts will reflect the actual ordering.

We hope you enjoy the Conference and look forward to seeing you in Amsterdam!

nbsp; David Groep,
nbsp; Conference Chair.

Contents

Software Engineering, Parallelism & Multi-Core / 428

The Telescope Array Fluorescence Detector Simulation on GPUs

Author: Tareq AbuZayyad[1]

[1] *University of Utah*

Corresponding Author: tareq@cosmic.utah.edu

The Telescope Array Cosmic Rays Detector located in the Western Utah Desert is used for the observation of ultra-high energy cosmic rays. The simulation of a fluorescence detector response to cosmic rays initiated air showers presents many opportunities for parallelization. In this presentation we report on the Monte Carlo program used for the simulation of the Telescope Array fluorescence detector located at the Middle Drum site. The program makes extensive use of GPU acceleration (CUDA) to achieve a 50x speed-up compared to running on a single CPU core.

The main design criteria for the simulation code is that it can be run with/without acceleration while eliminating code duplication as much as possible. This is seen as critical for long term maintainability of the source code. A CMake based build system allows the user to easily select to compile/run the code with or without CUDA, and to include external package, e.g. ROOT, dependent code in the build.

All of the physics simulation from shower development, light production and propagation with atmospheric attenuation, as well as, the realistic detector optics and electronics simulations are done on the GPU. A detailed description of the code implementation
is given, and results on the accuracy and performance of the simulation are presented as well.

Detector event reconstruction is performed using an inverse Monte Carlo method. We are in the process of porting that code to make use of the GPU based simulation. Results will be presented on this effort as well.

Event Processing, Simulation and Analysis / 408

Extending the FairRoot framework to allow for simulation and reconstruction of free streaming data

Authors: Alexey Rybalchenko[1]; Dennis Klein[1]; Florian Uhlig[2]; Mohammad Al-Turany[3]

[1] *GSI Darmstadt*

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

[3] *GSI*

Corresponding Author: mohammad.al-turany@cern.ch

The FairRoot framework is the standard framework for simulation, reconstruction and data analysis for the FAIR experiments. The framework, is designed to optimize the accessibility for beginners and developers, to be flexible and to cope with future developments. FairRoot enhances the synergy between the different physics experiments within the FAIR project. Moreover, the framework is meanwhile also used outside FAIR project by the MPD (NIKA) project at JINR Russia and the EIC project at BNL. As a first step toward simulation of free streaming data, the time based simulation was introduce to the framework. The next step is the event source simulation. This is achieved via a client server system. After digitization the so called "samplers" can be started, each sampler can read the data of the corresponding detector from the simulated files and keep it memory. The data is then made available for the reconstruction clients via the server. Such a system make it possible to develop and validate the online reconstruction algorithms. In this work, the design and implementation of this push architecture and the communication layer will be presented.

Summaries / 524

Summary of track 5

Corresponding Author: solveig.albrand@lpsc.in2p3.fr

Data Acquisition, Trigger and Controls / 124

Review of the LHCb Higher Level Trigger operations and performance during 2010-2012

Authors: Gerhard Raven[1]; Johannes Albrecht[2]; Vladimir Gligorov[3]

[1] *NIKHEF (NL)*

[2] *Technische Universitaet Dortmund (DE)*

[3] *CERN*

Corresponding Authors: johannes.albrecht@cern.ch, gerhard.raven@nikhef.nl

The LHCb experiment is a spectrometer dedicated to the study of heavy flavor at the LHC. The rate of proton-proton collisions at the LHC is 15 MHz, but resource limitations imply that only 5 kHz can be written to storage for offline analytsis. For this reason the LHCb data acquisition system – trigger – plays a key role in selecting signal events and rejecting background. In contrast to previous experiments at hadron colliders like for example CDF or D0, the bulk of the LHCb trigger is implemented in software and deployed on a farm of 20k parallel processing nodes. This system, called the High Level Trigger (HLT) is responsible for reducing the rate from the maximum at which the detector can be read out, 1.1 MHz, to the 5 kHz which can be processed offline, and has 20 ms in which to process and accept/reject each event. In order to minimize systematic uncertainties, the HLT was designed from the outset to reuse the offline reconstruction and selection code. This contribution describes the design, implementation, performance and evolution of the HLT from the initial commissioning to its present status.

Data Acquisition, Trigger and Controls / 78

Many-core applications to online track reconstruction in HEP experiments

Authors: Alessio Gianelle[1]; Donatella Lucchesi[2]; Marco Corvo[3]; Peter Wittich[4]; Ryan Rivera[5]; Silvia Amerio[6]; Stefano Gelain[6]; Stephen Poprocki[4]; Tiehui Ted Liu[7]; Wesley Ketchum[8]

[1] *Universita e INFN (IT)*

[2] *INFN Padova*

[3] *INFN*

[4] *Cornell University (US)*

[5] *Fermilab*

[6] *University of Padova & INFN*

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

[8] *Los Alamos National Laboratory*

Corresponding Authors: silvia.amerio@pd.infn.it, pw94@cornell.edu

One of the most important issues facing particle physics experiments at hadron colliders is real-time selection of interesting events for offline storage. Collision frequencies do not allow all events to be

written to tape for offline analysis, and in most cases, only a small fraction can be saved. Typical trigger systems use commercial computers in the final stage of processing. Much of the effort is focused on understanding the latency for trigger systems. In this talk we describe updates to a previous study of latencies in GPU for potential trigger applications, where we measured the latency to transfer data to/from the GPU, exploring the timing of different I/O technologies on different GPU models. Those studies, where a simplified track fitting algorithm was parallelized and run on a GPU, show that latencies of few tens of microseconds can be achieved to transfer and process packets of 4 kB of data, combining the modern Infiniband data transfer technology with direct access to GPU memory allowed by NVIDIA GPUDirect utilities. We now have expanded our latency studies to include other multi-core systems (Intel Xeon Phi and AMD GPUs, in addition to NVIDIA GPUs) and other software environments (OpenCL, in addition to NVIDIA CUDA). We also discuss the implementation of a scaled-up version of the algorithm used at CDF for online track reconstruction - the SVT algorithm - as a realistic test-case for low-latency trigger systems using new computing architectures for LHC experiments.

Poster presentations / **411**

Architectural improvements and 28nm FPGA implementation of the APEnet+ 3D Torus network for hybrid HPC systems

Authors: Alessandro Lonardo[1]; Andrea Biagioni[1]; Davide Rossetti[1]; Francesca Lo Cicero[1]; Francesco Simula[1]; Laura Tosoratto[1]; Ottorino Frezza[1]; Pier Stanislao Paolucci[1]; Piero Vicini[1]; Roberto Ammendola[2]

[1] *INFN Roma*

[2] *INFN Roma Tor Vergata*

Corresponding Author: roberto.ammendola@roma2.infn.it

Modern Graphics Processing Units (GPUs) are now considered accelerators for general purpose computation. A tight interaction between the GPU and the interconnection network is the strategy to express the full potential on capability computing of a multi-GPU system on large HPC clusters; that is why an efficient and scalable interconnect is a key technology to finally deliver GPUs for scientific HPC.
In this paper we show the latest architectural and performance improvement of the APEnet+ network fabric, a FPGA-based PCIe board with 6 fully bidirectional off-board links with 34 Gbps of raw bandwidth per direction, and X8 Gen2 bandwidth towards the host PC. The board implements a Remote Direct Memory Access (RDMA) protocol that leverages upon peer-to-peer (P2P) capabilities of Fermi and Kepler-class NVIDIA GPUs to obtain real zero-copy, low-latency GPU-to-GPU transfers. Finally we report on the development activities for 2013 focusing on the adoption of the latest generation 28 nm FPGAs and the preliminary results achieved with synthetic benchmarks exploiting the implementation of state-of-the-art signalling capabilities of PCI-express Gen3 host interface.

Poster presentations / **48**

GPU for Real Time processing in HEP trigger systems

Author: Gianluca Lamanna[1]

Co-authors: Alessandro Lonardo [2]; Andrea Biagioni [2]; Andrea Messina [1]; Davide Rossetti [3]; Francesco Simula [4]; Marco Rescigno [2]; Marco Sozzi [5]; Massimiliano Fiorini [1]; Piero Vicini [3]; Riccardo Fantechi [5]; Roberto Ammendola [6]; Stefano Giagu [2]

[1] *CERN*

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

[3] *INFN Rome Section*

[4] *Università e INFN, Roma I (IT)*

[5] *Sezione di Pisa (IT)*

[6] *INFN*

Corresponding Authors: roberto.ammendola@roma2.infn.it, gianluca.lamanna@cern.ch

We describe a pilot project for the use of GPUs (Graphics processing units) in online triggering applications for high energy physics experiments. Two major trends can be identified in the development of trigger and DAQ systems for particle physics experiments: the massive use of general-purpose commodity systems such as commercial multicore PC farms for data acquisition, and the reduction of trigger levels implemented in hardware, towards a pure software selection system (trigger-less). The very innovative approach presented here aims at exploiting the parallel computing power of commercial GPUs to perform fast computations in software both in early trigger stages and in high level triggers. General-purpose computing on GPUs is emerging as a new paradigm in several fields of science, although so far applications have been tailored to the specific strengths of such devices as accelerator in offline computation. With the steady reduction of GPU latencies, and the increase in link and memory throughputs, the use of such devices for real-time applications in high-energy physics data acquisition and trigger systems is becoming ripe.
We will discuss in details the use of online parallel computing on GPU for synchronous low level trigger with fixed latency. In particular we will show the preliminary results on a first field test in the CERN NA62 experiment. The use of GPUs in high level triggers will be also considered, the CERN ATLAS experiment (and in particular the muon trigger) will be taken as a study case of possible applications.

Facilities, Infrastructures, Networking and Collaborative Tools / **207**

Agile Infrastructure Monitoring

Authors: Ivan Fedorko[1]; Pedro Andrade[1]

Co-authors: Benjamin Fiorini [1]; Joao Ricardo Goncalves Lopes Ascenso [2]; Omar Pera Mira [3]; Sebastien Ponce [1]

[1] *CERN*

[2] *Universidade de Evora (PT)*

[3] *Valencia Polytechnic University (ES)*

Corresponding Authors: pedro.andrade@cern.ch, ivan.fedorko@cern.ch

At the present time computing centres are facing a massive rise in virtualization and cloud computing. The Agile Infrastructure (AI) project is working to deliver new solutions to ease the management of CERN Computing Centres. Part of the solution consists in a new common monitoring infrastructure which collects and manages monitoring data of all computing centre servers and associated software as well as additional environment and facilities data (e.g. temperature, power consumption, etc.).
The new monitoring system is addressing requirements for a very large scale. Performance measurement will be implemented by gathering metric data from the entire Computing Centre. Linux hosts data will be collected by improving the Lemon (LHC Era Monitoring System) client to forward metric data to a messaging layer responsible for data transport. Using the same messaging channel, other metrics data sources (windows servers, network data, non-hosts data) will also be collected on top of which different visualization and data analytics solutions will be implemented. Given the architecture similarities with the WLCG grid monitoring tools, such as the Service Availability Monitoring (SAM) system, the same AI monitoring model and technologies can also be applied to monitor grid resources.
Another important component of the new monitoring system is to directly notify system administrators and service managers about errors and problems. These situations are handled and processed by a new operations workflow, the General Notification Infrastructure (GNI). Using messaging technology for the transport of monitoring messages, GNI allows multiple entities (currently Lemon for linux servers, SCOM for windows servers, and other isolated clients) to produce monitoring notifications which are processed by an extensible number of notifications consumers. Today GNI provides a gateway to CERN event management system, part of CERN IT service management implemented in the Service-Now framework, and a notifications dashboard. In the future a notifications analysis

framework and other consumers (e.g. email SMS, etc.) will be added.
In this article, a high level architecture overview of the new monitoring infrastructure is provided. The GNI operational tools developed and deployed to monitor CERN Computing Centres (Meyrin and Wigner) are presented as well as the future plans towards large scale data analytics for CERN Computing Centres and the WLCG grid infrastructure.

Software Engineering, Parallelism & Multi-Core / **67**

GooFit: A massively-parallel fitting framework

Author: Rolf Edward Andreassen[1]

Co-authors: Brian Meadows [1]; Karen Tomko [2]; Michael Sokoloff [3]; Weeraddana De Silva [1]

[1] *University of Cincinnati (US)*

[2] *Ohio Supercomputer Center*

[3] *University of Cincinnati*

Corresponding Author: andrear@ucmail.uc.edu

We present a general framework for maximum-likelihood fitting, in which GPUs are used to massively parallelise the per-event probability calculation. For realistic physics fits we achieve speedups, relative to executing the same algorithm on a single CPU, of several hundred.

Poster presentations / **254**

AGIS: The ATLAS Grid Information System

Author: Alexey Anisenkov[1]

Co-authors: Alessandro Di Girolamo [2]; Alexei Klimentov [3]; Artem Petrosyan [4]; Danila Oleynik [4]

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

[2] *CERN*

[3] *Brookhaven National Laboratory (US)*

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

Corresponding Authors: alexey.anisenkov@cern.ch, alessandro.di.girolamo@cern.ch, alexei.klimentov@cern.ch, danila.oleynik@cern.ch, artem.petrosyan@cern.ch

In this paper we describe the ATLAS Grid Information System (AGIS), the system designed to integrate configuration and status information about resources, services and topology of the computing infrastructure used by ATLAS Distributed Computing (ADC) applications and services.

The Information system centrally defines and exposes the topology of the ATLAS computing infrastructure including various static, dynamic and configuration parameters collected both from independent sources like gLite BDII (Berkley Database Information Index), Grid Operations Centre Database (GOCDB), the Open Science Grid Information services (MyOSG), and from ATLAS specific ones like the ATLAS Distributed Data Management (DQ2) system and the ATLAS Production and Distributed Analysis (PanDA) workload management system.

Being an intermediate middleware system between clients and external information sources, AGIS automatically collects and keeps data up to date, caching information required by ATLAS, removing the source as a direct dependency for end-users, but without duplicating the source information itself: AGIS represents data objects in the way more convenient for ATLAS services, introduces additional object relations required by ATLAS applications, and exposes the data via the REST style

API and WEB front end services. For some types of information AGIS itself has become the primary repository.

We describe the evolution and new functionalities of WEB and API services implemented to integrate AGIS with ADC applications and provide user-friendly native WEB interface to manage data. We will explain how the AGIS flexibility allows the definition of new services still not integrated in production in WLCG but already used by ATLAS for specific use cases. In particular, it concerns the definitions of the FAX storage federation, which is based on XrootD storage services, and redirectors organized with a precise ATLAS specific topology, stored in AGIS. Special attention is given also to the implementation of user roles and privileges allowed to separate access within user groups using AGIS in production. Introducing various user access groups makes the Information System a sustainable way to keep a global and coherent view of the whole computing infrastructure used by ATLAS.

Poster presentations / **455**

DIRAC framework evaluation for the Fermi-LAT, CTA and LSST experiments

Author: Luisa Arrabito[1]

Co-authors: Andrei Tsaregorodtsev [2]; Johann Cohen-Tanugi [1]; Matthieu Renaud [1]; Matvey Sapunov [2]; Ricardo Graciani Diaz [3]; Stephan Zimmer [4]; Vincent Rolland [1]

[1] *LUPM Université Montpellier 2, IN2P3/CNRS*

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

[3] *University of Barcelona (ES)*

[4] *University of Stockholm*

Corresponding Author: arrabito@in2p3.fr

DIRAC (Distributed Infrastructure with Remote Agent Control) is a general framework for the management of tasks over distributed heterogeneous computing environments. It has been originally developed to support the production activities of the LHCb (Large Hadron Collider Beauty) experiment and today is extensively used by several particle physics and biology communities. Current (Fermi-LAT, Fermi-Large Area Telescope) and planned (CTA, Cherenkov Telescope Array, LSST, Large Synoptic Survey Telescope) with very large processing and storage needs, are currently investigating the usability of DIRAC in this context. Each of these use cases has some peculiarities: Fermi-LAT will interface DIRAC to its own workflow system to allow the access to the grid resources; CTA is using DIRAC as workflow management system for the Monte Carlo production on the grid; LSST is exploring DIRAC to access to heterogeneous resources, like local clusters, grid and cloud. We describe the prototype effort that we lead toward deploying a DIRAC solution for some aspects of Fermi-LAT, CTA, and LSST needs.

Poster presentations / **314**

Data Acquisition of A totally Active Scintillator Calorimeter of the Muon Ionization Cooling Experiment

Author: Ruslan Asfandiyarov[1]

Co-author: Yordan Ivanov Karadzhov [1]

[1] *Universite de Geneve (CH)*

Corresponding Authors: ruslan.asfandiyarov@cern.ch, yordan.karadzhov@cern.ch

The Electron-Muon Ranger (EMR) is a totally active scintillator detector which will be installed in the muon beam of the Muon Ionization Cooling Experiment (MICE), the main R&D project for a future neutrino factory. It is designed to measure the properties of a low energy beam composed of muons, electrons and pions, and to perform an identification on a particle by particle basis. The EMR is made up of 48 intersecting layers, each of which is made of 59 triangular bars. Wavelength shifting fibers incorporated into the bars trap and transfer light, generated by particles traversing the detector, to PMTs located at the ends of the bars. One side is read out by single-anode PHILIPS XP2972 PMTs and the other by 64-ch. HAMAMATSU R7600 PMTs.

Signals from 48 single-anode PMTs are read out by 8 fast ADCs (CAEN V1731) which store pulse shapes with 2ns time resolution. Each 64-ch. PMT is interfaced to a front-end-board which hosts a MAROC ASIC that amplifies, discriminates, and shapes all input signals. Pulse height information can be extracted at low rate and will be used during calibration and tests with cosmic rays. Fast discriminated signals from the front-end-board are directed to a piggy-back buffer board which stores all the signals created during one duty cycle of the accelerator. Six buffer boards are connected in a chain and read out by a dedicated VME card. Communication between the buffer and VME board is made through TLK chip to allow for fast data transfer. The front-end-boards also have corresponding VME readout boards which at the same time configure the MAROC ASICs. All electronics boards employ FPGA chips which allow for great customization of the detector behavior. The complete read out system of the EMR will be described, including the hardware, firmware, and results of a full system test using cosmic rays.

Poster presentations / **444**

Performance evaluation of a dCache system with pools distributed over a wide area

Authors: Eduardo Bach[1]; Rogerio Iope[1]

[1] *UNESP - Universidade Estadual Paulista (BR)*

Corresponding Author: eduardo.bach@cern.ch

Distributed storage systems have evolved from providing a simple means to store data remotely to offering advanced services like system federation and replica management. This evolution have made possible due to the advancement of the underlying communication technology, that plays a vital role in determining the communication efficiency of the distributed systems. The dCache system, which has a wide installation base in the High Energy Physics community, is a distributed data caching environment built over a large number of heterogeneous storage servers usually located at a single physical site that is able to offer to the end user a unified filesystem view of the entire data repository by means of several protocol standards. One of the key features of the dCache system is the dissociation of the file namespace of the repository from the actual location of the data files in the storage servers. This allows files to reside anywhere in the distributed system, facilitating data replication and improving data access. Another interesting feature is load balancing by means of hot-spot detection, live migration, and caching mechanisms. The latest versions of dCache are capable of providing access to data through the new Network File System version 4.1 (NFSv4.1) protocol, which extends the capabilities of its predecessors by supporting parallel I/O capabilities, thus increasing scalability and overall performance.

In this work we propose to evaluate the performance of the new dCache system implemented with NFSv4.1, with the goal of analyzing its behavior when combining distributed server pools spread over a wide area and interconnected by a dedicated optical network system. To establish reference points we will begin by analyzing dCache deployed as a central storage system, providing data to local clients through NFSv4.1. A simple local dCache system will be built, in which the server pools are basically low cost JBODs. The centralization of different storage systems into a single structure allows the reduction in the number of storage systems deployed in an enterprise datacenter, which consequently leads to cost reduction. Another interesting advantage is the decrease of the setup time for new storage servers, which basically consists of adding more JBODs to the already deployed centralized storage system. The dCache I/O performance will then be tested with Iozone, a commonly used benchmarking tool, and the behavior of the system will be evaluated in response

to increasing loads. During the tests we will evaluate system scalability (by adding new pools) and fault tolerance (by detaching some of the pool servers).

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / 474

Integrating multiple scientific computing needs via a Private Cloud Infrastructure

Authors: Dario Berzano[1]; Riccardo Brunetti[2]; Sara Vallero[3]; Stefano Bagnasco[3]; Stefano Lusso[4]

[1] *CERN*

[2] *Unknown*

[3] *Universita e INFN (IT)*

[4] *INFN-TO*

Corresponding Author: svallero@to.infn.it

In a typical scientific computing centre, diverse applications coexist and share a single physical infrastructure. An underlying Private Cloud infrastructure eases the management and maintenance of such heterogeneous applications (such as multipurpose or application-specific batch farms, Grid sites catering to different communities, parallel interactive data analysis facilities and others), allowing to dynamically and efficiently allocate resources to any application, precisely tailoring the virtual machines according to the applications' requirements. Furthermore, the maintenance of large deployments of complex and rapidly evolving middleware and application software is eased by the use of virtual images and contextualization techniques; for example, rolling updates can be performed easily and minimizing the downtime. In this contribution we describe the Private Cloud infrastructure at the INFN-Torino Computer Centre, that hosts a full-fledged WLCG Tier-2 centre, a dynamically expandable PROOF-based Interactive Analysis Facility for the ALICE experiment at the CERN LHC and several smaller scientific computing applications. The private cloud building blocks include the OpenNebula software stack, the GlusterFS filesystem (used in two different configurations for worker- and service-class hypervisors) and the OpenWRT Linux distribution (used for network virtualization); a future integration into a federated higher-level infrastructure is made possible by exposing commonly used APIs like EC2 and OCCI. In this talk we describe the operational experience and the latest developments in the integration with evolving experiment Computing Models.

Data Acquisition, Trigger and Controls / 12

The H.E.S.S. Phase II Data Acquisition System

Author: Arnim Balzer[1]

Co-authors: Anton Lopatin [2]; Christian Stegmann [3]; Daniel Göring [4]; Mathieu de Naurois [5]; Matthias Füßling [2]; Michael Gajdus [6]; Philipp Wagner [6]; Thomas Murach [6]; Ullrich Schwanke [6]

[1] *DESY, University Potsdam*

[2] *University Potsdam*

[3] *University Potsdam, DESY*

[4] *University Erlangen-Nürnberg*

[5] *LLR Ecole Polytechnique*

[6] *Humboldt University Berlin*

The High Energy Stereoscopic System (H.E.S.S.) is a system of five Imaging Atmospheric Cherenkov Telescopes (IACTs) located in the Khomas Highland in Namibia. It measures cosmic gamma-rays with very high energies (VHE; > 100 GeV) using the Earth's atmosphere as a calorimeter. The H.E.S.S. array has entered Phase II in September 2012 with the inauguration of a fifth telescope that is larger and more complex than the other four. The very large mirror area of 600 m2 in comparison to the 100m2 of the smaller telescopes results in a lower energy threshold as well as an increased overall sensitivity of the system. Moreover, the parabolic dish allows the utilization of timing information in the shower reconstruction and, together with the improved camera electronics, gener- ates a considerably higher data rate. This talk will give an overview of the current H.E.S.S. data acquisition and array control system (DAQ) with particular emphasis on the first year of operation with the full five telescope array. We present the various requirements for the DAQ and discuss the general design principles to fulfil these requirements. The performance, stability and reliability of the H.E.S.S. Phase II DAQ, which resulted in a DAQ-related observation time loss of less than 1 %, are shown.

Data Stores, Data Bases, and Storage Systems / **404**

Sequential Data access with Oracle and Hadoop: a performance comparison

Author: Zbigniew Baranowski[1]

Co-authors: Eric Grancher [1]; Luca Canali [1]

[1] *CERN*

Corresponding Author: zbigniew.baranowski@cern.ch

The Hadoop framework has proven to be an effective and popular approach for dealing with "Big Data"and, thanks to its scaling ability and optimised storage access, Hadoop Distributed File System-based projects such as MapReduce or HBase are seen as candidates to replace traditional relational database management systems whenever scalable speed of data processing is a priority. But do these projects deliver in practice? Does migrating to Hadoop's "shared nothing"architecture really improve data access throughput? And, if so, at what cost?

We answer these questions—addressing cost/performance as well as raw performance—based on a performance comparison between an Oracle-based relational database and Hadoop's distributed solutions like MapReduce or HBase for sequential data access. A key feature of our approach is the use of an unbiased data model as certain data models can significantly favour one of the technologies tested.

Poster presentations / **250**

The ATLAS EventIndex: an event catalogue for experiments collecting large amounts of data

Author: Dario Barberis[1]

Co-authors: Alvaro Fernandez Casani [2]; David Malon [3]; Gancho Dimitrov [4]; JOSE SALT [5]; Jack Cranshaw [3]; Javier Sanchez [6]; Julius Hrivnac [7]; Marcin Nowak [8]; Qizhi Zhang [3]; Roman Sorokoletov [9]; Santiago Gonzalez De La Hoz [10]

[1] *Università e INFN Genova (IT)*

[2] *Universidad de Valencia (ES)*

[3] *Argonne National Laboratory (US)*

[4] *CERN*

[5] *IFIC-VALENCIA*

[6] *IFIC*

[7] *Universite de Paris-Sud 11 (FR)*

[8] *Brookhaven National Laboratory (US)*

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

[10] *IFIC-Valencia*

Corresponding Authors: dario.barberis@cern.ch, cranshaw@anl.gov, gancho.dimitrov@cern.ch, julius.hrivnac@cern.ch, malon@anl.gov, marcin.nowak@cern.ch, roman.sorokoletov@cern.ch, qzhang@anl.gov, alvaro.fernandez.casani@cern.ch, santiago.gonzalez@ific.uv.es, jose.salt@ific.uv.es

Modern scientific experiments collect vast amounts of data that must be cataloged to meet multiple use cases and search criteria. In particular, high-energy physics experiments currently in operation produce several billion events per year. A database with the references to the files including each event in every stage of processing is necessary in order to retrieve the selected events from data storage systems. The ATLAS EventIndex project is studying the best way to store the necessary information using modern data storage technologies (Hadoop, HBase etc.) that allow saving in memory key-value pairs and select the best tools to support this application from the point of view of performance, robustness and ease of use. At the end of this development, a new technology that is inherently independent of the type of data that are stored in the database – and therefore directly applicable to all scientific experiments with large amounts of data – will be available and demonstrated by the example of the ATLAS experiment. This paper describes the initial design and performance tests and the project evolution towards deployment and operation in 2014.

Data Stores, Data Bases, and Storage Systems / **261**

ATLAS Replica Management in Rucio: Replication Rules and Subscriptions

Author: Martin Barisits[1]

Co-authors: Angelos Molfetas [2]; Armin Nairz [1]; Cedric Serfon [1]; Graeme Andrew Stewart [1]; Luc Goossens [1]; Mario Lassnig [1]; Ralph Vigne [3]; Thomas Beermann [4]; Vincent Garonne [1]

[1] *CERN*

[2] *University of Sydney (AU)*

[3] *University of Vienna (AT)*

[4] *Bergische Universitaet Wuppertal (DE)*

Corresponding Authors: martin.barisits@cern.ch, vincent.garonne@cern.ch, mario.lassnig@cern.ch, graeme.andrew.stewart@cern.
thomas.beermann@cern.ch, ralph.vigne@cern.ch, cedric.serfon@cern.ch, luc.goossens@cern.ch, armin.nairz@cern.ch, angelos.molfetas@cern.ch

The ATLAS Distributed Data Management system stores more than 140PB of physics data across 100 sites worldwide. To cope with the anticipated ATLAS workload of the coming decade, Rucio, the next-generation data management system has been developed. Replica management, as one of the key aspects of the system, has to satisfy critical performance requirements in order to keep pace with the experiment's high rate of continuous data generation. The challenge lies in meeting these performance objectives while still giving the users and applications a powerful toolkit to control their data workflows. In this work we present the concept, design and implementation of the replica management in Rucio. We will specifically introduce the workflows behind replication rules, their formal language definition, weighting and site selection. Furthermore we will present the subscription component, which offers functionality for users to proclaim interest in data that has not been created yet. 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.

Poster presentations / 38

The keys to CERN conference rooms - Managing local collaboration facilities in large organisations

Author: Thomas Baron[1]

Co-authors: Franck Jouberjean [1]; Guillaume Duran [1]; Joao Correia Fernandes [1]; Jose Benito Gonzalez Lopez [1]; Loic Lavrut [1]; Marek Domaracky [1]; Nicola Tarocco [2]; Pedro Ferreira [1]

[1] *CERN*

[2] *Universita degli Studi di Udine (IT)*

Corresponding Authors: thomas.baron@cern.ch, jose.benito.gonzalez@cern.ch, pedro.ferreira@cern.ch

For a long time HEP has been ahead of the curve in its usage of remote collaboration tools, like videoconference and webcast, while the local CERN collaboration facilities were somewhat behind the expected quality standards for various reasons. This time is now over with the creation by the CERN IT department in 2012 of an integrated conference room service which provides guidance and installation services for new rooms (either equipped for video-conference or not), as well as maintenance and local support. Managing now nearly half of the 250 meeting rooms available on the CERN sites, this service has been built to cope with the management of all CERN rooms with limited human resources. This has been made possible by the intensive use of professional software to manage and monitor all the room equipment, maintenance and activity. This paper will focus on presenting these packages, either off-the-shelf commercial products (asset and maintenance management tool, remote audiovisual equipment monitoring systems, local automation devices, new generation touch screen interfaces for interacting with the room) when available or locally developed integration and operational layers (generic audiovisual control and monitoring framework) and how they help overcoming the challenges presented by such a service. The aim is to minimise local human interventions while preserving the highest service quality and placing the end user back to the center of this collaboration platform.

Facilities, Infrastructures, Networking and Collaborative Tools / 209

Experience from the 1st Year running a Massive High Quality Videoconferencing Service for the LHC

Authors: Joao Correia Fernandes[1]; Thomas Baron[1]

Co-author: Bruno Luis Dos Santos Bompastor [2]

[1] *CERN*

[2] *ADI Agencia de Inovacao (PT)*

Corresponding Authors: thomas.baron@cern.ch, joao.fernandes@cern.ch

In the last few years, we have witnessed an explosion of visual collaboration initiatives in the industry. Several advances in video services and also in their underlying infrastructure are currently improving the way people collaborate globally. These advances are creating new usage paradigms: any device in any network can be used to collaborate, in most cases with an overall high quality.
To keep apace with this technology progression, the CERN IT Department launched a service based on the Vidyo product.
This new service architecture introduces Adaptive Video Layering, which dynamically optimises the video for each endpoint by leveraging the H.264 Scalable Video Coding (SVC)-based compression technology. It combines intelligent AV routing techniques with the flexibility of H.264 SVC video compression, in order to achieve resilient video collaboration over the Internet, 3G and WiFi.
We will present an overview of the results that have been achieved after this major change. In particular, the first year of operation of the CERN Vidyo service will be described in terms of performance and scale: The service became part of the daily activity of the LHC collaborations, reaching a monthly usage of more than 3200 meetings with a peak of 750 simultaneous connections.

We will also present some key features such as the integration with CERN Indico. LHC users can now join a Vidyo meeting either from their personal computer or a CERN videoconference room simply from an Indico event page, with the ease of a single click. The roadmap for future improvements, service extensions and core infrastructure tendencies such as cloud based services and virtualisation of system components will also be discussed.
Vidyo's strengths allowed us to build a universal service (it is accessible from PCs, but also videoconference rooms, traditional phones, tablets and smartphones), developed with 3 key ideas in mind: ease of use, full integration and high-quality.

Poster presentations / **2**

softinex, inlib, exlib, ioda, g4view, g4exa, wall

Author: Guy Barrand[1]

[1] *Universite de Paris-Sud 11 (FR)*

Corresponding Author: barrand@lal.in2p3.fr

Softinex names a software environment targeted to data analysis and visualization. It covers the C++ inlib and exlib "header only" libraries that permit, through GL-ES and a maximum of common code, to build applications deliverable on the AppleStore (iOS), GooglePlay (Android), traditional laptops/desktops under MacOSX, Linux and Windows, but also deliverable as a web service able to display in various web browsers compatible with WebGL (FireFox, Chrome, Safari). The ioda app permits (with fingertips on a tablet) to read files at various formats (xml-aida, cern-root, fits) and visualize some of their data such as images, histograms, ntuples and geometries. The g4view app permits to visualize Geant4 gdml files and do, for example on a tablet, some simple outreach particle physics. g4exa is a simple Geant4 template open source code for people wanting to create their own app done in the same spirit. The wall programs permit to visualize HEP data (plots, geometries, events) on a large display surface done with an assembly of screens driven by a set of computers. We want to present this software suite but also the grounding ideas, such as the "Software Least Action Principle", that led to their developments.

Facilities, Infrastructures, Networking and Collaborative Tools / **20**

Experience with procuring, deploying and maintaining hardware at remote co-location centre

Author: Olof Barring[1]

Co-authors: Afroditi Xafi ; Alain Gentit [1]; Anthony Grossir [1]; Benoit Clement [1]; Eric Bonfillou [1]; Miguel Coelho dos Santos [1]; Vincent Dore [1]; Wayne Salter [1]

[1] *CERN*

Corresponding Author: olof.barring@cern.ch

In May 2012 CERN signed a contract with the Wigner Data Centre in Budapest for an extension to the CERN's central computing facility beyond its current boundaries set by electrical power and cooling available for computing. The centre is operated as a remote co-location site providing rack-space, electrical power and cooling for server, storage and networking equipment acquired by CERN. The contract includes a 'remote-hands'services for physical handling of hardware (rack mounting, cabling, pushing power buttons, ⋯) and maintenance repairs (swapping disks, memory modules, ⋯). However, only CERN personnel have network and console access to the equipment for system administration. This report gives an insight to undertaken adaptations of hardware architecture, procurement and delivery procedures enabling remote physical handling of the hardware. We will

also describe tools and procedures developed for automating the registration, burn-in testing, acceptance and maintenance of the equipment as well as an independent but important change to the IT assets management (ITAM) developed in parallel as part of CERN IT Agile Infrastructure project. Finally, we will report on experience from the first large delivery of 400 servers and 80 SAS JBOD expansion units (24 bays) to Wigner in March 2013.

Poster presentations / 100

WLCG Transfers Dashboard: A unified monitoring tool for heterogeneous data transfers.

Authors: Alexandre Beche[1]; David Tuckett[1]

Co-authors: Ivan Kadochnikov [2]; Julia Andreeva [1]; Pablo Saiz [1]; Sergey Belov [2]

[1] *CERN*

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

Corresponding Authors: david.tuckett@cern.ch, alexandre.beche@cern.ch

The Worldwide LHC Computing Grid provides resources for the four main virtual organizations. Along with data processing, data distribution is the key computing activity on the WLCG infrastructure. The scale of this activity is very large, the ATLAS virtual organization (VO) alone generates and distributes more than 40 PB of data in 100 million files per year. Another challenge is the heterogeneity of data transfer technologies. Currently there are two main alternatives for data transfers on the WLCG: File Transfer Service (FTS) and XRootD protocol for transferring data on the XRootD federated storage. Each LHC VO has its own monitoring system which allows it to understand its own transfer activity but is limited to the scope of that particular VO. There is a need for a global system which would provide a complete cross-VO and cross-technology picture of all WLCG data transfers.

We present a unified monitoring tool - WLCG Transfers Dashboard - where all the VOs and technologies coexist and are monitored together. The scale of the activity and the heterogeneity of the system raises a number of technical challenges. Each technology comes with its own monitoring specificities and some of the VOs use several of these technologies. The presentation will describe the implementation of the system with particular focus on the design principles applied to ensure the necessary scalability and performance, and to easily integrate any new technology providing additional functionality which might be specific to that technology.

Poster presentations / 101

Monitoring of large-scale federated data storage: XRootD and beyond.

Author: Alexandre Beche[1]

Co-authors: Artem Petrosyan [2]; Daniel Diéguez Arias [3]; Danila Oleynik [2]; David Tuckett [1]; Domenico Giordano [1]; Ilija Vukotic [4]; Julia Andreeva [1]; Matevz Tadel [5]; Pablo Saiz [1]; Sergey Belov [2]

[1] *CERN*

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

[3] *CERN / University of Vigo (ES)*

[4] *University of Chicago (US)*

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

Corresponding Author: alexandre.beche@cern.ch

The computing models of the LHC experiments are gradually moving from hierarchical data models with centrally managed data pre-placement towards federated storage which provides seamless access to data files independently of their location and dramatically improved recovery due to failover mechanisms. Enabling loosely coupled data clusters to act as a single storage resource should increase opportunities for data analysis and should enable more effective use of computational resources at sites with limited storage capacities. Construction of the data federations and understanding the impact of the new approach to data management on user analysis requires complete and detailed monitoring. Monitoring functionality should cover the status of all components of the federated storage, measuring data traffic and data access performance, as well as being able to detect any kind of inefficiencies and to provide hints for resource optimization and effective data distribution policy. Data mining of the collected monitoring data provides a deep insight into new patterns of usage of the storage resources, beyond that provided by other monitoring strategies.

In the WLCG context, there are several federations currently based on the XRootD technology. The talk will focus on monitoring for the ATLAS and CMS XRootD federations (Federated Atlas XRootD (FAX) and Any Data, Any Time, Anywhere (AAA)) implemented in the Experiment Dashboard monitoring framework. Both federations consist of many dozens of sites accessed by many hundreds of clients and they continue to grow in size. Handling of the monitoring flow generated by these systems has to be well optimized in order to achieve the required performance.

The talk will demonstrate that though FAX and AAA Dashboards are being developed for XRootD federations, the implementation is generic and can be easily adapted for other technologies, such as HTTP/WebDAV federations.

Poster presentations / **222**

ROOT: native graphics on Mac OS X

Author: Timur Pocheptsov[1]

Co-author: Bertrand Bellenot [2]

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

[2] *CERN*

Corresponding Authors: bertrand.bellenot@cern.ch, timur.pocheptsov@cern.ch

In my poster I'll present a new graphical back-end for ROOT that has been developed for the Mac OS X operating system as an alternative to the more than 15 year-old X11-based version. It represents a complete implementation of ROOT's GUI, 2D and 3D graphics based on Apple's native APIs/frameworks, written in Objective-C++.

Poster presentations / **167**

ROOT I/O in JavaScript - Reading ROOT files in a browser

Author: Bertrand Bellenot[1]

[1] *CERN*

Corresponding Author: bertrand.bellenot@cern.ch

In order to be able to browse (inspect) ROOT files in a platform independent way, a JavaScript version of the ROOT I/O subsystem has been developed. This allows the content of ROOT files to be displayed

in most available web browsers, without having to install ROOT or any other software on the server or on the client. This gives a direct access to ROOT files from any new device in a light way. It is possible to display simple graphical objects such as histograms and graphs (TH1, TH2, TH3, TProfile, TGraph, ...). The rendering of 1D/2D histograms and graphs is done with an external JavaScript library (d3.js), and 2D & 3D histograms with another library (three.js). This poster will describe the techniques used to stream and display the content of a ROOT file, with a rendering being now very close to the one provided by ROOT.

Poster presentations / 301

Development and application of CATIA-GDML geometry builder

Author: Sergey Belogurov[1]

Co-authors: Andrey Chernogorov [2]; Egor Ovcharenko [1]; Peter Malzacher [3]; Vitaly Shchetinin [2]; Yuri Berchun [4]

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

[2] *ITEP*

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

[4] *BMSTU*

Corresponding Author: belogurov@itep.ru

Detector geometry exchange between CAD systems and physical Monte Carlo (MC), packages ROOT and Geant4 is a labor-consuming process necessary for fine design optimization. CAD and MC geometries have completely different structure and hierarchy. For this reason automatic conversion is possible only for very simple shapes.

CATIA-GDML Geometry Builder is a tool which allows to facilitate significantly creation of MC compatible geometry in GDML format from the CAD system CATIA v.5 which is the main design system in CERN, GSI, LNGS, and other scientific and industrial entities. The tool was introduced at CHEP2010 [1].We employ powerful measurement, design and VBA customization features of various CATIA workbenches for creation of GDML compatible representation of an existing engineering assembly. For many Root/Geant primitives we have developed parameterized CATIA User Defined Features. We have implemented in CATIA concepts of logical, physical and mother volumes. The Constructive Solid Geometry (CSG) Boolean tree can be optimized from point of view of simulation performance. At the end the CSG tree is exported into GDML.

For the last three years a number of detector models were transferred from CAD to MC and vice versa. Functionality of the tool was extended and usability was improved according to the practical experience.

.
The following novelties were the most important for improvement of usability: extension of the set of implemented primitives; introduction of polycone and polyhedra; introduction of a part template for MC compatible geometry; development of a correctness checker of a resulting model before export; automated positioning by clicking; generation of symmetry from an assembly; improved handling of materials.

The up-to-date functionality, analysis of nontrivial use cases, an approach to integration with automated recognition of simple shapes, examples and hints on the best practices will be reported at the conference.

[1] S. Belogurov, Yu. Berchun, et al. 'CATIA-GDML GEOMETRY BUILDER'.
Journal of Physics: Conference Series, Volume 331, Number 3, 2011 , pp. 32035-32040

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / 308

PROOF as a Service on the Cloud: a Virtual Analysis Facility based on the CernVM ecosystem

Author: Dario Berzano[1]

Co-authors: Georgios Lestaris [1]; Gerardo Ganis [1]; Ioannis Charalampidis [2]; Jakob Blomer [1]; Predrag Buncic [1]; Rene Meusel [1]

[1] *CERN*

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

Corresponding Author: dario.berzano@cern.ch

PROOF, the Parallel ROOT Facility, is a ROOT-based framework which enables interactive parallelism for event-based tasks on a cluster of computing nodes.

Although PROOF can be used simply from within a ROOT session with no additional requirements, deploying and configuring a PROOF cluster used to be not as straightforward. Recently great efforts have been spent to make the provisioning of generic PROOF analysis facilities with zero configuration, with the added advantages of positively affecting both stability and scalability, making the deployment operations feasible even for the end user.

Since a growing amount of large-scale computing resources are nowadays made available by Cloud providers in a virtualized form, we have developed the Virtual PROOF-based Analysis Facility: a cluster appliance combining the solid CernVM ecosystem and PoD (PROOF on Demand), ready to be deployed on the Cloud and leveraging some peculiar Cloud features such as elasticity.

We will show how this approach is effective both for sysadmins, who will have little or no configuration to do to run it on their Clouds, and for the end users, who are ultimately in full control of their PROOF cluster and can even easily restart it by themselves in the unfortunate event of a major failure. We will also show how elasticity leads to a more optimal and uniform usage of Cloud resources.

Summaries / 523

Summary of track 4 (Data Stores, Data Bases, and Storage Systems)

Corresponding Author: wahid.bhimji@cern.ch

Poster presentations / 231

Hepdoop

Author: Wahid Bhimji[1]

Co-authors: Andrew John Washbrook [1]; Timothy Michael Bristow [1]

[1] *University of Edinburgh (GB)*

Corresponding Author: wahid.bhimji@cern.ch

"Big Data"is no longer merely a buzzword, but is business-as-usual in the private sector. High Energy Particle Physics is often cited as the archetypal Big Data use case, however it currently shares very little of the toolkit used in the private sector or other scientific communities.

We present the initial phase of a programme of work designed to bridge this technology divide by both performing real HEP analysis workflows using predominately industry "Big Data"tools, formats and techniques, and the reverse: to perform real industry tasks with HEP tools. In doing so it will improve interoperation of those tools, reveal strengths and weakness and enable efficiencies within both communities.

The first phase of this work performs key elements of an LHC Higgs Analysis using very common Big Data tools. These elements include data serialization, filtering and data mining. They are performed with a range of tools chosen not just for performance but also for ease-of-use, maturity and size of user community. This includes technologies such as Protocol Buffers, Hadoop and Python Scikit and for each element we make comparisons with the same analysis performed using current HEP tools such as ROOT, Proof and TMVA.

Poster presentations / **70**

go-hist: a multi-threaded Go package for histogramming

Author: Sebastien Binet[1]

[1] *IN2P3/LAL*

Corresponding Author: sebastien.binet@cern.ch

Current HENP libraries and frameworks were written before multicore systems became widely deployed and used.
From this environment, a 'single-thread' processing model naturally emerged but the implicit assumptions it encouraged are greatly impairing our abilities to scale in a multicore/manycore world.

Thanks to C++11, C++ is finally slowly catching up with regard to concurrency constructs, at the price of further complicating the language and its standard library.
Instead, the approach of the Go language is to provide simple concurrency enabling building blocks, readily integrated into the language, in the hope that these goroutines and channels will help to design, write and maintain concurrent applications.

To further investigate whether Go is a suitable C++ and python replacement language for HENP, we developed go-hist, a multi-threaded Go package for histogramming.

We will first present the overall design of the go-hist package and the various building blocks (goroutines, channels and wait-groups) provided by the Go language which are then leveraged by this library. A special emphasis will be put on the current SIMD support available in Go and thus how vectorization can be leveraged in histogramming code: a cornerstone for automatic performance scalability on the ever-wider-registers architectures the future has in store.

Then, I/O considerations, such as read/write performances, ease of serialization and disk sizes will be discussed, for each of the currently implemented backends (protobuf, gob and json.)

Finally, comparisons with ROOT, Java ROOT and inlib/exlib (an AIDA implementation in C++) performances will be presented.

Poster presentations / **71**

A method to improve the electron momentum reconstruction for the PANDA experiment

Author: MA Binsong[1]

[1] *IPN Orsay France*

Corresponding Author: binsong@ipno.in2p3.fr

The PANDA (AntiProton ANnihilation at DArmstadt) experiment is one of the key projects at the future Facility for Antiproton and Ion Research (FAIR), which is currently under construction at Darmstadt. This experiment will perform precise studies of antiproton-proton and antiproton-nucleus annihilation reactions. The aim of the rich experimental program is to improve our knowledge of the strong interaction and of the structure of hadrons.

In particular, the study of electromagnetic processes, like (pbar p->e+e-, pbarp->e+e-π0, etc..), gives access to the proton structure (electric and magnetic form factors, Transition Distribution Amplitudes, etc..). In such channels, the electron and positron signal needs to be separated from the hadronic background, which is six orders of magnitude larger than the signal. Excellent electron particle identification and momentum reconstruction are therefore crucial for such studies.

The PandaRoot software, based on ROOT and Virtual MonteCarlo, is used as the simulation and analysis framework for the future PANDA experiment. A Kalman Filter provides the particle momenta deduced from the central tracker, with GEANE as track follower. This method is not well suited for electrons, for which the highly non-gaussian Bremsstrahlung process yields a tail in the momentum resolution distribution.

A new method was developed to solve this problem at least partially, in an event by event procedure, taking advantage of the possible detection of the Bremsstrahlung photon as a separate cluster in the Electromagnetic Calorimeter. We will show that this is possible for tracks with transverse momentum up to 1 GeV/c. The shape of the electron shower is also used to identify the Bremsstrahlung photon at higher electron momenta region.

The improvement of electron momentum reconstruction will be shown, as well as the gain on the signal to background ratio. In the presentation, the details about the technical implementation of the method in PANDARoot will also be given.

Poster presentations / 398

Synergy between the CIMENT tier-2 HPC centre in Grenoble (France) and the HEP community at LPSC (”Laboratoire de Physique Subatomique et de Cosmologie”)

Authors: Bruno Bzeznik[1]; Catherine Biscarat[2]

[1] *UJF/CIMENT/LIG*

[2] *LPSC/IN2P3/CNRS France*

Corresponding Author: catherine.biscarat@cern.ch

We describe the synergy between CIMENT (a regional multidisciplinary HPC centre) and the infrastructures used for the analysis of data recorded by the ATLAS experiment at the LHC collider and the D0 experiment at the Tevatron.
CIMENT is the High Performance Computing (HPC) centre developed by Grenoble University. It is a federation of several scientific departments and it is based on the gridification of a dozen HPC machines with iRODS storage. CIMENT is a medium scale centre, or a tier-2 in the HPC pyramidal scheme, of about 35 TFlops, placing it in the top-5 of the French tier-2 sites in 2012. The acquisition of an additional machine in spring 2013 is more than doubling its computing capacity and

hence consolidating CIMENT's leading role in the French HCP landscape. CIMENT aims at providing significant computing power for tests and algorithm developments before execution on national [tier-1] or european [tier-0] platforms. This profile of resource allocation necessarily implies that not all resources are used at all times.

The main goals of the LHC collider at CERN are the search for a subatomic particle called the Higgs boson and the search for new phenomena beyond the Standard Model of particle physics. The observation of a Higgs-like boson has been reported last year by ATLAS and CMS, the general-purpose experiments operating at the LHC.
Current research is focused on the characterisation of the newly discovered boson, and on the search for new phenomena.

Researchers at LPSC in Grenoble are leading the search for one type of new phenomena in ATLAS, namely the search for additional spatial dimensions which are likely to manifest themselves in LHC collisions. Given the rich multitude of physics studies proceeding in parallel in the ATLAS collaboration, one of the limiting factors in the timely analysis of ATLAS data is the availability of computing resources for physics analysis. Another LPSC team suffers from a similar limitation. This team is leading the ultimate precision measurement based on the data from the D0 experiment: the precise measurement of the W boson mass, which yields an indirect constraint on the mass of the Higgs boson and completes the direct search at the LHC. A relative precision of 10ˆ-4 in the measurement of the W boson mass is needed to help elucidate the nature of the newly discovered Higgs-like particle. Such a measurement requires simulations of unprecedented precision, and therefore considerable computing power.

The limitation in available computing power becomes problematic in the months that precede the international conferences where major results are released. The sharing of resources between different scientific fields, like the ones discussed in this article, constitutes a valuable synergy, because the spikes in need for computing resources are uncorrelated in time between different fields (HPC and HEP). The results of our collaboration between fields manifest themselves in the timely delivery of HEP results that had been eagerly awaited both by the particle physics community and by the general public.

Poster presentations / 88

A Comprehensive Approach to Tier-2 Administration

Authors: John Bland[1]; Robert Fay[1]

Co-authors: Mark Norman [1]; Stephen Jones [2]

[1] *University of Liverpool*

[2] *Liverpool University*

Corresponding Author: jbland@hep.ph.liv.ac.uk

Liverpool is consistently amongst the top Tier-2 sites in Europe in terms of efficiency and cluster utilisation. This presentation will cover the work done at Liverpool over the last six years to maximise and maintain efficiency and productivity at their Tier 2 site, with an overview of the tools used (including established, emerging, and locally developed solutions) for monitoring, testing, installation, configuration, ticketing, logging, and administration, along with the techniques for management, operations, and optimisation that tie them together into a comprehensive, scalable, and sustainable approach to Tier 2 administration.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / 213

Micro-CernVM: Slashing the Cost of Building and Deploying Virtual Machines

Author: Jakob Blomer[1]

Co-authors: Dario Berzano [1]; Georgios Lestaris [2]; Gerardo Ganis [1]; Ioannis Charalampidis [3]; Predrag Buncic [1]; Rene Meusel [1]

[1] *CERN*

[2] *University of Athens (GR)*

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

Corresponding Author: jakob.blomer@cern.ch

The traditional virtual machine building and and deployment process is centered around the virtual machine hard disk image. The packages comprising the VM operating system are carefully selected, hard disk images are built for a variety of different hypervisors, and images have to be distributed and decompressed in order to instantiate a virtual machine. Within the HEP community, the CernVM File System has been established in order to decouple the distribution from the experiment software from the building and distribution of the VM hard disk images.

We show how to get rid of such pre-built hard disk image altogether. Due to the high requirements on POSIX compliance imposed by HEP application software, CernVM-FS can also be used to host and boot a Linux operating system. This allows the use of a tiny bootable CD image that comprises only a Linux kernel while the rest of the operating system is provided on demand by CernVM-FS. This approach speeds up the initial boot time and reduces virtual machine image sizes by an order of magnitude. Furthermore, security updates can be distributed instantaneously through CernVM-FS and by leveraging the fact that CernVM-FS is a versioning file system, a historic analysis environment can be easily re-spawned by selecting the corresponding CernVM-FS file system snapshot.

Poster presentations / **62**

Security in the CernVM File System and the Frontier Distributed Database Caching System

Author: Dave Dykstra[1]

Co-author: Jakob Blomer [2]

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

[2] *CERN*

Corresponding Authors: jakob.blomer@cern.ch, dwd@fnal.gov

Both the CernVM File System (CVMFS) and the Frontier Distributed Database Caching System (Frontier) distribute centrally updated data worldwide for LHC experiments using http proxy caches. Neither system provides privacy or access control on reading the data, but both control access to updates of the data and can guarantee the integrity of the data transferred to clients over the internet. CVMFS has since its early days required digital signatures and secure hashes on all distributed data, and recently both CVMFS and Frontier have added X509-based integrity checking. In this paper we detail and compare the security models of CVMFS and Frontier.

Data Stores, Data Bases, and Storage Systems / **94**

CMS Use of a Data Federation

Author: Kenneth Bloom[1]

[1] *University of Nebraska (US)*

Corresponding Author: kenbloom@unl.edu

CMS is in the process of deploying an Xrootd based infrastructure to facilitate a global data federation. The services of the federation are available to export data from half the physical capacity and the majority of sites are configured to read data over the federation as a back-up. CMS began with a relatively modest set of use-cases for recovery of failed local file opens, debugging and visualization. CMS is finding that the data federation can be used to support small scale analysis and load balancing. Looking forward we see potential in using the federation to provide more flexibility in the location workflows are executed as the differenced between local access and wide area access are diminished by optimization and improved networking. In this presentation we will discuss the application development work and the facility deployment work, the use-cases currently in production, and the potential for the technology moving forward.

Poster presentations / **110**

CMS Data Analysis School Model

Author: Sudhir Malik[1]

Co-author: Ian Fisk [2]

[1] *University of Nebraska-Lincoln*

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

Corresponding Authors: kenbloom@unl.edu, ian.fisk@cern.ch

To impart hands-on training in physics analysis, CMS experiment initiated the concept of CMS Data Analysis School (CMSDAS). It was born three years ago at the LPC (LHC Physics Center), Fermilab and is based on earlier workshops held at the LPC and CLEO Experiment. As CMS transitioned from construction to the data taking mode, the nature of earlier training also evolved to include more of analysis tools, software tutorials and physics analysis. This effort epitomized as CMSDAS has proven to be a key for the new and young physicists to jump start and contribute to the physics goals of CMS by looking for new physics with the collision data. With over 400 physicists trained in six CMSDAS around the globe , CMS is trying to engage the collaboration discovery potential and maximize the physics output. As a bigger goal, CMS is striving to nurture and increase engagement of the myriad talents of CMS, in the development of physics, service, upgrade, education of those new to CMS and the career development of younger members. An extension of the concept to the dedicated software and hardware schools is also planned, keeping in mind the ensuing upgrade phase.

Plenaries / **487**

Big Data - Flexible Data - for HEP

Author: Brian Paul Bockelman[1]

[1] *University of Nebraska (US)*

Corresponding Author: brian.bockelman@cern.ch

The experience with the processing of large amounts of data results in changing data models and data access patterns, both locally as well as over the wide area. Dr. Brian Bockelman of the University of Nebraska will present the developments in big data for particle physics looking at data mining, extreme data bases, access to data storage, and the impact thereof on data modelling at different scales in many-core era.

Data Stores, Data Bases, and Storage Systems / 160

Optimizing High-Latency I/O in CMSSW

Author: Brian Paul Bockelman[1]

Co-author: Elizabeth Sexton-Kennedy [2]

[1] *University of Nebraska (US)*

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

Corresponding Authors: brian.bockelman@cern.ch, sexton@fnal.gov

To efficiently read data over high-latency connections, ROOT-based applications must pay careful attention to user-level usage patterns and the configuration of the I/O layer. Starting in 2010, CMSSW began using and improving several ROOT "best practice" techniques such as enabling the TTreeCache object and avoiding reading events out-of-order. Since then, CMS has been deploying additional improvements not part of base ROOT, such as the removal of the TTreeCache startup penalty and significantly reducing the number of network roundtrips for sparse event filtering. CMS has also implemented an algorithm for multi-source reads using Xrootd. This new client layer splits ROOT read requests between active source servers based on recent server performance and issues these requests in parallel.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / 164

Dynamic VM provisioning for Torque in a cloud environment

Author: Shunde Zhang[1]

Co-authors: Martin Sevior [2]; Paul Coddington [3]

[1] *eRSA, CoEPP*

[2] *University of Melbourne (AU)*

[3] *eRSA*

Corresponding Authors: lucien.boland@cern.ch, shunde.zhang@adelaide.edu.au, martines@unimelb.edu.au

The Nectar national research cloud provides compute resources to Australian researchers using OpenStack. CoEPP, a WLCG Tier2 member, wants to use Nectar's cloud resources for Tier 2 and Tier 3 processing for ATLAS and other experiments including Belle, as well as theoretical computation. CoEPP would prefer to use the Torque job management system in the cloud because they have extensive experience in managing Torque and users are familiar with running batch jobs using Torque. However, Torque was developed for static clusters, and worker nodes cannot easily be dynamically added or removed, and this also requires updating related services such as monitoring servers like Nagios and Ganglia, file service and package management service.

A number of projects have been looking for an easy way to run batch jobs in the cloud. The solution described here is inspired by two successful projects: ViBatch, which enables Torque to use a local KVM hypervisor to add and remove virtualized resources, and Cloud Scheduler, which dynamically allocates cloud VMs according to the workload of an associated Condor queue. This generalised solution combines the advantages of the above projects and enables cloud-based dynamic VM provisioning for Torque. It includes two parts: prologue, epilogue and a number of utility scripts working alongside Torque; and a service called VM Pool that maintains an elastic set of VMs in the cloud.

A special Torque worker node is configured with the aforementioned scripts in order to communicate with VM Pool for job execution. When a new job comes in, prologue script requests a VM from VM Pool, and does some initialization on it, including sending the job description file to the VM, creating

relevant credentials on the VM and setting up environment variables. Then Torque executes the job on that VM through an SSH connection. When finished, epilogue returns the VM back to VM Pool, and cleans up all relevant data.

VM Pool runs as a standalone service to orchestrate VMs in the cloud. The number of VMs in the pool changes according to the number of VM requests from the worker node but remains between a specified minimum and maximum value. Any VM in the pool has three states: Free, Unusable and Busy. Unusable VMs can be dead, in building state, in deleting state, or in active state but inaccessible due to network or other issues. Unusable VMs are checked periodically to see if the state is changed, if they remain in this state for longer than a specific time, they will be terminated.

Torque sees a single worker node with a fixed number of processors (the maximum number specified in VMPool), so resources do not need to be dynamically added or removed from Torque, this is handled in VM Pool.

Currently a Beta system has been implemented and tested on the Nectar cloud. Next it will be under User Acceptance Test, then once everything works fine it will become a production system for Tier 3 and Tier 2.

Data Acquisition, Trigger and Controls / **362**

The core trigger software framework of the ATLAS experiment

Authors: Dmitry Emeliyanov[1]; Rustem Ospanov[2]; Sami Kama[3]

[1] *STFC - Science & Technology Facilities Council (GB)*

[2] *University of Pennsylvania*

[3] *Southern Methodist University (US)*

Corresponding Authors: tomasz.bold@cern.ch, sami.kama@cern.ch, dmitry.emeliyanov@stfc.ac.uk

The high level trigger (HLT) of the ATLAS experiment at the LHC selects interesting proton-proton and heavy ion collision events for the wide ranging ATLAS physics program. The HLT examines events selected by the level-1 hardware trigger using a combination of specially designed software algorithms and offline reconstruction algorithms. The flexible design of the entire trigger system was critical for the success of the ATLAS data taking during the first run of the LHC. The flexibility of the HLT is due to a versatile core software which includes a steering infrastructure, responsible for configuration and execution of hundreds of trigger algorithms, and navigation infrastructure, responsible for storing trigger results for physics analysis and combining algorithms into multi-object triggers. The multi-object triggers are crucial for efficient selection of interesting physics events at high LHC luminosity while running within limited bandwidth budgets. A resource consumption by the software algorithms was minimized thanks to a sophisticated navigation interface which encapsulates trigger logic and caches results of trigger algorithms. Detailed description of the steering and navigation infrastructures will be presented together with details of the caching implementation and results of operating the system with and without the caching.
In preparation for future LHC running conditions, a new software interface has been developed to maximize benefit from new commodity computing hardware for CPU intensive parts of the HLT. The interface contains a software layer which provides hardware abstraction and handles the data communication between the existing software components and the optimized algorithms that are executed on the dedicated computing resource. This layer is extensible and suitable for multi-threaded, multi-process, multi-node environment utilizing diverse hardware resources. The results of the tests on various hardware platforms and with several programming systems will be presented.

Poster presentations / **319**

Performance evaluation and capacity planning for a scalable and

highly available virtulization infrastructure for the LHCb experiment

Authors: Enrico Bonaccorsi[1]; Francesco Sborzacchi[2]; Niko Neufeld[1]

[1] *CERN*

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

Corresponding Authors: francesco.sborzacchi@cern.ch, enrico.bonaccorsi@cern.ch, niko.neufeld@cern.ch

The virtual computing is often run to satisfy different needs: reduce costs, reduce resources, simplify maintenance and the last but not the least add flexibility.
The use of Virtualization in a complex system such as a farm of PCs that control the hardware of an experiment (PLC, power supplies ,gas, magnets..) put as in a condition where not only an High Performance requirements need to be carefully considered but also a deep analysis of strategies to achieve a certain level of High Availability.
We conducted a performance evaluation on different and comparable storage/network/virtulization platforms.
The performance is measured using a series of independent benchmarks , testing the speed an the stability of multiple VMs runnng heavy-load operations on the I/O of virtualized storage and the virtualized network.
The result from the benchmark tests allowed us to study and evaluate how the different workloads of Vm workloads interact with the Hardware/Software resource layers.

Poster presentations / **445**

FIFE-Jobsub: A Grid Submission System for Intensity Frontier Experiments at Fermilab

Author: Dennis Box[1]

[1] *F*

Corresponding Author: dbox@fnal.gov

The Fermilab Intensity Frontier Experiments use an integrated submission system known as FIFE-jobsub, part of the FIFE (Fabric for Frontier Experiments) initiative, to submit batch jobs to the Open Science Grid. FIFE-jobsub eases the burden on experimenters by integrating data transfer and site selection details in an easy to use and well documented format. FIFE-jobsub automates tedious details of maintaining grid proxies for the lifetime of the grid job. Data transfer is handled using the ifdh (Intensity Frontier Data Handling) tool suite, which facilitates selecting the appropriate data transfer method from many possibilities while protecting shared resources from overload. Chaining of job dependencies into Directed Acyclic Graphs (Condor DAGS) is well supported and made easier through the use of input flags and parameters.

We will present a talk describing these features, their implementation, and the benefits to users of this system.

Poster presentations / **377**

Design and Performance of the Virtualization Platform for Offline computing on the ATLAS TDAQ Farm

Authors: Alexandr Zaytsev[1]; Franco Brasolin[2]

Co-authors: Alessandro Di Girolamo [3]; Cristian Contescu [4]; Diana Scannicchio [5]; Mikel Eukeni Pozo Astigarraga [5]; Sergio Ballestrero [6]; Silvia Maria Batraneanu [5]

[1] *Brookhaven National Laboratory (US)*

[2] *Universita e INFN (IT)*

[3] *CERN*

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

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

[6] *University of Johannesburg (ZA)*

Corresponding Author: franco.brasolin@cern.ch

With the LHC collider at CERN currently going through the period of Long Shutdown 1 (LS1) there is a remarkable opportunity to use the computing resources of the large trigger farms of the experiments for other data processing activities.
In the case of ATLAS experiment the TDAQ farm, consisting of more than 1500 compute nodes, is particularly suitable for running Monte Carlo production jobs that are mostly CPU and not I/O bound. This contribution gives a thorough review of all the stages of "Sim@P1"project dedicated to the design and deployment of a virtualized platform running on the ATLAS TDAQ computing resources and using it to run the large groups of CernVM based virtual machines operating like a single "CERN--P1"Grid site. This platform has been designed to avoid interference with TDAQ usage of the farm and to guarantee the security and the usability of the ATLAS private network; Openstack has been chosen to provide a cloud management layer.
The approaches to organizing support for the sustained operation of the system on both infrastructural (hardware, virtualization platform) and logical (site support and job execution) levels are also discussed.
The project is a result of combined effort of the ATLAS TDAQ SysAdmins and NetAdmins teams, CERN IT ES Department and RHIC & ATLAS Computing Facility at BNL.

Poster presentations / **366**

Real-time flavor tagging selection in ATLAS

Author: Carlo Schiavi[1]

[1] *INFN Genova*

Corresponding Author: adrian.buzatu@glasgow.ac.uk

In high--energy physics experiments, online selection is crucial to reject most uninteresting collisions and to focus on interesting physical signals.
The b--jet selection is part of the trigger strategy of the ATLAS experiment and is meant to select hadronic final states with heavy--flavor content. This is important for the selection of physics channels with more than one b--jet in the final state, enabling to reject QCD light jets and maintain affordable trigger rates without raising jet energy thresholds. ATLAS introduced b--jet triggers in 2011 and deployed more complex and better performing tagging algorithms in 2012. An overview of the b--jet trigger menu and its performance on real data is presented in this contribution.
Data--driven techniques to extract the online b--tagging performance, a key ingredient for all analyses relying on such triggers, are also discussed and results presented.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / **277**

OASIS: a data and software distribution service for Open Science Grid

Authors: Brian Paul Bockelman[1]; John Hover[2]; John Steven De Stefano Jr[3]; Jose Caballero Bejar[3]; Rob Quick[4]; Scott Werner Teige[5]

[1] *University of Nebraska (US)*

[2] *Brookhaven National Laboratory (BNL)-Unknown-Unknown*

[3] *Brookhaven National Laboratory (US)*

[4] *OSG - Indiana University*

[5] *Indiana University (US)*

Corresponding Author: jose.caballero@cern.ch

The Open Science Grid (OSG) encourages the concept of software portability: a user's scientific application should be able to run in as many operating system environments as possible. This is typically accomplished by compiling the software into a single static binary, or distributing any dependencies in an archive downloaded by each job. However, the concept of portability runs against the software distribution philosophy of many Linux packages, and becomes increasingly difficult to achieve as the size of a scientist's software stack increases. Despite being a core philosophy, portability has become a deterrent to the adoption of the OSG.

It is necessary to provide a mechanism for OSG Virtual Organizations (VO) to install software at sites. Since its initial release, the OSG Compute Element has provided a application software installation directory to VOs, into which VOs assume they can create their own sub-directory, install software into that sub-directory, and have the directory shared on the worker nodes for their sites (typically via NFS). The OSG provides guidelines for the size and UNIX permissions of such directories.

The current model lacks the ability to manage the software are; there are shortcomings with regard to permissions, policies, versioning, and the lack of a unified, collective procedure or toolset for deploying software across all sites. Therefore, a new mechanism for data and software distributing is desirable. The proposed architecture for the OSG Application Software Installation Service (OASIS) is a server-client model: the software and data are installed only once in a single place (or a reduced number of places), and are automatically distributed to all client sites simultaneously.

Central file distribution offers other advantages, including server-side authentication and authorization, activity records, quota management, data validation and inspection, and well-defined versioning and deletion policies.

Currently the file transfer mechanism in OASIS is implemented using the CERN Virtual Machine Filesystem (CVMFS) as underlaying technology.

The proposed architecture, as well as a complete analysis of the current implementation, will be described in this paper.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / 191

Direct exploitation of a top500 supercomputer in the analysis of CMS data.

Author: Luis Cabellos[1]

Co-authors: Iban Jose Cabrillo Bartolome [2]; Isidro Gonzalez Caballero [3]; Javier Fernandez Menendez [3]; Jesus Marco [4]

[1] *IFCA, CSIC-UC*

[2] *Universidad de Cantabria (ES)*

[3] *Universidad de Oviedo (ES)*

[4] *IFCA CSIC-UC, Santander, Spain*

Corresponding Author: iban.jose.cabrillo.bartolome@cern.ch

The Altamira supercomputer at the Institute of Physics of Cantabria (IFCA) entered in operation in summer 2012.
Its last generation FDR Infiniband network used for message passing in parallel jobs, also supports the connection to General Parallel File System (GPFS) servers, enabling an efficient processing of multiple data demanding jobs at the same time.

Sharing a common GPFS system with the existing GRID clusters at IFCA, and a single LDAP-based identification for users in both systems (supercomputer and local grid), allows CMS researchers to exploit the large instantaneous capacity of this supercomputer to execute analysis jobs.

The detailed experience describing this opportunistic use for skimming and final analysis of CMS 2012 data for an specific physics channel, resulting in a reduction of the waiting time of an order of magnitude, is presented.

Poster presentations / 141

ARC SDK: A toolbox for distributed computing and data applications

Authors: David Cameron[1]; Jonas Lindemann[2]; Martin Skou Andersen[3]

[1] *University of Oslo (NO)*

[2] *Lund University*

[3] *University of Copenhagen (DK)*

Corresponding Author: david.cameron@cern.ch

Grid middleware suites provide tools to perform the basic tasks of job submission and retrieval and data access, however these tools tend to be low-level, operating on individual jobs or files and lacking in higher-level concepts. User communities therefore generally develop their own application-layer software catering to their specific communities' needs on top of the Grid middleware. It is thus important for the Grid middleware to provide a friendly, well documented and simple to use interface for the applications build on. The Advanced Resource Connector (ARC), developed by NorduGrid, provides a Software Development Kit (SDK) which enables applications to use the middleware for job and data management. This paper presents the architecture and functionality of the ARC SDK along with an example graphical application developed with the SDK. The SDK consists of a set of libraries accessible through Application Programming Interfaces (API) in several languages. It contains extensive documentation and example code and is available on multiple platforms. The libraries provide generic interfaces and rely on plugins to support a given technology or protocol and this modular design makes it easy to add a new plugin if the application requires supporting additional technologies. The ARC Graphical Clients package is a graphical user interface built on top of the ARC SDK and the Qt toolkit and it is presented here as a fully functional example of an application. It provides a graphical interface to enable job submission and management at the click of a button, and allows data on any Grid storage system to be manipulated using a visual file system hierarchy, as if it were a regular file system.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / 246

Evolution of the ATLAS Distributed Computing system during the LHC Long Shutdown

Author: Simone Campana[1]

[1] *CERN*

Corresponding Author: simone.campana@cern.ch

The ATLAS Distributed Computing project (ADC) was established in 2007 to develop and operate a framework, following the ATLAS computing model, to enable data storage, processing and bookkeeping on top of the WLCG distributed infrastructure. ADC development has always been driven by operations and this contributed to its success. The system has fulfilled the demanding requirements of ATLAS, daily consolidating worldwide up to 1PB of data and running more than 1.5 million payloads distributed globally, supporting almost one thousand concurrent distributed analysis users. Comprehensive automation and monitoring minimized the operational manpower required. The flexibility of the system to adjust to operational needs has been important to the success of the ATLAS physics program.

The LHC shutdown in 2013-2015 affords an opportunity to improve the system in light of operational experience and scale it to cope with the demanding requirements of 2015 and beyond, most notably a much higher trigger rate and event pileup. We will describe the evolution of the ADC software foreseen during this period. This includes consolidating the existing Production and Distributed Analysis framework (PanDA) and ATLAS Grid Information System (AGIS), together with the development and commissioning of next generation systems for distributed data management (DDM/Rucio) and production (PRODSYS2). We will explain how new technologies such as Cloud Computing and NoSQL databases, which ATLAS investigated as R&D projects in past years, will be integrated in production. Finally, we will describe more fundamental developments such as breaking job-to-data locality by exploiting storage federations and caches, and event level (rather than file or dataset level) workload engines.

Poster presentations / **330**

A GPU offloading mechanism for LHCb

Authors: Alexander Zvyagin[1]; Alexey Badalov[2]; Daniel Hugo Campora Perez[3]

Co-authors: Niko Neufeld [3]; Xavier Vilasis Cardona [2]

[1] *Fakultaet fuer Physik-Ludwig-Maximilians-Univ. Muenchen*

[2] *La Salle - Ramon Llull University*

[3] *CERN*

The LHCb Software Infrastructure is built around a flexible, extensible, single-process, single-threaded framework named Gaudi. One way to optimise the overall usage of a multi-core server, which is used for example in the Online world, is running multiple instances of Gaudi-based applications concurrently. For LHCb, this solution has been shown to work well up to 32 cores and is expected to scale up a bit further.

The appearance of many-core architectures such as GPGPUs and the Intel Xeon/Phi poses a new challenge for LHCb. Since the individual data sets are so small (about 60 kB raw event size), many events must be processed in parallel for optimum efficiency. This is, however, not possible with the current framework, which allows only a single event at a time. Exploiting the fact that we always have many instances of the same application running, we have developed an offloading mechanism, based on a client-server design.

The server runs outside the Gaudi framework and thus imposes no additional dependencies on existing applications. It asynchronously receives event data from multiple client applications, coalesces them, and returns the computed results back to callers. While this incurs additional inter-process communication overhead, it allows co-processors to be used within the existing large framework with minimal changes. We present our solution and describe the achieved performance, both at the single-instance and the server levels.

Poster presentations / **329**

Time structure analysis of the LHCb Online network

Authors: Daniel Hugo Campora Perez[1]; Gianni Antichi[2]; Guoming Liu[1]; Marc Bruyere[3]

Co-authors: Andrew Moore [4]; Niko Neufeld [1]; Philippe Owezarski [5]; Stefano Giordano [2]

[1] *CERN*

[2] *Department of Information Engineering, University of Pisa*

[3] *1 CNRS, LAAS, 2 Université de Toulouse, LAAS, 3 DELL Inc*

[4] *University of Cambridge*

[5] *1 CNRS, LAAS, 2 Université de Toulouse, LAAS*

The LHCb Online Network is a real time high performance network, in which 350 data sources send data over a Gigabit Ethernet LAN to more than 1500 receiving nodes. The aggregated throughput of the application, called Event Building, is more than 60 GB/s. The protocol employed by LHCb makes the sending nodes transmit simultaneously portions of events to one receiving node at a time, which is selected using a credit-token scheme. The resulting traffic is very bursty and sensitive to irregularities in the temporal distribution of packet-bursts to the same destination or region of the network.

In order to study the relevant properties of such a dataflow, a non-disruptive monitoring setup based on a networking capable FPGA (NetFPGA) has been deployed. The NetFPGA allows order of hundred nano-second precise time-stamping of packets. We study in detail the timing structure of the Event Building communication, and we identify potential effects of micro-bursts like buffer packet drops or jitter.

Plenaries / **489**

Future directions for key physics software packages

Author: Philippe Canal[1]

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

Corresponding Author: philippe.canal@cern.ch

Developments in many of our key software packages, such as Root 6 and the next generation Geant, will have a significant impact on the way analysis is done. Dr. Philippe Canal will present the birds-eye view on where these developments can lead us, on the way next generation ROOT and Geant can be combined, and on how for example the increased use of concurrency in these key software packages will impact physics analysis tomorrow.

Software Engineering, Parallelism & Multi-Core / **3**

High Energy Electromagnetic Particle Transportation on the GPU

Author: Soon Yung Jun[1]

Co-authors: Jim Kowalkowski [2]; John Apostolakis [3]; Marc Paterno [2]; Philippe Canal [1]; Victor Daniel Elvira [1]

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

[2] *Fermilab*

[3] *CERN*

Corresponding Authors: philippe.canal@cern.ch, soon.yung.jun@cern.ch

We will present massively parallel high energy electromagnetic particle transportation through a finely segmented detector in the Graphic Processor Unit (GPU). Simulating events of energetic particle decay in a general-purpose high energy physics (HEP) detector requires intensive computing resources, due to the complexity of the geometry as well as physics processes applied to particles copiously produced by primary collisions and secondary interactions. The recent advent of hardware architectures of many-core or accelerated
processors provides the variety of concurrent programming models applicable not only for the high performance parallel computing, but also for the conventional computing intensive application such as the HEP detector simulation. The component of the transportation prototype consists of a transportation process under a non-uniform magnetic field, a geometry navigation with a set of solid shapes and materials, electromagnetic physics processes for electrons and
photons, and an interface to a framework that dispatches bundles of tracks in a highly vectorized manner optimizing for spatial locality and throughput. Core algorithms and methods are excerpted from the Geant4 toolkit, and are modified and optimized for the GPU application. Programs written in C/C++ are designed to be compatible with CUDA and openCL and generic enough for future variations of programming models and hardware architectures. Used with multiple
streams, asynchronous kernel executions are overlapped with concurrent data transfers of streams of tracks to balance arithmetic intensity and memory bandwidth. Issues with floating point accuracy, random number generation, data structure, kernel divergences and register spills are also considered. Performance evaluation for the relative speedup compared to the corresponding sequential execution on CPU will be presented as well.

Software Engineering, Parallelism & Multi-Core / 476

The path toward HEP High Performance Computing

Author: Federico Carminati[1]

Co-authors: Andrei Gheata [1]; John Apostolakis [1]; Rene Brun [1]

[1] *CERN*

Corresponding Author: federico.carminati@cern.ch

High Energy Physics code has been known for making poor use of high performance computing architectures. Efforts in optimising HEP code on vector and RISC architectures have yield limited results and recent studies have shown that, on modern architectures, it achieves a performance between 10% and 50% of the peak one. Although several successful attempts have been made to port selected codes on GPUs, no major HEP code suite has a “High Performance” implementation. With LHC undergoing a major upgrade and a number of challenging experiments on the drawing board, HEP cannot any longer neglect the less-than-optimal performance of its code and it has to try making the best usage of the hardware. This activity is one of the foci of the SFT group at CERN, which hosts, among others, the Root and Geant 4 projects. The activity of the experiments is shared and coordinated via a Concurrency Forum, where the experience in optimising HEP code is presented and discussed. Another activity is centred on the development of a high-performance prototype for particle transport. Achieving a good concurrency level on the emerging parallel architectures without a complete redesign of the framework can be done only by parallelizing at the event level, or

with a much larger effort at the track level. Apart from the shareable data structures, this typically implies a multiplication factor in terms of memory consumption compared to the single threaded version, together with sub-optimal handling of event processing tails. Besides this, the low level instruction pipelining of modern processors cannot be used efficiently to speedup the program. We have implemented a framework that allows scheduling vectors of particles on an arbitrary number of computing resources in a fine grain parallel approach. The talk will review the current optimisation activities within the SFT group with a particular emphasis on the development perspectives towards a simulation framework able to profit best from the recent technology evolution in computing.

Event Processing, Simulation and Analysis / 153

Alignment and calibration of CMS detector during collisions at LHC

Author: Roberto Castello[1]

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

Corresponding Author: roberto.castello@cern.ch

Fast and efficient methods for the calibration and the alignment of the detector play a key role in ensuring reliable physics performance to an HEP experiment. CMS has set up a solid framework for alignment and calibration purpose, in close contact with the detector and physics needs. The about 200 types of calibration and alignment existing for the various sub-detectors are collected by relational databases (Oracle) flexible enough to allow easy access as requested by the various users. The CMS alignment and calibration infrastructure is also designed to ensure a proper handling of inter-dependencies among the different calibration workflows, and includes dedicated streams of data for a fast turnaround of the calibration process during the data taking. This presentation reviews the design of the system and reports on the experience gained during its operation including results from selected workflows contributing to main physics achievements of the CMS experiment during last years.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / 140

Recent and planned changes to the LHCb computing model

Author: Marco Cattaneo[1]

Co-authors: Peter Clarke [2]; Philippe Charpentier [1]; Stefan Roiser [1]

[1] *CERN*

[2] *University of Edinburgh (GB)*

Corresponding Author: marco.cattaneo@cern.ch

The LHCb experiment has taken data between December 2009 and February 2013. The data taking conditions and trigger rate have been adjusted several times to make optimal use of the luminosity delivered by the LHC and to extend the physics potential of the experiment.

By 2012, LHCb was taking data at twice the instantaneous luminosity and 2.5 times the high level trigger rate than originally foreseen. This represents a considerable increase in the amount of data to be handled compared to the original Computing Model from 2005, both in terms of compute power and in terms of storage.

In this paper we describe the changes that have taken place in the LHCb computing model during the last 2 years of data taking to process and analyse the increased data rates within limited computing resources. In particular a quite original change was introduced at the end of 2011 when LHCb started to use for reprocessing compute power that was not co-located with the RAW data, namely using Tier2 sites and private resources. The flexibility of the LHCbDirac Grid interware allowed to easily include these additional resources that in 2012 provided 40% of the compute power for the end-of-year reprocessing. Several changes were also implemented on the Data Management model in order to limit the need for accessing data from tape, as well as in the data placement policy in order to cope with a large imbalance in storage resources at Tier1 sites.

We also discuss changes that are being implemented during the LHC Long Shutdown 1 to prepare for a further doubling of the data rate when the LHC restarts at a higher energy in 2015.

Poster presentations / 501

External access to ALICE controls conditions data

Authors: Andre Augustinus[1]; Peter Chochula[1]

Co-authors: Anna Jadlovská [2]; Jakub Cerkala [3]; Ján Jadlovský [2]; Ján Sarnovský [2]; Matej Čopík [2]; Michal Kopčík [2]; Peter Papcun [2]; Radoslav Bielek [2]; Slávka Jadlovská [2]; Štefan Jajčišin [2]

[1] *CERN*

[2] *Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Košice*

[3] *Technical University of Košice*

Corresponding Authors: slavka.jadlovska@tuke.sk, jakub.cerkala@tuke.sk, peter.chochula@cern.ch

ALICE Controls data produced by commercial SCADA system WINCCOA is stored in ORACLE database on the private experiment network. The SCADA system allows for basic access and processing of the historical data. More advanced analysis requires tools like ROOT and needs therefore a separate access method to the archives.

The present scenario expects that detector experts create simple WINCC OA scripts, which retrieves and stores data in a form usable for further studies. This relatively simple procedure generates a lot of administrative overhead - users have to request the data, experts needed to run the script, the results have to be exported outside of the experiment network. The new mechanism profits from database replica, which is running on the CERN campus network. Access to this database is not restricted and there is no risk of generating a heavy load affecting the operation of the experiment.

The developed tools presented in this paper allow for access to this data. The users can use web-based tools to generate the requests, consisting of the data identifiers and period of time of interest. The administrators maintain full control over the data - an authorization and authentication mechanism helps to assign privileges to selected users and restrict access to certain groups of data. Advanced caching mechanism allows the user to profit from the presence of already processed data sets. This feature significantly reduces the time required for debugging as the retrieval of raw data can last tens of minutes. A highly configurable client allows for information retrieval bypassing the interactive interface. This method is for example used by ALICE Offline to extract operational conditions after a run is completed. Last but not least, the software can be easily adopted to any underlying database structure and is therefore not limited to WINCCOA.

Data Stores, Data Bases, and Storage Systems / 355

Fuzzy Pool Balance: An algorithm to achieve two dimensional balances in distribute storage systems

Author: Wenjing Wu[1]

Co-author: Gang CHEN [2]

[1] *IHEP, CAS*

[2] *INSTITUTE OF HIGH ENERGY PHYSICS*

Corresponding Authors: wuwj@ihep.ac.cn, gang.chen@ihep.ac.cn

The limitation of scheduling modules and the gradual addition of disk pools in distributed storage systems often result in imbalances among their disk pools in terms of both available space and number of files. This can cause various problems to the storage system such as single point of failure, low system throughput and imbalanced resource utilization and system loads. An algorithm named Fuzzy Pool Balance (FPB) is proposed here to solve this problem. The input of FPB is the current file distribution among disk pools and the output is a file migration plan indicating what files are to be migrated to which pools. FPB uses an array to classify the files by their sizes. The file classification array is dynamically calculated with a defined threshold named Tmax which defines the allowed available space deviations of disk pools. File classification is the basis of file migration. FPB also defines the Immigration Pool (IP) and Emigration Pool (EP) according to the available space of the disk pools and File Quantity Ratio (FQR) which indicates the percentage of each category of files in each disk pool, so files with higher FQR in an EP will be migrated to IP(s) with a lower FQR of this file category. To verify this algorithm, we implemented FPB on an ATLAS Tier2 dCache production system which hosts 12 distributed disk pools with 300TB of storage space. The results show that FPB can achieve a very good balance among the disk pools, and a tradeoff between available space and file quantity can be achieved by adjusting the threshold value Tmax and the correction factor to the average FQR.

Poster presentations / 49

gluster file system optimization and deployment at IHEP

Author: Yaodong CHENG[1]

[1] *Institute of High Energy Physics,Chinese Academy of Sciences*

Corresponding Author: chyd@ihep.ac.cn

Gluster file system adopts no metadata architecture, which theoretically eliminates both a central point of failure and a performance bottleneck of metadata server. Firstly, this talk will introduce gluster compared to lustre or hadoop. However, its some mechanisms are not so good in current version. For example, it has to read the extend attributes of all bricks to locate one file. And it is slow to list files in one directory when there are too many bricks or brick servers are busy. Some other functions, such as expand or shrink volume, file distribution, replication policy and so on, performs not so well in large scale storage system. This talk will analyze the advantages and disadvantages of gluster file system in high performance computing system. To solve these problems, we proposed some new methods, optimized or developed some modules, including modifying elastic hash algorithm implementation, introducing a new index module, designing a new replication layer, and on so. This talk will introduce these methods or functions. We already deployed a gluster file system in production data analysis environment. The talk finally describes the deployment scenario and some lessons we can learn from it.

Poster presentations / 23

A New Nightly Build System for LHCb

Author: Marco Clemencic[1]

Co-author: Ben Couturier [1]

[1] *CERN*

Corresponding Author: marco.clemencic@cern.ch

The nightly build system used so far by LHCb has been implemented as an extension on the system developed by CERN PH/SFT group (as presented at CHEP2010). Although this version has been working for many years, it has several limitations in terms of extensibility, management and ease of use, so that it was decided to develop a new version based on a continuous integration system.

In this paper we describe a new implementation of the LHCb Nightly Build System based on the open source continuous integration system Jenkins and report on the experience on the configuration of a complex build workflow in Jenkins.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / 98

The CMS openstack, opportunistic, overlay, online-cluster Cloud (CMSooooCloud)

Author: Jose Antonio Coarasa Perez[1]

Co-authors: Andre Georg Holzner ; Andrea Petrucci [1]; Andrei Cristian Spataru [1]; Attila Racz [1]; Aymeric Arnaud Dupont [1]; Carlos Nunez Barranco Fernandez [1]; Christian Deldicque [1]; Christian Hartl [1]; Christoph Paus [2]; Christoph Schwick [1]; Christopher Colin Wakefield [3]; Dominique Gigi [1]; Emilio Meschi [1]; Fabian Stoeckli [2]; Frank Glege [1]; Frans Meijers [1]; Gerry Bauer [2]; Giovanni Polese [4]; Hannes Sakulin [1]; James Gordon Branson [5]; Konstanty Sumorok [2]; Lorenzo Masetti [1]; Luciano Orsini [1]; Marc Dobson [1]; Marco Pieri [5]; Matteo Sani [5]; Olivier Chaze [1]; Olivier Raginel [2]; Petr Zejdl [1]; Remi Mommsen [6]; Robert Gomez-Reino Garrido [1]; Samim Erhan [7]; Sergio Cittolin [5]; Srecko Morovic [8]; Ulf Behrens [9]; Vivian O'Dell [10]; Wojciech Andrzej Ozga [11]

[1] *CERN*

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

[3] *Staffordshire University (GB)*

[4] *University of Wisconsin (US)*

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

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

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

[8] *Institute Rudjer Boskovic (HR)*

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

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

[11] *AGH University of Science and Technology (PL)*

Corresponding Author: jose.antonio.coarasa.perez@cern.ch

The CMS online cluster consists of more than 3000 computers. It has been exclusively used for the Data Acquisition of the CMS experiment at CERN, archiving around 20Tbytes of data per day.
An openstack cloud layer has been deployed on part of the cluster (totalling more than 13000 cores) as a minimal overlay so as to leave the primary role of the computers untouched while allowing an opportunistic usage of the cluster. This allows running offline computing jobs on the online infrastructure while it is not (fully) used.
We will present the architectural choices made to deploy an unusual, as opposed to dedicated, "overlaid cloud infrastructure". These architectural choices ensured a minimal impact on the running cluster configuration while giving a maximal segregation of the overlaid virtual computer infrastructure. Openvswitch was chosen during the proof of concept phase in order to avoid changes on

the network infrastructure. Its use will be illustrated as well as the final networking configuration used. The design and performance of the openstack cloud controlling layer will be also presented together with new developments and experience from the first year of usage.

Facilities, Infrastructures, Networking and Collaborative Tools / **121**

The CMS openstack, opportunistic, overlay, online-cluster Cloud (CMSooooCloud)

Author: Jose Antonio Coarasa Perez[1]

[1] *CERN*

Corresponding Author: jose.antonio.coarasa.perez@cern.ch

The CMS online cluster consists of more than 3000 computers. It has been exclusively used for the Data Acquisition of the CMS experiment at CERN, archiving around 20Tbytes of data per day. An openstack cloud layer has been deployed on part of the cluster (totalling more than 13000 cores) as a minimal overlay so as to leave the primary role of the computers untouched while allowing an opportunistic usage of the cluster. This allows running offline computing jobs on the online infrastructure while it is not (fully) used. We will present the architectural choices made to deploy an unusual, as opposed to dedicated, "overlaid cloud infrastructure". These architectural choices ensured a minimal impact on the running cluster configuration while giving a maximal segregation of the overlaid virtual computer infrastructure. Openvswitch was chosen during the proof of concept phase in order to avoid changes on the network infrastructure. Its use will be illustrated as well as the final networking configuration used. The design and performance of the openstack cloud controlling layer will be also presented together with new developments and experience from the first year of usage.

Poster presentations / **437**

An Agile Service Deployment Framework and its Application

Author: Matthew James Viljoen[1]

Co-author: Ian Collier [2]

[1] *STFC - Science & Technology Facilities Council (GB)*

[2] *UK Tier1 Centre*

Corresponding Authors: matthew.viljoen@cern.ch, ian.peter.collier@cern.ch

In this paper we shall introduce the service deployment framework based on Quattor and Microsoft HyperV at the RAL Tier 1. As an example, we will explain how the framework has been applied to CASTOR in our test
infrastructure and outline our plans to roll it out into full production. CASTOR is a relatively complicated open source hierarchical storage management system in production use at RAL for both WLCG and large scale scientific facilities data.

Finally, we will examine different approaches of virtualizing CASTOR using HyperV and we will present an optimal approach along with the advantages and disadvantages of each approach.

Poster presentations / **392**

CernVM-FS - Beyond LHC Computing

Author: Ian Peter Collier[1]

Co-author: Catalin Condurache [1]

[1] *STFC - Science & Technology Facilities Council (GB)*

Corresponding Author: ian.peter.collier@cern.ch

In the last three years the CernVM Filesystem (CernVM-FS) has transformed the distribution of experiment software to WLCG grid sites. CernVM-FS removes the need for local installations jobs and performant network fileservers at sites, in addition it often improves performance at the same time. Furthermore the use of CernVM-FS standardizes the computing environment across the grid and removes the need for software tagging at sites.

Now established and proven to work at scale, CernVM-FS is beginning to perform a similar role for non-LHC computing.

We discuss the deployment of a Stratum 0 'master' CernVM-FS repository at the RAL Tier 1 and the development of a network of Stratum 1 replicas somewhat modeled upon the infrastructure developed to support WLCG computing.

We include a case study of one small non-LHC virtual organisation, describing their use of the CernVM-FS Stratum 0 service. We examine the impact of using CernVM-FS upon their ability to utilize a wider range of resources across the UK GridPP network.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / 93

Usage of the CMS Higher Level Trigger Farm as a Cloud Resource

Authors: Claudio Grandi[1]; David Colling[2]; Mattia Cinquilli[3]

[1] *INFN - Bologna*

[2] *Imperial College Sci., Tech. & Med. (GB)*

[3] *CERN*

Corresponding Authors: d.colling@imperial.ac.uk, mattia.cinquilli@cern.ch

The Higher Level Trigger (HLT) farm in CMS is a more than ten thousand core processor farm that is heavily used during data acquisition and largely unused when the detector is off. In this presentation we will cover the work done in CMS to utilize this large processing resource with cloud resource provisioning techniques. This resource when configured with Open Stack and Agile Infrastructure techniques virtualization and resource provisions has many similarly attributes to large scale commercial clouds, but without the cost per core so it provides a unique resource to test. When the facility moves to production it will also represent a large increase in the production capacity available to CMS. We will cover the work on resource provisioning through the EC2 interface using the CMS pilot submission infrastructure, glide-in WMS; the configuration and contextualization of the virtual machines; and the configuration of the local environment and the execution of CMS reprocessing workflows.

Poster presentations / 215

Preserving access to ALEPH Computing Environment via Virtual Machines

Author: Simone Coscetti[1]

Co-authors: Andrea Domenici [2]; Cinzia Bernardeschi [2]; Marcello Maggi [3]; Tommaso Boccali [1]

[1] *Sezione di Pisa (IT)*

[2] *University of Pisa*

[3] *Universita e INFN (IT)*

Corresponding Author: simone.coscetti@cern.ch

The ALEPH Collaboration took data at the LEP (CERN) electron-positron collider in the period 1989-2000, producing more than 300 scientific papers. While most of the Collaboration activities stopped in the last years, the data collected still has physics potential, with new theoretical models emerging, and needing a check with data at the Z and WW production energies. An attempt to revive and preserve the ALEPH Computing Environment is presented; the aim is not only the preservation of the data files (usually called "bit preservation"), but of the full environment a physicist would need to perform brand new analyses. Technically, a Virtual Machine approach has been chosen, using the VirtualBox platform. Concerning simulated events, the full chain from event generators to physics plots is possible, and reprocessing of data events is also functioning. Interactive tools like the DALI event display can be used on both data and simulated events. The Virtual Machine approach seems suited for both interactive usage, and for massive computing using Cloud like approaches. Studies are now moving from technical functionality tests (which are positively concluded), to tests and development on how to guarantee an easy and transparent access to ALEPH data in the virtualized platform.

Event Processing, Simulation and Analysis / **177**

Geant4 - Towards major release 10

Author: Gabriele Cosmo[1]

[1] *CERN*

Corresponding Author: gabriele.cosmo@cern.ch

The Geant4 simulation toolkit has reached maturity in the middle of the previous decade, providing a wide variety of established features coherently aggregated in a software product which has become the standard for detector simulation in HEP and is used in a variety of other application domains.

We review the most recent capabilities introduced in the kernel, highlighting those which are being prepared for the next major release (version 10.0) that is scheduled for the end of 2013.

A significant new feature contained in this release will be the integration of multi-threading processing, aiming at targeting efficient use of modern many-cores system architectures and minimisation of the memory footprint for exploiting event-level parallelism.
We discuss its design features and impact on the existing API and user-interface of Geant4. Revisions are made to balance the need for preserving backwards compatibility and to consolidate and improve the interfaces, taking into account requirements from the multi-threaded extensions and from the evolution of the data processing models of the LHC experiments.

Software Engineering, Parallelism & Multi-Core / **475**

Semi-automatic SIMD-efficient data layouts for object-oriented programs

Author: Pascal Costanza[1]

[1] *ExaScience Lab, Intel, Belgium*

Corresponding Author: pascal.costanza@intel.com

Using Intel's SIMD architecture (SSE, AVX) to speed up operations on containers of complex class and structure objects is challenging, because it requires that the same data members of the different objects within a container have to be laid out next to each other, in a structure of arrays (SOA) fashion. Currently, programming languages do not provide automatic ways for arranging containers as structures of arrays. Instead, programmers have to change the data layout manually, which requires changed definitions of the classes involved that are at odds with object-oriented programming principles. This step is usually considered too invasive, and the common result is that programmers simply give up on the performance opportunity.

I will present a novel technique, made possible by new C++11 capabilities, which bridges the gap and allows the programmer to lay out data in a structure of array format, and yet access the data members using the standard object-oriented programming style. The main value of the presented technique is that it makes the efficient use of Intel's SIMD architecture with C++ classes and structures substantially easier and thus available to a much wider audience of software developers.

Poster presentations / 295

The Role of the Collaboratory in enabling Large-Scale Identity Management for HEP

Author: Von Welch[1]

Co-authors: Bob Cowles [2]; S. Craig Jackson [1]

[1] *University of Indiana / CACR*

[2] *BrightLite Information Security*

Corresponding Author: bob.cowles@gmail.com

As HEP collaborations grow in size (10 years ago, BaBar was 600 scientists; now, both CMS and ATLAS are on the order of 3000 scientists), the collaboratory has become a key factor in allowing identity management (IdM), once confined to individual sites, to scale with the number of members, number of organizations, and the complexity of the science collaborations. Over the past two decades (at least) there has been a great deal of applied research and success in implementing collaboratories, but there has also been a great deal of controversy and variety of implementations in the community. A common implementation, or even a model for contrasting different implementations, does not yet exist. This lack of common approach makes collaboration between existing collaboratories and establishment of new collaboratories a challenge.

The eXtreme Scale Identity Management (XSIM) project is addressing this short-coming by defining a model for IdM that captures existing and future collaboratory implementations. XSIM is first capturing an understanding of the trust relationships in today's scientific collaborations and their resource providers and analyzing how the trade-offs between the policies and trust relationships affect current IdM models. This understanding is being developed through a review of existing literature and one-on-one interviews with dozens of members of the communities involved to fully understand the motivations for the decisions and the lessons learned.

Building on this research, XSIM is proposing a model for identity management that describes the core trust relationships between HEP collaborations and resource providers, and the different choices for those relationships, both in terms of levels and types of trust, and implementation. The model must be sufficiently comprehensive to encompass the reality of the existing IdM architectures; be understandable and useful to future collaboratory developers who are not IdM experts; relate well to efforts in the HEP community; and be accepted by resource providers. Developing such a model will give the community a language in which to express differences in identity management solutions, and easily communicate and understand the impacts of changes in the trust relationships

involved with different choices. This in turn will expedite understanding and establishment of new collaborations.

The presentation will provide a summary of the interviews and literature, the resulting analysis, and a model that captures the core trust relationships, especially those relating to IdM.

Poster presentations / **369**

Changing the batch system in a Tier 1 computing center: why and how

Authors: Andrea Chierici[1]; Stefano Dal Pra[2]

[1] *INFN-CNAF*

[2] *Unknown*

Corresponding Authors: stefano.dalpra@cnaf.infn.it, chierici@cnaf.infn.it

At the Italian Tier1 Center at CNAF we are evaluating the possibility to change the current production batch system. This activity is motivated mainly because we are looking for a more flexible licensing model as well as to avoid vendor lock-in.

We performed a technology tracking exercise and among many possible solutions we chose to evaluate Grid Engine as an alternative because its adoption is increasing in the HEPiX community and because it's supported by the EMI middleware that we currently use on our computing farm.
Another INFN site evaluated Slurm and we will compare our results in order to understand pros and cons of the two solutions.

We will present the results of our evaluatin of Grid Engine, in order to understand if it can fit the requirements of a Tier 1 center, compared to the solution we adopted long ago.
We performed a survey and a critical re-evaluation of our farming infrastructure: many production softwares (accounting and monitoring on top of all) rely on our current solution and changing it required us to write new wrappers and adapt the infrastructure to the new system.

We believe the results of this investigation can be very useful to other Tier-1s and Tier-2s centers in a similar situation, where the effort of switching may appear too hard to stand.
We will provide guidelines in order to understand how difficult this operation can be and how long the change may take.

Poster presentations / **298**

Installation and configuration of an SDN test-bed made of physical switches and virtual switches managed by an Open Flow controller.

Author: Stefano Zani[1]

Co-authors: Donato De Girolamo [1]; Lorenzo Chiarelli [1]

[1] *INFN CNAF*

Corresponding Authors: stefano.zani@cnaf.infn.it, lorenzo.chiarelli@cnaf.infn.it, donato.degirolamo@cnaf.infn.it

The computing models of HEP experiments, starting from the LHC ones, are facing an evolution with the relaxation of the data locality paradigm: the possibility of a job accessing data files over the WAN is becoming more and more common.

One of the key factors for the success of this change is the ability to use the network in the most efficient way: in the best scenario, the network should be capable to change its behavior on the base of a per flow analysis.

The SDN (Software Defined Networks) are a promising technology to address this challenging requirement and OpenFlow is a candidate protocol to implement it. At CNAF, we have installed a Software Defined Networks test-bed functional to build a concrete layout where testing OpenFlow protocol and the interoperability between devices of different vendors.

The objective of this activity is to build a network with a control plane driven by a software layer in charge of defining the "route" for a specific flow, based on conditions verified inside a datacenter network, in order to solve routing problems not addressable with standard networking protocols.

Poster presentations / **151**

The CMS Data Quality Monitoring software: experience and future improvements

Author: Marco Rovere[1]

Co-author: Federico De Guio [1]

[1] *CERN*

Corresponding Authors: federico.de.guio@cern.ch, marco.rovere@cern.ch

The Data Quality Monitoring (DQM) Software proved to be a central tool in the CMS experiment. Its flexibility allowed its integration in several environments: Online, for real-time detector monitoring; Offline, for the final, fine-grained Data Certification; Release-Validation, to constantly validate our reconstruction software; in Monte Carlo productions. The central tool to deliver Data Quality information is a web site for browsing data quality histograms (DQMGUI). In this presentation 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. The main technical challenges and the adopted solutions to them will be also discussed with emphasis on functionality, long-term robustness and performance. Finally the experience in operating the DQM systems over the past years will be reported.

Poster presentations / **375**

A quasi-online distributed data processing on WAN: the ATLAS muon calibration system.

Author: Enrico Pasqualucci[1]

[1] *INFN Roma*

Corresponding Author: alessandro.de.salvo@cern.ch

In the Atlas experiment, the calibration of the precision tracking chambers of the muon detector is very demanding, since the rate of muon tracks required to get a complete calibration in homogeneous conditions and to feed prompt reconstruction with fresh constants is very high (several hundreds Hz for 8-10 hours runs). The calculation of calibration constants is highly CPU consuming. In order to fulfill the requirement of completing the cycle and having the final constants available within

24 hours, distributed resources at Tier-2 centers have been allocated.
The best place to get muon tracks suitable for detector calibration is the second level trigger, where the pre-selection of data sitting in a limited region by the first level trigger via the Region of Interest mechanism allows selecting all the hits from a single track in a limited region of the detector. Online data extraction allows calibration data collection without performing special runs. Small event pseudo-fragments (about 0.5 kB) built at the muon level-1 rate (2-3 kHz at the beginning of 2012 run, to become 10-12 kHz at maximum LHC luminosity) are then collected in parallel by a dedicated system, without affecting the main data taking, and sent to the Tier-0 computing center at CERN.
The computing resources needed to calculate the calibration constants are distributed through three calibration centers (Rome, Munich, Ann Arbor) for the tracking device and one (Napoli) for the trigger chambers. From Tier-0, files are directly sent to the calibration centers through the ATLAS Data Distribution Manager.
At the calibration centers, data is split per trigger tower and distributed to computing nodes for concurrent processing (˜250 cores are currently used at each center). A two-stage processing is performed, the first stage reconstructing tracks and creating ntuples, the second one calculating constants. The calibration parameters are then
stored in the local calibration database and replicated to the main condition database at CERN, which makes them available for data analysis within 24 hours from data extraction.
The architecture and performance of this system during the 2011-2012 data taking will be presented.
This system will evolve in the next future to comply with the new stringent requirements of the LHC and ATLAS upgrade. If for the WAN distribution part the availability of bandwidth is already much larger than needed for this task and the CPU power can be increased according to our need, the online part will follow the evolution of the ATLAS TDAQ architecture. In particular, the current model foresees the merging of the level-2 and event filtering processes on the same nodes, allowing the simplification of the system and a more flexible and dynamic resource distribution. Two possible architectures are possible to comply with this model; possible implementation will be discussed.

Poster presentations / **178**

Tier-1 Site Evolution in Response to Experiment Requirements

Authors: Jhen-Wei Huang[1]; Shaun De Witt[2]

[1] *ASGC*

[2] *STFC - Science & Technology Facilities Council (GB)*

Corresponding Author: shaun.de-witt@stfc.ac.uk

LHC experiments are moving away from a traditional HSM solution for Tier 1's in order to separate long term tape archival from disk only access, using the tape as a true archive (write once, read rarely). In this poster we present two methods by which this is being achieved at two distinct sites, ASGC and RAL, which have approached this change in very different ways.

Poster presentations / **169**

Managing and throttling federated xroot across WLCG Tier 1s

Author: Shaun De Witt[1]

Co-author: Andrew David Lahiff [1]

[1] *STFC - Science & Technology Facilities Council (GB)*

Corresponding Author: shaun.de-witt@stfc.ac.uk

WLCG is moving towards greater use of xrootd. While this will in general optimise resource usage on the grid, it can create load problems at sites when storage elements are unavailable. We present some possible methods of mitigating these problems and the results from experiments at STFC

Poster presentations / **172**

Analysis of Alternative Storage Technologies for the RAL Tier 1

Author: Shaun De Witt[1]

Co-authors: Brian Davies [2]; Ian Collier [3]; James Adams [4]; Matthew James Viljoen [1]; Robert Appleyard [5]

[1] *STFC - Science & Technology Facilities Council (GB)*

[2] *Lancaster University (GB)*

[3] *UK Tier1 Centre*

[4] *STFC RAL*

[5] *STFC*

Corresponding Author: shaun.de-witt@stfc.ac.uk

At the RAL Tier 1 we have successfully been running a CASTOR HSM instance for a number of years. While it performs well for disk-only storage for analysis and processing jobs, it is heavily optimised for tape usage. We have been investigating alternative technologies which could be used for online storage for analysis. We present the results of our preliminary selection and test results for selected storage technologies.

Event Processing, Simulation and Analysis / **290**

Concepts for fast large scale Monte Carlo production for the ATLAS experiment

Author: Chiara Debenedetti[1]

[1] *University of Edinburgh (GB)*

Corresponding Author: chiara.debenedetti@cern.ch

The huge success of Run 1 of the LHC would not have been possible without detailed detector simulation of the experiments. The outstanding performance of the accelerator with a delivered integrated luminosity of 25 fb-1 has created an unprecedented demand for large simulated event samples. This has stretched the possibilities of the experiments due to the constraint of their computing infrastructure and available resources. Modern, concurrent computing techniques optimized for new processor hardware are being exploited to boost future computing resources, but even the most optimistic scenarios predict that additional action needs to be taken to guarantee sufficient Monte Carlo production statistics for high quality physics results during Run 2.

In recent years, the ATLAS collaboration has put dedicated effort in the development of a new Integrated Simulation Framework (ISF) that allows running full and fast simulation approaches in parallel and even within one event. We present the main concepts of the ISF, which allows a fine-tuned detector simulation targeted at specific physics cases with a decrease in CPU time per event

by orders of magnitude. Additionally, we will discuss the implications of a customized simulation in terms of validity and accuracy and will present new concepts in digitization and reconstruction to achieve a fast Monte Carlo chain with a per event execution time of a few seconds.

Poster presentations / 311

Long Term Data Preservation for CDF at INFN-CNAF

Author: Luca dell'Agnello[1]

Co-authors: Pier Paolo Ricci [2]; Silvia Amerio [3]; Stefano Zani

[1] *INFN-CNAF*

[2] *INFN CNAF*

[3] *University of Padova & INFN*

Long-term preservation of experimental data (intended as both raw and derived formats) is one of the emerging requirements coming from scientific collaborations. Within the High Energy Physics community the Data Preservation in High Energy Physics (DPHEP) group coordinates this effort.
CNAF is not only one of the Tier-1s for the LHC experiments, it is also a computing center providing computing and storage resources to many other HEP and non-HEP scientific collaborations, including the CDF experiment. After the end of data taking in 2011, CDF is now facing the challenge to both preserve the large amount of data produced during several years of data taking and to retain the ability to access and reuse it in the future.
CNAF is heavily involved in the CDF Data Preservation activities, in collaboration with the FNAL computing sector. At the moment about 5 PB of data (raw data and analysis-level "ntuples") are being copied from FNAL to the CNAF tape library and the framework to subsequently access the data is being set up. In parallel to the data access system, a data analysis framework is being developed which allows to run the complete CDF analysis chain in the long term future, from raw data reprocessing to analysis-level "ntuple"production. In this contribution we illustrate the technical solutions we put in place to address the issues encountered as we proceeded in this activity.

Facilities, Infrastructures, Networking and Collaborative Tools / 318

WAN Data Movement Architectures at US-LHC Tier-1s

Author: Phil Demar[1]

Co-author: Scott Bradley [2]

[1] *Fermilab*

[2] *Brookhaven National Laboratory*

Corresponding Author: demar@fnal.gov

LHC networking has always been defined by high volume data movement requirements in both LAN and WAN. LAN network demands can typically be met fairly easily with high performance data center switches, albeit at high cost. LHC WAN data movement, on the other hand, presents a more complicated and difficult set of challenges. Typically, there are three high-level issues a high traffic volume LHC site needs to deal with in providing a quality LHC WAN service:

- Ensuring sufficient bandwidth capacity for the LHC data
- Protecting the site's other WAN traffic from being negatively impacted by LHC traffic flows
- Contending with the site's perimeter security policies and mechanisms

The emergence of alternate network paths specifically for LHC data movement has provided a means for many LHC sites to appropriately deal with the first two issues. The LHCOPN and LHCONE are examples of physical and virtual network infrastructure respectively that enable sites to direct their LHC WAN traffic over adequately provisioned, isolated network paths. However, the site must still deal with its local security policies to move that traffic through its perimeter. Historically, the firewalls and security tools used to implement local security policies have not been capable of keeping up with LHC WAN traffic loads. This problem normally necessitates use of perimeter bypass mechanisms. ESnet has pioneered in the development of the Science DMZ, a general architecture for separating high impact science data flows from a site's normal routed internet traffic. Like any architecture, implementation varies according to circumstances and conditions. This presentation will discuss the concept of the science DMZ architecture, with a focus on implementation of that concept at the two US Tier-1 facilities, Fermilab (CMS) and Brookhaven National Laboratory (Atlas). The talk will discuss how each US Tier-1 has structured and configured its network perimeter infrastructure to meet the demands of its LHC WAN data movement, while still maintaining a secure network perimeter consistent with its overall security policies. Particular emphasis will be given to deployment of 100GE WAN technology on the site perimeter. Both US Tier-1 facilities are currently in the process of deploying 100GE support for their LHC data movement, and implementation details will be covered.

Poster presentations / 63

Squid monitoring tools - a common solution for LHC experiments.

Authors: Alastair Dewhurst[1]; Dave Dykstra[2]

Co-authors: Alessandro Di Girolamo [3]; Andrea Valassi [3]; Barry Jay Blumenfeld [4]; Luis Emiro Linares Garcia [5]; Simone Campana [3]

[1] *STFC - Science & Technology Facilities Council (GB)*

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

[3] *CERN*

[4] *Johns Hopkins University (US)*

[5] *Universidad de los Andes (CO)*

Corresponding Authors: alastair.dewhurst@cern.ch, dwd@fnal.gov

During the early running of the LHC , multiple collaborations began to include Squid caches in their distributed computing models. The two main use cases are: for remotely accessing conditions data via Frontier, which is used by ATLAS and CMS; and serving collaboration software via CVMFS, which is used by ATLAS, CMS, and LHCb, and is gaining traction with some non-LHC collaborations. As a result, Squid cache deployment across the grid has rapidly increased, with hundreds of Squids now in production. While some effort has been made between collaborations to share monitoring tools, this had been done on an ad-hoc basis, and it was found that some sites had duplicated monitoring, while others had no monitoring of their cache service at all. The WLCG Squid Monitoring Task Force was established to decide how to improve and better deploy Squid monitoring with WLCG common operations, and produce an architecture for a common Squid monitoring system configuration for use by all collaborations. In this paper, we present the recommendations produced by the Task Force, and the subsequent work necessary to implement them.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / 8

Towards a Global Service Registry for the World-wide LHC Computing Grid

Authors: Alessandro Di Girolamo[1]; Laurence Field[1]; Maria Alandes Pradillo[1]

[1] *CERN*

Corresponding Authors: alessandro.di.girolamo@cern.ch, maria.alandes.pradillo@cern.ch

The WLCG information system is just one of the many information sources that are required to populate a VO configuration database. Other sources include central portals such as the GOCDB and the OIM from EGI and OSG respectively. Providing a coherent view of all this information that has been synchronized from many different sources is a challenging activity and has been duplicated to various extents by each of the LHC experiments.

The WLCG Global Service Registry address these issues by aggregating information from multiple information sources and presenting a consolidated view of both pledged and available resources. It aims to help the LHC experiments populate their own VO configuration databases, used for job submission and storage management, by providing them with a single point for obtaining information on WLCG resources. In addition, in-depth validation checks are incorporated into a wider system-wide strategy to ensure the information is of the highest quality.

It is hoped that this service will decouple the LHC experiments from the underlying blocks of the WLCG information system, making it easier to evolve the system in future. This paper presents the WLCG Global Service Registry architecture, its advantages compared to the current approach and how can be used to populate a VO configuration database.

DPHEP Workshop / 530

PREDON

Corresponding Author: diaconu@cppm.in2p3.fr

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / 354

The IceProd (IceCube Production) Framework

Author: Juan Carlos Diaz Velez[1]

[1] *University of Wisconsin-Madison*

Corresponding Author: juancarlos@icecube.wisc.edu

IceProd is a data processing and management framework developed by IceCube Neutrino Observatory for processing of Monte Carlo simulations and data. IceProd runs as a separate layer on top of middleware and can take advantage of a variety of computing resources including grids and batch systems such as GLite, Condor, NorduGrid, PBS and SGE. This is accomplished by a set of dedicated daemons which process job submission in a coordinated fashion through the use of middleware plug-ins that serve to abstract the details of job submission and job management. IceProd can also manage complex workflow DAGs across distributed computing grids in order to optimize usage of resources. We describe several aspects of IceProd's design including security, data integrity, scalability, throughput as well as the various challenges in each of these topics. We also discuss design aspects of a second generation IceProd, currently being tested in IceCube.

Data Stores, Data Bases, and Storage Systems / 238

Next generation database relational solutions for ATLAS distributed computing

Author: Gancho Dimitrov[1]

Co-authors: Tadashi Maeno [2]; Vincent Garonne [1]

[1] *CERN*

[2] *Brookhaven National Laboratory (US)*

Corresponding Authors: gancho.dimitrov@cern.ch, tmaeno@bnl.gov, vincent.garonne@cern.ch

The ATLAS Distributed Computing (ADC) project delivers production tools and services for ATLAS offline activities such as data placement and data processing on the Grid. The system has been capable of sustaining with high efficiency the needed computing activities during the first run of LHC data taking, and has demonstrated flexibility in reacting promptly to new challenges. Databases are a vital part of the whole ADC system. The Oracle Relational Database Management System (RDBMS) has been addressing a majority of the ADC database requirements for many years. Much expertise was gained through the years and without a doubt will be used as a good foundation for the next generation PanDA (Production ANd Distributed Analysis) and DDM (Distributed Data Management) systems.

In this paper we present the current production ADC database solutions and notably the planned changes on the PanDA system, and the next generation ATLAS DDM system called Rucio. Significant work was performed on studying different solutions to arrive at the best relational and physical database model for performance and scalability in order to be ready for deployment and operation in 2014.

Poster presentations / **469**

DCS Data Viewer, a Application that Access ATLAS DCS historical Data

Author: Charilaos Tsarouchas[1]

Co-authors: Gancho Dimitrov [1]; Stefan Schlenker [1]

[1] *CERN*

Corresponding Authors: gancho.dimitrov@cern.ch, charilaos.tsarouchas@cern.ch, stefan.schlenker@cern.ch

The ATLAS experiment at CERN is one of the four Large Hadron Collider experiments. The DCS Data Viewer (DDV) is an application that provides access to historical data of the ATLAS Detector Control System (DCS) parameters and their corresponding alarm information. It features a server-client architecture: the pythonic server serves as interface to the Oracle-based conditions database and can be operated stand alone using http requests; the client is developed with the Google Web Toolkit (GWT) and offers a user friendly browser independent web interface. The client data visualization is done using various output plugins such as java or javascript applets which are integrated using an open JSON interface allowing for easy plugin development. The default output provides charts, histograms or tables with broad filtering and sorting capabilities. Further, export to ROOT files is supported and smartphone compatibility is taken into consideration. A server-based configuration storage facility allows e.g. for sharing of resulting plots or embedding into other web applications invoking the tool with a single configuration URL. Web security constraints along with database dedicated protection mechanisms permit a successful exposure of the tool to hundreds of collaborators worldwide.

Poster presentations / 10

Handling Worldwide LHC Computing Grid Critical Service Incidents : The infrastructure and experience behind nearly 5 years of GGUS ALARMs

Author: Maria Dimou[1]

Co-authors: Guenter Grein [2]; Helmut Dres [2]; Oleg Dulov [3]

[1] *CERN*

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

[3] *KIT*

Corresponding Authors: maria.dimou@cern.ch, guenter.grein@kit.edu, helmut.dres@kit.edu, oleg.dulov@kit.edu

In the Wordwide LHC Computing Grid (WLCG) project the Tier centres are of paramount importance for storing and accessing experiment data and for running the batch jobs necessary for experiment production activities.
Although Tier2 sites provide a significant fraction of the resources a non-availability of resources at the Tier0 or the Tier1s can seriously harm not only WLCG Operations but also the experiments' workflow and the storage of LHC data which are very expensive to reproduce.
This is why availability requirements for these sites are high and committed in the WLCG Memorandum of Understanding (MoU).
In this talk we describe the workflow of GGUS ALARMs, the only 24/7 mechanism available to LHC experiment experts for reporting to the Tier0 or the Tier1s problems with their Critical Services.
Conclusions and experience gained from the detailed drills performed in each such ALARM for the last 4 years will be explained and the shift with time of Type of Problems met.
The physical infrastructure put in place to achieve GGUS 24/7 availability will be summarised.

Poster presentations / 135

Optimization of Italian CMS Computing Centers via MIUR funded Research Projects

Author: Tommaso Boccali[1]

Co-author: Giacinto Donvito [2]

[1] *Sezione di Pisa (IT)*

[2] *INFN-Bari*

Corresponding Authors: tommaso.boccali@cern.ch, giacinto.donvito@ba.infn.it

The Italian Ministry of Research (MIUR) funded in the past years research projects aimed to an optimization of the analysis activities in the Italian CMS computing Centers. A new grant started in 2013, and activities are already ongoing in 9 INFN sites, all hosting local CMS groups. Main focus will be on the creation of an italian storage federation (via Xrootd initially, and later HTTP) which allows all the italian CMS physicists to a privileged access to CMS data and simulations. Another task will focus on the optimization of the last step of a CMS analysis, via interactive access to resources; this will result in a number of small- to medium-sized analysis centers, where access will be granted at national level to multicore machines, PROOF facilities, high throughput local queues. An important part of this last activity will imply experimenting with on demand analysis machine instantiation via Clouds, using the experience and the resources INFN is building on the subject.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / 333

Testing SLURM open source batch system for a Tier1/Tier2 HEP computing facility

Authors: Alessandro Italiano Italiano[1]; Davide Salomoni[2]; Giacinto Donvito[2]

[1] *INFN-CNAF*

[2] *Universita e INFN (IT)*

Corresponding Author: giacinto.donvito@cern.ch

In this work the testing activities that were carried on to verify if the SLURM batch system could be used as the production batch system of a typical Tier1/Tier2 HEP computing center are shown. SLURM (Simple Linux Utility for Resource Management) is an Open Source batch system developed mainly by the Lawrence Livermore National Laboratory, SchedMD, Linux NetworX, Hewlett-Packard, and Groupe Bull. Testing was focused both on verifying the functionalities of the batch system and the performance that SLURM is able to offer.

We first describe our initial set of requirements. Functionally, we started configuring SLURM so that it replicates all the scheduling policies already used in production in the computing centers involved in the test, i.e. INFN-Bari and the INFN-Tier1 at CNAF, Bologna. Currently, the INFN-Tier1 is using IBM LSF (Load Sharing Facility), while INFN-Bari, an LHC Tier2 for both CMS and Alice, is using Torque as resource manager and MAUI as scheduler.

We show how we configured SLURM in order to enable several scheduling functionalities such as Hierarchical FairShare, Quality of Service, user-based and group-based priority, limits on the number of jobs per user/group/queue, job age scheduling, job size scheduling, and scheduling of "consumable resources". We then show how different job typologies, like serial, MPI, multi-thread, whole-node and interactive jobs can be managed. Tests on the use of ACLs on queues or in general other resources are then described. A peculiar SLURM feature we also verified is triggers on event, useful to configure specific actions on each possible event in the batch system.

We also tested highly available configurations for the master node. This feature is of paramount importance since a mandatory requirement in our scenarios is to have a working farm cluster even in case of hardware failure of the server(s) hosting the batch system.

Among our requirements there is also the possibility to deal with pre-execution and post-execution scripts, and controlled handling of the failure of such scripts. This feature is heavily used, for example, at the INFN-Tier1 in order to check the health status of a worker node before execution of each job. Pre- and post-execution scripts are also important to let WNoDeS, the IaaS Cloud solution developed at INFN, use SLURM as its resource manager. WNoDeS has already been supporting the LSF and Torque batch systems for some time; in this work we show the work done so that WNoDeS supports SLURM as well.

Finally, we show several performance tests that we carried on to verify SLURM scalability and reliability, detailing scalability tests both in terms of managed nodes and of queued jobs.

Data Stores, Data Bases, and Storage Systems / 332

Testing of several distributed file-system (HadoopFS, CEPH and GlusterFS) for supporting the HEP experiments analisys.

Authors: Domenico Diacono[1]; Giacinto Donvito[2]; Giovanni Marzulli[3]

[1] *INFN-Bari*

[2] *Universita e INFN (IT)*

[3] *GARR INFN*

Corresponding Author: giacinto.donvito@cern.ch

In this work we will show the testing activity carried on several distributed file-system in order to check the capability of supporting the HEP data analysis
In particular, we focused our attention and our test on HadoopFS, CEPH, and GlusterFS.

All are Open Source software.

HadoopFS is an Apache foundation software and is part of a more general framework, that contains: task scheduler, a NOSQL database, a data warehouse system, etc. It is used by several big company and institution (Facebook, Yahoo, Linkedin, etc).

CEPH is a quite young file-system that has very good design in order to guarantee great scalability, performance and very good high availability features. It is also the unique file-system that is able to provide three interface to storage: posix file-system, REST object storage and device storage. The support for CEPH was introduced as a native in the last release of the kernel.

GlusterFS is recently acquired by RedHat and this will ensure the long term support of the code. It has indeed a large user base both in HPC computing farms, and in several Cloud computing facilities. Indeed it support access to storage both in terms of posix file-system and via a REST gateway for object storage support.

All those file-system are capable of supporting high availability of the data and metadata in order to build a distributed file-system that could provide resilience to the hardware and/or software failure of one or more data server in the cluster.
We will describe each file-system in details providing the technical specification and reporting about the testing of the most interesting functionalities of each of the softwares.

We will focus our attention on the capabilities of recover from failures of both hardware and software and on how each software is able to provide those capabilities and describing the test carried on to prove them.

We will show also performance test carried on using data analysis application that reads data in standard ROOT format in order to better compare those software from a point of view of the HEP community.

In this work we will also present the results of tests that will highlight the scalability of each of those file systems.

We will show also the development that we have done to provide more powerful monitoring capabilities for HadoopFS. We have developed a web based monitoring system that is capable to show in details the information about the status of the data nodes or the status and the historical information about the location of each block.

We will also provide detailed information on automatic procedures and script developed in order to easily manage a big datacenter composed of hundreds of data node installed with HadoopFS
In this work we will also focus on the test executed in order to exploit the GlusterFS and CEPH file-system within an IaaS Cloud Infrastructure based on OpenStack thanks to the interfaces available in those storage technologies

Poster presentations / 136

An Xrootd Italian Federation for CMS

Author: Giacinto Donvito[1]

[1] *INFN-Bari*

Corresponding Author: giacinto.donvito@ba.infn.it

The italian community in CMS has built a geographically distributed network in which all the data stored in the italian region are available to all the users for their everyday work. This activity involves

at different level all the CMS centers: the Tier1 at CNAF, all the four Tier2s (Bari, Rome, Legnaro and Pisa), and few Tier3s (Trieste, Perugia, etc). The federation uses the new network connections as provided by our NREN, GARR, which provides a minimum of 10 Gbit/s to all the sites via the GARR-X project. The federation is currently based on Xrootd technology, and on a redirector aimed to seamlessly connect all the sites, giving the logical view of a single entity. A special configuration has been put in place for the Tier1, CNAF, where ad-hoc Xrootd changes have been implemented in order to protect the tape system from excessive stress, by not allowing WAN connections to access tape only files, on a file-by-file basis. We will describe in details the test carried on the authentication and authorization capabilities in the Xrootd code, in order to achieve a better fine grained authorization criteria. For example with this authentication mechanism it is possible to implement a protection on the base of VOMS attributes and the group and roles. In order to improve the overall performance while reading files, both in terms of bandwidth and latency, it is implemented a hierarchical solution for the xrootd redirectors. The solution implemented provides a dedicated redirector where all the INFN sites are registered, without considering their status (T1, T2, or T3 sites). This redirector is the first option for the end user where they can read files both in the CMSSW framework and using bare ROOT Macros. This redirector is used also to publish information of sites that do not provide official Service Level Agreement to CMS, so that could not join the official CMS redirector. An interesting use case were able to cover via the federation are disk-less Tier3s. For sites where local manpower and/or funding does not allow the operation of a storage system, CMS analysis is still allowed by serving all the input files via WAN; the option of a local frontend cache protects the infrastructure from excessive data transfers in this case. The caching solution allows to operate a local storage with minimal human intervention: transfers are automatically done on a single file basis, and the cache if maintained operational by automatic removal of old files.

Poster presentations / 385

Evaluation of Apache Hadoop for Parallel Data Analysis with ROOT

Author: Sebastian Lehrack[1]

Co-authors: Guenter Duckeck [2]; Johannes Ebke [3]

[1] *LMU Munich*

[2] *Experimentalphysik-Fakultaet fuer Physik-Ludwig-Maximilians-Uni*

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

Corresponding Authors: sebastian.lehrack@physik.uni-muenchen.de, johannes.ebke@physik.uni-muenchen.de, guenter.duckeck@physik.uni-muenchen.de

The Apache Hadoop software is a Java based framework for distributed
processing of large data sets across clusters of computers using the
Hadoop file system (HDFS) for data storage and backup and MapReduce as a processing platform.
Hadoop is primarily designed for processing large textual data sets
which can be processed in arbitrary chunks, and must be adapted to the use case of processing binary data files which can not be split
automatically. However, Hadoop offers attractive features in terms of
fault tolerance, task supervision and controlling, multi-user
functionality and job management.

For this reason, we have evaluated Apache Hadoop as an alternative
approach to PROOF for root data analysis. Two alternatives in
distributing analysis data are discussed: Either the data is stored in HDFS and processed with MapReduce, or the data is accessed via a
standard Grid storage system (dCache Tier-2) and MapReduce was used only as execution backend.

The focus in the measurements are on the one hand to safely store
analysis data on HDFS with reasonable data rates and on the other hand to process data fast and reliably with MapReduce.

For evaluation of the HDFS, data rates for writing to and reading from local Hadoop cluster have been measured and are compared to normal data
rates on the local NFS. For evaluation of MapReduce, realistic ROOT
analyses have been used and event rates were compared to PROOF.

Poster presentations / **97**

Towards a centralized Grid Speedometer

Authors: Edgar Fajardo Hernandez[1]; Ivan Antoniev Dzhunov[2]; Julia Andreeva[3]; Oliver Gutsche[4]; Pablo Saiz[3]; Sten Luyckx[5]

[1] *Universidad de los Andes (CO)*

[2] *University of Sofia*

[3] *CERN*

[4] *FERMILAB*

[5] *University of Antwerp (BE)*

Corresponding Author: ivan.antoniev.dzhunov@cern.ch

Given the distributed nature of the grid and the way CPU resources are pledged and scared around the globe, VO's are facing the challenge to monitor the use of these resources. For CMS and the operation of centralized workflows the monitoring of how many production jobs are running and pending in the Glidein WMS production pools is very important. The Dashboard SSB (Site Status Board) provides a very flexible framework to collect, aggregate and visualize data. The CMS production monitoring team uses the SSB to define the metrics that have to be monitored and the alarms that have to be set. During the integration of the CMS production monitoring into the SSB, several enhancements to the core functionality of the SSB were implemented; all in a generic way, so that other VOs using the SSB can use them as well. Alongside these enhancements, there were a few changes to the core of the SSB framework from which the CMS production team was able to benefit. We will present the details of the implementation and the advantages for current and future usage of the new features in the VO agnostics Dashboard.

Data Stores, Data Bases, and Storage Systems / **225**

The Drillbit column store

Authors: Johannes Ebke[1]; Peter Waller[2]

[1] *TNG Technology Consulting*

[2] *University of Liverpool (GB)*

Corresponding Author: johannes@ebke.org

In comparison to storing data packed by event, column data stores store event variables or sets of event variables in individual data packs. One well-known example is the CERN ROOT library's TTree, which has a mode where it behaves like a column store. Columnar data stores can offer fast processing of a subset of the event structure or individual variables.

In the experimental Drillbit column store we explore the encoding of Google protocol buffer data structures into columns, using a method used in the internal Google Dremel architecture. In addition, Drillbit aims to provide a robust mechanism to synchronize event variables stored in different files, providing a guarantee to the analyst that the event or partial event has been reassembled correctly. By using blockwise unique identifiers and enforcing event ordering in blocks of events, the performance problems usually associated with database joins are avoided. For reduced analysis

datasets, the Drillbit data structure allows efficient removal of events, object variables or subsets of objects, even while keeping the full alignment and compatibility with non-reduced datasets at all levels.

Preliminary studies on real-life ROOT analysis datasets have yielded exciting results, indicating a possible gain of about a quarter in storage space while using the same compression algorithm and settings. In addition, an experimental analysis library which is compatible with a subset of the TTree API showed performance on par with or exceeding the ROOT TTree.

Finally, in connection with an in-development dynamic event model, Drillbit could make it practical to do more cache-efficient computations on small numbers of variables, as well as providing several opportunities to use multiple cores. For analysts, Drillbit could allow fast and reliable retrieval of only the relevant analysis variables, and a simple way to share new data corrections and analysis objects.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / 292

Challenges of the ATLAS Monte Carlo production during run 1 and beyond

Author: Wolfgang Ehrenfeld[1]

Co-authors: Claire Gwenlan [2]; Jiahang Zhong [2]; John Derek Chapman [3]; Jose Enrique Garcia Navarro [4]; Sascha Mehlhase [5]

[1] *Universitaet Bonn (DE)*

[2] *University of Oxford (GB)*

[3] *University of Cambridge (GB)*

[4] *Universidad de Valencia (ES)*

[5] *University of Copenhagen (DK)*

Corresponding Authors: wolfgang.ehrenfeld@cern.ch, chapman@hep.phy.cam.ac.uk, jose.enrique.garcia@cern.ch, c.gwenlan1@physics.ox.ac.uk, sascha.mehlhase@cern.ch, jiahang.zhong@cern.ch

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 1 and Long Shutdown 1 will be presented, including 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 various specialised activities. These include 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.

Software Engineering, Parallelism & Multi-Core / 303

Explorations of the viability of ARM and Intel Xeon Phi for Physics Processing

Authors: David Abdurachmanov[1]; Gene Cooperman[2]; Giulio Eulisse[3]; Kapil Arya[4]; Peter Elmer[5]; Shahzad Malik Muzaffar[3]

[1] *Vilnius University (LT)*

[2] *Unknown*

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

[4] *Northeastern University*

[5] *Princeton University (US)*

Corresponding Author: peter.elmer@cern.ch

In the last decade power limitations led to the introduction of multicore CPU's. The cores on the processors were however not dramatically different from the processors just before the multicore-era. In some sense, this was merely a tactical choice to maximize compatibility and buy time. The same scaling problems that led to the power limit are likely to push processors in the direction of ever greater numbers of simpler, less performant lower-power cores. Without this architectural change, it is doubtful if the gains expected from a Moore's Law extrapolation will continue to be realized over the next five years.

Although it is very hard to predict where the market will wind up in the long run, we already see today a couple of concrete product examples which give indications as to the kinds of things that we will see going forward, namely Intel's Many Integrated Core (MIC) architecture and the ARM processor. The first MIC commercial products (Xeon Phi) are in the form of a coprocessor and aimed at the HPC market. 32bit ARM is ubiquitous today in mobile electronics and the 64bit version (ARMv8) is expected to make a debut in the server space this year.

In this presentation we report on our first investigations into the viability of the ARM processor and Intel Xeon Phi processors. We use benchmarks with real physics applications and performance profiles to explore the viability of these processors for production physics processing. We investigate what changes would be necessary for complete, large applications to perform efficiently and in a scalable way on these kind of processors.

Software Engineering, Parallelism & Multi-Core / 41

A taxonomy of scientific software applications - HEP's place in the world

Author: Peter Elmer[1]

[1] *Princeton University (US)*

Corresponding Author: peter.elmer@cern.ch

Modern HEP software stacks, such as those used by the LHC experiments at CERN, involve many millions of lines of custom code per experiment, as well as a number of similarly sized shared packages (ROOT, Geant4, etc.) Thousands of people have made contributions over time to these code bases, including graduate students, postdocs, professional researchers and software/computing professionals. Elaborate software integration, testing and validation systems are used to manage the resulting workflow.

HEP has also been a poster child for "Big Data" science with more than 100 PetaBytes of event data stored around the world. Its

applications however typically need event data of order 1MB in memory at a given time plus some tens of MB of calibration data. The resulting data processing, as well as its Monte Carlo simulations, is “embarrassing parallel”. The hardware needed for this type of “High Throughput” computing is relatively unspecialized, in the category of “low-end” CPU servers, although great numbers are needed. Most of the code executed is arguably non-numerical, with floating point operations corresponding to only a small fraction of the total time.

These are particular points in a large phase space of scientific applications. Other characteristics like MPI-style parallelism, small code bases, the need for “High Performance” computing (GPU’s, specialized interconnects or large memory needs) are not uncommon. Many are true numerical codes and the data requirements vary greatly. This presentation will cover the results of investigations of the software stacks and applications of a variety of other scientific fields. Where are the commonalities and with whom? Or do we inhabit a small niche in the world of scientific computing? Particular attention will be placed on the characteristics and needs of other scientific projects which will require similar or greater amounts of resources than HEP in the next ten years.

Poster presentations / **266**

Grid Site Testing for ATLAS with HammerCloud

Author: Johannes Elmsheuser[1]

Co-authors: Daniel van der Ster [2]; Federica Legger [1]; Francesco Giovanni Sciacca [3]; Ramon Medrano Llamas [2]

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

[2] *CERN*

[3] *Universitaet Bern (CH)*

Corresponding Authors: federica.legger@physik.uni-muenchen.de, ramon.medrano.llamas@cern.ch, daniel.vanderster@cern.ch, gsciacca@mail.cern.ch

With the exponential growth of LHC (Large Hadron Collider) data in 2012, distributed computing has become the established way to analyze collider data. The ATLAS grid infrastructure includes more than 130 sites worldwide, ranging from large national computing centers to smaller university clusters. HammerCloud was previously introduced with the goals of enabling VO- and site-administrators to run validation tests of the site and software infrastructure in an automated or on-demand manner. The HammerCloud infrastructure has been constantly improved to support the addition of new test workflows. These new workflows comprise e.g. tests of the ATLAS nightly build system, ATLAS MC production system, XRootD federation FAX and new site stress test workflows. We report on the development, optimization and results of the various components in the HammerCloud framework.

Poster presentations / **96**

di-EOS - ”distributed EOS”: Initial experience with split-site persistency in a production service

Authors: Andreas Joachim Peters[1]; Jan Iven[1]; Luca Mascetti[1]; Xavier Espinal Curull[1]

[1] *CERN*

Corresponding Authors: xavier.espinal@cern.ch, andreas.joachim.peters@cern.ch, luca.mascetti@cern.ch, jan.iven@cern.ch

After the strategic decision in 2011 to separate tier-0 activity from analysis, CERN-IT developed EOS as a new petascale disk-only solution to address the fast-growing needs for high-performance low-latency data access. EOS currently holds around 22PB usable space for the four big experiment (ALICE, ATLAS, CMS, LHCb), and we expect to grow to >30PB this year. EOS is one of the first production services to be running in CERN's new facility located in Budapest: we foresee to have about a third of total EOS storage capacity in the new facility in 2013, making it the largest storage service in the new CERN computer centre.

We report on the initial experience running EOS as a distributed service (via the new CERN IT Agile Infrastructure tools and a new "remote-hands" contract) in a production environment, as well as on the particular challenges and solutions. Among these solutions we are investigating the stochastic geo-location of data replicas and countermeasures against "split-brain" scenarios for the new high-availability namespace. In addition we are considering optimized clients and draining procedures to avoid link overloads across the two CERN sites as well as maximising data-access and operations efficiency.

Poster presentations / 148

Streamlining CASTOR to manage the LHC data torrent

Author: Giuseppe Lo Presti[1]

Co-authors: Andrea Ieri ; Benjamin Fiorini [2]; Elvin Alin Sindrilaru [1]; Eric Cano [1]; Sebastien Ponce [1]; Steven Murray [1]; Xavier Espinal Curull [1]

[1] *CERN*

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

Corresponding Authors: xavier.espinal@cern.ch, giuseppe.lopresti@cern.ch

This contribution describes the evolution of the main CERN storage system, CASTOR, as it manages the bulk data stream of the LHC and other CERN experiments, achieving nearly 100 PB of stored data by the end of LHC Run 1.

Over the course of 2012 the CASTOR service has addressed the Tier-0 data management requirements, focusing on a tape-backed archive solution, ensuring smooth operations of the required experiments' workflow (data taking, reconstruction, export to other sites), and guaranteeing data safety by enforcing strong authentication. This evolution was marked by the introduction of policies to optimize the tape sub-system throughput, going towards a cold storage system where data placement is managed by the experiments' production managers. More efficient tape migrations and recalls have been implemented and deployed where bulk metadata operations greatly reduce the overhead due to small files. A repack facility is now integrated in the system and it has been enhanced in order to automate the repacking of several tens of petabytes, required in 2014 in order to prepare for the next LHC run. Finally the scheduling system has been evolved to integrate the internal monitoring for a more efficient dynamic usage of the underlying disk resources.

To efficiently manage the service a solid monitoring infrastructure is required, able to analyze the logs produced by the different components (~1 kHz of log messages). A new system has been developed and deployed, which uses a transport messaging layer provided by the CERN-IT Agile Infrastructure and exploits technologies including Hadoop and HBase. This enables efficient data mining by making use of MapReduce techniques, and real-time data aggregation and visualization.

We will also present the outlook for the future. A further simplification of the CASTOR system is foreseen: xrootd will take the role of the main native protocol, and only few disk pools for each experiment would serve as staging areas for the Tier-0 workflow. The adoption of xrootd as main protocol opens the possibility to exploit its upcoming features, including for instance the support of xroot federations or the HTTP protocol. Directions and possible evolutions will be discussed in view of the restart of data taking activities.

Data Stores, Data Bases, and Storage Systems / 83

Disk storage at CERN: handling LHC data and beyond

Authors: Belinda Chan[1]; Jan Iven[1]; Luca Mascetti[1]; Massimo Lamanna[1]; Xavier Espinal Curull[1]

[1] *CERN*

Corresponding Author: xavier.espinal@cern.ch

Data Storage and Services (DSS) group at CERN stores and provides access to the data coming from the LHC and other physics experiments. We implement specialized storage services to provide tools for an optimal data management, based on the evolution of data volumes, the available technologies and the observed experiment and users usage patterns. Our current solutions are CASTOR for highly-reliable tape-backed storage for heavy-duty Tier-0 workflows and EOS for disk-only storage for full-scale analysis activities.
CASTOR has been the main physics storage system at CERN since 2001 and successfully caters for the LHC experiments' needs, storing 90 PB of data and more than 350 M files. During the last LHC run CASTOR was routinely storing 1 PB/week of data to tape. CASTOR is currently evolving towards a simplified disk layer in front of the tape robotics, focusing on recording the primary data from the detectors.
EOS is now a well established storage service used intensively by the four big LHC experiments, holding over 15 PB of data and more than 130M files (30 PB usable disk space expected at the end of the year). Its conceptual design based on multi-replica and in-memory namespace make it the perfect system for data intensive workflows and its usage will expand via a shared instance for non-LHC experiments. In the short term EOS usage will absorb most of the newly installed capacity at CERN and expand via a shared instance for the non-LHC experiments. An additional challenge will be to run this service across two geographically different sites (CERN Geneva and Budapest).
LHC-Long Shutdown 1 presents a window of opportunity to shape up both of our storage services and validate against the ongoing analysis activity in order to successfully face the new LHC data taking period in 2015. Besides summarizing the current state and foreseen evolutions, the talk will focus on the detailed analysis of the operational experience of both systems, in particular service efficiency, performance and reliability.

Software Engineering, Parallelism & Multi-Core / 223

The Rise of the Build Infrastructure

Author: Giulio Eulisse[1]

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

Corresponding Author: giulio.eulisse@cern.ch

CMS Offline Software, CMSSW, is an extremely large software project, with roughly 3 millions lines of code, two hundreds of active developers and two to three active development branches. Given the scale of the problem, both from a technical and a human point of view, being able to keep on track such a large project, bug free, and to deliver builds for different architectures is a challenge in itself. Moreover the challenges posed by the future migration of CMSSW to multithreading also require adapting and improving our QA tools. We present the work done in the last two years in our build and integration infrastructure, particularly in the form of improvements to our build tools, in the simplification and extensibility of our build infrastructure and the new features added to our QA and profiling tools. Finally we present our plans for the future directions for code management and how this reflects on our workflows and the underlying software infrastructure.

Poster presentations / 214

Continuous service improvement

Author: Maite Barroso Lopez[1]

Co-authors: Helge Meinhard [1]; Juan Manuel Guijarro [1]; Line Everaerts [1]; Nils Hoimyr [1]; Pierre Baehler [1]

[1] *CERN*

Corresponding Authors: line.gunther@cern.ch, maria.barroso.lopez@cern.ch

Using the framework of ITIL best practises, the service managers within CERN-IT have engaged into a continuous improvement process, mainly focusing on service operation. This implies an explicit effort to understand and improve all service management aspects in order to increase efficiency and effectiveness. We will present the requirements, how they were addressed and share our experiences. We will describe how we measure, report and use the data to continually improve both the processes and the services being provided. The focus is not the tool or the process, but the results of the continuous improvement effort from a large team of IT experts providing services to thousands of users, supported by the tool and its local team.

Poster presentations / **90**

Testnodes –a Lightweight Node-Testing Infrastructure

Author: Robert Fay[1]

Co-author: John Bland [1]

[1] *University of Liverpool*

Corresponding Author: fay@hep.ph.liv.ac.uk

A key aspect of ensuring optimum cluster reliability and productivity lies in keeping worker nodes in a healthy state. Testnodes is a lightweight node testing solution developed at Liverpool. While Nagios has been used locally for general monitoring of hosts and services, Testnodes is optimised to answer one question: is there any reason this node should not be accepting jobs? This tight focus enables Testnodes to inspect nodes frequently with minimal impact and provide a comprehensive and easily extended check with each inspection.

On the server side, Testnodes, implemented in python, interoperates with the Torque batch server to control the nodes production status. Testnodes remotely and in parallel executes client-side test scripts and processes the return codes and output, adjusting the node's online/offline status accordingly to preserve the integrity of the overall batch system. Testnodes reports via log, email and Nagios, allowing a quick overview of node status to be reviewed and specific node issues to be identified and resolved quickly.

This presentation will cover testnodes design and implementation, together with the results of its use in production at Liverpool, and future development plans.

Poster presentations / **89**

Tier-2 Optimisation for Computational Density/Diversity and Big Data

Authors: John Bland[1]; Robert Fay[1]

Co-authors: Mark Norman [1]; Stephen Jones [2]

[1] *University of Liverpool*

[2] *Liverpool University*

Corresponding Author: fay@hep.ph.liv.ac.uk

As the number of cores on chip continues to trend upwards and new CPU architectures emerge, increasing CPU density and diversity presents multiple challenges to site administrators.

These include scheduling for massively multi-core systems (potentially including GPU (integrated and dedicated) and many integrated core (MIC)) to ensure a balanced throughput of jobs while preserving overall cluster throughput, in addition to meeting data demands as both dataset sizes increase and as the rate of demand scales with increased computational power, along with the practical management of these resources.

In this report, we evaluate the current tools and technologies available to manage these emerging requirements, including cluster software (batch, scheduling), resource management solutions (VMs, clouds, containers) and infrastructure (hardware and network specification and optimisation, software services) in order to assess what options are available at the present time, the limits therein, and to identify issues remaining to be addressed.

Poster presentations / **210**

Dirac integration with a general purpose bookkeeping DB: a complete general suite

Authors: Armando Fella[1]; Bruno Santeramo[2]; Cristian De Santis[3]; Giacinto Donvito[4]; Marcin Jakub Chrzaszcz[5]; Milosz Zdybal[6]; Rafal Zbigniew Grzymkowski[7]

Co-authors: Alberto Gianoli [8]; Alessio Gianelle [9]; Andrea Di Simone [10]; Domenico Del Prete [11]; Eleonora Luppi [8]; Fabrizio Bianchi ; Francesco Giacomini [12]; Guido Russo [9]; Luca Tomassetti [13]; Luis Alejandro Perez Perez [14]; Marco Corvo [15]; Matteo Manzali [12]; Matteo Rama ; Paolo Franchini [9]; Roberto Stroili [16]; Silvio Pardi [15]; Stefano Longo [12]; Steffen Luitz [17]; Vincenzo Ciaschini [18]

[1] *INFN Pisa*

[2] *INFN Bari*

[3] *Universita degli Studi di Roma Tor Vergata (IT)*

[4] *INFN-Bari*

[5] *Polish Academy of Sciences (PL)*

[6] *Institute of Nuclear Physics, Polish Academy of Science*

[7] *P*

[8] *Universita di Ferrara (IT)*

[9] *Universita e INFN (IT)*

[10] *Universita e INFN Roma Tor Vergata (IT)*

[11] *I.N.F.N.*

[12] *INFN CNAF*

[13] *University of Ferrara and INFN*

[14] *INFN Sezione di Pisa*

[15] *INFN*

[16] *Università degli Studi di Padova & INFN*

[17] *SLAC*

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

Corresponding Authors: marcin.jakub.chrzaszcz@cern.ch, cristian.de.santis@cern.ch, giacinto.donvito@ba.infn.it, armando.fella@pi.infn.it, bruno.santeramo@ba.infn.it, milosz.zdybal@ifj.edu.pl

In HEP computing context, R&D studies aiming to the definition of the data and workload models were brought forward by the SuperB community beyond the experiment life itself.
This work is considered of great interest for a generic mid- and small size VO to fulfil Grid exploiting requirements involving CPU-intensive tasks.
We present the R&D line achievements in the design, developments and
test of a distributed resource exploitation suite based on DIRAC. The
main components of such a suite are the information system, the job
wrapper and the new generation DIRAC framework. The DB schema and the SQL logic have been designed to be able to be adaptive with respect to the VO requirements in terms of physics application, job environment and bookkeeping parameters. A deep and flexible integration with DIRAC features has been obtained using SQLAlchemy technology allowing mapping and interaction with the information system. A new DIRAC extension has been developed to include this functionality along with a new set of DIRAC portal interfaces aimed to the job, distributed resources, and metadata management. The results of the first
functionality and efficiency tests will be reported.

Poster presentations / **271**

R&D work for a data model definition: data access and storage system studies

Authors: Armando Fella[1]; Domenico Diacono[2]; Giacinto Donvito[3]; Giovanni Marzulli[4]; Paolo Franchini[5]; Silvio Pardi[6]

Co-authors: Alberto Gianoli [7]; Alessio Gianelle [8]; Andrea Di Simone [9]; Bruno Santeramo ; Cristian De Santis [10]; Domenico Del Prete [11]; Eleonora Luppi [7]; Elisa Manoni [12]; Fabrizio Bianchi ; Francesco Giacomini [13]; Guido Russo [5]; Luca Tomassetti [14]; Luis Alejandro Perez Perez [15]; Marcin Jakub Chrzaszcz [16]; Marco Corvo [6]; Matteo Manzali [17]; Matteo Rama ; Milosz ZDYBAL [18]; Rafal Zbigniew Grzymkowski [19]; Roberto Stroili [20]; Stefano Longo [13]; Steffen Luitz [21]; Vincenzo Ciaschini [17]

[1] *INFN Pisa*

[2] *INFN Bari*

[3] *INFN-Bari*

[4] *GARR*

[5] *Universita e INFN (IT)*

[6] *INFN*

[7] *Universita di Ferrara (IT)*

[8] *INFN Padova*

[9] *Universita e INFN Roma Tor Vergata (IT)*

[10] *Universita degli Studi di Roma Tor Vergata (IT)*

[11] *I.N.F.N.*

[12] *INFN Perugia*

[13] *INFN CNAF*

[14] *University of Ferrara and INFN*

[15] *INFN Sezione di Pisa*

[16] *Polish Academy of Sciences (PL)*

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

[18] *Institute of Nuclear Physics, Polish Academy of Science*

[19] *P*

[20] *Università degli Studi di Padova & INFN*

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

Corresponding Authors: domenico.diacono@ba.infn.it, giacinto.donvito@ba.infn.it, armando.fella@pi.infn.it, paolo.franchini@cnaf.
giovanni.marzulli@ba.infn.it, spardi@na.infn.it

In HEP computing context, R&D studies aiming to the definition of the data and workload models were brought forward by the SuperB community beyond the experiment life itself. This work is considered of great interest for a generic mid- and small size VO during its Computing Model definition phase.
Data-model R&D work we are presenting, starts with the general design description of the crucial components in terms of typical HEP use cases; a discussion on strategies and motivations for the taken choices in the fields of data access, mass data transfer and meta-data catalog system is provided firstly. In such a context we focused the evaluation, test and development work on storage systems enabled for geographically-distributed data management: data access, data replication, data recovery and backup in WAN scenarios. HadoopFS and GlusterFS distributed file-system have been mainly considered in this analysis.
Data availability in a distributed environment is a key point in the definition of the computing model for an HEP experiment. Among all the possible interesting data models, we identify the WAN direct access via reliable and robust protocols such as HTTP/WebDAV and xrootd as a viable option. The development of a dedicated library has been carried on allowing an optimized file access procedure on remote storage resources. The implemented features include read-ahead and data prefetching techniques, caching mechanism and optimized target file localization. The results of performance and efficiency tests will be presented for the treated subjects trying to describe in conclusion the general strategy lines and technologies for the drafting of a concrete data model design report.

Poster presentations / **45**

Cloud flexibility using Dirac Interware

Authors: Marcos Seco Miguelez[1]; Victor Manuel Fernandez Albor[1]

Co-authors: Juan Jose Saborido Silva [1]; Ricardo Graciani Diaz [2]; Tomas Fernandez [1]; Victor Mendez Muñoz [3]

[1] *Universidade de Santiago de Compostela (ES)*

[2] *University of Barcelona (ES)*

[3] *Port d'Informació Científica (PIC), Universitat Autònoma de Barcelona, ES-08193 Bellaterra (Barcelona) Spain*

Corresponding Author: victormanuel.fernandez@usc.es

Communities of different locations are running their computing jobs on dedicated infrastructures without the need to worry about software, hardware or even the site where their programs are going to be executed. Nevertheless, this usually implies that they are restricted to use certain types or versions of an Operating System because either their software needs an definite version of a system library or a specific platform is required by the collaboration to which they belong. On this scenario, if a data center wants to service software incompatible communities, it has to split its physical resources among those communities. This splitting will inevitably lead to an underuse of resources because the data centers are bound to have periods where one or more of its subclusters are idle.

It is in this situation where Cloud Computing provides the flexibility and reduction in computational cost that data centers are searching for. This paper describes a set of realistic tests that we ran on one of such implementations. The test comprise software from three different HEP communities (Auger, LHCb and QCD phenomelogists) and the Parsec Benchmark Suite running on one or more of three Linux flavors (SL5, Ubuntu 10.04 and Fedora 13). The implemented infrastructure has, at the cloud level, CloudStack that manages the virtual machines (VM) and the hosts on which they run, and, at

the user level, the DIRAC framework along with a VM extension that will submit, monitorize and keep track of the user jobs and also requests CloudStack to start or stop the necessary VM's. In this infrastructure, the community software is distributed via the CernVM-FS, which has been proven to be a reliable and scalable software distribution system. With the resulting infrastructure, users are allowed to send their jobs transparently to the Data Center.

The main purpose of this system is the creation of flexible cluster,
multiplatform with an scalable method for software distribution for several VOs. Users from different communities do not need to care about the installation of the standard software that is available at the nodes, nor the operating system of the host machine, which is transparent to the user.

Data Stores, Data Bases, and Storage Systems / **32**

Forming an ad-hoc nearby storage framework, based on the IKAROS architecture and social networking services

Author: Christos Filippidis[1]

Co-authors: Christos Markou [1]; Yiannis Cotronis [2]

[1] *Nat. Cent. for Sci. Res. Demokritos (GR)*

[2] *University of Athens*

Corresponding Author: christos.filippidis@cern.ch

Given the current state of I/O and storage systems in petascale systems, incremental solutions in most aspects are unlikely to provide the required capabilities in exascale systems. Traditionally I/O has been considered as a separate activity that is performed before or after the main simulation or analysis computation, or periodically for activities such as check-pointing, but still as separate overhead. I/O architectures, when designed in this way, have already shown not to be scalable as needed. At the same time, Grid computing implementations are not considered to be user friendly and the Virtual Organization (VO) approach is not very efficient for individual users and small groups.
We present an ad-hoc "nearby"storage framework, based on the IKAROS architecture and social networking services (such as Facebook), in order to address the above problems. IKAROS is, by design, able to add or remove nodes to or from the I/O system instance on the fly, without bringing everything down or losing data. Furthermore, it is capable to decide the file partition distribution schema, by taking into account user and application requests, as well as domain and VO policies. We are using existing social networking services, such as Facebook, in order to dynamically manage, share and publish meta-data. In this way, we do not need to build our own utilities for searching, sharing and publishing and at the same time we enable users to dynamically use the infrastructure, by creating ad-hoc storage formations. This results in a model which can scale both up and down and so can provide more cost effective infrastructures for both large scale and smaller size groups and consortiums. This approach gives us the opportunity to build a more efficient ad-hoc nearby storage: multiple instances of smaller capacity, higher bandwidth storage closer to the compute nodes.
To achieve this, IKAROS is using the JSON (JavaScript Object Notation) format for populating meta-data, which is very common within Web 2.0 technologies. Individual users or groups can dynamically change their meta-data management schema based on their needs. We use Facebook as a use case to show how easily IKAROS can connect with existing, successful infrastructures which already scales to millions of users. Although we are currently using Facebook, it should be emphasized that we are not constrained by this specific choice. IKAROS is responsible only for the core services and Facebook for the meta-data utilities. In this way, the responsibilities are kept separated and the two infrastructures can scale independently from each other.

Poster presentations / **381**

Featured "Single Sign-In" interface enabling Grid, Cloud and local resources for HEP

Authors: Gunter Quast[1]; Marian Zvada[1]; Max Fischer[1]

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

Corresponding Authors: max.fischer@cern.ch, marian.zvada@cern.ch

The CMS collaboration is successfully using glideInWMS for managing grid resources within the WLCG project. The GlideIn mechanism with HTCondor underneath provides a clear separation of responsibilities between administrators operating the service and users utilizing computational resources.
German CMS collaborators (dCMS) have explored modern capabilities of the glideInWMS and aiming at merging national grid resources, institutional CPU power and cloud resources into the set of pools with common sign-in interface presented towards HEP analysis users. The key goals of service development include ease of use, uniform access, load balancing and automated selection among different resource technologies. The approach shares experience of dCMS during the development and integration phases, and production operations and highly encourages other countries to follow. First experience with the production system and an outlook towards ongoing development will be presented.

Poster presentations / 138

Evaluating Tier-1 Sized Online Storage Solutions

Authors: Catalin Lucian Dumitrescu[1]; Ian Fisk[1]

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

Corresponding Authors: ian.fisk@cern.ch, catalin.lucian.dumitrescu@cern.ch

The Fermilab CMS Tier-1 facility provides processing, networking, and storage as one of seven Tier-1 facilities for the CMS experiment. The storage consists of approximately 15 PB of online/nearline disk managed by the dCache file system, and 22 PB of tape managed by the Enstore mass storage system. Data is transferred to and from computing centers worldwide using the CMS-developed PhEDEx transfer management system.

Prior to 2013, control over which files were staged on all Tier-1 nearline storage was provided on a site-by-site basis, despite the fact that the decisions were made centrally by the CMS Computing Operations team. We were required to change this model by mapping two separate PhEDEx transfer endpoints to our storage, allowing CMS to use this tool to manage Tier-1 nearline storage worldwide. In this paper, we evaluate various storage management solutions, most of which involve migrating the bulk of the disk (˜13 PB) to another system. Hadoop, Lustre, EOS, dCache 2.x were the systems under consideration. We will present the results from evaluating the performance of metadata services, storage services, and support and operations models.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / 109

Opportunistic Computing only knocks once: Processing at SDSC

Author: Ian Fisk[1]

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

Corresponding Authors: linacre@fnal.gov, ian.fisk@cern.ch

During Spring 2013, CMS processed 1 Billion RAW data events at the San Diego Super Computing Center (SDSC) that was nearly the size of half the CMS dedicated Tier-1 processing resources. This

facility has none of the permanent CMS services, service level agreements, or support normally associated with a Tier-1, and was assembled with a few weeks notice to process only a few workflows. The size of the facility was sufficiently large that even a few weeks of time had a major impact on the reprocessing schedule for CMS. We describe our experience with creating a dynamic processing center on the fly from service deployment, to integration, to operations.

Poster presentations / **137**

Event processing time prediction at the CMS Experiment of the Large Hadron Collider

Author: Samir Cury Siqueira[1]

Co-authors: Dorian Kcira [1]; Oliver Gutsche [2]

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

[2] *FERMILAB*

Corresponding Authors: ian.fisk@cern.ch, samir.cury.siqueira@cern.ch, oliver.gutsche@cern.ch

The physics event reconstruction in LHC/CMS is one of the biggest challenges for computing.

Among the different tasks that computing systems perform, the reconstruction takes most of the CPU resources that are available. The reconstruction time of a single event varies according to the event complexity. Measurements were done in order to find precisely this correlation, creating means to predict it based on the physics conditions of the input data.

Currently the data processing system do not account that when splitting a task in chunks(jobs), this can cause a considerable variation in the job length, thus a considerable increase into the workflow Estimated Time of Arrival.

The goal is to use this estimate on processing time to more efficiently split the work in chunks, considering the CPU time needed for each chunk and due to this, lowering the standard deviation of the job length distribution in a workflow.

Poster presentations / **112**

Evolution of the pilot infrastructure of CMS: towards a single glideinWMS pool

Authors: Ian Fisk[1]; Oliver Gutsche[1]

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

Corresponding Authors: ian.fisk@cern.ch, oliver.gutsche@cern.ch

CMS production and analysis job submission is based largely on glideinWMS and pilot submissions. The transition from multiple different submission solutions like gLite WMS and HTCondor-based implementations was carried out over years and is coming now to a conclusion. The historically explained separate glideinWMS pools for different types of production jobs and analysis jobs are being unified into a single global pool. This enables CMS to benefit from global prioritization and scheduling possibilities. It also presents the sites with only one kind of pilots and eliminates the need of having to make scheduling decisions on the CE level. This paper provides an analysis of the benefits of a unified resource pool, as well as a description of the resulting global policy. It will explain the technical challenges moving forward and present solutions to some of them.

Poster presentations / **237**

A J2EE based server for Muon Spectrometer Alignment monitoring in the ATLAS detector

Author: Andrea Formica[1]

Co-authors: Florian Bauer [1]; Pierre-Francois Giraud [1]

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

Corresponding Authors: andrea.formica@cern.ch, pierre-francois.giraud@cern.ch, florian.bauer@cern.ch

The ATLAS muon alignment system is composed of about 6000 optical sensors for the Barrel muon spectrometer and the same number for the 2 Endcaps wheels.

The system is acquiring data from every sensor continuously , with a whole read-out cycle of about 10 minutes. The read-out chain stores data inside an Oracle DB. These data are used as input from the alignment algorithms (C++ based) in order to compute alignment corrections that are then used by ATLAS muon reconstruction software.

An application deployed inside a J2EE server takes care of interactions between the DB and the alignment algorithms, delivering also functions for monitoring tools, handling scheduled tasks to launch alignment reconstruction software and checking and validating results before their migration to the official ATLAS Condition DB (COOL). The same application allows access to COOL database information and to another database containing Conditions and Configurations Metadata for ATLAS (COMA), giving thus the possibility to follow the full chain of the data flow of the Muon Alignment system.

We will describe the architecture of the J2EE application and the monitoring tools that have been developed.

Poster presentations / **287**

A tool for Conditions Tag Management in ATLAS

Author: Alexander Sharmazanashvili[1]

Co-authors: Andrea Formica [2]; Giorgi Batiashvili [1]; Giorgi Gvaberidze [3]

[1] *Computer Aided Design Center (GE)*

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

[3] *E. Andronikashvili Inst. of Phys.-Georgian Academy of Sciences*

Corresponding Authors: andrea.formica@cern.ch, lasha.sharmazanashvili@cern.ch, giorgi.batiashvili@cern.ch, gvaberidze@cadcam.ge

ATLAS Conditions data include about 2 TB in a relational database and 400 GB of files referenced from the database. Conditions data is entered and retrieved using COOL, the API for accessing data in the LCG Conditions Database infrastructure. It is managed using an ATLAS-customized python based tool set.

Conditions data are required for every reconstruction and simulation job, so access to them is crucial for all aspects of ATLAS data taking and analysis, as well as by preceding tasks to derive optimal corrections to reconstruction. Ensuring that the optimal alignment and calibration information is used in the reconstruction is a complex task: variations can occur even within a run and independently in time from one subsystem to the other. Optimized sets of conditions for processing are accomplished using strict version control on those conditions: a process which assigns COOL Tags to sets of conditions, and then unifies those conditions over data-taking intervals into a COOL Global Tag. This

Global Tag identifies the set of conditions used to process data so that the underlying conditions can be uniquely identified with 100% reproducibility should the processing be executed again.

Understanding shifts in the underlying conditions from one tag to another and ensuring interval completeness for all detectors for a set of runs to be processed is a complex task, requiring tools beyond the above mentioned python utilities. Therefore, a Java/php based utility called the Conditions Tag Browser (CTB) has been developed. CTB gives detector and conditions experts the possibility to navigate through the different databases and COOL folders; explore the content of given tags and the differences between them, as well as their extent in time; visualize the content of channels associated with leaf tags. This report describes the structure and implementation of the CTB, demonstrates its use during LHC Run 1, and describes plans for expanding its functionality in preparation for LHC Run 2.

Plenaries / **486**

Data archiving and data stewardship

Author: Pirjo-Leena Forsström[1]

[1] *CSC*

Developments in data preservation and data life cycle management are having great impact the computing and storage landscape. In this talk dr Pirjo-Leena Forsström of CSC (Helsinki) will describe trends and future developments in data services for science, humanities and culture, the way these developments are being addressed by at CSC, and how this could apply to physics data.

Event Processing, Simulation and Analysis / **36**

DD4hep: A General Purpose Detector Description Toolkit

Author: Markus Frank[1]

Co-authors: Frank Gaede [2]; Pere Mato Vila [1]

[1] *CERN*

[2] *DESY IT*

Corresponding Author: markus.frank@cern.ch

The geometry, and in general, the detector description is an essential component for the development of the data processing applications in high-energy physics experiments. We will present a generic detector description toolkit, describing the guiding requirements and the architectural design for the main components of the toolkit, as well as the main implementation choices. The design is strongly driven by easy use of the toolkit: developers of detector descriptions and applications using them should provide minimal information and minimal specific code to achieve the desired result. The toolkit has been built reusing already existing components such as the ROOT geometry package, the GDML interchange format and corresponding converters. The toolkit provides missing functional elements and interfaces to offer a complete and coherent detector description solution suitable for the simulation of particle collisions in a detector, the reconstruction and the physics analysis. A natural integration to Geant4, the detector simulation program used in high-energy physics is provided.

Data Acquisition, Trigger and Controls / **35**

Deferred High Level Trigger in LHCb: A Boost to CPU Resource Utilization

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 at least 1 TB of local storage space.
This work describes the architecture to treble the availible CPU power of the HLT farm given that the LHC collider in previous years delivered stable physics beams about 30 % of the time. The gain is achieved by splitting the event selection process in two, a first stage reducing the data taken during stable beams and buffering the preselected particle collisions locally. A second processing stage running constantly at lower priority will then finalize the event filtering process and benefits fully from the time when LHC does not deliver stable beams e.g. while preparing a new physics fill or during periods used for machine development.

Data Stores, Data Bases, and Storage Systems / **260**

Looking back on 10 years of the ATLAS Metadata Interface: Reflections on architecture, code design and development methods

Author: Jerome Fulachier[1]

Co-authors: Fabian Lambert [1]; Osman AIDEL [2]; Solveig Albrand [1]

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

[2] *CNRS / CC-IN2P3*

Corresponding Authors: jerome.fulachier@lpsc.in2p3.fr, solveig.albrand@lpsc.in2p3.fr, fabian.lambert@lpsc.in2p3.fr, oaidel@cc.in2p3.fr

The "ATLAS Metadata Interface"framework (AMI) has been developed in the context of ATLAS, one of the largest scientific collaborations. AMI can be considered to be a mature application, since its basic architecture has been maintained for over 10 years.

In this paper we will briefly describe the architecture and the main uses of the framework within the experiment (Tag Collector for release management and Dataset Discovery). These two applications, which share almost 2000 registered users, are superficially quite different, however much of the code is shared and they have been developed and maintained over a decade almost completely by the same team of 3 people.

We will discuss how the architectural principles established at the beginning of the project have allowed us to continue both to integrate the new technologies and to respond to the new metadata use cases which inevitably appear over such a time period.

These principles are:
- Integration of a schema description in the AMI databases enabling an advantageous use of generic code
- Modularity, in particular decoupling of generic database and application specific layers
- Benefiting from the "open/closed principle"a well-known concept of object-oriented

programming
- Development methods and use of tools which assure the stability of the project.

Software Engineering, Parallelism & Multi-Core / 173

Parallel track reconstruction in CMS using the cellular automaton approach

Authors: Daniel Funke[1]; Thomas Hauth[2]; Vincenzo Innocente[2]

Co-authors: Dennis Schieferdecker [3]; Gunter Quast [1]; Peter Sanders [3]

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

[2] *CERN*

[3] *Karlsruher Institut für Technologie*

Corresponding Authors: daniel.funke@cern.ch, thomas.hauth@cern.ch

The Compact Muon Solenoid (CMS) experiment at the Large Hadron Collider (LHC) at CERN near Geneva/Switzerland is a general-purpose particle detector which led, among many other results, to the discovery of a Higgs-like particle in 2012. It comprises the largest silicon-based tracking system built to date with 75 million individual readout channels and a total surface area of 205 mˆ2.

The precise reconstruction of particle tracks from this tremendous amount of input channels is a compute-intensive task and requires an elaborate set of algorithms. A large portion of the reconstruction runtime can be accounted to the Kalman filter-based, iterative track finding and fitting procedures. The combinatorial complexity of these algorithms depends on the number of particle tracks in one event, which itself scales with the number of simultaneous proton-proton collisions in the detector, named pile-up.

The foreseen LHC beam parameters for the next data taking period, starting in 2015, will result in an increase in the number of pile-up interactions. Due to the stagnating clock frequencies of individual CPU cores, new approaches to particle track reconstruction need to be evaluated in order to cope with the computational challenge of the increased number of tracks per event.

Track finding methods that are based on cellular automata (CA) offer a fast and parallelizable alternative to the well-established Kalman filter-based algorithms. The CA proceeds in three steps: i) identify compatible combinations of three energy deposits in adjacent layers of the tracking system; ii) join these fragments to complete track candidates; and iii) select the best tracks for further processing. All computations are rather simple and data local, therefore, they can be implemented in a highly data-parallel fashion. This makes CA a very promising method to be run on modern GPGPUs and hardware accelerators, as well as multi-core CPUs equipped with vector units.

We present a new cellular automaton based track reconstruction, which copes with the complex detector geometry of CMS. By combining the CA with a grid-based data structure, we enable fast access to measured positions on the silicon detectors. Furthermore, we detail the specific design choices made to allow for a high-performance computation on GPU and CPU devices, such as branch-avoidance and the use of computationally inexpensive cut criteria before employing more involved ones. To address the specifics of data parallel programming requirements, we further partition the steps of the CA into more fine-grained sub-tasks and use a two-pass approach for each of them, hence separating computation and memory write access. We conclude by evaluating the physics efficiency, as well as computational properties of our implementation on various hardware platforms.

Poster presentations / 195

An HTTP Ecosystem for HEP Data Management

Author: Fabrizio Furano[1]

Co-authors: Adrien Devresse [1]; Alejandro Alvarez Ayllon [1]; Andrea Manzi [1]; Ivan Calvet [1]; Martin Philipp Hellmich [2]; Oliver Keeble [1]; Ricardo Brito Da Rocha [1]

[1] *CERN*

[2] *University of Edinburgh (GB)*

Corresponding Author: fabrizio.furano@cern.ch

In this contribution we present a vision for the use of the HTTP protocol for data management in the context of HEP, and we present demonstrations of the use of HTTP-based protocols for storage access & management, cataloguing, federation and transfer.

The support of HTTP/WebDAV, provided by frameworks for scientific data access like DPM, dCache, STORM, FTS3 and foreseen for XROOTD, can be seen as a coherent ensemble –an ecosystem –that is based on a single, standard protocol, where the HEP-related features are covered, and the door is open to standard solutions and tools provided by third parties, in the context of the Web and Cloud technologies.

The application domain for such an ecosystem of services goes from large scale Cloud and Grid-like computing to the data access from laptops, profiting from tools that are shared with the Web community, like browsers, clients libraries and others. Particular focus was put into emphasizing the flexibility of the frameworks, which can interface with a very broad range of components, data stores, catalogues and metadata stores, including the possibility of building high performance dynamic federations of endpoints that build on the fly the feeling of a unique, seamless and efficient system. The overall goal is to leverage standards and standard practices, and use them to provide the higher level functionalities that are needed to fulfil the complex problem of Data Access in HEP. In this context we explain how the subset of SRM functionality relevant to disk system could be offered over HTTP. Other points of interest are about harmonizing the possibilities given by the HTTP/WebDAV protocols with existing frameworks like ROOT and already existing Storage Federations based on the XROOTD framework. We also provide quantitative evaluations of the performance that is achievable using HTTP for remote transfer and remote I/O in the context of HEP data, with reference to the recent implementation of HTTP support in FTS3.

Poster presentations / **258**

Next Generation HEP Networks at Supercomputing 2012

Authors: Harvey Newman[1]; Ian Gable[2]; Randy Sobie[2]; Shawn Mc Kee[3]

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

[2] *University of Victoria (CA)*

[3] *University of Michigan (US)*

Corresponding Author: igable@uvic.ca

We review the demonstration of next generation high performance 100 Gbps networks for HEP that took place at the Supercomputing 2012 (SC12) conference in Salt Lake City. Three 100 Gbps circuits were established from the California Institute of Technology, the University of Victoria and the University of Michigan to the conference show floor. We were able to to efficiently utilize these circuits using limited set of hardware surpassing previous records established at SC11. Highlights include a record overall disk to disk rate using the three links of 187 Gbps, a unidirectional transfer between storage systems in Victoria and Salt Lake on one link of 96 Gbps, an 80 Gbps transfer from Caltech to a single server with two 40GE interfaces at Salt Lake with nearly 100% use of the servers' interfaces at both ends, and a transfer using Remote Data Memory Access (RDMA) over Ethernet between Pasadena and Salt Lake that sustained 75 Gbps with a CPU load on the servers of only 5%. A total of 3.8 Petabytes was transferred over the three days of the conference exhibit, including 2 Petabytes on the last day. Three different storage setup were used during the demonstration: an conventional disk Lustre system, 2U rack servers containing solid state disks and systems containing PCI Express 3.0 Solid State storage cards.

Poster presentations / 418

Dynamic web cache publishing for IaaS clouds using Shoal

Authors: Ian Gable[1]; Randy Sobie[1]

[1] *University of Victoria (CA)*

Corresponding Author: igable@uvic.ca

It has been shown possible to run HEP workloads on remote IaaS cloud resources. Typically each running Virtual Machine (VM) makes use of the CERN VM Filesystem (CVMFS), a caching HTTP file system, to minimize the size of the VM images, and to simplify software installation. Each VM must be configured with a HTTP web cache, usually a Squid Cache, in proximity in order to function efficiently. Regular grid sites which use CVMFS to host worker node software areas use one or more Squid servers for their site. However, each IaaS cloud site has none of the static infrastructure associated with a typical HEP site. Each cloud VM must be configured to use a particular Squid server. We have developed a method and software application called Shoal for publishing squid caches which are dynamically created on IaaS clouds. The Shoal server provides a simple REST interface which allows clients to determine their closest Squid cache. Squid servers advertise their existence by running Shoal agent which uses the Advanced Message Queuing Protocol (AMQP)to publish information about the Squid server to the Shoal server. Having a method for exchanging the information rapidly allows for Squid servers to be instantiated in the cloud in response to load and for clients to quickly learn of their existence. In this work we describe the design of Shoal and evaluate its performance at scale in an operational system.

Event Processing, Simulation and Analysis / 204

Track Reconstruction at the ILC

Author: Frank-Dieter Gaede[1]

Co-authors: Christoph Rosemann [2]; Georgios Gerasimos Voutsinas [3]

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

[2] *DESY*

[3] *Institut Pluridisciplinaire Hubert Curien (FR)*

Corresponding Author: frank-dieter.gaede@cern.ch

One of the key requirements for Higgs physics at the International Linear Collider ILC is excellent track reconstruction with very good momentum and impact parameter resolution. ILD is one of the two detector concepts at the ILC.
Its central tracking system comprises of a highly granular TPC, an intermediate silicon tracker and a pixel vertex detector, and it is complemented by silicon tracking discs in the forward direction.
Large hit densities from beam induced incoherent electron-positron pairs at the ILC pose an additional challenge to the pattern recognition algorithms.
We present the ILD tracking algorithms that are using clustering techniques, cellular automatons and kalman filter based track extrapolation. The performance of the tracking reconstruction is evaluated using a realistic geant4 simulation including dead material, gaps and imperfections, that recently has been used for a large Monte Carlo production for the Detailed Baseline Design of the ILD detector concept.
The algorithms are written to a large extend in a framework independent way, with the eventual goal of providing a generic tracking toolkit. Parallelization techniques for some of the algorithms are under investigation.

Data Stores, Data Bases, and Storage Systems / 289

The future of event-level information repositories, indexing and selection in ATLAS

Author: Jack Cranshaw[1]

Co-authors: Armin Nairz [2]; Dario Barberis [3]; David Malon [1]; Donnchadha Quilty [4]; Elizabeth Gallas [5]; Gancho Dimitrov [2]; Julius Hrivnac [6]; Marcin Nowak [7]; Peter Van Gemmeren [1]; Qizhi Zhang [1]; Roman Sorokoletov [8]; Thomas Doherty [9]

[1] *Argonne National Laboratory (US)*

[2] *CERN*

[3] *Università e INFN Genova (IT)*

[4] *University of Glasgow (GB)*

[5] *University of Oxford (GB)*

[6] *Universite de Paris-Sud 11 (FR)*

[7] *Brookhaven National Laboratory (US)*

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

[9] *Department of Physics and Astronomy-University of Glasgow*

Corresponding Authors: elizabeth.gallas@physics.ox.ac.uk, cranshaw@anl.gov, malon@anl.gov, dario.barberis@cern.ch, gancho.dimitrov@cern.ch, tdoherty@physics.gla.ac.uk, julius.hrivnac@cern.ch, armin.nairz@cern.ch, marcin.nowak@cern.ch, donnchadha.quilty@cern.ch, roman.sorokoletov@cern.ch, peter.van.gemmeren@cern.ch, qzhang@anl.gov

ATLAS maintains a rich corpus of event-by-event information that provides a global view of virtually all of the billions of events the collaboration has seen or simulated, along with sufficient auxiliary information to navigate to and retrieve data for any event at any production processing stage. This unique resource has been employed for a range of purposes, from monitoring, statistics, anomaly detection, and integrity checking to event picking, subset selection, and sample extraction. Recent years of data-taking provide a foundation for assessment of how this resource has and has not been used in practice, of the uses for which it should be optimized, of how it should be deployed and provisioned for scalability to future data volumes, and of the areas in which enhancements to functionality would be most valuable.

This paper describes how ATLAS event-level information repositories and selection infrastructure are evolving in light of this experience, and in view of their expected roles both in wide-area event delivery services and in an evolving ATLAS analysis model in which the importance of efficient selective access to data can only grow.

Data Stores, Data Bases, and Storage Systems / 251

Utility of collecting metadata to manage a large scale conditions database in ATLAS

Author: Elizabeth Gallas[1]

Co-authors: Andrea Formica [2]; Misha Borodin [3]; Solveig Albrand [4]

[1] *University of Oxford (GB)*

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

[3] *Moscow State Engineering Physics Institute (RU)*

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

Corresponding Authors: elizabeth.gallas@physics.ox.ac.uk, solveig.albrand@lpsc.in2p3.fr, mikhail.borodin@cern.ch, andrea.formica@cern.ch

The ATLAS Conditions Database, based on the LCG Conditions Database infrastructure, contains a wide variety of information needed in online data taking and offline analysis. The total volume of ATLAS conditions data is in the multi-Terabyte range.

Internally, the active data is divided into 65 separate schemas (each with hundreds of underlying tables) according to overall data taking type, detector subsystem, and whether the data is used offline or strictly online. While each schema has a common infrastructure, each schema's data is entirely independent of other schemas, except at the highest level, where sets of conditions from each subsystem are tagged globally for ATLAS event data reconstruction and reprocessing.

The partitioned nature of the conditions infrastructure works well for most purposes, but metadata about each schema is problematic to collect in global tools from such a system because it is only accessible via LCG tools schema by schema. This makes it difficult to get an overview of all schemas, collect interesting and useful descriptive and structural metadata for the overall system, and connect it with other ATLAS systems. This type of global information is needed for time critical data preparation tasks for data processing and has become more critical as the system has grown in size and diversity.

Therefore, a new system has been developed to collect metadata for the management of the ATLAS Conditions Database. The structure and implementation of this metadata repository will be described. In addition, we will report its usage since its inception in LHC Run 1, how it will be exploited in the process of conditions data evolution during LS1 (the current LHC long shutdown) in preparation for Run 2, and long term plans to incorporate more of its information into future ATLAS Conditions Database tools and the overall ATLAS information infrastructure.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / **313**

Evaluating Google Compute Engine with PROOF

Authors: Gerardo Ganis[1]; Sergey Panitkin[2]

[1] *CERN*

[2] *Brookhaven National Laboratory (US)*

Corresponding Authors: gerardo.ganis@cern.ch, panitkin@bnl.gov

The advent of private and commercial cloud platforms has opened the question of evaluating the cost-effectiveness of such solution for computing in High Energy Physics .
Google Compute Engine (GCE) is a IaaS product launched by Google as an experimental platform during 2012 and now open to the public market.
In this contribution we present the results of a set of CPU-intensive and I/O-intensive tests we have run with PROOF on a GCE resources made available by Google for test purposes.
We have run tests on large scale PROOF clusters (up to 1000 workers) to study the overall scalability of coordinated multi-process jobs.
We encountered the known scalability limitation of the PROOF technology in single master mode and we have investigated solution to lift this limitation.
We have studied and compared the performance of ephemeral and persistent storage with PROOF-Lite on the single machines and of standard PROOF on the whole cluster.
We will discuss our results in perspective, in particular with respect to the typical analysis needs of an LHC experiment.

Data Acquisition, Trigger and Controls / **359**

The evolution of the Trigger and Data Acquisition System in the ATLAS experiment

Authors: Francesca Pastore[1]; Taylor Childers[2]

[1] *University of London (GB)*

[2] *CERN*

Corresponding Authors: nicoletta.garelli@cern.ch, john.taylor.childers@cern.ch

The ATLAS experiment, aimed at recording the results of LHC proton-proton collisions, is upgrading its Trigger and Data Acquisition (TDAQ) system during the current LHC first long shutdown. The purpose of such upgrade is to add robustness and flexibility to the selection and the conveyance of the physics data, simplify the maintenance of the infrastructure, exploit new technologies and, overall, make ATLAS data-taking capable of dealing with increasing event rates.

The TDAQ system used to date is organised in a three-level selection scheme, including a hardware-based first-level trigger and second- and third-level triggers implemented as separate software systems distributed on commodity hardware nodes. The second-level trigger operates over limited regions of the detector, the so-called Regions-of-Interest (RoI). The third-level trigger deals instead with complete events.

While this architecture was successfully operated well beyond the original design goals, the accumulated experience stimulated interest to explore possible evolutions. With higher luminosities, the required number and complexity of Level-1 triggers will increase in order to satisfy the physics goals of ATLAS, while keeping the total Level-1 rates at or below 100kHz. The Central Trigger Processor will be upgraded to increase the number of manageable inputs and accommodate additional hardware for improved performance, and a new Topological Processor will be included in the slice. This latter will apply selections based either on geometrical information, like angles between jets/leptons, or even more complex observables to further optimize the selection at this trigger stage.

Concerning the high-level trigger, the main step in the current plan is to deploy a single homogeneous system, which merges the execution of the second and third trigger levels, still logically separated, on a unique hardware node. This design has many advantages, among which: the radical simplification of the architecture, the flexible and automatically balanced distribution of the computing resources, the sharing of code and services on nodes. Furthermore, the full treatment of the HLT selection on a single node enables both further optimisations, e.g. the caching of event fragments already collected for RoI-based processing, and new approaches giving better balancing of the selection steps before and after the event building. Prototyping efforts already demonstrated many of these benefits.

In this paper, we report on the design and the development status of the upgraded trigger system, with particular attention to the tests currently on-going to identify the required performance and to spot its possible limitations.

Poster presentations / **255**

Geant4 application in a web browser

Author: Laurent Garnier[1]

[1] *LAL-IN2P3-CNRS*

Corresponding Author: garnier@lal.in2p3.fr

Geant4 application in a web browser

Geant4 is a toolkit for the simulation of the passage of particles through matter. The Geant4 visualization system supports many drivers including OpenGL, OpenInventor, HepRep, DAWN, VRML, RayTracer, gMocren and ASCIITree, with diverse and complementary functionalities.

Web applications have an increasing role in our work, and thanks to emerging frameworks such as Wt [1], we are now able to build a web application on top of a C++ application without rewriting all the code. Because the Geant4 toolkit's visualization and user interface modules are well decoupled from the rest of Geant4, it is straightforward to adapt these modules to render in a web application instead of a computer's native window manager. The API of the Wt framework closely matches

that of Qt[3] so we can benefit from our experience in developing the Geant4 Qt driver. Rendering is through the WebGL[2] framework.

In this presentation, we will show how we ported the Geant4 interface to a Web application and how, with minimal effort, other Geant4 users can replicated this process to share their own Geant4 applications in a web browser.

[1] http://www.webtoolkit.eu
[2] http://www.khronos.org/webgl/
[3] http://geant4.web.cern.ch/geant4/UserDocumentation/UsersGuides/ForApplicationDeveloper/html/ch08s03.html#sect.VisI

Data Stores, Data Bases, and Storage Systems / **262**

Rucio - The next generation of large scale distributed system for ATLAS Data Management

Author: Vincent Garonne[1]

Co-authors: Angelos Molfetas [2]; Armin Nairz [1]; Cedric Serfon [1]; Graeme Andrew Stewart [1]; Luc Goossens [1]; Mario Lassnig [1]; Martin Barisits [1]; Ralph Vigne [3]; Thomas Beermann [4]

[1] *CERN*

[2] *University of Sydney (AU)*

[3] *University of Vienna (AT)*

[4] *Bergische Universitaet Wuppertal (DE)*

Corresponding Authors: vincent.garonne@cern.ch, mario.lassnig@cern.ch, graeme.andrew.stewart@cern.ch, martin.barisits@cern.ch, thomas.beermann@cern.ch, ralph.vigne@cern.ch, cedric.serfon@cern.ch, luc.goossens@cern.ch, armin.nairz@cern.ch, angelos.molfetas@cern.ch

Rucio is the next-generation Distributed Data Management (DDM) system benefiting from recent advances in cloud and "Big Data" computing to address HEP experiments scaling requirements. Rucio is an evolution of the ATLAS DDM system Don Quijote 2 (DQ2), which has demonstrated very large scale data management capabilities with more than 140 petabytes spread worldwide across 130 sites, and accesses from 1,000 active users. However, DQ2 is reaching its limits in terms of scalability, requiring a large number of support staff to operate and being hard to extend with new technologies. Rucio will address these issues by relying on a conceptual data model and new technology to ensure system scalability, address new user requirements and employ new automation framework to reduce operational overheads.

We present the key concepts of Rucio, including its data organization/representation and a model of how ATLAS central group and user activities will be managed. The Rucio design, and the technology it employs, is described, specifically looking at its RESTful architecture and the various software components it uses. We show also the performance of the system. We describe the strategy to roll-out the system during the first LHC long shutdown and how the transition from DQ2 to Rucio will be handled, including how Rucio will take advantage of the commissioning of new services, and how it will be integrated by external applications.

Facilities, Infrastructures, Networking and Collaborative Tools / **26**

Big Data over a 100G Network at Fermilab

Authors: Dave Dykstra[1]; Gabriele Garzoglio[2]; Hyunwoo Kim[3]; Marko Slyz[3]; Parag Mhashilkar[4]

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

[2] *FERMI NATIONAL ACCELERATOR LABORATORY*

[3] *Fermilab*

[4] *Fermi National Accelerator Laboratory*

Corresponding Author: garzoglio@fnal.gov

As the need for Big Data in science becomes ever more relevant, networks around the world are upgrading their infrastructure to support high-speed interconnections. To support its mission, the high-energy physics community as a pioneer in Big Data has always been relying on the Fermi National Accelerator Laboratory to be at the forefront of storage and data movement. This need was reiterated in recent years with the data taking rate of the major LHC experiments reaching tens of Petabytes per year. At Fermilab, this resulted regularly in peaks of data movement on the WAN in and out of the laboratory of about 30 Gbits/s and on the LAN between storage and computational farms of 160 Gbits/s. To address these ever increasing needs, as of this year Fermilab is connected to the Energy Sciences Network (ESNet) through a 100 Gbit/s link.

To understand the optimal system- and application-level configuration to interface computational systems with the new high-speed interconnect, Fermilab has deployed a Network R&D facility connected to the ESNet 100G Testbed. For the past two years, the High Throughput Data Program has been using the Testbed to identify gaps in data movement middleware when transferring data at these high-speeds. The program has published evaluations of technologies typically used in High Energy Physics, such as GridFTP, XrootD, and Squid. This work presents the new R&D facility and the continuation of the evaluation program.

Poster presentations / **386**

Grids, Virtualization and Clouds at Fermilab

Author: Keith Chadwick[1]

Co-authors: Gabriele Garzoglio [2]; Seo-Young Noh [3]; Steven Timm [1]

[1] *Fermilab*

[2] *FERMI NATIONAL ACCELERATOR LABORATORY*

[3] *KISTI*

Corresponding Authors: garzoglio@fnal.gov, chadwick@fnal.gov

Fermilab supports a scientific program that includes experiments and scientists located across the globe. To better serve this community, in 2004, the (then) Computing Division undertook the strategy of placing all of the High Throughput Computing (HTC) resources in a Campus Grid known as FermiGrid, supported by common shared services. In 2007, the FermiGrid Services group deployed a service infrastructure that utilized Xen virtualization, LVS network routing and MySQL circular replication to deliver highly available services that offered significant performance, reliability and serviceability improvements. This deployment was further enhanced through the deployment of an distributed redundant network core architecture and the physical distribution of the systems that host the virtual machines across multiple buildings on the Fermilab Campus.

In 2010, building on the experience pioneered by FermiGrid in delivering production services in a virtual infrastructure, the Computing Sector commissioned the FermiCloud, GPCF and Virtual Services projects to serve as platforms for support of scientific computing (FermiCloud & GPCF) and core computing (Virtual Services).

This work will present the evolution of the Fermilab Campus Grid, Virtualization and Cloud Computing infrastructure together with plans for the future.

Poster presentations / **284**

Job Scheduling in Grid Farms

Author: Andreas Gellrich[1]

[1] *DESY*

Corresponding Author: andreas.gellrich@desy.de

The vast majority of jobs in the Grid are embarrassingly parallel. In particular HEP tasks are divided into atomic jobs without need for communication between them. Jobs are still neither multi-threaded nor multi-core capable. On the other hand, resource requirements reach from CPU-dominated Monte Carlo jobs to network intense analysis jobs.

The main objective of any Grid site is to stably operate its Grid farm while achieving a high job slot occupancy, an optimal usage of the computing resources (network, CPU, memory, disk space) and guaranteed shares for the VOs and groups. In order to optimize the utilization of resources, jobs must be distributed intelligently over the slots, CPUs, and hosts. Although the jobs resource requirements cannot be deduced directly, jobs are mapped to POSIX user/group ID based on their VOMS-proxy. The user/group ID allows to distinguish jobs, assuming VOs make use of the VOMS group and role mechanism.

The multi-VO Tier-2 site at DESY (DESY-HH) supports ˜20 VOs on federated computing resources, using an opportunistic resource usage model. As at many EGI/WLCG sites, the Grid farm is based on the queuing system PBS/TORQUE, which was deployed from the EMI middleware repositories. Initially, the scheduler MAUI was used. It showed severe scalability problems with 4000 job slots as soon as the number of running plus queued jobs approached 10000. Job scheduling became slow or even blocked. In addition, MAUI's many configuration options appeared to be hard to control.

To be able to further increase the number of worker nodes as requested by the VOs (to currently 8000 job slots), DESY-HH needed a scalable and performing scheduler, which runs in conjunction with PBS/TORQUE. In the course of studying alternative scheduling models, a home-made scheduler was developed (working title: MySched), which is tailored to then needs of the DESY-HH Grid farm and uses the C-API of PBS/TORQUE. It is based on a simple scheduling model without support for multi-core jobs and job parallelism and is optimized for high job slot occupancy and intelligent distribution of jobs to the worker nodes. Furthermore, it allows for a fine-grained adjustment of limits and parameters on VO and group level.

In the contribution to CHEP 2013 we will discuss the impact of a classification of jobs according to their potential resources requirements on scheduling strategies. Subsequently, we will describe our home-made implementation and present operational results.

Software Engineering, Parallelism & Multi-Core / 453

Vectorizing the detector geometry to optimize particle transport

Author: Andrei Gheata[1]

[1] *CERN*

Corresponding Author: andrei.gheata@cern.ch

Among the components contributing to particle transport, geometry navigation is an important consumer of CPU cycles. The tasks performed to get answers to "basic" queries like locating a point within a geometry hierarchy or computing accurately the distance to the next boundary can become very computing intensive for complex detector setups. Among several optimization methods already in use by the existing geometry algorithms, like caching or solution finding based on topological constraints, the usage of modern processors SIMD capabilities is maybe the least explored. While the potential gain by vectorizing loops is important and the technology trends push for larger vector units and more CPU pipes, applying this technology to the highly hierarchical multiple branched geometry code is a difficult challenge. The techniques used for producing vectorized code range from simple transformations allowing for compiler auto-vectorization to usage of intrinsic SIMD instructions or external helper libraries.
The work done to vectorize an important part of the critical navigation algorithms in ROOT geometry will be described, as well as a detailed benchmark of the benefits. We describe from a critical point of view the different techniques that were used. We comment on the estimated efforts to extend this work to the full geometry, as well as the large potential gains coming from using a vector geometry navigator as client of a future vector-based particle transport engine.

Poster presentations / 234

An Infrastructure in Support of Software Development

Authors: Francesco Giacomini[1]; Marco Bencivenni[2]; Matteo Manzali[3]; Riccardo Veraldi[2]; Stefano Antonelli[4]; Stefano Longo[2]

[1] *INFN CNAF*

[2] *INFN*

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

[4] *CNAF - INFN*

Corresponding Author: francesco.giacomini@cern.ch

The success of a scientific endeavor depends, often significantly, on the ability to collect and later process large amounts of data in an efficient and effective way. Despite the enormous technological progress in areas such as electronics, networking and storage, the cost of the computing factor remains high. Moreover the limits reached by some historical directions of hardware development, such as the saturation of the CPU clock speed with the consequent strong shift towards hardware parallelization, has made the role of software more and more important.

In order to support and facilitate the daily activities of software developers within INFN, an integrated infrastructure has been designed, comprising several tools, each providing a function: project tracking, code repository, continuous integration, quality control, knowledge base, dynamic provisioning of virtual machines, services, or clusters thereof. When applicable, access to the services is based on the INFN-wide Authentication and Authorization Infrastructure. The system is being installed and progressively made available to INFN users belonging to tens of INFN sites and laboratories and will represent a solid foundation to the software development efforts of the many experiments and projects that see the involvement of INFN. The infrastructure will be beneficial especially for small- and medium-size collaborations, which often cannot afford the resources, in particular in terms of know-how, needed to set up such services.

This contribution describes the design of the infrastructure, the components that implement it, how they integrate with each other, in particular from the Authentication point of view, and how it is expected to evolve in the future.

Event Processing, Simulation and Analysis / 155

The Fast Simulation of the CMS detector

Author: Andrea Giammanco[1]

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

Corresponding Author: andrea.giammanco@cern.ch

A framework for Fast Simulation of particle interactions in the CMS detector has been developed and implemented in the overall simulation, reconstruction and analysis framework of CMS. It produces data samples in the same format as the one used by the Geant4-based (henceforth Full) Simulation and Reconstruction chain; the output of the Fast Simulation of CMS can therefore be used in the analysis in the same way as data and Full Simulation samples. The Fast Simulation has been used already for several physics analyses in CMS, in particular those requiring a generation of many samples to scan an extended parameter space of the physics model (e.g. SUSY) or for the purpose of estimating systematic uncertainties. Comparisons of the Fast Simulation results both with the Full Simulation and with the LHC data collected in the years 2010 and 2011 at the center of mass energy of 7 TeV will be shown, to demonstrate the level of accuracy achieved so far. In addition, a description of recent developments: a tighter integration with the Full Simulation in the simulation of the electronic read-out ("digitization") and of the pileup of events from other proton-proton collisions, both in-time and out-of-time.

Data Acquisition, Trigger and Controls / 33

Implementation of a PC-based Level 0 Trigger Processor for the NA62 Experiment

Authors: Marcello Pivanti[1]; Marco Sozzi[2]; Pietro Dalpiaz[3]; Sebastiano Fabio Schifano[1]

Co-author: Alberto Gianoli [3]

[1] *University of Ferrara and INFN Ferrara*

[2] *Sezione di Pisa (IT)*

[3] *Universita di Ferrara (IT)*

Corresponding Authors: alberto.gianoli@cern.ch, marcello.pivanti@cern.ch

The performance of "level 0" (L0) triggers is crucial to reduce and appropriately select the large amount of data produced by detectors in high energy physics experiments. This selection must be accomplished as fast as possible, since data staging within detectors is a critical resource. For example, in the NA62 experiment at CERN, the event rate is estimated at around 10 MHz, and the L0-trigger should reduce it by a factor of 10 within a time budget limit of 1ms.
So far, the most common approach to the development of an L0 trigger system has been based on custom hardware processors, so event filtering has been performed by algorithms implemented in hardware. More recently, the implementation of custom processors has been based on FPGA devices, whose hardware functionalities can be configured using specific programming languages.
The use of FPGAs offers greater flexibility in maintaining, modifying, improving filter algorithms, however even small changes require a hardware re-configuration of the systems, and changes to the algorithm logic can be limited by hardware constraints that have not been foreseen at the development time.

So, even if this approach guarantees fast processing, strong limitations still remains in the available flexibility when changing filtering algorithms on the fly or testing more filtering conditions at the same time could be an “added-value”, as required during the data-taking phase of the experiment.
In this contribution we present an innovative approach in the implementation of an L0-trigger system based on the use commodity PC, describing the architecture that we are developing for the NA62 experiment at CERN.
Data streams coming from the detectors are collected by an FPGA installed on a PCI-Express board plugged on a commodity PC. The FPGA receives data from detectors via giga-bit data channels, and stores them into the main memory of the PC. The PC then performs the filter algorithms on data available on its own memory, and writes back results to the FPGA for routing to the appropriate destination.
In our case we have used a commodity board with an Altera Stratix IV FPGA, 4 gigabit channels and a X8 Gen2 PCI-Express link delivering a peak bandwidth of 4 GB/s per direction.
In this presentation we focus on the description of the logic inside the FPGA to interface with the PCI-Express bus, and on the software organization including the Linux driver that allows the software filtering algorithm to read and write data to and from the FPGA.
We also analyze performances, and investigate ways to move quickly data to and from the FPGA.
Since the filtering program runs on a commodity PC, algorithm changes are much simpler, as they do not impact on the rest of the hardware system, and do not require to re-configure the FPGA.

Data Stores, Data Bases, and Storage Systems / **120**

Data Bookkeeping Service 3 - Providing event metadata in CMS

Authors: Manuel Giffels[1]; Yuyi Guo[2]

[1] *CERN*

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

Corresponding Authors: manuel.giffels@cern.ch, yuyi@fnal.gov

The Data Bookkeeping Service 3 (DBS 3) provides an improved event meta data catalog for Monte Carlo and recorded data of the CMS (Compact Muon Solenoid) experiment at the Large Hadron Collider (LHC). It provides the necessary information used for tracking datasets, like data processing history, files and runs associated with a given dataset on a scale of about 10ˆ5 datasets and more than 10ˆ7 files. All kinds of data processing in CMS are relying on the information stored in DBS. It is widely used within CMS, in Monte Carlo production, processing of recorded data as well as in physics analysis done by users.

DBS 3 has been completely re-designed and re-implemented in Python using a CherryPy based environment, utilizing RESTful (Representational State Transfer) web services, commonly used within the data management and workload management (DMWM) group of CMS. DBS 3 is using the Java Script Object Notation (JSON) dataformat for interchanging information and Oracle as database backend. Main focuses during the process of development were an adaptation of the database schema to better match the evolving CMS data processing model, the introduction of the Data Aggregation System in CMS, which is combining the information of a variety of database services (PhEDEx, SiteDB, DBS, etc.) in one user interface and the achievement of a better scalability to match the growing demands even in the future.

This contribution covers the design of the service, the results of recent stress and scale testing, as well as first experiences with the system during daily operations.

Plenaries / **494**

Horizon 2020: an EU perspective on data and computing infrastructures for research

Author: Kostas Glinos[1]

[1] *European Commission*

Through joint efforts between the HEP community in the early days of the EU DataGrid project, through EGEE, and via EGI-InSPIRE today, the European Commission has had a profound impact in the way computing and data management for high energy physics is done.

Kostas Glinos, Head of Unit eInfrastructures of the European Commission, has been with the European Commission since 1992. He leads the eInfrastructures unit of the Directorate General for Communications, Networks, Content and Technology since 1 January 2009. From 2003 to 2008 he was Head of the Embedded Systems and Control unit and interim Executive Director of the ARTEMIS Joint Undertaking. Previously he was deputy head of Future and Emerging Technologies. Before joining the Commission Kostas worked with multinational companies and research institutes in the U.S., Greece and Belgium. He holds a diploma in Chemical Engineering from the University of Thessaloniki, a PhD from the University of Massachusetts and a MBA in investment management from Drexel University.

As identified e.g. by the High Level Expert Group in their "Riding the Wave" report (October 2010), the emergence of "big data" in research triggered by ever more powerful instruments (in HEP, bioinformatics, astronomy, etc.) demands advanced computing resources and software to increase the available capacity to manage, store and analyse extremely large, heterogeneous and complex datasets. We should support integrated, secure, permanent, on-demand service-driven and sustainable e-infrastructures to tackle this challenge.
Our vision is a pan-European e-infrastructure that will cover the whole lifecycle of scientific data and provide the necessary computing resources needed by researchers to process information. All components of the e-infrastructure should expose standard interfaces to enable interoperation with other relevant e-infrastructures.

Poster presentations / 450

Hangout With CERN - Reaching the Public with the Collaborative Tools of Social Media

Author: Steven Goldfarb[1]

Co-authors: Achintya Rao [2]; Kate Kahle [3]

[1] *University of Michigan (US)*

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

[3] *CERN*

Corresponding Author: steven.goldfarb@cern.ch

On July 4, 2012, particle physics became a celebrity. Around 1,000,000,000 people (yes, 1 billion) saw rebroadcasts of two technical presentations announcing discovery of a new boson. The occasion was a joint seminar of the CMS and ATLAS collaborations, and the target audience were members of those collaborations plus interested experts in the field of particle physics. Yet, the world ate it up like a sporting event.

Roughly two days later, in a parallel session of ICHEP in Melbourne, Australia, a group of physicists decided to explain the significance of this discovery to the public. They started up a tool called "Hangout", part of the new Google+ social media platform, to converse directly with the public via videoconference and webcast. The demand to join this Hangout overloaded the server several times. In the end, a compromise involving Q&A via comments was set up, and the conversation was underway.

I present a new project born from this experience, called Hangout With CERN, and discuss its success in creating an effective conversational channel between the public and high-energy physicists. I

review earlier efforts by both CMS and ATLAS contributing to this development, and then describe the current programme, involving nearly all aspects of CERN, and some topics that go well beyond that. I conclude by discussing the potential of the program both to improve our accountability to the public and to train our community for public communication.

Poster presentations / **457**

FwWebViewPlus: integration of web technologies into WinCC-OA based Human-Machine Interfaces at CERN.

Author: Piotr Golonka[1]

[1] *CERN*

Corresponding Author: piotr.golonka@cern.ch

Rapid growth of popularity of web applications gives rise to a plethora of reusable graphical components, such as Google Chart Tools or
jQuery Sparklines, implemented in JavaScript and running inside a web browser. In the paper we describe the tool that allows for
seamless integration of web-based widgets into WinCC Open Architecture, the SCADA system used commonly at CERN to build complex Human-Machine Interfaces. Reuse of widely available widget libraries and pushing the development efforts to a higher abstraction layer based on scripting language allow for significant reduction in maintenance of the code in multi-platform environment, when compared to currently used C++ visualization plugins. Adequately designed interfaces allow for rapid integration of new web widgets into WinCC-OA. At the same time the mechanisms familiar to WinCC-OA HMI developers are preserved, making the use of new widgets "native". Perspectives for further integration between the realms of WinCC-OA/CTRL and Web/JavaScript development are also discussed.

Facilities, Infrastructures, Networking and Collaborative Tools / **59**

Indico 1.0

Authors: Jose Benito Gonzalez Lopez[1]; Pedro Ferreira[1]

Co-authors: Alberto Resco Perez [1]; Jakub Piotr Trzaskoma [2]; Matthew Alexander Pugh [3]; Thomas Baron [1]

[1] *CERN*

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

[3] *Aberystwyth University / CERN*

Corresponding Author: jose.benito.gonzalez@cern.ch

Indico has evolved into the main event organization software, room booking tool and collaboration hub for CERN. The growth in its usage has only accelerated during the past 9 years, and today Indico holds more that 215,000 events and 1,100,000 files. The growth was also substantial in terms of functionalities and improvements. In the last year alone, Indico has matured considerably in 3 key areas: enhanced usability, optimized performance and additional features, especially those related to meeting collaboration. Along the course of 2012, much activity has centered around consolidating all this effort and investment into "version 1.0", recently released in 2013.

Version 1.0 brings along new features, such as the Outlook calendar synchronization for participants, many new and clean interfaces (badges and poster generation, list of contributions, abstracts, etc) and so forth. But most importantly, it brings a message, Indico is now stable, consolidated and mature after more than 10 years of non-stop development. This message is addressed not only to

CERN users but also to the many organisations, in or outside HEP which have already installed the software and to others who might soon join this community. In this presentation we will describe the current state of the art of Indico, and how it was built.

This does not mean that the Indico software is complete, far from it! We have plenty of new ideas and projects that we are working on and we will also share them during CHEP 2013.

Poster presentations / **470**

The ILC Detailed Baseline Design Exercise: Simulating the Physics capabilities and detector performance of the Silicon Detector using the Grid

Author: Norman Anthony Graf[1]

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

Corresponding Author: norman.graf@slac.stanford.edu

The International Linear Collider (ILC) physics and detector
community recently completed an exercise to demonstrate the
physics capabilities of detector concepts. The Detailed
Baseline Design (DBD) involved the generation, simulation, reconstruction and analysis of large samples of Monte Carlo datasets. The detector simulations utilized extremely detailed Geant4 implementations of engineered detector elements. The datasets incorporated the full set of Standard Model physics backgraounds, as well as machine backgrounds, all overlaid with the correct time structure of overlapping events. We describe how this exercise was undertaken using Grid computing and storage elements, as well as the submission, bookkeeping and cataloging infrastructure developed to support this international endeavor.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / **102**

CMS Computing Model Evolution

Authors: Claudio Grandi[1]; David Colling[2]

Co-author: Ian Fisk [3]

[1] *INFN - Bologna*

[2] *Imperial College Sci., Tech. & Med. (GB)*

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

Corresponding Authors: claudio.grandi@cern.ch, ian.fisk@cern.ch

The CMS Computing Model was developed and documented in 2004. Since then the model 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 changes planned for the restart of the LHC program in 2015. We will discuss the changes planning in the use and definition of the computing tiers, that were defined with the MONARC project. We will present how we intend to use new services and infrastructure to provide more efficient and transparent access to the data. We will discuss the computing plans to make better use of the computing capacity by scheduling more of the processor nodes, making better use of the disk storage, and more intelligent use of the networking.
Speakers

Poster presentations / **259**

Toward the Cloud Storage Interface of the INFN CNAF Tier-1 Mass Storage System

Author: Pier Paolo Ricci[1]

Co-authors: Daniele Gregori [2]; Luca dell'Agnello [3]; Tommaso Boccali [4]; Vincenzo Vagnoni [5]; Vladimir Sapunenko [6]

[1] *INFN CNAF*

[2] *Istituto Nazionale di Fisica Nucleare (INFN)*

[3] *INFN-CNAF*

[4] *Sezione di Pisa (IT)*

[5] *INFN Bologna*

[6] *INFN*

Corresponding Authors: pierpaolo.ricci@cnaf.infn.it, tommaso.boccali@cern.ch, vincenzo.vagnoni@bo.infn.it, vladimir.sapunenko@cern.ch, gregori@bo.infn.it

The Mass Storage System installed at the INFN CNAF Tier-1 is one of the biggest hierarchical storage facilities in Europe. It currently provides storage resources for about 12% of all LHC data, as well as to other High Energy Physics experiments.
The Grid Enabled Mass Storage System (GEMSS) is the present solution implemented at the INFN CNAF Tier-1 and it is based on a custom integration between a high performance parallel file system (General Parallel File System, GPFS) and a tape management system for long-term backend storage on magnetic media (Tivoli Storage Manager, TSM). Data access to Grid users is being granted since several years by the Storage Resource Manager (StoRM), an implementation of the standard SRM interface, widely adopted in the WLCG (Worldwide Large Hadron Collider Computing Grid) collaboration.
Requirements from the experiments at the (LHC) Large Hadron Collider and for other High Energy Physics cases, are leading to investigate the adoption of more flexible and user-friendly methods for accessing the storage over the WAN. These ideas include both the storage federation implementation (that is an approach where computing sites are divided in geographic federations permitting the direct file-access of the "federated"storage between sites) and, in general, promising cloud-like approach of sharing data storage. In particular at CNAF a specific integration between GEMSS and Xrootd has been developed in order to match the requirements of the CMS experiment. This was already the case for ALICE, using ad-hoc Xrootd modifications; CMS changes have been validated and are already available in the official Xrootd integration builds. This integration is currently under pre-production and appropriate large scale tests are under way. Moreover, an alternative approach for the storage federations based on http/webdav, in particular for the Atlas use case is under development.
In this paper we present the emerging methods and technologies, with particular attention to the cloud data access protocols like WebDAV the Xrootd and http data federations approach namespace . We also discuss the solutions adopted to increase the availability and to optimize the overall performance of the services behind the system, and we provide a short summary and comparison between results obtained using different data access protocols over a cloud-like environment.

Plenaries / **529**

Logistics update and tour programme

Corresponding Author: davidg@nikhef.nl

Plenaries / **497**

CHEP in Amsterdam: from 1985 to 2013

Author: David Groep[1]

[1] *NIKHEF (NL)*

Corresponding Author: davidg@nikhef.nl

Panel discussion / **531**

Dinner Cruise directions

Corresponding Author: davidg@nikhef.nl

Conference closing / **483**

Closing

Corresponding Author: davidg@nikhef.nl

Poster presentations / **7**

FPGA-based 10-Gbit Ethernet Data Acquisition Interface for the Upgraded Electronics of the ATLAS Liquid Argon Calorimeters

Authors: Benjamin Trocme[1]; Johannes Philipp Grohs[2]

Co-author: Arno Straessner [2]

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

[2] *Technische Universitaet Dresden (DE)*

Corresponding Authors: philipp.grohs@cern.ch, arno.straessner@cern.ch

The readout of the trigger signals of the ATLAS Liquid Argon (LAr) calorimeters is foreseen to be upgraded in order to prepare for operation during the first high-luminosity phase of the Large Hadron Collider (LHC). Signals with improved spatial granularity are planned to be received from the detector by a Digitial Processing System (DPS) in ATCA technology and will be sent in real-time to the ATLAS trigger system using custom optical links. These data are also sampled by the DPS for monitoring and will be read out by the regular Data Acquisition (DAQ) system of ATLAS which is a network-based PC-farm.

The bandwidth between DPS module and DAQ system is expected to be in the order of 10 Gbit/s per module and a standard Ethernet protocol is foreseen to be used. DSP data will be prepared and sent by a modern FPGA either through a switch or directly to a Read-Out System (ROS) PC serving as buffer interface of the ATLAS DAQ.

In a prototype setup, an ATCA blade equipped with a Xilinx Virtex-5 FPGA is used to send data via an ATCA switch to a server PC which has 10 Gbit dual-port Myricom network interface cards installed. The FPGA is implementing a 10 Gbit Ethernet with a XAUI interface and UDP protocol. After tuning of the network parameters, transfer speeds of up to 9.94 Gbit/s were achieved. The 10

Gbit Ethernet link is also used for configuration of the FPGA. Data is stored in a ring-buffer on the server PC for further random access by the DAQ system according to a trigger ID.

The talk presents the overall concept of a 10 Gbit Ethernet readout link between a FPGA-based Data Processing System and a PC-based buffer and DAQ system, compatible with the existing ATLAS DAQ. Experience from the prototype system in ATCA technology will be reported including performance and technical implementation, which may also be useful for other DAQ applications of particle detectors.

Poster presentations / **417**

The CC1 system –a solution for private cloud computing.

Author: Mariusz Witek[1]

Co-authors: Bartlomiej Henryk Zabinski [1]; Janusz Chwastowski [1]; Krzysztof Danielowski [2]; Maciej Nabozny [2]; Milosz Zdybal [2]; Oleksandr Gituliar [2]; Piotr Wojcik [2]; Rafal Zbigniew Grzymkowski [3]; Tomasz Sosnicki [2]; Tomasz Wojton [2]; Zofia Sobocinska [2]

[1] *Polish Academy of Sciences (PL)*

[2] *Institute of Nuclear Physics PAN, Krakow, Poland*

[3] *P*

Corresponding Author: mariusz.witek@cern.ch

In the multidisciplinary institutes the traditional way of computations is highly ineffective. A computer cluster dedicated to a single research group is typically exploited at a rather low level. The private cloud model enables various groups to share computing resources. It can boost the efficiency of the infrastructure usage by a large factor and at the same time reduce maintenance costs. The complete cloud computing system has been developed in the Institute of Nuclear Physics PAN, Cracow. It is based on the Python programming language and a low level virtualization toolkit –libvirt. The CC1 system provides resources within the Infrastructure as a Service (IaaS) model. The main features of the system are the following:
• custom web-based user interface,
• automatic creation of virtual clusters equipped with a preconfigured batch system,
• groups of users with the ability to share resources,
• permanent virtual storage volumes that can be mounted to a VM,
• distributed structure –a federation of clusters running as a uniform cloud,
• quota for user resources,
• monitoring and accounting.
Emphasis was put on the simplicity of user access, administrative tasks and installation procedure. The self-service access to the system is provided via the intuitive Web interface. The administration module contains a rich set of tools for user management and system configuration. Particular attention was paid to the preparation of the automatic installation procedure based on a standard package management system of Linux distributions. This way the system can be setup quickly within one hour and operated without the need of a deep understanding of the underlying cloud computing technology. One of the crucial features is easy creation of clusters of virtual machines equipped with a preconfigured batch system. This way the intensive calculations can be performed on demand without the need of time-consuming configuration of clusters. When finished, the resources can be released and made accessible to other users.
The stable system was put into operation at the beginning of 2012 and was extensively used for calculations by various research teams, in particular by HEP groups. The CC1 software is distributed under the Apache License 2.0. The web site of the project is located at http://cc1.ifj.edu.pl.

Facilities, Infrastructures, Networking and Collaborative Tools / **30**

Network architecture and IPv6 deployment at CERN

Author: David Gutierrez Rueda[1]

Co-authors: Carles Kishimoto Bisbe [1]; Edoardo Martelli [1]

[1] *CERN*

Corresponding Author: david.gutierrez@cern.ch

The network infrastructure at CERN has evolved with the increasing service and bandwidth demands of the scientific community. Analysing the massive amounts of data gathered by the experiments requires more computational power and faster networks to carry the data. The new Data Centre in Wigner and the adoption of 100Gbps in the core of the network are the latest answers to these demands. In this presentation, the network architecture at CERN and the technologies deployed to support a reliable, manageable and scalable infrastructure will be described.

The status of the IPv6 deployment at CERN, from a network perspective, will also be covered, describing the mechanisms used to provide configuration to network clients and to give service managers the ability to decide when and how they provide their services with IPv6. The deployment of network services like DNS or DHCP for IPv6 will also be described, together with the lessons learnt during this deployment.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / 99

CMS Computing Operations During Run1

Authors: Christoph Paus[1]; Christoph Wissing[2]; Daniele Bonacorsi[3]; Ian Fisk[4]; Oliver Gutsche[4]

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

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

[3] *University of Bologna*

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

Corresponding Authors: oliver.gutsche@cern.ch, christoph.wissing@desy.de, paus@mit.edu, ian.fisk@cern.ch, daniele.bonacorsi@bo.infn.it

During the first run, CMS collected and processed more than 10B data events and simulated more than 15B events. Up to 100k processor cores were used simultaneously and 100PB of storage was managed. Each month petabytes of data were moved and hundreds of users accessed data samples. In this presentation we will discuss the operational experience from the first run. We will present the workflows and data flows that were executed, we will discuss the tools and services developed, and the operations and shift models used to sustain the system. Many techniques were followed from the original computing planning, but some were reactions to difficulties and opportunities. In this presentation we will also address the lessons learned from an operational perspective, and how this is shaping our thoughts for 2015.

Software Engineering, Parallelism & Multi-Core / 11

Evaluation of the flow-based programming (FBP) paradigm as an alternative to standard programming practices in physics data processing applications

Author: Vardan Gyurjyan[1]

Co-authors: Bryan Moffit [1]; Carl Timmer [1]; David Abbott [1]; Edd Jastrzemski [1]; Graham Heyes [1]; William Gu [1]

[1] *Jefferson Lab*

Corresponding Author: gurjyan@jlab.org

The majority of developed physics data processing applications (PDP) are single, sequential processes that start at a point in time, and advance one step at a time until they are finished. In the current era of cloud computing and multi-core hardware architectures this approach has noticeable limitations.
In this paper we present a detailed evaluation of the FBP-based Clas12 event reconstruction program that was deployed and operated both in cloud and in batch processing environments. We demonstrate the programming methodology and discuss some of the issues and optimizations affecting performance. We will also discuss our choice of using the Petri-Net process modeling formalism for the representation of the Clas12 PDP application building blocks which exhibit concurrency, parallelism, and synchronization.

Poster presentations / **4**

Phronesis, a diagnosis and recovery tool for system administrators

Author: Christophe Haen[1]

Co-authors: Niko Neufeld [2]; Vincent Barra [3]

[1] *Univ. Blaise Pascal Clermont-Fe. II (FR)*

[2] *CERN*

[3] *LIMOS, UMR 6158 CNRS, Univ. Blaise Pascal.*

Corresponding Author: christophe.denis.haen@cern.ch

The backbone of the LHCb experiment is the Online system, which is a very large and heterogeneous computing center. Making sure of the proper behavior of the many different tasks running on the more than 2000 servers represents a huge workload for the small expert-operator team and is a 24/7 task. At the occasion of CHEP 2012, we presented a prototype of a framework that we designed in order to support the experts. The main objective is to provide them with always improving diagnosis and recovery solutions in case of misbehavior of a service, without having to modify the original applications. Our framework is based on adapted principles of the Autonomic Computing model, on reinforcement learning algorithms, as well as innovative concepts such as Shared Experience. While the presentation made at CHEP 2012 showed the validity of our prototype on simulations, we here present a version with improved algorithms, manipulation tools, and report on experience with running it in the LHCb Online system.

Data Acquisition, Trigger and Controls / **447**

State Machine Operation of the MICE Cooling Channel

Author: Pierrick Hanlet[1]

[1] *Illinois Institute of Technology*

Corresponding Author: hanlet@fnal.gov

The Muon Ionization Cooling Experiment (MICE) is a demonstration experiment to prove the feasibility of cooling a beam of muons for use in a Neutrino Factory and/or Muon Collider. The MICE cooling channel is a section of a modified Study II cooling channel in which we will measure a 10% reduction in beam emittance. In order to ensure

a reliable measurement, MICE will measure the beam emittance before and after the cooling channel at the level of 1%, or an absolute measurement of 0.001. This renders MICE a precision experiment which requires strict controls and monitoring of all experimental parameters in order to control systematic errors. The MICE Controls and Monitoring system is based on EPICS and integrates with the DAQ, Data monitoring systems, and a configuration database. The cooling channel for MICE has between 12 and 18 superconductnig solenoid coils in 3 to 7 magnets, depending on the staged development of the experiment. The magnets are coaxial and in close proximity, thus requiring coordinated operation of magnets when ramping, responding to quench conditions, and quench recovery. To reliably manage the operation of the magnets, MICE is implementing state machines for each magnet and an over arching state machine for the magnets integrated in the cooling channel. The state machine transitions and operating parameters are stored/restored to/from the configuration database and coupled with MICE Run Control. Proper implementation of the state machines will not only ensure safe operation of the magnets, but will help ensure reliable data quality. A description of MICE, details of the state machines, and lessons learned from use of the state machines in recent magnet training tests will be discussed.

Poster presentations - Board: P1.01 / **446**

The MICE Run Control System

Author: Pierrick Hanlet[1]

[1] *Illinois Institute of Technology*

Corresponding Author: hanlet@fnal.gov

The Muon Ionization Cooling Experiment (MICE) is a demonstration experiment to prove the feasibility of cooling a beam of muons for use in a Neutrino Factory and/or Muon Collider. The MICE cooling channel is a section of a modified Study II cooling channel which will provide a 10% reduction in beam emittance. In order to ensure a reliable measurement, MICE will measure the beam emittance before and after the cooling channel at the level of 1%, or an absolute measurement of 0.001. This renders MICE a precision experiment which requires strict controls and monitoring of all experimental parameters in order to control systematic errors. The MICE Controls and Monitoring system is based on EPICS and integrates with the DAQ, Data monitoring systems, and a configuration database. The new MICE Run Control has been developed for to ensure proper sequencing of equipment and use of system resources to protect data quality. A description of this system, its implementation, and performance during recent muon beam data collection will be discussed.

Poster presentations / **345**

The Reconstruction Software for the Muon Ionisation Cooling Experiment Trackers

Author: Adam Dobbs[1]

Co-authors: Christopher Heidt [2]; David Adey [3]; Edward Santos [4]; Pierrick Hanlet [5]

[1] *Imperial College London*

[2] *University of California Riverside*

[3] *Fermilab*

[4] *Imperial College*

[5] *Illinois Institute of Technology*

Corresponding Authors: hanlet@fnal.gov, a.dobbs07@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. In order to measure the change in beam emittance, MICE is equipped with a pair of high precision scintillating fibre trackers. The trackers are required to measure a 10% change in beam emittance to 1% accuracy (giving an overall emittance measurement of 0.1%).

This paper describes the tracker reconstruction software, as a part of the overall MICE software framework, MAUS. The process of producing fibre digits is described for both the GEANT4 based Monte Carlo case, and for real data. Fibre clustering is described, proceeding to the formation of spacepoints, which are then associated with particle tracks using pattern recognition algorithms. Finally a full custom Kalman track fit is performed, to account for energy loss and multiple scattering. Exemplar results are shown for both Monte Carlo and cosmic ray data.

Data Acquisition, Trigger and Controls / 390

The IceCube Neutrino Observatory DAQ and Online System

Author: Kael Hanson[1]

[1] *Université Libre de Bruxelles*

Corresponding Author: kael.hanson@icecube.wisc.edu

The IceCube Neutrino Observatory is a cubic kilometer-scale neutrino detector built into the ice sheet at the geographic South Pole. The online system for IceCube comprises subsystems for data acquisition, online filtering, supernova detection, and experiment control and monitoring. The observatory records astrophysical and cosmic ray events at a rate of approximately 3 kHz and selects the most interesting events for transmission via satellite to the experiment's data warehouse in the northern hemisphere. The system has been designed to run in a remote environment with minimal operator intervention. Its user interface permits remote control and monitoring of the experiment both locally and via satellite. Despite the remote location and complexity of the various subsystems interoperating, the system as a whole achieves an uptime in excess of 98%. We describe the design and implementation of the core detector online systems: the Data Acquisition Software (DAQ), including the in-ice and surface optical modules, the triggering system, and event builder; the distributed Processing and Filtering (PnF) system; the IceCube Live control and monitoring system; and SPADE, the data archival and transport system.

Poster presentations / 201

Strategies for preserving the software development history in LCG Savannah

Authors: Benedikt Hegner[1]; Victor Diez Gonzalez[2]

[1] *CERN*

[2] *Univ. Rov. i Virg., Tech. Sch. Eng.-Unknown-Unknown*

Corresponding Author: benedikt.hegner@cern.ch

For more than ten years, the LCG Savannah portal has successfully served the LHC community to track issues in their software development cycles. In total, more than 8000 users and 400 projects use this portal. Despite its success, the underlying infrastructure that is based on the open-source project "Savane" did not keep up with the general evolution of web technologies and the increasing need for information security. Thus, LCG Savannah is about to be replaced by a new service based on the Jira issue and project tracking software.
During the many years of preparation for LHC running, a huge amount of project specific data were collected in LCG Savannah. We discuss the importance of these historical data and the valuable knowledge they represent for the collaborations. We present our strategy for preserving the LCG Savannah data, the tools we have developed that support the migration process, and we describe the current status of the project.

Event Processing, Simulation and Analysis / **203**

Introducing Concurrency in the Gaudi Data Processing Framework

Authors: Benedikt Hegner[1]; Danilo Piparo[1]; Pere Mato Vila[1]

[1] *CERN*

Corresponding Author: benedikt.hegner@cern.ch

In the past, the increasing demands for HEP processing resources could be fulfilled by distributing the work to more and more physical machines. Limitations in power consumption of both CPUs and entire data centers are bringing an end to this era of easy scalability. To get the most CPU performance per Watt, future hardware will be characterised by less and less memory per processor, as well as thinner, more specialized and more numerous cores per die, and rather heterogeneous resources. To fully exploit the potential of the many cores, HEP data processing frameworks need to allow for parallel execution of reconstruction or simulation algorithms on several events simultaneously.
We describe our experience in introducing concurrency related capabilities into Gaudi, a generic data processing software framework, which is currently being used by several HEP experiments, including the ATLAS and LHCb experiments at the LHC. After a description of the concurrent framework and the most relevant design choices driving its development, we demonstrate its projected performance emulating data reconstruction workflows of the LHC experiments. As a second step, we describe the behaviour of the framework in a more realistic environment, using a subset of the real LHCb reconstruction workflow, and present our strategy and the used tools to validate the physics outcome of the parallel framework against the results of the present, purely sequential LHCb software. We then summarize the measurement of the code performance of the multithreaded application in terms of memory and CPU usage and I/O load.

Data Stores, Data Bases, and Storage Systems / **335**

Cloud storage performance and first experience from prototype services at CERN

Authors: Dirk Duellmann[1]; Maitane Zotes Resines[1]; Rainer Toebbicke[1]; Seppo Sakari Heikkila[1]

[1] *CERN*

Corresponding Authors: seppo.heikkila@cern.ch, dirk.duellmann@cern.ch

Cloud storage is an emerging architecture aiming to provide increased scalability and access performance, compared to more traditional solutions. CERN is evaluating this promise using Huawei UDS

and OpenStack storage deployments, focusing on the needs of high-energy physics. Both deployed setups implement S3, one of the protocols that are emerging as standard in the cloud storage market. A set of client machines has been used to generate I/O load patterns to evaluate the performance of both storage systems.

In this contribution we present scalability results for meta-data and data throughput tests and analyse the performance impact of internal caches. Further we evaluated the system performance under random access patterns, which are typical for later stages of physics analysis.We conclude by summarising the results of a total cost of ownership evaluation of several prototype services based on this new storage technology.

Data Stores, Data Bases, and Storage Systems / **146**

DPM - efficient storage in diverse environments

Author: Martin Philipp Hellmich[1]

Co-authors: Alejandro Alvarez Ayllon [2]; Andrea Manzi [2]; David Smith [2]; Fabrizio Furano [2]; Ivan Calvet [2]; Oliver Keeble [2]; Ricardo Brito Da Rocha [2]

[1] *University of Edinburgh (GB)*

[2] *CERN*

Corresponding Author: martin.hellmich@cern.ch

Recent developments, including low power devices, cluster file systems and cloud storage, represent an explosion in the possibilities for deploying and managing grid storage. In this paper we present how different technologies can be leveraged to build a storage service with differing cost, power, performance, scalability and reliability profiles, using the popular DPM/dmlite storage solution as the enabling technology.

The storage manager DPM is designed for these new environments, allowing users to scape up and down as they need it, and optimizing their computing centers energy efficiency and costs. DPM runs on high-performance machines, profiting from multi-core and multi-CPU setups. It supports separating the database from the head node, largely reducing its hard disk requirements. Since version 1.8.6, DPM is released in EPEL and Fedora, simplifying distribution and maintenance, but also supporting the ARM architecture beside i386 and x86_64, allowing it to run the smallest low-power machines such as the raspberry pi or the CuBox. This usage is facilitated by the possibility to scale horizontally using a main database and a distributed memcached-powered namespace cache. Additionally, DPM supports a variety of storage pools in the backend, most importantly HDFS, S3-enabled storage, and cluster file systems, allowing users to fit their DPM installation exactly to their needs.

In this paper, we investigate the power-efficiency and total cost of ownership of various DPM configurations. We develop metrics to evaluate the expected performance of a setup both in terms of namespace and disk access considering the overall cost including equipment, power consumptions, or data/storage fees. The setups tested range from the lowest scale using raspberry pies with only 700MHz single cores and a 100Mbps network connections, over conventional multi-core servers to typical virtual machine instances in cloud settings. We evaluate the combinations of different name server setups, for example load-balanced clusters, with different storage setups, from using a classic local configuration to private and public clouds.

Data Stores, Data Bases, and Storage Systems / **216**

Data and Software Preservation for Open Science (DASPOS)

Authors: Mike Hildreth[1]; Mike Hildreth[2]

Co-authors: Gordon Watts [3]; Kenneth Bloom [4]; Mark Neubauer [5]; Mark Stephen Neubauer [6]; Robert William Gardner Jr [7]

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

[2] *Department of Physics-College of Science-University of Notre Da*

[3] *University of Washington (US)*

[4] *University of Nebraska (US)*

[5] *Univ. Illinois at Urbana-Champaign (US)*

[6] *Univ. Illinois at Urbana-Champaign*

[7] *University of Chicago (US)*

Corresponding Author: mikeh@omega.hep.nd.edu

Data and Software Preservation for Open Science (DASPOS), represents a first attempt to establish a formal collaboration tying together physicists from the CMS and ATLAS experiments at the LHC and the Tevatron experiments with experts in digital curation, heterogeneous high-throughput storage systems, large-scale computing systems, and grid access and infrastructure. Recently funded by the National Science Foundation, the project is organizing multiple workshops aimed at understanding use cases for data, software, and knowledge preservation in High Energy Physics and other scientific disciplines, including BioInformatics and Astrophysics. The goal of this project is the development and specification of an architecture for curating HEP data and software to the point where the repetition of a physics analysis using only the archived data, software, and analysis description is possible. The novelty of this effort is this holistic approach, where not only data but also software and frameworks necessary to use the data are part of the preservation effort, making it true "physics preservation"rather than merely data preservation. This effort is an exploration of the problems to be solved at the technical, sociological, and policy levels, in order for integrated data preservation as envisioned by DPHEP to be possible. This work will provide the solid foundation necessary for the next steps of preservation infrastructure development. The research is a combination of two overlapping activities: a "horizontal"coordination and consensus-forming activity, both internal to HEP and including other disciplines, to agree on prototype metadata definitions and other common aspects of data preservation, and the more technical construction of the "vertical"slice of archival infrastructure. One measure of success, the so-called "Curation Challenge"will be a small-scale but full system test of a particular archiving solution enabling the discovery and enumeration of the critical issues in establishing preservation architectures. A key aspect of this work will be the inclusion of different scientific disciplines in the discussions of research use cases, archival strategies, metadata definitions, and policy considerations. Through this extended dialogue, we will be able to establish elements of commonality that can lead to shared technical and architectural solutions across disciplines. We also expect to outline branch points throughout the preservation architecture specification where policy choices will dictate technical outcomes, leading to a blueprint for any discipline approaching the problems of large-scale data preservation and open access. We aim for these common solutions and principles established to serve a similar role within the HEP community that the OAIS (Open Archival Information System) model plays for Trusted Digital Repositories. Of equal importance to these broad-ranging policy and technology issues will be the training of a team of graduate students in the technical aspects of large data set preservation, global grid-based access tools, and other facets of this multi-disciplinary problem. Finally, the development of technologies for the preservation of large scientific data archives opens up the possibility of future scientific opportunities and insights not otherwise available.

Event Processing, Simulation and Analysis / 159

CMS Full Simulation: Evolution Toward the 14 TeV Run

Author: Vladimir Ivantchenko[1]

Co-authors: Elizabeth Sexton-Kennedy [2]; Mike Hildreth [3]

[1] *CERN*

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

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

Corresponding Authors: mikeh@omega.hep.nd.edu, vladimir.ivantchenko@cern.ch, sexton@fnal.gov

The total amount of Monte Carlo events produced for CMS in 2012 is about 6.5 billion. In the future run at 14 TeV larger datasets, higher particle multiplicity and higher pileup are expected. This is a new challenge for the CMS software. In particular, increasing the speed of Monte Carlo production by a significant factor without compromising the physics performance is a highly-desirable goal. In this work we present the current status of the CMS full simulation software and perspectives for improvements.

The CMS full simulation is based on the Geant4 toolkit. For the production in 2012 Geant4 9.4 was used. In this work we report on the physics performance of the new Geant4 9.6 version, currently in development. Comparisons between 2012 data and Monte Carlo predictions will be shown, and validation software will be discussed.

Several methods have been studied that might allow a significant increase in speed of the CMS full simulation: fast mathematical libraries, GFlash, and Geant4 biasing options. In this presentation effects of these methods will be discussed and validation results will be presented. The most significant CPU improvement comes from the Russian roulette method in the new Geant4, which will be described in detail.

Event Processing, Simulation and Analysis / 161

Strategies for Modeling Extreme Luminosities in the CMS Simulation

Author: Mike Hildreth[1]

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

Corresponding Author: mikeh@omega.hep.nd.edu

Within the last year, design studies for LHC detector upgrades have begun to reach a level of detail that requires the simulation of physics processes with simulation performance at the level provided by Geant4. Full detector geometries for potential upgrades have been designed and incorporated into the CMS software. However, the extreme luminosities expected during the lifetimes of the upgrades must also be simulated. The use of many individual minimum-bias interactions to model the pileup poses several challenges to the CMS Simulation framework, including huge memory consumption, increased computation time, and the necessary handling of large numbers of event files during Monte Carlo production.

Recently, CMS has re-engineered the Simulation framework to allow the addition of pileup events using a dramatically smaller memory footprint. An alternate framework has been designed that can take the additional interactions from the data itself, obviating the need for hundreds of simulated minimum bias interactions to populate a single hard-scatter Monte Carlo event. Both of these developments are expected to have a dramatic impact on the efficiency of Monte Carlo production for the upcoming 14 TeV running and future CMS upgrade studies, and both can be used by the CMS Fast Simulation and the Full Geant4-based code. The structure of these reforms and the problems faced in their implementations will be discussed.

Software Engineering, Parallelism & Multi-Core / 387

A well-separated pairs decomposition algorithm for kd-trees implemented on multi-core architectures

Author: raul lopes[1]

Co-authors: Ivan Reid [2]; Peter Hobson [2]

[1] *School of Design and Engineering - Brunel University, UK*

[2] *Brunel University (GB)*

Corresponding Authors: raul.lopes@brunel.ac.uk, peter.hobson@brunel.ac.uk

Variations of kd-trees represent a fundamental data structure frequently used in geometrical algorithms, Computational Statistics, and clustering. They have numerous applications, for example in track fitting, in the software of the LHC experiments and in physics in general. Computer simulations of N-body systems, for example, have seen applications in the study of dynamics of interacting galaxies, particle beam physics, and molecular dynamics in biochemistry. The many-body tree methods devised by Barnes and Hutt in the 1980s and the Fast Multipole Method introduced in 1987 by Greengard and Rokhlin use variants of kd-trees to reduce the computation time upper bound to O(n log n) or even O(n) from the
O(nˆ2) demanded by Particle in Cell algorithms. Naive approaches to
kd-trees can, however, lead to algorithms that produce uncompressed trees and use O(nˆ2) work to build a tree for n items. We present an
algorithm that uses the principle of well-separated pairs decomposition to always produce compressed trees in O(n log n) work. We present and evaluate parallel implementations for the algorithm that can take advantage of multi-core architectures.

Facilities, Infrastructures, Networking and Collaborative Tools / **424**

Setting up collaborative tools for a 1000-member community

Authors: Adrien Rivière[1]; Dirk Hoffmann[1]

Co-author: Consortium CTA [2]

[1] *Centre de Physique des Particules de Marseille, CNRS/IN2P3*

[2] *Heidelberg*

Corresponding Author: dirk.hoffmann@cern.ch

The CTA (Cherenkov Telescope Array) consortium is developing a next generation ground-based instrument for very high energy gamma-ray astronomy, made up of approximately 100 telescopes of at least three different sizes. It counts presently more than 1000 members, out of which almost 800 have a computer account to use the “CTA web services”.

CTA decided in 2011 to use a SharePoint 2010 “site collection” operated by a subcontractor, as standard framework for collaborative tools. This system completed pre-existing installations of an InDiCo server, a MailMan mailing list management system, a MediaWiki instance and a Drupal content management system used by a community of 400 official users at that time. First actions were the unification of user logins in view of the expected increase of activity, by means of an LDAP directory, followed by the creation of an administrative user database for the consortium reflecting the up to date memberships of the groups from 178 institutes. Migration of existing data had to be carried out, as well as careful and persuasive user training.

Working groups and ad-hoc communities (countries, institutes) have parametrised the SharePoint “sub-sites” system for their needs, while the CTA-IT-support team, spread over several sites in Europe, has developed two major applications that were not available from SharePoint out of the box:

- a document management system for the consortium (RecordsCentre) and
- a conference and publication management system named after the committee in charge - SAPO (Speakers and Publications Office).

CTA uses a mixed “LAMP+WS” (Linux, Apache, MySQL, PHP, Perl, Windows and SharePoint) system at present to respond to user needs in an optimal way. Extensions by SVN and Redmine are

foreseen and partly operational in a transparent way for CTA users. Further specific implementations like approval workflows for the engineering process are planned in SharePoint.

After two years of operation and interaction with an outsourced provider for the SharePoint services, we are trying to strike a balance of our choices and the resources needed to implement new requests and keep the existing system up, running and safe.

Poster presentations / **427**

PLUME –FEATHER

Author: Dirk Hoffmann[1]

Co-author: Technical Committee PLUME [2]

[1] *Centre de Physique des Particules de Marseille, CNRS/IN2P3*

[2] *CNRS*

Corresponding Author: dirk.hoffmann@cern.ch

PLUME - FEATHER is a non-profit project created to Promote economicaL, Useful and Maintained softwarE For the Higher Education And THE Research communities. The site references software, mainly Free/Libre Open Source Software (FLOSS) from French universities and national research organisations, (CNRS, INRA...), laboratories or departments. Plume means feather in French. The main goals of PLUME –FEATHER are:
• promote the community's own developments,
• contribute to the developement and sharing FLOSS (Free/Libre Open Source Software) information, experiences and expertise in the community,
• bring together FLOSS experts and knowledgeable people to create a community,
• foster and facilitate FLOSS use, deployment and contribution in the higher education and the research communities.
PLUME - FEATHER was initiated by the CNRS unit UREC, which has been integrated into the CNRS computing division DSI in 2011. Various resources are provided by the main partners involved in the project, coming from many french research institutes. The French PLUME server contains more than 1000 software reference cards, edited and peer-reviewed by more than 909 contributors, out of 2000 overall members from the research and education community. The site http://projet-plume.org/ is online since November 2007, and the first English pages have been published in April 2009. Currently there are 91 software products referenced in the PLUME-FEATHER area, 14 of them having been published since January 2012.
The PLUME project is presented regularly in national conferences and events of the FLOSS community, and at CHEP 2012. Therefore we renew our proposal to expose a poster about the services available from the PLUME project on the international level to find not only users, but also contributors: editors and reviewers of frequently used software in our domain.

Poster presentations / **46**

Prototyping a Multi-10-Gigabit Ethernet Event-Builder for a Cherenkov Telescope Array

Authors: Dirk Hoffmann[1]; Julien Houles[2]

Co-author: The CTA Consortium [3]

[1] *Centre de Physique des Particules de Marseille, CNRS/IN2P3*

[2] *Centre de Physique des Particules de Marseilles / CNRS*

[3] *CTA Consortium*

Corresponding Author: dirk.hoffmann@cern.ch

We are developing the prototype of a high speed data acquisition (DAQ) system for the Cherenkov Telescope Array. This experiment will be the next generation ground-based gamma-ray instrument. It will be made up of approximately 100 telescopes of at least three different sizes, from 6 to 24 meters in diameter.

Each camera equipping the telescopes is composed of hundreds of light detecting modules pushing out data through gigabit Ethernet links. They will generate a total data flow of up to 24 Gb/s. Merging and handling such data rates with a single off the shelf computer and switches without any data loss require well designed and tested software and hardware. In a first stage, the software receives and reconstructs the incoming partial events. In a second stage, it performs on-line calculations with the data in order to improve event selection and sends the remaining data to the central DAQ. For the purpose of testing and stimulating our DAQ system, we designed and started to build a full scale dummy camera cluster with single board computers. This cluster provides 300 physical gigabit Ethernet ports which will send pre-calculated simulation data to the DAQ system, reproducing the timing and the instantaneous flow of a real camera. With this equipment, we are able to validate our hardware and software architectures. We will present our approach for the development of such a high data rate system, first measurements and solutions that we have applied to solve the problems we encountered to sustain the maximum dataflow reliably.

Poster presentations / **227**

The upgrade and re-validation of the Compact Muon Solenoid Electromagnetic Calorimeter Control System

Author: Oliver Holme[1]

Co-authors: Diogo Raphael Da Silva Di Calafiori [2]; Dragoslav Jovanovic [3]; Guenther Dissertori [2]; Lubomir Djambazov [2]; Peter Adzic [3]; Serguei Zelepukin [4]; Werner Lustermann [2]

[1] *ETH Zurich, Switzerland*

[2] *Eidgenoessische Tech. Hochschule Zuerich (CH)*

[3] *University of Belgrade (RS)*

[4] *University of Wisconsin (US)*

Corresponding Authors: oliver.holme@cern.ch, diogo.di.calafiori@cern.ch, peter.adzic@cern.ch

The Electromagnetic Calorimeter (ECAL) is one of the sub-detectors of the Compact Muon Solenoid (CMS) experiment of the Large Hadron Collider (LHC) at CERN. The Detector Control System (DCS) that has been developed and implemented for the CMS ECAL was deployed in accordance with the LHC schedule and has been supporting the detector data-taking since LHC physics runs started in 2009. During these years, the control system has been regularly adapted according to operational experience and new requirements, always respecting the constraints imposed on significant changes to a running system. Several hardware and software upgrades and system extensions were therefore deferred to the first LHC Long Shutdown (LS1). This paper presents the main architectural differences between the system that supported the CMS ECAL during its first years and the new design for the coming physics runs after LS1. Details on the upgrade planning, including the certification methods performed in the CMS ECAL DCS laboratory facilities, reports on the implementation progress and the expectations for the post-LS1 system are highlighted.

Data Acquisition, Trigger and Controls / **139**

The new CMS DAQ system for LHC operation after 2014 (DAQ2)

Author: Frans Meijers[1]

Co-author: Andre Georg Holzner [2]

[1] *CERN*

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

Corresponding Authors: andre.georg.holzner@cern.ch, frans.meijers@cern.ch

The Data Acquisition system of the Compact Muon Solenoid experiment at CERN assembles events at a rate of 100 kHz, transporting event data at an aggregate throughput of 100 GB/s. By the time the LHC restarts after the 2013/14 shut-down, the current compute nodes, networking, and storage infrastructure infrastructure will have reached the end of their lifetime. In order to handle higher LHC luminosities / or event pile-up, a number of sub-detectors will be upgraded, increase the number of read-out channels and replace the off-detector read-out electronics with an implementation in the microTCA architecture. The second generation DAQ system, foreseen for 2014, will need to accommodate the readout of both existing and new off-detector electronics and provide an increased throughput capacity. Advances in storage technology could make it feasible to write the output of the event builder to (ram or SSD) disks and implement the HLT processing entirely file based. We are presenting the design of the 2nd generation DAQ system, including studies of the event builder based on advanced networking technologies such as 10 and 40 Gb/s Ethernet and 56 Gb/s FDR Infiniband and exploitation of multi-core CPU architecture.

Poster presentations / **280**

AutoPyFactory and the Cloud: Flexible, scalable, and automatic management of virtual resources for ATLAS

Author: John Hover[1]

Co-authors: Jose Caballero Bejar [2]; Peter Love [3]

[1] *Brookhaven National Laboratory (BNL)-Unknown-Unknown*

[2] *Brookhaven National Laboratory (US)*

[3] *LANCASTER UNIVERSITY*

Corresponding Authors: john.hover@cern.ch, peter.love@cern.ch, jose.caballero@cern.ch

AutoPyFactory (APF) is a next-generation pilot submission framework that has been used as part of the ATLAS workload management system (PanDA) for two years. APF is reliable, scalable, and offers easy and flexible configuration. Using a plugin-based architecture, APF polls for information from configured information and batch systems (including grid sites), decides how many additional pilot jobs are needed, and submits them.

With the advent of cloud computing, providing resources goes beyond submitting pilots to grid sites. Now, the resources on which the pilot will run also need to be managed. Handling both pilot submission and controlling the virtual machine life cycle (creation, retirement, and termination) from the same framework allows robust and efficient management of the process.

In this paper we describe the design and implementation of these virtual machine management capabilities of APF. Expanding on our plugin-based approach, we allow cascades of virtual resources associated with a job queue. A single workflow can be directed first to a private, facility-based cloud, then a free academic cloud, then spot-priced EC2 resources, and finally on-demand commercial clouds. Limits, weighting, and priorities are supported, allowing free or less expensive resources to be used first, with costly resources only used when necessary. As demand drops, resources are drained and terminated in reverse order. Performance plots and time series will be included, showing how the implementation handles ramp-ups, ramp-downs, and spot terminations.

Poster presentations / **233**

Integration of g4tools in Geant4

Author: Ivana Hrivnacova[1]

[1] *Universite de Paris-Sud 11 (FR)*

Corresponding Author: ivana.hrivnacova@cern.ch

g4tools, that is originally part of the inlib and exlib packages [1], provides a very light and easy to install set of C++ classes that can be used to perform analysis in a Geant4 batch program. It allows to create and manipulate histograms and ntuples, and write them in supported file formats (ROOT, AIDA XML, CSV and HBOOK).

It is integrated in Geant4 through analysis manager classes, thus providing a uniform interface to the g4tools objects and also hiding the differences between the classes for different supported output formats. Moreover, additional features, such as for example histogram activation or support for Geant4 units, are implemented in the analysis classes following users requests. A set of Geant4 user interface commands allows the user to create histograms and set their properties interactively or in Geant4 macros. g4tools was first introduced in the Geant4 9.5 release where its use was demonstrated in one basic example, and it is already used in a majority of the Geant4 examples within the Geant4 9.6 release.

In this presentation, we will give an overview and the present status of the integration of g4tools in Geant4 and report on upcoming new features.

[1] http://inexlib.lal.in2p3.fr/

Software Engineering, Parallelism & Multi-Core / 175

Collaboration platform @CERN : Self-service for software development tools

Authors: Gautam Botrel[1]; Michal Husejko[2]

Co-authors: Alvaro Gonzalez Alvarez [2]; Georgios Koloventzos [2]; Jasiek Otto [3]; Nils Hoimyr [2]

[1] *Universite Paul Sabatier (FR)*

[2] *CERN*

[3] *University of Warsaw (PL)*

Corresponding Authors: michal.husejko@cern.ch, gautam.botrel@cern.ch, nils.hoimyr@cern.ch

This contribution describes how CERN has designed and integrated multiple essential tools for agile software development processes, ranging from a version control (Git) to issue tracking (Jira) and documentation (Wikis).

Running such services in a large organisation like CERN requires many administrative actions both by users and the service providers, such as creating software projects, managing access rights management, managing users and groups (which at CERN is tightly related to our LDAP eGroup implementation and Shibboleth based Single Sign On system), and performing tool-specific customisations.

Dealing with these requests manually would be a time-consuming task. This contribution illustrates how we have optimised the process from the conception stage (through the Git service example) to the design of an end-user facing platform acting as portal into all related services, inspired by popular portals for open-source developments such as Sourceforge, GitHub and others.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / 440

The Fermilab SAM data handling system at the Intensity Frontier

Author: Robert Illingworth[1]

[1] *Fermilab*

Corresponding Author: illingwo@fnal.gov

Fermilab Intensity Frontier experiments such as Minerva, NOvA, and MicroBooNE are now using an improved version of the Fermilab SAM data handling system. SAM was originally used by the CDF and D0 experiments for Run II of the Fermilab Tevatron to provide file metadata and location cataloguing, uploading of new files to tape storage, dataset management, file transfers between global processing sites, and processing history tracking. However SAM was heavily tailored to the Run II environment and required complex and hard to deploy client software, which made it hard to adapt to new experiments.

The Fermilab Computing Sector has progressively updated SAM to use modern, standardized, technologies in order to more easily deploy it for current and upcoming Fermilab experiments, and to support the data preservation efforts of the Run II experiments. We will describe these solutions, their technical implementation, and their deployment and integration with the experiments.

Poster presentations / 235

Geant4 Electromagnetic Physics for LHC Upgrade

Author: Vladimir Ivantchenko[1]

Co-authors: Alexander Bagulya [2]; Alexey Bogdanov [3]; Andreas Schaelicke [4]; Anton Ivantchenko [5]; Daren Lewis Sawkey ; John Apostolakis [1]; Laszlo Urban [6]; Luciano Pandola [7]; Michael Schenk ; Michel Maire [8]; Sebastien Laurent Incerti [9]; Vladimir Grichine [2]

[1] *CERN*

[2] *Russian Academy of Sciences (RU)*

[3] *Moscow State Engineering Physics Institute (RU)*

[4] *University of Edinburgh (GB)*

[5] *Geant 4 Associates International Experts in Radiation Simulatio*

[6] *Unknown*

[7] *INFN-LNGS*

[8] *LAPP*

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

Corresponding Author: vladimir.ivantchenko@cern.ch

Electromagnetic physics sub-package of the Geant4 Monte Carlo toolkit is an important component of LHC experiment simulation and other Geant4 applications. In this work we present recent progress in Geant4 electromagnetic physics modeling, with an emphasis on the new refinements for the processes of multiple and single scattering, ionisation, high energy muon interactions, and gamma induced processes. These developments affect the results of ongoing analysis of LHC data, in particular, electromagnetic shower shape parameters used for analysis of H->gg and Z->ee decays.

The LHC upgrade to future 14 TeV run will bring new requirements regarding the quality of electromagnetic physics simulation: energy, particle multiplicity, and statistics will be increased. To

address new requirements high energy electromagnetic models and cross sections are improved. Geant4 testing suite for electromagnetic physics is extended and new validation results will be presented. An evolution of CPU performance and developments for Geant4 multi-threading connected with Geant4 electromagnetic physics sub-packages will also be discussed.

Poster presentations / **439**

Data Preservation at the CDF Experiment

Author: Bodhitha Jayatilaka[1]

[1] *Fermilab*

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 experiment has 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 CDF 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 a talk on the status, benefits, and challenges of data preservation efforts within the CDF collaboration at Fermilab.

Poster presentations / **270**

Upgrades for Offline Data Quality Monitoring at ATLAS

Author: Peter Onyisi[1]

Co-authors: Anna Sfyrla [2]; Iurii Ilchenko [1]; James Frost [3]; Jessica Leveque [4]; Joerg Stelzer [5]; Sami Kama [6]

[1] *University of Texas (US)*

[2] *CERN*

[3] *University of Cambridge (GB)*

[4] *LAPP (Annecy-Le-Vieux)*

[5] *Michigan State University (US)*

[6] *Southern Methodist University (US)*

Corresponding Authors: m.dj@cern.ch, ponyisi@utexas.edu, yuriy.ilchenko@cern.ch, sami.kama@cern.ch, joerg.stelzer@cern.ch, anna.sfyrla@cern.ch, frost@hep.phy.cam.ac.uk, leveque@lapp.in2p3.fr

The ATLAS offline data quality monitoring infrastructure functioned successfully during the 2010-2012 run of the LHC. During the 2013-14 long shutdown, a large number of upgrades will be made in response to user needs and to take advantage of new technologies - for example, deploying richer web applications, improving dynamic visualization of data, streamlining configuration, and moving applications to a common messaging bus. Additionally consolidation and integration activities will occur. We will discuss lessons learned so far and the progress of the upgrade project, as well as associated improvements to the data reconstruction and processing chain.

Facilities, Infrastructures, Networking and Collaborative Tools / 472

Fabric Management (R)Evolution at CERN

Authors: Ben Jones[1]; Gavin Mccance[1]; Nacho Barrientos Arias[None]; Steve Traylen[1]

Co-authors: Akos Hencz [2]; Alex Lossent [1]; Daniel Lobato Garcia [3]; Ignacio Reguero [1]; Juan Manuel Guijarro [1]; Vitor Emanuel Gomes Gouveia [4]

[1] *CERN*

[2] *Tampere University of Technology (FI)*

[3] *University Carlos III (ES)*

[4] *Universidade de Lisboa (PT)*

Corresponding Authors: steve.traylen@cern.ch, gavin.mccance@cern.ch, ignacio.barrientos.arias@cern.ch, ben.dylan.jones@cern.c juan.manuel.guijarro@cern.ch

For over a decade CERN's fabric management system has been based on home-grown solutions. Those solutions are not dynamic enough for CERN to face its new challenges such as significantly scaling out, multi-site management and the Cloud Computing model, without any additional staff. This presentation will illustrate the motivations for CERN to move to a new tool-set in the context of the Agile Infrastructure project; it will explain the criteria that were applied both when selecting and when implementing the different components of this new infrastructure and the interactions between them, particularly focusing on services required for configuring servers. Finally, the current status of this ongoing migration will be exposed, together with lessons learned.

Data Acquisition, Trigger and Controls / 468

The Data Acquisition System for DarkSide-50

Authors: Alessandro Razeto[1]; Kurt Biery[2]

Co-authors: Stephen Foulkes [3]; stephen pordes [4]

[1] *LNGS*

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

[3] *Fermi National Accelerator Lab. (Fermilab)*

[4] *Fermilab*

Corresponding Authors: cdj@fnal.gov, biery@fnal.gov, alessandro.razeto@lngs.infn.it

The DarkSide-50 dark matter experiment has recently been constructed and commissioned at the Laboratori Nazionali del Gran Sasso (LNGS). The data acquisition system for the experiment was jointly constructed by members of the LNGS Research Division and the Fermilab Scientific Computing Division, and it makes use of commercial, off-the-shelf hardware components and the artdaq DAQ software toolkit.

This toolkit provided many core functions for the data acquisition system, including data transfer, event building, process control, system performance monitoring, and the framework for data compression and online data quality monitoring. Using the toolkit, experiment-specific functionality for reading out the commercial digitizers, trigger cards, and time-to-digital converters was developed, as were the experiment-specific data compression algorithms and data quality monitoring software.

We will present the overall design and implementation of the DarkSide-50 data acquisition system, the advantages of using the artdaq toolkit, the experiment-specific components that were developed, and the performance of the system.

Poster presentations / 171

VomsSnooper - a tool for managing VOMS records

Author: Stephen Jones[1]

[1] *Liverpool University*

Corresponding Author: sjones@hep.ph.liv.ac.uk

VomsSnooper is a tool that provides an easy way to keep documents and sites up to date with the newest VOMS records from the Operations Portal, and removes the need for manual edits to security configuration files.

Yaim is used to configure the middle-ware at grid sites. Specifically, Yaim processes variables that define which VOMS services are used to authenticate users of any VO. The data for those variables is administered centrally at the Operations Portal, and it is made available in XML format.

It was necessary at each site to manually convert the XML data to make it suitable for Yaim. At Liverpool, we wrote VomsSnooper to partly automate this process by checking and creating new VOMS records directly from the portal, providing a bridge between the Operations Portal and the site configuration. The tool is also used to automatically obtain and publish online the VOMS data of thirty GridPP Approved VOs in the GridPP wiki.

This tool solves several problems for Liverpool and the wider GridPP community.

Firstly, the VOMS records for GridPP Approved VOs were available from two sources that were not necessarily consistent, i.e. the Operations Portal and the GridPP wiki. So the first use case for VomsSnooper was to periodically synchronise the data on the GridPP wiki to the canonical source of the data in the Operations Portal, making both data sources compatible. The Approved VOs wiki now receives reliable, accurate, semi-automatic updates on a weekly basis, and long-term staleness has been eliminated. Sites can now update their records from either data source with more confidence.

Once a process was developed to automatically extract and format the VOMS records from the XML, it was apparent that the intermediate step of reading the Approved VOs wiki could be eliminated altogether. To this end, use cases were developed to both check the VOMS records at any site, and create new records directly from the Operations Portal. Sites who choose this approach can keep their records update to date in a semi-automatic manner, without reference to the Approved VOs wiki, and without manual edits to the security configuration files.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / 28

Optimization of data life cycles

Author: Christopher Jung[1]

Co-authors: Achim Streit [2]; Andre Giesler [3]; Fabian Rigoll [2]; Jörg Meyer [2]; Kilian Schwarz [4]; Marcus Hardt [2]; Martin Gasthuber [5]; Rainer Stotzka [2]

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

[2] *KIT*

[3] *FZ Jülich*

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

[5] *DESY*

Corresponding Author: christopher.jung@kit.edu

Data play a central role in most fields of Science. In recent years, the amount of data from experiment, observation, and simulation has increased rapidly and the data complexity has grown. Also, communities and shared storage have become geographically more distributed. Therefore, methods and techniques applied for scientific data need to be revised and partially be replaced, while keeping the community-specific needs in focus.
The German Helmholtz Association project "Large Scale Data Management and Analysis" (LSDMA) aims to maximize the efficiency of data life cycles in different research areas, ranging from high energy physics to system biology. In its five Data Life Cycle Labs (DLCLs), data experts closely collaborate with the communities in joint research and development to optimize the respective data life cycle. In addition, the Data Services Integration Team (DSIT) provides data analysis tools and services which are common to several DLCLs.
This presentation describes the various activities within LSDMA and focuses on the work performed in the DLCLs.

Data Acquisition, Trigger and Controls / **429**

An Event Building scenario in the trigger-less PANDA experiment

Author: Radoslaw Karabowicz[1]

[1] *GSI*

Corresponding Author: r.karabowicz@gsi.de

The PANDA experiment will be running up to 2.10ˆ7 antiproton-proton collisions per second at energies reaching 15 GeV.
The lack of simple features distinguishing the interesting events from background, as well as strong pileup of events' data streams make the use of a hardware trigger impossible. As a consequence the whole data stream of about 300 GB/s has to be analyzed online, i.e: tracking, vertex finding, particle identification and event building.

The GEM Tracker covers polar angles from 4 to 20 degrees in the forward direction, and can be used for the event building in a large fraction of events. In this work the analysis chain and the implemented algorithms will be presented. Moreover, the event builder prototype based on GEM Tracker data will be presented.

Poster presentations / **307**

MICE Experiment Data Acquisition system

Author: Yordan Ivanov Karadzhov[1]

[1] *Universite de Geneve (CH)*

Corresponding Author: yordan.karadzhov@cern.ch

The Muon Ionization Cooling Experiment (MICE) is under development at the Rutherford Appleton Laboratory (UK). The goal of the experiment is to build a section of a cooling channel that can demonstrate the principle of ionization cooling and to verify its performance in a muon beam. The final setup of the experiment will be able to measure a 10% reduction in emittance (transverse phase space size) of the beam with a relative precision of 1%.
The Data Acquisition (DAQ) system of the MICE experiment must be able to acquire data for ˜600 muons, crossing the experiment during 1 ms long beam spill. To fulfill this requirement, the Front-End Electronics (FEE) must digitize the analog signals in less than 500 ns and store the digitized data

in buffer memory. The time before the arrival of the next spill (˜1 s) is used to read out the buffers and record the data.
The acquisition of the data coming from the detectors is based on VME FEE interfaced to Linux PC processors. The event building, and the DAQ user interface software has been developed from the DATE package, originally developed for the ALICE experiment. The DAQ system communicates with the Control and Monitoring System of the experiment, using the EPICS (Experimental Physics and Industrial Control System) platform. This communication is used to exchange large number of parameters, describing the run conditions and the status of the data taking process. All these parameters are recorded in a Configuration Data Base.
The Detector DAQ is strongly dependent on the Trigger System, which is divided into two parts. The first part is responsible for the generation of the so called "Particle Trigger", which triggers the digitization of the analog signals received from the detectors. The second part of the Trigger System generates the so called "DAQ Trigger". This signal is generated after the end of the spill and causes the readout and storage of the digital data, corresponding to all the particle triggers received during a spill. A new Trigger System for the MICE experiment, based on programmable FPGA logic is now under development and tests.

Poster presentations / 240

Common accounting system for monitoring the ATLAS Distributed Computing resources

Author: Edward Karavakis[1]

Co-authors: I Ueda [2]; Jaroslava Schovancova [3]; Julia Andreeva [1]; Laura Sargsyan [4]; Pablo Saiz [1]; Simone Campana [1]; Stavro Gayazov [5]; Stephane Jezequel [6]

[1] *CERN*

[2] *University of Tokyo (JP)*

[3] *Brookhaven National Laboratory (US)*

[4] *ANSL (Yerevan Physics Institute) (AM)*

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

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

Corresponding Authors: edward.karavakis@cern.ch, julia.andreeva@cern.ch, simone.campana@cern.ch, stavro.gayazov@cern.ch, stephane.jezequel@cern.ch, pablo.saiz@cern.ch, laura.sargsyan@cern.ch, jaroslava.schovancova@cern.ch, i.ueda@cern.ch

The ATLAS Experiment at the Large Hadron Collider has been collecting data for three years. The ATLAS data are distributed, processed and analysed at more than 130 grid and cloud sites across the world. The total throughput of transfers is more than 5 GB/s and data occupies more than 120 PB on disk and tape storage. At any given time, there are more than 100,000 concurrent jobs running and more than a million jobs are submitted on a daily basis.

The large scale activity of the ATLAS Distributed Computing (ADC) increases the level of complexity of the system and, thus, increases the probability of failures or inefficiencies in the involved components. Effective monitoring provides a comprehensive way to identify and address any issues with the infrastructure. It is also a key factor in the effective utilisation of the system.

A significant effort has been invested over the last three years within the Experiment Dashboard project to assure effective and flexible monitoring. The Experiment Dashboard system provides generic solutions that cover all areas of ADC activities, such as data distribution and data processing over a large number of sites, and these solutions are extensively used by different categories of ATLAS users ranging from daily operations to resource management.

This talk covers a common accounting system used to monitor the utilisation of the available computational and storage resources of ATLAS. This system provides quality and scalable solutions that are flexible enough to support the constantly evolving requirements of the ATLAS user community.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / 106

Processing of the WLCG monitoring data using NoSQL.

Author: Edward Karavakis[1]

Co-authors: Alexandre Beche [1]; David Tuckett [1]; Ivan Antoniev Dzhunov [2]; Ivan Kadochnikov [3]; Jaroslava Schovancova [4]; Julia Andreeva [1]; Pablo Saiz [1]; Sergey Belov [3]

[1] *CERN*

[2] *University of Sofia*

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

[4] *Acad. of Sciences of the Czech Rep. (CZ)*

Corresponding Authors: edward.karavakis@cern.ch, alexandre.beche@cern.ch

The Worldwide LCG Computing Grid (WLCG) today includes more than 170 computing centres where more than 2 million jobs are being executed daily and petabytes of data are transferred between sites. Monitoring the computing activities of the LHC experiments, over such a huge heterogeneous infrastructure, is extremely demanding in terms of computation , performance and reliability. Furthermore, the generated monitoring flow is constantly increasing , which represents another challenge for the monitoring systems. While existing solutions are traditionally based on ORACLE for data storage and processing, recent developments evaluate NoSQL for processing large-scale monitoring datasets. NoSQL is an increasingly popular framework for processing datasets at the terabyte and petabyte scale using commodity hardware. In this contribution, we describe the integration of NoSQL data processing in the Experiment Dashboard framework and the first experience of using this technology for monitoring the LHC computing activities.

Poster presentations / 200

Offline software for the PANDA Luminosity Detector

Author: Anastasia Karavdina[1]

Co-authors: Achim Denig [2]; Florian Feldbauer [1]; Heinrich Leithoff [1]; Mathias Michel [3]; Miriam Fritsch [1]; Prometeusz Jasinski [1]; Stefan Pfluege [1]; Tobias Weber [1]

[1] *University Mainz*

[2] *Univ. Mainz*

[3] *Helmholtz-Institut Mainz*

Corresponding Author: karavdin@kph.uni-mainz.de

Precise luminosity determination is crucial for absolute cross-section measurements and scanning experiments with the fixed target PANDA experiment at the planned antiproton accelerator HESR (FAIR, Germany). For the determination of the luminosity we will exploit the elastic antiproton-proton scattering. Unfortunately there are no or only a few data with large uncertainties available in the momentum range we need for PANDA. Therefore in order to minimize the large systematic uncertainty from theory and former experiments we will perform measurements at very small momentum transfer (and thus very small scattering angle), where the Coulomb part of the elastic cross section can be calculated very precisely.
To achieve the precision needed for the luminosity determination the detector should have full angular acceptance and good spatial resolution. The current design are four planes of sensors with distances of 10 or 20 cm in between. The sensors themselves are HV-MAPS (High Voltage Monolithic Active Pixel Sensor). The whole set-up is placed in vacuum to minimize multiple scattering. In parallel to the prototype construction this design is under study with Monte Carlo based simulation. The luminosity detector is located outside of any magnetic field and will register hits of the straight tracks of the scattered antiprotons. Our reconstruction software includes hit reconstruction, track search, track fitting and software alignment. Moreover there are specific procedures like luminosity

extraction or background treatment.
Nowadays tracking systems usually measure up to hundred hits per track, which demands quite complicated pattern recognition algorithms, but makes track reconstruction more robust against loosing of hits in a track. In our case we have only a little amount of hits per track and have to be sure of using most of them. Therefore we developed two different algorithms for track search and did a performance study for the comparison of the algorithms.
In order to achieve the best resolution of the scattering angle measurement even a misalignment of 50 um has to be avoided. The Millipede algorithm will adjust the alignment corrections for all sensors simultaneously. The limitation of the Millipede algorithm (originally developed for a large number of layers) is studied in detail.
In this presentation the basic concept will be introduced before focusing on Monte Carlo based performance studies of
each reconstruction step.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / **239**

Reliability Engineering analysis of ATLAS data reprocessing campaigns

Author: Alexandre Vaniachine[1]

Co-authors: Dmitri Golubkov [2]; Dmytro Karpenko [3]

[1] *ANL*

[2] *Institute for High Energy Physics (IHEP)-Unknown-Unknown*

[3] *University of Oslo (NO)*

Corresponding Authors: dmytro.karpenko@cern.ch, vaniachine@anl.gov, dmitri.golubkov@cern.ch

During three years of LHC data taking, the ATLAS collaboration completed three petascale data reprocessing campaigns on the Grid, with up to 2 PB of data being reprocessed every year. In reprocessing on the Grid, failures can occur for a variety of reasons, while Grid heterogeneity makes failures hard to diagnose and repair quickly. As a result, Big Data processing on the Grid must tolerate a continuous stream of failures, errors and faults. While ATLAS fault-tolerance mechanisms improve the reliability of Big Data processing in the Grid, their benefits come at costs and result in delays making the performance prediction difficult.

Reliability Engineering provides a framework for fundamental understanding of the Big Data processing on the Grid, which is not a desirable enhancement but a necessary requirement. In ATLAS, cost monitoring and performance prediction became critical for the success of the reprocessing campaigns conducted in preparation for the major physics conferences. In addition, our Reliability Engineering approach supported continuous improvements in data reprocessing throughput during LHC data taking. The throughput doubled in 2011 vs. 2010 reprocessing, then quadrupled in 2012 vs. 2011 reprocessing.

We present the Reliability Engineering analysis of ATLAS data reprocessing campaigns providing the foundation needed to scale up the Big Data processing technologies beyond the petascale.

Poster presentations / **473**

Sustainable Software LifecycManagement for Grid Middleware: Moving from central control to the open source paradigms

Author: Alberto Aimar[1]

Co-authors: Markus Schulz [1]; Oliver Keeble [1]

[1] *CERN*

Corresponding Authors: oliver.keeble@cern.ch, alberto.aimar@cern.ch

In the recent years, with the end of the EU Grid projects such as EGEE and EMI in sight, the management of software development, packaging and distribution has moved from a centrally organised approach to a collaborative one, across several development teams. While selecting their tools and technologies, the different teams and services have gone through several trends and fashion of product and techniques. In general the software tools and technologies can be divided in three categories: (1) the in-house development of tools and services, (2) the participation and use of open source solutions and (3) the adoption and purchase of commercial tools and technologies.

The European projects spanned a period of more than a decade. Since the initial choices have been made new tools and paradigms have emerged. The initially adopted more central approach to the overall organization helped the process of a coherent migration to adequate tool chains.

The contribution will show, with concrete examples (mock/koji, JIRA, Bamboo, etc.), how we moved from centralized in-house services and repositories to distributed open source and commercial solutions. We will compare, based on experience, the benefits, shortcomings, costs and the risks of the approaches and the lessons learned from the different solutions.

In addition to the change of technologies also a more loosely coupled development of the different software components has emerged. In this case the adoption of popular public repositories (EPEL, Debian, Maven) together with their policies and standards for packaging and release management allows a less tight synchronization of the release of different packages. Each component can be updated at its own pace because it has to be compliant to the widely- adopted and enforced procedures for building testing and releasing into these repositories.

This presentation will provide some examples and insight on how a change of the core paradigms lead to the move to new tool chains and a loosely coupled collaboration with the focus on the impact on the development, build, release and deployment activities.

Poster presentations / **187**

GLUE 2 deployment: Ensuring quality in the EGI/WLCG information system

Authors: Oliver Keeble[1]; Stephen Burke[2]

Co-authors: Laurence Field [1]; Maria Alandes Pradillo [1]

[1] *CERN*

[2] *egi.eu*

Corresponding Authors: oliver.keeble@cern.ch, stephen.burke@stfc.ac.uk

The GLUE 2 information schema is now fully supported in the production EGI/WLCG information system. However, to make the schema usable and allow clients to rely on the information it is important that the meaning of the published information is clearly defined, and that information providers and site configurations are validated to ensure as far as possible that what they publish is correct. In this paper we describe the definition of a detailed schema usage profile, the implementation of a software tool to validate published information according to the profile and the use of the tool in the production Grid, and also summarise the overall state of GLUE 2 deployment.

Facilities, Infrastructures, Networking and Collaborative Tools / **336**

WLCG Security: A Trust Framework for Security Collaboration among Infrastructures

Author: Dave Kelsey[1]

Co-authors: Christos Kanellopoulos [2]; David Groep [3]; Irwin Gaines [4]; James Marsteller [5]; Jules Wolfrat [6]; Keith Chadwick [7]; Ralph Niederberger [8]; Romain Wartel [9]; Urpo Kaila [10]; Vincent Ribaillier [11]; Willy Weisz [12]

[1] *STFC - Science & Technology Facilities Council (GB)*

[2] *GRNET*

[3] *NIKHEF (NL)*

[4] *DOE/FNAL*

[5] *PSC*

[6] *SURFsara*

[7] *Fermilab*

[8] *FZ-Juelich*

[9] *CERN*

[10] *CSC - IT Center for Science Ltd.*

[11] *CNRS*

[12] *University of Vienna*

Corresponding Author: david.kelsey@stfc.ac.uk

The Security for Collaborating Infrastructures (SCI) group (http://www.eugridpma.org/sci/) is a collaborative activity of information security officers from several large-scale distributed computing infrastructures, including EGI, OSG, PRACE, WLCG, and XSEDE. SCI is developing a framework to enable interoperation of collaborating Grids with the aim of managing cross-Grid operational security risks and to build trust and develop policy standards for collaboration especially in cases where we cannot just share identical security policy documents. This assists in building the trust required for cooperation in operational security within WLCG.

Each infrastructure consists of distributed computing resources, users, and a set of policies and procedures all managed by different organisations. Even when such an infrastructure considers itself to be decoupled from other infrastructures, it is in fact subject to many of the same threats and vulnerabilities as other infrastructures because of the use of common software and technologies. Moreover, in WLCG there are users who use resources in more than one infrastructure and are thus potential vectors that can spread infection from one infrastructure to another. In each of these situations, the infrastructures can benefit from working together and sharing information on security issues.

We will present, based on current best practices and a long real-world experience, the current SCI activities including our documented requirements in 6 areas (operational security, incident response, traceability, participant responsibilities, legal issues and data protection) that each infrastructure must address in relation to being considered a trusted partner. We will also present an analysis method for showing the extent to which the infrastructures comply with the requirements.

Facilities, Infrastructures, Networking and Collaborative Tools / 360

WLCG and IPv6 - the HEPiX IPv6 working group

Authors: Bruno Heinrich Hoeft[1]; Dave Kelsey[2]

Co-authors: Andreas Pfeiffer [3]; Andrew Elwell [3]; Armin Nairz [3]; Costin Grigoras [3]; Duncan Rand [4]; Edoardo Martelli [3]; Fernando Lopez Munoz [5]; Francesco Prelz [6]; Gang CHEN [7]; Jiri Chudoba [8]; Kars Ohrenberg [9]; Keith Chadwick [10]; Luc Goossens [3]; Marek Elias [11]; Mario Reale [12]; Mark Mitchell [13]; Peter Clarke [14]; Qi Fazhi [15]; Ramiro Voicu [16]; Sandor Rozsa [16]; Simon Fayer [4]; Thomas Finnern [9]; Tomas Kouba [8]; Tony Wildish [17]

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

[2] *STFC - Science & Technology Facilities Council (GB)*

[3] *CERN*

[4] *Imperial College*

[5] *PIC*

[6] *Università degli Studi e INFN Milano (IT)*

[7] *INSTITUTE OF HIGH ENERGY PHYSICS*

[8] *Acad. of Sciences of the Czech Rep. (CZ)*

[9] *DESY*

[10] *Fermilab*

[11] *FZU ASCR*

[12] *GARR*

[13] *University of Glasgow*

[14] *University of Edinburgh (GB)*

[15] *IHEP*

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

[17] *Princeton University (US)*

Corresponding Author: david.kelsey@stfc.ac.uk

The HEPiX (http://www.hepix.org) IPv6 Working Group has been investigating the many issues which feed into the decision on the timetable for the use of IPv6 networking protocols in HEP Computing, in particular in WLCG. RIPE NCC, the European Regional Internet Registry, ran out of IPv4 addresses in September 2012. The North and South America RIRs are expected to run out in 2014. In recent months it has become more clear that some WLCG sites, including CERN, are running short of IPv4 address space, now without the possibility of applying for more. This has increased the urgency for the switch-on of dual-stack IPv4/IPv6 on all outward facing WLCG services to allow for the eventual support of IPv6-only clients.

The activities of the group include the analysis and testing of the readiness for IPv6 and the performance of many required components, including the applications, middleware, management and monitoring tools essential for HEP computing. Many WLCG Tier 1 and Tier 2 sites are participants in the group's distributed IPv6 testbed and the major LHC experiment collaborations are fully engaged in the testing. We have worked closely with similar activities elsewhere, such as EGI and EMI. We are constructing a group web/wiki which will contain useful information for sites on the IPv6 readiness of the various software components. This includes advice on IPv6 configuration and deployment issues for sites (https://w3.hepix.org/ipv6-bis/doku.php?id=ipv6:siteconfig).

This paper will describe the work done by the HEPiX IPv6 working group since CHEP2012. This will include detailed reports on the testing of various WLCG services on IPv6 including data management, data transfer, workload management and system/network monitoring. It will also present the up to date list of those applications and services which function correctly in a dual-stack environment together with those that still have open issues. The plan for more testing on the production infrastructure with a dual-stack IPv4/IPv6 setup and the work required before the support of IPv6-only clients is realised will be described.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / 436

The Fabric for Frontier Experiments Project at Fermilab

Author: Michael Kirby[1]

Co-author: Adam Lyon [2]

[1] *Fermi National Accelerator Laboratory*

[2] *Fermilab*

Corresponding Author: kirby@fnal.gov

The Fabric for Frontier Experiments (FIFE) project is a new far-reaching, major-impact initiative within the Fermilab Scientific Computing Division to drive the future of computing services for Fermilab Experiments. It is a collaborative effort between computing professionals and experiment scientists to produce an end-to-end, fully integrated set of services for computing on the grid and clouds, managing data, accessing databases, and collaborating within experiments. FIFE includes 1) easy to use job submission services for processing physics tasks on the Open Science Grid and elsewhere; 2) an extensive data management system for managing local and remote caches, cataloging, querying, moving, and tracking the use of data; 3) custom and generic database applications for calibrations, beam information, and other purposes; 4) collaboration tools including an electronic log book, speakers bureau database, and experiment membership database. All of these aspects will be discussed in detail. FIFE sets the direction of computing at Fermilab experiments now and in the future, and therefore is a major driver in the design of computing services world wide.

Poster presentations / **442**

Data Preservation at the D0 Experiment

Author: Kenneth Richard Herner[1]

Co-author: Michael Kirby [2]

[1] *Fermi National Accelerator Laboratory (US)*

[2] *Fermi National Accelerator Laboratory*

Corresponding Authors: kirby@fnal.gov, kenneth.richard.herner@cern.ch

The Tevatron experiments have entered their post-data-taking phases but are still producing physics output at a high rate.
The D0 experiment has initiated efforts to preserve both data access and full analysis capability for the collaboration members through at least 2020. 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.

D0 is making a number of changes to its computing infrastructure to retain analysis capability and maintain long-term data access. These changes include transitioning to newer versions of data handling tools, virtualizing database servers and batch worker nodes, modifying job submission scripts, migrating to new data storage technology, and developing tools for automated validation of physics software on future OS platforms. We will present a talk describing the benefits of long-term data preservation, the status of changes to the D0 computing infrastructure, the challenges of data preservation efforts, and plans for the next several years within the D0 collaboration at Fermilab.

Event Processing, Simulation and Analysis / **340**

FLES: First Level Event Selection Package for the CBM Experiment

Author: Ivan Kisel[1]

Co-authors: Igor Kulakov [2]; Maksym Zyzak [2]; Valentina Akishina [3]

[1] *GSI, Gesellschaft fuer Schwerionenforschung mbH*

[2] *Uni-Frankfurt, FIAS*

[3] *Uni-Frankfurt*

Corresponding Author: i.kisel@gsi.de

The CBM (Compressed Baryonic Matter) experiment is an experiment being prepared to operate at the future Facility for Anti-Proton and Ion Research (FAIR, Darmstadt, Germany). Its main focus is the measurement of very rare probes, which requires interaction rates of up to 10 MHz. Together with the high multiplicity of charged tracks produced in heavy-ion collisions, this leads to huge data rates of up to 1 TB/s. Most trigger signatures are complex (short-lived particles, e.g. open charm decays) and require information from several detector sub-systems.

First Level Event Selection (FLES) in the CBM experiment will be performed on-line on a dedicated processor farm. This requires the development of fast and precise reconstruction algorithms suitable for on-line data processing. The algorithms have to be intrinsically local and parallel and thus require a fundamental redesign of traditional approaches to event data processing in order to use the full potential of modern many-core CPU/GPU architectures. Massive hardware parallelization has to be reflected in mathematical and computational optimization of the algorithms.

The Cellular Automaton (CA) algorithm is used for track reconstruction. The CA algorithm creates short track segments (triplets) in each three neighboring stations, then links them into track-candidates and selects them according to the maximum length and minimum chiˆ2 criteria. The algorithm is optimized with respect to time, vectorized, fully implemented in single precision and robust with respect to the detector geometry and inefficiency. Reconstruction of minimum-bias heavy-ion collisions shows 98% efficiency for most of signal particles and speed of 11 ms per event per core. The Kalman filter (KF) based track fit is used for precise estimation of track parameters.

The KFParticle package for short-lived particles reconstruction, based on the Kalman filter, has rich functionality: the complete particle reconstruction with momentum and covariance matrix calculation; reconstruction of decay chains; daughter particles can be added one by one; simple access to parameters of the particle, such as mass, lifetime, decay length, rapidity, and their errors; transport of the particle; estimation of the distance between particles etc. The KFParticle package has been also vectorized using the SIMD instructions set.

An overview of the on-line FLES processor farm concept, different levels of parallel data processing in the farm from the supervisor down to the multi-threading and the SIMD vectorization, implementation of the algorithms in single precision, memory optimization, scalability on up to 80 CPU cores, efficiency, precision and speed of the FLES algorithms with respect to track multiplicity are presented and discussed.

Poster presentations / **409**

Geant4 Based Simulations for Novel Neutron Detector Development

Author: Thomas Kittelmann[1]

Co-authors: Irina Stefanescu [2]; Kalliopi Kanaki [1]; Karl Zeitelhack [2]; Mirko Boin [3]; Richard Hall-Wilton [1]

[1] *European Spallation Source ESS AB*

[2] *FRM2, Technische Universität München*

[3] *Helmholtz Zentrum Berlin*

The construction of the European Spallation Source ESS AB, which will become the worlds most powerful source of cold and thermal neutrons (meV scale), is about to begin in Lund, Sweden, breaking ground in 2014 and coming online towards the end of the decade. Currently 22 neutron-scattering instruments are planned as the baseline suite at the facility, and a crucial part of each such beam-line will be the detector at which neutrons are detected after undergoing scattering in a given sample under study. Historically, the technological choices for neutron detection at thermal energies have been Helium-3 based, a gas which in recent years has become unavailable for all but the smallest of detectors, due to the effect of a rapidly dwindling worldwide supply at the same time the demand is increasing heavily. This makes novel neutron detectors a critical technology for ESS to develop,

and also neutron detection itself presently is a hot topic to a range of disciplines and industrial applications.

Thus, an extensive international R&D programme is currently underway at ESS and in European partner institutes, and worldwide (under the auspices of the International Collaboration for the Development of Neutron Detectors, icnd.org) in order to develop efficient and cost-effective detectors based on alternative isotopes such as Boron-10 based thin-film detectors or Lithium-6 doped scintillators.

In this contribution we present the Geant4-based python/C++ simulation and coding framework, which has been developed and used within the ESS Detector Group and in collaboration with FRM-II, in order to aid in these R&D efforts. We show specific examples of results from investigations of specific proposed detector designs, and discuss the extensions to Geant4 which have been implemented in order to include the (at this energy scale) very significant effects of low-energy phenomena such as coherent scattering (Bragg diffraction) in the polycrystalline support materials of the detector. We also present a custom object oriented output file format with meta-data, GRIFF ("Geant4 Results In Friendly Format"), developed in order to facilitate a faster turn-around time when analysing simulation results by enabling high-level whole-event analysis in addition to the usual benefits of a persistified output format (such as multi-processing and fast re-analysis of the same data).

Whilst these simulations have been implemented specifically for neutron detectors, it has potential for wider applications in neutron scattering, and in other disciplines.

Plenaries / **510**

The KPMG Challenge

Corresponding Author: sander.klous@gmail.com

Plenaries / **515**

The KPMG Challenge: the Awarding

Corresponding Author: sander.klous@gmail.com

Poster presentations / **168**

HS06 benchmark values for an ARM based server

Author: Stefan Kluth[1]

[1] *Max-Planck-Institut fuer Physik (Werner-Heisenberg-Institut) (D*

Corresponding Author: skluth@mpp.mpg.de

We benchmarked an ARM Cortex A9 based server system with a four-core CPU running at 1.1 GHz. The system used Ubuntu 12.04 as operating system and the hepspec 2006 (HS06) benchmarking suite was compiled natively with gcc-4.4 on the system. The benchmark was run for various settings of the relevant gcc compiler options. We did not find significant influence from the compiler options on the benchmark result. The final HS06 benchmark result is 10.4.

Software Engineering, Parallelism & Multi-Core / 353

Monte Carlo Simulations of the IceCube Detector with GPUs

Authors: Claudio Kopper[None]; David Schultz[1]; Dmitry Chirkin[1]; Juan Carlos Diaz Velez[1]

[1] *University of Wisconsin-Madison*

Corresponding Authors: claudio.kopper@icecube.wisc.edu, juancarlos@icecube.wisc.edu

The IceCube Neutrino Observatory is a cubic kilometer-scale neutrino detector built into the ice sheet at the geographic South Pole. Light propagation in glacial ice is an important component of IceCube detector simulation that requires a large number of embarrassingly parallel calculations. The IceCube collaboration recently began using GPUs in order to simulate direct propagation of Cherenkov photons in the antarctic ice as part of our detector simulation. GPU computing is now being utilized in large scale Monte Carlo productions involving computing centers distributed across the world. We discuss practical issues of our implementation involving mixed CPU and GPU resources in the simulation chain and our efforts to optimize the utilization of these resources in a grid environment.

Poster presentations / 356

Enabling IPv6 at FZU - WLCG Tier2 in Prague

Authors: Jiri Chudoba[1]; Marek Elias[1]; Tomas Kouba[1]

[1] *Acad. of Sciences of the Czech Rep. (CZ)*

Corresponding Author: koubat@fzu.cz

The production usage of the new IPv6 protocol is becoming reality in the HEP community and the Computing Centre of the Institute of Physics in Prague participates in many IPv6 related activities. Our contribution will present experience with monitoring in HEPiX
distributed IPv6 testbed which includes 11 remote sites. We use Nagios
to check availability of services and Smokeping for monitoring the network latency. It is not always trivial to setup DNS in a dual stack
environment properly therefore we developed a Nagios plugin for checking
whether a domain name is resolvable when using only IP protocol version
6 and only version 4. We will also present local area network monitoring
and tuning related to IPv6 performance. One of the most important software for a grid site is a batch system for job execution. We will present our experience with configuring and running Torque and Slurm batch systems in dual stack environment. And we will discuss the steps needed to run VO specific jobs in our IPv6 testbed.

Poster presentations / 407

Performance of most popular open source databases for HEP related computing problems

Authors: Dmytro Kovalskyi[1]; Frank Wuerthwein[2]; Igor Sfiligoi[3]

[1] *Univ. of California Santa Barbara (US)*

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

[3] *University of California San Diego*

Corresponding Author: dmytro.kovalskyi@cern.ch

Databases are used in many software components of the HEP computing, from monitoring and task scheduling to data storage and processing. While the database design choices have a major impact on the system performance, some solutions give better results out of the box than the others. This paper presents detailed comparison benchmarks of the most popular Open Source systems for a typical class of problems relevant to HEP computing.

Plenaries / **490**

Data processing in the wake of massive multicore and GPU

Author: Jim Kowalkowski[1]

[1] *Fermilab*

Corresponding Author: jbk@fnal.gov

Developments in concurrency (massive multi-core, GPU, and architectures such as ARM) are changing the physics computing landscape. In this talk dr Jim kowalkowski of Fermilab will describe on the use of GPU, massive multi-core, and the changes that result from massive parallelization and how this impacts data processing and models.

Data Acquisition, Trigger and Controls / **466**

The artdaq Data Acquisition Software Toolkit

Authors: Jim Kowalkowski[1]; Kurt Biery[2]

Co-authors: Christopher Green [3]; Marc Paterno [1]; Stephen Foulkes [4]

[1] *Fermilab*

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

[3] *Department of Physics*

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

Corresponding Authors: jbk@fnal.gov, biery@fnal.gov

The artdaq data acquisition software toolkit has been developed within the Fermilab Scientific Computing Division to meet the needs of current and future experiments. At its core, the toolkit provides data transfer, event building, and event analysis functionality, the latter using the art event analysis framework.

In the last year, functionality has been added to the toolkit in the areas of state behavior, process control, data quality monitoring, and system monitoring with the goal of providing a complete DAQ software toolkit for future experiments. In addition, the toolkit has been used in the construction of the DAQ system of the DarkSide-50 dark matter experiment and the prototype DAQ for the Mu2e rare decay experiment.

We will present the design and features of the toolkit, the advantages of using the toolkit to construct the DAQ software for an experiment, representative performance results, and future plans.

Software Engineering, Parallelism & Multi-Core / **461**

Improving robustness and computational efficiency using modern C++ (video conference)

Author: Marc Paterno[1]

Co-authors: Christopher Green [1]; Jim Kowalkowski [1]

[1] *Fermilab*

Corresponding Authors: jbk@fnal.gov, paterno@fnal.gov

For nearly two decades, the C++ programming language has been the dominant programming language for experimental HEP. The publication of ISO/IEC 14882:2011, the current version of the international standard for the C++ programming language, makes available a variety of language and library facilities for improving the robustness, expressiveness, and computational efficiency of C++ code. However, much of the C++ written by the experimental HEP community does not take advantage of the features of the language to obtain these benefits, either due to lack of familiarity with these features or concern that these features must somehow be computationally inefficient.

In this paper, we address some of the features of modern C++, and show how they can be used to make programs that are both robust and computationally efficient. We compare and contrast simple yet realistic examples of some common implementation patterns in C, currently-typical C++, and modern C++, and show (when necessary, down to the level of generated assembly language code) the quality of the executable code produced by recent C++ compilers, with the aim of allowing the HEP community to make informed decisions on the costs and benefits of the use of modern C++.

Poster presentations / 206

Compute Farm Software for ATLAS IBL Calibration

Authors: Andreas Kugel[1]; Moritz Kretz[1]

Co-authors: Joern Grosse-Knetter [2]; Karolos Potamianos [3]; Marcello Bindi [4]; Paolo Morettini [5]; Timon Heim [6]; Tobias Flick [6]; Yosuke Takubo [7]

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

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

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

[4] *University of Bologna and INFN (IT)*

[5] *INFN Genova*

[6] *Bergische Universitaet Wuppertal (DE)*

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

Corresponding Author: moritz.kretz@cern.ch

In 2014 the Insertable B-Layer (IBL) will extend the existing Pixel Detector of the ATLAS experiment at CERN by 12 million additional pixels. As with the already existing pixel layers, scanning and tuning procedures need to be employed for the IBL to account for aging effects and guarantee a unified response across the detector. Scanning the threshold or time-over-threshold of a front-end module consists of two main steps: gathering histograms of the detector response for different configurations and then fitting a target function to these histograms. Despite of the currently used method of performing the computationally demanding fits on DSPs located on the read-out hardware, it was

decided to abandon this approach for IBL and realize the functionality on an external computing farm for easier development and greater flexibility.

This not only requires the fast transfer of histogram data from the read-out hardware to the computing farm via Ethernet, but also the integration of the fit farm software and hardware into the already existing data-acquisition and calibration framework (TDAQ and PixelDAQ) of the ATLAS experiment and the current Pixel Detector.
It is foreseen to implement the software running on the compute cluster with an emphasis on modularity, allowing for flexible adjustment of the infrastructure and a good scalability with respect to the number of network interfaces, available CPU cores, and deployed machines. By using a modular design we are able to not only employ CPU based fitting algorithms, but also have the possibility to take advantage of the performance offered by a GPU-based approach to fitting.

We present the compute farm software architecture and report on the status of the implementation of the IBL calibration architecture into the ATLAS hardware and software framework. We discuss used test methods and point out obstacles that were encountered along the way.

Facilities, Infrastructures, Networking and Collaborative Tools / 107

Opportunistic Resource Usage in CMS

Authors: Dirk Hufnagel[1]; Peter Kreuzer[2]

Co-author: Ian Fisk [1]

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

[2] *Rheinisch-Westfaelische Tech. Hoch. (DE)*

Corresponding Authors: peter.kreuzer@cern.ch, dirk.hufnagel@cern.ch, ian.fisk@cern.ch

CMS is using a tiered setup of dedicated computing resources provided by sites distributed over the world and organized in WLCG. These sites pledge resources to CMS and are preparing them specially for CMS to run the experiment's applications. But there are more resources available opportunistically both on the GRID and in local university and research clusters which can be used for CMS applications. We will present CMS' strategy to use opportunistic resources and prepare them dynamically to run CMS applications. CMS is able to run it's applications on resources that can be reached through the GRID, through EC2 compliant cloud interfaces. Even resources that can be used through ssh login nodes can be harnessed. All of these usage modes are integrated transparently into the glideIn WMS submission infrastructure which is the basis of CMS' opportunistic resource usage strategy. Technologies like Parrot to mount the software distribution via CVMFS and xrootd for access to data and simulation samples via the WAN are used and will be described. We will summarize the experience with opportunistic resource usage and give an outlook for the restart of LHC data taking in 2015.

Poster presentations / 22

Transactional Aware Tape Infrastructure Monitoring System

Author: Fotios Nikolaidis[1]

Co-authors: Daniele Francesco Kruse [2]; German Cancio Melia [2]

[1] *University of Crete (GR)*

[2] *CERN*

Corresponding Authors: daniele.francesco.kruse@cern.ch, fotios.nikolaidis@cern.ch, german.cancio.melia@cern.ch

Administrating a large-scale, multi-protocol, hierarchical tape storage infrastructure like the one at CERN, which stores around 30PB / year, requires an adequate monitoring system for quick spotting of malfunctions, easier debugging and on demand report generation. The main challenges for such system are: to cope with log format diversity and its information scattered among several log files, the need for long term information archival, the strict data consistency requirements and the group based GUI visualization. For this purpose, we have designed, developed and deployed a centralized system consisting of four independent layers: a Log Transfer layer for collecting log lines from all tape servers to a single aggregation server, a Data Mining layer for combining log data into transactional context, a Storage layer for archiving the resulting transactions and finally a Web UI layer for accessing the information. Having flexibility, extensibility and maintainability in mind, each layer is designed to work as a message broker for the next layer, providing a clean and generic interface while ensuring consistency, redundancy and ultimately fault tolerance. This system unifies information previously dispersed over several monitoring tools into a single user interface, using Splunk, which also allows us to provide information visualization based on access control lists (ACL). Since its deployment, it has been successfully used by CERN tape operators for quick overview of transactions, performance evaluation, malfunction detection and by managers for report generation. In this paper we present our design principles, problems with corresponding solutions, disaster cases and how we handle them, comparison with other solutions and future work that can be done.

Poster presentations / **9**

The Repack Challenge

Author: Daniele Francesco Kruse[1]

[1] *CERN*

Corresponding Author: daniele.francesco.kruse@cern.ch

Physics data stored in CERN tapes is quickly reaching the 100 PB milestone. Tape is an ever-changing technology that is still following Moore's law in terms of capacity. This means we can store every year more and more data in the same amount of tapes. However this doesn't come for free: the first obvious cost is the new higher capacity media. The second less known cost is related to moving the data from the old tapes to the new ones. This activity is what we call repack. Repack is vital for any large tape user: without it, one would have to buy more tape libraries and more floor space and, eventually, data on old non supported tapes would become unreadable and be lost forever. The challenge is not an easy one. First, to make sure we won't need any more tape slots in the near future, we will have to repack 120 PB from 2014 to 2015, this in turn means that we will have to be able to cope with peaks of 3.5 GB/s smoothly. Secondly, all the repack activities will have to run concurrently and in harmony with the existing experiment tape activities. Making sure that this works out seamlessly implies careful planning of the resources and the various policies for sharing them fairly and conveniently. Our previous setup allowed for an average repack performance of only 360 MB/s. Our needs demand this figure increase tenfold by 2013. To tackle this problem we needed to fully exploit the speed and throughput of our modern tape drives. This involved careful dimensioning and configuration of the disk arrays (the middle step between an old source tape and a new higher capacity destination tape) and all the links between them and the tape servers (the machines responsible for managing the tape drives). We also planned a precise schedule and provided a visual monitoring tool to check the progress over time. The new repack setup we deployed brought an average 80% increase in the throughput of tape drives, allowing them to perform closer to their design specifications. This improvement in turn meant a 40% decrease in the number of drives needed to achieve the 3.5 GB/s goal. CERN is facing its largest data migration challenge yet. By restructuring the repack infrastructure we allowed the vital repack and LHC experiments activities to coexist without the need for new expensive tape drives.

Poster presentations / **477**

System level traffic shaping in diskservers with heterogeneous

protocols

Author: Eric Cano[1]

Co-author: Daniele Francesco Kruse [1]

[1] *CERN*

Corresponding Authors: daniele.francesco.kruse@cern.ch, eric.cano@cern.ch

Disk access and tape migrations compete for network bandwidth in CASTOR's disk servers, over various protocols: RFIO, Xroot, root and GridFTP. As there are a limited number of tape drives, it is important be keep them busy all the time, at their nominal speed. With potentially 100s of user read streams per server, the bandwidth for the tape migrations has to be guaranteed to a controlled level, and not the default fair share the system gives by default. Xroot provides a prioritization mechanism , but using it implies moving exclusively to the Xroot protocol, which is not possible in short to mid-term time frame, as users are equally using all. The greatest commonality of all those protocols is not more that usage of TCP/IP. We investigated the Linux kernel traffic shaper to control TCP/IP bandwidth. The performance and limitations of the traffic shaper have been understood in test environment, and satisfactory working point has been found for production. Notably, TCP offload engines' negative impact on traffic shaping, and the limitations of the length of the traffic shaping rules were discovered and measured. A suitable working point has been found and the traffic shaping is now successfully deployed in the CASTOR production systems at CERN. This system level approach could be transposed easily to other environments.

Event Processing, Simulation and Analysis / 184

The Role of Effective Event Reconstruction in the Higgs Boson Discovery at CMS

Author: Slava Krutelyov[1]

[1] *Texas A & M University (US)*

Corresponding Author: vyacheslav.krutelyov@cern.ch

In 2012 the LHC increased both the beam energy and intensity. The former made obsolete all of the simulation data generated for 2011; the latter increased the rate of multiple proton-proton collisions (piluep) in a single event, significantly increasing the complexity of both the reconstructed and matching simulated events. Once the pileup surpassed 10, the resources needed for the software to function created significant strain on CMS computing facilities. Problems with increasing memory and CPU use had to be alleviated in a way that did not sacrifice the physics performance of the reconstruction. In 2012 this was particularly important as the prompt calibration system was fully commissioned, making the data produced in the prompt reconstruction, the primary datasets used in 2012 physics publications on 8Tev data. This paper summarizes the changes applied to the CMS data reconstruction software, which was deployed successfully and delivered high quality data used in the Higgs boson discovery and many other physics results from CMS.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / 182

First Production with the Belle II Distributed Computing System

Authors: Hideki Miyake[1]; Martin Sevior[2]; Takanori HARA[1]; Thomas Kuhr[3]

[1] *KEK*

[2] *University of Melbourne (AU)*

[3] *KIT*

The Belle II experiment, a next-generation B factory experiment at KEK, is expected to record a two orders of magnitude larger data volume than its predecessor, the Belle experiment. The data size and rate are comparable to or more than the ones of LHC experiments and requires to change the computing model from the Belle way, where basically all computing resources were provided by KEK, to a more distributed scheme. We exploit existing grid technologies, like DIRAC for the management of jobs and AMGA for the metadata catalog, to build a distributed computing infrastructure for the Belle II experiment. The system provides an abstraction layer for collections of jobs, called a project, and collections of files in a dataset. This year we could demonstrate for the first time the viability of our system in a generation, simulation, and reconstruction of 60M events on several grid sites. The results of this Monte-Carlo production campaign and the further plans for the distributed computing system of the Belle II experiment are presented in this talk.

Poster presentations / **456**

Simulation of the PANDA Lambda disks

Author: Ajay Kumar[1]

Co-author: Ankhi Roy [1]

[1] *Indian Institute of Technology Indore*

Corresponding Author: ajayk@iiti.ac.in

Ajay Kumar and Ankhi Roy
For the PANDA collaboration
Indian Institute of Technology Indore, Indore-4520017, India
Email- ajayk@iiti.ac.in

The PANDA experiment is one of the main experiments at the future accelerator facility FAIR which is currently under construction in Darmstadt, Germany. Experiments will be performed with intense, phase space cooled antiproton beams incident on a hydrogen or heavy nuclear target. The main physics motivation of PANDA is to explore the non-perturbative regime of QCD and to study hadronic states. In this context, here is a possibility to include hyperon studies in the PANDA physics program. Hyperons travel a large distance before they decay into other particles. In order to increase the acceptance to measure these particles, there is a concept to include an additional so-called "Lambda Disk"detector.

The Micro Vertex Detector (MVD) is the innermost tracking detector of PANDA. It consists of four barrel layers and six forward disk layers. It is made up of two types of silicon sensors –silicon hybrid pixels and double sided silicon strips. The last layer of the MVD forward disk is situated at 23 cm downstream of the interaction point and the first layer of GEM tracking station is located 110 cm downstream from the interaction point. Hence, there is a large region without tracking information. Therefore, it is proposed to place two additional disks known as the Lambda disks in this region. One layer is at 40 cm and the other is at 60 cm downstream from the interaction point. The detector will enhance the reconstruction probability for hyperons. As a starting geometry, it has been proposed for the Lambda disks to be made up of only double-sided silicon strip sensor. In this conceptual design, the outer ring has been kept similar to the outermost layers of the MVD forward disks and inner layer of Lambda disks has been designed using silicon strip sensor but of different size.

At present, we are involved in simulation studies of the Lambda disks detector with proton antiproton to lambda anti-lambda to calculate reconstruction efficiency and resolution of this channel. This channel provides essential inputs in understanding the vertex reconstruction of hyperon pairs. We have also started to study different parameters related to the development of the Lambda disks detector.

In this presentation we will report about the reconstruction efficiency and identification probability of Lambda and Anti-Lambda particles with and without the Lambda disks detector. In addition, simulation study of detector coverage, material budget, radiation damage and rate estimation with the Lambda disks detector which are essential for the development of the detector will be presented.

Poster presentations / 205

Tier-1 experience with provisioning virtualized worker nodes on demand

Authors: Andrew David Lahiff[1]; Ian Collier[2]

[1] *STFC - Science & Technology Facilities Council (GB)*

[2] *UK Tier1 Centre*

Corresponding Author: andrew.david.lahiff@cern.ch

While migration from the grid to the cloud has been gaining increasing momentum in recent times, WLCG sites are currently still expected to accept grid job submission, and this is likely to continue for the foreseeable future. Furthermore, sites which support multiple experiments may need to provide both cloud and grid-based access to resources for some time, as not all experiments may be ready to move to the cloud at the same time. In order to make optimal use of resources, a site with a traditional batch system as well as a cloud resource may want to make opportunistic use of their cloud at times when there are idle jobs and all worker nodes in the batch system are busy, by extending the batch system into the cloud. We present two implementations, one based on HTCondor and the other based on SLURM as a batch system, in which virtualized worker nodes are provisioned on demand using a StratusLab cloud.

Event Processing, Simulation and Analysis / 286

Preparing the Track Reconstruction in ATLAS for a high multiplicity future

Author: Robert Johannes Langenberg[1]

Co-authors: Andreas Salzburger [2]; Anthony Morley [3]; Markus Elsing [2]; Niels Van Eldik [2]; Ruslan Mashinistov [4]

[1] *Technische Universitaet Muenchen (DE)*

[2] *CERN*

[3] *KTH Royal Institute of Technology (SE)*

[4] *Russian Academy of Sciences (RU)*

Corresponding Authors: robert.langenberg@cern.ch, anthony.morley@cern.ch, andreas.salzburger@cern.ch, markus.elsing@cern.ch, niels.van.eldik@cern.ch, ruslan.mashinistov@cern.ch

The track reconstruction algorithms of the ATLAS experiment have demonstrated excellent performance in all of the data delivered so far by the LHC. The expected large increase in the number of interactions per bunch crossing in the future introduce new challenges both in the computational aspects and physics performance of the algorithms. With the aim of taking advantage of modern CPU design and optimising memory and CPU usage in the reconstruction algorithms a number of projects are being pursued. These include rationalisation of the event data model, vectorisation of the core components of the algorithms, and removing algorithm bottlenecks by using modern code analysis tools. We will discuss the technologies and techniques that are being investigated to improve the track reconstruction at ATLAS to make the best use developments of computing technology and to handle the expected increase in event multiplicity.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / 60

Virtualised data production infrastructure for NA61/SHINE based on CernVM

Author: Dag Larsen[1]

[1] *University of Silesia (PL)*

Corresponding Author: dag.larsen@cern.ch

Currently, the NA61/SHINE data production is performed on the CERN shared batch system, an approach inherited from its predecessor NA49. New data productions are initiated by manually submitting jobs to the batch system. An effort is now under way to migrate the data production to an automatic system, on top of a fully virtualised platform based on CernVM. There are several motivations for this. From a practical perspective, it will make it easier to both initiate new data productions, and to utilise computing resources available outside CERN, e.g. from institutes participating in NA61/SHINE, as well as commercial clouds. In addition, there is a data preservation perspective.

CernVM is a Linux distribution created by CERN specifically for the needs of virtual machines. It features a small images size (˜300MB), and comes in flavours tailored to specific use-cases, e.g. developers' desktop and batch systems. It supports all popular hypervisors and clouds. Additional software components beyond the basic system can be easily installed via a package manager either through automatic boot-time contextualisation or by hand. Data production software, calibration data and other experiment-specific files are easily distributed globally via the HTTP-based CernVM file system. CernVM images are built from a software database using a special build system. This will allow legacy versions of CernVM to be adapted and built for future hypervisors and other technology. This will make it possible to run legacy versions of the data production software on the legacy CernVM versions they were originally running on, but on modern hardware, thus enabling a new approach to data preservation. An additional CernVM service, CernVM On-line, is a tool to set up virtual clusters seamlessly spanning multiple physical clouds, thus allowing for cloud-based distributed computing.

The effort to create a virtualised data production infrastructure for NA61/SHINE builds on top of these tools. The NA61/SHINE data production software has been adapted to run under CernVM through CernVM file system. Both raw and reconstructed data is stored on CERN Castor, and is accessed through Xrootd. Databases are used to keep track of the data. A web-based, graphical user interface for the data production will be available. This will allow the system to present lists of both raw data as well as existing data productions. If a new data production is needed, the privileged user may choose the data, software versions, and calibrations to be used. At the start of the processing, the system will automatically create a virtual cluster matching the requirements, and at the end of the processing, it will be terminated. This is needed to free up the resources for other users. CernVM-online will be used for the management of the virtual clusters. In case a legacy version of the software is selected, a matching version of CernVM will be used. Finished jobs will be scanned for errors, and automatically resubmitted for processing if needed. Finally, the relevant databases will be updated to reflect the freshly produced data.

Poster presentations / **134**

Implementing long-term data preservation and open access in CMS

Authors: Kati Lassila-Perini[1]; Mike Hildreth[2]

[1] *Helsinki Institute of Physics (FI)*

[2] *Department of Physics-College of Science-University of Notre Da*

Corresponding Author: katri.lassila-perini@cern.ch

Implementation of the CMS policy on long-term data preservation, re-use and open access has started. Current practices in providing data additional to published papers and distributing simplified data-samples for outreach are promoted and consolidated. The first measures have been taken for the analysis and data preservation for the internal use of the collaboration and for the open access to part of the data. Two complementary approaches are followed. First, a virtual machine environment, which will pack all ingredients needed to compile and run a software release with which the legacy data was reconstructed. Second, a validation framework, maintaining the capability not only to read the old raw data, but also to reconstruct them with more recent releases to guarantee long-term reusability of the legacy data.

Software Engineering, Parallelism & Multi-Core / **242**

The ATLAS Data Management Software Engineering Process

Author: Mario Lassnig[1]

Co-authors: Angelos Molfetas [2]; Armin Nairz [1]; Cedric Serfon [1]; Graeme Andrew Stewart [1]; Luc Goossens [1]; Martin Barisits [1]; Ralph Vigne [3]; Thomas Beermann [4]; Vincent Garonne [1]

[1] *CERN*

[2] *University of Sydney (AU)*

[3] *University of Vienna (AT)*

[4] *Bergische Universitaet Wuppertal (DE)*

Corresponding Authors: mario.lassnig@cern.ch, vincent.garonne@cern.ch, graeme.andrew.stewart@cern.ch, martin.barisits@cern.ch, thomas.beermann@cern.ch, ralph.vigne@cern.ch, cedric.serfon@cern.ch, luc.goossens@cern.ch, armin.nairz@cern.ch, angelos.molfetas@cern.ch

Rucio is the next-generation data management system supporting ATLAS physics workflows in the coming decade. The software engineering process to develop Rucio is fundamentally different to existing software development approaches in the ATLAS distributed computing community. Based on a conceptual design document, development takes place using peer-reviewed code in a test-driven environment, where nothing enters production if it has not been approved by at least one other engineer. The main objectives are to ensure that every engineer understands the details of the full project, even components usually not touched by them, that the design and architecture are coherent, that temporary contributors can be productive without delay, that programming mistakes are prevented before being committed to the source code, and that the source is always in a fully functioning state. This contribution will illustrate the workflows and products used, and demonstrate the typical development cycle of a component from inception to deployment within this software engineering process. Next to the technological advantages, this contribution will also highlight the social aspects and implications of an environment where every action by an engineer is subject to scrutiny from colleagues.

Poster presentations / **272**

ATLAS DDM Workload Emulation

Author: Ralph Vigne[1]

Co-authors: Armin Nairz [2]; Cedric Serfon [2]; Erich Schikuta [3]; Graeme Andrew Stewart [2]; Luc Goossens [2]; Mario Lassnig [2]; Martin Barisits [2]; Thomas Beermann [4]; Vincent Garonne [2]

[1] *University of Vienna (AT)*

[2] *CERN*

[3] *University of Vienna*

[4] *Bergische Universitaet Wuppertal (DE)*

Corresponding Authors: mario.lassnig@cern.ch, ralph.vigne@cern.ch, vincent.garonne@cern.ch, graeme.andrew.stewart@cern.ch, martin.barisits@cern.ch, thomas.beermann@cern.ch, cedric.serfon@cern.ch, luc.goossens@cern.ch, armin.nairz@cern.ch

Rucio is the successor of the current Don Quijote 2 (DQ2) system for the distributed data management (DDM) system of the ATLAS experiment. The reasons for replacing DQ2 are manifold, but besides high maintenance costs and architectural limitations, scalability concerns are on top of the list.

The data collected so far by the experiment adds up to about 115 Peta bytes spread over 270 million distinct files. Current expectations are that the amount of data will be three to four times as it is today by the end of 2014. Further is the expansion of the WLCG computing resources pushing additional pressure on the DDM system by adding more powerful computing resources, which subsequently increases the demands on data provisioning. Although DQ2 is capable of handling the current workload, it is already at its limits. To ensure that Rucio will be up to the expected quality of service, a way to emulate the expected workload is needed. To do so, first the current workload observed in DQ2 must be understood in order to scale it up to future expectations.

This paper presents an overview of the theory behind workload emulation and discusses how selected core concepts are applied to the workload of the experiment. Further a detailed discussion is provided on how knowledge about the current workload is derived from central file catalogue logs, PanDA dashboard, etc. The discussion also addresses how this knowledge is utilized in the context of the emulation framework. Finally a description of the implemented emulation framework used for stress-testing Rucio is given.

Poster presentations / **243**

The DMLite Rucio Plugin: ATLAS data in a filesystem

Author: Mario Lassnig[1]

Co-authors: Alejandro Alvarez Ayllon [1]; Daan Van Dongen [2]; Philippe Calfayan [3]; Ricardo Brito Da Rocha [1]

[1] *CERN*

[2] *C*

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

Corresponding Authors: mario.lassnig@cern.ch, philippe.calfayan@cern.ch, rocha@cern.ch, alejandro.alvarez.ayllon@cern.ch, daan.van.dongen@cern.ch

Rucio is the next-generation data management system supporting ATLAS physics workflows in the coming decade. Historically, clients interacted with the data management system via specialised tools, but in Rucio additional methods are provided. To support filesystem-like interaction with all ATLAS data a plugin to the DMLite software stack has been developed. It is possible to mount Rucio as a filesystem, and execute regular filesystem operations in a POSIX fashion. This is exposed via various protocols like HTTP/WebDAV or NFS, which then removes any dependency on Rucio for client software. The main challenge for this work is the mapping of the set-like ATLAS namespace into a hierarchical filesystem, whilst preserving the high performance features of the former. This includes listing and searching for data, creation of files, datasets and containers, and the aggregation of existing data - all within POSIX directories with potentially millions of entries. This contribution will detail the design and implementation of the plugin, and demonstrate how physicist can interact with ATLAS data via commonly available and standard tools. Furthermore, an evaluation of the performance characteristics is given, to show that this approach can scale to the requirements of ATLAS physics analysis.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / **394**

User Centric Job Monitoring –a redesign and novel approach in the STAR experiment

Authors: Dmitry Arkhipkin[1]; Jerome LAURET[2]; Yulia Zoulkarneeva[3]

[1] *Brookhaven National Laboratory*

[2] *BROOKHAVEN NATIONAL LABORATORY*

[3] *None*

Corresponding Author: jlauret@bnl.gov

User Centric Monitoring (or UCM) has been a long awaited feature in STAR, whereas programs, workflows and system "events"could be logged, broadcast and later analyzed. UCM allows to collect and filter available job monitoring information from various resources and present it to users in a user-centric view rather than and administrative-centric point of view. The first attempt and implementation of "a"UCM approach was made in STAR 2004 using a log4cxx plug-in back-end and then further evolved with an attempt to push toward a scalable database back-end (2006) and finally using a Web-Service approach (2010, CSW4DB SBIR). The latest showed to be incomplete and not addressing the general (evolving) needs of the experiment where streamlined messages for online (data acquisition) purposes as well as the continuous support for the data mining needs and event analysis need to coexists and unified in a seamless approach. The code also revealed to be hardly maintainable.
This work will present the next evolutionary step of the UCM toolkit, a redesign and redirection of our latest attempt acknowledging and integrating recent technologies and a simpler, maintainable and yet scalable manner. The extended version of the job logging package is built upon three-tier approach based on Task, Job and Event, and features a Web-Service based logging API, responsive AJAX-powered user interface, and database back-end relying on MongoDB, which seems to be uniquely suited for STAR needs. In addition, we present details on integration of this logging package with STAR offline and online software frameworks. Leveraging on the reported experience and work from the ATLAS and CMS experience on using the ESPER engine, we will discuss and show how such approach has been implemented in STAR for meta-data event triggering stream processing and filtering. An ESPER based solution seems to fit well into the online data acquisition system where many systems are monitored.

Poster presentations / 509

Experience with Intel's Many Integrated Core Architecture in ATLAS Software

Author: Wim Lavrijsen[1]

Co-authors: Manuel Neumann [2]; Roberto Agostino Vitillo [1]; Sami Kama [3]; Sebastian Fleischmann [2]

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

[2] *Bergische Universitaet Wuppertal (DE)*

[3] *Southern Methodist University (US)*

Corresponding Author: wim.lavrijsen@cern.ch

Intel recently released the first commercial boards of its Many Integrated Core (MIC) Architecture. MIC is Intel's solution for the domain of throughput computing, currently dominated by general purpose programming on graphics processors (GPGPU). MIC allows the use of the more familiar x86 programming model and supports standard technologies such as OpenMP, MPI, and Intel's Threading Building Blocks. This should make it possible to develop for both throughput and latency devices using a single code base. In ATLAS Software, track reconstruction has been shown to be a good candidate for throughput computing on GPGPU devices. In addition, the newly proposed offline parallel event-processing framework, GaudiHive, uses TBB for task scheduling. The MIC is thus, in principle, a good fit for this domain. In this presentation, we report our experiences of porting to and optimizing ATLAS tracking algorithms for the MIC, comparing the programmability and relative cost/performance of the MIC against those of current GPGPUs and latency-optimized CPUs.

Software Engineering, Parallelism & Multi-Core / 61

Parallelization of Common HEP patterns with PyPy (cancelled)

Author: Wim Lavrijsen[1]

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

Corresponding Author: wim.lavrijsen@cern.ch

The Python programming language brings a dynamic, interactive environment to physics analysis. With PyPy high performance can be delivered as well, when making use of its tracing just in time compiler (JIT) and cppyy for C++ bindings, as cppyy is able to exploit common HEP coding patterns. For example, ROOT I/O with cppyy runs at speeds equal to that of optimized, hand-tuned C++.
Python does not, however, offer an easy way to exploit computational parallelization, because of the global interpreter lock (GIL). In PyPy this could be solved using software transactional memory (STM). With STM in place, the patterns in cppyy can be employed to automatically parallelize user code when the interpreter deems them, and the underlying libraries, safe. The work described in this paper takes the existing ROOT I/O patterns in cppyy and shows how they can be parallelized using STM.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / 293

Derived Physics Data Production in ATLAS: Experience with Run 1 and Looking Ahead

Author: Paul James Laycock[1]

Co-authors: Matthew Beckingham [2]; Nurcan Ozturk [3]; Robert Henderson [4]

[1] *University of Liverpool (GB)*

[2] *University of Washington (US)*

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

[4] *Lancaster University (GB)*

Corresponding Authors: paul.james.laycock@cern.ch, matthew.beckingham@cern.ch, robert.henderson@cern.ch, nurcan@uta.edu

While a significant fraction of ATLAS physicists directly analyse the AOD (Analysis Object Data) produced at the CERN Tier 0, a much larger fraction have opted to analyse data in a flat ROOT format. The large scale production of this Derived Physics Data (DPD) format must cater for both detailed performance studies of the ATLAS detector and object reconstruction, as well as higher level and generally lighter-content physics analysis. The delay between data-taking and DPD production allows for software improvements, while the ease of arbitrarily defined skimming/slimming of this format results in an optimally performant format for end-user analysis.

Given the diversity of requirements, there are many flavours of DPDs, which can result in large peak computing resource demands. While the current model has proven to be very flexible for the individual groups and has successfully met the needs of the collaboration, the resource requirements at the end of Run 1 are much larger than planned. In the near future, ATLAS plans to consolidate DPD production, optimising resource usage vs flexibility such that the final analysis format will be more homogeneous across ATLAS while still keeping most of the advantages enjoyed during Run 1.

The ATLAS Run 1 DPD Production Model is presented along with the resource usage statistics at the end of Run 1, followed by an outlook for future plans.

Event Processing, Simulation and Analysis / **327**

Selected event reconstruction algorithms for the CBM experiment at FAIR

Author: Semen Lebedev[1]

Co-authors: Andrey Lebedev [2]; Claudia Hoehne [1]; Gennady Ososkov [3]

[1] *Justus-Liebig-Universitaet Giessen (DE)*

[2] *IKF Frankfurt University / LIT JINR*

[3] *Joint Institute for Nuclear Research, Dubna, Russia*

Corresponding Author: s.lebedev@gsi.de

Development of fast and efficient event reconstruction algorithms is an important and challenging task in the Compressed Baryonic Matter (CBM) experiment at the future FAIR facility. The event reconstruction algorithms have to process terabytes of input data produced in particle collisions. In this contribution, several event reconstruction algorithms, which use available features of modern processors, namely, SIMD execution model, are presented. Optimization and vectorization of the algorithms in the following CBM detectors are discussed: Ring Imaging Cherenkov (RICH) detector, Transition Radiation Detectors (TRD) and Muon Chamber (MUCH). In RICH event reconstruction includes ring finding (based on Hough Transform method), fitting (based on circle or ellipse fit methods) and association of reconstructed rings and tracks. In TRD and MUCH track reconstruction algorithms are based on track following and Kalman Filter methods. All algorithms were significantly optimized to achieve maximum speed up and minimum memory consumption. Obtained results showed that a significant speed up factor for all algorithms was achieved and the reconstruction efficiency stays at high level.

Poster presentations / **326**

Quality Assurance for simulation and reconstruction software in CBMROOT

Authors: Andrey Lebedev[1]; Florian Uhlig[2]; Semen Lebedev[3]

[1] *IKF Frankfurt University / LIT JINR*

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

[3] *Justus-Liebig-Universitaet Giessen (DE)*

Corresponding Author: s.lebedev@gsi.de

The software framework of the CBM experiment at FAIR - CBMROOT - has been continuously growing over the years. The increasing complexity of the framework and number of users require improvements in maintenance, reliability and in overall software development process. In this report we address the problem of the software quality assurance (QA) and testing. Two main problems are considered in our test suit. First, test of the build process (configuration and compilation) on different systems. Second, test of correctness of the simulation and reconstruction results. The build system and QA infrastructure are based on CMake, CTest and CDash. The build process is tested using the standard above-mentioned set of tools. For the simulation and reconstruction tests a set of tools was developed, which includes base classes for reports, histogram management, a simulation and reconstruction QA classes and scripts. Test results in form of the user-friendly reports are published on the CDash and on the dedicated web-server where developer can browse, for example, the tracking performance two weeks ago in order to fix the bug. Described QA system considerably improves the development process and leads to a faster development cycles of CBMROOT.

Poster presentations / 357

dCache Billing data analysis with Hadoop

Author: Kai Leffhalm[1]

Co-author: Andreas Knoepke [2]

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

[2] *DESY*

Corresponding Author: kai.leffhalm@desy.de

The dCache storage system writes billing data into flat files or a relational database.
For a midsize dCache installation there are one million entries - representing 300 MByte - per day.
Gathering accounting information for a longer time interval about transfer rates per group, per file type or per user results in increasing load on the servers holding the billing information.
Speeding up these requests renders new approaches to performance optimization worthwhile.

Hadoop is a framework for distributed processing of large data using multiple computer nodes.
The essential point in our context is the scalability for big data.
Data is distributed over many nodes in the Hadoop Distributed File System (HDFS).
Queries are processed in parallel on every node to extract the information and combine it in another step.
This is called a MapReduce algorithm.
As the dCache storage is distributed over many storage nodes combining both on every node is obvious.

The transformation of the billing information into the HDFS structure is done by a small script.
The MapReduce functions to create the results to the most important queries are implemented for each request.
We will present the system's setup and performance comparisons of the created queries using Postgresql, flat files and Hadoop.
The overall gain in performance and its dependence on both the amount of analysed data and available machines for paralleling the request will be demonstrated.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / 275

The ATLAS Distributed Analysis System

Author: Federica Legger[1]

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

Corresponding Author: federica.legger@physik.uni-muenchen.de

In the LHC operations era, analysis of the multi-petabyte ATLAS data sample by globally distributed physicists is a challenging task. To attain the required scale the ATLAS Computing Model was designed around the concept of grid computing, realized in the Worldwide LHC Computing Grid (WLCG), the largest distributed computational resource existing in the sciences. ATLAS currently stores over 140 PB of data and runs about 140,000 concurrent jobs continuously at WLCG sites. During the LHC´s first run, the ATLAS Distributed Analysis (DA) service has operated stably and scaled well. More than 1600 users submitted jobs 2012, with 2 million or more analysis jobs per week, peaking at about a million jobs per day. The system dynamically distributes popular data to expedite processing and maximally utilize resources. The reliability of the DA service is high but steadily improving; grid sites are continually validated against a set of standard tests, and a dedicated team of expert shifters provides user support and communicates user problems to the sites. Both the user support techniques and the direct feedback of users have been effective in improving the success rate and user experience when utilizing the distributed computing environment. In this contribution

a description of the main components, activities and achievements of ATLAS distributed analysis is given. Also several future improvements being undertaken will be described.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / 185

CernVM Online and Cloud Gateway: a uniform interface for CernVM contextualization and deployment

Authors: Georgios Lestaris[1]; Ioannis Charalampidis[1]

Co-authors: Dario Berzano [1]; Gerardo Ganis [1]; Jakob Blomer [1]; Predrag Buncic [1]; Rene Meusel [1]

[1] *CERN*

Corresponding Author: george.lestaris@cern.ch

In a virtualized environment, contextualization is the process of configuring a VM instance for the needs of various deployment use cases. Contextualization in CernVM can be done by passing a handwritten context to the "user data"field of cloud APIs, when running CernVM on the cloud, or by using CernVM web interface when running the VM locally. CernVM online is a publicly accessible web interface that unifies these two procedures. A user is able to define, store and share CernVM contexts using CernVM Online and then apply them either in a cloud by using CernVM Cloud Gateway, or on a local VM with the single-step pairing mechanism. CernVM Cloud Gateway is a distributed system that provides a single interface to use multiple and different clouds (by location or type, private or public). Cloud gateway has been so far integrated with OpenNebula, CloudStack and EC2 tools interfaces. A user, with access to a number of clouds, can run CernVM cloud agents that will communicate with these clouds using their interfaces, and then use one single interface to deploy and scale CernVM clusters. CernVM clusters are defined in CernVM Online and consist of a set of CernVM instances that are contextualized and can communicate with each other. In this contribution we present these new components, their status and some common use cases, as well as possible future work. We will also show how the combination of CernVM Online and Cloud Gateway turns out to be an effective way to federate clouds.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / 220

Public Storage for the Open Science Grid

Author: Tanya Levshina[1]

Co-author: Ashu Guru [2]

[1] *FERMILAB*

[2] *University of Nebraska, Lincoln*

Corresponding Author: tlevshin@fnal.gov

The Open Science Grid (OSG) Public Storage project is focused on improving and simplifying the management of OSG Storage. Currently, OSG doesn't provide efficient means to manage public storage offered by participating sites. A Virtual Organization (VO) that relies on opportunistic storage has difficulties finding appropriate storage, verifying its availability, and monitoring its utilization. The involvement of the Production Manager, site administrators and VO support personnel is required to allocate or rescind storage space. One of the main requirements for Public Storage implementation is that it should use SRM or GridFTP protocols to access the Storage Elements(SE) provided by the OSG Sites and doesn't put any additional burden on sites. No new services related to Public Storage could be installed and run on OSG sites.
Opportunistic users also have difficulties in accessing the distributed storage during the execution

of jobs. The typical users' data management workflow includes pre-staging common data on sites before a job's execution, then somehow storing output data produced by a job on a worker node for a subsequent download to a local institution. When the amount of data is significant, the only means to temporarily store the data is to upload it to one of the Storage Elements. In order to do that, a user's job should be aware of the storage location, availability, and free space. After a successful data upload, users should somehow keep track of the data's location for future access. In this presentation we proposed solutions for storage management and data handling issues in the OSG. We are investigating the feasibility of using the integrated Rule-Oriented Data System (iRODS) developed at RENCI as a front-end service to the OSG SEs. The current architecture, state of deployment and performance test results will be discussed. We will also provide examples of current usage of the system by beta-users.

Event Processing, Simulation and Analysis / **331**

Development of Bayesian analysis program for extraction of polarisation observables at CLAS

Author: Stefanie Lewis[None]

Co-authors: David Ireland [1]; Wim Vanderbauwhede [2]

[1] *University of GLasgow*

[2] *University of Glasgow*

Corresponding Author: s.lewis.glasgow@gmail.com

At the mass of a proton, the strong force is not well understood. Various quark models exist, but it is important to determine which quark model(s) are most accurate. Experimentally, finding resonances predicted by some models and not others would give valuable insight into this fundamental interaction. Several labs around the world use photoproduction experiments to find these missing resonances. The aim of this work is to develop a robust Bayesian data analysis program for extracting polarisation observables from pseudoscalar meson photoproduction experiments using CLAS at Jefferson Lab. This method, known as nested sampling, has been compared to traditional methods and has incorporated data parallelisation and GPU programming. It involves an event-by-event likelihood function, which has no associated loss of information from histogram binning, and results can be easily constrained to the physical region. One of the most important advantages of the nested sampling approach is that data from different experiments can be combined and analysed simultaneously. Results on both simulated and previously analysed experimental data for the K-Lambda channel will be discussed.

Poster presentations / **402**

An efficient data protocol for encoding preprocessed clusters of CMOS Monolithic Active Pixel Sensors

Author: Qiyan Li[1]

Co-authors: Joachim Stroth [2]; Michael Deveaux [3]; Samir Amar-Youcef [4]

[1] *Goethe University Frankfurt*

[2] *Goethe-University and GSI*

[3] *University Frankfurt*

[4] *Uni Frankfurt*

Corresponding Author: pigeon7736@gmail.com

CBM aims to measure open charm particles from 15-40 AGeV/c heavy ion collisions by means of secondary vertex reconstruction. The measurement concept includes the use of a free-running DAQ, real time tracking, primary and secondary vertex reconstruction and a tagging of open charm candidates based on secondary vertex information. The related detector challenge will be adressed with an ultra-light and highly granular Micro Vertex Detector (MVD) based on CMOS Monolithic Active Pixel Sensors (MAPS).

Performing the real time vertex reconstruction at collision rates of ˜1e5 coll./s will introduce a substantial CPU-load to the computing system (FLES) of CBM. To reduce this load, we consider to perform pre-processing steps like cluster finding already in DAQ-instances upstream the FLES. A successful pre-processing concept should be FPGA-compatible. Moreover, it should provide a lossless encoding of the original information as much as the newly computed information on the cluster position and shape without extending the data volume.

To fulfill those requirements, we developped a cluster encoding concept which may encode the above mentioned information in a single 32-bit word. This concept is introduced and its validity is discussed based on data from the recent beam test of the MVD-prototype at CERN-SPS.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / **481**

Implementation of grid Tier 2 and Tier 3 facilities on a Distributed OpenStack Cloud

Author: Martin Sevior[1]

Co-authors: Antonio Limosani [2]; Joanna Huang [1]; Ross Wilson [3]; Shunde Zhang [4]

[1] *University of Melbourne*

[2] *University of Melbourne (AU)*

[3] *University of Adelaide*

[4] *eRSA*

Corresponding Authors: antonio.limosani@cern.ch, msevior@gmail.com

The Australian Government is making a $AUD 100 million investment in Compute and Storage for the academic community. The Compute facilities are provided in the form of 24,000 CPU cores located at 8 nodes around Australia in a distributed virtualized Infrastructure as a Service facility based on OpenStack. The storage will eventually consist of over 100 petabytes located at 6 nodes. All will be linked via a 100 Gbs network.

This presentation will describe the development of a fully connected WLCG Tier-2 grid site as well as a general purpose Tier-3 computing cluster based on this architecture.

The facility employs an extension to Torque to enable dynamic allocations of virtual machine instances. Storage is provided by a federation of DPM installations at each storage node. A base Scientific Linux VM image is deployed in the OpenStack cloud and automatically configured as required using Puppet. Custom scripts are used to launch multiple VMs, integrate them into the dynamic Torque cluster and to mount remote file systems.

We will report on our experience in implementing this nation-wide ATLAS and Belle II Tier 2 and Tier 3 computing infrastructure using the
national Research Cloud and storage facilities. In particular we will describe how we have addressed the challenges of using OpenStack VMs in a Torque cluster, automated configuration of VM instances, federated

authentication across multiple institutions and supported access to remote file systems.

Poster presentations / 6

An SQL-based approach to Physics Analysis

Author: Maaike Limper[1]

[1] *CERN*

Corresponding Author: maaike.limper@cern.ch

As part of the CERN Openlab collaboration, an investigation has been made into the use of an SQL-based approach for physics analysis with various up-to-date software and hardware options.
Currently physics analysis is done using data stored in customised root-ntuples that contain only the variables needed for a specific analysis. Production of these ntuples is mainly done by accessing the centrally produced analysis data through the LHC computing grid and can take several days to complete.
We'll present an alternative approach to physics data analysis where analysis data is stored in a database, removing the need for customized ntuple production, and allowing calculations that are part of the analysis to be done on the database side. An example implementation of such a database will be shown, demonstrating how physics analysis in a database can be done via ROOT.
The use of an Oracle database in this setup is compared to use of a Hadoop-based data structure. The advantages and drawbacks of the different database setups are presented, with a detailed analysis of the CPU and I/O usage, in particular for the case when many users need to access the database at the same time.

Poster presentations / 352

Dataset-based High-Level Data Transfer System in BESDIRAC

Authors: Tao Lin[1]; Xiaomei Zhang[1]

Co-authors: Andrei Tsaregorodtsev [2]; Weidong Li [1]

[1] *Institute of High Energy Physics*

[2] *Marseille*

Corresponding Author: lintao@ihep.ac.cn

Data Transfer is an essential part in grid. In the BESIII experiment, the result of Monte Carlo Simulation should be transfered back from other sites to IHEP and the DST files for physics analysis should be tranfered from IHEP to other sites. A robust transfer system should
make sure all data are transfered correctly.
DIRAC consists of cooperation distributed services and light-weight agents delivering the workload to the Grid Resources. We reuse the most basic functionalities supplied by DIRAC. BESDIRAC is an extension to DIRAC for BES specified. In BESDIRAC, a Dataset-based Data Transfer System is developed. In this paper, we present the design of this system and its implementation details. A Transfer Request Service is used for creating and monitoring
the transfer requests. A Transfer Agent is used for transfering data from one SE to another.
For flexibility and reuse of the current low-level transfer systems, we have designed a transfer worker factory to create transfer workers with different protocols. A transfer worker is the wrapper of the low-level file transfer commands. The Transfer Agent uses the Async I/O to manage the transfer workers.

Plenaries / **498**

Nikhef, the national institute for subatomic physics

Author: Frank Linde[1]

[1] *NIKHEF (NL)*

Corresponding Author: f.linde@nikhef.nl

Poster presentations / **188**

T2K-ND280 Computing Model

Author: Thomas Lindner[1]

[1] *T*

Corresponding Author: lindner@triumf.ca

ND280 is the off-axis near detector for the T2K neutrino experiment. ND280 is a sophisticated, multiple sub-system detector designed to characterize the 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. The first key challenge has been dealing with very different computing infrastructure in different continents; a second challenge has been dealing with the relatively large ˜1PB ND280 MC dataset. We will describe the software, data storage and data distribution solutions developed to meet these challenges. We will also briefly discuss the database infrastructure developed to support ND280.

Poster presentations / **413**

Matrix Element Method with Graphics Processing Units (GPUs)

Authors: Robert Duane Harrington Jr[1]; Stephen Lloyd[1]

Co-author: Andy Buckley [2]

[1] *University of Edinburgh*
[2] *University of Edinburgh (GB)*

Corresponding Author: robert.duane.harrington.jr@cern.ch

The Matrix Element Method has been used with great success in the past several years, notably for the high precision top quark mass determination, and subsequently the single top quark discovery, at the Tevatron. Unfortunately, the Matrix Element method is notoriously CPU intensive due to the complex integration performed over the full phase space of the final state particles arising from high energy interactions. At the Large Hadron Collider (LHC), high luminosities mean much larger numbers of events to analyse than we had at the Tevatron. This makes the Matrix Element method difficult to use, particularly now that computing resources are being utilised already at capacity. We have studied the feasibility of using Graphics Processing Units (GPUs) to reduce the computing time required for the measurement of Higgs decaying to two muons at the LHC. We present the technical approach followed, the speed-up gained, and the prospects for future improvements in the analysis through the use of GPUs.

Poster presentations / 39

Monitoring System for the GRID Monte Carlo Mass Production in the H1 Experiment at DESY

Authors: Alexander Fomenko[1]; Bogdan Lobodzinski[2]; Lena Bystritskaya[3]; Nelly Gogitidze[1]

[1] *Lebedev Institute, Moscow, Russia*

[2] *DESY, Hamburg, Germany*

[3] *ITEP Moscow, Russia*

Corresponding Author: bogdan@mail.desy.de

Small Virtual Organizations (VO) employ all components of the EMI or gLite Middleware. In this framework, a monitoring system is designed for the H1 Experiment to identify and recognize within the GRID the best suitable resources for execution of CPU-time consuming Monte Carlo (MC) simulation tasks (jobs). Monitored resources are Computer Elements (CEs), Storage Elements (SEs), WMS-servers (WMSs), CernVM File System (CVMFS) available to the VO "hone" and local GRID User Interfaces (UIs).

The general principle of monitoring of the GRID elements is based on the execution of short test jobs on different CE queues using submission through various WMSs and directly to the CREAM-CEs as well. Real H1 MC Production jobs with a small number of events are used to perform the tests. Test jobs are periodically submitted into GRID queues, the status of these jobs is checked, output files of completed jobs are retrieved, the result of each job is analyzed and the waiting time and run time are derived. Using this information, the status of the GRID elements is estimated and the most suitable ones are included in the automatically generated configuration files for use in the H1 MC production. Monitored information is stored in a MySQL database and is presented in detail on web pages and the MonAlisa visualisation system.

The monitoring system allows for identification of problems in the GRID sites and promptly reacts on it (for example by sending GGUS trouble tickets). The system can easily be adapted to identify the optimal resources for tasks other than MC production, simply by changing to the relevant test jobs. The monitoring system is written mostly in Python and Perl with insertion of a few shell scripts.

In addition to the test monitoring system we additionally use information from real production jobs to monitor the availability and quality of the GRID resources. The monitoring tools register the number of job resubmissions, the percentage of failed and finished jobs relative to all jobs on the CEs and determine the average values of waiting and running time for the involved GRID queues. CEs which do not meet the set criteria can be removed from the production chain by including them in an exception table. All of these monitoring actions lead to a more reliable and faster execution of MC requests.

Software Engineering, Parallelism & Multi-Core / 339

Systematic profiling to monitor and specify the software refactoring process of the LHCb experiment

Authors: Ben Couturier[1]; Emmanouil Kiagias[2]; Stefan Lohn[1]

[1] *CERN*

[2] *University of Athens (GR)*

Corresponding Author: stefan.lohn@cern.ch

Software optimization is a complex process, where the intended improvements have different effects on different platforms, with multiple operating systems and an ongoing introduction of new

hardware. In addition several compilers produce differing object-code as result of different internal optimization procedures. To trace back the impact of the optimizations is going to become more difficult. To obtain precise information of the general performance, to make profiling results comparable and to verify the influences of improvements in the framework or of specific algorithms, it is important to rely on standardized profiling and regression tests. Once done, software metrics can be created from the profiling results to monitor the changes in performance and to create reports about on a regular basis if modifications lead to significant performance degradations.
The LHCb collaboration develops and maintains large software frameworks for the LHCb experiment, and for the HEP community like Gaudi. In the upcoming years a big refactoring effort is planned to introduce or optimize the utilization of features like vectorization, parallelization, to reduce the influences of hotspots and to evaluate strategies to reduce the impact of bottlenecks. It is crucial to guide the refactoring with a profiling system that gives hints for necessary source-code reengineering and how these optimizations behave with different configurations. To achieve this a system for systematic profiling is set up to run test jobs along with the new build system based on the open-source project Jenkins. Summary data are abstracted from the detailed profiling results to be visualized on a web based analysis platform and more detailed information can be accessed through a public network file system.
In order to improve the optimization process and to focus the labor intensive developments on crucial issues it is necessary to evaluate the benefits of such a profiling service, to point out limitations, and to reuse features of already widespread profiling software.

Data Acquisition, Trigger and Controls / **430**

NaNet: a low-latency NIC enabling GPU-based, real-time low level trigger systems.

Authors: Alessandro Lonardo[1]; Andrea Biagioni[2]; Davide Rossetti[3]; Francesca Locicero[4]; Francesco Simula[4]; Laura Tosoratto[4]; Ottorino Frezza[4]; Pier Stanislao Paolucci[4]; Piero Vicini[5]; Roberto Ammendola[4]

Co-authors: Felice Pantaleo [6]; Gianluca Lamanna [7]; Marco Sozzi [8]; Riccardo Fantechi [8]

[1] *INFN, Roma I (IT)*

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

[3] *U*

[4] *INFN*

[5] *INFN Rome Section*

[6] *CERN - University of Pisa*

[7] *CERN*

[8] *Sezione di Pisa (IT)*

Corresponding Author: alessandro.lonardo@cern.ch

The integration of GPUs in trigger and data acquisition systems is currently being investigated in several HEP experiments.
At higher trigger levels, when the efficient many-core parallelization of event reconstruction algorithms is possible, the benefit of reducing significantly the number of the farm computing nodes is evident.
At lower levels, where tipically severe real-time constraints are present and custom hardware is used, the advantages of GPUs adoption is less straightforward.
A pilot project within the CERN NA62 experiment is investigating the usage of GPUs in the central Level 0 trigger processor, exploiting their computing power to implement efficent, high throughput event selection algorithms while retaining the real-time requisites of the system. One of the project preliminary results was that data transfer over GbE links from readout boards to GPU memories using commodity NICs and vanilla software stack consumed the biggest part of the time budget and was the main source of fluctuations in the global system response time.
In order to reduce data transfer latency and its fluctuations we envisioned the usage of the GPUDirect RDMA technology, injecting readout data directly from the NIC into the GPU memories without any intermediate buffering, and the offloading of the network stack protocol management from the CPU, eliminating OS contribution to latency and jitter.

We implemented these two features in the NaNet FPGA-based NIC: the first was inherited from the APEnet+ 3D NIC development, while the second was realized integrating an Open IP provided by the FPGA vendor.
We will provide a deep description of the NaNet architecture and a detailed performance analysis of the integrated system on the NA62 RICH detector GPU-based L0 trigger processor case study, along with an insight of future developments.

Poster presentations / **263**

Use of VMWare for providing cloud infrastructure for the Grid

Author: Robin Eamonn Long[1]

[1] *Lancaster University (GB)*

Corresponding Author: r.long@cern.ch

The need to maximize computing facilities whilst maintaining versatile and flexible setups leads to the need for on demand virtual machines through the use of cloud computing. GridPP is currently investigating the role that Cloud Computing, in the form of Virtual Machines, can play in supporting Particle Physics analyses. As part of this research we look at the ability of VMWare's ESXi hypervisors to provide such Virtual Machines; the advantages of such systems and their overall performance. We conclude with a contrast between the ability of VMWare and other major OpenSource alternatives such as Openstack and StratusLab in fulfilling this role.

Poster presentations / **388**

Efficient computation of hash functions

Author: raul lopes[1]

Co-authors: Peter Hobson [2]; Virginia Franqueira [3]

[1] *School of Design and Engineering - Brunel University, UK*

[2] *Brunel University (GB)*

[3] *University of Central Lancashire, UK*

Corresponding Authors: raul.lopes@brunel.ac.uk, peter.hobson@brunel.ac.uk

The performance of hash function computations can impose a significant workload on SSL/TLS authentication servers. In the WLCG this workload shows also in the computation of data transfers checksums. It has been shown in the EGI grid infrastructure that the checksum computation can double the IO load for large file transfers leading to an increase in re-transfers and timeout errors. Storage managers like SToRM try to reduce that impact by computing the checksum during the transfer. That may not be feasible, however, when multiple transfer streams are combined with the use of hashes like MD-5 or SHA-2.

We present two alternatives to reduce the hash computation load.
First we introduce implementations for the Fast SHA-256 and SHA-512 that can reduce the number of cycles per second of a hash computation from 15 to under 11. Secondly we introduce and evaluate parallel implementations for two novel hash tree functions:
NIST SHA-3 Keccak and Skein. These functions were conceived to take advantage of parallel data transfers and their deployment can significantly reduce the timeout and re-transfer errors mentioned above.

Software Engineering, Parallelism & Multi-Core / **422**

Synergia-CUDA: GPU Accelerated Accelerator Modeling Package (video conference)

Author: Qiming Lu[1]

Co-author: James Amundson [1]

[1] *Fermi National Accelerator Laboratory*

Corresponding Author: qlu@fnal.gov

Synergia is a parallel, 3-dimensional space-charge particle-in-cell code that is widely used by the accelerator modeling community. We present our work of porting the pure MPI-based code to a hybrid of CPU and GPU computing kernels. The hybrid code uses the CUDA platform, in the same framework as the pure MPI solution. We have implemented a lock-free collaborative charge-deposition algorithm for the GPU, as well as other optimizations, including local communication avoidance for GPUs, customized FFT, and fine-tuned memory access patterns. On a small GPU cluster (up to 4 Tesla C1070 GPUs), our benchmarks exhibit both superior peak performance and better scaling, when compared to a CPU cluster with 16 nodes and 128 cores. We have further compared the code performance on different GPU architectures, including C1070 Tesla, M2070 Fermi, and K20 Kepler. We show 10 to 20% performance increases with optimizations addressing each specific hardware architectures.

Data Stores, Data Bases, and Storage Systems / **229**

Integration of S3-based cloud storage in BES III computing environment

Authors: Fabio Hernandez[1]; Wang Lu[2]; Ziyan Deng[2]

[1] *IN2P3/CNRS and Institute of High Energy Physics,CAS*

[2] *Institute of High Energy Physics,CAS*

Object storage systems based on Amazon's Simple Storage Service (S3) have substantially developed in the last few years. The scalability, durability and elasticity characteristics of those systems make them well suited for a range of use cases where data is written, seldom updated and frequently read. Storage of images, static web sites and backup systems are some of the use cases where S3 systems have proven effective. Experimental data for high-energy physics research can also benefit from storage systems optimized for write-once read-many operational models.
The BES III experiment studies physics in the tau-charm energy region from 2GeV to 4.6 GeV, at the Institute of High Energy Physics (IHEP) in Beijing, China. Since spring 2009, BES III has been recording and accumulating a significant amount of experimental data, in the order of 1 PB per year. Organized around the central data repository operated by IHEP's computing center, the experiment's computing environment is composed of sites located in several countries.
In this contribution we present an ongoing work, which aims to evaluate the suitability of S3-based cloud storage as a supplement to the Lustre file system for storing experimental data for BES III. In particular, we discuss our findings regarding the integration of S3-based storage in the software stack of the experiment. We report on our development work that improves the support of CERN's ROOT data analysis framework and allows efficient remote access to data through the S3 protocol. We also discuss our results providing the experiment with efficient command line tools for interacting with S3-based data repositories from interactive sessions and grid jobs. The FUSE-based file system interface for a S3 storage backend that we developed is also presented and our efforts for providing tools for easily navigating the experiment's data repository and making it seamlessly accessible in particular from the researcher's personal computer.
This work is being validated through real use cases of production BES III jobs by using two different storage backends: a hardware-based solution around Huawei UDS appliance and a software-based solution around OpenStack Swift. We compare the performance of those systems with the Lustre file

system for local and grid jobs and also for transferring data from and to remote sites participating in the BES collaboration.

Plenaries / 493

Software engineering for science at the LSST

Author: Robert Lupton[1]

[1] *Princeton*

Many of the scientific computing frameworks used in 'big science' have several million lines of source code, and software engineering challenges are amongst the most prominent challenges, be it in high-energy physics, astronomy, or other sciences. Dr Robert Lupton of Princeton University will talk the software engineering challenges that face scientific computing and how large scale systems like the Large Synoptic Survey Telescope LSST and others deal with these challenges.

Event Processing, Simulation and Analysis / 348

ArtG4: A generic framework for Geant4 simulations

Author: Adam Lyon[1]

[1] *Fermilab*

Corresponding Author: lyon@fnal.gov

Flexibility in producing simulations is a highly desirable, but difficult to attain feature. A simulation program may be written for a particular purpose, such as studying a detector or aspect of an experimental apparatus, but adapting that program to answer different questions about that detector or apparatus under different situations may require recoding or a separate fork of the program. The Fermilab Muon g-2 collaboration faced this problem and has written a generic framework for Geant4 simulations that provides the desired flexibility and has had a large positive impact on our ability to quickly produce valuable and effective simulation code.

A stand-alone Geant4 simulation of the Muon g-2 storage ring and detectors, called g2migtrace, had been written previously to test muon storage acceptance and detector response for different apparatus options [1]. This simulation is extremely detailed and contained enormously valuable code describing ring and detector geometry as well as detector response, especially for the crystal calorimeters. Last year a test beam was performed to test different calorimeter designs. In an effort to simulate the test beam conditions but utilize the extensive library of Geant4 code in g2migtrace, a collaborator introduced #ifdef compilation switches to turn the storage ring simulation into a test beam simulation, and checked that code back in to the main branch of the source. These switches make the code extremely difficult to maintain and inflexible. How do we simulate future different test beams? We now have to do tests with these switches. In fact, when built with full compiler optimizations, the code suffered a segmentation fault from a variable left uninitialized due to a poorly placed #ifdef.

The right way to introduce flexibility into the simulation code is to build it in from the start. We have written "ArtG4", a completely generic framework on top of the Fermilab Art infrastructure [2] used by many Intensity Frontier experiments at Fermilab. Modules defining geometry and response for particular detectors as well as user written Geant4 actions may be arranged into a simulation program with a configuration file. With this system, we can write simulations of the entire muon storage ring as well as test individual detectors in a test beam situation all with the same library of code. This system allows us to work together and rapidly produce many simulations, answering a

variety of questions much more easily than before. This system will be fully explained and several examples will be provided.

[1] g2migtrace: Kevin Lynch and Zach Hartwig
[2] C Green et. al., 2012 J. Phys.: Conf. Ser. 396 022020 (CHEP 2012)

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / **434**

The ”Last Mile” of Data Handling - Fermilab’s IFDH tools

Authors: Adam Lyon[1]; Marc Mengel[1]

[1] *Fermilab*

Corresponding Author: lyon@fnal.gov

IFDH (Intensity Fronter Data Handling), is a suite of tools for data movement tasks for Fermilab experiments and is an important part of the FIFE (Fabric for Frontier Experiments) initiative described at this conference. IFDH encompasses moving input data from caches or storage elements to compute nodes (the “last mile” of data movement) and moving output data potentially to those caches as part of the journey back to the user. IFDH also involves throttling and locking to ensure that large numbers of jobs do not cause data movement bottlenecks. IFDH is realized as an easy to use layer that users call in their job scripts (e.g. “ifdh cp”), hiding the low level data movement tools. One advantage of this layer is that the underlying low level tools can be selected or changed without the need for the user to alter their scripts. Logging and performance monitoring can also be added easily. This system will be presented in detail as well as its impact on the ease of data handling at Fermilab experiments.

Poster presentations / **378**

Control functionality of DAQ-Middleware

Author: Hiroyuki Maeda[1]

Co-authors: Eiji Inoue [2]; Hiroshi Sendai [2]; Masaki Wada [3]; Noriaki Ando [4]; Shuhei Ajimura [5]; Tetsuo Kotoku [4]; Yasushi Nagasaka [1]

[1] *Hiroshima Institute of Technology*

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

[3] *Bee Beans Technologies Co.Ltd*

[4] *The National Institute of Advanced Industrial Science and Technology (AIST)*

[5] *Osaka University*

Corresponding Author: m161203@cc.it-hiroshima.ac.jp

DAQ-Middleware is a software framework for a network-distributed data acquisition (DAQ) system that is based on the Robot Technology Middleware (RTM). The framework consists of a DAQ-Component and a DAQ-Operator. The basic functionalities such as transferring data, starting and stopping the system, and so on, are already prepared in the DAQ-Components and DAQ-Operator. The DAQ-Component is especially used as a core component to read, transfer, and record data. And the DAQ-Operator is used as controlling the DAQ-Components. The system works as a state machine and it has four states, i.e., LOADED, CONFIGURED, RUNNING, PAUSED states. The states changed by the command to be transferred from the DAQ-Operator.
The DAQ system can be easily implemented with using the framework. But the control functionalities, such as transferring parameters from DAQ-Operator to DAQ-Components, were not prepared,

except when the state changes from LOADED to CONFIGRED. Then we developed the functionality to transfer data from DAQ-Operator to DAQ-Components at any time.
The new functionality enables us to transfer any parameters that are stored in a configuration file on the DAQ-Operator to each DAQ-Component. In order to add this functionality, the new state, SET, is introduced. The DAQ-Operator transfers the specified configuration file to the specified DAQ-Components in this SET state. On the other hand, the DAQ-Components receive the configuration data from the DAQ-Operator in the SET state.
Any parameters can be transferred and set with the DAQ-Middleware without restarting the system with using this functionality.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / **294**

Evolution of the ATLAS PanDA Workload Management System for Exascale Computational Science

Author: Tadashi Maeno[1]

Co-authors: Alexandre Vaniachine [2]; Alexei Klimentov [1]; Dantong Yu [3]; Kaushik De [4]; Paul Nilsson [4]; Sergey Panitkin [1]; Torre Wenaus [1]

[1] *Brookhaven National Laboratory (US)*

[2] *ANL*

[3] *BROOKHAVEN NATIONAL LABORATORY*

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

Corresponding Authors: tmaeno@bnl.gov, alexei.klimentov@cern.ch, kaushik@uta.edu, paul.nilsson@cern.ch, panitkin@bnl.gov, vaniachine@anl.gov, wenaus@gmail.com, dtyu@bnl.gov

An important foundation underlying the impressive success of data processing and analysis in the ATLAS experiment at the LHC is the Production and Distributed Analysis (PanDA) workload management system. PanDA was designed specifically for ATLAS and proved to be highly successful in meeting all the distributed computing needs of the experiment. However, the core design of PanDA is not experiment specific. The PanDA workload management system is capable of meeting the needs of other data intensive scientific applications. Alpha-Magnetic Spectrometer, an astro-particle experiment on the International Space Station, and the Compact Muon Solenoid, an LHC experiment, have successfully evaluated PanDA and are pursuing its adoption. In this talk, a description of the new program of work to develop a generic version of PanDA will be given, as well as the progress in extending PanDA's capabilities to support supercomputers and clouds and to leverage intelligent networking. PanDA has demonstrated at a very large scale the value of automated dynamic brokering of diverse workloads across distributed computing resources. The next generation of PanDA will allow other data-intensive sciences and a wider exascale community employing a variety of computing platforms to benefit from ATLAS' experience and proven tools.

Poster presentations / **368**

The performance of the ATLAS tau trigger in 8 TeV collisions and novel upgrade developments for 14TeV

Author: Phillip Urquijo[1]

[1] *Universitaet Bonn (DE)*

Corresponding Author: joern.mahlstedt@cern.ch

The LHC is the world's highest energy and luminosity proton-proton (p-p) collider. During 2012 luminosities neared 10^34 cm-2 s-1, with bunch crossings occurring every 50 ns. The online event selection system of the ATLAS detector must reduce the event recording rate to only a few hundred Hz and, at the same time, selecting events considered interesting. This presentation will specifically cover the online selection of events with tau leptons decaying to hadronic final states. The "hadronic tau trigger" has been operated successfully since 2009. Tau leptons provide a unique window for understanding physics at the Tera scale. They are the most crucial signature for measuring the fermion couplings of the Higgs boson, an important next step in identifying the recently discovered Higgs-like boson. Many theories also predict new particles beyond the Standard Model that have large couplings to tau leptons. The first step in their identification is the online event trigger.
ATLAS employs a 3-level trigger scheme. The level-1 (L1) hadronic tau trigger system measures energy deposited in electromagnetic (EM) and hadronic (HAD) calorimeter trigger towers to select taus based on energy in core and isolation regions. The remaining two levels, are software-based. At the second level, further identification is done within regions of interest identified by L1, taking into account track and calorimeter information with dedicated fast algorithms. At the third and final level, the event filter algorithms closely match the algorithms used for offline reconstruction. Taus are often mimicked by similarly behaved quark jets, which have large rates. The rate at which events must be selected severely limits the complexity of reconstruction algorithms, compounded by the high probability of overlap (pileup) between bunch crossings. ATLAS has a program dedicated to the continued improvement of the tau trigger to address these problems, including: Multivariate identification algorithms using high granularity calorimeter and tracking information, fast track reconstruction methods for accurate determination of impact parameters reducing pile up, and new topological criteria using multiple event features. The latter involves simultaneous triggering on taus with muons, electrons, and missing energy. These developments have given ATLAS the potential to explore a large program of tau physics analysis.
This presentation gives a full overview of the ATLAS tau trigger system, summarising the running experience over the past 3 years. We demonstrate that the ATLAS tau trigger performed remarkably well throughout its operation, and discuss computational innovations in 2012. Results of the performance of the tau trigger from the full 22 fb-1 2012 p-p data taking period will be shown, including measurements of the trigger efficiency using Z->tau tau and W->tau nu events and the application to searches for tau tau resonances, such as the Higgs boson. We also outline the upgrade plan to 2015 for 14(13) TeV LHC proton-proton collisions, which includes the use of a novel associative memory trigger for track finding throughout the full detector.

Poster presentations / **395**

Redundant Web Services Infrastructure for High Performance Data Access

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 REST HTTP-based data access architecture to provide access to various types of data. We have built a simple yet versatile infrastructure which allows us to use redundant copies of web application servers to increase service performance and availability. It is designed to handle both state-less and state-full data access methods using distributed web server.

In most cases, the 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.

Poster presentations / 396

Scientific Collaborative Tools Suite at FNAL

Author: Igor Mandrichenko[1]

Co-authors: Margherita Vittone-Wiersma [2]; Vladimir Podstavkov [2]

[1] *Fermilab*

[2] *FNAL*

Corresponding Author: ivm@fnal.gov

Over several years, we have developed a number of collaborative tools used by groups and collaborations at FNAL, which is becoming a Suite of Scientific Collaborative Tools. Currently, the suite includes:

- Electronic Logbook (ECL),
- Shift Scheduler,
- Speakers Bureau and
- Members Database.

These product organize and help run the collaboration at every stage of its life cycle from detector building to simulation, data acquisition, processing, analysis and publication of the results of the research.

These projects share common technologies and architecture solutions. Mobile computing opens whole new area of exciting opportunities for collaborative tools and new vision of the human/computer interface.

The presentation will cover history, current status and future development plans for the Scientific Collaborative Tools Suite at FNAL.

Software Engineering, Parallelism & Multi-Core / 278

ATLAS Offline Software Performance Monitoring and Optimization

Author: Rocco Mandrysch[1]

Co-authors: Andreas Salzburger [2]; Elmar Ritsch [3]; Graeme Andrew Stewart [2]; Gunjan Kabra [4]; Neelima Chauhan [4]; Niels Van Eldik [2]; Robert Johannes Langenberg [5]; Roberto Agostino Vitillo [6]; Rolf Seuster [7]; Thomas Kittelmann [8]

[1] *University of Iowa (US)*

[2] *CERN*

[3] *University of Innsbruck (AT)*

[4] *Mody Institute of Technology and Science (IN)*

[5] *Technische Universitaet Muenchen (DE)*

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

[7] *TRIUMF (CA)*

[8] *ESS - European Spallation Source (SE)*

Corresponding Authors: rocco.mandrysch@cern.ch, graeme.andrew.stewart@cern.ch, gunjan.kabra@cern.ch, robert.langenberg@cern.ch, neelima.neelima@cern.ch, elmar.ritsch@cern.ch, andreas.salzburger@cern.ch, rolf.seuster@cern.ch, niels.van.eldik@cern.ch, roberto.agostino.vitillo@cern.ch

In a complex multi-developer, multi-package software environment, such as the ATLAS offline Athena framework, tracking the performance of the code can be a non-trivial task in itself. In this paper we describe improvements in the instrumentation of ATLAS offline software that have given considerable insight into the performance of the code and helped to guide optimisation.

Code can be instrumented firstly using the PAPI tool, which is a programing interface for accessing hardware performance counters. PAPI events can count floating point operations, cycles and instructions and cache accesses. Triggering PAPI to start/stop counting for each algorithm and processed event gives a good understanding of the whole algorithm level performance of ATLAS code.

Further data can be obtained using pin, a dynamic binary instrumentation tool. Pintools can be used to obtain similar statistics as PAPI, but advantageously without requiring recompilation of the code. Fine grained routine and instruction level instrumentation is also possible. Pintools can additionally interrogate the arguments to functions, like those in linear algebra libraries, so that a detailed usage profile can be obtained.

These tools have charaterised the extensive use of vector and matrix operations in ATLAS tracking. Currently, CLHEP is used here, which is not an optimal choice. To help evaluate replacement libraries a testbed has been setup allowing comparison of the performance of different linear algebra libraries (including CLHEP, Eigen and SMatrix/SVector). Results are then presented via the ATLAS Performance Management Board framework, which runs daily with the current development branch of the code and monitors reconstruction and Monte-Carlo jobs. This framework analyses the CPU and memory performance of algorithms and an overview of results are presented on a web page.

These tools have provided the insight necessary to plan and implement performance enhancements in ATLAS code by identifying the most common operations, with the call parameters well understood, and allowing improvements to be quantified in detail.

Poster presentations / **374**

High-Level Trigger Performance for Calorimeter based algorithms during LHC Run 1 data taking period

Authors: Alex Mann[1]; Alexander Mann[2]; Denis Oliveira Damazio[3]

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

[2] *Ludwig-Maximilians-Universität*

[3] *Brookhaven National Laboratory (US)*

Corresponding Authors: mann@cern.ch, a.mann@lmu.de

The ATLAS detector operated during the three years of the run 1 of the Large Hadron Collider collecting information on a large number of proton-proton events. One the most important results obtained so far is the discovery of one Higgs boson. More precise measurements of this particle must be performed as well as there are other very important physics topics still to be explored. One of the key components of the ATLAS detector is its trigger system. It is composed of three levels : one (called Level 1 - L1) built on custom hardware and the two others based on software algorithms - called Level 2 (L2) and Event Filter (EF) – altogether referred to as the ATLAS High Level Trigger. The ATLAS trigger is responsible for reducing almost 20 million of collisions per second produced by the accelerator to less than 1000. The L2 operates only in the regions tagged by the first hardware level as containing possible interesting physics while the EF

operates in the full detector, normally using offline-like algorithms to reach a final decision about recording the event.
Amongst the ATLAS subdetectors, there is a complex set of calorimeter specialized to detect and measure the energy of electrons, photons, taus, jets and even measure global event missing transverse energy.
The present work describes the performance of the ATLAS High-Level Calorimeter Trigger. Algorithms for detecting electrons and taus were able to reconstruct clusters of calorimeter cells, measure their energy and shower shape variables for particle classification with quite high efficiency. Since the beginning of the operation only minor modification on cut values were necessary given the wide range of instantaneous luminosity explored (over 5 order of magnitude). Another class of algorithms studies jets and can detect and estimate the energy of these using different cone sizes and jet finding techniques. Finally, an algorithm only for the Event Filter was design to measure with high precision the total transverse energy using only cells that are above their own noise level.
As the luminosity increased, it was fundamental to create an option for reducing the rate of events from the L1 into for EF missing E_{T} chains. So, the ATLAS L2 trigger operating paradigm of only using regions around the L1 based had to be broken for the special case of the missing E_{T}. The paper describes the parallel chain built in the ATLAS data acquisition system to accomplish this task. Also, it seemed increasingly important to work on a similar topic for the L2 jets. The usage of full detector information so that jets could be found in full scans of the detector - not limited by the L2 regions - is quite relevant for the efficiency of multiple jet algorithms. The option taken will also be described in the paper. Finally, we will describe the future plans of operation for the LHC run 2 data taking period.

Poster presentations / 197

Experience with a frozen computational framework from LEP age

Author: Rafael Marco De Lucas[1]

Co-authors: David Rodriguez [2]; Jesus Marco [3]; Miguel Angel Nuñez [4]

[1] *IFCA (CSIC-UC) Santander SPAIN*

[2] *IFCA (CSIC-UC) (now at University of Edinburgh)*

[3] *IFCA (CSIC-UC) Santander Spain*

[4] *IFCA (CSIC-UC), Santander SPAIN*

Corresponding Authors: marco@ifca.unican.es, rmarco@ifca.unican.es

The strategy at the end of the LEP era for the long term preservation of physics results and data processing framework was not obvious.
One of the possibilities analyzed at the time, previously to the generalization of virtualization techniques, was the setup of a dedicated farm, to be conserved in its original state for medium-long term, at least until the new data from LHC could indicate the need to reanalyze LEP data, the most significant example being the Higgs boson search.
Such an infrastructure was setup at IFCA in 2001, including 80 equal servers where the software of the DELPHI experiment was installed and tested, and analysis ntuples and code were stored. This set of servers have been periodically restarted and tested, to examine the feasibility of this approach for complete preservation, and allow a detailed comparison with the approach based on the use of virtual machines.

In parallel, all DELPHI data at DST (Data Summary Tapes) level were copied to IFCA and stored in tape in LTO-1 format.

The latest results at LHC indicate that there will be likely no need to reanalyze LEP data.

This contribution describes this experience, the results obtained after more than 10 years of "freezing", and concludes with the lessons learnt in this cycle across two generation of experiments.

Poster presentations / **57**

CDS Multimedia Services and Export

Author: Ludmila Marian[1]

Co-author: Jean-Yves Le Meur [1]

[1] *CERN*

Corresponding Authors: ludmila.marian@cern.ch, jean-yves.le.meur@cern.ch

The volume of multimedia material produced by CERN is growing rapidly, fed by the increase of dissemination activities carried out by the various outreach teams, such as the central CERN Communication unit and the Experiments Outreach committees. In order for this multimedia content to be stored digitally for the long term, to be made available to end-users in the best possible conditions and finally to be easily re-usable in various contexts e.g. by web site creators or by media professionals, various new services and export mechanisms have been set up in the CERN Document Server Multimedia collections.

This talk will explain which technologies have been selected, the challenges satisfying all these needs together, and how users can take advantage of the newly developed facilities. In particular we will present the developments in video and image embedding, image slideshow generation, embedding via oEmbed, image processing, and image and video retrieval and ranking. The underlying architecture of the CDS multimedia platform (Invenio based), responsible for ensuring a robust, flexible and scalable service will also be described.

Event Processing, Simulation and Analysis / **291**

Simulation of Pile-up in the ATLAS Experiment

Author: Zachary Louis Marshall[1]

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

Corresponding Author: zach.marshall@cern.ch

In the 2011/12 data the LHC provided substantial multiple proton-proton collisions within each filled bunch-crossing and also multiple filled bunch-crossings within the sensitive time window of the ATLAS detector. This will increase in the near future during the run beginning in 2015. Including these effects in Monte Carlo simulation poses significant computing challenges. We present a description of the standard approach used by the ATLAS experiment and details of how we manage the conflicting demands of keeping the background dataset size as small as possible while minimizing the effect of background event re-use. We also present details of the methods used to minimize the memory footprint of these digitization jobs, to keep them within the grid limit, despite combining the information from thousands of simulated events at once. We also describe an alternative approach, known as Overlay, where the actual detector conditions are sampled from raw data using a special zero-bias trigger, and the simulated physics events are overlaid on top of this zero-bias data. This

gives a realistic simulation of the detector response to physics events. The overlay simulation runs in time linear in the number of events and consumes memory proportional to the size of a single event, with small overhead.

Poster presentations / **372**

Performance and development plans for the Inner Detector trigger algorithms at ATLAS

Author: Emily Laura Nurse[1]

[1] *University of London (GB)*

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

We present a description of the algorithms and the performance of the ATLAS Inner Detector trigger for LHC run I, as well as prospects for a redesign of the tracking algorithms in run 2. The Inner Detector trigger algorithms are vital for many trigger signatures at ATLAS. The performance of the algorithms for muons, electrons, taus and b-jets is presented.
The ATLAS trigger software after will be restructured from 2 software levels into a single stage which poses a big challenge on the trigger algorithms in terms of execution time and maintaining the physics performance. Expected future improvements in the timing and efficiencies of the Inner Detector triggers are discussed, utilising the planned merging of the current two-stage software of the ATLAS trigger.

Poster presentations / **358**

MICE Data Handling on the Grid

Author: Janusz Martyniak[1]

[1] *Imperial College London*

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.

The configuration/condition of the experiment during each run is stored in a Postgres-based "Configuration Database", that has a read-only replica with a publicly-accessible web-services interface. Meanwhile the raw data from the DAQ (based on the DATE framework from ALICE) is written to a storage system close to the Control Room as a series of tarballs, one per run, each containing checksum information about the contents.

There are two main data handling projects for the MICE experiment which involve data distribution on the Grid:
• The RAW Data Mover
• Off-line and batch data reconstruction

The aim of the Data Mover is to upload RAW data files onto a safe tape storage as soon as the data have been written out by the DAQ system and marked as ready to be uploaded. The Data Mover actively watches the directories where new RAW files are to appear and copies them to an intermediate disk storage. This step is useful mainly to avoid bottlenecks in access to the DAQ disk, to which the DAQ has write priority.

After the initial copy is made and internal file integrity is verified, each file is uploaded to RAL Tier1 Castor Storage Element (SE) and placed on 2 tapes for redundancy. We also make another copy at a separate disk-based SE at this stage to make it easier for users to access data quickly. Both copies are check-summed and the replicas are registered with an instance of LCG File Catalog (LFC). On success a record with basic file properties is added to the Mice Metadata DB.

MICE standard reconstruction software (MAUS) has been installed on the Grid. The reconstruction process is triggered by new RAW data records filled in the Mice Metadata DB by the mover system described above. Off line reconstruction jobs for new RAW files are submitted to RAL Tier1 and the output is stored on a Castor disk. We are currently working on distributed submissions to MICE enabled Tier 2 sites . This is mainly aimed for data reprocessing when a new version of MAUS becomes available. It this case output files will be shipped back to RAL using File Transfer Service (FTS) based system.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / 104

The Common Analysis Framework Project

Authors: Alexei Klimentov[1]; Ian Fisk[2]; Maria Girone[3]

Co-author: Marco Mascheroni [4]

[1] *Brookhaven National Laboratory (US)*

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

[3] *CERN*

[4] *Universita & INFN, Milano-Bicocca (IT)*

Corresponding Authors: marco.mascheroni@cern.ch, maria.girone@cern.ch, ian.fisk@cern.ch

ATLAS, CERN-IT, and CMS embarked on a project to develop a common system for submitting analysis jobs to the distributed computing infrastructure based on elements of PANDA. After an extensive feasibility study and development of a proof-of-concept prototype, the project has a basic infrastructure that can be used to support the analysis use case of both experiments with common services. In this presentation we will discuss the process for organizing a common project between large stakeholders, the current state of the proof-of-concept prototype, and the operations plans to deploy a common service.

Poster presentations / 111

CMS users data management service integration and first experiences with its NoSQL data storage

Author: Hassen Riahi[1]

Co-authors: Ian Fisk [2]; Marco Mascheroni [3]

[1] *Universita e INFN (IT)*

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

[3] *Universita & INFN, Milano-Bicocca (IT)*

Corresponding Authors: marco.mascheroni@cern.ch, hassen.riahi@cern.ch, ian.fisk@cern.ch

The distributed data analysis workflow in CMS assumes that jobs run in a different location to where their results are finally stored. Typically the user outputs must be transferred from one site to another by a dedicated CMS service, AsyncStageOut. This new service is originally developed to address

the inefficiency in using the CMS computing resources when transferring the analysis job outputs, synchronously, once they are produced in the job execution node to the remote site.

The AsyncStageOut is designed as a thin application relying only on the NoSQL database (CouchDB) as input and data storage. It has progressed from a limited prototype to a highly adaptable service which manages and monitors the whole user files steps, namely file transfer and publication.

The AsyncStageOut is integrated with the Common CMS/Atlas Analysis Framework. It foresees the management of nearly 200k users files per day of close to 1000 individual users per month with minimal delays, and providing a real time monitoring and reports to users and service operators, while being highly available. The associated data volume represents a new set of challenges in the areas of database scalability and service performance and efficiency. In this paper, we present an overview of the AsyncStageOut model and the integration strategy with the Common Analysis Framework. The motivations for using the NoSQL technology are also presented, as well as data design and the techniques used for efficient indexing and monitoring of the data. We describe deployment model for the high availability and scalability of the service. We also discuss the hardware requirements and the results achieved as they were determined by testing with actual data and realistic loads during the commissioning and the initial production phase with the Common Analysis Framework.

Poster presentations / **75**

Optimising query execution time in LHCb Bookkeeping System using partition pruning and partition wise joins

Author: Zoltan Mathe[1]

Co-author: Philippe Charpentier [1]

[1] *CERN*

Corresponding Author: zoltan.mathe@cern.ch

The LHCb experiment produces a huge amount of data which has associated metadata such as run number, data taking condition (detector status when the data was taken), simulation condition, etc. The data are stored in files, replicated on the Computing Grid around the world. The LHCb Bookkeeping System provides methods for retrieving datasets based on their metadata. The metadata is stored in a hybrid database model, which is a mixture of Relational and Hierarchical database models and is based on the Oracle Relational Database Management System (RDBMS). The database access has to be reliable and fast. In order to achieve a high timing performance, the tables are partitioned and the queries are executed in parallel. When we store large amounts of data the partition pruning is essential for database performance, because it reduces the amount of data retrieved from the disk and optimises the resource utilisation. This research presented here is focusing on the extended composite partitioning strategy such as range-hash partition, partition pruning and usage of the partition wise joins. The system has to serve thousands of queries per minute, the performance and capability of the system is measured when the above performance optimisation techniques are used.

Poster presentations / **196**

INFN Pisa scientific computation environment (GRID HPC and Interactive analysis)

Author: Enrico Mazzoni[1]

Co-authors: Alberto Ciampa [2]; Andrea Carboni [3]; Silvia Arezzini [4]; Simone Coscetti [5]; Simone Piras [6]; Tommaso Boccali [5]

[1] *INFN-Pisa*

[2] *Universita degli Studi di Pisa-INFN, Sezione di Pisa*

[3] *Unknown*

[4] *Univ. + INFN*

[5] *Sezione di Pisa (IT)*

[6] *INFN Pisa*

Corresponding Author: enrico.mazzoni@pi.infn.it

The INFN-Pisa Tier2 infrastructure is described, optimized not only for GRID CPU and Storage access, but also for a more interactive use of the resources in order to provide good solutions for the final data analysis step. The Data Center, equipped with about 5000 production cores, permits the use of
modern analysis techniques realized via advanced statistical tools (like RooFit and RooStat) implemented in multi core systems.
In particular a POSIX file storage access integrated with standard srm access is provided. Therefore the unified storage infrastructure is described, based on GPFS and Xrootd, used both for SRM data repository and interactive POSIX access. Such a common infrastructure allows a transparent access to the Tier2 data to the users for their interactive analysis.
The organization of a specialized many cores CPU facility devoted to interactive analysis is also described along with the login mechanism integrated with the INFN-AAI (National INFN Infrastructure) to extend the site access and use to a geographical distributed community.
Such infrastructure is used also for a national computing facility in use to the INFN theoretical community, it enables a synergic use of computing and storage resources. Our Center initially developed for the HEP community is now growing and includes also HPC resources fully integrated. In recent years has been installed and managed a cluster facility (1000 cores, parallel use via InfiniBand connections) and we are now updating this facility that will provide resources for all the intermediate level HPC computing needs of the theoretical national INFN community.

Facilities, Infrastructures, Networking and Collaborative Tools / 154

Deployment of a WLCG network monitoring infrastructure based on the perfSONAR-PS technology

Authors: Aaron Brown[None]; Alessandra Forti[1]; Alvaro Fernandez Casani[2]; Anthony Hesnaux[3]; Daniele Bonacorsi[4]; Donato De Girolamo[5]; Duncan Rand[6]; Fernando Lopez Munoz[7]; Ian Gable[8]; Jason Zurawski[9]; Jose Flix Molina[10]; Kashif Mohammad[11]; Marek Zielinski[12]; Mario Reale[13]; Nicolo Magini[3]; Oliver Gutsche[14]; Shawn Mc Kee[15]; Si Liu[14]; Simone Campana[3]; Stefan Roiser[3]; Vincenzo Capone[16]

[1] *University of Manchester (GB)*

[2] *Universidad de Valencia (ES)*

[3] *CERN*

[4] *University of Bologna*

[5] *INFN*

[6] *Imperial College*

[7] *PIC*

[8] *University of Victoria (CA)*

[9] *Internet2*

[10] *Centro de Investigaciones Energ. Medioambientales y Tecn. - (ES*

[11] *Aligarh Muslim University (IN)*

[12] *University of Rochester (US)*

[13] *GARR*

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

[15] *University of Michigan (US)*

[16] *Universita e INFN (IT)*

Corresponding Authors: simone.campana@cern.ch, shawn.mckee@cern.ch

The WLCG infrastructure moved from a very rigid network topology, based on the MONARC model, to a more relaxed system, where data movement between regions or countries does not necessarily need to involve T1 centers. While this evolution brought obvious advantages, especially in terms of flexibility for the LHC experiment's data management systems, it also opened the question of how to monitor the increasing number of possible network paths, in order to provide a global reliable network service. The perfSONAR network monitoring system has been evaluated and agreed as a proper solution to cover the WLCG network monitoring use cases: it allows WLCG to plan and execute latency and bandwidth tests between any instrumented endpoint through a central scheduling configuration, it allows archiving of the metrics in a local database, it provides a programmatic and a web based interface exposing the tests results; it also provides a graphical interface for remote management operations. In this contribution we will present our activity to deploy a perfSONAR based network monitoring infrastructure, in the scope of the WLCG Operations Coordination initiative: we will motivate the main choices we agreed in terms of configuration and management, describe the additional tools we developed to complement the standard packages and present the status of the deployment, together with the possible future evolution.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / 119

Running jobs in the Vacuum

Author: Andrew McNab[1]

Co-authors: Federico Stagni [2]; Mario Ubeda Garcia [2]

[1] *University of Manchester (GB)*

[2] *CERN*

Corresponding Author: andrew.mcnab@cern.ch

We present a model for the operation of computing nodes at a site using virtual machines, in which the virtual machines (VMs) are created and contextualised for virtual organisations (VOs) by the site itself. For the VO, these virtual machines appear to be produced spontaneously "in the vacuum" rather than in response to requests by the VO. This model takes advantage of the pilot job frameworks adopted by many VOs, in which pilot jobs submitted via the grid infrastructure in turn start job agents which fetch the real jobs from the VO's central task queue. In the vacuum model, the contextualisation process starts a job agent within the virtual machine and real jobs are fetched from the central task queue as normal. This is similar to ongoing cloud work where job agents are also run inside virtual machines, but where VMs are created by the virtual organisation itself using cloud APIs.

An implementation of the vacuum scheme, vac, is presented in which a VM factory runs on each physical worker node to create and contextualise its set of virtual machines.

With this system, each node's VM factory can decide which VO's virtual machines to run, based on site-wide target shares and on a peer-to-peer protocol in which the site's VM factories query each other to discover which virtual machine types they are running, and

therefore identify which virtual organisations' virtual machines should be started as nodes become available again, and which virtual organisations' virtual machines should be signaled to shut down. A property of this system is that there is no gate keeper service, head node, or batch system accepting and then directing jobs to particular worker nodes, avoiding several central points of failure.

Finally, we describe tests of the vac system using jobs from the central LHCb task queue, using the same contextualisation procedure for virtual machines developed by LHCb for clouds.

Facilities, Infrastructures, Networking and Collaborative Tools / 17

Testing as a Service with HammerCloud

Authors: Andrea Sciaba[1]; Daniel van der Ster[1]; Federica Legger[2]; Francesco Giovanni Sciacca[3]; Johannes Elmsheuser[2]; Quentin Barrand[None]; Ramon Medrano Llamas[1]

[1] *CERN*

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

[3] *Universitaet Bern (CH)*

Corresponding Author: ramon.medrano@cern.ch

HammerCloud was designed and born under the needs of the grid community to test the resources and automate operations from a user perspective. The recent developments in the IT space propose a shift to the software defined data centers, in which every layer of the infrastructure can be offered as a service.

Testing and monitoring is an integral part of the development, validation and operations of big systems, like the grid. This area is not escaping the paradigm shift and we are starting to perceive as natural the Testing as a Service (TaaS) offerings, which allow to test any infrastructure service, such as the Infrastructure as a Service (IaaS) platforms being deployed in many grid sites, both from the functional and stress perspectives.

This work will review the recent developments in HammerCloud and its evolution to a TaaS conception, in particular its deployment on the Agile Infrastructure platform at CERN and the testing of many IaaS providers across Europe in the context of experiment requirements. The first section will review the architectural changes that a service running in the cloud needs, such an orchestration service or new storage requirements in order to provide functional and stress testing. The second section will review the first tests of infrastructure providers on the perspective of the challenges discovered from the architectural point of view. Finally, the third section will evaluate future requirements of scalability and features to increase testing productivity.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / 86

Commissioning the CERN IT Agile Infrastructure with experiment workloads

Authors: Fernando Harald Barreiro Megino[1]; Katarzyna Kucharczyk[2]; Marek Kamil Denis[2]; Mattia Cinquilli[1]; Ramon Medrano Llamas[1]

[1] *CERN*

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

Corresponding Authors: ramon.medrano@cern.ch, mattia.cinquilli@cern.ch

In order to ease the management of their infrastructure, most of the WLCG sites are adopting cloud based strategies. In the case of CERN, the Tier 0 of the WLCG, is completely restructuring the resource and configuration management of their computing center under the codename Agile Infrastructure. Its goal is to manage 15,000 Virtual Machines by means of an OpenStack middleware in order to unify all the resources in CERN's two datacenters: the one placed in Meyrin and the new on in Wigner, Hungary.

During the commissioning of this infrastructure, CERN IT is offering an attractive amount of computing resources to the experiments (800 cores for ATLAS and CMS) through a private cloud interface. ATLAS and CMS have joined forces to exploit them by running stress tests and simulation workloads since November 2012.

This work will describe the experience of the first deployments of the current experiment workloads on the CERN private cloud testbed. The paper is organized as follows: the first section will explain the integration of the experiment workload management systems (WMS) with the cloud resources. The second section will revisit the performance and stress testing performed with HammerCloud in order to evaluate and compare the suitability for the experiment workloads. The third section will go deeper into the dynamic provisioning techniques, such as the use of the cloud APIs directly by the WMS. The paper finishes with a review of the conclusions and the challenges ahead.

Poster presentations / 84

Helix Nebula and CERN: A Symbiotic Approach to Exploiting Commercial Clouds

Authors: Bob Jones[1]; Daniel van der Ster[1]; Katarzyna Kucharczyk[2]; Ramon Medrano Llamas[1]

[1] *CERN*

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

Corresponding Authors: ramon.medrano@cern.ch, robert.jones@cern.ch, daniel.vanderster@cern.ch

The recent paradigm shift toward cloud computing in IT, and general interest in "Big Data" in particular, have demonstrated that the computing requirements of HEP are no longer globally unique. Indeed, the CERN IT department and LHC experiments have already made significant R&D investments in delivering and exploiting cloud computing resources. While a number of technical evaluations of interesting commercial offerings from global IT enterprises have been performed by various physics labs, further technical, security, sociological, and legal issues need to be address before their large-scale adoption by the research community can be envisaged.

Helix Nebula - the Science Cloud is an initiative that explores these questions by joining the forces of three European research institutes (CERN, ESA and EMBL) with leading European commercial IT enterprises. The goals of Helix Nebula are to establish a cloud platform federating multiple commercial cloud providers, along with new business models, which can sustain the cloud marketplace for years to come.

This contribution will summarize the participation of CERN in Helix Nebula. We will explain CERN's flagship use-case and the model used to integrate several cloud providers with an LHC experiment's workload management system. During the first proof of concept, this project contributed over 40.000 CPU-days of Monte Carlo production throughput to the ATLAS experiment with marginal manpower required. CERN's experience, together with that of ESA and EMBL, is providing a great insight into the cloud computing industry and highlighted several challenges that are being tackled in order to ease the export of the scientific workloads to the cloud environments.

Summaries / 525

Summary of track 6

Corresponding Author: helge.meinhard@cern.ch

Poster presentations / 465

Integrating the Network into LHC Experiments: Update on the ANSE (Advanced Network Services for Experiments) Project

Authors: Alan Tackett[1]; Andrew Malone Melo[2]; Artur Jerzy Barczyk[3]; Azher Mughal[3]; Ben Meekhof[4]; Harvey Newman[3]; Jorge Batista[4]; Kaushik De[5]; Paul Sheldon[2]; Ramiro Voicu[3]; Robert Ball[6]; Tony Wildish[7]

[1] *Vanderbilt University*

[2] *Vanderbilt University (US)*

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

[4] *University of Michigan*

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

[6] *University of Michigan (US)*

[7] *Princeton University (US)*

Corresponding Author: andrew.malone.melo@cern.ch

The LHC experiments have always depended upon a ubiquitous, highly-performing network infrastructure to enable their global scientific efforts. While the experiments were developing their software and physical infrastructures, parallel development work was occurring in the networking communities responsible for interconnecting LHC sites. During the LHC's Long Shutdown \#1 (LS1) we have an opportunity to incorporate some of these network developments into the LHC experiment's software.

The ANSE (Advance Network Services for Experiments) Project, an NSF CC-NIE funded effort, is targeting the creation, testing and deployment of a set of software tools (network services and an associated API) suitable for use by the LHC experiments. This library will explicitly enable obtaining both knowledge of the networks and provide mechanisms for interacting with those networks. The project will leverage existing infrastructure from DYNES\cite{dynes}, AutoBahn\cite{autobahn}, perfSONAR-PS\cite{psps}, OpenFlow\cite{OpenFlow}, MonALISA\cite{MonaLISA} and Software Defined Networking\cite{SDN} as well as experiment-specific applications to enable these capabilities.

We will report on the progress we have made to date, present our initial architecture and show some examples of the kinds of functionality we are developing in the context of the ATLAS and CMS experiments.

@misc{dynes,
title = "{{MRI}-{R}2 {C}onsortium: {D}evelopment of {D}ynamic {N}etwork {S}ystem ({DYNES})}",
key = "dynes",
howpublished = "\url{http://www.internet2.edu/ion/dynes.html}"
}

@misc{autobahn,
title = "{{GEANT}2 Auto{BAHN}}",
key = "autobahn",
howpublished = "\url{http://www.geant2.net/server/show/ConWebDoc.2544}"
}

@misc{fdt,
title = "{Fast Data Transfer (FDT)}",

key = “fdt”,
howpublished = “\url{http://monalisa.cern.ch/FDT}”
}

@misc{psps,
title = “{perf{SONAR}-PS}”,
key = “psps”,
howpublished = “\url{http://psps.perfsonar.net}”
}

@misc{SDN,
title = “{{S}oftware {D}efined {N}etworking}”,
key = “SDN”,
howpublished = “\url{http://en.wikipedia.org/wiki/Software-defined_networking}”
}

@misc{OpenFlow,
title = “{OpenFlow Layer 2 Communications Protocol}”,
key = “OpenFlow”,
howpublished = “\url{http://en.wikipedia.org/wiki/OpenFlow}”
}
@misc{MonALISA,
title = “{MonALISA: An agent based, dynamic service system to monitor, control and optimize distributed system}”,
key = “MonALISA”,
howpublished = “\url{http://dx.doi.org/10.1016/j.bbr.2011.03.031}”
}

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / 147

ArbyTrary, a cloud-based service for low-energy spectroscopy

Author: Dario Menasce[1]

Co-authors: Ezio Previtali [2]; Massimiliano Clemenza [3]; Massimiliano Nastasi [1]; Oliviero Cremonesi [4]; Silvia Capelli [1]; maura pavan [5]

[1] *INFN Milano-Bicocca*

[2] *I*

[3] *Università di Milano Bicocca*

[4] *INFN*

[5] *Universita’ di Milano Bicocca*

Corresponding Author: dario.menasce@mib.infn.it

Radiation detectors usually require complex calibration procedures in order to provide reliable activity measurements. The Milano-Bicocca group has developed, over the years, a complex simulation tool, based on GEANT4, that provide the functionality required to compute the correction factors necessary for such calibrations in a broad range of use-cases, considering various radioactive source types in different geometrical arrangements, paired with a large variety of possible detector setups. This simulation tool, so far, has only been used interactively in a command-line mode with very limited flexibility for what concerns user interaction, input and output. We present a prototype of a novel approach to the simulation, via a dedicated web-server, provided with a sophisticated web interface based on a public-domain standard Javascript framework (EXTJS). The current prototype features a desktop-like environment placed in a browser window, allowing users to drive the simulation as if they were running interactively on their desktop/laptop computers. The actual code, instead, runs in a remote, cloud-like environment, shielding users from the complexity of the underlying computing structure. The project aims at developing an infrastructure that allows a complete decoupling between the simulation code (a compiled, static framework with a library of plugin services) and the graphical user interface. By adopting the MVC (Model View Controllers) Architecture

we were able to provide this system with a high degree of flexibility, making it suitable for a generic interface to a service of simulation programs hosted on a cloud-based system.

Poster presentations / 44

Migration of the CERN IT Data Center Support System to ServiceNow

Authors: Anthony Grossir[1]; Eric Bonfillou[1]; Fabio Trevisani[1]; Gregory Arneodo[1]; Ivan Fedorko[1]; John Hefferman[1]; Mats Moller[1]; Miguel Coelho dos Santos[1]; Olof Barring[1]; Omar Pera Mira[2]; Patricia Mendez Lorenzo[1]; Roberto Alvarez Alonso[3]; Veronique Lefebure[1]; Vincent Dore[1]; Wayne Salter[1]; Zhechka Toteva[1]

[1] *CERN*

[2] *Valencia Polytechnic University (ES)*

[3] *Universidad de Oviedo (ES)*

Corresponding Author: patricia.mendez@cern.ch

The large potential and flexibility of the ServiceNow infrastructure based on "best practices" methods is allowing the migration of some of the ticketing systems traditionally used for the tracing of the servers and services available at the CERN IT Computer Center. This migration enables a standardization and globalization of the ticketing and control systems implementing a generic system extensible to other departments and users. One of the activities of the Service Management project together with the IT-CF group has been the migration of the ITCM structure based in Remedy to ServiceNow within the context of one of the ITIL processes called Event Management.
From the technical point of view, the adaptation of ServiceNow to the well known ITCM Remedy based structure has been instrumental to ensure a controlled and smooth migration. In addition, the versatility of ServiceNow has allowed the implementation of generic facilities based in best practices processes to extend the potential provided by ITCM including Service Level Management facilities for the tracking of external contractors , specific forms, etc.
From the operational point of view a complete reporting infrastructure in ServiceNow has been created with a particular emphasis on the end users training.
The migration has meant a first step towards the interface with other systems (such as SDB and Agile) and a generalization of the Event Management process to all services at CERN.
In this talk we will present the full architecture, the workflow and the facilities implemented in ServiceNow for this migration. The experience gained during the first months of operation will be discussed together with the interface of the current Event Management structure to other monitoring systems including Spektrum and the future Agile structure and other service databases such as SDB and CDB. The usage of this structure is extended to the service tracking at the Wigner Center in Budapest.

Poster presentations / 179

Archival Services and Technologies for Scientific Data

Author: Jörg Meyer[1]

Co-authors: Achim Streit [2]; Jos Van Wezel [2]; Marcus Hardt [3]

[1] *KIT - Karlsruher Institute of Technology*

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

[3] *Kalrsruhe Institute of Technology*

Corresponding Author: joerg.meyer2@kit.edu

After analysis and publication, there is no need to keep experimental data online on spinning disks. For reliability and costs inactive data is moved to tape and put into a data archive. The data archive

must provide reliable access for at least ten years following a recommendation of the German Science Foundation (DFG), but many scientific communities wish to keep data available much longer. Data archival is on the one hand purely a bit preservation activity in order to ensure the bits read are the same as those written years before. On the other hand enough information must be archived to be able to use and interpretate the content of the data. The latter is depending on many also community specific factors and remains an areas of much debate among archival specialists.
The paper describes the current practice of archival and bit preservation in use for different science communities at KIT for which a combination of organisational services and technical tools are required. The special monitoring to detect tape related errors, the software infrastructure in use as well as the service certification are discussed. Plans and developments at KIT also in the context of the Large Scale Data Management and Analysis (LSDMA) project are presented. The technical advantages of the T10 SCSI Stream Commands (SSC-4) and the Linear Tape File System (LTFS) will have a profound impact on future long term archival of large data sets.

Event Processing, Simulation and Analysis / **56**

A Common Partial Wave Anaylsis Framework for PANDA

Author: Mathias Michel[1]

Co-authors: Anastasia Karavdina [2]; Bertram Kopf [3]; Florian Feldbauer [4]; Klaus Goetzen [5]; Klaus Peters [6]; Matthias Steinke [7]; Miriam Fritsch [8]; Prometeusz Jasinski [4]

[1] *Helmholtz-Institut Mainz*

[2] *University Mainz*

[3] *Ruhr-Universität Bochum*

[4] *Universität Mainz*

[5] *GSI Darmstadt*

[6] *Institut fuer Experimentalphysik I*

[7] *RUHR-UNIVERSITäT BOCHUM*

[8] *Universitaet Mainz*

Corresponding Author: michel@kph.uni-mainz.de

A large part of the physics program of the PANDA experiment at FAIR deals with the search for new conventional and exotic hadronic states like e.g. hybrids and glueballs. In a majority of analyses PANDA will need a Partial Wave Analsis (PWA) to identify possible candidates and for the classification of known states. Therefore, a new, agile and efficient PWA-Framework will be developed. It will be modularized to provide easy extension with models and formalisms as well as fitting of multiple datasets, even from different experiments. Experience from existing PWA programs was used to fix the requirements of the framwork and to prevent it from restrictions. It will provide various estimation and optimization routines like Minuit2 and the Geneva library. The challenges involve parallelization, fitting with a high number of free parameters, managing complex meta-fits and quality assurance / comparability of fits. To test the software, it will be used with data from running experiments like BaBar or BESIII. The presentation will show the status of the framework implementation as well as first tests.

Data Stores, Data Bases, and Storage Systems / **376**

dCache: Big Data storage for HEP communities and beyond

Authors: Albert Rossi[1]; Antje Petersen[2]; Christian Bernardt[3]; Dmitry Litvintsev[1]; Karsten Schwank[None]; Patrick Fuhrmann[2]; Paul Millar[3]; Tigran Mkrtchyan Mkrtchyan[4]

[1] *FNAL*

[2] *DESY*

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

[4] *Deutsches Elektronen-Synchrotron DESY*

Corresponding Author: paul.millar@desy.de

Storage is a continually evolving environment, with new solutions to both existing problems and new challenges. With over ten years in production use, dCache is also evolving to match this changing landscape. In this paper, we present three areas in which dCache is matching demand and driving innovation.

Providing efficient access to data that maximises both streaming and random-access workloads has always been a challenge. Various proprietary protocols, such as dcap and xrootd, have been developed to meet this challenge. Maintaining a property protocol incurs costs as the client must be written and supported on different platforms. With the arrival of NFS v4.1/pNFS, there is an industry standard that maximises throughput, so the need for property protocols has gone. Realising this, the dCache team has been involved with NFS v4.1 standard since its initial draft stage and all available dCache versions provide NFS support. We present the results of running NFS in production environment.

Managing storage is an evolving field. Formally, the SRM protocol was key; however, more recently HEP communities have shunned SRM in favour of other protocols, such as WebDAV. In a related move, various groups have investigated using cloud infrastructures, either wholly or using a cloud-bursting model to satisfy surges in demand. While continuing to support WebDAV, dCache is introducing support for CDMI, the ISO standard protocol for managing cloud storage. In common with WebDAV, CDMI is based on HTTP but provides better support for common management operations. In this paper, we present a comparison between WebDAV and CDMI, the current status of support in dCache, early results from experiments and future plans for this protocol.

Finally, one constant complaint from people using the grid is related to X.509 certificates. Various solutions have been presented, including EMI's Security Token Service to allow portal-like access to grid resources and to simplify the "grid login" step. Within Germany, the LSDMA project is investigating a more radical solution. In addition to the established grid methods, storage services may be
accessed directly using an identity asserted by the user's home institute. This would allow scientists to use dCache securely by logging in with the user-name and password from their home university or research institute. Details of this development work are presented along with possible deployment scenarios.

Poster presentations / 183

Monitoring in a grid cluster

Author: David Crooks[1]

Co-authors: David Britton [1]; Gareth Roy [2]; Mark Mitchell [3]; Samuel Cadellin Skipsey ; Stuart Purdie [3]

[1] *University of Glasgow (GB)*

[2] *U*

[3] *University of Glasgow*

Corresponding Authors: mark.mitchell@glasgow.ac.uk, david.crooks@cern.ch

The monitoring of a grid cluster (or of any piece of reasonably scaled IT infrastructure) is a key element in the robust and consistent running of that site. There are several factors which are important to the selection of a useful monitoring framework, which include ease of use, reliability,

data input and output. It is critical that data can be drawn from different instrumentation packages and collected in the framework to allow for a uniform view of the running of a site. It is also very useful to allow different views and transformations of this data to allow its manipulation for different purposes, perhaps unknown at the initial time of installation. In this context, we firstly present the findings of an investigation of the Graphite monitoring framework and its use at the Scotgrid Glasgow site. In particular, we examine the messaging system used by the framework and means to extract data from different tools, including the existing framework Ganglia which is in use at many sites, in addition to adapting and parsing data streams from external monitoring frameworks and websites. We also look at different views in which the data can be presented to allow for different use cases. We report on the installation and maintenance of the framework from a system manager perspective.

This work is relevant to site managers and anyone interested in high level, adaptable site monitoring.

Plenaries / **516**

Inside numerical weather forecasting - Algorithms, domain decomposition, parallelism

Author: Toon Moene[1]

[1] *KNMI*

Weather forecasting is both one of the most visible as well as one of the more demanding applications of computing in the world we know today. The development of forecasting models draws heavily on parallelization and efficient exploitation of many-core systems to get predictions done in near real-time. Because the computational domain is very large when using high resolution models, domain decomposition over many nodes using distributed memory parallel programming paradigms are a necessity.

Toon Moene of the Royal Dutch Meteorological Institute KNMI will tell us how modern models such as HIRLAM and HARMONIE exploit advances in computing, having run these models also on his home PC - catching several meteorological errors in the process. Toon Moene is a researcher at KNMI, member of the Technical Advisory Committee to the Council of the European Centre for Medium Range Weather Forecasts ECMWF, and member of the GCC Steering Committee and the Fortran Standardization Committee.

Poster presentations / **19**

Solving Small Files Problem in Enstore

Author: Alexander Moibenko[1]

Co-authors: Alexander Kulyavtsev [2]; Dmitry Litvintsev [2]; Gene Oleynik [3]; John Hendry [2]; Stan Naymola [2]

[1] *Fermi NAtiona Accelerator Laboratoy*

[2] *FNAL*

[3] *Fermilab*

Corresponding Author: moibenko@fnal.gov

Enstore is a tape based Mass Storage System originally designed for Run II Tevatron experiments at FNAL (CDF, D0). Over the years it has proven to be reliable and scalable data archival and delivery solution, which meets diverse requirements of variety of applications including US CMS Tier 1, High Performance Computing, Intensity Frontier experiments as well as data backups. Data intensive

experiments like CDF, D0 and US CMS Tier 1 generally produce huge amount of data stored in files with the average size of few Gigabytes, which is optimal for writing and reading data to/from tape. In contrast, much of the data produced by Intensity Frontier experiments, Lattice QCD and Cosmology is sparse, resulting in accumulation of large amounts of small files.
Reliably storing small files on tape is inefficient due to file marks writing which takes significant amount of the overall file writing time (few seconds). There are several ways of improving data write rates, but some of them are unreliable, some are specific to the type of tape drive and still do not provide transfer rates adequate to rates offered by tape drives (20% of the drives potential rate).
In order to provide good rates for small files in a transparent and consistent manner, the Small File Aggregation (SFA) feature has been developed to provide aggregation of files into containers which are subsequently written to tape. The file aggregation uses reliable internal Enstore disk buffer. File grouping is based on policies using file metadata and other user defined steering parameters.
If a small file, which is a part of a container, is requested for read, the whole container is staged into internal Enstore read cache thus providing a read ahead mechanism in anticipation of future read requests for files from the same container. SFA is provided as service implementing file aggregation and staging transparently to user.
The SFA is has been successfully used since April 2012 by several experiments. Currently we are preparing to scale up write/read SFA cache.
This paper describes Enstore Small Files Aggregation feature and discusses how it can be scaled in size and transfer rates.

Data Acquisition, Trigger and Controls / **72**

Prototype of a File-Based High-Level Trigger in CMS

Author: Remi Mommsen[None]

Co-authors: Andre Georg Holzner [1]; Andrea Petrucci [2]; Andrei Cristian Spataru [2]; Attila Racz [2]; Aymeric Arnaud Dupont [2]; Carlos Nunez Barranco Fernandez [2]; Christian Deldicque [2]; Christian Hartl [2]; Christoph Paus [3]; Christoph Schwick [2]; Christopher Colin Wakefield [4]; Dominique Gigi [2]; Emilio Meschi [2]; Fabian Stoeckli [3]; Frank Glege [2]; Frans Meijers [2]; Gerry Bauer [3]; Giovanni Polese [5]; Hannes Sakulin [2]; James Gordon Branson [1]; Jose Antonio Coarasa Perez [2]; Konstanty Sumorok [3]; Lorenzo Masetti [2]; Luciano Orsini [2]; Marc Dobson [2]; Marco Pieri [1]; Matteo Sani [1]; Olivier Chaze [2]; Olivier Raginel [3]; Petr Zejdl [2]; Robert Gomez-Reino Garrido [2]; Samim Erhan [6]; Sergio Cittolin [1]; Srecko Morovic [7]; Ulf Behrens [8]; Vivian O'Dell [9]; Wojciech Andrzej Ozga [10]

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

[2] *CERN*

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

[4] *Staffordshire University (GB)*

[5] *University of Wisconsin (US)*

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

[7] *Institute Rudjer Boskovic (HR)*

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

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

[10] *AGH University of Science and Technology (PL)*

Corresponding Author: remigius.mommsen@cern.ch

The DAQ system of the CMS experiment at the LHC is redesigned during the accelerator shutdown in 2013/14. To reduce the interdependency of the DAQ system and the high-level trigger (HLT), we investigate the feasibility of using a file-system-based HLT. Events of ˜1 MB size are built at the level-1 trigger rate of 100 kHz. The events are assembled by ˜50 builder units (BUs). Each BU writes the raw events at ˜2GB/s to a local file system shared with O(10) filter-unit machines (FUs) running the HLT code. The FUs read the raw data from the file system, select O(10ˆ-3) of the events, and write the selected events together with monitoring metadata back to a disk. This data is then aggregated over several steps and made available for offline reconstruction and online monitoring. We present the challenges, technical choices, and performance figures from the prototyping phase. In addition, the steps to the final system implementation will be discussed.

Plenaries / 485

Software defined networking and bandwidth-on-demand

Author: Inder Monga[1]

[1] *ESnet*

Networking is one of the important factors in getting physics done, and the flows between data sources, data centres and physicists have reached an unprecedented scale. To make the next step, the network itself has to become more flexible and a schedulable resource. In this talk dr. Inder Monga of ESNet will talk about software defined networking, the protocols and services to describe and configure networks dynamically, and how this can provide bandwidth 'on-demand'.

Data Stores, Data Bases, and Storage Systems / 82

Rethinking how storage services are delivered to end-users at CERN: prototyping a file sharing and synchronisation platform with ownCloud

Author: Jakub Moscicki[1]

[1] *CERN*

Corresponding Author: jakub.moscicki@cern.ch

Individual users at CERN are attracted by external file hosting services such as Dropbox. This trend may lead to what is know as the "Dropbox Problem": sensitive organization data stored on servers outside of corporate control, outside of established policies, outside of enforceable SLAs and in unknown geographical locations. Mitigating this risk also provides a good incentive to rethink how our storage services are delivered to end-users: a file syncing and sharing platform which would allow offline work for mobile devices, seamlessly integrate with major desktop environments and provide convenience and functionality of commercial competitors. This would not only allow us to stay aligned with the expectations of the users in the rapidly evolving area but ultimately it should allow us to keep the data under control. As the market of open source projects capable of meeting such requirements begins maturing we think it is a good moment to evaluate potential technologies. This work will present the outcome of evaluating ownCloud at CERN including functionality and scalability testing. The goal of the prototype is to understand if ownCloud may be used to serve a community of 10^4 users and provide storage space on par or exceeding the one offered by storage systems currently in production. We will also discuss how a file sharing and synchronisation platform could fit into existing service architecture at CERN: storage backends, shared filesystems and interfaces. Additionally, we evaluate how this platform could take advantage of emerging storage technologies.

350

Using Solid State Disk Array as a Cache for LHC ATLAS Data Analysis

Author: Richard Philip Mount[1]

Co-authors: Andrew Hanushevsky [2]; Wei Yang [1]

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

[2] *STANFORD LINEAR ACCELERATOR CENTER*

Corresponding Authors: richard.mount@slac.stanford.edu, yangw@slac.stanford.edu, abh@stanford.edu

User data analysis in high energy physics presents a challenge to spinning-disk based storage systems. The analysis is data intense, yet reads are small, sparse and covers a large volume of data files. It is also unpredictable due to users' response to storage performance. We describe here a system with an array of Solid State Disk as a non-conventional, standalone file level cache in front of the spinning disk storage to help improve the performance of LHC ATLAS user analysis at SLAC. The system uses a long period of data access records to make caching decisions. It can also use information from other sources such as a work-flow management system. We evaluate the performance of the system both in terms of caching and its impact on user analysis jobs. The system currently uses Xrootd technology, but the technique can be applied to any storage system.

Poster presentations / **312**

A browser based multi-user working environment for physicists

Authors: Christian Glaser[1]; Dennis Klingebiel[1]; Gero Müller[1]; Marcel Rieger[1]; Martin Erdmann[1]; Martin Urban[1]; Matthias Komm[1]; Robert Fischer[1]; Tobias Winchen[1]

[1] *RWTH Aachen University*

Corresponding Author: gero.mueller@physik.rwth-aachen.de

Many programs in experimental particle physics do not yet have a graphical interface, or demand strong platform and software requirements. With the most recent development of the VISPA project, we provide graphical interfaces to existing software programs and access to multiple computing clusters through standard web browsers. The scalable client-server system allows analyses to be performed in sizable teams, and disburdens the individual physicist from installing and maintaining a software environment. The VISPA graphical interfaces are implemented in HTML, JavaScript and extensions to the Python webserver. The webserver uses SSH and RPC to access user data, code and processes on remote sites. As example applications we present graphical interfaces for steering the reconstruction framework OFFLINE of the Pierre-Auger experiment, and the analysis development toolkit PXL. The browser based VISPA system was field-tested in bi-weekly homework of a third year physics course by more than 100 students. We discuss the system deployment and the evaluation by the students.

Data Acquisition, Trigger and Controls / **503**

Algorithms, performance, and development of the ATLAS High-level trigger

Author: Enrico Pasqualucci[1]

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

Corresponding Authors: kunihiro.nagano@cern.ch, enrico.pasqualucci@cern.ch

The ATLAS trigger system has been used for the online event selection for three years of LHC data-taking and is preparing for the next run. The trigger system consists of a hardware level-1 (L1) and a software high-level trigger (HLT). The high-level trigger is currently implemented in a region-of-interest based level-2 (L2) stage and a event filter (EF) operating after even building with offline-like software. During the past three years, the luminosity and pile-up (number of collisions per beam crossing) has increased significantly placing escalating demands on the rejection and timing performace. The HLT algorithms advanced during this period to maintain and even improve performance. For the next run, the boundary between the L2 and EF will be removed, so that there is only one

high-level trigger which can operate either on regions of interest or on the full event depending on the objects found in the event either by the L1 or by the HLT itself.

This talk will discuss the algorithms, performance and ongoing development work on the reconstruction of calorimeter objects electrons, photons, taus, jets, and missing energy), inner detector tracking, and muon reconstruction. Among the improvements is a new missing energy trigger which uses specialized sums of the calorimeter cells to access the calorimeter readout earlier than was previously possible with strict region-of-interest only L2 system. Another improvement is a jet scan algorithm which operates at L2 using the information from the L1 digitization, but applies a clustering algorithm (anti-kT) similar to that used in the offline software. The jet and b-jet algorithms have been further developed to more closely resemble and include improvements from the offline software. Also discussed will be the work towards the merging of the two HLT levels in to a single level HLT, as well as operational experiences from the first LHC run.

Poster presentations / **371**

Towards more stable operation of the Tokyo Tier2 center

Author: Tomoaki Nakamura[1]

Co-authors: Hiroshi Sakamoto [1]; Ikuo Ueda [1]; Nagataka Matsui [1]; Tetsuro Mashimo [1]

[1] *University of Tokyo (JP)*

Corresponding Author: tomoaki.nakamura@cern.ch

The Tokyo Tier2 center, which is located at International Center for Elementary Particle Physics (ICEPP) in the University of Tokyo, was established as a regional analysis center in Japan for the ATLAS experiment. The official operation with WLCG was started in 2007 after the several years development since 2002. In December 2012, we have replaced almost all hard wares as the third system upgrade to deal with analysis for further growing data of the ATLAS experiment. The number of CPU cores are increased by factor of two (9984 cores in total), and the performance of individual CPU core is improved by 14% according to the HEPSPEC06 benchmark test at 32bit compile mode. It is estimated as 17.06 per core by using Intel Xeon E5-2680 2.70GHz. Since all worker nodes are made by 16 CPU cores configuration, we deployed 624 blade servers in total. They are connected to 6.7PB of disk storage system with non-blocking 10Gbps internal network backbone by using two center network switches (NetIron MLXe-32). The disk storage is made by 102 of RAID6 disk arrays (Infortrend DS S24F-G2840-4C16DO0) and served by equivalent number of 1U file servers with 8G-FC connection to maximize the file transfer throughput per storage capacity. As of February 2013, 2560 CPU cores and 2.00PB of disk storage have already been deployed for the WLCG. Currently, the remaining non-grid resources for both CPUs and disk storages are used as dedicated resources for the data analysis by the ATLAS Japan collaborators. Since all HWs in the non-grid resources are made by same architecture with Tier2 resource, they will be able to be migrated as the Tier2 extra resource on demand of the ATLAS experiment in the future. In addition to the upgrade of computing resources, we expect the improvement of connectivity on the wide area network. Thanks to the Japanese NREN (NII), another 10Gbps trans-Pacific line from Japan to Washington will be available additionally with existing two 10Gbps lines (Tokyo to NY and Tokyo to LA). The new line will be connected to the LHCONE for the more improvement of the connectivity. In this circumstance, we are working for the further stable operation. For instance, we have newly introduced GPFS (IBM) for the non-grid disk storage, while Disk pool manager (DPM) are continued to be used as Tier2 disk storage from the previous system. Since the number of files stored in a DPM pool will be increased with increasing the total amount of data, the development of stable database configuration is one of the crucial issues as well as scalability. We have started some studies on the performance of asynchronous database replication so that we can take daily full backup. In this presentation, we would like to introduce several improvements in terms of the performances and stabilities of our new system, and also present the status of the wide area network connectivity from Japan to US and/or EU with LHCONE.

Poster presentations / **144**

Arby, a general purpose, low-energy spectroscopy simulation tool

Authors: Massimiliano Nastasi[1]; Silvia Capelli[1]

Co-authors: Ezio Previtali [1]; Massimiliano Clemenza [1]; Maura Pavan [1]; Oliviero Cremonesi [1]

[1] *INFN Milano-Bicocca*

Measurements of radioactive sources, in order to reach an optimum level of accuracy, require an accurate determination of the detection efficiency of the experimental setup. In gamma ray spectroscopy, in particular, the high level of sensitivity reached nowadays implies a correct evaluation of the detection capability of source emitted photons. The standard approach, based
on an analytical extension of calibrated source measurements usually introduces large uncertainties related to shapes, geometrical setup, compositions and details of the nuclear decay emissions. To overcome the limitations intrinsic in the standard approach a different and more appropriate methodology is needed, specifically the possibility to simulate a virtual experiment featuring the same characteristics as the real measurements (geometry, materials, physical process involved). The GEANT4 toolkit offers all the ingredients needed to build such a simulator: the standard approach is to write specialized code to simulate a specific problem using the toolkit components and assembling them in a compiled program that contains the full specification of the experimental setup. Our approach, at INFN Milano-Bicocca has been, instead, to build a general purpose program capable of tackling a wide range of use-cases by reading the complete specification of the experiment to simulate from external files. This decoupling between simulation algorithm and experimental setup description allows to maintain and validate a single program, making it easy to add features and components by leveraging on an already existing body of functionality. This code, called Arby, was designed based on our experience in very low-background experiments: it's generality stems from the complete decoupling between its simulation capabilities and the actual specification of the experiment's setup. Different materials in different geometries (source and detector) can be specified externally, as well as a wide variety of different physical process to describe the interaction of radiation with matter. Even pile-up phenomena are correctly taken into account by the code, allowing for a fine-tuning and an arbitrary level of accuracy in determining the efficiency of the experimental setup. In this talk we will describe the architecture of the system and show some applications to real radioactive data analysis.

Plenaries / **496**

C++ evolves!

Author: Axel Naumann[1]

[1] *CERN*

Corresponding Author: axel.naumann@cern.ch

High Energy Physics is unthinkable without C++. But C++ is not the language it used to be: today it evolves continuously to respond to new requirements, and to benefit from the streamlined delivery process of new language features to compilers. How should HEP react?

After a short, subjective overview of parallel languages and extensions, the main features of C++11 will be presented, including the new concurrency model. A simple migration strategy for HEP will be discussed. A second theme will focus on structs-of-arrays and limits of auto-vectorization. The evolution of C++ including vectorization and concurrency will be presented.

Poster presentations / **125**

The LHCb Trigger Architecture beyond LS1

Authors: Gerhard Raven[1]; Johannes Albrecht[2]; Vladimir Gligorov[3]

[1] *NIKHEF (NL)*

[2] *Technische Universitaet Dortmund (DE)*

[3] *CERN*

Corresponding Authors: sebastian.neubert@cern.ch, vladimir.gligorov@cern.ch, gerhard.raven@nikhef.nl

The LHCb experiment is a spectrometer dedicated to the study of heavy flavor at the LHC. The rate of proton-proton collisions at the LHC is 15 MHz, but resource limitations mean that only 5 kHz can be written to storage for offline analytsis. For this reason the LHCb data acquisition system – trigger – plays a key role in selecting signal events and rejecting background. In contrast to previous experiments at hadron colliders like for example CDF or D0, the bulk of the LHCb trigger is implemented in software and deployed on a farm of 20k parallel processing nodes. This system, called the High Level Trigger (HLT) is responsible for reducing the rate from the maximum at which the detector can be read out, 1.1 MHz, to the 5 kHz which can be processed offline,and has 20 ms in which to process and accept/reject each event. In order to minimize systematic uncertainties, the HLT was designed from the outset to reuse the offline reconstruction and selection code. During the long shutdown it is proposed to extend this principle and enable the HLT to access offline quality detector alignment and calibration, by buffering events on the HLT nodes for long enough for this alignment and calibration to be performed and fed into the HLT algorithms. This will in turn allow the HLT selections to be tightened and hence will significantly increase the purity of the data being written for offline analysis. This contribution describes the proposed architecture of the HLT beyond LS1 and the technical challenges of implementing a real-time detector alignment and calibration in the LHC environment.

Summaries / 521

Summary of track 1 (Data acquisition, trigger and controls)

Corresponding Author: niko.neufeld@cern.ch

Software Engineering, Parallelism & Multi-Core / 224

Measurements of the LHCb software stack on the ARM architecture

Authors: Niko Neufeld[1]; Vijay Kartik Subbiah[1]

Co-authors: Ben Couturier [1]; Marco Clemencic [1]

[1] *CERN*

Corresponding Authors: niko.neufeld@cern.ch, vijay.kartik@cern.ch

The ARM architecture is a power-efficient design that is used in most processors in mobile devices all around the world today since they provide reasonable compute performance per watt. The current LHCb software stack is designed (and expected) to build and run on machines with the x86/x86_64 architecture. This paper outlines the process of measuring the performance of the LHCb software stack on the ARM architecture - specifically, the ARMv7 architecture on Cortex-A9 processors from NVIDIA, and also on full-fledged ARM servers with Calxeda chipsets - and makes comparisons with the performance on x86_64 architectures on the Intel Xeon 5650 and AMD Opteron 6272. The paper emphasises the aspects of performance per core with respect to the power drawn by the compute

nodes for the given performance - this ensures a fair real-world comparison with much more 'powerful' Intel/AMD processors. The comparisons of these real workloads in a HEP context are also complemented with standard synthetic benchmarks like HEPSPEC, LMBench and Coremark.

The pitfalls and solutions for the non-trivial task of porting the source code to build for the ARMv7 instruction set are presented. The specific changes in the build process needed for ARM-specific portions of the software stack are described, to serve as pointers for further attempts taken up by other groups in this direction. Cases where architecture-specific tweaks at the assembler lever (both in ROOT and the LHCb software stack) were needed for a successful compile are detailed - these cases are good indicators of where/how the software stack as well as the build system can be made more portable and multi-arch friendly. The experience gained from the tasks described in this paper are intended to i) assist in making an informed choice about ARM-based server solutions as a feasible low-power alternative to the current compute nodes, and ii) revisit the software design and build system for portability and generic improvements.

Poster presentations / **54**

DAQ Architecture for the LHCb Upgrade

Author: Niko Neufeld[1]

[1] *CERN*

Corresponding Author: niko.neufeld@cern.ch

LHCb will have an upgrade of its detector in 2018. After the upgrade, the
LHCb experiment will run at a high luminosity of 2x 10ˆ33 cmˆ-2 . sˆ-1. The upgraded detector will be read out at 40 MHz with a highly flexible software-based triggering strategy. The Data Acquisition (DAQ) system of HCb reads out the data fragments from the Front-End Electronics and transports them to the High-Lever Trigger farm at an aggregate throughput of ˜32 Tbit/s. The DAQ system will be based on high speed network technologies such as InfiniBand and/or 10/40/100 Gigabit Ethernet. Independent of the network technology, there are different possible architectures for the DAQ system.

In this paper, we present our studies on the DAQ architecture, where we analyze size, complexity and (relative) cost. We evaluate and compare several data-flow schemes for a network-based DAQ: push, pull and push with barrel-shifter traffic shaping. We also discuss the requirements and overall implications of the data-flow schemes on the DAQ system.

Plenaries / **484**

Trends in Advanced Networking

Author: Harvey Newman[1]

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

Corresponding Author: harvey.newman@cern.ch

Optical networking plays a key role in high-speed data transport, but the technology is developing at a fast pace. These developments are having direct impact not only for local and wide area data transport, but also in 'on-line'systems. Dr. Harvey Newman of Caltech will talk about the future trends not only in topical network, but also looking beyond to what advanced networking can enable tomorrow.

Event Processing, Simulation and Analysis / 208

Improvement of the ALICE Online Event Display using OO patterns and parallelization techniques

Author: Mihai Niculescu[1]

[1] *ISS - Institute of Space Science (RO) for the ALICE Collaboration*

Corresponding Author: mihai.niculescu@cern.ch

The visualization applications called event displays, are used in every high energy physics experiment as a fast quality assurance method for the entire process flow: starting from data acquisition, data reconstruction & calibration and finally obtaining the global view: a 3D view.
In this paper, we present a method that parallelizes this process flow and show how it is used for the ALICE online event display. This method presents how the offline reconstruction is parallelized at the event level and constructed as a mini server in order to serve the reconstructed event data to clients: event displays. This method incorporates an object orientated pattern called MVC - Model, View, Controller and brings the main advantage: the complete separation of the data (coming from DAQ in RAM), of the controller (reconstruction server) and of the view (event display). Another advantage is the improvement of the responsiveness of the event display while the acquisition/reconstruction process.
Separating the whole process in this manner, brings also the possibility to more easily parallelize its components. This fits perfectly with the upgrade plans for the long shutdown.

Poster presentations / 321

Intergrating configuration workflows with project management system

Authors: Dmitry Nilsen[1]; Pavel Weber[2]

[1] *Karlsruhe Institute of Technology*

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

Corresponding Authors: pavel.weber@cern.ch, dimitri.nilsen@kit.edu

The complexity of the heterogeneous computing resources, services and recurring infrastructure changes at the GridKa WLCG Tier-1 computing center require a structured approach to configuration management and optimization of interplay between functional components of the whole system. A set of tools deployed at GridKa, including Puppet, Redmine, Foreman, SVN and Icinga, provides the administrative environment giving the possibility to define and develop configuration workflows, reduce the administrative effort and improve sustainable operation of the whole computing center.

In this presentation we discuss the developed configuration scenarios
implemented at GridKa, which we use for host installation, service deployment, change management procedures, service retirement etc. The integration of Puppet with a project management tool like Redmine provides us with the opportunity to track problem issues, organize tasks and automate these workflows. The interaction between Puppet and Redmine results in automatic updates of the issues related to the executed workflow performed by different system components.

The extensive configuration workflows require collaboration and interaction between different departments like network, security, production etc. at GridKa. Redmine plugins developed at GridKa and integrated in its administrative environment provide an effective way of collaboration within the GridKa team. We present the structural overview of the software components, their connections, communication protocols and show a few working examples of the workflows and their automation.

Poster presentations / **276**

Next Generation PanDA Pilot for ATLAS and Other Experiments

Author: Paul Nilsson[1]

Co-authors: Fernando Harald Barreiro Megino [2]; John Hover [3]; Jose Caballero Bejar [4]; Kaushik De [1]; Peter Love [5]; Ramon Medrano Llamas [2]; Rodney Walker [6]; Tadashi Maeno [4]; Torre Wenaus [4]

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

[2] *CERN*

[3] *Brookhaven National Laboratory (BNL)-Unknown-Unknown*

[4] *Brookhaven National Laboratory (US)*

[5] *LANCASTER UNIVERSITY*

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

Corresponding Authors: paul.nilsson@cern.ch, fernando.harald.barreiro.megino@cern.ch, jose.caballero@cern.ch, kaushik@uta.edu, john.hover@cern.ch, peter.love@cern.ch, tmaeno@bnl.gov, ramon.medrano.llamas@cern.ch, rodney.walker@physik.uni-muenchen.de, wenaus@gmail.com

The Production and Distributed Analysis system (PanDA) has been in use in the ATLAS Experiment since 2005. It uses a sophisticated pilot system to execute submitted jobs on the worker nodes. While originally designed for ATLAS, the PanDA Pilot has recently been refactored to facilitate use outside of ATLAS. Experiments are now handled as plug-ins, and a new PanDA Pilot user only has to implement a set of prototyped methods in the plug-in classes, and provide a script that configures and runs the experiment specific payload.

We will give an overview of the Next Generation PanDA Pilot system and will present major features and recent improvements including live user payload debugging, data access via the federated xrootd system, stage-out to alternative storage elements, support for the new ATLAS DDM system (Rucio), and an improved integration with glExec, as well as a description of the experiment specific plug-in classes. The performance of the pilot system in processing LHC data on the OSG, LCG and NorduGrid infrastructures used by ATLAS will also be presented. We will describe plans for future development on the time scale of the next few years.

Data Acquisition, Trigger and Controls / **443**

Synchronization of a the 14 kTon Neutrino Detector with the Fermilab Beam

Author: Andrew Norman[1]

Co-authors: Gregory Deuerling [1]; Neal Wilcer [1]; Philip Adamson [1]; Richard Kwarciany [1]

[1] *Fermilab*

Corresponding Authors: eniner@indiana.edu, anorman@fnal.gov

The NOvA detector utilizes not only a high speed streaming readout system which capable of reading out the waveforms of over 368,000 detector cells, but a distributed timing system that is able drive and program the frontend clock systems of each of these readout to allow each hit in the detector to be time stamped with a universal wall clock time. This system is used to perform an absolute synchronization of the time across the entire far detector to a timing resolution of better than 10ns. This fine timing resolution allows for the data taken with the far detector to be precisely correlated with the extraction of beam from the Fermilab Main Injector and allows for precise determination of the time at which the neutrino beam cross the far detector.

The NOvA Timing and Synchronization system began production operations in January 2013 and has been used to drive the front end data acquisition system to synchronize the detector readout with the newly upgrade NuMI (Neutrinos at the Main Injector) neutrino beamline. This paper will cover the performance of the timing system during the first six months of detector operations and will cover the specialized diagnostic system that were put in place to validate and monitor offsets and drifts in the master clocks and GPS system that are used as reference sources for the time transfers between the detector sites.

Poster presentations / **449**

Data Driven Trigger Algorithms to Search for Exotic Physics in the NOvA Detector

Author: Zukai Wang[1]

Co-authors: Andrew Norman [2]; E. Craig Dukes [1]; Evan Niner [3]; Martin Frank [4]; Robert Group [1]

[1] *University of Virginia*

[2] *Fermilab*

[3] *Indiana University*

[4] *U*

Corresponding Authors: zw4vm@virginia.edu, eniner@indiana.edu, anorman@fnal.gov

The NOvA experiment at Fermi National Accelerator Lab, due to its unique readout and buffering design, is capable of accessing physics beyond the core neutrino oscillations program for which it was built. In particular the experiment is able to search for evidence of relic cosmic magnetic monopoles and for the signs of the neutrino flash from a near by supernova through uses of a specialized triggering system that is able to be self driven off of the raw readout stream.
These types of "data driven triggers"require fast track reconstruction and rejection over a wide range of time windows and detector conditions. The amount of data that the NOvA detector produces and the continuous streaming nature of the data make this type of real time reconstruction is a challenge for modern computing.
To meet these challenges a fast track reconstruction algorithm has been developed based on N-dimensional Hough Transform which is able to meet the latency requirements of the NOvA data acquisition system. This algorithm forms the basis of both the searches for magnetic monopoles, where it is able to separate out the slow moving monopoles from the prompt cosmic backgrounds, as well as for the supernova detection trigger, where it is used to subtract out the cosmic ray background and leave the low energy EM shower candidates.
In this paper, we will discuss in details of this algorithm and show how it can be used to distinguish and reconstruct slow magnetic monopole tracks verse fast cosmic ray tracks through an expansion of the Hough space to include velocity and pulse height information. We will discuss the scaling of the algorithm and its performance when run on real data. Examples of cosmic rate rejection and the improvements to supernova detection will be discussed and examples of the algorithm and trigger performance will shown for the early running of the NOvA far detector.

Poster presentations / **149**

ValDb: an aggregation platform to collect reports on the validation of CMS software and calibrations

Author: Antanas Norkus[1]

[1] *Vilnius University (LT)*

Corresponding Author: antanas.norkus@cern.ch

The scrutiny and validation of the software and of the calibrations used to simulate and reconstruct the collision events, have been key elements to the physics performance of the CMS experiment.

Such scrutiny is performed in stages by approximately one hundred experts who master specific areas of expertise, ranging from the low-level reconstruction and calibration which specific to a subdetector, to the reconstruction of higher level quantities such as particle candidates and global event properties.

In this paper we present ValDb, a web-based aggregation platform which collects from the all validation experts reports consisting of a concise write-up and links to pictures and plots. The reports are organized in campaigns, each targeting one specific software release or calibration update, and are all marked with a final summary icon (pass, fail, changes expected). ValDb is integrated with the CMS hypernews mailing system, where reports are sent to concerned fora.

Data Acquisition, Trigger and Controls / **426**

A First Look at the NOvA Far Detector Data Driven Trigger System

Author: Andrew Norman[1]

Co-authors: Alec Habig [2]; Evan Niner [3]; Gavin Davies [4]; Jan Zirnstein [5]; Martin Frank [6]; Matthew Cleary Tamsett [7]; Zukai Wang [8]

[1] *Fermilab*

[2] *Univ. of Minnesota Duluth*

[3] *Indiana University*

[4] *Iowa State University*

[5] *University of Minnesota*

[6] *U*

[7] *Louisiana Tech University (US)*

[8] *University of Virginia*

Corresponding Authors: anorman@fnal.gov, matthew.cleary.tamsett@cern.ch

The NOvA experiment is unique in its stream readout and triggering design. The experiment utilizes a sophisticated software triggering system that is able to select portions of the raw data stream to be extracted for storage, in a manner completely asynchronous to the actual readout of the detector. This asynchronous design permits NOvA to tolerate trigger decision latencies ranging from milliseconds to minutes and allows the experiment to reconstruct data in real time to search for phenomena outside of the neutrino beam window. The NOvA data driven trigger (DDT) is a high speed, low overhead, modular system based upon the ARTDAQ analysis framework. It is capable of dynamically managing the execution of complex pattern recognition and physics reconstruction algorithms as well as interfacing seamlessly with the primary DAQ readout chain and the NOvA offline processing and Monte Carlo chains.

The DDT system has been deployed to the NOvA trigger farms for the near and far detector, where it has been used successfully to detect neutrino interactions associated with the NuMI beam, acquire "rare"event topologies for detector calibration, search for supernova neutrino bursts, and perform searches for exotic physics signatures in the NOvA far detector.

The NOvA far detector and DDT system started commissioning in February 2013. This paper presents the overall design of the data driven triggering system and examines the performance of the trigger over the first six months of detector operations. It examines the scaling of the real time analysis system in both the distributed environment of the NOvA data acquisition clusters and in the multicore environment of the buffer nodes under which the core algorithms run.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / 448

Using the CVMFS for Distributing Data Analysis Applications for the Fermilab Intensity Frontier

Author: Andrew Norman[1]

Co-author: Adam Lyon [1]

[1] *Fermilab*

Corresponding Author: anorman@fnal.gov

The Cern Virtual File System (CVMFS) provides a technology for efficiently distributing code and application files to large and varied collections of computing resources. The CVMFS model and infrastructure has been used to provide a new, scalable solution to the previously difficult task of application and code distribution for grid computing.

At Fermilab, a new CVMFS based application and code distribution system has been deployed as an alternative for previous central storage systems. This centralized server system has been successfully adopted by the Intensity Frontier experiments (NOvA, g-2, Mu2e, Minerva) to distribute their application code efficiently to both Fermilab grid resources running thousands of concurrent jobs and to individual experiments who develop and run analysis code on their laptops and desktop system. The new system removes many of the issues related to system stability and to I/O bottlenecks previously encountered with other central storage and code distribution systems during large-scale production running.

The Fermilab solution simplifies the process of code and release management for the experiments while simultaneously addressing security concerns related to code integrity and application consistency by providing a single centrally managed and secure server for the experiments to push their updates to. The system has made the previously unthinkable possibility of experiments being able to work transparently both on their laptops and on the grid a reality.

Data Acquisition, Trigger and Controls / 43

FPGA based data acquisition system for COMPASS experiment

Author: Josef Novy[1]

Co-authors: Dmytro Levit [2]; Igor Konorov [2]; Martin Bodlak [3]; Miroslav Virius [1]; Richard Salac [3]; Stefan Huber [2]; Stephan Paul [4]; Vladimir Frolov [5]; Vladimir Jary [1]

[1] *Czech Technical University (CZ)*

[2] *Technische Universitaet Muenchen (DE)*

[3] *Charles University (CZ)*

[4] *Physik Department - Technische Universitaet Muen*

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

Corresponding Author: josef.novy@cern.ch

The COMPASS is a fixed target experiment, situated at the Super Proton Synchrotron (SPS) accelerator in the north area of the CERN laboratory, in Geneva, Switzerland. The experiment was commissioned during 2001, data-taking started in 2002. The data acquisition system of the experiment is based on the DATE soft-ware package, originally developed for the ALICE experiment. In 2011, after the physics program of the COMPASS experiment was ap-proved for the next 6 years, it was decided to build a new data acquisition system, based on modern FPGAs.
The new data acquisition system uses new FPGA based modules in two different configurations: multiplexer and switch. These modules allow whole events to be collected and built purely by hardware

and to be received by the readout nodes; therefore no further software event building is required. Deployment of these FPGA modules significantly decreases the amount of components involved in the data acquisition chain, mainly by removal of the event-building network based on Ethernet. Due to removal of the software part of event-building, the performance and reliability of data acquisition system is improved.
The new software architecture consists of Master, SlaveReadout, SlaveControl, graphical user interface, database, Message logger, and Message browser components. All processes are programmed in the C++ language with use of the QT framework and are designed to operate under the Scientific Linux CERN operating system. Communication between processes is based on the Distributed Information Management System (DIM) library which was originally developed for the Delphi experiment at CERN. The database service is based on the MySQL software. Behavior of the Master, the SlaveReadout, and the SlaveControl processes is described by finite state machines. In contrast to the present data acquisition system the new one takes advantage of modern technologies including multithreading.
The aim of this contribution is to analyze the original COMPASS data acquisition system, to present the new one, to compare both architectures, and to discuss their advantages and disadvantages. Furthermore, applications of state-of-the art technologies and modern approaches to the data acquisition software development based on usage of frameworks are presented. The design and results of performance tests and further development steps are also discussed.

Software Engineering, Parallelism & Multi-Core / 212

Is the Intel Xeon Phi processor fit for HEP workloads?

Authors: Alfio Lazzaro[1]; Andrzej Nowak[2]; Liviu Valsan[2]; Sverre Jarp[2]

Co-authors: Julien Leduc [2]; Mirela Madalina Botezatu

[1] *CRAY Research*

[2] *CERN*

Corresponding Author: andrzej.nowak@cern.ch

This paper summarizes the five years of CERN openlab's efforts focused on the Intel Xeon Phi co-processor, from the time of its inception to public release. We consider the architecture of the device vis a vis the characteristics of HEP software and identify key opportunities for HEP processing, as well as scaling limitations. We report on improvements and speedups linked to parallelization and vectorization on benchmarks involving software frameworks such as Geant4 and ROOT. Finally, we extrapolate current software and hardware trends and project them onto accelerators of the future, with the specifics of offline and online HEP processing in mind.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / 37

Evolution of interactive Analysis Facilities: from NAF to NAF 2.0

Authors: Andreas Haupt[1]; Friederike Nowak[2]; Yves Kemp[3]

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

[2] *DESY*

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

In 2007, the National Analysis Facility (NAF) was set up within the framework of the Helmholtz Alliance "Physics at the Terascale", and is located at DESY. Its purpose was the provision of an analysis infrastructure for up-to-date research in Germany, complementing the Grid by offering a interactive access to the data. It has been well received within the physics community, and has

proven to be a highly successful concept.
In this contribution, we will review experiences with the original NAF, and discuss both the resulting motivation and constraints for the transition to an evolved model. We call this new facility the NAF 2.0. We will present a new setup including its building blocks and user handling, and give an overview of the current status. The integration of new communities has broadened the range of the analysis facility beyond its primary focus on LHC and ILC experiments. To finish, an outlook on further developments like the adoption of new technologies will be given.

Poster presentations / **403**

Using Puppet to contextualize computing resources for ATLAS analysis on Google Compute Engine

Author: Carl Henrik Ohman[1]

Co-authors: Sergey Panitkin [2]; Valerie Cork Hendrix [3]

[1] *Uppsala University (SE)*

[2] *Brookhaven National Laboratory (US)*

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

Corresponding Author: henrik.ohman@cern.ch

With the advent of commercial as well as institutional and national clouds, new opportunities for on-demand computing resources for the HEP community become available. With the new cloud technologies come also new challenges, and one such is the contextualization of cloud resources with regard to requirements of the user and his experiment. In particular on Google's new cloud platform Google Compute Engine (GCE) upload of user's virtual machine images is not possible, which precludes application of ready to use technologies like CernVM and forces users to build and contextualize their own VM images from scratch. We investigate the use of Puppet to facilitate contextualization of cloud resources on GCE, with particular regard to ease of configuration, dynamic resource scaling, and high degree of scalability.

Poster presentations / **230**

CMS geometry through 2020

Author: Ianna Osborne[1]

Co-authors: David Lange [2]; Elizabeth Sexton-Kennedy [1]; Eric Charles Brownson [3]

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

[2] *Lawrence Livermore Nat. Laboratory (US)*

[3] *University of Puerto Rico (US)*

Corresponding Author: ianna.osborne@cern.ch

CMS faces real challenges with upgrade of the CMS detector through 2020. One of the challenges, from the software point of view, is managing upgrade simulations with the same software release as the 2013 scenario. We present the CMS geometry description software model, its integration with the CMS event setup and core software. The CMS geometry configuration and selection is implemented in Python. The tools collect the Python configuration fragments into a script used in CMS workflow. This flexible and automated geometry configuration allows choosing either transient or persistent version of the same scenario and specific version of the same scenario. We describe how the geometries are integrated and validated, how we define and handle different geometry scenarios in simulation and reconstruction. We discuss how to transparently manage multiple incompatible

geometries in the same software release. Several examples are shown based on current implementation assuring consistent choice of scenario conditions. The consequences and implications for multiple/different code algorithms are discussed.

Poster presentations / **300**

Parallelization of particle transport with INTEL TBB

Authors: Egor Ovcharenko[1]; Sergey Belogurov[1]

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

Corresponding Author: e.ovcharenko@gsi.de

One of the current problems in HEP computing is the development of particle propagation algorithms capable of efficient work at parallel architectures. An interesting approach in this direction has been recently introduced by the GEANT5 group at CERN [1]. Our report will be devoted to realization of similar functionality using Intel Threading Building Blocks (TBB) library.
In the prototype implemented by the GEANT5 group the independent part of job to be processed by a single computing node is transport of a group of particles located in the same volume. Such a group of particles is called a basket. Processing of a single basket by a single core allows to exploit most efficiently the principle of data locality. The implementation is done using traditional pthreads library. This tool requires manual handling of jobs'mapping onto the threads and hence a special manager-thread code has been written. This manager always occupies one thread and is a potential bottleneck causing low scalability of the code.
Intel TBB is a library of parallel programming templates which is in some sense an extension of STL. Along with standard thread-programming features like locks, mutexes and thread-safe concurrent containers, Intel TBB states a new standard in concurrent programming - task-based programming. The idea of task-based programming is to allow the developer to write tasks and to send them to the task scheduler. The last will automatically solve the core load balancing problem. Intel TBB is considered as a candidate for being a standard tool in various domains of HEP computations.
We are going to present at the CHEP2013 conference an adaptation of the basket oriented track propagation algorithm to the possibilities and immanent logic of TBB. Our prototype is based on an interplay between two types of tasks: PropagationTask which contains job taking care of particle transport within a basket, the result of a PropagationTask is a collection of tracks located on the boundaries of the volume; and DispatcherTask into which the job distributing one or more collections of tracks between several new baskets is packed.
The basic implementation starts a PropagationTask for each basket produced by the DispatcherTask and a DispatcherTask for a group of track collections produced by the PropagationTasks. A number of dispatching policies is under investigation. Intel TBB concurrent
queues are used for data exchange between the tasks.
Performance and scalability tests for the basic algorithm with different dispatching policies will be presented. Results will be
compared with original GEANT 5 implementation. Possible improvements and further development towards compatibility with vectorized basket processing will be discussed.

The authors acknowledge stimulating and instructive discussions with Federico Carminati and Andrei Gheata (CERN).

[1] Rethinking particle transport in the many-core era towards GEANT 5
Apostolakis, John; Brun, Rene; Carminati, Federico; Gheata, Andrei
J. Phys.: Conf. Ser. 396 (2012) 022014

Data Acquisition, Trigger and Controls / **21**

The readout and control system of the mid-size telescope prototype of the Cherenkov Telescope Array

Authors: Bagmeet Behera[1]; David Melkumyan[1]; Ekrem Oguzhan Anguner[2]; Emrah Birsin[2]; Igor Oya[3]; Matthias Fuessling[4]; Peter Wegner[5]; Ronny Sternberge[1]; Stephan Wiesand[5]; Torsten Schmidt[1]; Ullrich Schwanke[6]

[1] *DESY, Zeuthen*

[2] *Humboldt-Universitaet zu Berlin*

[3] *Humboldt University*

[4] *Universtitaet Potsdam*

[5] *DESY*

[6] *Humboldt University Berlin*

Corresponding Author: oya@physik.hu-berlin.de

The Cherenkov Telescope Array (CTA) is one of the major ground-based astronomy projects being pursued and will be the largest facility for ground-based gamma-ray observations ever built. CTA will consist of two arrays: one in the Northern hemisphere composed of about 20 telescopes, and the other one in the Southern hemisphere composed of about 100 telescopes, both arrays containing telescopes of several sizes. A prototype for the Mid-Size Telescope (MST) with a diameter of 12 m has been installed in Berlin and is currently being commissioned. This prototype is composed of a mechanical structure, a drive system and mirror facets mounted with powered actuators to enable active control. Five Charge-Coupled Device (CCD) cameras, and a wide set of sensors allow the evaluation of the performance of the instrument.

The design of the control software is following concepts and tools under evaluation within the CTA consortium in order to provide a realistic test-bed for the middleware: 1) The readout and control system for the MST prototype is implemented with the Atacama Large Millimeter/submillimeter Array (ALMA) Common Software (ACS) distributed control middleware; 2) the OPen Connectivity-Unified Architecture (OPC UA) is used for hardware access; 3) MySQL databases are used for archiving the slow control monitoring data and operation configuration parameters storage; and 4) the document oriented MongoDB database is used for an efficient storage of CCD images, logging and alarm information. In this contribution, the details on the implementation of the control system for this MST prototype telescope are described.

Poster presentations / 324

A Validation Framework to faciliatate the Long Term Preservation of High Energy Physics Data (The DESY-DPHEP Group)

Author: David South[1]

Co-author: Dmitry Ozerov [2]

[1] *DESY*

[2] *D*

Corresponding Authors: dmitri.ozerov@desy.de, david.south@cern.ch

In a future-proof data preservation scenario, the software and environment employed to produce and analyse high energy physics data needs to be preserved, rather than just the data themselves. A software preservation system will be presented which allows analysis software to be migrated to the latest software versions and technologies for as long as possible, substantially extending the lifetime of the software, and hence also the data. Contrary to freezing the environment and relying on assumptions about future virtualisation standards, we propose a rolling model of preservation of the software. Technically, this is realised using a virtual environment capable of hosting an arbitrary number of virtual machine images, built with different configurations of operating systems and the relevant software, including any necessary external dependencies. A significant fraction of the work involved requires a deep level of validation of the experimental software and environment, and in particular the progress made by the participating experiments will be presented. Such a system is by design expandable and able to host and validate the requirements of multiple experiments, and can be thought of as a tool to aid migration that will detect problems and incoherence, helping to identify and solve them by the join efforts of experiments and computer experts.

Summaries / 520

Summary of track 3B

Corresponding Author: nurcan@uta.edu

Event Processing, Simulation and Analysis / 361

An exact framework for uncertainty quantification in Monte Carlo simulation

Author: Saracco Paolo[1]

Co-author: Maria Grazia Pia [2]

[1] *INFN Genova (Italy)*

[2] *Universita e INFN (IT)*

Corresponding Author: paolo.saracco@ge.infn.it

Uncertainty Quantification (UQ) addresses the issue of predicting non-statistical errors affecting the results of Monte Carlo simulations, deriving from uncertainties in the physics data and models they embed. In HEP it is relevant to particle transport in detectors, as well as to event generators.

We summarize recent developments, which have established the mathematical ground of an exact framework for UQ calculation. This study assessed that in the case of a single uncertainty and under wide hypotheses a simple general relation exists, which relates the probability density function (PDF) of the input to Monte Carlo simulation, and of the output it produces. This result has been empirically verified in a conceptually simplified Monte Carlo simulation environment.

In this contribution we address the problem of extending this approach to the multi-variate case. A typical scenario in this context consists of predicting the dependence of simulation results on input cross section tabulations. We show that for a wide class of probability distributions of the input unknowns it is possible to determine analytically the expected output PDF for any required observable, both in the case of independent variations and in the case of linear correlations among the input variables. This class includes normal distributions, flat (and in general finite interval distributions), and all the Levy stable distributions - a four parameter family of heavy-tailed distributions, which includes the Breit-Wigner one.

For all these distributions it is possible to evaluate exactly the confidence intervals for the physical observables of experimental interest produced by the Monte Carlo simulation.

This is a powerful environment to perform UQ in many physical cases of interest to HEP and low energy nuclear physics experiments.

We present the mathematical methods for uncertainty quantification and some applications to relevant use cases.

Poster presentations / **349**

The GridKa Tier-1 Computing Center within the ALICE Grid Framework

Author: WooJin Park[1]

Co-authors: Andreas Heiss [2]; Andreas Petzold [2]; Christopher Jung [2]; Kilian Schwarz [3]

[1] *KIT*

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

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

The GridKa computing center, hosted by Steinbuch Centre for Computing at the Karlsruhe Institute for Technology (KIT) in Germany, is serving as the largest Tier-1 center used by the ALICE collaboration at the LHC. In 2013, GridKa provides 30k HEPSEPC06, 2.7 PB of disk space, and 5.25 PB of tape storage to ALICE. The 10Gbit/s network connections from GridKa to CERN, several Tier-1 centers and the general purpose network are used by ALICE intensively. In 2012 a total amount of ~1 PB was transferred to and from GridKa. As Grid framework, AliEn (ALICE Environment) is being used to access the resources, and various monitoring tools including the MonALISA (MONitoring Agent using a Large Integrated Services Architecture) are always running to alert in case of any problem. GridKa on-call engineers provide 24/7 support to guarantee minimal loss of availability of computing and storage resources in case of hardware or software problems. We introduce the GridKa Tier-1 center from the viewpoint of ALICE services.

Poster presentations / **192**

Geo-localization in CERN's underground facilities

Author: Aurelie Pascal[1]

[1] *CERN*

Corresponding Author: aurelie.pascal@cern.ch

CERN has recently renewed its obsolete VHF firemen's radio network and replaced it by a digital one based on TETRA technology. TETRA already integrates an outdoor GPS localization system, but it appeared essential to look for a solution to also locate TETRA users in CERN's underground facilities.
The system which answers this problematic and which has demonstrated a good resistance to radiation effects, is based on autonomous beacons placed in strategic locations and broadcasting specific identification numbers. The radios are able to decode these identification numbers and transmit this information through the TETRA network to the fire brigade Control Center. An application dedicated to the indoor localization is then able to locate the TETRA terminal on a map.

Poster presentations / **391**

SPADE : A peer-to-peer data movement and warehousing orchestration

Author: Simon Patton[1]

Co-author: Cindy Mackenzie [2]

[1] *LAWRENCE BERKELEY NATIONAL LABORATORY*

[2] *University of Wisconsin - Madison*

Corresponding Author: sjpatton@lbl.gov

The SPADE application was first used by the IceCube experiment to move its data files from the South Pole to Wisconsin. Since then is has been adapted by the DayaBay experiment to move its data files from its experiment, just outside Hong Kong, to both Beijing an LBNL. The aim of this software is to automate much of the data movement and warehousing that is often done by hand or home-grown script on smaller experiments.

The latest generation of this software has been developed to be experiment independent and fully peer-to-peer. This means that it can be used anywhere, not only to move raw experiment data to computing centers but also as a means of exchanging data between centers. For example, gathering Monte Carlo production data from member institutions to an experiment's main warehouse and distributing copies back out to the users who need them.

Special attention has been paid to keeping administration and infrastructure time and effort to a minimum so that this software can be easily used by experiments that do not have large computer support resources. Additional features of the software include facilities to run prompt analysis on files as they arrive in the warehouse, the archiving of files to HPSS and the option to create local copies of files before they are moved to enable immediate assessment of the data by computers based at the experiment.

This paper will examine the requirement of the new version of the software. It will then discuss the main architecture of the application and show how this satisfies those requirements. This will be followed by a quick tour of what is needed to install and run the basic version, after which an explanation of now experiment specific customization can be made.

Finally, the DayaBay use case will be summarized to show a practical deployment of the application.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / **423**

Dayabay Offline processing chain: data to paper in 20 Days.

Author: Simon Patton[1]

[1] *LAWRENCE BERKELEY NATIONAL LABORATORY*

Corresponding Author: sjpatton@lbl.gov

In March 2012 the Dayabay Neutrino Experiment published the first measurement of the theta_13 mixing angle. The publication of this result occurred 20 days after the last data that appeared in the paper was taken, during which time normal data taking and processing was continuing. This achievement used over forty thousand 'core hours' of CPU and handled eighteen thousand files totaling 16 TBs. While these numbers are not in the same league as those seen by the larger LHC experiment, they were achieved on a much smaller infrastructure than is available to those experiments.

This paper will provide an overview of the Dayabay Offline processing chain that made this possible. It will follow the data's progress from when the DAQ files were closed up to the point where fully calibrated and reconstructed data is supplied to the physicists so that they can execute their own

analysis algorithms. The tools used throughout the chain, such as Sentry and PSquared, will be described as well as how the chain is linked together.

Also included in this paper will be the improvement to the processing chain since the first results were published, improvements such as the automation of gain calibrations and energy scaling. The purpose of this paper is to demonstrate what can be done by experiments on the scale of Dayabay with limited software and computing resources.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / 105

CMS Multicore Scheduling Strategy

Authors: Antonio Maria Perez Calero Yzquierdo[1]; Ian Fisk[2]; Jose Hernandez Calama[1]

[1] *Centro de Investigaciones Energ. Medioambientales y Tecn. - (ES*

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

Corresponding Authors: antonio.perez.calero.yzquierdo@cern.ch, chema.hernandez.calama@gmail.com, ian.fisk@cern.ch

In the next years, processor architectures based on much larger numbers of cores will be most likely the model to continue "Moore's Law" style throughput gains. This not only results in many more jobs in parallel running the LHC Run 1 era monolithic applications. Also the memory requirements of these processes push the workernode architectures to the limit. One solution is parallelizing the application itself, through forking and memory sharing or through threaded frameworks. CMS is following all of these approaches and has a comprehensive strategy to schedule multi-core jobs on the GRID based on the glideIn WMS submission infrastructure. We will present the individual components of the strategy, from special site specific queues used during provisioning of resources and implications to scheduling; to dynamic partitioning within a single pilot to allow to transition to multi-core or whole-node scheduling on site level without disallowing single-core jobs. In this presentation, we will present the experiences made with the multi-core scheduling modes and give an outlook of further developments working towards the restart of the LHC in 2015.

Poster presentations / 317

GPU Implementation of Bayesian Neural Networks in SUSY Studies

Author: Michelle Perry[1]

Co-author: Harrison Prosper [1]

[1] *Florida State University*

Corresponding Author: mep03e@my.fsu.edu

The search for new physics has typically been guided by theoretical models with relatively few parameters. However, recently, more general models, such as the 19-parameter phenomenological minimal supersymmetric standard model (pMSSM), have been used to interpret data at the Large Hadron Collider. Unfortunately, due to the complexity of the calculations, the predictions of these models are available at a discrete set of parameter points, which makes the use of analysis techniques that require smooth maps between the parameters and a given prediction problematic. It would be useful, therefore, to have a computationally routine way to construct such mappings. We propose to construct the mappings using Bayesian neural networks (BNN). Bayesian neural networks have been used in a few high-profile analyses in high energy physics for both classification and functional approximation. The main limitation to their widespread use is the time required to construct

these functions. In this talk, we describe an efficient Graphical Processing Unit (GPU) implementation of the construction of BNNs using the Hybrid Markov-Chain Monte Carlo (MCMC) method. We describe our implementation of the MCMC algorithm on the GPU, including the speedups we have achieved so far and illustrate the effectiveness of our implementation by mapping the pMSSM parameter space to some of its key predictions.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / 410

Data Processing for the Dark Energy Survey

Author: Donald Petravick[1]

Co-authors: Adam Lyon [2]; Elizabeth Buckley-Geer [3]; Margaret Gelman [4]; Michael Johnson [4]; Michael Wang ; Michelle Gower [4]; Todd Tomashek [4]

[1] *U*

[2] *Fermilab*

[3] *Fermi National Accelerator Laboratory*

[4] *National Center for Supercomputing Applications*

Corresponding Author: petravic@illinois.edu

The Dark Energy Survey (DES) is designed to probe the origin of the accelerating universe and help uncover the nature of dark energy by measuring the 14-billion-year history of cosmic expansion with high precision. More than 120 scientists from 23 institutions in the United States, Spain, the United Kingdom, Brazil, Switzerland and and Germany are working on the project. This collaboration has built an extremely sensitive 570-Megapixel digital camera, DECam, and has mounted it on the Blanco 4-meter telescope at Cerro Tololo Inter-American Observatory high in the Chilean Andes.

The survey has completed an initial season of science verification. The survey will start in September 2013 and run for 5 years. DES data are transferred by network to the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign. The processing system there supports quick-turn-around processing for super nova science and final processing of data into catalogs.

We describe the processing software system which is in place for the five year data taking period. The system is capable of processing data on mid scale super computers and the Open Science Grid. The software structure is oriented towards wrapping community codes and custom codes, in a way that provides for uniform handling and common operational characteristics for 10 processing pipelines. The system is supported by a 100TB oracle database, which is used to store object catalogs as well as extensive operational and file system meta-data. Provenance data is stored in in a uniform schema derived from the Open Provenance Model.

Data Stores, Data Bases, and Storage Systems / 363

A Preview of a Novel Architecture for Large Scale Storage

Author: Andreas Petzold[1]

Co-authors: Christoph Erdmann Pfeiler [2]; Jos Van Wezel [1]

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

[2] *Forschungszentrum Karlsruhe GmbH (FZK)*

Corresponding Author: andreas.petzold@cern.ch

The need for storage continues to grow at a dazzling pace and science and society have become dependent on access to digital data. First sites storing an exabyte of data will be reality in a few years. The common storage technology in small and large computer centers continues to be magnetic disks because of their very good price performance ratio. Storage class memory and solid state disk (ssd) storage is entering the data center but currently foremost in combination with traditional magnetic storage in which it bridges the gap between access time and large storage capacity. The storage building block usually consists of tens to hundreds of magnetic disks connected to a dedicated storage controller that logically aggregates multiple disks into large RAID groups. The storage controller connects to the storage fabric and the application servers via fibre channel, infiniband or tcp/ip and runs proprietary firmware on special purpose hardware. The advent of powerful multicore processors and high bandwidth shared memory of the uniform Intel platform, allows sharing of storage management applications and storage controller ops on the same server. Alternatively the storage management application, normally running on the storage server can run on the storage controller when both have the same architecture. In this presentation we evaluate solutions to be used for large scale storage environments such as the German WLCG Tier 1 center GridKa or the Large Scale Data Facility, hosted by Steinbuch Centre for Computing at Karlsruhe Institute of Technology. On common hardware the storage controller functionality is shared with popular storage management applications like xrootd and dCache. The financial and operational benefits are discussed and first experiences are presented.

Poster presentations / **306**

Deploying an IPv6-enabled grid testbed at GridKa

Author: Bruno Heinrich Hoeft[1]

Co-author: Andreas Petzold [2]

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

[2] *KIT*

Corresponding Authors: andreas.petzold@cern.ch, bruno.hoeft@kit.edu

GridKa, the German WLCG Tier-1 site hosted by Steinbuch Centre for Computing at Karlsruhe Institute of Thechnology, is a collaboration partner in the HEPIX-IPv6 testbed. A special IPv6-enabled gridftp server has been installed previously. In 2013, the IPv6 efforts will be increased. Already the installation of a new Mini-Grid site has been started. This Mini-Grid installation is planned as a dual-stack IPv4/IPv6 environment and will contain the current services of GridKa. Mainly the following EMI services BDII, Cream-CE, WorkerNode, dCache, xrootd as well as a Grid Engine job scheduler will be deployed. The Mini-Grid setup has initially been started with IPv4 only. Hence the IPv6 readiness of each service is to be evaluated.

There are several other initiatives analyzing the IPv6 readiness of WLCG software. There is EMI evaluating the middleware, the HEPIX-IPv6 working group evaluating the readiness of transport protocols, EGI IPv6 working group evaluating non-WLCG Grid middleware (e.g. unicore). Our testbed is meant to offer an installation basis for the initiatives in the Grid framework and enable them to use it with as little effort as possible.

The paper shows the setup of the IPv6 testbed in detail. It illustrates the IPv6 readiness of the different application and the services offered as well. It also highlights the problems that occurred during the deployment of the testbed and how these obstacles were overcome by thorough investigation and evaluation of all included programs.

Poster presentations / 176

Testing and Open Source installation and server provisioning tool for the INFN-CNAF Tier1 Storage system

Author: michele pezzi[1]

Co-authors: Daniele Gregori [2]; Pier Paolo Ricci [3]

[1] *Infn-cnaf*

[2] *Istituto Nazionale di Fisica Nucleare (INFN)*

[3] *INFN CNAF*

Corresponding Author: michele.pezzi@cnaf.infn.it

In large computing centers, such as the INFN-CNAF Tier1, is essential to be able to set all the machines, depending on use, in an automated way. For several years at the Tier1 has been used Quattor, a server provisioning tool, which is currently used in production.

Nevertheless we have recently started a comparison study involving other tools able to provide specific server installation and configuration features and also to offer a proper full customizable solution as an alternative to Quattor. Our choice at the moment fell on integration between two well-known tools: Cobbler for the installation phase and Puppet for the server provisioning and management operation.

The tool should provide the following properties in order to replicate and gradually improve the actual system features:
1) Implement a system check for storage specific constrain such as kernel modules black list at boot time to avoid undesired SAN access during disk partitioning.
2) A simple and effective mechanism for kernel upgrade and downgrade.
3) The ability of setting package provider using yum, rpm or apt.
4) Easy to use Virtual Machine installation support including bonding and specific Ethernet configuration.
5) Scalability for managing thousands of nodes and parallel installation.

This paper describes the results of the comparison and the experiments carried to verify the above requirements and if the new system is suitable for INFN-T1 storage system will be also described in details.

Facilities, Infrastructures, Networking and Collaborative Tools / 344

Scholarly literature and the media: scientific impact and social perception of HEP computing

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] *Oak Ridge National Laboratory*

Corresponding Author: maria.grazia.pia@cern.ch

The broad coverage of the search for the Higgs boson in the mainstream media is a relative novelty for HEP research, whose achievements have traditionally been limited to scholarly literature. This

presentation illustrates the results of a scientometric analysis of HEP computing in scientific literature, institutional media and the press, and a comparative overview of similar metrics concerning recent particle physics measurements.

The picture emerging from these scientometric data documents the scientific impact and social perception of HEP computing.

This analysis intends to open a discussion in the software-oriented community for improved communication of the scientific and social role of HEP computing.

Poster presentations / 296

New physics and old errors: validating the building blocks of major Monte Carlo codes

Authors: Chan Hyeong Kim[1]; Gabriela Hoff[2]; Han Sung Kim[1]; Maria Grazia Pia[3]; Matej Batic[4]; Mincheol Han[5]; Paolo Saracco[6]

[1] *Hanyang Univ., Seoul, Korea*

[2] *PUCRS, Brazil*

[3] *Universita e INFN (IT)*

[4] *Jozef Stefan Institute*

[5] *H*

[6] *INFN Genova, Italy*

Corresponding Author: maria.grazia.pia@cern.ch

A large-scale project is in progress, which validates the basic constituents of the electromagnetic physics models implemented in major Monte Carlo codes (EGS, FLUKA, Geant4, ITS, MCNP, Penelope) against extensive collections of experimental data documented in the literature. These models are responsible for the physics observables and the signal generated in particle detectors, including those originating from the products of hadronic interactions.

Total and differential cross sections, angular distributions and other basic physics features are examined. In addition to currently implemented models, theoretical calculations, semi-empirical models and physics data libraries not yet exploited in Monte Carlo codes are evaluated and compared to those currently in use. The results of this analysis identify and quantify the state of the art achievable in Monte Carlo simulation based on the available body of knowledge.

This validation analysis highlights similarities and differences in the modeling choices adopted by major Monte Carlo codes, and quantitatively documents their accuracy based on rigorous statistical methods. Categorical data analysis techniques are applied to quantify objectively the significance of the differences in compatibility with experiment among the examined Monte Carlo codes.

Duplicated models are identified, not only across different Monte Carlo systems, but also between packages of the same system, and even within the same package. Guidelines for pruning duplicated functionality in Geant4 are discussed.

In parallel to the evaluation of physics accuracy, the computational performance of alternative physics models and calculation approaches is estimated. Quantitative results show that there is no univocal winner between analytical calculations and data library interpolation in terms of computational speed.

The interplay of this validation methodology with epistemic uncertainties embedded in Monte Carlo codes and its applicability also to hadronic physics models are discussed.

Poster presentations / 34

Distributed storage and cloud computing: a test case

Author: Stefano Piano[1]

Co-author: Giuseppe Della Ricca [2]

[1] *INFN (IT)*

[2] *Universita e INFN (IT)*

Corresponding Authors: stefano.piano@cern.ch, giuseppe.della-ricca@cern.ch

Since 2003 the computing farm hosted by the INFN T3 facility in Trieste supports the activities of many scientific communities. Hundreds of jobs from 45 different VOs, including those of the LHC experiments, are processed simultaneously. The currently available shared disk space amounts to about 300 TB, while the computing power is provided by 712 cores for a total of 7400 HEP-SPEC06. Given that normally the requirements of the different computational communities are not synchronized, the probability that at any given time the resources owned by one of the participants are not fully utilized is quite high. A balanced compensation should in principle allocate the free resources to other users, but there are limits to this mechanism. In fact, the Trieste site may not hold the amount of data needed to attract enough analysis jobs, and even in that case there could be a lack of bandwidth for their access. The Trieste ALICE and CMS computing groups, in collaboration with other Italian groups, aim to overcome the limitations of existing solutions using two approaches. Sharing the data among all the participants, avoiding data duplication and taking full advantage of GARR-X wide area networks (10 GB/s) allows to distribute more efficiently the jobs according to the CPU availability, irrespective of the storage system size. Integrating the resources dedicated to batch analysis with the ones reserved for dynamic interactive analysis, through modern solutions as cloud computing, can further improve the use of the available computing power. The first tests of the investigated solutions for both distributed storage on wide area network and cloud computing approaches will be presented.

Event Processing, Simulation and Analysis / **502**

Track extrapolation and muon identification using GEANT4E in event reconstruction in the Belle II experiment

Author: Leo Piilonen[1]

Co-authors: Takanori HARA [2]; Thomas Kuhr [3]

[1] *Virginia Tech*

[2] *KEK*

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

Corresponding Author: piilonen@vt.edu

I will describe the charged-track extrapolation and the muon identification modules in the Belle II data analysis code library. These modules use GEANT4E to extrapolate reconstructed charged tracks outward from the Belle II Central Drift Chamber into the outer particle-identification detectors, the electromagnetic calorimeter, and the K-long and muon (KLM) detector embedded in the iron yoke surrounding the Belle II solenoid. using the detailed detector geometry that was developed for the simulation module. The extrapolation module propagates the position, momentum, 6-dimensional covariance matrix, and time of flight from the interaction point to permit comparison of the extrapolated track with the hits detected in the outer detectors. In the course of track extrapolation into the KLM, a Kalman fitting procedure is applied that adjusts the track parameters using the matching hits in each of the crossed detectors. The muon identification procedure then compares the longitudinal and transverse profiles of the extrapolation and the matched hits in the KLM and, for the low-momentum tracks, the extrapolated and matched crystals in the electromagnetic calorimeter, to distinguish between the muon and hadron-like hypotheses. Several modifications were made to permit GEANT4E to interoperate with GEANT4 and to expand the number of particle species that can be extrapolated.

Data Acquisition, Trigger and Controls / 480

K-long and muon trigger in the Belle II experiment

Author: Leo Piilonen[1]

Co-author: Yoshihito Iwasaki [2]

[1] *Virginia Tech*

[2] *KEK*

Corresponding Author: piilonen@vt.edu

I will describe the first-level trigger in the Belle II experiment that examines the hit patterns in the K-long and muon (KLM) detector to find evidence for
compact clusters (indicative of a K-long meson hadronic shower) or tracks (indicative of a charged particle from the interaction point or of a cosmic ray).
The algorithm is implemented in a VIRTEX6 FPGA on a Universal Trigger Module (used throughout the Belle II experiment) that receives raw inputs from each of the KLM sectors via 32 fiber-optic cables, sorts these data, performs rudimentary cluster-finding and track-fitting, and delivers its output to the Global Decision Logic module. I also describe the offline KLM-trigger simulation module in the Belle II data analysis code library.

Software Engineering, Parallelism & Multi-Core / 202

Preparing HEP Software for Concurrency

Authors: Benedikt Hegner[1]; Danilo Piparo[1]; Pere Mato Vila[1]

[1] *CERN*

Corresponding Authors: danilo.piparo@cern.ch, benedikt.hegner@cern.ch

The necessity for really thread-safe experiment software has recently become very evident, largely driven by the evolution of CPU architectures towards exploiting increasing levels of parallelism, For high-energy physics this represents a real paradigm shift, as concurrent programming was previously only limited to special, well-defined domains like control software or software framework internals. This paradigm shift, however, falls into the middle of the successful LHC programme and many million lines of code have already been written without the need for parallel execution in mind. In this presentation we will have a closer look at the offline processing applications of the LHC experiments and their readiness for the many-core era. We will review how previous design choices impact the move to concurrent programming. We present our findings on transforming parts of the LHC experiments' reconstruction software to thread-safe code, and the main design patterns that have emerged during the process. A plethora of parallel-programming patterns are well known outside the HEP community, but only a few have turned out to be straight forward enough to be suited for non-expert physics programmers. Finally, we propose a potential strategy for the migration of existing HEP experiment software to the many-core era.

Software Engineering, Parallelism & Multi-Core / 219

Speeding up HEP experiments' software with a library of fast and autovectorisable mathematical functions

Author: Danilo Piparo[1]

Co-authors: Thomas Hauth [1]; Vincenzo Innocente [1]

[1] *CERN*

Corresponding Author: danilo.piparo@cern.ch

During the first four years of data taking at the Large Hadron Collider (LHC), the simulation and reconstruction programs of the experiments proved to be extremely resource consuming. In particular, for complex event simulation and reconstruction applications, the impact of evaluating elementary functions on the runtime is sizeable (up to one fourth of the total), with an obvious effect on the power consumption of the hardware dedicated to their execution. This situation clearly needs improvement, especially considering the even more demanding data taking scenarios after this first LHC long shut down. A possible solution to this issue is the VDT (VectorisD maTh) mathematical library. VDT provides the most common mathematical functions used in HEP in an open source product. The functions' implementations are fast, can be inlined, provide an approximate accuracy and usable in vectorised loops. Their implementation is portable across platforms: x86 and ARM processors, Xeon Phi coprocessors and GPGPUS. In this contribution, we describe the features of the VDT mathematical library, showing significant speedups with respect to the Libm library and comparable accuracies. Moreover, taking as examples simulation and reconstruction workflows ran in production by the experiments, we show the benefits of the usage of VDT in terms of runtime reduction and stability of physics output.

Poster presentations / **351**

Status and new developments of the Generator Services project

Authors: Mikhail Kirsanov[1]; Witold Pokorski[2]

Co-authors: Anton Karneyeu [1]; Dima Konstantinov [3]; Kirill Lugovskiy [4]; Oleg Zenin [5]; Pere Mato Vila [2]; Peter Skands [2]

[1] *Russian Academy of Sciences (RU)*

[2] *CERN*

[3] *IHEP*

[4] *I*

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

Corresponding Authors: witold.pokorski@cern.ch, mikhail.kirsanov@cern.ch

The LCG Generator Services project provides validated, LCG compliant Monte Carlo generators code for both the theoretical and experimental communities at the LHC. It collaborates with the generators authors, as well as the experiments software developers and the experimental physicists.

In this paper we present the recent developments and the future plans of the project. We start with reporting on the current status of the generators repository. Due to the increased number of dependencies on external packages the generators repository was reorganized to follow the releases of the repository "LCG external".

In order to perform the GENSER installation in a way consistent with LCG external, it was decided to migrate to the new cmake - based system, LCGCMAKE. This migration is planned to be finished in the first half of 2013. cmake - based build system is now provided for several generators.

Some LHC experiments directly use generator libraries installed in the GENSER repository, others use source tarballs prepared by GENSER and also put in the repository.

We discuss different testing and physics validation procedures. The regression tests system automatically checks that several physical observables does not change from one version to another and for dirrerent computer platforms supported by LCG (there are 5 such platforms now). It checks also the source tarballs provided by GENSER. This ensures that all LHC experiments use the same generator code. More detailed physical tests are performed with HepMC Analysis Tool (mainly comparison of different versions) and Rivet (mainly comparison of generation results with data).

We present a new activity, within Generator Services, MCPLOTS. This subproject is intended as a simple browsable repository of MC (Monte Carlo) plots comparing High Energy Physics event generators to a wide
variety of available experimental data, for tuning and reference purposes. Apart from individual plots contained in papers and presentations, there has so far not been any central database where people can quickly see how tune X of version Y of generator Z looks on distribution D. The idea with mcplots is to provide such a repository.

Poster presentations / 181

Recent Developments in the Geant4 Hadronic Framework

Author: Witold Pokorski[1]

Co-author: Alberto Ribon [1]

[1] *CERN*

Corresponding Author: witold.pokorski@cern.ch

In this paper we present the recent developments in the Geant4 hadronic framework, as well as in some of the existing physics models.

Geant4 is the main simulation toolkit used by the LHC experiments and therefore a lot of effort is put into improving the physics mod
els in order for them to have more predictive power. As a consequence, the code complexity increases, which requires constant improvement and optimisation on the programming side. At the same time, we would like to review and eventually reduce the complexity of the hadronic software framework.

As an example, a factory design pattern has been applied in Geant4 to avoid duplications of objects, like cross sections, which can be used by several processes or physics models. This approach has been applied also for physics lists, to provide a flexible configuration mechanism at run-time, based on macro files. Moreover, these developments open the future possibility to build Geant4 with only a specified sub-set of physics models.

Another technical development focused on the reproducibility of the simulation, i.e. the possibility to repeat an event once the random generator status at the beginning of the event is known. This is crucial for debugging rare situations which may occur after long simulations. Moreover, reproducibility in normal, sequential Geant4 simulation is an important prerequisite to verify the equivalence with multi-threaded Geant4 simulations.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / 252

Task Management in the New ATLAS Production System

Author: Maxim Potekhin[1]

Co-authors: Alexandre Vaniachine [2]; Alexei Klimentov [1]; Dmitri Golubkov [3]; Kaushik De [4]

[1] *Brookhaven National Laboratory (US)*

[2] *ATLAS*

[3] *Institute for High Energy Physics (IHEP)-Unknown-Unknown*

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

Corresponding Authors: maxim.potekhin@cern.ch, kaushik@uta.edu, dmitri.golubkov@cern.ch, alexei.klimentov@cern.ch, sasha.vanyashin@cern.ch

The ATLAS Production System is the top level workflow manager which translates physicists' needs for production level processing into actual workflows executed across about a hundred processing sites used globally by ATLAS. As the production workload increased in volume and complexity in recent years (the ATLAS production tasks count is above one million, with each task containing hundreds or thousands of jobs) there is a need to upgrade the Production System to meet the challenging requirements of the next LHC run while minimizing the operating costs. Providing a front-end and a management layer for petascale data processing and analysis, the new Production System contains generic subsystems that can be used in a wider range of applications. The main subsystems are the Database Engine for Tasks (DEfT) and the Job Execution and Definition Interface (JEDI). Based on users' requests, the DEfT subsystem manages inter-dependent groups of tasks (Meta-Tasks) and generates corresponding data processing workflows. The JEDI subsystem dynamically translates the task definitions from DEfT into workload jobs executed in the PanDA Workload Management System.

We present the requirements, design parameters, object model and concrete solutions utilized in building the DEfT subsystem, such as Component Based Software Engineering. We also explain how the use of standard software modules and data formats led to reduction of development and maintenance costs.

Software Engineering, Parallelism & Multi-Core / **194**

Using Cling/LLVM and C++11 for parametric function classes in ROOT

Author: Lorenzo Moneta[1]

Co-authors: Fons Rademakers [1]; Maciej Zimnoch [2]

[1] *CERN*

[2] *University of Wroclaw*

Corresponding Authors: fons.rademakers@cern.ch, lorenzo.moneta@cern.ch

The parametric function classes of ROOT (TFormula and TF1) have been improved using the capabilities of Cling/LLVM. We will present how formula expressions can now be compiled on the fly using the just-in-time capabilities of LLVM/Cing. Furthermore using the new features of C++ 11, one can build complex function expressions by re-using the existing mathematical functions. We will show also the possibility of implementing auto-differentiation for having an automatic derivative computation of the functions.

Poster presentations / **384**

Upgrading HFGFlash for Faster Simulation at Super LHC

Author: Rahmat Rahmat[1]

[1] *University of Mississippi (US)*

Corresponding Author: rahmat.rahmat@cern.ch

HFGFlash is a very fast simulation of electromagnetic showers using parameterizations of the profiles in Hadronic Forward Calorimeter. HF GFlash has good agreement to Collision Data and previous Test Beam results. In addition to good agreement to Data and previous Test Beam results,

HFGFlash can simulate about 10000 times faster than Geant4. We will report the latest development of HFGFlash for Faster Simulation at Super LHC.

Poster presentations / 50

Preparing the Gaudi-Framework and the DIRAC-WMS for Multi-core Job Submission

Author: Nathalie Rauschmayr[1]

[1] *CERN*

Corresponding Author: nathalie.rauschmayr@cern.ch

Due to the continuously increasing number of cores on modern CPUs, it is important to adapt HEP applications. This must be done at different levels: the software which must support parallelization and the scheduling has to differ between multicore and singlecore jobs. The LHCb software framework (GAUDI) provides a parallel prototype (GaudiMP), based on the multiprocessing approach. It allows a reduction of the overall memory footprint and a coordinated access to data via separated reader and writer processes. A comparison between the parallel prototype and multiple independent Gaudi jobs in respect to CPU-time and memory consumption will be shown. In the context of parallelization speedup is the most important metric, as it shows how software scales with the number of cores. It is influenced by many factors, due to software limitations like synchronization, but also due to hardware configurations, like frequency scaling. Those limitations and their dependencies will be discussed and the influence of hardware features will be evaluated, in order to forecast the spread in CPU-time. Furthermore, speedup must be predicted in order to find the limit beyond which the parallel prototype (GaudiMP) does not support further scaling. This number must be known as it indicates the point, where new technologies must be introduced into the software framework. In order to reach further improvements in the overall throughput, scheduling strategies for mixing parallel jobs can be applied. It allows overcoming limitations in the speedup of the parallel prototype. Those changes require modifications at the level of the workload management system (DIRAC). Results will be presented for the reconstruction, simulation and analysis software of the LHCb experiment.

Poster presentations / 459

Round-tripping DIRAC: Automated Model-Checking of Concurrent Software Design Artifacts

Authors: Adrian Casajus Ramo[1]; Daniela Remenska[2]; Jeff Templon[2]

[1] *University of Barcelona (ES)*

[2] *NIKHEF (NL)*

Corresponding Author: danielar@nikhef.nl

A big challenge in concurrent software development is early discovery of design errors which can lead to deadlocks or race-conditions. Traditional testing does not always expose such problems in complex distributed applications. Performing more rigorous formal analysis, like model-checking, typically requires a model which is an abstraction of the system. For object-oriented software, UML is the industry-adopted modeling language, offering behavioral views that capture the dynamics of the system with features for modeling code-like structures, such as loops, conditions, and referring to existing interactions. We present an automatic procedure for translating UML into mCRL2 process algebra models, amenable to automatic model-checking. Our prototype is able to produce a formal model, and feed model-checking traces back into any UML modeling tool, without the user having to leave the UML domain. We apply our methodology to the newly developed Executors framework,

part of DIRAC's WMS system responsible for orchestrating the workflow steps of jobs submitted to the grid. Executors process any task sent to them by a Dispatcher, each one responsible for a different step (such as resolving the input data for a job). The Dispatcher takes care of persisting the jobs states and distributing them among the Executors. Occasionally, tasks submitted in the system would not get dispatched out of the queue, despite the fact that their responsible Executors were idle at the moment. The root cause of this problem could not be identified by certification testing with different workload scenarios, nor by analysis of the generated logs. We used our toolset to generate an mCRL2 model of this system, based on the reverse-engineered sequence diagrams. Model-checking the generated mCRL2 model discovered a trace violating the desired progress requirement, the bug being localized in the Dispatcher component implementation. The trace was automatically translated back to the UML domain, showing an intuitive view of the communication between components which leads to the faulty behavior.

Event Processing, Simulation and Analysis / 166

Geant4 studies of the CNAO facility system for hadrontherapy treatment of uveal melanomas

Author: Adele Rimoldi[1]

Co-authors: Andrea Fontana [2]; Pierluigi Piersimoni [3]

[1] *Universita e INFN (IT)*

[2] *Universita degli Studi di Pavia*

[3] *Universita de Pavia and INFN*

Corresponding Authors: adele.rimoldi@cern.ch, pierluigi.piersimoni@pv.infn.it

The Italian National Centre of Hadrontherapy for Cancer Treatment (CNAO –Centro Nazionale di Adroterapia Oncologica) in Pavia, Italy, has started the treatment of selected cancers with the first patients in late 2011. In the coming months at CNAO plans are to activate a new dedicated treatment line for irradiation of uveal melanomas using the available active beam scan. The beam characteristics and the experimental setup should be tuned in order to reach the necessary precision required for such treatments. Collaboration between CNAO, University of Pavia and INFN has started in 2011 for studying the feasibility of these specialized treatments with the aim of implementing a detailed simulation of the beam-line and comparing the obtained simulation results with the test measurements at CNAO. The goal is to optimize a new dedicated beam-line and to find the best conditions for an optimal patient irradiation. The application studied in this paper describes the simulation with the Geant4 tool of the CNAO setup with the passive/active components on the expected beam-line as well as a modelized human eye with a tumour inside. The simulation tool could be also used to test possible treatment planning systems. The results illustrate the possibility to adapt the CNAO standard transport beam line. With the suggested modifications studied in this paper for dose delivery, uveal melanoma in human eye could be treated by optimizing the positioning of the present passive elements on the standard line and with the addition of new passive elements to better shape the beam for this dedicated study.

Facilities, Infrastructures, Networking and Collaborative Tools / 217

Production Large Scale Cloud Infrastructure Experiences at CERN

Author: Tim Bell[1]

Co-authors: Belmiro Rodrigues Moreira [2]; Jan van Eldik [1]; Jose Castro Leon [3]; Ulrich Schwickerath [1]

[1] *CERN*

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

[3] *Universidad de Oviedo (ES)*

Corresponding Authors: belmiro.daniel.rodrigues.moreira@cern.ch, tim.bell@cern.ch, belmiro.moreira@cern.ch

CERN's Infrastructure as a Service cloud is being deployed in production across the two data centres in Geneva and Budapest.

This talk will describe the experiences of the first six months of production, the different uses within the organisation and the outlook for expansion to over 15,000 hypervisors based on OpenStack by 2015.

The open source toolchain used, accounting and scheduling approaches and scalability challenges will be covered.

Poster presentations / **127**

Distributing CMS Data between the Florida T2 and T3 Centers using Lustre and Xrootd-fs

Authors: Dimitri Bourilkov[1]; Jorge Luis Rodriguez[2]

[1] *University of Florida (US)*

[2] *UNIVERSITY OF FLORIDA*

We have developed remote data access for large volumes of data over the Wide Area Network based on the Lustre filesystem and Kerberos authentication for security. It this paper we explore a prototype for two-step data access from worker nodes at Florida T3 centers, located behind a firewall and using a private network, to data hosted on the Lustre filesystem at the University of Florida CMS T2 center. The T2-T3 links are 10 Gigabit per second, and the typical round trip times are 10-15 msec. For each T3 center we use a client which mounts securely the Lustre filesystem and hosts a Xrootd server. The worker nodes access the data from the T3 client using POSIX compliant tools via the Xrootd-fs filesystem. We perform scalability tests with up to 200 jobs running in parallel on the T3 worker nodes.

Poster presentations / **405**

Automatic Tools for Enhancing the Collaborative Experience in Large Projects

Authors: Dimitri Bourilkov[1]; Jorge Luis Rodriguez[2]

[1] *University of Florida (US)*

[2] *UNIVERSITY OF FLORIDA*

Corresponding Author: bourilkov@phys.ufl.edu

With the explosion of big data in many fields, the efficient management of knowledge about all aspects of the data analysis gains in importance. A key feature of collaboration in large scale projects is keeping a log of what and how is being done - for private use and reuse and for sharing selected parts with collaborators and peers, often distributed geographically on an increasingly global scale. Even better if this log is automatic, created on the fly while a scientist or software developer is working in a habitual way, without the need for extra efforts. This saves human time and enables a team to do more with the same resources. The CODESH - COllaborative

DEvelopment SHell - and CAVES - Collaborative Analysis Versioning Environment System projects address this problem in a novel way. They build on the concepts of virtual states and transitions to enhance the collaborative experience by providing automatic persistent virtual logbooks. CAVES is designed for sessions of distributed data analysis using the popular ROOT framework, while CODESH generalizes the same approach for any type of work on the command line in typical UNIX shells like bash or tcsh. Repositories of sessions can be configured dynamically to record and make available the knowledge accumulated in the course of a scientific or software endeavor. Access can be controlled to define logbooks of private sessions or sessions shared within or between collaborating groups. A typical use case is building working scalable systems for analysis of Petascale volumes of data as encountered in the LHC experiments. Our approach is general enough to find applications in many fields.

Summaries / 519

Summary of track 3A

Corresponding Author: stefan.roiser@cern.ch

Poster presentations / 1

A Voyage to Arcturus

Authors: Gareth Roy[1]; Mark Mitchell[2]

Co-authors: David Britton [3]; David Crooks [3]; Samuel Cadellin Skipsey ; Stuart Purdie [4]

[1] *U*

[2] *University of Glasgow*

[3] *University of Glasgow (GB)*

[4] *University of Glasgow-Unknown-Unknown*

Corresponding Authors: mark.mitchell@glasgow.ac.uk, gareth.roy@glasgow.ac.uk

With the current trend towards “On Demand Computing” in big data environments it becomes crucial that the deployment of services and resources becomes increasingly automated. With open-source projects such as Canonicals MaaS and Redhats Spacewalk; automated deployment is available for large scale data centre environments but these solutions can be too complex and heavyweight for smaller, resource constrained WLCG Tier-2 sites. Along with a greater desire for bespoke monitoring and the collection of more Grid related metrics, a more lightweight and modular approach is desired.

In this paper work carried out on the test cluster environment at the Scotgrid site of the University of Glasgow is presented. Progress towards a lightweight automated framework for building WLCG grid sites is presented, based on “off the shelf” software components such as Cobbler and Puppet, the building blocks of the larger open source projects mentioned before.

Additionally the test cluster is used to investigate these components in a mixed IPv4/IPv6 environment, as well as using emerging OpenFlow technologies for software service provisioning.

As part of the research into an automation framework the use of IPMI and SNMPv2 for physical device management will be included, as well as the possibility of SNMPv2 as a monitoring/data sampling layer such that more comprehensive decision making can take place and potentially be

automated. This could lead to reduced down times and better performance as services are recognised to be in a non-functional state by autonomous systems.

Finally, through the use of automated service provisioning and automated device management the building blocks of a fully automated expert system will be touched upon.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / 401

Running a typical ROOT HEP analysis on Hadoop/MapReduce

Author: Stefano Alberto Russo[1]

Co-authors: Marina Cobal [1]; Michele Pinamonti [1]

[1] *Universita degli Studi di Udine (IT)*

Corresponding Author: stefano.alberto.russo@cern.ch

Hadoop/MapReduce is a very common and widely supported distributed computing framework. It consists in a scalable programming model named MapReduce, and a locality-aware distributed file system (HDFS). Its main feature is to implement data locality: through the fusion of computing and storage resources and thanks to the locality-awareness of HDFS, the computation can be scheduled on the nodes where data resides, therefore completely avoiding network bottlenecks and congestion. Thanks to this feature a Hadoop cluster can be easily scaled out. If one takes into account also its wide diffusion and support, it is clear that managing and processing huge amounts of data becomes an easier task.

The main difference between the original goal of Hadoop and High Energy Physics (HEP) analyses is that the first consists in analysing plain text files with simple programs, while in the latter data is highly structured and complex programs are required to access the physics information and perform the analysis.

We investigated how a typical HEP analysis can be performed using Hadoop/MapReduce, in a way which is completely transparent to ROOT, the data and the user. The method we propose relies on a "conceptual middleware" that allows to run ROOT without any modification, to store the data on HDFS in its original format, and to let the user deal with Hadoop/MapReduce in a classical way which does not require any specific knowledge about this new model. Hadoop's features, and in particular data locality, are fully exploited. The developed workflow and solutions can be easily adopted for almost any HEP analysis code, and in general for any complex code working on binary data relying on independent sub-problems.

The proposed approach has been tested on a real case, an analysis to measure the top quark pair production cross section with the ATLAS experiment. The test worked as expected, bringing great benefits in terms of reducing by several orders of magnitude the network transfers required to perform the analysis with respect to a classic computing model.

Poster presentations / 315

LHC Grid Computing in Russia- present and future

Authors: V. Ilyin[1]; V. Korenkov[2]; V. Velikhov[1]

Co-authors: A. Dolbilov [2]; E. Tikhonenko [2]; Eygene Ryabinkin [3]; F. Checherov [1]; I. Lyalin [1]; I. Tkachenko [1]; R. Kolchin [1]; S. Shmatov [2]; T. Strizh [2]; V. Mitsyn [2]; V. Trofimov [2]; V. Zhiltsov [2]; Y. Lazin [1]

[1] *National Research Centre "Kurchatov Institute", Moscow*

[2] *Joint Institute for Nuclear Research, Dubna*

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

Corresponding Authors: rea@grid.kiae.ru, ilyin@theory.sinp.msu.ru, korenkov@cv.jinr.ru, velikhovve@kiae.ru, elena.tikhonenko@cern.ch

The review of the distributed grid computing infrastructure for LHC experiments in Russia is given. The emphasis is placed on the Tier-1 site construction at the National Research Centre "Kurchatov Institute" (Moscow) and the Joint Institute for Nuclear Research (Dubna).
In accordance with the protocol between CERN, Russia and the Joint Institute for Nuclear Research (JINR) on participation in LCG Project approved in 2003 and Memorandum of Understanding (MoU) on Worldwide LHC Computing Grid (WLCG) signed in October of 2007. Russia and the Joint Institute for Nuclear Research bear responsibility for nine Tier-2 centers. Here and now Russia and JINR computing Tier-2 infrastructure fully satisfies the WLCG Computing Requirements and provides proper support of the LHC experiments' Data Processing and Analysis Tasks.
In March of 2011 the proposal to create the LCG Tier1 center as an integral part of the central data handling service of the LHC Experiments in Russia was expressed in the official letter by Minister of Science and Education of Russia to CERN Director General.
In 2011 The Federal Target Programme Project: «Creation of the automated system of data processing for experiments at the Large Hadron Collider of Tier-1 level and maintenance of Grid services for distributed analysis of this data» was approved for the period 2011-2013.
The Project is aimed at the creation of a Tier-1 computer-based system in Russia and JINR for the processing of experimental data received from LHC and provisioning of Grid services for a subsequent analysis of this data at the distributed centers of the LHC computing Grid. It is shared that the National Research Centre "Kurchatov Institute" (Moscow) is responsible primarily for support of ALICE, ATLAS, and LHC-B experiments while the JINR (Dubna) provides Tier-1 services for the CMS experiment.
The master construction plan consists of two phases. The first phase is the construction of prototype in the middle of 2013 and the second one is building of full-scale fully functional Tier-1 which has to be completed in 2014.

Data Stores, Data Bases, and Storage Systems / 29

ECFS: A decentralized, distributed and fault-tolerant FUSE filesystem for the LHCb online farm

Author: Tomasz Rybczynski[1]

Co-authors: Enrico Bonaccorsi [2]; Niko Neufeld [2]

[1] *AGH University of Science and Technology (PL)*

[2] *CERN*

Corresponding Author: tomasz.rybczynski@cern.ch

The LHCb experiment records millions of proton collisions every second, but only a fraction of them are useful for LHCb physics.
In order to filter out the "bad events" a large farm of x86-servers (˜2000 nodes) has been put in place. These servers boot from and run from NFS, however they use their local disk to temporarily store data, which cannot be processed in real-time ("data-deferring"). These events are subsequently processed, when there are no live-data coming in. The effective CPU power is thus greatly increased. This gain in CPU power depends critically on the availability of the local disks. For cost and power-reasons, mirroring (RAID-1) is not used, leading to a lot of operational headache with failing disks and disk-errors or server failures induced by faulty disks.

To mitigate these problems and increase the reliability of the LHCb farm, while at same time keeping cost and power-consumption low, an extensive research and study of existing highly available and distributed file systems has been done. While many distributed file systems are providing reliability by "file replication", none of the evaluated one supports erasure algorithms.

A decentralised, distributed and fault-tolerant “write once read many” file system has been designed and implemented as a proof of concept providing fault tolerance without using expensive - in terms of disk space - file replication techniques and providing a unique namespace as a main goals.

This paper describes the design and the implementation of the Erasure Codes File System (ECFS) and presents the specialised FUSE interface for Linux.
Depending on the encoding algorithm ECFS will use a certain number of target directories as a back-end to store the segments that compose the encoded data. When target directories are mounted via nfs/autofs - ECFS will act as a file-system over network/block-level raid over multiple servers.

Poster presentations / **241**

ATLAS software configuration and build tool optimisation

Author: Grigori Rybkin[1]

[1] *Universite de Paris-Sud 11 (FR)*

Corresponding Author: grigori.rybkine@cern.ch

The ATLAS software code base is over 7 million lines organised in about 2000 packages. It makes use of some 100 external software packages, is developed by more than 400 developers and used by more than 2500 physicists from over 200 universities and laboratories in 6 continents. To meet the challenge of configuration and building of this software, the Configuration Management Tool (CMT) is used. CMT expects each package to describe its build targets, build and environment setup parameters, dependencies on other packages in a text file called requirements, and each project (group of packages) to describe its policies and dependencies on other projects in a text project file. Based on the effective set of configuration parameters read from the requirements files of dependent packages and project files, CMT commands build the packages, generate the environment for their use,
or query the packages.

The main focus was on build time performance that was optimised within several approaches:
- reduction of the number of reads of requirements files that are now read once per package by a CMT build command that generates cached requirements files for subsequent CMT build commands;
- introduction of more package level build parallelism, i.e., dependent applications and libraries are compiled in parallel;
- code optimisation of CMT commands used for build;
- introduction of project level build parallelism, i.e., parallelise the build of independent packages.

By default, CMT launches NUMBER-OF-PROCESSORS build commands in parallel. The other focus was on CMT commands optimisation in general that made them about 2 times faster.
CMT can generate a cached requirements file for the environment setup command, which is especially useful for deployment on distributed file systems like AFS or CERN VMFS.

The use of parallelism, caching and code optimisation significantly - by several times - reduced software build time, environment setup time, increased the efficiency of multi-core computing resources utilisation, and considerably improved software developer and user experience.

Poster presentations / **341**

ILCDIRAC, a DIRAC extension for the Linear Collider community

Authors: Andre Sailer[1]; Christian Grefe[1]; Stephane Guillaume Poss[2]

Co-author: Andrei Tsaregorodtsev [2]

[1] *CERN*

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

Corresponding Authors: andre.philippe.sailer@cern.ch, christian.grefe@cern.ch

ILCDIRAC was initially developed in the context of the CLIC Conceptual
Design Report (CDR), published in 2012-2013. It provides a convenient interface for the mass production of the simulation events needed for the physics performance studies of the two detectors concepts considered, ILD and SID. It was since used in the ILC Detailed Baseline Detector (DBD) studies of the SID detector concept, and is currently moving towards a complete unification of the production systems for the whole Linear Collider community. The CALICE collaboration uses part of the ILC software tools and can transparently use ILCDIRAC for its activities.
ILCDIRAC extends the core functionality of DIRAC for multiple aspects: There are currently 14 applications supported, which have very different inherent interfaces.
ILCDIRAC simplifies this situation for user convenience and maximum flexibility. For that purpose the relationship between application and
job definition was reviewed and they are completely separated. The base application and job classes are designed in a generic way, allowing for simple extension. This design is independent of the Linear Collider use case and can be applied in other contexts. Another specificity of ILCDIRAC is the management of the so called Overlay. The Linear Collider experiments will be subject to intense machine induced backgrounds and physics backgrounds like gamma gamma -> hadrons. For realistic studies these backgrounds need to be included in the simulation.
Instead of repeating the simulation of these background events, they are overlaid to the signal events during digitisation. The overlay files are randomly selected from a pool of available files and supplied to the job.
The constraint in that system is the fact that many events are needed as input (up to 200 background events per signal event), so the overlay files represent a very large sample per job to be obtained from the Storage Elements. To avoid data access issues, the system prevents too many concurrent jobs to query the Storage Elements at the same time, avoiding problems due to Storage Element overload.
The design of the software management of ILCDIRAC ensures maximum availability by using the shared area when/where possible. In case the shared area has no “Role” protection, typical for OSG sites, a locking procedure is implemented to ensure that no job will overwrite an existing software being installed. The procedure also validates the input software tar balls and their content via md5 check sum verifications.
These ILCDIRAC specific aspects rely heavily on the DIRAC features, in particular the File Catalog, which was mostly developed for ILCDIRAC. In addition, the Transformation System is used to produce all the events, in particular for the CLIC CDR and the SID DBD. It successfully generated, simulated, reconstructed more than 100 million events in 2.5 years, not counting for the user activities, and the File Catalog contains nearly 7 million files corresponding to approximately 1 PB.

Conference closing / **482**

CHEP2015: Okinawa

Author: Hiroshi Sakamoto[1]

[1] *University of Tokyo (JP)*

Corresponding Authors: sakamoto@icepp.s.u-tokyo.ac.jp, tomoaki.nakamura@cern.ch

Data Acquisition, Trigger and Controls / **74**

Automating the CMS DAQ

Author: Hannes Sakulin[1]

Co-authors: Andre Georg Holzner [2]; Andrea Petrucci [1]; Andrei Cristian Spataru [1]; Attila Racz [1]; Aymeric Arnaud Dupont [1]; Carlos Nunez Barranco Fernandez [1]; Christian Deldicque [1]; Christian Hartl [1]; Christoph Paus [3]; Christoph

Schwick [1]; Christopher Colin Wakefield [4]; Dominique Gigi [1]; Emilio Meschi [1]; Fabian Stoeckli [3]; Frank Glege [1]; Frans Meijers [1]; Gerry Bauer [3]; Giovanni Polese [5]; James Gordon Branson [2]; Jose Antonio Coarasa Perez [1]; Juan Pablo Gomez [6]; Konstanty Sumorok [3]; Lorenzo Masetti [1]; Luciano Orsini [1]; Marc Dobson [1]; Marco Pieri [2]; Matteo Sani [2]; Olivier Chaze [1]; Olivier Raginel [3]; Petr Zejdl [1]; Remi Mommsen [7]; Robert Gomez-Reino Garrido [1]; Samim Erhan [8]; Sergio Cittolin [2]; Srecko Morovic [9]; Ulf Behrens [10]; Vivian O'Dell [11]; Wojciech Andrzej Ozga [12]

[1] *CERN*

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

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

[4] *Staffordshire University (GB)*

[5] *University of Wisconsin (US)*

[6] *IFIC*

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

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

[9] *Institute Rudjer Boskovic (HR)*

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

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

[12] *AGH University of Science and Technology (PL)*

Corresponding Author: hannes.sakulin@cern.ch

We present the automation mechanisms that have been added to the Data Acquisition and Run Control systems of the Compact Muon Solenoid (CMS) experiment during Run 1 of the LHC, ranging from the automation of routine tasks to automatic error recovery and context-sensitive guidance to the operator. These mechanisms helped CMS to maintain a data taking efficiency above 90% and to even improve it to 95% towards the end of Run 1, despite an increase in the occurrence of single-event upsets in sub-detector electronics at high LHC luminosity.

Poster presentations / 156

Detector and Event Visualization with SketchUp at the CMS Experiment

Author: Tai Sakuma[1]

Co-author: Thomas Mc Cauley [2]

[1] *Texas A & M University (US)*

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

Corresponding Authors: tai.sakuma@cern.ch, thomas.mccauley@cern.ch

We describe the creation of 3D models of the CMS detector and events using SketchUp, a 3D modelling program. SketchUp provides a Ruby API with which we interface with the CMS Detector Description, the master source of the CMS detector geometry, to create detailed 3D models of the CMS detector. With the Ruby API we also interface with the JSON-based event format used for the iSpy event display to create 3D models of events. These models have many applications related to 3D representation of the CMS detector and events. Figures produced based on the models were used in conference presentations, paper publications, technical design reports for the detector upgrade, and other formal and informal documentation. Furthermore, several art projects, exhibitions, and outreach programs using the models are being planned. We describe technical implementation, show models created, and discuss present and future applications.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / 40

FTS3 –Robust, simplified and high-performance data movement service for WLCG

Author: Michail Salichos[1]

Co-authors: Alejandro Alvarez Ayllon [1]; Michal Kamil Simon [1]; Oliver Keeble [1]

[1] *CERN*

Corresponding Author: michail.salichos@cern.ch

FTS is the service responsible for distributing the majority of LHC data across the WLCG infrastructure. From the experiences of the last decade supporting and monitoring FTS, reliability, robustness and
high-performance data transfers has proved to be of high importance in the Data Management world. We are going to present the current status and features of the new File Transfer Service (FTS3), which address the problems that the previous FTS version face with: static channel model, configuration and scalability problems, new protocols support, more database back-ends support, etc. We present the solution we proposed and the design of the new tools as well the reliability, stability, scalability and performance requirements of a data movement middle-ware in the grid environment. The service has already undergone extensive pre-production validation and we report the results of high volume production transfers performed on the pilot service.

Anticipating the upcoming data movement needs of WLCG, and building on the lessons learned during the first run, we present a new, scalable and highly-optimized data movement service, which provides a simple interface for transfer job submission, status retrieval, advanced monitoring capabilities, multiple access and transfer protocols support and simplified configuration.

Transfer auto-tuning (dynamically adjusting the number of active
transfers based on success/failure rate and achieved throughput),
endpoint-centric VO share configuration, multiple replicas support,
REST-style interface for transfer submission and status retrieval,
staging files from archive, support for Oracle and MySQL database
back-ends, multiple transfer and access protocols support using gfal2
plug-in mechanism (namely SRM, gsiftp, http and xroot are already
implemented) and session/connection reuse (gridftp, ssl, etc), are only some of the new features and functionality that FTS3 has been delivered with. In order to be a credible long-term platform for data transfer, FTS3 has been designed to exploit upcoming developments in networking, such as integrating monitoring data from perfsonar for further transfer optimization, resource management and monitoring network state.

FTS3 aims to become the new data movement service for the WLCG
infrastructure.

Poster presentations / 342

Distributed cluster testing using new virtualized framework for XRootD

Authors: Justin Lewis Salmon[1]; Lukasz Janyst[2]

[1] *University of the West of England (GB)*

[2] *CERN*

Corresponding Authors: justin.lewis.salmon@cern.ch, lukasz.janyst@cern.ch

The Extended ROOT Daemon (XRootD) is a distributed, scalable system for low-latency clustered data access. XRootD is mature and widely used in HEP, both standalone and as core functionality for the EOS system at CERN, and hence requires extensive testing to ensure general stability. However, there are many difficulties posed by distributed testing, such as cluster initialization, synchronization, orchestration, inter-cluster communication and controlled failure handling.

A three-layer master/hypervisor/slave model is presented to ameloirate these difficulties by utilizing libvirt and QEMU/KVM virtualization technologies to automate spawning of configurable virtual clusters and orchestrate multi-stage test suites. The framework also incorporates a user-friendly web interface for scheduling and monitoring tests.

The framework has been used successfully to build new test suites for XRootD and EOS with existing unit test integration. It is planned for the future to sufficiently generalize the framework to encourage usage by potentially any distributed system.

Software Engineering, Parallelism & Multi-Core / 174

Evaluating Predictive Models of Software Quality

Authors: Davide Salomoni[1]; Elisabetta Ronchieri[1]; Marco Canaparo[1]; Vincenzo Ciaschini[1]

[1] *INFN CNAF*

Corresponding Authors: vincenzo.ciaschini@cnaf.infn.it, marco.canaparo@cnaf.infn.it, elisabetta.ronchieri@cnaf.infn.it, davide.salomoni@cnaf.infn.it

Software packages in our scientific environment are constantly growing in size, and are written by any number of developers. This implies a strong churn on the code itself, and an associated risk of bugs and stability problems. This risk is unavoidable as long as the software undergoes active evolution, as it always happens with software that is still in use. However, the necessity of having production systems goes against this. In fact, in this case stability and predictability are promoted over most other factors; in addition, a short turn-around time for the bug discovery-correction-deployment cycle is normally required.
We suggest that a way to address these two opposite needs is to evaluate models that define software quality. These models should offer a reasonable approximation of the "risk" associated to a program, specifically in relation to stability and maintainability. The final goal would then be to assess software maturity and, for example, to identify release milestones when a risk lower than an agreed threshold is achieved.

In this article we evaluate several quality predictive models, such as ARMOR to automatically identify the operational risks of software program modules, ComPare to give an overall quality prediction, and Case-based reasoning (CBR) to predict the quality of software components. We then apply these models to the development history of some packages released in the context of the European Middleware Initiative (EMI) with the intent to discover the risk factor associated by each model to a given program, to compare it with its real history. Finally, we attempt to determine which of the models best maps reality for the applications under evaluation, and conclude suggesting directions for further studies.

Poster presentations / 236

Lessons learned from the ATLAS performance studies of the Iberian Cloud for the first LHC running period

Author: Victoria Sanchez Martinez[1]

Co-authors: Alexey Sedov [2]; Andreu Pacheco Pages [3]; Carlos Borrego Iglesias [4]; Goncalo Borges ; Helmut Wolters [5]; Jorge Oliveira Gomes [6]; Jose Del Peso [7]; Jose Salt [8]; Manuel Delfino Reznicek [9]; Miguel Villaplana Perez [8]; Santiago Gonzalez De La Hoz [8]

[1] *Instituto de Fisica Corpuscular (IFIC) UV-CSIC (ES)*

[2] *Universitat Autònoma de Barcelona*

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

[4] *Universidad Politecnica de Madrid (ES)*

[5] *LIP Coimbra, Portugal*

[6] *LIP Laboratorio de Instrumentaco e Fisica Experimental de Particulas*

[7] *Universidad Autonoma de Madrid (ES)*

[8] *Universidad de Valencia (ES)*

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

Corresponding Authors: victoria.sanchez.martinez@cern.ch, goncalo.borges@cern.ch, carlos.borrego@cern.ch, jose.delpeso@uam.es, delfino@pic.es, jorge@lip.pt, santiago.gonzalezdelahoz@cern.ch, pacheco@ifae.es, jose.salt@cern.ch, miguel.villaplana.perez@cern.ch, helmut@coimbra.lip.pt

In this contribution we expose the performance of the Iberian (Spain and Portugal) ATLAS cloud during the first LHC running period (March 2010-January 2013) in the framework of the GRID Computing and Data Model. The evolution of the resources for CPU, disk and tape in the Iberian Tier1 and Tier2s is summarized. The data distribution over all ATLAS destinations is shown, focusing in the number of files transferred and the size of the data. The status and distribution of simulation and analysis jobs within the cloud are discussed. The distributed analysis tools used to perform physics analysis are explained as well. Cloud performance in terms of the availability and reliability of its sites is discussed. The effect of the changes in the ATLAS Computing Model on the cloud is analyzed. Finally, the readiness of the Iberian cloud towards the 1st Long Shut Down (LS1) is evaluated and an outline of the foreseen actions to take in the coming years is given. The shutdown will be a good opportunity to improve and evolve the ATLAS Distributed Computing system to prepare for the future challenges of the LHC operation.

Poster presentations / **190**

The LHCb Silicon Tracker - Control system specific tools and challenges

Author: Sandra Saornil Gamarra[1]

[1] *Universitaet Zuerich (CH)*

Corresponding Author: sandra.saornil@cern.ch

The experiment control system of the LHCb experiment is continuously evolving and improving. The guidelines and structure initially defined are kept, and more common tools are made available to all sub-detectors. Although the main system control is mostly integrated and actions are executed in common for the whole LHCb experiment, there is some degree of freedom for each sub-system to implement the control system using these tools or by creating new ones.

The implementation of the LHCb Silicon Tracker control system was extremely disorganized and with little documentation. This was due to either lack of time and manpower, and/or to limited experience and specifications. Despite this, the Silicon Tracker control system has behaved well during the first LHC run. It has continuously evolved since the start of operation and been adapted to the needs of operators with very different degrees of expertise. However, improvements and corrections have been made on a best effort basis due to time constraints placed by the need to have a fully operating detector. The system will be transformed by an ambitious rework of the code which will take place in the first months of the LS1. Performance issues with the regard to configuration and monitoring will be addressed, and the maintainability and use of the system will be improved.

This work describes the main tools which have been created specifically for the Silicon Tracker operation including the safety tree and the automated safety actions which are implemented to prevent damage to the detector electronics. In addition, the automation of recurrent tasks related mostly to data mining, error detection and recovery will be discussed. It describes also the new features and improvements that will be introduced after the code rework during the LS1, and a summary of the main tasks needed to accomplish this.

Poster presentations / 249

Experiment Dashboard Task Monitor for managing ATLAS user analysis on the Grid

Author: Laura Sargsyan[1]

Co-authors: David Tuckett [2]; Edward Karavakis [2]; Jaroslava Schovancova [3]; Julia Andreeva [2]; Lukasz Kokoszkiewicz [2]; Manoj Jha [4]; Pablo Saiz [2]

[1] *ANSL (Yerevan Physics Institute) (AM)*

[2] *CERN*

[3] *Brookhaven National Laboratory (US)*

[4] *Purdue University (US)*

Corresponding Authors: laura.sargsyan@cern.ch, julia.andreeva@cern.ch, manoj.jha@cern.ch, edward.karavakis@cern.ch, lukasz.kokoszkiewicz@cern.ch, pablo.saiz@cern.ch, jaroslava.schovancova@cern.ch, david.tuckett@cern.ch

The organization of the distributed user analysis on the Worldwide LHC Computing Grid (WLCG) infrastructure is one of the most challenging tasks among the computing activities at the Large Hadron Collider. The Experiment Dashboard offers a solution that not only monitors but also manages (kill, resubmit) user tasks and jobs via a web interface. The ATLAS Dashboard Task Monitor provides analysis users with a tool that is operating system and Grid environment independent. This contribution describes the functionality of the application and its implementation details, in particular authentication, authorization and audit of the management operations.

Software Engineering, Parallelism & Multi-Core / 244

Computing on Knights and Kepler Architectures

Authors: Davide Salomoni[1]; Francesco Giacomini[2]; Gaetano Maron[1]; Marcello Pivanti[3]; Marco Caberletti[4]; Matteo Manzali[4]; Raffaele Tripiccione[4]; Sebastiano Schifano[5]

[1] *Universita e INFN (IT)*

[2] *INFN CNAF*

[3] *Universita di Ferrara (IT)*

[4] *INFN*

[5] *U*

Corresponding Authors: sebastiano.schifano@cern.ch, davide.salomoni@cnaf.infn.it

An interesting evolution in scientific computing is represented by the streamline introduction of co-processor boards that were originally built to accelerate graphics rendering and that are now being used to perform general computing tasks. A peculiarity of these boards (GPGPU, or General Purpose Graphic Processing Units, and many-core boards like the Intel Xeon Phi) is that they normally ship on the one hand with a

limited amount of on-board memory and, on the other hand, with several tens or even several thousands of processing cores. These facts normally require specific considerations when writing or adapting software so that it can efficiently run on them. In addition, programmability of these boards is often multi-faceted and needs to be carefully evaluated as well.

An INFN project called Computing on Knights and Kepler Architectures, involving several INFN sites and collaborations, has been set up to investigate the suitability of these boards for scientific computation in a range of physics-related fields. The hardware targets for these investigations are the recently released x86-based Intel Xeon Phi and the K20-based NVIDIA GPGPU Tesla boards.

We present the results of the investigations performed by this project using production samples of these boards. In particular, we will show adaptability, portability considerations, configuration tips and performance analysis for Xeon Phi and K20 cards applied to real use cases linked to the processing of data produced by experimental physics and to problems typical of theoretical physics, in single and multiple Phi and GPGPU configurations. We will also provide several micro benchmarks related to basic operations like memory copy and use of vector extensions.

Poster presentations / **288**

ATLAS Distributed Computing Monitoring tools during the LHC Run I

Author: Jaroslava Schovancova[1]

Co-authors: Alessandro Di Girolamo [2]; I Ueda [3]; Simone Campana [2]; Stephane Jezequel [4]; Torre Wenaus [1]

[1] *Brookhaven National Laboratory (US)*

[2] *CERN*

[3] *University of Tokyo (JP)*

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

Corresponding Authors: jaroslava.schovancova@cern.ch, simone.campana@cern.ch, alessandro.di.girolamo@cern.ch, stephane.jezequel@cern.ch, i.ueda@cern.ch, wenaus@gmail.com

The ATLAS Distributed Computing (ADC) Monitoring targets three groups of customers: ADC Operations, ATLAS Management, and ATLAS sites and ATLAS funding agencies. The main need of ADC Operations is to identify malfunctions early and then escalate issues to an activity or a service expert. The ATLAS Management use visualisation of long-term trends and accounting information about the ATLAS Distributed Computing resources. The ATLAS sites and the ATLAS funding agencies utilize both real-time monitoring and long-term measurement of the performance of the provided computing resources.

During the LHC Run I a significant development effort has been invested in standardization of the monitoring and accounting applications in order to provide an extensive monitoring and accounting suite. ADC Monitoring applications separate the data layer and the visualisation layer. The data layer exposes data in a predefined format. The visualisation layer is designed bearing in mind visual identity of the provided graphical elements, and reusability of the visualisation elements across the different tools. A rich family of filtering and searching options enhancing available user interfaces comes naturally with the data and visualisation layer separation.

With a variety of reliable monitoring data accessible through standardized interfaces, the possibility of automating actions under well defined conditions, correlating multiple data sources, has become

feasible. In this contribution we also discuss the automated exclusion of degraded resources and their automated recovery in different activities.

Data Acquisition, Trigger and Controls / 165

The LHCb Data Acquisition during LHC Run 1

Author: Rainer Schwemmer[1]

Co-authors: Alexander Zvyagin [2]; Beat Jost [1]; Christophe Haen [3]; Clara Gaspar [1]; Enrico Bonaccorsi [1]; Eric van Herwijnen [1]; Federico Alessio [1]; Guoming Liu [1]; Loic Brarda [1]; Markus Frank [1]; Mohamed Chebbi [1]; Niko Neufeld [1]; Richard Jacobsson [1]; Vijay Kartik Subbiah [1]

[1] *CERN*

[2] *Fakultaet fuer Physik-Ludwig-Maximilians-Univ. Muenchen*

[3] *Univ. Blaise Pascal Clermont-Fe. II (FR)*

Corresponding Author: rainer.schwemmer@cern.ch

The LHCb Data Acquisition system reads data from over 300 read-out boards and distributes them to more than 1500 event-filter servers. It uses a simple push-protocol over Gigabit Ethernet. After filtering, the data is consolidated into files for permanent storage using a SAN-based storage system. Since the beginning of data-taking many lessons have been learned and the reliability and robustness of the system has been greatly improved. We report on these changes and improvements, their motivation and how we intend to develop the system for Run 2. We also will report on how we try to optimise the usage of CPU resources during the running of the LHC ("deferred triggering") and the implications on the data acquisition.

Data Acquisition, Trigger and Controls / 18

A PCIe GEn3 based readout for the LHCb upgrade.

Authors: Beat Jost[1]; Guoming Liu[1]; Ignazio Lax[2]; Niko Neufeld[1]; Paolo Durante[1]; Rainer Schwemmer[1]; domenico galli[3]; umberto marconi[2]; vincenzo vagnoni[2]

[1] *CERN*

[2] *INFN Bologna*

[3] *Università di Bologna and INFN*

Corresponding Authors: rainer.schwemmer@cern.ch, umberto.marconi@bo.infn.it

The architecture of the data acquisition for the LHCb upgrade is designed to allow for data transmission from the front-end electronics directly to the readout boards synchronously with the bunch crossing at the rate of 40 MHz. To connect the front-end electronics to the readout boards the upgraded detector will require order of 12000 GBT based (3.2 Gb/s radiation hard CERN serializers) optical links, for a corresponding aggregate throughput of about 38 Tb/s. The readout boards act as event buffers and the data format converters for the injection of the event fragments into the network of the High Level Trigger (HLT) computing farm. The connection between the readout boards and the HLT farm has to be designed to be capable to be seamlessly scaled up to the full readout of 40 MHz bunch-crossings. The data transfer rate will be tuned by means of a new Low Level Trigger (LLT) based on custom hardware, which will allow varying the HLT input frequency in a range between 10 to 40 MHz. A readout board consists of an ATCA compliant carrier-board, hosting up to four active AMC40-card pluggable modules (mezzanines). Each AMC40-card is equipped with a single powerful FPGA (likely a last generation Stratix V by ALTERA) used for establishing high-speed serial connections and for data processing. The AMC40-card as proposed today has 24

GBT input-links and 12 output-links. All the Stratix V FPGA serializers are 10 Gb/s. The 24 input-links deliver a maximum amount of user-data of 77 Gbit/s in the GBT standard mode. The baseline for the AMC40 foresees to implement a local area network protocol (LAN) directly in the FPGA. The candidate technologies considered so far are Ethernet and InfiniBand. An alternative solution for the read-out system is to send data from the FPGAs to the HLT farm via PCIe Gen3 bus extension/expansion (at the link-level PCIe does not look very different from the LAN protocols). Data in this approach would be pushed over a suitable physical link (optical fibre for instance) from the FPGA into a PCIe custom receiver card plugged to a HLT server motherboard. PCIe Gen3 would use 8 Gb/s on the serializers. The 12 output-links of the FPGA allows to set up two PCIe devices of varying lane-count (x4 and x8) for data transmission. The PCIe hard IP blocks available in the ALTERA FPGAs are very efficient: one 8-lane block uses less than 1% of the resources. The PCIe custom receiver card consists of an optical-to-electrical transducer plus a PCIe switch chip used to adapt to the PCIe slot of the HLT server. The main architectural advantage of using PCIe Gen3 is that the LAN protocol and link-technology can be left open until very late to profit from the most cost-effective industry technology available by the time of LS2.

Facilities, Infrastructures, Networking and Collaborative Tools / **73**

Operating the Worldwide LHC Computing Grid: current and future challenges

Authors: Alessandra Forti[1]; Andrea Sciaba[2]; José Flix[None]; Maria Girone[2]

[1] *University of Manchester (GB)*

[2] *CERN*

Corresponding Author: andrea.sciaba@cern.ch

The Wordwide LHC Computing Grid project (WLCG) provides the computing and storage resources required by the LHC collaborations to store, process and analyse their data. It includes almost 200,000 CPU cores, 200 PB of disk storage and 200 PB of tape storage distributed among more than 150 sites. The WLCG operations team is responsible for several essential tasks, such as the coordination of testing and deployment of Grid middleware and services, communication with the experiments and the sites, followup and resolution of operational issues and medium/long term planning. In 2012 WLCG critically reviewed all operational procedures and restructured the organisation of the operations team as a more coherent effort in order to improve its efficiency. In this paper we describe how the new organisation works, its recent successes and the changes to be implemented during the long LHC shutdown in preparation for the LHC Run 2.

Facilities, Infrastructures, Networking and Collaborative Tools / **77**

Nagios and Arduino integratition for monitoring

Authors: Marcos Seco Miguelez[1]; Victor Manuel Fernandez Albor[1]

Co-authors: Antonio Pazos Alvarez [1]; Juan Jose Saborido Silva [1]

[1] *Universidade de Santiago de Compostela (ES)*

Corresponding Authors: victormanuel.fernandez@usc.es, marcos.seco@usc.es

The Datacenter at the Galician Institute of High Energy Physics(IGFAE) of the Santiago de Compostela University (USC) is a computing cluster with about 150 nodes and 1250 cores that hosts the LHCb Tiers 2 and 3. In this small datacenter, and of course in similar or bigger ones, it is very important to keep optimal conditions of temperature, humidity and pressure. Therefore, it is a necessity to monitor the environment and be able to trigger alarms when operating outside the recommended settings.

There are currently a plenty of tools and systems developed for Datacenter monitorization, but until recent years all of them were of comercial nature and expensive. In recent years there has been and increasing interest in the use of technologies based on Arduino due to its open hardware licensing and the low cost of this type of components. In this article we describe the system developed to monitorize IGFAE's Datacenter, which integrates an Arduino controlled sensor network with the Nagios monitoring software.
Sensors of several types, temperature, humidity or pressure, are connected to the Arduino board. The nagios software is in charge of monitoring the different sensors, with the help of nagiosgraph to keep track of the historic data and to produce the plots. An arduino program, developed in house, provides the nagios sensor with the readout of one or several sensors depending on the sensor's request. The nagios temperature sensor also broadcasts an SNMP trap when the temperature gets out of the allowed operating range.

Poster presentations / **257**

ATLAS Distributed Computing Operation Shift Teams experience during the discovery year and beginning of the Long Shutdown 1

Author: Alexey Sedov[1]

Co-authors: Alessandro Di Girolamo [2]; Armen Vartapetian [3]; Guidone Negri [2]; Hiroshi Sakamoto [4]; Iouri Smirnov [5]; Jae Yu [3]; Jaroslava Schovancova [5]

[1] *Universitat Autònoma de Barcelona*

[2] *CERN*

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

[4] *University of Tokyo (JP)*

[5] *Brookhaven National Laboratory (US)*

Corresponding Authors: asedov@pic.es, alessandro.di.girolamo@cern.ch, jaehoonyu@uta.edu, guido.negri@cern.ch, sakamoto@icepp.s.u-tokyo.ac.jp, jaroslava.schovancova@cern.ch, iouri.smirnov@cern.ch, armen.vartapetian@cern.ch

ATLAS Distributed Computing Operation Shifts were evolved to meet new requirements. New monitoring tools as well as new operational changes led to modifications in organization of shifts. In this paper we describe the roles and the impacts of the shifts to smooth operation of complex computing grid employed in ATLAS, the influence of Discovery of Higgs like particle on shift operations, the achievements in monitoring and automation that made possible to focus more on tasks that leaded to the Discovery, as well as influence of the Long Shutdown 1 and operational changes related to no beam period.

Poster presentations / **268**

ATLAS DQ2 to Rucio renaming infrastructure

Author: Cedric Serfon[1]

Co-authors: Angelos Molfetas [2]; Armin Nairz [1]; Graeme Andrew Stewart [1]; Luc Goossens [1]; Mario Lassnig [1]; Martin Barisits [1]; Ralph Vigne [3]; Thomas Beermann [4]; Vincent Garonne [1]

[1] *CERN*

[2] *University of Sydney (AU)*

[3] *University of Vienna (AT)*

[4] *Bergische Universitaet Wuppertal (DE)*

Corresponding Authors: cedric.serfon@cern.ch, martin.barisits@cern.ch, thomas.beermann@cern.ch, vincent.garonne@cern.ch, luc.goossens@cern.ch, mario.lassnig@cern.ch, angelos.molfetas@cern.ch, armin.nairz@cern.ch, graeme.andrew.stewart@cern.ch, ralph.vigne@cern.ch

The current ATLAS Distributed Data Management system (DQ2) is being replaced by a new one called Rucio. The new system has many improvements, but it requires a number of changes. One of the most significant ones is that no local file catalog like the LFC, which was a central component in DQ2, will be used by Rucio. Instead of querying a file catalogue that stores the association of files with their corresponding locations, in Rucio the physical path of a file can be ascertained by deriving it from the file's Logical File Name (LFN) via a deterministic function. Therefore, all file replicas produced by ATLAS have to be renamed before migrating to the new system. It represents about 300M files split between about 120 sites with six different storage technologies. An infrastructure to perform this is needed: it should be automated, robust, transparent for the users, fault tolerant, storage technology agnostic and require as little work as possible from site administrative personnel. It needs also to be fast enough to rename everything before the final switch to Rucio in 2014. An infrastructure following all these requirements has been developed and is described here. The technologies that have been used are also presented as well as the performance of the system.

Event Processing, Simulation and Analysis / **158**

Stitched Together: Transitioning CMS to a Hierarchical Threaded Framework

Author: Christopher Jones[1]

Co-author: Elizabeth Sexton-Kennedy [1]

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

Corresponding Authors: sexton@fnal.gov, cdj@fnal.gov

Modern computing hardware is transitioning from using a single high frequency complicated computing core to many lower frequency simpler cores. As part of that transition, hardware manufacturers are urging developers to exploit concurrency in their programs via operating system threads. We will present CMS' effort to evolve our single threaded framework into a highly concurrent framework. We will outline the design of the new framework and how the design was constrained by the initial single threaded design. Then we will discuss the tools we have used to identify and correct thread unsafe user code. Finally we will end with a description of the coding patterns we found useful when converting code to being thread safe.

Poster presentations / **114**

CMS experience of running glideinWMS in High Availability mode

Author: Igor Sfiligoi[1]

Co-author: Ian Fisk [2]

[1] *University of California San Diego*

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

Corresponding Authors: isfiligoi@ucsd.edu, ian.fisk@cern.ch

The CMS experiment at the Large Hadron Collider is relying on the HTCondor-based glideinWMS batch system to handle most of its distributed computing needs. In order to minimize the risk of

disruptions due to software and hardware problems, and also to simplify the maintenance procedures, CMS has set up its glideinWMS instance to use most of the attainable High Availability (HA) features. The setup involves running services distributed over multiple nodes, which in turn are located in several physical locations, including Geneva, Switzerland, Chicago, Illinois and San Diego, California. This paper describes the setup used by CMS, the HA limits of this setup, as well as a description of the actual operational experience spanning many months.

Poster presentations / **132**

Estimating job runtime for CMS analysis jobs

Author: Igor Sfiligoi[1]

[1] *University of California San Diego*

Corresponding Author: isfiligoi@ucsd.edu

The basic premise of pilot systems is to create an overlay scheduling system on top of leased resources. And by definition, leases have a limited lifetime, so any job that is scheduled on such resources must finish before the lease is over, or it will be killed and all the computation wasted. In order to effectively schedule jobs to resources, the pilot system thus requires the expected lifetime of the jobs. Past studies have shown that relying on user provided estimates is not a valid strategy, so the system should try to make an estimate by itself. This paper provides a description of a system that makes estimates using machine learning based on past behavior. The work was performed in the context of physics analysis jobs of the CMS experiment at the Large Hadron Collider, using the domain knowledge to improve the accuracy. The attained results are presented in the paper.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / **47**

Minimizing draining waste through extending the lifetime of pilot jobs in Grid environments

Authors: Brian Paul Bockelman[1]; Frank Wurthwein[2]; Igor Sfiligoi[2]; Terrence Martin[2]

[1] *University of Nebraska Lincoln*

[2] *University of California San Diego*

Corresponding Author: isfiligoi@ucsd.edu

The computing landscape is moving at an accelerated pace to many-core computing.
Nowadays, it is not unusual to get 32 cores on a single physical node.
As a consequence, there is increased pressure in the pilot systems domain to move from purely single-core scheduling and allow multi-core jobs as well.
In order to allow for a gradual transition from single-core to multi-core user jobs, it is envisioned that pilot jobs will have to handle both kinds of user jobs at the same time, by requesting several cores at a time from Grid providers and then partitioning them between the user jobs at runtime.
Unfortunately, the current Grid ecosystem only allows for relatively short lifetime of pilot jobs, requiring frequent draining, with the relative waste of compute resources due to varying lifetimes of the user jobs. Significantly extending the lifetime of pilot jobs is thus highly desirable, but must come without any adverse effects for the Grid resource providers.
In this paper we present a mechanism, based on communication between the pilot jobs and the Grid provider, that allows for pilot jobs to run for extended periods of time when there are available resources, but also allows the Grid provider to reclaim the resources in a short amount of time when needed. We also present the experience of running a prototype system using the above mechanism on a couple US-based Grid sites.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / 76

Cloud Bursting with Glideinwms: Means to satisfy ever increasing computing needs for Scientific Workflows

Author: Parag Mhashilkar[1]

Co-authors: Anthony Tiradani [2]; Burt Holzman [3]; Igor Sfiligoi [4]; Krista Larson [3]; Mats Rynge [5]

[1] *Fermi National Accelerator Laboratory*

[2] *Fermilab*

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

[4] *University of California San Diego*

[5] *University of South California*

Corresponding Authors: isfiligoi@ucsd.edu, parag@fnal.gov

Scientific communities have been in the forefront of adopting new technologies and methodologies in the computing. Scientific computing has influenced how science is done today, achieving breakthroughs that were impossible to achieve several decades ago. For past decade several such communities in the Open Science Grid (OSG) and the European Grid Infrastructure (EGI) have been using the Glideinwms system to run complex application work-flows to effectively share computational resources over the Grid. Glideinwms is a pilot-based workload management system (WMS) that creates on demand, dynamically-sized overlay Condor batch system on Grid resources. At present, the computational resources shared over the grid are just adequate to sustain the computing needs. We envision that the complexity of the science driven by "Big Data" will further push the need for computational resources. To fulfill their increasing demands and/or to run specialized workflows, some of the big communities like CMS are investigating the use of Cloud Computing as Infrastructure-As-A-Service (IAAS) with Glideinwms as a potential alternative to fill the void. Similarly, communities with no previous access to computing resources can use Glideinwms to setup up a batch system on the Cloud Infrastructure. To enable this architecture of Glideinwms has been extended to enable support for interfacing Glideinwms with different Scientific and commercial cloud providers like HLT, FutureGrid, FermiCloud and Amazon EC2. In this paper, we describe a solution for cloud bursting with Glideinwms. The paper describes the approach, architectural changes and lessons learned while enabling support for Cloud infrastructures in Glideinwms.

Poster presentations / 25

Using enterprise-class software to monitor the Grid - The CycleServer experience

Authors: Frank Wurthwein[1]; Igor Sfiligoi[1]; Miron Livny[2]

[1] *University of California San Diego*

[2] *University of Wisconsin - Madison*

Corresponding Author: isfiligoi@ucsd.edu

Monitoring is an important aspect of any job scheduling environment, and Grid computing is no exception. Writing quality monitoring tools is however a hard proposition, so the Open Science Grid decided to leverage existing enterprise-class tools in the context of the glideinWMS pilot infrastructure, which powers a large fraction of its Grid computing. The product chosen is the CycleServer, created and maintained by Cycle Computing LLC, for which OSG negotiated a very advantageous licensing deal.
In this poster we present an overview of its features, alongside the experience of its use by several OSG VOs.

Poster presentations / **322**

Using ssh and sshfs to virtualize Grid job submission with rcondor

Authors: Igor Sfiligoi[1]; Jeffrey M. Dost[1]

[1] *University of California San Diego*

Corresponding Author: isfiligoi@ucsd.edu

The HTCondor based glideinWMS has become the product of choice for exploiting Grid resources for many communities. Unfortunately, its default operational model expects users to log into a machine running a HTCondor schedd before being able to submit their jobs. Many users would instead prefer to use their local workstation for everything.
A product that addresses this problem is rcondor, a module delivered with the HTCondor package. RCondor provides command line tools that simulate the behavior of a local HTCondor installation, while using ssh under the hoods to execute commands on the remote node instead. RCondor also interfaces with sshfs, virtualizing access to remote files, thus giving the user the impression of a truly local HTCondor installation.
This paper presents a detailed description of rcondor, as well as comparing it to the other methods currently available for accessing remote HTCondor schedds.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / **113**

Using ssh as portal - The CMS CRAB over glideinWMS experience

Authors: Igor Sfiligoi[1]; Stefano Belforte[2]

Co-author: Ian Fisk [3]

[1] *University of California San Diego*

[2] *Universita e INFN (IT)*

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

Corresponding Authors: isfiligoi@ucsd.edu, ian.fisk@cern.ch

The User Analysis of the CMS experiment is performed in distributed way using both Grid and dedicated resources. In order to insulate the users from the details of computing fabric, CMS relies on the CRAB (CMS Remote Analysis Builder) package as an abstraction layer. CMS has recently switched from a client-server version of CRAB to a purely client-based solution, with ssh being used to interface with either HTCondor or glideinWMS batch systems. This switch has resulted in significant improvement of user satisfaction, as well as in significant simplification of the CRAB code base. This paper presents the architecture of the ssh-based CRAB package, the rationale behind it, as well as the operational experience running both the client-server and the ssh-based versions in parallel for several months.

Poster presentations / **232**

The Legnaro-Padova distributed Tier-2: challenges and results

Authors: Alberto Crescente[1]; Fulvia Costa[1]; Gaetano Maron[2]; Massimo Biasotto[2]; Massimo Sgaravatto[1]; Michele Gulmini[2]; Michele Michelotto[1]; Nicola Toniolo[2]; Roberto Ferrari[1]; Sergio Fantinel[2]; Simone Badoer[2]

[1] *INFN Padova*

[2] *INFN LNL*

Corresponding Author: massimo.sgaravatto@pd.infn.it

The Legnaro-Padova Tier-2 is a computing facility serving the
ALICE and CMS LHC experiments. It also supports other High Energy Physics experiments and other virtual organizations of different disciplines, which can opportunistically harness idle resources if
available.

The unique characteristic of this Tier-2 is its topology: the computational resources are spread in two different sites, about 15 km
apart: the INFN Legnaro National Laboratories and the INFN Padova unit,
connected through a 10 Gbps network link (it will be soon updated to 20 Gbps).
Nevertheless these resources are seamlessly integrated and are exposed
as a single computing facility.
Despite this intrinsic complexity, the Legnaro-Padova Tier-2 ranks among the best Grid sites for what concerns reliability and availability.

The Tier-2 comprises about 190 worker nodes, providing about
26000 HS06 in total.
Such computing nodes are managed by the LSF local resource management system, and are accessible using a Grid-based interface implemented through multiple CREAM CE front-ends.

dCache, xrootd and Lustre are the storage systems in use at the Tier-2:
about 1.5 PB of disk space is available to users in total, through multiple access protocols.

A 10 Gbps network link, planned to be doubled in the next months, connects the Tier-2 to WAN. This link is used for the LHC Open Network Environment (LHCONE) and for other general purpose traffic.

In this paper we discuss about the experiences at the Legnaro-Padova Tier-2: the problems that had to be addressed, the lessons learned, the implementation choices.
We also present the tools used for the daily management
operations. These include DOCET, a Java-based webtool designed, implemented and maintained at the Legnaro-Padova Tier-2,
and deployed also in other sites, such as the LHC Italian T1.
DOCET provides an uniform interface to manage all the information
about the physical resources of a computing center. It is also used as
documentation repository available to the Tier-2 operations team.

Finally we discuss about the foreseen developments of the existing
infrastructure.
This includes in particular the evolution from a Grid-based resource
towards a Cloud-based computing facility.

Poster presentations / 145

Towards Provenance and Traceability in CRISTAL for HEP

Author: Andrew Branson[1]

Co-authors: Jetendr Shamdasani [1]; Richard Mcclatchey [1]

[1] *University of the West of England (GB)*

Corresponding Authors: jetendr.shamdasani@cern.ch, andrew.branson@cern.ch

Efficient, distributed and complex software is central in the analysis of high energy physics (HEP) data. One area that has been somewhat overlooked in recent years has been the tracking of the development of the HEP software and of its use in data analyses and its evolution over time. This area of tracking analyses to provide records of actions performed, outcomes achieved and (re-)design decisions taken is an active part of computer science research known as provenance data capture and management. In recent years there has been a wealth of research conducted in the computer science community on this topic, however very little work has been done to address the requirements that have emerged from the HEP domain in the LHC era.
This paper discusses a system known as CRISTAL which has been in development and active use at CERN for the past decade. CRISTAL is a mature and very stable system which was originally developed to track the construction of the ECAL element of CMS. The current usage is discussed in this paper in the context of its application at CMS. CRISTAL has also been commercialised by two external companies. The first company, M1i (Annecy France), has developed a purely BPM (Business Process Management) solution and has sold the product (Agilium) in the retail, logistics and manufacturing sectors; in the second company, Technoledge (Geneva, Switzerland) it is being applied to fuel cell production lines with a focus on provenance data capture and management, therefore demonstrating its maturity as a provenance system.
CRISTAL is currently being moved towards an open source license, and is being used in several EC projects, one example of which is N4U (or neuGRID for Users) where a so-called Analysis Service is being developed to enable neuroimaging researchers to record and track their complex workflows and analyses. This Analysis Service allows for the reuse of clinical research analysis workflows by scientists proving its generic application as a tool for the management of scientific data. We are currently aiming to apply CRISTAL for the indexing of previous HEP data with our team based at CERN. We also feel that CRISTAL can be applied to aid scientists at CERN in creating their experiments through use of the N4U Analysis Service built on top of CRISTAL, allowing them to share, reuse or amend past HEP analyses.
In addition to analysis reuse and sharing, CRISTAL's unique approach to provenance capture provides a means for scientists to log errors and to audit which analyses can be used in conjunction with various datasets. Consequently, CRISTAL provides an unique viewpoint for investigators to see where and more importantly why their experiments may have failed and to store their results. Some initial ideas for the use of CRISTAL in HEP are outlined in detail in this paper. Currently we are investigating the feasibility of using the N4U Analysis Service or a derivative along with CRISTAL to address the requirements of data and analysis logging and provenance capture within the HEP environment.

Data Stores, Data Bases, and Storage Systems / **53**

ARIADNE: a Tracking System for Relationships in LHCb Metadata

Author: Illya Shapoval[1]

Co-authors: Marco Cattaneo [2]; Marco Clemencic [2]

[1] *CERN, KIPT*

[2] *CERN*

Corresponding Authors: illya.shapoval@cern.ch, marco.clemencic@cern.ch

The computing model of the LHCb experiment implies handling of an evolving set of heterogeneous metadata entities and relationships between them. The entities range from software and databases states to architecture specificators and software/data deployment locations. For instance, there is an important relation between the LHCb Conditions Database (CondDB), which provides versioned, time dependent geometry and conditions data, and the LHCb software, which is the data processing applications (used for simulation, high level triggering, reconstruction and analysis of physics data). The evolution of CondDB and of the LHCb applications is a weakly-homomorphic process. It means that relationships between a CondDB state and LHCb application state may not be preserved across different database and application generations. These issues may lead to various kinds of problems in the LHCb production, varying from unexpected application crashes to incorrect data processing results. In this paper we present the ARIADNE - a generic metadata relationships tracking system based on the novel NoSQL Neo4j graph database. Its aim is to track and store many thousands

of evolving relationships for the cases such as the one described above, and several others, which would otherwise remain unmanaged and potentially harmful. The highlights of the paper include the system's implementation and management details, infrastructure needed for running it, security issues, first experience of usage in the LHCb production and potential of the system to be applied to a wider set of LHCb tasks.

Poster presentations / **58**

The role of micro size computing clusters for small physics groups

Author: Andrey SHEVEL[1]

[1] *Petersburg Nuclear Physics Institute*

A small physics group (3-15 persons) might use a number of computing facilities for the analysis/simulation, developing/testing, teaching. It is discussed different types of computing facilities: collaboration computing facilities, group local computing cluster (including colocation), cloud computing. The author discuss the growing variety of different computing options for small groups and does emphasize the role of the group owned computing cluster of micro size.

Data Stores, Data Bases, and Storage Systems / **323**

DPHEP: From Study Group to Collaboration (The DPHEP Collaboration)

Author: Jamie Shiers[1]

[1] *CERN*

Corresponding Author: jamie.shiers@cern.ch

The international study group on data preservation in high energy physics, DPHEP, achieved a milestone in 2012 with the publication of its eagerly anticipated large scale report, which contains a description of data preservation activities from all major high energy physics collider-based experiments and laboratories. A central message of the report is that data preservation in HEP is not possible without long term investment in not only hardware but also human resources, and with this in mind DPHEP will evolve to a new collaboration structure in 2013. The DPHEP study group and the major conclusions from the report will be presented as well as an outline of the future working directions of the new collaboration.

Poster presentations / **365**

The Design and Performance of the ATLAS jet trigger

Author: Steven Robertson[1]

[1] *McGill*

Corresponding Author: shima.shimizu@cern.ch

The ATLAS jet trigger is an important element of the event selection process,
providing data samples for studies of Standard Model physics and searches for new

physics at the LHC. The ATLAS jet trigger system has undergone substantial modifications over the past few years of LHC operations, as experience developed with triggering in a high luminosity and high event pileup environment. In particular, the region-of-interest (ROI) based strategy has been replaced by a full scan of the calorimeter data at the third trigger level, and by a full scan of the level-1 trigger input at level-2 for some specific trigger chains. Hadronic calibration and cleaning techniques are applied in order to provide improved performance and increased stability in high luminosity data taking conditions. In this presentation we discuss the implementation and operational aspects of the ATLAS jet trigger during 2011 and 2012 data taking periods at the LHC.

Poster presentations / **400**

A data parallel digitizer for a time-based simulation of CMOS Monolithic Active Pixel Sensors with FairRoot

Author: Philipp Sitzmann[1]

Co-authors: Joachim Stroth [2]; Michael Deveaux [3]

[1] *Goethe University Frankfurt*

[2] *Goethe-University and GSI*

[3] *University Frankfurt*

Corresponding Author: philipp.sitzmann@stud.uni-frankfurt.de

CMOS Monolithic Active Pixel Sensors (MAPS) have demonstrated excellent performances as tracking detectors for charged particles. Their outstanding spatial resolution (few µm), ultra-light material budget (50 µm) and advanced radiation tolerance (> 1Mrad, >1e13 neq/cm²). They were therefore chosen for the vertex detectors of STAR and CBM and are foreseen to equip the upgraded ALICE-ITS. They also constitute a valuable option for tracking devices at future e+e- colliders.

MAPS were initially developed as sensors for photographic devices and the data is readout with a rolling shutter. The readout time of an individual frame lasts typically 10-100µs. In high rate experiments like CBM, the pixels matrix may sum particle signals generated by several particle collisions during this integration time. Powerful tracking codes are needed to disentangle those collisions based on the data obtained from the faster tracking detectors located more downstream the collision point. Developing this code requires a realistic and fast digitizer software, which represents the properties of the sensors within GEANT-based simulation frameworks like FairRoot.

We introduce the challenges related to representing collision pile-up in an event based simulation environment and discuss our simulation strategy. Moreover, we introduce our concept for data parallelism, which aims to allow for a parallel code execution in a near future.

Poster presentations / **186**

Analysis and improvement of data-set level file distribution in Disk Pool Manager

Author: Samuel Cadellin Skipsey[None]

Co-authors: David Britton [1]; David Smith [2]; Mark Mitchell [3]; Stuart Purdie [3]; Wahid Bhimji [4]

[1] *University of Glasgow (GB)*

[2] *CERN*

[3] *University of Glasgow*

[4] *University of Edinburgh (GB)*

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

Of the three most widely used implementations of the WLCG Storage Element specification, Disk Pool Manager (DPM) has the simplest implementation of file placement balancing (StoRM doesn't attempt this, leaving it up to the underlying filesystem, which can be very sophisticated in itself). DPM uses a round-robin algorithm (with optional filesystem weighting), for placing files across filesystems and servers. This does a reasonable job of evenly distributing files across the storage array provided to it. However, it does not offer any guarantees of the evenness of distribution of that subset of files associated with a given "dataset" (which often maps onto a "directory" in the DPM namespace (DPNS)). It is useful to consider a concept of 'balance', where an optimally balanced set of files indicates that the files are distributed evenly across all of the pool nodes. The best case performance of the round robin algorithm is to maintain balance, it has no mechanism to improve balance.

In the past year or more, larger DPM sites have noticed load spikes on individual disk servers, and suspected that these were exacerbated by excesses of files from popular datasets on those servers.
We present here a software tool which analyses file distribution for all datasets in a DPM SE, providing a measure of the poorness of file location in this context. Further, the tool provides a list of file movement actions which will improve dataset-level file distribution, and can action those file movements itself.
We present results of such an analysis and movement on the UKI-SCOTGRID-GLASGOW DPM.

Panel discussion / **500**

The end of HEP-specific computing as we know it?

There is an emerging trend in computing for HEP, namely, that it spills outside the traditional laboratory boundaries and benefits from becoming less HEP-specific. As all forms of research are becoming ICT-dependent, what is the high energy and nuclear physics community doing to encourage mainstream software? Or do we do exactly the opposite? Dr. Oxana Smirnova of Lund University will chair a colourful plenary panel discussion discussing these issues to look at the role of HEP computing in the coming decade.

Poster presentations / **433**

Automated Configuration Validation with Puppet & Nagios

Author: Jason Alexander Smith[1]

[1] *Brookhaven National Laboratory (US)*

Corresponding Author: smithj4@bnl.gov

Running a stable production service environment is important in any field. To accomplish this, a proper configuration management system is necessary along with good change management policies. Proper testing and validation is required to protect yourself against software or configuration changes to production services that can cause major disruptions. In this paper, we discuss how we extended our Puppet configuration management system to automatically run all configuration changes through a validation environment, which replicates all critical services and warns us of potential problems before applying these changes to our production servers.

Poster presentations / **432**

Automated Cloud Provisioning Using Puppet & MCollective

Author: Jason Alexander Smith[1]

[1] *Brookhaven National Laboratory (US)*

Corresponding Author: smithj4@bnl.gov

Public clouds are quickly becoming cheap and easy methods for dynamically adding more computing resources to your local site to help handle peak computing demands. As cloud use continues to grow, the HEP community is looking to run more than just simple homogeneous VM images, which run basic data analysis batch jobs. The growing demand for heterogeneous server configurations demands better provisioning systems that can quickly and automatically instantiate various service VMs. In this paper, we discuss how we extend our local Puppet configuration management system to help us easily provision various service instances using the Marionette Collective framework.

Facilities, Infrastructures, Networking and Collaborative Tools / **64**

The Effect of FlashCache and Bcache on I/O Performance

Author: Christopher Hollowell[1]

Co-authors: Alexandr Zaytsev [2]; Jason Alexander Smith [2]; Richard Hogue [1]; Tony Wong [1]; William Strecker-Kellogg [3]

[1] *Brookhaven National Laboratory*

[2] *Brookhaven National Laboratory (US)*

[3] *Brookhaven National Lab*

Corresponding Authors: smithj4@bnl.gov, hollowec@bnl.gov, tony@bnl.gov

Solid state drives (SSDs) provide significant improvements in random I/O performance over traditional rotating SATA and SAS drives. While the cost of SSDs has been steadily declining over the past few years, high density SSDs continue to remain prohibitively expensive when compared to traditional drives. Currently, 1TB SSDs generally cost more than USD $1,000$, $while 1TB SAT Adrive stypically ret ail for u$ With ever-increasing x86_64 server CPU core counts, and therefore job slot counts, local scratch space density and random I/O performance have become even more important for HEP/NP applications.

FlashCache and Bcache are Linux kernel modules which implement caching of SATA/SAS hard drive data on SSDs, effectively allowing one to create hybrid SSD drives using software. In this presentation, we discuss our experience with FlashCache and Bcache, and the effects of this software on local scratch storage performance.

Facilities, Infrastructures, Networking and Collaborative Tools / **274**

ATLAS Cloud Computing R&D

Author: Sergey Panitkin[1]

Co-author: Fernando Harald Barreiro Megino [2]

[1] *Brookhaven National Laboratory (US)*

[2] *CERN*

Corresponding Authors: rsobie@uvic.ca, panitkin@bnl.gov, fernando.harald.barreiro.megino@cern.ch

The computing model of the ATLAS experiment was designed around the concept of grid computing and, since the start of data taking, this model has proven very successful. However, new cloud computing technologies bring attractive features to improve the operations and elasticity of scientific distributed computing. ATLAS sees grid and cloud computing as complementary technologies that will coexist at different levels of resource abstraction, and two years ago created an R&D working group to investigate the different integration scenarios. The ATLAS Cloud Computing R&D has been able to demonstrate the feasibility of offloading work from grid to cloud sites and, as of today, is able to integrate transparently various cloud resources into the PanDA workload management system. The ATLAS Cloud Computing R&D is operating various PanDA queues on private and public resources and has provided several hundred thousand CPU days to the experiment. The HammerCloud grid site testing framework is used to evaluate the performance of cloud resources and, where appropriate, compare it with the performance of bare metal to measure virtualization penalties. As a result, the ATLAS Cloud Computing R&D group has gained a deep insight into the cloud computing landscape and has identified points that still need to be addressed in order to fully profit from this young technology.

This contribution will explain the cloud integration models that are being evaluated and will discuss ATLAS' learning during the collaboration with leading commercial and academic cloud providers.

Poster presentations / 310

An integrated framework for the data quality assessment and database management for the ATLAS Tile Calorimeter

Authors: Carlos Solans Sanchez[1]; David Calvet[2]; Raffaela De Castro Cunha[3]

[1] *CERN*

[2] *Univ. Blaise Pascal Clermont-Fe. II (FR)*

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

Corresponding Authors: carlos.solans@cern.ch, calvet@in2p3.fr, raffaela.de.castro.cunha@cern.ch

The Tile calorimeter is one of the sub-detectors of ATLAS. In order to ensure its proper operation and assess the quality of data, many tasks are to be performed by means of many tools which were developed independently to satisfy different needs. Thus, these systems are commonly implemented without a global perspective of the detector and lack basic software features. Besides, in some cases they overlap in the objectives and resources with another one. It is therefore evident the necessity of an infrastructure to allow the implementation of any functionality without having to duplicate the effort while being possible to integrate with an overall view of the detector status.

Tile-in-ONE is intended to meet these needs, by providing a unique system, which integrates all the web systems and tools used by the Tile calorimeter, with a standard development technology and documentation. It also intends to abstract the user from knowing where and how to get the wanted data by providing a user friendly interface. It is based in a server containing a core, which represents the basic framework that loads the configuration, manages user settings and loads plugins at run-time; a set of services, which provide common features to be used by the plug-ins, such as connectors to different databases and resources; and the plug-ins themselves which provide features at the top level layer for the users. Moreover, a web environment is being designed to allow collaborators develop their own plug-ins, test them and add them to the system. To make it possible, an API is used allowing any kind of application to be interpreted and displayed in a standard way.

Poster presentations / 309

Computing challenges in the certification of ATLAS Tile Calorimeter front-end electronics during maintenance periods

Author: Carlos Solans Sanchez[1]

[1] *CERN*

Corresponding Author: carlos.solans@cern.ch

After two years of operation of the LHC, the ATLAS Tile Calorimeter is undergoing the consolidation process of its front-end electronics. The first layer of certification of the repairs is performed in the experimental area with a portable test-bench which is capable of controlling and reading out all the inputs and outputs of one front-end module through dedicated cables. This testbench has been redesigned to improve the quality assessment of the data until the end of Phase I. It
is now possible to identify low occurrence errors due to its increased read-out bandwidth and perform more sophisticated quality checks due to its enhanced computing power. Improved results provide fast and reliable feedback to the user.

Plenaries / **491**

Designing the computing for the future experiments

Author: Stefano Spataro[1]

[1] *University of Turin*

Corresponding Author: spataro@to.infn.it

For many experiments, e.g. those at the LHC, design choices made a very long time ago for the compute and trigger model are still used today. The incoming experiments have the opportunity to make new choices based on the current state of computing technology and novel ways to design the reconstruction frameworks, using the experience from previous experiments as well as already existing software packages developed outside the collaboration and used and by a larger community. This talk will show the computing decisions and the current developments of the Panda experiment at FAIR - Darmstadt (Germany), which will take data starting from 2018. One of the key features is a modular software framework with a dynamic data structure based on ROOT, in common to other experiments outside FAIR, with the possibility to run simulation and analysis on grid but open to different middleware technologies such as the cloud. Due to the high data rate, a special attention is also given to the online reconstruction of the continuous data stream coming from a trigger-less DAQ system, where the pre-processing for event selection is done online and concurrency is the key feature to achieve the requested high performances; all the efforts to develop a modular multi-core architecture, supporting also FPGA and GPU units, will be here presented together with the results obtained so far.

Poster presentations / **150**

A flexible monitoring infrastructure for the simulation requests

Author: Vincenzo Spinoso[1]

[1] *Universita e INFN (IT)*

Corresponding Author: vincenzo.spinoso@ba.infn.it

Running and monitoring simulations usually involves several different aspects of the entire workflow: the configuration of the job, the site issues, the software deployment at the site, the file catalogue, the transfers of the simulated data. In addition, the final product of the simulation is often the result of several sequential steps. This project tries a different approach to monitoring the simulation requests. All the necessary data are collected from the central services which lead the submission of the requests and the data management, and stored by a backend into a NoSQL-based data cache; those data can be queried through a Web Service interface, which returns JSON responses, and allows users, sites, physics groups to easily create their own web frontend, aggregating only the needed information. As an example, it will be shown how it is possible to monitor the CMS services (ReqMgr, DAS/DBS, PhEDEx) using a central backend and multiple customized cross-language frontends.

Poster presentations / **85**

Self managing experiment resources

Authors: Federico Stagni[1]; Mario Ubeda Garcia[1]

[1] *CERN*

Corresponding Authors: mario.ubeda.garcia@cern.ch, federico.stagni@cern.ch

Within this paper we present an autonomic Computing resources management system used by LHCb for assessing the status of their Grid resources. Virtual Organizations Grids include heterogeneous resources. For example, LHC experiments very often use resources not provided by WLCG and Cloud Computing resources will soon provide a non-negligible fraction of their computing power.
The lack of standards and procedures across experiments and sites generated the appearance of multiple information systems, monitoring tools, ticket portals, etc... which nowadays coexist and represent a very precious source of information for running HEP experiments Computing systems as well as sites.
These two facts lead to many particular solutions for a general problem: managing the experiment resources. In this paper we present how LHCb, via the DIRAC interware addressed such issues. With a renewed Central Information Schema hosting all resources metadata and a Status System (Resource Status System) delivering real time information, the system controls the resources topology, independently of the resource types. The Resource Status System applies data mining techniques against all possible information sources available and assesses the status changes, that are then propagated to the topology description. Obviously, giving full control to such an automated system is not risk-free. Therefore, in order to minimise the probability of misbehavior, a battery of tests has been provided in order to certify the correctness of its assessments.

We will demonstrate the performance and efficiency of such a system in terms of cost reduction and reliability.

Poster presentations / **273**

Popularity Prediction Tool for ATLAS Distributed Data Management

Author: Thomas Beermann[1]

Co-authors: Angelos Molfetas [2]; Armin Nairz [3]; Cedric Serfon [3]; Erich Schikuta [4]; Graeme Andrew Stewart [3]; Luc Goossens [3]; Mario Lassnig [3]; Martin Barisits [3]; Ralph Vigne [5]; Vincent Garonne [3]

[1] *Bergische Universitaet Wuppertal (DE)*

[2] *University of Sydney (AU)*

[3] *CERN*

[4] *University of Vienna*

[5] *University of Vienna (AT)*

Corresponding Authors: graeme.andrew.stewart@cern.ch, thomas.beermann@cern.ch, mario.lassnig@cern.ch, vincent.garonne@cern.ch, martin.barisits@cern.ch, ralph.vigne@cern.ch, cedric.serfon@cern.ch, luc.goossens@cern.ch, armin.nairz@cern.ch, angelos.molfetas@cern.ch

This paper describes a popularity prediction tool for data-intensive data management systems, such as the ATLAS distributed data management (DDM) system. The tool is fed by the DDM popularity system, which produces historical reports about ATLAS data usage and provides information about the files, datasets, users and sites where data was accessed. The tool described in this contribution uses this historic information to make a prediction about the future popularity of data. It finds trends in the usage of data using a set of neural networks and a set of input parameters and predicts the number of accesses in the near term future. This information can then be used in a second step to improve the distribution of replicas at sites, taking into account the cost of creating new replicas (bandwidth and load on the storage system) compared to the gain of having new ones (faster access of data for analysis). The tool ensures that the total amount of space available on the grid is not exceeded. This information can then help to make a decision about adding and also removing data from the grid to make a better use of the available resources. The design and architecture of the popularity prediction tool is described, examples of its use are shown and an evaluation of its performance is presented.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / 264

ATLAS Job Transforms: A Data Driven Workflow Engine

Author: Graeme Andrew Stewart[1]

Co-authors: Bjorn Sarrazin [2]; Harvey Jonathan Maddocks [3]; Marisa Sandhoff [4]; Torsten Harenberg [5]; William Dmitri Breaden Madden [6]

[1] *CERN*

[2] *Universitaet Bonn (DE)*

[3] *Lancaster University (GB)*

[4] *Bergische Universitaet Wuppertal (DE)*

[5] *UNIVERSITY OF WUPPERTAL*

[6] *University of Glasgow (GB)*

Corresponding Authors: graeme.andrew.stewart@cern.ch, harenberg@physik.uni-wuppertal.de, marisa.sandhoff@cern.ch, bjorn.sarrazin@cern.ch, harvey.jonathan.maddocks@cern.ch, william.dmitri.breaden.madden@cern.ch

The need to run complex workflows for a high energy physics experiment such as ATLAS has always been present. However, as computing resources have become even more constrained, compared to the wealth of data generated by the LHC, the need to use resources efficiently and manage complex workflows within a single grid job have increased.

In ATLAS, a new Job Transform framework has been developed that we describe in this paper. This framework manages the multiple execution steps needed to 'transform' one data type into another (e.g., RAW data to ESD to AOD to final ntuple) and also provides a consistent interface for the ATLAS production system.

The new framework uses a data driven workflow definition which is both easy to manage and powerful. After a transform is defined, jobs are expressed simply by specifying the input data and the desired output data. The transform infrastructure then executes only the necessary substeps to produce the final data products. The global execution cost of running the job is minimised and the transform can adapt to scenarios where data can be produced along different execution paths. Transforms for specific physics tasks which support over 60 individual substeps have been successfully run.

As the new transforms infrastructure has been deployed in production many features have been added to the framework which improve reliability, quality of error reporting and also provide support for multi-threaded and multi-process jobs.

Poster presentations / **412**

Implementation of the twisted mass fermion operator on accelerators

Author: Alexei Strelchenko[1]

[1] *FNAL*

Corresponding Author: astrel@fnal.gov

Lattice Quantum Chromodynamics (LQCD) simulations are critical for understanding the validity of the Standard Model and the results of the High-Energy and Nuclear Physics experiments. Major improvements in the calculation and prediction of physical observables, such as nucleon form factors or flavor singlet meson mass, require large amounts of computer resources, of the order of hundreds of Tflop/s of sustained performance. For the first part of our study, we extended the QUDA library, an open source library for performing calculations in LQCD on NVIDIA GPUs, to include kernels for the non-degenerate twisted mass fermion operator. Next, we implemented the operator on the Intel MIC architecture. A detailed performance analysis for both cases is provided.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / **16**

BESIII physical analysis on hadoop platform

Author: Gongxing Sun[1]

Co-authors: Dongsong Zang [2]; Jing Huo [3]; xiaofeng lei [3]

[1] *INSTITUE OF HIGH ENERGY PHYSICS*

[2] *Chinese Academy of Sciences (CN)*

[3] *IHEP*

Corresponding Authors: sungx@mail.ihep.ac.cn, donal.zang@cern.ch

This paper brings the idea of MapReduce parallel processing to BESIII physical analysis, gives a new data analysis system structure based on HADOOP framework; Optimizes the process of data processing, by establish an event level metadata(TAG) database and do event pre-selection based on TAGs, significantly reduce the number of events that need to do further analysis by 2-3 classes, which reduces the I/O volume and improves the efficiency of data analysis jobs; The event storage structure in DST files are re-organized to optimize the selective reading pattens with event pre-selection. Designs the MapReduce models for TAG generation, TAG based event pre-selection and event analysis, and develop proper MapReduce libs that fit for the ROOT framework to do things such as data splitting, event fetching and result merging. An 8-nodes cluster is used for system test, the testing result shows that the new system shortens the data analyzing time by 80%, and the cluster system shows great scalability when adding more worker nodes.

Facilities, Infrastructures, Networking and Collaborative Tools / **397**

Beyond core count: a look at new mainstream computing platforms for HEP workloads

Authors: Andrzej Nowak[1]; Liviu Valsan[2]; Pawel Szostek[2]; Sverre Jarp[2]

[1] *CERN openlab*

[2] *CERN*

Corresponding Authors: pawel.szostek@cern.ch, andrzej.nowak@cern.ch

As Moore's Law continues to deliver more and more transistors, the mainstream processor industry is preparing to expand its investments in areas other than simple core count. These new interests include deep integration of on-chip components, advanced vector units, memory, cache and interconnect technologies. We examine these moving trends with parallelized and vectorized High Energy Physics workloads in mind. In particular, we report on practical experience resulting from experiments with scalable HEP benchmarks on the Intel "Ivy Bridge-EP"and "Haswell"processor families. In addition, we examine the benefits of the new "Haswell"microarchitecture and its impact on multiple facets of HEP software. Finally, we report on the power efficiency of new systems.

Poster presentations / 116

XRootd, disk-based, caching-proxy for optimization of data-access, data-placement and data-replication

Author: Matevz Tadel[1]

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

Corresponding Author: matevz.tadel@cern.ch

Following the smashing success of XRootd-based USCMS data-federation, AAA project investigated extensions of the federation architecture by developing two sample implementations of an XRootd, disk-based, caching-proxy. The first one simply starts fetching a whole file as soon as a file-open request is received and is suitable when completely random file access is expected or it is already known that a whole file be read. The second implementation supports on-demand downloading of partial files. Extensions to the Hadoop file-system have been developed to allow for an immediate fallback to network access when local HDFS storage fails to provide the requested block. Tools needed to analyze and to tweak block replication factors and to inject downloaded blocks into a running HDFS installation have also been developed. Both cache implementations are in operation at UCSD and several tests were also performed at UNL and UW-M. Operational experience and applications to automatic storage healing and opportunistic computing, especially on T3 sites and campus resources, will be discussed.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / 91

Experience of a low-maintenance distributed data management system

Authors: Adil Hasan[1]; Francesca Di Lodovico[2]; Wataru Takase[3]; Yoshimi Matsumoto[3]

Co-authors: Takashi Sasaki [4]; Yoshiyuki Watase [4]

[1] *University of Liverpool, UK*

[2] *Queen Mary College, University of London, UK*

[3] *High Energy Accelerator Research Organization (KEK), Japan*

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

Corresponding Authors: wataru.takase@kek.jp, f.di.lodovico@qmul.ac.uk, yoshimi.matsumoto@kek.jp, adilhasan2@gmail.com

In this paper we report on the setup, deployment and operation of a low-maintenance, policy-driven distributed data management system for scientific data based on the integrated Rule Oriented Data System (iRODS). The system is located at KEK, Tsukuba, Japan with a satellite system at QMUL, London, UK. The system has been running stably in production for more than two years with minimal management overhead. We also report on the experience gained and lessons learnt during the setup and operation.
The management tools that were developed to support the production system are also described. In addition we describe a simple XOR-based approach to file replication that reduces the amount of storage space consumed. In situations of large data volumes this approach can be of great benefit.

Poster presentations / **52**

The ALICE DAQ infoLogger

Author: Sylvain Chapeland[1]

Co-authors: Adriana Telesca [1]; Alexandru Grigore [2]; Barthelemy Von Haller [1]; Charles Delort [3]; Costin Ionita [1]; Csaba Soos [1]; Ervin Denes [4]; Filippo Costa [1]; Franco Carena [1]; Giuseppe Simonetti [5]; Pierre Vande Vyvre [1]; Roberto Divia [1]; Ulrich Fuchs [1]; Vasco Chibante Barroso [1]; Wisla Carena [1]

[1] *CERN*

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

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

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

[5] *Universita e INFN (IT)*

Corresponding Authors: adriana.telesca@cern.ch, sylvain.chapeland@cern.ch

ALICE (A Large Ion Collider Experiment) is a heavy-ion detector studying the physics of strongly interacting matter and the quark-gluon plasma at the CERN LHC (Large Hadron Collider). The ALICE DAQ (Data Acquisition System) is based on a large farm of commodity hardware consisting of more than 600 devices (Linux PCs, storage, network switches). The DAQ reads the data transferred from the detectors through 500 dedicated optical links at an aggregated and sustained rate of up to 10 Gigabytes per second and stores at up to 2.5 Gigabytes per second.

The infoLogger is the log system which collects centrally the messages issued by the thousands of processes running on the DAQ machines. It allows to report errors on the fly, and to keep a trace of runtime execution for later investigation.

More than 500000 messages are stored every day in a MySQL database, in a structured table keeping track for each message of 16 indexing fields (e.g. time, host, user, ...). The total amount of logs for 2012 exceeds 75GB of data and 150 million rows.

We present in this paper the architecture and implementation of this distributed logging system, consisting of a client programming API, local data collector processes, a central server, and interactive human interfaces. We review the operational experience during the 2012 run, in particular the actions taken to ensure shifters receive manageable and relevant content from the main log stream. Finally, we present the data mining and reporting tools developed to analyze and make best use of this vast amount of information.

Facilities, Infrastructures, Networking and Collaborative Tools / **13**

System performance monitoring of the ALICE Data Acquisition System with Zabbix

Author: Adriana Telesca[1]

Co-authors: Alexandru Grigore [2]; Barthelemy Von Haller [1]; Charles Delort [3]; Costin Ionita [1]; Csaba Soos [1]; Ervin Denes [4]; Filippo Costa [1]; Franco Carena [1]; Giuseppe Simonetti [5]; Pierre Vande Vyvre [1]; Roberto Divia [1]; Sylvain Chapeland [1]; Ulrich Fuchs [1]; Vasco Chibante Barroso [1]; Wisla Carena [1]

[1] *CERN*

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

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

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

[5] *Universita e INFN (IT)*

Corresponding Author: adriana.telesca@cern.ch

ALICE (A Large Ion Collider Experiment) is a heavy-ion detector studying the physics of strongly interacting matter and the quark-gluon plasma at the CERN LHC (Large Hadron Collider). The ALICE Data-AcQuisition (DAQ) system handles the data flow from the sub-detector electronics to the permanent data storage in the CERN computing center. The DAQ farm consists of about 1000 devices of many different types ranging from direct accessible machines to storage arrays and custom optical links. The system performance monitoring tool used during the LHC run 1 will be replaced by a new tool for run 2.

This presentation shows the results of an evaluation that has been conducted on six existing and publicly available monitoring tools. The evaluation has been carried out by taking into account selection criteria such as scalability, flexibility, reliability as well as data collection methods and display. All the tools have been prototyped and evaluated according to those criteria.

We will describe the considerations that have brought to the selection of the Zabbix monitoring tool for the DAQ farm. The results of the tests conducted in the ALICE DAQ laboratory will be presented.

In addition, the deployment of the software on the DAQ machines in terms of metrics collected and data collection methods will be described. We will illustrate how remote nodes are monitored with Zabbix by using SNMP-based agents and how DAQ specific metrics are retrieved and displayed. We will also show how the monitoring information is accessed and made available via the graphical user interface and how Zabbix communicates with the other DAQ online systems for notification and reporting.

Poster presentations / 118

A Scalable Infrastructure for CMS Data Analysis Based on Openstack Cloud and Gluster File System

Authors: John White White[1]; Lirim Osmani[2]; Oscar Kraemer[3]; Paula Eerola[4]; Salman Toor[1]; Sasu Tarkoma[2]; Tomas Lindén[5]

[1] *Helsinki Institute of Physics (FI)*

[2] *University of Helsinki*

[3] *Helsinki Institute of Physics (HIP),*

[4] *Helsinki Institute of Physics (HIP)*

[5] *HELSINKI INSTITUTE OF PHYSICS*

Corresponding Author: salman.toor@helsinki.fi

The challenge of providing a resilient and scalable computational and data management solution for massive scale research environments, such as the CERN HEP analyses, requires continuous exploration of new technologies and techniques. In this article we present a hybrid solution of an open source cloud with a network file system for CMS data analysis. Our aim has been to design a scalable and resilient infrastructure for CERN HEP data analysis. The infrastructure is based on Openstack components for structuring a private cloud together with the Gluster filesystem. The Openstack cloud platform is one of the fastest growing solutions in the cloud world. The Openstack components provide solutions for computational resource (NOVA), instance (GLANCE), network (QUANTUM) and security (Keystone) management all underneath an API layer that supports global applications, web clients and large ecosystems. NOVA and GLANCE manage Virtual Machines (VMs) and image repository respectively. QUANTUM provides network as a service (NAAS) and advanced network management capabilities. The virtual network layer provided by QUANTUM supports seamless migration of the VMs while preserving the network configuration. One important component that is currently not part of the Openstack suite is a network file system. To overcome this limitation we have used GlusterFS, a network-based file system for high availability. GlusterFS uses FUSE and may be scale to petabytes. In our experiments, 1TB is used for instance management and 2TB for the data related to CMS jobs within GlusterFS. We integrate the above state-of-the-art cloud technologies with the traditional Grid middleware infrastructure. This approach implies no changes for the end-user, while the production infrastructure is enriched by the high-end resilient and scalable components. To achieve this, we have run Advance Resource Connector (ARC) as a meta-scheduler. Both Computing Elements (CE) and Worker Nodes (WN) are running on VM instances inside the Openstack cloud. Currently we consider our approach as semi-static, as the instance management is manual yet provides scalability and performance. In near future we are aiming for a comprehensive elastic solution by including the EMI authorization service (Argus) and the Execution Environment Service (Argus-EES). In order to evaluate the strength of the infrastructure, four test cases have been selected for experimentation and analysis. (i) The first test case is based on instance performance, the boot time of customized images using different hypervisors and the performance of multiple instances with different configurations. (ii) The second test case focuses on I/O-related performance analysis based on GlusterFS. This test also presents the performance of instances running on GlusterFS compared with the local file system. (iii) The third test case examines system stability with live migration of VM instances based on QUANTUM. (iv) In the fourth test we will present long-term system performance both at the level of VMs running CMS jobs and physical hosts running VMs. Our test results show that the adopted approach provides a scalable and resilient solution for managing resources without compromising on performance and high availability.

Data Acquisition, Trigger and Controls / 211

The CMS High Level Trigger

Author: Daniele Trocino[1]

[1] *Northeastern University (US)*

Corresponding Author: daniele.trocino@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 to cope with the increasing LHC pile-up, reducing its impact on the physics performance.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / 379

DIRAC Distributed Computing Services

Author: Andrei Tsaregorodtsev[1]

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

Corresponding Author: atsareg@in2p3.fr

DIRAC is a framework for building general purpose distributed computing systems. It was developed originally for the LHCb HEP experiment at CERN and now it is used in several other HEP and astrophysics experiments as well as for user communities in other scientific domains.

There is a large interest from smaller user communities to have a simple to use tool for accessing grid and other types of distributed computing resources like DIRAC. However, small experiments can not afford to install and maintain dedicated community services.
Therefore, several grid infrastructure projects are providing DIRAC services for their respective user communities. The DIRAC services are used for the user tutorials as well as to help porting the applications to the grid for a practical day-to-day work. The services are giving access typically to several grid infrastructures as well as to standalone computing clusters accessible by the target user communities.

In the paper we will present the experience of running DIRAC services provided by the France-Grilles NGI and other national grid infrastructure projects. Specific features needed to support multiple user communities by a single instance of the DIRAC service will be discussed. The outlook for provisioning of the new types of computing resources, like clouds or desktop grids, will be given.

Poster presentations / 467

Sustainable software and the Xenon 1 T high-level trigger

Author: Christopher Tunnell[None]

Corresponding Author: ctunnell@nikhef.nl

In the coming years, Xenon 1 T, a ten-fold expansion of Xenon 100, will further explore the dark matter WIMP parameter space and must be able to cope with correspondingly higher data rates. With a focus on sustainable software architecture, and a unique experimental scale compared to collider experiments, a high-level trigger system is being designed for the next many years of Xenon 1 T running. As a high level trigger, processing time must be minimized while maintaining efficient event selection. By using Python as a 'glue' language around quicker C routines, the complexity of the system can be minimized and testing facilitated. After discussing the data flow and parallelization, the talk will focus on testing techniques and software sustainability. For example, in addition to unit tests, in situ (i.e., control room) and ex situ (i.e., cloud based) system tests use old data to demonstrate the reliability of and benchmark the system. Strategies appropriate to this experimental scale that ensure sustainability will also be discussed, including the documentation and training policies, testing requirements, style guides, and general project management. In this way, the high level trigger system will be maintainable, robust, and reliable for the foreseen lifetime of the experiment.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / 269

Automating usability of ATLAS Distributed Computing resources

Author: Salvatore Tupputi[1]

Co-authors: Alessandro Di Girolamo [2]; Jaroslava Schovancova [3]; Tomas Kouba [4]

[1] *Universita e INFN (IT)*

[2] *CERN*

[3] *Brookhaven National Laboratory (US)*

[4] *Acad. of Sciences of the Czech Rep. (CZ)*

Corresponding Authors: salvatore.tupputi@cern.ch, alessandro.di.girolamo@cern.ch, tomas.kouba@cern.ch, jaroslava.schovancova@cern.ch

The automation of ATLAS Distributed Computing (ADC) operations is essential to reduce manpower costs and allow performance-enhancing actions which improve the reliability of the system. In this perspective a crucial case is the automatic exclusion/recovery of ATLAS computing sites storage resources, which are continuously exploited at the edge of their capabilities.

It is challenging to adopt unambiguous decision criteria for storage resources which feature non-homogeneous types, sizes and roles. The recently developed Storage Area Automatic Blacklisting (SAAB) tool has provided a suitable solution, by employing an inference algorithm which processes SAM (Site Availability Test) site-by-site SRM test outcomes. SAAB accomplishes both the tasks of providing global monitoring as well as automatic operations on single sites.

The implementation of the SAAB tool has been the first step in a comprehensive review of the storage areas monitoring and central management at all levels. Such review has involved the reordering and optimization of SAM tests deployment and the inclusion of SAAB results in the ATLAS Site Status Board with both dedicated metrics and views. The final structure allows monitoring the storage resources status with fine time-granularity and automatic actions to be taken in foreseen cases, like automatic exclusion/recovery and notifications to sites. Hence, the human actions are restricted to ticket tracking and exchanging, where and when needed.

In this work we show SAAB working principles and features. We present also the decrease of human interactions achieved within the ATLAS Computing Operations team. The automation results in a prompt reaction to failures, which grants the optimization of resource exploitation.

Distributed Processing and Data Handling A: Infrastructure, Sites, and Virtualization / **31**

Integration of Cloud resources in the LHCb Distributed Computing

Authors: Mario Ubeda Garcia[1]; Victor Mendez Munoz[2]

Co-authors: Adrian Casajus Ramo [3]; Andrei Tsaregorodtsev [4]; Federico Stagni [5]; Joel Closier [1]; Philippe Charpentier [1]; Ricardo Graciani Diaz [3]; Stefan Roiser [1]; Victor Manuel Fernandez Albor [6]

[1] *CERN*

[2] *PIC*

[3] *University of Barcelona (ES)*

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

[5] *INFN Ferrara*

[6] *Universidade de Santiago de Compostela (ES)*

Corresponding Authors: vmendez@pic.es, mario.ubeda.garcia@cern.ch

This contribution describes how Cloud resources have been integrated in the LHCb Distributed Computing. LHCb is using Dirac and its LHCb-specific extension LHCbDirac as an interware for its Distributed Computing. So far it was seamlessly integrating Grid resources and Computer clusters. The cloud extension of Dirac (VMDIRAC) extends it to the integration of Cloud computing infrastructures.

Several computing resource providers in the eScience environment are planning to deploy IaaS in production during 2013. VMDIRAC is able to interface to multiple types of infrastructures in commercial and institutional clouds, supported by multiple interfaces (Amazon EC2, OpenNebula, OpenStack and CloudStack). It instantiates, monitors and manages Virtual Machines running on this

aggregation of Cloud resources. These VMs then create an overlay of the computing resources in the same way as pilot jobs do on the Grid: jobs submitted to the LHCbDirac infrastructure can be executed seamlessly either on standard Grid resources or on Cloud resources.

This work addresses the specifications for institutional Cloud resources proposed by HEPIX and WLCG. The WLCG Cloud approach defines an instance framework on service level basis similar to the spot instances of commercial clouds. VMDIRAC is in particular able to deal with the agreed constraints on the VM lifetime: based on default limits defined by the resources provider as well as short notice on demand requests for shutdown. It also allows to instantiate VMs running on multiple cores, under control of the VO. In a first instance, the WLCG scenario considers the static assignment of cloud slots to each VO, which is good enough for a starting point of WLCG cloud deployment, but the VMDIRAC implementation also makes provision for more dynamic, job-driven, VM management.

We describe the solution implemented by LHCb and VMDIRAC for the contextualisation of the VMs, and how job agents are instantiated on these VMs. We report on operational experience of using in production several institutional Cloud resources that are thus becoming integral part of the Distributed Computing resources used by LHCb. We present a comparison of the performance of those Cloud resources with Grid traditional resources, in particular for what concerns data access and memory footprint. An outlook is also given on optimizing the memory footprint of VMs running on multiple cores by using parallel processing applications based on GaudiMP.

Summaries / **522**

Summary of track 2 (Event Processing, Simulation and Analysis)

Corresponding Author: f.uhlig@gsi.de

Poster presentations / **247**

ATLAS Nightly Build System Upgrade

Author: Alexander Undrus[1]

Co-authors: Brinick Simmons [2]; Emil Obreshkov [3]; Gancho Dimitrov [4]

[1] *Brookhaven National Laboratory (US)*

[2] *Department of Physics and Astronomy - University College London*

[3] *University of Innsbruck (AT)*

[4] *CERN*

Corresponding Authors: undrus@bnl.gov, gancho.dimitrov@cern.ch, emil.obreshkov@cern.ch, brinick.simmons@gmail.com

The ATLAS Nightly Build System is a facility for automatic production of software releases. Being the major component of ATLAS software infrastructure, it supports more than 50 multi-platform branches of nightly releases and provides vast opportunities for testing new packages, for verifying patches to existing software, and for migrating to new platforms and compilers. The Nightly System testing framework runs several hundred integration tests of different granularity and purposes. The nightly releases are distributed and validated, and some are transformed into stable releases used for data processing worldwide. The first LHC long shutdown (2013-2015) activities will elicit increased load on the Nightly System as additional releases and builds are needed to exploit new programming techniques, languages, and profiling tools. This talk describes the program of the ATLAS Nightly Build System Long Shutdown upgrade. It brings modern database and web technologies into the Nightly System, improves monitoring of nightly build results, provides new tools for offline release shifters. We will also outline our long term plans for distributed nightly releases builds and testing.

Poster presentations / **253**

ATLAS Experience with HEP Software at the Argonne Leadership Computing Facility

Author: Thomas Le Compte[1]

Co-authors: Doug Benjamin [2]; Tom Uram [3]; Venkat Vishwanath [3]

[1] *Argonne National Laboratory (US)*

[2] *Duke University (US)*

[3] *ANL*

Corresponding Authors: lecompte@anl.gov, douglas.benjamin@cern.ch

A number of HEP software packages used by the ATLAS experiment, including GEANT4, ROOT and ALPGEN, have been adapted to run on the IBM Blue Gene supercomputers at the Argonne Leadership Computing Facility. These computers use a non-x86 architecture and have a considerably less rich operating environment than in common use in HEP, but also represent a computing capacity an order of magnitude beyond what ATLAS is presently using via the LCG. The status and potential for making use of leadership-class computing, including the status of integration with the ATLAS production system, is discussed.

Event Processing, Simulation and Analysis / **419**

The Belle II Physics Analysis Model

Author: Phillip Urquijo[1]

[1] *Universitaet Bonn (DE)*

Corresponding Author: phillip.urquijo@cern.ch

The Belle II experiment is a future flavour factory experiment at the intensity frontier SuperKEKB e+e- collider, KEK Japan. Belle II is expected to go online in 2015, and collect a total of 50 ab-1 of data by 2022. The data will be used to study rare flavour phenomena in the decays of B- and D- mesons and tau-leptons, as well as heavy meson spectroscopy. Owing to the record breaking luminosity of SuperKEKB, L=8x10ˆ35 cm-2s-1, and to the large number of detector channels, particularly in the silicon tracking detector (4 strip + 2 pixel layers) the data output rate is expected to rival that of the LHC experiments.

In this presentation we will present the Belle II physics analysis model, designed to cope with both precision measurements, and searches for rare processes on this prodigious quantity of data. We will discuss how the new framework is being constructed to robustly use Grid computing, with common analysis tools, common data preparation tools, and a steering mechanism for automated reconstruction of particle cascades. At the core of the framework is a newly developed root based physics analysis data model, a layer for particle reconstruction and persistency. It will facilitate centralised computation of multitudes of particle decay channels for analysis groups.

A key aspect for particle reconstruction at SuperKEKB will be the use of recoil techniques, which rely on precisely known initial beam energies, to study decays with neutrinos and multiple neutral particles in the final state. We will discuss how we are integrating various recoil techniques into the design of the analysis framework, and the particle persistency layer. We also discuss integration of other complex analysis tools, such as continuum suppression and tag vertexing for CP violation analyses. The analysis data model, and its optimisation for Grid facilities will be discussed. Finally the reconstruction capabilities of the framework will be shown, including Monte Carlo studies for the analysis of physics benchmark channels.

Poster presentations / **117**

CORAL and COOL during the LHC long shutdown

Author: Andrea Valassi[1]

Co-authors: Andrey Salnikov [2]; Dave Dykstra [3]; Marco Clemencic [1]; Martin Wache [4]; Neha Goyal [5]; Raffaello Trentadue [6]

[1] *CERN*

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

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

[4] *Institut fur Physik-Johannes-Gutenberg-Universitaet-Unknown*

[5] *Mody Institute of Technology and Science (IN)*

[6] *Universita e INFN (IT)*

Corresponding Author: andrea.valassi@cern.ch

CORAL and COOL are two software packages that are widely used by the LHC experiments for the management of conditions data and other types of data using relational database technologies. They have been developed and maintained within the LCG Persistency Framework, a common project of the CERN IT department with ATLAS, CMS and LHCb. The project used to include the POOL software package, providing a hybrid store for C++ objects, which is now maintained by ATLAS. This presentation will report on the status of CORAL and COOL at the time of CHEP2013. It will cover the new features in CORAL and COOL (such as Kerberos authentication to Oracle and the prototyping of CORAL server monitoring), as well as the changes and improvements in the software process infrastructure (better integration with various tools, new code repository, new platforms). It will also review the usage of the software in the experiments and the outlook for the project during the LHC long shutdown (LS1) and beyond.

Plenaries / **488**

Probing Big Data for Answers using Data about Data

Author: Edwin Valentijn[1]

[1] *Kapteijn Institute University of Groningen*

Large amounts of data are now streaming daily from large astronomical survey telescopes, such as Lofar and the new generation of wide field imagers at ESO's Paranal Observatory, but also from DNA scanners, text scanners etc etc. In the future the volumes will only increase with ESA's Euclid all sky deep imaging survey mission and SKA. Prof Dr Edwin A. Valentijn of the Kapteyn Institute will talk how data handling systems tend to focus more and more on enabling users to find their way in Big Data by means of datacentric modelling approaches, such as employed in the AstroWise information system and its WISE derivatives used outside astronomy.

Poster presentations / **297**

GPU Enhancement of the High Level Trigger to extend the Physics Reach at the LHC

Author: Halyo Valerie[None]

Corresponding Author: valerie.halyo@cern.ch

Significant new challenges are continuously confronting the High Energy Physics (HEP) experiments in particular the Large Hadron Collider (LHC) at CERN who does not only drive forward theoretical, experimental and detector physics but also pushes to limits computing. LHC delivers proton-proton collisions to the detectors at a rate of 40 MHz. This rate must be significantly reduced to comply with the performance limitations of the mass storage hardware, and the capabilities of the computing resources to process the collected data in a timely fashion for physics analysis. At the same time, the physics signals of interest must be retained with high efficiency.

The quest for rare new physics phenomena at the LHC and the flexibility of the HLT allows us to propose a GPU enhancement of the conventional computer farm that provides faster and more efficient events selection at the trigger level.The proposed enhancement is made possible due to rising hybrid systems consisting of processors with multiple cores integrated with Graphics processing units (GPU) as a modified form of stream processor

A new tracking algorithm will be executed on the hybrid computer farm that will permit optimal use of the tracker information to reconstruct not only all the prompt tracks but also tracks not originating from the interaction point. One of its benefit, It will allow for the first time to reconstruct tracks originating from very long lived particles in the tracker system in the trigger system, hence extending the search for new physics at the LHC. Preliminary results will be presented comparing the algorithm performance between Nvidia K20 and Intel xeon-phi chip.

Poster presentations / **69**

Toward a petabyte-scale AFS service at CERN

Authors: Arne Wiebalck[1]; Daniel van der Ster[1]; Jakub Moscicki[1]

[1] *CERN*

Corresponding Authors: daniel.vanderster@cern.ch, jakub.moscicki@cern.ch

AFS is a mature and reliable storage service at CERN, having worked for more than 20 years as the provider of Linux home directories and application areas. Recently, our AFS service has been growing at unprecedented rates (300% in the past year), thanks to innovations in both the hardware and software components of our file servers.

This work will present how AFS is used at CERN and how the service offering is evolving with the increasing storage needs of its local and remote user communities. In particular, we will demonstrate the usage patterns for home directories, workspaces and project spaces, as well as show the daily work which is required to rebalance data and maintaining stability and performance. Finally, we will highlight some recent changes and optimisations made to the AFS Service, thereby revealing how AFS can possibly operate at all while being subjected to frequent —almost DDOS-like —attacks from its users.

Data Stores, Data Bases, and Storage Systems / **68**

Building an organic block storage service at CERN with Ceph

Authors: Arne Wiebalck[1]; Daniel van der Ster[1]

[1] *CERN*

Corresponding Author: daniel.vanderster@cern.ch

Emerging storage requirements, such as the need for block storage for both OpenStack VMs and file services like AFS and NFS, have motivated the development of a generic backend storage service for

CERN IT. The goals for such a service include (a) vendor neutrality, (b) horizontal scalability with commodity hardware, (c) fault tolerance at the disk, host, and network levels, and (d) support for geo-replication. Ceph is an attractive option due to its native block device layer RBD which is built upon its scalable, reliable, and performant object storage system, RADOS. It can be considered an "organic" storage solution because of its ability to balance and heal itself while living on an ever-changing set of heterogeneous disk servers.

This work will present the outcome of a petabyte-scale test deployment of Ceph by CERN IT. We will first present the architecture and configuration of our cluster, including a summary of best practices learned from the community and discovered internally. Next the results of various functionality and performance tests will be shown: the cluster has been used as a backend block storage system for AFS and NFS servers as well as a many-thousand node OpenStack cluster at CERN. Finally, we will discuss the next steps and future possibilities for Ceph at CERN.

Software Engineering, Parallelism & Multi-Core / 180

Experiences with moving to open source standards for building and packaging

Author: Dennis Van Dok[1]

Co-authors: Mischa Salle [1]; Oscar Arthur Koeroo [1]

[1] *Nikhef (NL)*

Corresponding Author: dennisvd@nikhef.nl

The LCMAPS family of grid middleware has improved in the last years by moving from a custom build system to open source community standards for building, packaging and distributing. This contribution outlines what improvements were made and the benefits they rendered.

LCMAPS, gLExec and related middleware components were developed under a series of European framework program projects, starting in 2001 (Datagrid) and lasting more than a decade (EGEE-I, -II, and -III, with EMI and IGE ending in 2013).

Many interdependent components were made by these projects; as building and packaging of the ever growing collection of software became more complex, so did the build system that was used to manage it. Eventually ETICS was developed which became the central build system for all of EGEE-III and EMI.

Building outside ETICS proved hard, which raised concerns about long term sustainability and adoption outside the original community, for instance by OSG.

In the past years we upgraded the software we maintained to become independent of any build system, using common patterns in popular open source software as our template. The result is that any package can be downloaded in tarball form, configured to find dependencies on the local system, and build to install in a custom location. But our software is also packaged according to the rigid guidelines of the Fedora and Debian projects, and easily deployed on systems running Debian, Ubuntu, Red Hat (with EPEL) or Fedora.

The required adaptations were not all trivial, but all were worthwhile in retrospect. They helped to uncover several bugs that would have gone unnoticed and offered additional unanticipated benefits.

The adaptations cover the areas of version control, adoption of standards (especially POSIX and ANSI C), backward and forward compatibility, dependency resolution, use of GNU autotools, developing build and release procedures, packaging for EPEL and Debian, automated builds with Koji, and repository generation.

Some of the added benefits are improved portability, fewer bugs, increased serviceability, improved sustainability, a closer match with common open source skill sets, easier configuration, easier installation, easier packaging, shorter release cycles, and inclusion in mainstream distributions.

In this presentation we share our experiences concerning this transition.

Poster presentations / 248

Next-Generation Navigational Infrastructure and the ATLAS Event Store

Author: Peter Van Gemmeren[1]

Co-authors: David Malon [1]; Marcin Nowak [2]

[1] *Argonne National Laboratory (US)*

[2] *Brookhaven National Laboratory (US)*

Corresponding Authors: peter.van.gemmeren@cern.ch, malon@anl.gov, marcin.nowak@cern.ch

The ATLAS event store employs a persistence framework with extensive navigational capabilities. These include real-time back navigation to upstream processing stages, externalizable data object references, navigation from any data object to any other both within a single file and across files, and more. The 2013-2014 shutdown of the Large Hadron Collider provides an opportunity to enhance this infrastructure in several ways that both extend these capabilities and allow the collaboration to better exploit emerging computing platforms. Enhancements include redesign with efficient file merging in mind, content-based indices in optimized reference types, and support for forward references. The latter provide the potential to construct valid references to data before those data are written, a capability that is useful in a variety of multithreading, multiprocessing, distributed processing, and deferred processing scenarios.

This paper describes the architecture and design of the next generation of ATLAS navigational infrastructure.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / 425

A modern web based data catalog for data access and analysis

Authors: Brian Van Klaveren[1]; Tony Johnson[1]

[1] *SLAC*

Corresponding Authors: bvan@slac.stanford.edu, tonyj@slac.stanford.edu

The SLAC Computing Applications group (SCA) has developed a general purpose data catalog framework, initially for use by the Fermi Gamma-Ray Space Telescope, and now in use by several other experiments. The main features of the data catalog system are:

- Ability to organize datasets in a virtual hierarchy without regard to physical location or access protocol
- Ability to catalog datasets stored at multiple locations and with multiple versions
- Ability to attach arbitrary meta-data to datasets and folders
- Web based and command line interfaces for registering, viewing and searching datasets
- A data “crawler” to verify catalog integrity and automate meta-data extraction

- A download manager for reliable download of collections of files

In this paper we will describe a recent project to update the data catalog to current web standards, in particular to:

- Isolate the database back-end from the server-side middle-ware by use of a file abstraction layer
- Develop Restful interfaces to make the server side functionality accessible to many tools and languages
- Develop a modern HTML5 based web client which also communicates with the server using Restful interfaces, and provides dynamic functionality such as drag and drop file upload/download.

These improvement open the way to integrating components of the data catalog with different back-end systems, and to provide a portal to support not only access to data, but to be able to operate on and analyze data remotely.

Data Acquisition, Trigger and Controls / 389

O2: a new combined online and offline > computing for ALICE after 2018

Authors: Pierre Vande Vyvre[1]; Predrag Buncic[1]; Thorsten Kollegger[2]

Co-authors: . Ananya [3]; Aditi Udupa [4]; Adriana Telesca [1]; Alexandru Grigore [5]; Alina Gabriela Grigoras [1]; Andreas Morsch [1]; Andrei Gheata [1]; Anju Bhasin [6]; Anshul AVASTHI [4]; Barthelemy Von Haller [1]; Basanta Kumar Nandi [7]; Camilo Lara [8]; Charles Delort [9]; Chiara Zampolli [10]; Costin Grigoras [1]; Costin Ionita [1]; David Michael Rohr [2]; Dominic Eschweiler [2]; Ervin Denes [11]; Falco Francesco Vennedey [2]; Filippo Costa [1]; Franco Carena [1]; Gergely Barnafoldi [12]; Giuseppe Simonetti [10]; Gyorgy Rubin [11]; Heiko Engel [2]; Iosif Legrand [13]; Ivan Kisel [14]; Ivana Hrivnacova [15]; Jan Marian Pluta [16]; Jouri Belikov [17]; Latchezar Betev [1]; Lukasz Kamil Graczykowski [18]; Maciej Pawel Szymanski [18]; Malgorzata Anna Janik [18]; Marian Ivanov [14]; Matthias Jakob Bach [2]; Matthias Kretz [2]; Mihaela Gheata [19]; Mirko Planinic [20]; Nikola Poljak [21]; Raffaele Grosso [22]; Rishabh KHANDELWAL [4]; Roberto Divia [1]; Ruben Shahoyan [1]; Sadhana Dash [7]; Sebastian Kalcher [23]; Sergey Gorbunov [2]; Sylvain Chapeland [1]; Timo Gunther Breitner [2]; Tivadar Kiss [11]; Udo Wolfgang Kebschull [2]; Ulrich Fuchs [1]; Vasco Chibante Barroso [1]; Vincent Claude Lafage [15]; Volker Lindenstruth [2]

[1] *CERN*

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

[3] *IIT - Indian Institute of Technology*

[4] *IIT- Indian Institute of Technology*

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

[6] *University of Jammu (IN)*

[7] *IIT- Indian Institute of Technology (IN)*

[8] *University of Heidelberg*

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

[10] *Universita e INFN (IT)*

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

[12] *Wigner RCP Hungarian Academy of Sciences (HU)*

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

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

[15] *Universite de Paris-Sud 11 (FR)*

[16] *Faculty of Physics-Warsaw University of Technology*

[17] *Institut Pluridisciplinaire Hubert Curien (FR)*

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

[19] *ISS - Institute of Space Science (RO)*

[20] *University of Zagreb (HR)*

[21] *Institute Rudjer Boskovic (HR)*

[22] *University of Houston (US)*

[23] *FIAS - Institute for Computer Science-Johann-Wolfgang-Goethe Uni*

Corresponding Author: pierre.vande.vyvre@cern.ch

for the ALICE O2 Collaboration

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

After the second long shutdown of the LHC, the ALICE detector will be upgraded in order to make high precision measurements of rare probes at low pT, which cannot be selected with a trigger, and therefore require a large sample of events recorded on tape. The online computing system will be entirely redesigned to address the major challenge of sampling the full 50 kHz Pb-Pb interaction rate increasing by a factor 100 the present limit. This upgrade will also include the continuous un-triggered read-out of two detectors (ITS and TPC) producing a sustained throughput of 1 TB/s.

This unprecedented data volume will be reduced by an entirely new strategy including the online calibration and reconstruction which will result in storing only the reconstruction results and discarding the raw data. This system, demonstrated in production on the TPC data since 2011, will have to be optimized for the online usage of reconstruction algorithms. It implies much tighter coupling between online and offline computing systems.

We present in this paper the R&D program put in place to address this huge challenge and the first results of this program.

Poster presentations / **142**

Reconstruction of the Higgs mass in H -> tautau Events by Dynamical Likelihood techniques

Author: Christian Veelken[1]

[1] *Ecole Polytechnique (FR)*

Corresponding Author: christian.veelken@cern.ch

An algorithm for reconstruction of the Higgs mass in $H \rightarrow \tau\tau$ decays is presented. The algorithm computes for each event a likelihood function $P(M_{\tau\tau})$ which quantifies the level of compatibility of a Higgs mass hypothesis $M_{\tau\tau}$, given the measured momenta of visible tau decay products plus missing transverse energy reconstructed in the event. The algorithm is used in the CMS $H \rightarrow \tau\tau$ analysis. It is found to improve the sensitivity for the Standard Model Higgs boson in this decay channel by about 30%.

Event Processing, Simulation and Analysis / **454**

RooFit and RooStats - a framework for advanced data modeling and statistical analysis

Author: Wouter Verkerke[1]

Co-authors: Gena Kukartsev [2]; Giovanni Petrucciani [3]; Gregory Alfred Schott [4]; Kyle Stuart Cranmer [5]; Lorenzo Moneta [6]; Sven Kreiss [5]

[1] *NIKHEF (NL)*

[2] *Brown University (US)*

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

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

[5] *New York University (US)*

[6] *CERN*

Corresponding Author: verkerke@nikhef.nl

RooFit is a library of C++ classes that facilitate data modeling in the ROOT environment. Mathematical concepts such as variables, (probability density) functions and integrals are represented as C++ objects. The package provides a flexible framework for building complex fit models through classes that mimic math operators. For all constructed models RooFit provides a concise yet powerful interface for fitting, plotting and toy Monte Carlo generation as well as sophisticated tools to manage large scale projects.
RooFit has been used in countless published B-factory and LHC results. We will review recent developments such as the ability to persist models in ROOT files in container classes, which enables the concept of digital publishing of analytical likelihood functions with an arbitrary number of parameters. Persistent models enable completely ways to perform physics analyses: Complex fit results can be trivially shared between groups and experiments for validation and detailed combinations of physics results, such as the combination of Higgs decay channels, can be constructed in a matter of hours. Finally, model persistence simplifies the streaming of tasks to other hosts to parallelize calculation of computing intensive problems that are common in statistical techniques.

RooStats is a project providing advanced statistical tools required for the analysis of LHC data, with emphasis on discoveries, confidence intervals, and combined measurements in the both the Bayesian and Frequentist approaches. The tools are built on top of the RooFit data modeling language and core ROOT mathematics libraries and persistence technology. These tools have been developed in collaboration with the LHC experiments and used by them to produce numerous physics results, the discovery of the the Higgs boson by ATLAS and CMS Higgs, using models with more than 1000 parameters. We will review new developments which have been included in RooStats and the performance optimizations, required to cope with such complex models used by the LHC experiments. We will show as well the parallelization capability of these statistical tools using multiple-processors via PROOF.

Distributed Processing and Data Handling B: Experiment Data Processing, Data Handling and Computing Models / 256

PROOF-based analysis on the ATLAS Grid facilities: first experience with the PoD/PanDa plugin

Author: Elisabetta Vilucchi[1]

Co-authors: Alessandra Doria [2]; Alessandro De Salvo [3]; Anar Manafov [4]; Antonio Salvucci [5]; Arturo Sanchez Pineda [6]; Camilla Di Donato [2]; David Rebatto [7]; Francesco Prelz [7]; Gerardo Ganis [8]; Giada Mancini [9]; Roberto Di Nardo [10]; Simone Michele Mazza [7]

[1] *Laboratori Nazionali di Frascati (LNF) - Istituto Nazionale di F*

[2] *Universita e INFN (IT)*

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

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

[5] *INFN-Laboratori Nazionali di Frascati, Frascati (RM)*

[6] *Universita di Napoli Federico II-Universita e INFN*

[7] *Università degli Studi e INFN Milano (IT)*

[8] *CERN*

[9] *Istituto Nazionale Fisica Nucleare - Laboratori Nazionali di Frascati (IT) and Università di Roma "Tor Vergata"*

[10] *Istituto Nazionale Fisica Nucleare - Laboratori Nazionali di Frascati (IT)*

Corresponding Authors: elisabetta.vilucchi@cern.ch, arturo.sanchez.pineda@cern.ch

In the ATLAS computing model Grid resources are managed by the PanDA system, a data-driven workload management system designed for production and distributed analysis. 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. The ROOT way to analyze data provide users the possibility to use PROOF to exploit the computing power of multi-core machines or to dynamically manage analysis facilities. Since analysis facilities are, in general, not dedicated to PROOF only, PROOF-on-Demand (PoD) is used to enable PROOF on top of an existing resource management system.

In a previous work we investigated the usage of PoD to enable PROOF-based analysis on Tier-2 facilities using the PoD/gLite plug-in interface. Our study focused in particular on the startup time and on the aggregate read-out rate, showing promising results for both cases.

In this paper we present the status of our investigations using the recently developed PoD/PanDA plug-in to enable PROOF, and a real end-user ATLAS physics analysis as payload. For this work, data were accessed using two different protocols: XRootD and file protocol, the former in the site where the SRM interface is Disk Pool Manager (DPM) and the latter where the SRM interface is StoRM with GPFS file system. Using PanDA also gives the possibility to test more realistic scenarios, where users belong to different groups and roles and are in real competition for the resources.

We will first describe the configuration and setup details and the results of some benchmark tests we run on the Italian Tier-2 sites and at CERN. Then, we will compare the results of different types of analysis, comparing performances accessing data in relation to different types of SRM interfaces and accessing data with XRootD in the LAN and in the WAN using the ATLAS storage federation infrastructure.

Finally we will discuss the behavior of the system in the presence of concurrent users.

Poster presentations / 152

MCM : The Evolution of PREP. The CMS tool for Monte-Carlo Request Management.

Author: Jean-Roch Vlimant[1]

[1] *CERN*

Corresponding Author: jean-roch.vlimant@cern.ch

The analysis of the LHC data at the CMS experiment requires the production of a large number of simulated events. In 2012, CMS has produced over 4 Billion simulated events in about 100 thousands of datasets. Over the past years a tool (PREP) has been developed for managing such a production of thousands of samples.

A lot of experience working with this tool has been gained, and conclusions on its limitations have been drawn. For better interfacing with the CMS production infrastructure and data book-keeping system, new database technology (couchDB) has been adopted. More recent server infrastructure technology (cherrypy + java) has been set as the new platform for an evolution of PREP. The operational limitations encountered over the years of usage have been solved in the new system. The aggregation of the production information of samples has been much improved for a better traceability and prioritization of work.

This contribution will cover the description of the functionalities of this major evolution of the software for managing samples of simulated events for CMS.

Poster presentations / **42**

The ALICE Data Quality Monitoring: qualitative and quantitative review of 3 years of operations

Author: Barthelemy Von Haller[1]

Co-authors: Adriana Telesca [1]; Francesca Bellini [2]; Yiota Foka [3]

[1] *CERN*

[2] *Universita e INFN (IT)*

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

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

ALICE (A Large Ion Collider Experiment) is a detector designed to study the physics of strongly interacting matter and the quark-gluon plasma produced in heavy-ion collisions at the CERN Large Hadron Collider (LHC). Due to the complexity of ALICE in terms of number of detectors and performance requirements, Data Quality Monitoring (DQM) plays an essential role in providing an online feedback on the data being recorded. It intends to provide shifters with precise and complete information to quickly identify and overcome problems, and as a consequence to ensure acquisition of high quality data.

This paper presents a review of the DQM operations during the first three years of data taking from a quantitative and qualitative point of view. We will start by presenting the DQM software and tools before moving on to the various analyses carried out and the underlying physics. An overview of the produced monitoring quantities will be given, presenting the diversity of usage and the flexibility of the DQM system.

Well-prepared shifters and experts, in addition to a precise organisation, were required to ensure smooth and successful operations. The description of the measures taken to ensure both aspects and an account of the DQM shifter's job are followed by a summary of the evolution of the system. We will then give a quantitative review of the final setup of the system used during the whole year 2012. We conclude the paper with real world cases when the DQM proved to be very valuable, scalable and efficient and with the plans for the coming years.

Data Stores, Data Bases, and Storage Systems / **265**

Data Federation Strategies for ATLAS Using XRootD

Author: Robert William Gardner Jr[1]

[1] *University of Chicago (US)*

Corresponding Authors: ilija.vukotic@cern.ch, robert.w.gardner@cern.ch

In the past year the ATLAS Collaboration has accelerated its program to federate data storage resources using an architecture based on XRootD with its attendant redirection and storage integration services. The main goal of the federation is an improvement in the data access experience for the end user while allowing for more efficient and intelligent use of computing resources by monitoring and optimizing for observed data access patterns. Along with these advances come integration with existing ATLAS production services (PanDA and its pilot services) and data management services (DQ2, and in the next generation, Rucio). A system which tests functionality of the federation has been

integrated into the standard ATLAS and WLCG monitoring frameworks and a dedicated set of tools provides high granularity information on its current and historical usage. We use a federation topology designed to search from the site's local storage outward to its region and then more globally. We describe programmatic testing of various federation access modes including direct access over the wide area network or staging in of remote data files to local disk. To support job brokering decisions, a time-dependent cost-of-data-access matrix is made taking into account network performance and other key site performance indicators. The system's response to production-scale physics analysis workloads, either from individual end-users or ATLAS analysis services, is discussed.

Poster presentations / **328**

Exploring virtualization tools with a new virtualization provisioning method to test dynamic grid environments for ALICE grid jobs over ARC grid middleware

Authors: Bjarte Kileng[1]; Boris Wagner[2]

[1] *Bergen University College (NO)*

[2] *University of Bergen (NO) for the ALICE Collaboration*

Corresponding Author: boris.wagner@cern.ch

The Nordic Tier-1 for the LHC is distributed over several, sometimes smaller, computing centers. In order to minimize administration effort, we are interested in running different grid jobs over one common grid middleware. ARC is selected as the internal middleware in the Nordic Tier-1. The AliEn grid middleware, used by ALICE has a different design philosophy than ARC. In order to use most of the AliEn infrastructure and available software deployment methods for running ALICE grid jobs on ARC, we are investigating different possible virtualization technologies.

Software deployment is a critical a part in grid computing. Ideally a contributing computing center doesn't need to change their infrastructure or install additional software, except at the grid entry point. One solution to achieve this is a cloud infrastructure that provides (dynamically) virtual machines with the required software environments. It seems that no easy solution exists to run grid jobs on top of those cloud providers. They are missing tools that submit a job into a virtual machine, that may be created dynamically and collect the results afterwards. Therefore a specialized, virtualization provisioning method was developed. This method could be developed further into a system that is used by smaller sites, which don't want to run a more general and more complex private cloud solution.

The new provisioning system is used to test different combinations of virtualization backends and virtual machines.

The CernVM project is developing a virtual machine that can provide a common analysis environment for all LHC experiments. One of our interests is to investigate the use of CernVM as a base setup for a dynamical grid environment capable of running grid jobs. For this, performance comparisons between different virtualization technologies have been conducted.

CernVM needs an existing virtualization infrastructure, which is not always existing or wanted at some computing sites. To increase the possible application of dynamical grid environments to those sites, we describe several possibilities that are less invasive and have less specific Linux distribution requirements, at the cost of lower performance.

Different tools like user-mode Linux (UML), micro Linux distributions, a new software packaging project by Stanford university (CDE) and CernVM are under investigation for their invasiveness, distribution requirements and performance. Comparisons between the different methods with solutions that are closer to the hardware will be presented.

Poster presentations / 383

Optimising network transfers to and from QMUL, a large WLCG Tier-2 Grid site

Author: Christopher John Walker[1]

Co-authors: Daniel Peter Traynor [1]; Duncan Rand [2]; Steve Lloyd [1]

[1] *University of London (GB)*

[2] *Imperial College*

Corresponding Author: christopher.john.walker@cern.ch

The WLCG, and high energy physics in general, relies on remote Tier-2 sites to analyse the large quantities of data produced. Transferring this data in a timely manner requires significant tuning to make optimum usage of expensive WAN links.

In this paper we describe the techniques we have used at QMUL to optimise network transfers. Use of the FTS with settings and appropriate TCP window sizes allowed us to saturate a 1 Gbit link for 24 hours - whilst still achieving acceptable download speeds. Source based routing and multiple gridftp servers allowed us to use an otherwise unused “resilient” link.

After the WAN link was upgraded to 10Gbit/s, a significant reduction in transfer rates was observed from some sites - due to suboptimal routing resulting in packet loss on congested links. Solving this dramatically improved performance.

The use of jumbo frames (MTU=9000) offers potential improvements in performance, particularly for latency limited links. Whilst much of the Internet backbone is capable of supporting this, most sites are not, and path MTU discovery fails at some sites. We describe our experience with this.

Poster presentations / 189

A novel dynamic event data model using the Drillbit column store

Authors: Johannes Ebke[1]; Peter Waller[2]

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

[2] *University of Liverpool (GB)*

Corresponding Author: peter.waller@gmail.com

The focus in many software architectures of the LHC experiments is to deliver a well-designed Event Data Model (EDM). Changes and additions to the stored data are often very expensive, requiring large amounts of CPU time, disk storage and man-power. At the ATLAS experiment, such a reprocessing has only been undertaken once for data taken in 2012.

However, analysts have to develop and apply corrections and do computations after the final official data processing or re-processing has taken place. The current practice at ATLAS is to distribute software tools to apply these corrections at analysis runtime, requiring any necessary input data to be present and manually supplied to the tool, and requiring manual application of the tool output to the event data by each analyst.

This approach has proven to be very expensive in terms of man-power, especially since verifying that the tools have been applied correctly (or at least consistently) is very time consuming.

Drillbit enables dynamic addition of event data, stored in and read from external files. This would make it possible to forego a fixed EDM class structure and instead collate validated variables and objects in a dynamically defined event. Corrections could be computed once or twice by experts, versioned, and then made available to others directly.

The technical basis for this architecture is currently being prototyped, and initial versions of the underlying Drillbit column store are available.

Poster presentations / **438**

Leveraging HPC resources for High Energy Physics

Author: Andrew John Washbrook[1]

Co-author: Rodney Walker [2]

[1] *University of Edinburgh (GB)*

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

Corresponding Author: andrew.washbrook@cern.ch

High Performance Computing (HPC) provides unprecedented computing power for a diverse range of scientific applications. As of November 2012, over 20 supercomputers deliver petaflop peak performance with the expectation of "exascale" technologies available in the next 5 years. Despite the sizeable computing resources on offer there are a number of technical barriers that limit the use of HPC resources for High Energy Physics applications. HPC facilities have traditionally opted for specialised hardware architectures and favoured tightly coupled parallel MPI-based workloads rather than the high throughput commodity computing model typically used in HEP.

However, more recent HPC facilities use x86-based architectures managed by Linux-based operating systems which could potentially allow unmodified HEP software to be run on supercomputers. There is now renewed interest from both the LHC experiments and the HPC community to accommodate data analysis and event simulation production on HPC facilities either from a dedicated resource share or from opportunistic use during low utilisation periods. If existing job scheduling and execution frameworks used by the LHC experiments could be successfully adapted for HPC use it would significantly increase the total amount of computing resources available.

A feasibility study into the use of LHC software in an HPC environment will be presented. The HECToR supercomputer in the UK and the SuperMUC supercomputer in Germany will be used as demonstrators. The challenges faced from the perspective of software execution and the interaction with existing LHC distributed computing environments will be highlighted using software typical in data analysis workflows. In particular, we will discuss how tighter restrictions on HPC worker node access could limit the functionality of existing middleware and how this can be potentially adapted for external job submission and more efficient job scheduling.

Poster presentations / **431**

Many-core on the Grid: From Exploration to Production

Author: Andrew John Washbrook[1]

Co-authors: John Walsh [2]; Liliana Salvador [3]; Matthew Doidge [4]; Thomas Doherty [3]; Tiziana Ferrari [5]

[1] *University of Edinburgh (GB)*

[2] *Unknown*

[3] *University of Glasgow*

[4] *Lancaster University*

[5] *INFN CNAF*

Corresponding Author: andrew.washbrook@cern.ch

A number of High Energy Physics experiments have successfully run feasibility studies to demonstrate that many-core devices such as GPGPUs can be used to accelerate algorithms for trigger systems and data analysis. After this exploration phase experiments on the Large Hadron Collider are now investigating how these devices can be incorporated into key areas of their software framework in advance of the significant increase in data volume expected in the next phase of LHC operations.

A recent survey performed by the EGI GPGPU Resource Centre indicates that there is increasing community interest for GPUs to be made available on the existing grid infrastructure. However, despite this anticipated usage there is no standard method available to run GPU-based applications in distributed computing environments. Before GPU resources are available on the grid operational issues such as job scheduling and resource discovery will have to be addressed. For example, software developed for many-core devices is often optimised for a particular device and therefore specific architecture details - such as GPU shared memory capacity and thread block size - will be required for job resource matching.

The key technical challenges for grid-enabling many-core devices will be discussed. Consideration will be given to how jobs can be scheduled to maximise device occupancy, how they can be advertised using methods familiar to the grid user-community and how their usage can be accurately evaluated for accounting purposes. A selection of GPU devices will be made available at two Tier-2 Grid sites in the UK to demonstrate resource usage. Functionality will be evaluated using software from existing many-core feasibility studies in addition to examples provided by the EPIC project who will use grid-enabled GPUs as a critical part of their workflow.

Poster presentations / **338**

The STAR ”Plug and Play” Event Generator Framework

Authors: Jason Webb[1]; Jerome LAURET[2]; John Novak[3]; Victor Perevoztchikov[1]

[1] *Brookhaven National Lab*

[2] *BROOKHAVEN NATIONAL LABORATORY*

[3] *Michigan State University*

Corresponding Authors: jwebb@bnl.gov, jlauret@bnl.gov

The STAR experiment pursues a broad range of physics topics in pp, pA and AA collisions produced by the Relativistic Heavy Ion Collider (RHIC). Such a diverse experimental program demands a simulation framework capable of supporting an equally diverse set of event generators, and a flexible event record capable of storing the (common) particle-wise and (varied) event-wise information provided by the external generators. With planning underway for the next round of upgrades to exploit ep and eA collisions from the electron-ion collider (or eRHIC), these demands on the simulation infrastructure will only increase and requires a versatile framework.
STAR has developed a new event-generator framework based on the best practices in the community (a survey of existing approach had been made and the “best of all worlds”kept in mind in our design). It provides a common set of base classes which establish the interface between event generators and the simulation and handles most of the bookkeeping associated with a simulation run. This streamlines the process of integrating and configuring an event generator within our software chain. Developers implement two classes: the interface for their event generator, and their event record. They only need to loop over all particles in their event and push them out into the event record. The framework is responsible for vertex assignment, stacking the particles out for simulation, and event persistency. Events from multiple generators can be merged together seamlessly,

with an event record which is capable of tracing each particle back to its parent generator.
We present our work and approach in detail and illustrate its usefulness by providing examples of event generators implemented within the STAR framework covering for very diverse physics topics. We will also discuss support for event filtering, allowing users to prune the event record of particles which are outside of our acceptance, and/or abort events prior to the more computationally expensive digitization and reconstruction phases. Event filtering has been supported in the previous framework and showed to save enormous amount of resources –the approach within the new framework is a generalization of filtering.

Poster presentations / **337**

The Abstract geometry Modeling Language (AgML): Experience and Road map toward eRHIC

Authors: Jason Webb[1]; Jerome LAURET[2]; Victor Perevoztchikov[1]

[1] *Brookhaven National Lab*

[2] *BROOKHAVEN NATIONAL LABORATORY*

Corresponding Authors: jwebb@bnl.gov, jlauret@bnl.gov

The STAR experiment has adopted an Abstract Geometry Modeling Language (AgML) as the primary description of our geometry model. AgML establishes a level of abstraction, decoupling the definition of the detector from the software libraries used to create the concrete geometry model. Thus, AgML allows us to support both our legacy GEANT3 simulation application and our ROOT/TGeo based reconstruction software from a single source, which is demonstrably self-consistent. While AgML was developed primarily as a tool to migrate away from our legacy FORtran-era geometry codes, it also provides a rich syntax geared towards the rapid development of detector models. AgML has been successfully employed by users to quickly develop and integrate the descriptions of several new detectors in the RHIC/STAR experiment including the Forward GEM Tracker (FGT) and Heavy Flavor Tracker (HFT) upgrades installed in STAR for the 2012 and 2013 runs. AgML has furthermore been heavily utilized to study future upgrades to the STAR detector as it prepares for the eRHIC era. With its track record of practical use in a live experiment in mind, we present the status, lessons learned and future of the AgML language as well as our experience in bringing the code into our production and development environments. We will discuss the path toward eRHIC and pushing the current model to accommodate for detector miss-alignment and high precision physics.

Poster presentations / **282**

Negative improvements

Authors: Gabriela Hoff[1]; Georg Weidenspointner[2]; Maria Grazia Pia[3]; Matej Batic[4]

[1] *Pontifícia Universidade do Rio Grande do Sul, Porto Alegre, Brazil*

[2] *MPE Garching*

[3] *Universita e INFN (IT)*

[4] *Jozef Stefan Institute*

Corresponding Authors: georg.weidenspointner@hll.mpg.de, maria.grazia.pia@cern.ch

An extensively documented, quantitative study of software evolution resulting in deterioration of physical accuracy over the years is presented. The analysis concerns the energy deposited by electrons in various materials produced by Geant4 versions released between 2007 and 2013.

The evolution of the functional quality of the software is objectively quantified by means of a rigorous statistical analysis, which combines goodness-of-fit tests and methods of categorical data testing

to validate the simulation against high precision experimental measurements. Significantly lower compatibility with experiment is observed with the later Geant4 versions subject to evaluation; the significance level of the test is 0.01.

Various issues related to the complexity of appraising the evolution of software functional quality are considered, such as the dependence on the experimental environment where the software operates and its sensitivity to detector characteristics.

Methods and techniques to mitigate the risk of "negative improvements"are critically discussed: they concern various disciplines of the software development process, including not only testing and quality assurance, but also domain decomposition, software design and change management. Concrete prototype solutions are presented.

This study is intended to provide a constructive contribution to identify possible causes of the deterioration of software functionality, and means to address them effectively. It is especially relevant to the HEP software environment, where widely used tools and experiments' software are expected to stand long life-cycles and are necessarily subject to evolution.

Poster presentations / **226**

Accessing opportunistic resources with Bosco

Author: Derek John Weitzel[1]

Co-authors: Brian Paul Bockelman [1]; Dan Fraser [2]; David Swanson [3]; Frank Wuerthwein [4]; Igor Sfiligoi [4]; Jaime Frey [5]

[1] *University of Nebraska (US)*

[2] *Argonne National Laboratory*

[3] *University of Nebraska - Lincoln*

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

[5] *University of Wisconsin - Madison*

Corresponding Author: derek.weitzel@cern.ch

Bosco is a software project developed by the Open Science Grid to help scientists better utilize their on-campus computing resources. Instead of submitting jobs through a dedicated gatekeeper, as most remote submission mechanisms use, it uses the built-in SSH protocol to gain access to the cluster. By using a common access method, SSH, we are able to simplify the interaction with the cluster, making the submission process more user friendly. Additionally, it does not add any extra software to be installed on the cluster making Bosco an attractive option for the cluster administrator.

In this paper, we will describe Bosco, the personal supercomputing assistant, and how Bosco is used by researchers across the U.S. to manage their computing workflows. In addition, we will also talk about how researchers are using it, including an unique use of Bosco to submit CMS reconstruction jobs to an opportunistic XSEDE resource.

Poster presentations / **221**

Grid Accounting Service: State and Future Development

Author: Tanya Levshina[1]

Co-authors: Ashu Guru [2]; Brian Paul Bockelman [3]; Chander Sehgal [1]; Derek John Weitzel [3]

[1] *FERMILAB*

[2] *University of Nebraska, Lincoln*

[3] *University of Nebraska (US)*

Corresponding Authors: derek.weitzel@cern.ch, tlevshin@fnal.gov

During the last decade, large-scale federated distributed infrastructures have continually developed and expanded. One of the crucial components of a cyber-infrastructure is an accounting service that collects data related to resource utilization and identity of users using resources. The accounting service is important for verifying pledged resource allocation per particular groups and users, providing reports for funding agencies and resource providers, and understanding hardware provisioning requirements. It can also be used for end-to-end troubleshooting as well as billing purposes. In this work we describe Gratia, a federated accounting service jointly developed at Fermilab and University of Nebraska Holland Computing Center (HCC). It has been used in production by the Open Science Grid, Fermilab, HCC, and several other institutions for several years. The current development activities include Virtual Machines provisioning information, XSEDE allocation usage accounting, and Campus Grids resource utilization.
We also identify the directions of future work: improvement and expansion of Cloud accounting, persistent and elastic storage space allocation, and the incorporation of WAN and LAN network metric.

Plenaries / **492**

Computing for the LHC: the next step up

Author: Torre Wenaus[1]

[1] *Brookhaven National Laboratory (US)*

Corresponding Author: wenaus@gmail.com

The computing for the LHC experiments has resulted in spectacular physics during the first few years of running. Now, the long shutdown offers the possibility to re-think some of the underlying concepts, look back to the lessons learned from this first run, and at the same work on revised models for the next after LS1. Dr Torre Wenaus of Brookhaven National Lab will talk about the revisions made during LS1, and what the impact might be of these changes on the next run, and what this could means for the future at a high-luminosity LHC.

Event Processing, Simulation and Analysis / **415**

Simulation and analysis of the LUCID experiment in the Low Earth Orbit radiation environment

Author: Tom Whyntie[1]

Co-author: The LUCID Collaboration [2]

[1] *Queen Mary, University of London/The Langton Star Centre*

[2] *The Langton Star Centre*

The Langton Ultimate Cosmic ray Intensity Detector (LUCID) experiment [1] is a satellite-based device that uses five Timepix hybrid silicon pixel detectors [2] to make measurements of the radiation environment at an altitude of approximately 660km, i.e. in Low Earth Orbit (LEO). The experiment is due to launch aboard Surrey Satellite Technology Limited's (SSTL's) TechDemoSat-1 in Q3 of 2013. The Timepix detectors, developed by the Medipix Collaboration [3], are arranged to form the five sides of a cube enclosed by a 0.7 mm thick aluminium covering, and will be operated in Time-over-Threshold mode to allow the flux, energy and directionality of incident ionising radiation to be

measured. To understand the expected detector performance with respect to these measurements, the LUCID experiment has been modelled using the Allpix package, a generic simulation toolkit for silicon pixel detectors built upon the GEANT4 framework [4]. Furthermore, the anticipated data rates for differing space radiation environments (for example, during polar passes or when passing through the South Atlantic Anomaly) have been estimated. The UK's GridPP infrastructure was used to run the simulations and store the resultant datasets; a web portal was also developed to allow members of the LUCID Collaboration to easily specify the space radiation environment of interest, request the necessary simulation jobs and retrieve the results for local analysis. The results obtained have been used to confirm that the LUCID's data transmission allowance is sufficient, and also to validate the data transmission protocols that will be used when LUCID starts transmitting data towards the end of 2013.

Keywords:
LUCID
Pixel detector
Silicon detector
Space radiation
Timepix
GEANT4
Grid computing
GridPP

References:
[1] L. Pinsky et al., Radiation Measurements 46 (2011) 1610-1614
[2] X. Llopart et al., Nucl. Instr. Meth. A 581 (2007) 485-494
[3] http://medipix.web.cern.ch/
[4] A. Agostinelli et al., Nucl. Instr. Meth. A 506 (2003) 250-303

Poster presentations / **325**

Data Preservation activities at DESY (The DESY-DPHEP Group)

Author: David South[1]

Co-author: katarzyna wichmann [1]

[1] *DESY*

Corresponding Authors: katarzyna.wichmann@desy.de, david.south@cern.ch

The data preservation project at DESY was established in 2008, shortly after data taking ended at the HERA ep collider, soon after coming under the umbrella of the DPHEP global initiative. All experiments are implementing data preservation schemes to allow long term analysis of their data, in cooperation with the DESY-IT division. These novel schemes include software validation and migration techniques as well as archival storage of the data themselves. In addition to the preservation of data and software, a consolidation programme of all digital and non-digital documentation, some of which dates back to the 1980s, is being performed, including projects with the INSPIRE initiative. The activities of the group and their relevance and portability to other HEP experiments will be presented.

Event Processing, Simulation and Analysis / **373**

A new Scheme for ATLAS Trigger Simulation using Legacy Code

Author: Gorm Galster[1]

Co-authors: Joerg Stelzer [2]; Werner Wiedenmann [3]

[1] *University of Copenhagen (DK)*

[2] *Michigan State University (US)*

[3] *University of Wisconsin (US)*

Corresponding Author: werner.wiedenmann@cern.ch

An accurate simulation of the trigger response is necessary for high quality data analyses. This poses a challenge. For event generation and simulated data reconstruction the latest software is used to be in best agreement with the reconstructed data. Contrary the trigger response simulation needs to be in agreement with when the data was taken. The approach we follow is to use trigger software and conditions data that matches the simulated data-taking period - potentially dating many years back. Having a strategy for running old software in a modern environment thus becomes essential when data simulated for past years start to present a sizable fraction of the total.
We examined the requirements and possibilities for such a simulation scheme within and beyond the existing ATLAS software framework and successfully implemented a proof-of-concept simulation chain. One of the greatest challenges has been that of bridging old and new file formats, as most of the file formats and data representations used by ATLAS are changing with time. Over the time periods envisaged data format incompatibilities are likely to emerge in databases and other external storage services as well. Software availability is an issue. The support for the underlying operating system might stop. In this
talk we will present the encountered problems and developed solutions, and will discuss proposals for future development. These ideas will have reach beyond the retrospective trigger simulation scheme at ATLAS as they are applicable in other areas of data preservation.

Data Stores, Data Bases, and Storage Systems / **103**

The CMS Data Management System

Authors: Nicolo Magini[1]; Tony Wildish[2]

[1] *CERN*

[2] *Princeton University (US)*

Corresponding Authors: tony.wildish@cern.ch, nicolo.magini@cern.ch

The data management elements in CMS are scalable, modular, and designed to work together. The main components are PhEDEx, the data transfer and location system; the Dataset Booking System (DBS), a metadata catalogue; and the Data Aggregation Service (DAS), designed to aggregate views and provide them to users and services. Tens of thousands of samples have been cataloged and petabytes of data have been moved since the run began. The modular system has allowed the optimal use of appropriate underlying technologies. In this presentation we will discuss the use of both Oracle and nonSQL databases to implement the data management elements as well as the individual architectures chosen. We will discuss how the data management system functioned during the first run, and what improvements are planned in preparation for 2015.

Poster presentations / **128**

CMS Space Monitoring

Author: Natalia Ratnikova[1]

Co-author: Tony Wildish [2]

[1] *Fermilab*

[2] *Princeton University (US)*

Corresponding Authors: tony.wildish@cern.ch, natasha@fnal.gov

During the first LHC run, CMS saturated one hundred petabytes of storage resources with data. Storage accounting and monitoring help to meet the challenges of storage management, such as efficient space utilization, fair share between users and groups, and further resource planning. We present newly developed CMS space monitoring system based on the storage dumps produced at the sites. Storage contents information is aggregated and uploaded to the central database. Web based data service is provided to retrieve the information for a given time interval and a range of sites, so it can be further aggregated and presented in the desired format. The system has been designed based on the analysis of CMS monitoring requirements and experiences of the other LHC experiments. In this paper, we demonstrate how the existing software components of the CMS data placement system PhEDEx have been re-used, reducing dramatically the development effort.

Facilities, Infrastructures, Networking and Collaborative Tools / **92**

Challenging data and workload management in CMS Computing with network-aware systems

Authors: Daniele Bonacorsi[1]; Tony Wildish[2]

[1] *University of Bologna*

[2] *Princeton University (US)*

Corresponding Authors: tony.wildish@cern.ch, daniele.bonacorsi@bo.infn.it

After a successful first run at the LHC, and during the Long Shutdown (LS1) of the accelerator, the workload and data management sectors of the CMS Computing Model are entering into an operational review phase in order to concretely assess area of possible improvements and paths to exploit new promising technology trends. In particular, since the preparation activities for the LHC start, the Networks have constantly been of paramount importance for the execution of CMS workflows, exceeding the original expectations - as from the MONARC model - in terms of performance, stability and reliability. The low-latency transfers of PetaBytes of CMS data among dozens of WLCG Tiers worldwide using the PhEDEx dataset replication system is an example of the importance of reliable Networks. Another example is the exploitation of WAN data access over data federations in CMS. A new emerging area of work is the exploitation of "Intelligent Network Services", including also bandwidth on demand concepts. In this paper, we will review the work done in CMS on this, and the next steps.

Poster presentations / **131**

Request for All - Generalized Request Framework for PhEDEx

Author: Tony Wildish[1]

[1] *Princeton University (US)*

Corresponding Author: tony.wildish@cern.ch

PhEDEx has been serving CMS community since 2004 as the data broker. Every PhEDEx operation is initiated by a request, such as request to move data, request to delete data, and so on. A request has it own life cycle, including creation, approval, notification, and book keeping and the details depend on its type. Currently, only two kinds of requests, transfer and deletion, are fully integrated

in PhEDEx. They are tailored specifically to the operations' workflows. To be able to serve a new type of request it generally means a fair amount of development work.

After several years of operation, we have gathered enough experience to rethink the request handling in PhEDEx. "Generalized Request Project" is set to abstract such experience and come up with a request system which is not tied into current workflow yet it is general enough to accommodate current and future requests.

The challenges are dealing with different stages in a request's life cycle, complexity of approval process and complexity of the ability and authority associated with each role in the context of the request.

We start with a high level abstraction driven by a deterministic finite automata, followed by a formal description and handling of approval process, followed by a set of tools that make such system friendly to the users. As long as we have a formal way to describe the life of a request and a mechanism to systematically handle it, to server a new kind of request is merely a configuration issue, adding the description of the new request rather than development effort.

In this paper, we share the design and implementation of a generalized request framework and the experience of taking an existing serving system through a re-design and re-deployment.

Poster presentations / **122**

Integration and validation testing for PhEDEx, DBS and DAS with the PhEDEx LifeCycle agent

Author: Tony Wildish[1]

[1] *Princeton University (US)*

Corresponding Author: tony.wildish@cern.ch

The ever-increasing amount of data handled by the CMS dataflow and workflow management tools poses new challenges for cross-validation among different systems within CMS experiment at LHC. To approach this problem we developed an integration test suite based on the LifeCycle agent, a tool originally conceived for stress-testing new releases of PhEDEx, the CMS data-placement tool. The LifeCycle agent provides a framework for customising the test workflow in arbitrary ways, and can scale to levels of activity well beyond those seen in normal running. This means we can run realistic performance tests at scales not likely to be seen by the experiment for some years, or with custom topologies to examine particular situations that may cause concern some time in the future.

The LifeCycle agent has recently been enhanced to become a general purpose integration and validation testing tool for major CMS services (PhEDEx, DBS, DAS). It allows cross-system integration tests of all three components to be performed in controlled environments, without interfering with production services.

In this paper we discuss the design and implementation of the LifeCycle agent. We describe how it is used for small-scale debugging and validation tests, and how we extend that to large-scale tests of whole groups of sub-systems. We show how the LifeCycle agent can emulate the action of operators, physicists, or software agents external to the system under test, and how it can be scaled to large and complex systems.

Poster presentations / **123**

Re-designing the PhEDEx security model

Author: Tony Wildish[1]

[1] *Princeton University (US)*

Corresponding Author: tony.wildish@cern.ch

PhEDEx. the data-placement tool used by the CMS experiment at the LHC, was conceived in a more trusting time. The security model was designed to provide a safe working environment for site agents and operators, but provided little more protection than that. CMS data was not sufficiently protected against accidental loss caused by operator error or software bugs or from loss of data caused by deliberate manipulation of the database. Operations staff were given high levels of access to the database, beyond what should have been needed to accomplish their tasks. This exposed them to the risk of suspicion should an incident occur. Multiple implementations of the security model led to difficulties maintaining code, which can lead to degredation of security over time.

In order to meet the simultaneous goals of protecting CMS data, protecting the operators from undue exposure to risk, increasing monitoring capabilities and improving maintainability of the security model, the PhEDEx security model was redesigned and re-implemented. Security was moved from the application layer into the database itself, fine-grained access roles were established, and tools and procedures created to control the evolution of the security model over time. In this paper we describe this work, we describe the deployment of the new security model, and we show how the resulting enhancements have improved security on several fronts simultaneously.

Facilities, Infrastructures, Networking and Collaborative Tools / 14

Operating dedicated data centers - Is it cost-effective?

Author: Tony Wong[1]

Co-authors: Alexandr Zaytsev [2]; Christopher Hollowell [1]; Michael Ernst [3]; Richard Hogue [1]; William Strecker-Kellogg [4]

[1] *Brookhaven National Laboratory*

[2] *Brookhaven National Laboratory (US)*

[3] *Unknown*

[4] *Brookhaven National Lab*

Corresponding Authors: tony@bnl.gov, alezayt@bnl.gov

The advent of cloud computing centers such as Amazon's EC2 and Google's Computing Engine has elicited comparisons with dedicated computing clusters. Discussions on appropriate usage of cloud resources (both academic and commercial) and costs have ensued. This presentation discusses a detailed analysis of the costs of operating and maintaining the RACF (RHIC and ATLAS Computing Facility) compute cluster at Brookhaven National Lab and compares them with the cost of cloud computing resources under various usage scenarios. An extrapolation of likely future cost effectiveness of dedicated computing resources is also presented.

Poster presentations / 15

Disaster Recovery and Data Center Operational Continuity

Authors: Alexandr Zaytsev[1]; Carlos Fernando Gamboa[1]; Christopher Hollowell[2]; Costin Caramarcu[1]; David Yu[3]; Hironori Ito[1]; Jason Alexander Smith[1]; John Hover[4]; John Peter Fetzko[1]; John Steven De Stefano Jr[1]; Jose Caballero Bejar[1]; Michael Ernst[5]; Mizuki Karasawa[3]; Ofer Rind[6]; Saroj Kandasamy[1]; Shigeki Misawa[2]; Tomasz Wlodek[2]; Tony Wong[2]; William Strecker-Kellogg[7]; Xin Zhao[1]; Zhenping Liu[2]

[1] *Brookhaven National Laboratory (US)*

[2] *Brookhaven National Laboratory*

[3] *BNL*

[4] *Brookhaven National Laboratory (BNL)-Unknown-Unknown*

[5] *Unknown*

[6] *BROOKHAVEN NATIONAL LABORATORY*

[7] *Brookhaven National Lab*

Corresponding Authors: tony@bnl.gov, alezayt@bnl.gov

The RHIC and ATLAS Computing Facility (RACF) at Brookhaven Lab is a dedicated data center serving the needs of the RHIC and US ATLAS community. Since it began
operations in the mid-1990's, it has operated continuously with few unplanned downtimes. In the last 24 months, Brookhaven Lab has been affected by two hurricanes and a record-breaking snowstorm. In
this presentation, we discuss lessons learned regarding (natural or man-made) disaster preparedness, operational continuity, remote access and safety protocols, including overall operational procedures developed as a result of these recent events.

Poster presentations / **367**

The ATLAS Muon Trigger

Author: Kunihiro Nagano[1]

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

Corresponding Author: martin.woudstra@cern.ch

CERN's Large Hadron Collider (LHC) is the highest energy proton-proton collider, providing also the highest instantaneous luminosity as a hadron collider. Bunch crossings occurred every 50 ns in 2012 runs. Amongst of which the online event selection system should reduce the event recording rate down to a few 100 Hz, while events are in a harsh condition with many overlapping proton-proton collisions occurring in a same bunch crossing. Muons often provide an important and clear signature of physics processes that are searched for, for instance as in the discovery of Higgs particle in year 2012.
The ATLAS experiment deploys a three-levels processing scheme at online. The level-1 muon trigger system gets its input from fast muon trigger detectors. Fast sector logic boards select muon candidates, which are passed via an interface board to the central trigger processor and then to the High Level Trigger (HLT). The muon HLT is purely software based and encompasses a level-2 (L2) trigger followed by an event filter (EF) for a staged trigger approach. It has access to the data of the precision muon detectors and other detector elements to refine the muon hypothesis. Trigger-specific algorithms were developed and are used for the L2 to increase processing speed for instance by making use of look-up tables and simpler algorithms, while the EF muon triggers mostly benefit from offline reconstruction software to obtain most precise determination of the track parameters. There are two algorithms with different approaches, namely inside-out
and outside-in tracking, which was used in trigger with conditional-OR to obtain maximum efficiency with least processing time.
This presentation gives a full overview of the ATLAS muon trigger system, summarizes the 3 years running experiences and reports about online performances for instance processing time and trigger rates as well as trigger efficiency, resolution, and other general performance.

Facilities, Infrastructures, Networking and Collaborative Tools / **320**

Tool for Monitoring and Analysis of Large-Scale Data Movement in (Near) Real Time

Author: Wenji Wu[1]

Co-author: Phil Demar [2]

[1] *Fermi National Accelerator Laboratory*

[2] *Fermilab*

Corresponding Authors: wenji@fnal.gov, demar@fnal.gov

Fermilab is the US-CMS Tier-1 Centre, as well as the main data centre for several other large-scale research collaborations. As a consequence, there is a continual need to monitor and analyse large-scale data movement between Fermilab and collaboration sites for a variety of purposes, including network capacity planning and performance troubleshooting. To meet this need, Fermilab designed and implemented a network traffic characterization system for our large-scale bulk data transfers. The original version of the system simply analysed flow data sequentially on a conventional multi-core system. That design had two significant limitations. First, there was an appreciable delay in the analysis. While the results were still useful for network characterization studies and capacity planning purposes, they were not available in near real-time. Such a capability would open up more operationally-oriented uses. Second, the system would not scale well to 40GE and 100GE network environments, due to the sequential analysis.

Fermilab is currently developing an enhanced system which remedies those limitations. The system consists of two major components, a flow-data collection component and a data analysis engine based on GPU (graphic processing unit) technology. Our system exploits the data parallelism that exists within network flow data to provide accurate monitoring and near real-time analysis of bulk data movement. The system is designed to work in 40GE/100GE network environments. In this paper, we discuss the architecture and design of our system, including some of the applications available. Results from analysis on production traffic will be incorporated into the presentation.

Poster presentations / **27**

The Design and Realization of the Distributed Data Sharing System of the Detector Control System of the Daya Bay Neutrino Experiment

Author: Mei YE[1]

Co-authors: Shuhua ZHANG [1]; Xiaofeng DU [1]

[1] *IHEP*

Corresponding Author: yem@ihep.ac.cn

The Daya Bay reactor neutrino experiment is designed to determine precisely the neutrino mixing angle θ13 with the sensitivity better than 0.01 in the parameter sin22θ13 at the 90% confidence level. To achieve this goal, the collaboration has built eight functionally identical antineutrino detectors. The detectors are immersed in water pools that provide active and passive shielding against backgrounds. The experiment has been taking data for almost 1.5 years and making steady progress. Eight antineutrino detectors are taking data now in 3 experimental halls. And the first results have already been released. The detector control and monitoring system(DCS) was developed to support the running experiment. And according to the difference of different hardware systems, such as high voltage crates, front end electronic crates, water system, gas system, low voltage crates, temperature and Humidity of the environment system etc., different data acquisition(DAQ) modules are developed. A global control system is developed to monitor and control the entire running status of the whole experiment. Sharing data from the subsystems are most important for both equipment monitoring and data analysis. This paper will present the design and the realization of a distributed data sharing system used in the Detector Control System of the experiment. The interface of the embedded DAQ of the sensors and the communication logic will be show in details. The integration of the developed remote control framework will be introduced as well.

Key words : Daya Bay, Neutrino, Detector Control System, Embedded System, Data Acquisition, Distributed Sharing Data System, Remote Control

Data Acquisition, Trigger and Controls / **435**

The NOvA Far Detector Data Acquisition System

Author: Jaroslav Zalesak[1]

Co-authors: Alec Habig [2]; Denis Perevalov [3]; Jonathan Paley [4]; Kurt Biery [5]; Mathew Muether [6]; Peter Shanahan [6]; Ronald Rechenmacher [7]; Susan Kasahara [8]

[1] *Acad. of Sciences of the Czech Rep. (CZ)*

[2] *Univ. of Minnesota Duluth*

[3] *Fermi National Accelerator Laboratory*

[4] *Argonne National Laboratory*

[5] *CMS/Fermilab*

[6] *Fermilab*

[7] *Fermi National Accelerator Lab. (Fermilab)*

[8] *University of Minnesota*

Corresponding Author: jaroslav.zalesak@cern.ch

The NOvA experiment has developed a data acquisition system that is able to continuously digitize and produce a zero bias streaming readout for the more than 368,000 detectors cells that constitute the 14 kTon far detector. The NOvA DAQ system combines custom built frontend readout and data aggregation hardware, with advances in enterprise class networking to continuously deliver data to large commodity computing farm where the data can be buffered, examined and extracted to durable storage.

In addition to the unique design of the acquisition hardware, the NOvA DAQ system is novel in its use of a sophisticated hierarchy of software components that handle resource discovery, configuration and management as well as the formal event building and processing. This broad suite of software allows for the dynamic allocation of multiple instances of the DAQ chain. This allows the experiment to be used simultaneously for detector commissioning tasks, dedicated studies of detector performance and for production data taking and has allowed the experiment to begin production data taking while the detector is still being built.

The NOvA DAQ system was deployed to the far detector in January 2013 and has been used to successfully commission the far detector and being production data taking. This paper will present the overall design of the core data acquisition and timing systems and will examine the performance of the DAQ over the first six months of detector operations. It examines the performance issues and scaling behaviors that have been encountered in increasing the size of the readout, networking and data processing by over an order of magnitude to handle the physical size of the detector. It will present some of the first looks at the NOvA far detector neutrino data.

Facilities, Infrastructures, Networking and Collaborative Tools / **228**

SynapSense Wireless Environmental Monitoring System of the RHIC & ATLAS Computing Facility at BNL

Authors: Alexandr Zaytsev[1]; Kevin CASELLA[1]

Co-authors: Antonio WONG [1]; Christopher Hollowell [1]; Enrique GARCIA [1]; Richard HOGUE [1]; William Strecker-Kellogg [1]

[1] *Brookhaven National Laboratory (US)*

Corresponding Authors: alezayt@bnl.gov, kac@bnl.gov

RHIC & ATLAS Computing Facility (RACF) at BNL is a 15000 sq. ft. facility hosting the IT equipment of the BNL ATLAS WLCG Tier-1 site, offline farms for the STAR and PHENIX experiments operating at the Relativistic Heavy Ion Collider (RHIC), BNL Cloud installation, various Open Science Grid (OSG) resources, and many other small physics research oriented IT installations. The facility originated in 1990 and grew steadily up to the present configuration with 4 physically isolated IT areas with a combined rack capacity of about 2000 racks and the total peak power consumption of 1.5 MW. Since the infrastructural components of the RACF were deployed over such a long period of time (the oldest parts of physical infrastructure were built in late 1960s while the newest ones were added in 2010) a multitude of various environmental monitoring systems were eventually inherited in different areas. These various groups of equipment that in the end required costly maintenance and support were lacking a high level integration mechanism as well as a centralized web interface. In June 2012 a project was initiated with the primary goal to replace all these environmental monitoring systems with a single commercial hardware and software solution by SynapSense Corp. based on wireless sensor groups and proprietary SynapSense MapSense (TM) software that offers a unified solution for monitoring the temperature and humidity within the rack/CRAC units as well as pressure distribution underneath the raised floor across the entire facility. The new system also supports a set of additional features such as capacity planning based on measurements of total heat load, power consumption monitoring and control, CRAC unit power consumption optimization based on feedback from the temperature measurements and overall power usage efficiency estimations that are not currently implemented within RACF but may be deployed in the future.
This contribution gives a detailed review of all the stages of deployment of the system and its integration with the existing personnel notification mechanisms and emergency management/disaster recovery protocols of the RACF. The experience gathered while operating the system is summarized and a comparative review of functionality/maintenance costs for the RACF environmental monitoring system before and after transition is given.

Data Acquisition, Trigger and Controls / 87

10Gbps TCP/IP streams from the FPGA for High Energy Physics

Author: Petr Zejdl[1]

Co-authors: Andre Georg Holzner [2]; Andrea Petrucci [1]; Andrei Cristian Spataru [1]; Attila Racz [1]; Aymeric Arnaud Dupont [1]; Carlos Nunez Barranco Fernandez [1]; Christian Deldicque [1]; Christian Hartl [1]; Christoph Paus [3]; Christoph Schwick [1]; Christopher Colin Wakefield [4]; Dominique Gigi [1]; Emilio Meschi [1]; Fabian Stoeckli [3]; Frank Glege [1]; Frans Meijers [1]; Gerry Bauer [3]; Giovanni Polese [5]; Hannes Sakulin [1]; James Gordon Branson [2]; Jose Antonio Coarasa Perez [1]; Konstanty Sumorok [3]; Lorenzo Masetti [1]; Luciano Orsini [1]; Marc Dobson [1]; Marco Pieri [2]; Matteo Sani [2]; Olivier Chaze [1]; Olivier Raginel [3]; Remi Mommsen [6]; Robert Gomez-Reino Garrido [1]; Samim Erhan [7]; Sergio Cittolin [2]; Srecko Morovic [8]; Ulf Behrens [9]; Vivian O'Dell [10]; Wojciech Andrzej Ozga [11]

[1] *CERN*

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

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

[4] *Staffordshire University (GB)*

[5] *University of Wisconsin (US)*

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

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

[8] *Institute Rudjer Boskovic (HR)*

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

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

[11] *AGH University of Science and Technology (PL)*

Corresponding Author: petr.zejdl@cern.ch

The CMS data acquisition (DAQ) infrastructure collects data from more than 600 custom detector Front End Drivers (FEDs). In the current implementation data is transferred from the FEDs via 3.2 Gbs electrical links to custom interface boards, which transfer the data to a commercial Myrinet network based on 2.5 Gbps optical links. During 2013 and 2014 the CMS DAQ system will undergo a major upgrade to face the new challenges expected after the upgrade of the LHC accelerator and various detector components. The interface to the FED readout links will be implemented with a custom card based on FPGAs. The Myrinet network will be replaced by a 10Gbps Ethernet network running a TCP/IP protocol. This allows us to reliably aggregate several streams at the destination. To limit the implementation complexity we designed a stripped down version of the TCP/IP protocol. We preserved the compliance with the RFC 793. Therefore we can use a PC with the standard Linux TCP/IP stack as a receiver. We present the hardware challenges and architectural choices made to the TCP/IP protocol in order to simplify its FPGA implementation. We also describe and discuss the interaction between hardware and software TCP/IP stacks. The performance measurements of the current prototype will be presented.

Poster presentations / **399**

Keyword Search over Data Service Integration for Accurate Results

Authors: Valentin Y Kuznetsov[1]; Vidmantas Zemleris[2]

Co-author: Peter Kreuzer [3]

[1] *Cornell University (US)*

[2] *Vilnius University (LT)*

[3] *Rheinisch-Westfaelische Tech. Hoch. (DE)*

Corresponding Author: vidmantas.zemleris@cern.ch

Background: The goal of the virtual data service integration is to provide a coherent interface for querying a number of heterogenous data sources (e.g., web services, web forms, proprietary systems, etc.) in cases where accurate results are necessary. This work explores various aspects of its usability.

Problem: Querying is usually carried out through a structured query language, such as SQL, which forces the users to learn the language and to get acquainted with data organization (i.e. the schema) thus negatively impacting the system's usability. Limited access to data instances as well as users' concern with accurate results of arbitrary queries present additional challenges to traditional approaches (such as query forms, information retrieval, keyword search over relational databases) making them not applicable.

Solution: This paper presents a keyword search system which deals with the above discussed problem by operating on available information: the metadata, such as the constraints on allowed values, analysis of user queries, and certain portions of data. Given a keyword query, it proposes a ranked list of structured queries along with the explanations of their meanings. Unlike previous implementations, the system is freely available and makes no assumptions about the input query, while maintaining its ability to leverage the query's structural patterns - in case they exist. The system is discussed in the context of CMS data discovery service where the simplicity and capabilities of the search interface play a crucial role in the ability of its users to satisfy their information needs.

Poster presentations / **451**

Maximising job throughput using Hyper-Threading

Authors: Alastair Dewhurst[1]; DIMITRIOS ZILASKOS[2]

[1] *STFC - Science & Technology Facilities Council (GB)*

[2] *STFC*

Corresponding Authors: dimitrios.zilaskos@stfc.ac.uk, alastair.dewhurst@cern.ch

The WLCG uses HEP-SPEC as its benchmark for measuring CPU performance. This provides a consistent and repeatable CPU benchmark to describe experiment requirements, lab commitments and existing resources. However while HEP-SPEC has been customized to represents WLCG applications it is not a perfect measure.

The Rutherford Appleton Laboratory (RAL), is the UK Tier 1 site and provides CPU and disk resources for the four largest LHC experiments as well as to numerous other experiments.

Recent generations of hardware procurement at RAL have included CPUs with Hyper-Threading. Previous studies have shown that as the number of logical cores being used increases, the measured HEP-SPEC will also increase but by increasingly smaller amounts. The more jobs that are running, the higher the chance that their will be contention on other resources which will cause jobs to slow down. It is therefore not obvious what is the optimal number of jobs to run on a Hyper-Threaded machine.

This paper details the work done to maximize job throughput at RAL. Over the course of several months different machine configurations were tested at RAL to see the impact on real job throughput. The results have allowed RAL to maximize job throughput while also accurately reporting the available HEP-SPEC and provided useful information for future procurements.

Poster presentations / **285**

Self-Organizing Map in ATLAS Higgs Searches

Author: Giovanni Zurzolo[1]

Co-authors: Arturo Sanchez Pineda [2]; Claudio Savarese [1]; Francesco Conventi [1]

[1] *Universita e INFN (IT)*

[2] *Universita di Napoli Federico II-Universita e INFN*

Corresponding Authors: giovanni.zurzolo@cern.ch, francesco.conventi@cern.ch, arturo.sanchez.pineda@cern.ch, claudio.savarese@cern.ch

Artificial Neural Networks (ANN) are widely used in High Energy Physics, in particular as software for data analysis. In the ATLAS experiment that collects proton-proton and heavy ion collision data at the Large Hadron Collider, ANN are mostly applied to make a quantitative judgment on the class membership of an event, using a number of variables that are supposed to discriminate between different classes. The discrimination between quark-initiated and gluon-initiated jets is an example of the possible applications in the ATLAS experiment and it potentially has a great chance to improve many physics analyses.

In this work the application of the unsupervised Self-Organizing Map (SOM) is proposed as quark-gluon tagging for events with two jets in the ATLAS experiment. The performance of the SOM application for quark-gluon discrimination are shown in different ranges of jet pT , confirming the feasibility of the quark-gluon tagging down to very low pT values. The application of the SOM technique to the Higgs searches is described and the results are shown and discussed.

Poster presentations / **526**

Posters (villages 2, 4, and 6)

Poster presentations / **527**

Posters (roam free)

Summaries / **528**

Lightning Talks

Plenaries / **499**

Welcome

DPHEP Workshop / **518**

Round-table of Experiment / Lab activities

DPHEP Workshop / **508**

DPHEP: Where do we want to be in Okinawa?

DPHEP Workshop / **506**

DPHEP Portal - what should it cover?

DPHEP Workshop / **507**

DPHEP Common Projects cont.

DPHEP Workshop / **504**

DPHEP: HEPiX ”Bit Preservation” Working Group

DPHEP Workshop / **505**

DPHEP: ”CERNLIB consortium”

Facilities, Infrastructures, Networking and Collaborative Tools / **245**

Application Performance Evaluation and Recommendations for the DYNES

Co-authors: Aaron Brown ; Alan Tackett [1]; Andrew Malone Melo [2]; Artur Jerzy Barczyk [3]; Azher Mughal [3]; Ben Meekhof [4]; Dale Finkelson [5]; Eric Boyd [5]; Harvey Newman [3]; Jason Zurawski [5]; Mathew Binkley [6]; Paul Sheldon [2]; Ramiro Voicu [3]; Robert Ball [7]; Robert Brown [8]; Sandor Rozsa [3]; Stephen Wolff [5]

[1] *VANDERBILT UNIVERSITY*

[2] *Vanderbilt University (US)*

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

[4] *University of Michigan*

[5] *Internet2*

[6] *V*

[7] *University of Michigan (US)*

[8] *Vanderbilt University*

Computing and networking infrastructures across the world continue to grow to meet the increasing needs of data intensive science, notably those of the LHC and other large high energy physics collaborations. The LHC’s large data volumes challenge the technology used to interconnect widely-separated sites (and their available resources) and lead to complications in the overall process of end-to-end data distribution, analysis and management. A delicate balance is required to serve both long-lived, high capacity network flows, as well as more traditional end-user activities using general purpose infrastructure.

R&E networks have experimented with Virtual Circuits (VC) for a number of years as a mechanism that affords greater control over network capacity and traffic management [Oscars, ION, SDN, Autobahn]. This connection-oriented concept emulates a physical point-to-point connection, using the underlying technology of common packet-switched networks. In contrast to a physical circuit, VC technology allows for variable duration, guaranteed bandwidth channels, and fosters efficient use of common network infrastructures.

The DYNES instrument, an NSF funded cyberinfrastructure project designed to facilitate end-to-end dynamic circuit services, is built using this VC technology, as well as other common open source software packages for network monitoring and data movement [DYNES, FDT, perfSONAR-PS]. Dynamic circuits have been used in production for the last 6 years among a limited number of major laboratory and university sites. DYNES is extending this capability to many campuses with the goal of increasing the number of sites able to easily participate as end-points of virtual circuits.

A key observation during installation and testing of DYNES was related to the performance of standard data movement tools over virtual circuits: the observed performance did not match expectations related to the bandwidth reserved in circuits. In many cases the data movement reality was an order of magnitude lower than the initial bandwidth request; investigation as to a possible cause centered on the QoS mechanisms of the underlying network. In most cases bandwidth reservations are significantly below the “wire-speed” of the host’s network interface card. This was deemed a likely source of at least part of the low performance typically observed.

Our study focused on factors commonly responsible for degrading network performance: buffer overruns due to lack of available memory in relation to application burst behavior and queuing on network devices that introduce out of order behavior in TCP streams. After experimenting with these behaviors we explored various techniques, some of which do not require modification of legacy

applications, which can be used to mitigate these concerns at the end hosts.
We will present the results of our testing and list the benefits and shortcomings of the various options we explored. We will discuss our experiences with kernel network stack tuning, application pacing, tc, RoCE and TCP variants. When implemented correctly these mechanisms will improve (sometimes significantly) the end-to-end flow of traffic across VC resources.

DPHEP Workshop / 514

DPHEP Common Projects

Talks on ongoing projects:
LHCb, CDF, PREDON, H2020 perspective