# International Symposium on Grids & Clouds 2018 (ISGC 2018) in conjunction with Frontiers in Computational Drug Discovery (FCDD)

Friday 16 March 2018 - Friday 23 March 2018

Academia Sinica

# Book of Abstracts

# Contents

**0**

# Seismic indicators based Earthquake Early Warning System using Genetic Programming and Artificial Neural Networks

**Author:** Asim Khawaja Muhammad[1]

**Co-authors:** Francisco Martinez-Alvarez [2]; Talat Iqbal [1]

[1] *Centre for Earthquake Studies, National Centre for Physics, Islamabad, Pakistan*

[2] *Department of Computer Science, Pablo de Olavide University of Seville, Spain*

**Corresponding Author:** asim.khawaja@ncp.edu.pk

Earthquake Early Warning System (EEWS) holds potential for saving human lives and consequent financial damages. EEWS can be classified into two categories depending upon the methodology: a) P-wave arrival time based, and b) Seismic indicators or parameters based. The former approach generates warnings a few seconds before an earthquake while the later methodology predicts earthquakes a few days or weeks earlier. In this study Genetic Programming (GP) and Artificial Neural Networks (ANN) based methodology is employed with seismic indicators to develop an EEWS. Seismic indicators are based upon geophysical facts and laws, such as Gutenberg-Richter's law, seismic energy release, foreshock frequency, seismic rate changes and total recurrence time. These indicators are computed using the temporal sequence of past seismicity. The indicators are meant to express the internal seismic state of the region in numeric form, which is further modeled with the subsequent seismic activity through Computational Intelligence (CI) based algorithms. GP and ANN are concurrently employed to model the relationship between seismic indicators and earthquakes, thereby developing an EEWS, capable of generating warnings a few weeks before an earthquake.

In this novel combination of two different CI techniques, the role of ANN is to support GP through developing initial estimation of upcoming seismic activity. The approximation of ANN is further provided to GP as an auxiliary prediction for construction of a refined EEWS. The proposed model is designed in a way to generate earthquake warnings for the horizon period of 15 days. Initially, the model is trained for the regions of Hindukush and Chile. The prediction model is based on integration of the searching capabilities of GP and strong classification capabilities of ANN, in two layers. The first layer is complementing the succeeding layer, which results into an enhanced model for a challenging problem of earthquake prediction.

The proposed model also tackles the issue of predicting multiple earthquakes during single horizon period by computing seismic indicators for every earthquake. In this way, the capability of prediction model is enhanced for dealing with multiple seismic events during single horizon period. Upon occurrence of an earthquake before the completion of preceding horizon period, the model computes new seismic indicators inclusive of recent earthquake and generates renewed prediction without any delay. Such an approach enables this GP-ANN based model for predicting multiple successive earthquakes.

**1**

# Digitization of Bhutan's literary heritage

**Author:** Sonam Tobgay[1]

[1] *National Library & Archives of Bhutan*

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

The National Library and Archives of Bhutan (NLAB) was established in 1967 and functions under the Department of Culture, Ministry of Home and Cultural Affairs, Bhutan. NLAB was initially established with the aim of safeguarding and preserving the cultural and literary heritage. Gradually it started to collect and maintain written documents and scriptural resources of the kingdom.

As Bhutan slowly and steadily moves into the era of knowledge-based society, the NLAB as information centre play a greater role in an emerging society and promoting a well-read and well-informed society. The NLAB thrives to be the most dynamic resource centre for preservation, protection, and promotion of the literary heritage of Bhutan, and manage the diverse literary resources and ensure their sustainable development. Pursuant to the commitment to its mandate, the National Library of Bhutan is in the process of disseminating information to wider audiences electronically with the automation of library. Although the use of information technology such as Online Public Access Catalogue (OPAC) and the internet to deliver library service is relatively new in Bhutan, NLAB is working to make documents more accessible with digitization of materials.

The NLAB endeavors to discover Bhutan's rich literary heritage and preserve important and rare documents for the nation. The NLAB team is in the process of carrying out survey of rare text and manuscripts throughout the country. The team travel to the Dzong (Fortress), Monasteries, temples, private houses and temples with permission throughout the country. The main objective of the documentation is to register the textual possession and location of rare documents owned by the different monasteries, temples, and households. This will help the owners or the bearers obtain a copy of the documents in case of the natural calamities like fire, earthquake and flood in the future. The survey includes digitization and registration of rare texts, and recording of details of their location and condition for future reference. These rare texts are then catalogue, digitized and stored in external hard drive for future use.

**Humanities, Arts & Social Sciences Session / 2**

## Trust Building as a Magnet of Online Survey Participation: How an Open Survey Data Platform Assists Theory Development in Social Sciences

**Author:** Frank Liu[1]

[1] *National Sun Yat-Sen University*

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

Conventional practice of public opinion data collection is conducting face-to-face or telephone surveys. Recently scholars have turned to online survey for reducing cost and increasing efficiency. The problems of collecting public opinion data online, however, are that (1) data quality cannot be assured, that (2) respondents may not like to be (re)contacted, and that (3) panel data where respondents are willing to be re-contacted are stored privately. This paper demonstrates a practice of how an online-survey data collection platform overcomes these challenges. Since 2012, the platform smilepoll.tw has been serving as the only one public, open-data survey platform that provides respondents full datasets, which are released to their account right after the survey period. This transparency, as well as other practices of trust building, provides researchers a new tool for both qualitative interviewing and quantitative analysis. This paper will also give an example about how such publicized survey data help exploring attitudinal and behavioral patterns among the respondents. Social scientists in Taiwan are empowered with a resources of pattern recognition and theory formation.

**Infrastructure Clouds & Virtualisation Session / 3**

## Curriculum in the Cloud: Using OpenStack to Transform Computer Science Education

**Author:** Brent Seales[1]

**Co-authors:** Charles Pike [1]; Cody Bumgardner [1]

[1] *University of Kentucky*

**Corresponding Author:** pike@netlab.uky.edu

We explore the question of relying on OpenStack as critical infrastructure for the instructional requirements of the ABET-accredited Bachelor of Computer Science program in the Computer Science department at the University of Kentucky.

The Computer Science department (established in 1966) has seen its instructional capabilities change dramatically over the past fifty years of computing, from time-shared access to mainframe computers in the 1970s to the personal computer labs of the 1980s to the now ubiquitous use of portable and mobile computing.

In some ways OpenStack harks back to the centralized main frames of the 1970s but with many evolved user conveniences. Key among these are the access through portable and widely distributed computing platforms connected to the OpenStack cloud services through broadband network infrastructure, and the evolution of the OpenStack "Big Tent" project model to provide a wide array of self provisioned capabilities to users.

At last the OpenStack architecture promises real advances for instructional programs and the institutions responsible for running them, offering freedom from the dedicated space and hardwired computer lab environments of the past, and leveraging all the convenience of mobile and personal computing with the power of large scale virtualized compute, storage, and networking resources.

But perhaps even more important is the tailored experience that the OpenStack transformation promises for the student: a virtual machine with resources and software allocated specifically to their course load and their place in the timeline of a curriculum designed for progressive instruction and learning.

We will discuss the issues in moving to an instructional program that is fully dependent on OpenStack for teaching a computer science curriculum, including provisioning requirements and goals, and monitoring issues and solutions that lead to improved scaling and tuning. We share our experiences integrating with the broader institutional authentication requirements and tailoring environments at per-student granularity for almost 1000 students.

Finally, we will discuss the practical issues of roll-out, proprietary environments, and the continuous onboarding of new faculty, researchers, and students.

**Summary**:

We explore adopting OpenStack for instruction with the goal of making reliance on instructional computer lab space and equipment a thing of the past. The self service cloud architecture presents unique infrastructure decisions to provide every student with their own virtual machines for course labs, projects, and research. We discuss the issues in moving to an instructional program that is fully dependent on OpenStack for teaching curriculum.

**Networking, Security, Infrastructure & Operation Session** / **4**

## A Simplified SDN-Driven All-Campus Science DMZ

**Author:** Jacob Chappell[1]

**Co-authors:** Brent Seales [1]; Charles Pike [1]; Cody Bumgardner [1]; James Griffioen [1]

[1] *University of Kentucky*

**Corresponding Author:** pike@netlab.uky.edu

Data-intensive computational techniques such as machine learning, data analytics, and visualization, increasingly require data sets at unprecedented scale - massive sizes that are orders of magnitude larger than previous work. This presents challenges for computer networks. This problem is particularly acute for universities where researchers are increasingly using big data in their research, but

the campus network infrastructure is not designed for high-throughput communication. Moreover, north/south traffic to cloud storage providers such as Amazon and Google is growing at an explosive rate, and it is now common for researchers to move terabytes of data to/from the cloud. The result is nothing short of a dire need for high-throughput campus network infrastructure.

To complicate matters, campus networks, which are designed to support common institutional business functions (web browsing, email) are also intended to support high-end research endeavors. Is it possible for the network to do both? As designed, these networks are almost always littered with so-called middle-boxes that perform services such as intrusion detection, rate limiting, firewalling, and other forms of deep packet inspection. These middleboxes play an important role in ensuring security and stability of the campus network, but sacrifice network performance. Though the resulting decline in network performance may be acceptable for common low-bandwidth applications, large data transfers that rely on higher bandwidth suffer greatly.

To meet this challenge, we proposed a new approach to the design of campus networks based on software defined networks (SDN) — specifically OpenFlow. We began by replacing certain building distribution routers with OpenFlow-enabled switches that operate in a hybrid mode, providing normal routing and switching to our standard campus core by default. However, using OpenFlow, we are able to redirect high-throughput flows from approved researchers to an all-new SDN core. The SDN core then forwards packets directly to our campus edge router, bypassing all middleboxes north of the standard campus core infrastructure. A major benefit of our approach is that individual flows from a machine can receive high-speed (middlebox free) paths while all other flows from the same machine travel the standard campus path through policy-enforcing middleboxes. Consequently, transparent to the end user, a host can perform a high-speed file transfer to the cloud while at the same time streaming rate-limited video content. This effectively creates a virtual all-campus DMZ, granular to protocol port level, that can be turned on or off programmatically as needed by researchers.

**Summary**:

University researchers increasingly must choose between connectivity to science DMZs or to the traditional campus network. The campus network provides convenient access to campus services and security policy enforcement whereas the science DMZ allows unfettered network throughput needed for large data sets. We present an SDN-driven approach that permits individual flows from a machine over either path, effectively creating a granular all-campus DMZ that can be enabled or disabled programmatically as needed by researchers.

**5**

## Comparison of GNSS-TEC and IRI-2012 TEC in different regions of Pakistan during the years (2015–2016)

**Author:** Tariq Muhammad Arslan[1]

**Co-authors:** Hernández-Pajares Manuel [2]; Iqbal Talat [1]; Tariku Yekoye Asmare [3]

[1] *Centre for Earthquake Studies, National Centre for Physics, Pakistan*

[2] *IonSAT, Universitat Politècnica de Catalunya, Barcelona, Spain*

[3] *University of Ambo, Department of Physics, Ambo, Ethiopia*

**Corresponding Author:** arslan@ncp.edu.pk

This paper investigates the performance of International Reference Ionosphere (IRI-2012) model in estimating the variations of Vertical Total Electron Content (VTEC) in different regions of Pakistan during the descending phase of solar cycle 24 (2015–2016). The study has been accomplished by comparing the IRI-2012 model and measured VTEC deriving from permanent dual frequency Global Network Satellite System (GNSS) receivers at Islamabad (geographic latitude 33.74°N, longitude 73.16°E), Multan (geographic latitude 30.26°N, longitude 71.50°E) and Quetta (geographic latitude 30.20°N, longitude 67.02°E). We have analyzed the diurnal, monthly and seasonal variability in the measured VTEC and compared with IRI-2012 model VTEC. The highest peaks of the measured and model

VTEC are observed during the April equinoctial month, whereas the lowest values are registered in December solstice month at each station. The diurnal variability of measured VTEC is found to be maximum at 06:00 – 12:00 UT and minimum nearly at 21:00 – 24:00 UT (02:00 – 05:00 LT). Moreover, maximal and minimal monthly mean measured VTEC for each station have also been observed in April and December, receptively. For equinoxes and winter solstices the highest and lowest seasonal mean measured VTEC are found, respectively. The model predictions generally follow the diurnal variability of the measured VTEC with minimum at predawn hours and maximum at noontime hours (06:00 – 09:00 UT). The measured and model monthly mean VTEC from each station are in good agreement during the equinoctial and winter solstice months. In the summer solstice months at Islamabad station, the measured monthly and seasonal VTEC are larger than the corresponding model VTEC by about 34% and 30%, respectively. Similarly, at Multan and Quetta stations, the summer solstice months have a difference in values between the measured and model VTEC about 27% and 25%, respectively. The large discrepancies are observed in diurnal model and measured VTEC during the equinoctial months for Islamabad station at the time interval 05:00 – 16:00 UT. The outcomes of this study might be helpful to understand the ionospheric dynamics and its effects on radio propagation on different regions of Pakistan.

**Infrastructure Clouds & Virtualisation Session / 6**

## Data intensive ATLAS workflows in the Cloud

**Author:** Gerhard Rzehorz[1]

**Co-authors:** Arnulf Quadt [2]; Gen Kawamura [2]; Oliver Keeble [3]

[1] *CERN, University of Göttingen*

[2] *University of Göttingen*

[3] *CERN*

**Corresponding Author:** g.rzehorz@cern.ch

From 2025 onwards, the ATLAS collaboration at the Large Hadron Collider (LHC) at CERN will experience a massive increase in data quantity as well as complexity (High-Luminosity LHC). Including mitigating factors, the prevalent computing power by that time will only fulfil one tenth of the requirement.

This contribution will focus on Cloud computing as an approach to help overcome this challenge by providing flexible hardware that can be configured to the specific needs of a workflow. Experience with Cloud computing exists, but there is a large uncertainty if and to which degree it can be able to reduce the burden by 2025. In order to understand and quantify the benefits of Cloud computing, the "Workflow and Infrastructure Model" was created. It estimates the viability of Cloud computing by combining different inputs from the workflow side with infrastructure specifications. The model delivers metrics that enable the comparison of different Cloud configurations as well as different Cloud offerings with each other. A wide range of results can be obtained - from the required bandwidth over the workflow duration to the cost per workflow - making the model useful for fields outside of physics as well. In the High Energy Physics (HEP) use case, a workload is quantifiable by individual bunch crossings within the detector ('events'). A powerful metric that can be derived from that is EC = 'Events per Cost'. Comparing EC values with each other immediately points to the best Cloud offering for HEP workflows, maximising the physics throughput while minimising the cost.

Instead of using generic benchmarks, the model uses reference workflows in order to obtain infrastructure parameters. The workflow parameters are obtained from running the workflow on a reference machine. The model linearly combines different job aspects such as the machine specific CPU time in order to get the results for one workflow on one machine, which is then extrapolated to a whole Cloud infrastructure. Limiting factors to the predictions of the model are therefore fluctuations within the workflows (varying data complexity, software updates) as well as within the infrastructure ("noisy neighbours").

Finally the usefulness and accuracy of the model will be demonstrated by the real-world experience gathered during the latest CERN Cloud procurement, which included several commercial Cloud

providers. The results encompass recommendations regarding the desirability to commission storage in the Cloud, in conjunction with a simple analytical model of the system, and correlated with questions about the network bandwidth and the type of storage to utilise.

**Physics & Engineering Session / 7**

## Exploit the massive Volunteer Computing resource for HEP computation

**Author:** Wenjing wu[1]

[1] *IHEP*

**Corresponding Author:** wuwj@ihep.ac.cn

Exploit the massive Volunteer Computing resource for HEP computation

**Summary**:

It has been over a decade since the HEP community initially started to explore the possibility of using the massively available Volunteer Computing resource for its computation. The first project LHC@home was only trying to run a platform portable FORTRAN program for the SixTrack application in the BOINC traditional way. With the development and advancement of a few key technologies such as virtualization and the BOINC middleware which is commonly used to harness the volunteer computers, it not only became possible to run the platform heavily dependent HEP software on the heterogeneous volunteer computers, but also yielded very good performance from the utilization. With the technology advancements and the potential of harvesting a large amount of free computing resource to fill the gap between the increasing computing requirements and the flat available resources, more and more HEP experiments endeavor to integrate the Volunteer Computing resource into their Grid Computing systems based on which the workflows were designed. Resource integration and credential are the two common challenges for this endeavor. In order to address this, each experiment comes out with their own solutions, among which some are lightweight and put into production very soon while the others require heavier adaptation and implementation of the gateway services due to the complexity of their Grid Computing platforms and workflow design. Among all the efforts, the ATLAS experiment is the most successful example by harnessing several tens of millions of CPU hours from its Volunteer Computing project ATLAS@home each year.
In this paper, we will retrospect the key phases of exploring Volunteer Computing in HEP, and compare and discuss the different solutions that experiments coming out to harness and integrate the Volunteer Computing resource, finally based on the production experience and successful outcomes, we envision the future challenges in order to sustain, expand and more efficiently utilize the Volunteer Computing resource. Furthermore, we envision common efforts to be put together in order to address all these current and future challenges and to achieve a full exploitation of Volunteer Computing resource for the whole HEP computing community.

**8**

## WISE Recommendations for wiser e-Infrastructures

**Authors:** David Kelsey[1]; Hannah Short[2]; Romain Wartel[2]

[1] *STFC-RAL*

[2] *CERN*

**Corresponding Authors:** david.kelsey@stfc.ac.uk, romain.wartel@cern.ch, hannah.short@cern.ch

As most are fully aware, cybersecurity attacks are an ever-growing problem as larger parts of our lives take place on-line. Distributed digital infrastructures are no exception and action must be taken to both reduce the security risk and to handle security incidents when they inevitably happen. These activities are carried out by the various e-Infrastructures and it has become very clear in recent years that collaboration with others both helps to improve the security and to work more efficiently.

The WISE (Wise Information Security for collaborating E-infrastructures [1]) community was born as the result of a workshop in late 2015, which was jointly organised by the GÉANT group SIG-ISM (Special Interest Group on Information Security Management) and SCI, the 'Security for Collaboration among Infrastructures' group of staff from several large-scale distributed computing infrastructures. All agreed at the workshop that collaboration and trust is the key to successful information security in the world of federated digital infrastructures for research.

WISE provides a trusted forum where security experts can share information on topics such as risk management, experiences about certification processes and threat intelligence. With participants from e-Infrastructures such as EGI, EUDAT, PRACE, XSEDE, NRENs and more, WISE focuses on standards, guidelines and practices, and promotes the protection of critical infrastructure. To date WISE has published two documents; a risk management template and a second version of the SCI framework, endorsed by multiple, large-scale e-Infrastructures.

We present an overview of the available WISE recommendations, and extend an invitation to participate in our working groups.

[1] https://wise-community.org

**Humanities, Arts & Social Sciences Session** / **9**

## Complex Negotiations in Agent-Based Modelling (ABM): Insights from Model Building

**Author:** Shih-Chieh Wang[1]

[1] *Taiwan Institute of Economic Research*

**Corresponding Author:** r95322031@ntu.edu.tw

Negotiation is one of the most common things in the world. Everyone negotiates something every day. People negotiate even when they don't think that they are doing so. As negotiation is part of social life, it is also undeniable of its complex essence. Complexity arises when there are a lot of variables interacting, such as the variety of interests, the number of agents and issues.

Complex negotiations actually have been studied for quite a long time. However, only recently the complex negotiation's fundamental nature, essential elements, and the causal effect relations among elements, etc., are discussed through a "descriptive five-level framework", namely, (1) negotiation structure, (2) context, (3) structure and relationships, (4) process and (5) decision making. Nevertheless, the "descriptive five-level framework" only can provide basic understanding of the complex nature such as the background, critical timing, elements and structures of a negotiation case.

With this regard, this paper will introduce a new approach – Agent-Based Modelling (ABM) – to further analyze complex negotiations such as how agents interact and how a consensus is reached. The three-year work experience as the coordinator of Taiwan to the Asia-Pacific Economic Cooperation (APEC) SME Working Group (SMEWG) allows me to deeply involve in international negotiations and collective decision-making process. Therefore, a specific complex negotiation case from APEC SMEWG will be introduced and examined by the built ABM model through NetLogo.

To sum up, in this preliminary study, I will firstly review the academic literature of complex negotiation and display how an ABM model can be designed for analyzing a complex negotiation. After that, a systematically review of the case study will be presented for further discussions regarding insights from model building of ABM.

Keywords: Complex Negotiations, Agent-Based Modelling (ABM), NetLogo, Model Building, Asia-Pacific Economic Cooperation (APEC), Small and Medium Enterprise Working Group (SMEWG)

**Networking, Security, Infrastructure & Operation Session / 10**

## Explore New Computing Environment for LHAASO offline data analysis

**Author:** Qiulan Huang[1]

**Co-author:** Gongxing Sun [2]

[1] *Institute of High Energy of Physics, Chinese Academy Sciences*

[2] *Institute of High Energy of Physics,Chinese Academy of Sciences*

**Corresponding Author:** huangql@ihep.ac.cn

The exploitation of a new computing environment has become an urgent practice to overcome a series of challenges with the development of the new generation of High Energy Physics(HEP). LHAASO(Large High Altitude Air Shower Observatory) is expected the most sensitive project to studies the problems in Galactic cosmic ray physics, and requires massive storage and computing power. Efficient parallel algorithms/frameworks and High IO throughput are key to meet the scalability and performance requirements of LHAASO offline data analysis. Though Hadoop has gained a lot of attention from scientific community for its scalability and parallel computing framework for large data sets, it is still difficult to make LHAASO data processing tasks run directly on Hadoop. In this paper we explore ways to build a new computing environment using Hadoop to make LHAASO jobs run on it transparently. Particularly, we discuss a new mechanism to support LHAASO software to random access data in HDFS. Because HDFS is streaming data stored only supporting sequential write and append. It cannot satisfy LHAASO jobs to random access data. This new feature allows the Map/Reduce tasks to random read/write on the local file system on data nodes instead of using Hadoop data streaming interface. This makes HEP jobs run on Hadoop possible. We also develop diverse MapReduce model for LHAASO jobs such as Corsika simulation, ARGO detector simulation (Geant 4) and MK2A data processing. And we wrap the models to make them transparent to users. In addition, we provide the real-time cluster monitoring in terms of cluster healthy, number of running jobs, number of finished jobs and number of killed jobs. Also the accounting system is included. This work has been in production for LHAASO offline data analysis to gain about 40,000 CPU hours per month since September, 2016. The results show the efficiency of IO intensive job can be improved about 46%. Finally, we describe our current work of data migration tools to serve the data move between HDFS and other storage system or Tape.

**Networking, Security, Infrastructure & Operation Session / 11**

## Design and Development of the Platform for Network Traffic Statistics and Analysis

**Author:** Hao Hu[1]

**Co-authors:** Fazhi QI [2]; Qi Luo [3]

[1] *Institute of High Energy Physics*

[2] *Institute of High Energy Physics,CAS*

[3] *IHEP*

**Corresponding Author:** huhao@ihep.ac.cn

Huge amount of experimental data are produced by large scientific facilities of IHEP, such as Daya Bay, JUNO, LHASSO and CSNS. The performance and efficiency of the data exchange are playing an important role in these scientific research activities.

The quality of data exchange among the members relies heavily on the stability and reliability of the network. The statistics and analysis for the network traffic is a way to know the network status and also very useful for the network performance optimization and the strategy plan for network architecture.

The paper describes the design and development a network traffic statistics and analysis platform based on PMACCT and KAFKA. The functional modules of this platform include data acquisition, data storage, data analysis and visualization. The data acquisition module is developed based on an open source software 'PMACCT', which collects network traffic data from the routers. For data storage module, KAFKA is used as a message queue to subscribe traffic records from PMACCT and save these records to MongoDB with high efficiency. Data analysis module is responsible for the classification and statistics of traffic data according to the requirement parameters pre-configured. Visualization module serves for network administrator which provides customized graphic reports via HTTP.

Partial functions of the platform have been finished and deployed to analysis IHEP network traffic, and more features will be released in the future.

**Summary**:

The paper describes the design and development a network traffic statistics and analysis platform based on PMACCT and KAFKA. The functional modules of this platform include data acquisition, data storage, data analysis and visualization.Part functions of the platform has been finished and deployed to analysis IHEP network traffic, and more features will be released in the future.

**Physics & Engineering Session** / **12**

## Studies on Job Queue Health and Problem Recovery

**Author:** Xiaowei Jiang[1]

**Co-authors:** Hongnan Tan [2]; Jiaheng Zou [3]; Jingyan Shi [2]; Qingbao Hu [2]; Ran Du [1]; Zhenyu Sun [2]

[1] *Institute of High Energy Physics, Chinese Academy of Sciences*

[2] *IHEP*

[3] *IHEP, Chinese Academy of Sciences*

**Corresponding Author:** jiangxw@mail.ihep.ac.cn

In a batch system, the job queue is in charge of a set of jobs. Job queue health is determined by the health status of these jobs. The job state can be queuing, running, completed, error or held, etc. Generally jobs can move from one state to another. However, if one job keeps in one state for too long, there might be problems, such as worker node failure and network blocking. In a large-scale computing cluster, problems cannot be avoided. Then some jobs will be blocked in one state, and cannot be completed in time. This will delay the progress of the computing task.
For the previous situation, this paper studies on the abnormal job state reason, problem handling and job queue stability. We aim to improve the job queue health, so that we can raise job success rate and speed up users' task progress. Abnormal reasons can be found from job attributes, queue information and logs, which can be analyzed in detail to acquire better solutions. These solutions are grouped into two categories. The first one is automatic job recovering that associated with the monitor system. When a job is recovered, it can be rescheduled in time. The second one is automatically informing users to recover jobs by themselves. Depending on the analysis results, feasible recommendations are pushed to users for quick recovering.
As described above, a queue health system is designed and implemented at IHEP. We define a series of standards to determine abnormal jobs. Various information is collected and analyzed in association. According to the analysis results, automatic recovery measures are applied to abnormal jobs. In case

of invalid automatic recovery, recommendations are sent to users by emails, WeChat, etc. The status of this system shows that it's able to improve the job queue health in most conditions.

**Heritage Science Session** / **13**

## Document Management with ConservationSpace

**Author:** Mervin Richard[1]

[1] *National Gallery of Art*

ConservationSpace is a web-based, open source software application to create, manage, and preserve conservation documents. With support of The Andrew W. Mellon Foundation, the National Gallery of Art, Washington, has led the development effort in partnership with the Courtauld Institute of Art, Denver Art Museum, Indianapolis Museum of Art, Metropolitan Museum of Art (withdrew in late 2015), Statens Museum for Kunst, and Yale University.

The conservation community realized many years ago that a digital solution for managing documents would increase access, improve discoverability, expand research opportunities, support workflow procedures, and reduce document loss. Several institutions sought grants to develop document management software tailored to the specific needs of their institution; however, it quickly became clear that this was an expensive approach that did not facilitate collaboration between institutions. In response to these concerns, the Mellon Foundation organized several meetings to discuss challenges and collaborative solutions for conservation documentation. These efforts resulted in the formation of the ConservationSpace partnership.

In 2012, the National Gallery awarded a contract for software development to Sirma Enterprise Systems (formerly Sirma ITT). ConservationSpace utilizes the Sirma Enterprise Platform, a combination of multiple open source components designed to allow rapid development for diverse business requirements. While the system can be installed on institutional servers, it is primarily intended for a Software-as-a-Service (SaaS) approach in which the application is centrally hosted and delivered over the Internet as a service. For many users, web-based software reduces costs and simplifies software maintenance.
The database management system is constructed on semantic technology. Semantic repositories are an alternative to more traditional relational databases for storing, querying, and handling structured data. The semantic graph database utilizes an ontology that provides easier integration and reasoning capabilities with a large volume of diverse data. It enables searches that find complex relationships—ones that current standard data models would not discover. This benefit is particularly important for an application that must access and link highly heterogeneous data scattered in diverse systems, making it the ideal approach to gathering and storing information for conservators. In the summer of 2016 the National Gallery of Art was the first partner institution to implement ConservationSpace to create and manage new documents. The Gallery is now in the process of scanning hard copy legacy documents, with optical character recognition (OCR), and migrating existing digital records into ConservationSpace. The other partners are in various stages of implementation.

The ultimate success of ConservationSpace will be measured both by the effectiveness of its document-management solutions and the ease in which the software application can be used and adapted by conservators and scientists in other institutions and private practice.

**Data Management & Big Data Session** / **14**

## Smart Policy Driven Data Management and Data Federations, enabled by the H2020 eXtreme DataCloud project.

**Authors:** Giacinto Donvito[1]; Lukasz Dutka[2]; Oliver Keeble[3]; Patrick Fuhrmann[4]; daniele cesini[5]

[1] *INFN/Bari*

[2] *CYFRONET*

[3] *CERN*

[4] *DESY/dCache.org*

[5] *CNAF/INFN*

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

In November 2017, the H2020 "eXtreme DataCloud" project will be launched, developing scalable technologies for federating storage resources and managing data in highly distributed computing environments. The project will last for 27 months and combines the expertise of 8 large European organizations. The targeted platforms are the current and next generation e-Infrastructures deployed in Europe, such as the European Open Science Cloud (EOSC), the European Grid Infrastructure (EGI), the Worldwide LHC Computing Grid (WLCG) and the computing infrastructures that will be funded by H2020 EINFRA-12 calls.
One of the core activities within XDC is the policy-driven orchestration of federated and heterogeneous data management. The high-level objective of this work is the semi or fully automated placement of scientific data in the Exabyte region on the site (IaaS), as well as on the federated storage level. In the context of this Work Package, placement may either refer to the media the data is stored on, to guarantee a requested Quality of Service, or the geographical location, to move data as close to the compute facility as possible to overcome latency issues in geographically distributed infrastructures. The solutions will be based on already well established data management components as there are dCache, EOS, FTS and the INDIGO PaaS Orchestrator, Onedata and many more. The targeted scientific communities, represented within the project itself, are from a variety of domains, like astronomy (CTA), Photon Science (European X-FEL), High Energy Physics (LHC), Life Science (LifeWatch) and others.
The work will cover "Data Lifecycle Management" and smart data placement on meta-data, including storage availability, network bandwidth and data access patterns. The inevitable problem of network latency is planned to be tackled by smart caching mechanisms or, if time allows, using deep learning algorithms to avoid data transfers in the first place. Furthermore, data ingestion and data movement events, reported to the centralized INDIGO PaaS orchestration engine can trigger automated compute processes, starting from meta-data extraction tools to sophisticated workflows, e.g. to pre-analyze images from Photon Science or Astronomy detectors.
This presentation will present the overall architecture of this activity inside the eXtreme DataCloud project and will elaborate on the work plan and expected outcome for the involved communities.

**Summary**:

The core activity within the newly created H2020 project "eXtreme DataCloud" will be the policy-driven orchestration of federated data management for data intensive sciences like High Energy Physics, Astronomy, Photon and Life Science. Well known experts in this field will work on combining already established data management and orchestration tools to provide a highly scalable solution supporting the computing models the entire European Scientific Landscape. The work will cover "Data Lifecycle Management" as well as smart data placement based on domain specific and technical meta-data, including storage availability, network bandwidth and data access patterns. Mechanisms will be but in place to trigger computational resources based on data ingestion and data movements. This presentation will present the first architecture of this endeavor.

**15**

## VCondor: a dynamic cloud scheduler with HTCondor for multi-queue in cloud based environment

**Author:** Haibo Li[1]

**Co-authors:** Yaodong CHENG [2]; Zhenjing Cheng [3]

[1] *Chinese*

[2] *IHEP, CAS*

[3] *Insititue of High Enery Physics, Chinese Academy of Sciences*

**Corresponding Author:** lihaibo@ihep.ac.cn

As a new approach to manage computing resource, virtualization technology is more and more widely applied in the high-energy physics field. A virtual computing cluster based on Openstack was built at IHEP, using HTCondor as the job queue management system. In a traditional static cluster, a fixed number of virtual machines are pre-allocated to the job queue of different experiments. However this method cannot be well adapted to the volatility of computing resource requirements. To solve this problem, VCondor: an elastic computing resource management system in cloud based environment has been designed. This system performs unified management of virtual computing nodes on the basis of job queue in HTCondor, and based on dual resource thresholds as well as the quota service (VMquota). A VM will be created automatically when a job is waiting to run. It will be destroyed when the job is finished and there is no more job in HTCondor queue. The job queue is checked in a period of time such as 10 minutes, so a VM will continue to run if there are new jobs in the period of time. The system is consisted of four loosely-coupled components, including job status monitoring, computing node management, load balance system and the daemon. Job status monitoring system communicates with HTCondor by command lines or APIs to get the current status of each job queue. Computing node management component communicates with Openstack to launch or destroy virtual machines. After a VM is created, it will be added to the resource pool of corresponding experiment group. Then the VM can get a job to run. After the job finishes, the virtual machine will be shutdown. When the VM shutdown in Openstack, it will be removed from the resource pool. Meanwhile, the computing node management system provides an interface to query virtual resources usage. Load balance system provides an interface to get the information of available virtual resources for each experiment from VM Quota. The VM Quota tells load balance system how many virtual machines one experiment can use and reserve them for a period of time such as 30 minutes. The daemon component asks load balance system to decide the number of available virtual resources. It also communicates with job status monitoring system to get the number of queued jobs. Finally, it calls computing node management system to launch or destroy a few of virtual computing nodes.
The practical run shows virtual computing resource dynamically expanded or shrunk while computing requirements change. Additionally, the CPU utilization ratio of computing resource was significantly increased when compared with traditional resource management. VCondor also has good performance when there are multiple condor schedulers and multiple job queues.

**Data Management & Big Data Session / 16**

## Automatic Extraction of Extended Named Entities from Wikipedia for Conversation Analysis

**Author:** Daiki Ishizuka[1]

**Co-author:** Minoru Nakazawa [1]

[1] *Kanazawa Institute of Technology*

**Corresponding Author:** b6700531@planet.kanazawa-it.ac.jp

In recent years, AI assistants have been developed and these are increasingly helping humans at work and at home. Among them, voice assistances technologies are particularly attractive. They are not limited only to smart phone applications, such as Apple Siri and Google Assistance. Recently Amazon Alexa is used at work and Google Home and LINE Wave are used at home. These voice assistance technologies also help in the business places as well as at home.
As described above, a voice assistant is active in various situations, but there are many operating steps to complete. These steps are quite complicated because they require the use of complex techniques. The steps can be roughly divided into two parts. The first step is to get a voice assistant to understand the human language and to recognize the conversational situation and context. The second step is to create the contents of the next dialogue to continue conversation based on the results obtained in the first step, and then choose a suitable response among from possible choices of responses.

In this research, focusing on the first step we propose a method for capturing and structuring human words in more detail. To explain more concretely, this method automatically extracts the extended named entity from the Wikipedia articles, and it structures the contents of the sentences and the dialogues based on the extracted information.
There are seven types of named entities. They are human, organization, location, date, time, money and rate described by MUC (Message Understanding Conference) and used for classification and analysis of sentences. However, the scope of these classifications is ambiguous, and it does not reach the practical level in actual sentence classification and analysis. Therefore, an extended named entity which improved the named entity has been proposed. Nevertheless, even though it has expanded from seven types to more than 200 types of named entities, it is necessary to deal with vast datasets manually in order to actually perform the classification work. To solve this problem, this research introduces a way to construct datasets of extended named entity from Wikipedia and classify sentences with more detailed granularity.

**17**

## Jasmine : A Cluster Job Test Suite for Multiple Scheduling Systems

**Author:** Ran Du[1]

**Co-authors:** Jiaheng Zou [2]; Jingyan Shi [3]; Xiaowei Jiang [1]

[1] *Institute of High Energy Physics, Chinese Academy of Sciences*

[2] *IHEP, Chinese Academy of Sciences*

[3] *IHEP*

**Corresponding Author:** duran@ihep.ac.cn

Job Scheduling Systems are crutial to Clusters. Different Scheduling Systems have different suitable applications. With the cluster scaling up and different requirements from experiments, there are two local clusters in IHEP at present. One is the HTCondor cluster scheduled by HTCondor, the other is the SLURM cluster managed by SLURM. To provide better services, researches are necessary including resource management policy, scheduling algorithms, and supporting systems. To make it more convenient for users, a job test suite named Jasmine is developed. Jasmine can generate customized test job pools based on the configuration file, and a job submitter to submit test jobs to the target cluster. Jasmine has been used for research and production purpose, which is proved to be convenient and powerful.

**Networking, Security, Infrastructure & Operation Session / 18**

## Towards cross infrastructure Operational Security in EOSC-hub

**Author:** Sven Gabriel[1]

[1] *Nikhef/EGI*

**Corresponding Author:** sveng@nikhef.nl

The EOSC-hub proposes a new vision to data-driven science, where researchers from all disciplines have easy, integrated and open access to the advanced digital services, scientific instruments, data, knowledge and expertise they need to collaborate to achieve excellence in science, research and innovation.

The process towards the integration of the different security activities will be supported through the development of harmonized policies and procedures, to

ensure consistent and coordinated security operations across the services provided in the catalogue.

Coordinating the Operational Security in such a broad environment is a challenge. At the same time it offers many possibilities of a closer collaboration of the already existing security teams active in the distributed infrastructures.

The expertise built, and tools developed in response to specific problems in the different infrastructures can be used in cross-infrastructure co-operations. In this presentation we will present examples for possible collaborations in:
* Incident Prevention
* Incident Handling/Coordination
* Security Training and Exercises

In our presentation we will also share information on some actual cases of vulnerability management and incident coordination within and across infrastructures. A discussion on lessons learned will include review on how comprehensively critical vulnerabilities and incidents have been identified and how efficiently risks and incidents have been contained. As usual, the results from the debriefings give us pointers to which tools and procedures need further development to further improve our cross infrastructure operational security capabilities.

**Networking, Security, Infrastructure & Operation Session / 19**

## Best practises and experiences in user support –A case study at GARUDA

**Author:** Divya MG[1]

**Co-authors:** Henrysukumar S [1]; Santhosh J [2]

[1] *C-DAC*

[2] *C_DAC*

**Corresponding Authors:** santhoshj@cdac.in, divyam@cdac.in

To improve quality of user support GARUDA Help Desk (GHD) extends single point gateway to Indian national grid computing initiative – GARUDA users. All operational, usage and computing related issues from scientists, academicians, system administrators and network service providers are effectively managed and addressed.

GHD has unified portal to receive queries from users. Over all focus of GHD is to provide suitable and efficient support structure. Objective of this paper is to share our experience/best practices which is successfully being adopted in user support. Various facts and figures are presented.

User can report to GARUDA Help Desk (GHD) through any mode of communication like telephone, email and web portal. All the reported issues are converted as ticket at web portal named Request Tracker. Within 24 hours, a user will receive an e mail notification about ticket owner and the status. Same will be updated in web portal also, intermediate status of the ticket will be updated in the web portal until it is closed.

The challenges involved in GARUDA user support and how GHD (ticketing system) helps to better services is discussed in this paper. In GARUDA, the resources are heterogeneous in nature and spread across India. Maintaining grid availability and reliability is highly challenging in a federated environment. Some of the challenges to mention are: remote ssh access, sudden cluster down time, faster support to users, provisioning of large home areas for user, lack of cooperation from remote system administrators, slow response time for reported ticket, routing the tickets to right affiliate, tracking the progress of ticket, tracking the ticket status, so on.

Most of the issues are related to remote cluster, where remote system administrators' support was required to resolve the issues. To mitigate suitable tools are deployed, but hard to find a tool to work as per the need. Hence in-house developed automatic scripts were deployed. Another best practice is all the issues reported by users are recorded. Like nature of ticket, its category, methodology used in solving, actual solution etc are documented, preserved and made available for reference. We have found that proper monitoring methodology, highly transparent and interactive work culture and adhering to quality standards gives good results in effective user support. Another key practice followed is collecting "Annual user feedback", which helps us to restructuring of GHD guidelines and processes. Hence we are able to progressively improve user support. It is proudly said that transformed ordinary user support to extraordinary successfully.

GHD handles 400 to 500 requests (tickets) per month. Average ticket handling time by respective owner is 8 Hrs. All the ticket handling operations are adhering to ISO 9001:2008 quality management standards. GHD has reduced redundancy of ticketing and improved the capability of sharing problem and solutions among the administrators.

**Summary**:

In this paper authors have shared best practices and experiences in user support through GARUDA help desk. Also the paper has covered GARUDA help desk operational process, procedures, and the challenges.

**20**

## Visual analysis system for large-scale data storage

**Authors:** Qingbao Hu[1]; Wentao Zhang[1]; Yaodong CHENG[2]

[1] *IHEP*

[2] *IHEP, CAS*

**Corresponding Author:** zhangwt@ihep.ac.cn

With the development of new generation high energy physics experiment facilities, a large amount of data has been produced, which brings great challenge to the storage management of large-scale data. Traditional storage system can't find what data need to be cleaned up in time. Idle data takes up a lot of storage space, which leads to poor utilization of storage system. Storage system often adopts hierarchical storage architecture, the higher the hierarchy, the faster the access speed, the smaller the capacity. Traditional system can't know which level the file to be stored should be placed on. When people suddenly find out that the data is placed in the wrong storage location , they often need to spend a large amount of computing resources, manpower and time to transfer large amounts of data. To solve these problems, this paper put forward to establish a visual analysis system for large-scale data storage, which could manage files dynamically. The machine learning is used to accurately predict the storage location of the files to be stored by using the file information. Fast index mechanism is established, which supports fuzzy check for file information. Collecting file information is critical for these tasks. There are several ways to collect information: accessing log, EOS(EOS open storage) metadata dump and multi-thread scan. This paper discusses the advantages and disadvantages of these methods. File information is written in time series database in bulk, and InfluxDB is used here. By writing in database, data can be quickly extracted for analysis and processing compared to getting file information from disk. Using the file information collected, the monitoring system can display the usage of the storage system in real time according to various attributes, so as to manage the storage data timely. Finally, this paper introduces how machine learning is used to improve the efficiency of the storage system.

**Summary**:

This paper put forward to establish a visual analysis system for large-scale data storage, which could manage files dynamically. The machine learning is used to accurately predict the storage location of the files to be stored by using the file information. Fast index mechanism is established, which supports fuzzy check for file information.

**Earth, Environmental Science & Biodiversity Session / 21**

## Towards a crowdsourcing platform for labelling remote sensing images online

**Author:** Jianghua Zhao[1]

**Co-authors:** Qinghui Lin [1]; Xuezhi Wang [1]; xiaohua zhou [1]; yuanchuan zhou [1]

[1] *Computer Network Information Center, Chinese Academy of Sciences*

**Corresponding Author:** zjh@cnic.cn

Vast quantities of remote sensing data are becoming available at an ever-accelerating rate, and it is transforming geosciences today. When investigating global-scale environmental phenomena, large scale of temporal and spatial remote sensing data need to be processed to extract useful information, especially the land use types. As most powerful remote sensing image processing methods need labeled training dataset, and there is no such open remote sensing image training dataset, researchers have to label the image all by themselves. Sometimes, researcher have to download and label large volume of remote sensing images to satisfy their machine learning methods, such as deep learning. This is not only time-consuming, but also a waste of researchers energy.
Based on our past work of having built a cloud-based platform for massive remote sensing data retrieving, downloading and processing, we have accumulated a large volume of remote sensing data, and more than 187 thousand users. Crowdsourcing is a way of utilizing the human intelligence of citizens. As it has been proved to be an effective way to tackle tasks that are difficult to be automated processed, such as Wikipedia, we decide to try to collect training datasets for remote sensing image processing through crowdsourcing. It not only can fully utilize the human intelligence of the large number of users, but also can build a large remote sensing labeled dataset of great value.
In this paper, we propose an on-line remote sensing image labeling platform. Through it, users neither have to install any software, nor download large remote sensing images. Remote sensing images are displayed on the web, and users can directly label the images by just drawing polygons and add tags. Tags are predefined. All the labeled dataset can be downloaded and shared with other people. Moreover, standards of the training dataset have been studied. And two components of the labeled data are stored, which are the true image values of the labeled region and their metadata, such as the bounding box, tag, and location information. In this way, the training data can be used directly for a variety of machine learning models. By implementing such an online remote sensing labelling platform, participants not only can visualize, label, download, and share training datasets only with a simple web-connection, but also will have a better understanding of the land types on our planet.

**VRE / 22**

## Unified Account Management for High Performance Computing as a Service with Microservice Architecture

**Author:** Rongqiang Cao[1]

**Co-authors:** Haili Xiao [1]; Shasha Lu [1]; Xiaoning Wang [1]; Xuebin Chi [1]

[1] *Computer Network Information Center, Chinese Academy of Sciences*

**Corresponding Author:** caorq@sccas.cn

In recent years, High Performance Computing (HPC) has developed rapidly in China. From Chinese Academy of Sciences (CAS) level, Scientific Computing Grid (ScGrid), is a general-purpose computing platform started from 2006 in CAS, which provided a problem solving environment for comput-

ing users through grid computing and cloud computing technologies. Then ScGrid becomes Supercomputing Cloud, an important port of Chins Science Cloud from 2011. From national level, China National Grid (CNGrid) has integrated massive HPC resources from several national supercomputing centers and other large centers distributed geographically, and been providing efficient computing services for users in diverse disciplines and research areas. During more than 10 years, CNGrid and ScGrid has integrated tens of HPC resources distributed geographically across China, comprising 6 National Supercomputer Centers of Tianjin, Jinan, Changsha, and Shenzhen, Guangzhou, Wuxi, and also dozens of teraflops-scale HPC resources belong to universities and institutes. In total, the computing capability is more than 200PF and the storage capacity is more than 160PB in CNGrid.

As worked in the operation and management center of CNGrid and ScGrid for many years, we notice that users prefer to manage their jobs at different supercomputers and clusters via a global account on different remote clients such as science gateways, desktop applications and even scripts. And they don't like to apply for an account to each supercomputer and login into the supercomputer in specific way.

Therefore, we described Unified Account Management as a Service (UAMS) to access and use all HPC resources via a global account for each user in this paper. We addressed and solved challenges for mapping a global account to many local accounts, and provided unified account registration, management and authentication for different collaborative web gateways, command toolkits and other desktop applications. UAMS was designed in accordance with the core rules of simplicity, compatibility and reusability. In architecture design, we focused on loosely-coupled style to acquire good scalability and update internal modules transparently. In implementation, we applied widely accepted knowledge for the definitions of the RESTful API and divided them into several isolated microservices according to their usages and scenarios. For security, all sensitive data transferred in wide-network is protected by HTTPS with transport layer security outside of CNGrid and secure communication channels provided by OpenSSH inside of CNGrid. In addition, all parameters submitted to RESTful web services are strictly checked in format and variable type.

By providing these frequently important but always challenging capabilities as a service, UAMS allows users to use tens of HPC resources and clients via only an account, and makes it easy for developers to implement clients and services related HPC with advantages of numerous users and single sign-on capability. Based on UAMS, representative clients are introduced and reviewed combined with different authentication schemes. Finally, analysis and test of UAMS shows that it can support authentication in milliseconds level and has good scalability. In future, we plan to implement federated account service that enable a local HPC account similar to a global account to login the national HPC environment, access and use all HPC resources in CNGrid.

**Summary**:

In this paper, UAMS is proposed to provide account management and authentication schemes for accessing and using national HPC environment via a global account of CNGrid. UAMS is implemented in self-contained microservices and solved challenges for mapping a global account to many local accounts, and provides unified account registration, management and authentication for administrators and users. It is creative for users to update the requirements of HPC in the full life-cycle of accounts. Correspondingly, administrators adjust assign available privileges to satisfy the demands.

For developers, a group of RESTful API is implemented in several microservices and provides full life-cycle account services as a whole for developers. Besides, UAMS provides the basic username and password authentication scheme and its variants for developers to integrate the authentication service into their clients. With UAMS, multiple clients, consisting of command-line toolkits, web gateways and scripts, have integrated authentication service based different authentication schemes.

For performance, the result of test shows that the UAMS can provide account management and authentication services at milliseconds level. For scalability, UAMS could elastically stretch the capacity at the granularity of a microservice to deal with massive requests when the load is heavy and release resource when the load becomes slight.

**Humanities, Arts & Social Sciences Session / 23**

## Educational Quality Assurance and Accountability of ICT-Enhanced Higher Education - Developing an SOP for Educational Informatics -

**Author:** Tosh YAMAMOTO[1]

**Co-author:** Masaki Watanabe [1]

[1] *CTL, Kansai University*

**Corresponding Authors:** soetosh@gmail.com, masaki@igroupjapan.com

This proposal is to put the higher education in a bigger picture and looks at the education from the institutional level or Chief Information Officer's perspective. Target audience will be school administrators as well as decision-makers for the deployment and implementation of ICT-enhanced curriculum or education.
In order to assure the quality of education enhanced with ICT-enhanced learning to better design the future society, no one denies that education plays the essential role. However, there has not been any quality assurance model to visually explain the quality of education or any specification or standard for an effective role of ICT in education discussed in an institution or presented in public. Further, there has not been any demonstration of accountability of such issues to stakeholders. As the results, school administrators or IT staff in charge of implementing ICT in education did not have any chance to quality control the education. Teachers and instructors has been implementing ICT-enhanced education with trials and errors.
This paper purports to propose a robust standard operating procedures for the use of ICT in education, especially in the field of learning rather than teaching, i.e., educational Informatics. In the presentation, the SOP in consideration with the following components are discussed: (i) the educational paradigm as well as the learning environment to fortify the ICT-enhanced learning, (ii) the strategies for effective ICT-enhanced learning, (iii) effective methodologies for learning, (iv) effective assessment strategies for ICT-learning in conjunction with required ICT literacy level for learners and educators.

**24**

## New Curriculum Design for Liberal Arts in the Global Era - Future Education Model for Liberal Arts for Lifelong and Life-wide Learning and Career in Collaboration with Corporations-

**Author:** Tosh YAMAMOTO[1]

**Co-authors:** Maki OKUNUKI [2]; Masanori TAGAMI [3]

[1] *CTL, Kansai University*

[2] *Hands-on Learning Center, Kwansei Gakuin University*

[3] *Otemon Gakuin University*

**Corresponding Authors:** soetosh@gmail.com, okufield2013@gmail.com

This presentation reports a successful hybrid liberal arts program for both company employees and university students, which has been conducted with the collaboration of corporations and universities for the last five years. In other words, it is a success report for a "win-win relationship" between the lifelong and life-wide corporate training and the liberal arts education at the university to nurture the lifelong and life-wide learning attitude.
In an effort to implement education for lifelong and life-wide learning and career, Kansai University and well-known corporations such as Fuji Xerox Corporation, IBM Japan, and ANA collaboratively engaged in series of negotiation practicum for advanced communication for long-lasting trust building with empathy. It is believed that such training will change the mindset of the employees in a long journey of their career and orienting students to a well- balanced adulthood.
So far, most educational programs as well as the corporate trainings have been conducted targeting at the same age groups, the same gender groups, or groups with common background/demographic features. As the result, the expected effectiveness in learning has not been reached and there was no change in attendees' performance and attitude even after the program.
In order to remedy such disadvantages, the workshop proposed here has double-barrel goals. For students, the program will bring the young generation ready for their career for life by learning the communicative negotiation process to build a long-lasting trust through learning working with the 5 to 10 years older generation, which makes them think of their future scenario planning. For corporate employees, they tend to lose sight of the career goal after working for 5 to 10 years and

thus need reorientation in their career. They are also struggling in communication with different generations within the company as younger generations become populated in the workplace. This program offers them opportunities to reflect back their career through meta-cognition by sharing with students the experiences of successes and mistakes in life as well as values gained by such experiences. At the same time, both company employees and students have opportunities to acquire better communication skills that may lead to build and maintain a long-lasting trust.
This program offers opportunities for learning how to build trust through communication, which seems to be the fundamental drive for the lifelong and life-wide education. While elaborating on the intention of the program, this presentation includes the following:
1. Educational Model Explained
2. Negotiation Practicum for Building Trust with Empathy
3. Hands-on & Heads-on Workshop
4. What has been done: Workshop Programs, Artifacts
5. Reflection: Gained Experiences: Pros and Cons
6. Future planning

**25**

## The research of High-Performance Computing infrastructure for Artificial Intelligence and Big Data

**Author:** Jue Wang[1]

**Co-authors:** Chen Li [1]; Fang Liu [1]; Kun Sun [1]; Tengteng Hu [1]; Yangang Wang [1]; Yongze Sun [1]; Zhonghua Lu [1]

[1] *Computer Network Information Center, Chinese Academy of Sciences*

Since the difference of software stack between traditional HPC applications, big data applications and artificial intelligence applications, infrastructure for each usually adopted completely different method and system to build and manage.
To meet the rapidly growing demand for computing and storage resources raised by big data and artificial intelligence application, and take fully advantage of infrastructure's computing and storage ability, we build our infrastructure based on traditional HPC technology and infrastructure, and make improvements on different layers of the infrastructure to meet the requirements of big data and artificial intelligence application, realizing a high-performance computing infrastructure with high efficiency and user-friendly.

For storage layer, besides the traditional high-performance parallel file system, we specifically equipped high-volume SSD storage for each computing node, making the infrastructure compatible with distributed storage file system like HDFS, and meet the data-locality requirements. Based on that, the infrastructure can handle big data applications built on top of Hadoop, Spark and other related framework. To manage the data of different I/O pattern efficiently, we build multiple data management interface to upload and write data, satisfying different application mode.

To adapt to the computing pattern of artificial application, like deep learning application, each computing node equipped with 8 Nvidia Tesla P100 GPUs, and inter-connected by 56Gbps InfiniBand network. We use high efficient and powerful scheduler system respectively based on LSF and Apache Yarn. For both system, leveraging the container and other resource isolation and technologies, the infrastructure implement CPU affinity, GPU affinity and other features to improve computing and communication efficiency. Multiple scheduling policy and queue management policy are supported by the infrastructure, so application with different resource requirement characteristics can be managed reasonably and fully utilize the computing resource. Based on the container technology, users can build their own software stack image, manage their software and framework by a more easy-to-use way than the traditional HPC user environment.

Specifically, we build interface between common deep learning framework and the scheduler system, like tensorflow, caffe, pytorch, mxnet. For distributed deep-learning applications, user need not to explicitly allocate resource detailly in their code, the scheduler will handle most of the work. Furthermore, we pay attention to applying MPI and NCCL technology in common deep-learning

frameworks on the infrastructure, aiming to make a full use of the InfiniBand network and improving the communication performance.

To make the infrastructure more easy to use by artificial intelligence developers, different kinds of assistance service are being built, including visualization tools and training monitoring tools for the common deep learning frameworks, that uses can view their model structure and training process from web.

The infrastructure has been adopted by users from different fields. An atmospheric Science application using deep-learning networks to predict the weather, the test results shows that the infrastructure make a significant performance improvement than their own deep-learning clusters. Other applications include a research of applying reinforcement learning to poker games, a research of building algorithmic trader based on reinforcement learning leveraging large scale of financial trading data, and application of face recognition.

In the future, the infrastructure will be connected with other commercial computing service platform through grid technology, like Kingsoft's Cloud Service. Furthermore, the infrastructure will connected to the China Nation Grid, to offer service for more academic and enterprise users.

**Networking, Security, Infrastructure & Operation Session** / **26**

## Creating a trust-group for security information sharing

**Author:** Romain Wartel[1]

[1] *CERN*

**Corresponding Author:** romain.wartel@cern.ch

A growing number of organisations realise the only cost-effective mean to protect themselves against cybercrime is to join forces with peer organisations and operate a community-based response to common security threats.
However, organising such a joint response is challenging both from the trust and technical aspects. This presentation provides guidance and advice on how to setup a trust group, how to enable new participants to join, how to encourage active participation and information sharing, how to provide added value for the members, and how to connect such a trust group to existing forums in the Internet community.

**Infrastructure Clouds & Virtualisation Session** / **27**

## Harvesting dispersed computational resources with Openstack: a Cloud infrastructure for the Computational Science community

**Author:** Mirko Mariotti[1]

**Co-authors:** Antonio Guerra [2]; Daniele Spiga [3]; Giuseppe Vitillaro [4]; Loriano Storchi [5]; Manuel Ciangottini [3]; Matteo Duranti [3]; Matteo Mergé [6]; Paolo D'Angeli [2]; Roberto Primavera [2]; Valerio Formato [3]

[1] *Department of Physics and Geology, University of Perugia*

[2] *Space Science Data Center at the Italian Space Agency.*

[3] *INFN sezione di Perugia*

[4] *Istituto di Scienze e Tecnologie Molecolari, Consiglio Nazionale delle Ricerche*

[5] *Dipartimento di Farmacia, Universita' degli Studi "G. D'Annunzio", Chieti.*

[6] *INFN sezione Roma 2*

**Corresponding Author:** mirko.mariotti@unipg.it

Harvesting dispersed computational resources is nowadays an important and strategic topic especially in an environment, like the computational science one, where computing needs constantly increase. On the other hand managing dispersed resources might not be neither an easy task not costly effective. We successfully explored the usage of OpenStack middleware to achieve this objective, aiming not only at harvesting resource but also at providing a modern paradigm of computing and data usage access.
The deployment of a multi-sites Cloud [1] implies the management of multiple computing services with the goal of sharing resources between more sites, allowing local data centers to scale out with external assets. The cited scenario immediately points out portability, interoperability and compatibility issues, but on the other side this opens really interesting scenarios and use cases especially in load balancing, disaster recovery and advanced features such as cross-site networks, cross-site storage systems, cross-site scheduling and placement of remote VMs. Moreover, a cloud infrastructure allows the exploitation of Platform as a Service, and software as a server solutions which in the end it is what really matter for the science.
The main goal of the present work is to illustrate a real example on how to build a geographically distributed cloud to share and manage computing and storage resources, owned by heterogeneous cooperating entities. We put together four different entities namely: Department of Physics and Geology - I.N.F.N. Sez. Perugia, Department of Chemistry (UNIPG) also located in Perugia, Department of Pharmacy (UdA) located in Chieti and ASI-SSDC (Space Science Data Center at the Italian Space Agency) located in Rome. The involved sites are, not only geographically dislocated but also a good representatives of the research interest diversity we want to explore in our Cloud infrastructure. We report about the porting of concrete use cases exploiting the available PaaS solutions provided by INDIGO-DataCloud project and following the upcoming EOSC directives. We identified: ab-initio quantum chemistry applications and AMS-02 (Alpha Magnetic Spectrometer on the International Space Station) experiments.
The sites are connected via a SDN[2] technology through encrypted point-to-point channels being the “backbone” for the overlay networks. Specifically openVPN and IPsec have been used, depending on specific networking constraints. Over the point-to-point mesh, OSI L2 Ethernet frames of both the OpenStack management and projects VLAN networks have been encapsulated using the VXLAN [3] protocol. The networking structure has been deployed for the specific purpose of creating a distributed overlay L2 Ethernet to be used by OpenStack.
Finally, to fine tune the network setup and avoid cross site unnecessary traffic, different sites have been logically separated using different availability zones, with VMs on every zone having the capability of independently accessing the Internet through the site where they are physically running.

Mell, Peter, and Tim Grance. “The NIST definition of cloud computing.” 28 (2011).

Kreutz, Diego et al. “Software-defined Networking: A comprehensive Survey” arXiv:1406.0440v3.

Mahalingam et al. “Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks” - RFC 7348

**Physics & Engineering Session / 28**

## A brief history of distributed computing at the LHC

**Author:** Philippe Charpentier[1]

[1] *CERN*

**Corresponding Author:** philippe.charpentier@cern.ch

Since the inception of the MONARC model and of the Grid paradigm at the turn of the millenium, our views on how to organize distributed computing for the LHC experiments has considerably evolved. There have been several generations of tools (Grid middleware) as well as a lot of conceptual and

technical developments within the experiments in order to optimize the efficiency of data management and data processing. The rigid MONARC model has evolved towards a more versatile usage of computing resources, and new resource providers have emerged, in particular cloud providers, institutional clusters outside the Grid and volunteer computing.
in this presentation we shall review the main stages of evolution of distributed computing at the LHC during these 18 years and try and give an outlook on the possible further evolution, in view of the upcoming upgrades of the LHC and of the experiments during the next 10 years.

**VRE** / **29**

## DODAS: How to effectively exploit heterogeneous clouds for scientific computations

**Author:** Daniele spiga[1]

**Co-authors:** Cristina Duma [2]; Davide Salomoni [2]; Marica Antonacci [3]; Matteo Duranti [4]; Tommaso Boccali [5]; Valerio Formato [4]; diego ciangottini [4]; giacinto donvito [3]

[1] *INFN-PG*

[2] *INFN-CNAF*

[3] *INFN-Bari*

[4] *INFN-Perugia*

[5] *INFN-Pisa*

**Corresponding Author:** spiga@pg.infn.it

Dynamic On Demand Analysis Service (DODAS) is a Platform as a Service tool built combining several solutions and products developed by the INDIGO-DataCloud H2020 project. DODAS allows on-demand generation of a container-based HTCondor batch system over cloud-based infrastructures implementing the “Batch System as a Service” paradigm. As such, it is a cloud enabler designed for scientists seeking to easily exploit distributed and heterogeneous clouds to process data.
Aiming to reduce the learning curve as well as the operational cost of managing community specific services running on distributed cloud, DODAS completely automates the process of provisioning, creating, managing and accessing a pool of heterogeneous computing and storage resources.
DODAS has a modular architecture providing the following major features: the abstraction of the underlying compute and data resources, which is crucial for the interoperability across IaaS providers; the security toolkit, a key aspect for managing community services on opportunistic resources; the automated resource provisioning and configuration, to simplify the setup of the complex experiment computing environments.
The resource abstraction and the full automation are implemented combining together the INDIGO PaaS Orchestrator and the INDIGO Infrastructure Manager (IM). IM provides the support for several types of IaaS like, for example, OpenStack, OpenNebula, Amazon AWS and Microsoft Azure, while the PaaS Orchestrator represents the DODAS entrypoint. As such it is directly exposed to the end user, who is required to provide properly configured TOSCA templates, which will be then processed by the INDIGO PaaS layer. The cluster setup and the services configuration are automated using Ansible recipes. The TOSCA and Ansible combination guarantees an easy procedure to describe complex computing infrastructures.
The Identity Access Manager (IAM) and Token Translation Service (TTS) are used to manage user authentication/authorization both to grant access to the resources and to protect experiment services such as Workload Management Systems and data storages.
Apache Mesos is the baseline solution used by DODAS to abstract CPU, RAM and storage, while Marathon is adopted as container orchestration platform on top of Mesos. Marathon takes care of setting up both HTCondor and any additional services (e.g squid proxy, proxy certificate cache) that experiments might require. Such architecture provides high scaling capabilities and self-healing support that results in a drastic reduction of time and cost, through setup and operational efficiency increase.
The high level of modularity of DODAS is a key to its generic applicability, providing the ability to easily customize the workflow depending on the community computational requirements. Although originally designed for the Compact Muon Solenoid (CMS) Experiment at LHC, DODAS has been quickly adopted by the Alpha Magnetic Spectrometer (AMS) astroparticle physics experiment

mounted on the ISS as a solution to exploit opportunistic computing, nowadays an extremely important topic for research domains where computing needs constantly increase. Due to its flexibility and efficiency, DODAS was selected as one of the Thematic Services that will provide multi-disciplinary solutions in the EOSC-hub project, an integration and management system of the European Open Science Cloud starting in January 2018.
The main goals of this contribution are to provide a comprehensive overview of the overall technical implementation of DODAS, as well as to illustrate two distinct real examples of usage: the integration within the CMS Workload Management System and the extension of the AMS computing model.

**Heritage Science Session** / **30**

## E-RIHS' DIGILAB: A data and service infrastructure

**Authors:** Athanasios Koutoupas[1]; Franco Niccolucci[2]; Luca Pezzati[3]; Sorin Hermon[1]

[1] *The Cyprus Institute*

[2] *Vast-Lab Pin S.C.R.L.*

[3] *Consiglio Nazionale Delle Ricerche*

**Corresponding Authors:** luca.pezzati@cnr.it, franco.niccolucci@gmail.com, s.hermon@cyi.ac.cy, a.koutoupas@cyi.ac.cy

The European Research Infrastructure for Heritage Science (E-RIHS) incorporating and taking advantage of the long-term tradition of the heritage science research, the ability to combine with innovation, and the integration promoted by EU-funded projects such as ARIADNE and PARTHENOS exploits the synergy of the cooperation among the academy, research centers, and cultural institutions. E-RIHS will provide state-of-the-art tools and services to support cross-disciplinary research communities of users through its four access platforms: MOLAB, FIXLAB, ARCHLAB, and DIGILAB.
The DIGILAB platform will rely on a network of federated repositories where researchers, professionals, managers and other heritage-related professionals deposit the digital results of their work. These data will not be kept internally. On the contrary, the platform will provide access to the original repositories where the data are stored. DIGILAB inspired by the FAIR principles (Findable-Accessible-Interoperable-Reusable) will provide online access and remote services to the heritage science research community. It will enable finding data through an advanced search system operating on a registry containing metadata describing each individual dataset. It will support access to the data through a federated identity system, while data access grants will be local to each repository. It will guarantee data interoperability by requiring the use of a standard data model. It will foster re-use by making services available to users, to process the data according to their own research questions and use requirements.
The platform includes and enables access to searchable registries of specialized digital resources (datasets, reference collections, thesauri, ontologies, etc.). Furthermore, DIGILAB will set up guidelines for dataset recovery and assist researchers and research institutions, while as cloud-based infrastructure will support data interoperability through the creation of shared knowledge organization systems and provides tools to process them according to researchers' needs and research questions.
Overall DIGILAB facilitates virtual access to tools and data hubs for heritage research and is designed to be the privileged gateway to European scientific knowledge in heritage.

**Biomedicine & Life Science Session** / **31**

## Building bridges between services and e-infrastructure in structural biology

**Author:** Alexandre M.J.J. Bonvin[1]

**Co-authors:** Antonio Rosato [2]; Chris Morris [3]; Joerg Schaarschmidt [1]; Jose Maria Carazo Garcia [4]; Marco Verlato [5]; Martyn Winn [3]; Mikael Trellet [1]; Tomas Kulhanek [3]

[1] *Utrecht University*

[2] *University of Florence*

[3] *STFC*

[4] *National Center for Biotechnology - CNB - CSIC*

[5] *INFN*

**Corresponding Author:** a.m.j.j.bonvin@uu.nl

Structural biology deals with the characterization of the structural (atomic coordinates) and dynamic (fluctuation of atomic coordinates over time) properties of biological macromolecules and adducts thereof. Since 2010, the WeNMR project has implemented numerous web-based services to facilitate the use of advanced computational tools by researchers in the field, using the grid computational infrastructure provided by EGI [1]. These services have been further developed in subsequent initiatives, such as the MoBrain competence center within EGI-ENGAGE.

Currently, in the context of the West-Life Virtual Research Environment H2020 project [2], we have been working on the integration of different pre-existing services to enable the combined application of tools on the same given dataset. Here we will demonstrate how a user can exploit these integrated services without the need to download partial data, entirely via a browser interface. Examples involve the calculation of protein-protein complexes from nuclear magnetic resonance (NMR) data using the FANTEN web server, followed by the optimization of the initial model using HADDOCK and fitting structures into cryo-EM maps with the PowerFit web server using maps obtained from the Scipion web portal.

In addition, we will show the implementation of a cloud storage solution developed by the West-Life H2020 project [2], called VirtualFolder [3], which allows the user to connect to her/his account on for example B2DROP or on public clouds (e.g. Dropbox). In the future, this will allow users to access datasets acquired in different experimental facilities directly from the web interfaces of the services connected to West-Life.

1. Wassenaar TA, et al. WeNMR: Structural biology on the Grid. J. Grid. Computing 10:743-767, 2012
2. https://www.west-life.eu/
3. https://portal.west-life.eu/virtualfolder/

**Summary**:

Since 2010, the WeNMR project has implemented numerous web-based services to facilitate the use of advanced computational tools by researchers in structural biology, using the grid computational infrastructure provided by EGI. These services have been further developed in subsequent initiatives, such as the MoBrain competence center within EGI-ENGAGE. In the context of the West-Life Virtual Research Environment H2020 project, we have been working on the integration of different pre-existing services to enable their combined application to the same given dataset. For this purpose, we have built bridges between web portals and with storage infrastructure to allow users to exploit these services and access their data in an integrated manner entirely via a browser interface.

**Networking, Security, Infrastructure & Operation Session / 32**

## Simulation approach for improving the computing network topology and performance of the China IHEP Data Center

**Authors:** Fazhi Qi[1]; Li Wang[1]

**Co-authors:** Andrey Nechaevskiy [2]; Darya Pryahina [2]; Gennady Ososkov [2]; Weidong Li [1]

[1] *IHEP*

[2] *JINR*

**Corresponding Author:** wangli320@ihep.ac.cn

The goal of the project is to improve the computing network topology and performance of the China IHEP Data Center taking into account growing numbers of hosts, experiments and computing resources including cloud computing environment.The analysis of the computing performance of the IHEP Data Center in order to optimize its distributed data processing system is a really hard problem due to the great scale and complexity of shared computing and storage resources between various IHEP divisions. Therefore, we decide to utilize the simulation approach and adopt as a simulation tool the simulation program CloudSim Plus [2] which enables researchers rapidly evaluate the efficiency, performance and reliability of their computing network. The simulation uses input parameters from the data base of the IHEP computing infrastructure, besides we use some data of the BESIII [3] experiments to indicate workflow and data flow parameters. The first simulation results show that the proposed approach allows us to make an optimal choice of the network topology improving its performance and saving resources.

**Data Management & Big Data Session** / **33**

## Management of Cost Effective Mass Storage Environments

**Author:** Tim Chou[1]

[1] *Brookhaven National Laboratory*

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

The Scientific Data and Computing Center (SDCC) at Brookhaven National Laboratory (BNL) is a major scientific data storage site with more than 100 PB of archived experimental data, which amount is supposed to grow exponentially during the next several years. BNL is a Tier 1 data center for the ATLAS collaboration.

The challenge is to meet the data storage and processing requirements of the Relativistic Heavy Ion Collider (RHIC) experiments at BNL and especially the newly emerging enormous storage needs of ATLAS with limited budget.

In this past year SDCC has developed several tools to lower operating costs. One of the cost saving implementations is the introduction of JBOD in addition to existing RAID systems and LTO tape systems. The SDCC-developed software gathers statistical information to provide adaptive methods and techniques for device management and rigorous performance tuning and optimization of the entire storage environment. The software suite provides an intuitive user-friendly web-based interface enabling dynamic parameter visualization based on a priority-driven device-monitoring paradigm and addresses the following major technical issues:

- proprietary JBOD monitoring and management software enhancing the scalability of JBOD systems. It can monitor, configure and optimize the disk systems at 50% of the cost of commercial RAID systems.
- developed in-house data retrieval management and optimization application. ERADAT is SDCC developed software that manages and optimizes data retrievals from mass storage media, such as tape cartridges. This application will optimize data retrievals based on the data locations on the media, and manage the device allocations for multiuser concurrent operations.
- predict device failures before they actually happen and service the devices. This self-developed software periodically queries all mass storage devices for usage history and predicts failure before it actually happens. It alerts administrators when a device usage is nearing marginal limits and when a media error or device error is detected.

- detect and protect storage devices and media from unauthorized access. All media access and data retrievals are monitored and logged. The self-developed application will alert system administrators of any unusual data access/deletion patterns.
- encryption and duplication of data as requested by users.

**Summary**:

The Scientific Data and Computing Center (SDCC) at Brookhaven National Laboratory (BNL) strives to meet the data storage and processing requirements while lowering the costs at the mean time.

**Infrastructure Clouds & Virtualisation Session** / **35**

## Integration and optimization of cloud computing within the BNL workload management system

**Author:** iris wu[1]

[1] *Brookhaven National Lab*

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

The Scientific Data and Computing Center (SDCC) oversees all scientific computing activities at BNL, and a primary goal is to provide resources to a heterogeneous and geographically dispersed user community. The SDCC currently supports HEP, NP, Astrophysics, Photon Sciences, Material Sciences, Biology and other communities. This presentation describes the SDCC on-going activities in the high-throughput and high-performance computing (HTC and HPC) domains at BNL, including workload management changes that allow HTC applications to use HPC resources (local and remote), integration of virtualization containers technologies to enable seamless access to institutional resources, BNLBox (a cloud-like storage service that allows users to share and synchronize data across devices) and enabling cost-optimized access to cloud (commercial and private) resources for time-sensitive applications. The presentation also discusses timelines and near-term activities to increase flexible and timely access to computing resources for the BNL user community.

**Infrastructure Clouds & Virtualisation Session** / **36**

## Dynamic extension of INFN-CNAF Tier1 Data Center

**Author:** Antonio Falabella[1]

**Co-author:** tommaso boccali [1]

[1] *INFN*

**Corresponding Author:** antonio.falabella@cnaf.infn.it

INFN CNAF hosts the INFN Tier-1, the main data center of INFN, the Italian National Institute for Nuclear Physics; it provides resources and services to more than 30 scientific collaborations, each utilizing a different computing model.
The largest supported collaborations are the four WLCG experiments while the remaining are mainly astro-particle experiments.
INFN CNAF currently deploys resources in excess of 200 kHS06 of Computing Power, 25 PB of disk

and 50 PB of tape, the latter two interconnected via a GPFS-TSM SAN.
Recently, we has started to elastically extend the data center farm to resources provided by commercial clouds. This will prepare the center for future upgrades and could allow to cope with burst activities. The adopted approach is via the use of VPN tunnels to remote resources, in order to serve users in a completely transparent way.
The storage can be accessed via a XrootD fallback on CNAF storage when available or using the GPFS/AFM caching system.
In this talk, we are presenting our experimentation with Microsoft Azure public cloud for the CMS specific use case, using an academic grant for experimentation. We will report on the system setup, from testing to its deployment into production, and on performance obtained while utilizing Azure's remote resources.

**Networking, Security, Infrastructure & Operation Session** / **37**

## Security Situation Assessment Method Based On States Transition

**Authors:** Chun Long[1]; Hanji Shen[1]; Jing Zhao[1]; Peng Gao[1]; Wei Wan[1]

[1] *Computer Network Information Center of Chinese Academy of Sciences*

**Corresponding Author:** wanwei@cnic.cn

With the development of demands in the network security operation, how to assess the network security situation becomes a research hotspot. In order to solve the problem that the security situation of current network cannot be reflected by the alarm information from security equipment, the security situation assessment model based on state transition was built with HMM, by re-searching hosts states and analysing events affected states transition. This method is effective in training the parameters of the model, and it can analyse the security situation quantitatively and qualitatively. At last, the result validates the method by the historical security data in CSTNET.

**Networking, Security, Infrastructure & Operation Session** / **38**

## Building a large scale Intrusion Detection System using Big Data technologies

**Author:** Pablo Panero[1]

**Co-authors:** Cristian Schuszter [1]; Liviu Valsan [1]; Romain Wartel [1]; Vincent BRILLAULT [2]

[1] *CERN*

[2] *CERN/EGI*

**Corresponding Authors:** liviu.valsan@cern.ch, pablo.panero@cern.ch

Computer security threats have always been a major concern and continue to increase in frequency and complexity. The nature and techniques of the attacks evolve rapidly over time, making the detection of attacks more difficult, therefore the means and tools used to deal with them need to evolve at the same pace if not faster.

In this paper a system for intrusion detection (IDS) both at the network (NIDS) and host (HIDS) level is presented. The system is currently processing in real time approximately half a TB of data per day, with the final goal of coping with 2.5 TB. In order to accomplish this goal firstly an infrastructure to collect data from sources such as system logs, web server logs and the network based Intrusion Detection System logs has been developed making use of technologies such as Apache Flume and Apache Kafka. Once the data is collected it needs to be processed in search of malicious activity: the

data is consumed by Apache Spark jobs which compare in real time this data with known signatures of malicious activities. These are known as IoC or Indicator of Compromise, they are published by many security experts and centralized in a local MISP (Malware Information Sharing Platform) instance.

Nonetheless, detecting an intrusion is not enough. There is a need to understand what happened and why. In order to gain knowledge on the context of the detected intrusion the data is also enriched in real time
when it is passing through the pipeline. For example DNS resolution and IP geolocation are applied to the it. Therefore, a system generic enough to process any kind of data in JSON format is enriching the data in order to get full context of what is happening and finally looking for indicators of compromise to detect possible intrusions, making use of the latest technologies in the Big Data ecosystem.

**39**

## CSTCloud: A Cloud Computing Platform Designed for Scientific Researcher

**Author:** Honghai Zhang[1]

**Co-authors:** Jun Li [1]; LeiLei Zhang [1]; Ting Wei [1]

[1] *Computer Network Information Center, Chinese Academy of Sciences*

**Corresponding Author:** zhh@cnic.cn

Many cloud enterprises appear in recent years such as Aliyun, Jinshanyun. While these cloud enterprises mainly serve to business companies. Their cloud resources and software services may not satisfy the need of some scientists and researchers. In order to especially support scientists and researchers to use and manage various kinds of scientific and technological resources and services consistently, transparently and on myself, we develop a cloud platform exclusively for scientists and researchers based on our superiority on distributed computing, distributed storage, big data analysis, information services, software defined networks and so on. The platform shall have following features: (1) The platform deeply integrates public and exclusive base information resources inside. In addition, it also joints the resources outside so that various kinds of information resources such as network, data, cloud computing, high performance computing and storage can be integrated to improve resource utilization and sharing level. Then the scientists and researchers can obtain rich information resources and service environment. (2) All these information resources are loosely coupled so that they can join and quit the cloud platform dynamically. Researchers can use local resources preferentially as well as share their idle resources to the platform so that all the kinds of resources can be managed uniformly and scheduled dynamically. (3) As the resources are distributed, heterogeneous and dynamical, the relative services and applications are diverse and the management need are different, we develop integrative and distributed monitoring system which can uniformly manage and schedule information resources and provide support to multi-layered resources sharing on demand and coordination on self. (4) We shall provide diverse toolkits in the platform so as to provide more stable and mature IT services for users.In a word, we will build an safe and reliable resource integrated system which is oriented for various application needs and which can support resource sharing on demand and schedule dynamically.

**Networking, Security, Infrastructure & Operation Session / 40**

## CernVM-FS - status and latest developments from RAL Tier-1 perspective

**Author:** Catalin Condurache[1]

[1] *STFC Rutherford Appleton Laboratory*

**Corresponding Author:** catalin.condurache@stfc.ac.uk

The CernVM File System (CernVM-FS) was developed to assist WLCG High Energy Physics (HEP) collaborations to deploy software on the worldwide distributed computing infrastructure used to run data processing applications. CernVM-FS has been the primary method for distributing WLCG experiment software and condition data for the last 7 years, and in the same period the use of CernVM-FS outside WLCG has been growing steadily and an increasing number of Virtual Organizations (VOs), both within the HEP and in other communities (i.e. Space, Natural and Life Sciences), have identified this technology as a more efficient way of maintaining and accessing software across Grid and Cloud computing environments.

This presentation will give an overview of the CernVM-FS infrastructure deployed at RAL Tier-1 as part of the WLCG Stratum-1 network, but also as the facility provided to setup a complete service - the Release Manager Machine, the Replica Server and a customized uploading mechanism - for the non-LHC communities within EGI and that can be used as a proof of concept for other research infrastructures and communities looking to adopt a common software repository solution.

The latest developments to widen and consolidate the CernVM-FS infrastructure as a global facility (with main contributors in Europe, North America and Asia) are reviewed, such as the implementation of the 'confidential' CernVM-FS repositories, a requirement for academic communities willing to use CernVM-FS technology. The presentation will include a case study of a Life Science research community, describing the design of their production models around 'public' and 'confidential' CernVM-FS repositories and we examine how a common mechanism to access the latter is developed using Robot X.509 Grid certificates.

**41**

## MOVIDA 2.0: A digital tool for documentation and analysis of artworks scientific data

**Author:** Anna Amat[1]

**Co-authors:** Brunetto G. Brunetti [2]; Costanza Miliani [2]; Franco Niccolucci [3]; Luca Pezzati [4]

[1] *Istituto di Scienze e Tecnologie Molecolari del CNR (CNR-ISTM); SMAArt c/o Dipartimento di Chimica di Perugia*

[2] *Istituto di Scienze e Tecnologie Molecolari del CNR (CNR-ISTM)*

[3] *PIN, Prato,Italy*

[4] *Istituto Nazionale di Ottica INO CNR, Italy*

**Corresponding Author:** anna.amatalb@gmail.com

The European Research Infrastructure for Heritage Science[1], E-RIHS is a distributed research infrastructure, organised in national networks and coordinated by national hubs that is preparing to be launched as a standalone research infrastructure consortium in 2021. E-RIHS will provide state-of-the-art tools and services to support cross-disciplinary research communities of users through four access platforms. Two pillars of E-RIHS are DIGILAB and the MOLAB access.

MOLAB offers access to advanced mobile analytical instrumentation for diagnostics of heritage objects while DIGILAB[2] is designed to be the privileged gateway to European scientific knowledge in heritage. DIGILAB will enable the heritage science research community to access information of analyses, conservation and restoration by providing searchable registries containing metadata that describes individual datasets. Moreover, it fosters re-use by making a series of services available for the discovery and access of relevant information.

In this framework, Movida 2.0 intends to be a bridge between DIGILAB and MOLAB access, providing a service for the data management and analysis of complex multi-technique diagnostic projects. The software is developed in Java and the data stored in a SQLite DBMS that has been designed to comply with the semantic graph built in the DIGILAB registry. Movida 2.0 datasets are thus directly importable to the registry making results of the MOLAB investigations re-usable by all the community.

The software has been completely redisigned from its former version, MOVIDA 1.0 that was mainly devoted to puntcual spectral data.[3,4] This new version has been developed exploiting exisiting java gis libraries to manage more complex data as that comming from multi and hyper spectral techniques.

Movida 2.0 has been tailored to the needs of all the MOLAB providers of IPERION-CH[5], gathering the state-of-the-art instrumentation in heritage science. All the information generated in a MOLAB campaign (raw and elaborated data, annotations, metadata) can be managed within the same application and the information can be easily consulted, compared and analyzed. The software is self-comprehensive and user-friendly and can be used by all the people involved heritage science and preservation.

Movida 2.0, not only allows for the digital preservation of all the information and knowledge acquired in an analytic campaign, but can also be used as an analytical tool to inquiry and make cross-examination of the data. Moreover, as an integrating part of a digital infrasturcture as DIGILAB, it will encourage and facilitate knowledge sharing and collaboration between researchers and conservators.

References:

[1] http://www.e-rihs.eu/

[2] L. Pezzati and A. Felicetti ERCIM NEWS 111 October 2017, p. 26

[3] A. Amat et al,. J Cult Herit. 2013; 14, 23-30

[4] F. Rosi et al. Heritage Science, 2016, 4, doi:10.1186/s40494-016-0089-y

[5] http://www.iperionch.eu/molab/

**Summary**:

Movida 2.0 is a java-based software for the data management and analysis of complex multi-technique diagnostic projects.
The software will be integrated in the DIGILAB platform as a service for the MOLAB investigations within E-RIHS, facilitating knowledge sharing and collaboration between researchers and conservators.

**Networking, Security, Infrastructure & Operation Session** / **42**

# Harnessing the Power of Threat Intelligence in Grids and Clouds: WLCG SOC Working Group

**Authors:** David Crooks[1]; Liviu Valsan[2]

[1] *University of Glasgow*

[2] *CERN*

**Corresponding Author:** liviu.valsan@cern.ch

The modern security landscape affecting Grid and Cloud sites is evolving to include possible threats from a range of avenues, including social engineering as well as more direct approaches. An effective strategy to defend against these risks must include cooperation between security teams in different contexts. It is essential that sites have the ability to share threat intelligence data with confidence, as well as being able to act on this data in a timely and effective manner.

As reported at ISGC 2017, the WLCG [1] Security Operations Centres Working Group has been working with sites across the WLCG to develop a model for a Security Operations Centre reference design. This work includes not only the technical aspect of developing a security stack appropriate for sites of different sizes and topologies, but also the more social aspect of sharing data between

groups of different kinds. In particular, since many Grid and Cloud sites operate as part of larger University or other Facility networks, collaboration between Grid and Campus/Facility security teams is an important aspect of maintaining overall security.

We discuss recent work on sharing threat intelligence, particularly involving the WLCG MISP [2] instance located at CERN. In addition, we examine strategies for the use of this intelligence, as well as considering recent progress in the deployment and integration of the Bro Intrusion Detection System at contributing sites.

An important part of this work is a report on the first WLCG SOC WG Workshop/Hackathon, a workshop planned at time of writing for December 2017. This workshop is planned to assist participating sites in the deployment of these security tools as well as giving attendees the opportunity to share experiences and consider site policies as a result. This workshop is hoped to play a substantial role in shaping the future goals of the working group.

[1] Worldwide LHC Computing Grid
[2] Malware Information Sharing Platform

**Networking, Security, Infrastructure & Operation Session** / **43**

# FIM4R version 2 - Federated Identity Requirements for Research

**Author:** David Kelsey[1]

**Co-authors:** Hannah Short [2]; Peter Gietz [3]; Scott Koranda [4]; Tom Barton [5]

[1] *STFC-RAL*

[2] *CERN*

[3] *DAASI*

[4] *LIGO*

[5] *UChicago & Internet2*

**Corresponding Author:** david.kelsey@stfc.ac.uk

Federated identity management (FIM) is an arrangement that can be made among multiple organisations that lets subscribers use the same identification data to obtain access to the secured resources of all organisations in the group. In many research communities there is an increasing interest in a common approach to FIM as there is obviously a large potential for synergies. FIM4R [1] provides a forum for communities to share challenges and ideas, and to shape the future of FIM for our researchers. Current participation covers life sciences, humanities, and physics, to mention but a few. In 2012 FIM4R converged on a common vision for FIM, enumerated a set of requirements and proposed a number of recommendations for ensuring a roadmap for the uptake of FIM is achieved [2]. FIM4R is currently working on an updated version of this paper, to be published in spring 2018.

How can research communities leverage identity federations for the benefit of their researchers? What are the specific challenges faced by virtual organisations in heterogeneous computing environments? The experience gained by FIM4R over the past 5 years has led to a general acceptance of the necessity of proxies, to act as a mediator between identity providers and the services leveraged by research disciplines. Although many aspects of a federated Authentication and Authorization Infrastructure (AAI) are simplified by the use of such a proxy, several new complexities have been introduced. The adoption of proxies represents a shift in the approach to identity federation and provides the opportunity to reflect on whether our requirements for research environments can be adequately served by existing technologies.

The second whitepaper to be authored by FIM4R is currently under draft and will document the progress made in FIM for Research, in addition to the current challenges. Preliminary requirements gathering activities at FIM4R's 11th Workshop in Montreal and at Internet2's Technology Exchange

2017 has identified multiple areas of focus. It is hoped that FIM4R version 2 will be a source of input for federation operators, technology providers, and collaborative projects such as AARC [3] and GN4 [4] as they consider their work plans for the coming years.

During this presentation we will share the conclusions of this second FIM4R whitepaper and present a summary of the identified requirements and recommendations.

[1] https://fim4r.org
[2] https://fim4r.org/documents/
[3] https://aarc-project.eu
[4] https://www.geant.org/Projects/GEANT_Project_GN4

**Biomedicine & Life Science Session** / **44**

## Bridging artificial intelligence and physics-based docking for better modelling of biomolecular complexes

**Author:** Li Xue[1]

**Co-author:** Alexandre M.J.J. Bonvin [1]

[1] *Utrecht University*

**Corresponding Author:** l.xue@uu.nl

Proteins and other biomolecules, such as DNA and RNA, are the minimal functional entities that realize life. Understanding how they execute their functions through their 3D structures and interaction dynamics provides a fundamental view of what defines life. This knowledge also allows us to exploit or modify these elegant molecules for a wide variety of purposes, such as gene therapy, drug design, immunotherapy, novel enzymes and others.

Since it is still challenging to experimentally study protein interactions at high-resolution, we combine the computational power of data-driven machine learning with physics-based molecular docking to better model 3D protein complexes (static) and their binding affinity (thermodynamic). Specifically, we exploit deep learning and graph theory to tackle the major challenge in docking, namely scoring, the identification of correct conformations from a large pool of docked conformations, which still suffers from a low success rate. Another challenge in the field of biomolecular interaction is to quantitatively characterize the impact of a mutation on the binding affinity of a complex. For this purpose, we developed iSEE, a fast and reliable predictor for binding affinity changes upon single point mutation. iSEE is trained and cross-validated over a diverse dataset consisting of 1102 mutations in 57 protein-protein complexes, and it outperforms state-of-the-art methods on two independent test datasets. Reliably predicting binding affinity changes upon mutations has wide applications in, for example, the study of the relation between coding variants and phenotype, design of antigen-binding CDR loops of antibodies, and engineering binding specificity.

Our projects integrate statistical models learned from the huge wealth of heterogeneous experimental data with fundamental physics rules governing protein interactions at atomic level. By combining machine learning with physics-based modelling we aim to markedly enhance our capability to reliably model biomolecular complexes. This will help molecular biologists formulate testable hypotheses to answer important questions about cells' molecular machinery and aid the development of new therapeutics.

**Heritage Science Session** / **45**

## A DIGITAL MAPPING OF THE MALAY PENINSULA: ISLAM, HINDU AND BUDDHIST PLACES OF WORSHIP

**Author:** Faridah Mohd. Noor[1]

**Co-author:** Andrew Howard [2]

[1] *University of Malaya*

[2] *National University of Australia*

At the University of Malaya a digital humanities project was conducted to locate and map places of worship built in the Malay Peninsula before 1960. This presentation is part of a major project on constructing a digital cultural atlas featuring the location of mosques, Hindu and Buddhist temples that were constructed before the independence of Malaya in 1957. Fieldwork was conducted to obtain and verify GIS information at sites around the states of West Malaysia. Information gathered from document analysis, interviews and oral history formed part of the methodology employed and transferred to a digital map of the Peninsular Malaysia.

The digital map system used in this project is based on the widely used Open Source Leaflet mapping toolkit which supports using a range of publicly available base maps and overlay layers. The mapping component can operate either as a standalone interface or be integrated into other services, can be easily adapted to display a range of information and styled to match the project. Using a simple template for information entry, styled markers are used to represent each site with colour designating the time period.

The coming of Hinduism, Buddhism and Islam to the Malay Peninsula has been established to have spread throughout the Malay Archipelago even before the 1st Century. Based on the dates these religious sites were built, we hope to show the pattern of distribution of these places of worship and investigate the direction how these religions spread in the Malay Peninsula based on the mapping of these sites.

**Summary**:

Keywords: Religions, Digital humanities, Malaysia, Cultural heritage, Cultural map, Leaflet

**Networking, Security, Infrastructure & Operation Session** / **46**

## Provenance as a Building Block for an Open Science Infrastructure

**Author:** Andreas Schreiber[1]

[1] *German Aerospace Center*

**Corresponding Author:** andreas.schreiber@dlr.de

In science, results that are not reproducible by peer scientists are valueless and of no significance. Good practices for reproducible science are to publish used codes under Open Source licenses, perform code reviews, save the computational environments with containers (e.g., Docker), use open data formats, use a data management system, and record the provenance of all actions.

This talk focuses on provenance of scientific processes as a foundation of open reproducible science and a building block of a distributed trustful open science infrastructure.

The concept of provenance is introduced and the W3C standard model PROV is presented. PROV gives an ontology, a data model, and specifications for provenance notations, for accessing and querying, and for mapping to other standards. For practical use of provenance in science processes, the talk gives strategies for recording and storing provenance from scientific workflow systems. Also recording provenance from scripts and from the generation of documents are presented.

For storing provenance, the talk shows how to facilitate graph databases (such as Neo4j), since the provenance of processes is a directed acyclic graph (DAG). Based on graph query languages such as

Cypher or GraphQL, these provenance information can be analyzed. For example, the provenance can be used to proof the compliance of a scientific process, to collect all data that contributed to to scientific results such a journal paper, or detect any issues regarding privacy, security, and trust of the scientific data.

To assure trust in the scientific process, we present a technology for storing provenance and related data in blockchains and blockchain-like databases. Using blockchains for storing (provenance) graphs, one can detect if the graphs have been manipulated. This gives a higher degree of confidence that the scientific data has been produced using the described processes. We describe how provenance graphs are mapped to blockchains and show how to used blockchains implementations for storing this provenance graphs practically.

**Biomedicine & Life Science Session** / **47**

## First-principle-based and data-driven design of therapeutic peptides

**Author:** Lee-Wei Yang[1]

**Co-authors:** Cheng-Yu Tsai [2]; Hongchun Li [3]; Hui-Yuan Yu [4]; Jya-Wei Cheng [4]

[1] *National Tsing Hua University*

[2] *NTU*

[3] *University of Pittsburgh*

[4] *NTHU*

**Corresponding Author:** lwyang@life.nthu.edu.tw

To combat antibiotic resistance demands timely improvement of drug efficacy. Here we introduce a computational approach that systematically creates variants from known antimicrobial peptides (AMPs) and selects the ones with improved potency based on physical quantities computed from a short, so called “immerse and surface” (IAS), molecular dynamics (MD) simulation. The designed in-silico platform was used to evaluate the efficacy of a series of tryptophan-rich AMPs, targeting Gram-positive and negative lipid membranes without a mediating receptor. Structures of three AMPs were solved by nuclear magnetic resonance (NMR) spectroscopy revealing their formation of α-helices on membrane, supported by circular dichroism (CD) data. Insertion orientations of AMPs predicted by MD simulations were validated by Paramagnetic relaxation enhancement (PRE) experiments. It was found, from both NMR experiments and computation, that antimicrobial efficacy of AMPs increased with the characterized insertion depth. Simulations reveal atomistic details of the AMP insertion over time and the importance of spatial arrangement of key residues synergistically mediating the insertion. To preserve important insertion patterns, peptide X, an AMP with already characterized potency, has its sequence circularly shuffled to create 13 variants. Partition free energies of all these AMPs are calculated from the SAS simulations and found to highly correlate with their minimal inhibitory concentration (MIC) values. The resulting correlation coefficient (>0.84) is found much favorably compare to those obtained from data-driven (machine learning) algorithms implemented in published web servers. Our computational platform capable of predicting AMP potency for peptides with similar primary sequences lends itself well to the development of new AMP design as well as improving existing AMPs in our continuous battle against pathogenic microbes. If time allows, I will also mention how accumulated structural data could help the design of anti-cancer peptides.

**Summary**:

A new computationl platform to design and screen therapeutic peptides will be reported

**Physics & Engineering Session** / **48**

## Construction of real-time monitoring system for Grid services based on log analysis at the Tokyo Tier-2 center

**Author:** Tomoe Kishimoto[1]

**Co-authors:** Hiroshi Sakamoto [2]; Nagataka Matsui [2]; Tetsuro Mashimo [2]; Tomoaki Nakamura [3]

[1] *University of Tokyo*

[2] *The University of Tokyo*

[3] *KEK*

**Corresponding Author:** tomoe@icepp.s.u-tokyo.ac.jp

The Tokyo Tier-2 center, which is located in the International Center for Elementary Particle Physics (ICEPP) at the University of Tokyo, is providing computer resources for the ATLAS experiment in the Worldwide LHC Computing Grid (WLCG). The official site operation in the WLCG was launched in 2007 after several years of development. The site has been achieving a stable and reliable operation since then.

We replaced almost all hardware devices in every three years in order to satisfy the requirement of the ATLAS experiment. The latest hardware upgrade was done in January 2016, and the new system (so-called 4th system) is stably running. In the 4th system, 6144 CPU cores (256 worker nodes) and 7392 TB disk storages are reserved for the ATLAS experiment. For the Grid middlewares, the ARC-CE is deployed as the computing element in front of the HTCondor batch job scheduler. The disk storage consists of 48 sets of a disk array and a file server, which is managed by the Disk Pool Manager (DPM).

Logs produced by these Grid services provide useful information to determine whether the services are working correctly. For example, the job slot occupancy, the job success rate, the job duration and so on can be measured by parsing log files in the ARC-CE + HTCondor system. Therefore, we are constructing a new real-time monitoring system based on log analysis using the ELK stack in order to detect the problem of the Grid services. The ELK stack provides an efficient way of log processing, storing, query and visualization. In this poster, the status of construction of this new real-time monitoring system based on log analysis at the Tokyo Tier-2 center will be described. Improvements in terms of flexibility and reliability of the site operation by introducing the new monitoring system will also be discussed.

**Earth, Environmental Science & Biodiversity Session / 49**

## Monitoring of coral reef ecosystem: an integrated approach of marine soundscape and machine learning

**Author:** Tzu-Hao Lin[1]

**Co-authors:** Frederic Sinniger [2]; Saki Harii [2]; Tomonari Akamatsu [3]; Yu Tsao [1]

[1] *Research Center for Information Technology Innovation, Academia Sinica*

[2] *Tropical Biosphere Research Center, University of the Ryukyus*

[3] *National Research Institute of Fisheries Science, Japan Fisheries Research and Education Agency*

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

Coral reefs represent the most biologically diverse marine ecosystem, however, they are vulnerable to environmental changes and impacts. Therefore, information on the variability of environment and biodiversity is essential for the conservation management of coral reefs. In this study, a soundscape monitoring network of shallow and mesophotic coral reefs was established in Okinawa, Japan. Three autonomous sound recorders were deployed in water depths of 1.5 m, 20 m, and 40 m since May 2017. To investigate the soundscape variability, we applied the periodicity-coded non-negative matrix factorization to separate biological sounds and the other noise sources displayed

on long-term spectrograms. The separation results indicate that the coral reef soundscape varied among different locations. At 1.5 m depth, biological sounds were dominated by snapping shrimp sounds and transient fish calls. Although not knowing the specific source, noises were clearly driven by tidal activities. At 20 m and 40 m depths, biological sounds were dominated by nighttime fish choruses and noises were primary related to shipping activities. Furthermore, the clustering result indicates the complexity of biological sounds was higher in mesophotic coral reefs compare to shallow-water coral reefs. Our study demonstrates that the integration of machine learning in the analysis of soundscape is efficient to interpret the variability of biological sounds, environmental and anthropogenic noises. Therefore, the conservation management of coral reefs, especially those rarely studied such as mesophotic coral reefs, can be facilitated by the long-term monitoring of coral reef soundscape.

**Data Management & Big Data Session** / **50**

## EOS Open Storage - the CERN storage ecosystem for scientific data repositories

**Author:** Andreas-Joachim Peters[1]

[1] *CERN*

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

EOS Open Storage platform is a software solution for central data recording, user analysis and data processing.

In 2017 the EOS system at CERN provided 250 PB of disk storage on more than 50k disks. Most of the stored data originates from the Large Hadron Collider and various other experiments at CERN.

Originally developed as a pure disk storage system, EOS has been extended with interfaces to support data life cycle management and tiered storage setups. A work flow engine allows to trigger chained work flows on predefined storage events to notify external services on arrival, retrieval or deletion of data. It is planned to connect EOS during 2018 to the CERN tape archive system (currently archiving 200 PB) to optimize data durability and costs.

Unlike many classical storage systems, EOS is also designed and used for distributed deployments in WAN environments. The CERN storage setup is distributed over two computer centers in Geneva and Budapest. Another example is a distributed setup in Australia for the CloudStor service of the australian research network AARNet.

EOS participates in the eXtreme Data Cloud project with the goal to use EOS as an overlay layer to implement a data lake concept where various non-uniform storage systems can be virtualized into a centrally managed distributed storage system. The described deployment model allows to optimize redundancy parameters on a higher level for cost reduction and availability optimization - in contrast to a simplistic file replication model where local redundancy within servers or a storage site is not taken into account.

To enable scientific collaboration and interactive data analysis EOS is used as the back-end implementation for CERNBox with file synchronization, sharing and collaborative editing capabilities - and SWAN as a service for web-based data analysis (Jupyter notebook interface). The CERN OpenData digital repository based on Invenio uses EOS as its storage back-end. The enabling functionality for these front-end services is a very versatile quota and access control system built-in to EOS and a large variety of access protocols optimized for LAN and WAN usage.

Another major development area in the EOS ecosystem is a FUSE based file system client providing low-latency/high-throughput access to EOS data via a file system interface. The third generation allows kerberos or certificate based authentication with similar performance levels of distributed file systems like AFS and allows transparent re-exporting via NFS or CIFS protocol.

**Earth, Environmental Science & Biodiversity Session / 51**

# A Deep Learning Approach to Tropical Cyclone Intensity Estimation Using Satellite-based Infrared Images

**Author:** Jays Samuel Combinido[1]

**Co-authors:** Jeffrey Aborot [1]; John Robert Mendoza [1]

[1] *Department of Science and Technology – Advanced Science and Technology Institute*

**Corresponding Author:** jaysamuel@asti.dost.gov.ph

The standard method for estimating tropical cyclone (TC) intensity is by interpreting TC structure from geostationary satellite images, except when aircraft reconnaissance flights are routine. Since much of a TC life cycle occurs over an ocean, satellites are an excellent source of meteorological information for the task as it offers coverage much better than traditional observations. While satellite imagers do not directly measure variables related to TC strength such as winds and central pressure, TC intensity can be gleaned based on the extent and organization of its cloud patterns. One of the first comprehensive pattern recognition methods for estimating TC intensity using satellite-based images is known as the "Dvorak technique" introduced by Vernon Dvorak. The technique uses visible and infrared (IR) brightness temperature images to extract specific cloud patterns to detect rotation, deep thunderstorms, and eye; then relates them to the deepening or weakening of a TC.

Satellite-based TC intensity estimation is tackled as a feature extraction and pattern recognition task. In its basic principle, the detection of a well-formed eye is indicative of an organized TC structure. A well-formed eye is likely a sign of a strong TC. Likewise, visually disorganized cloud signatures are regarded as an attribute of weaker TCs. The original Dvorak technique is a rule-based algorithm that looks for central dense overcasts (CDO) and outer banding features of a TC to determine its intensity. Several objective methodologies developed, thereafter, also involved prior manual extraction of satellite image features that relate to TC intensity. These features included cloud organization properties, inner core characteristics, the radius of curvature and the rain rate. Recent approaches to image-based classification tasks though enabled implicit extraction of features from images. Particularly, artificial neural networks (ANN) designed for image classification such as the Convolutional Neural Network (CNN) learns about relevant low to high-level features through several layers of artificial neurons.

In this work, we posit that: since the features are directly engineered from TC satellite images, TC intensity estimation can be performed directly over the images without the need for explicit feature extraction. This idea transforms the task of satellite-based TC intensity estimation from a feature extraction and pattern recognition task into an image classification problem.

In this study, we investigate transfer learning of a pre-trained CNN architecture for estimating TC intensity based on grayscale IR images of TCs in the Western North Pacific basin; and identify relevant TC features relating to TC intensity. Transfer learning is the process of extending the knowledge of a pre-trained ANN to learn additional new knowledge. We particularly use the VGG19 network architecture for this task in this study. Furthermore, we qualify how much information IR image signatures offer in the estimation of TC maximum wind speeds. Different from previous works, the study explores a different approach without explicitly extracting TC features. To our knowledge, this is by far the first attempt use CNN transfer learning to estimate TC intensities solely based on grayscale IR images of TCs.

Specifically, we make the following contributions

- We construct a CNN model using transfer learning with VGG19 network architecture for the purpose of TC intensity estimation through satellite IR image classification.
- We show that our model performs well on 2015-2016 TC satellite images and is comparable to previous methods which require explicit feature extraction.
- We show that our model is also able to learn TC features which are previously found to be strong indicators of TC intensity, such as cloud formation and presence or absence of a well-formed eye.

**Summary**:

The standard method for estimating tropical cyclone (TC) intensity is by interpreting TC structure from geostationary satellite images, except when aircraft reconnaissance flights are routine. Since much of a TC life cycle occurs over an ocean, satellites are an excellent source of meteorological information for the task as it offers coverage much better than traditional observations. While satellite imagers do not directly measure variables related to TC strength such as winds and central pressure, TC intensity can be gleaned based on the extent and organization of its cloud patterns. One of the first comprehensive pattern recognition methods for estimating TC intensity using satellite-based images is known as the "Dvorak technique" introduced by Vernon Dvorak. The technique uses visible and infrared (IR) brightness temperature images to extract specific cloud patterns to detect rotation, deep thunderstorms, and eye; then relates them to the deepening or weakening of a TC.

Satellite-based TC intensity estimation is tackled as a feature extraction and pattern recognition task. In its basic principle, the detection of a well-formed eye is indicative of an organized TC structure. A well-formed eye is likely a sign of a strong TC. Likewise, visually disorganized cloud signatures are regarded as an attribute of weaker TCs. The original Dvorak technique is a rule-based algorithm that looks for central dense overcasts (CDO) and outer banding features of a TC to determine its intensity. Several objective methodologies developed, thereafter, also involved prior manual extraction of satellite image features that relate to TC intensity. These features included cloud organization properties, inner core characteristics, the radius of curvature and the rain rate. Recent approaches to image-based classification tasks though enabled implicit extraction of features from images. Particularly, artificial neural networks (ANN) designed for image classification such as the Convolutional Neural Network (CNN) learns about relevant low to high-level features through several layers of artificial neurons.

In this work, we posit that: since the features are directly engineered from TC satellite images, TC intensity estimation can be performed directly over the images without the need for explicit feature extraction. This idea transforms the task of satellite-based TC intensity estimation from a feature extraction and pattern recognition task into an image classification problem.

In this study, we investigate transfer learning of a pre-trained CNN architecture for estimating TC intensity based on grayscale IR images of TCs in the Western North Pacific basin; and identify relevant TC features relating to TC intensity. Transfer learning is the process of extending the knowledge of a pre-trained ANN to learn additional new knowledge. We particularly use the VGG19 network architecture for this task in this study. Furthermore, we qualify how much information IR image signatures offer in the estimation of TC maximum wind speeds. Different from previous works, the study explores a different approach without explicitly extracting TC features. To our knowledge, this is by far the first attempt use CNN transfer learning to estimate TC intensities solely based on grayscale IR images of TCs.

Specifically, we make the following contributions

- We construct a CNN model using transfer learning with VGG19 network architecture for the purpose of TC intensity estimation through satellite IR image classification.
- We show that our model performs well on 2015-2016 TC satellite images and is comparable to previous methods which require explicit feature extraction.
- We show that our model is also able to learn TC features which are previously found to be strong indicators of TC intensity, such as cloud formation and presence or absence of a well-formed eye.

**Humanities, Arts & Social Sciences Session** / **53**

## Skill-based Occupation/Job Recommendation system

**Author:** Ankhtuya Ochirbat[1]

**Co-author:** Timothy K.Shih [2]

[1] *National University of Mongolia*

[2] *National Central University*

**Corresponding Author:** ankhaa3@yahoo.com

A mass of adolescents has decided their occupations/jobs/majors out of proper and professional advice from school services. For instance, adolescents do not have adequate information about occupations/jobs, what occupations can be reached by which majors, and what kind of education and training are needed for particular jobs. On the other hand, major choices of adolescents are influenced by a society and their family. They receive occupational information in common jobs from the environment. But they are a lack of information in professional occupations.
Furthermore, the choice of major has become increasingly complex due to the existence of multiple human skills which mean each person has their ability at the certain area and can be applied to multiple jobs/occupations. Their major choices are influenced by society, education environment, and mostly their families. Those pitfalls are potentially the causes of a mismatch major between academic achievements, personality, interest and skills of students. It would be useful to understand how students' choice of the academic majors depends on personal characteristics, competencies, and vocational interests. Most of the students do not possess adequate information about meaning of occupations/majors, what careers can be reached by which majors, and what kind of skills and abilities are needed for a particular occupation/major. For those reasons, students need an automatic counselling system according to their values.
To do this, occupation recommendation system is implemented with a variety of IT and soft skills. The main goal of this research is to build an occupation recommendation system (ORS) by using data mining and natural language processing (NLP) methods on open educational resource (OER) and skill dataset, in order to help adolescents. The system can provide different variety of academic programs, related online courses (e.g., MOOCs), required skills, ability, knowledge, and job tasks, and jobs currently announced as well as relevant occupational descriptions. The system can assist adolescents in major selection and career planning. Furthermore, the system incorporates a set of searching results, which are recommended using similarity measurements and hybridization recommendation techniques. These methods serve as a base for recommending occupations that meet interests and competencies of adolescents.

**VRE / 54**

## Workload management for heterogeneous multi-community grid infrastructures

**Author:** Andrei Tsaregorodtsev[1]

[1] *CPPM-IN2P3-CNRS*

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

Grid infrastructures are providing access to computing resources using transparent uniform tools to multiple user communities of different sizes exploiting applications with different properties and requirements. Traditional grid computing resources, the so-called High Throughput Computing (HTC) clusters can be complemented with virtualized cloud and High Performance Computing (HPC) resources. The user workflows can require access to either of these types of resources or can use all of them. This can be done with job scheduling systems that can submit user payloads to various types of the computing resources transparently.
The DIRAC project Workload Management System is providing a job scheduler that spans various heterogeneous computing resources and provides means for the users to select dynamically those resources that correspond to their application requirements. Various policies can be applied to activities of different users and groups to allow fair sharing of the communal resources. In particular HTC and cloud resources can be used together for the same user payloads. Special attention is paid to the use of the cloud resources respecting quotas for different user groups, which allows to share those resources efficiently.
In this contribution we describe the use of the DIRAC Workload Management System (WMS) based on the example of the European Grid Infrastructure (EGI). We outline the general architecture of the system and give details on its practical operations. Experience with the replacement of the traditional grid middleware WMS will be presented as well as the necessary developments to meet the requirements of different user communities and applications.

**Networking, Security, Infrastructure & Operation Session** / **55**

## WLCG Tier-2 site at NCP, Status Update and Future Direction

**Author:** Saqib Haleem[1]

**Co-author:** Muhammad Imran [2]

[1] *National Centre for Physics, Islamabad, Pakistan*

[2] *National Centre for Physics*

**Corresponding Authors:** muhammad.imran@ncp.edu.pk, saqib.haleem@ncp.edu.pk

National Centre for Physics (NCP) in Pakistan, maintains a large computing infrastructure for scientific community. Major portion of computing and storage resources are reserved for CMS experiment of WLCG project, and small portion of the computing resources are reserved for other non EHEP scientific experiments. For efficient utilization of resources, most of the scientific organizations have migrated their resources on Cloud. NCP has also taken initiative last year, and migrated most of their resources on scientific cloud. HT-condor based batch system has been deployed for local experimental high energy physics community, to perform their analysis task. Recently we deployed HT-Condor Compute element (CE) as a gateway for CMS jobs. On a network side, our Tier-2 site is completely accessible and operational on IPv6. Moreover, we recently deployed Perfsonar node, to actively monitor the throughput and latency issues between WLCG sites. This paper discusses status of NCP Tier-2 site, current challenges and future direction.

**Distributed & Parallel Computing Applications Session** / **56**

## Science Gateway on GARUDA GRID for Open Source Drug Discovery (OSDD) community

**Author:** Karuna Prasad[1]

**Co-authors:** Janaki CH [1]; Mangala N [1]

[1] *C-DAC*

**Corresponding Author:** karunap@cdac.in

Scientific applications are moving to the distributed environment like grid computing and cloud to take advantage of the high-end computational power availability with reduced cost. Large datasets and compute intensive analysis needs compute infrastructure. The researchers/users want a platform/gateway where they have ease of submitting their experiments accessing a web browser to run them on high end computational resources and if required to share them with other researchers. Science gateways allow science & engineering communities to access shared data, software, computing services, instruments, educational materials, and other resources specific to their disciplines. In this paper we describe, how the open source Galaxy Workflow is customized to provide a science gateway for Open Source Drug Discovery community on GARUDA GRID. GARUDA grid is an aggregation of heterogeneous resources which aims to bridge gaps between the researchers and accessibility to high computational power. Galaxy is an open source web based workflow system for data intensive biomedical research. Open Source Drug Discovery (OSDD) is a CSIR led team India Consortium and platform for drug discovery which brings together informaticians, scientists and research organizations.

Galaxy based Science Gateway for OSDD is enabled with secure access to GARUDA through in-house developed Login Service, job runner for the Gridway metascheduler, seamless access to the storage server, web services of bio-informatics tools that are added as tools and deployed and hosted on GARUDA over internet.

Firstly, we describe extending the pluggable architecture of Galaxy workflow for execution on the grid with the grid middleware-Gridway. Gridway metascheduler is the GARUDA workload manager that performs job execution management and resource brokering. For managing jobs submission

with Garuda grid, a Gridway runner was developed which will be managing the execution of jobs. It will include, preparing the jobs for submission and creating a job wrapper, putting it in a queue to be submitted, capturing the Job Id, monitoring the Job Id – watches the jobs currently in the queue and deals with the state change (queued to running and job completion), and finishing the job - Get the output/error for a finished job, pass to JobWrapper class finish method and cleanup all the temporary files.
The paper also talks about the integration of Garuda Login service with the Science Gateway and how the authentication is managed by the Digital certificates for rights delegation.
Science Gateway for OSDD community is designed to support the large distributed community involved in drug discovery. The users are freed from the issues of maintenance and cost and can focus on large data analysis experiments more. There has been substantial increase of resources after the release of gateway mainly due to ease of usage. Also, the implementation of job runner was specifically done for the Gridway metascheduler, but can be extended to support other types of grid systems.

**Summary**:

Science Gateway for OSDD community is designed to support the large distributed community involved in drug discovery. The users are freed from the issues of maintenance and cost and can focus on large data analysis experiments more. There has been substantial increase of resources after the release of gateway mainly due to ease of usage. Also, the implementation of job runner was specifically done for the Gridway metascheduler, but can be extended to support other types of grid systems.

**Networking, Security, Infrastructure & Operation Session** / **58**

## Optical Interconnects for Cloud Computing Data Centers: Recent Advances and Future Challenges

**Author:** Muhammad Imran[1]

**Co-author:** Saqib Haleem [2]

[1] *National Centre for Physics*

[2] *National Centre for Physics, Islamabad, Pakistan*

**Corresponding Author:** muhammad.imran@ncp.edu.pk

Internet traffic has been increasing exponentially over the last few years due to the emergence of new end user applications which are based on cloud computing infrastructure. These applications run on the servers deployed in the data centers and require huge network bandwidths. The data centers are getting more and more importance in our lives because the cloud computing has shifted computation and storage away from desktops to large scale datacenters. Traditional cloud computing data centers architecture is based on a hierarchical design and comprises several layers of electrical switches at the edge and the core. There are significant challenges to meet the growing performance requirements with current data center architectures. For example, high power consumption, high traffic locality, support of higher data rates, scalability, latency and network oversubscription. Effective optical interconnect is a fundamental requisite to realize Internet-scale data centers due to the capabilities and benefits of optical devices. An optical interconnection system can meet the above mentioned challenges due to the properties of optical components. Basic fundamental elements in optical networks are optical switches, optical transceivers and optical fibers. Optical switches are power efficient and consume less power than electrical switches. Optical interconnects can provide high capacity links to meet requirements for traffic locality and higher bit rates by using optical fibers and optical transceivers. This paper presents a brief overview on optical interconnects for data centers. Furthermore, the paper provides a qualitative categorization and comparison of the proposed schemes based on their main features. Moreover, various types of optical switches and optical switching techniques that can be considered in designing an optical interconnection system for data center networks are presented. In the end, future research direction, challenges and opportunities of optical interconnect for data centers are discussed.

**Networking, Security, Infrastructure & Operation Session / 59**

# A Study of Credential Integration Model in Academic Research Federation Supporting a Wide Variety of Services

**Author:** Eisaku Sakane[1]

**Co-authors:** Kento Aida [1]; Motonori Nakamura [1]; Takeshi Nishimura [1]

[1] *National Institute of Informatics*

**Corresponding Author:** sakane@nii.ac.jp

Single Sign-on mechanism raises usability of Information Communication Technology (ICT) service and is currently an essential technology. It is an ideal situation for users to be able to receive the desired services with only one credential. However, there is still a situation where they need to use each appropriate credential according to services.

The purpose of this paper is to investigate the situation where users must use each credential according to the desired services, and to clarify the problems in the situation and the addressed issues. Then, a credential integration model is considered.

In Japan, there is the GakuNin which is an academic access management federation. In the federation, if the negotiation between an identity provider (IdP) and a service provider (SP) have been completed regarding a use contract for the service which the SP offers, all constituent members of the academic institution which operates the IdP will be able to receive the services with the credential issued by the IdP. The restricted services, for example, a service for the staff of the academic institution only, can be offered by specifying the attributes which the IdP manages.

Also, Japan has the HPCI project that is a national project and offers high performance computing infrastructure (HPCI) to not only academic researchers but also industrial ones. In order to use HPCI, first, researchers must apply a research project proposal. If the proposal is accepted, the researchers will obtain HPCI credential after initial identity vetting based on a face-to-face meeting. The HPCI credential is issued by an IdP in the HPCI federation. Since the HPCI opens the door to not only academia but also industry the IdPs in the HPCI federation cannot be simply replaced with the IdPs in the GakuNin. However, if the HPCI user belongs to an academic instituion the user will be compelled to manage both the GakuNin and the HPCI credentials. Such credential management burden is one of the issues addressed in this paper.

In this paper, based on the situation in Japan mentioned above, we discuss a credential integration model in order to more efficiently use a wide variety of services. We first characterize services in an academic federation from point of view of authorization and investigate the problem that users must use each credential issued by different IdPs. Then, we discuss the issues to integrate user's credentials, and consider a model that solves the issues.

**60**

# Deep Learning with Evolutionary Algorithms

**Author:** Ruediger Berlich[1]

**Co-authors:** Ariel Garcia [2]; Sven Gabriel [2]

[1] *Gemfony scientific UG (haftungsbeschraenkt)*

[2] *Gemfony scientific*

**Corresponding Author:** r.berlich@gemfony.eu

The presentation discusses the application of evolutionary algorithms to deep learning. It shortly introduces the topic of deep learning, before discussing

common approaches to training “deep” neural networks. It then compares these training algorithms with Evolutionary Algorithms as a means of powerful, large-scale minimization
applied to neural networks and identifies fields of opportunity as well as difficulties. compared to standard algorithms. The presentation is based on experiences made with the Geneva (Grid-enabled evolutionary algorithms) toolkit. Geneva implements Evolutionary Algorithms, Swarm Algorithms, Simulated Annealing and Gradient Descents with configurable backends ranging from GPGPU to large scale HPC clusters.

**Physics & Engineering Session / 61**

## Extending WLCG Tier-2 Resources using HPC and Cloud Solutions

**Author:** Jiri Chudoba[1]

**Co-author:** Michal Svatos [1]

[1] *Institute of Physics of the CAS, Prague*

**Corresponding Author:** jiri.chudoba@cern.ch

Available computing resources limit data simulation and processing of LHC experiments. WLCG Tier centers connected via Grid provide majority of computing and storage capacities, which allow relatively fast and precise analyses of data. Requirements on the number of simulated events must be often reduced to meet installed capacities. Projection of requirements for future LHC runs shows a significant shortage of standard Grid resources if a flat budget is assumed. There are several activities exploring other sources of computing power for LHC projects. The most significant are big HPC centers (supercomputers) and Cloud resources provided both by commercial and academic institutions.
The Tier-2 center hosted by the Institute of Physics (IoP) in Prague provides resources for ALICE and ATLAS collaborations on behalf of all involved Czech institutions. Financial resources provided by funding agencies and resources provided by IoP do not allow to buy enough servers to meet demands of experiments. We extend storage resources by two distant sites with additional finance sources. Xrootd servers in the Institute of Nuclear Physics in Rez near Prague store files for the ALICE experiment. CESNET data storage group operates dCache instance with a tape backend for ATLAS (and Pierre Auger Observatory) collaboration. Relatively big computing capacities could be used in the national supercomputing center IT4I in Ostrava. Within the ATLAS collaboration, we explore two different solutions to overcome technical problems arising from different computing environment on the supercomputer. The main difference is that individual worker nodes do not have an external network connection and cannot directly download input and upload output data. One solution is already used for HPC centers in the USA, but until now requires significant adjustments of procedures used for standard ATLAS production. Another solution is based on ARC CE hosted by the Tier-2 center at IoP and resubmission of jobs remotely via ssh. We will also report on our experience with resource extensions via Open Nebula Cloud provided by CESNET.

**Data Management & Big Data Session / 62**

## Progress on Machine and Deep Learning applications in CMS Computing

**Authors:** A Repečka[1]; Christian Contreras[2]; Daniele Abercrombie[3]; Daniele Bonacorsi[4]; Jean Roch Vlimant[5]; K Kančys[1]; Luca Giommi[6]; Tommaso Diotalevi[6]; Valentin Kuznetsov[7]; Z Matonis[1]

[1] *Univ. Vilnius*

[2] *DESY*

[3] *MIT*

[4] *University of Bologna*

[5] *Caltech*

[6] *Univ. Bologna*

[7] *Univ. Cornell*

**Corresponding Author:** daniele.bonacorsi@unibo.it

Machine and Deep Learning techniques are being used in various areas of CMS operations at the LHC collider, like data taking, monitoring, processing and physics analysis. A review a few selected use cases shows the progress in the field, with highlight on most recent developments, as well as an outlook to future applications in LHC Run III and towards the High-Luminosity LHC phase.

**Humanities, Arts & Social Sciences Session** / **64**

## Authorship recognition and disambiguation of scientific papers using a neural networks approach

**Authors:** Luca Tomassetti[1]; Sebastiano Fabio Schifano[1]; Tommaso Sgarbanti[2]

[1] *University of Ferrara and INFN*

[2] *University of Ferrara*

**Corresponding Authors:** sebastiano.schifano@unife.it, luca.tomassetti@unife.it

One of the main issues affecting the quality and reliability of bibliographic records retrieved from digital libraries – such as Web of Science, Scopus, Google Scholar and many others – is the autorship recognition and author names disambiguation. So far these problems have been faced using methods mainly based on text-pattern-recognition for specific datasets, with high-level degree of errors.

In this paper, we propose an approach using neural networks to learn features automatically for solving authorship recognition and disambiguation of author names. The network learns for each author the set of co-writers, and from this information recovers authorship of papers. In addition, the network can be trained taking into account other features such as author affiliations, keywords, projects and research areas.

The network has been developed using the TensorFlow framework, and run on recent Nvidia GPUs and multi-core Intel CPUs. Test datasets have been selected from records exported in RIS format from the Scopus digital library, for several groups of authors working in the fields of computer science, environmental science and physics. The proposed methods achieves accuracies above 99% in authorship recognition and is able to effectively disambiguate homonyms.

We have taken into account several network parameters, such as training-set size and batch size, number of levels and hidden units, threshold and weight initialization, back-propagation algorithms, and analyzed the impact on accuracy of results.

This approach can be easily extended to any dataset and any bibliographic records provider.

**Distributed & Parallel Computing Applications Session** / **65**

## DevOps adoption in scientific applications: DisVis and PowerFit cases

**Authors:** Alexandre M.J.J. Bonvin[1]; Mikael Trellet[1]; Pablo Orviz[2]

[1] *Utrecht University*

[2] *CSIC*

**Corresponding Authors:** a.m.j.j.bonvin@uu.nl, orviz@ifca.unican.es

The DevOps methods emphasize the commitment to software Quality Assurance (QA) procedures during the development phase to avoid disruptions when the new software releases are deployed into production. Automation drives the whole process of testing, distribution and, eventually, deployment to allow prompt and frequent software releases, resulting in more reliable software.

INDIGO-DataCloud project progressively adopted DevOps practices that ruled the development (continuous integration) and distribution (continuous delivery) of the core services. The successful experience was exported to the scientific communities participating in the project, by prototyping a continuous deployment scenario for the Python-based DisVis and PowerFit applications. The implementation provides a test and validation suite that ensures the viability of any change in the application's source code, by the automatic and sequential execution of the 1) code inspection, 2) software packaging, and 3) application validation tasks.

The source code inspection tries to detect issues at early stages in the development phase, otherwise tougher to resolve once the software has been released. Any type of source code testing (style, unit, functional, integration) can fit at this stage.The QA checks executed for the prototype are devoted to guarantee the readability and sanity of the code by making it compliant with Python's PEP8 standard. The application packaging phase deploys the application, packages it as a container image and publishes it to the INDIGO-DataCloud's online repository. The uploaded image is tagged with the current code version, so it does not overwrite any previous one. The recently created version is then fetched and tested as the final step in the pipeline. The application validation phase executes the container with a set of defined inputs and the results obtained are programmatically compared with reference results.

The adoption of DevOps methods has proven to be an effective and suitable solution for the development and distribution of scientific software. DisVis and PowerFit application developers can detect promptly design issues or bugs in their code, while the users of these applications are provided with more frequent software versions, validated through the automated pipeline, being able to safely explore the latest features of the software.

**Data Management & Big Data Session / 66**

## Experiences of hard disk management in a erasure coded 10 petabyte-scale Ceph cluster

**Authors:** George Vasilakakos[1]; Tom Byrne[1]

**Co-author:** Rob Appleyard [1]

[1] *STFC*

**Corresponding Author:** rob.appleyard@stfc.ac.uk

RAL has developed a disk storage solution based on Inktank and Red Hat's 'Ceph' object storage platform to support the UK's WLCG Tier One centre. This solution, known as 'Echo' (Erasure-Coded High-throughput Object store), is now providing ~40% of the disk storage available to the LHC experiments and wider WLCG at the UK Tier 1.

Data is stored in the cluster using erasure coding to reduce the space overhead while keeping the data safety and availability high in the face of disk failures. An erasure coding profile of 8 data shards plus 3 erasure coding shards is used to provide data safety levels comparable to a system with 4-fold replication at a fraction of the cost, keeping the space overheads down to 37.5%.

In large-scale Ceph clusters, deployment of storage nodes without RAID controllers is preferred, mainly for performance reasons. This comes at a cost however, as any and all disk errors become visible to Ceph and will cause inconsistencies to be noticed that must be dealt with, generating

administrative workload for the service's operators. Factors from disk age and disk placement in the physical machine to disk utilisation and load contribute to create and operational environment that requires constant attention and administrative intervention.

In this paper, we present the history of disk-related problems that have affected Echo. Echo consists of approximately 2200 disks with a raw capacity approaching 13 PiB. A certain quantity of disk problems are to be expected when running this quantity of hard disks. However, we have also experienced a problem where undetected bad sectors on disks have a destructive interactaction with a Ceph bug that can trigger cascading failures during routine data movement operations. We present methods the Ceph team have utilised to cope with this workload, as well as how the disk management tooling provided by the Ceph project has evolved.

**Data Management & Big Data Session** / **67**

## What goes up must go down: A case study from RAL on the process of shrinking an existing storage service

**Author:** Rob Appleyard[1]

**Co-author:** George Patargias [1]

[1] *STFC*

**Corresponding Author:** rob.appleyard@stfc.ac.uk

Much attention is paid to the process of how new storage services are deployed into production that the challenges therein. Far less is paid to what happens when a storage service is approaching the end of its useful life. The challenges in rationalising and de-scoping a service that, while relatively old, is still critical to production work for both the UK WLCG Tier 1 and local facilities are not to be underestimated.

RAL has been running a disk and tape storage service based on CASTOR (Cern Advanced STORage) for over 10 years. CASTOR must cope with both the throughput requirements of supplying data to a large batch farm and the data integrity requirements needed by a long-term tape archive. A new storage service, called 'Echo' is now being deployed to replace the disk-only element of CASTOR, but we intend to continue supporting the CASTOR system for tape into the medium term. This, in turn, implies a downsizing and redesign of the CASTOR service in order to improve manageability and cost effectiveness. We will give an outline of both Echo and CASTOR as background.

This paper will discuss the project to downsize CASTOR and improve its manageability when running both at a considerably smaller scale (we intend to go from around 140 storage nodes to around 20), and with a considerably lower amount of available staff effort. This transformation must be achieved while, at the same time, running the service in 24/7 production and supporting the transition to the newer storage element. To achieve this goal, we intend to transition to a virtualised infrastructure to underpin the remaining management nodes and improve resilience by allowing management functions to be performed by many different nodes concurrently ('cattle' as opposed to 'pets'), and also intend to streamline the system by condensing the existing 4 CASTOR 'stagers' (databases that record the state of the disk pools) into a single one that supports all users.

**Data Management & Big Data Session** / **68**

## dCache - running a fault-tolerant storage over public networks

**Author:** Tigran Mkrtchyan[1]

**Co-authors:** Albert Rossi [2]; Dmitry Litvintsev [2]; Gerd Behrmann [3]; Juergen Starek [1]; Marina Sahakyan [1]; Olufemi Adeyemi [1]; Patrick Fuhrmann [4]; Paul Millar [1]

[1] *DESY*

[2] *FNAL*

[3] *NDGF*

[4] *DESY/dCache.org*

**Corresponding Authors:** patrick.fuhrmann@desy.de, tigran.mkrtchyan@desy.de

As a robust and scalable storage system, dCache has always allowed the number of storage nodes and user accessible endpoints to be scaled horizontally, providing several levels of fault tolerance and high throughput. Core management services like the POSIX name space and central load balancing components however are merely vertically scalable. This greatly limits the scalability of the core services as well as provides single points of failures. Such single points of failures are not just a concern for fault tolerance, but also prevent zero downtime rolling upgrades. For large sites, redundant and horizontally scalable services translate to higher uptime, easier upgrades, and higher maximum request rates. This becomes more important with growing demand in muti-site distributed deployments.

Since version 2.16 dCache team have made a big effort to move towards redundant services in dCache. The low level UDP based service discovery is replaced with widely-adopted Apache-ZooKeeper, which is a redundant, persistent, hierarchical directory service with strong ordering guarantees. As ZooKeeper itself run in a fault-tolerant cluster mode with strong consistency guarantee, it become a natural place to keep shared state or take a role of service coordination.

Many dCache internal services are updated and can run in replicated mode by providing truly fault-tolerant deployment. However, as any distributed service, dCache is affected by a network partitioning. In terms of so-called CAP theorem, dCache will prefer consistency over availability and return an error or timeout if data consistency can not be guaranteed.

Yet another aspect of distributed deployments, is the security. The different components must be authenticated and communication must be secure. The latest dCache versions provide a mechanism to use standard PKI infrastructure to achieve secure network communication as well as inter-component authentication.

With this presentation we will show how to deploy distributed, fault-tolerant dCache to provide reliable storage. We will discuss some technical details, share experience and lessons learned.

**Humanities, Arts & Social Sciences Session / 69**

## Responding to Environmental Change: Research Collaborations, Integrated Data Systems, and Deep Mapping

**Author:** David J. Bodenhamer[1]

[1] *Indiana University-Purdue University*

**Corresponding Author:** intu100@iupui.edu

In 2016, Indiana University created a multi-campus interdisciplinary team and charged it to understand how climate change has affected and will affect the citizens of Indiana. Six interdisciplinary research clusters are examining the biologic, environmental, and human and cultural dimensions of climate change. The Polis Center is developing a distributed spatial data platform to allow the research and findings to be integrated and visualized within and across the geography of the state. The aim is to produce a deep map of the data, both qualitative and quantitative, that allows researchers to examine problems within a spatio-temporal framework that scales easily from small units such as neighborhoods (or smaller) to the state, as well as to any number of intermediate geographies. Tools will exit within the platform to permit easier transformation, management, and visualization of the data, with the aim to make data available in as near real-time fashion as possible. More importantly, the deep map that results must be flexible, supporting different perspectives and capable

of managing both expert and native knowledge, the later contributed by citizen scientists who will be an important part of the initiative.

The presentation will outline the requirements for the integrated data spatial data platform as well as the resulting system schema. It will suggest how technologies such as GIS may be linked with other non-spatial modules to construct a dynamic interdisciplinary virtual research environment for experts as well as a system that invites volunteered information and citizen participation. It will outline as well the schema for producing on-the-fly advanced visualizations for purposes of research and public consumption. Finally, the presentation will invite both ideas and collaboration from attendees at the conference. Ultimately, grand challenges are collaborative ventures, and this presentation will seek to model how such collaborations may cross continents and not simply disciplines.

**70**

## Efficient Energy Utilization in Fog Computing Based Wireless Sensor Networks

**Author:** Arslan Rafi[1]

**Co-authors:** Adeel Rehman [2]; Ilyas Ahmad [3]

[1] *Student*

[2] *Co-Supervisor*

[3] *Supervisor*

**Corresponding Author:** arslanrafi@ymail.com

Abstract

Recent technology trends in proficient utilization of resources and powerful computing as well as storage requirements of cloud networks inject new vitality into wireless sensor networks (WSNs). In its basic form, a wireless sensor network includes a large number of sensing nodes that are deployed at some remote locations in order to collect useful data by continuously monitoring the surrounding environment. Since each node is equipped with a fixed battery source that have limited lifetime, the network needs to use its energy requirement in an efficient way in order to maximize its lifetime. Development of energy efficient protocol is an important challenge in WSNs for which different techniques have been proposed in the literature. Among these, the classical low-energy adaptive clustering hierarchy (LEACH) protocol is well used in various applications along with its variants. Recently it is observed that the LEACH protocol can be combined with the Dijkstra's algorithm so that the nodes transmit data from node to node in an optimal way until it reaches the cluster head, which then transmit it to the base station. In this paper, we proposed a modified form of the combined settings, LEACH protocol with Dijkstra's algorithm, in order to reduce the energy requirements of the network. In addition, the proposed settings can improve the packet latency of some application specific data travelling between nodes and cloud end. The idea is to introduce a model of edge or fog computing that act as an intermediate network between sensor nodes and the cloud network. Strengthening the stability of the fog-supported wireless sensor network while conserving energy consumption of the underlying nodes increases the lifetime of the overall network. Different functionality tests have been carried out and results of the proposed framework are compared with the performance of the classic LEACH implementation. The results show that the proposed framework improves the network lifetime by efficiently utilizing the energy of the network.

KEYWORDS: Wireless Sensor Networks, LEACH protocol, Dijkstra's Algorithm, Fog computing

**Data Management & Big Data Session / 71**

# Next Generation Data Management Services: the eXtreme DataCloud project

**Author:** Daniele Cesini[1]

**Co-authors:** Alessandro Costantini [1]; Christian Ohmann [2]; Cristina Duma [1]; Fernando Aguilar [3]; Giacinto Donvito [4]; Lukasz Dutka [5]; Patrick Fuhrmann [6]; Serena Battaglia [2]; Vincent Poireau [7]; de Lucas Jesus Marco [3]; matthew Viljoen [8]; oliver keeble [9]; rachid lemrani [10]

[1] *INFN-CNAF*

[2] *ECRIN*

[3] *University of Cantabria*

[4] *INFN-BARI*

[5] *AGH-CYFRONET*

[6] *DESY/dCache.org*

[7] *CNRS-LAPP*

[8] *EGI*

[9] *CERN*

[10] *CNRS*

**Corresponding Author:** daniele.cesini@cnaf.infn.it

The eXtreme DataCloud (XDC) is a EU H2020 funded project aimed at developing scalable technologies for federating storage resources and managing data in highly distributed computing environments. The services provided will be capable of operating at the unprecedented scale required by the most demanding, data intensive, research experiments in Europe and Worldwide. XDC will be based on existing tools, whose technical maturity is proved and that the project will enrich with new functionalities and plugins already available as prototypes (TRL6+) that will be brought at the production level (TRL8+) at end of XDC. The targeted platforms are the current and next generation e-Infrastructures deployed in Europe, such as the European Open Science Cloud (EOSC), the European Grid Infrastructure (EGI), the Worldwide LHC Computing Grid (WLCG) and the computing infrastructures funded by other public and academic initiatives. The main high-level topics addressed by the project include: i) federation of storage resources with standard protocols, ii) smart caching solutions among remote locations, ii) policy driven data management based on Quality of Service, data lifecycle management, metadata handling and manipulation, data preprocessing and encryption during ingestion, optimized data management based on access patterns.
All the developments will be community-driven and tested against real life use cases provided by the consortium partners representing research communities belonging to a variety of scientific domains: Life Science, Astrophysics, High Energy Physics, Photon Science and Clinical Research. The XDC project aims at opening new possibilities to scientific research communities in Europe and worldwide by supporting the evolution of e-Infrastructure services for Exascale data resources.
The XDC software will be released as Open Source platforms available for general exploitation.

**72**

# GridPP wider community support and the UKT0

**Author:** Jeremy Coles[1]

[1] *University of Cambridge*

**Corresponding Author:** jeremy.coles@cern.ch

GridPP has been a core member of the WLCG since its inception. In recent years various considerations have led to an increased focus on sharing GridPP resources and supporting new communities outside of the LHC experiments and beyond HEP. The first part of this paper brings together a review of the current approaches within GridPP to harness traditional resources for these 'other' Virtual Organisations, and also to bring in new resources such as HPC and commercial cloud providers. The

paper presents several case studies and highlights common difficulties being encountered and the solutions being proposed. The second part of the paper gives an overview of the needs of a diverse set of science communities, including GridPP, that have come together in the UK to form the "UKT0", a project that has just been awarded £1.5M to provide the first components of a joined up eInfrastructure and provision production ready capacity by March 2018. The paper describes the scope of work being undertaken in the first phase of the UKT0 and its aspirations.

**73**

## The SKA Science Data Processor and SKA Regional Centre studies

**Author:** Jeremy Coles[1]

[1] *University of Cambridge*

**Corresponding Author:** jeremy.coles@cern.ch

A community whose impacts will be strongly felt in the coming years is astrophysics. Within this domain, the Square Kilometre Array (SKA) project presents some of the biggest computing challenges. This paper presents the background to SKA computing work being taken forward by the Science Data Processor (SDP) Consortium and outlines studies informing the design (many of which parallel HEP challenges in areas such as exploiting parallelism). A summary of the challenge for SKA Regional Science Centres and current progress on the design and costing considerations of a European RSC will be given. Potential synergies with HEP distributed computing are discussed.

**Earth, Environmental Science & Biodiversity Session / 74**

## Improving biodiversity monitoring through soundscape information retrieval

**Author:** Yu Tsao[1]

**Co-authors:** Chia-Yun Lee [2]; Chiou-Ju Yao [3]; Joe Chun-Chia Huang [2]; Mao-Ning Tuanmu [2]; Tzu-Hao Lin [1]

[1] *Research Center for Information Technology Innovation, Academia Sinica*

[2] *Biodiversity Research Center, Academia Sinica*

[3] *National Museum of Natural Science*

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

Passive acoustic monitoring has been suggested as an effective tool for investigating the dynamics of biodiversity. For instance, automatic detection and classification of sounds can acquire information of species occurrences and behavioral activities of vocalizing animals. However, current methods of automatic acoustic identification of species remain uncertain for most taxa, which constrains the application of remote acoustic sensing in biodiversity monitoring. One challenge is that most of the training samples more-or-less contains undesired sound signals from non-target sources. To overcome this issue, we developed a source separation algorithm based on a deep version of non-negative matrix factorization (NMF). Using multiple layers of convolutive NMF to learn spectral features and temporal modulation of sound signals from a spectrogram, vocalizations of different species can be effectively separated in an unsupervised manner. Based on the pre-trained features, acoustic activities of target species can be efficiently separated from long-duration field recordings. Besides, spectral features of each vocalizing species can also be archived for further utilizations. In this presentation, we will demonstrate the application of deep NMF on separating sounds from different species for both birds and bats. Our results show that the proposed deep NMF approach can be used to establish recognition database of vocalizing animals for soundscape-based biodiversity monitoring, confirming its promising applicability for the field of soundscape information retrieval.

**Heritage Science Session / 75**

# Spectral Database Application for Color Compensation Process in Painting

**Author:** M. James Shyu[1]

**Co-author:** Yuan-Feng Chang [2]

[1] *Chinese Culture University*

[2] *National Normal University*

Color compensation is an important step in painting conservation process that actual pigments are re-applied on the original art works. The process of color compensation is a challenging task since most media used in East Asian paintings are water-soluble and are not protected by any surface coating. To maintain the originality of the art works, it is desirable to make the color compensation work identifiable and reversible. Consequently, the area of the newly compensated colors should be easily identified, yet still visually d agreed with the original colors. There are new synthetic pigments that showing similar colors as the traditional mineral pigments in the visible band (380-730 nm), however exhibiting totally different spectral reflectance characteristics in the infrared region than the natural mineral pigments. It is then possible to maintain the color similarity visually and to establish the identifiable difference for reversible purpose if necessary in the color compensation process. To accomplish this unique task, a spectral database is proposed to record the spectral characteristics of the synthetic and natural pigments both in visible and near-infrared bands. Through the indexing in the CIELAB color space, a color match can be searched for the natural mineral pigment to the corresponding synthetic pigment in the visible band. Furthermore, the difference among spectral characteristics can also be assured and observed through the spectral data in the infrared region in the database. With such information in the database, it is much easier to achieve the identifiability and reversibility in color compensation for East Asian paints.

**76**

# E-RIHS DIGILAB: starting a digital platform for the European Research Infrastructure for Heritage Science

**Author:** Luca Pezzati[1]

[1] *CNR INO*

**Corresponding Author:** luca.pezzati@cnr.it

E-RIHS[1] is a research infrastructure project of the European strategic roadmap (ESFRI[2] Roadmap) since 2016. E-RIHS support research on heritage interpretation, preservation, documentation and management. Both cultural and natural heritage are addressed: collections, buildings, archaeological sites, digital and intangible heritage. E-RIHS is a distributed research infrastructure: it includes facilities from many Countries, organized in national networks and coordinated by National Hubs. The E-RIHS Headquarters will be seated in Florence, Italy. E-RIHS will provide state-of-the-art tools and services to cross-disciplinary research communities of users through its four access platforms:

- MOLAB: access to advanced mobile analytical instrumentation for diagnostics of heritage objects, archaeological sites and historical monuments.
- FIXLAB: access to large-scale and specific facilities with unique expertise in heritage science, for cutting-edge scientific investigation on samples or whole objects.
- ARCHLAB: physical access to archives and collections of prestigious European museums, galleries, research institutions and universities containing non-digital samples and specimens and organized scientific information.

- DIGILAB: virtual access to tools and data hubs for heritage research – including measurement results, analytical data and documentation – from large academic as well as research and heritage institutions.

The DIGILAB platform will provide remote services to the heritage science research community, reaching out beyond the European boundaries. DIGILAB will enable access to research information as well as to general documentation of analyses, conservation, restoration and any other kind of relevant information about heritage research and background references. DIGILAB will rely on a network of federated repositories where researchers, professionals, managers and other heritage-related professionals deposit the digital results of their work. DIGILAB will not keep those data internally: instead, it will provide access to the original repositories where the data are stored. DIGILAB is inspired by the FAIR data principles: it will enable Finding the data through an advanced search system operating on a registry containing metadata describing each individual dataset; it will support Accessing the data through a federated identity system, while access grants will be local to each repository; it will guarantee data Interoperability by requiring the use of a standard data model; it will foster Re-use by making services available to users, to process the data according to their own research questions or use requirements.
E-RIHS will help the preservation of the World's Heritage by enabling cutting-edge research in heritage science, liaising with governments and heritage institutions to promote its constant development and, finally, raising the appreciation of the large public for cultural and natural heritage and the recognition of its historic, social and economic significance.

1 http://www.e-rihs.eu/
2 http://www.esfri.eu/

**Summary**:

**The European Research Infrastructure for Heritage Science, E-RIHS ['ɪris], is committed to launch its new platform in 2018. DIGILAB will be the new data and service infrastructure for the heritage science research community.**

**77**

## CHNET, the INFN initiative for Cultural Heritage

**Authors:** Andrea Ceccanti[1]; Francesco Giacomini[1]; Francesco Taccetti[1]; Lisa Castelli[1]; Luca dell'Agnello[1]

[1] *INFN*

**Corresponding Author:** luca.dellagnello@cnaf.infn.it

In the framework of the INFN initiative for Cultural Heritage (CHNet), a first implementation of a central repository of digital images has been deployed and is hosted at the INFN-CNAF data center. Access to the objects stored in the repository is granted based on attributes, e.g. belonging to a certain group of people, obtained from the Identity and Access Management service developed in the INDIGO-DataCloud project and integrated, among others, with the IDEM authentication federation. Currently the repository stores several objects, including photos and data files obtained with X-Ray Fluorescence (XRF) scans.
On top of the repository, an application is being developed in order to allow the analysis of the stored XRF objects.
The application, now available as a prototype, is mostly based on modern web technologies and doesn't require any local installation of software, lowering the access barrier; it allows also the upload of remote data sets present on the client.
In the near future the data sets will be extended to other imaging techniques such as digital radiography, tomography, etc.; also news functionalities will be added to the application in order to take into account these new data formats.

**Earth, Environmental Science & Biodiversity Session** / **81**

## Introduction of Soundscape Project

**Corresponding Author:** eric.yen@twgrid.org

**Infrastructure Clouds & Virtualisation Session** / **82**

## Integration of GPU and Container with Distributed Cloud for Scientific Applications

**Earth, Environmental Science & Biodiversity Session** / **83**

## Impacts of Horizontal Resolution and Air–Sea Flux Parameterizations on the Intensity and Structure of Tropical Cyclone

This study investigates the impacts of horizontal resolution and surface flux formulas on typhoon intensity and structure simulations through the case study of the Super Typhoon Haiyan (2013). Three different surface flux formulas in the Weather Research and Forecasting Model are tested for grid spacing from 6 to 1 km. Both increased resolution and more reasonable flux formulas can improve the typhoon intensity simulation, but their impacts on wind structures are different. Sufficiently high resolution is more conducive to the positive effect of flux formulas. Reduce the grid spacing to 1 km yields deeper and stronger, more upright and contracted eyewall. As resolution increases, the size of updraft cores in the eyewall shrinks and the region of downdraft increases, and both updraft and downdraft become more intense. In the finer resolution of simulations, the convective cores are driven by more intense updrafts within a rather small fraction of spatial area. This resolution dependence of the spatial scale of updrafts is attributed to the model effective resolution, which is determined by grid spacing, not the flux formulas. While the use of more reasonable flux formulas can increase the simulated storm intensity to some extent, the positive effect of surface flux formulas cannot be effectively enhanced unless the grid spacing is properly reduced to efficiently yield intense and contracted eyewall structure.

**Closeing Keynote & Ceremony** / **84**

## Non-Linear Earthquake Simulation on Sunway TaihuLight

**Closeing Keynote & Ceremony** / **85**

## Closing Ceremony

**Keynote Session** / **86**

## Challenging Einstein's Theory of General Relativity by Gravitational Waves with Advanced Computing Technologies

The recent historic discovery of Gravitational Waves (GW) by LIGO won the Nobel Prize in physics in 2017 and opened a new era of GW astronomy. After about 100 years since the completion of General Relativity by Einstein, for the first time we can test this theory at extreme gravity conditions by using GW signals. The advanced computing technologies such as GRID and GPU play important roles to achieve these goals. The summary of world-wide GW detection network and data analysis approaches will be discussed.

**Keynote Session** / **87**

## Preparing Applications for the New Era of Computing

We are entering a new Era of computing with large scale HPC systems soon reaching performance levels of Exaflops and large scale cloud infrastructures providing unprecedented levels of compute power. These greatly enhanced computational capabilities will enable new science, pushing both the boundaries of existing computational science and enhancing new domains like artificial intelligence and high performance data analytics.

These enormous advances are being enabled by new hardware technologies, particularly changing CPU and memory design. Unfortunately, these changes are not transparent to applications and call for revised algorithms and implementations; appropriate programming environments, workflows, and data management systems; and a deep understanding of the hardware developments to come.

Dedicated effort have started in Europe to address these issues and in this talk we review some of the key challenges and present Europe's approach to them. Highlights of some selected projects, like the BioExcel Centre of Excellence for Computational Biomolecular Research will also be presented.

**Keynote Session** / **88**

## Grid computing and cryo-EM

Throughout the last 25 years, single particle cryo-electron microscopy (cryo-EM) has continuously evolved into a powerful modality for determining the 3D structure of radiation-sensitive biological macromolecules, culminating in the award of the recent Nobel prices in Chemistry 2017. This development has been enabled by constant maturation of image processing algorithms in concert with the emergence of direct electron detectors, improvements in electron optics, stage stability and microscope automation, resulting in ever-growing image data volumes. I will give a brief overview of the state-of-the-art single particle cryo-EM workflow and the process of 3D reconstruction from images. Although the data volume generated by a single modern cryo electron microscope is still far from what detectors in particle physics produce each day, the growing interest and combined word-wide deployment of these facilities have already created bottlenecks in data storage and processing capacity. I will outline the resulting challenges faced by cryo-EM labs, describe the current modus operandi and a road map to grid computing in this field, which may enable institutions without expensive local infrastructure to benefit from distributed compute resources.

**Keynote Session** / **89**

## The European Open Cloud for e-Science towards automation, service composition, big data analytics and new frontiers in data management

**Corresponding Author:** davide.salomoni@cnaf.infn.it

**Keynote Session** / **90**

## Applying deep learning to the prediction of protein structure and function

**Keynote Session** / **91**

## Opening Remarks

**Corresponding Authors:** ludek@ics.muni.cz, simon.lin@twgrid.org

**CryoEM Workshop** / **92**

## De-demystifying 2017 Nobel Prize in Chemistry from a structural biologist view

Since the year of 2013, Low temperature transmission electron microscope operated at high voltage (cryo-EM) has suddenly emerged as a powerful of tool for elucidating virtually “solution” structures of many proteins to near atomic resolution, allowing for the building of PDB model directly from the cryo-EM “Coulomb Map”. However, the last year Nobel Prize has gone to three scholars that seems not to have direct contributions as to the technology breakthroughs that transformed cryo-EM. In this lecture, I will brief transverse the landscape the development of cryo-EM since 1974 to address milestones in the advance of cryo-EM including the pioneering effort by those three laureates, and the final removal of the key barriers of resolution by technology advance in (1) direct electron camera; (2) microscope automation; and (3) Baysian-based algorithm for 3D reconstruction, despite none of the inventor of the above has been rewarded with Nobel prize.

**CryoEM Workshop** / **93**

## EMAN2 (part 1)

**CryoEM Workshop** / **94**

## EMAN2 (part 2)

**CryoEM Workshop** / **95**

## Relion (part 1)

**CryoEM Workshop** / **96**

## Relion (part 2)

**CryoEM Workshop** / **97**

## cryoSPARC

**CryoEM Workshop** / **98**

## Introduction to Appion and Leginon tools

**CryoEM Workshop** / **99**

## Computation resources for cryoEM in Academia Sinica and their benchmark

**Corresponding Authors:** eric.yen@twgrid.org, felix@twgrid.org

**Workshop on Frontiers in Computational Drug Discovery** / **100**

## Registration & overview of the workshop

**Workshop on Frontiers in Computational Drug Discovery** / **101**

## Introduction to Protein Data Bank (PDB) and molecular graphics (PyMOL)

**Workshop on Frontiers in Computational Drug Discovery** / **102**

## Fundamentals in structure biology

**Workshop on Frontiers in Computational Drug Discovery / 103**

## Molecular graphics (UCSF Chimera) and analytics for biomolecule-drug interactions (LigPlot+, PDB2PQR, etc.)

**Workshop on Frontiers in Computational Drug Discovery / 104**

## Quantum chemical calculations of drug-like molecules

**Workshop on Frontiers in Computational Drug Discovery / 105**

## Hands-on tutorials of quantum chemical calculation with Gaussian and visualization of molecular orbitals and chemical spectra (GaussView)

**Workshop on Frontiers in Computational Drug Discovery / 106**

## Principle of molecular docking

**Workshop on Frontiers in Computational Drug Discovery / 107**

## Hands-on tutorials of AutoDock 4.0 and AutoDock vina

**Workshop on Frontiers in Computational Drug Discovery / 108**

## Deep learning approaches in computation drug discovery

**Workshop on Frontiers in Computational Drug Discovery / 109**

## Hands-on Tutorial of DeepChem and Gnina

**Workshop on Frontiers in Computational Drug Discovery / 110**

## Molecular dynamics simulations for drug-target complexes

**Workshop on Frontiers in Computational Drug Discovery / 111**

## Hands-on Tutorial of AMBER16 (xLEaP, sander, pmemd, cpptraj)

**Workshop on Frontiers in Computational Drug Discovery / 112**

## Quantum mechanical/molecular mechanical molecular dynamics simulations

**Workshop on Frontiers in Computational Drug Discovery / 113**

## Hands-on Tutorial of AMBER16 (sqm, sander, pmemd)

**Workshop on Frontiers in Computational Drug Discovery / 114**

## Gaussian accelerated molecular dynamics simulation (GaMD)

**Workshop on Frontiers in Computational Drug Discovery / 115**

## Hand-on Tutorial of AMBER16 (sander, pmemd, WHAM, UI) and Gaussian accelerated molecular dynamics (GaMD)

**Symposium on Frontiers in Computational Drug Discovery / 116**

## HADDOCK goes small molecules. Integrative modelling of biomolecular interactions from fuzzy data

**Corresponding Author:** a.m.j.j.bonvin@uu.nl

**Symposium on Frontiers in Computational Drug Discovery / 117**

## Computaional Modulator Design to Target Protein-Protein Interactions

**Symposium on Frontiers in Computational Drug Discovery / 118**

## Accelerated Computer Simulations and Drug Discovery of G-Protein-Coupled Receptors

**Symposium on Frontiers in Computational Drug Discovery / 119**

## Principles governing biological Processes: Applications to drug design and drug target identification

**Symposium on Frontiers in Computational Drug Discovery / 120**

## Harnessing structures and dynamics of biomolecules for polypharmacology-based computational drug design”

**Security Workshop / 121**

## Introduction

**Corresponding Author:** david.kelsey@stfc.ac.uk

**Security Workshop / 122**

## The Threat Landscape: Introducing terms in context of ENISA’s Threat Landscape Underground economy

**Security Workshop / 123**

## Malware Techniques

**Security Workshop / 124**

## Demonstration of typical attacks on FedCLoud Virtual Machines

**Corresponding Author:** sveng@nikhef.nl

**Security Workshop / 125**

## Discussion and Hands-on of operational security for Cloud User Communities – session 1

**Corresponding Author:** sveng@nikhef.nl

**Security Workshop** / **126**

## Discussion and Hands-on – session 2

**Corresponding Author:** sveng@nikhef.nl

**Security Workshop** / **127**

## Wrap up/conclusions

**Corresponding Author:** david.kelsey@stfc.ac.uk

**128**

## Resilience of cultural heritage to natural disasters: preventive conservation and enhancement

The theme of resilience has been widely circulated, constituting a real discipline in many cases. Its definition originates in the field of ecology and refers to the ability of a system, after disturbance, to return with sufficient rapidity, to a state close to the initial one. The general application of this concept is evident, and it has recently been extended to cultural heritage too.
The protection of cultural heritage as a priority, within the framework of international politics and disaster risk reduction programs, was included for the first time in the "Sendai Framework for Disaster Risk Reduction 2015-2030" by UNISDR (The United Nations Office for Disaster Risk Reduction).
In this context, the Accademia Nazionale dei Lincei has contributed with the proposition of the "Charter of Rome on the Resilience of Art Cities to Natural Disasters", approved by the global network of Academies IAP (the interacademy partnership) and the 'Statement on Cultural Heritage : Building Resilience to Disasters", approved by the G7 Science.
A first way to follow the guidelines contained in these strategic documents has been the recent establishment of CERHER - Center of Resilience on Heritage - an integrated skills center operating in the macro-region of central Italy: Umbria, Tuscany and Marche, which aims to develop the resilience of art cities to natural disasters. CERHER's primary objective is to act in the context that surrounds cultural heritage, building a network of active protection and risk mitigation, capable of optimizing the resilience of the art cities.
In conclusion, as a practical example of applicating the above cited guidelines, will be presented the measures of preventive conservation and enhancement of the Villa Farnesina in Rome, a splendid Renaissance suburban villa, with frescoes by Raphael.

**Earth, Environmental Science & Biodiversity Session** / **129**

## Taiwan Earthquake Model: Bring in Earthquake Science to Society

**eScience Activities in Asia Pacific Session / 130**

## eScience Activities in Japan

**Corresponding Author:** aida@nii.ac.jp

**eScience Activities in Asia Pacific Session / 131**

## eScience Activities in China

**Corresponding Author:** gang.chen@ihep.ac.cn

**eScience Activities in Asia Pacific Session / 132**

## eScience Activities in Korea

**eScience Activities in Asia Pacific Session / 133**

## eScience Activities in Taiwan

**Corresponding Author:** eric.yen@twgrid.org

**eScience Activities in Asia Pacific Session / 134**

## eScience Activities in Mongolia

**eScience Activities in Asia Pacific Session / 135**

## Panel Discussion

**eScience Activities in Asia Pacific Session / 136**

## eScience Activities in Thailand

**eScience Activities in Asia Pacific Session / 137**

## eScience Activities in Indonesia

**eScience Activities in Asia Pacific Session / 138**

## eScience Activity in Malaysia

**eScience Activities in Asia Pacific Session / 139**

## eScience Activities in Vietnam

**eScience Activities in Asia Pacific Session / 140**

## eScience Activities in Philippine

**Corresponding Author:** jaysamuel@asti.dost.gov.ph

**eScience Activities in Asia Pacific Session / 141**

## Panel Discussion

**eScience Activities in Asia Pacific Session / 142**

## Singapore Smart Nation Initiatives Using HPC

**eScience Activities in Asia Pacific Session / 143**

## eScience Activity in India

**144**

## eScience Activity in Australia

**Corresponding Author:** glenn.moloney@nectar.org.au

**eScience Activities in Asia Pacific Session / 145**

## eScience Activity in Pakistan

**Corresponding Author:** muhammad.imran@ncp.edu.pk

**eScience Activities in Asia Pacific Session / 146**

## Panel Discussion

**APGridPMA & IGTF Meeting / 147**

## Introduction

**Corresponding Author:** eric.yen@twgrid.org

**APGridPMA & IGTF Meeting / 148**

## TAGPMA Update

**APGridPMA & IGTF Meeting / 149**

## EUGridPMA & IGTF Update

**Corresponding Author:** davidg@nikhef.nl

**APGridPMA & IGTF Meeting / 150**

## Remote Vetting

**Corresponding Author:** sakane@nii.ac.jp

**APGridPMA & IGTF Meeting / 151**

## MICS CA Audit Guideline

**Corresponding Author:** sakane@nii.ac.jp

**APGridPMA & IGTF Meeting / 152**

## Self Audit Report

**APGridPMA & IGTF Meeting / 153**

## CA Report

**APGridPMA & IGTF Meeting / 154**

## Future Meeting

**APGridPMA & IGTF Meeting / 155**

## AoB

**Environmental Computing Workshop / 156**

## Regional Collaboration on Disaster Mitigation

**Corresponding Author:** eric.yen@twgrid.org

**Environmental Computing Workshop / 157**

## Storm Surge Modeling and Case Study of 2013 Super Typhoon Haiyan

**Environmental Computing Workshop / 158**

## Case Study of Philippine

**Corresponding Author:** jaysamuel@asti.dost.gov.ph

**Environmental Computing Workshop / 159**

## Case Study of Vietnam (Remote Presentation)

**Environmental Computing Workshop / 160**

## Science & Technology – Hydroinformatics Implementation for Water Related Disasters in Thailand (Remote Presentation)

**Environmental Computing Workshop** / **161**

## Case Study of Malaysia

**Environmental Computing Workshop** / **162**

## Case Study of Indonesia

**Networking, Security, Infrastructure & Operation Session** / **163**

## Modern security landscape in Taiwan

**Environmental Computing Workshop** / **164**

## Introduction

**Environmental Computing Workshop** / **165**

## Interoperation with EOSC-Hub Framework

**Corresponding Author:** ludek@ics.muni.cz

**Environmental Computing Workshop** / **166**

## Remote Sensing for Disaster Mitigation in Taiwan

**Workshop on Frontiers in Computational Drug Discovery** / **167**

## Potential of mean force and free energy calculations

**Workshop on Frontiers in Computational Drug Discovery** / **168**

## Improving Scoring-Docking-Screening Powers of Protein–Ligand Scoring Functions using Random Forest

**Workshop on Learning Analytics** / **169**

## Defining the Realm of Learning in Life

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

**Workshop on Learning Analytics** / **170**

## Concepts surrounding Learning Analytics

**Workshop on Learning Analytics** / **171**

## Demo

**Workshop on Learning Analytics** / **172**

## Second Language Writing for Meta Cognitive Reflection

**Workshop on Learning Analytics** / **173**

## Reflective Writing enhanced with ICT

**Workshop on Learning Analytics** / **174**

## Learning Analytics or Assessment for Active Learning (PBL/TBL)

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

**Workshop on Learning Analytics** / **175**

## Wrapping Up: Summary

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

**Environmental Computing Workshop** / **176**

# Environmental Computing

**Environmental Computing Workshop / 177**

## EOSC-Hub Project (Remote Presentation)

**Environmental Computing Workshop / 178**

## DRIHM WRF Portal (Remote Presentation)

**Environmental Computing Workshop / 179**

## Discussion

**Earth, Environmental Science & Biodiversity Session / 180**

## Taiwan Earthquake Science Information System and Crowdsourcing-based Earthquake Reporting Systems

**Earth, Environmental Science & Biodiversity Session / 181**

## Development of advanced infrastructure for earthquake science in Taiwan: Taiwan Earthquake Research Center

**GDB Meeting / 182**

## Introduction

**GDB Meeting / 183**

## Storage Accounting Update

**GDB Meeting / 184**

## IHEP Grid Report

**Corresponding Author:** wuwj@ihep.ac.cn

**GDB Meeting / 185**

## Security Report

**Corresponding Author:** romain.wartel@cern.ch

**GDB Meeting / 186**

## IPv6 Status

**Author:** Duncan Rand[None]

**Corresponding Authors:** david.kelsey@stfc.ac.uk, duncan.rand@imperial.ac.uk

**GDB Meeting / 187**

## Belle II Report

**GDB Meeting / 188**

## Middleware Report

**GDB Meeting / 189**

## LHCOPN/LHCONE Report

**GDB Meeting / 190**

## EISCAT 3D Report

**GDB Meeting / 191**

## AMS Report

**Workshop on Frontiers in Computational Drug Discovery / 192**

## Hands-on Tutorial of Δvina

**ECAI/ApSTi Workshop / 193**

## Our ApSTi Approach: Physical Heritage Conservation through Spatial Humanities

**ECAI/ApSTi Workshop / 194**

## Story Maps of Taipei City

**ECAI/ApSTi Workshop / 195**

## Visualizing Historical Stories Using Virtual Reality

**196**

## Crossing the Taiwan Strait: The life Cycles and Functions of Invented Epigraphic Traditions

**ECAI/ApSTi Workshop / 197**

## Tombs Research in the Ryukyus: Crossing the border between Okinawa and Kagoshima

**ECAI/ApSTi Workshop / 198**

## Variance or Uniformity: Muslim Epigraphs in Macau and Hong Kong

**199**

## Panel Discussion

**ECAI/ApSTi Workshop / 200**

## Endangered Languages and the Flow of Ethnicity: State Policies and Language Ideology among the Thao of Taiwan

**ECAI/ApSTi Workshop / 201**

## Comments on the Political Obstacles and Meaning of Cultural Preservation in Taiwan

**ECAI/ApSTi Workshop / 202**

## Cultural Asset Preservation Efforts at Xindian First Public Cemetery 2015-2017

**ECAI/ApSTi Workshop / 203**

## Update on the Research Data Repository data.depositar.io

**ECAI/ApSTi Workshop / 204**

## In Search of Agents and Routes: Mapping Stone Carving Practices in the Penghu Archipelago

**ECAI/ApSTi Workshop / 205**

## Discussion “On Cultural Conservation – Physical and Digital”

**ECAI/ApSTi Workshop / 206**

## Eternal Resting Place: Conservation of the Variety in the Muslim Section of Kaohsiung Fudingjin Public Cemetery