GridKa School 2013: Big Data, Clouds and Grids

Report of Contributions

https://indico.kit.edu/e/26

Contribution ID: **2** Type: **not specified**

Cloud Computing: Expanding Humanity's Limits to Planet Mars

Thursday, August 29, 2013 9:00 AM (40 minutes)

As another tool that Humanity has used for expanding its limits, cloud computing was born and evolved in consonance with the different challenges where it has been applied.

Due to its seamless provision of resources, dynamism and elasticity, this paradigm has been brought into the spotlight by the Space scientific community and in particular that devoted to the exploration of Planet Mars. This is the case of Space Agencies in need of great amounts of on demand computing resources and with a budget to take care of.

The Red Planet represents the next limit to be reached by Humanity, attracting the attention of many countries as a destination for the next generation manned spaceflights. However, theres is still much research to do on Planet Mars and many computational needs to fulfill.

This talk will review the cloud computing approach by NASA and then it will focus on the Mars MetNet Mission, with which the speaker is actively collaborating. This Mission is being put together by Finland, Russia and Spain, and aims to deploy several tens of weather stations on the Martian surface. The Atmospheric Science research is a crucial area in the exploration of the Red Planet and represents a great opportunity for harnessing and improving current computing tools, and establish interesting collaborations between countries.

Author: Dr VAZQUEZ-POLETTI, Jose Luis (Universidad Complutense de Madrid (Spain))

Presenter: Dr VAZQUEZ-POLETTI, Jose Luis (Universidad Complutense de Madrid (Spain))

Session Classification: Plenary talks

Track Classification: Cloud&Grid Technologies

Contribution ID: **3** Type: **not specified**

Effective Analysis Programming with C++

The language C++ supports multiple programming paradigms and is often the first choice for applications where performance matters. It is widely being used by scientific communities including high energy physics. The course covers basic software design patterns, simple best practice rules, examples from the Standard Template Library, and selected topics from object oriented and generic programming. The goal is to help scientists to efficiently use C++ in order to improve the quality and to ease the maintenance of their software. Participants are required to have basic knowledge of C++ and the concepts of object oriented programming.

Authors: Dr MEYER, Jörg (KIT); Dr HECK, Martin (KIT)

Presenters: Dr MEYER, Jörg (KIT); Dr HECK, Martin (KIT)

Track Classification: Effective programming and multi-core computing

Contribution ID: **4** Type: **not specified**

Multi-threaded Programming

During this session, the participants will learn the basic concepts of multi-threaded programming. In particular, they will apply this paradigms to well known and widely used data-processing algorithms. Available software solutions will be introduced and specific functionalities they offer will be discussed. The second, hands-on part of this session will give the participants the opportunity to implement multi-threaded algorithms and benchmark their profitability.

Desirable Prerequisite:
Basic knowledge of C++
C++ templates will be used

Author: Mr HAUTH, Thomas (CERN)

Presenter: Mr HAUTH, Thomas (CERN)

Track Classification: Effective programming and multi-core computing

Contribution ID: **5** Type: **not specified**

Handling Big Data - an Overview of Mass Storage Technologies

Monday, August 26, 2013 4:00 PM (40 minutes)

According to IDC forecasts, Big Data-related IT spending is to rise 40% each year between 2012 and 2020, and the total amount of information stored world-wide will about double every two years. It means that the, so called, digital universe will explode from 2.8 zettabytes in 2012 to 40ZB, or 40 trillion GB, in 2020. This is more than 5200 gigabytes for every man, woman, and child alive in 2020. It is therefore crucial to build storage systems robust and scalable enough to not only safely hold fast-growing amount of data but also capable of making all the information they hold available for access in an efficient manner.

This presentation will introduce basic principles, characteristics and fundamental technologies used in today's storage systems. It will compare distributed POSIX-like filesystem approach to the web-scale systems implementing only a reduced set of operations, like GET, PUT & DELETE. There will be a discussion of advantages and disadvantages of both technologies and their applicability to various requirements. This will lead to a characterization of efficient access protocols, in particular the proposed HTTP 2.0 standard based on Google's SPDY protocol.

Author: JANYST, Lukasz (CERN)

Presenter: JANYST, Lukasz (CERN)

Session Classification: Plenary talks

Track Classification: Big Data and Large Storage Systems

Contribution ID: **6** Type: **not specified**

Enabling eScience

Monday, August 26, 2013 2:30 PM (1 hour)

By today, especially in the natural sciences, computers have become indispensible tools and instruments for research. Recently, due to progress in digital measurement technology, researchers acquire vast amounts of data in ALL domains of science. Not only the amount of data, but also its complexity is continuously increasing. And to top it off, the data needs to be shared within large scientific collaborations among many people located at various research institutions all across the globe. The term eScience has been coined in the early 2000s to describe research that heavily relies on computational technologies - today it is applicable to almost all sciences.

However, the new organisational and engineering tasks to deal with the new complexity cannot be expected to be managed by the researchers alone. There is the need for a new kind of IT, engineering and research support that focuses on the methods and technologies enabling eScience. This includes large-scale data management, data storage, data policies and data lifecycle considerations, as well as automated validation, processing and analysis of data. It also includes the creation and usage of large-scale infrastructures for eScience, like distributed Grid or Cloud infrastructures. And it requires domain-specific know-how, for example in bioinformatics, chemoinformatics, medical informatics, etc. Ultimately is all about integration of several layers of infrastructure and software so that the scientists can extract the necessary information from large amounts of raw data, to be used by to propose new insights and theories, and to plan the next round of experimentation and observation.

I will show based on the example of the SystemsX.ch SyBIT project and the Sloan Digital Sky Survey how modern eScience operates, what kinds of problems already have reasonable solutions and where the current challenges are.

Author: Dr KUNSZT, Peter (ETH Zürich)

Presenter: Dr KUNSZT, Peter (ETH Zürich)

Session Classification: Plenary talks

Track Classification: Big Data and Large Storage Systems

Contribution ID: **7** Type: **not specified**

Welcome to Karlsruhe Institute of Technology

Monday, August 26, 2013 2:00 PM (5 minutes)

Presenter: Prof. SCHMECK, Hartmut (KIT, COMMputation)

Session Classification: Plenary talks

Contribution ID: **8** Type: **not specified**

GridKa School - Event Overview

Monday, August 26, 2013 2:20 PM (10 minutes)

Presenter: Dr WEBER, Pavel (SCC-KIT)

Session Classification: Plenary talks

Contribution ID: **9** Type: **not specified**

Security Challenges in Distributed Environments

Monday, August 26, 2013 4:40 PM (40 minutes)

The European Grid Infrastructure (EGI, http://egi.eu/) is a distributed environment, spanning roughly 270,000 logical CPUs, 140 PB of disk, and 130 PB of tape storage at 352 sites in 54 countries. More than 20,000 users, organised in more than 200 virtual organisations, from all over the world are currently running approximately 1.4 million jobs per day using this infrastructure.

This presentation will cover the many challenging tasks of maintaining operational security within the EGI infrastructure. The activities of the EGI CSIRT include detailed monitoring of the sites and measurement of their incident response capability, as exercised through large-scale security drills, closely simulating real-world incidents. The results of these simulations stimulated the set up of a forensics training frame work for cluster administrators and also addresses the particularities of grid technology.

The interactions with other national and infrastructure CSIRTs will also be presented.

Author: NIXON, Leif (Linköping University)

Presenter: NIXON, Leif (Linköping University)

Session Classification: Plenary talks

Track Classification: Cloud&Grid Technologies

Contribution ID: **10** Type: **not specified**

From Milliwatts to PFLOPS - High-Performance and Energy Efficient General Purpose x86 Multi/Many-Core Architecture

Tuesday, August 27, 2013 9:00 AM (40 minutes)

As we see Moore's Law alive and well, more and more parallelism is introduced into all computing platforms and on all levels of integration and programming to achieve higher performance and energy efficiency. We will discuss the new Intel® Many Integrated Core (MIC) architecture for highly-parallel workloads with general purpose, energy efficient TFLOPS performance on a single chip. This also includes the challenges and opportunities for parallel programming models, methodologies and software tools to archive high efficiency, highly productivity and sustainability for parallel applications. At the end we will discuss the journey to ExaScale including technology trends for high-performance computing and look at some of the R&D areas for HPC and Technical Computing at Intel.

Author: Dr CORNELIUS, Herbert (INTEL)

Presenter: Dr CORNELIUS, Herbert (INTEL)

Session Classification: Plenary talks

Track Classification: Effective programming and multi-core computing

Contribution ID: **11** Type: **not specified**

Hadoop in Complex Systems Research (A Review on Tools, Best Practices & Applications)

Tuesday, August 27, 2013 9:40 AM (40 minutes)

A Hadoop cluster is the tool of choice for many large scale analytics applications and a large variety of commercial tools is available for Data Warehouses and for typical SQL-like applications.

But how to deal with networks and time series? How to collect data for complex systems studies and what are good practices for working with libraries like Mahout and Giraph?
The sample use case deals with a data set from Wikipedia to illustrate one can combine multiple public data sources with own personal data collections, e.g. from Twitter, intranet servers or even personal mailboxes. Efficient approaches for time series (pre)-processing and time dependent graph analysis will be presented.

Author: KÄMPF, Mirko (Cloudera)

Presenter: KÄMPF, Mirko (Cloudera)

Session Classification: Plenary talks

Track Classification: Big Data and Large Storage Systems

Contribution ID: **12** Type: **not specified**

Cloud Computing Patterns –Fundamentals to Design, Build, and Manage Cloud Applications

Tuesday, August 27, 2013 10:50 AM (40 minutes)

The functionality found in different products in the cloud computing market today is often similar, but hidden behind different product names and other provider-specific terminology. We analyzed this multitude of cloud-related products to extract the common underlying behavior as well as the common architectural best practices that developers using these cloud technologies should follow. The goal of this abstraction was to create a set of architectural patterns that capture the provider-independent sustainable knowledge about how to design, build, and manage cloud applications. The resulting cloud computing patterns help to characterize cloud environments, describe abstract functionality of cloud offerings, and guide application developers during the design time, deployment, and runtime of their applications.
This talk gives an overview on the book of the same name (Fehling, C., Leymann, F., Retter, R., Schupeck, W., Arbitter, P.: Cloud Computing Patterns, Springer, 2013. ISBN: 978-3-7091-1567-1). The covered patterns describe how to build cloud-native applications and how to select suitable cloud infrastructure offerings and platform offerings by employing practical use-case scenarios.

Author: FEHLING, Christoph (Uni Stuttgart)

Presenter: FEHLING, Christoph (Uni Stuttgart)

Session Classification: Plenary talks

Track Classification: Cloud&Grid Technologies

Contribution ID: **13** Type: **not specified**

GPU Acceleration Benefits for Scientific Applications

Tuesday, August 27, 2013 11:30 AM (40 minutes)

Computational researchers, scientists and engineers are rapidly shifting to computing solutions running on GPUs as this offers significant advantages in performance and energy efficiency.

This presentation will provide a short overview about GPU Computing and NVIDIA's parallel computing platform. It will show how features of the latest Kepler GPU architecture (eg. Hyper-Q, GPU-aware MPI and GPUDirect RDMA) improve the performance, scalability and GPU utilization of scientific applications. In addition an outlook about future GPU developments will be presented.

Author: KOEHLER, Axel (NVIDIA)

Presenter: KOEHLER, Axel (NVIDIA)

Session Classification: Plenary talks

Track Classification: Effective programming and multi-core computing

Contribution ID: **14** Type: **not specified**

Smart Microscopy Platforms For Efficient In Vivo Small Molecule Screens

Wednesday, August 28, 2013 9:00 AM (40 minutes)

Modern robotic microscopy platforms (High content screening platforms) are ideal instruments for large scale genome studies. The image based read outs often generate 10s of TByte data sets per single experiment. 10.000s of experiments are waiting to be done in the next years in hundreds of labs worldwide. Besides cell based assays , transgenic model organism like zebrafish or drosophila allow more detailed HCS studies in vivo in 4D, opening an entire new field of large scale research. We present a cutting edge technology overview including in vivo compound screening strategies for pharmaceutically relevant readouts (e.g. inflammatory effects and parkinson model systems) and discuss the data challenge and solutions.
Next Generation robotic microscopes offer a huge potential for efficient life science research but require novel screening technologies for higher throughput approaches. Novel microscope types, easy to use data storage systems, high speed data processing, data integration, 4D visualization, search engine technologies fpr distributed data sinks will be the challenge for Next-Generation-High-Content-Screening.

Authors :
Gehrig J.*, Grabher C., Westhoff J.*, Asmi Shah*, Dominic Lütjohann*, Liebel U.*

- Karlsruhe Institute of Technology KIT - Accelerator Lab, Karlsruhe Germany ** KIT Karlsruhe, Institute of Toxicology and Genetics, Karlsruhe Germany *** Children's Hospital – Medical University Heidelberg, Germany Corresponding author's e-mail urban.liebel@kit.edu

Keywords: Screening, in vivo, 4D, intelligent microscopy, image processing, data analysis

References:
J Vis Exp. 2012 Jul 16;(65):e4203. doi: 10.3791/4203.
Biotechniques http://www.ncbi.nlm.nih.gov/pubmed/21548893. 2011 May;50(5):319-24. doi: 10.2144/000113669
Nat Methods. 2011 Mar;8(3):246-9. doi: 10.1038/nmeth.1558. Epub 2011 Jan 23

Author: Dr LIEBEL, Urban (Accelerator-lab)

Presenter: Dr LIEBEL, Urban (Accelerator-lab)

Session Classification: Plenary talks

Track Classification: Big Data and Large Storage Systems

Contribution ID: **15** Type: **not specified**

Multi-core Computing in High Energy Physics

Wednesday, August 28, 2013 9:40 AM (40 minutes)

Even though the miniaturization of transistors on chips continues like predicted by Moore's law, computer hardware starts to face scaling issues, so-called performance 'walls'. The probably best known one is the 'power wall', which limits clock frequencies. The best way of increasing processor performance remains now to increase the parallelization of the architecture. Soon standard CPUs will contain many dozen cores on the same die. In addition, vector units become again standard. To not to waste the available resources, application developers are forced to re-think their traditional ways of software design.

This talk will explain some of the common problems, and some ways of solving them. It will summarize the on-going parallelization activities in the field of high-energy physics software and as well give an outlook for what to expect in the coming decade.

Author: Dr HEGNER, Benedikt (CERN)

Presenter: Dr HEGNER, Benedikt (CERN)

Session Classification: Plenary talks

Track Classification: Effective programming and multi-core computing

Contribution ID: **16** Type: **not specified**

SmartCloud Technology for a Smarter Planet

Thursday, August 29, 2013 9:40 AM (40 minutes)

coming soon

Author: Dr OBERST, Oliver (IBM)

Presenter: Dr OBERST, Oliver (IBM)

Session Classification: Plenary talks

Track Classification: Cloud&Grid Technologies

Contribution ID: **17** Type: **not specified**

Formation of the Software Defined Data Center

Thursday, August 29, 2013 10:50 AM (40 minutes)

The It infrastructure of today's datacenters are getting more and more complex while at the same time the demand of ease of use is changing the whole industry. Petabyte scale datacenters don't allow traditional operations where administrators and technicians need to investigate failures for singe users or applications at scale. A change towards a policy driven architecture is required that considers outages right from the beginning in the architecture. Failures and outages need to be seen as normal behaviors rather than rare events. This view is one of the ideas of the Software Defined Datacenters where Business Managers and Architects define policies which comply with specific SLAs. The presentation will outline the need for SDDCs by presenting analogies to the real world which will help to understand the transformation even for non-technical IT personnel. It'll then give a very brief introduction to EMC's strategy in that area.

Author: Dr RADTKE, Stefan (EMC<b>2</b>)

Presenter: Dr RADTKE, Stefan (EMC<b>2</b>)

Session Classification: Plenary talks

Track Classification: Cloud&Grid Technologies

Contribution ID: **18** Type: **not specified**

The Evolution of the Information System in EGI/WLCG

Friday, August 30, 2013 9:40 AM (40 minutes)

In a distributed system it's necessary to be able to get information about the available services and resources. This includes the existence and properties of Grid services and details about their current state. The information is structured according to a schema, which needs to be flexible enough to represent the variety of services in the Grid but simple enough to be usable. It is collected by information providers running on the services, transported and aggregated to provider a Grid-wide view, and accessed via queries from client tools. An information system comprises all these aspects.

The general structure of the information system used in the EGI/WLCG Grid was defined more than a decade ago. The representation of the information uses the so-called "GLUE schema", implemented using the LDAP technology, and information is aggregated in LDAP servers providing a view of the entire Grid. This system has been stable for some years, but is currently evolving in various ways. In this talk I will discuss the structure of the system as it is now, the changes which are currently under way, and possible directions in the near future.

Author: Dr BURKE, Stephen (European Grid Infrastructure (EGI))

Presenter: Dr BURKE, Stephen (European Grid Infrastructure (EGI))

Session Classification: Plenary talks

Track Classification: Cloud&Grid Technologies

Contribution ID: **19** Type: **not specified**

Helix Nebula –the Science Cloud: a public-private partnership building a multidisciplinary cloud platform for data intensive science

Friday, August 30, 2013 10:50 AM (1 hour)

The feasibility of using commercial cloud services for scientific research is of great interest to research organisations such as CERN, ESA and EMBL, to the suppliers of cloud-based services and to the national and European funding agencies. Through the Helix Nebula - the Science Cloud [1] initiative and with the support of the European Commission, these stakeholders are driving a two year pilot-phase during which procurement processes and governance issues for a framework of public/private partnership will be appraised. Three initial flagship use cases from high energy physics, molecular biology and earth-observation are being used to validate the approach, enable a cost-benefit analysis to be undertaken and prepare the next stage of the Science Cloud Strategic Plan [2] to be developed and approved.

The power of Helix Nebula lies in a shared set of services for initially 3 very different sciences each supporting a global community and thus building a common e-Science platform. CERN is exploring how commercial cloud services could serve its high energy physics experiments [9] while EMBL is developing a portal for cloud-supported analysis of large and complex genomes. This will facilitate genomic assembly and annotation, allowing a deeper insight into evolution and biodiversity across a range of organisms [10]. ESA is developing the Geohazard Supersites project to advance scientific understanding of the physical processes which control earthquakes and volcanic eruptions as well as those driving tectonics and Earth surface dynamics.[11].

The work of Helix Nebula and its recent architecture model [6] has shown that is it technically feasible to allow publicly funded infrastructures, such as EGI [7] and GEANT [8], to interoperate with commercial cloud services. Such hybrid systems are in the interest of the existing users of publicly funded infrastructures and funding agencies because they will provide "freedom of choice"over the type of computing resources to be consumed and the manner in which they can be obtained.

But to offer such freedom-of choice across a spectrum of suppliers, various issues such as intellectual property, legal responsibility, service quality agreements and related issues need to be addressed. Investigating these issues is one of the goals of the Helix Nebula initiative.

The next generation of researchers will put aside the historical categorisation of research as a neatly defined set of disciplines and integrate the data from different sources and instruments into complex models that are as applicable to earth observation or biomedicine as they are to high-energy physics. This aggregation of datasets and development of new models will accelerate scientific development but will only be possible if the issues of data intensive science described above are addressed. The culture of science has the possibility to develop with the availability of Helix Nebula as a "Science Cloud"because:
Large scale datasets from many disciplines will be accessible
Scientists and others will be able to develop and contribute open source tools to expand the set of services available
Collaboration of scientists will take place around the on-demand availability of data, tools and services
Cross-domain research will advance at a faster pace due to the availability of a common platform.

References:

1 http://www.helix-nebula.eu/

2 http://cdsweb.cern.ch/record/1374172/files/CERN-OPEN-2011-036.pdf

3 http://www.helix-nebula.eu/index.php/helix-nebula-use-cases/uc3.html
4 http://www.nsf.gov/geo/earthcube/
5 http://www.oceanobservatories.org/
6 http://cdsweb.cern.ch/record/1478364/files/HelixNebula-NOTE-2012-001.pdf
7 http://www.nsf.gov/geo/earthcube/
8 http://www.geant.net/
9 http://www.helix-nebula.eu/index.php/helix-nebula-use-cases/uc1.html
10 http://www.helix-nebula.eu/index.php/helix-nebula-use-cases/uc2.html
11 http://www.helix-nebula.eu/index.php/helix-nebula-use-cases/uc3.html

Author: Dr JONES, Bob (CERN)

Presenter: Dr JONES, Bob (CERN)

Session Classification: Plenary talks

Track Classification: Cloud&Grid Technologies

Contribution ID: **20** Type: **not specified**

BIG DATA in Science, Best Practices

Friday, August 30, 2013 9:00 AM (40 minutes)

Whilst Big Data is often characterised in terms of its volume in bytes: Tera, Peta, Zeta, there is also the crucial aspect regarding the degree of complexity within the data set to consider. Such complexity means that good data management is an essential element in the creation of high quality research data, without which researchers who collect the data will themselves be unable to realise the full scientific potential of the data set. Research data needs to be well organised, documented, preserved and accessible if their accuracy and validity is to be controlled.

The advent of Big Data by definition only makes these challenges harder, meaning that along with advances in data processing technology and applications, policies regarding data handling need to be evolved and adhered to. Increasingly funding agencies are now making demands for data management plans to be included within research grant applications. This becomes particularly prevalent where data sharing and open access to scientific data are made preconditions for access to research funding.

This presentation will discuss the main issues concerning best practice for handling large scientific data sets, as well as trying to look ahead to see how funding agencies are increasingly attempting to influence how scientists handle their data, by themselves defining such best practice.

Author: Dr APLIN, Steve (DESY)

Presenter: Dr APLIN, Steve (DESY)

Session Classification: Plenary talks

Track Classification: Big Data and Large Storage Systems

Contribution ID: **21** Type: **not specified**

Summary and Evolution

Friday, August 30, 2013 11:50 AM (10 minutes)

Presenter: Dr WEBER, Pavel (SCC-KIT)

Session Classification: Plenary talks

Contribution ID: **22** Type: **not specified**

GPU Programming using CUDA

All computing systems, from mobile to supercomputers, are becoming heterogeneous parallel computers using both multi-core CPUs and many-thread GPUs for higher power efficiency and computation throughput.
While the computing community is racing to build tools and libraries to ease the use of these heterogeneous parallel computing systems, effective and confident use of these systems will always require knowledge about the low-level programming interfaces in these systems.
This lecture is designed to introduce through examples and hands-on exercises, based on the CUDA programming language, the three abstractions that make the foundations of GPU programming:

- Thread hierarchy
- Synchronization
- Memory hierarchy/Shared Memory

Author: Mr PANTALEO, Felice (University of Pisa)

Presenter: Mr PANTALEO, Felice (University of Pisa)

Track Classification: Effective programming and multi-core computing

Contribution ID: **23** Type: **not specified**

Amazon Cloud Computing Tutorial

In the last couple of years cloud computing has achieved an important status in the IT scene. The renting of computing power, storage and applications according to requirements is regarded as future business.

This tutorial course gives an introduction of the basic concepts of the Infrastructure-as-a-Service (IaaS) model based on the cloud offerings provided by Amazon, one of the present leading commercial cloud computing providers.

Authors: KRAUSS, Peter (KIT); Mr KURZE, Tobias (KIT); Mr MAUCH, Viktor (KIT)

Presenters: KRAUSS, Peter (KIT); Mr KURZE, Tobias (KIT); Mr MAUCH, Viktor (KIT)

Track Classification: Cloud&Grid Technologies

Contribution ID: **24** Type: **not specified**

Relational and non-relational databases

This session will be an introduction to relational and non-relational database management systems, with a hands-on approach.

1) Theory Session

Introduction to relational databases including terminology, relations, constraints, and operations.

2) Practice Session

Development of a simple application with a relational database backend using Python and SQLite.

3) Discussion

Typical pitfalls when building applications with a database backend.

4) Theory Session

Introduction to non-relational databases including characteristics, scalability, consistency, mapreduce, and operations.

5) Practice Session

Development of a simple application with a non-relational database backend using Python and MongoDB.

6) Discussion

Comparison of the developed applications with both types of backends.

Author: LASSNIG, Mario (CERN)

Presenter: LASSNIG, Mario (CERN)

Track Classification: Big Data and Large Storage Systems

Contribution ID: **25** Type: **not specified**

Python for Scientific programming

Python is a high-level, dynamic, general-purpose programming language. It is remarkable for the clarity and expressive power it offers in exchange for a relatively low learning investment.

Python is designed to be extensible with low-level languages. SciPy is a collection of efficient tools for scientific programming, exposed as Python modules. Cython is a compiler for (an extended version of) Python which makes it possible to turn Python code in to highly efficient low-level extension modules, or to link Python code to existing low-level libraries.

Combining Python with packages such a SciPy and Cython, provides the programmer with the best of both worlds: the high productivity and ease of use of the Python language combined with the efficiency of low-level components.

This session introduces the Python language, highlighting its flexibility and expressivity and contrasting it to more static and low-level languages such as C++. It goes on to explore how highly performant programs can be developed in Python with the help of SciPy and Cython.

Author: Dr GENEROWICZ, Jacek (CERN)

Presenter: Dr GENEROWICZ, Jacek (CERN)

Track Classification: Effective programming and multi-core computing

Contribution ID: **26** Type: **not specified**

Cluster security tournament - Hands-on incident response and forensics in a realistic environment

In this workshop the participants will take on the role as security teams being responsible for the operational security of simulated grid sites running in a virtualized environment.

The sites will face attacks very similar to those seen in real life. The teams' task is to respond to these attacks and keep their services up and running as far as possible.

A running score will be kept, and at the end of the workshop the winners will receive fabulous prizes.

Maximum number of participants is 18.

Be prepared

A security incident always puts you in a challenging situation. You have to do many things correctly, quickly and in the correct order. What to do and when during incident response is usually formulated in an incident response procedure. We will start from the general Grid Incident Response Procedure available from EGI-CSIRT and discuss how to adapt it to local regulations.

Have a view of your site

Usually the information you initially get in a security incident will be relatively sparse and the amount of logs quite large. Therefore it is crucial to quickly get an initial overview of the problem, i.e. which systems are affected and which systems are at risk. Here we will discuss and use tools you can easily set up at your site as well, like a central syslog facility, a grid systems log analyzer and EGI CSIRT's vulnerability scanner Pakiti.

The heat is on

Now it is time to put your newly acquired knowledge to the test. As administrator of a simulated cluster, you will have to defend yourself against a determined attacker.

Hands-on forensics

Investigating a compromised system is a delicate situation. It is easy to lose crucial information if you are not careful enough. We will discuss several levels of volatile information and do and dont's in how to collect it. The creation and analysis of memory and disk images will be discussed.

Wrap-up, lessons learned

At the end of the work shop we will discuss the findings. The crucial point here is to find how the site was attacked and which steps could be taken to prevent this from happening again. This should result in some best practices on how to reduce the attack surface of your site.

Authors: VERSTEGEN, Aram; REESE, Heiko (KIT); NIXON, Leif (Linköping University); GABRIEL, Sven (Nikhef); DUSSA, Tobias (KIT); EPTING, Ursula (KIT)

Presenters: VERSTEGEN, Aram; NIXON, Leif (Linköping University); GABRIEL, Sven (Nikhef)

Contribution ID: **27** Type: **not specified**

Training Session on Openstack

OpenStack is currently one of the most evolving open IaaS solutions available. Every new release comes with a huge set of new features. It can be hard to hold pace with such changes. Starting from scratch also proves difficult due to the complexity of the several components interacting with each other.

The proposed training targets system administrators with little or no knowledge on cloud infrastructure, interested in learning how deploy and operate Openstack.
The training is organised in two full days. Main topics of the training will be:

- a general introduction into OpenStack (Folsom/Grizzly) and its core components
- an overview of the supporting software and choices (e.g. database, messaging queue and so on)
- a hands-on install guide on Keystone (OpenStack identity management)
- a hand-on install guide on Glance (OpenStack image-store)
- a hands-on install guide on Nova (OpenStack compute)
- a hands-on install guide on Quantum (OpenStack Networking)

If time is left, a short introduction on OpenStack compatible storage systems may be added (Swift/Ceph)

Authors: MESSINA, Antonio (University of Zurich / GC3); CASUTT, Joël (SWITCH)

Track Classification: Cloud&Grid Technologies

Contribution ID: **28** Type: **not specified**

dCache Workshop

Christian Bernardt (DESY)
Christoph Anton Mitterer (Ludwig Maximilian University of Munich)
Oleg Tsigenov (RWTH Aachen)

dCache is one of the most used storage solutions in the WLCG consisting of 94 PB of storage distributed world wide on 77 sites. Depending on the Persistency Model, dCache provides methods for exchanging data with backend (tertiary) Storage Systems as well as space management, pool attraction, dataset replication, hot spot determination and recovery from disk or node failures. Beside HEP specific protocols, data in dCache can be accessed via NFSv4.1 (pNFS) as well as through WebDav. The workshop includes theoretical sessions and practical hands-on sessions such as installation, configuration of its components, simple usage and monitoring. The basic knowledge of Unix systems is required. Please familiarise yourself with a Linux terminal and the peculiarities of a linux text editor (vi, emacs etc.).

Authors: BERNARDT, Christian (DESY); MITTERER, Christoph (LMU); TSIGENOV, Oleg (Uni Aachen)

Presenters: BERNARDT, Christian (DESY); MITTERER, Christoph (LMU); TSIGENOV, Oleg (Uni Aachen)

Track Classification: Big Data and Large Storage Systems

Contribution ID: **29** Type: **not specified**

ROOT Tutorial

The ROOT software framework provides all the functionality needed to store and analyze large amounts of data in an efficient way.
We will provide an introduction to the ROOT system and its tools for data analysis and visualisation.
The main features of ROOT such as histogramming, data visualization, object I/O and advanced statistical analysis techniques will be presented. We will also introduce RooFit and RooStats, dedicated tools for advance fitting, which are currently used by the LHC experiments. The participants will have the opportunity also to directly practise and learn the ROOT tools via hands-on exercises in C++, covering some of the main functionality of the ROOT framework.

Author: MONETA, Lorenzo (CERN)

Presenter: MONETA, Lorenzo (CERN)

Track Classification: Effective programming and multi-core computing

Contribution ID: **30** Type: **not specified**

gLite Middleware Administration

This gLite middleware administration workshop gives students a chance to perform installation and configuration of some of the EMI compute components. The goal of the workshop is to install a minimal Grid site using the EMI CREAM Computing Element (CE) and a Worker Node (WN) using the PBS-Torque batch system. Students will be shown how to install and configure these services using YAIM, and how to troubleshoot problems that may occur.
After the manual installation, an example on a puppet automated installation will be provided.

Author: BERTOCCO, Sara (INFN)

Presenter: BERTOCCO, Sara (INFN)

Track Classification: Cloud&Grid Technologies

Contribution ID: **31** Type: **not specified**

Combining Grid, Cloud and Volunteer Computing

It's well known that the developments environments used in Grid, Volunteer computing (VC) and Cloud are very different. The key differences between these three platforms are based on theoretical concepts as well as implementa¬tions.

The aim of this tutorial is to propose a set of concepts and tools used to bridge these three large-scale distributed systems: Grid, Cloud and Volunteer Computing.
Concretely speaking, we propose a common library used to develop high performance applications that could be deployed on Grid, VC and Cloud without any re-writing. The following platforms/middlewares will be used during the practical part:
1.the Advanced Resource Connector (ARC) middleware
2.the XtremWeb-CH volunteer computing platform (XWCH: www.xtremwebch.net)
3.the cloud platforms: Amazon, Azure and Venus-C

The tutorial is composed of theoretical and practical parts. The theoretical part will deal with the following aspects:
- Grid and Volunteer computing vs. Cloud computing
- Overview of ARC, XWCH, Amazon, Venus-C and Azure platforms
- How to develop applications for ARC, XWCH, Amazon, Venus-C and Azure platforms
- A common high-level API for large scale distributed systems

During the practical part, the students will be able to:
- Write his/her own application
- Deploy his/her application by using one or several of these bridges: ARC/XWCH, XWCH/Amazon and XWCH/Azure.

Author: Dr ABDENNADHER, Nabil (SwiNG)

Presenter: Dr ABDENNADHER, Nabil (SwiNG)

Track Classification: Cloud&Grid Technologies

Contribution ID: **32** Type: **not specified**

Evening Lecture: Nature Inspired Computing

Tuesday, August 27, 2013 7:00 PM (1h 30m)

Computers - the high point of technology. Our omnipresent slaves and sometimes masters. But thousands of years before the first vacuum tube lit up biological computing machines existed that would outmatch our contemporary silicon companions in nearly every aspect. If in doubt just try to build a machine doing what a simple ordinary house fly does. Soon you will realize, that this simple creature processes and integrates large amounts of various sensory data, infers decisions to sustain its life and adapts to the environment. How?

In the talk we will go through some theory and practice of computing methods inspired by the nature.Despite we are slowly becoming capable to build the computing hardware of the desired complexity we often fail to program that hardware to our liking. Partially, that's why we want the hardware to learn by itself. We will see how artificial neural networks are build and used. We will also see how it is possible to find solutions to complex problems by means of simulated evolutionary optimization. And finally, the memory-prediction framework a progressive new theory trying to explain how the mammalian brain works will be presented. The practical examples to be shown include land use categorization using artificial neural networks, evolutionary optimization of a mechanical structure, object identification in sequences of images using a method based on memory-prediction framework and some more.

Author: Dr BUNDZEL, Marek (Technical University Kosice)

Presenter: Dr BUNDZEL, Marek (Technical University Kosice)

Contribution ID: **33** Type: **not specified**

GridKa School Dinner

Thursday, August 29, 2013 8:00 PM (2 hours)

Contribution ID: **36** Type: **not specified**

Introduction to the Steinbuch Centre for Computing

Monday, August 26, 2013 2:05 PM (15 minutes)

Presenter: Prof. STREIT, Achim (KIT-SCC)

Session Classification: Plenary talks