

# **ACAT 2011**

Monday, September 5, 2011 - Friday, September 9, 2011

## **Book of Abstracts**



# Contents

DRA method: Powerful tool for the calculation of the loop integrals. . . . .	1
Multiloop calculations in supersymmetric theories with the higher covariant derivative regularization . . . . .	1
Application of Symbolic Regression to Mass Measurement in H->WW Dilepton Channels	1
HYPERDIRE: HYPERgeometric Differential REDuction HYPERDIRE: HYPERgeometric Differential REDuction Mathematica based programs for differential reduction of hypergeometric functions and its application to Feynman Diagrams Calculation. . . . .	2
ATLAS data processing with GridPilot on NorduGrid and WLCG . . . . .	3
Full Reconstruction based on NeuroBayes at the Belle Experiment . . . . .	3
Application of Remote Debugging Techniques in User-Centric Job Monitoring . . . . .	4
Regularization Schemes and Higher Order Corrections . . . . .	4
Advanced event reweighting for MVA training. . . . .	4
Integrating Amazon EC2 with the CMS Production Framework . . . . .	5
Tau identification using multivariate techniques in ATLAS . . . . .	5
Real time algorithms in the ATLAS tau trigger system at 7 TeV center of mass energy . .	6
ADAM Project – A generic web interface for retrieval and display of ATLAS TDAQ information. . . . .	6
A Persistent Back-End for the ATLAS TDAQ Online Information Service (P BEAST) . . .	7
Online Measurement of LHC Beam Parameters with the ATLAS High Level Trigger . . .	8
DRUID: Event Display for the Linear Collider Detector . . . . .	8
Fractal dimension analysis in a highly granular calorimeter . . . . .	9
The AAL project: Automated monitoring and intelligent AnaLysis for the ATLAS data taking infrastructure . . . . .	9
Alignment of the ATLAS Inner Detector . . . . .	10
Automated Quality Monitoring and Validation of the CMS Reconstruction Software . . .	11

Visual Physics Analysis (VISPA) - From Desktop Towards Physics Analysis at Your Fingers . . . . .	11
Track reconstruction and b-jet identification for the ATLAS trigger system . . . . .	11
Monitoring the Grid at local, national, and global levels . . . . .	12
Dynamic deployment of a PROOF-based analysis facility for the ALICE experiment over virtual machines using PoD and OpenNebula . . . . .	13
Mass production of air showers using Grid for Auger . . . . .	13
Unparametrized multi-dimensional kernel density- and likelihood ratio estimator . . . . .	14
Continuous simulation of Beyond-Standard-Model processes with multiple parameters . . . . .	14
A Linear Iterative Unfolding Method . . . . .	14
SecDec: a tool for numerical multi-loop/leg calculations . . . . .	15
Three-Loop Calculation of the Higgs Boson Mass in Supersymmetry . . . . .	15
FormCalc 7 . . . . .	16
A Validation System for Data Preservation in HEP . . . . .	16
LHCb distributed computing operations . . . . .	17
GELATIO - The GERDA framework for digital signal analysis . . . . .	17
Numerical evaluation of one-loop QCD amplitudes . . . . .	18
Feynman integrals, polylogarithms and symbols . . . . .	18
Strange Bedfellows: Quantum Mechanics and Data Mining . . . . .	18
The LHCb DIRAC-based production and data management operations systems . . . . .	19
DIRAC Framework for Distributed Computing Systems . . . . .	19
Can 'Go' address the multicore issues of today and the manycore problems of tomorrow ?	20
A CMake-based build and configuration framework . . . . .	20
Track finding using GPUs . . . . .	21
One-loop integrations with Hypergeometric functions . . . . .	21
Gibbs sampler for background discrimination in particle physics . . . . .	22
Challenges in using GPUs for the reconstruction of digital hologram images. . . . .	22
Non-parametric comparison of histogrammed two-dimensional data distributions using the Energy Test. . . . .	23
Offloading peak processing to Virtual Farm by STAR experiment at RHIC . . . . .	23
Data Preservation in High Energy Physics . . . . .	24

10 Years of Object-Oriented Analysis on H1 . . . . .	24
Advances in Service and Operations for ATLAS Data Management . . . . .	25
An Alternative Method for Tilecal Signal Detection and Amplitude Estimation . . . . .	26
Development of noSQL data storage for the ATLAS PanDA Monitoring System . . . . .	26
An adaptive Monte-Carlo Markov chain algorithm for counting muons in Auger water Cherenkov detector signals . . . . .	27
Reweighting NNPDFs. . . . .	27
PROOF Performance Measurements Using PROOF Benchmark Suite . . . . .	28
Progress on the Direct Computation Method . . . . .	28
Regularization of IR-divergent loop integrals . . . . .	29
The EOS disk storage system at CERN . . . . .	29
Self-Organizing Maps Parametrization of Deep Inelastic Structure Functions with Error Determination . . . . .	30
Different forms of the generalized Crewther relation in QCD and QED: concrete conse- quences of analytical multiloop calculations . . . . .	30
Semi-Supervised Anomaly Detection - Towards Model-Independent Searches of New Physics . . . . .	31
Comparison of the CPU and memory performance of StatPatternRecognition (SPR) and Toolkit for MultiVariate Analysis (TMVA) . . . . .	31
Multicore in Production: Advantages and Limits of the Multi-process Approach. . . . .	32
The toolbox of modern multi-loop computations: novel analytic and semi-analytic tech- niques . . . . .	33
One-loop tensor Feynman integral reduction with signed minors . . . . .	33
Status of TMVA, the toolkit for multivariate analysis . . . . .	33
Online Particle Detection by Neural Networks Based on Topologic Calorimetry Information . . . . .	34
GPU Linear algebra extensions for GNU/Octave . . . . .	34
An Exploration of SciDB in the Context of Emerging Technologies for Data Stores in Parti- cle Physics and Cosmology . . . . .	35
Lessons from Static Analysis on HEP Software . . . . .	36
Moving ROOT Forward. . . . .	36
Multivariate analysis and data mining: statistics in the computer age . . . . .	36
Evaluation of likelihood functions on CPU and GPU devices . . . . .	36

Do regions of ALICE matter? (Social relationships and data exchanges in the Grid) . . .	37
Multivariate Correlated Sampling Using Extended Alias Techniques . . . . .	37
Efficient Pseudo-Random Number Generation for Monte-Carlo Simulations Using Graphic Processors . . . . .	38
Making distributed ALICE analysis simple using the GRID plug-in . . . . .	39
Modeling Fake Missing Transverse Energy with Bayesian Neural Networks . . . . .	39
One click dataset transfer: toward efficient coupling of distributed storage resources and CPUs. . . . .	39
The alignment of the CMS Silicon Tracker . . . . .	40
The five dimensions of the genome . . . . .	40
Status of parallelization of FORM . . . . .	41
Progress in Automated Next-to-Leading Order calculations . . . . .	41
Modern actions, algorithms, and computers for Lattice QCD . . . . .	42
Automated one-loop calculations with Golem/Samurai . . . . .	42
Polynomial Algebra in Form 4 . . . . .	42
An analytical solution for a non-planar massive double box diagram . . . . .	43
Reduze 2 . . . . .	43
New approaches for numerical techniques in higher order calculations . . . . .	43
NFS 4.1/pNFS, the final step . . . . .	44
Where do we go from here? - The next phase of computing in HEP . . . . .	44
Computing On Demand: Analysis in the Cloud . . . . .	45
Using machine learning techniques in classification problems in Astrophysics . . . . .	45
SALAMI project . . . . .	46
Building an Outsourcing Ecosystem for Science . . . . .	46
Summary - Computing Technology for Physics Research . . . . .	47
Summary - Data Analysis – Algorithms and Tools . . . . .	47
Summary - Computations in Theoretical Physics – Techniques and Methods . . . . .	47
ACAT 2011 - Summary and outlook . . . . .	47
Welcome - Prof. Geoff Rodgers, Pro-Vice Chancellor for Research, Brunel University . .	48

Tuesday 06th - Computations in Theoretical Physics / 2

## **DRA method: Powerful tool for the calculation of the loop integrals.**

**Author:** Roman Lee<sup>1</sup>

<sup>1</sup> *Budker Institute of Nuclear Physics*

**Corresponding Author:** r.n.lee@inp.nsk.su

The method of calculation of the loop integrals based on the dimensional recurrence relation and analyticity of the integrals as functions of  $d$  is reviewed. Special emphasis is made on the possibility to automatize many steps of the method. New results obtained with this method are presented.

Monday 05th - Computations in Theoretical Physics / 3

## **Multiloop calculations in supersymmetric theories with the higher covariant derivative regularization**

**Author:** konstantin stepanyantz<sup>1</sup>

<sup>1</sup> *Moscow State University*

**Corresponding Author:** stepanyantz@mail.ru

Most calculations of quantum correction in the supersymmetric theories are made with the dimensional reduction, which is a modification of the dimensional regularization. However, it is well known that the dimensional reduction is not self-consistent. A consistent regularization, which does not break the supersymmetry is the higher covariant derivative regularization. However, the integrals obtained with this regularization can not be usually calculated analytically. We discuss application of this regularization to the calculations in supersymmetric theories. In particular, it is demonstrated that integrals defining the beta-function are possibly integrals of total derivatives. This feature allows to explain the origin of the exact NSVZ beta-function, relating the beta-function with the anomalous dimensions of the matter superfields. However, integrals for the anomalous dimension should be calculated numerically.

Thursday 08 - Data Analysis – Algorithms and Tools / 4

## **Application of Symbolic Regression to Mass Measurement in $H \rightarrow WW$ Dilepton Channels**

**Author:** Su Yong Choi<sup>1</sup>

<sup>1</sup> *Korea University*

**Corresponding Author:** su.yong.choi@cern.ch

We derive a kinematic variable that is sensitive to the mass of the Standard Model Higgs boson ( $M_H$ ) in the  $H \rightarrow WW^* \rightarrow l \bar{\nu} \nu$  channel using symbolic regression method. Explicit mass reconstruction is not possible in this channel due to the presence of two neutrinos which escape detection. Mass determination problem is that of finding a mass-sensitive function that depends on the measured observables. We use symbolic regression, which is an analytical approach to the problem of non-linear regression, to derive an analytic formula sensitive to  $M_H$  from the two lepton momenta and the missing transverse momentum. Using the newly-derived mass-sensitive variable,

we expect Higgs mass resolutions between 1 to 4 GeV for  $M_H$  between 130 and 190 GeV at the LHC with  $10 \text{ fb}^{-1}$  of data.

**Tuesday 06th - Computations in Theoretical Physics / 7**

## **HYPERDIRE: HYPERgeometric Differential REDuction HYPER-DIRE: HYPERgeometric Differential REDuction Mathematica based programs for differential reduction of hypergeometric functions and its application to Feynman Diagrams Calculation.**

**Author:** Bytev Vladimir<sup>1</sup>

**Co-authors:** B Ward<sup>2</sup>; Berndt Kniehl<sup>3</sup>; Mikhail Kalmykov<sup>3</sup>; Scott Yost<sup>4</sup>

<sup>1</sup> *JINR*

<sup>2</sup> *Baylor University*

<sup>3</sup> *II. Institut für Theoretische Physik, Universität Hamburg*

<sup>4</sup> *Citadel Military Coll.*

**Corresponding Author:** bvv@jinr.ru

The differential reduction algorithm allows to change the values of parameters of any Horn-type hypergeometric functions on arbitrary integers numbers. The description of mathematical part of algorithm have been presented on ACAT08 by M.Kalmykov [6].

We will describe the status of project and will present a new version of MATHEMATICA based package including a several important hypergeometric functions of one and two variables.

Interrelation between Differential Reduction algorithm and Integration-by-Parts technique is discussed.

We illustrate the procedure in the context of generalized hypergeometric functions, and give an example for a type of bubble and propagator type diagram.

Another application of HYPERDIRE is the construction of epsilon-expansion of Horn-type Hypergeometric Functions.

Talk is based on the following publications:

1. "HYPERDIRE: HYPERgeometric functions Differential REDuction MATHEMATICA based packages for differential reduction of generalized hypergeometric functions: now with pFq, F1,F2,F3,F4"  
by V.V.Bytev, M.Yu.Kalmykov,B.A.Kniehl,  
[arXiv:1105.3565]
2. "Differential Reduction Techniques for the Evaluation of Feynman Diagrams"  
by S.A. Yost, V.V. Bytev, M.Yu. Kalmykov, B.A. Kniehl, B.F.L. Ward  
PoS ICHEP2010:135,2010  
[arXiv:1101.2348]
3. "Differential reduction of generalized hypergeometric functions from Feynman diagrams: One-variable case",  
by V.V.Bytev, M.Yu.Kalmykov,B.A.Kniehl,  
Nucl.Phys.B836:129-170, 2010 [arXiv:0904.0214]
4. "Counting master integrals: integration by parts vs. differential reduction"



by Mikhail Yu. Kalmykov, Bernd A. Kniehl  
[arXiv:1105.5319]

5.  
“Differential Reduction Algorithms for Hypergeometric Functions  
Applied to Feynman Diagram Calculation”  
by V.V.Bytev, M.Kalmykov, B.A.Kniehl, B.F.L.Ward, S.A.Yost  
[arXiv:0902.1352]

[6]  
“Feynman Diagrams, Differential Reduction, and Hypergeometric Functions”  
by M. Yu. Kalmykov, V. V. Bytev, Bernd A. Kniehl, B.F.L. Ward, S.A.Yost  
PoS ACAT08:125,2009 [arXiv:0901.4716]

8

## ATLAS data processing with GridPilot on NorduGrid and WLCG

**Author:** Frederik Orellana<sup>1</sup>

**Co-authors:** Ask Emil Jensen<sup>1</sup>; Jacob Debel<sup>1</sup>; Jørgen Beck Hansen<sup>1</sup>; Morten Badensø<sup>1</sup>; Peter kadlecik; Simon Heisterkamp<sup>1</sup>

<sup>1</sup> *University of Copenhagen*

**Corresponding Author:** frederik.orellana@cern.ch

We present a novel tool for managing data processing on grid resources. The tool provides a graphical user interface that offers new ATLAS users a quick and gentle start with computing, using a library of applications built up by previous users.

Tuesday 06th - Data Analysis – Algorithms and Tools / 9

## Full Reconstruction based on NeuroBayes at the Belle Experiment

**Authors:** Anze Zupanc<sup>1</sup>; Daniel Zander<sup>1</sup>; Michael Feindt<sup>1</sup>; Sebastian Neubauer<sup>1</sup>; Thomas Kuhr<sup>1</sup>

<sup>1</sup> *Karlsruhe Institute of Technology*

**Corresponding Author:** zander@ekp.uni-karlsruhe.de

Full Reconstruction is an important analysis technique utilized at B factories where B mesons are produced in  $e^+e^- \rightarrow Y(4S) \rightarrow B\bar{B}$  processes. By reconstructing one of the two B mesons in an event fully in a hadronic final state, the properties of the other B meson are determined using momentum conservation. Therefore, it allows to measure or perform searches for rare B meson decays involving one or more neutrinos in the final state.

This ansatz is complicated in practice by huge combinatorics and large amounts of background. With over 1000 exclusively reconstructed B decay channels the Full Reconstruction utilizes a hierarchical reconstruction procedure and probabilistic calculus instead of classical selection cuts. In this approach, the decision to accept or reject a candidate is delayed to a later stage in order to make the most use of all available information. The multivariate analysis software package NeuroBayes was used extensively to hold the balance between highest possible efficiency and acceptable consumption of CPU time.

As a result of applying this ansatz, the number of fully reconstructed B mesons was increased by a factor of 2 after 10 years of successful data taking. The new full reconstruction algorithm will thus allow for more precise measurements of rare B meson decays.

**Monday 05th - Computing Technology for Physics Research / 10**

## **Application of Remote Debugging Techniques in User-Centric Job Monitoring**

**Author:** Tim dos Santos<sup>1</sup>

**Co-authors:** Frank Volkmer<sup>1</sup>; Nikolaus Wulff<sup>2</sup>; Peter Maettig<sup>1</sup>; Raphael Ahrens<sup>1</sup>; Sergey Kalinin<sup>1</sup>; Thomas Beermann<sup>1</sup>; Torsten Harenberg<sup>1</sup>

<sup>1</sup> *Bergische Universitaet Wuppertal*

<sup>2</sup> *Fachhochschule Muenster*

With the Job Execution Monitor, a user-centric job monitoring software developed at the University of Wuppertal and integrated into the Pilot-based “PanDA” job brokerage system of the WLCG, job progress and grid worker node health can be supervised in real time. Imminent error conditions can thusly be detected early by the submitter and countermeasures taken. Grid site admins can access aggregated data of all monitored jobs to infer the site status and to detect job misbehaviour. To remove the last “blind spot” from this monitoring, a remote debugging technique based on the GNU C compiler suite was developed and integrated into the software; its design concept and architecture will be described and its application discussed.

**Monday 05th - Computations in Theoretical Physics / 11**

## **Regularization Schemes and Higher Order Corrections**

**Author:** William Kilgore<sup>1</sup>

<sup>1</sup> *Brookhaven National Lab*

**Corresponding Author:** kilgore@bnl.gov

I apply commonly used regularization schemes to a multiloop calculation to examine the properties of the schemes at higher orders. I find complete consistency between the conventional dimensional regularization scheme and dimensional reduction, but I find that the four-dimensional helicity scheme produces incorrect results at next-to-next-to-leading order and singular results at next-to-next-to-next-to-leading order. It is not, therefore, a unitary regularization scheme.

**Tuesday 06th - Data Analysis – Algorithms and Tools / 14**

## **Advanced event reweighting for MVA training.**

**Authors:** Daniel Martschei<sup>1</sup>; Michael Feindt<sup>1</sup>

**Co-authors:** Jeannine Wagner-Kuhr<sup>1</sup>; Simon Honc<sup>1</sup>

<sup>1</sup> *Inst. für Experimentelle Kernphys.-Universitaet Karlsruhe-KIT*

**Corresponding Author:** daniel.martschei@cern.ch

Title: Advanced event reweighting for MVA training.

Multivariate discrimination techniques, such as Neural Networks, are key ingredients to modern data analysis and play an important role in high energy physics. They are usually trained on simulated Monte Carlo (MC) samples to discriminate signal from background and are then applied to data. This has in general some side effects which we address in this talk.

One is that the discriminator behaviour on real data depends on the agreement between the MC training sample and data. We present ways of re-weighting MC samples on a per event basis to make them more look like data.

In some cases it is even possible to become completely independent from MC simulations by using the sPlot technique, which also makes extensive use of weights during the training and is a sort of advanced background subtraction procedure.

Another issue is that a cut on the discriminator can change the distribution of variables which discriminate signal from background themselves. This becomes an issue if one wants to see and fit a clear signal peak in this distribution on data as a final result, e.g. in the invariant mass of decay particles. Our approach uses a neural network which is trained to discriminate between signal and background while explicitly disallowing any influence on the distribution the variable of interest to be used for template fits in the end.

We will give examples of the application of these three techniques performed with the NeuroBayes package in different physics analysis.

**Monday 05th - Computing Technology for Physics Research / 15**

## **Integrating Amazon EC2 with the CMS Production Framework**

**Author:** Andrew Malone Melo<sup>1</sup>

<sup>1</sup> *Vanderbilt University*

**Corresponding Author:** andrew.malone.melo@cern.ch

As cloud middleware (and cloud providers) have become more robust, various experiments with experience in Grid submission have begun to investigate the possibility of taking previously Grid-Enabled applications and making them compatible with Cloud Computing, which will allow for dynamic scaling of the available hardware resources on a dynamic basis, providing access to peak-load handling capabilities and possibly resulting in lower costs to the experiment. Here we discuss current work within the CMS collaboration at the LHC to both perform computation on EC2, both for production and analysis use-cases.

**Tuesday 06th - Data Analysis – Algorithms and Tools / 16**

## **Tau identification using multivariate techniques in ATLAS**

**Author:** Dugan O'Neil<sup>1</sup>

<sup>1</sup> *Simon Fraser University (SFU)*

**Corresponding Author:** dugan.o'neil@cern.ch

Tau leptons will play an important role in the physics program at the LHC. They will be used in electroweak measurements and in detector related studies like the determination of the missing transverse energy scale, but also in searches for new phenomena like the Higgs boson or Supersymmetry.

Due to the huge background from QCD processes, efficient tau identification techniques with large fake rejection are essential. Tau object appear as collimated jets with low track multiplicity and single variable criteria are not enough to efficiently separate them from jets and electrons. This can be achieved using modern multivariate techniques which make optimal use of all the information available. They are particularly useful when the discriminating variables are not independent and no single variable provides good signal and background separation.

In ATLAS several advanced algorithms are applied to identify taus, in particular a projective likelihood estimator and boosted decision trees. All multivariate methods applied to the ATLAS simulated data perform better than the baseline cut analysis. Their performance is shown using high energy data collected at the ATLAS experiment. The strengths and weaknesses of each technique are also discussed.

17

## Real time algorithms in the ATLAS tau trigger system at 7 TeV center of mass energy

**Authors:** Kadlecik Peter<sup>1</sup>; Pilar Casado Lechuga<sup>2</sup>

<sup>1</sup> *Theoretical High Energy Phys. Dept. (NBI)-Niels Bohr Inst. Astr*

<sup>2</sup> *IFAE - Institut de Física d'Altes Energies-Universitat Autònoma*

**Corresponding Authors:** peter.kadlecik@cern.ch, casado@ifae.es

The ATLAS tau trigger system runs very challenging real time algorithms on commodity computers. Whilst in the second level trigger (L2) fast and specialized algorithms are used, in the third level trigger (Event Filter -EF-) sophisticated and detailed reconstruction algorithms run. The performance of both types of algorithms can be decoupled because they both start from the information provided by first level (L1) hardware-based system. For both cases, data from the whole detector can be used, and in fact there are dedicated separate algorithms processing the calorimeter data and the data from the tracking detectors.

In this contribution we focus on the online performance of the L2 and EF algorithms during 2011 data taking period at the LHC, with special emphasis on the fast calorimeter selection. We present the overall performance and robustness of the operation of such algorithms during its use at the LHC. Finally, we outline the plans for future operations in light of the experience accumulated during this year's running.

18

## ADAM Project – A generic web interface for retrieval and display of ATLAS TDAQ information.

**Author:** Adam Harwood<sup>1</sup>

**Co-authors:** Giovanna Lehmann Miotto<sup>2</sup>; Luca Magnoni<sup>2</sup>

<sup>1</sup> *University of the West of England*

<sup>2</sup> *CERN*

**Corresponding Authors:** adam.james.harwood@cern.ch, luca.magnoni@cern.ch

This paper describes a new approach to the visualization of stored information about the operation of the ATLAS Trigger and Data Acquisition system.

ATLAS is one of the two general purpose detectors positioned along the Large Hadron Collider at CERN.

Its data acquisition system consists of several thousand computers interconnected via multiple gigabit Ethernet networks, that are constantly monitored via different tools. Operational parameters ranging from the temperature of the computers, to the network utilization are stored in several databases for a posterior analysis. Although the ability to view these data-sets individually is already in place, there currently is no way to view this data together, in a uniform format, from one location.

The ADAM project has been launched in order to overcome this limitation. It defines a uniform web interface to collect data from multiple diversely structured providers. It is capable of aggregating and correlating the data according to user defined criteria. Finally it visualizes the collected data using a flexible and interactive web front-end system.

Structurally, the project comprises of 3 main levels of the data collection 'cycle':

The Level 0 represents the information sources within ATLAS. These providers do not store information in a uniformed fashion. The first step of the project was to define a common interface with which to expose stored data. The interface designed for the project originates from the Google Data Protocol API. The idea is to allow read-only access to data providers, through HTTP requests similar in format to the SQL query structure. This provides a standardized way to access this different information sources within ATLAS.

The Level 1 can be considered the engine of the system. The primary task of the Level is to gather data

from multiple data sources via the common interface, to correlate this data together, or over a defined time series, and expose the combined data as a whole to the Level 2 web interface.

The Level 2 is designed to present the data in a similar style and aesthetic, despite the differing origins of the data. Pages can be constructed, edited and personalized by users to suit the specific data being shown. Pages can show a collection of graphs displaying data potentially coming from multiple sources.

The project as a whole has a great amount of scope thanks to the uniformed approach chosen for exposing data, and the flexibility of the Level 2 in presenting results.

The paper will describe in detail the design and implementation of this new tool. In particular we will go through the project architecture, the implementation choices and the examples of usage of the system in place within the ATLAS TDAQ infrastructure.

19

## **A Persistent Back-End for the ATLAS TDAQ Online Information Service (P BEAST)**

**Author:** Alexandru Dan Sicoe<sup>1</sup>

**Co-authors:** Giovanna Lehmann Miotto<sup>1</sup>; Igor Soloviev<sup>2</sup>; Luca Magnoni<sup>1</sup>; Serguei Kolos<sup>2</sup>

<sup>1</sup> *CERN*

<sup>2</sup> *University of California Irvine (US)*

**Corresponding Author:** alexandru.dan.sicoe@cern.ch

ATLAS is the largest of several experiments built along the Large Hadron Collider at CERN, Geneva. Its aim is to measure particle production when protons collide at a very high center of mass energy, thus reproducing the behavior of matter a few instants after the Big Bang. The detecting techniques

used for this purpose are very sophisticated and the amount of digitized data created by the sensing elements requires a very large data acquisition system, based on thousands of interconnected computers.

The experiment is successfully taking data since the end of 2008 and the trigger and data acquisition are now in a production stage. The main development efforts of the ATLAS Controls and Configurations work group are guided towards adding easy to use and intuitive tools to aid experts monitor different components or subsystems. P BEAST is an example of such a tool. It facilitates the storage of vast amounts of operational information which is otherwise lost. With this data at hand, long term analysis is possible.

This paper describes the work done and the results obtained in implementing a persistent system for the ATLAS Online Information Service. The novelty of this research consists of using a modern key-value storage technology (Cassandra) to satisfy the massive time series data rates, flexibility and scalability requirements entailed by the project. Cassandra's horizontally scalable system accounts for the fact that the the TDAQ hardware and software infrastructure will continue growing in the next few years and thus the amount operational information data to store will increase. The loose schema allows for seamless evolution of the stored information with the one flowing in the system, creating a permanent footprint of the data coming into the Information Service. This allows past issues to be investigated without needing to reproduce the problem in order to observe behavior of different components.

## Monday 05th - Computing Technology for Physics Research / 20

### Online Measurement of LHC Beam Parameters with the ATLAS High Level Trigger

**Author:** Emanuel Alexandre Strauss<sup>1</sup>

<sup>1</sup> *SLAC National Accelerator Laboratory*

**Corresponding Author:** [estrauss@slac.stanford.edu](mailto:estrauss@slac.stanford.edu)

We present an online measurement of the LHC beam parameters in ATLAS using the High Level Trigger (HLT). When a significant change is detected in the measured beamspot, it is distributed to the HLT. There, trigger algorithms like b-tagging which calculate impact parameters or decay lengths benefit from a precise, up-to-date set of beamspot parameters. Additionally, online feedback is sent to the LHC operators in real time. The measurement is performed by an algorithm running on the Level 2 trigger farm, leveraging the high rate of usable events. Dedicated algorithms perform a full scan of the silicon detector to reconstruct event vertices from registered tracks. The distribution of these vertices is aggregated across the farm and their shape is extracted through fits every 60 seconds to determine the beamspot position, size, and tilt. The reconstructed beam values are corrected for detector resolution effects, measured in situ using the separation of vertices whose tracks have been split into two collections. Furthermore, measurements for individual bunch crossings have allowed for studies of single-bunch distributions as well as the behavior of bunch trains. This talk will cover the constraints imposed by the online environment and describe how these measurements are accomplished with the given resources. The algorithm tasks must be completed within the time constraints of the Level 2 trigger, with limited CPU and bandwidth allocations. This places an emphasis on efficient algorithm design and the minimization of data requests.

21

### DRUID: Event Display for the Linear Collider Detector

**Author:** Manqi Ruan<sup>1</sup>

<sup>1</sup> *Laboratoire Leprince-Ringuet (LLR)-Ecole Polytechnique*

**Corresponding Author:** manqi.ruan@llr.in2p3.fr

Based on the ROOT TEve/TGeo classes and the standard Linear Collider data format (LCIO), a general linear collider event display has been developed. It supports the latest detector models for both the International Linear Collider (ILC) and Compact Linear Collider (CLIC) as well as test beam prototypes. It can be used to visualise various informations at the generation, simulation and reconstruction levels. Many options are provided, allowing an intuitive interface to the display of different levels of information. It has been heavily employed in a variety of analyses.

**Monday 05th - Data Analysis – Algorithms and Tools / 22**

## Fractal dimension analysis in a highly granular calorimeter

**Author:** Manqi Ruan<sup>1</sup>

<sup>1</sup> *Laboratoire Leprince-Ringuet (LLR)-Ecole Polytechnique*

**Corresponding Author:** manqi.ruan@llr.in2p3.fr

The concept of “particle flow” has been developed to optimise jet energy resolution by best separating the different components of hadronic jets. A highly granular calorimetry is mandatory and provides an unprecedented level of detail in the reconstruction of showers. This enables new approaches to shower analysis. Here the measurement and use of showers’ fractal dimension is described.

The fractal dimension is a characteristic number that measures the global density of the shower. This property is highly dependent on the type of interaction and the particle energy. Its use in identifying particles and estimating their energy is described in the context of the semi-digital hadron calorimeter for the ILLD concept (International Large Detector for the International Linear Collider)

**Monday 05th - Computing Technology for Physics Research / 23**

## The AAL project: Automated monitoring and intelligent Analysis for the ATLAS data taking infrastructure

**Author:** Luca Magnoni<sup>1</sup>

**Co-authors:** Andrei Kazarov<sup>2</sup>; Giovanna Lehmann Miotto<sup>1</sup>

<sup>1</sup> *Conseil Europeen Recherche Nucl. (CERN)*

<sup>2</sup> *St. Petersburg, INP*

**Corresponding Author:** luca.magnoni@cern.ch

The Trigger and Data Acquisition (TDAQ) system of the ATLAS experiment at CERN is the infrastructure responsible for filtering and transferring ATLAS experimental data from detectors to the mass storage system. It relies on a large, distributed computing environment, including thousands of computing nodes with thousands of application running concurrently.

In such a complex environment, information analysis is fundamental for controlling applications behavior, error reporting and operational monitoring. During data taking runs, streams of messages sent by applications via the message reporting system together with data published from applications via information services are the main sources of knowledge about correctness of running operations. The huge flow of data produced (with an average rate of O(1-10KHz)) is constantly monitored by experts to detect problem or misbehavior. This require strong competence and experience in understanding and discovering problems and root causes, and often the meaningful information is not in the single message or update, but in the aggregated behavior in a certain time-line.

The AAL project is meant at reducing the man power needs and at assuring a constant high quality

of problem detection by automating most of the monitoring tasks and providing real-time correlation of data-taking and system metrics.

This project combines technologies coming from different disciplines, in particular it leverages on an Event Driven Architecture to unify the flow of data from the ATLAS infrastructure, on a Complex Event Processing (CEP) engine for correlation of events and on a machine learning module to detect anomaly and problems that cannot be defined in advance.

The project is composed of 3 main components: a core processing engine, responsible for correlation of events through expert-defined queries, a machine learning module to detect anomalies in an unsupervised manner and a web based front-end to present real-time information and interact with the system. All components works in a loose-coupled event based architecture, with a message broker to centralize all communication between modules.

The result is an intelligent system able to extract and compute relevant information from the flow of operational data to provide real-time feedback to human experts who can promptly react when needed. The paper presents the design and implementation of the AAL project, together with the results of its usage as automated monitoring assistant for the ATLAS data taking infrastructure.

**Monday 05th - Data Analysis – Algorithms and Tools / 24**

## Alignment of the ATLAS Inner Detector

**Authors:** Anthony Morley<sup>1</sup>; Jike Wang<sup>2</sup>; Pawel Bruckman De Renstrom<sup>3</sup>; Salvador Marti I Garcia<sup>4</sup>

<sup>1</sup> CERN

<sup>2</sup> High Energy Group-Institute of Physics-Academia Sinica

<sup>3</sup> Polish Academy of Sciences (PL)

<sup>4</sup> Universidad de Valencia

**Corresponding Authors:** jike.wang@cern.ch, martis@ific.uv.es

Atlas is a multipurpose experiment that records the LHC collisions. In order to reconstruct the trajectories of charged particles, ATLAS is equipped with a tracking system built using distinct technologies: silicon planar sensors (both pixel and microstrips) and drift-tubes (the Inner Detector). The tracking system is embedded in a 2 T solenoidal field. In order to reach the track parameter accuracy requested by the physics goals of the experiment, the ATLAS tracking system requires to determine accurately its almost 700,000 degrees of freedom. The demanded precision for the alignment of the silicon sensors is below 10 micrometers.

The implementation of the track based alignment within the ATLAS software framework unifies different alignment approaches and allows the alignment of all tracking subsystems together. The alignment software counts of course on the tracking information (track-hit residuals) but also includes the capability to set constraints on the beam spot and primary vertex for the global positioning, plus constraints on the track parameters as the momentum measured by the Muon System or the E/p using the calorimetry information. The assembly survey data can be used as constraint to the alignment corrections.

The alignment chain starts at the trigger level where a stream of high pT and isolated tracks is selected online. Also a cosmic ray trigger is enabled while ATLAS is recording collision data, but only during those short intervals where there are no LHC beams inside ATLAS. Thus a stream of cosmic-ray tracks is recorded exactly with the same detector operating conditions as the normal collision tracks.

As the alignment algorithms are based on the minimization of the track-hit residuals, one needs to solve a linear system with large number of degrees of freedom. The solving involves the inversion or diagonalization of a large matrix that may be dense. The alignment jobs can be executed either at the CERN Analysis Facility or using the GRID infrastructure. The event processing is run in parallel in many jobs (for both collision data and cosmic ray tracks). Then all output matrices and vectors are added together before the linear algebra solving. The alignment procedure can also be run either offline (to reprocess old data) or quasi-online at the Tier0 in the calibration loop. With the latter alignment constants are computed before the bulk reconstruction of the ATLAS data.

We will present results of the alignment of the ATLAS tracker using the 2011 collision data. The validation of the alignment is performed first using its own observables (track-hit residuals) as well as using many other physics observables, notably the resonance invariant masses in a wide energy



range ( $K_0$ s,  $J/\psi$  and  $Z$  decays in to  $\mu+\mu^-$ ) and the effect of the detector systematic distortions on the reconstructed invariant mass and on the  $\mu$  momentum. Also the electrons  $E/p$  has been studied mainly in the  $W \rightarrow e\nu$  channel. The results of the alignment with real data reveals that the attained precision for the alignment parameters is approximately 5 micrometers.

25

## Automated Quality Monitoring and Validation of the CMS Reconstruction Software

**Author:** Danilo Piparo<sup>1</sup>

<sup>1</sup> *Conseil Europeen Recherche Nucl. (CERN)*

**Corresponding Author:** danilo.piparo@cern.ch

A crucial component of the CMS Software is the reconstruction, which translates the signals coming from the detector's readout electronics into concrete physics objects such as leptons, photons and jets. Given its relevance for all physics analyses, the behaviour and quality of the reconstruction code must be carefully monitored. In particular, the compatibility of its outputs between subsequent releases and the impact of the usage of new algorithms must be carefully assessed. The automated procedure adopted by CMS to accomplish this ambitious task and the innovative tools developed for that purpose are presented. The whole chain of steps is illustrated, starting from the application testing over large ensembles of datasets emulating Tier-0, Tier-1 and Tier-2 environments, to the collection of the produced physical quantities in the form of several hundred thousand histograms, to the estimation of their compatibility between releases, to the final production and publication of reports characterised by an efficient representation of the information.

**Monday 05th - Data Analysis – Algorithms and Tools / 27**

## Visual Physics Analysis (VISPA) - From Desktop Towards Physics Analysis at Your Fingertips

**Authors:** Andreas Hinzmann<sup>1</sup>; Dennis Klingebiel<sup>1</sup>; Gero Müller<sup>1</sup>; Hans-Peter Bretz<sup>1</sup>; Jan Steggemann<sup>1</sup>; Joschka Lingemann<sup>1</sup>; Marcel Rieger<sup>1</sup>; Martin Erdmann<sup>1</sup>; Matthias Komm<sup>1</sup>; Robert Fischer<sup>1</sup>; Tobias Winchen<sup>1</sup>

<sup>1</sup> *RWTH Aachen University, III. Physikalisches Institut A*

Visual Physics Analysis (VISPA) is an analysis development environment with applications in high energy as well as astroparticle physics. VISPA provides a graphical steering of the analysis flow, which is comprised of self-written C++ and Python modules. The advances presented in this talk extend the scope from prototyping to the execution of analyses. A novel concept of analysis layers has been integrated in VISPA. On top of a base layer, it is possible to derive additional layers in which options are adjustable and modules can be activated or deactivated. This enables the creation of different stages already within the design phase of a single analysis, e.g. the event selection and the statistical analysis, or the optimization of settings for different types of input data such as electrons and muons which are to be processed within the same analysis flow. Furthermore, analysis execution in VISPA has been extended to include a graphical interface for parameter sets that are handled within a back-end independent design. This allows for direct job submission from VISPA to local computing clusters as well as to the LHC Computing Grid.

31

## Track reconstruction and b-jet identification for the ATLAS trigger system

**Author:** Andrea Coccaro<sup>1</sup>

<sup>1</sup> *Sezione di Genova (INFN)-Sezione di Genova (INFN)-Universita e*

**Corresponding Author:** andrea.coccaro@cern.ch

A sophisticated trigger system, capable of real-time track and vertex reconstruction, is in place in the ATLAS experiment, to reject most of the events containing uninteresting background collisions while preserving as much as possible the interesting physics signals. In this contribution we present the strategy adopted by the ATLAS collaboration for fast reconstruction of charged tracks and vertexing in the trigger framework. Their application to different online selections and in particular to the b-jet selection is discussed. Performance is reviewed with data from the 2011 LHC running period and particular emphasis is given to the new challenges of the 2011 data-taking campaign, where a high number of interaction per bunch crossing occurs, and to the b-jet trigger plans to enhance the ATLAS physics potential for 2011 and 2012.

**Tuesday 06th - Computing Technology for Physics Research / 32**

## Monitoring the Grid at local, national, and global levels

**Author:** Peter Gronbech<sup>1</sup>

<sup>1</sup> *Particle Physics-University of Oxford*

**Corresponding Author:** p.gronbech1@physics.ox.ac.uk

Monitoring the Grid at local, national, and global levels  
The GridPP Collaboration

The World-wide LHC Computing Grid is the computing infrastructure setup to process the experimental data coming from the experiments at the Large Hadron Collider located at CERN. GridPP is the project that provides the UK part of this infrastructure across 19 sites in the UK. To ensure that these large computational resources are available and reliable requires many different monitoring systems. These range from local site monitoring of, for example, the hardware and of batch system utilization, to UK-wide monitoring of Grid functionality and ultimately the worldwide monitoring of resource provision and usage. In this paper we describe the monitoring systems used for the many different aspects of the system, and how some of them are being integrated together. Local site monitoring covers, cluster load, batch system utilization, network bandwidth monitoring and fault condition monitoring. The most common software used to monitor a cluster is Ganglia, this system can be easily installed on all clients allowing data to be collected on a master node and displayed via a web server. Monitoring specific to the batch system used at a site is also typically used. Many GridPP sites use the torque batch system (developed from PBS). This can be monitored with pbswebmon, which provides a graphical way to monitor the occupancy of the cluster, and the different user's job shares and efficiencies. Another tool is Nagios, which provides a very powerful framework that can be used to monitor the status of systems. The Nagios system can be configured to run tests at intervals and carry out actions dependant on the results. This can be emailing a warning message or running an event handler that takes remedial action to solve a problem. One of the advantages of Nagios is that if all is well it does not bother you and there is no need to actually look at a status Web page. It can let you know (via email, web or SMS) when there is a problem. Network health, usage and bandwidth is monitored at many sites with cacti and/ or Network Weathermap. Available bandwidth between sites in the UK is monitored by each site having a dedicated 'Gridmon' test box that performs a matrix of iperf and other tests between the UK sites. The results are stored on a central database with a web frontend. Other UK wide testing includes a GridPP developed summation of relevant WLCG tests coupled with dedicated UK tests developed by Prof. S. Lloyd at QMUL and the UK regional Nagios based Service Availability Monitoring (SAM). This service queries a central database (GOCDB) and Grid information services to create a list of sites and systems to be tested. The services offered are tested

and the results of the tests are sent via an active MQ message bus to the EGI Central Operations Dashboard. Each region has an operator on duty that can raise alarm tickets against sites that have failed critical tests.

Systems Administrators are often overwhelmed by the number of different web sites and monitoring systems they should track. Attempts to integrate output from several systems into a site dashboard have been made at the Tier 1 and some of the larger sites. These systems will be described.

**Monday 05th - Computing Technology for Physics Research / 33**

## **Dynamic deployment of a PROOF-based analysis facility for the ALICE experiment over virtual machines using PoD and OpenNebula**

**Authors:** Berzano Dario<sup>1</sup>; Brunetti Riccardo<sup>2</sup>; Lusso Stefano<sup>3</sup>; Stefano Bagnasco<sup>2</sup>

<sup>1</sup> *Sezione di Torino (INFN)-Universita e INFN*

<sup>2</sup> *INFN Torino (INFN)*

<sup>3</sup> *INFN Torino*

**Corresponding Author:** [dario.berzano@to.infn.it](mailto:dario.berzano@to.infn.it)

The conversion of existing computing centres to cloud facilities is becoming popular also because of a more optimal usage of existing resources. Inside a medium to large cloud facility, many specific virtual computing facilities might concur for the same resources based on their usage and destination elastically, i.e. by expanding or reducing allocated resources for currently running VMs, or by turning them on and off. In the ALICE experiment PROOF, a parallel processing infrastructure, has become very popular for interactive analysis. The locality of PROOF-based analysis facilities forces site admins to scavenge enough resources to dedicate, yet the chaotic nature of user-written analysis tasks would deem these resources to be unstable and used intensively only in small bursts typically during working hours, making PROOF a typical use-case for HPC cloud computing. Currently, a solution named PROOF-on-Demand (PoD) does exist to dynamically and quickly provide a PROOF-enabled cluster by enqueueing agents to a job scheduler. In a medium-sized computing centre, namely a Tier-2, sharing a queue between PROOF and ordinary Grid jobs is not viable due to the very large time to wait in order to get enough workers ready: however, an elastic cloud approach will enable existing machines currently running Grid jobs to temporarily make room for many personal PoD-provided PROOF clusters on the same hardware in near-real-time, with no stability issues for long-running Grid jobs, through the perfect sandboxing intrinsically offered by virtual machines. We will show a usable prototype of a dynamically-deployed PROOF-based analysis facility by using existing tools, such as PoD and OpenNebula, orchestrated by a simple and lightweight control daemon.

34

## **Mass production of air showers using Grid for Auger**

**Author:** Julio Lozano-Bahilo<sup>1</sup>

<sup>1</sup> *Universidad de Granada*

**Corresponding Author:** [julio@ugr.es](mailto:julio@ugr.es)

The Pierre Auger Collaboration studies ultra high energy cosmic rays which induce extensive air showers when they interact at the top of the atmosphere. The generation of simulated showers involves tracking billions of particles as the shower develops through the atmosphere. The CPU time consumption of the complete simulation of a single shower is enormous but there are techniques to reduce it without losing much physical accuracy.

The high demand of computing power has driven our collaboration to adopt Grid technology for the production of our official simulations. A framework based on a set of scripts has been developed to automate all the tasks related to the handling of jobs on Grid. Additional tools allow us to monitor frequently our productions and react to changes on the Grid infrastructure. In this way we have been able to sustain a good job rate which allowed us to generate massive amounts of air showers.

**Thursday 08 - Data Analysis – Algorithms and Tools / 35**

## **Unparametrized multi-dimensional kernel density- and likelihood ratio estimator**

**Author:** Peter Kovesarki<sup>1</sup>

**Co-authors:** Adriana Elizabeth Nuncio Quiroz<sup>1</sup>; Ian C. Brock<sup>1</sup>

<sup>1</sup> *Physikalisches Institut-Universitaet Bonn*

A novel method to estimate probability density functions, suitable for multivariate analyses will be presented. The implemented algorithm can work on relatively large samples, iteratively finding a non-parametric density function with adaptive kernels. With increasing number of sample points the resulting function converges to the real probability density. Specifically, we discuss a classification example, showing the optimal separation of signal and background events based on likelihood ratios. Unlike traditional classification methods, such as neural networks, this method is free from classical overtraining effects. Furthermore, as it is possible to calculate likelihood ratios depending on signal and background cross section, the method is suitable for small signal searches at LHC.

**Thursday 08 - Data Analysis – Algorithms and Tools / 36**

## **Continuous simulation of Beyond-Standard-Model processes with multiple parameters**

**Author:** Jiahang Zhong<sup>1</sup>

**Co-author:** Shih-Chang Lee<sup>1</sup>

<sup>1</sup> *Institute of Physics-Academia Sinica*

**Corresponding Author:** [jiahang.zhong@cern.ch](mailto:jiahang.zhong@cern.ch)

We present a new approach to simulate Beyond-Standard-Model (BSM) processes which are defined by multiple parameters. In contrast to the traditional grid-scan method where a large number of events are simulated at each point of a sparse grid in the parameter space, this new approach simulates only a few events at each of a selected number of points distributed randomly over the whole parameter space. In subsequent analysis, we rely on the fitting by the Bayesian Neural Network (BNN) technique to obtain accurate estimation of the acceptance distribution. With this new approach, the signal yield can be estimated continuously, while the required number of simulation events is greatly reduced.

**Thursday 08 - Data Analysis – Algorithms and Tools / 37**

## **A Linear Iterative Unfolding Method**

**Author:** Andras Laszlo<sup>1</sup>

<sup>1</sup> *CERN, Geneva (on leave of absence from KFKI Research Institute for Particle and Nuclear Physics, Budapest)*

**Corresponding Author:** andras.laszlo@cern.ch

A frequently faced task in experimental physics is to measure the probability distribution of some quantity. Often this quantity to be measured is smeared by a non-ideal detector response or by some physical process. The procedure of removing this smearing effect from the measured distribution is called unfolding, and is a delicate problem in signal processing. Due to the numerical ill-posedness of this task, various methods were invented which, given some assumptions on the initial probability distribution, try to regularize the problem. Most of these methods definitely introduce bias on the estimate of the initial probability distribution. We propose a linear iterative method (motivated by the Neumann series known in functional analysis), which has the advantage that no assumptions on the initial probability distribution is needed. Since it is a linear scheme, statistical error propagation can be performed in an exact manner. Convergence is proved under certain quite general conditions, and in that case the method can be seen to be asymptotically unbiased. On the other hand, as a price, the approach is relatively statistics demanding. We provide a numerical C and C++ library for the implementation of the method.

**Monday 05th - Computations in Theoretical Physics / 38**

## **SecDec: a tool for numerical multi-loop/leg calculations**

**Author:** Jonathon Carter<sup>1</sup>

**Co-author:** Gudrun Heinrich<sup>2</sup>

<sup>1</sup> *University of Durham*

<sup>2</sup> *Max Planck Institute for Physics*

**Corresponding Authors:** j.p.carter@durham.ac.uk, gudrun@mpp.mpg.de

Sector decomposition is a method to extract singularities from multi-dimensional polynomial parameter integrals in a universal way. Integrals of this type arise in perturbative higher order calculations in multi-loop integrals as well as in phase space integrals involving unresolved massless particles.

The program 'SecDec' will be presented, which applies iterated sector decomposition in an automated way, to produce a Laurent series in the regularisation parameter. The coefficients of this series are finite parameter integrals which are integrated numerically by Monte Carlo techniques.

The power of the program is illustrated by presenting results and timings for a number of cutting edge multi-loop integrals, e.g. 2-loop box integrals entering top quark pair production at NNLO or 4-loop propagators. Applications to integrals occurring in calculations of real radiation at higher perturbative orders will also be presented.

**Monday 05th - Computations in Theoretical Physics / 39**

## **Three-Loop Calculation of the Higgs Boson Mass in Supersymmetry**

**Author:** Philipp Kant<sup>1</sup>

**Co-authors:** Luminita Mihaila<sup>2</sup>; Matthias Steinhauser<sup>2</sup>; Robert V. Harlander<sup>3</sup>

<sup>1</sup> *Humboldt-Universität zu Berlin*

<sup>2</sup> *KIT*

<sup>3</sup> *Bergische Universität Wuppertal*

**Corresponding Author:** philipp.kant@physik.hu-berlin.de

A Key feature of the minimal supersymmetric extension of the Standard Model (MSSM) is the existence of a light Higgs boson, the mass of which is not a free parameter but an observable that can be predicted from the theory. Given that the LHC is able to measure the mass of a light Higgs with very good accuracy, a lot of effort has been put into a precise theoretical prediction.

We present a calculation of the SUSY-QCD corrections to this observable to three-loop order. We perform multiple asymptotic expansions in order to deal with the multi-scale three-loop diagrams, making heavy use of computer algebra and keeping a keen eye on the numerical error introduced.

We provide a computer code in the form of a Mathematica package that combines our three-loop SUSY-QCD calculation with the literature of one- and two-loop corrections to the Higgs mass, providing a state-of-the-art prediction for this important observable.

**Monday 05th - Computations in Theoretical Physics / 40**

## FormCalc 7

**Author:** Thomas Hahn<sup>1</sup>

<sup>1</sup> *MPI f. Physik*

**Corresponding Author:** hahn@feynarts.de

The talk presents the new features in FormCalc 7 (and some in LoopTools), such as analytic tensor reduction, inclusion of the OPP method, and the interface to FeynHiggs.

**Tuesday 06th - Computing Technology for Physics Research / 41**

## A Validation System for Data Preservation in HEP

**Author:** Yves Kemp<sup>1</sup>

**Co-authors:** Hermann Heßling<sup>2</sup>; Marco Strutz<sup>2</sup>

<sup>1</sup> *Deutsches Elektronen-Synchrotron (DESY)*

<sup>2</sup> *HTW Berlin*

**Corresponding Author:** yves.kemp@cern.ch

Preserving data from past experiments and preserving the ability to perform analysis with old data is of growing importance in many

domains of science, including High Energy Physics (HEP). A study group on this issue, DPHEP, has been established in this field to provide guidelines and a structure for international collaboration on data preservation projects in HEP.

This contribution presents a framework that allows experimentalists to validate their software against a previously defined set of tests in an automated way. The framework has been designed with a special focus for longevity, as it makes use of open protocols, has a modular design and is based on simple communication mechanisms. On the fabrics side, tests are carried out in a virtual environment using a cloud infrastructure. Within the framework, it is easy to run validation tests on different hardware platforms, or different major or minor versions of operating systems. Experts from IT or the experiments can automatically detect failures in the test procedure by the help of reporting tools. Hence, appropriate actions can be taken in a timely manner. The design and important implementation aspects of the framework are shown and first experiences from early-bird-users will be presented.

42

## LHCb distributed computing operations

**Author:** Federico Stagni<sup>1</sup>

**Co-author:** Roberto Santinelli<sup>1</sup>

<sup>1</sup> *Conseil Europeen Recherche Nucl. (CERN)*

**Corresponding Author:** federico.stagni@cern.ch

The proliferation of tools for monitoring both activities and infrastructure, together with the pressing need for prompt reaction in case of problems impacting data taking, data reconstruction, data reprocessing and user analysis brought to the need of better organizing the huge amount of information available. The monitoring system for the LHCb Grid Computing relies on many heterogeneous and independent sources of information offering different views for a better understanding of problems while an operations team and defined procedures have been put in place to handle them. This work summarizes the state-of-the-art of LHCb Grid operations emphasizing the reasons that brought to various choices and what are the tools currently in use to run our daily activities. We highlight the most common problems experienced across years of activities on the WLCG infrastructure, the services with their criticality, the procedures in place, the relevant metrics and the tools available and the ones still missing.

**Monday 05th - Data Analysis – Algorithms and Tools / 43**

## GELATIO - The GERDA framework for digital signal analysis

**Authors:** Luciano Pandola<sup>1</sup>; Matteo Agostini<sup>2</sup>; Paolo Zavarise<sup>3</sup>

<sup>1</sup> *INFN LNGS*

<sup>2</sup> *Munich Technical University*

<sup>3</sup> *INFN LNGS, Aquila University*

**Corresponding Author:** matteo.agostini@ph.tum.de

We present the concept, the implementation and the performance of a new software framework developed to provide a flexible and user-friendly environment for advanced analysis and processing of digital signals. The software has been designed to handle the full data analysis flow of GERDA, a

low-background experiment which searches for the neutrinoless double beta decay of Ge-76 by using high-purity germanium detectors at the INFN Gran Sasso underground Laboratory. The framework organizes the data into a multi-tier structure, from the raw traces of the Ge detectors up to the condensed analysis parameters, and includes tools and utilities to handle the data stream between the different tiers. It supports a multi-channel modular and flexible analysis, widely customizable by the user either via human-readable initialization files or via a graphical interface. The framework is designed to be solid, maintainable over a long lifetime and scalable to the future phases of the experiment. To ensure flexibility and good computational performances, the framework includes both compiled and interpreted code (C++, Python and Bash). It relies upon ROOT and its extension TAM, which provides compatibility with PROOF, enabling the software to run in parallel on clusters of computers or multi-core machines. The software was tested on different platforms and benchmarked in several GERDA-related applications. A stable version is presently available for the collaboration and it is used to provide the reference analysis of the GERDA data. A few applications of the framework to real GERDA data are presented and discussed.

**Thursday 08th - Computations in Theoretical Physics / 44**

## Numerical evaluation of one-loop QCD amplitudes

**Author:** Benedikt Biedermann<sup>1</sup>

**Co-authors:** Peter Uwer<sup>1</sup>; Simon Badger<sup>2</sup>

<sup>1</sup> *Humboldt Universität zu Berlin*

<sup>2</sup> *Niels Bohr Institute, Copenhagen*

**Corresponding Author:** benedikt.biedermann@physik.hu-berlin.de

We present the publicly available program NGLUON allowing the numerical evaluation of colour-ordered amplitudes at one-loop order in massless QCD.

The program allows the evaluation of one-loop amplitudes for an arbitrary number of gluons. We discuss in detail the speed as well as the numerical stability. In addition the packages allows the evaluation of one-loop scattering amplitudes using extended floating point precision.

Furthermore we discuss the extension to one-loop amplitudes including massless quarks and show some phenomenological applications.

**Wednesday 07th - Morning session / 45**

## Feynman integrals, polylogarithms and symbols

**Author:** Vittorio Del Duca<sup>1</sup>

<sup>1</sup> *Laboratori Nazionali di Frascati (INFN)*

**Corresponding Author:** vittorio.del.duca@cern.ch

We suppose that a solution to a given Feynman integral is known in terms of multiple polylogarithms, and address the question of how to find another solution which is equivalent to the former, but with a simpler analytic structure.

**Wednesday 07th - Morning session / 46**



## Strange Bedfellows: Quantum Mechanics and Data Mining

**Author:** Marvin Weinstein<sup>1</sup>

<sup>1</sup> *SLAC National Accelerator Laboratory*

**Corresponding Author:** niv@slac.stanford.edu

All fields of scientific research have experienced an explosion of data. Analyzing this data to extract unexpected patterns presents a computational challenge that requires new, advanced methods of analysis. DQC (Dynamic Quantum Clustering), invented by David Horn (Tel Aviv University), is a novel, interactive and highly visual approach to this problem. Studies are already underway at SLAC to apply this technology to, among other things, discovering hard-to-find events in particle physics data, analyzing Fermi/Glast data and implementing large scale SSRL XAF studies of the in-situ chemistry of macroscopic heterogeneous samples. The method has also been applied to problems in medicine, bio-informatics and even the stock market. My talk will provide a brief introduction to the distinction between supervised and unsupervised methods in data mining (clustering in particular). Then, I will, very briefly, discuss the theory of DQC and show a simple application. Finally I will review some of the problems that have been studied to date. This part of the discussion will, as an aside, present a very simple visualization technique that makes it possible to see very small features in two-dimensional data (think Dalitz plots).

**Tuesday 06th - Computing Technology for Physics Research / 47**

## The LHCb DIRAC-based production and data management operations systems

**Authors:** Federico Stagni<sup>1</sup>; Philippe Charpentier<sup>1</sup>

<sup>1</sup> *Conseil Europeen Recherche Nucl. (CERN)*

**Corresponding Authors:** federico.stagni@cern.ch, philippe.charpentier@cern.ch

The LHCb computing model was designed in order to support the LHCb physics program, taking into account LHCb specificities (event sizes, processing times etc...). Within this model several key activities are defined, the most important of which are real data processing (reconstruction, stripping and streaming, group and user analysis), Monte-Carlo simulation and data replication. In this contribution we detail how these activities are managed by the LHCbDIRAC Data Transformation System. The LHCbDIRAC Data Transformation System leverages the workload and data management capabilities provided by DIRAC, a generic community grid solution, to support data-driven workflows (or DAGs). The ability to combine workload and data tasks within a single DAG allows to create highly sophisticated workflows with the individual steps linked by the availability of data. This approach also provides the advantage of a single point at which all activities can be monitored and controlled. To highlight the versatility of the system we present in more detail experience with real data of the 2010 and 2011 LHC run.

While several interfaces are currently supported (including python API and CLI), we will present the ability to create LHCb workflows through a secure web interface, control their state in addition to creating and submitting jobs. To highlight the versatility of the system we present in more detail experience with real data of the 2010 and 2011 LHC run.

48

## DIRAC Framework for Distributed Computing Systems

**Author:** Andrei Tsaregorodtsev<sup>1</sup>

<sup>1</sup> *Centre de Physique de Particules de Marseille (CPPM)-Faculte de*

**Corresponding Author:** atsareg@in2p3.fr

Many modern applications need large amounts of computing resources both for calculations and data storage. These resources are typically found in the computing grids but also in commercial clouds and computing clusters. Various user communities have access to different types of resources. The DIRAC project provides a solution for an easy aggregation of heterogeneous computing resources for a given user community. It helps also to organize the work of the users by applying policies regulating the usage of common resources. DIRAC was initially developed for the LHCb Collaboration - large High Energy Physics experiment on the LHC accelerator at CERN, Geneva. The project now offers a generic platform for building distributed computing systems. The design principles, architecture and main characteristics of the DIRAC software will be described and examples of its usage by various user communities will be presented.

**Tuesday 06th - Computing Technology for Physics Research / 49**

## **Can 'Go' address the multicore issues of today and the manycore problems of tomorrow ?**

**Author:** Sebastien Binet<sup>1</sup>

<sup>1</sup> *Laboratoire de l'Accelérateur Lineaire (LAL)-Universite de Pari*

**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.

While parallel programming - still in an intensive phase of R&D despite the 30+ years of literature on the subject - is an obvious topic to consider, other issues (build scalability, code clarity, code deployment and ease of coding) are worth investigating when preparing for the manycore era. Moreover, if one wants to use another language than C++, a language better prepared and tailored for expressing concurrency, one also needs to ensure a good and easy reuse of already field-proven libraries.

We present the work resulting from such investigations applied to the 'Go' programming language. We first introduce the concurrent programming facilities 'Go' is providing and how its module system addresses the build scalability and dependency hell issues.

We then describe the process of leveraging the many (wo)man-years put into scientific Fortran/C/C++ libraries and making them available to the Go ecosystem.

The ROOT data analysis framework, the C-BLAS library and the Herwig-6 MonteCarlo generator will be taken as examples.

Finally, performances of a small analysis written in Go and using Fortran and C++ libraries will be discussed.

references:

Go: <http://golang.org>

ROOT: <http://root.cern.ch>

C-BLAS: <http://www.netlib.org/clapack/cblas/>

Herwig-6: <http://hepwww.rl.ac.uk/theory/seymour/herwig/>

50

## A CMake-based build and configuration framework

**Author:** Marco Clemencic<sup>1</sup>

**Co-authors:** Hubert Degaudenzi<sup>2</sup>; Pere Mato<sup>1</sup>

<sup>1</sup> CERN

<sup>2</sup> EPFL

**Corresponding Author:** marco.clemencic@cern.ch

The LHCb experiment has been using the CMT build and configuration tool for its software since the first versions, mainly because of its multi-platform build support and its powerful configuration management functionality. Still, CMT has some limitations in terms of build performance and the increased complexity added to the tool to cope with new use cases added latterly. Therefore, we have been looking for a viable alternative to it and we have investigated the possibility of adopting the CMake tool, which does a very good job for building and is getting very popular in the HEP community. The result of this study is a CMake-based framework which provides most of the special configuration features available natively only in CMT, with the advantages of better performances, flexibility and portability.

**Tuesday 06th - Computing Technology for Physics Research / 51**

## Track finding using GPUs

**Author:** Christian Schmitt<sup>1</sup>

<sup>1</sup> Institut fuer Physik-Johannes-Gutenberg-Universitaet Mainz

**Corresponding Author:** christian.schmitt@cern.ch

The reconstruction and simulation of collision events is a major task in modern HEP experiments involving several ten thousands of standard CPUs. On the other hand the graphics processors (GPUs) have become much more powerful and are by far outperforming the standard CPUs in terms of floating point operations due to their massive parallel approach. The usage of these GPUs could therefore significantly reduce the overall reconstruction time per event or allow for the usage of more sophisticated algorithms.

In this contribution the track finding in the ATLAS experiment will be used as an example on how the GPUs can be used in this context: the seed finding alone shows already a speed increase of one order of magnitude compared to the same implementation on a standard CPU. On the other hand the implementation on the GPU requires a change in the algorithmic flow to allow the code to work in the rather limited environment on the GPU in terms of memory, cache, and transfer speed from and to the GPU.

**Thursday 08th - Computations in Theoretical Physics / 52**

## One-loop integrations with Hypergeometric functions

**Author:** Toshiaki Kaneko<sup>1</sup>

<sup>1</sup> KEK**Corresponding Author:** toshiaki.kaneko@kek.jp

Numerically stable analytic expression of a one-loop integration is one of the most important elements of the accurate calculations of one-loop corrections to the physical processes. It is known that these integrations are expressed by some generalized classes of Gauss hypergeometric functions. Power series expansions, differential equations, contiguous and many other identities are known for them. For Lauricella  $F_D$  functions, analytic properties are studied in detail, which provide useful information for the numerical stabilities.

We show that two- and three-point functions are exactly expressed in terms of  $F_D$  for arbitrary combinations of mass parameters in any space-time dimensions. We also show the relation between four-point functions and Aomoto-Gelfand hypergeometric functions.

**Tuesday 06th - Data Analysis – Algorithms and Tools / 54**

## Gibbs sampler for background discrimination in particle physics

**Author:** Federico Colechia<sup>1</sup><sup>1</sup> *University College London***Corresponding Author:** federico@hep.ucl.ac.uk

Background properties in experimental particle physics are typically estimated from large collections of events. This usually provides precise knowledge of average background distributions, but inevitably hides fluctuations. To overcome this limitation, an approach based on statistical mixture model decomposition is presented. Events are treated as heterogeneous populations comprising particles originating from different processes, and individual particles are mapped to a process of interest on a probabilistic basis. When used to discriminate against background, the proposed technique based on the Gibbs sampler allows some features of the background distributions to be estimated directly from the data without training on high-statistics samples. A feasibility study on Monte Carlo is presented, together with a comparison with existing techniques. Finally, the prospects for the development of the Gibbs sampler into a tool for intensive offline analysis of interesting events at the Large Hadron Collider are discussed.

**Tuesday 06th - Computing Technology for Physics Research / 55**

## Challenges in using GPUs for the reconstruction of digital hologram images.

**Author:** Ivan D Reid<sup>1</sup>**Co-authors:** Jindrich Nebrensky<sup>1</sup>; Peter R Hobson<sup>1</sup><sup>1</sup> *Brunel University***Corresponding Authors:** peter.hobson@brunel.ac.uk, ivan.reid@cern.ch

In-line holography has recently made the transition from silver-halide based recording media, with laser reconstruction, to recording with large-area pixel detectors and computer-based reconstruction. This form of holographic imaging is used for small particulates, such as cloud or fuel droplets, marine plankton and alluvial sediments, and enables a true 3D object field to be recorded at high resolution over a considerable depth. To reconstruct a digital hologram a 2D FFT must be calculated for every depth slice desired in the replayed image volume. A typical hologram of ~100 micrometre particles over a depth of a few hundred millimetres will require  $O(1000)$  2D FFT operations to be performed on an hologram of typically a few million pixels. With the growing use of video-rate recording and the desire to reconstruct fully every frame the computational challenge becomes considerable. In previous work (<http://bura.brunel.ac.uk/handle/2438/2823>) we have reported on our experiences with reconstruction on a computational grid. In this paper we discuss the technical challenges in converting our reconstruction code to make efficient use of the NVIDIA CUDA based GPU cards and show how near real-time video slice reconstruction can be obtained with holograms as large as 4K by 4K pixels. We also discuss the issues surrounding the reconstruction of holograms which are larger than 50% of the GPU memory where a different approach to reconstruction will be needed. Finally we consider the implications for grid and cloud computing, and the extent to which GPU can replace these approaches, when the important step of locating focussed objects within a reconstructed volume is included.

56

## Non-parametric comparison of histogrammed two-dimensional data distributions using the Energy Test.

**Author:** Ivan D Reid<sup>1</sup>

**Co-authors:** Peter R Hobson<sup>1</sup>; Raul H C Lopes<sup>1</sup>

<sup>1</sup> *Brunel University*

**Corresponding Author:** [ivan.reid@cern.ch](mailto:ivan.reid@cern.ch)

When monitoring complex experiments, comparison is often made between regularly acquired histograms of data and reference histograms which represent the ideal state of the equipment. With the larger HEP experiments now ramping up, there is a need for automation of this task since the volume of comparisons would overwhelm human operators. However, the two-dimensional histogram comparison tools currently available in ROOT have been noted in the past to exhibit shortcomings (<http://bura.brunel.ac.uk/handle/2438/2763>). We discuss a newer comparison test for 2D histograms, based on the Energy Test of Aslan and Zech (<http://arxiv.org/abs/hep-ex/0203010>), which provides more decisive discrimination between histograms of data coming from different distributions, and compare it with recent ROOT releases.

**Tuesday 06th - Computing Technology for Physics Research / 58**

## Offloading peak processing to Virtual Farm by STAR experiment at RHIC

**Author:** jan balewski<sup>1</sup>

<sup>1</sup> *MIT*

**Corresponding Author:** [balewski@mit.edu](mailto:balewski@mit.edu)

In recent years, Cloud computing has become a very attractive “notion” and popular model for accessing distributed resources and has emerged as the next big trend after the so-called Grid computing approach. The onsite STAR computing resources amounting to about 3000 CPU slots have

been extended by additional 1000 slots using opportunistic resources from pilot DOE/Magellan and DOE/Nimbus projects. The Virtual Machine (VM) framework was used to assemble the STAR-computing environment, which is independent on specific hardware. STAR VM was validated once, deployed on over 100 8-core VMs at NERSC and Argon National Lab, and used as homogenous Virtual Farm processing in real time events acquired by STAR detector located at Brookhaven National Lab. To provide time dependent calibration constants to the large number of isolated VMs, a database snapshot scheme was devised and used for this exercise. It allows periodic synchronization of VM DB with the master DB without the need for frequent DB client connections to the master DB from multiple jobs running on every VM. The two high capacity disks localized at the opposite coasts of US and interconnected via Globus-Online protocol were used in this setup, which resulted with highly scalable Cloud-based extension of STAR computing resources. The STAR Virtual Farm scaled up between February and May of 2011 from 160 to 1300 CPU slots. It has been used to process fraction of events STAR in real time and later to reanalyze past STAR events to providing key arguments for changing the course of ongoing STAR data taking

59

## Data Preservation in High Energy Physics

**Author:** Roman Kogler<sup>1</sup>

**Co-authors:** Cristinel Diaconu<sup>2</sup>; David South<sup>1</sup>

<sup>1</sup> DESY

<sup>2</sup> CPPM / DESY

**Corresponding Authors:** roman.kogler@cern.ch, david.south@cern.ch

Data from high-energy physics experiments are collected with significant financial and human effort and are mostly unique.

However, until recently no coherent strategy existed for data preservation and re-use, and many important and complex data sets have simply been lost. While the current focus is on the LHC at CERN, in the current period several important and unique experimental programs at other facilities are coming to an end, including those at HERA, b-factories and the Tevatron. To address this issue, an inter-experimental study group on HEP data preservation and long-term analysis (DPHEP) was convened at the end of 2008. The group now aims to publish a full and detailed review of the present status of data preservation in high energy physics. This contribution summarises the results of the DPHEP study group, describing the challenges of data preservation in high energy physics and the group's first conclusions and recommendations. The physics motivation for data preservation, generic computing and preservation models, technological expectations and governance aspects at local and international levels are examined.

**Monday 05th - Data Analysis – Algorithms and Tools / 60**

## 10 Years of Object-Oriented Analysis on H1

**Author:** Paul Laycock<sup>1</sup>

**Co-author:** David South<sup>2</sup>

<sup>1</sup> *University of Liverpool*

<sup>2</sup> *DESY*

**Corresponding Authors:** paul.james.laycock@cern.ch, david.south@cern.ch

Over a decade ago, the H1 Collaboration decided to embrace the object-oriented paradigm and completely redesign its data analysis model and data storage format. The event data model, based on the RooT framework, consists of three layers - tracks and calorimeter clusters, identified particles and finally event summary data - with a singleton class providing unified access. This original solution was then augmented with a fourth layer containing user-defined objects.

This contribution will summarise the history of the solutions used, from modifications to the original design, to the evolution of the high-level end-user analysis object framework which is used by H1 today. Several important issues are addressed - the portability of expert knowledge to increase the efficiency of data analysis, the flexibility of the framework to incorporate new analyses, the performance and ease of use, and lessons learned for future projects.

**Monday 05th - Computing Technology for Physics Research / 61**

## **Advances in Service and Operations for ATLAS Data Management**

**Authors:** Graeme Andrew Stewart<sup>1</sup>; Vincent Garonne<sup>1</sup>

<sup>1</sup> *CERN*

**Corresponding Author:** graeme.andrew.stewart@cern.ch

ATLAS has recorded almost 5PB of RAW data since the LHC started running at the end of 2009. Many more derived data products and complimentary simulation data have also been produced by the collaboration and, in total, 55PB is currently stored in the Worldwide LHC Computing Grid by ATLAS. All of this data is managed by the ATLAS Distributed Data Management system, called Don Quixote 2 (DQ2).

DQ2 has evolved rapidly to help ATLAS Computing operations to manage these large quantities of data across the many grid sites at which ATLAS runs and to help ATLAS physicists get access to this data. In this paper we describe new and improved DQ2 services:

- Popularity service, which measures usage of data across ATLAS.
- Space monitoring and accounting at sites.
- Automated blacklisting service.
- Cleaning agents, which trigger deletion of unused data at sites.
- Deletion agents, to reliably delete unwanted data from sites.

We describe the experience of data management operation in ATLAS computing, showing how these services enable management of petabyte scale computing operations.

We illustrate the coupling of data management services to other parts of the ATLAS computing infrastructure, in particular showing how feedback from the distributed analysis system in ATLAS has enabled

dynamic placement of the most popular data, helping users and groups to analyse the increasing data volumes on the grid.

**Thursday 08 - Data Analysis – Algorithms and Tools / 63**

## **An Alternative Method for Tilecal Signal Detection and Amplitude Estimation**

**Author:** Peralva Sotto-Maior<sup>1</sup>

**Co-author:** ATLAS Tile Calorimeter group

<sup>1</sup> *Universidade Federal do Rio de Janeiro (UFRJ)*

The Barrel Hadronic calorimeter of ATLAS (Tilecal) is a detector used in the reconstruction of hadrons, jets, muons and missing transverse energy from the proton-proton collisions at the Large Hadron Collider (LHC). It comprises 10,000 channels in four readout partitions and each calorimeter cell is made of two readout channels for redundancy. The energy deposited by the particles produced in the collisions is read out by the several readout channels and its value is estimated by an optimal filtering algorithm, which reconstructs the amplitude and the time of the digitized signal pulse sampled every 25 ns.

This work deals with signal detection and amplitude estimation for the Tilecal under low signal-to-noise ratio (SNR) conditions. It explores the applicability (at the cell level) of a Matched Filter (MF), which is known to be the optimal signal detector in terms of the SNR. Moreover, it investigates the impact of signal detection when summing both signals from the same cell before estimating the amplitude, instead of performing it afterwards as it is currently done. The signal of interest is electronically conditioned to have a well-defined shape (the Tilecal reference pulse shape) and the electronic noise distribution is a Gaussian-like, for which decorrelation can be handled by estimating the whitening transformation of the process. As a result, the MF method implements a finite impulse response (FIR) filter whose coefficients are the Tilecal reference pulse shape.

The MF method is compared to the Optimal Filter (OF) algorithm currently implemented in the Tilecal DSP, which performs the signal reconstruction online. To this end, two classes of data have been used: the noise dataset, which comprises noise signals taken from a pedestal run during nominal Tilecal operation, and the signal dataset, which is constructed from Tilecal reference pulse shape in convolution with added noise. In order to simulate realistic conditions, amplitude and time shifting distributions were taken into account to generate the signal dataset. The results showed that for conditions where the signal pedestal could be considered stationary, the MF filter technique achieves a better SNR performance compared to the OF technique for the tested simulated signals. Current studies include analyzing the behavior of the MF method in conditions where the signal pulse is distorted by the pile-up from additional interactions to the primary collision.

66

## **Development of noSQL data storage for the ATLAS PanDA Monitoring System**

**Author:** Maxim Potekhin<sup>1</sup>

<sup>1</sup> *Brookhaven National Laboratory*

**Corresponding Author:** maxim.potekhin@cern.ch

For several years the PanDA Workload Management System has been the basis for distributed production and analysis for the ATLAS experiment at the LHC. Since the start of data taking PanDA usage has ramped up steadily, typically exceeding 500k completed jobs per day by June 2011. The associated monitoring data



volume has been rising as well, to levels that present a new set of challenges in the areas of database scalability and monitoring system performance and efficiency. These challenges are being met with a R&D effort aimed at implementing a scalable and efficient monitoring data storage based on a noSQL solution (Cassandra). We present our motivations for using this technology, as well as data design and the techniques used for efficient indexing of the data. We also discuss the hardware requirements as they were determined by testing with actual data and realistic loads.

**Thursday 08 - Data Analysis – Algorithms and Tools / 69**

## **An adaptive Monte-Carlo Markov chain algorithm for counting muons in Auger water Cherenkov detector signals**

**Authors:** Balázs Kégl<sup>1</sup>; Rémi Bardenet<sup>1</sup>

<sup>1</sup> *Linear Accelerator Laboratory*

**Corresponding Author:** balazs.kegl@gmail.com

Adaptive Metropolis (AM) is a powerful recent algorithmic tool in numerical Bayesian data analysis. AM builds on a well-known Markov Chain Monte Carlo (MCMC) algorithm but optimizes the rate of convergence to the target distribution by automatically tuning the design parameters of the algorithm on the fly. In our data analysis problem of counting muons in the water Cherenkov signal of the surface detectors in the Pierre Auger Experiment, the signal is modeled by a mixture distribution. Label switching is a major problem in inference on such models because of the invariance to symmetries. The simplest (non-adaptive) solution is to modify the prior in order to make it select a single permutation of the variables, introducing an identifiability constraint. This solution is known to cause artificial biases by not respecting the topology of the posterior. In this paper we describe a new online relabeling procedure which can be incorporated into the AM algorithm. We state the convergence of the algorithm and identify the link between its modified target measure and the original posterior distribution of interest.

Our long-term goal in the Pierre Auger Experiment is to develop a comprehensive generative model for the surface detector signal and use MCMC techniques to estimate the parameters. The first step of this program is the development of a generative model of the response of an Auger water tank and an adaptive reversible jump MCMC algorithm that can deal with the unknown number of muonic components in the signal. In the second part of this paper we discuss the algorithmic and computational issues of implementing MCMC techniques for large-scale data analysis.

**Tuesday 06th - Computations in Theoretical Physics / 70**

## **Reweighting NNPDFs.**

**Author:** Francesco Cerutti<sup>1</sup>

<sup>1</sup> *Universitat de Barcelona*

**Corresponding Author:** francesco.cerutti@gmail.com

I present a method, elaborated within the NNPDF Collaboration, that allows the inclusion of the information contained in new datasets into an existing set of parton distribution functions without the need for refitting.

The method exploits bayesian inference in the space of PDF replicas, computing for each replica a

chisquare with respect to the new dataset and a weight associated to this. These weights are then applied to the ensemble of parton densities, producing a reweighted set of replicas. The reweighting method may be used to assess the impact of any new data or pseudodata on parton densities and thus on their predictions.

**Tuesday 06th - Computing Technology for Physics Research / 71**

## **PROOF Performance Measurements Using PROOF Benchmark Suite**

**Author:** Sangsu Ryu<sup>1</sup>

**Co-author:** Gerardo Ganis<sup>2</sup>

<sup>1</sup> *KiSTi Korea Institute of Science & Technology Information (KiS*

<sup>2</sup> *CERN*

**Corresponding Authors:** sangsu.ryu@cern.ch, gerardo.ganis@cern.ch

PROOF (Parallel ROOT Facility) is an extension of ROOT enabling interactive analysis in parallel on clusters of computers or a many-core machine. PROOF has been adopted and successfully utilized as one of main analysis models by LHC experiments including ALICE and ATLAS. ALICE has seen growing number of PROOF clusters around the world, CAF at CERN, SKAF in Slovakia, GSIAF at Darmstadt being the main ALICE PROOF service farms. KIAF at KISTI is also planning on PROOF farm service in 2011.

The PROOF benchmark suite is a new utility suite of PROOF to measure the performance and scalability of PROOF. The primary goal of benchmark suite is to determine the optimal configuration parameters for a set of machines to be used as PROOF cluster.

The suite measures the performance of the cluster for a set of standard tasks, CPU-intensive task and IO-intensive task which are 2 distinctive styles of analysis in typical HEP application, as a function of the number of effective processes. From these results, indications about the optimal number of concurrent processes can be derived. For large facilities, the suite should also give indications about the optimal number of sub-masters into which the cluster should be partitioned.

Site administrators of PROOF cluster can use the suite to measure the performance of the cluster and optimize the configuration of their cluster. PROOF developers can also utilize the suite to help them measure, identify problems and improve their software.

Performance of PROOF cluster measured with the benchmark suite will be presented including real use cases at ALICE experiment.

**Thursday 08th - Computations in Theoretical Physics / 72**

## **Progress on the Direct Computation Method**

**Author:** Fukuko Yuasa<sup>1</sup>

**Co-authors:** Elise de Doncker<sup>2</sup>; Kiyoshi Kato<sup>3</sup>; Nobuyuki Hamaguchi<sup>1</sup>; Tadashi Ishikawa<sup>1</sup>; Yoshimitsu Shimizu<sup>1</sup>

<sup>1</sup> *KEK*

<sup>2</sup> *Western Michigan University*

<sup>3</sup> *Kogakuin University*

**Corresponding Author:** fukuko.yuasa@kek.jp

We report our progress on the development of the Direct Computation Method (DCM), which is a fully numerical method for the computation of Feynman diagrams.

Based on a combination of a numerical integration tool and a numerical extrapolation technique, all steps in the computation are carried out in a fully numerical way. The combined method is applicable to one-, two- and multi-loop diagrams with arbitrary masses including complex masses.

In this talk we show numerical results of a scalar one-loop pentagon and hexagon without any analytical treatment, neither reducing to a sum of box diagrams nor sector decomposition. Further we discuss the possibility of handling ultraviolet divergence using DCM.

**Monday 05th - Computations in Theoretical Physics / 73**

## Regularization of IR-divergent loop integrals

**Author:** Elise de Doncker<sup>1</sup>

**Co-authors:** Fukuko Yuasa<sup>2</sup>; Yoshimasa Kurihara<sup>2</sup>

<sup>1</sup> *Western Michigan University*

<sup>2</sup> *High Energy Accelerator Research Organization (KEK), Japan*

**Corresponding Author:** elise.dedoncker@wmich.edu

We report results of a new regularization technique for infrared (IR) divergent loop integrals using dimensional regularization, where a positive regularization parameter (epsilon, satisfying that the dimension  $d = 4 + 2 \cdot \text{epsilon}$ ) is introduced in the integrand to keep the integral from diverging as long as  $\text{epsilon} > 0$ .

Based on an asymptotic expansion of the integral we construct a linear system of equations, which incorporates values of the integral for varying epsilon in the right hand side of the system. The linear system is extended by one equation at a time for decreasing epsilon, and solved for the leading coefficients of the Laurent expansion of the integral. This gives rise to an extrapolation as epsilon tends to zero. The solutions can be obtained by solving the systems directly or by a recursive method.

We will outline the computations and the evaluation of the integrals for various problems. An analysis involves the condition and truncation error of the method. All computations are kept numerical and performed with automatic code, including a possible reduction of the integral to a form without entangled singularities.

The basic technique can be applied to IR divergent integrals without (threshold) singularities in the interior of the domain. For non-IR divergent integrals with threshold singularities, the same method reduces to a linear extrapolation for a calculation of the integral. We outline an extension of the technique for integrals which have both types of singularities by resorting to a double extrapolation or regularization.

**Monday 05th - Computing Technology for Physics Research / 75**

## The EOS disk storage system at CERN

**Author:** Andreas Joachim Peters<sup>1</sup>

<sup>1</sup> *CERN*

**Corresponding Author:** andreas.joachim.peters@cern.ch

EOS was designed to fulfill generic requirements on disk storage scalability and IO scheduling performance for LHC analysis use cases following the strategy to decouple disk and tape storage as individual storage systems.

The project was setup in April 2010. Since October 2010 EOS was evaluated by ATLAS as a disk only storage pool at CERN for analysis use cases in the context of various WLCG demonstrator projects.

Since May 2011 analysis data has been migrated to the EOSCMS and EOSATLAS production instances. Each instance contains several thousand disks and provides few petabytes of storage capacity individually managed by EOS.

In this paper we summarize features available in the first release version of EOS and highlight some of the benefits as a user analysis disk pool in comparison with other storage solutions.

In the second part we will describe the current deployment and operation model of EOS in the CERN computer centre and its usage by the CMS and ATLAS experiments. We will conclude with a roadmap and future directions of EOS development and operations at CERN.

**Tuesday 06th - Computations in Theoretical Physics / 76**

## **Self-Organizing Maps Parametrization of Deep Inelastic Structure Functions with Error Determination**

**Author:** simonetta liuti<sup>1</sup>

<sup>1</sup> *university of virginia*

**Corresponding Author:** sl4y@virginia.edu

We will present a method to extract parton distribution functions from hard scattering processes based on an alternative type of neural networks, the Self-Organizing Maps (SOMs). Quantitative results including a detailed treatment of uncertainties will be presented within a Next to Leading Order analysis of both unpolarized and polarized inclusive deep inelastic scattering data. With a fully working procedure in hand, we are capable to extend our analysis to the Generalized Parton Distribution (GPD) case, thus exploiting the “classification” and “visualization” properties of the SOMs.

Work supported by US D.O.E. grant DE-FG02-01ER41200. We thank for computer time the University of Virginia Alliance for Computational Science and Engineering, and the HPC group at Jefferson Lab.

**Monday 05th - Computations in Theoretical Physics / 77**

## **Different forms of the generalized Crewther relation in QCD and QED: concrete consequences of analytical multiloop calculations**

**Author:** Andrei Kataev<sup>1</sup>

<sup>1</sup> *INR, Moscow, Russia*

**Corresponding Author:** kataev@ms2.inr.ac.ru

Different forms of the generalized Crewther relation in QED and QCD are discussed. They follow from application of the method of OPE to the AVV triangle amplitude in the limit when conformal symmetry is valid and broken by the procedure of renormalizations in the various variants of MS scheme, including 't Hooft prescription for defining beta-function. Special features of the consequences of the advanced  $\alpha_s^4$ -order analytical calculations of the Bjorken polarized sum rule and non-singlet

contribution to the Adler D-function are discussed. The results of application of conformal symmetry and the original Crewther relation for getting QED-type analytical contributions to the Ellis-Jaffe sum rule in the 4-th order of PT is also demonstrated.

**Tuesday 06th - Data Analysis – Algorithms and Tools / 78**

## **Semi-Supervised Anomaly Detection - Towards Model-Independent Searches of New Physics**

**Authors:** Eric Malmi<sup>1</sup>; Mikael Kuusela<sup>2</sup>; Tommi Vatanen<sup>1</sup>

**Co-authors:** Risto Orava<sup>2</sup>; Tapani Raiko<sup>1</sup>

<sup>1</sup> *Aalto University*

<sup>2</sup> *Helsinki Institute of Physics (HIP)*

**Corresponding Author:** mikael.kuusela@cern.ch

Most classification algorithms used in high energy physics fall under the category of supervised machine learning. Such methods require a training set containing both signal and background events and are prone to classification errors should this training data be systematically inaccurate for example due to the assumed MC model. To complement such model-dependent searches, we propose an algorithm based on anomaly detection techniques, which does not require a MC training sample for the signal data. We first model the MC background using multivariate mixtures of Gaussians. We then search for deviations from the background model by fitting to the observations a mixture of the background model and a number of additional Gaussians using a variant of the EM algorithm. This allows us to perform pattern recognition of any excess over the background. We show by comparison to neural networks that such a semi-supervised approach is a lot more robust against misspecification of the signal MC than supervised classification. In cases where there is an unexpected signal, a neural network fails to correctly identify it while anomaly detection does not suffer from such a limitation. On the other hand, when there are no systematic errors in the signal MC, both methods perform comparably. Due to its fully probabilistic nature, the anomaly detection model has a number of additional advantages as well. Firstly, the mixing proportion of the anomalous excess immediately gives an estimate for its cross section and secondly, the statistical significance of the excess can easily be estimated using a bootstrapping-based likelihood-ratio test.

82

## **Comparison of the CPU and memory performance of StatPatternRecognition (SPR) and Toolkit for MultiVariate Analysis (TMVA)**

**Author:** Giulio Palombo<sup>1</sup>

<sup>1</sup> *California Institute of Technology*

**Corresponding Author:** giulio@justanswer.com

High Energy Physics data sets are often characterized by a huge number of events. Therefore, it is extremely important to use statistical packages able to efficiently analyze these unprecedented amounts of data. We compare the performance of the statistical packages StatPatternRecognition (SPR) and Toolkit for MultiVariate Analysis (TMVA). We focus on how CPU time and memory usage of the learning process scale versus data set size. As classifiers, we consider Random Forests, Boosted Decision Trees and Neural Networks. For our tests, we employ a data set widely used in the machine learning community, “Threenorm” data set, as well as data tailored for testing various edge cases. For each data set, we constantly increase its size and check CPU time and memory needed to build the classifiers implemented in SPR and TMVA.

We show that SPR is often significantly faster and consumes significantly less memory. For example, the SPR implementation of Random Forest is by an order of magnitude faster and consumes an order of magnitude less memory than TMVA on Threernorm data.

**Thursday 08th - Computing Technology for Physics Research / 83**

## **Multicore in Production: Advantages and Limits of the Multi-process Approach.**

**Authors:** Paolo Calafiura<sup>1</sup>; Vakhtang Tsulaia<sup>1</sup>

**Co-authors:** Andy Washbrook<sup>2</sup>; Charles Leggett<sup>1</sup>; David Lesny<sup>3</sup>; Douglas Smith<sup>4</sup>; Horst Severini<sup>5</sup>; Manoj Kumar Jha<sup>6</sup>; Mous Tatarckhanov<sup>1</sup>; Peter van Gemmeren<sup>7</sup>; Scott Snyder<sup>8</sup>; Sebastien Binet<sup>9</sup>; Wim Lavrijsen<sup>1</sup>

<sup>1</sup> *LBL*

<sup>2</sup> *Univ of Edinburgh*

<sup>3</sup> *Univ of Illinois Urbana-Champaign*

<sup>4</sup> *SLAC*

<sup>5</sup> *Univ of Oklahoma*

<sup>6</sup> *INFN Napoli*

<sup>7</sup> *ANL*

<sup>8</sup> *BNL*

<sup>9</sup> *LAL*

**Corresponding Authors:** vakhtang.tsulaia@cern.ch, pcalafiura@lbl.gov

The shared memory architecture of multicore CPUs provides HENP developers with the opportunity to reduce the memory footprint of their applications by sharing memory pages between the cores in a processor. ATLAS pioneered the multi-process approach to parallelizing HENP applications. Using Linux fork() and the Copy On Write mechanism we implemented a simple event task farm which allows to share up to 50% memory pages among event worker processes with negligible CPU overhead.

By leaving the task of managing shared memory pages to the operating system, we have been able to run in parallel large reconstruction and simulation applications originally written to be run in a single thread of execution with little to no change to the application code. In spite of this, the process of validating athena multi-process for production took ten months of concentrated effort and is expected to continue for several more months. In general terms, we had two classes of problems in the multi-process port: merging the output files produced by the event workers, and assuring the reproducibility of the results, especially of Montecarlo simulations, when running with different configurations, in particular with different number of event workers.

Besides validating the software itself, an important and time-consuming aspect of running multicore applications in production is to configure the production system to handle multicore jobs. This entails defining multicore batch queues, where the unit resource is not a core, but a whole computing node; monitoring the output of many event workers; and adapting the job definition layer to handle computing resources with very different event throughputs (depending on the number of cores used).

To conclude, we will present scalability and memory usage studies, based on data gathered both on dedicated hardware and on ATLAS production nodes. From these it should become apparent that the most promising development to improve performance will be to transition from a simple, flat, event task farm in which all processes handle events independently to a task farm with specialized worker processes, which will be in charge of event I/O. This approach will further reduce the memory footprint of our multicore applications, and at the same time address the issue of merging event worker outputs, at the cost of some increase in the complexity of the ATLAS core software.

Monday 05th - Morning session / 84

## The toolbox of modern multi-loop computations: novel analytic and semi-analytic techniques

**Author:** Alexey Pak<sup>1</sup>

<sup>1</sup> *TTP KIT Karlsruhe*

**Corresponding Author:** apak@particle.uni-karlsruhe.de

After a short introduction, sketching the structure of a typical calculation of higher-order quantum corrections, I will discuss a few examples illustrating ideas that were instrumental in obtaining some recent novel results. Attention will be given to the tools facilitating those techniques and the technical challenges. In particular, the talk will cover the progress in sector decomposition method, gluing relations, and dimensional recurrence relations. Finally, I will mention some very promising theoretical developments in understanding the mathematical structure of Feynman integrals that are yet to yield new results.

Thursday 08th - Computations in Theoretical Physics / 85

## One-loop tensor Feynman integral reduction with signed minors

**Authors:** Edward Glover<sup>1</sup>; Tord Riemann<sup>2</sup>

**Co-authors:** Jochem Fleischer<sup>3</sup>; Valery Yundin<sup>4</sup>

<sup>1</sup> *University of Durham (GB)*

<sup>2</sup> *DESY*

<sup>3</sup> *Univ. Bielefeld*

<sup>4</sup> *Univ. Katowice*

**Corresponding Author:** tord.riemann@desy.de

The algebraic tensor reduction of one-loop Feynman integrals with signed minors has been further developed.

There is now available the

C++ package PJFry by V. Yundin for the reduction of 5-point 1-loop tensor integrals up to rank 5.

Special care is devoted to vanishing or small Gram determinants.

Further, we derived

extremely compact expressions for the contractions of the tensor integrals with external momenta.

They are based on sums over signed minors weighted with scalar products of the external momenta.

Tuesday 06th - Data Analysis – Algorithms and Tools / 86

## Status of TMVA, the toolkit for multivariate analysis

**Authors:** Andreas Hoecker<sup>1</sup>; Eckhard von Toerne<sup>2</sup>; Helge Voss<sup>3</sup>; Jan Therhaag<sup>2</sup>; Joerg Stelzer<sup>4</sup>; Peter Speckmayer<sup>1</sup>

<sup>1</sup> *CERN*

<sup>2</sup> *University of Bonn*

<sup>3</sup> *MPI Heidelberg*

<sup>4</sup> *Michigan State University*

**Corresponding Author:** evt@physik.uni-bonn.de

The toolkit for multivariate analysis, TMVA, provides a large set of advanced multivariate analysis techniques for signal/background classification and regression problems. These techniques are embedded in a framework capable of handling input data preprocessing and the evaluation of the results, thus providing a simple and convenient tool for multivariate techniques. The analysis techniques implemented in TMVA can be easily invoked and the direct comparison of their performance allows the user to choose the most appropriate for a particular data analysis. This talk presents recently developed features, such as improved preprocessing, option tuning and an extended unit test framework to ensure code stability. We also discuss the performance of our most important multivariate techniques on example data and a comparison with theoretical performance limits.

**Thursday 08 - Data Analysis – Algorithms and Tools / 87**

## Online Particle Detection by Neural Networks Based on Topologic Calorimetry Information

**Author:** Thiago Ciodaro Xavier<sup>1</sup>

**Co-authors:** Denis Damazio<sup>2</sup>; Dhiana Deva<sup>1</sup>; José Manoel de Seixas<sup>1</sup>

<sup>1</sup> *Univ. Federal do Rio de Janeiro (UFRJ)*

<sup>2</sup> *Brookhaven National Lab. BNL*

**Corresponding Author:** seixas@lps.ufrj.br

Electrons and photons are among the most important signatures in ATLAS. Their identification against jets background by the online trigger system relies very much on calorimetry information. The ATLAS online trigger comprises three cascaded levels and the Ringer is an alternative set of algorithms that uses calorimetry information for electron detection at the second trigger level (L2). It is split into two parts: the feature extraction algorithm (FEX), which represents particle interaction as a set of concentric ring sums, and the hypothesis test (HYPO), which implements a multilayer perceptron neural network to perform final particle identification. The neural network may also be used to implement a Fisher discriminant, in case linear processing is desired in this stage.

The Ringer FEX starts by searching the most energetic cell (hot cell) in each calorimeter layer from the Region of Interest (RoI) previously selected by the ATLAS level-1 trigger. The hot cell energy becomes the first ring and it is also considered the center of all further rings, which are formed as the sum of the energies from the outer cells of the inner ring. A total of 100 rings are computed. The Ringer HYPO normalizes the ring values in order to fit them to the neural network dynamic range. After propagating the rings through the network, a single output node provides the incoming event classification.

Optimizations, guided by detailed time performance analysis, were made at the Ringer algorithm core, in order to make it prepared for operation in ATLAS. Studies showed that the execution time was improved by a factor of 50, while its payload necessary to store the Ringer information represents only 1.2% of the present total HLT amount. Also, Monte Carlo simulations of 14 TeV proton-proton collisions at  $2 \times 10^{34}$  luminosity were used to evaluate the Ringer performance over pile-up.

**Thursday 08th - Computations in Theoretical Physics / 89**

## GPU Linear algebra extensions for GNU/Octave

**Author:** Attilio Santocchia<sup>1</sup>



**Co-authors:** Leone Bosi <sup>2</sup>; Mirko Mariotti <sup>3</sup>

<sup>1</sup> *Universita e INFN Perugia*

<sup>2</sup> *INFN Perugia*

<sup>3</sup> *Dipartimento di Fisica Perugia*

**Corresponding Author:** attilio.santocchia@cern.ch

Octave is one of the most used open source tools for numerical analysis and linear algebra. Our project wants to improve Octave introducing the support for GPU computing, in order to speed up some linear algebra operations. The core of our work is a C library that executes on GPU some BLAS operations concerning vector-vector, vector-matrix and matrix-matrix functions. OpenCL functions are used to program GPU kernels, which are bound within the GNU/octave framework. We report the project implementation design and some preliminary results about performances.

**Thursday 08th - Computing Technology for Physics Research / 90**

## **An Exploration of SciDB in the Context of Emerging Technologies for Data Stores in Particle Physics and Cosmology**

**Authors:** David Malon<sup>1</sup>; Jack Weinstein<sup>2</sup>; Peter van Gemmeren<sup>2</sup>

<sup>1</sup> *High Energy Physics Division-Argonne National Laboratory (ANL)*

<sup>2</sup> *Argonne National Laboratory*

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

Traditional relational databases have not always been well matched to the needs of data-intensive sciences, but efforts are underway within the database community to attempt to address many of the requirements of large-scale scientific data management. One such effort is the open-source project SciDB. Since its earliest incarnations, SciDB has been designed for scalability in parallel and distributed environments, with a particular emphasis upon native support for array constructs and operations. Such scalability is of course a requirement of any strategy for large-scale scientific data handling, and array constructs are certainly useful in many contexts, but these features alone do not suffice to qualify a database product as an appropriate technology for hosting particle physics or cosmology data. In what constitutes its 1.0 release in June 2011, SciDB has extended its feature set to address additional requirements of scientific data, with support for user-defined types and functions, for data versioning, and more.

This paper describes an evaluation of the capabilities of SciDB for two very different kinds of physics data: event-level metadata records from proton collisions at the Large Hadron Collider, and the output of cosmological simulations run on very-large-scale supercomputers. This evaluation exercises the spectrum of SciDB capabilities in a suite of tests that aim to be representative and realistic, including, for example, definition of four-vector data types and natural operations thereon, and computational queries that match the natural use cases for these data.

**Thursday 08th - Computing Technology for Physics Research / 91**

## **Lessons from Static Analysis on HEP Software**

**Author:** Axel Naumann<sup>1</sup>

<sup>1</sup> *CERN*

**Corresponding Author:** axel.naumann@cern.ch

Coverity's static analysis tool has been run on most of the LHC experiments' frameworks, as well as several of the packages provided to them (e.g. ROOT, Geant4). I will present how static analysis works and why it is complimentary to dynamic checkers like valgrind or test suites; typical issues discovered by static analysis; and lessons learned.

**Thursday 08th - Computing Technology for Physics Research / 92**

## **Moving ROOT Forward.**

**Author:** Fons Rademakers<sup>1</sup>

<sup>1</sup> *CERN*

**Corresponding Author:** fons.rademakers@cern.ch

Now that the LHC has started the LHC experiments crave for stability in ROOT, however progress in computing technology is not stopping and to keep ROOT up to date and compatible with new technologies requires a lot of work. In this presentation we will show what we are currently working on and what new technologies we try to exploit.

**Tuesday 06th - Morning session / 94**

## **Multivariate analysis and data mining: statistics in the computer age**

**Author:** David Hand<sup>1</sup>

<sup>1</sup> *Imperial College London*

**Corresponding Author:** d.j.hand@imperial.ac.uk

For very sound reasons, including the central limit theorem and mathematical tractability, classical multivariate statistics was heavily based on the multivariate normal distribution. However, the development of powerful computers, as well as increasing numbers of very large data sets, has led to a dramatic blossoming of research in this area, and the development of entirely new tools for multivariate analysis. The talk will present an overview of such developments, illustrating with ideas, tools, and methods such as empirical Bayes, false discovery rate, bootstrap methods, anomaly detection methods, and streaming data analysis.

**Thursday 08th - Computing Technology for Physics Research / 95**

## **Evaluation of likelihood functions on CPU and GPU devices**

**Authors:** Alfio Lazzaro<sup>1</sup>; Yngve Sneen Lindal<sup>2</sup>

**Co-authors:** Andrzej Nowak<sup>1</sup>; Jarp Sverre<sup>1</sup>; Julien Leduc<sup>1</sup>

<sup>1</sup> *CERN openlab*

<sup>2</sup> *Norges Teknisk-Naturvitens. Univ. (NTNU) and CERN openlab*

**Corresponding Authors:** yngve.sneen.lindal@cern.ch, alfio.lazzaro@cern.ch

In this work we present the parallel implementations of an algorithm used to evaluate the likelihood function of the data analysis. The implementations run on CPU and GPU, respectively, and both devices cooperatively (hybrid). Therefore the execution of the algorithm can take full advantage from users commodity systems, like desktops and laptops, using entirely the hardware at disposal. CPU and GPU implementations are based on OpenMP and OpenCL, respectively. For the hybrid case, we implemented a scheduler of the tasks so that the workload can be split and balanced in the two devices. Initially the scheduler determines the workloads for each device, so that the corresponding execution times are balanced. From this phase a ratio of the workloads is obtained. Then it starts the likelihood function evaluations, keeping fixed the previously determined ratio of the workloads. We show the results of the scalability when running on CPU. Then we show the comparison of the performance of the GPU implementation on different hardware systems from different vendors, and the performance when running in the hybrid case. The tests are based on likelihood functions from real data analysis carried out in the high energy physics community.

**Thursday 08th - Computing Technology for Physics Research / 96**

## **Do regions of ALICE matter? (Social relationships and data exchanges in the Grid)**

**Authors:** Costin Grigoras<sup>1</sup>; Eric Widmer<sup>2</sup>; Federico Carminati<sup>1</sup>; Gil Viry<sup>2</sup>; Giuliana Galli Carminati<sup>3</sup>

<sup>1</sup> *CERN, Geneva, Switzerland*

<sup>2</sup> *Dpt of Sociology, University of Geneva*

<sup>3</sup> *University Hospital of Geneva*

**Corresponding Author:** federico.carminati@cern.ch

Following a previous publication, this study aims at investigating the impact of regional affiliations of centres on the organisation of collaboration within the Distributed Computing ALICE infrastructure, based on social networks methods. A self-administered questionnaire was sent to all centre managers about support, email interactions and wished collaborations in the infrastructure. Several additional measures, stemming from technical observations were produced, such as bandwidth, data transfers and Internet Round Trip Time (RTT) were also included. Information for 50 centres were considered (60% response rate). Empirical analysis shows that despite the centralisation on CERN, the network is highly organised by regions. The results are discussed in the light of policy and efficiency issues.

**Monday 05th - Data Analysis – Algorithms and Tools / 97**

## **Multivariate Correlated Sampling Using Extended Alias Techniques**

**Author:** Siddhant Ajit Mohanty<sup>1</sup>

**Co-authors:** Ajit Kumar Mohanty<sup>1</sup>; Dushyant Goyal<sup>1</sup>; Federico Carminati<sup>1</sup>; Subho Sankar Banerjee<sup>1</sup>

<sup>1</sup> *CERN*

**Corresponding Author:** federico.carminati@cern.ch

Monte-Carlo technique enables one to generate random samples from distributions with known characteristics and helps to make probability based inferences of the underlying physical processes. Fast and efficient Monte-Carlo particle transport code particularly for high energy nuclear and particle physics experiments has become an important tool starting from the design and fabrication of detectors to the modelling of the physics outcome as close as the reality. Quite often Monte-Carlo simulations require multivariate random numbers to be generated from correlated data both from normal and non-normal distributions. Although several techniques exist for multivariate correlated samplings of varying degrees of success, the most elegant method is the technique that uses the principal component analysis of the given correlation matrix  $R$  for generating multivariate random numbers with specified inter-correlations. While the component analysis is suitable for multivariate normal distribution, it may not work always particularly when the distribution is non Gaussian. In this work, we propose an extended alias sampling which was originally proposed by A. J. Walker in 1977 to sample from an one dimensional distribution. This method is quite fast, efficient and reproduces the original distributions quite accurately (verified through chi-square as well co-variance test). It may be mentioned here that this method is quite robust and is applicable to all type of multivariate distribution irrespective of whether the distribution is Gaussian or Non-Gaussian. Although this method is quite general and can be applied to any dimensions, in this work we have restricted sampling only from a two dimensional correlated distribution. The motivation behind this study has been to develop a ROOT based Monte-Carlo application package for low energy neutron transport down in energy to a few keV using the evaluated nuclear data file (ENDF) which is available in ROOT format. Work is in progress to apply this new method of alias technique to the ENDF data set where the angle and energy distributions are strongly correlated.

**Thursday 08th - Computing Technology for Physics Research / 98**

## **Efficient Pseudo-Random Number Generation for Monte-Carlo Simulations Using Graphic Processors**

**Author:** Siddhant Ajit Mohanty<sup>1</sup>

**Co-authors:** Ajit Kumar Mohanty<sup>1</sup>; Federico Carminati<sup>1</sup>

<sup>1</sup> CERN

**Corresponding Author:** federico.carminati@cern.ch

The future of high power computing is evolving towards the efficient use of highly parallel computing environment. The class of devices that has been designed having parallelism features in mind is the Graphics Processing Units (GPU) which are highly parallel, multithreaded computing devices. One application where the use of massive parallelism comes instinctively is Monte-Carlo simulations where a large number of independent events have to be simulated. At the core of the Monte-Carlo simulation lies the random number generators. For GPU programming, the random number generator should have (a) good statistical properties (b) high computational speed (c) low memory use, and (d) a large period. The most commonly used Mersenne Twister generator has very good statistical properties with a long period of  $2^{19937}-1$ , but not suitable for implementation in the GPU as it has a large state that must be updated serially. Each GPU thread must have an individual state in global RAM and requires multiple access per generator. The relatively large number of computation per generated number makes the generator too slow for GPU programming except in cases where the ultimate in quality is needed. In this paper, we have used a hybrid approach as used in NVIDIA CUDA library. The suggestion is to use a combination of three Tausworthe generator with different parameters along with a simple Linear Congruential Generator (LCG) where the mod operation is not performed explicitly. The period of these combinations is quite high ( $2^{121}$ ) and has good statistical properties as the defects of one generator gets compensated by other. This hybrid generator requires four random seeds which can be supplied using a CPU-side random number generator. We have carried out alias Monte-Carlo sampling using this hybrid generator where each GPU thread is used to generate random variable in parallel. This would mean each thread needs to be provided a random seed independently. In the present work, we have implemented alias sampling with NVIDIA GeForce GTX 480 GPU card using both CUDA and OpenCL kernels. It is noticed that the kernel execution in both cases is about 1000 times faster as compared to the CPU whereas the total code execution is only 10 times faster. This is due to the fact that memory copy

from host to device or vice-versa is very slow. Therefore, we try to minimise memory access time and implement a simple scheme to generate random seed per thread on the fly from the formulae  $\text{seed} = 1099087573 * \text{id}$  where  $\text{id}$  is the thread index. This is known as quick and dirty LCG which has a period of 232 and mod operation is not explicitly needed due to overflow of unsigned integer. It is shown that this hybrid generator which takes seed on the fly is quite fast, reproduces the statistical properties reasonably well and can easily be implemented on each thread of GPU as well as CPU in an efficient way.

99

## Making distributed ALICE analysis simple using the GRID plug-in

**Authors:** Andrei Gheata<sup>1</sup>; Mihaela Gheata<sup>2</sup>

<sup>1</sup> CERN

<sup>2</sup> ISS Bucharest

**Corresponding Author:** andrei.gheata@cern.ch

The presentation will describe an interface within the ALICE analysis framework that allows transparent usage of the experiment's distributed resources. This analysis plug-in makes it possible to configure back-end specific parameters from a single interface and to run with no change the same custom user analysis in many computing environments, from local workstations to PROOF clusters or GRID resources. The tool is used now extensively in the ALICE collaboration for both end-user analysis and large scale productions.

Tuesday 06th - Data Analysis – Algorithms and Tools / 101

## Modeling Fake Missing Transverse Energy with Bayesian Neural Networks

**Authors:** Harrison Prosper<sup>1</sup>; Silvia Tentindo<sup>1</sup>

<sup>1</sup> Department of Physics-Florida State University

**Corresponding Authors:** silvia.tentindo@cern.ch, harry.prosper@cern.ch

Neural networks (NN) are universal approximators. Therefore, in principle, it should be possible to use them to model any reasonably smooth probability density such as the probability density of fake missing transverse energy (MET). The modeling of fake MET is an important experimental issue in events such as

$Z \rightarrow l^+ l^- + \text{jets}$ , which is an important background in high-mass Higgs searches at the Large Hadron Collider. We describe how Bayesian neural networks (BNN) can be used to model the MET in  $\gamma + \text{jets}$  events and how, in turn, the resulting BNN function can be used to model the missing transverse energy distribution in samples other than  $\gamma + \text{jets}$  in which the MET is largely due to instrumental effects.

Monday 05th - Computing Technology for Physics Research / 102

## One click dataset transfer: toward efficient coupling of distributed storage resources and CPUs.

**Authors:** Jérôme Lauret<sup>1</sup>; Michal Zerola<sup>2</sup>; Roman Barták<sup>3</sup>

**Co-author:** Michal Šumbera<sup>2</sup>

<sup>1</sup> *Brookhaven National Laboratory*

<sup>2</sup> *Academy of Sciences, Czech Republic*

<sup>3</sup> *Charles University*

**Corresponding Author:** zerola@matfyz.cz

The massive data processing in a multi-collaboration environment with geographically spread diverse facilities will be hardly “fair” to users and hardly using network bandwidth efficiently unless we address and deal with planning and reasoning related to data movement and placement. The needs for coordinated data resource sharing and efficient plans solving the data transfer paradigm in a dynamic way are being more required. We will present the work which purpose is to design and develop an automated planning system acting as a centralized decision making component with emphasis on optimization, coordination and load-balancing.

We will describe the most important optimization characteristic and modeling approach based on “constraints”. Constraint-based approach allows for a natural declarative formulation of what must be satisfied, without expressing how. The architecture of the system, communication between components and execution of the plan by underlying data transfer tools will be shown. We will emphasize the separation of the planner from the “executors” and explain how to keep the proper balance between being deliberative and reactive. The extension of the model covering full coupling and reasoning about computing resources will be shown.

The system has been deployed within STAR experiment over several Tier sites and has been used for data movement in the favour of user analyses or production processing. We will present several real use-case scenario and performance of the system with a comparison to the “traditional” - solved by hands methods. The benefits in terms of indispensable shorter data delivery time due to leveraging available network paths and intermediate caches will be revealed. Finally, we will outline several possible enhancements and avenues for future work.

**Monday 05th - Data Analysis – Algorithms and Tools / 104**

## **The alignment of the CMS Silicon Tracker**

**Author:** Gero Flucke<sup>1</sup>

<sup>1</sup> *DESY (Hamburg)*

**Corresponding Author:** gero.flucke@cern.ch

The CMS all-silicon tracker consists of 16588 modules. In 2010 it has been successfully aligned using tracks from cosmic rays and pp-collisions, following the time dependent movements of its innermost pixel layers. Ultimate local precision is now achieved by the determination of sensor curvatures, challenging the algorithms to determine about 200000 parameters. Remaining alignment uncertainties are dominated by systematic effects that can bias track parameters by an amount relevant for physics analyses. These effects are controlled by adding further information, e.g. the mass of decaying resonances. The orientation of the TK respect to the magnetic field of CMS is determined with a stand-alone  $\chi^2$  minimization procedure. The geometries are finally validated with several tools, the monitored quantities include the basic track quantities (for both tracks from collisions and cosmics) and physics resonances.

**Thursday 08th - Morning session / 106**

## The five dimensions of the genome

**Author:** Jacques Rougemont<sup>1</sup>

<sup>1</sup> EPFL

**Corresponding Author:** jacques.rougemont@epfl.ch

Thanks to large sequencing initiatives of the last 10 years we now have access to full genome sequences in digital form, in particular for laboratory species such as the mouse whose genome is about 3.5 billion letters in size. Recent high-throughput technologies allow to then probe the function of this genome in many different experimental conditions by sampling the genome at the rate of 2-3 billion letters per experiment, distributed with strong bias towards particular regions of the genome sharing a given biochemical property. The analysis of these large datasets is a fascinating challenge. I will illustrate this with two situations where time, space and chemical state of the DNA are inter-related: I will first present data on the circadian (24h) rhythms in the mouse liver: many biological functions must be activated synchronously at certain times of the day and are coupled to an internal (biochemical) clock within each cell. The second example comes from embryonic development, where the correct body patterning relies on a complex network of interactions within the genome and in particular on a tight control of the 3D folding of the DNA molecule within the cell's nucleus. I will show how we reconstruct such 5D configurations from the statistical analysis of the genome samples relative to the known full genome sequence, and how we can make inferences about cellular machineries from these data.

**Tuesday 06th - Computations in Theoretical Physics / 107**

## Status of parallelization of FORM

**Author:** Takahiro Ueda<sup>1</sup>

<sup>1</sup> Karlsruhe Institute of Technology

**Corresponding Author:** ueda@particle.uni-karlsruhe.de

We report on the current status of the development of parallel versions of the symbolic manipulation system FORM. Currently there are two parallel versions of the FORM: one is TFORM which is based on the POSIX threads and for running on multicore machines, and the other is ParFORM which uses the MPI and can run on computer clusters. By using these versions, most of existing FORM programs can benefit from the parallelization without any modifications.

**Wednesday 07th - Morning session / 108**

## Progress in Automated Next-to-Leading Order calculations

**Author:** Francesco Tramontano<sup>1</sup>

<sup>1</sup> CERN

**Corresponding Author:** francesco.tramontano@cern.ch

With the beginning of the experimental programs at the LHC, the need of describing multi particle scattering events with high accuracy becomes more pressing. On the theoretical side, perturbative calculation within leading order precision cannot be sufficient, therefore accounting for effects due to Next-to-Leading Order (NLO) corrections becomes mandatory. In the last few years we observed a tremendous progress in the computation of one-loop virtual corrections for processes involving many particles. The new ideas based on the universal four-dimensional decomposition for the numerator of the integrand for any one-loop scattering amplitudes, the four-dimensional unitarity-cuts, and unitarity-cuts in  $d$ -dimension, yielding the complete determination of dimensionally regulated one-loop amplitudes, give the possibility to develop automated multi-process evaluators for scattering amplitudes at NLO.

**Tuesday 06th - Morning session / 109**

## **Modern actions, algorithms, and computers for Lattice QCD**

**Author:** Peter Boyle<sup>1</sup>

<sup>1</sup> *University of Edinburgh*

**Corresponding Author:** paboyle@ph.ed.ac.uk

I discuss recently developed formulations of lattice Fermions possessing near-exact chiral symmetry. These are particularly appropriate for the simulation of complex weak matrix elements. I also discuss the state of the art of supercomputing for Lattice simulation

**Thursday 08th - Computations in Theoretical Physics / 110**

## **Automated one-loop calculations with Golem/Samurai**

**Author:** Gudrun Heinrich<sup>1</sup>

**Co-authors:** Francesco Tramontano ; Gavin Cullen ; Gionata Luisoni ; Giovanni Ossola ; Pierpaolo Mastrolia ; Thomas Reiter

<sup>1</sup> *Max Planck Institute Munich*

**Corresponding Author:** gudrun@mpp.mpg.de

A program package will be presented which aims at the automated calculation of one-loop amplitudes for multi-particle processes. The program offers the possibility to optionally use either unitarity cuts or traditional tensor reduction of Feynman diagrams, or a combination of both. It can be used to calculate one-loop corrections to both QCD and electro-weak theory. Beyond the Standard Model theories can be interfaced using FeynRules or LanHep. A standard interface to programs calculating real radiation is also included. It will further be described how the program detects and deals with numerical instabilities, and how the rational terms can be computed efficiently.

**Tuesday 06th - Computations in Theoretical Physics / 111**



## Polynomial Algebra in Form 4

**Author:** Jan Kuipers<sup>1</sup>

<sup>1</sup> *Nikhef*

New features of the symbolic algebra package Form 4 are discussed. Most importantly, these features include polynomial factorization and polynomial GCD computation. Examples of their use are shown. One of them is an exact version of Mincer which gives answers in terms of rational polynomials and 5 master integrals.

**Monday 05th - Computations in Theoretical Physics / 112**

## An analytical solution for a non-planar massive double box diagram

**Author:** Andreas von Manteuffel<sup>1</sup>

**Co-author:** Cedric Studerus<sup>2</sup>

<sup>1</sup> *University of Zurich*

<sup>2</sup> *University of Bielefeld*

An analytical calculation of a non-planar 2-loop box diagram is presented. This diagram appears in the computation of higher order corrections to top-quark pair production and contains one internal massive line. The corresponding integrals are solved with differential equation and Mellin-Barnes techniques.

**Tuesday 06th - Computations in Theoretical Physics / 113**

## Reduze 2

**Authors:** Andreas von Manteuffel<sup>1</sup>; Cedric Studerus<sup>2</sup>

<sup>1</sup> *von Manteuffel*

<sup>2</sup> *University of Bielefeld*

**Corresponding Author:** cedricstuderus@gmail.com

Reduze is a computer program for reducing Feynman Integrals to master integrals employing the Gauss/Laporta algorithm. Reduze is written in C++ and uses the GiNaC library to perform simplifications of the algebraic prefactors in the system of equations.

In this talk, the new version, Reduze 2, is presented. The program supports fully parallelised computations with MPI and allows to resume aborted reductions with the use of the Berkeley database. The user inputs are standardized with the YAML file format. Reduze 2 also provides an interface to use the computer algebra system Fermat.

**Thursday 08th - Morning session / 114**

## New approaches for numerical techniques in higher order calculations

**Author:** Michal Czakon<sup>1</sup>

<sup>1</sup> *RWTH Aachen*

**Corresponding Author:** mczakon@physik.rwth-aachen.de

It has become customary to think of higher order calculations as analytic, in the sense that the result should be presented in the form of known functions or constants. If such a result is obtained, numerical evaluation for practical applications or expansion in asymptotic regimes should not pose any problem. There are, however, many problems of interest, where the analytic structure, due to the number of involved variables, does not make it possible to express predictions through known functions. One strategy is to extend the class of functions, as for example in the case of harmonic and generalized harmonic polylogarithms. On the other hand, if the aim is to provide results quickly and with moderate effort, then there are much more efficient approaches, which involve numerical methods at earlier stages of the calculation. In this talk, I will review methods for the evaluation of virtual corrections, such as contour deformation in Feynman-parametric and Mellin-Barnes representations, as well as the method of differential equations. Subsequently, I will present recent advances in the calculation of real radiation contributions with non-analytic evaluation of integrals over the unresolved phase space.

**Tuesday 06th - Morning session / 115**

## NFS 4.1/pNFS, the final step

**Author:** Patrick Fuhrmann<sup>1</sup>

<sup>1</sup> *DESY*

**Corresponding Author:** patrick.fuhrmann@desy.de

With the introduction of clustered storage, combining a set of hosts to a single storage system, a very successful standard data access protocol, NFS2/3 became obsolete. One of the reasons was that NFS 2/3 assumes the name service part of the protocol being severed from the same host as the actual data, which is of course no longer true for clustered systems. As a result, high performance storage systems e.g. Panasas, GPFS, Lustre and many more, designed their own file system network protocols, with the obvious advantage of an extremely optimized use of the underlying network and storage resources, as the server and client software are provided by the same source. The drawbacks however were that proprietary software had to be installed on all client machines, with the hassle of kernel and driver dependencies and maintenance issues, particularly annoying when operating large compute farms. In order to catch up on that development, well-known storage providers decided to invest into a standard network file system protocol supporting clustered storage services, the Parallel Network File System (pNFS). The activity is organized by the Center for Information Technology Integration (CITI) at the University of Michigan. At the time being, all partners in this group have the NFS 4.1/pNFS server software integrated into their storage systems, however, except for dCache.org, companies seem to be reluctant making it available to customers. NFS 4.1/pNFS client drivers are available for the Linux 2.6.38 kernel and are slowly approaching standard Linux distributions. This presentation will elaborate on the advantages of NFS 4.1/pNFS as well as on the availability of the different components and possibly on missing bits and pieces. Furthermore it will provide details on the stability and performance evaluation done in the context of the European Middleware Initiative (EMI) and at dCache.org.

**Monday 05th - Morning session / 116**

## Where do we go from here? - The next phase of computing in HEP

**Author:** Sverre Jarp<sup>1</sup>

<sup>1</sup> *CERN*

**Corresponding Author:** sverre.jarp@cern.ch

The speaker will start by reviewing the dominant technologies chosen for the LHC Computing Grid and briefly discuss their suitability. He will then go on to look at technologies that have emerged since, but are not being seriously used. Some of these technologies are being or have been evaluated by the CERN openlab. In the last part of the talk the speaker will argue for the adoption of certain of these technologies for the direct benefit of the LCG/HEP community.

**Thursday 08th - Morning session / 117**

## Computing On Demand: Analysis in the Cloud

**Author:** Anar Manafov<sup>1</sup>

<sup>1</sup> *GSI - Helmholtzzentrum für Schwerionenforschung GmbH*

**Corresponding Author:** a.manafov@gsi.de

Constant changes in computational infrastructure like the current interest in Clouds, imply conditions on the design of applications. We must make sure that our analysis infrastructure, including source code and supporting tools, is ready for the on demand computing (ODC) era.

This presentation is about a new analysis concept, which is driven by users needs, completely disentangled from the computational resources, and scalable.

What does it take for an analysis code to be performed on any resource management system?

How can one achieve goals of on demand analysis, using PROOF on Demand (PoD)?

These questions and such topics as preferable location of data files as well as tools and software development techniques for on demand data analysis are covered. Also analysis implementation requirements and comparisons of traditional and “on demand” facilities will be discussed during this talk.

**Tuesday 06th - Morning session / 119**

## Using machine learning techniques in classification problems in Astrophysics

**Author:** Somak Raychaudhury<sup>1</sup>

<sup>1</sup> *University of Birmingham*

Multivariate datasets in astrophysics can be large, with the increasing volume of information now becoming available from a range of observations, from ground and Space, across the electromagnetic spectrum. The observations are in the form of raw images and/or spectra, and tables of derived quantities, obtained at multiple epochs in time. Large archives of images, spectra and catalogues are now

being assembled into publicly-available databases: one example is the emerging global effort towards the Virtual Observatory. This necessitates the development of techniques that will allow fast, automated classification and extraction of key physical properties for very large datasets, and the ability to visualise the structure of highly multi-dimensional data, for extracting and studying substructures in a flexible way. Automated algorithms for clustering and outlier detection are necessary for a wide range of Astrophysical problems involving these growing datasets.

The applicability of commercial data mining tools is limited, since these do not incorporate the handling of errors in a principled manner, which is central to the analysis of Astronomical data, as it is in other branches of Physics. I will review how techniques used in the field of machine learning are being adapted for use in classification and clustering problems. Examples will include the use of topographic mapping to classify light curves of eclipsing binary stars, showing that this is an efficient way of searching for transiting extrasolar planets in large datasets, and robust density modelling for determining clusters and outliers, resulting in finding high-redshift quasars.

**Thursday 08th - Morning session / 120**

## **SALAMI project**

**Author:** David De Roure<sup>1</sup>

<sup>1</sup> *Oxford e-Research Centre*

**Monday 05th - Morning session / 121**

## **Building an Outsourcing Ecosystem for Science**

**Author:** Kate Keahey<sup>1</sup>

<sup>1</sup> *Argonne National Laboratory*

**Corresponding Author:** keahey@mcs.anl.gov

Infrastructure-as-a-Service (IaaS) cloud computing is revolutionizing the way we acquire and manage computational and storage resources: by allowing on-demand resource leases and supporting user control over those resources it enables us to treat resource acquisition as an operational consideration rather than capital investment. The emergence of this new model raises many questions, in particular for special requirements groups such as scientific computing. Can cloud computing be used by scientific applications? Does it, or will it ever, provide sufficient capabilities for high-performance applications? How will it change our work patterns? What challenges need to be overcome, and what is its overall potential for accelerating science?

In this talk, I will give an overview of the challenges and potential of cloud computing projects in scientific community. I will describe what attracted various scientific communities to cloud computing, give examples of how they integrated this new model into their work, and describe the challenges they encountered while doing so. I will then discuss how those challenges drove the development of Nimbus Infrastructure, which allows users to provide cycle outsourcing via their clouds, as well as the Nimbus Platform, which provides ecosystem tools allowing users to leverage infrastructure cloud resources across different academic and commercial platforms ranging from proprietary (Amazon Web Services) to open source (Nimbus, OpenStack, Eucalyptus and others). I will also discuss

challenges and issues – related to performance, logistics, utilization, and privacy that need to be overcome to make the benefits of cloud computing available to an ever larger set of scientific applications. Finally, I will discuss the emerging technology trends and discuss how they can benefit science.

Bio: Kate Keahey is a Scientist in the Distributed Systems Lab at Argonne National Laboratory and a Fellow at the Computation Institute at the University of Chicago. Kate pioneered the use of cloud computing for scientific applications and created and leads the open source Nimbus project which provides an Infrastructure-as-a-Service cloud computing implementation as well as a set of higher-level services allowing users to build elastic application by combining on-demand commercial and scientific cloud resources.

**Friday 09th - Morning session / 122**

## **Summary - Computing Technology for Physics Research**

**Author:** Jerome Lauret<sup>1</sup>

<sup>1</sup> *BNL*

**Corresponding Author:** jlauret@bnl.gov

**Friday 09th - Morning session / 123**

## **Summary - Data Analysis – Algorithms and Tools**

**Corresponding Author:** pushpalatha.bhat@cern.ch

**Friday 09th - Morning session / 124**

## **Summary - Computations in Theoretical Physics – Techniques and Methods**

**Author:** Nigel Glover<sup>1</sup>

<sup>1</sup> *IPPP Durham*

**Corresponding Author:** e.w.n.glover@durham.ac.uk

**Friday 09th - Morning session / 125**

## **ACAT 2011 - Summary and outlook**

**Author:** Denis Perret-Gallix<sup>1</sup>

<sup>1</sup> *CNRS/IN2P3*

**Corresponding Author:** denis.perret-gallix@cern.ch

**Monday 05th - Morning session / 126**

**Welcome - Prof. Geoff Rodgers, Pro-Vice Chancellor for Research,  
Brunel University**