PRACE 2017 Spring School joint event with VI-SEEM, Cyprus - System Administration and Data/Computational Services for Scientific Communities

Tuesday 25 April 2017 - Thursday 27 April 2017

The Cyprus Institute, Nicosia

Scientific Programme

The aim of the seasonal school is to serve the training needs of different persons of the PRACE, VI-SEEM and European communities and will thus have two parallel sessions - a Developer Track and a System Administration Track along with a Keynote Address.
Developer Track

The developer track focuses on various aspects which are of relevance to computational scientists from various disciplines. HPC compute sessions will teach participants about Accelerator Programming - and more specifically GPU and Xeon Phi programming. Data management sessions will teach participants about Big Data Projects, Data Management and Computation, Data Generation and Data Processing and Data Analysis and Processing Services. There will also be Data Visualisation sessions and presentation of case studies from the VI-SEEM and EUDAT projects. More details of sessions can be found below (missing descriptions will appear soon):
Introduction to Parallel and Scalable Machine Learning – Basics

Speaker: Morris Riedel - University of Iceland (short bio)

The first course part offers basics of analysing data with machine learning and data mining algorithms in order to understand foundations of learning from large quantities of data. This tutorial is especially oriented towards beginners that have no previous knowledge of machine learning techniques. The course consists of general methods for data analysis in order to understand clustering, classification, and regression. This includes a thorough discussion of test datasets, training datasets, and validation datasets required to learn from data with a high accuracy. Easy application examples will foster the theoretical course elements that also will illustrate problems like overfitting followed by mechanisms such as validation and regularisation that prevent such problems. The tutorial will start from a very simple application example in order to teach foundations like the role of features in data, linear seperability, or decision boundaries for machine learning models. In particular this course will point to key challenges in analysing large quantities of data sets (aka ‘big data’) in order to motivate the use of parallel and scalable machine learning algorithms. After this course participants will have a general understanding how to approach data analysis problems in a systematic way.
Introduction to Parallel and Scalable Machine Learning – Parallelization Benefits

Speaker: Morris Riedel - University of Iceland (short bio)

The second course part targets specific challenges in analysing large quantities of datasets that can not be analysed with traditional serial methods provided by tools such as R, SAS, or Matlab. This includes several challenges as part of the machine learning algorithms, the distribution of data, or the process of performing validation. The course will introduce selected solutions to overcome these challenges using parallel and scalable computing techniques based on the Message Passing Interface (MPI) and OpenMP that run on massively parallel High Performance Computing (HPC) platforms. Complementary to these solutions the course introduces algorithms and approaches that take advantage of parallel High Throughput Computing (HTC) resources and tools such as Apache Spark and Hadoop with map-reduce that are available in modern Cloud platforms today. In particular this course will provide insights into key benefits of parallelization such as during the n-fold cross-validation process where significant speed-ups can be obtained compared to serial methods. After this course participants will have a detailed understanding why and how parallelization provides benefits to a scalable data analyzing process using machine learning methods for big data.
Introduction to Deep Learning with Convolutional Neural Networks & Applications

Speaker: Morris Riedel - University of Iceland (short bio)

The third course part focusses on a more recent machine learning method known as deep learning that emerged as a promising disruptive approach, allowing knowledge discovery from large datasets in an unprecedented effectiveness and effiiency. It is particularly relevant in research areas, which are not accessible through modeling and simulation. Traditional learning, which was introduced in the 1950s and became a data-driven paradigm in the 90s, is usually based on an iterative process of feature engineering, learning, and modeling. Although successful on many tasks, the resulting models are often hard to transfer to other datasets and research areas. This course provides an introduction into the evolving field of deep learning that provides methods with the inherent ability to derive optimal and often quite generic problem representations from the data (aka ‘feature learning’). This includes concrete algorithms such as Convolutional Neural Networks (CNNs) that will be applied to real datasets of applications using known frameworks such as Tensorflow, Keras, or Torch. As the learning process with CNNs is extremely computational-intensive the course will cover aspects of how parallel platforms can be leveraged in order to speed-up the learning process using general purpose computing on graphics processing units (GPGPUs). After this course participants will have a general understanding for which problems deep learning algorithms are useful and how parallel and scalable computing is facilitating the learning process when facing big datasets.
SESAME an International Research lab

Speakers: Salman Matalgah - SESAME (short bio)

SESAME (Synchrotron-light for Experimental Science and Applications in the Middle East) is a third-generation synchrotron light source under commissioning in Allan (Jordan). It will be the Middle East's first major international research center. The main aim is to foster scientific and technological excellence in the Middle East and neighboring countries (and prevent or reverse the brain drain) by enabling world-class scientific research in subjects ranging from biology, archaeology and medical sciences through basic properties of materials science, physics, chemistry, and life sciences; and Build scientific and cultural bridges between diverse societies, and contribute to a culture of peace through international cooperation in science.

In each experimental lab at SESAME (Beamline), scientists, including graduate students, from universities and research institutes will typically visit the Centre for a week or two to carry out experiments, frequently in collaboration with scientists from other centers, and then return home to analyze their data, which were obtained on SESAME's Beamline.

The amount of experimental data depends on each experiment where in some experiments the produced raw data could reach a few Tera Bytes. This leads users to be more demanding for both the computing processing power and data connectivity to operate in-site/off-site data processing.

To meet the community concerns, the Computing group at SESAME is securing the facility with an IT backbone infrastructure to assure smooth daily operation of the machine and data acquisition systems at each beamline to meet the scientists and the community demands.
The VI-SEEM EU E-Infrastructures Project

Speakers: Andreas Athenodorou - The Cyprus Institute (short bio)

In this session a short description of the VI-SEEM project will be given.

VI-SEEM aims at creating a unique Virtual Research Environment (VRE) in Southeast Europe and the Eastern Mediterranean (SEEM), with special focus on the scientific communities of Life Sciences, Climatology and Digital Cultural Heritage.
GPU Accelerator Programming

Speaker: Kyriacos Hadjiyiannakou - The Cyprus Institute (short bio)

In this session, training on GPU Accelerator Programming, and more specifically CUDA programming will be delivered. The level of difficulty of this session (whether it is an intermediate or advanced session to CUDA) will depend on the CUDA experience of the participants - which they are kindly requested to identify during registration for this track.
Xeon Phi Accelerator Programming

Speaker: Jacob Finkenrath - The Cyprus Institute (short bio)

In this session, an introductory session to Xeon Phi Accelerator Programming will be delivered.
Data Management Plans – EUDAT best practices and case study

Speaker: Stéphane Coutin - CINES (short bio)

Science and more specifically projects using HPC is facing a digital data explosion. Instruments and simulations are producing more and more volume; data can be shared, mined, cited, preserved… They are a great asset, but they are facing risks: we can miss storage, we can lose them, they can be misused,… To start this session, we will review why it is important to manage research data and how to do this by maintaining a Data Management Plan. This will be based on the best practices from EUDAT H2020 project and European Commission recommendation. During the second part we will interactively draft a DMP for a given use case.
HPC and Big Data – Use Cases from the VI-SEEM project

Speakers: Georgios Artopoulos - The Cyprus Institute (short bio), Theodoros Christoudias - The Cyprus Institute (short bio), Zoe Cournia - Biomedical Research Foundation Academy Of Athens (short bio)

VI-SEEM aims at creating a unique Virtual Research Environment (VRE) in Southeast Europe and the Eastern Mediterranean (SEEM), with special focus on the scientific communities of Life Sciences, Climatology and Digital Cultural Heritage. VI-SEEM provides functions to facilitate data management for the selected Scientific Communities, engage the full data management lifecycle, link data across the region, and provide data interoperability across disciplines.

The session deals with full data lifecycle support for the scientific community, including data storage (live, staged), data archiving, data manipulation, collaborative access, domain specific interfaces to storage, data annotation and citation, metadata, PIDs, etc. All these state-of-the-art services enable users to conduct high-quality research with the relevant data.
Data Visualisation

Speaker: Roberto Sisneros - NCSA, University of Illinois at Urbana-Champaign (short bio)

Our visual perception is able to find seeming patterns in data which may be worthy of further investigation. In the visualization training sessions we will cover basic concepts across scientific visualization, information visualization, and visual analytics. These fields encompass techniques for visualizing and analyzing a continuity of data from scientific simulation output to data lacking any inherent spatial coordinates. We will introduce production workflows and analysis pipelines as used by the visualization community and discuss the special considerations for utilizing such at HPC scales. When possible we will stay grounded by focusing on practical applications and highlighting typical libraries and software suites. We will conclude with a functionality deep-dive, including a demonstration using actual scientific simulation data, of a data analysis and visualization suite designed for at-scale deployment, VisIt.
The OPEN SESAME Horizon2020 project

Speakers: Mitchell Peter Edward - OpenSESAME (short bio)

The OPEN SESAME project is aiming to ensure optimal exploitation of the Synchrotron light for Experimental Science and Applications in the Middle East (SESAME) light source. With this aim, OPEN SESAME has three key objectives:
To train SESAME staff in the storage ring and beamline instrumentation technology, research techniques and administration for the optimal use of a modern light source facility.
To build-up human capacity in Middle East researchers to optimally exploit SESAME's infrastructure.
To train SESAME staff and its user community in public outreach and corporate communications, and to support SESAME and its stakeholders in building awareness and demonstrating its socio-economic impact to assure longer-term exploitation.

Each objective is tackled by a work package. Firstly, SESAME staff training is addressed by 65 staff exchanges planned between SESAME and the European partners. Secondly, capacity-building is targeted by five training schools, a short-term fellowship programme and an industrial workshop. Finally, a proactive communications strategy will be created, including an educational "roadshow" to all of the SESAME Members, and a training programme in research infrastructure administration and their economic role and impact for young science managers of SESAME Member stakeholders.
Keynote Address
Biomedical Research and Future Challenges

Speaker: Zoe Cournia - Biomedical Research Foundation Academy Of Athens (short bio)

High Performance Computing is becoming an essential tool in assisting fast and cost-efficient lead discovery and optimization. The application of rational, structure-based drug design is proven to be more efficient than the traditional way of drug discovery since it aims to understand the molecular basis of a disease and utilizes the knowledge of the three-dimensional structure of the biological target in the process. In this talk, we focus on the principles and applications of methodologies applied to computer-aided drug design. We discuss Molecular Dynamics (MD) simulations to discover allosteric pockets in proteins, Virtual Screening to discover novel small molecule inhibitors of protein function and Free Energy Perturbation calculations to optimize the binding affinity of small molecule inhibitors. We examine different procedures ranging from the initial stages of the process that include receptor and library pre-processing, to docking, scoring and post-processing of top-scoring hits and how HPC resources have been used to achieve the desired results.

System Administration Track

The System Administration Track will bring expert administrators to act as trainers on System Administration Tools, Security aspects of HPC Centers, System monitoring, Data Services and Cloud Services - specifically Open Stack. Further to these, the new proposed PRACE Network structure will be presented and there will also be "Discussion on Future of HPC Administration" between both trainers and participants. More details of sessions can be found below (missing descriptions will appear soon):
Modern Scientific Software Management using EasyBuild & co

Speaker: Kenneth Hoste - Ghent University (short bio)

In these sessions, we will introduce several open source projects that significantly facilitate managing scientific software stacks on HPC systems.

The focus of this session is on EasyBuild (http://hpcugent.github.io/easybuild/), a framework for building and installing scientific software on HPC systems. We will motivate the need for a tool like EasyBuild, show how it works, what it can do, how to write (additional) build recipes, how to extend its capabilities and how to share your efforts with the EasyBuild community.

In addition, we will briefly cover:
Lmod (https://www.tacc.utexas.edu/research-development/tacc-projects/lmod) - a modern environment modules tool
Singularity (http://singularity.lbl.gov/) - a container solution developed for HPC systems that can be used by non-privileged users
clustershell (http://cea-hpc.github.io/clustershell/) - a tool for running remote commands on large Linux clusters
Security Aspects of HPC Centers

Speaker: Alexander Withers - NCSA, University of Illinois at Urbana-Champaign (short bio)

In this first session we approach the challenges in securing open science networks and high performance computing centers. The session will take a layered approach to security, demonstrating the need for in-depth security. We will cover how to effectively monitor both the systems and networks, manage and remediate vulnerabilities. We will cover the log analysis life cycle of monitoring, event management, analysis and response. Architecting and scaling effective Intrusion detection using the Bro IDS for high throughput networks. Configuring and securing publically accessible systems and ensuring that those systems can be effectively audited. Many of the examples in the training will include real world case studies on systems at the NCSA. A laptop is recommended to examine and view example log files.

In the second session we continue with techniques and tactics used to mitigate against threats. This includes active response and ensuring an up-to-date Bro IDS instance. The second half of this session will focus on topics in Public Key Infrastructure (PKI). The session will cover PKI basics to help ensure working systems function properly and to better address certificate problems. The session will also cover various authentication technologies: ssh keys, passwords, and certifications--when, where, and what is best used.
System Integration at the Cyprus Institute

Speaker: George Tsouloupas - The Cyprus Institute (short bio)

This talk will address current technical developments at the Cyprus Institute High Performance Computing Facility focusing on deployment strategies and tools. The talk will provide an overview of the systems currently in production at the HPCF including HPC, Storage (hardware and filesystems) and Cloud (Openstack) including the rationale behind the directions taken as well as the deployment tools employed.
PRACE Network and Monitoring Issues

Speaker: Ralph Niederberger - Juelich Supercomputing Centre (short bio)

Since the first days, PRACE has offered an internal network connecting the different supercomputer systems via a dedicated infrastructure.

Having started with a physical infrastructure, new technologies came up, which allow realizing this network in a virtual scenario. This new infrastructure allows cheaper connectivity, faster access for new sites and much more flexibility for temporary solutions.

This session will provide a detailed overview about the new infrastructure, the current status, and transition paths from the old to the new infrastructure. Further on we will describe how new partners can connect to it and how monitoring of the infrastructure is implemented.

Detailed information will be given, how to become part of the monitoring environment as well. Additionally it will be describe, which IT security policies have been defined concerning data transmissions and how they are realized in this multi domain environment.
System Monitoring

Speaker: Brett Bode - NCSA, University of Illinois at Urbana-Champaign (short bio)

Monitoring HPC systems can take many forms - from very basic instantaneous status monitoring to in depth tracking of system metrics and status over time. This session will cover the various methods used on the Blue Waters system to monitor performance and detect problems which all combined record over 8 billion datums per day. While the scale of Blue Waters drives part of this volume, a system of any size can generate a data flow sufficient to overwhelm the administration staff. The methods used by Blue Waters to analyze the data flow for both faults and application issues will be presented.
Discussion on Future of HPC Administration

Chair: Brett Bode - NCSA, University of Illinois at Urbana-Champaign

This will be an informal discussion between the trainers of the System Administration track and the attendees about what they foresee will be the future trends and tools in system administration, as well as what they believe they will have to offer their users.
Data Services

Speaker: Brett Bode - NCSA, University of Illinois at Urbana-Champaign (short bio)

Satisfying the need for data storage and services is becoming the single biggest challenge in the data center. The challenges range from the traditional issues of managing parallel file systems for traditional large-scale HPC applications, to meeting the needs of science domains that are new to HPC and assume the parallel file system behaves like the SSD in their laptop. This session will cover issues of file systems and in meeting the needs for non-traditional software on HPC resources through the use of containers and other new technologies.
Cloud Services - OpenStack

Speaker: David Power - vScaler (short bio), Gaetano Pastore - Datera (short bio), Philip Paul Hofstad - Mellanox Technologies (short bio)

OpenStack is the leading open source IaaS platform, powering many of the world's most notable science and research organisations. Surprisingly, research and science disciplines comprise some of the most prevalent use cases for OpenStack clouds, and OpenStack has provided compelling solutions for many of the challenges of delivering flexible infrastructure for research computing.

This talk is intended for HPC system architects and research computing managers that are exploring the benefits of cloud and how to bring those benefits to HPC centres. We discuss the need for traditional HPC systems to be augmented with self-service research capabilities to cope with the exploding demands on researchers today across traditional sciences, big data analytics and deep learning and deep dive in to how OpenStack can address those challenges.

Our speakers will cover a range of topics including:
OpenStack and HPC, the crossroads of scientific research
How to build your first OpenStack Cloud
Considerations for maximising the performance of OpenStack
RDMA enabled fabrics in an OpenStack environment
IOPS in the cloud – impossible! Or is it?
Case studies from HPC centres using OpenStack for HPC workloads.