GridKa School 2014: Big Data, Cloud Computing and Modern Programming

Report of Contributions

https://indico.scc.kit.edu/e/46

Contribution ID : **0** Type : **not specified**

Welcome to Karlsruhe Institute of Technology

Summary

Contribution ID : **1** Type : **not specified**

Introduction to the Steinbuch Centre for Computing

Summary

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

GridKa School - Event Overview

Summary

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

LHC Computing: towards clouds and agile infrastructures

Monday, 1 September 2014 14:30 (60)

CERN is undergoing a major transformation in how computing services are delivered with the addition of a second data centre to help process over 35PB/year from the Large Hadron Collider.

Within the constraints of fixed budget and manpower, agile computing techniques and common open source tools are being adopted to support over 11,000 physicists.

By challenging special requirements and understanding how other large computing infrastructures are built, we have deployed a 50,000 core cloud based infrastructure building on tools such as Puppet, OpenStack and Kibana.

In moving to a cloud model, this has also required close examination of the IT processes and culture. Finding the right approach between Enterprise and DevOps techniques has been one of the greatest challenges of this transformation.

This talk will cover the requirements, tools selected, results achieved so far and the outlook for the future.

==========

The talk is presented by Tim Bell. Tim Bell is responsible for the CERN IT Operating System and Infrastructure Group which supports Windows, Mac and Linux across the site along with virtualisation, printing, E-mail and web services. Prior to working at CERN, Tim worked for Deutsche Bank managing private banking infrastructure for Europe and with IBM as a Unix kernel developer and deploying large scale technical computing solutions. Tim is also an elected individual member of the OpenStack management board since 2012 and a member of the OpenStack user committee.

Summary

Primary author(s) : BELL, Tim (CERN)

Presenter(s) : BELL (CERN), Tim

Session Classification : Plenary talks

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

Brain Pathologies and Big Data

Monday, 1 September 2014 16:00 (60)

We now know that a single gene mutation may present with multiple phenotypes, and vice versa, that a range of genetic abnormalities may cause a single phenotype. These observations lead to the conclusion that a deeper understanding is needed of the way changes at one spatial or temporal level of organisation (e.g., genetic, proteomic or metabolic) integrate and translate into others, eventually resulting in behaviour and cognition. The traditional approach to determining disease nosology- eliciting symptoms and signs, creating clusters of like individuals and defining diseases primarily on those criteria has not generated fundamental breakthroughs in understanding sequences of pathophysiology mechanisms that lead to the repertoire of psychiatric and neurological diseases.

It is time to radically overhaul our epistemological approach to such problems. We now know a great deal about brain structure and function. From genes, through functional protein expression, to cerebral networks and functionally specialised areas defined via physiological cell recording, microanatomy and imaging we have accumulated a mass of knowledge about the brain that so far defies easy interpretation. Advances in information technologies, from supercomputers to distributed and interactive databases, now provide a way to federate very large and diverse datasets and to integrate them via predictive data-led analyses.

Human functional and structural brain imaging with MRI continues to revolutionise tissue characterisation from development, through ageing and as a function of disease. Multi-modal and multi-sequence imaging approaches that measure different aspects of tissue integrity are leading to a rich mesoscopic-level characterisation of brain tissue properties. Novel image classification techniques that capitalise on advanced machine learning techniques and powerful computers are opening the road to individual brain analysis. Data-mining methods, often developed in other data-rich domains of science, especially particle and nuclear physics, are making it possible to identify causes of disease or its expression from patterns derived by exhaustive analysis of combinations of genetic, molecular, clinical, behavioural and other biological data. Imaging is generating data that links molecular and cellular levels of organisation to the systems that subtend, action, sensation, cognition and emotion. These ideas will be illustrated with reference to the human dementias.

Summary

Primary author(s) : Prof. FRACKOWIAK, Richard (Department of Clinical Neuroscience, University of Lausanne)

Presenter(s) : Prof. FRACKOWIAK (UNIVERSITY OF LAUSANNE), Richard

Session Classification : Plenary talks

Track Classification : Big Data and Storage Systems

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

Evolution of Security Threats and Models

Monday, 1 September 2014 17:00 (40)

In a computing environment in constant evolution, the security management of our systems need to adapt: cyber-criminals use new attack angles, new technologies and architectures are introduced, old security models are weakened, etc.

This presentation will cover such recent evolutions from a security point of view and discuss new or future security challenges.

Summary

Primary author(s) : BRILLAULT, Vincent (CERN)

Presenter(s) : BRILLAULT (CERN), Vincent

Session Classification : Plenary talks

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

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

Tuesday, 2 September 2014 09:00 (40)

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.

Summary

Presenter(s) : Dr. KLEMM (INTEL), Michael

Session Classification : Plenary talks

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

SAP's Big Data Platform HANA - Technology and Business Innovation

Thursday, 4 September 2014 09:40 (40)

Summary

Presenter(s) : Dr. HAGEDORN (SAP), Jürgen

Session Classification : Plenary talks

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

Processing Big Data with modern applications

Tuesday, 2 September 2014 10:50 (25)

We will present two real-world data warehousing projects we solved using Hadoop. Both projects resulted in hybrid data warehouses, with Hadoop in the backend and a relational database as the interface for both BI tools and business users. We describe the architecture as well as the data sources and data volume involved.

Summary

Primary author(s) : SPREYER, Kathrin (inovex GmbH)

Presenter(s) : SPREYER (INOVEX GMBH), Kathrin

Session Classification : Plenary talks

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

Hadoop in Complex Systems Research

Tuesday, 2 September 2014 11:15 (25)

I am planning to shed light onto the theme of 'Metadata Management' in Hadoop. The Hive-Metastore exists for a long time and complementary to to it, there is HCatalog.
With this Pig users and MapReduce developers can access those Metadata as well. But how do we handle time-dependent aspects of Complex Systems that consist of multiple interrelated layers represented as graphs?
To handle such aspects efficiently, a new methodology that uses a semantic Wiki is proposed and demonstrated. The triple store is used as a centralized database and as an automatic system integration layer which works with a SPARQL-like query language.
Researchers and analysts can concentrate on system modeling aspects while developers focus on efficient I/O operations - whereby the content of the data is of minor importance.
I demonstrate the concept with an example using Apache Giraph and Gephi. Such analysis workflows can span numerous distributed clusters and all dependencies are documented in the Semantic Wiki. So we maintain a meta model for an arbitrary analysis-workflow which can be split into separate 'local Oozie workflows.'

Summary

Presenter(s) : KÄMPF (CLOUDERA), Mirko

Session Classification : Plenary talks

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

Parallel Programming using FastFlow

Tuesday, 2 September 2014 11:40 (40)

FastFlow is an open-source C++ research framework to support the development of multi-threaded applications in modern multi/many-core heterogeneous platforms.
The framework provides well-known stream-based algorithm skeleton constructs such as pipeline, task-farm and loop that are used to build more complex and powerful pattern: parallel_for, map, reduce, macro data-flow interpreter, genetic-computation, etc.

During the talk we introduce the structured parallel programming framework FastFlow and we discuss problems and issues related to the run-time implementation of the patterns. In particular we will discuss:

- algorithmic skeleton approaches and the associated static (template based) or dynamic (macro-data-flow based) implementation
- management of non functional features, with particular focus on performance
- different optimisations aimed at targeting clusters of multi-core
- heterogeneous architecture targeting (including GPGPUs, Intel Xeon PHI and Tilera Tile64)

Summary

Presenter(s) : Dr. TORQUATI (UNIVERSITY OF PISA), Massimo

Session Classification : Plenary talks

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

Outlier Detection and Description in Complex Databases

Wednesday, 3 September 2014 09:00 (40)

Outlier analysis is an important data mining task that aims to detect unexpected, rare, and suspicious objects in large and complex databases. Consistency checks in sensor networks, fraud detection in financial transactions, and emergency detection in health surveillance are only some of today's application domains for outlier analysis. As measuring and storing of data has become cheap, in all of these applications, objects are described by a large variety of different measures and relationships between objects. However, out of these complex databases, for each object only a small subset of relevant measures and relationships provides the meaningful information for outlier detection. The residual information is irrelevant for this object, and with the growing amount of irrelevant information traditional outlier mining approaches fail to detect outliers.

To address this problem, recent subspace search techniques focus on a selection of subspace projections. The objective is to find multiple subsets (i.e. subspaces) of the given attributes, which show a significant deviation between an outlier and regular objects. Thus, subspace search allows: (1) A clear distinction between clustered objects and outliers. (2) A description of outlier reasons by the selected subspaces. However, it lacks flexibility in handling different outlier characteristics that have been invented for different application domains and proposed as formal outlier models in the literature.

This talk will cover a flexible subspace selection scheme allowing instantiations with different outlier models. We utilize the differences of outlier scores in random subspaces to perform a combinatorial refinement of relevant subspaces. Our refinement allows an individual selection of subspaces for each outlier, which is tailored to the underlying outlier model. This flexibility ensures that the approach directly benefits from any research progress in future outlier models. It allows search for relevant subspaces individually for each outlier, and hence, enables to describe each outlier by its specific outlier properties.

Summary

Presenter(s) : Dr. MÜLLER (KIT), Emmanuel

Session Classification : Plenary talks

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

Multi-core Computing in High Energy Physics

Wednesday, 3 September 2014 09:40 (40)

Summary

Presenter(s) : Dr. HEGNER (CERN), Benedikt

Session Classification : Plenary talks

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

Can HPCclouds supersede traditional high performance computing?

Wednesday, 3 September 2014 10:50 (40)

With the advent of cloud computing, flexible and scalable services have been provided with the ambition to utilize bare metal resources in a more efficient way. The base technology for cloud computing is represented by virtualization; hence servers can contain several virtualized operating systems in a single physical box. As a small example, most of the servers offering web services are virtualized, from elastic e-business applications controlled by introduced user traffic through to virtual storage offerings managed by user's individual disk space demands. These encapsulated virtual machines are the key to flexibility and scalability, but due to fully virtualized operating systems the overall performance of those various resources decreases.

In contrast to flexible and scalable traditional cloud operation models, high performance computing requires a maximum of performance in computational power as well as I/O. Thus, performance dropping virtualization is not regarded at all even if it would provide beneficial capabilities. Within this talk, innovative approaches for high performance clouds will be introduced and elaborated in order to compare execution performance with configurability and flexibility.

Summary

Presenter(s) : GIENGER (UNIVERSITY OF STUTTGART), Michael

Session Classification : Plenary talks

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

Big Data Analytics - Use Cases & Strategy

Thursday, 4 September 2014 09:00 (40)

Big Data Analytics: Strategy and Use-Cases

The presentation by Christian Dornacher covers Hitachi's strategy for Big Data Analytics solutions based on existing know-how from solutions like predictive maintenance and log-analytics. It also shows different customer use-cases and how these customers plan to get better insight in their data.

About the presenter

Christian Dornacher has more than 22 years of IT experience. He worked as engineer, consultant, pre-sales, Alliance Manager, Sales and Business Development roles at Digital Equipment, Megabyte (Distributor), bdata systems (SI), Paralan, McDATA and prior to joining HDS at BlueArc where he was responsible for the OEM Sales / Business Development in EMEA and APAC. At HDS he was part of the EMEA Channel team focused on File and Content Solutions and since April 2013 focuses on Business Development for the File, Content and Cloud solutions as well as Big Data Analytics solutions in EMEA. He acts as the GEO-Lead within the EMEA team and works with internal teams like Product Management and Engineering as well as sales teams, partners and end-users.

Summary

Primary author(s) : DORNACHER, Christian (HITACHI DATA SYSTEMS GmbH)

Presenter(s) : DORNACHER (HITACHI DATA SYSTEMS GMBH), Christian (HITACHI DATA SYSTEMS GmbH)

Session Classification : Plenary talks

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

Next Generation of Monitoring - Predictive Analytics

Tuesday, 2 September 2014 09:40 (40)

In today's smarter planet whether it's smart meters in an electric grid, escalators and security cameras in office buildings, signals and switches from railroad networks or Wi-Fi in airplanes or the software systems that support them, our world is filled with devices that are instrumented and interconnected.

There was a time when a person walking through a building and checking meters individually was enough. Manual checks of the IT infrastructure also could be sufficient when the infrastructure was simple. But as complexity has grown, monitoring has required more powerful and sophisticated tools. Operational centers now face the problem of doing more with less, an increasing array of devices and systems that can be monitored coupled with larger and larger systems of increasing complexity .

IBM's Cloud and Smarter Infrastructure has been at the forefront of assisting organisations manage their operations centers. As the volume of data going though operations centers has exploded these centers face an increasing need to apply analytical techniques to prevent data blindness. As complex as these large systems may be, they are tied together by physical infrastructure and man made components and software, this provides a signal with which we can learn and build patterns. This talk will introduce some of the data that operations centers collect and work with, It will highlight how statistical patterns can be applied back to the operations center to reduce costs and drive operational efficiency.

Summary

Primary author(s) : Dr. BREW, Anthony (IBM)

Presenter(s) : Dr. BREW (IBM), Anthony

Session Classification : Plenary talks

Track Classification : Cloud Computing

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

Parallel Programming using the PGAS Approach

Friday, 5 September 2014 09:00 (40)

The two most common approaches for parallel programming are message passing (for example using MPI, the message passing interface) and threading (for example using OpenMP or Pthreads). Threading is generally considered an easier and more straightforward solution for parallel programming but it can generally only be used on a single shared memory node. MPI, on the other hand, scales to the full size of today's machines, but it requires a more complex planning and orchestration of data distribution and movement.

PGAS (Partitioned Global Address Space) approaches try to combine the best of both worlds, providing a threading abstraction for programming large distributed memory machines. Data locality is made explicit in order to be able to take advantage of it for performance and energy efficiency reasons. The talk will give an introduction to the concept of PGAS programming and provide examples using UPC (unified parallel C). The research project DASH, which provides a realization of the PGAS model in the form of a C++ template library, will also be introduced in the talk.

Summary

Presenter(s) : FÜRLINGER (UNIVERSITY OF MUNICH), Karl

Session Classification : Plenary talks

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

Identity challenges in a Big Data world

Friday, 5 September 2014 09:40 (40)

Proving who you are is a prerequisite for using computer resources, but the explosion of big data resources has resulted in users who are more likely to be remote and use the resources briefly. This tension has provided the opportunity for fresh solutions that are better suited to modern scientific methods. In this talk, such challenges are presented along with their solutions, using the international laboratory DESY and the dCache software collaboration as motivation.

Summary

Presenter(s) : Dr. MILLAR (DESY), Paul

Session Classification : Plenary talks

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

Cloud computing in Europe for Science and industry. First experience and current trends

Friday, 5 September 2014 10:50 (60)

The talk will discuss the current transformation in the computing landscape. The advent of Virtualization have made possible highly scalable and affordable distributed computing systems such as those offered by Cloud providers, public or private. This poses new challenges and problems to do with latency in accessing the data, SLAs, privacy and security issues. At the same time the explosion of data has generated the emergence of new computing paradigms such as MapReduce and Hadoop and the need for new computing storage hierarchies for HPC and distributed computing.

The talk will review some practical experience drawn from the recently concluded FP7 project Venus-C and discuss current issues and trends.

Summary

Presenter(s) : Dr. GAGLIARDI (UNIVERSITY OF CATALONIA), Fabrizio

Session Classification : Plenary talks

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

Robotics & Artificial Intelligence

Robotics & Artificial Intelligence

In recent years robotics has gained a lot of interest also in the area of artificial intelligence. While systems for a long time have been used as tools to implement classical AI approaches in the area of object recognition, environment representation, path and motion planning etc., researchers now begin to understand that the system (robot) itself is part of the question and has to be taken into account when teaching AI questions. This talk might survey the state of the art in robotics and outline ways to tackle the question of AI in the light of the systems as an integral part of the approach. Future milestones an key achievements will be discussed as a proposal to tackle the big question of AI.

Summary

Presenter(s) : Prof. KIRCHNER (UNIVERSITY OF BREMEN), Frank

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

Relational Databases

Throughout the course, the students will implement a full database application with safe and efficient methods, based on the concepts learned. Additionally, where necessary, pointers to the NoSQL/non-relational database sessions with MongoDB and Hadoop are given.Basic understanding of Linux and programming (at least C or Python) is required for this session.
The agenda is as follows:
Part 1: The basics
Database management systems - What/How/Why
The relational data model - Modeling languages
Structured Query Language (SQL) - The basics
Part 2: Safe use of databases
ACID - Making sure your data stays safe
Transactions, race conditions, deadlocks
SQL Injection - Malicious user requests
Part 3: Efficient use of databases
Query plans
Indexing
Partitioning
Part 4: Finishing up
Application development with a database backend
Questions/Answers

Summary

Presenter(s) : LASSNIG (CERN), Mario

Track Classification : Big Data and Storage Systems

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

OpenStack Workshop

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 but also due to the lack of exhaustive documentation. The proposed training targets system administrators with little or no knowledge on cloud infrastructure, interested in learning how to deploy and operate Openstack. The training is organised in three full days. Main topics of the training will be:

```
a general introduction to OpenStack (IceHouse) and its core components, with particular attenti
an overview of the supporting software, available choices and limitations (database, messaging
hands-on installation of the baisc components:
    MySQL
    RabbitMQ
    Keystone (identity service)
    Nova (compute service), using nova-network
    Glance (image service)
    Cinder (block storage service)
    Horizon (web interface)
```

The last day will be dedicated to Neutron, the OpenStack network service, and will include:

```
an overview of Neutron, its network providers and plugins
hands-on installation of Neutron
```

Summary

Primary author(s) : MESSINA, Antonio (S3IT, University of Zurich); ALEKSIEV, Tyanko (S3IT, University of Zurich)

Presenter(s) : MESSINA, Antonio; ALEKSIEV (S3IT, UNIVERSITY OF ZURICH), Tyanko (S3IT, University of Zurich)

Track Classification : Cloud Computing

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

Amazon Cloud Workshop

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.

Summary

Presenter(s) : MAUCH (KIT), Viktor

Track Classification : Cloud Computing

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

From C++03 to C++11

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. With the new C++11 Standard the language becomes simpler and at the same time it provides new methods to gain performance. The course will introduce new language features and will give an overview of extensions of the Standard Template Libraries. The targeted audience are people with some experience in C++(03) programming, who would like to get the best out of the new features provided by the C++(11) standard.

Summary

Presenter(s) : Dr. MEYER (KIT), Jörg

Track Classification : Modern Programming

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

Data Analysis in Python

Python is a high-level dynamic object-oriented programming language. It is easy to learn, intuitive, well documented, very readable and extremely powerful. Python is packaged with an impressive standard library following the so called “batteries included” philosophy. Together with the large number of additionally available scientific packages like NumPy, SciPy, pandas, matplotlib, etc., Python becomes a very well suited programming language for data analysis.

One more thing to mention is the possibility to easily integrate C, C++ or even FORTRAN code into Python, which can be used to optimize computational bottlenecks by moving the code to a lower-level compiled language. Cython, a compiler for Python code, is one of the standard ways to transform Python code into fast compiled low-level extensions and to interface already existing C/C++ code.

This hands-on session introduces the pythonic way of programming, demonstrates the power of Python in data analysis and gives a brief glimpse of developing performant code in Python using Cython.

Summary

Presenter(s) : Dr. GIFFELS (KIT), Manuel

Track Classification : Modern Programming

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

Programming Multi-core using FastFlow

During this tutorial session, the participants will learn how to build application structured as a combination of stream-based parallel pattern like pipeline, task-farm loops and their combinations. Then more high-level patterns will be introduced such as parallel_for, map and reduce, and we will see how to mix stream and data-parallel patterns to build simple (and not so simple) applications. During the tutorial different possible implementations will be discussed. Finally we will give the participants the opportunity to implement multi-threading algorithms and simple benchmark and to evaluate their performance.

Desirable Prerequisite:

```
Good knowledge of C
Basic knowledge of C++ templates (basic C++11 features will be also used)
Basic knowledge of multi-threading programming
```

Summary

Presenter(s) : Dr. MASSIMO (UNIVERSITY OF PISA), Torquati

Track Classification : Modern Programming

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

Hadoop for beginners

In the last couple of years Hadoop established itself as the de facto standard for dealing with large and very large datasets. However, Hadoop does introduce quite a lot of challenges for developers with a background of classical data analytics. One example is handling raw data (e.g., logfiles) which works quite differently in Hadoop than in classical, data warehouse focused architectures. Another example is developing MapReduce jobs, which differs from standard object-oriented or procedural paradigms. In addition to this, Hadoop has grown from a "simple" MapReduce tool to a complex ecosystem of technologies, covering a large variety of use cases: from distributed storage, data exploration and data analysis to automatic classification and prediction. This course covers Hadoop MapReduce and HDFS in great detail and enables the participants to be able to develop complex MapReduce algorithms on their own. The resulting in-depth understanding of the architecture allows for easier evaluation and selection of appropriate tools from the Hadoop ecosystem in future projects.

Prerequisites: basic knowledge of Java

Summary

Presenter(s) : SPREYER (INOVEX GMBH), Kathrin

Track Classification : Big Data and Storage Systems

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

MongoDB Workshop

This session is an introduction to a particular NoSQL database, MongoDB. MongoDB is an open-source database with document-oriented storage approach. Since it doesn't enforce any schema on data and because of its good performance, Mongo is nowadays widely used especially where unstructured data storage is needed. In addition, Mongo scales well and even provides partitioning over cluster of nodes. So, it is ideal for Big Data use cases. This session will provide theoretical basic knowledge about Mongo and support it with hands-on activities to get to know Mongo in practice. The agenda will cover the followings:

```
Getting familiar with Mongo terminologies
Executing CRUD operations
Indexing
Getting to know replication and Sharding mechanisms
```

Basic Linux knowledge and some background knowledge about relational databases might be helpful in this session.

Summary

Presenter(s) : Dr. SZUBA (KIT), Marek; AMERI, Parinaz (KIT)

Track Classification : Big Data and Storage Systems

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

Hadoop Workshop

Usage of Apache Hadoop in large scale Data Analysis Projects are on the way to become mainstream. But what are the required skills and how do I start with an Apache Hadoop project? The workshop shows and compares several aspects which should be considered in the beginning of large projects. How do I start with a POC and how works this: “scale out”? What data is stored how and how do I access data in my Hadoop cluster? What programming skills are required and what are the processing paradigms I should know in the beginning? Such questions are discussed and possible solutions are presented during this interactive hands on session. The example use case is a data driven market study, which combines social media, time series data, and network analysis in one project.

Participants will receive a download link for the latest Workshop-VM and a preparation survey two weeks before the workshop.

Summary

Presenter(s) : KÄMPF (CLOUDERA), Mirko

Track Classification : Big Data and Storage Systems

Contribution ID : **34** Type : **not specified**

Configuration Management with Puppet

Puppet is a configuration management tool adopted by many institutions in academia and industry of different size.Puppet can be used to configure many different operating systems and applications. Puppet integrates well with other tools e.g. Foreman, MCollective, …

The workshop will feature a hands-on tutorial on Puppet allowing users to write simple manifests themselves and managing them using Git.A selection of useful tools around Puppet will be presented.

Basic knowledge of the Linux operating system is required. The detailed agenda for the course is following:
1st day:

```
Introduction to Git
Setup \& technical infrastructure
    Explanation for the setup of the infrastructure, login to the machines
Write manifests
    Puppet language, resource types, modules, etc.
```

2nd day:

```
Leftovers from previous day, and/or some more advanced configuration
Series of small presentations and walk-throughs: Hiera, Facter, Foreman, MCollective, GitLab, .
```

Prerequisites:

```
Attendants should familiarize themselves with a Linux terminal and the peculiarities of a Linux
No knowledge of Puppet or Git is required.
```

Summary

Presenter(s) : JONES (CERN), Ben Dylan; STERNBERGER (DESY), Sven; Dr. KEMP (DESY), Yves

Track Classification : Cloud Computing

Contribution ID : **35** Type : **not specified**

OwnCloud Workshop

ownCloud provides universal access to your files via the web, your computer or your mobile devices — wherever you are. It also provides a platform to easily view & sync your contacts, calendars and bookmarks across all your devices and enables basic editing right on the web.

In this Workshop we will setup a basic ownCloud installation, extend it with apps, set up synchronization with various clients and - if time permits - dive into the development of ownCloud apps. Topics

```
Hello \& Welcome
Installation and Configuration
    Minimal setup
        apache, php \& sqlite->kinda meh
    Adding Apps
    Solid Single Machine setup
        nginx
        php-fpm
        apc
        mysql
    Scaling to multiple machines?
        +memcached
        mysql with replication
        load balancer
    What is different in the Enterprise Edition?
Synchronization \& Access Protocols
    WebDAV to access Files
        Desktop Client
        network drive mount in Win, Linux \& MacOS
        owncloud iOS \& Android Apps
    CalDAV to access Calendar
        Thunderbird via Lightning
        Evolution
        iOS natively
        Android via CalDAV Sync App
    CardDaV to access Contact
        iOS natively
        Android via CardDAV Sync App
    Notes
Your first owncloud app
    Setting up a development environment
```

Summary

Presenter(s) : BÖHM (OWNCLOUD), Felix

Track Classification : Cloud Computing

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

dCache Workshop

dCache is one of the most used storage solutions in the WLCG consisting of over 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. dCache has steadily improved its functionality up to the point that we are becoming the DESY storage cloud provider. This means that dCache users can now access data using the OwnCloud client software with its synchronisation functionality. In addition to that users can access their data by using the same user over NFSv41, WebDAV and gridFTP, which allows for a wide range of use cases from traditional HEP storage to even HPC application.

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.).

Summary

Presenter(s) : DELLE FRATTE (RZG), Cesare (Rechenzentrum Garching (RZG)); BERNARDT (DESY), Christian; MITTERER (UNIVERSITY OF MUNICH), Christoph Anton; MAZZAFERRO (RZG), Luca (Rechenzentrum Garching (RZG)); TSIGENOV (AACHEN), Oleg

Track Classification : Big Data and Storage Systems

Contribution ID : **37** Type : **not specified**

Security Workshop

In this security workshop the participants will change ends and take the role of a hacker attacking servers and services within a virtualized environment. We focuses on common real-life vulnerabilities and attacks - the ones that have great impact on both company networks and individuals using the Internet.

Every part of the workshop starts with a condensed introduction of the basics of the topic. We present vulnerabilities, exploits, and tools. After that, it's your turn! You have the opportunity to replay our demos and explore further techniques and possibilities of the exploit tools. Finally, you can attack and try to "pwn" servers with varying levels of difficulty in our lab environment. At the end of every unit we will discuss your findings and experiences together. This will lead to interesting insights on how to better protect yourself and your network.

During the workshop you will play with different web applications waiting to be hacked. Many web apps have striking bugs that in real-life threaten the data of millions of users. You will learn about SQL injection, scripting issues, request forgery and more.

Encrypted connections like HTTPS/SSL are safe, aren't they? Unfortunately, reality is not that easy: You will conduct an active man-in-the-middle attack and manipulate even encrypted connections to obtain the clear text of the conversation. There are powerful tools available that make man-in-the-middle attacks easy.

Finally, you will explore and use the Metasploit Framework, a tool that aids the hacker at choosing and running exploits against one or many targets.
Requirements for participants

The workshop targets everyone interested in IT security who wants to extend his knowledge by hacking vulnerable applications and playing with exploit tools. You should be familiar with the Unix command line and the concept of manpages. A thorough understanding of common web technologies and the ability to read scripting languages is necessary. Basic knowledge of TCP/IP and network services is also recommended.

The participants are required to bring their own device, preferably a laptop running some kind of Linux/Unix, but Windows-based computers are fine too.

Summary

Presenter(s) : SPECHT (GENUA), Raimund

Contribution ID : **38** Type : **not specified**

Getting started with Android and App Engine

This workshop is for Java developers, that want to get started with Android development. It covers the basics in Android programming and usage of the new Android build system. You will create your first application during the workshop and will create a simple cloud backend for synchronization of your data. We will learn about basic Android concepts like Activities, Services or the Android resource system. Since this workshop targets total Android beginners, we won't cover topics as native (C/C++) coding in Android or responsive user-interface designs.
Requirements for participation

Basic programming knowledge in Java is required to attend the workshop. You should know the following concepts and be able to implement them:

```
Coding basics (if, switch, loops, ...)
Object oriented programming and patterns:
    classes and objects
    static
    inner classes
    anonymous classes
    generic classes
You do NOT need any knowledge in Android programming.
```

What should you prepare for the workshop?

You should have your laptop with installed software for applications development. Also if possible bring your Android phone.

We will send all attendees an email around 2 - 4 weeks before the workshop with additional information on what software you should install beforehand.

Summary

Presenter(s) : ROES (INOVEX GMBH), Tim

Track Classification : Modern Programming

Contribution ID : **39** Type : **not specified**

Concurrent Programming in C++

In this course we will introduce how to program for concurrency in C++, taking advantage of modern CPUs ability to run multi-threaded programs on different CPU cores. Firstly, we will explore the new concurrency features of C++11 itself, which will also serve as a general introduction to multi-threaded programming. Students will learn the basics of asynchronous execution, thread spawning, management and synchronisation. Some elementary considerations about deadlocks and data races will be introduced, which will illustrate the common problems that can arise when programming with multiple threads. After this the Threaded Building Block template library will be introduced. We shall see how the features of this library allow programers to exploit multi-threading at a higher level, not needing to worry about so many of the details of thread management.

Students should be familiar with C++ and the standard template library. Some familiarity with makefiles would be useful.

Summary

Presenter(s) : Dr. STEWART (CERN), Graeme

Track Classification : Modern Programming

Contribution ID : **40** Type : **not specified**

Microsoft Azure Cloud Computing Workshop

Microsoft Azure is a general, open, and flexible global cloud platform supporting any language, tool, or framework - including Linux, Java, Python, and other non-Microsoft technologies. It is ideally suited to researchers' needs across disciplines. The workshop is intended specifically for active scientists who can code, who will soon code, or are interested in coding in a modern computing context.

Attendees will be able to access Microsoft Azure on their own laptop during the training and for evaluation purposes for up to six months after the event. The attendee's laptop does not need to have the Windows operating system installed—Microsoft Azure is accessed via your Internet browser.

This workshop will allow you to :

```
Gain an understanding of cloud computing and why and when you would use it in scientific or oth
Acquire hands-on experience in the major design patterns for successful cloud applications, inc
Develop the skills to run your own application/services on Microsoft Azure
```

Summary

Presenter(s) : Dr. TAKEDA (MICROSOFT), Kenji

Track Classification : Cloud Computing

Contribution ID : **41** Type : **not specified**

ROOT 6 Workshop

ROOT is the software framework used in High Energy Physics and other Big Data environments to store, statistically analyze and visualize large amounts of data in a reliable, efficient way. The new major release ROOT 6, published right before the school, brings several major improvements. ROOT 6 is expected to be the standard ROOT version for instance for ATLAS, CMS and LHCb for Run 2. Its new interpreter cling replaces CINT; it adds support for C++11, drastically improves error messages even compared to GCC and fixes the use of templates. It enables for instance a much simplified TTree access called TTreeReader only available in ROOT 6. Further major improvements are in the graphics and math area. This GridKa School workshop will be the first ever ROOT 6 tutorial, focusing on the improvements since ROOT 5 but also giving a general introduction to data analysis with ROOT.

Summary

Presenter(s) : Dr. NAUMANN (CERN), Axel

Track Classification : Modern Programming

Contribution ID : **42** Type : **not specified**

CUDA GPU Programming Workshop

While the computing community is racing to build tools and libraries to ease the use of 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 workshop is designed to introduce the CUDA programming language, through examples and hands-on exercises so as to enable the user to recognize CUDA friendly algorithms and completely exploit the computing potential of a heterogeneous parallel system.

Summary

Primary author(s) : Mr. PANTALEO (UNIVERSITY OF PISA), Felice

Presenter(s) : Mr. PANTALEO (UNIVERSITY OF PISA), Felice

Track Classification : Modern Programming

Contribution ID : **43** Type : **not specified**

Workshop – Introduction to HPC for Life-Science Researchers

While the percentage of females in computer science and other technical areas is still relatively small, in the life- and bio sciences, females comprise 50% and more of students and early-stage researchers. This, in combination with the trend of an ever increasing application of HPC (high performance computing) in these "non-traditional" fields of health, bio- and life-sciences, leads to a dire need to bring women into the field of HPC.

To address this situation, the DFG-funded DASH project http://www.dash-project.org hosts an introductory HPC workshop targeted at female early career researchers from the life sciences, health and bio-sciences. The workshop will provide the participants with an introduction to high performance computing, covering computing platforms and parallel programming. We will also invite experts to give a talk about the gender issues, especially career for women. Other topics will be included based on the interest of the participants.

Funding and Support
This workshop is supported by the gender incentives program of the German Priority Program "Software for Exascale Computing" (SPPEXA) funded by the German Research Foundation (DFG).
While everyone interested in the workshop is welcome to attend, we provide up to ten stipends (for female participants only) for travel assistance of the participants and the registration payment if the participants register GridKa School for joining other events. Please go to the workshop homepage for stipend application.

Summary

Presenter(s) : Dr. TAO (KIT), Jie

Track Classification : Cloud Computing

Contribution ID : **45** Type : **not specified**

Welcome to Karlsruhe Institute of Technology

Monday, 1 September 2014 14:00 (15)

Summary

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

Session Classification : Plenary talks

Contribution ID : **46** Type : **not specified**

GridKa School - Event Overview

Monday, 1 September 2014 14:15 (15)

Summary

Presenter(s) : Dr. WEBER (KIT-SCC), Pavel

Session Classification : Plenary talks