GridKa School 2015: Big Data, Virtualization and Modern Programming

GridKa School 2015
Big Data Virtualization Modern Programming

Report of Contributions

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

Hadoop in Complex Systems Research

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.'

Presenter: KÄMPF (CLOUDERA), Mirko

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

Parallel Programming using FastFlow

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)

Presenter: Dr TORQUATI (UNIVERSITY OF PISA), Massimo

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

Welcome to Karlsruhe Institute of Technology

Monday, September 7, 2015 2:00 PM (30 minutes)

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

Session Classification: Plenary talks

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

GridKa School - Event Overview

Monday, September 7, 2015 2:30 PM (15 minutes)

Presenter: Dr HARTMANN, Thomas (SCC)

Session Classification: Plenary talks

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

Multi-core Computing in High Energy Physics

Presenter: Dr HEGNER (CERN), Benedikt

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

Identity challenges in a Big Data world

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.

Presenter: Dr MILLAR (DESY), Paul

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

Welcome to Karlsruhe Institute of Technology

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

GridKa School - Event Overview

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

The Icosahedral Nonhydrostatic (ICON) model: Scalability on Massively Parallel Computer Architectures

Friday, September 11, 2015 10:15 AM (45 minutes)

Simulation in numerical weather prediction and climate forecasting has a fast-growing demand for memory capacity and processing speed. For the last decade, however, computer technology has shifted towards multi-core chip designs while at the same time on-chip clock rates have increased only moderately. The parallel implementation of DWD's operational forecast model ICON therefore follows a hybrid distributed/shared memory approach, based on the Message Passing Interface (MPI) and the OpenMP API.

The ICON code couples the different components of the earth system model, e.g. dynamics, soil and radiation, with high-level language constructs. Its communication characteristics and programming patterns take the unstructured triangular grid into account and are designed to meet the main challenges in high performance computing, i.e. load balancing, cache efficiency, and low-latency networking. The implementation employs special domain decomposition heuristics, parallel range-searching algorithms, and makes use of asynchronous I/O servers to deal with the potentially prohibitive amount of data generated by earth system models. This facilitates the ICON code to extract an adequate level of performance on a wide range of HPC platforms, targeting large scalar cluster systems with thousands of cores as well as vector computers.

Author: PRILL, Florian (DWD)

Presenter: PRILL, Florian (DWD)

Session Classification: Plenary talks

Track Classification: Simulation

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

Docker for a ROOT based data analysis flow

Wednesday, September 9, 2015 1:00 PM (5 hours)

Linux containers (LXC) is a technology that provides operating system-level virtualisation not via a virtual machines but rather by using a single kernel to run multiple instances on the same OS. Linux namespaces and control groups (cgroups) represent the foundation on which LXC are built. Containers are fast to deploy, they introduce no overhead or indirection as in the case of traditional virtual machines and also have the added design benefit of ensuring complete isolation between processes. Containers are great for running multiple instances of the same service in parallel either as part of a scaling out strategy or just for testing purposes. Docker is built around Linux containers and offers an intuitive way of managing them by abstracting and automating some of the configuration details. Besides being an open-source project, Docker has enabled the development of an entire "ecosystem" of tools and products targeting container technology.

Author: Mr SINDRILARU, Elvin (CERN (CH))

Presenter: Mr SINDRILARU, Elvin (CERN (CH))

Session Classification: Docker for a ROOT based data analysis flow

Track Classification: Virtualization

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

Linux containers and Docker

Wednesday, September 9, 2015 9:40 AM (40 minutes)

Linux containers (LXC) is a technology that provides operating system-level virtualisation not via a virtual machines but rather by using a single kernel to run multiple instances on the same OS. Linux namespaces and control groups (cgroups) represent the foundation on which LXC are built. Containers are fast to deploy, they introduce no overhead or indirection as in the case of traditional virtual machines and also have the added design benefit of ensuring complete isolation between processes. Containers are great for running multiple instances of the same service in parallel either as part of a scaling out strategy or just for testing purposes. Docker is built around Linux containers and offers an intuitive way of managing them by abstracting and automating some of the configuration details. Besides being an open-source project, Docker has enabled the development of an entire "ecosystem" of tools and products targeting container technology.

Author: SINDRILARU, Elvin (CERN)

Presenter: SINDRILARU, Elvin (CERN)

Session Classification: Plenary talks

Track Classification: Virtualization

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

HPCCloud 101: HPC on Cloud Computing for newcomers

Tuesday, September 8, 2015 1:00 PM (5 hours)

Never been into Cloud Computing before? Do you think that an extra computing power is crucial for your research? Do you have some neat parallel codes that your institution doesn't allow you to execute because the cluster is full? Maybe this tutorial is for you!

<p/>

The tutorial will cover the following topics:

 Infrastructure as a Service clouds (user level) with OpenNebula (http://opennebula.org/) and Amazon EC2 (http://aws.amazon.com/es/ec2/). Virtual Clusters on cloud with Star-Cluster (http://star.mit.edu/cluster/).

As Virtual Clusters deployed by StarCluster have Sun Grid Engine and OpenMPI installed you are more than welcome to bring your own codes and give them a try!

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

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

Session Classification: HPC Cloud 101: HPC on Cloud Computing for newcomers

Track Classification: Virtualization

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

Linux containers and Docker

Linux containers (LXC) is a technology that provides operating system-level virtualisation not via a virtual machines but rather by using a single kernel to run multiple instances on the same OS. Linux namespaces and control groups (cgroups) represent the foundation on which LXC are built. Containers are fast to deploy, they introduce no overhead or indirection as in the case of traditional virtual machines and also have the added design benefit of ensuring complete isolation between processes. Containers are great for running multiple instances of the same service in parallel either as part of a scaling out strategy or just for testing purposes. Docker is built around Linux containers and offers an intuitive way of managing them by abstracting and automating some of the configuration details. Besides being an open-source project, Docker has enabled the development of an entire "ecosystem" of tools and products targeting container technology.

Author: Mr SINDRILARU, Elvin (CERN (CH))

Presenter: Mr SINDRILARU, Elvin (CERN (CH))

Track Classification: Virtualization

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

From Mars to Earth through Cloud Computing

Monday, September 7, 2015 4:45 PM (45 minutes)

Our society has benefited from Space exploration in many ways. Many of the inventions we use nowadays have their origin in or have been improved by Space research. Computer Science is not an exception.

This talk will introduce the application of Cloud Computing done by the speaker in the context of different Mars missions: Mars MetNet (Spain-Russia-Finland), MSL Curiosity (NASA) and Exo-Mars2016 (ESA). The achieved know-how allowed the optimization of other areas at Planet Earth, such as weather forecast and agricultural wireless sensor networks processing.

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: Plenary Talks

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

IT security in an IPv6 world

Monday, September 7, 2015 4:00 PM (45 minutes)

Unused IPv4 network addresses are a scarce resource. The
deployment of IPv6 networking across the world is well underway. Some
large IT distributed infrastructures, such as the Worldwide Large Hadron
Collider Computing Grid, are starting to deploy dual-stack IPv6/IPv4
services to support IPv6-only clients. New networking protocols, such as
IPv6, always bring new challenges for operational IT security. We have
spent many decades understanding and fixing security problems and concerns
in the IPv4 world. We have only just started with IPv6! Its lack of
maturity together with all the additional complexities, particularly in a
dual-stack environment, bring many challenges. This talk will consider
some of the security concerns in an IPv6 world and consider best practices
for system administrators who manage (or will manage) IT services on
distributed infrastructures and also for their related security teams.

Author: Dr KELSEY, David (Rutherford Appleton Laboratory,)

Presenter: Dr KELSEY, David (Rutherford Appleton Laboratory,)

Session Classification: Plenary talks

Track Classification: Plenary Talks

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

Apache Spark: The next Generation of Hadoop Processing

Thursday, September 10, 2015 9:00 AM (40 minutes)

Apache Spark is known as the "Next Generation Framework" of Hadoop based data processing. Why, and what Apache Spark offers to the scientific community is explained in this talk. The convergence of different analysis techniques into one flexible and highly efficient processing engine allows completely new interdisciplinary analysis methods beside cheap analysis prototypes. In this presentation I shown examples in Scala and Python. Beside fundamental techniques and the core features of Apache Spark we look into development practices and data analysis techniques. Therefore we recap the theoretical background about Map-Reduce- and Bulk-Synchronous-Parallel processing before I introduce the machine learning library MLlib and the graph processing framework GraphX. Apache Spark uses the concept of data frames, and allows SQL operations on data sets, after this presentation you know how this works and how you can save a lot of time. Finally, you can see how data can be collected and analyzed on the fly, using Spark Streaming.

Author: Mr KAEMPF, Mirko (Cloudera)

Presenter: Mr KAEMPF, Mirko (Cloudera)

Session Classification: Plenary talks

Track Classification: Plenary Talks

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

Apache Spark in Scientific Applications [B]

Thursday, September 10, 2015 1:00 PM (5 hours)

This tutorial is limited to 12 participants. Another session of this tutorial is also available
</p>

The workshop Spark in Scientific Applications covers fundamentale development and data analysis techniques using Apache Hadoop and Apache Spark. Beside an introduction into the theoretical background about Map-Reduce- and Bulk-Synchronous-Parallel processing, also the machine learning library MLlib and the graph processing framework GraphX are used.
</p>
We work on sample data sets from Wikipedia, financial market data, and from a generic data generator. During the tutorial sessions we illustrate the Data Science Workflow and present the right tools for the right task.
</p>
All practical exercises are well prepared in a pre-configured virtual machine. Participants get access to required data sets on a „one node pseudo-distributed"cluster with all tools inside. This VM is also a starting point for further experiments after the workshop.

Author: Mr MIRKO, Kaempf (Cloudera)

Presenter: Mr MIRKO, Kaempf (Cloudera)

Session Classification: Apache Spark in Scientific Applications

Track Classification: Data Analysis

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

Puppet Workshop

Tuesday, September 8, 2015 1:00 PM (5 hours)

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.

<p/>

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

<p/>

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.

<p/>

2nd day:
• Leftovers from previous day, and/or some more advanced configuration
• Series of small presentations and walk-throughs: Hiera, Facter, Foreman, MCollective, GitLab, …

<p/>

Prerequisites:
• Attendants should familiarize themselves with a Linux terminal and the peculiarities of a Linux text editor (vi, emacs etc.).
• No knowledge of Puppet or Git is required.

Author: Mr STERNBERGER, Sven (DESY)

Co-authors: Mr JONES, Ben (CERN (CH)); Mr KEMP, Yves (DESY)

Presenters: Mr JONES, Ben (CERN (CH)); Mr STERNBERGER, Sven (DESY); Mr KEMP, Yves (DESY)

Session Classification: Puppet

Track Classification: Data Center Management

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

Software-Defined Networking for the Data Center

Tuesday, September 8, 2015 9:40 AM (40 minutes)

Software Defined Networking (SDN) is a paradigm shift in the networking domain. It has recently become a hot topic and it is expected to change the way we think about networks and how we architect them. In this talk we will look at what SDN is, how it can be realized, and what the impact on networking, mainly in the data center, is. The SDN architecture will be explained, the abstractions used by OpenFlow will be introduced, and some use cases as well as some SDN implementations will be described.

Author: Dr HASSELMEYER, Peer (NEC)

Presenter: Dr HASSELMEYER, Peer (NEC)

Session Classification: Plenary talks

Track Classification: Data Center Management

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

AngularJS workshop

Thursday, September 10, 2015 1:00 PM (5 hours)

This workshop will focus on creating modern web applications and store their data into the cloud. AngularJS is a framework that has been growing popular during the last years, due to its flexibility, its power to build rich web applications and yet its ease to use.
</p>
During this workshop we will build a web application from scratch and we'll connect it to the cloud
to easily sync your (users) data over the web.
</p>
Participants of this workshop are required to bring their own notebook.
If possible install Node.js from https://nodejs.org in preparation.
To work you will also require a text editor of your choice (we recommend SublimeText from http://www.sublimetext.com).
</p>
Since the workshop targets AngularJS beginners, you don't need any experience in AngularJS (or even have heard about it).
But to follow the workshop you will need at least beginners knowledge in JavaScript (or optional TypeScript).

Author: Mr ROES, Tim (Inovex)

Presenters: REUTER, Matthias (Inovex); Mr ROES, Tim (Inovex)

Session Classification: AngularJS

Track Classification: Programming Techniques

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

SDN: Software-Defined Networks

Thursday, September 10, 2015 1:00 PM (5 hours)

Today's communication networks are designed around the original mechanisms of Ethernet and TCP/IP. Because of the success of these early technologies, networks grew bigger and more complex, which led to a need for more complex control options, such as VLANs and ACLs. A variety of heterogeneous network appliances such as firewalls, load balancers, IDS, optimizers, and so on, each implement their own proprietary control stack. Reciprocal communication is handled by other complex protocols such as Spanning Tree, Shortest Path Bridging, Border Gateway, or similar. Each additional component thus increases the complexity and complicates integrated network management. The consequences are often low network utilization, poor manageability, lack of control options in cross-network configurations, and vendor lock-in.

<p/>

One way out of this dilemma is Software Defined Networks (SDNs) and OpenFlow. OpenFlow is an Open Networking Foundation (ONF) standard protocol that abstracts the complex details of a fast and efficient switching architecture. Today, OpenFlow offers an open control interface that is now implemented in hardware by all major network component manufacturers. Several vendors even offer software switches that support virtualized datacenters. OpenFlow also supports the concept of separating the data and control paths, which lets a central control point oversee a variety of OpenFlow-enabled network components. The SDN controller could even be a distributed application to provide additional security, fault-tolerance, or load balancing.

<p/>

This presentation focuses on a general introduction to Software Defined Networking and OpenFlow. We shed light on various aspects of today's network management and its challenges and elaborate on possible solutions offered by SDN. Moreover, the hands-on tutorial addresses the OpenDaylight SDN controller. To this end, we install, configure, and run OpenDaylight. We emulate a small network using the MiniNet Network Emulator and have OpenDaylight manage the data flows in that network. We will experience the beauty of such a centralized solution and discuss further areas of application, such as cloud computing and OpenStack, for instance.

Author: Mr MICHAEL, Bredel (FH Kufstein)

Presenter: Mr MICHAEL, Bredel (FH Kufstein)

Session Classification: Software Defined Networks

Track Classification: Data Center Management

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

Data Preservation

Tuesday, September 8, 2015 9:00 AM (40 minutes)

The long-term preservation of information is a crucial requirement for
scientific progress in every research community. In former times hammer
and chisel were the tools of choice to preserve the cultural heritage,
nowadays the digital world introduces additional and novel challenges.
Obsolete formats and technologies, a quick decay of storage media or power
outages are just a few examples of threats that need to be faced in order
to ensure sustainable access to valuable data.
</p>
This talk will give an introduction into the broad field of data
preservation and its basic principles. Various examples from the arts and
humanities community will illustrate how heterogeneous preservation
demands are and how they can be supported by a research infrastructure.

Author: Ms TONNE, Danah (KIT)

Presenter: Ms TONNE, Danah (KIT)

Session Classification: Plenary talks

Track Classification: Plenary Talks

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

Big data in critical infrastructure: Production and failover infrastructure in DWD's central data management

Thursday, September 10, 2015 10:40 AM (40 minutes)

The German Weather Service (DWD) provides a wide variety of services for
the protection of life and property in the form of weather and climate
information. One core task is safeguarding aviation, marine safety and
terrestrial traffic. Another is warning before meteorological events
that could endanger public safety and order. Additionally, we monitor
the climate and are active in multiple research fields, from ensemble
numerical weather forecasting to applications of weather data in new
areas. Data is recorded, processed and transformed into time-critical
products and securely archived 24 hours per day, 365 days per year.
</p>
The DWD maintains a high productivity, redundant infrastructure in order
to provide these services reliably and on demand. We ensure
deliverability with multiple tiers of failover strategies, enabling us
to manage and monitor production even when faced with major hardware or
software failures.
</p>
Specialized systems allow rapid access to large, cross-sectional binary
files in file system caches for near-real-time applications, while an
automated tape archive provides short-term access to long-term archival
data. Simultaneously, observational data is processed and stored in
relational databases in order to allow comfortable processing of long
time series data. Various application layers are used to post-process
products in order to refine them for domain-specific queries.
</p>
Demands for weather and climate based data and services, as well as the
associated needs for processing power, network transfer capabilities and
storage capacity are constantly increasing. It is the DWD's goal not
only to maintain a production infrastructure with high quality and
availability, but also to continue to evolve to meet these demands.
Doing so while maintaining our tradition of quality, speed and
reliability is one of the major challenges facing the DWD. Some current
projects designed to meet these goals are introduced in the outlook.

Author: Mr LEE, Daniel (DWD)

Presenter: Mr LEE, Daniel (DWD)

Session Classification: Plenary talks

Track Classification: Plenary Talks

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

Climate simulations

Wednesday, September 9, 2015 9:00 AM (40 minutes)

Climate change as a result of increased emissions of greenhouse gases is a global scale challenge for
today's society and future generations. Climate model simulations are important tools to test our
scientific knowledge of the processes involved, and to provide projections of future changes and their
impacts. After a general introduction this talk will focus on recent advances in atmospheric
chemistry-climate modelling, including a discussion of the technical challenges of climate model
simulations.

Author: Dr SINNHUBER, Björn-Martin (KIT/IMK-ASF)

Presenter: Dr SINNHUBER, Björn-Martin (KIT/IMK-ASF)

Session Classification: Plenary talks

Track Classification: Plenary Talks

Contribution ID: **38**

Type: **not specified**

Mongo DB Tutorial

Thursday, September 10, 2015 1:00 PM (5 hours)

This session is an introduction to a particular NoSQL database, MongoDB.<p/>
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.

<p/>

This session will provide theoretical basic knowledge about Mongo and support it with hands-on
activities to get to know Mongo in practice.

<p/>

The agenda will cover the followings:

 Getting familiar with Mongo terminologies Executing CRUD operations
Indexing Schema design Use of Mongo to make a small web application
Authentication and authorization possibilities Getting to know replication and Shard-
ing mechanisms (optional) Analyzing data stored in Mongo using R

<p/>

Basic Linux knowledge and some background knowledge about relational databases will be helpful
in this session, but is not mandatory.

Author: Ms AMERI, Parinaz (KIT/SCC)

Co-author: Mr SZUBA, Marek (KIT\SCC)

Presenters: Mr SZUBA, Marek (KIT\SCC); Ms AMERI, Parinaz (KIT/SCC)

Session Classification: MongoDB

Track Classification: Big Data

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

R for Large Scale Data Analysis

Wednesday, September 9, 2015 1:00 PM (5 hours)

The R programming language is a free software environment for statistical computing and graphics. It is widely used among statisticians and data miners.
But especially the huge variety of available packages will make the introduction into daily business far from easy. This tutorial focuses on using R for large amounts of data. It deals with three different topics: managing data, analysing data, as well as basic and intermediate plotting. Each topic is accompanied by a short introduction, overview of experiences, as well as recommended packages. The tutorial itself is hands-on. We will look at different possibilities and solutions.
The tutorial targets participants who already have some basic experiences with programming but do not necessarily know much about R.

Author: Ms KÜHN, Eileen (KIT\SCC)

Presenter: Ms KÜHN, Eileen (KIT\SCC)

Session Classification: R for Large Scale Data Analysis

Track Classification: Data Analysis

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

FastFlow: Parallel Programming using parallel patterns and the FastFlow frameworks

Thursday, September 10, 2015 1:00 PM (5 hours)

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.
</p>
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, stencil-reduce
and we will see how to mix stream and data-parallel patterns to build parallel applications and algorithms.
Different possible implementations will be discussed and tested. Participants will have the opportunity
to implement multi-threading algorithms and simple benchmarks to evaluate performance (considering
also energy consumption).
</p>
Desirable prerequisite:

 Good knowledge of C programming Knowledge of multi-threading programming and concurrency problems. Knowledge of C++ templates. Features of C++11 standard will be also used. Basic Knowledge of OpenCL.

</p>
Expected participants: 10/15

Author: Mr TORQUATI, Massimo (Universita di Pisa)

Presenter: Mr TORQUATI, Massimo (Universita di Pisa)

Session Classification: FastFlow: Parallel Programming using parallel patterns and the Fast-Flow frameworks

Track Classification: Programming Techniques

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

Concurrent Programming in C++

Wednesday, September 9, 2015 1:00 PM (4 hours)

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

Author: Mr STEWART, Graeme (CERN (CH))

Presenter: Mr STEWART, Graeme (CERN (CH))

Session Classification: Concurrent Programming in C++

Track Classification: Programming Techniques

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

Programming Templates Tutorial

Tuesday, September 8, 2015 1:00 PM (5 hours)

Programming Templates

Author: Dr HECK, Martin (KIT/EKP)

Presenter: Dr HECK, Martin (KIT/EKP)

Session Classification: Programming Templates

Track Classification: Programming Techniques

Contribution ID: **44**   Type: **not specified**

CEPH Tutorial

Wednesday, September 9, 2015 1:00 PM (5 hours)

Ceph is an open-source, software-defined distributed storage system that strives to achieve scalability and reliability through an innovative decentralised design.
</p>
Distributed file systems nowadays face multiple challenges: scaling to peta-byte capacity and providing high performance, while protecting against failures. Moreover, file systems should be able to adapt to dynamic distributed workloads to provide the best performance.
Ceph tries to tackle these challenges with a completely decentralised architecture that has no single point of failure. Reliability is achieved
through distributed data placement and replication. Ceph's dynamic metadata partitioning feature helps deal with dynamic workloads.
</p>
This session is an introduction to Ceph and it will cover theoretical background on Ceph's architecture, as well as hands-on exercises, such as installation and configuration of a Ceph cluster, simple usage and monitoring. After completing this session, you should be able to understand and discuss Ceph concepts, and deploy and manage a Ceph Storage Cluster.
</p>
Basic knowledge of Linux and storage concepts is required.

Author: Ms GUDU, Diana (KIT\SCC)

Presenter: Ms GUDU, Diana (KIT\SCC)

Session Classification: CEPH

Track Classification: Big Data

Contribution ID: 45

Type: **not specified**

Software Defined Data Center

Thursday, September 10, 2015 1:00 PM (5 hours)

Running traditional data centers, engineers have to face many challenges, such as running multiple different workloads. In this workshop we will have a look at such a data center and identify the challenges that have to be faced when using these architectures. After this we look at a basic use case and implement it for a traditional data center. In the next step we will have a look at new technologies for Software Defined Data Centers (SDDC). This enables us to compare how these new technologies cope with the problems found earlier and help make data centers more flexible.

<p/>

A big benefit of SDDCs is running dynamic and flexible workloads while archiving high resource utilization.

<p/>

A SDDC contains these (and more) technologies:

Software defined Networking (SDN)
Software defined Storage (SDS)
Data Center Operating System (DCOS)

<p/>

The goal of this workshop is to build your own mini SDDC with a reference software stack based on:

CentOS / CoreOS
Docker
Mesos and Mesosphere
(Quobyte)
(OpenVswitch)

<p/>

Requirement for participation:

Basic knowledge of data centers

An additional tutorial for deepening the topic is available on Friday
http://indico.kit.edu/indico/event/89/session/35/contribution/53

Author: Mr SCHEUERMANN, Johannes (Inovex\KIT)

Presenter: Mr SCHEUERMANN, Johannes (Inovex\KIT)

Session Classification: Software Defined Data Center

Track Classification: Data Center Management

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

CUDA GPU Programming Workshop

Tuesday, September 8, 2015 1:00 PM (5 hours)

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.
<\p>
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.

Author: PANTALEO, Felice (CERN)

Presenter: PANTALEO, Felice (CERN)

Session Classification: CUDA GPU Programming Workshop

Track Classification: Programming Techniques

Contribution ID: **47**   Type: **not specified**

Application development with relational and non-relational databases

Tuesday, September 8, 2015 1:00 PM (5 hours)

In this workshop, the students will learn how to use relational and non-relational databases to build multi-threaded applications. The focus of the workshop is to teach efficient, safe, and fault-tolerant principles when dealing with high-volume and high-throughput database scenarios.
</p>
A basic understanding of the following things is required:</br>
- A programming language (preferably Python or any C-like)</br>
- Basic SQL (CREATE, DROP, SELECT, UPDATE, DELETE)</br>
- Linux shell scripting (bash or zsh)</br>
</p>
The course will cover the following three topics:
</p>
- When to use relational databases, and when not</br>
* Relational primer</br>
* Non-relational primer</br>
* How to design the data model</br>
</p>
- Using SQL for fun and profit</br>
* Query plans and performance analysis</br>
* Transactional safety in multi-threaded environments</br>
* How to deal with large amounts of sparse metadata</br>
* Competetive locking and selection strategies</br>
</p>
- Building a fault-tolerant database application</br>
* Distributed transactions across relational and non-relational databases</br>
* SQL injection and forceful breakage</br>
* Application-level mitigation for unexpected database issues</br>

Author: Dr LASSNIG, Mario (CERN)

Presenter: Dr LASSNIG, Mario (CERN)

Session Classification: Application development with relational and non-relational databases

Track Classification: Big Data

Contribution ID: 48

Type: **not specified**

Apache Spark in Scientific Applications [A]

Wednesday, September 9, 2015 1:00 PM (5 hours)

This tutorial is limited to 12 participants. Another session of this tutorial is also available
</p>
The workshop Spark in Scientific Applications covers fundamentale development and data analysis techniques using Apache Hadoop and Apache Spark. Beside an introduction into the theoretical background about Map-Reduce- and Bulk-Synchronous-Parallel processing, also the machine learning library MLlib and the graph processing framework GraphX are used.
</p>
We work on sample data sets from Wikipedia, financial market data, and from a generic data generator. During the tutorial sessions we illustrate the Data Science Workflow and present the right tools for the right task.
</p>
All practical exercises are well prepared in a pre-configured virtual machine. Participants get access to required data sets on a „one node pseudo-distributed"cluster with all tools inside. This VM is also a starting point for further experiments after the workshop.

Author: Mr KÄMPF, Mirko (Cloudera)

Presenter: Mr KÄMPF, Mirko (Cloudera)

Session Classification: Apache Spark in Scientific Applications

Track Classification: Big Data

Contribution ID: **49**   Type: **not specified**

Scientific Python

Thursday, September 10, 2015 1:00 PM (5 hours)

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 additional available scientific packages like NumPy, SciPy, pandas, matplotlib, scikit-learn, etc., Python becomes a very well suited programming language for data analysis. This hands-on session aims towards advanced Python beginners, who have already gained some knowledge about Python (Scripting experience and knowing the term list comprehension should be sufficient). This course gives an introduction and demonstrates the power of Python in data analysis using NumPy and pandas.

Author: Dr GIFFELS, Manuel (KIT\EKP)

Presenter: Dr GIFFELS, Manuel (KIT\EKP)

Session Classification: Python

Track Classification: Data Analysis

Contribution ID: **50**   Type: **not specified**

Research Data Alliance - Research Data Sharing without barriers

Thursday, September 10, 2015 11:20 AM (40 minutes)

Even examples from Psycholinguistics –a humanities discipline –show that data intensive science is changing all scientific disciplines dramatically posing unprecedented challenges in data management and processing. A recent survey in Europe showed clearly that most of the research departments are not prepared for this step and that the methods that are used to manage, curate and process data are inefficient and too costly. Despite a wide agreement on some obvious trends with respect to data and on principles about data sharing such as those formulated by the G8 ministers, we lack clear guidelines and strategies of how to move ahead.

Therefore, Research Data Alliance as a bottom-up organized global and cross-disciplinary initiative has been established to accelerate the process of changing data practice. After only two years RDA produced its first concrete results, which have to demonstrate their practicality. In particular the infrastructure builders are requested to act as early adopters. RDA as an initiative to specify interfaces, protocols, guidelines, etc. needs to be seen as a chance for us to discuss how we can move ahead. Infrastructure builders need to put results in place to test the results. All three - researchers, infrastructure builders and RDA experts - need to remain in a close discussion process to achieve the fast progress we are waiting for.

The talk will address all aspects which have been mentioned.

Author: Mr WITTENBURG, Peter (RDA)

Presenter: Mr WITTENBURG, Peter (RDA)

Session Classification: Plenary talks

Track Classification: Plenary Talks

Contribution ID: **51**   Type: **not specified**

EUDAT - The European Data Infrastructure

Friday, September 11, 2015 9:00 AM (45 minutes)

This talk will provide an overview of the EUDAT initiative which has laid out the foundation of a new Collaborative Data Infrastructure (CDI) providing solutions for finding, sharing, preserving and performing computations with primary and secondary research data on a pan-European level. By addressing the accelerated proliferation of data and the resulting challenges faced by the research communities through a cross-disciplinary approach, and by identifying and proposing solutions to barriers to the development of an efficient pan-European e-infrastructure ecosystem, research e-infrastructures like EUDAT make concrete contributions to eliminating barriers to cross national and cross disciplinary collaboration and reinforcing the level playing field for European researchers and data managers.

Author: Dr LECARPENTIER, Damien (EUDAT)

Presenter: Dr LECARPENTIER, Damien (EUDAT)

Session Classification: Plenary talks

Track Classification: Plenary Talks

Contribution ID: **53**   Type: **not specified**

Software Defined Data Center in detail - Addon

Friday, September 11, 2015 1:00 PM (5 hours)

This is an additional session for deepening the tutorial on SDDC
http://indico.kit.edu/indico/event/89/session/35/contribution/45

basic knowledge of SDDCs as presented in the general tutorial are required

Author: Mr SCHEUERMANN, Johannes (Inovex)

Presenter: Mr SCHEUERMANN, Johannes (Inovex)

Session Classification: Software Defined Data Center

Track Classification: Data Center Management

Contribution ID: **54**
Type: **not specified**

Big Data and Data Protection

Big Data - a new challenge for privacy?
Besides improved techniques, e. g. Deep Learning, Big Data is characterised by huge volumes of data and numerous data categories. The algorithms and their results are not understandable for everyone. Furthermore, the data are analysed in different contexts. In contrast, the data protection laws require compliance with the key privacy principles as data minimization, purpose limitation and transparency. How can this conflict be resolved?
Some considerations of legal requirements and technical solutions will be presented.

Author: Mr MANNY, Klaus (LfD BW)

Presenter: Mr MANNY, Klaus (LfD BW)

Track Classification: Plenary Talks

Contribution ID: **55** Type: **not specified**

UNICORE Summit 2015

Monday, September 7, 2015 11:00 AM (6 hours)

UNICORE Summit 2015 homepage
</p>
The UNICORE Summit is the annual meeting of the UNICORE community. It provides a unique opportunity for UNICORE users, developers, administrators, researchers, service providers, and managers to meet.
</p>
Participate to share your experience, present recent and planned developments, learn about the latest UNICORE features, and get new ideas for interesting and prosperous collaborations.
</p>
Please register at http://unicore.eu/summit/2015/registration.php

Program overview

Session Classification: Unicore Summit

Track Classification: UNICORE Summit 2015

Contribution ID: **56**   Type: **not specified**

ElasticSearch and the ELK stack for monitoring and data analysis

Wednesday, September 9, 2015 10:40 AM (40 minutes)

Elasticsearch, Logstash and Kibana, known as the ELK stack, are three open source projects designed to ship, parse, search, analyze and visualize data, from Apache logs to Twitter streams. The Web-based Information Systems (WebIS) group of the Institute for Applied Computer Science (IAI) of the Karlsruhe Institute of Technology (KIT) uses the ELK stack in different large scale web information system projects as central components for data aggregation, analysis and search driven data access. Therefore, besides giving a rough overview on ELK features, this talk will explore possibilities and scenarios of using the ELK stack in web applications like community web portals, environmental information systems, smart energy management systems, or for log data analysis. Common data storage and analysis capabilities of ElasticSearch will be explained and examples given, how the ELK stack could be (programmatically) integrated into own software solutions.

Author: Dr DÜPMEIER, Clemens (KIT/IAI)

Presenter: Dr DÜPMEIER, Clemens (KIT/IAI)

Session Classification: Plenary talks

Track Classification: Plenary Talks

Contribution ID: **57**   Type: **not specified**

Elastic Search, Logstash and Kibana

Wednesday, September 9, 2015 1:00 PM (5 hours)

Elasticsearch, Logstash and Kibana, known as the ELK stack, are three open source projects designed to ship, parse, search, analyse and visualize your data, from Apache logs to Twitter streams. A short description of the components is the following:

 Logstash allows you to ship and parse your data using a great variety of plugins. It is highly scalable. Elasticsearch is a search server based on Apache Lucene. It is distributed and highly scalable. Kibana is the visualization platform available through a web browser with a nice interface and easy to customize directly from the browser.

In this course we will explain to you these three components and we will guide you through their installation and configuration. Several different data logs will be analyzed in order to finally create your own Kibana dashboards.

<p/>

Basic Linux knowledge and be familiar with vim is required. Some regular expressions knowledge would be a plus.

Authors: Mr PATHOMKEERATI, Kajorn (KIT/IAI); Mr AMBROJ PEREZ, Samuel (KIT/SCC)

Presenters: Mr PATHOMKEERATI, Kajorn (KIT/IAI);   Mr AMBROJ PEREZ, Samuel (KIT/SCC)

Session Classification: Elastic Search, Logstash, Kibana

Track Classification: Data Center Management

Contribution ID: **58**

Type: **not specified**

Puppet Workshop session 2

Wednesday, September 9, 2015 1:00 PM (5 hours)

Second session of the puppet workshop.
For details, please see the description of the first session

Session Classification: Puppet

Contribution ID: **59**   Type: **not specified**

Storage Technologies

Monday, September 7, 2015 2:45 PM (45 minutes)

Software-Defined Storage

Author: Dr MILLAR, Paul (DESY)

Presenter: Dr MILLAR, Paul (DESY)

Session Classification: Plenary talks

Track Classification: Plenary Talks

Contribution ID: **60**   Type: **not specified**

Digital Transformation, Big Data ⋯and the changing role of IT

Tuesday, September 8, 2015 10:40 AM (40 minutes)

How does the Digital Transformation change business models and which new business models arise? How do processes and business segments have to change and what is the role of IT within this development? To answer these questions we will look at the new technologies in a comprehensive way with a special focus on Big Data.
Get to know SAP as the global market leader for business software and see how SAP seizes new opportunities by leveraging technologies like cloud, in-memory, and mobile computing. Selected SAP customers of different size and from various industries show how to manage the digital change and remain competitive.
See also what possibilities SAP offers for graduates and professionals to work within those new technological fields.

<p/>

Links for further information:

 Keynotes on Big Data: http://events.sap.com/sapphirenow/en/home SAP HANA Could Platform for students: http://hcp.sap.com/students.html SAP Careers for students and graduates: https://www.sap.com/careers/index.html

Author: Dr QUACK, Engelbert (SAP SE, Head of Consulting Area Data & Technology)

Presenter: Dr QUACK, Engelbert (SAP SE, Head of Consulting Area Data & Technology)

Session Classification: Plenary talks

Track Classification: Plenary Talks

Contribution ID: **61** Type: **not specified**

SAP CodeJam

Joint the SAP CodeJam on Friday for hands-on experiance!<p/>
SAP CodeJam is a 5 to 6 hour hands-on coding and networking event where attendees share their knowledge and collaboratively develop with SAP technologies, platforms, and tools in a fun and casual environment. The events are developer community focused, supported by SAP, and explore technologies available through the Developer Center such as SAP HANA, Mobile, and Cloud.

<p/>

More details at http://scn.sap.com/community/events/codejam

Contribution ID: **62**   Type: **not specified**

Technical Computing on OpenPower

Thursday, September 10, 2015 9:40 AM (40 minutes)

Since 2013 the OpenPower Foundation grew steadily to over 130 Members so far. The goal of the OpenPower Foundation is to enable a joint development and integration of different technologies around the IBM Power CPU architecture to speed up innovation. Within the foundation, Technical computing (HPC, HTC) is a focus topic for several members like NVIDIA, Mellanox, IBM and others to enable accelerated computing based on OpenPower e.g. with GPGPUs or FPGAs. This talk will outline the technical computing future with OpenPower in DataCentric environments.

Author: Dr OBERST, Oliver (IBM)

Presenter: Dr OBERST, Oliver (IBM)

Session Classification: Plenary talks

Track Classification: Plenary Talks

Contribution ID: **63**   Type: **not specified**

IBM SPSS Data Mining Workshop

Friday, September 11, 2015 1:00 PM (5 hours)

This hands-on IBM SPSS Data Mining Workshop is an instructor-led session using IBM's data mining and predictive modeling software and is designed for those who are familiar with predictive analytics. Through this workshop you will experience first hand how IBM SPSS Modeler works and how easy it is to implement predictive analytics.

Introduction in Predictive Analytics<p/>
Exercise: IBM SPSS Modeler

 Predictive in 20 min. Association Modelling Segmentation Modelling Classification Modeling Deployment

Author: LEDDIN, Henrik (IBM)

Presenter: LEDDIN, Henrik (IBM)

Session Classification: IBM BootCamp

Track Classification: Big Data

Contribution ID: **64**
Type: **not specified**

Shinkansen or why trains can arrive on time

Tuesday, September 8, 2015 11:20 AM (40 minutes)

Big Data ist eines der treibenden Themen unserer Zeit. Wie sieht jedoch die Praxis aus?
HDS versucht in ihrem Vortrag den Spagat zwischen Theorie und der realen Anwendung zu schlagen.

Author: KREBS, Jürgen (Hitachi Data Systems)

Presenter: KREBS, Jürgen (Hitachi Data Systems)

Session Classification: Plenary talks

Track Classification: Big Data

Contribution ID: **65**   Type: **not specified**

Conclusions

Friday, September 11, 2015 11:00 AM (15 minutes)

Presenter: Dr HARTMANN, Thomas (SCC)

Session Classification: Plenary talks

Contribution ID: 66 Type: **not specified**

visit of the SCC CS computing centre

Tuesday, September 8, 2015 6:30 PM (1 hour)

For interested participants, we are organizing a short excursion to the SCC computing center at Campus South

Session Classification: School Social Event

Contribution ID: **67**   Type: **not specified**

Tarte Flambee

Tuesday, September 8, 2015 6:30 PM (3h 30m)

Social evening with Tarte Flambee, beer and drinks in the courtyard of building 30.22

Session Classification: School Social Event