

# **International Symposium on Grids & Clouds 2021 (ISGC 2021)**

## **Report of Contributions**

Contribution ID: 0

Type: **Oral Presentation**

## Structural biology in the clouds: The HADDOCK/WeNMR-EOSC Ecosystem

*Tuesday, 23 March 2021 11:30 (30 minutes)*

Structural biology aims at characterizing the structural (atomic coordinates) and dynamic (fluctuation of atomic coordinates over time) properties of biological macromolecules. Gaining insight into 3D structures of biomolecules is critical for understanding the vast majority of cellular processes, with direct application in health and food sciences.

Since 2010, the **WeNMR** project ([www.wenmr.eu](http://www.wenmr.eu)) has implemented numerous web-based services to facilitate the use of advanced computational tools by researchers in the field, using the HTC infrastructure provided by EGI. These services have been further developed in subsequent initiatives under H2020 projects are a now operating as Thematic Services in the EOSC-Hub project ([www.eosc-portal.eu](http://www.eosc-portal.eu)), with the **HADDOCK** portal ([haddock.science.uu.nl](http://haddock.science.uu.nl)) sending >10 millions of jobs and using ~4000 CPU-years per year.

We will summarize 11 years of successful use of e-infrastructure solutions to serve a large world-wide community of users (>18'000 to date), providing them with user-friendly, web-based solutions that allow to run complex workflows in structural biology. We will in particular share our experience operating and deploying those services from the provider point of view with a focus on sustainability and easy of deployment of the services, making use of containerization technologies.

**Primary author:** Dr JIMÉNEZ-GARCÍA, Brian (University of Utrecht)

**Co-author:** Dr BONVIN, Alexandre M.J.J. (Utrecht University)

**Presenter:** Dr JIMÉNEZ-GARCÍA, Brian (University of Utrecht)

**Session Classification:** Biomedicine & Life Science

**Track Classification:** Health & Life Sciences (including COVID-19) Applications

Contribution ID: 2

Type: **Oral Presentation**

## **Enabling HPC systems for HEP: the INFN-CINECA experience**

*Thursday, 25 March 2021 15:00 (30 minutes)*

We want to report about a successful integration exercise between CINECA (PRACE Tier-0) Marconi KNL system and LHC processing.

A production-level system has been deployed using a 30 Mhours grant from the 18th Call for PRACE Project Access; this has needed modifications at multiple levels: experiments' WMS layers, site level access policies and routing, virtualization.

The success of the integration process paves the way to integration with additional local systems, and in general shows how the requirements of a HPC center can coexist with the needs from data intensive, complex distributed workflows.

**Primary author:** Dr BOCCALI, tommaso (INFN)

**Presenter:** Dr DAL PRA, Stefano (INFN)

**Session Classification:** Converging High Performance infrastructures: Supercomputers, clouds, accelerators Session

**Track Classification:** Converging High Performance infrastructures: Supercomputers, clouds, accelerators

Contribution ID: 4

Type: **Oral Presentation**

## Making Identity Assurance and Authentication Strength Work for Federated Infrastructures

*Thursday, 25 March 2021 11:30 (20 minutes)*

In both higher Research and Education (R&E) as well as in research-/ e-infrastructures (in short: infrastructures), federated access and single sign-on by way of national federations (operated in most cases by NRENs) are used as a means to provide users access to a variety of services. Whereas in national federations institutional accounts (e.g. provided by an university) are typically used to access services, many infrastructures also accept other sources of identity: provided by 'community identity providers', social identity providers, or governmental IDs. Hence, the quality of a user identity, for example in regard to identity proofing, enrollment and authentication, may differ - which has an impact on the service providers risk perception and thus their authorization decision.

In order to communicate qualitative information on both identity vetting and on the strength of the authentication tokens used between the identity providers and service providers, assurance information is used - with the strength being expressed by different Levels of Assurance (LoA) or 'assurance profiles' combining the various elements in community-specific ways. While in the commercial sector assurance frameworks such as NIST 800-63-3 or Kantara IAF have been established, these are often considered as too heavy with strict requirements, and not appropriate for the risks encountered in the R&E community. This is why in the R&E space a more lightweight solution is necessary.

The REFEDS Assurance Suite comprises orthogonal components on identity assurance (the REFEDS Assurance Framework RAF) and authentication assurance (Single Factor Authentication Profile, Multi Factor Authentication Profile) and provides profiles for low and high risk use cases. The Suite is applicable in many scenarios, like identity interfederations (cross-national collaborations) or for exchanging assurance information between identity providers and Infrastructure Proxies (according to AARC Blueprint Architecture). This presentation serves as a guidance on how the assurance values can be assessed and implemented with standard products (e.g. Shibboleth IdP, SimpleSAMLphp, SaToSa), and how that enables new use cases in the research infrastructures.

This talk starts with a short overview of existing assurance frameworks showing the relationships and dependencies between commercial frameworks such as NIST 800-63 and Kantara and the standards introduced in the R&E sector. Following that, use cases of the REFEDS Assurance Suite will be presented to show how the REFEDS specifications can be used to exchange identity and authentication assurance in cross-collaborative scenarios. The focus of this talk lies in providing guidance for operators to facilitate the adoption of exchanging assurance information, also in regard to proxy scenarios where assurance elements may arise from different sources.

**Primary author:** Mrs ZIEGLER, Jule A. (Leibniz Supercomputing Centre)

**Co-authors:** Dr GROEP, David (Nikhef); Dr KELSEY, David (STFC-RAL); Mr NEILSON, Ian (STFC-RAL); Mr KREMERS, Maarten (SURF); Dr STEVANOVIC, Uros (Karlsruhe Institute of Technology)

**Presenter:** Mrs ZIEGLER, Jule A. (Leibniz Supercomputing Centre)

**Session Classification:** Network, Security, Infrastructure & Operations Session

**Track Classification:** Network, Security, Infrastructure & Operations

Contribution ID: 6

Type: **Oral Presentation**

## Study of Dark Matter at e+ e- collider using KISTI-5 supercomputer

*Friday, 26 March 2021 09:30 (30 minutes)*

The Standard Model (SM) of particle physics is well established with the discovery of Higgs boson which is the last particle to be discovered in SM. Since SM can not give a description of dark matter, dark matter is barely known and under research through various method. The importance of computational science, which comprises theory, experiment and simulation in science, have been emerging in the field of high energy physics (HEP), especially, in searching for dark matter. Because the cross section of dark matter is very small compared to that of the Standard Model (SM), huge amount of calculation is required. Hence, to optimize Central Processing Unit (CPU) time is crucial to increase the efficiency of research in HEP.

We have studied the dark matter of the decay mode,  $e^+ e^- \rightarrow \mu^+ \mu^- A'$ . The signal process is dark photon which couples only to heavy leptons. We focus on the case in which dark photon decays into two muons. Therefore, this is four muon final state. In this work, using MadGraph5 as a simulation tool kit, we have studied CPU time as well as cross section according to various parameters such as center of mass (CM) energy, dark photon ( $A'$ ) mass and coupling constant. The imported theoretical model in MadGraph5 is the Simplified Model which covers SM particles, dark matter and dark photon particles. We have studied background using the SM of  $e^+ e^- \rightarrow \mu^+ \mu^- \mu^+ \mu^-$ . This result will be helpful for searching for dark matter at e+ e- collider of experiments such as Belle II, CEPC and ILC experiments.

To compare CPU time of calculation, we have used the KISTI-5 supercomputer [Nurion Knight Landing (KNL) and Skylake (SKL)] and the local Linux machine. Nurion KNL is equipped with processor Intel Xeon Phi 7250 and 68 cores per node. Nurion SKL is equipped with Intel Xeon 6148 and 40 cores per node. We have used one core or more cores to compare CPU time at machines. We have checked the optimization of the simulation tool kit by comparing the CPU time consumed for various physics modes. We have checked the physics simulation including MadGraph5 only and checked the full simulation including Pythia8, Delphes and MadAnanlysis5 as well as physics simulation. We have also checked efficiencies of parallel processing among the machines. The CPU time and the wall clock time of the KISTI-5 supercomputer and the local Linux machine with one or multiple cores have been studied for the optimization and parallelization with large number of cores of KISTI-5 supercomputer. The results will help to optimize HEP software using high performance computing (HPC).

### Summary

We have studied the dark matter of the decay mode,  $e^+ e^- \rightarrow \mu^+ \mu^- A'$ . The signal process is dark photon which couples only to heavy leptons. We focus on the case in which dark photon decays into two muons. In this work, using MadGraph5 as a simulation tool kit, we have studied CPU time as well as cross section according to various parameters such as center of mass (CM) energy, dark photon ( $A'$ ) mass and coupling constant. To compare CPU time of calculation, we have used the KISTI-5 supercomputer and the local Linux machine. The CPU time and the wall clock time of the KISTI-5 supercomputer and the local Linux machine with one or multiple cores have been studied for the optimization and parallelization with large number of cores of KISTI-5 supercomputer. The results will help to optimize HEP software using high performance computing (HPC).

**Primary authors:** Mr PARK, Kihong (KISTI/UST); Prof. CHO, Kihyeon (KISTI/UST)

**Presenters:** Mr PARK, Kihong (KISTI/UST); Prof. CHO, Kihyeon (KISTI/UST)

**Session Classification:** Physics & Engineering Session

**Track Classification:** Physics (including HEP) and Engineering Applications

Contribution ID: 7

Type: **Oral Presentation**

## Russian National Data Lake Prototype

*Wednesday, 24 March 2021 13:20 (20 minutes)*

The evolution of the computing facilities and the way storage will be organized and consolidated will play a key role in how this possible shortage of resources will be addressed by the LHC experiments. The need for an effective distributed data storage has been identified as fundamental from the beginning of LHC, and this topic has become particularly vital in the light of the preparation for the HL-LHC run. WLCG has started an R&D within DOMA project and in this contribution we will report the recent results related to the Russian federated data storage systems configuration and testing. We will describe different system configurations and various approaches to test data storage federation. We are considering EOS and dCache storage systems as a backbone software for data federation and xCache for data caching. We'll also report about synthetic tests and experiments specific tests developed by ATLAS and ALICE for federated storage prototype in Russia. Data Lake project has been launched in Russian Federation in 2019 to set up a National Data Lake prototype for HENP and to consolidate geographically distributed data storage systems connected by fast network with low latency, we will report the project objectives and status.

**Primary author:** Mr KIRYANOV, Andrey (NRC "Kurchatov Institute")

**Presenter:** Mr KIRYANOV, Andrey (NRC "Kurchatov Institute")

**Session Classification:** Data Management & Big Data Session

**Track Classification:** Data Management & Big Data

Contribution ID: 8

Type: **Oral Presentation**

## Scientific data management at HEPS

*Wednesday, 24 March 2021 13:00 (20 minutes)*

The High Energy Photon Source (HEPS) in Beijing is the first national light source of high-energy synchrotron radiation in China, and will be one of the world's brightest fourth-generation synchrotron radiation facilities.

Doubtless data are of crucial importance for the scientific discoveries made in the experiments at HEPS. According to the estimated data rates, we predict 30 PB raw experimental data will be produced per month from 14 beamlines at the first stage of HEPS, and the data volume will be even greater after over 90 beamlines are available at the second stage in the near future.

Therefore, successful data management is critically important for the present and future scientific productivity of HEPS. The data management system is responsible for automating the organization, transfer, storage, distribution of the data collected at HEPS. First of all, the main features of the scientific data acquired from all the beamlines and the possible problems exists in data managing and data sharing is explained in this paper. Second, the architecture and data flow of the HEPS data management system are described from the perspective of facility users and IT. Furthermore, key techniques implemented in this system are introduced. Finally, the progress and the effect of the data management system deployed as a pilot test at BSRF are given.

**Primary author:** Ms HU, Hao (Institute of High Energy Physics)

**Presenter:** Ms HU, Hao (Institute of High Energy Physics)

**Session Classification:** Data Management & Big Data Session

**Track Classification:** Data Management & Big Data

Contribution ID: 9

Type: **Oral Presentation**

## A possible solution for HEP processing on network secluded Computing Nodes

*Thursday, 25 March 2021 15:30 (30 minutes)*

The computing needs of LHC experiments in the next decades (the so-called High Luminosity LHC) are expected to increase substantially, due to the concurrent increases in the accelerator luminosity, in the selection rates and in the detectors' complexity. Many Funding Agencies are aiming to a consolidation of the national LHC computing infrastructures, via a merge with other large scale computing activities, such as HPC and Cloud centers. The LHC Experiments have started long ago tests and production activities on such centers, with intermittent success. The biggest obstacle with some centers comes from stricter network policies with respect to our standard centers, which do not allow an easy merge with the distributed LHC computing infrastructure. A possible solution for such centers is presented here, able to satisfy three main goals: be user deployable, be a catch-all solution for all the protocols and services, and be transparent to the experiment software stack. It is based on the integration of existing tools like tsocks, tunsocks, openconnect, cvmfsexec and singularity. We present results from an early experimentation, which positively show how the solution is indeed usable. Large scale testing on thousands of nodes is the next step in our agenda.

**Primary authors:** Mr MARIOTTI, Mirko (Department of Physics and Geology, University of Perugia); Dr SPIGA, daniele (INFN-PG); Dr BOCCALI, tommaso (INFN)

**Presenter:** Mr MARIOTTI, Mirko (Department of Physics and Geology, University of Perugia)

**Session Classification:** Converging High Performance infrastructures: Supercomputers, clouds, accelerators Session

**Track Classification:** Converging High Performance infrastructures: Supercomputers, clouds, accelerators

Contribution ID: 10

Type: **Oral Presentation**

## Reinforcement Learning for Smart Caching in the CMS experiment

*Tuesday, 23 March 2021 14:00 (20 minutes)*

In the near future, High Energy Physics experiments' storage and computing needs will go far above what can be achieved by only scaling current computing models or current infrastructures. Considering the LHC case, for 10 years a federated infrastructure (Worldwide LHC Computing Grid, WLCG) has been successfully developed. Nevertheless, the High Luminosity (HL-LHC) scenario is forcing the WLCG community to dig for innovative solutions. In this landscape, one of the initiatives is the exploitation of Data Lakes as a solution to improve the Data and Storage management. The current Data Lake model foresees data caching to play a central role as a technical solution to reduce the impact of latency and network load. Moreover, even higher efficiency can be achieved through a smart caching algorithm: this motivates the development of an AI-based approach to the caching problem.

In this work, a Reinforcement Learning-based cache model (named QCACHE) is applied in the CMS experiment context. More specifically, we focused our attention on the optimization of both cache performances and cache management costs.

The QCACHE system is based on two distinct Q-Learning (or Deep Q-Learning) agents seeking to find the best action to take given the current state. More explicitly, they try to learn a policy that maximizes the total reward (i.e. hit or miss occurring in a given time span). While the Addition Agent is taking care of all the cache writing requests, clearly the Eviction Agent deals with the decision to keep or to delete files in the cache.

We will present an overview of the QCACHE framework with results comparing standard replacement policies using "Real-world" data (i.e. historical data requests aggregation used to predict dataset popularity filtered for Italy and US regions). Moreover, we will show the planned subsequent evolutions of the framework.

**Primary authors:** TRACOLLI, Mirco (University of Perugia, INFN Perugia and University of Florence); TEDESCHI, Tommaso (University and INFN, Perugia (Italy))

**Co-authors:** CIANGOTTINI, Diego (INFN Perugia); STORCHI, Lorian (INFN Perugia); Dr SPIGA, daniele (INFN-PG)

**Presenter:** TEDESCHI, Tommaso (University and INFN, Perugia (Italy))

**Session Classification:** Data Management & Big Data Session

**Track Classification:** Data Management & Big Data

Contribution ID: 11

Type: **Oral Presentation**

## Proposing an Assessment Strategy in Data Science Approach

*Tuesday, 23 March 2021 13:00 (20 minutes)*

Academia has been promoting Academic Integrity. With the Pandemic, Academic Integrity has been more and more important due to online education. Students have more opportunities to write term papers or reports as evidence or artifact of their learning. It is now essential that students write papers guaranteed by originality. There is no question about it. On the other hand, papers that passed the rigorous process of Academic Integrity have been still evaluated by hand by a professor who is in charge of the course. In here, the evaluation has been solely based on the expertise and professional experience, i.e., by human hands, which is time-consuming and not reliable. Some innovative professors employ the rubric assessment approach to be fair with grading. And yet, the grading is by hand. In most Asian universities, a professor usually is in charge of several courses, each of which has more than several hundred students. It is almost impossible to monitor all students' progress while nurturing their learning throughout the course. This paper proposes an assessment strategy for actively engaged students in such large classrooms based on educational innovation. It is meant for the assessment strategy making use of Data Science in mind.

Based on the concept that the reflective writing by a learner is the mirror of the active learning mind, the reflective writing can be a valid and reliable corpus filled with active vocabulary in the mind of the learner. By having students opportunities to reflect on their learning at the milestones in the course, the corpus is accumulated with active vocabulary. By categorizing such active vocabulary making use of machine learning or deep learning approach with reference to Bloom's Taxonomy Matrix, is it possible to visualize in the form of a heatmap, not only the result but the progress of learning in the course. In this way, the learner can visually view his/her own progress in learning immediately.

While the Learning Analytics approach makes use of the activity log archived in the Learning Management System managed by the university, the proposed assessment approach makes use of the reflective writing by the actively engaged learner as the corpus. The main goal of the Learning Analytics approach is to identify a group of students who have a tendency to drop out or quit school and to offer academic advising before it is too late. On the other hand, the proposed approach is for all levels of actively engaged learners to view their own progress of learning as well as the result of learning at the end of the course.

The presentation includes demonstrations of the proposed assessment strategy, followed by the theoretical description as well as the rationale behind the proposal.

### Summary

Academia has been promoting Academic Integrity. With the Pandemic, Academic Integrity has been more and more important due to online education. Students have more opportunities to write term papers or reports as evidence or artifact of their learning. It is now essential that students write papers guaranteed by originality. There is no question about it. On the other hand, papers that passed the rigorous process of Academic Integrity have been still evaluated by hand by a professor who is in charge of the course. In here, the evaluation has been solely based on the expertise and professional experience, i.e., by human hands, which is time-consuming and not

reliable. Some innovative professors employ the rubric assessment approach to be fair with grading. And yet, the grading is by hand. In most Asian universities, a professor usually is in charge of several courses, each of which has more than several hundred students. It is almost impossible to monitor all students' progress while nurturing their learning throughout the course. This paper proposes an assessment strategy for actively engaged students in such large classrooms based on educational innovation. It is meant for the assessment strategy making use of Data Science in mind.

Based on the concept that the reflective writing by a learner is the mirror of the active learning mind, the reflective writing can be a valid and reliable corpus filled with active vocabulary in the mind of the learner. By having students opportunities to reflect on their learning at the milestones in the course, the corpus is accumulated with active vocabulary. By categorizing such active vocabulary making use of machine learning or deep learning approach with reference to Bloom's Taxonomy Matrix, is it possible to visualize in the form of a heatmap, not only the result but the progress of learning in the course. In this way, the learner can visually view his/her own progress in learning immediately.

While the Learning Analytics approach makes use of the activity log archived in the Learning Management System managed by the university, the proposed assessment approach makes use of the reflective writing by the actively engaged learner as the corpus. The main goal of the Learning Analytics approach is to identify a group of students who have a tendency to drop out or quit school and to offer academic advising before it is too late. On the other hand, the proposed approach is for all levels of actively engaged learners to view their own progress of learning as well as the result of learning at the end of the course.

The presentation includes demonstrations of the proposed assessment strategy, followed by the theoretical description as well as the rationale behind the proposal.

**Primary author:** Dr YAMAMOTO, Tosh (Kansai University, CTL)

**Co-author:** Mr IKEZAWA, Tomoya (Asia University)

**Presenters:** Mr IKEZAWA, Tomoya (Asia University); Dr YAMAMOTO, Tosh (Kansai University, CTL)

**Session Classification:** Humanities, Arts & Social Sciences Session

**Track Classification:** Humanities, Arts, and Social Sciences (HASS) Applications

Contribution ID: 12

Type: **Oral Presentation**

## **Distributed filesystems (GPFS, CephFS and Lustre-ZFS) deployment on Kubernetes/Docker clusters**

*Friday, 26 March 2021 12:00 (30 minutes)*

Nowadays Kubernetes has become a powerful tool to deploy and manage containerized applications.

Modern datacenters need distributed filesystems to provide user applications with access to stored data on a large number of nodes.

The possibility to mount a distributed filesystem and exploit its native application programming interfaces in a Docker container,

combined with the advanced orchestration features provided by Kubernetes, may enhance flexibility in data management and transfer services' installation, running and monitoring, allowing the execution of dedicated services on different nodes, in isolated and automatically replicable environments, this way improving deployment efficiency and fail-safeness.

The goal of this work is to demonstrate the feasibility of using Kubernetes and Docker to setup clients capable to access a distributed filesystem from existing clusters and to create clusters based on containerized servers.

Although this is just a proof of concept, the effort has shown the possibility of using different types of distributed filesystems (GPFS, CephFS, Lustre-ZFS) with equally positive results. Read/write performances with these filesystem setups have been tested and compared to each other.

### **Summary**

The goal of this work is to demonstrate the feasibility of using Kubernetes and Docker to setup clients capable to access a distributed filesystem from existing clusters and to create clusters based on containerized servers.

**Primary author:** Dr FORNARI, Federico (INFN-CNAF)

**Presenter:** Dr FORNARI, Federico (INFN-CNAF)

**Session Classification:** Infrastructure Clouds and Virtualisation Session

**Track Classification:** Infrastructure Clouds and Virtualisation

Contribution ID: 13

Type: **Oral Presentation**

## Expanding the climateprediction.net project: the creation of a multi model distributed computing infrastructure for climate science

*Thursday, 25 March 2021 13:30 (30 minutes)*

climateprediction.net (CPDN) is a citizen science project that uses the paradigm of volunteer distributed computing via the Berkeley Open Infrastructure for Network Computing (BOINC) software platform. This platform enables the computation of very large ensembles of climate simulations using the spare resources of the machines of volunteering members of the public. The CPDN project has been running, without interruption for over 15 years. It is a unique facility in the climate science community, enabling studies to be performed involving extremely large ensembles of climate simulations at a low cost, answering questions on uncertainty, attribution and other key problems which would otherwise be much more difficult (or impossible in some cases) to study using traditional HPC techniques.

To date the CPDN project has exclusively used models derived from the UK Meteorological Office Hadley Centre Coupled Model version 3 (HadCM3). In this work we describe the expansion of the CPDN project to allow computations to be performed using the OpenIFS and WRF models. OpenIFS is a portable research version of the European Center for Medium Range Weather Forecasts (ECMWF) developed Integrated Forecasting System (IFS), their main tool for operational forecasting. WRF (Weather Research and Forecast) model is a numerical weather prediction system produced by a collaboration of NCAR, NOAA and AFWA. Both OpenIFS and WRF have large communities of engaged scientific users. This expansion of the CPDN project capabilities expands the possible user to a new community of climate scientists, enabling a range of new studies to be performed for the first time involving very large ensembles these newly supported models. As an added benefit this will also allow inter-model comparisons to be performed between OpenIFS, WRF and HadCM3 based models.

We describe the technical challenges and design considerations necessary to enable the inclusion of the OpenIFS and WRF models into the CPDN project. We also describe the workflow and the successful implementation on both the server-side and client-side of the project, together with preliminary results that have been performed to scientifically validate these new models in the CPDN environment.

### Summary

A presentation of the technical challenges and design considerations necessary to enable the inclusion of the OpenIFS and WRF models into the climateprediction.net (CPDN) volunteer distributed computing project, including the workflow and the successful implementation on both the server-side and client-side of the project, together with preliminary results that have been performed to scientifically validate these new models in the CPDN environment.

**Primary author:** Mr BOWERY, Andy (University of Oxford)

**Co-authors:** Dr WEISHEIMER, Antje (Oxford University); Prof. WALLOM, David (University of Oxford); Dr LIN, Dongqi (Canterbury University, Christchurch, New Zealand); Dr CARVER, Glenn (ECMWF); Dr KOEHLER, Marcus (ECMWF); Dr KATURJI, Marwan (University of Canterbury, Christchurch,

New Zealand); Dr SPARROW, Sarah (Oxford University); Dr ROSIER, Suzanne (NIWA, New Zealand)

**Presenter:** Mr BOWERY, Andy (University of Oxford)

**Session Classification:** Infrastructure Clouds and Virtualisation Session

**Track Classification:** Infrastructure Clouds and Virtualisation

Contribution ID: 14

Type: **Oral Presentation**

## unroutable LHCONE traffic

*Friday, 26 March 2021 10:00 (30 minutes)*

This talk explores the methods and results confirming the baseline assumption that LHCONE traffic is science traffic. The LHCONE (LHC Open Network Environment) is a network conceived to support globally distributed collaborative science. The LHCONE connects thousands of researchers to LHC data sets at hundreds of universities and labs performing analysis within the global collaboration. It is “Open” to all levels of the LHC as well as a short list of approved non-LHC science collaborations. It is distinct from the smaller, tightly integrated and private LHCOPN (Optical Private Network) network which is strictly for “Tier 1” compute centers and used in support of the engineered workflow for LHC data processing, distribution and storage of the baseline datasets. LHCONE satisfies the need for a high performance global data transfer network of networks supporting scientific analysis at universities and science labs.

**Science traffic separation is the hard part**

The separation of science flows from non-science flows, an essential first step in traffic engineering high performance science networks. Before resources or preference can be applied to more effectively move science data, it is essential to identify and separate the science from non-science traffic. This talk explores the methods and results in detecting traffic in the LHCONE network that does not comply with the Appropriate Use Policy established by the global LHC collaboration.

**LHCONE hosts are high performance**

Through integration of the Science DMZ network model and collaborative software platforms. The data transfer nodes connected to LHCONE are high performing data movers placed on the network edge/Science DMZ and secured precisely according to the applications they support and the purpose they serve.

**LHCONE is at risk of unauthorized use**

Unauthorized use of LHCONE places both the network and the sites using it at risk. The risk takes two forms:- Science flows mixing with non-science flows- Unauthorized traffic being dropped inside LHCONE

**Identifying unauthorized traffic**

An EDUgain authenticated portal displaying unauthorized usage will be demonstrated. Since LHCONE is growing and changing quite frequently the underlying database will get collaborative maintained and administered.

**Primary author:** Mr HOEFT, Bruno (Karlsruhe Institute of Technology)

**Presenter:** Mr HOEFT, Bruno (Karlsruhe Institute of Technology)

**Session Classification:** Network, Security, Infrastructure & Operations Session

**Track Classification:** Network, Security, Infrastructure & Operations

Contribution ID: 15

Type: **Oral Presentation**

# HPC-Cloud-Big Data Convergent Architectures and Research Data Management: The LEXIS Approach

*Thursday, 25 March 2021 16:00 (30 minutes)*

The LEXIS project (Large-scale EXecution for Industry & Society, H2020 GA825532) provides a platform for optimized execution of Cloud-HPC workflows, reducing computation time and increasing energy efficiency. The system will rely on advanced, distributed orchestration solutions (Bull Ystia Orchestrator, based on TOSCA and Alien4Cloud technologies), the High-End Application Execution Middleware HEAppE, and new hardware capabilities for maximizing efficiency in data processing, analysis and transfer (e.g. Burst Buffers with GPU- and FPGA-based data reprocessing).

LEXIS handles computation tasks and data from three Pilots, based on representative and demanding HPC/Cloud-Computing use cases in Industry (SMEs) and Science: i) Simulations of complex turbomachinery and gearbox systems in Aeronautics, ii) Tsunami simulations and earthquake loss assessments which are time-constrained to enable immediate warnings and to support well-informed decisions, and iii) Weather and Climate simulations where massive amounts of in-situ data are assimilated to improve forecasts. A user-friendly LEXIS web portal, as a unique entry point, will provide access to data as well as workflow-handling and remote visualization functionality.

As part of its back-end, LEXIS builds an elaborate system for the handling of input, intermediate and result data. At its core, a Distributed Data Infrastructure (DDI) ensures the availability of LEXIS data at all participating HPC sites, which will be federated with a common LEXIS AAI (with unified security model, user database and authorization policy). The DDI leverages best of breed data-management solutions from EUDAT, such as B2SAFE (based on iRODS) and B2HANDLE. REST APIs on top of it will ensure a smooth interaction with LEXIS workflows and the orchestration layer. Last, but not least, the DDI will provide functionalities for Research Data Management following the FAIR principles ("Findable, Interoperable, Accessible, Reusable"), e.g. DOI acquisition, which helps to publish and disseminate open data products.

## Summary

This talk presents the LEXIS ("Large-Scale Execution for Industry and Society") platform for running orchestrated compute-intensive workflows in Science, Engineering, Industry, Governmental Applications and beyond. These workflows handle Big Data with a hybrid, geographically distributed Cloud/HPC computing environment and an EUDAT/iRODS-based Distributed Data Infrastructure (DDI). Particular emphasis in the presentation is put on the DDI architecture and its integration with European (EUDAT/EOSC) research data management. Currently, the DDI connects two large Supercomputing Centres (IT4I/CZ and LRZ/DE), with additional significant European computing and data sites being added to the data federation.

**Primary author:** Dr HACHINGER, Stephan (Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities, Garching, Germany)

**Co-authors:** Dr SCIONTI, Alberto (Advanced Computing and Applications, LINKS Foundation, Torino, Italy); Dr PARODI, Antonio (CIMA Research Foundation, Savona, Italy); Mr D'AMICO, Carmine

(Advanced Computing and Applications, LINKS Foundation, Torino, Italy); Mr KOCH-HOFER, Cédric (ATOS, Paris, France); Prof. SCHORLEMMER, Danijel (GFZ German Research Centre for Geosciences, Potsdam, Germany); Dr MAGARIELLI, Donato (AvioAero, Torino, Italy); Mr APOPEI, Florin Ionut (TESEO, Torino, Italy); Dr MARTINOVIČ, Jan (IT4Innovations, VŠB - Technical University of Ostrava, Ostrava-Poruba, Czech Republic); Dr LEVRIER, Marc (ATOS, Paris, France); Mr GOLASOWSKI, Martin (IT4Innovations, VŠB - Technical University of Ostrava, Ostrava-Poruba, Czech Republic); Mr HAYEK, Mohamad (Leibniz Supercomputing Centre (LRZ), Bavarian Academy of Sciences and Humanities, Garching, Germany); Dr TERZO, Olivier (Advanced Computing and Applications, LINKS Foundation, Torino, Italy); Dr HARSH, Piyush (Cyclops Labs, Zurich, Switzerland); Dr MURPHY, Seán (School of Engineering, Zürcher Hochschule für Angewandte Wissenschaften (ZHAW), Zurich, Switzerland); Dr CICCIA, Simone (Advanced Computing and Applications, LINKS Foundation, Torino, Italy); Dr GOUBIER, Thierry (CEA, LIST, Paris, France)

**Presenter:** Dr HACHINGER, Stephan (Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities, Garching, Germany)

**Session Classification:** Converging High Performance infrastructures: Supercomputers, clouds, accelerators Session

**Track Classification:** Converging High Performance infrastructures: Supercomputers, clouds, accelerators

Contribution ID: 17

Type: **Oral Presentation**

# Performance and Cost Evaluation of Public Cloud Cold Storage Services for Astronomy Data Archive and Analysis

*Friday, 26 March 2021 11:00 (30 minutes)*

Currently major cloud providers provide cold storage services as a part of their public IaaS offerings, targeting users who need to store data with relatively low access frequency for long periods. The adoption of cold storage services should be considered in order to reduce the total cost of ownership and the labor of storage management of maintaining large amounts of scientific research data over a long period of time. However, performance and cost of public cold storage services in scientific applications have not been well studied, and the following issues arise:

<ol> <li>It is difficult to determine whether cold storage services meet the performance requirements of research applications.</li> <li>It is also difficult to assess the feasibility of storing and accessing research data in cold storage services in terms of cost.</li> </ol>

In order to address the issues mentioned above and to validate feasibility of adopting cold storage services in the astronomical research area, we present evaluation of cloud cold storage services using astronomical research data and applications. We stored the observation and analysis data of the ALMA radio telescope project<sup>1</sup> in S3 Infrequent Access and Glacier provided by Amazon Web Services (AWS) and we ported the data archive software used in the ALMA project, Next Generation Archive System (NGAS), to AWS.

To solve the first issue, we measured the performance of data retrieval operations of NGAS on AWS. In addition, we conducted performance benchmark tests such as uploading up to 60TB ALMA data. We also conducted the same benchmark tests on other commercially available cold storage services such as Google, Azure, and Oracle to validate the performance requirements can be generally fulfilled.

To solve the second issue, we proposed a cost estimation model of NGAS based on the AWS payment required for storing and retrieving data and estimated the yearly expense of the NGAS on AWS by using the actual values of data amounts and the accesses frequency statistics. Our estimation shows that retrieving data from a cold storage service and analyzing the data outside of the cloud (e.g. an on-premise system) increase the cost because data transfer cost outward the cloud is significantly high. We designed the architecture to analyze the retrieved data inside cloud and estimated cost for running common analysis applications, Common Astronomy Software Applications package (CASA)<sup>2</sup>, with NGAS on a variety of instances of AWS.

From those experiments, the following findings are obtained:

<ol> <li>We can obtain practically acceptable performance of data access in the cold storage services by configuring the archive system with appropriate sizing of instances and tuning the system. Although some cold storage services require hours to start data access, this disadvantage can be mitigated by adopting an appropriate tiered storage architecture.</li> <li>The proposed cost estimation model enables to estimate total cost of data archive in the cloud cold storage services and data analysis on the cloud services. The model is also capable to estimate cost on a hybrid system organized by clouds and on-premise systems. Additionally, the practical information which can be used to determine the optimal configuration of the analysis system such as sizing information of AWS instances are acquired.</li> </ol>

<sup>1</sup> <https://www.nao.ac.jp/en/research/project/alma.html>

<sup>2</sup> <https://casa.nrao.edu/>

**Primary author:** Mr YOSHIDA, Hiroshi (National Institute of Informatics)

**Co-authors:** Dr YOHEI, Hayashi (National Astronomical Observatory of Japan); Prof. AIDA, Kento (National Institute of Informatics); Prof. GEORGE, Kosugi (National Astronomical Observatory of Japan); Dr RENAUD, Miel (National Astronomical Observatory of Japan); Dr EISUKE, Morita (National Astronomical Observatory of Japan); Dr TAKESHI, Nakazato (National Astronomical Observatory of Japan)

**Presenter:** Mr YOSHIDA, Hiroshi (National Institute of Informatics)

**Session Classification:** Infrastructure Clouds and Virtualisation Session

**Track Classification:** Infrastructure Clouds and Virtualisation

Contribution ID: 18

Type: **Oral Presentation**

## Architecture of Job Scheduling Simulator for Demand Response Based Resource Provision

*Friday, 26 March 2021 13:30 (30 minutes)*

In High Performance Computing (HPC) center that operates large-scale cluster system for providing computing resources for fee-paying users, it is important to reduce the power consumption. In order to save power of large-scale cluster system, the administrator of the HPC center stops a part of the computing nodes or changes them in a power-saving state with degradation of execution performance. These operations are collectively called fallback. Since the fallback generally causes the decrease of computing resources and degradation of execution performance of computing nodes, it makes the job waiting time and the job execution time increase. However, many existing resource provision service do not consider which jobs are affected by fallback.

In this situation, a new resource provision service called Demand Response (DR) based resource provision has been attracting attention. In DR based resource provision, a user declares whether an influence caused by fallback is allowed or not in a job. Job scheduler that allocates computing resources to jobs controls the execution of the job so that undeclared jobs are not affected by fallback as possible. Incentive are established to users for encouraging to accept the declaration in DR based resource provision. How to design incentive for users is important in order to make use of DR based resource provision in actual operations.

For appropriately designing an incentive, the energy consumption of computing nodes, the increase of the job waiting time and the job execution time are essential as index. These information enables to measure the effect of fallback. However, there is no mechanism to present such index according to the applied fallback methods in the various structures of computing nodes and submitted jobsets. Although a job scheduling simulator is widely used as a mechanism to evaluate how submitted jobs are processed on cluster system, most job scheduling simulators don't have functionalities to provides the amount of energy consumption of computing nodes and the increase of the job waiting time and the job execution time in DR based resource provision. Therefore, we aims to realize a new job scheduling simulator which enables to provide the amount of power consumption of computing nodes, the job waiting time and the job execution time in DR based resource provision.

In this study, we propose a job scheduling simulator with the functionality for managing jobs in DR based resource provision manner and outputting indexes to design incentive. For deriving the required functionalities, a process flow of DR resource provision is analyzed by comparing with a traditional job processing flow. Based on the analysis, we design and develop DR based resource provision execution module composed of three functions. The proposed job scheduling simulator is implemented by linking DR resource provision execution module to Simulus, an existing event-driven simulator. In the evaluation, we conducted experiments to observe the behavior of the proposed job scheduling simulator under several conditions.

Keywords: Job Scheduling Simulator, Demand Response (DR) Based Resource Provision

**Primary author:** MATSUI, shogo (Graduate School of Information Science and Technology, Osaka University, Japan)

**Co-authors:** Prof. LIU, Jason (School of Computing and Information Sciences, Florida International University, USA); Prof. HARUMOTO, Kaname (Institute for Dataability Science, Osaka University, Japan); Prof. SHIMOJO, Shinji (Cybermedia Center, Osaka University, Japan); Dr DATE, Susumu

(Cybermedia Center, Osaka University, Japan); Dr WATASHIBA, Yasuhiro (Cybermedia Center, Osaka University, Japan)

**Presenter:** MATSUI, shogo (Graduate School of Information Science and Technology, Osaka University, Japan)

**Session Classification:** Converging High Performance infrastructures: Supercomputers, clouds, accelerators Session

**Track Classification:** Converging High Performance infrastructures: Supercomputers, clouds, accelerators

Contribution ID: 20

Type: **Oral Presentation**

## **IRIS Identity and Access Manager: UK Experiences with token based AAI**

*Thursday, 25 March 2021 11:10 (20 minutes)*

Driven by the physics communities supported by UKRI-STFC (UK Research and Innovation Science and Technology Facilities Council) the eInfrastructure for Research and Innovation for STFC, or IRIS, is a collaboration of UKRI STFC, science activities and provider entities in the UK. Over the last few years the UK's IRIS collaboration and IRIS 4x4 project (£16m as £4m p.a. for four years) has worked to deploy hardware and federating tools across the range of physics supported by STFC.

The IRIS IAM (Identity and Access Manager) is IRIS' answer to federated access, providing a coherent framework for accessing a range of different computing resources. An identity proxy based on the AARC blueprint architecture and utilising the INDIGO IAM software (originally developed as part of INDIGO Data cloud), the IRIS IAM service provides federated access to IRIS resources and aims to remove friction for scientific communities and promises to facilitate a new generation of workflows across diverse resources. As well as acting as providing federated access to users affiliated with eduGAIN institutions, the IRIS IAM is also able to act as an IdP-of-last-resort for members of the IRIS community who do not have a home institution IdP. This is achieved through provisioning of locally managed IAM accounts, with new accounts vetted by a trusted community representative. The IRIS IAM has also been an important technical driver for ongoing parallel development of the IRIS Trust Framework security policy set.

The IRIS IAM also provides community attributes in the form of group memberships, which can then be used for authorization purposes. Delegation of management of these groups is possible, and this responsibility will usually be assigned to the administrators of the services behind the IAM or to trusted representatives of the science activities whose researchers need access. These group managers can then approve or deny membership request, which correspond to confirming access to controlled resources and services.

Development of the IRIS IAM service has been in parallel to other community Authentication and Authorization activities, such as FIM4R and the WLCG authorization project, in order to ensure that the IRIS solution aligns with and supports the work undertaken elsewhere. Three years since the project began, The IRIS IAM is now an established production service, providing access to a number of IRIS services, including OpenStack clouds, accounting dashboards and security portals and with work underway to connect both Rucio and Dynafed services, among others. Recent work has focused on enhancing the service's operational performance and features, including work investigating how best to utilise the OpenID Connect workflow over command line/ssh sessions. However, work is still underway to enhance the service's offering, including the range and scope of clients the IAM provides access too. This talk shall report on work to support web-based and command-line based services, progress thus far, notable challenges, and next steps and plans for the IRIS IAM service.

### **Summary**

Over the last few years the UK's IRIS collaboration and IRIS 4x4 project (£16m as £4m p.a. for four years) has worked to deploy hardware and federating tools across the range of physics supported by STFC. Providing a coherent framework for accessing HTC, HPC and Open Stack cloud resources, the IRIS IAM (Identity and Access Management) service provides federated access to resources, removing friction for scientific communities and facilitates a new generation of workflows across

diverse resources.

**Primary author:** Mr DACK, Tom (STFC UKRI)

**Co-author:** Mr COLLIER, Ian (STFC-RAL)

**Presenter:** Mr DACK, Tom (STFC UKRI)

**Session Classification:** Network, Security, Infrastructure & Operations Session

**Track Classification:** Network, Security, Infrastructure & Operations

Contribution ID: 21

Type: **Oral Presentation**

## Building an Open Source Security Operations Center at Nikhef

*Wednesday, 24 March 2021 13:30 (30 minutes)*

Centralizing (and visualizing) what is happening at a compute site enjoys ongoing effort from both industry and the community. An established example is include the ELK stack, importing, storing and visualizing logs, while more recently traffic analysis has gotten more and more attention - especially from WLCG.

At Nikhef, we have attempted to create a reasonable, recreatable SOC design based on output of the WLCG SOC WG. After experiments with an IBM POWER8 (S822L) system and various network setups, we landed on an AMD 64-core system that handles 100Gbit, potentially 200Gbit pending experimentation.

The result is a small cluster of Open Source software that receives between 40 and 80 million syslog messages per day, and around 600 and 750 million traffic logs from Zeek. Much of this cluster is set up using Ansible and is intended to be released to the community.

During this talk, we will discuss our SOC setup, and discuss some practical findings with regards to a high-performing setup.

**Primary author:** ROORDA, Jouke (Nikhef)

**Presenter:** ROORDA, Jouke (Nikhef)

**Session Classification:** Network, Security, Infrastructure & Operations Session

**Track Classification:** Network, Security, Infrastructure & Operations

Contribution ID: 22

Type: **Oral Presentation**

## IPv6-only networking for WLCG

*Friday, 26 March 2021 09:30 (30 minutes)*

The use of IPv6 on the general internet continues to grow. Several Broadband/Mobile-phone companies, such as T-Mobile in the USA and BT/EE in the UK, now use IPv6-only networking with connectivity to the IPv4 legacy world enabled by the use of NAT64/DNS64/464XLAT. Large companies, such as Facebook, use IPv6-only networking within their internal networks, there being good management and performance reasons for this.

The transition of WLCG central and storage services to dual-stack IPv4/IPv6 has gone well, thus enabling the use of IPv6-only CPU resources as agreed by the WLCG Management Board. The use of dual-stack services is, however, a complex environment not only to configure and manage but also for trouble-shooting during observation of network performance and operational problems. WLCG is therefore actively considering when and where it can move to the much simpler environment of IPv6-only networking.

The HEPiX IPv6 working group has been encouraging and supporting the WLCG transition to IPv6 over many years. We last reported on our work to an ISGC conference in 2015. During 2020, the HEPiX IPv6 working group has been chasing and supporting the remaining transition to dual-stack storage services. We have investigated and fixed the reasons for the use of IPv4 between two dual-stack endpoints when IPv6 should be preferred. We present this work and the tests that have been made of IPv6-only CPU showing the successful use of IPv6 protocols in accessing WLCG services.

The dual-stack deployment, as mentioned above, does however result in a networking environment which is much more complex than when using just IPv4 or just IPv6. Some services, e.g. the EOS storage system at CERN, are using IPv6-only for internal communication, where possible. The IPv6 working group has been investigating the drivers for using data transfers over IPv6 and the removal of the IPv4 protocol in more places. We will present the areas where this could be useful and possible and be even so bold as to suggest a potential timetable for the end of support for IPv4 within WLCG.

There are many lessons we learned along the way which should be of interest to other research communities who have not yet started their transition to IPv6. Even more importantly for new research communities just starting to plan their distributed IT Infrastructure, there is a clear message to consider the use of IPv6-only right from the start.

**Primary author:** Dr KELSEY, David (STFC-RAL)

**Presenter:** Dr KELSEY, David (STFC-RAL)

**Session Classification:** Network, Security, Infrastructure & Operations Session

**Track Classification:** Network, Security, Infrastructure & Operations

Contribution ID: 23

Type: **Oral Presentation**

## The WISE Community and updated baseline policies for collaborating research infrastructures

*Wednesday, 24 March 2021 13:00 (30 minutes)*

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 in recent years a successful collaboration between Infrastructures has helped improve the security and allowed all to work more efficiently.

For more than 5 years, “the WISE community” has enhanced best practice in information security for IT infrastructures for research. WISE fosters a collaborative community of security experts and builds trust between IT infrastructures, i.e. all the various types of distributed computing, data, and network infrastructures in use today for the benefit of research, including cyberinfrastructures, e-infrastructure and research infrastructures. Through membership of working groups and attendance at workshops these experts participate in the joint development of policy frameworks, guidelines, and templates.

With participants from e-Infrastructures such as EGI, EUDAT, GEANT, EOSC-hub, PRACE, XSEDE, HBP, OSG, NRENs and more, the actual work of WISE is performed in focussed working groups, each tackling different aspects of collaborative security and trust. Two virtual community meetings have been held during the year of COVID-19 restrictions, one in April and one in October 2020.

The WISE Security for Collaborating Infrastructures (SCI) working group, in collaboration with Trust and Security activities in EOSC-hub, the GEANT GN4-3 project and the UK IRIS Infrastructure, is working on some new baseline/template security policies. This builds on the Policy Development Kit, an output of the EU Horizon 2020 projects Authentication and Authorisation for Research Collaborations (AARC/AARC2) and is aiming to produce new WISE policy templates. SCI working group has also been producing guidance for assessment of an Infrastructure’s maturity against the SCI Trust Framework.

This talk will present a report on recent activities and the future plans of WISE, together with details of the work to produce new security baseline policy templates, guidelines and recommendations.

**Primary author:** Dr KELSEY, David (STFC-RAL)

**Presenter:** Dr KELSEY, David (STFC-RAL)

**Session Classification:** Network, Security, Infrastructure & Operations Session

**Track Classification:** Network, Security, Infrastructure & Operations

Contribution ID: 24

Type: **Oral Presentation**

## **ccPortal: KEK Computing Research Center online application portal**

*Wednesday, 24 March 2021 11:20 (25 minutes)*

The Computing Research Center at High Energy Accelerator Research Organization (KEK) provides various IT services, such as email, WiFi network, data analysis system, etc. A user submits paper application forms to apply those services. Moreover, some services require endorsement by a KEK staff. From the Computing Research Center point of view, we receive applications from many users every day. These forms are checked, processed, and handed out among the responsible persons to complete applications.

To improve these circumstances, we have developed an online application portal system named "ccPortal." This portal allows users to apply new services and modify account information online. The portal also supports the center staff's online workflow, checking validity, creating an account, DB registration, and sending a completion notice. The ccPortal service was rolled out to KEK staff from the middle of September and gradually opened to all others by January.

We will describe the portal service and share our experience of improving work efficiency in IT service operations.

**Primary author:** TAKASE, Wataru (KEK)

**Co-authors:** MURAKAMI, Koichi (KEK); KANEKO, Sari (KEK); OKADA, Shogo (KEK)

**Presenter:** TAKASE, Wataru (KEK)

**Session Classification:** Network, Security, Infrastructure & Operations Session

**Track Classification:** Network, Security, Infrastructure & Operations

Contribution ID: 25

Type: **Oral Presentation**

## Exploring the awareness of health and life promotion from the differences in the life styles of young people before and after the epidemic

*Tuesday, 23 March 2021 13:40 (20 minutes)*

During the outbreak of COVID-19, strict preventive measures have been taking in all countries, such as city lockdown that makes the working, production, school activities being suspended, as well as the flight restrictions in some areas, while people are encouraged to implementing home isolation to prevent the further spread of the epidemic. Therefore, this condition has affected people's daily activities habits, especially for the young generation who has more diversified lives style. A study showed that abnormal changes in daily life habits can harm human health. As a result, this study investigates the life patterns differences of young people before and during the pandemic. It also makes an analysis study of the changes in related life patterns that may affect the health factors. Nonetheless, this study hopes to explore how to raise young people's awareness of health and life promotion during epidemics.

In this study, in order to increase the interest of research and the participation of young people, a "Life Stripe Project" method created by Japanese designers: Kenai Yamada and Hiro Kobayashi are introduced in this research study, then based on it, we revealed 21 kinds of different pattern activities such as sleeping, eating, reading, shopping then specified by different corresponding colors. After that, the Variable stripes of thicknesses are used as the representative of each activity's length in a day that would analyzable in visual information. Put this design method as a research template, the life stripe activity questionnaire was derived through the online survey tool to 57 students in Mainland China as the object participants to measure their habits due to pre and during the pandemic time condition. The color life stripe provided the visualization of the changes in target participants' personal life pattern routine to be examined. Besides, the six factors of health-promotion are integrated into some daily routine, for instance: sports and leisure, nutrition, stress management, interpersonal relationship, health responsibility, spiritual growth, etc. which contains the 21 activities defined in the life stripe used as the comparison of life patterns data before and during the pandemic, together to understand the health risks of affection. In the end, a service design methodology is used to provide viable solutions for guiding the group of young people to tackle the various health risks. On the other hand, this research wants to design the plan, and Adjustment activities for the young generation in order to promote better health awareness of their life.

### Summary

Keyword: Epidemic, Life styles, Awareness of health and life promotion, Life Stripe, Difference analysis, Service design

**Primary author:** Mr PAN, Wei-Jie (National Taipei University of Technology)

**Co-author:** Prof. WANG, Sheng-Ming (National Taipei University of Technology)

**Presenter:** Mr PAN, Wei-Jie (National Taipei University of Technology)

**Session Classification:** Humanities, Arts & Social Sciences Session

**Track Classification:** Humanities, Arts, and Social Sciences (HASS) Applications

Contribution ID: 26

Type: **Oral Presentation**

## Concept of the Social Design in Public Space – case of Berlin and Taipei

*Tuesday, 23 March 2021 15:20 (20 minutes)*

Path, Edges, Districts, Nodes, and Landmark are the five elements of legibility that contribute to our ability to create, maintain, images of a place suggested by Lynch. Districts are the medium-to-large sections of the city conceived as having two-dimensional extents, which the observer mentally enters inside of and which are recognizable as having some common identifying character. Our research focuses on the medium-scale levels of the spatial patterns and urban individuality, which is the neighborhood or district. We figured out that people with the same preferences tend to congregate in a similar area of the city, as well as the stores, and this area will have the difference from others, from the buildings to the pedestrians' dressing style, all represent indicators of the area. The study aims to come up with an innovative design service blueprint with these patterns in the areas of the cities, let the people who are unfamiliar with the areas will have a better sense of what the city is made of, and maybe they will be able to understand the city more quickly and deeply.

With observation diary and data collection, we can use the information we found to build a local network to visualize the regional differences observed. Our research is mainly separate into four parts, (1) Field research, (2) Thematic cartography Analysis, (3) Concept design, and (4) Evaluation. The observation method has been used in the very first part, observe the area on the map and create aesthetic structures, visualize the street network as well, try to find out underlying principles and patterns. Next, through the method "Mental Map Interview" to capture the human perception of the environment and make the cognitive(mental) mapping which is the mental construct of the environment seen from multiple viewpoints, try to find out the characteristic of each area. Mental Mapping is the process by which an individual acquires, codes, stores, and recalls information about the relative locations and attributes of phenomena in his everyday spatial environment. After that, combined the research and the result of analysis to come up with an innovative design service blueprint which is the concept of social design. Lastly, use the Kano Model for evaluation, because it's the future design concept, through this method, we can know which pattern is more important for the people.

This research is currently in the process, but we expect the result can be the pilot study of this kind of research.

Keywords: Service Design, Social Design, Mental Mapping, Image of The City, Kano Model

**Primary authors:** Dr WANG, Ryan Sheng-Ming (Department of Interaction Design, National Taipei University of Technology); Ms WANG, Shiau-Ting (Department of Interaction Design, National Taipei University of Technology)

**Presenter:** Ms WANG, Shiau-Ting (Department of Interaction Design, National Taipei University of Technology)

**Session Classification:** Humanities, Arts & Social Sciences Session

**Track Classification:** Humanities, Arts, and Social Sciences (HASS) Applications

Contribution ID: 28

Type: **Oral Presentation**

## Serious Game Design For Playful Exploratory Urban Simulation

*Tuesday, 23 March 2021 14:00 (20 minutes)*

The year 2020 strikes with the ferocious COVID-19, regulation, and restriction are active to mitigate the spread but simultaneously causing massive behavioral change globally. This year, the lack of tourists has caused critical damage to tourism, and people lose an entertainment activity although it is temporary. New habit formation after the pandemic requires investigation and simulation, so this research proposes using Minecraft and its gameplay to design a playful simulation to explore Dihua Street, a tourist hotspot in Taipei, Taiwan. This paper's main objective is to deliver to serious game developers and Minecraft players an approach to have an in-depth analysis of the holistic features that may cause behavior change and the influence of an exploratory game-based simulation. A serious game's primary benefit is the learning or training opportunity that has extended humanity's spaces to work, play, and learn. In particular, the virtual environment can simulate the real world and its scenarios to provide a learning platform or a playground for the citizens to explore and learn through experimentation. More than just a learning platform, this game-based solution also engages the learners to participate in urban planning tasks and thus increase public space awareness by researching, recreating, and reconstructing the historical old Dihua Street and its cultural background. During the pandemic time, the human-computer relationship becomes more reliable, and the foundation for such a relationship is the human-computer interaction. Over the years, simulation is well-known for its cost-efficient, risk-free, and accurate in achieving results. However, the game may inevitably suspend the impact to maintain the motivation to engage in these virtual environments without any interaction. To overcome such limitations, users need to interact through play, and a playful simulation can expand its boundary to attain desirable outcomes and discover the urban areas' concerning information. Therefore, this research proposed three tasks to engage the users with the urban by involving them in building, role-playing, and exploring to increase intrinsic motivation and engagement of a platform. This paper applies Don Norman's 3-Level of Emotional Design before applying Jobs To Be Done to examine the three tasks' critical elements thoroughly. After that, this game-based solution continues with the Kano Model analysis to verify the users' satisfaction with the game features offering within this 3D virtual world. This paper contributes to humanity's good health and well-being, together with sustainable cities and community development, to serve the worst scenarios and prepare action.

**Primary author:** Mr VU HONG, Lan (National Taipei University of Technology)

**Co-author:** Prof. WANG, Sheng-ming (National Taipei University of Technology)

**Presenter:** Mr VU HONG, Lan (National Taipei University of Technology)

**Session Classification:** Humanities, Arts & Social Sciences Session

Contribution ID: 29

Type: **Oral Presentation**

## Feasibility study on MPTCP ACK via alternative path in real network environment

*Friday, 26 March 2021 09:00 (30 minutes)*

Improving network performance is a key for grid computing system consisting of computational resources and data sources. There have been studies about network performance between them such as file transfer schedulers and network protocols. MPTCP is one of network protocols which has the potential to improve network performance. This protocol treats one TCP flow as one subflow and handle two or more TCP flows as one MPTCP flow. If there are multiple network paths between machines, it can be expected high network performance by MPTCP. These days, cloud computing services such as Azure, AWS and GCP are popular for deploying system instead of adopting on-premises system. These providers have been deployed all over the world, and they have many computational resources and data storages in various regions. Therefore, it can be also expected to compose a high-performance computer system by clustering them using multiple network paths in various countries.

Morikoshi et al. proposed MPTCP with HayACK for the purpose of improving data transfer throughput when multiple network paths are available, whose RTT are quite different. In original MPTCP, the sender transmits a data packet via the path and the receiver transmits ACK packet via same path. In MPTCP with HayACK, the sender can select appropriate path, regardless of which path the data packets came through. In that study, they simulated HayACK comparing original MPTCP. As a result, HayACK improved MPTCP throughput when RTTs of the paths are quite different.

However, there are concerns about using HayACK in real network environment. According to a study by Honda et al. in 2011, it was confirmed that 33% of packets having inconsistent sequence number and ACK number were discarded by middlebox which exists between a source and a destination. Middlebox include various devices such as NAT (Network Address Translators) router and FW(Firewall). Middlebox not only relays packets, but also changes or discards the contents of packets for the purposes of improving performance, changing destination, or hardening security. When MPTCP with HayACK using an alternative subflow, HayACK returns MPTCP ACK without modification from receiver. Therefore, HayACK packets may be discarded by middlebox in the real network environment.

The purpose of this study is to support the usefulness of HayACK for high-performance network which has multiple paths. Particularly, we focus on how middlebox affect HayACK packets.

Our investigation conducted comprehensively via various network path by preparing many IP address clients and one server. The influence from the middlebox is decided by checking probe packets passing through, changed or discarded. Assuming HayACK response, we crafted some packets such as TCP SYN packets and TCP ACK packets.

**Primary author:** Mr KOIBUCHI, Hiroyuki (Graduate school of Systems and Information Engineering University of Tsukuba, Japan)

**Co-authors:** Dr ABE, Hirotake (University of Tsukuba); Dr KATO, Kazuhiko (University of Tsukuba)

**Presenter:** Dr ABE, Hirotake (University of Tsukuba)

**Session Classification:** Network, Security, Infrastructure & Operations Session

**Track Classification:** Network, Security, Infrastructure & Operations

Contribution ID: 30

Type: **Oral Presentation**

## Application of OMAT in HTCONDOR resource management

*Wednesday, 24 March 2021 10:30 (25 minutes)*

At IHEP Computing Center, there are thousands of nodes managed by the htcondor scheduler, with about 12,000 cores, and these nodes provide computing services for multiple experimental groups. In the process of job scheduling, some work nodes will cause jobs abnormal due to some service exception. Under the traditional scheduling method, these abnormal nodes will continue to devour jobs, like “black holes”, resulting in a large number of job errors, affecting the service quality of the computing cluster. In addition, in the process of job scheduling, in order to quickly locate the unknown abnormal information in the job or the operating environment, we often isolate part of the work nodes for the specific experimental group. How to quickly isolate the node and record the change history is also an urgent problem to be solved. OMAT is short for open operation analysis toolkits, it is applied to cluster computing center in 2017 operational monitoring, providing rapid acquisition of abnormal data, correlation analysis, strategy alarm and visualization, etc. In this report, we will introduce how to use OMAT monitoring IHEP computing cluster, and feedback the node service status, help htcondor scheduler rapid convenient management of computing resources, and thus to minimize the effects of abnormal service for user operation, improve computing cluster service quality.

**Primary author:** Mr HU, Qingbao (IHEP)

**Co-authors:** Dr SHI, Jingyan (IHEP); Mr ZHENG, Wei (Institute of High Energy Physics, CAS); Mr JIANG, Xiaowei (Institute of High Energy Physics, Chinese Academy of Sciences)

**Presenter:** Mr HU, Qingbao (IHEP)

**Session Classification:** Network, Security, Infrastructure & Operations Session

**Track Classification:** Network, Security, Infrastructure & Operations

Contribution ID: 31

Type: **Oral Presentation**

## **An anomaly detection model based on DNS log data in IHEP**

*Thursday, 25 March 2021 10:30 (20 minutes)*

The Institute of high energy physics is operating and launching many large science facilities in China, such as BEPCII in Beijing, CSNS in Guangdong and JUNO in Shenzhen. These large science facilities are facing many network security threats. How to detect and prevent these threats is becoming important.

The domain name system is the cornerstone of Internet services, and most network applications rely on it. Malicious software, taking botnets as an example, usually uses DNS to complete communication with C&C servers and carry out illegal actions. The mainstream detection work mainly relies on the way of finding failures or matching signatures, and some work is based on wide-area network traffic, which aims to find active botnets and other malicious codes, and lacks attention to the terminal. We propose an anomaly detection model based on DNS log data. By extracting DNS fingerprints (a set of data describing the DNS behavior of the host per hour) from the log data, the IP address is the core, and the anomaly detection algorithm is used to automatically identify the abnormal terminal that is completely different from the normal terminal behavior. In addition, in order to prevent some functional servers, such as mail servers, from being incorrectly identified, the whitelist function is designed and supported.

**Primary author:** Mr LIANG, Zhongtian (Institute of High Energy Physics)

**Co-authors:** Mr QI, Fazhi (Institute of High Energy Physics,CAS); Dr YAN, Tian (IHEP)

**Presenter:** Mr LIANG, Zhongtian (Institute of High Energy Physics)

**Session Classification:** Network, Security, Infrastructure & Operations Session

**Track Classification:** Network, Security, Infrastructure & Operations

Contribution ID: 32

Type: **Oral Presentation**

## The Integration of Deep Learning and 360-Degree Video in Capturing Spatial Pedestrian Movements

*Tuesday, 23 March 2021 15:00 (20 minutes)*

People are the main users of cities, the development of cities is mainly to meet people's service needs and create different user experiences. Therefore, understanding users' behaviors, movements, directions, observation and data collection of different aggregation modes in urban space has always been an important issue in developing smart cities in the future. Under the background of the development of science and technology, people's observation in cities has been introduced into scientific and technological applications from man-made on-the-spot investigation, for example, from monitoring systems or aerial drone images, through deep learning, the interaction between people and space can be detected and interpreted. Or through the method of machine learning, from the collection of a small amount of local data, deduce the experience movement of a large number of people in space. Although the application of the above-mentioned related technologies has gradually matured, it is only in certain specific fields, such as indoor space and low forest; Or the shooting characteristics of the applied image, there are many limitations and undetectable characteristics. Therefore, this research study proposes to use 360 cameras to capture images containing the interaction between human and space in different spaces, integrate pre-processing, image optimization processing, and deep learning of image recognition, trying to analyze the position of human in space, basic data reading, moving line and directional interpretation, and aggregation mode. It is hoped that through the collection and analysis of a large amount of data from 360 landscape images, the application system of near-timely semi-automatic recognition of user's characteristics and behavior patterns in future cognitive space can be developed.

In this research study, the process is divided into three parts, namely, the shooting and processing of the 360-degree landscape video, the identification and path tracking of the user characteristics in the video, the output and analysis and interpretation of the user path. In the part of image data collection, this research study will use a camera with 360 panoramic view. In the aspect of user detection, YOLOv4 services are integrated as detection algorithms. In addition, in order to allow the identified pedestrians to add the same label for analysis, Deepsort algorithm is also used to track the behavior of specific people by this labeling method. The coordinate results obtained from the 360-degree video are converted into the top view and the moving track of pedestrians is superimposed on the plane map of the research field to judge the relationship between users and space in the field.

The final output will include pedestrian path records, pedestrian aggregation patterns and space use probability. These data can be applied to the simulation of pedestrian behavior patterns in space, so as to predict the behavior of some special events. Combined with the output data and analysis, we can predict the use state closer to the space, and then improve the reliability and availability of the prediction. Finally, a clear and standardized data format will be formulated for coordinate output, future coordinate data can be extended to more existing spatial analysis systems.

**Primary author:** Mr LIN, Cheng-Yen (National Taipei University of Technology)

**Co-author:** Prof. WANG, Sheng-Ming (National Taipei University of Technology)

**Presenter:** Mr LIN, Cheng-Yen (National Taipei University of Technology)

**Session Classification:** Humanities, Arts & Social Sciences Session

**Track Classification:** Infrastructure Clouds and Virtualisation

Contribution ID: 33

Type: **Oral Presentation**

## Simulate the social distancing in the new normal

*Tuesday, 23 March 2021 11:00 (30 minutes)*

In order to reduce the spread of COVID-19 since it broke out globally, the related studies have recommended maintaining a social distance of 1.5 meters, which avoids mass gatherings in order to reduce the spread of the epidemic by reducing contact between infected and uninfected persons. The implementation of relevant strategies not only affects the behavior pattern of the crowd in the open public space but also gives rise to the New Normal life pattern of the zero-contact economy. In addition, in the age of data, data acquisition, analysis, interpretation, and simulation can provide the assessment of epidemic spread in different situations according to different regional characteristics and population behavior patterns. In view of the above research requirements, this study firstly understands the characteristics of the research sample area through data analysis, then, it makes use of relevant characteristics to draw up different situation design, and uses the tool of flow simulation to present different situations. And they are also used to analyze the clustering behavior and dynamic lines of the population in the research sample area, so as to provide a reference for the strategy of reducing contact to reduce the spread of the epidemic.

This study takes "Ximen Ding" in Taipei city as the research sample area. First, it uses web crawlers to retrieve comments from social media sites in the study sample, namely, it refers to the number of times that a business facility is discussed on the Internet. Instead of using the positive and negative aspects of the comments as evaluation criteria, the discussion frequency is recorded and translated into the network share of voice of the region. And the potential population clusters in the research sample area are evaluated by thermal analysis of the spatial distribution of all commercial facilities in the field and the share of voice of its social media network. Finally, this study uses GAMA-Platform, a simulation analysis tool that has been widely used in recent years to simulate urban phenomena or predict the appearance of future cities. Through the agent-based analysis method, you can define the activity in each agent such as speed, born of location. etc it integrates the design and application of information visualization to present in the situation simulation. The crowd gathering places and open spaces tend to cause conflict points. And the study puts forward the reference for formulating spatial social distance related epidemic prevention strategies in the new epidemic situation in the future.

**Primary author:** Mr WANG, Yu-Chen (National Taipei University of Technology, Taipei Tech)

**Co-author:** Prof. WANG, Sheng-Ming (National Taipei University of Technology, Taipei Tech)

**Presenter:** Mr WANG, Yu-Chen (National Taipei University of Technology, Taipei Tech)

**Session Classification:** Biomedicine & Life Science

**Track Classification:** Health & Life Sciences (including COVID-19) Applications

Contribution ID: 34

Type: **Oral Presentation**

## EGI Workload Manager Service

*Wednesday, 24 March 2021 10:55 (25 minutes)*

The EGI Workload Manager Service (WMS) is based on the DIRAC Interware and is part of the EOSC Project service catalog. The service provides access to various computing resources of the EGI infrastructure to various scientific communities in Europe and in the world. Different kinds of computing resources can be connected to the Manager: HTC/grid resources, cloud resources or standalone computing clusters including HPC centers. The DIRAC WMS provides tools for submitting jobs with a detailed description of their requirements, job execution on the matching resources, monitoring and accounting of the consumed computing power. It assists users in construction and execution of complex workflows consisting of very large numbers of jobs automatically submitted as soon as all the necessary prerequisites are available.

The EGI WMS ensures user support helping to adapt their applications for the efficient use on the currently available resources. This required multiple developments to meet user's needs in accessing new computing technologies, e.g. GPUs, containers, cloud clusters, etc. New developments were also needed for managing user communities using Authentication/Authorization systems based on OAuth2/OIDC technologies and SSO Federated Identity Provides solutions.

In this contribution we will present the experience with setting up and running the EGI Workload Manager Service and we will describe the new developments carried out to fulfill requirements of the EOSC users.

**Primary author:** Dr TSAREGORODTSEV, Andrei (Aix Marseille Univ, CNRS/IN2P3, CPPM, Marseille, France)

**Co-author:** Mr LYTOVCHENKO, Andrii (Aix Marseille Univ, CNRS/IN2P3, CPPM, Marseille, France)

**Presenter:** Dr TSAREGORODTSEV, Andrei (Aix Marseille Univ, CNRS/IN2P3, CPPM, Marseille, France)

**Session Classification:** Network, Security, Infrastructure & Operations Session

**Track Classification:** Network, Security, Infrastructure & Operations

Contribution ID: 37

Type: **Oral Presentation**

## **A comprehensive security operation center based on big data intelligent detection and threat intelligence sharing**

*Thursday, 25 March 2021 10:50 (20 minutes)*

The continued growth of cybersecurity incidents calls for effective cybersecurity monitoring solutions. The operation of security operation centers (SOCs) is the recommended best practice to which large and medium-size enterprises rely for the detection, notification, and ultimately response to cybersecurity incidents. However, traditional SOCs using passive defense can not meet the current cybersecurity situation. In this talk, a comprehensive SOC is introduced which employs big data intelligent detection and threat intelligence sharing technology to detect and respond cybersecurity incidents rapidly and proactively. First, the SOC collects a wide variety of data including network traffic, server logs, security incidents, assets and vulnerability information. The collected data is stored in a big data storage platform for the following threat detection. Second, the SOC provides scalable cybersecurity incidents detection framework which can combine the detection performance of multiple detectors. Every detector can analyze behavior anomaly based on the data on the big data storage platform and multiple detectors can be correlated for further analysis. Third, the SOC can uniformly manage and respond the incidents identified from detection framework. At the same time, visualization is adopted to reveal the cybersecurity situation of entire enterprise. Besides, the SOC can share threat intelligence between multiple threat intelligence sharing instances and enrich threat intelligence by incorporating cybersecurity incidents from security incident response platform. The framework of SOC is referred to CERN and is customized to make it is practical and deployable for the Institute of High Energy Physics to discover, identify, understand, analyze, and respond to cybersecurity incidents from a global perspective.

**Primary author:** Dr JIARONG, Wang (IHEP)

**Co-authors:** Mr DEHAI, An (IHEP); Mr QI, Fazhi (Institute of High Energy Physics,CAS); Ms HU, Hao (Institute of High Energy Physics); Mrs LANXIN, Ma (IHEP); Dr YAN, Tian (IHEP)

**Presenter:** Dr JIARONG, Wang (IHEP)

**Session Classification:** Network, Security, Infrastructure & Operations Session

**Track Classification:** Network, Security, Infrastructure & Operations

Contribution ID: 38

Type: **Oral Presentation**

# Machine Learning Infrastructure on the Frontier of Virtual Unwrapping

*Tuesday, 23 March 2021 15:40 (20 minutes)*

Virtual unwrapping is a software pipeline for the noninvasive recovery of texts inside damaged manuscripts or scrolls via the analysis of three dimensional tomographic data, typically X-ray micro-CT. Recent advancements to the virtual unwrapping pipeline include the use of trained models to perform the “texturing” phase, where the content written upon a surface is extracted from the 3D volume and projected onto a surface mesh representing that page. Trained models are critical for their ability to discern subtle changes that indicate the presence or absence of writing at a given point on the surface.

The unique datasets and computational pipeline required to train and make use of these models make it a challenge to develop succinct, reliable, and reproducible research infrastructure. This paper presents our response to that challenge, and outlines our framework designed to support the ongoing development of machine learning models to advance the capability of virtual unwrapping.

This framework is designed around a set of principles chosen to reflect our team’s needs from this research area. We emphasize the following as the primary design principles, and briefly discuss our approach to each:

1. Visualization: Direct visualization of a trained model’s ability to reveal text from inside a scanned artifact is the most important tool we have to evaluate the model performance. The result of applying a trained model to a mesh inside a volume should ideally be a clear image revealing the cultural object’s content. The nature of the datasets requires that these images are created by our own custom Python scripts rather than image operations built into existing frameworks. For example, we generate training videos for each job to show how the output improves throughout training. We also often split an image into k regions for k-fold cross validation during training and make use of SLURM job arrays to perform these in parallel.

1. Automation: These complex custom pipelines need to be executed across hardware environments, often training in the University compute cluster on GPUs and then being used on desktop machines. We make use of Singularity containers to ensure a reliable computing environment. SLURM job dependencies enable summary and results upload jobs to be executed after the primary training job(s) has completed.
2. Data: We work with large datasets that are also frequently updated, and generate many output files for each experiment run. Google Drive is used to store both the datasets and results, and rclone is used to perform automated transfers ensuring job results are immediately visible to all team members.
3. Metadata: Our own methods and code are constantly changing, so we have to keep careful metadata to ensure our experiment results are clear to others or to ourselves in the future. All information required to reproduce any of our experiments is automatically saved to a JSON file which is uploaded upon job termination along with the job results.
4. Benchmarks: There are no existing performance benchmarks for these models, so it has been critical to develop our own benchmark datasets to compare our model performance across different methods and implementations.

We discuss the research improvements enabled by this framework, and in-progress and future improvements to the pipeline. We additionally discuss how this pipeline has prepared our team for performing active research in the midst of the unique challenges posed to the global research

community in the past year. Finally, we discuss the ways in which this framework can be applied to a variety of other research domains.

**Primary authors:** Mr PIKE, Charles (University of Kentucky); CHAPPELL, Jacob (University of Kentucky); PARSONS, Stephen (University of Kentucky); Prof. SEALES, William B. (University of Kentucky)

**Presenter:** PARSONS, Stephen (University of Kentucky)

**Session Classification:** Humanities, Arts & Social Sciences Session

**Track Classification:** Humanities, Arts, and Social Sciences (HASS) Applications

Contribution ID: 39

Type: **Oral Presentation**

## Proposed Hybrid Text and Image Compression Parallel Algorithm

*Tuesday, 23 March 2021 13:20 (20 minutes)*

Compression is a course used by numerous applications and techniques to make data transfer laidback and swift. Similarly, there are different types of data shared across such as text and image almost every second of our lives. However, with the growing number of working applications and users the compression techniques are limited and yet to explore and deploy. Hence, we tried to come up with a new compression route using LZSS and Huffman Coding for text compression. Also, we plan on incorporating this hybrid text compression using Parallel Processing that can drastically improve the time complexity of our algorithm. Deep convolution neural network is an underestimated and extensive tool for addressing image compression. We have tried to use the above technologies to achieve higher compression ratio than typical compression algorithms available. Also, we have used hybrid CNN to generate a better image which is unlike from existing techniques. Correspondingly including parallelization technique has helped us to optimize the output and reduce the responding time.

**Primary author:** Prof. WAGASKAR, Kalpita Ajinkya (Don Bosco Institute of Technology)

**Co-author:** Mr BITLA, Yash (Don Bosco Institute of Technology)

**Presenter:** Mr BITLA, Yash (Don Bosco Institute of Technology)

**Session Classification:** Data Management & Big Data Session

**Track Classification:** Data Management & Big Data

Contribution ID: 40

Type: **Oral Presentation**

## **WLCG SOC WG: Use of threat intelligence and deployment of SOC components at WLCG Tier-1 centres**

*Wednesday, 24 March 2021 14:00 (30 minutes)*

The information security threats currently faced by the research community are not only sophisticated, but also in many instances highly profitable for the actors involved. Evidence suggests that targeted organisations take on average more than six months to detect a cyber attack; the more sophisticated the attack, the more likely it is that it will pass undetected for longer. Enabling sites to receive appropriate, technical and actionable threat intelligence is a cornerstone of the WLCG security strategy. The mandate of the WLCG SOC working group primarily includes the exploration of the technical aspects of developing operational security capabilities, through the design and implementation of various components of Security Operations Centers. The technological means of sharing threat intelligence is provided by the Malware Information Sharing Platform (MISP) which allows for considerable flexibility in the design of an information sharing network. The topology adopted by the working group focuses in the first instance on a purpose deployed central MISP instance hosted at CERN, which leverages existing trust partnerships. In addition, MISP allows for a number of methods to access threat intelligence data, including synchronising events to peer instances as well as direct access via a REST API.

Particular attention has recently been paid to the rollout of threat intelligence capabilities for the WLCG Tier 1 sites. This rollout represents a key milestone not only in providing an important security function for WLCG in its own right, but also in an operational rollout sense, gaining experience that can then be used for the benefit of Tier 2 sites.

This effort covers two aspects. The first is establishing the capability of Tier 1 sites to access the threat intelligence available through the CERN-hosted WLCG academic MISP instance. This will allow the sites to become familiar with the threat intelligence available and consider where it can be integrated within their existing Intrusion Detection and Intrusion Prevention Systems. The second aspect is the deployment, where necessary, of a minimally viable SOC following the SOC working group reference design. The approach taken is to work with individual sites following their specific circumstances. The goal of these activities is to provide sites with the threat intelligence and intrusion detection capabilities allowing them to quickly detect and defend against typical attacks.

We report on specific experiences of sites deploying SOC components and gaining access to threat intelligence, in addition to reporting on experience gained in integrating threat intelligence with existing Intrusion Detection and Intrusion Prevention systems.

**Primary authors:** Dr CROOKS, David (UKRI STFC); Mr VALSAN, Liviu (CERN)

**Presenters:** Dr CROOKS, David (UKRI STFC); Mr VALSAN, Liviu (CERN)

**Session Classification:** Network, Security, Infrastructure & Operations Session

**Track Classification:** Network, Security, Infrastructure & Operations

Contribution ID: 41

Type: **Oral Presentation**

## **Sentinel Asia: A space-based disaster management support in the Asia-Pacific region**

*Tuesday, 23 March 2021 13:00 (20 minutes)*

The Sentinel Asia (SA) is a voluntary initiative led by the Asia-Pacific Regional Space Agency Forum (APRSAF), established in 2005 as a collaboration between space agencies and disaster management agencies, applying remote sensing and Web-GIS technologies to assist disaster management in the Asia-Pacific region. It provides and shares disaster information in near-real-time across the Asia-Pacific region using Web-GIS platform (<https://sentinel-asia.org/>).

SA is promoted with cooperation amongst the space community (APRSAF), the international community (UNWFP, UNESCAP, UNOOSA, ASEAN, the Asian Institute of Technology (AIT), and so on), and the disaster management community (the Asian Disaster Reduction Center (ADRC) and its member organizations). To support the implementation of the SA project, a Joint Project Team (JPT) was organized. Membership in the JPT is open to all disaster prevention organizations and regional/international organizations that are prepared to contribute their experiences and technical capabilities and wish to participate in technical aspects of disaster information sharing activities. JPT consists of 111 organizations from 28 countries/region and international organizations.

JAXA is a secretariat of the JPT. In particular, space agencies in the Asia-Pacific region—the Japan Aerospace Exploration Agency (JAXA), the Indian Space Research Organization (ISRO), the Geoinformatics and Space Technology Development Agency (GISTDA), the National Applied Research Laboratories (NARLabs), the Vietnamese Academy of Science and Technology (VAST), Mohammad bin Rashid Space Centre (MBRSC) and the Center for Remote Imaging, Sensing and Processing (CRISP) National University of Singapore—are supporting the SA's emergency observation of damaged areas by disasters.

SA has responded over 330 emergency requests from JPT and ADRS members since 2007. SA is promoting to make stronger cooperation with end-users used by satellite-based emergency mapping.

**Primary author:** Dr KAWAKITA, Shiro (Japan Aerospace Exploration Agency)

**Presenter:** Dr KAWAKITA, Shiro (Japan Aerospace Exploration Agency)

**Session Classification:** Data Management & Big Data Session

**Track Classification:** Data Management & Big Data

Contribution ID: 42

Type: **Oral Presentation**

## Air quality prediction of Ulaanbaatar using machine learning approach

*Tuesday, 23 March 2021 15:00 (30 minutes)*

The capital city Ulaanbaatar is one of the most polluted city in the world. The government are looking for suited solution decreasing the air pollution, but it could not work out efficiency. Air pollution is contained from various components like particulate matter (PM), gaseous pollutants like ozone, nitrogen dioxide (NO<sub>2</sub>), carbon monoxide (CO), sulfur dioxide (SO<sub>2</sub>) and other organic compounds. Ulaanbaatar city have 15 air quality monitoring stations. There are some air quality monitoring data of Ulaanbaatar, that PM<sub>10</sub> particulate matter in January 2019 was 5% higher than air quality in 2018 and tolerance level is 2.5 times higher than air quality standard (WHO), the average amount of PM<sub>2.5</sub> in January was 12% higher compared to the same period of the previous year and 3.9 times higher than the air quality standard, also the average amount of nitrogen dioxide in January was 10% higher than 2018 and 1.1 times higher than the air quality standard, the amount of sulfur dioxide in January 2019 was 1.2 times higher than the air quality standard compared to the same period in 2018. The air quality is affected by Ger district, vehicle, furnace, industry, thermal power station and other.

Predicting and forecasting air quality is the one of the most essential activity in the Smart City. Recently, there are many study to use the machine learning approaches for evaluating and predicting air quality using big data. The aim of this study is to obtain machine learning model for air quality forecasting using previous air quality station data and the weather data. The air quality depends on multi-dimensional factors including location, time, weather parameters, such as temperature, humidity, wind direction and force, air pressure, etc. There are many machine learning approaches, but artificial neural Network model tries to simulate the structures and networks within human brain. It is convenient for working to find relation between multi parameters. If the neural network could determine the relation of the air quality using the weather and air quality data of last year, it is possible to predict approximately air quality of Ulaanbaatar city. We used input layers including parameters of temperature, humidity, wind direction, air pressure, PM<sub>2.5</sub> and PM<sub>10</sub>, NO<sub>2</sub>, CO, SO<sub>2</sub> and measuring time to build neural network model. We are working on the neural network model and this time we are going to introduce the machine learning approach predicting air quality of Ulaanbaatar. In the further work we are going to do experiment of neural network algorithm of the air quality prediction and to discuss machine learning results.

**Primary author:** Mrs BADRAKH, Otgonsuud (Mongolian Academy of Sciences)

**Presenter:** Mrs BADRAKH, Otgonsuud (Mongolian Academy of Sciences)

**Session Classification:** Earth & Environmental Sciences & Biodiversity Applications

**Track Classification:** Earth/Environmental Sciences & Biodiversity Applications

Contribution ID: 43

Type: **Oral Presentation**

## **SOME STUDY RESULTS OF COLOR CHANGES DEPENDING ON MONGOLIAN ENVIRONMENTAL CONDITION**

*Tuesday, 23 March 2021 13:20 (20 minutes)*

We started a color experiment, which is tried to study the quality of colorants and color fading in ancient buildings. This long-term color measurement can be used to determine how the colorants change by environmental influence, and to develop the methodology for such analysis will predict the initial object color. This will allow the cultural heritage to be restored by its original color, as well as to determine which colorants have the least changing and the most durable. Therefore, in this study we prepared a measurement object to analyze color fading, performed long-term and short-term measurements weekly for five months. Processing of the measurement results shows that most of the pigments have faded and most of the dye were stable, but the white dye changed the most. It is possible that the change depends on the absorption of the color spectrum. In the future, we need to increase the measurement data, so the data analysis will be more precise, and the aim is to forecast the initial color of the object using a machine learning approach based on regression analysis by more color measurement values.

**Primary author:** Mrs BADRAKH, Otgonsuud (Mongolian Academy of Sciences)

**Presenter:** Mrs BADRAKH, Otgonsuud (Mongolian Academy of Sciences)

**Session Classification:** Humanities, Arts & Social Sciences Session

**Track Classification:** Humanities, Arts, and Social Sciences (HASS) Applications

Contribution ID: 45

Type: **Oral Presentation**

## Event classification using Graph Neural Network

*Friday, 26 March 2021 09:00 (30 minutes)*

After the Higgs boson discovery, the main interest in elementary particle physics is the discovery of beyond the Standard Model. LHC, which is the most energetic collider in the world, continues to be the leading experiment in the energy frontier. While LHC does not increase the center of mass energy beyond 14 TeV over a few decades, the amount of collision data will be significantly increased by the upgrade to High-Luminosity LHC. We need to leverage the observed data to the fullest extent in such an era.

One of the most attractive ways to utilize observed collision data is deep learning. Deep learning can represent a complex correlation between the input variables, and it is known to have better sensitivity than the traditional analysis method in several practical examples. Although deep learning has a very significant capability to represent any complicated functions, such a representation ability leads to overfitting of the training data. To use deep learning more effectively, we need to embed our domain knowledge in the deep learning model as inductive bias. Graph neural network could accomplish such a requirement. Graph neural network handles a graph, which has nodes and edges. Using a graph structure, we can assign a grouped element as a node and the known relation between elements as an edge. Additionally, a graph structure could produce interpretable results than the typical one-row representation.

We apply the graph neural network to an event classification problem,  $t\bar{t}Hbb$  process, to embed a domain knowledge as an inductive bias in the model. Implementation of inductive bias is expected to avoid the unphysical calculation of the input variables in the neural network. We expect to avoid overfitting and accomplish better classification performance than the traditional multilayer perceptron model, in particular, when the number of training data is insufficient. We construct the graph neural network model for the event classification for a typical physics process and compare the performance between the graph neural network and traditional analysis methods such as multilayer perceptron, Long-Short Term Memory, and boosted decision tree.

Additionally, it is important to consider the known constraint in the experimental data. Usually, experimental data has symmetries, e.g., spatial rotation invariance. A generic deep learning model does not consider such a rule, resulting in the over-fitting. By giving such a known rule, we can improve the performance of deep learning.

We will report the latest results and the issues in the implementation of the high-level domain knowledge.

**Primary author:** Mr SAITO, Masahiko (ICEPP, The University of Tokyo)

**Co-authors:** Prof. TANAKA, Junichi (University of Tokyo); TERASHI, Koji (University of Tokyo); SAWADA, Ryu (University of Tokyo); Mr KISHIMOTO, Tomoe (University of Tokyo); IYAMA, Yutaro (University of Tokyo)

**Presenter:** Mr SAITO, Masahiko (ICEPP, The University of Tokyo)

**Session Classification:** Physics & Engineering Session

**Track Classification:** Physics (including HEP) and Engineering Applications

Contribution ID: 46

Type: **Oral Presentation**

## HIFIS transfer service: FTS for Helmholtz

*Wednesday, 24 March 2021 13:40 (20 minutes)*

The German Helmholtz Association consists of 19 research institutes distributed over Germany, covering a wide variety of research topics ranging from particle and material physics over cancer research to marine biology. In order to stimulate collaborations between different centres, Helmholtz established so-called incubator platforms. One of those platforms, HIFIS, aims to provide the infrastructure for federating IT services offered by the different Helmholtz centres. The other platforms provide thematically bound resources for domain scientists in the forms of project funding for cross-centre and cross-domain collaborations, computing resources, and consulting services.

Use cases of interdisciplinary research are arising in the platforms and show that there is a definitive need to transfer a significant amount of large data sets between centres. This results from the fact that primary data acquisition and the subsequent processing steps are increasingly distributed over multiple centers. As data processing is generally sensitive to network latencies, remote data access is not efficient in those cases and consequently data needs to be transferred from the primary institution to another, where the data processing is taking place.

In order to cater to those needs, a file transfer service is being established by HIFIS for convenient and automated data transfers between the sites of the aforementioned cross-centre research groups. After evaluating alternative solutions like Globus Online and Onedata, we agreed to use FTS3 for reasons we will elaborate on during the presentation. FTS3 is a file transfer service that can commission data transfers between storage endpoints and has been developed at CERN for the transfer of WLCG research data between CERN and several hundred LHC Tier centres. Those endpoints need to be able to communicate with FTS and each other using a third-party copy (TPC) extension of the HTTP protocol to transfer data directly. In order to facilitate an easy installation of endpoints, we provide an Apache web server extension that complies with the needs of FTS3 and can thus act as an endpoint for data transfers via HTTP-TPC.

We will present the necessary prerequisites and setup variants for such an endpoint, its configuration details, as well as a brief overview of the modifications applied to the Apache modules. Adding to that, we will present insights into the access possibilities, performance and reliability of the data transfers.

### Summary

We present details of and use cases for a storage endpoint solution based on the Apache web server capable of making use of FTS3 for transferring large data sets.

**Primary authors:** Dr FUHRMANN, Patrick (DESY/dCache.org); Dr MILLAR, Paul (DESY); Mr WETZEL, Tim (Deutsches Elektronen-Synchrotron DESY); Dr JANDT, Uwe (DESY)

**Presenter:** Mr WETZEL, Tim (Deutsches Elektronen-Synchrotron DESY)

**Session Classification:** Data Management & Big Data Session

**Track Classification:** Data Management & Big Data

Contribution ID: 47

Type: **Oral Presentation**

## Recent Highlights of the Belle II Computing

*Friday, 26 March 2021 11:00 (30 minutes)*

The Belle II is the next-generation flavor factory experiment at the SuperKEKB accelerator in KEK (High Energy Accelerator Research Organization), Tsukuba, Japan. The first physics run with the full Belle II detector started in 2019, then we are now increasing the luminosity gradually. We have already exceeded the highest world record of the luminosity, which was achieved by Belle, the predecessor of our experiment. The operation will be continued until 50 times more data than Belle is collected. Such a huge amount of data allows us to explore the new physics possibilities through a large variety of analyses in quark sectors as well as in tau physics and to deepen understanding of nature.

The Belle II computing system is expected to manage the process of massive raw data, production of copious simulation and many concurrent user analysis jobs. For the Belle II is a worldwide collaboration of about 1,000 scientists working in 26 countries and region, we adopted a distributed computing model with DIRAC as a workload management system. Recently, we decided to integrate Rucio in BelleDIRAC (an extension of the basic DIRAC) system as a distributed data management (DDM) tool. We are now testing the Rucio functionality and preparing the transition from the current own made DDM to Rucio, which will happen in January 2021.

We also have established the automated procedures to transfer the Raw data from the online storage in the data acquisition system to the permanent storage at the offline computing facility located 1.5km away from the detector. In these procedures, not only the data copy but also format conversion, data validation, quality check and its performance monitoring are included. Thanks to these procedure, the Raw data is available on the KEK offline computing system within half a day after the data taking. And this allows physicists to perform the detector calibration in a timely manner.

For near future, we are planning to execute the detector calibration process on the distributed computing system to accelerate the speed of calibration performance. This also mitigates the heavy load of the KEK offline computing system. However, to realize this, we need to make the Raw data available on the Grid. We have developed an automatized Raw data registration and replication tool, so-called BelleRawDIRAC (the another extension of DIRAC) and started to use this from this year.

We will present the highlights of the recent achievements in the Belle II computing as well as the general status and plans of the Belle II experiment in this report.

**Primary author:** Dr HARA, Takanori (High Energy Accelerator Research Organization (KEK))

**Presenter:** Dr HARA, Takanori (High Energy Accelerator Research Organization (KEK))

**Session Classification:** Physics & Engineering Session

**Track Classification:** Physics (including HEP) and Engineering Applications

Contribution ID: 48

Type: **Oral Presentation**

## Performance estimation of deep learning methods for change detection on satellite images with a low-power GPU

*Tuesday, 23 March 2021 15:30 (30 minutes)*

In the last few years, Earth observation has known a significant growth due to both the increased number of satellite launched into orbit and the augmented data production capabilities provided by new sensors on board. However, processing this huge amount of data is becoming a crucial problem that require innovative solutions both on board, where the amount of mass memory storage is very limited, and on the ground stations, where data are often transmitted after long time due to the bottleneck effect resulting from the limited bandwidth.

In order to provide an optimization of the big data processing, two concurrent possible solutions are being investigated: transmitting to the ground station only a few selected data (containing interesting information) and exploiting heterogeneous computing on board the satellites, to speed up computational expensive operations of data analysis.

Change detection is an interesting task in the field of remote sensing, with many useful applications, from land cover studies to anomalies' observation (landslips, snowslides, wide firewoods, floods, oilspills into the sea, etc.). Many satellites like Sentinel-2 provide a full coverage of our planet every few days, but transmitting multispectral images of the same area multiple times within a small time interval might not be efficient for certain tasks. At the same time, the analysis of each image on ground requires a considerable amount of time and efforts that might be reduced if knowing in advance that some data do not provide additional information with respect to those acquired few days before. Therefore, the idea of comparing on board a new image with the older one corresponding to the same area on Earth can be a very powerful tool to both reduce the bottleneck effect and organize the analysis on the ground stations more efficiently by adding redundant data on a low-priority queue.

In this study, a deep learning approach is used to accomplish the change detection task on Sentinel-2 multispectral images. A pre-existing dataset focused on urban changes is exploited for training and validation purposes, with ground truths available in the form of binary change maps that provide pixel-wise target labels. Two different neural networks are tested. The first features an encoder-decoder structure to perform a segmentation of the two images of each pair and provide the exact location of each change happened in the processed area; for speed and computational load reasons, it was designed for the execution on the ground stations. The second, instead, is a classical convolutional neural network that aims at classifying each pair as hosting a considerable number of changes or not; therefore, its speed and smaller architecture satisfy the main requirements for the execution on board the satellite, exploiting dedicated hardware.

Finally, benchmark tests are also conducted on a low-power GPU, the NVIDIA Jetson AGX Xavier, to investigate throughput and speed results with both TensorFlow and NVIDIA TensorRT on this energy-efficient device. Indeed, the space scenario has inevitable constraints on the power consumption and energy efficiency of the instruments that can be used, and this platform shows important features that make it suitable for the installation on board the satellites in future missions.

**Primary author:** Mr DI PILATO, Antonio (University of Bari and INFN Bari)

**Co-authors:** IACOBELLIS, Michele (Planetek Italia); TAGGIO, Nicolo (Planetek Italia)

**Presenter:** Mr DI PILATO, Antonio (University of Bari and INFN Bari)

**Session Classification:** Earth & Environmental Sciences & Biodiversity Applications

**Track Classification:** Earth/Environmental Sciences & Biodiversity Applications

Contribution ID: 49

Type: **Oral Presentation**

## PCJ Java Library as a solution to integrate HPC, Big Data and Artificial Intelligence workloads

*Friday, 26 March 2021 14:30 (30 minutes)*

With the development of peta- and exascale size computational systems there is growing interest in running Big Data and Artificial Intelligence (AI) applications on them.

Big Data and AI applications are implemented in Java, Scala, Python and other languages that are not widely used in High-Performance Computing (HPC) which is still dominated by C and Fortran. Moreover, they are based on dedicated environments such as Hadoop or Spark which are difficult to integrate with the traditional HPC management systems.

We have developed the PCJ library (Parallel Computing in Java), a tool for scalable high-performance computing and big data processing in Java. In this paper, we present the basic functionality of the PCJ library with examples of highly scalable applications running on the large resources. The performance results are presented for a different classes of applications including traditional computational intensive (HPC) workloads (e.g. stencil), as well as communication intensive algorithms such as Fast Fourier Transform (FFT). We present implementation details and performance results for Big Data type processing running on petascale size systems. The examples of large scale AI workloads parallelized using PCJ are presented.

**Primary author:** Dr NOWICKI, Marek (N. copernicus University)

**Co-authors:** Prof. BAŁA, Piotr (University of Warsaw); Dr GÓRSKI, Łukasz (ICM University of Warsaw)

**Presenters:** Dr NOWICKI, Marek (N. copernicus University); Prof. BAŁA, Piotr (University of Warsaw)

**Session Classification:** Converging High Performance infrastructures: Supercomputers, clouds, accelerators Session

**Track Classification:** Virtual Reserach Environment (including Middleware, tools, services, workflow, ... etc.)

Contribution ID: 50

Type: **Oral Presentation**

## Using HEP experiment workflows for the benchmarking of WLCG computing resources

*Friday, 26 March 2021 11:30 (30 minutes)*

The HEP-SPEC06 (HS06) benchmarking suite has been used for over a decade in the accounting and procurement of WLCG resources. HS06 is stable, accurate and reproducible, but it is an old benchmark and it has become clear that its performance and that of typical HEP applications have started to diverge. After evaluating several alternatives for the replacement of HS06, the HEPIX benchmarking WG has chosen to focus on the development of a HEP-specific benchmarking suite based on actual software workloads of the LHC experiments. This approach, based on container technologies, is designed to provide by construction a better correlation between the new benchmark and the throughputs of the experiment production workloads. It also offers the possibility to separately explore and describe the independent architectural features of different computing resource types. This is very important in view of the growing heterogeneity of the HEP computing landscape, where the role of non-traditional computing resources such as HPCs and GPUs is expected to increase significantly. This presentation will review the status and outlook of development of the new benchmarking suite, and in particular of the efforts to value HPC resources and GPUs, at the time of ISGC2021. It will also describe the ongoing effort to agree on a common WLCG policy for the deployment of the new benchmark in production, for which a dedicated task force has recently been set up.

**Primary author:** Dr VALASSI, Andrea (CERN)

**Presenter:** Dr SOUTHWICK, David (CERN)

**Session Classification:** Physics & Engineering Session

**Track Classification:** Physics (including HEP) and Engineering Applications

Contribution ID: 51

Type: **Oral Presentation**

## A Big Data Platform for heterogeneous data collection and analysis in large-scale data centers

*Tuesday, 23 March 2021 13:40 (20 minutes)*

With the upcoming start of Run-3, and especially Run-4, the amount of data managed by the WLCG data centres is expected to massively increase. In this context the importance of using efficient routines for the analysis of the incoming data is crucial to ensure the effectiveness of the HEP experiments. The actionable data that is available for the analysis also include the logging data produced by each computing node of the WLCG infrastructure.

Much information could be extracted from these sources, regarding the state of the system, and proper analysis of this data could result in better insight on its working state and eventually predict and prevent faults and errors. While efforts towards the extraction of these insight has enormous potential advantages, having an effective and efficient monitoring system is a critical asset, and a continuous concern for administrators of the data centre.

The INFN Tier-1 data centre hosted at CNAF has used various monitoring tools over the years all replaced, a few years ago, by a system common to all CNAF functional units, based on Sensu, Influxdb, and Grafana to acquire, store and visualize facility metrics (e.g. CPU load, memory usage, IO requests). Given the complexity of the inter-dependencies of the several services running at the data centre and the foreseen large increase of resource usage, a more powerful and versatile monitoring system is needed.

In this context, INFN-CNAF is promoting effort to introduce a log ingestion and analytics platform across its services, with the purpose of exploring possible solutions for the development of Predictive Maintenance model to detect and anticipate failures. This new monitoring system should provide a framework to provide and be able to correlate log files and metrics coming from heterogeneous sources and devices. Due to the novelty of this approach, and to the variety and heterogeneity of the involved data sources, identifying, extracting, and processing valuable information in an efficient and effective way is a challenging task.

We present a layered scalable big data infrastructure for data aggregation, exploration, and analysis for large data centers. Leveraging open source and de facto standard technologies, the proposed infrastructure follows a general approach ensuring its portability to different frameworks. On the bottom level, forwarding tools such as Filebeat, Fluentd and Thingsboard perform data ingestion from different sources (e.g. syslog, time series databases, IoT sensors). Timely distribution to processing nodes is carried out at a higher level by the topic-based publish-subscribe engine Apache Kafka. Processing is performed through Apache Spark and Spark Streaming instances. Data persistence is achieved through MinIO's high performance Object Storage, that allows long term storage for both raw data and the results of the analysis. The topmost layer includes visualization and alerting tools such as Kibana and Grafana through which the final users (i.e., sysadmins) may monitor the status of the system, be notified about predicted faults or explore new data.

The platform will satisfy the requirements of the Tier-1 group and the CNAF departments for more actionable insight from monitoring data, and provide the data center with an advanced infrastructure able to explore and analyze heterogeneous data in all the projects the center is involved.

**Primary authors:** Dr MARTELLI, Barbara (INFN - CNAF); CESINI, Daniele (INFN-CNAF); Mr ROSSI TISBENI, Simone (INFN-CNAF)

**Co-authors:** Dr FALABELLA, Antonio (INFN); CARBONE, Arianna (INFN-CNAF); CAVALLARO,

Claudia (INFN-CNAF); MICHELOTTO, Diego (INFN-CNAF); DUMA, Doina Cristina (INFN - CNAF); FURLAN, Elisabetta (INFN-CNAF); Dr RONCHIERI, Elisabetta (INFN CNAF); SERGI, Giusy (INFN-CNAF); GAS-PARETTO, Jacopo (INFN-CNAF); MORGANTI, Lucia (INFN-CNAF); GALLETTI, Matteo (INFN-CNAF)

**Presenter:** Mr ROSSI TISBENI, Simone (INFN-CNAF)

**Session Classification:** Data Management & Big Data Session

**Track Classification:** Data Management & Big Data

Contribution ID: 52

Type: **Oral Presentation**

## Deep Learning fast inference on FPGA for CMS Muon Level-1 Trigger studies

*Friday, 26 March 2021 14:00 (30 minutes)*

Machine and Deep Learning techniques experienced an explosion in the adoption of a variety of HEP applications, ranging from event selection in trigger operations to end-user physics data analysis, as well as computing metadata based optimisations.

The range of applicability of such techniques in the High Energy Physics (HEP) context – with a particular accent on the experiments at the Large Hadron Collider (LHC) at CERN in Geneva – will extend to an even larger variety of applications, if low-latency hardware solutions are added to the ones usually exploited (CPUs/GPUs - or even Google TPUs).

An area of application in particle physics, for instance, is the domain of FPGA-based Trigger/DAQ, characterized by even sub-microsecond latency requirements: stringent prerequisites towards this solution come from the upcoming Run-3 needs and even more, from the evolution towards the operational conditions foreseen for the High-Luminosity LHC (HL-LHC) phase.

Crucial ingredients required to be prepared for this future are the availability of adequate hardware resources and expertise, as well as the capability to streamline the process of building and testing ML/DL models into FPGA firmware.

This paper presents and discusses the activity running at the University of Bologna and INFN-Bologna: pioneered by members of the CMS experiment together with a cross-experiment working group, including LHC physicists and electronics experts, the work is focused on the development of Neural Network models for HEP use-cases as well as prototyping the exploitation of various FPGA resources for the model inference. In particular, the target hardware used consisted in a Xilinx ZCU102 Evaluation Board featuring a Zynq Ultrascale+ MPSoC, mounted in a test bed controlled environment. To synthesize the high-level network architectures, written in Python, in a language oriented to a more circuital description (HDL), the High-Level Synthesis (HLS) toolkit was used - called HLS4ML - developed very closely to the HEP needs.

Hardware and software set-up, and performances on various baseline models used as benchmarks, will be presented. In particular, a specific real-life case study on the CMS Muon Level-1 Trigger will be presented, implementing a Deep Neural Network for the assignment of muon transverse momentum ( $p_T$ ) in an FPGA readable firmware. The possible advantages of performing such neural networks on dedicated programmable hardware, in terms of latency reduction as well as the tuning of inference performances with parallelisation and quantisation capabilities, for this class of problems will also be discussed.

As a future development for this study, a comparison in terms of performance and resource usage can be performed, using different hardware solutions; in particular, a newly acquired Vadatech ATC136 board hosting a Xilinx Virtex-7 FPGA, mounted on a ATCA crate and installed in the INFN-CNAF Tier-1 data center.

**Primary author:** Dr DIOTALEVI, Tommaso (University of Bologna and INFN)

**Co-authors:** Dr BATTILANA, Carlo (University of Bologna and INFN); Prof. BONACORSI, Daniele (University of Bologna and INFN); Dr LORUSSO, Marco (University of Bologna); Dr TRAVAGLINI, Riccardo (INFN Bologna)

**Presenter:** Dr DIOTALEVI, Tommaso (University of Bologna and INFN)

**Session Classification:** Converging High Performance infrastructures: Supercomputers, clouds, accelerators Session

**Track Classification:** Converging High Performance infrastructures: Supercomputers, clouds, accelerators

Contribution ID: 53

Type: **Oral Presentation**

## Machine Learning as a Service for High Energy Physics on heterogeneous computing resources

*Thursday, 25 March 2021 13:00 (30 minutes)*

Machine Learning (ML) techniques in the High-Energy Physics (HEP) domain are ubiquitous and will play a significant role also in the upcoming High-Luminosity LHC (HL-LHC) upgrade foreseen at CERN: a huge amount of data will be produced by LHC and collected by the experiments, facing challenges at the exascale. Despite ML models are successfully applied in many use-cases (online and offline reconstruction, particle identification, detector simulation, Monte Carlo generation, just to name a few) there is a constant seek for scalable, performant, and production-quality operations of ML-enabled workflows. In addition, the scenario is complicated by the gap among HEP physicists and ML experts, caused by the specificity of some parts of the HEP typical workflows and solutions, and by the difficulty to formulate HEP problems in a way that match the skills of the Computer Science (CS) and ML community and hence its potential ability to step in and help. Among other factors, one of the technical obstacles resides in the difference of data-formats used by ML-practitioners and physicists, where the former use mostly flat-format data representations while the latter use to store data in tree-based objects via the ROOT data format. Another obstacle to further development of ML techniques in HEP resides in the difficulty to secure the adequate computing resources for training and inference of ML models, in a scalable and transparent way in terms of CPU vs GPU vs TPU vs other resources, as well as local vs cloud resources. This yields a technical barrier that prevents a relatively large portion of HEP physicists from fully accessing the potential of ML-enabled systems for scientific research.

In order to close this gap, a Machine Learning as a Service for HEP (MLaaS4HEP) solution is presented as a product of R&D activities within the CMS experiment. It offers a service that is capable to directly read ROOT-based data (both local and remote), apply the ML solution developed by the user, and ultimately serve predictions by pre-trained ML models “as a service” accessible via HTTP protocol.

This solution can be used by physicists or experts outside of HEP domain and it provides access to local or remote data storage without requiring any modification or integration with the experiment’s specific framework. Moreover, MLaaS4HEP is built with a modular design allowing independent resource allocation that opens up a possibility to train ML models on PB-size datasets remotely accessible from the WLCG sites without physically downloading data into local storage.

To prove the feasibility and utility of the MLaaS4HEP service with large datasets and thus be ready for the next future when an increase of data produced is expected, an exploration of different hardware resources is required. In particular, this work aims to provide the MLaaS4HEP service transparent access to heterogeneous resources, which opens up the usage of more powerful resources without requiring any effort from the user side during the access and use phase. We show the comparison of performance using different kinds of resources, both local and remote, for typical physics use-cases, e.g. in signal vs background discrimination problems.

**Primary author:** Dr GIOMMI, Luca (INFN and University of Bologna)

**Co-authors:** Prof. BONACORSI, Daniele (University of Bologna); Dr KUZNETSOV, Valentin (Cornell University, Ithaca, USA)

**Presenter:** Dr GIOMMI, Luca (INFN and University of Bologna)

**Session Classification:** Infrastructure Clouds and Virtualisation Session

**Track Classification:** Infrastructure Clouds and Virtualisation

Contribution ID: 55

Type: **Oral Presentation**

## **Piloting Interactive Data Science Learning Platforms through the Development of Cloud-based Computational Digital Notebooks**

*Tuesday, 23 March 2021 16:00 (20 minutes)*

In a world presently mired in cascading lockdowns resulting from the spread of the COVID-19 pandemic, physical learning, communication, and collaboration have taken a colossal hit. Virtual access has become the need-of-the-hour, and the uses of cloud-based course content delivery, distance learning, and document collaboration are becoming increasingly ubiquitous. This paper introduces a novel method to allow students and faculty in the Humanities, Arts, and Social Sciences (HASS) to collaborate and interact through data analytic technologies using computational notebooks.

We demonstrate this approach using a digitally archived Legacy of Slavery (LoS) dataset from the Maryland State Archives (MSA) and illustrate the socio-technical challenges faced in establishing this learning environment. We provide a step-by-step process involved in accessing, developing, and integrating different infrastructure elements. The LoS in Maryland is a major initiative of the MSA. The program seeks to preserve and promote the vast universe of experiences that have shaped the lives of Maryland's African American population. Over the last 18 years, some 420,000 individuals have been identified, and data assembled into 16 major databases. These databases contain information unique to enslaved people's lives such as manumission records, certificates of freedom, census data, penitentiary records, etc.

One of this paper's primary objectives is to enable the digital representation of these culturally rich and sensitive collections ready to be analyzed and studied through contemporary scholars' lenses. This project aims to achieve this goal by making these databases available and accessible so that users can generate individual stories, glean insights, and possibly recover "erased" memories of the enslaved people. To achieve this goal, as a first step, individual dataset collections were prepared by downloading the databases and put through rigorous exploration, cleaning, and visualization process through coordination with interdisciplinary scholars composed of archivists, historians, computer scientists, and technology analysts. This project also illustrates the importance of a multidisciplinary approach to a unique set of digitized archival data with a specific focus on contextual aspects due to the data's historical value and sensitivity. The collaborative process used open-source and readily-accessible tools to create meaningful visualizations as an arrangement that flows together conducive for educators to teach. The visualizations use the spatial and temporal characteristics of the datasets to produce graphs and charts for a graphical view of the datasets. The visualizations constructed are responsive to present the data by instant connections to the datasets dynamically. Integration of these digital artifacts obtained from each dataset followed next on editors called "digital notebooks" that allow text and software code to co-exist and render in a single document coherently for instructors and students to follow the text with visual representations back-to-back. The "digital notebooks" are equipped with live examples of machine learning models and natural language processing on certain text-rich features of these dataset collections. The open-source nature of this project's setup and cloud-based distribution of these digital notebooks pave the way for students from underserved communities to take advantage of a unique way of learning and to perform hands-on work on marketable software tools, preparing them for a successful career.

The contributions of this paper to the field of digital humanities lie in the idea of providing an "always-on" cloud-based pedagogical environment for aspiring archivists and researchers worldwide to analyze, learn and unearth stories through data and fact-driven approach on the Legacy of Slavery in the State of Maryland.

**Primary author:** GNANASEKARAN, Rajesh Kumar (The University of Maryland)

**Co-author:** Prof. MARCIANO, Richard (The University of Maryland)

**Presenter:** GNANASEKARAN, Rajesh Kumar (The University of Maryland)

**Session Classification:** Humanities, Arts & Social Sciences Session

**Track Classification:** Humanities, Arts, and Social Sciences (HASS) Applications

Contribution ID: 56

Type: **Oral Presentation**

## Dynamic storage provisioning for elastic cloud services with dCache

*Friday, 26 March 2021 11:30 (30 minutes)*

DESY contributes computing and storage infrastructure to the EGI Federated Cloud that is used by Photon and Neutron Research Infrastructures (PaN RIs) in European H2020 projects, working towards the European Open Science Cloud (EOSC). Building on this federated Infrastructure-as-a-Service (IaaS), DESY - as a research centre of the Helmholtz association - provides integrated Platform-as-a-Service (PaaS) solutions to EGI users and to Helmholtz Federated IT Services (HIFIS). Serving for scientific cloud computing applications, which need dynamically provisioned local data access for an optimized resource usage, dCache adds robust, elastic storage for private and hybrid cloud solutions.

The PaN RIs are producing petabytes of data with free electron laser, synchrotron light and neutron sources and provide FAIR data services to a wide scientific user community. The Photon and Neutron Data Service (ExPaNDS) project expands and accelerates access to this data through the EOSC. Use-cases include the integration with high performance computing infrastructure at site and span to data lake type scenarios with federated storage systems. The development of the Photon and Neutron Open Science Cloud (PaNOSC) comprises a spectrum of use-cases that reach from environments for development and testing of containerized applications to running compute intensive simulations.

To support these use-cases for federated cloud computing resources, DESY integrates services for the whole life cycle of user supplied virtual machines and containers, adopting solutions from the EOSC and providing incentives for scaling to the enormous demands of the PaN RIs.

This builds on enabling efficient and secure distribution of private and public container images as shared ready-to-use environments, that can auto-scale user jobs. Users and service providers, who realize their work leveraging Kubernetes' orchestration capabilities can provision virtualized compute nodes, abstracting away most of the required interaction with the underlying Cloud Platform and benefiting from additional building blocks for monitoring, logging and alerting functionality as well as an application catalog. This service adds centralized authentication and role-based access control (RBAC).

This talk will highlight how the dCache storage system tailors for this platform and use cases running on federated cloud projects at DESY, effectively enabling federated AAI, token-based access, and dynamically provisioning access to the storage for elastic virtual scientific computing environments on hybrid cloud and container orchestration systems.

**Primary author:** Mr SCHUH, Michael (DESY)

**Co-authors:** Dr VOSS, Christian (DESY Hamburg); REPPIN, Johannes (DESY); Dr FUHRMANN, Patrick (DESY/dCache.org); Dr MILLAR, Paul (DESY); Mr MKRTCHYAN, Tigran (DESY); Mr WETZEL, Tim (Deutsches Elektronen-Synchrotron DESY)

**Presenter:** Mr SCHUH, Michael (DESY)

**Session Classification:** Infrastructure Clouds and Virtualisation Session

**Track Classification:** Infrastructure Clouds and Virtualisation

Contribution ID: 57

Type: **Oral Presentation**

## Using Natural Language Processing to Extract Information from Unstructured code-change version control data: lessons learned

*Friday, 26 March 2021 10:00 (30 minutes)*

Natural Language Processing (NLP) is a branch of artificial intelligence that extracts information from language. It has been applied to a wide range of fields 1, such as voice recognition, email categorization, social network analysis and others. In the field of software engineering, NLP has been adopted to extract key information from free-form text, to generate models from the analysis of text or to categorize code changes according to their commit messages 2.

In the present study, NLP has been applied to software code and its documentation in order to detect any defects that may be omitted by other techniques. NLP has helped in the identification of patterns of software code modifications [3, 4], to determine how it has evolved over time and to understand the complexity faced by developers (who have implemented new features and fixed faults).

The wide adoption of version control systems, such as github, to manage software code opened the access to different unstructured information. Each modification entry, in fact, is characterized by specific elements such as the creation date of a file, the date of the modification, the name of the developer who performed the change, the lines of code that were added or removed, and a message that explains the reasons for the change. According to the content of the message, it is possible to identify key terms that can be used during the classification of the entries.

Willing to define useful instruments to support developers in their daily activities, in this study we present its key terms and the lessons learned from the assessment of the change history of WLCG software available on github to the HEP community [5]. Adopting NLP techniques, we have cleaned the messages and extracted some key terms to categorize software problems and other code changes performed by developers like the integration of a third-party dependency or the control sequence introduced for a given service. By combining the information gathered from the WLCG software and its change history present in the github repository [5] with the one in already existing literature [6] we have built a code change dictionary. Finally, we have applied some Machine Learning (ML) techniques to investigate possible, unconventional connections between code changes and software defects.

### References

- 1 S. Falkenstine, A. Thornton, B. Meiners, "Natural Language Processing for Autonomous Identification of Impactful Changes to Specification Documents", in Proceedings of the 2020 AIAA/IEEE 39th Digital Avionics System Conference (DASC), doi: 10.1109/DASC50938.2020.9256611
- 2 F. Gilson and D. Weyns, "When Natural Language Processing Jumps into Collaborative Software Engineering", in Proceedings of the 2019 IEEE International Conference on Software Architecture Companion (ICSA-C), 2019. doi: 10.1109/ICSA-C.2019.00049.
- 3 C. Rosen, B. Grawi, and E. Shihab, "Commit guru: analytics and risk prediction of software commits", in Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering, pp. 966-969, ACM, 2015. doi: 10.1145/2786805.2803183
- [4] S. Majumder, J. Chakraborty, A. Agrawal, and T. Menzies, "Why Software Projects need Heroes (Lessons Learned from 1100+ Projects)", in cs.SE, 2020. arXiv:1904.09954v2
- [5] E. Ronchieri, Y. Yang, M. Canaparo, A. Costantini, D. C. Duma, D. Salomoni, "A new code change prediction dataset: a case study based on HEP software", under publication in Proceedings

of IEEE NSS MIC 2020

[6] R. Ferenc, P. Gyimesi, G. Gyimesi, Z. Toth and T. Gyimothy, "An automatically created novel bug dataset and its validation in bug prediction", in *The Journal of Systems & Software*, vol. 169, 2020. doi: 10.1016/j.jss.2020.110691

**Primary authors:** Dr COSTANTINI, Alessandro (INFN-CNAF); Dr SALOMONI, Davide (INFN); DUMA, Doina Cristina (INFN - CNAF); Dr RONCHIERI, Elisabetta (INFN CNAF); Mr CANAPARO, Marco (INFN); YANG, Yue (Department of Statistical Sciences, University of Bologna)

**Presenter:** Dr RONCHIERI, Elisabetta (INFN CNAF)

**Session Classification:** Physics & Engineering Session

**Track Classification:** Physics (including HEP) and Engineering Applications

Contribution ID: 58

Type: **Oral Presentation**

## **A new method for geomorphological studies and land-cover classification using Machine Learning techniques**

*Tuesday, 23 March 2021 16:00 (30 minutes)*

The processing of aerial high-resolution images is key for territorial mapping and change detection analysis in hydro-geomorphological high-risk areas. A new method has been developed in the context of "CLOSE (Close to the Earth)" project, resulting in a workflow based on open source MicMac photogrammetric suite and on High-Performance Computing. The workflow allowed to process a sequence of more than 1000 drone images captured along a reach belonging to the Basento River in Basilicata (Italy) during one single run.

The workflow optimisation aims to extract the orthophotomosaic, the point cloud and the Digital Surface Model (DSM) of selected areas. The high quality of the image details can be used for land-cover classification and extrapolating features useful to mitigate the hydro-geomorphological hazard, through machine learning models trained with satellite public data. Several Convolutional Neural Networks have been tested using progressively more complex layer sequences, data augmentation and callback techniques for training procedures. The results are given in terms of model accuracy and loss.

**Primary authors:** Dr MINIELLO, Giorgia (University of Bari "Aldo Moro", Italy); Dr LASALANDRA, Marco (University of Bari "Aldo Moro", Italy)

**Presenter:** Dr MINIELLO, Giorgia (University of Bari "Aldo Moro", Italy)

**Session Classification:** Earth & Environmental Sciences & Biodiversity Applications

**Track Classification:** Earth/Environmental Sciences & Biodiversity Applications

Contribution ID: 59

Type: **Oral Presentation**

## The European Open Science Cloud on a global stage

*Tuesday, 23 March 2021 09:10 (40 minutes)*

Over the past years, numerous policy makers from around the world have articulated a clear and consistent vision of global Open Science as a driver for enabling a new paradigm of transparent, data-driven science as well as accelerating innovation. In Europe, this vision is being realised through an ambitious programme under the heading of the European Open Science Cloud (EOSC). The EOSC will offer 1.7 million European researchers and 70 million professionals in science, technology, the humanities and social sciences a virtual environment with open and seamless services for storage, management, analysis and re-use of research data, across borders and scientific disciplines by federating existing scientific data infrastructures, currently dispersed across disciplines and the EU Member States. The research communities that EOSC aims to serve span the globe and so EOSC cannot operate in isolation but must find ways to be part of the broader, international open science movement. This presentation will give an update on the status of EOSC and the approaches being considered for how EOSC can engage at a global level.

**Presenter:** Dr JONES, Bob (CERN)**Session Classification:** Opening Ceremony & Keynote Speech I

Contribution ID: 60

Type: **Oral Presentation**

## **Constructing glocal monitoring system for safe and secure society**

*Wednesday, 24 March 2021 09:00 (40 minutes)*

In 2016, Tokai University initiated a project call “Constructing glocal monitoring system for safe and secure society”. “Glocal” is the coined word of “global” and “local”. The main concept of the project is to connect the global monitoring system using satellite observation with the local monitoring system using SNS for monitoring disasters and environmental changes. The project was approved by the Ministry of Education, Culture, Sports, Science and Technology(MEXT) of Japan as one of the Research Branding Project of Private Universities in Japan. Tokai University is receiving MODIS, VIIRS and AVHRR data at it’s ground stations one in Shonan Campus and another in Kyushu Campus. The data are automatically processed and archived in near real time. On the other hand, the authors have been operating Disaster Information Tweeting System(DITS) for disaster information collection. The authors are constructing the Glocal Monitoring System by connecting the satellite data system with DITS. The concept and progress of the Glocal Monitoring Project will be presented.

**Presenter:** Prof. CHO, Kohei (Tokai University)

**Session Classification:** Keynote Speech

Contribution ID: **61**

Type: **not specified**

## Opening Remarks

*Tuesday, 23 March 2021 09:00 (10 minutes)*

**Presenters:** Dr MATYSKA, Ludek (CESNET); LIN, Simon C. (ASGC); Prof. CHANG, Yuan-Hann (Institute of Physics, Academia Sinica)

**Session Classification:** Opening Ceremony & Keynote Speech I

Contribution ID: 62

Type: **Oral Presentation**

## Incident Detection

*Monday, 22 March 2021 09:00 (1h 30m)*

A typical situation security teams encounter is to find out whether they are possibly a victim of an ongoing security incident spreading in, or across infrastructures.

To decide whether your resources are affected by the incident you basically face two problems:

- How to share sensitive information, like Indicators of Compromise (IoCs).
- How to efficiently use the available information (IoCs).

In particular the problem of matching reported IoCs to the available data can be very time consuming. Starting from low level parsing of central syslog data of the monitored systems, to matching patterns in network flows need to be automated as far as possible.

The WLCG SOC WG activity addresses both challenges Resource Center (RC) security teams are facing in their daily life.

The framework developed there was presented at earlier Security Workshops at ISGC and at other conferences. After an introduction to the SOC, the focus will be on deployment of the framework and a demonstration how the SOC can help to share threat intelligence (for example) IoCs detected by a partner RC, and consequently give an answer to the question if your RC is also a victim to the same attack.

**Presenters:** Dr CROOKS, David (UKRI STFC); ROORDA, Jouke (Nikhef); Dr GABRIEL, Sven (Nikhef/EGI)

**Session Classification:** Security Workshop

Contribution ID: 63

Type: **Oral Presentation**

## Incident Handling

*Monday, 22 March 2021 11:00 (1h 30m)*

With the uptake of different virtualization technologies also in traditional data processing workflows the security landscape gets increasingly heterogeneous. Container technology allows users and user communities to easily ship complex data processing environments to the federated resources. While using containers adds a lot of flexibility for the resource usage, it also increases the possible attack surface of the infrastructure. The goal of this training session is to discuss selected aspects related to containers and present potential security threats. We will focus mainly on Docker and its typical use-cases. The main principles, however, will be applicable also for other container technologies. The workshop is not meant as an exhaustive training session covering every possible aspect of the technology. Its purpose is to point out some typical problems that are important to consider, some of which are inspired by real-world security incidents.

The workshop will be organized as a mixture of technical presentations, interleaved with shorter sessions where attendees will be able to practice the ongoing topics on a couple of hands-on exercises.

**Presenter:** Dr KOURIL, Daniel (Masaryk University)

**Session Classification:** Security Workshop

Contribution ID: 64

Type: **Oral Presentation**

## **ESCAPE, next generation management of exabytes of cross discipline scientific data**

*Wednesday, 24 March 2021 14:00 (20 minutes)*

The European-funded ESCAPE project will prototype a shared solution to computing challenges in the context of the European Open Science Cloud. It targets Astronomy and Particle Physics facilities and research infrastructures and focuses on developing solutions for handling Exabyte scale datasets.

The DIOS work package aims at delivering a Data Infrastructure for Open Science. Such an infrastructure would be a non HEP specific implementation of the data lake concept elaborated in the HSF Community White Paper and endorsed in the WLCG Strategy Document for HL-LHC.

The science projects in ESCAPE are in different phases of evolution. While HL-LHC can leverage 15 years of experience of distributed computing in WLCG, other sciences are building now their computing models. This contribution describes the architecture of a shared ecosystem of services fulfilling the needs in terms of data organisation, management and access for the ESCAPE community. The backbone of such a data lake will consist of several storage services operated by the partner institutes and connected through reliable networks. Data management and organisation will be orchestrated through Rucio. A layer of caching and latency hiding services, supporting various access protocols will serve the data to heterogeneous facilities, from conventional Grid sites to HPC centres and Cloud providers. The authentication and authorisation system will be based on tokens.

For the success of the project, DIOS will integrate open source solutions which demonstrated reliability and scalability as at the multi petabyte scale. Such services will be configured, deployed and complemented to cover the use cases of the ESCAPE sciences which will be further developed during the project.

**Presenter:** Dr DI MARIA, Riccardo (CERN)

**Session Classification:** Data Management & Big Data Session

Contribution ID: 65

Type: **Oral Presentation**

## **General introduction to integrative modelling and HADDOCK**

*Monday, 22 March 2021 09:00 (1h 30m)*

**Presenter:** Dr BONVIN, Alexandre M.J.J. (Utrecht University)

**Session Classification:** Workshop on Integrative modelling of biomolecular complexes using HADDOCK

Contribution ID: 66

Type: **Oral Presentation**

## **Modelling antibody-antigen complexes with HADDOCK**

*Monday, 22 March 2021 11:00 (1h 30m)*

**Presenter:** Dr BONVIN, Alexandre M.J.J. (Utrecht University)

**Session Classification:** Workshop on Integrative modelling of biomolecular complexes using HADDOCK

Contribution ID: 67

Type: **Oral Presentation**

## **HADDOCK advanced topics**

*Monday, 22 March 2021 13:30 (45 minutes)*

**Presenter:** Dr BONVIN, Alexandre M.J.J. (Utrecht University)

**Session Classification:** Workshop on Integrative modelling of biomolecular complexes using HADDOCK

Contribution ID: 68

Type: **Oral Presentation**

## **LightDock: An artificial intelligence-based integrative docking platform**

*Monday, 22 March 2021 14:30 (30 minutes)*

**Presenter:** Dr JIMÉNEZ-GARCÍA, Brian (University of Utrecht)

**Session Classification:** Workshop on Integrative modelling of biomolecular complexes using HADDOCK

Contribution ID: 69

Type: **Oral Presentation**

## **Integrative modelling of membrane complexes with LightDock and HADDOCK**

*Monday, 22 March 2021 15:00 (1 hour)*

**Presenter:** Dr BONVIN, Alexandre M.J.J. (Utrecht University)

**Session Classification:** Workshop on Integrative modelling of biomolecular complexes using HADDOCK

Contribution ID: 70

Type: **not specified**

## Question and answer session and wrap up

*Monday, 22 March 2021 17:00 (30 minutes)*

**Presenter:** Dr BONVIN, Alexandre M.J.J. (Utrecht University)

**Session Classification:** Workshop on Integrative modelling of biomolecular complexes using HADDOCK

Contribution ID: 71

Type: **Oral Presentation**

## Introduction

*Wednesday, 24 March 2021 10:30 (20 minutes)*

**Session Classification:** Deeper Understanding of Natural Disasters: Joint DMCC, UMD & Environmental Computing Workshop

Contribution ID: 72

Type: **Oral Presentation**

## **Regional Collaboration on Disaster Mitigation**

*Wednesday, 24 March 2021 10:50 (10 minutes)*

**Presenter:** YEN, Eric (ASGC)

**Session Classification:** Deeper Understanding of Natural Disasters: Joint DMCC, UMD & Environmental Computing Workshop

Contribution ID: 73

Type: **Oral Presentation**

## **Numerical simulation on extreme weather, air pollution and environmental changes**

*Wednesday, 24 March 2021 11:00 (1 hour)*

**Presenter:** Dr LIN, Chuan-Yao (Academia Sinica)

**Session Classification:** Deeper Understanding of Natural Disasters: Joint DMCC, UMD & Environmental Computing Workshop

Contribution ID: 74

Type: **Oral Presentation**

## **Disaster Volunteer Activities for 2021 Fukushima-ken Oki Earthquake**

*Friday, 26 March 2021 13:30 (20 minutes)*

**Presenter:** Prof. SATO, Dai (Tohoku University)

**Session Classification:** Deeper Understanding of Natural Disasters: Joint DMCC, UMD & Environmental Computing Workshop

Contribution ID: 75

Type: **Oral Presentation**

## **Updates from the Bavarian electronic Pollen Information Network ePIN**

*Wednesday, 24 March 2021 13:20 (15 minutes)*

**Presenter:** BUTERS, Jeroen (TUM)

**Session Classification:** Deeper Understanding of Natural Disasters: Joint DMCC, UMD & Environmental Computing Workshop

Contribution ID: 76

Type: **Oral Presentation**

## Development and Management of IT Infrastructure for a Citizen Science Project on Climate Research and Science Communication

*Wednesday, 24 March 2021 13:35 (25 minutes)*

Citizen Science is nowadays used in various fields, such as history, health, natural disaster management and ecology. BAYSICS (Bavarian Citizen Science Information Platform for Climate Research and Science Communication) is a multidisciplinary project from Bavaria, in the south-eastern part of Germany. Leibniz Supercomputing Centre (LRZ) is responsible for the development and maintenance of the IT infrastructure for the project. There are three main online tools offered to citizens: Website, web app and mobile app, the latter being implemented in the form of Progressive Web App based on web technologies.

On the project's web app / mobile app, citizens can submit observations in four main research areas: animals, plants, allergenic species, tree lines. Data collection started in spring 2020, and so far more than 630 observations have been submitted (as of the end of February 2021). The observation data have relevance to the Sustainable Development Goals (SDGs), such as the SDGs 11 and 15. The data can also be used to validate remote sensing information (Batsaikhan et al, 2020).

The observation data can be downloaded in Excel/CSV format along with the metadata. Application Programming Interface (API) is also available for interaction with an external database. In future, in order to process increasing amounts of data collected, high performance data analytics methods and tools can be useful. Artificial Intelligence (AI) can be applied to develop trust metrics for the observation data.

**Presenter:** BATSAIKHAN, Anudari (LRZ)

**Session Classification:** Deeper Understanding of Natural Disasters: Joint DMCC, UMD & Environmental Computing Workshop

Contribution ID: 77

Type: **Oral Presentation**

## **Forging deep learning-based and physics-based approaches for computational drug discovery**

*Tuesday, 23 March 2021 10:30 (30 minutes)*

**Presenter:** Dr LIN, Jung-Hsin (Academia Sinica)

**Session Classification:** Biomedicine & Life Science

Contribution ID: 78

Type: **Oral Presentation**

## Using Dynamic Data-Driven Cyberinfrastructure for Next Generation Disaster Intelligence

Wildland fires and related hazards are increasing globally. A common observation across these large events is that fire behavior is changing to be more destructive, making applied fire research more important and time-critical. Significant improvements towards modeling the extent and dynamics of the evolving plethora of fire-related environmental hazards and their socio-economic and human impacts can be made through intelligent integration of modern data and computing technologies with techniques for data management, machine learning, and fire modeling. However, there are still challenges and opportunities in integration of the scientific discoveries and data-driven methods for hazards with the advances in technology and computing in a way that provides and enables different modalities of sensing and computing. The WIFIRE cyberinfrastructure took the first steps to tackle this problem with a goal to create an integrated system, data and visualization services, and workflows for wildfire monitoring, simulation, and response. Today, WIFIRE provides an end-to-end management infrastructure from data sensing and collection to artificial intelligence and modeling efforts using a continuum of computing methods that integrate edge, cloud, and high-performance computing. Through this cyberinfrastructure, the WIFIRE project provides data-driven knowledge for a wide range of public and private sector users enabling scientific, municipal, and educational use. This talk will review some of our recent work on building this dynamic data-driven cyberinfrastructure and impactful application solution architectures that showcase the integration of a variety of existing technologies and collaborative expertise.

İlkay Altıntaş, Ph.D., University of California San Diego: Chief Data Science Officer, San Diego Supercomputer Center / Founder & Director, WorDS Center of Excellence and WIFIRE Lab / Founding Fellow, Halicioglu Data Science Institute (<https://words.sdsc.edu/ilkay/>)

**Presenter:** Prof. ALTINTAS, İlkay (San Diego Supercomputer Center)

Contribution ID: 79

Type: **Oral Presentation**

## **Tsunami and Storm Surge**

*Thursday, 25 March 2021 10:30 (1h 30m)*

**Presenter:** Prof. WU, Tso-Ren (National Central University)

**Session Classification:** Deeper Understanding of Natural Disasters: Joint DMCC, UMD & Environmental Computing Workshop

Contribution ID: 80

Type: **Oral Presentation**

## **LEXIS Earthquake and Tsunami Pilot**

*Thursday, 25 March 2021 13:45 (45 minutes)*

**Presenter:** Dr GOUBIER, Thierry (CEA List)

**Session Classification:** Deeper Understanding of Natural Disasters: Joint DMCC, UMD & Environmental Computing Workshop

Contribution ID: **81**

Type: **Oral Presentation**

## **LEXIS Weather and Climate Pilot**

*Thursday, 25 March 2021 13:00 (45 minutes)*

**Presenter:** Dr PARODI, Antonio (CIMA/IT)

**Session Classification:** Deeper Understanding of Natural Disasters: Joint DMCC, UMD & Environmental Computing Workshop

Contribution ID: 82

Type: **Oral Presentation**

## Case study on Flood

*Friday, 26 March 2021 09:00 (45 minutes)*

**Presenter:** Prof. LIEW, JN (UKM)

**Session Classification:** Deeper Understanding of Natural Disasters: Joint DMCC, UMD & Environmental Computing Workshop

Contribution ID: 83

Type: **Oral Presentation**

## **Thailand's fires and smoke haze: current capabilities, lessons learned and needs**

*Friday, 26 March 2021 09:45 (45 minutes)*

**Presenter:** Dr TANPIPAT, Veerachai (HII)

**Session Classification:** Deeper Understanding of Natural Disasters: Joint DMCC, UMD & Environmental Computing Workshop

Contribution ID: 84

Type: **Oral Presentation**

## **Case study and status report from Phillipines**

*Friday, 26 March 2021 11:00 (20 minutes)*

**Primary author:** Mr ROEL, Cruz (ASTI)

**Presenter:** Mr ROEL, Cruz (ASTI)

**Session Classification:** Deeper Understanding of Natural Disasters: Joint DMCC, UMD & Environmental Computing Workshop

Contribution ID: 85

Type: **Oral Presentation**

## **Case study and status report from Viet Nam**

*Friday, 26 March 2021 11:20 (20 minutes)*

**Presenter:** Mrs THI NGUYEN, Thanh (Vietnam Institute of Meteorology Hydrology and Climate Change)

**Session Classification:** Deeper Understanding of Natural Disasters: Joint DMCC, UMD & Environmental Computing Workshop

Contribution ID: 86

Type: **Oral Presentation**

## **Case study and status report from Indonesia**

*Friday, 26 March 2021 11:40 (20 minutes)*

**Session Classification:** Deeper Understanding of Natural Disasters: Joint DMCC, UMD & Environmental Computing Workshop

Contribution ID: 87

Type: **Oral Presentation**

## **The Case Study: Disaster Events in Myanmar**

*Friday, 26 March 2021 12:00 (20 minutes)*

**Presenter:** Dr SEIN, Myint Myint (UCSY)

**Session Classification:** Deeper Understanding of Natural Disasters: Joint DMCC, UMD & Environmental Computing Workshop

Contribution ID: 89

Type: **Oral Presentation**

## **Precision agriculture and regional collaborations (CRADR)**

*Wednesday, 24 March 2021 13:00 (20 minutes)*

**Presenter:** Dr ADINARAYANA, J. (IITB)

**Session Classification:** Deeper Understanding of Natural Disasters: Joint DMCC, UMD & Environmental Computing Workshop

Contribution ID: **90**

Type: **Oral Presentation**

## **The EGI Advanced Computing for EOSC (EGI-ACE) project**

*Friday, 26 March 2021 14:15 (25 minutes)*

**Presenter:** Dr SIPOS, Gergely (EGI)

**Session Classification:** Deeper Understanding of Natural Disasters: Joint DMCC, UMD & Environmental Computing Workshop

Contribution ID: 91

Type: **Oral Presentation**

## **Workflows in Environmental Research with Containers**

*Friday, 26 March 2021 13:50 (25 minutes)*

**Presenter:** PAUW, Viktoria (LRZ)

**Session Classification:** Deeper Understanding of Natural Disasters: Joint DMCC, UMD & Environmental Computing Workshop

Contribution ID: 92

Type: **not specified**

## Discussion and Wrap up

*Friday, 26 March 2021 14:50 (10 minutes)*

**Session Classification:** Deeper Understanding of Natural Disasters: Joint DMCC, UMD & Environmental Computing Workshop

Contribution ID: 93

Type: **Oral Presentation**

# Machine Learning in High-Energy Physics

*Thursday, 25 March 2021 09:00 (40 minutes)*

**Presenter:** Prof. BONACORSI, Daniele (University of Bologna)

**Session Classification:** Keynote Speech

Contribution ID: 94

Type: **Oral Presentation**

## **Scientific computing in the coming decade: trends and collaboration opportunities for Europe and Asia**

*Thursday, 25 March 2021 09:40 (40 minutes)*

The presentation will provide a perspective on how distributed computing has been instrumental to make groundbreaking scientific discoveries possible, and how the opening of computing infrastructures at international level has been effective in delivering unprecedented compute capacity and advance data analytics tools to international research collaborations.

The presentation will provide examples of the enormous scientific impact produced by the international collaboration of cyber infrastructures in all continents, and will explain the federated organisational model adopted by European countries to leverage national ICT investments and mobilise them through the EGI Federation, and the paradigm shifts that need to be embraced to effectively support the digital needs of science in the coming decade.

The presentation will offer an overview of the present and future technical and organisational challenges of data-driven research in various scientific domains and the opportunities of collaboration that these challenges offer to Europe and Asia.

**Presenter:** Dr FERRARI, Tiziana (EGI Foundation)

**Session Classification:** Keynote Speech

Contribution ID: 95

Type: **Oral Presentation**

## **Overview and Outlook – Environmental Computing at LRZ**

*Friday, 26 March 2021 14:40 (10 minutes)*

**Presenter:** KURTZ, Wolfgang (LRZ)

**Session Classification:** Deeper Understanding of Natural Disasters: Joint DMCC, UMD & Environmental Computing Workshop

Contribution ID: 96

Type: **Oral Presentation**

## **Integrative modelling of membrane complexes with LightDock and HADDOCK (continued)**

*Monday, 22 March 2021 16:15 (45 minutes)*

**Presenter:** Dr BONVIN, Alexandre M.J.J. (Utrecht University)

**Session Classification:** Workshop on Integrative modelling of biomolecular complexes using HADDOCK

Contribution ID: 97

Type: **Oral Presentation**

## **eScience Report from TW**

*Tuesday, 23 March 2021 10:30 (15 minutes)*

**Presenter:** Mr LEE, Felix (ASGC)

**Session Classification:** e-Science Activities in Asia Pacific Session

Contribution ID: 98

Type: **Oral Presentation**

## **eScience Report from KR**

*Tuesday, 23 March 2021 10:45 (15 minutes)*

**Presenter:** Dr AHN, Sang Un (Korea Institute of Science and Technology Information)

**Session Classification:** e-Science Activities in Asia Pacific Session

Contribution ID: 99

Type: **Oral Presentation**

## **eScience Report from SG**

*Tuesday, 23 March 2021 11:00 (15 minutes)*

**Presenter:** Dr TAN, Tinwee (National Supercomputing Center of Singapore)

**Session Classification:** e-Science Activities in Asia Pacific Session

Contribution ID: **100**

Type: **Oral Presentation**

## **eScience Report from CN**

*Tuesday, 23 March 2021 11:15 (15 minutes)*

**Presenter:** Dr CHEN, Gang (Institute Of High Energy Physics)

**Session Classification:** e-Science Activities in Asia Pacific Session

Contribution ID: **101**

Type: **Oral Presentation**

## **eScience Report from TH**

*Tuesday, 23 March 2021 11:30 (15 minutes)*

**Presenter:** Dr VORAKULPIPAT, Chalee (NECTEC)

**Session Classification:** e-Science Activities in Asia Pacific Session

Contribution ID: **102**

Type: **not specified**

## **Discussion**

*Tuesday, 23 March 2021 11:45 (15 minutes)*

**Session Classification:** e-Science Activities in Asia Pacific Session

Contribution ID: **103**

Type: **Oral Presentation**

## **eScience Report from AU**

*Wednesday, 24 March 2021 10:30 (15 minutes)*

**Presenter:** Dr WALSH, Carmel (Australian Research Data Commons, ARDC)

**Session Classification:** e-Science Activities in Asia Pacific Session

Contribution ID: **104**

Type: **Oral Presentation**

## **eScience Report from JP**

*Wednesday, 24 March 2021 11:00 (15 minutes)*

**Presenter:** Prof. AIDA, Kento (National Institute of Informatics)

**Session Classification:** e-Science Activities in Asia Pacific Session

Contribution ID: **106**

Type: **Oral Presentation**

## **eScience Report from ID**

*Wednesday, 24 March 2021 10:45 (15 minutes)*

**Presenter:** Dr SUHARDIMAN, Basuki (ITB)

**Session Classification:** e-Science Activities in Asia Pacific Session

Contribution ID: **107**

Type: **not specified**

## **Discussion & Wrap-up**

*Wednesday, 24 March 2021 11:15 (30 minutes)*

**Session Classification:** e-Science Activities in Asia Pacific Session

Contribution ID: **108**

Type: **Oral Presentation**

## **APGridPMA Update**

*Tuesday, 23 March 2021 10:30 (15 minutes)*

**Presenter:** YEN, Eric (ASGC)

**Session Classification:** APGridPMA Meeting

Contribution ID: **109**

Type: **Oral Presentation**

## **IAM in Asia**

*Tuesday, 23 March 2021 10:45 (30 minutes)*

**Presenter:** Dr SCHOFIELD, Brook

**Session Classification:** APGridPMA Meeting

Contribution ID: **110**

Type: **Oral Presentation**

## **EUGridPMA and AAI in EOSC**

*Tuesday, 23 March 2021 11:15 (45 minutes)*

**Presenter:** Dr GROEP, David (Nikhef)

**Session Classification:** APGridPMA Meeting

Contribution ID: 111

Type: **Oral Presentation**

## **TAGPMA and Token-based AAI**

*Tuesday, 23 March 2021 13:00 (45 minutes)*

**Presenter:** SIMMEL, Derek

**Session Classification:** APGridPMA Meeting

Contribution ID: 112

Type: **Oral Presentation**

## Running Identity Federation Services on Containers and K8s

*Tuesday, 23 March 2021 13:45 (20 minutes)*

Since 2018, SIFULAN Malaysian Access Federation (SIFULAN) has been operating in production mode by using Virtualization Machine (VM) to run several identity federation services. As the federation grows, SIFULAN plans to offer IdP-as-a-service to the existing and potential future members as an additional service. However, the current infrastructure setup has some limitations to support the plan as multi-tenant services were not part of the initial infrastructure design. Hence, SIFULAN migrated its federation infrastructure from a VM based to a container-based infrastructure and use Kubernetes (K8s) as the orchestration manager for the containers. This talk will share how SIFULAN migrated and runs Identity Federation services on containers/K8s and provides some ideas on how a multi-tenant IdP could be run on containers/K8s.

**Presenter:** Mr SJAUGI, Muhammad Farhan (Perdana University)

**Session Classification:** APGridPMA Meeting

Contribution ID: 113

Type: **Oral Presentation**

## **Token-based AuthN/Z for command-line applications**

*Tuesday, 23 March 2021 14:05 (20 minutes)*

**Presenter:** Dr SAKANE, Eisaku (National Institute of Informatics)

**Session Classification:** APGridPMA Meeting

Contribution ID: **114**

Type: **not specified**

## **Discussion & AOB**

*Tuesday, 23 March 2021 14:25 (5 minutes)*

**Session Classification:** APGridPMA Meeting

Contribution ID: 115

Type: **Oral Presentation**

## **Optimization of the Common Wealth Planning and Sharing based on Knowledge Value Transformation Theory**

*Thursday, 25 March 2021 14:00 (30 minutes)*

Since the advent of ICT, information and communication resources become the common wealth and the method dealing with the resource planning and sharing becomes the fundament of the ICT services. For the common wealth planning and sharing, I have carried out the paradigm of Organic Ecology, viz. HyQVIS (Hyper Quality and Value of ICT Systematics), which is inspired by the EU FP3-ESPRIT Project HyQIS. Within HyQVIS, the governmental authority and the participants are organic elements of the ecosystem. Based on the systematics, the Knowledge Value Transformation Theory is proposed.

According to the theory, the task of common wealth planning and sharing is to optimize the behavior of the organic elements and the ecosystem as a whole with respect to appropriate laws or de facto consensus. And, the whole system is a transformer from knowledge on semantic level to the value on the pragmatic level. For the cases up-to-date, three kinds of most significant optimization techniques, viz. Random Optimization(RO), Simulated Annealing(SA) and Genetic Algorithm(GA) are adopted. Since the techniques comes from different scientific disciplines, they have different moral background, so they represent different strategies for resource planning and sharing. We have aggregated the techniques to optimize the organic behaviors that the ecosystem can take.

In recent years, one of the most important issues of resource sharing and utilization is the spectrum splitting and sharing of mobile communication. Optimally splitting spectrum can lead the mobile operators to increase communication capacity and allow users to take advantage from Quality of Services(QoS). 5G/6G mobile technique can accomplish enhanced devices and communication capabilities with higher throughput, higher efficiency and lower latency for mobile services. This study starts from finding available 5G spectrum and the regulation law in Taiwan/ROC, and estimate the values of the bands in the available spectrum. Then, the study copes with the market share of the operators, the major investing factors as the financial constraints of operators and the construction cost as input variables, and then simulate different bidding rounds in the auction process and combinations of segments to maximize the gain of the 5G ecosystem. The result proves that spectrum splitting and auction is an optimization process through the dynamic organic behaviors of the operators in the auction process. Further, the result reveals the differences of two resource sharing strategies based on different backgrounds, i.e. SA from Physics, which seeks for the equilibrium of the system, and GA from biology, that emphasizes mutation and elitism. The 2nd case deals with the collaboration framework for the API and Platform Economy. At first, a framework for the API and Platform Economy is presented. Further, we analysed the complexity of different topologies of the ecosystem. The result shows the federation of multiple centralized groups, e.g. Grids, is most efficient as well as less of complexity, more robust for security and thus federated computing becomes the next edge of the collaboration and information resource sharing.

The theory and the systematic approach presented in this article can serve as a well-formed referral guideline for planning and sharing of common wealth in the regions of different cultures.

**Presenter:** Prof. CHIANG, Johannes K. (NCCU)

**Session Classification:** Infrastructure Clouds and Virtualisation Session

Contribution ID: **116**

Type: **not specified**

## **CA Self-Audit Reports**

*Tuesday, 23 March 2021 07:00 (2 hours)*

**Session Classification:** APGridPMA Meeting

Contribution ID: 117

Type: **not specified**

## Using Dynamic Data-Driven Cyberinfrastructure for Next Generation Disaster Intelligence

*Wednesday, 24 March 2021 14:00 (1 hour)*

**Presenter:** Prof. ALTINTAS, Ilkay (San Diego Supercomputer Center)

**Session Classification:** Deeper Understanding of Natural Disasters: Joint DMCC, UMD & Environmental Computing Workshop