International Symposium on Grids & Clouds 2019 (ISGC 2019)

Sunday 31 March 2019 - Friday 05 April 2019

Academia Sinica

Book of Abstracts

Contents

Earth & Environmental Sciences & Biodiversity Application / 0

Ensuring Data Readiness within Climateprediction.net

Author: David Wallom[1]

Co-authors: Mamun Rashid [2]; Peter Uhe [3]; Sarah Sparrow [1]; Sihan Li [1]

[1] *University of Oxford*

[2] *Kings College London*

[3] *Univeristy of Bristol*

Corresponding Author: david.wallom@oerc.ox.ac.uk

Climate change is both one of the grand scientific challenges of the moment and also one of the most politically charged. As such it is essential that research in this area operates in as transparent a manner as possible. This takes on extra relevance when considering studies that either lead towards political or social impact such as those feeding into the IPCC special report on 1.5degrees or impact studies of extreme event attribution, where some parties are even now trying to approach 'blaming' other parties for climate change.

Within the Climateprediction.net program we have long considered it essential, due to the nature of how we obtain our simulation results, to make our data open access. Until now this has mostly been through personal contact with the co-ordinating team located at the University of Oxford by any researcher that wishes to use the data. Whilst this is still an effective method, it is clear that we must disseminate more than just the final raw data, but also for that data to be truly effective we must include the experiment design so that users are able to fully understood how the data was created. As such the full data chain must be understood and accessible, from the persistent capture of experimental design, simulation management, steps in data readiness for results and finally archiving and publishing.

We have developed a number of these processes independently but are now in a position where we can describe the interconnection of these such that we can show end to end provenance on all studies now undertaken using CPDN;

1. Management of experiment definition Experiments are defined within CPDN using a number of configuration files, both model specific and XML based which manage all aspects of the experiment. They are managed via a repository system as well as their ingestion into the main database of the CPDN experimental setup as well as using the Trello project management tool. This allows remote submitters of work to interact with the core team within Oxford as well as keep all communication about the batch between the researchers stored for future reference. With these combination of steps and systems we have established the first steps in the provenance chain.
2. Simulation Management Once the workunits are submitted they are subsequently picked up by BOINC clients registered with the project and executed. As these are long running applications then to ensure that computational resource is not wasted the capability of check pointing and uploading intermediate results from the workunit back to the project servers is used extensively. The presence of these results and all the information about the systems which processed each workunit are kept within the main CPDN project database. This allows us to investigate any possibility of machine dependant results etc as well as understanding if there are any platform specific performance issues. The data that is returned is sent using the community standard, netcdf. This has the advantage of being a self-describing data format, where the content, diagnostics etc. that are contained within are all described within the headers of the files themselves.
3. Readiness of Results Data Once we have reached the end of the workunit then the results are all uploaded to the specified upload server which will curate and manage the results data for that workunit. Alongside this the client communicates with the core system scheduler to notify it that a task has completed and that results have (normally) completed successfully. The results data is presented to the scientist as a large number of individual files, one per workunit. As such utilising these large number of files could be severely challenging it has been necessary to develop a set of CPDN specific tools to allow simple interaction by the data using scientist with their data. These tools are managed within the Github software management platform. The processed data is again created using the netcdf data standard and at this point that the data is ready for analysis and subsequent publication of the science results normally occurs.

4. Archive and publication With many publishers as well as funding agencies it is necessary to curate your open data for a period beyond its last use of publication date. As such we have also determined that to increase the possible utilisation of the dataset that it would be welcome to make the data available using community repositories and descriptors which the research who may reuse the data greater insight into how it has been created. As such working with repositories such as the UK NERC Centre for Environmental Analysis repository and publishing venues such as Elsevier Data in Brief or Nature Scientific Data. This allows greater depth in the description of the available datasets as well as linkage to scientific results that have been already generated using them.

Summary:

Climateprediction.net is the largest climate simulation facility available to the general research community currently. As such it has partners in eleven different countries and has performed over 200 million years of simulation of both the whole earth system using HadCM3 global coupled model and using Atmosphere only regional models within the Wether@Home sub experiments. Generating in excess of 0.5PB of open access climate data we have needed to create a robust pipeline to ensure that provenance of the data is maintained. This has become especially necessary within the extreme event attribution experiments of Weather@Home projects where we can use up to 120k member ensembles to generate robust attribution statements. This pipeline must operate from experiment definition through to final archiving of the data when the main experiment has completed it utilisation and at all times we must be able to return to the previous step to understand where a particular piece of data has come from.

Physics & Engineering Application / 2

Test QUDA with AMD GPUs on ROCm Platform

Author: YUJIANG BI[1]

[1] *Institution of High Energy Physics, CAS*

Corresponding Author: biyujiang@ihep.ac.cn

Lattice QCD is a no-perturbative approach to solving the Quantum Chromodynamics (QCD) theory of quarks and gluons, using Monte Carlo method and technique analogous to statistic physics. QUDA, which is written mainly for NVIDIA GPUs in CUDA, is an open source library for Lattice QCD aimed to accelerating computing. And ROCm is a new open GPU Computing platform, compatible with AMD GPUs, as well as NVIDIAs. We are porting QUDA from CUDA to ROCm platform, and testing its correctness and performance, to examining the feasibility of study Lattice QCD on AMD GPUs via QUDA, and providing a reference for further Lattice QCD software deployment.

Data Management & Big Data / 3

ESCAPE: a multi-science data infrastructure for the 2020s

Author: Simone Campana[1]

[1] *CERN*

Corresponding Authors: patrick.fuhrmann@desy.de, simone.campana@cern.ch

The ESCAPE project aims at delivering 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 large sets of data. One key aspect of ESCAPE is prototyping and implementing a shared system for FAIR data management and, in this

contribution, we will present the effort in building such data infrastructure for open science. The infrastructure will be based on the idea of the WLCG Data Lake proposal presented as evolution of the facilities and middleware in preparation for HL-LHC. It will generalize however the components to cover the use cases of other data intensive sciences on the same physical facilities. We will describe how different work packages of the project will evaluate and prototype various components of the Data Lake architecture, such as a content delivering and caching service, storage and storage orchestration, data transfer services and access to compute resources. Dedicated effort will focus also on network R&D and the evolution of the authentication, authorization and identity management layer. The project will run for three years and intends to integrate adiabatically new solutions in the Open Science Cloud by the end date of the project.

VRE / 4

Implementation of a small-scale desktop grid computing infrastructure in a commercial domain

Author: Andy Bowery[1]

Co-author: David Wallom [1]

[1] *University of Oxford*

Corresponding Author: andy.bowery@oerc.ox.ac.uk

Distributed computing in the context of desktop grid computing, has been successfully utilized for a number of years in academia for a range of citizen science projects spread across a variety of different scientific domains. The Berkeley Open Infrastructure for Network Computing (BOINC) software is used as the framework of choice for these projects. BOINC is an L-GPL licensed software that has been specifically designed for distributed computing. To-date however, there has not been a single successful application of this distributed computing software within the commercial domain.

In this work we describe the successfully design and deployment of a small-scale desktop grid computing infrastructure within a commercial domain with the goal of building an infrastructure that would act as a test project to prove the viability of the potential construction of a much larger, production, desktop grid computing environment within the commercial domain.

In the commercial domain the specification, purchasing and setup of a new HPC system can take a significant amount of time, this acts as a limit to the number of computations that can be run at any one time and forces the reduction of the resolution of those computations through the contention of the limited resources. Desktop grid computing offers a low cost, quick to set up, alternative to obtaining a new HPC system. Desktop grid computing utilizes existing desktop computers when they are not being actively used, computers that have not been specifically obtained for the purpose of providing dedicated computational resources, but are nevertheless capable.

The test desktop grid computing infrastructure that was built in this work was built to prove that additional computational resources could be provided to augment an already overloaded in-house HPC system, providing an additional resource that could be used for computing and analysis. The desktop grid would enable the utilization of pre-existing in-house desktop computing resources more efficiently and would encompass minimal outlay for new infrastructure.

In this work we set up a small-scale test computing grid and implemented two different applications across the infrastructure. In the first application, a large number of simulations were run to perform a parameter sweep with a range of different parameters, this enabled the response of the application to be plotted over the parameter ranges, and provided the optimum solution to the problem studied. These simulations were of sufficient length such that they would complete within an overnight time window and would provide results in a timely fashion for the start of the next business day. In the second application, the test desktop grid was used to demonstrate that the overflow of CAD analysis computational jobs could be run in a timely fashion on the infrastructure.

In addition, we examined the additional energy that would be consumed and the heat that would be produced by the running the computational work on the desktop computing resources in a typ-

ical office environment. This was in order to quantify both the change in temperature produced and the additional cost of the energy consumed. The results from this analysis show that the raise in temperature produced was minimal and the additional cost of the energy consumed was also minimal.

In this work we proved that a test desktop grid computing infrastructure in a commercial domain could produce useful, timely results on a representation scale that with further scaling would produce commercially useful results that would represent a significantly saving of money compared to the cost of the equivalent computational time utilizing either HPC or cloud computing resources. We have shown that such a setup can have a significant business-relevant impact at a very low cost for the additional infrastructure obtained and for the energy consumed.

Networking, Security, Infrastructure & Operations / **5**

Applications of SDN in IHEP network environment

Author: Zhihui Sun[1]

Co-authors: SHAN ZENG [1]; Tao Cui [1]

[1] *IHEP*

Corresponding Author: sunzh@ihep.ac.cn

Software Defined Network(SDN) is a flexible and programmable network architecture, the controller of SDN uses the south API(openflow or netconf) to deploy the network policies into the network devices, and also provides the north API for the use-defined applications. There are two scenarios in IHEP network environment using SDN technologies and architecture. For the new generation of IHEP campus network, the users and devices access control system developed in-house is using as a north application of SDN controller to launch the users access policies to controller based on the network access procedures. Another scenario is the network security application, we are using the services chain function for SDN to separate the different application traffic into different security devices to do the traffic protection and also the white list traffic will be by-pass by the security devices to increase the network performance.

Infrastructure Clouds and Virtualisation / **6**

Building an elastic batch system with private and public clouds

Author: Wataru Takase[1]

Co-authors: Koichi Murakami [1]; Takashi Sasaki [1]; Tomoaki Nakamura [1]

[1] *KEK*

Corresponding Author: wataru.takase@kek.jp

The Computing Research Center at High Energy Accelerator Research Organization (KEK) provides a Linux cluster consisting of 350 physical servers with 10000 CPU cores for the scientific data analysis and numerical simulation. All of the computing resources are shared by using a batch system among the various projects supporting at KEK. We have adopted IBM Spectrum LSF as the batch system for many years, and users in KEK are well familiar with the system. However, when the computing cluster is becoming congested by the lack of resource, user jobs make a long stay in a job queue. As a result, the turnaround time often becomes longer for each job to be finished. Furthermore, we have to consider the different requirements for processing environments depending on users/groups.

Providing flexible resources both in terms of computing power and environment, we have investigated the possibility of Cloud computing technology. IBM Spectrum LSF has an optional functionality so-called Resource Connector which enables us to utilize additional computing resources in external cloud providers, such as Amazon Web Service (AWS), Microsoft Azure, OpenStack. By using that functionality, we have succeeded to integrate our batch system with on-premise OpenStack cloud and Amazon Web Service by the collaboration with National Institute of Informatics (NII), Japan.

Users jobs submitted into the dedicated job queue are going to be dispatched to dynamically provisioned instances at on-premise OpenStack, AWS, as well as static local computing node. Any kind of job processing environments can be utilized by choosing a different virtual machine according to the user's request. The OpenStack based private cloud has been integrated with our LDAP service and GPFS (General Parallel File System) for user authentication and data sharing on our batch system, respectively. Amazon Simple Storage Service is utilized for the data exchange between KEK and AWS. Both the physical servers on the batch system at KEK and provisioned instances on AWS mount the S3 bucket via FUSE for transparent data access. We conducted some performance tests on our hybrid batch system for investigation of the scalability and succeeded to execute a Deep Learning job using 3500 cores on AWS.

In this talk, we would like to present the detailed configuration of the hybrid system and some results of the performance tests.

Supercomputing, High Throughput Computing, Accelerator Technologies, and their Integration / 7

Enhanced Utilization and Allocation of Distributed Computing Resources

Author: Nam Beng Tan[1]

[1] *Nanyang Polytechnic*

Corresponding Author: tan_nam_beng@nyp.edu.sg

This paper describes an effective mean of determining the optimum resource for job execution in a distributed environment that will further improve the performance of job computation which Grid Computing offers.

Grid computing systems comprise of several resource nodes in a networked system. Each resource node may further comprise of a single machine or several machines forming another network. By utilizing the computing power of all these machines in this grid computing system, calculations that usually require super computers may be realized by the combined computing capability of all the machines in the system.

Presently, grid computing has become one of the leading ways for the sharing of resources and computing power over the internet. Several resource nodes of a grid computing system may be able to utilize each other's spare or idle computing power. An example of such sharing is when a resource node is located in a differing time zone, say at night where the computing power of the machines in that node are idle. Another resource node currently requiring extra computing power may be able to send jobs to this idle node to speed up the calculations or even to free up computing power in its own node for more crucial applications. Furthermore, present methods for assigning jobs to nodes are mainly based on a round robin queuing system that is disadvantageously arbitrary and takes no consideration of the characteristics of the idle or available resource nodes.

Our proposed solution thus provides a method and system for monitoring and managing grid resources in a grid computing environment with a plurality of resource nodes and at least one resource broker/supervisor node. The resource broker serves to retrieve information stored in a attribute repository about the resource nodes. All the resource nodes would send and update this information on the attributes and status of the resource nodes to the attribute repository periodically. The step of job brokering starts with the step of checking the attribute repository for attributes and status for all resource nodes in the grid computing environment. The information retrieved by the resource broker from the attribute repository may be attributes such as CPU utilization, Random Access memory (RAM) capacity or Virtual Memory (VM) Capacity. The status of the resource nodes may also be retrieved.

Next, the step of receiving a first priority attribute and a second priority attribute for the matching of resource nodes that may be best suitable for performing the job. These first and second priority attributes may be decided by a user based on the requirements of the job. For example, the job submitted may require very large RAM capacity for handling large amounts of variable data, the first priority attribute will then be RAM. The second priority may then be CPU utilization, depending on the user's requirements for the job. The step of matching available resource nodes corresponding to the first and second priority attributes is then performed. Following which the step of recommending the matching resource node to the job submitter is then determined.

Summary:

In summary, we proposed a solution with the following features :.

• An improved method for Grid job submission, monitoring and management is proposed.

• Monitoring Grid processes and switching between Grid resources in the event of a process failure

• Job brokering to determine the "Best match" grid resource to perform the job request

• Distributing jobs to available resources whereby the number of jobs exceeded the number of resources

• Notification in the event of a process failure through SMS, Email and messaging

Earth & Environmental Sciences & Biodiversity Application / 8

EISCAT_3D Data Solutions

Author: John White[1]

[1] *NeIC*

Corresponding Author: john.white@cern.ch

EISCAT, originally the European Incoherent Scatter Scientific Association, was established in 1975 as a
collaboration between Norway, Sweden, Finland, UK, Germany and France.
The purpose was to develop an
incoherent scatter radar for the Northern auroral zone.
EISCAT has been operational since 1981
and has grown to a globally used research infrastructure.
The present members are Norway, Sweden, Finland, UK, China and Japan.
Many research groups world-wide have used EISCAT for studies of the middle atmosphere,
ionosphere and magnetosphere, either through collaboration with member institutes or by applying for time
through the peer-review access programme.

The existing EISCAT radars are single beam systems with parabolic dish antennas.
EISCAT has now started to construct the next generation imaging incoherent scatter radar, EISCAT_3D.
This will be a system of distributed antenna arrays with fully digital signal processing that will
enable comprehensive three-dimensional vector observations of the atmosphere
and ionosphere above Northern Fenno-Scandinavia.
Through the use of new technology based on the latest advances within digital signal processing,

it aims to achieve ten times higher temporal and spatial resolution compared to present radars.
It will also offer continuous measurement capabilities through fully remote operation.

On each of the three EISCAT_3D radar receiver sites will be 109 sub-arrays of 91 antennas each, each with a
first stage beamformer, based on FPGA technology, to form 10 wide-angle beams at two polarizations.
Subsequently, the wide-angle beam data packets from each sub-array will be stored in a fast RAM memory
buffer and combined into 100 narrow-angle beams by a second stage beamformer on the site level.
Data from each site will be sent to a storage buffer at a central site.
At this central site, computing capacity connected to the storage buffer will combine the site data to form
spatially resolved data products.
The operational modes of the EISCAT_3D radars may be affected
or controlled by results produced at the central site.
Quality-checked final data products will be sent to
long-term storage at data centres.
This data is subsequently served to the EISCAT_3D users, under FAIR principles, for further analysis.

The Nordic e-Infrastructure Collaboration (NeIC) facilitates development
and operation of high-quality e-Infrastructure solutions in areas of joint Nordic interest.
NeIC is a distributed organisation consisting of technical experts from academic high-performance
computing centres across the Nordic countries.

The NeIC EISCAT_3D Data Solutions (E3DDS) project follows from EISCAT_3D support project (E3DS).
This project cooperates with national e-infrastructure providers in the EISCAT_3D participating countries in
order to simulate the data flows at the radar receive sites and from the sites to the central data storage and
computing.
A scale simulation of the on-site processing will be deployed in national e-infrastructures in collaboration with the providers.
The simulations of on-site processing will generate data and send to a test storage (simulating the central disk buffer).
Also the networking configurations between the future EISCAT_3D sites can be tested in collaboration with Nordic network providers.

The EISCAT_3D data will be accessible according to the FAIR principles.
Users will access data primarily through a portal using their federated identity.
EISCAT does have internal policies regarding data access, but overall the principle of data as
findable, accessible, interoperable and re-usable is followed.

Summary:

The EISCAT_3D data chain is simulated as the construction of the experiment proceeds. In order that data can be served to the users in a timely manner the production of data on the remote sites and transfer to data centres must be simulated. Some of the questions to be answered with this simulation include the data format, addition of persistent identifiers in order to serve data in a FAIR manner.

Data Management & Big Data / 9

Storage-Events and dCache: new frontiers in storage

Authors: Albert Rossi[1]; Dmitry Litvintsev[1]; Jürgen Starek[2]; Marina Sahakyan[2]; Olufemi Adeyemi[2]; Patrick Fuhrmann[3]; Paul Millar[2]; Tigran Mkrtchyan[2]

[1] *FNAL*

[2] *DESY*

[3] *DESY/dCache.org*

Corresponding Authors: patrick.fuhrmann@desy.de, paul.millar@desy.de

dCache has introduced a new paradigm for scientific storage: storage events. In traditional interactions there is always the same pattern: the client requests some operation and the server replies with the result of that request. If the client wishes to learn the current status of some resource, it makes a request for this information. To discover when the state of that resource has changed, a client polls the service by making periodic requests that query the current status of the resource. Note that the time between successive queries must be carefully chosen: querying too often may place considerable strain on the storage system. Therefore, polling inevitably introduces latency.

It may happen that it is impossible to avoid placing unacceptable load on the storage system while providing a sufficiently quick response. In such situations, a typical solution is for the agent that modifies the storage system to inform interested parties through some communication side-channel. This coupling between the client that interacts with the storage and the agent reacting to the changes brings several disadvantages: the client must be custom (potentially a maintenance and platform availability problem), and it must know which agents it must inform (potentially a discovery problem).

Moreover, some events are apparently spontaneous and are not a result of client interactions. As an example, consider a client that wishes to learn when a file currently stored on a high-latency media (such as tape) is available for reading with low latency. This event is not triggered as the result of external client interaction, but rather from the activity of the tape system.

In the storage-event model, clients are alerted to activity within the storage system that they are interested in. The key distinction is that this happens without polling: when an interesting event happens, dCache informs the client with minimal delay. This allows the client to learn of changes in near realtime, without placing load on the server.

Storage events are widely applicable. They allow more robust and scalable solutions to several existing infrastructure challenges. For example, file catalogues may use storage events to learn of changes autonomously, without requiring custom clients; clients can request a large number of files be staged back from tape, and start processing as soon as files become available.

Storage events also allow for complex and innovative solutions. By breaking down the connection between the agent modifying storage and the agent reacting to those changes, new services are possible. For example, it is possible to couple computational workflows with data ingest events; domain-specific portals can react to data ingest quickly, so are always up-to-date with the available data.

dCache provides two mechanisms that support storage events: Kafka and Server-Sent Events (SSE). These provide complementary solutions that target different use-cases. Kafka storage-events are more suited for site-level integration, where the institute running dCache wishes to augment its behaviour by integrating it with other services. SSE, a standard web protocol for notification, gives ordinary dCache users access to storage events, allowing ad-hoc innovation.

In this paper, we will present a review of the storage events concept, an update on the current support for storage event support in dCache, and an overview of how various projects are using or plan to use storage events.

Networking, Security, Infrastructure & Operations / 10

The design and development of Vulnerability management system

Author: Hao Hu[1]

Co-authors: Chengcai Zhao [2]; Manman Cheng [2]; Tian Yan [2]

[1] *Institute of High Energy Physics*

[2] *IHEP*

Corresponding Author: huhao@ihep.ac.cn

There are always many vulnerabilities in the operation system, applications and network devices, and vulnerabilities are great threats for security. The vulnerability management and lifecycle tracking is very important and necessary for the security team.
The paper describes the design and development of the vulnerability management system. The functional modules of the system includes information system asserts management, vulnerability detection, vulnerability handling and tracking, vulnerability statistics and visualization. Asserts management modules as a basic part of this system, discovers and stores the existing IT asserts (Hardware information, Operation information, IP address, TCP port services, URLs and so on). Vulnerability detection module discovers the vulnerabilities in the IT asserts using some open-source and commercial tools. The vulnerability handling and tracking module distributes and tracks the vulnerability until all the vulnerability tickets are closed. The statistics and visualization module serves for the security team, which provides data analysis results diagrammatically to show the proportion of each vulnerability category and the vulnerabilities trend during a period of time.
Partial functions of the system have been developed and deployed in IHEP information environment.

Summary:

The paper describes the design and development of the vulnerability management system. The functional modules of the system includes information system asserts management, vulnerability detection, vulnerability handling and tracking, vulnerability statistics and visualization. Asserts management modules as a basic part of this system, discovers and stores the existing IT asserts (Hardware information, Operation information, IP address, TCP port services, URLs and so on). Vulnerability detection module discovers the vulnerabilities in the IT asserts using some open-source and commercial tools. The vulnerability handling and tracking module distributes and tracks the vulnerability until all the vulnerability tickets are closed. The statistics and visualization module serves for the security team, which provides data analysis results diagrammatically to show the proportion of each vulnerability category and the vulnerabilities trend during a period of time.

Infrastructure Clouds and Virtualisation / 11

Function-as-a-Service and event-driven automation for the European Open Science Cloud

Author: Michael Schuh[1]

Co-authors: Jürgen Starek [1]; Patrick Fuhrmann [1]; Paul Millar [1]; Schlünzen Frank [1]; Tigran Mkrtchyan [1]; Volker Gülzow [1]

[1] *DESY*

Corresponding Author: michael.schuh@desy.de

The Photon and Neutron science community (PaN) is pushing frontiers with ground breaking research and technologies in molecular imaging at the atomic level. State of the art Photon and Neutron sources, like the European XFEL and the European Spallation Source (ESS) will create hundreds of Petabytes of data per year, challenging established data processing strategies. Leveraging cloud computing methodologies, DESY develops innovative flexible and scalable storage and compute service, covering the entire data life cycle from experiment control to long term archival, with a particular focus on re-usability by the long tail of science.

Contributing to the European Open Science Cloud (EOSC) pilot, DESY, XFEL and ESS demonstrated cloud based solutions for FAIR access to large volumes of scientific data. The reproducibility of methods and results requires an integrated approach that bundles publications, data, workflows and functions. Fine grained access to functions-as-a-service (FaaS) infrastructures enables scientists to develop and deploy micro-services within minutes. Through shared container registries, those services will become available immediately as new cloud functions on the DESY compute cloud and be run by federated resources in the European Open Science Cloud.

The FaaS approach leads to evolving libraries and catalogues for standard functions, enhanced efficient resource provisioning and enables auto-scaling for compute intensive and repetitively used codes. For highly specialized applications, the platform preserves software environments, configurations and algorithm implementations, citable via DOIs.

We see the integration of our cloud functions into the European Open Science Cloud as an important incentive to further focus on metadata and data interoperability, feeding products from photon science into domain specific analysis and simulation tools e.g. in structural biology and material sciences. Well-defined interfaces enable users to route data through sequences of functions from various frameworks. Where data connectors or format converters are needed, they can be deployed as functions implementing additional micro-services. Providing scientists with the means to host and develop underlying codes, DESY runs collaborative platforms like GitLab and JupyterHub on auto-scaling clusters.

The interactive usage in scientific analysis is complemented by the directly achievable automation, which is closely integrated and designed to scale-up to high throughput use cases. Building on storage events, generated by the underlying dCache storage system, analysis pipelines can be triggered automatically for new, incoming data. The fully automated code execution in response to storage events directly extracts metadata, updates data catalogues, feeds into monitoring and accounting systems and creates derived data sets. In a decoupled design, storing derived data can trigger subsequent actions.

In the eXtreme-DataCloud (XDC) project, DESY demonstrates that event-driven code execution as a service adds a flexible building block to smart data placement strategies, enforcing machine actionable "Data Management Plans". For this, the FaaS system interacts with rule-based data management engines and file transfer systems, e.g. to create replicas of data sets with respect to data locality and Quality of Service for storage. On data ingestion, files can be copied to cloud storage elements, which act as buffers next to strong clusters of compute elements, which carry out Function-as-a-Service pipelines and update data placement rules on success. This automatically deregisters files in the buffer next to the compute clusters, and triggers their distribution to offline storage and long term archival.

In our presentation, we will discuss both, the user and the provider perspective of a computing infrastructure, described above. We will elaborate on the benefits of such a decoupled cloud based service oriented architecture and we will provide a live demonstration, illustrating how to interactively install developed functions in an automated data processing pipeline.

Summary:

The Photon and Neutron science community (PaN) will create hundreds of Petabytes of data per year and develops cloud based solutions for FAIR access and reproducibility of methods and results by bundling publications, data, workflows and functions-as-a-service (FaaS) infrastructures. This approach leads to evolving libraries and catalogues for standard functions, enhanced efficient resource provisioning and

enables auto-scaling on cloud resources. Contributing to the European Open Science Cloud (EOSC) pilot, the platform preserves software environments, configurations and algorithm implementations.

Triggered in response to storage events from the underlying dCache storage system, analysis pipelines can be fully automated and event-driven code execution as a service adds a flexible building block to smart data placement strategies, enforcing machine actionable "Data Management Plans".

This presentation, will discuss both, the user and the provider perspective of a decoupled cloud based service oriented architecture and will provide a live demonstration, illustrating how to interactively install developed functions in an automated data processing pipeline.

Networking, Security, Infrastructure & Operations / 12

Infrastructure-as-Code: How to make the Textual Representation of Scientific Infrastructures FAIR

Authors: Dieter Kranzlmuller[1]; Tobias Weber[2]

[1] *LMU Munich*

[2] *Leibniz Supercomputing Centre*

Corresponding Author: weber@lrz.de

Background:

The approach of infrastructure-as-code allows to efficiently manage large infrastructures, for instance to support FAIR data management. A canonical and machine-actionable description of these infrastructures can itself be an item of research and an essential component in handling reproducibility challenges for the results achieved on the infrastructures. Such a description would include all necessary steps and dependencies in order to provision the infrastructure described in a machine-actionable and human-readable way.

Objective:

The textual representations of the infrastructures need both to be handled in a FAIR spirit and to be compliant to common DevOps quality management standards: On the one hand, they should be Findable, Accessible, Interoperable and Reusable. This can be achieved by annotating the infrastructure-as-code with rich metadata, including a permanent identifier that can be used to retrieve the textual representation. On the other hand, they should be integrated into a workflow of continuous integration and deployment and managed by a version control system.

Method:

Guidelines and best practices to meet these requirements are presented, derived from experience with several research infrastructures hosted at the Leibniz Supercomputing Centre. The infrastructures considered span from small, but specialized systems to large-scale HPC clusters. Where beneficial, interviews with administrators were conducted to enrich the guidelines with their experiences.

Results:

The major recommendation is to stick to those standards and tools that are already existent and relevant for the continuous and scalable management of infrastructures in an industrial context. Ansible is an example for a well-established configuration and provisioning tool that finds applicaiton in both scientific and industrial service proliferation. DataCite as a generic metadata scheme is well-equipped to meet the domain-specific requirements with regard to infrastructure description. Beyond that the crucial part is to integrate both domains with qualified references and show their potential with a proof-of-concept.

Evaluation:

The guidelines are exemplified by a proof-of-concept to provision a generic research data infrastructure which runs on a kubernetes cluster. This infrastructure includes services to harvest data providers, a search engine and several backend services to facilitate common workflows such as bookmarking of search results, data staging and platforms to process and analyze the data. The infrastructure as well as the code deployed on it are open source and can be used to reproduce the presented findings.Conclusion: The adoption of the recommendations presented cannot only make infrastructure setups citeable, but might boost best practices and facilitate the federation of distributed services in the context of science.

Summary:

The approach of infrastructure-as-code allows to efficiently manage large infrastructures, for instance to support FAIR data management. The canonical description of these infrastructures can itself be an item of research and an essential component in handling reproducibility challenges with regard to the results achieved on the infrastructures. Such a canonical description typically includes a machin-actionable set of instructions to setup machines, services and seed data. The textual representations of the infrastructures need both to be handled in a FAIR spirit and to be compliant to common DevOps quality management standards. Guidelines and best practices to meet these requirements are presented, derived from experience with large-scale research infrastructures hosted at the Leibniz Supercomputing Centre. The major recommendation is to stick to those standards and tools that are already existent and relevant for the continuous and scalable management of infrastructures in an industrial context. The only modifications necessary of these can be found where scientific context adds requirements that typically cannot be found in a business context. The adoption of these recommendations cannot only make infrastructure setups citeable, but might boost best practices and facilitate the federation of distributed services in the context of science.

Humanities, Arts, and Social Sciences Application / **13**

Deep Learning and Augmented Reality as a tool to explore the naturalistic richness of urban areas

Author: Giulio Bianchini[1]

Co-authors: Daniele Spiga [2]; Diego Perugini [3]; Francesca Vercillo [3]; Laura Melelli [3]; Livio Fanò [4]; Luisa Liucci [3]; Sabrina Nazzareni [3]

[1] *University of Perugia*

[2] *INFN Perugia*

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

[4] *University of Perugia, Department of Physics and Geology ; INFN Perugia*

Corresponding Author: giulio.bianchini@hotmail.it

In the past two decades increasing efforts have been devoted to diversify the tourism industry, such is the case with urban trekking and geotourism, which has become an important channel for promoting geological knowledge (Del Monte at al., 2013). The recent advancements in Augmented Reality technologies as well as the increasing availability of 'born digital' data like those gathered from social media, create the basis for the development of immersive and customized touristic experiences. Urban scientific tourism, Augmented Reality, and data mining are the key elements of the HUSH project. Its first focus is the identification of the naturalistic components in a given urban area (flora, fauna and geological features), through literature surveys and scientific research. These components come to be Points of Interest (PoIs) along touristic paths, where they are connected to the historical and artistic components of the area. Augmented Reality is the mechanism by which the user can access these contents, by means of the HUSH mobile application. In the geodatabase, each PoI is defined by a target image. This allows the users to access the augmented content by framing the target element for the component with their mobile device. The contents are delivered as videos, text, images, or interactive 3D models. The access to the PoIs can be performed in different ways. The user has the possibility to choose between predefined paths, paths suggested according to a keyword-based search and "intelligent paths" based on a Deep Neural Network (DNN). From an infrastructure point of view, the latter represents the most innovative element of the HUSH project. The users can access the application in several ways. One of these is the social login, which allows them to volunteer their social information in order to customize their touristic experience taking into account their general interests and preferences. The information acquired from their social profiles are stored in an anonymous form in a dedicated server. Due to the heterogeneous nature of the data collected, the use of NoSQL database is preferable, as it allows achieving scalable performance, strong resilience and wide availability. Once properly stored, user data are preprocessed and used to train an unsupervised neural network model. Such a type of model represents the future of deep learning, since phenomena like human behavior are largely unsupervised (LeCun et al., 2015). Once properly trained, the model helps identify and select the PoIs that could be of major interest for the user, in order to suggest a customized touristic experience. To achieve this task, the PoIs are properly classified in the database according to their characteristics (e.g., semantic and geographic). The use of the application and the associated increase of user data will improve the accuracy of the model over time. Moreover, based on the consideration that not all users may have or want to use their social profile to login to the application, information regarding the application usage (i.e., PoIs visited by the users) are also collected and used as additional training data for the model.

Acknowledgments

The project 'HUSH – Hiking in Urban Scientific Heritage. Realtà aumentata per un nuovo approccio al patrimonio culturale e ambientale urbano di Perugia', 2017.0103.021 Ricerca Scientifica Tecnologica, is supported by the Fondazione Cassa di Risparmio di Perugia.

References

Del Monte M., Fredi P., Pica A., Vergari F. 2013. Geosites within Rome City center (Italy): a mixture of cultural and geomorphological heritage. Geografia Fisica e Dinamica Quaternaria, 36, 241-257.
LeCun Y., Bengio Y., Hinton G. 2015. Deep learning. Nature, 521, 436-444.

Networking, Security, Infrastructure & Operations / 14

Building a minimum viable Security Operations Centre for the modern grid environment

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

[1] *UKRI STFC*

[2] *CERN*

Corresponding Author: david.crooks@stfc.ac.uk

The modern security landscape affecting grid and cloud sites is constantly evolving, with threats being seen from a range of avenues, including social engineering as well as more direct approaches. It is vital to build up operational security capabilities across the Worldwide LHC Computing Grid (WLCG) in order to improve the defense of the community as a whole. As reported at ISGC 2017 and 2018, the WLCG Security Operations Centres (SOC) Working Group (WG) has been working with sites across the WLCG to develop a model for a Security Operations Centre reference design.

We present the current status of a minimum viable SOC design applicable to a range of different WLCG sites, centered around a few key components.

The design uses the Zeek Intrusion Detection System for monitoring what is happening at the network level in strategic locations: for example at border between the local cluster and external networks, the border between different local network domains or at core infrastructure nodes. The Malware Information Sharing Platform (MISP) is used to share information regarding relevant security events and the associated Indicators of Compromise (IoCs). By feeding IoCs from MISP into Zeek we have a platform that allows the community to share threat intel that is immediately actionable across the entire grid.

The logs produced by Zeek are processed using the Elasticsearch, Logstash, Kibana (elastic) stack for real time indexing and visualisation. This provides sites with a powerful tool for incident response and network forensics. The alerts raised by Zeek are further aggregated, correlated and enriched by an advanced notification processing engine. This ensures that most false positives are automatically whitelisted while at the same time reducing the total number of raised alerts that need to be managed by the computer security team of each site. By enriching these alerts and adding context of what happened around the moment the malicious activity was detected, the time needed to handle these alerts is greatly reduced.

We present possible deployment strategies for all these components in a grid context as well as the integration between them. We also report on the current status of work on integrating other sources of data, in particular using netflow/sflow, into this model.

Lastly we discuss how making use of these SOC capabilities distributed across the participating sites can lead to increasing the operational security across the entire grid.

Data Management & Big Data / 15

Russian academic institutes participation in WLCG DataLake project

Author: Andrey Kiryanov[1]

[1] *CERN*

WLCG DataLake R&D project aims at exploring an evolution of distributed storage while bearing in mind very high demands of HL-LHC era. Its primary objective is to optimize hardware usage and operational costs of a storage system deployed across distributed centers connected by fat networks and operated as a single service. Such storage could host a large fraction of the WLCG and other mega-science experiment's data while eliminating inefficiencies due to various levels of fragmentation. In this talk we will explain Russia's and in particular NRC "Kurchatov Institute's" role in the DataLake project with highlight on our goals, achievements and future plans.

Infrastructure Clouds and Virtualisation / 16

Cloud-based Distributed Computing System for LHAASO Experiment

Author: Qiulan Huang[1]

Co-authors: Haibo Li [2]; Qingbao Hu [3]; Tao Cui [1]; Wei Zheng [1]; Yaodong Cheng [3]

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

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

[3] *Institute of High Energy Physics*

Corresponding Author: huangql@ihep.ac.cn

The LHAASO(Large High Altitude Air Shower Observatory) experiment of IHEP which is located in Daocheng, Sichuan province (at the altitude of 4410 m), which will generate a huge large amount of data and requires massive storage and computing power. With the rapid growth of the High Energy Physics(HEP) experiments data, a single data center in Institute of High Energy Physics (IHEP) has been unable to meet the resource and cooperation requirements from experiments. And the computing resource of LHAASO distributes in several sites like Beijing, Daocheng, Chengdu and so on. It's very necessary to integrate resources across regions to provide the distributed computing service. This paper we mainly discuss the LHAASO distributed computing system based on virtualization and cloud computing technologies. Particularly, we introduce the cloud-based computing architecture to construct the data center located in the observation base Daocheng. Openstack and Docker container orchestration are used to greatly reduce the operation and maintenance cost as well as to make sure the system availability and stability. Also we discuss the key points of federating distributed resources. Firstly, the integration solution of cross-domain resources is proposed, which adopt the HTCondor-C to make the distributed resource work like a whole resource pool. Then a flexible resource scheduling strategy and a job scheduling policy are presented to realize the resource expansion on demand and the efficient job scheduling across regions transparently, so as to improve the overall resource utilization. Finally, the distributed monitoring and secure certification of the distributed computing system are illustrated.

Data Management & Big Data / 17

dCache - data access optimization in a hybrid cloud

Authors: Albert Rossy[1]; Dmitry Litvintsev[1]; Jürgen Starek[2]; Marina Sahakyan[2]; Olufemi Adeyemi[2]; Patrick Fuhrmann[3]; Paul Millar[2]; Sibel Yasar[4]; Tigran Mkrtchyan[2]

[1] *FNAL*

[2] *DESY*

[3] *DESY/dCache.org*

[4] *desy*

Corresponding Authors: michael.schuh@desy.de, tigran.mkrtchyan@desy.de

For over a decade commercial provides offer their computer resources as public clouds. With a couple of clicks one can build and run a full data center without any local infrastructure. While this is a very attractive approach, high costs and/or legal aspects force sites to run their own infrastructure, often as a private cloud, though. However, such local resources may be insufficient if user demand can't be foreseen in advance.

One option to cope with such peak loads is to build a hybrid cloud by combining the local instance with external resources or public clouds. However, jobs running on additional CPU resources may provide low efficiency when remote data access is required. Moreover, additional network traffic may produce higher network utilization, thus higher overall costs or even network bandwidth starvation.

For years, dCache.ORG has provided robust software, called dCache, that is used at more than 80 Universities and research institutes around the world, allowing these sites to provide reliable storage services for the WLCG experiments and many other scientific communities. The flexible architecture of dCache allows running it in a wide variety of configurations and platforms. In this presentation

we will show how to deploy a distributed dCache instance, that dynamically adds caching nodes to provide local data access and an optimized resource usage in a hybrid cloud.

Networking, Security, Infrastructure & Operations / 18

Orchestrating Dynamic High-Speed Network Flows for Secure Research Data Transfers Using an SDN-enabled Infrastructure

Authors: Charles Pike[1]; James Griffioen[1]; Mami Hayashida[1]; Sergio Rivera[1]; Zongming Fei[1]

[1] *University of Kentucky*

Corresponding Author: pike@netlab.uky.edu

Big data is now key to nearly all research disciplines. Even research areas historically unrepresented in high performance computing must now cope with vast data sets that need to be analyzed, processed, and transferred over the network. Network choke points can create significant delay during transmission of these large data sets to and from the cloud, where they often reside. Campus and enterprise networks depend on middleboxes (e.g., firewalls, NAT, load balancers, IDS/IDP) to provide essential services or enforce network policies, yet these middleboxes often contribute to the negative network performance affecting big data transfers. The transmission delays can cause troubling workflow design issues for researchers and raise a dire need for high-throughput networks as part of the campus cyberinfrastructure.

The traditional fix to the above problem is to create a science DMZ, a strictly-administered segment of the campus network that is positioned outside the firewalls and other policy-enforcing middleboxes. Privileged research systems needing high-speed data transfer capabilities are moved from the general campus network to the science DMZ. This exposes the entire machine to attack, and necessitates careful bastion-host configurations. Researchers are confronted with the dilemma of choosing between high-speed networking with the risks of the DMZ, or the negative performance impact that comes with the services and security provided by the general purpose campus network.

To meet this challenge, we proposed a new approach to the design of campus networks based on software defined networking (SDN), specifically OpenFlow. We began by replacing certain building distribution routers with OpenFlow-enabled switches that operate in a hybrid mode, providing normal routing and switching to our standard campus core by default. Using OpenFlow, flows can be redirected to a new SDN core. The SDN core then forwards packets directly to our campus edge router, bypassing all middleboxes in the campus infrastructure. A benefit of this design is that individual flows from a given machine can receive high-speed, middlebox free paths while all other flows from the same machine travel the standard campus path through policy-enforcing middleboxes. This effectively creates a virtual all-campus DMZ, granular to protocol port level, that can be turned on or off programmatically as needed by researchers.

Here we will present a system we call “VIP Lanes” that leverages the SDN-enabled network to provide the authentication, delegation, authorization, and orchestration to dynamically create and teardown trusted research flows either transparently or as requested by researchers. Unlike a science DMZ where privilege is granted to individual machines, VIP Lanes authorization is given out on a per-flow basis. VIP Lanes provides the ability for pre-authorized, trusted users or applications to create flows that bypass the normal campus route, thereby enabling those flows to achieve substantially better performance while maintaining security and policy compliance for other network traffic. The system dynamically calculates routes that bypass bottlenecks using a graph database approach that computes custom paths and inserts OpenFlow rules that modify packet forwarding hop-by-hop. Additional functionality is added to the path when needed using network function virtualization (NFV) techniques (e.g., NAT). We present the VIP Lanes abstraction and services. We describe our current production implementation that not only shows the viability of the VIP Lanes approach, but also demonstrates the types of performance improvements achieved - in some cases approaching a two order of magnitude reduction in transmission times.

Summary:

We present the VIP Lanes system that leverages SDN-enabled networks to provide the authentication, delegation, authorization, and orchestration to dynamically create and teardown trusted research flows either transparently or as requested by researchers. Unlike a science DMZ where privilege is granted to individual machines, VIP Lanes authorization is given out on a per-flow basis. VIP Lanes provides the ability for pre-authorized, trusted users or applications to create flows that bypass the normal campus routing, thereby achieving substantial performance improvements while maintaining security and policy compliance for other network traffic.

Data Management & Big Data / 19

"Faux"-tomography: Changing the Tomographic Paradigm via Deep Learning

Authors: Seth Parker[1]; Stephen Parsons[1]; William B. Seales[1]

Co-author: Charles Pike [1]

[1] *University of Kentucky*

Corresponding Author: pike@netlab.uky.edu

The acceleration of advances in machine learning (ML) as applied to image-based problems has produced robust solutions to challenges often considered too difficult - or even impossible - to solve by computer: face recognition, 3D object recognition, landmark detection, image-based geo-location. This project re-imagines the accepted imaging paradigm in the context of computed tomography and a cloud-based platform for at-scale machine learning. We focus on x-ray computed tomography as an imaging method, and the inclusion of ML techniques running at-scale in the cloud as a standard way to enhance, improve, and transform the details that tomography captures. We call the approach "photomography" (or "faux" tomography), and the idea is to incorporate ML as a crucial part of the tomographic imaging pipeline for the purpose of enhancing and amplifying signals in the data that are impossible for a human to perceive or for other algorithms to detect. The specific types of signals we seek to amplify include changes in topological patterns that give strong clues to material composition and integrity. These signals, which manifest tomographically in small but statistically significant variations in intensity, are present but not readily visible to the unaided human eye.

This paper presents a systematic paradigm for how to apply ML - convolutional neural networks (CNNs) and autoencoders - in the context of tomography; and points to results from specific demonstrations of the power of this approach to amplify important signals. We emphasize the following four technical areas as primary components:

1. **Automated approaches to massive data acquisition:** Tomographic systems typically discard information that we believe should be kept and used at the ML stage.
2. **Reference libraries from photographs:** Sets of large-scale reference libraries must be constructed and organized into CNNs from longitudinal data and supervised examples. This is computationally very expensive in terms of data sizes and required cycles.
3. **Amplification:** Reference to a pre-computed library delivers a result - an estimate of the degree to which a specific signal is present – using cloud-based architectures. The resulting estimate is pushed back into the original data for subsequent algorithmic analysis or human decision-making.
4. **Multi-modal rendering:** Through cloud-based ML reference libraries it is possible to acquire data with one modality (e.g., tomography), and render a realistic result that simulates a different modality (e.g., photograph).

We discuss the computational and structural framework for collecting tomographic data, including additional cues such as multi-power responses, fluorescence measurements, and phase shift estimates that are typically never centralized in the capture of tomographic data. Using the acquired data, we construct specific reference libraries - at scale - by training CNNs and informing the process through autoencoders. These libraries enable us to deliver predictive results in the data for the purpose of subsequent algorithmic processing and for human visualization to see, recognize, and quantify patterns previously thought to be invisible in tomography. Finally, we demonstrate the overall value of this technique in example areas where tomography is being used to solve new and difficult problems: analysis of bone density phenomena, and the analysis of antiquities (inks and fibers).

Summary:

We discuss the computational and structural framework for collecting tomographic data, including additional cues such as multi-power responses, fluorescence measurements, and phase shift estimates that are typically never centralized in the capture of tomographic data. Using the acquired data, we construct specific reference libraries - at scale - by training CNNs and informing the process through autoencoders. These libraries enable us to deliver predictive results in the data for the purpose of subsequent algorithmic processing and for human visualization to see, recognize, and quantify patterns previously thought to be invisible in tomography. Finally, we demonstrate the overall value of this technique in example areas where tomography is being used to solve new and difficult problems: analysis of bone density phenomena, and the analysis of antiquities (inks and fibers).

VRE / **20**

CSTCloud: A Cloud Computing Platform Designed for Scientific Researcher

Author: Honghai Zhang[1]

Co-authors: Leilei Zhang [1]; Ting Wei [1]; Yan Wang [1]

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

Corresponding Author: zhh@cnic.cn

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

Networking, Security, Infrastructure & Operations / **21**

Cyber security monitoring and data analysis at IHEP

Author: Tian Yan[1]

Co-authors: An Dehai [1]; Fazhi QI [2]; Hao Hu [3]; SHAN ZENG [1]

[1] *IHEP*

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

[3] *Institute of High Energy Physics*

Corresponding Author: yant@ihep.ac.cn

In recent years, along with the rapid development of large scientific facilities and e-science worldwide, various cyber security threats has becoming a noticeable challenge in many data centers for scientific research, such as DDoS attack, ransomware, crypto currency mining, data leak, etc.

Intrusion and abnormality detection by collecting and analyzing security data is an important measure for enhancing the sensitivity of security status perception, level of security protection, and agility of security incident response. However, as the scale of data center growing, it's difficult to use a single security box to process the large volume of various data generated by network traffic, device and host logs, threat intelligence, and so on.

In high energy physics (HEP) community, people are trying to establish a security operation center (SOC) for handle this problem. There is a SOC working group for the Worldwide LHC Computing Grid (WLCG). With help of this working group, we are building a cyber security monitoring and analysis framework at Institute of High Energy Physics (IHEP), Chinese Academy of Sciences. At IHEP, we have 4x10Gbps IPv4 and IPv6 dual-stacked internet connection, and 4x40Gbps inner data center network. There are also hundreds of web information servers, thousands of PC clients and thousands of computing nodes. It's really a challenge for us to handle the security related data generated by such a set of information assets.

In this framework, Malware Information Sharing Platform (MISP) is deployed for threat intelligence exchanging with collaborated HEP institutes and universities. Network traffic is collected from switches and firewalls by a 10Gbps network shunt, and then flows to a Bro instance for traffic analysis. Bro logs and hosts/web logs, security device logs, along with vulnerability scanning results and assets detection results, etc., are defined as cyber security data. All of these data are collected by Flume/Logstash/Syslog to a data pipeline named Kafka cluster. In this cluster, there are some Spark jobs running for stream processing, which are aimed at rapid intrusion and abnormality detection as well as data correlation and enrichment. Then all the processed data are written to Elasticsearch, MySQL and InfluxDB, and then visualized by Kibana and Grafana. At the same time, the processed data can be written to local storage, HDFS, or tap storage for backup and long-term analysis.

With help of this security data collection an analysis framework, it is possible for us to handle the large amount of security data generated at IHEP, and it's also very flexible and scalable for even larger amounts of and different kinds of data in future.

Humanities, Arts, and Social Sciences Application / 22

A Smart City Approach of Visitor's Spatial Experience in Night Markets of Taipei, Taiwan using Space Syntax Analysis

Author: Camilo Ariel Jaime Gomez[1]

Co-authors: Sara Sanchez Alquijay [1]; Sheng-Ming Wang [1]

[1] *National Taipei University of Technology*

Stated as one of the most visited top attractions in Taiwan, night markets are well-recognized for their diversity and very distinct service model. When regular working hours have finished, night markets become an essential component of cities' urban fabrics. Thus, the understanding of these unconventional commercial areas has emerged as a very important research topic within multiple fields. Learning about night market's dynamics enhances in different ways and professional disciplines the improvement of the spatial experience regarding user-oriented design. It also contributes to those who find in this retail shops their means of employment. Shopping and recreation seen as the main activities source of income in night markets are also changing. Following how technology progresses every day, computing automation and convenient internet accessibility are stimulating shopping habits to evolve, consumer's spending behavior gradually become aligned with new developments, such developments play a very important role to bring data visualization forward to be interpreted creating a variety of utilization possibilities in near future.

This study first provides a theoretical and conceptual background that illustrates Taiwanese night markets' design logic and conceptual urban sense. It pretends as well, to identify their necessities of retail space management and appropriate preparation to deal with urgent situations with all available resources. Having such information can provide an updated analysis of the influences between visitors, vendors and the physical environment where they interact with when using these retail shops during their business hours. Surveillance and site observation's data are intended to be supported and accompanied by the integration of space syntax analysis model (heat maps), this pursues to proof how computing technology provides valuable information that can be integrated with IoT to further analysis, and subsequently its interpretation causes a tremendous impact on every aspect of service providers and customers' behavior.

Urban formations also shape customer's shopping behavior in a very direct way, the layout within night market areas taken as reference for this study is constantly being read by people while exploring these cityscapes. This exploration phenomenon is generated as outcome of a mental map navigation involvement, like an emotional picture of the exterior physical world that is held by every individual's intuition, this physical world is directly consequence of everything that a person observes and relates to the urban pattern where it is located. It's proven that visitor's sensorial experience is directly influenced by the environment they explore, all these elements produce a very unique spatial experience in visitors and locals, resulting in users unconsciously engaging into the experience so-called: "way-finding practice".

Very few studies have conducted a comprehensive experimental observation model of such spatial experience, there's no preceding record of any study taking into account the usage of new technologies and internet clouds' applications affecting customer's behavior in night markets. Moreover, integrating this knowledge to the relationship between space morphology and functioning forces offers a satisfactory support to the hypotheses that IoT and resulting data visualization can help to discover service innovation methods from social generated activities such as consumers big data through the utilization of mapping visitor's shopping behavior, tracking vendor's positions and then addressing the impact caused on business when introducing such technologies, this proposal also attempts to become an instrument to measure and diagnose the functionality, accessibility and usability of the services offered in night markets.

Summary:

Keywords: service design, smart environment, smart city, space syntax, internet clouds data, service innovation value, service evaluation, social network mapping, space experience

Supercomputing, High Throughput Computing, Accelerator Technologies, and their Integration / 23

Relevance of Grid Computing for India in the era of Cloud Computing (remote presentation)

Author: Vineeth Arackal[1]

Co-author: Mangala N [1]

[1] *Centre for Development of Advanced Computing*

Corresponding Author: vineeth@cdac.in

Abstract: Grid Computing is quite often implemented by relying on the concept of sharing of available resources in order to improve economies of scale and utilization. However, after the advent of Cloud Computing, the adoption of Grid Computing has been rather slow, and people wonder if it is still relevant, and whether Grid Computing truly matters. In this abstract, the authors explain that, Grid Computing provides services that are complementary to Cloud Computing, rather than competing with Cloud Computing. In a country like India where computer, smart phone and internet penetration is still low when compared to developed countries, Grid Computing can make a huge difference in the adoption of scientific computing by making niche resources available for free or nominal charges to the relevant community of researchers.

Both Grid Computing and Cloud Computing aim to give users access to third-party systems that they do not own. Here the users can be independent users or institutions. A public cloud gives users access to third-party systems that are generally commercially operated. So the user will typically be charged on a pay-per-use model. An example is Amazon Elastic Compute Cloud (Amazon EC2). In contrast, the underlying resources in a private cloud are owned by the users. So, in a private cloud, better utilization of the available resources is the goal, possibly by employing virtualization. However, the pay-per-use model may still remain valid in a private cloud, depending on company policies. For example, one department may lease its resources to another one using the pay-per-use model.

For our discussion, the main difference between Grid Computing and Cloud Computing is that, in the case of a Grid institutions come together to share resources, so that they can work on solving larger problems together, or can have access to more systems. This model can be said to be somewhere between public cloud and private cloud. In fact, this has some resemblance to a community cloud, where users from related usage communities come together to share resources. But in contrast to a community cloud, a Grid is more likely to have high-end parallel systems, whereas systems making up the community cloud need not necessarily be parallel systems. The workload in a typical Grid environment is heavily skewed towards the scientific community. Moreover, supporting a particular community is just one aspect of Grid Computing, and is typically achieved by using the concept of virtual organizations (VOs). Providing access to users from different administrative domains, but without charging them, while at the same time properly verifying their credentials, is an important requirement in Compute Grids.

India's representation in the list of top 500 supercomputers of the world is still minimal. In spite of being the second largest country in terms of population, the fastest of supercomputer in India ranks a dismal 39th in the 51st edition of top 500 list that was released on June 2018. In such a scenario, it is imperative that the available resources be utilized properly. Grid Computing provides such a framework. While projects like the National Supercomputing Mission aims to bridge the gap between what the researchers want and what they currently have, it is important to have mechanisms that can leverage the currently available resources for solving grand challenge problems, and Grid Computing provides the proper mechanism for achieving this goal, through efficient security mechanisms, metascheduling capabilities, file transfer mthodologies, among others. The authors feel that Grid Computing is here to stay, though it may not become as common as Cloud Computing, since it targets a comparatively niche set of users.

Summary:

Abstract: Grid Computing is quite often implemented by relying on the concept of sharing of available resources in order to improve economies of scale and utilization. However, after the advent of Cloud Computing, the adoption of Grid Computing has been rather slow, and people wonder if it is still relevant, and whether Grid Computing truly matters. In this abstract, the authors explain that, Grid Computing provides services that are complementary to Cloud Computing, rather than competing with Cloud Computing. In a country like India where computer, smart phone and internet penetration is still low when compared to developed countries, Grid Computing can make a huge difference in the adoption of scientific computing by making niche resources available for free or nominal charges to the relevant community of researchers.

Both Grid Computing and Cloud Computing aim to give users access to third-party systems that they do not own. Here the users can be independent users or institutions. A public cloud gives users access to third-party systems that are generally commercially operated. So the user will typically be charged on a pay-per-use model. An example is Amazon Elastic Compute Cloud (Amazon EC2). In contrast, the underlying resources in a private cloud are owned by the users. So, in a private cloud, better utilization of the available resources is the goal, possibly by employing virtualization. However, the pay-per-use model may still remain valid in a private cloud, depending on company policies. For example, one department may lease its resources to another one using the pay-per-use model.

For our discussion, the main difference between Grid Computing and Cloud Computing is that, in the case of a Grid institutions come together to share resources, so that they can work on solving larger

problems together, or can have access to more systems. This model can be said to be somewhere between public cloud and private cloud. In fact, this has some resemblance to a community cloud, where users from related usage communities come together to share resources. But in contrast to a community cloud, a Grid is more likely to have high-end parallel systems, whereas systems making up the community cloud need not necessarily be parallel systems. The workload in a typical Grid environment is heavily skewed towards the scientific community. Moreover, supporting a particular community is just one aspect of Grid Computing, and is typically achieved by using the concept of virtual organizations (VOs). Providing access to users from different administrative domains, but without charging them, while at the same time properly verifying their credentials, is an important requirement in Compute Grids.
India's representation in the list of top 500 supercomputers of the world is still minimal. In spite of being the second largest country in terms of population, the fastest of supercomputer in India ranks a dismal 39th in the 51st edition of top 500 list that was released on June 2018. In such a scenario, it is imperative that the available resources be utilized properly. Grid Computing provides such a framework. While projects like the National Supercomputing Mission aims to bridge the gap between what the researchers want and what they currently have, it is important to have mechanisms that can leverage the currently available resources for solving grand challenge problems, and Grid Computing provides the proper mechanism for achieving this goal, through efficient security mechanisms, metascheduling capabilities, file transfer mthodologies, among others. The authors feel that Grid Computing is here to stay, though it may not become as common as Cloud Computing, since it targets a comparatively niche set of users.

Networking, Security, Infrastructure & Operations / 24

CILogon: Enabling Federated Identity and Access Management for Scientific Collaborations

Author: Jim Basney[1]

[1] *NCSA*

Corresponding Author: jbasney@illinois.edu

CILogon provides a software platform that enables scientists to work together to meet their identity and access management (IAM) needs more effectively so they can allocate more time and effort to their core mission of scientific research. The platform builds on open source Shibboleth and COmanage software to provide an integrated IAM platform for science, federated worldwide via eduGAIN. CILogon serves the unique needs of research collaborations, namely to dynamically form collaboration groups across organizations and countries, sharing access to data, instruments, compute clusters, and other resources to enable scientific discovery.

We operate CILogon via a software-as-a-service model to ease integration with a variety of science applications, while making all CILogon software components publicly available under open source licenses to enable re-use. Since CILogon operations began in 2010, our service has expanded from a federated X.509 certification authority (CA) to an OpenID Connect provider, SAML Attribute Authority, and multi-tenant collaboration platform.

In this presentation, we describe new applications of the CILogon platform, our recent operational experiences and lessons learned, and future plans.

Humanities, Arts, and Social Sciences Application / 25

East and West Elections: Mapping the Voice of the People

Author: Fazleen Md Ruslan[1]

Co-author: Faridah Noor Mohd Noor [2]

[1] *University of Malaya, Kuala Lumpur, Malaysia*

[2] *University of Malaya, Kuala Lumpur, Malaysia.*

Corresponding Author: fazleen48@gmail.com

Social media has taken the world by storm with netizens expressing their feelings and opinions on a variety of topics from emerging crisis events to political preferences. Digital media content, thus, has presented researchers the opportunity to investigate various phenomena in natural occurring setting. The subject of investigation for the present study involves a magnitude of sentiments regarding the shock victory elections in the US and Malaysia, namely, the discourse of Twitter users over the US Presidential Elections 2016 and the Malaysian 14th General Election (GE14) 2018. FAIR (ISGC 2019) refers to findable, accessible, interoperable and re-useable data. For this study, a multitude of positive and negative sentiments are used as data, inclusive of profanity, towards the contesting leaders. Data is findable through the usage of relevant hashtags, that is searchable terms made visible to an audience who are interested in reading and posting about the topic. This presentation aims to share the application of the freeware tool TAGS (Hawksey, 2010) and it's interoperability with Twitter (Search API) in allowing an automated collection of the intended search results. Selection is based on tweets containing the most used hashtags for both elections. Annotation and concordance of the data is processed through AntConc (Anthony, 2009) and then mapped according to a discourse semantic region of judgment and appreciation using the Appraisal Theory (Martin and White, 2005) framework. It is hoped that investigating the collective phenomena carries the purpose of developing a better understanding of the use of social media in political discourses. As Twitter not only provides a feasible avenue for engaging public opinion, it also provides practical uses for politicians and the society in becoming informed citizens. Thus, accessibility to information enables researchers to explore and track these political and policy preferences, which in turn could possibly be re-usable in predicting future election outcomes.

Summary:

This presentation is a PhD study in the area of discourse. It is hoped the presentation will hopefully provide feedback to improve the study.

VRE / 26

Authentication and Authorization for RESTful WEB API in Scientific Computing Environment

Author: Rongqiang Cao[1]

Co-authors: Rong He [1]; Shasha Lu [1]; Xiaoning Wang [1]; Xuebin Chi [1]; Yangang Wang [1]

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

Corresponding Author: caorq@sccas.cn

Through grid computing and cloud computing technologiesSCE (Scientific Computing Environment, previously also known as ScGrid) integrates massive computing, storage and application resources. AS a general-purpose computing platform started from 2006 in CAS (Chinese Academy of Sciences), SCE is designed as a pyramidal structure. At present, the top layer is a centralized massive computing environment named ERA which is a heterogeneous cluster, building a 2.3 petaflops (CPU: 700 teraflops, GPU and MIC: 1.6 petaflops) computer and software support platform. The middle layer is distributed among China, in which 9 branch centres are chose and connected. In addition, SCE has 18 sub-branch centres and 11 GPU centres in the bottom layer.

All resources in SCE are packaged as easy-to-use open APIs in RESTful web services. These APIs are used to develop client softwares for multi-disciplinary and cross-scenario. Around authentication and authorization issues among users, open APIs and clients, simplified authentication and authorization services are proposed and implemented in this paper. There are 3 problems to be solved as

follows: (1) how an account of SCE securely to login into a client developed by third-party without disclosing sensitive credential information such as password? (2) How a user of SCE authorizes a client access computing resources and private data in limited scope and block illegal actions beyond the approved scope? (3) How an administrator to manage privileges of open APIs and assign different permissions to any client dynamically?

Regarding background and issues above, the proposed services in this paper provide single sign-on in multiple WEB communities for SCE accounts, support users to authorize terminal softwares that could access massive resources and personal private data in proxy mode, and also help administrators determine which open APIs a client could access. Several micro-services were implemented or deployed to provide easy-to-use and simple authentication and authorization for RESTful WEB API in SCE. The single sign-on (SSO) micro-service is used to provide simple login service for clients and web gateways via account of SCE. The open API authorization is used to provide simple authorization workflow for user, perform delegated actions in permission scope, and forbid illegal and malicious attacks to computing resources and private data. The large files transfer micro-service also was implemented to transport large data in isolate service instances protected by the proposed authentication and authorization services. In addition, two-phrased authentication was proposed to enhance security and improve usability. The first authentication phase was used to provide the single sign-on or the simple username and password login service for users to login clients and web gateways developed by third-parties. The second authentication phase was used to provide authentication and permission validation service for clients and web gateways to perform delegated actions in user's authorization scope.

Atop the proposed services, all related people, consisting of users, developers and administrators, they no longer need to worry about and solve complex problems in authentication and authorization. What they need to pay much attention on are specific business logics and application scenarios for their interested areas.

Summary:

This paper designed and implemented simple authentication and authorization services for RESTful WEB API in SCE. The proposed services have been applied to general computing portal, operation and management portal in national high-performance computing environment, and also WEB communities for computational chemistry, bioinformatics, etc. These examples show that the proposed services have achieved positive effects and good results. In future, we will continue to improve the authorization service to support more types and sources of account, and extend the log analysis tools to discovery illegal events timely even real-time online. This work was partially supported by National Natural Science Foundation of China under grant No. 61702476.

Earth & Environmental Sciences & Biodiversity Application / 27

Air quality monitoring issue and some study results of Ulaanbaatar city

Author: Otgonsuvd Badrakh[1]

[1] *Institute of Physics and Technology, MAS*

Corresponding Author: otto_hi@yahoo.com

This study is discussed to air quality monitoring issues and challenges of Ulaanbaatar city which is a capital city of Mongolia. Air monitoring is one technique used to measure and assess the status of ambient air quality. Air pollutants are all very different in terms of chemical composition, reaction properties, emission sources, and fate and transport in the environment. Six of these pollutants are well studied and ubiquitous in our daily lives, including carbon monoxide (CO), nitrogen dioxide (NO2), ground level ozone (O3), sulfur dioxide (SO2), particulate matter (PM) and lead (Pb). Currently, the air pollution data at locations without monitoring stations are obtained by air quality models or estimations. However, the data from the air quality models lack of cross-validation and verification. The low-cost portable ambient sensors provide a huge opportunity in increasing the spatio-temporal

resolution of the air pollution information and are even able to verify, fine-tune or improve the existing ambient air quality models.

VRE / 28

Information Service for High Performance Computing Environment based on Message Bus

Author: Can Wu[1]

Co-authors: Haili XIAO [2]; xiaoning wang [1]; xuebin chi [1]; yining zhao [1]

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

[2] *Supercomputing Center, Chinese Academy of Sciences*

Corresponding Authors: zhaoyn@sccas.cn, wucan@sccas.cn

The High Performance Computing Environment in China (a.k.a. China National Grid, or CNGrid) aggregates the majority of China's supercomputers including Sunway TaihuLight and Milkyway-2 with 200PF aggregated computing resource and 167PB aggregated storage resource, and provides unified and convenient high-performance computing services for users. With the advent of the Big Data Era, the information service in the High Performance Computing Environment is facing a huge challenge. On one hand, the development of high-performance computer manufacturing brings more supercomputers into the High Performance Computing Environment. And a single supercomputer can provide more resource scale and support more application tasks; On the other hand, as the combination of Big Data technology and High Performance Computing technology, more and more Big Data applications are running in High Performance Computing Environment. The amount of resource information is surging, which brings a huge challenge to the information service. SCE is the middleware software in the High Performance Computing Environment and it is developed by Computer Network Information Center, Chinese Academy of Sciences (CAS). After the arrival of the Big Data Era, the resource information in SCE increases rapidly. How to quickly update a huge number of resource information for job scheduling and accurately provide real-time information for users becomes a key problem. To provide efficient and stable information service for users, an optimized information service is necessary.

This paper proposes an optimized information service based on message bus, which is efficient, scalable and simple. The optimized information service standardizes resource information which makes it easier to scale out. The message bus uses ZooKeeper cluster and Kafka cluster to transfer information, which improves the throughput greatly and supports multiple management centers to share resource information and to provide information management services at the same time, which reduces the time cost effectively. The security system in message bus is consisted of authority management, identity authentication and data backup, which ensures the system security. The optimized information service improves information update accuracy and security which gives users better experience. As the experimental experiment results shown, the optimized information service provides efficient information service with lower cost and lower system load by standardizing information and transferring information concurrently. Meanwhile, the security system and simple interfaces makes the information service more available and much easier to use. The optimized information service is an efficient, available and simple information service under the Big Data Era environment.

Infrastructure Clouds and Virtualisation / 31

Latest advancements in EGI operations for improved cloud federations

Author: Gergely Sipos[1]

Co-authors: Baptiste Grenier [2]; Enol Fernandez [2]; Giuseppe La Rocca [2]; Nicolas Liampotis [3]

[1] *EGI*

[2] *EGI Foundation*

[3] *GRNET*

Corresponding Author: gergely.sipos@egi.eu

EGI uses and offers a portfolio of services for communities to federate and operate distributed compute and storage sites. These services - called operation services - include both online technical elements and distributed support teams. The technical elements comprise of a Configuration Database, a Monitoring system, an Operations portal, various Security systems, an Information discovery tool, a Helpdesk, a Software validation pipeline and repository and a Messaging broker.

The talk will cover the recent advancements in these technical elements, with a focus on the Check-in security service, a recent addition to simplify security for both resource providers and users. Check-in is a security proxy service that operates as a central hub to connect federated Identity Providers (IdPs) with service providers. Check-in allows users to select their preferred IdP so that they can access and use EGI and external services in a uniform and easy way.

Check-in is a response to a long-standing need of research communities: be able to interact with distributed computing infrastructures using username-passwords that carry the same level of trust as X.509 certificates. Check-in was designed according to the AARC blueprint architecture and it's compatible with various academic and social identity providers, as well as various types of service providers. One of these service providers is OpenStack, a key building block of the EGI Cloud service.

The EGI Cloud service was designed in 2014 and was put into production in 2015. The service is implemented as a 'federated Cloud' of Openstack sites. The federation is built of Infrastructure as a Service (IaaS) cloud providers, where each IaaS is operated by different institutes according to collaboratively agreed principles and operational regulations. These principles and regulations require OpenStack providers to connect their site with the EGI Operation services, and expose their cloud to users through commonly agreed interfaces. During the 2018 - thanks to the new Check-in service - the EGI Cloud was made accessible with username-password that not only lowered the barrier of access, but also opened up possibilities for new types of interfaces.

One of these new interfaces is the AppDB VMOps Dashboard, a web interface that can be used to instantiate virtualised applications on any connected OpenStack cloud site through a single GUI. The VMOps Dashboard complements the existing features of AppDB and now serves as a front-end for both application providers and application managers & users. Through AppDB application providers can replicate virtualised applications (Virtual Machine Images) to the federated cloud sites, while application managers & users can intantite and use those applications.

Another new interface that Check-in enabled in the EGI Cloud is the Jupyter-based EGI Notebooks. The EGI Notebooks service provides browser-based, scalable tool for interactive data analysis supporting different programming languages and computational software by using the Open Source Jupyter and JupyterHub software packages. EGI Notebooks is a multi-user environment that offers communities a one-click experience without software setup to run data analysis tasks (e.g. data cleaning and transformation, numerical simulation, statistical modelling, data visualisation, machine learning). Jupyter is a cross-domain computational platform that has raised interest from various research communities, such as environmental sciences, life sciences, food research, mathematics and physics.

The presentation will describe these new EGI elements and will show to resource providers and researchers in the Asia Pacific region how the Check-in, Federated Cloud and Notebooks services can be adopted to local use.

Summary:

This presentation introduces the EGI Operations services with a focus on their recent advancements in the area of security, and showing how these advancements improved the operation of the EGI cloud federation. The EGI Cloud federation is a collaboration of cloud providers with a growing emphasis on federating OpenStack IaaS resources as well as higher level services offered on top.

Humanities, Arts, and Social Sciences Application / 32

Proposing an ICT-enhanced curriculum for global learning environment for Social Entrepreneurship for universities in Asia

Author: Tosh YAMAMOTO[1]

Co-authors: Benson ONG [2]; Maki OKUNUKI [3]

[1] *CTL Kansai University*

[2] *Dept. of Business, NYP Singapore*

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

This paper presents a curriculum development endeavor for Social Entrepreneurship including the simulation of preparing for a venture business to the Mezzanine stage. It must be emphasized that the curriculum is not limited to business majors but also for all university students of various majors. In such a curriculum, Creative/Creative Thinking in the global team was the focus.
The curriculum has been pilot-tested on the Kansai University campus for five years with Japanese, Korean, Chinese, Taiwanese, Vietnamese, English, Dutch, German, and French students. The students formed international teams and developed their own brand/product after market research and planned and designed their dream project to come true. The teams with global team members considered target consumers, the fund for initiating a business, and the plan to bring the project to the Mezzanine stage.
The course was enhanced with the state-of-the-art ICT making the entire class, as well as all the team members, be on the same page of the learning process throughout the course. Such ICT tools for globally collaborative learning are show-cased in the presentation.

Summary:

Although an oral presentation is requested. The proposed presentation could be a poster presentation.

Supercomputing, High Throughput Computing, Accelerator Technologies, and their Integration / 33

The BondMachine toolkit: Enabling Machine Learning on FPGA

Authors: Daniele Bonacorsi[1]; Davide Salomoni[2]; Loriano Storchi[3]; Mirko Mariotti[4]; daniele spiga[5]; tommaso boccali[6]

[1] *University of Bologna*

[2] *INFN-CNAF*

[3] *Dipartimento di Farmacia, Universitá degli Studi G. D'Annunzio, Chieti*

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

[5] *INFN-PG*

[6] *INFN*

Corresponding Author: mirko.mariotti@unipg.it

Future systems will be characterized by the presence of **many computing cores** in a single device, by **heterogeneous architectures** built to optimize power and "silicon" consumption as much as possible and by **re-configurable hardware** technologies. These concepts have been demonstrated, both in software programming and hardware evolution, by the multi-core, GPGPU, OpenCL and re-programmable logic devices which populate the spectrum from small devices up to large-scale

data centers. A key to the success in the era of hybrid computing will be how coherently HW/SW systems will take all these components into account.

The BondMachine (BM) is an innovative prototype software ecosystem aimed at creating facilities where both hardware and software are co-designed, guaranteeing a **full exploitation of fabric capabilities** (both in terms of concurrency and heterogeneity) with the smallest possible power dissipation. The disruptive innovation of the BM is to provide **a new kind of computer architecture**, where hardware **dynamically** adapts to the specific computational problem, rather than being static and generic, as in standard CPUs synthesized in silicon.

In order to exploit the dynamic nature of the BM, its main goal is to create the described heterogeneous and flexible architectures on top of re-configurable technology devices (such as FPGAs). Moreover, the overall BM vision is based on the reduction of the number of hardware/software layers, which as a byproduct guarantees a simpler software development.This is precisely why the BM project has been thought as a complete re-configurable computing ecosystem, that starting from a high-level description creates both the hardware and the software that runs on it.

The BM uses Go as main language for the codesign. Its concurrency primitives are perfect to be mapped in the BM architecture and to allow writing concurrent applications on FPGA with a small overhead compared to the HDL code.
The flexibility of the BM makes it possible the implementation of any computing system, ranging from networks of small agents, like IoT (Internet of Things), to high performance devices for ML (Machine Learning) or real time data analysis, and even systems that mix all these different characteristics together.

In the context of the HEP domain, we are developing new BM components to **deploy complex AI systems on hardware**, providing a high-level mechanism to translate into silicon Deep Learning networks, created via standard Tensorflow and Keras toolkits.
For what regards deployment models, the BM provides several solutions, such as a standalone FPGA, accelerators coupled to workstations, as well as a BM as a Service running on hybrid Clouds.

In this talk we will provide a technical overview of the key aspects of the BondMachine toolkit, highlighting the advancements brought about by the porting of Go code in hardware. We will then show a cloud-based BM as a Service deployment. Finally, we will focus on Tensor Flow, and in this context we will show how we plan to benchmark the system with a ML tracking reconstruction from pp collision at the LHC.

Supercomputing, High Throughput Computing, Accelerator Technologies, and their Integration / 34

Evolution of the INDIGO-DataCloud architecture toward an advanced open source Cloud platform integrating AI-based workflows exploiting large-scale Big Data facilities

Authors: Davide Salomoni[1]; Giacinto Donvito[1]

[1] *INFN*

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

In this contribution, we describe an innovative open source Cloud platform evolving the architecture and outcomes of the successful INDIGO-DataCloud project (INDIGO, https://www.indigo-datacloud.eu) and of the two INDIGO follow-on projects DEEP-HybridDataCloud (DEEP, https://deep-hybrid-datacloud.eu) and eXtreme-DataCloud (XDC, http://www.extreme-datacloud.eu).

INDIGO developed a modular open source data and computing platform targeted at scientific communities, deployable over public or private resources. By filling many technology gaps at the PaaS and SaaS levels, INDIGO helped developers, resource providers, e-infrastructures and scientific communities to overcome key challenges in the Cloud computing, storage and network areas.
XDC is developing scalable technologies for federating storage resources and managing massive

amount of data in highly distributed computing environments, as required by the most demanding, data intensive research experiments in Europe and worldwide.
DEEP is evolving intensive computing services offered via Cloud infrastructures and exploiting specialised hardware components, such as GPUs, low-latency interconnects, and others typically accessed as “bare metal” resources.

The proposed architecture described here is the evolution of these three solutions, which are either already available or planned to be made available shortly. This new platform will extend, leverage and achieve convergence of services used to access extreme large datasets over distributed environments and to exploit specialised hardware for high-level, easy to use support of AI workloads in Cloud and HPC environments.

Some of the key characteristics of the proposed architecture are:

- Integration of third-party storage services, allowing to automatically stage in data required by applications into dynamically-created clusters, running on any IaaS Cloud infrastructure.
- Support for experiment reproducibility, providing the possibility to e.g. select or playback given versions of applications or containers, PIDs for the input files, etc.
- Advanced QoS support on heterogeneous and specialized hardware, including e.g. SSD storage systems vs. other storage types. This will make it possible to build hybrid computing clusters with local dynamic data caches tailored to actual workloads.
- Provisioning of high-level interfaces for the exploitation of metadata used to find data to be processed. Based on this metadata and without having to know the details of data locations, file names, etc., users will be able to request analysis code to be automatically executed on distributed resources.
- Support for user-level workflows via standard languages such as CWL or similar, extending toward end users the work on service-oriented TOSCA templates that was started in INDIGO. Through these workflows, users will not only be able to ask for the creation and set-up of complex clusters of services, but also to co-design the execution of entire multi-step distributed analysis processes.
- Support for building overlay HTCondor-based computing clusters running on top of distributed heterogeneous resources, which can range from large Cloud or HPC allocations, down to opportunistic resources such as a single server or user desktops.

The talk will describe how we intend to drive and implement the proposed architecture, which existing components will need further development to support the features mentioned above, and eventually how services for these features can be deployed into real infrastructures.

Networking, Security, Infrastructure & Operations / **35**

Recent Developments and Plans on CernVM-FS at RAL

Author: Catalin Condurache[1]

[1] *STFC Rutherford Appleton Laboratory*

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

Firmly established as an extremely effective mechanism for providing scalable, POSIX like, access to experiment software and conditions data for the LHC experiments and many other research groups at Grid sites, the CernVm File System (CernVM-FS) continued to present increased interest to many other High Energy Physics (HEP) and non-HEP (i.e. Space, Natural and Life Sciences) communities activating within and making use of various Cloud computing environments.

This presentation will give an overview of the CernVM-FS infrastructure deployed at RAL Tier-1 as part both of the WLCG Stratum-1 network and the CernVM-FS facility that provides a complete service for the non-LHC communities within EGI and that can be used as a proof of concept for other research infrastructures and organizations looking to adopt a common software repository solution.

Focus of the presentation will be on the latest developments to widen the scope of the CernVM-FS technology usage within various research communities. The status of implementing the 'confidential' CernVM-FS repositories - a requirement for academic communities willing to use CernVM-FS technology - is reviewed, including a case study that describes the design of a production model around 'standard' and 'protected' repositories.

We will also describe the recent work undertaken at RAL with Large Scale CernVM-FS and DynaFed. Large Scale CernVM-FS is an extension of data distribution mechanism developed by Open Science Grid collaboration that allows access to files from any storage offering http access. DynaFed (Dynamic Federation) is a system that can federate http storage endpoints and is able to present a huge distributed repository as a single one and the presentation will discuss the steps taken to prototype and build a global file system for data access using Large Scale CernVM-FS and DynaFed.

Physics & Engineering Application / 36

R&D for the expansion of the Tokyo regional analysis center using Google Cloud Platform

Author: Michiru Kaneda[1]

Co-authors: Junichi Tanaka [2]; Nagataka Matsui [3]; Sawada Ryu [3]; Tetsuro Mashimo [3]; Tomoe Kishimoto [2]

[1] *ICEPP, the University of Tokyo*

[2] *University of Tokyo*

[3] *The University of Tokyo*

Corresponding Author: michiru.kaneda@cern.ch

Tokyo regional analysis center at the International Center for Elementary Particle Physics (ICEPP), the University of Tokyo, is a computing center for the ATLAS experiment at Large Hadron Collider (LHC) and one of the Worldwide LHC Computing Grid (WLCG) Tier2 sites supporting ATLAS VO. The center provides 10,000 CPU cores and 10 PB disk storage. A part of resources is dedicated to the local usage of the ATLAS Japan member. Hardware devices in the center are supplied by the three years rental. The current system is the 4th system which contract will end in this year. Currently, a migration to the next 5th system is ongoing. In the next system, the number of CPU cores is almost the same as the current system while the performance will be improved about 9% per core based on SPECint. The file storage will be increased to 15TB.

LHC plans the High-Luminosity LHC project starting from 2026. The peak luminosity will be 5 times higher. This will require more than 10 times of computing resources for the experiment. Such a requirement is 5 times higher than expected resources under the assumption of the flat budget scenario. Although many software improvements have been achieved, the gap between the requirement and the expectation is still large. To fill the gap, the availability of computing resources must be improved. One of the possibilities is to use GPGPU, which requires additional software development. There are also some R&D to use external resources of commercial cloud, HPC, or volunteer computing resources.

To expand Tokyo regional analysis center, R&D project using commercial cloud resource was launched. The batch system of the center for WLCG is managed by HTCondor. The first R&D was started to deploy worker nodes of HTCondor on Google Cloud Platform. The system is a hybrid system that worker nodes are on the cloud while header nodes and file storages consist of on-premises resources. To reduce the cost, Google Cloud Platform provides preemptible instances, which has a running time limit of 24 hours. To use preemptible resources, a load balancer called GCP_Condor_Pool_Manager (GCPM) has been developed. GCPM checks HTCondor's waiting queues and create new instances on demand. Instances are deleted after one job is executed. By this procedure, the system can use preemptible instances effectively. The system is set up by Puppet and it is easy to set up a similar system in other places.

In this presentation, the current status of the R&D project and our experiences of usage of Google Cloud Platform will be presented.
.

Networking, Security, Infrastructure & Operations / **37**

WISE Information Security for collaborating e-Infrastructures

Author: David Kelsey[1]

[1] *STFC-RAL*

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

The WISE community enhances 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-infrastructures 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, WLCG, 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. This year we have some new working groups which have recently started their work. While many of the working group activities are performed by conference calls and e-mail, experience has shown that we can also make very good progress by holding face to face WISE events. These events, which typically attract between 20 and 40 participants, are held at least twice a year.
This talk will present an overview of the active working groups together with details of published guidelines and recommendations. Activities currently include the trust framework called Security for Collaborating Infrastructures, recent work on a new baseline Acceptable Use Policy, other security policy templates, the sharing of threat intelligence and issues related to the security of high throughput data transfers.

Humanities, Arts, and Social Sciences Application / **39**

Digital Encyclopedias, Federated Data, Crowd Sourcing, and Deep Mapping

Author: David J Bodenhamer[1]

[1] *The Polis Center, IUPUI*

In 1994, the Polis Center at Indiana University-Purdue University Indianapolis published the Encyclopedia of Indianapolis. The web was in its infancy, with the first web browser, Mosaic, appearing the same year. The web is now robust, ubiquitous, and mature enough to justify an online version, and the city's approaching bicentennial in 2020-21 provides the occasion to develop one. Discussions with community leaders, teachers, researchers, public officials, and citizens suggests that the

city needs more than an updated EOI, however. It also needs a way to integrate and access the explosion and fragmentation of knowledge created both as born- digital information and as large new digital archives currently siloed in the city's heritage, cultural, and educational institutions. Also, residents no longer look to experts to tell them about their city and its history; they now demand to participate in creating the information that will tell their stories and be useable in various place-making and redevelopment initiatives. A digital encyclopedia must accommodate this need as well, which means it must be a multifaceted knowledge platform.

To meet this need, the Polis Center is developing a distributed data platform to federate existing repositories and draw from them dynamically, as well as to allow citizen contributions to enrich what scholars and others report about the city. The aim is to produce a deep map of Indianapolis, with qualitative and quantitative data presented in a way that allows researchers and citizens alike to view the city in all its facets. Tools will exit within the platform to permit easier transformation, management, and visualization of the archived and contributed data. The deep mapping engine will present results within their spatial context in a way that supports different perspectives and manages both expert and native knowledge.

The presentation will outline the requirements for the integrated data platform as well as the resulting system schema. It will suggest how technologies such as GIS may be linked with other non-spatial modules to construct a dynamic interdisciplinary virtual research environment for experts as well as a system that invites volunteered information and citizen participation. Finally, the presentation will invite both ideas and collaboration from attendees at the conference. Ultimately, digital encyclopedias are collaborative ventures, and this presentation will seek to model how such collaborations may cross organizations and disciplines to create a dynamic and sustainable knowledge platform.

Networking, Security, Infrastructure & Operations / 40

A Blueprint of Log Based Monitoring and Diagnosing Framework in Large Distributed Environments

Author: Yining Zhao[1]

Co-authors: Haili XIAO [2]; Xiaodong Wang [1]; Xuebin Chi [1]

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

[2] *Supercomputing Center, Chinese Academy of Sciences*

Corresponding Author: zhaoyn@sccas.cn

Distributed systems have grown larger and larger since this concept appears, and they soon evolve to environments that contain heterogeneous components playing different roles, e.g. data centers and computing units. From security point of view, it is a difficult task to get an idea of how such large environment works or if any undesired matters happened. Logs, produced by devices, sub-systems and running processes, are a very important source to help system maintainers to get relative security knowledge. But there are too many logs and too many kinds of logs to deal with, which makes manual checking impossible.

CNGrid is a good example of a large distributed environment. It is composed of 19 HPC clusters contributed by many research institutes and universities throughout China. Computer Network Information Center of Chinese Academy of Sciences is the operation and management center of CNGrid, and is responsible for keeping the environment running smoothly and efficiently. The maintenance team has been working for years on using logs generated in CNGrid to help us analyzing system behaviors and addressing sources of failures.

In this work we will share some of our experiences by representing a general framework that monitors events, analyzes hidden information and diagnoses the healthy state for large distributed computing environments bases on logs. There are 4 major steps in this framework: 1) identifying necessary logs that are produced by key modules in the environment; 2) classifying these logs either using variables or non-variable contents to obtain a set of log types; 3) performing analyses from various

angels such as user behaviors, type associations, etc., using data mining and machine learning techniques with the help of the obtained set of log types; 4) gathering results from previous analyses and find some metrics that can numerically represents the vulnerability level of the environment, and produce a diagnosis report.

We will use CNGrid as an example to demonstrate these 4 steps. First, we pick up OS system logs and SCE middleware logs as the target to be analyzed, according to our demands of responding to system failures, defending on malicious attacks and providing user services. Second, log pattern extraction algorithms are used to classify types for system logs, and SCE logs are grouped by usernames. We organize these log types as the log library and implement interfaces to access it. Third, by applying the log library, a number of log analyzing applications are developed to perform analyses, including log flow detection, fault prediction bases on log type associations and user behavior analysis. Finally, results of these analyses would be turned into some measurable factors, so that they can be combined to reach an overall conclusion. We plan to weight these factors so that the produced diagnosis report is more adequate to describe the vulnerability level of the environment.

Although the framework we initially designed was for the maintenance for CNGrid, it is believed that the process is adaptable to other distributed computing environments.

VRE / 41

Apache Airavata: A Science Gateways Framework

Author: Rob Quick[1]

[1] *Indiana University*

Apache Airavata is a distributed component-based software system used to build and operate science gateways, sometimes referred to as virtual research environments (VREs) or science portals. Science gateways provide science-centric user environments and cyberinfrastructure middleware that enable broader and more effective use of scientific computing resources, applications, and data to accelerate scientific discovery.

Apache Airavata is deployed as a secured, multi-tenanted, scalable, fault-tolerant platform serving over thirty science gateways in a production environment operated by the Science Gateways Research Center at Indiana University. Airavata follows an API-first approach implemented using Apache Thrift, which gives Airavata a strongly typed, language independent way of defining its programming interfaces. Airavata's execution engine, recently re-implemented with Apache Helix, can manage pipelines and graphs for the remote execution of data analysis and scientific applications on computing clouds, supercomputers, clusters, and computational grids. Airavata security infrastructure enables users to use the gateways with authentication systems federated across the world and empowers collaboration of data, computational resources enforced through a controlled sharing environment.

The Airavata framework has evolved over 15 years with frequent design improvements leading to an evolutionary architecture that allows new capabilities to be added while ensuring backward support and reliable production operations. Airavata gained experience from integrating and operating gateways through the full life cycle of development, integration and operation.

Use of the Airavata framework allows gateways to be created and maintained with shared operational effort, thus allowing more than thirty production instances to be operated within a single unit at Indiana University with minimal effort per instance. The sharing of operational load within a single software framework allows quick spin-up of new gateways on request at a low cost to researchers while maintaining production quality stability and reliability with minimal operational overhead.

In this talk we will illustrate Airavata capabilities through the discussion of usage vignettes, which highlights the wide diversity of science disciplines benefiting from Airavata-based services. These include, but are not limited to, computational chemistry, experimental biophysics data analysis, human vascular system modeling, transcriptomics, food-energy-water systems modeling, and geological sciences. As an Apache Software Foundation project, Airavata's open community governance model is as important as its software base. We discuss how this works within Airavata and how it may be applicable to other distributed computing infrastructure and cyberinfrastructure efforts. We will also discuss Airavata service management and business model implemented by the Science Gateways Research Center.

Humanities, Arts, and Social Sciences Application / 42

Internet of Things technology and applications for Smart cities

Author: Tumen-Ulzii NARANMANDAKH[1]

[1] *Chief-Secretary, Communications Regulatory Commission of Mongolia (CRC) and Assoc.Prof. of Mongolian University of Science and Technology (MUST)*

Corresponding Author: naran@crc.gov.mn

1. Introduction and Digital Economy
2. Key emerging technologies and IoT (Internet of Things) for building Smart cities
3. Policy and Regulatory frameworks for Smart cities
4. Study on international activities and best practices
5. Common challenges and key strategies on IoT for Smart cities
6. Way to forward-summary of Recommendations

Summary:

Now more than ever, the world is looking into cities to help achieve a sustainable future. In light of projections stating that two-thirds of humanity will be living in cities by 2050 and the anticipated challenges this brings about, cities will definitely have to lead the way in addressing present needs without compromising the future.
Building Smart City has become an important trend globally. In the past, telecom operators were already crucial partners in terms of city infrastructure. ICT sector is including IoT technology and applications often recognized as one of the enablers of Smart cities (SC) and its strategic use help cities become inclusive, safe, and resilient.
The potential of Smart cities is nearly limitless. The capabilities of SCs will be enacted not only by traditional ICTs, but also by advanced emerging trends 4-th industrial revolution, 3-rd wave of digital transformation including key technology such as Internet of Things (IoT), AI, Radio-Frequency Identification (RFID), Machine-to-Machine (M2M) communications, Bluetooth®, Cloud Computing and Big Data. The IoT is fundamentally changing the business and drives convergence between ICT and industries. Quite simply, SCs use IoT devices such as connected sensors, lights, and meters to collect and analyze data. The cities then use this data to improve infrastructure, public utilities and services, and more. With IoT becoming basic communication facilities in every city, the role of telecom operators cannot be overemphasized in Smart City development.

This paper has several aims: a) the presentation of a critical analysis of the terms "smart sustainable cities" and "Internet of things IoT" b)paper also attempts to define key emerging technology with IoT technology development trends, and challenges of national developments from various aspects ranging from standards and KPIs for measuring Smart Cities to design and implementing architectures, and IoT services & applications for Smart Cities. c) Analysis of policy and regulatory frameworks for planning and building SC d) Study on activities and best practices related to the ITU and international organizations, partnership projects for working to develop the tremendous potential ICTs have to help build smarter, more sustainable cities with other. U4SSC develops international standards and guidelines to enable the coordinated development of IoT technologies and their application in SCs. Recently, many

cities including Dubai, Singapore, Moscow, Taipei, Valencia, Montevideo, Maldonado, Pully and Rimini have asked ITU for assistance in the implementation of the U4SSC KPIs.
The creation of SCs require a trusted infrastructure capable of supporting an enormous volume of ICT-based applications and services by using IoT, which in turn requires coordinated adherence to common standards that ensure openness and interoperability. It is essential that next-generation urban systems including IoT are conceived with cybersecurity and data protection.
Finally, paper proposed the way to forward-set of recommendations on improvement of Government and public service delivery mechanisms including Smart city initiatives, ICT development trend impacts (IoT), common challenges on key strategies (i.e.integrated management, safety net-cyber security, the open and inclusive architecture) for building Smart cities.

Supercomputing, High Throughput Computing, Accelerator Technologies, and their Integration / 43

Resource Access System in High-performance Computing Environment on Third-party Application Platform

Author: Rong He[1]

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

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

Corresponding Author: herong@sccas.cn

The national-level supercomputing environment has integrated 19 supercomputing centers such as major domestic supercomputing centers, with a total computing capacity of 200+ PF and storage capacity of 200+ PB. Supported by the national key R&D plan, the environment has a certain foundation in computing platform, environmental monitoring and other aspects. Among them, software application resources are various, including chemistry, mathematics and so on. Computing platform has its own account management system. With the rapid development of high-performance computing (hpc), more and more users aware that using hpc can solve their problems. Moreover, the needs of accounts are also expanding. Therefore, third-party application platforms want to use hpc environment and resources in it. Because the account management system of each platform and the way of using it are different, it is necessary to research and develop an access system to make the platform fit the environment well. This paper focuses on how the accounts in the platform are used in the environment and how they make use of the resources. At present, the environment provides API interfaces for external invocation. Thus third-party platforms need to consider the authorization of API interfaces. In order to solve this problem, this paper proposes a simple resource access technology and develops an access system using this technology. As for account correspondence, we set a mapping between platform account and grid account. Accounts that already have grid accounts are directly bound with their grid accounts. Those without grid accounts can quickly create a grid account by filling in the account name and contact information. It is bound to the platform account at the meaning while. Accounts can directly enter the high-performance computing environment platform to submit jobs and use resources. To authorize the use of resources, third-party platforms need to apply for API interface access, focusing on APPID and APPKEY. Then, according to the permissions of each API interface, platforms will open query permissions default. Finally, according to different levels of platforms and the source of platform accounts different permissions will be set. This access system has been validated in China's Science and Technology Cloud (CSTCloud). Users who have logged through the CSTCloud can directly enter the environmental platform. This system shows that the access system we designed is feasible and the account problem and resource authorization are validated effectively. In the following educational platform will also use this access system to dock with our hpc environment. In short, the resource access system proposed in this paper is feasible. It plays an important role in accessing the hpc environment on third-party application platform and also makes the hpc to develop sustainably。

Networking, Security, Infrastructure & Operations / 44

Smooth migration of a feature-rich vulnerability analysis engine within a security portal for users with divergent skill levels

Author: Tadashi Murakami[1]

[1] *High Energy Accelerator Research Organization (KEK)*

Corresponding Author: tadashi.murakami@kek.jp

Vulnerability management is useful for maintaining security with keeping the flexibility of the network environment, especially for the DMZ network that allows connections from the Internet.
We have been operating a vulnerability management portal site named DMZ User's Portal for 13 years.
In KEK, all of the host administrators in DMZ network (DMZ admin) have their own accounts for the portal site, and they can manage the vulnerabilities by themselves.
Moreover, the portal site was also adopted to other two sites with different operation policy, and they are also now in operation.

In DMZ User's Portal, we have adopted the same series of vulnerability analysis engine which has many advantages, and it has offered a lot of benefits to us for 13 years.
Now, all of DMZ admins came to deal with the serious vulnerabilities detected by the engine promptly.
On the other hand, the inspection performance of the engine gradually came to be degraded.
Now, we decided to replace the engine into a more powerful and complex one.

In the replacement, it is desirable to continue the successful experiences and contributions within the portal site.
One of the important contributions covers divergent skill levels among DMZ admins in KEK, by simplification of the intricate usage of the vulnerability analysis engine that is designed for network security experts.
The other is that the portal site has accumulated a lot of know-how gained from the operation of the portal site that includes user-response of DMZ admins.
It is a difficult task to continue these contributions of the portal site when replacing the vulnerability analysis engine, without the design and development of the modules carefully in advance.

This talk presents the design and methods for smooth migration of a feature-rich vulnerability analysis engine within the security portal site mentioned above.
The key point is that module dependency has been carefully considered among essential functions of vulnerability analysis engine, web interface, email notifier, and database.
To achieve the lower degree of module dependency, the techniques of O/R mapping, code generation, wrapper architecture, template engine consolidation, and test case were leveraged.
The combinational use of these techniques brought not only modularity but also maintainability to the modules for the essential functions above.
Consequently, it enabled us to replace the vulnerability analysis engine with minor modifications of user interfaces within web and email.
In this way, we can continue to operate the portal site with inheriting the successful experiences, along with gaining the benefits of a new powerful vulnerability analysis engine.

Physics & Engineering Application / 45

Integration of the Italian cache federation within CMS computing model

Authors: Antonio Falabella[1]; Daniele Cesini[2]; Diego Ciangottini[3]; Enrico Mazzoni[4]; Giacinto Donvito[1]; Giuseppe Bagliesi[4]; Massimo Biasotto[5]; Mirco Tracolli[3]; daniele spiga[6]; tommaso boccali[1]

[1] *INFN*

[2] *INFN-CNAF*

[3] *INFN Perugia*

[4] *INFN Pisa*

[5] *INFN Legnaro*

[6] *INFN-PG*

The next decades at HL-LHC will be characterized by a huge increase of both storage and computing requirements. A factor 20 is expected for the storage, while on computing the estimation is about a 30x CPUs. Moreover, we foresee a shift on resources provisioning towards the exploitation of dynamic (on private or public cloud and HPC facilities) solutions. In this scenario, the computing model of the CMS experiment is pushed towards an evolution for the optimization of the amount of space that is managed centrally and the CPU efficiency of the jobs that run on “storage-less” resources. In particular, the computing resources of the “Tier2” sites layer, for the most part, can be instrumented to read data from a geographically distributed cache storage based on unmanaged resources, reducing, in this way, the operational efforts by a large fraction and generating additional flexibility.
One of the benefits behind a distributed cache space is the possibility to leverage the national high bandwidth network from NRENs to reduce the number of file replica around the WLCG sites. The cache system will appear as a distributed and shared file system populated with the most requested data; in case of missing information data access will fallback to the remote access.

Moreover in a possible future scenario based on the data-lake model, it is reasonable to imagine that many satellite computing centers might appear and disappear dynamically. In this sense, a protection layer against a central managed storage might be a key factor along with the control of the data access latency. A Content Delivery Network has many affinities with the cache architecture desired and, in each region, duplicate files within its federated servers are not foreseen. The cache storages will be by definition “non-custodial”, thus reducing the overall operational costs.

The objective of this contribution is to present the first implementation of an INFN federation of cache servers, developed also in collaboration with the eXtreme-DataCloud EU project. The CNAF Tier-1 plus Bari and Legnaro Tier-2s provide unmanaged storages which have been organized under a common namespace. This distributed cache federation has been seamlessly integrated with the CMS computing infrastructure.

The technical implementation of this solution is based on XRootD, largely adopted in the CMS computing model under the “Anydata, Anytime, Anywhere project” (AAA). The technology is compliant with the most common storage protocol present in the WLCG and several activities already demonstrated the effective management of cold and warm cache scenarios. Moreover, the possibility to plug custom caching algorithms will provide the ability of playing with historical and online data access metrics for smart data placement decision.
The results in terms of CMS workflows performances will be shown. In addition, a complete simulation of the effects of the described model under several scenarios, including dynamic hybrid cloud resource provisioning, will be discussed. Finally, a plan for the upgrade of such a prototype towards a stable INFN setup seamlessly integrated with the production CMS computing infrastructure will be discussed.

Biomedicine & Life Science Application / 46

CryoEM image processing using ASGC integrated environment

Author: Kuen-Phon Wu[1]

Co-authors: Chia-Yang Hsu [1]; Eric Yen [1]; Felix Lee [1]; Tsung-Hsun Wu [1]

[1] *Academia Sinica*

Corresponding Author: kpwu@gate.sinica.edu.tw

CryoEM single particle analysis often needs millions individual particles to successfully reconstruct the 3-dimentional map of interest target. High performance computational resources play crucial roles in accelerating 3D EM map reconstructions. With the implementation of powerful Nvidia

graphic cards in popular cryoEM software such as RELION and cryoSPARC, time to obtain refined and sharpened 3D map has been significantly reduced. Although both software and hardware are highly improved to reduce time for whole reconstruction process, accessing high-end GPU workstation or large-scale CPU cluster remains as major problem for users. It might be affordable for one laboratory to build its own GPU workstations or min-sized clusters, however, substantial routine management and issues of cyber security will be the burden for a 5-10 person laboratory. Moreover, monster-scale data transfer and storage of cryoEM dataset (1-4TB) will be the headache of each laboratory to maintain high-speed bandwidth and enough backup space.

The Academia Sinica Grid Center (ASGC), the first tier 1 level grid center in Asia, is proud to provide cloud computation for the cryoEM society in Taiwan. The ASGC team closely collaborates with cryoEM users at Academia Sinica to build grid instances for cryoEM-related computation including 3 most popular software: RELION, cisTEM and cryoSPARC, supported by 200+ CPUs, 64+ GPUs and 1TB+ memory in total. Currently, 8 RELION 2, 8 RELION 3, 10 cisTEM, and 2 cryoSPARC instances are available for public. There will be more grid instances for cryoEM-related software built in the near future. This talk will briefly describe the system environment, user application, workflow and benchmarks of cryoEM software at ASGC.

Infrastructure Clouds and Virtualisation / 48

Container Practices at IHEP

Author: Wei Zheng[1]

Co-authors: Jingyan Shi [2]; Qiulan Huang [3]; Xiaofei Yan [2]; Yaodong CHENG [4]

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

[2] *IHEP*

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

[4] *IHEP, CAS*

Corresponding Author: zhengw@ihep.ac.cn

Based on our experiments demands, especially for BES and LHAASO, Institute of High Energy Physics, Chinese Academy of Sciences (IHEP) has researched and deployed container technology based on singularity and docker. This presentation will introduce the current status of practices at IHEP□
We developed a container program that provides users with uniform access to the container, the location of the containers is transparent to the user, which is more than advantageous for the deployment and update of the container and is convenient for the user to use. Additionally, we developed and optimized different container image and storage mounting strategies for different experimental needs,. For example, the experimental SL55 container offered instead of the physical machine for BES experiments. BES Users can compile and submit jobs in the container just like physical machines. In LHAASO experimental cluster, which is located in Daocheng, Sichuan province (at the altitude of 4410 m), in order to provide a more stable and highly available login environment, we built a SL7-based login node docker container image, and through Kubernetes implements dynamic expansion and load balancing of the login node. And we started to support using containers with batch systems, the user's job can be dispatched to the container ,so that jobs can have exactly the same environment no matter what site they're running on. Finally some site services, such as distributed site monitor, are using docker to deploy and run more stable.

Data Management & Big Data / 49

The eXtreme-DataCloud project: advanced data management services for distributed e-infrastructures

Authors: Alessandro Costantini[1]; Christian Ohmann[2]; Daniele Cesini[1]; Doina Cristina Duma[3]; Fernando Aguilar Gomez[4]; Giacinto Donvito[5]; Luca dell'Agnello[5]; Lukasz Dutka[6]; Matthew Viljolen[7]; Oliver Keeble[8]; Patrick Fuhrmann[9]; Rachid Lemrani[10]; Serena Battaglia[2]; Vincent Poireau[11]

[1] *INFN-CNAF*

[2] *ECRIN*

[3] *INFN - CNAF*

[4] *IFCA*

[5] *INFN*

[6] *AGH/CYFRONET*

[7] *EGI Foundation*

[8] *CERN*

[9] *DESY/dCache.org*

[10] *IN2P3*

[11] *IN2P3-LAPP*

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

The eXtreme-DataCloud project (XDC) is a software development initiative aimed at implementing data management scalable services to address the following high level topics: policy driven data management based on Quality-of-Service, Data Life-cycle management, storage federations creation, smart placement of data with caching mechanisms, meta-data with no predefined schema handling, execution of pre-processing applications during ingestion, data management and protection of sensitive data in distributed e-infrastructures. The project is driven by user communities belonging to different scientific domains: High Energy Physics (WLCG), Astronomy (CTA and LSST), Photon and Life Science (XFEL and LifeWatch), Medical research (ECRIN). XDC is funded by the European Commission under the Horizon 2020 framework program, it started in November 2017 and its first major release was launched at the end of 2018. This release is based on a toolbox composed by well known, production quality services that have been enriched with new functionalities. The list of services include dCache, ONEDATA, EOS, FTS, Indigo-Orchestrator, Indigo-CDMI server and Dynafed. All the newly implemented functionalities can be easily plugged into the existing e-infrastructures but, given the use of standard protocols and authorization mechanisms, can be used also as building blocks of the new generation scientific infrastructures. This contribution will introduce the project, its overall architecture and the first release main features. Some use cases addressed by the project developments will also be presented.

VRE / **50**

Workflow management in DIRAC interware

Authors: Andrei Tsaregorodtsev[1]; Fabio Hernandez[2]; Johan Bregeon[3]; Pierre Gay[4]; Sorina Pop[5]; Vanessa Hamar[2]; luisa arrabito[6]

[1] *CPPM-IN2P3-CNRS*

[2] *CC-IN2P3/CNRS France*

[3] *LUPM IN2P3/CNRS France*

[4] *Université de Bordeaux France*

[5] *CREATIS/CNRS France*

[6] *LUPM IN2P3/CNRS*

Corresponding Authors: arrabito@in2p3.fr, atsareg@in2p3.fr

DIRAC interware is a layer between users communities and computing infrastructures. Scientific communities of different domains (high energy physics, astrophysics and biomedical) adopted DIRAC as workload and data management system in a distributed computing environment, mainly grids and

clouds. The DIRAC Workload Management System handles the whole job life cycle, allowing an efficient usage of distributed and heterogeneous resources through the pilot mechanism.
Moreover, the DIRAC Transformation System is dedicated to the management of large productions, i.e. several hundreds of thousands jobs, processing large amounts of data, through a data-driven mechanism.
However, scientific workflows are usually composed of several processing steps that are treated independently by the Transformation System.
In order to automatize the execution of whole workflows, we have developed a high-level fully data-driven DIRAC system, called Production System. In this contribution we describe the Production System architecture as well as its first application to the Monte Carlo productions of the Cherenkov Telescope Array.

Physics & Engineering Application / 51

Simulation of the cache hit rate for data readout at the Tokyo Tier-2 center

Author: Tomoe Kishimoto[1]

Co-authors: Junichi Tanaka [1]; Michiru Kaneda [2]; Nagataka Matsui [1]; Tetsuro Mashimo [1]

[1] *University of Tokyo*

[2] *ICEPP, the University of Tokyo*

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

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

We replaced almost all hardware devices in every three years in order to satisfy the requirement of the ATLAS experiment. The next hardware replacement will be performed in December 2018. In the current system, 6144 CPU cores (256 worker nodes) and 7392 TB disk storages are reserved for the ATLAS experiment. The disk storage consists of 48 sets of a disk array and a file server, where each disk array consists of 24 SATA HDDs. The worker nodes and the file servers are connected to a central network switch. The internal network bandwidth between the worker nodes and the central switch is 1040 Gbps in total (10 Gbps × 104 links). The file servers are also connected to the central switch by 480 Gbps in total (10 Gbps × 48 links).

We recently observe that the total throughput of data readout from the disk storage to the worker nodes is limited to about 100 Gbps in the current system even though the enough internal network bandwidth is available. The I/O performance of the disk arrays is one of the reasons of this throughput limitation because the disk array I/O utilization, which is measured by iostat command in Linux, is saturated during the heavy readout. The load of the data readout should increase in the next system because the number of CPU cores will increase, while the number of disk arrays will be the same with the current system due to the limited space of server racks. Therefore, we are discussing about a possibility to introduce a cache hierarchy for the storage system using SSDs to improve the I/O performance in the future system. However, we can not build an efficient caching system without the knowledge of the data readout patterns. The cache hit rate, which is defined as the number of data readout from cache area divided by the total number of data readout, is a good guideline to evaluate the data readout patterns for the caching system. In this presentation, simulation results of the cache hit rate will be reported. The simulation is performed using the real data access logs in the storage element at the Tokyo Tier2 center. We will discuss whether we can build an efficient cache system based on the simulation results.

Supercomputing, High Throughput Computing, Accelerator Technologies, and their Integration / 52

A Resource-saving Job Monitoring System of High-Performance Computing using Parent and Child Process

Author: Kajornsak Piyoungkorn[1]

Co-authors: Chalee Vorakulpipat [1]; Phithak Thaenkaew [1]

[1] *NECTEC*

Corresponding Author: chalee.vorakulpipat@nectec.or.th

High-performance computing has been more important in the past decade. In the present day, data used for processing becomes enormous where a high-performance computing resource is needed to help process the data. Some scientific experiments involving big data which requires high-speed data processing cannot be done by an ordinary computer system. Also, there is a need for support of parallel processing. The solution starts by dividing the job into a number of sections, and then the system sends the calculated result back to the compiled. This mechanism will speed up the processing time to complete the task and generate more output at the same time. Therefore, a solution in this study is to maximize efficiency when using the resources of the computer which involves the processing power of the processor (CPU-Core).

The mechanism can be explained in a scenario as follows. In Thailand, there are government agencies that provide free-of-charge high-performance computing services to facilitate researchers to conduct their research.It is usual that many users request a resource overrun or uses computing resources in an inefficient way.This is because they are not aware of the over-consumption of the resources, leading to unnecessary high costs. For example, a user requests computing resources that does not match the actual usage.Resource requests are calculated in high numbers for maximum processing speed that does not correspond to actual usage, resulting in resource wasting. Thus, a negative effect will go to the hardware system and it will be a hindrance to other users who have to lose an opportunity to use it.

In our previous study, demonstrated in Figure 1, we start checking all running jobs in the system from Job ID number. The CPU-Load that represents the CPU utilization is over or not 100%. Process ID is then analyzed by the Job ID.Displays user details of how many CPU-Core requests are made.Then, it compares with the current CPU-Load that works exactly as requested.We use the tolerance of 20%. If the current resource has a CPU-Load greater than 120% or less than 80%, it will be assumed that the work is running, and the CPU-Core request does not match the actual use. It means that the use of resources is not effective. The system alerts users via email or eliminates the process. However, using this mechanism, the CPU-Load is not correctly analyzed because some software uses unstable CPU-Core. As a result, the decision of the system is wrong.

Networking, Security, Infrastructure & Operations / 53

Architecture of Resource Manager for Software-Defined IT Infrastructure

Authors: Shinji SHIMOJO[None]; Susumu DATE[None]; Yasuhiro Watashiba[1]; Yoshiyuki KIDO[None]; Yuki Matsui[2]

[1] *Nara Institute of Science and Technology*

[2] *Osaka-University,japan*

The concept of Information-as-a-Service (InfaaS) is a critical concept for disaster management applications. In the disaster management, the flow of information and the synchronization of data between different sites must be maintained to facilitate the decision making process. Thus, in order for everyone involved to see the same information at the same time, an application is needed to visualize different types of data and synchronize the information among different sites. The visualization system requires an IT Infrastructure that can compensate partial lacks of components and can adapt to

suddenly changing environments caused by a disaster. As a technology that satisfies these requirements, we have focused on Software-Defined techniques that is to flexibly control a system by software. Software-Defined techniques have potential to bring the functionalities of the needed flexibility and resilience to existing IT Infrastructures. Therefore, we have been studying and developing the Software-Defined IT Infrastructure technology to realize a distributed visualization system for disaster management applications.

The distributed visualization system with the Software-Defined IT Infrastructure is composed of three components: servers to visualize various data, large-scale display systems for presenting information to users such as Tiled Display Wall and network for connecting among sites. These components need to sustain their functionalities even when they are suffered by a disaster. When a disaster makes a certain site inoperable, it is necessary to migrate the resources, servers and the large-scale display systems, running on the site to another site that did not suffer a disaster. Moreover, network connections have to be reconfigured for connecting to the migrated resources. In order to build a system with the functionalities, we have adopted two technologies in our distributed visualization system with the Software-Defined IT Infrastructure. For the purpose of facilitating migrate and deploy, the servers are virtualized by Docker. In order to be reconfigured easily, the network is constructed as an overlay network by Software-Defined Networking (SDN). SDN simplified the logical configuration change of the network by supervising the dataflow.

Currently, such components have behaved independently. However, for appropriately performing the functionalities for various situation caused by a disaster, it is essential to link each the components mutually. The servers are necessary to control migrating and deploying in consideration from availability of each the sites. The overlay network need to be set to maintain dataflow by an SDN controller, which is software that acts as a dataflow control point in the SDN. Therefore, a resource manager dealing with behavior of these components comprehensively is required. In this poster, we show an architecture of the resource manager for building the distributed visualization system with Software-Defined IT Infrastructure applied.

Keywords: Software-Defined IT infrastructure, resource manager, disaster management application, InfaaS

Networking, Security, Infrastructure & Operations / 54

Collection and harmonization of system logs and prototypal Analytics services with the Elastic (ELK) suite at the INFN-CNAF computing centre

Author: Tommaso Diotalevi[1]

Co-authors: Antonio Falabella [2]; Daniele Bonacorsi [3]; Diego Michelotto [2]

[1] *INFN and University of Bologna*

[2] *INFN-CNAF*

[3] *University of Bologna*

The distributed Grid infrastructure for High-Energy Physics experiments at the Large Hadron Collider (LHC) in Geneva comprises a set of computing centers, as part of the Worldwide LHC Computing Grid (WLCG). The Tier-1 level functionalities in Italy are served by the INFN-CNAF data centre, which actually serves also more than twenty non-LHC experiments. A key challenge is the modernisation of the center to be able to cope with the increasing flux of data expected in the near future.

High-level standards of operation require a continuous work towards full understanding of service behaviours and a constant seek for higher level of automation and optimization. Data centers worldwide witness the use of Artificial Intelligence (AI) to push data centers into a new phase, in which tasks traditionally managed by operators could be more efficiently managed by human-supervisioned algorithms and techniques. Besides, CNAF collects a high amount of logs every day from various sources, which are highly heterogeneous and difficult to harmonise: such log data are archived but almost never used, except for specific internal debugging and hardware monitoring operations.

In this contribution, a working implementation of a system that collects, parses and displays the log information from CNAF data sources, with analytics functionalities, is presented.

The open source ELK software suite, including Elasticsearch, is used for the log ingestion and transformation, as well as for creating a centralised and structured database to organize in a clean and ordered manner the CNAF logs, including the creation of new visualisations and dashboards that offer online monitoring functionalities. This infrastructure is then vital for the CNAF long-term goal of modernizing the centre via machine learning based predictive maintenance approaches, moving away from preventive replacements of equipment, which is highly expensive and far from optimal efficiency.

Physics & Engineering Application / 55

Boosting the CMS computing efficiency for data taking at higher luminosities

Author: Leonardo Cristella[1]

Co-authors: Giacinto Donvito [2]; Giorgio Pietro Maggi [3]

[1] *INFN Section of Bari*

[2] *INFN*

[3] *Politecnico di Bari & INFN Section of Bari*

Thousands of physicists continuously analyze data collected by the Compact Muon Solenoid (CMS) experiment at the Large Hadron Collider (LHC) using the CMS Remote Analysis Builder (CRAB) and the CMS global pool to exploit the resources of the World LHC Computing Grid. Efficient use of such an extensive and expensive system is crucial and the previous design needs to be upgraded for the next data taking at unprecedented luminosities. Supporting varied workflows while preserving efficient resource usage poses special challenges, like: scheduling of jobs in a multicore/pilot model where several single core jobs with an undefined runtime run inside pilot jobs with a fixed lifetime; avoiding that too many concurrent reads from same storage push jobs into I/O wait mode making CPU cycles go idle; monitor user activity to detect low-efficiency workflows and automatically provide them with advices for a smarter usage of the resources tailored on the use case.

In this talk we report on two novel complementary approaches adopted in CMS to improve the scheduling efficiency of user analysis jobs: job automatic splitting, and job automatic estimated running time tuning. They both aim at finding an appropriate value for the scheduling runtime. With the automatic splitting mechanism, an estimation of the runtime of the jobs is performed upfront so that an appropriate value can be estimated for the scheduling runtime. With the automatic time tuning mechanism instead, the scheduling runtime is dynamically modified by analyzing the real runtime of jobs after they finish. We also report on how we used the flexibility of the global computing pool to tune the amount, kind, and running locations of jobs exploiting remote access to the input data. We discuss the strategies, concepts, details, and operational experiences, highlighting the pros and cons, and we show how such efforts have helped to improve the computing efficiency in CMS.

Keyword
Computing efficiency, High luminosity, CMS

Data Management & Big Data / 56

Towards Predictive Maintenance with Machine Learning at the INFN-CNAF computing centre

Author: Luca Giommi[1]

Co-author: Daniele Bonacorsi [2]

[1] *INFN and University of Bologna*

[2] *University of Bologna*

The INFN-CNAF computing center, one of the Worldwide LHC Computing Grid Tier-1 sites, is serving a large set of scientific communities, in High Energy Physics and beyond. In order to increase efficiency and to remain competitive in the long run, CNAF is launching various activities aiming at implementing a global predictive maintenance solution for the site.

This requires a site-wide effort in collecting, cleaning and structuring all possibly useful data coming from log files of the various Tier-1 services and systems, as a necessary step prior to designing machine learning based approaches for predictive maintenance.

Among the Tier-1 services, efficient storage systems are one of the key ingredients of Tier-1 operations. CNAF uses StoRM as a Grid Storage Resource Manager solution: its operations are logged in a very complex manner, as the log content is deeply unstructured and hard to be exploited for analytics purposes. Despite such difficulty, the StoRM logs are a precious source of information for operators (real-time monitoring, anomaly detection), for developers (debugging, service stability, code improvements) and for site managers (service optimization, storage usage efficiency, time and money saving ways to spot and prevent unwanted behaviours).

This work describes how the StoRM logs can be handled and parsed to extract the relevant information, how such log handling can be designed to work automatically, how to define and implement metrics to tag critical states of the service, how to correlate StoRM events with external services' events, and ultimately how to contribute to the future CNAF-wide predictive maintenance system.

First steps in this activity are presented and discussed, and a mention to complementary work in progress by other teams at the CNAF centre is also mentioned.

Networking, Security, Infrastructure & Operations / 57

A Study of Certificate Management Mechanism Suitable for Virtual Machine with Arbitrary Lifecycle

Author: Eisaku Sakane[1]

Co-author: Kento Aida [1]

[1] *National Institute of Informatics*

Virtual machine can flexibly meet user's demands for computing resources and be freely created and deleted. The virtual machine validation and secure communication as network entity are required as well as a physical machine. This paper investigates an X.509 certificate issuing mechanism to virtual machine with arbitrary lifetime.

Let us consider a service that offers virtual machines as computing resources. The provisioning service can create a virtual machine immediately according to the demands of users and boot up it. If user requires an X.509 certificate for secure communication over TLS in the Internet, the service will not be able to arrange the certificate issued by an ordinary certificate authority unless the service prepares the certificate in advance. It is necessary to carry out procedure for certificate issuance

separately. When the virtual machine shut down, certificate revocation will be needed unless the user plan to reuse it. The quick revocation must be needed even if the period of use is very shorter than the validity of the certificate. Thus, because of the flexibility of lifetime of virtual machine, certificate management following the ordinary certificate authority does not often fit the lifetime of virtual machine. Therefore it is worthwhile to investigate certificate management mechanism suitable for the flexibility of virtual machine.

In this paper, we consider a mechanism for restrictedly issuing a kind of intermediate CA certificate. The subject possessing the certificate does not act as an ordinary certificate authority and shall issue server certificates only to virtual machines managed by itself. By means of the mechanism an organization that offers virtual machines can issue certificates suitable for the flexible lifetime of the virtual machines. Moreover it is able to manage the certificate lifecycle based on the lifecycle of virtual machine. We also discuss the proposed mechanism, comparing with Let's encrypt approach that can be considered as another solution to the issues.

Infrastructure Clouds and Virtualisation / 58

Toward Single Sign-on Establishment for Inter-Cloud Environment

Author: Eisaku Sakane[1]

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

[1] *National Institute of Informatics*

As diversification of cloud services is making progress, a service-oriented approach is more important, in which services are chosen from multiple cloud vendors according to the demands of users. The purpose of this paper is to investigate a mechanism that establishes single sign-on for inter-cloud computing environment built as the optimized result of the needs of users.

Single sign-on mechanism is indispensable for inter-cloud computing environment that is composed of various services – each service is provided by different cloud vendor. In general a single sign-on mechanism works within cloud environment provided by a cloud vendor, however, a single sign-on mechanism that extends across multiple clouds is not established at the beginning of use. For example, an academic researcher who can obtain a SAML assertion from the home organization should be enabled to access a public cloud with the assertion. Since cloud vendors, of course, already supports major authentication technologies such as SAML and OAuth, technically the credential issued by the home organization will be usable for access to public clouds. However, it is often hard for the identity provider operated by the home organization to manage the user attributes that the cloud vendor requires, because the operating department of the IdP is responsible only for attributes that are assigned naturally in terms of the constitute member of organization. Therefore, establishment of single sign-on to public clouds is being hampered. Based on the credential issued by the home organization, a mechanism for handling necessary attributes information is needed, which should not impose a burden on administrators of the identity provider.

GakuNin Cloud Gateway Service (CGS) is a service portal that enables users to manage services available in academic research or education based on the user authentication information. It is developed and maintained by National Institute of Informatics in Japan. In cooperation with identity providers and service providers that participate in the GakuNin, an academic access management federation, the GakuNin CGS checks service providers that the identity provider allows of sending SAML assertions to based on user's SAML assertion provided by the identity provider. Thus it displays users the list of available services.

In this paper, we design a single sign-on mechanism for the inter-cloud environment. The basic idea is to delegate responsibility for suitable attribute assignment to trusted third party. The trusted third party assumes a role of sender that sends necessary attributes for the service combined with fundamental authentication information. For various services we consider requirements to the attribute assignment system. Thus, we discuss a model that realizes single sign-on in the inter-cloud environment by designing such functions on the GakuNin CGS.

Data Management & Big Data / 59

Big Data and A General Theory of Concept Lattice

Authors: Simon C. Lin[1]; Tsong-Ming Liaw[1]

[1] *ASGC*

Big data is instrumental for the triumph of deep machine learning, it generates enormous number of quasi-ground states for the learning algorithms to converge. In the Big Data Analytics, it was recognized that prediction and description are two generic goals in its field. Prediction often appears with the use of attributes to predict the membership of a particular object in some object set with similar attributes. Description looks for the attributes that describes the object, often it involves identifying a set of attributes that are shared by all objects in the set. However, general frameworks in this direction, such as Formal Concept Analysis and Rough Set Theory, are not only suffering from the size of the object-attributes pair but also the challenges to sort out the complex relationships within the objects and attributes in order to generate implication rules.

We have developed a General Concept Lattice (GCL) theory that is capable of flexible knowledge representation, discovery, Big Data Analytics and logic reasoning. The GCL also emerges as a Galois lattice, on which simultaneous partial ordering is stablished among 2-tuples represented by General_Concept = (General_Extent, General_Intent). Each General Extent is referred to as an Object Class recognized as distinct by the categorization and the General Intent its corresponding description of properties. In practice, on particular General Concepts of GCL one achieves that the Formal Concept Lattice-Intents contributes to a part of the lower bound and/or Rough Set Lattice the upper bound for the General Intents. The set inclusion relation between two General Extents induces the logic implications between members of the two corresponding Intents and implies that the property of an object set is essentially a part of the features of its super set. In particular, attributes grouped into the same Intent are logically equivalent since they correspond to the property of the same object class.

Such a general theory that is capable of treating object-attribute-logic of the data set is fundamental to the Data Science and Machine Learning. Achievements of this study will establish a new fundamental theory and framework for the future research and development of Computational Intelligence. Various fields of Data Analytics and Data Science may be revolutionized by the novel applications of GCL theory.

Physics & Engineering Application / 60

Overview of Database Framework for GEM Detector at CERN

Author: Muhammad Imran[1]

Co-authors: Adeel-ur Rehman [1]; Rao Atif Shad [1]

[1] *National Centre for Physics*

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

In this paper, we give an overview of the database framework which we have developed for the Gas Electron Multiplier (GEM) Detector at CERN. The GEM constitutes a powerful addition to the family of fast radiation detectors; originally developed for particle physics experiments, and has spawned a large number of developments and applications. The GEM database framework comprises four components. The first component is the database itself. There are two instances of the database which have been deployed. One is for the development purpose which has test data, and other one is for the production purpose which has real data. The database further comprises various schemas and each schema has different tables in it. We use separate schemas for various types of tables. The second component of the database framework is called DB Loader. The DB Loader is used to load

data into database. This has been written in the java language. The data is prepared in the predefined format which is in the XML form. Then the xml file is copied into a spool area of a server in which DB loader is running. Once the file is copied, the DB Loader loads the file into the database. The loader returns status codes after performing database insertion/updation operations. The status of the database operations is checked with the status code which is returned by the loader. The loader also accepts zip files and extract XML files from the zip file and loads them into database for batch data upload.

The third component of the database framework comprises graphical user interface (GUI). This is a web-based interface which can be accessed from the internet. This interface is used to generate XML files and send them to the DB Loader for data loading. This interface is basically used for the detector construction and to perform various quality control tests on the detector and its components. In the first stage, individual components of the detector are registered such as foils, electronic boards, readout boards, drift boards, VFATS, external frames, opto hybrids, cooling plate circuits, temperature sensors and radmon sensors etc. In the second stage, the chamber is constructed using these individual components. In the next step, super chamber is constructed using two chambers. The various quality control (QC) tests are performed on individual components, chambers and super chambers. The GUI is used to load the data for various QC tests. Currently, the GUI has the data loading facility from QC1 to QC8. The last component of the database framework is called online monitoring system (OMS). The OMS is data visualization framework for the various detectors of the CMS experiment at CERN. It is also used to display data for the GEM. It enables users to view and retrieve database contents without having to learn database specifics.

62

First steps towards light-weight storage

Author: Jiri Chudoba[1]

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

One of the difficulties with administration and operation of grid Storage Elements is a relatively large number of protocols used for communication with the SE and for data transfers. A long term activity leading towards reduction of used protocols is reflected by a recent version of DPM storage, which is one of the most widely used in WLCG. Three protocols are now used for data transfers: gridftp, http and xrootd. The SRM protocol, which is used for negotiation of transfer parameters, can be avoided in case where all data are online, i.e. stored on disks, which is the case for DPM instances, because DPM does not support a tape backend. The current DPM version 1.11 with enabled DOME component still supports SRM, but only in legacy mode; and developers announced end of SRM support in 2019. Several activities were started to test third party transfers using only http or xrootd protocol. We report our experience of one of the first relatively big Tier-2 site using DPM 1.10 (and later 1.11) version with DOME enabled. The test installation on a smaller site with lower traffic were not able to uncover several issues visible in the production environment where several TB of data are transferred every day.

Networking, Security, Infrastructure & Operations / 64

Divide and Conquer: Distributing delivery of large Security Challenge payloads

Authors: Jouke Roorda[1]; Sven Gabriel[2]

[1] *Nikhef*

[2] *Nikhef/EGI*

In Security Service Challenges the readiness of an infrastructure's incident response capability is assessed. Here we simulate a situation where a legitimate credential is used for activities violating various policies, requiring the involved security teams to take action in order to resolve the incident. An important part here is the containment of the malware, which would be easily doable if the attacker would use the standard grid job submission systems. Therefore, an external delivery system is needed.

For the EGI Security Service Challenge, we explored using BitTorrent as a way to reliably deliver the software needed to connect to the Tor network. A number of initial 'seedboxes' were set up in different countries, which announced their availability to a private 'tracker'.

This tracker, hidden as a Tor service, could be reached through either a public Tor proxy, or through the Tor network. As a fallback to the peer to peer delivery system, the torrent file also contained a number of 'web seeds', http(s) mirrors that are being served for the Tor Project.

As the BitTorrent client checks for data integrity, there is some level of guarantee that the software has been delivered correctly and the web seeds are valid. We used the aria2 download utility on the to-be infected machines as it is light weight, well maintained, hosted on the generally available GitHub, and written in such a way that it relies solely on generally available libraries and technologies. Aria2 was built on the machines themselves to account for subtle differences among the grid infrastructures across the sites.
In an attempt to hide most of the infected nodes, not all the systems will start seeding after they have finished downloading. This makes the forensics more of a challenge the infected nodes cannot just be requested from the tracker.

Networking, Security, Infrastructure & Operations / 69

Assessment of the incident response processes in a Distributed Infrastructure

Authors: Christophe HAEN[1]; Sven Gabriel[2]; Vincent BRILLAULT[3]

[1] *CERN*

[2] *Nikhef/EGI*

[3] *CERN/EGI*

EGI CSIRT provides operational security to distributed compute infrastructures coordinated by EGI. One of EGI CSIRTs activities is to assess the overall incident response capabilities, which is done through security exercises, so called Security Service Challenges (SSCs).
Operational security in an agile environment with different job management systems, logging information at different locations and entities coordination of the involved security teams is key.

The used services need to provide sufficient traceability of user actions as well as interfaces to the systems that offer methods needed to contain an incident, i.e. suspending of credentials found in activities violating approved security policies.

In an assessment of the overall incident response capabilities one of the core aspects will be the junctions between the acting security teams, each having a different view on the situation and different tools available to contain the incident.

To be able to run Security Service Challenges (SSCs) targeting multiple Resource Centres (RCs) and Workload Management Systems (WMS) a framework, the SSC-Monitor, was developed, that allows for central management of the malicious activities as well as for recording and evaluating the expected actions of the participating CSIRTs. This SSC-Monitor was already used for earlier SSCs. While large parts of this framework remains constant several adoptions are needed to implement the Virtual Organisations WMS. Also to be able to measure the incident response of the participants various metrics have to be developed and made available to the SSC-Monitor, and in order to realistically hide the malicious activities, alternative methods for getting the payload to the compute nodes had to be researched and implemented into the SSC-Monitor. Details are presented in separate contribution to this conference.

In the SSC presented here we focused on the WMS Dirac developed and used by the LHCB VO. Incidents involving a VOs WMS require actions and information flow from/to the VOs security team, the RCs security teams and, eventually, additional entities providing authentication frameworks. The information flow and the orchestrated incident response activities are coordinated by EGI CSIRTs Incident Response Task Force (IRTF).

To assess the readiness of the above mentioned security teams, EGI CSIRT together with the VO LHCB created a realistic incident scenario, where valid user credentials are used to submit jobs to the infrastructure using LHCBs workload management system DIRAC as well as using generic services available for job submission directly to the sites.

In this presentation we will show the detailed incident scenario created by the SSC-Monitor, the expected actions described in the incident response procedures as well as the efficiency of the actions described in the developed metrics.

Physics & Engineering Application / 70

Machine Learning Techniques for Software Analysis of Unlabelled Program Modules

Author: Elisabetta Ronchieri[1]

Co-authors: Davide Salomoni [2]; Marco Canaparo [2]

[1] *INFN CNAF*

[2] *INFN*

Corresponding Author: elisabetta.ronchieri@cnaf.infn.it

Software analysis is of vital importance in the assessment of software characteristics. It is usually based on software measurement, defect data and techniques derived from both statistics and machine learning.

Machine learning has been widely adopted in the field of Software Engineering (SE). For typical SE tasks, machine learning helps computer engineers e.g. extract requirements from natural language text 1, generate source code [2] and predict defects in software [3]. To accomplish these tasks, data have to be collected and properly preprocessed before the application of machine learning techniques. Typical data preprocessing operations may include replacement of missing values and/or removal of inconsistencies. The obtained datasets are therefore composed of a set of instances (i.e. modules, such as files, classes and functions) and features (i.e. software metrics and defect data) and are used to train software quality models by using supervised learning approaches. In SE practice, software dataset may lack some information such as the software module defectiveness, which is mandatory for the application of supervised learning techniques.

Typical datasets employed in machine learning, known as labelled datasets, are related to a single software project whose features have been extracted over time. Among features, defect data may be difficult to collect especially when dealing with new projects or projects with partial historical data. The datasets without defect data are known as unlabelled and represent the majority of software datasets: the production of a labelled software dataset requires effort and time, penalizing a real application of machine learning techniques to predict modules defectiveness. Only in the last decade the unlabelled datasets have been explored with the purpose of conducting analysis and predictions [4, 5].

Machine learning problems have been increasing in complexity over the years leading to a greater resources' consumption. Therefore, average systems and platforms are not considered suitable for performing machine learning algorithms: the number of permutations requested to perform a prediction analysis is extremely time consuming. Cloud computing services have given the chance to overcome these limitations by giving access to large-scale computing and storage resources with little effort.

In this study, we are going to present the analysis of existing unlabelled datasets by implementing models in different available frameworks, such as TensorFlow, Theano and Keras [6], and running in Python. We have evaluated these frameworks on considering three aspects: extensibility, hardware utilization and speed. For this work, we have exploited (also GPU-equipped) cloud computing infrastructure to determine the best models according to common intrinsic evaluation metrics.

Due to the lack of a comprehensive study about practical aspects of software analytic models, we aim at providing a procedure to perform software defect prediction in the scientific environment in order to minimize human effort. Furthermore, we intend to reduce the distance between theory and practice by providing strengths and limitations of the considered frameworks to enable users to assess suitability according to their requirements.

1 Madala, K., Gaither, D., Nielsen, R., & Do, H. "Automated identification of component state transition model elements from requirements," in International Requirements Engineering Conference Workshops, pp.386-392, 2017.

[2] Joulin, A., & Mikolov, T. "Inferring algorithmic patterns with stack-augmented recurrent nets," NIPS'15, pp. 190-198, 2015.

[3] Tong, H., Liu, B., & Wang, S. "Software defect prediction using stacked denoising autoencoders and two-stage ensemble learning," IST'17, 2017.

[4] Catal, C., Sevim, U., & Diri, B. "Clustering and metrics thresholds based software fault prediction of unlabeled program modules," in Information Technology: New Generations, ITNG '09, Sixth International Conference on, pp. 199–204, April 2009.

[5] Zhong, S., Khoshgoftaar, T., & Seliya, N. "Unsupervised learning for expert-based software quality estimation," in High Assurance Systems Engineering, Proceedings. Eighth IEEE International Symposium on, pp. 149–155, March 2004.

[6] Bahrampour, S., Ramakrishnan,N., Schott, L., & Shah, M. "Comparative Study of Deep Learning Software Frameworks," https://arxiv.org/pdf/1511.06435.pdf

Earth & Environmental Sciences & Biodiversity Application / 71

New Seismic Database from the Continuous Recording of the Taiwan Strong Motion Instrument Program and its Potential Application

Author: Bor-Shouh Huang[1]

Co-authors: Juen-Shi Jiang [2]; Wen-Gee Huang [1]; Yen-Ling Chen [2]

[1] *Institute of Earth Sciences, Academia Sinica*

[2] *Central Weather Bureau*

The Taiwan Strong Motion Instrument Program (TSMIP) of Central Weather Bureau (CWB) has installed more than 700 seismometers near two decades which covered the entire Taiwan to record strong motions for seismic hazards, wave propagation and earthquake source physics investigation. Traditionally, those instruments are operated as the trigger mode to detect local events with significant ground shacking. Usually, regional and teleseismic events are rare reported by TSMIP. However, after several generation's upgrade, new TSMIP seismometers have equipped GPS and provided real time data transmission function through internet. In 2018, TSMIP has selected 356 stations (at least one station for each country) to require seismic signal continuous transmission to CWB through the campus internet network for the purpose of quick report seismic intensity of an earthquake. The continuous recording string of TSMIP created a new database of seismic events based on earthquake catalog. Regional and teleseismic events can be archived through global seismic catalogs without to rely on triggering. In this study, we will present several successful cases to demonstrate the well recording of large regional and teleseismic events from TSMIP. Although the raw data do not show clear seismic signal as local event, however, its signals can be well presented after de-spike and law pass filter processing. For the dene spatial distribution of TSMIP stations, the recording data provided a new data set of seismic travel times and waveforms which was never reported in Taiwan and proposed many potential applications to study the crustal structure and wave propagation of Taiwan and the same as deep earth model.

Data Management & Big Data / 72

Building Infrastructure for Big Earth Data and Cloud Services

Author: Xuebin Chi[1]

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

Big Earth Data, as a new type of strategic resource for all nations, brings impetus to Earth Science, and will be a new essential tool to understand the world. To improve the capability of effectively collecting, storing, managing and analyzing Big Earth Data, and to create groundbreaking discoveries, it is imperative to build a Cloud Service Platform and to develop innovative technologies and methodologies for Big Earth Data research. Chinese Academy of Sciences has accumulated vast amount of long-term multi-source scientific data resources and the data acquisition abilities from space to land. Based on these advantages, we propose to construct world leading infrastructure for Big Earth Data and Cloud Service. The infrastructure will support scientific discovery and decision making by integrating and composing a multidisciplinary fusing Big Earth Data Repository, and providing data products and capabilities of computing, storing and analyzing for CASEarth, and.
The program consists of four main parts: (1) The Big Earth Data Infrastructure: Building an extensible Big Earth Data platform to provide advanced computing and storage capabilities by deploying specific computing and storage clusters which converge high-performance computing, high-throughput computing and large-scale data storage. It also aggregates existing computing facilities including China National Grid, China Science & Technology Cloud. (2) The Big Earth Data Repository: Constructing a Big Earth Data Repository to provide data capability by collecting multi-source data of many domains like biology, ecology, environtology, and ingesting data products produced by Big Earth Data research projects. (3) Big Earth Data System Software: Developing systems, including storage management system, computing and processing system, data mining system, grid data engine to provide technical support for Big Earth Data research. (4) A Cloud Service Portal: Developing a Cloud Service Portal to provide unified public services of big data storing, computing, processing, analyzing, and application integrating for users from many research areas.
According to our plan, the infrastructure consisting of specific computing and storage clusters will be deployed before 2019, providing 2PF computing capability and 50PB storage capability. Besides, plenty of services and applications will be deployed on the infrastructure to improve the capabilities of data collecting, storing, managing and analyzing for Earth Science in CAS.

eScience Activities in Asia Pacific / 73

eScience Activities in Philippines

eScience Activities in Asia Pacific / 74

eScience Activities in Mongolia

eScience Activities in Asia Pacific / 75

eScience Activities in Thailand

eScience Activities in Asia Pacific / 76

eScience Activities in Indonesia

eScience Activities in Asia Pacific / 77

eScience Activities in Malaysia

eScience Activities in Asia Pacific / 78

eScience Activities in Japan

Corresponding Author: aida@nii.ac.jp

eScience Activities in Asia Pacific / 79

eScience Activities in China

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

eScience Activities in Asia Pacific / 80

eScience Activities in Taiwan

Corresponding Author: eric.yen@twgrid.org

Opening Ceremony & Keynote Speech I / 81

Scientific Computing for large Infrastructure in Europe - a personal view

Technology development for existing and new scientific infrastructures lead to drastic increases of computing, storage and network demands. This goes along with the request for professional software development in science and with a new class of people with new skills, often called the "data scientist". New challenges for scientific computing are coming at every step of the workflow for scientific data, starting with the data ingest, the analysis and the management of the data according to the FAIR principles and to finally archive the data and in case the software. In this talk these topics will be covered and the European and national view from a personal perspective will be described.

Keynote Speech / 82

FAIR in an unfair world: cybersecurity, data breaches, data integrity, and open science

Data breaches have evolved from routinely making headlines, to making headlines only when they involved large fractions of the world's population, to being routine to the point of no longer being newsworthy. In the U.S., there is increasing emphasis on cybersecurity to assure the confidentiality of data, emphasis that is reaching into higher education and research through contractual requirements. However, as we consider shared data and large scientific datasets, we need cybersecurity that appropriately respects broadly shared data whose greatest challenges may lie with integrity and availability and that is applicable to projects of all science domains and sizes. This talk will cover work in cybersecurity for science by the NSF Cybersecurity Center of Excellence and other NSF projects.

Keynote Speech / 83

Digital Humanities in the Cloud: Reading the Invisible Library

Progress over the past decade in the digitization and analysis of text found in cultural objects (inscriptions, manuscripts, scrolls) has led to new methods for reading the "invisible library". This talk explains the development of non-invasive methods, showing results from restoration projects on Homeric manuscripts, Herculaneum material, and Dead Sea scrolls. Premised on "virtual unwrapping" as an engine for discovery, the presentation culminates in a new approach - Reference-Amplified Computed Tomography (RACT) – where machine learning and cloud computing becomes a crucial part of the imaging pipeline. You will leave this talk considering that RACT may indeed be the pathway for rescuing still-readable text from some of the most stubbornly damaged materials, like the enigmatic Herculaneum scrolls.

Soundscape Conference / 84

Exploring ecosystem dynamics by hypotheses-driven soundscape information retrieval

Authors: Tzu-Hao Lin[1]; Yu Tsao[2]

[1] *Japan Agency for Marine-Earth Science and Technology*

[2] *Academia Sinica*

Soundscape information retrieval represents the technique to extract meaningful information relevant to geophysical, biological, and anthropogenic activities from field recordings. Supervised source separation and audio recognition techniques have been widely employed in the past, but the performance depends on the quantity of training database and the complexity of testing data. To counter this issue, unsupervised learning approaches, such as blind source separation and clustering tools, have been recently introduced to analyze the dynamics of marine and terrestrial soundscapes. However, the performance of unsupervised learning relies on a proper hypothesis for the input data. Until now, it remains a challenge for ecological researchers to integrate proper hypotheses in soundscape information retrieval. In this presentation, we will demonstrate the integration of acoustic niche hypothesis, which predicts soniferous species will avoid acoustical competition by shifting acoustic niche in time or frequency domains, in the analysis of soundscape dynamics in the shallow waters off western Taiwan. In the future, more advanced techniques of source information retrieval

are necessary to facilitate the soundscape-based ecosystem monitoring. The domain knowledge of ecological science and bioacoustics will be essential for the future development of soundscape information retrieval.

Biomedicine & Life Science Application / 85

X-ray imaging of brain

Comprehensive mapping of neural networks of animal brains is a formidable but very exciting challenge. The complexity of the complete network is beyond the current technology to describe, analyze and understand. It is now a consensus that the first step towards understanding brain functions is to construct a basic map – a connectome – showing the neural network at the level of single neurons and connections. As one of the six "high priority challenges" in the US BRAIN Initiative: "Maps at multiple scales: Generate circuit diagrams that vary in resolution from synapses to the whole brain", we believe our imaging strategy using synchrotron x-rays will transform this vision onto reality.

The key element in our technology arsenal is the phase contrast micro- and nano-tomography. The high-performance x-rays photons provided by the new facilities such as synchrotrons and x-ray free electron laser opens the door for x-ray microscopy to an unprecedented level of performance. The technology to focus hard-x-rays photons has made great progress in the past decade. However, practically achieving nanometer scale resolution remains a formidable technology challenge. The development of nanotechnology to fabricate nanostructured device impact unexpectedly x-ray microscopy by providing the long sought optics required to achieve high resolution and high contrast. Using the same x-ray photons with the nanotomography instrument, the fine details of the same specimens can be imaged in 3D with <20 nm resolution. This allows us to examine the smallest network features, such as dendrites and dendritic splines, within specific regions, important features to understand how the whole brain network functions.

Soundscape Conference / 86

Changing the Lens on Soundscape Ecological Research: What We Can Do to Address Current Grand Environmental Challenges

Sound is a universal measure of change and is one of the most emotional senses humans possess. How can we take advantage of these two fundamental notions to help improve ecological and social well-being of this planet? I address this question with a summary of my vision for how the paradigm of soundscape ecology is uniquely positioned to advance the critical understanding of our changing world. New pathways of scholarship are clearly needed but they are within our reach even today. I conclude with seven rules we must follow to fulfill this vision so that we may achieve a more sustainable planet.

Soundscape Conference / 87

Listening to Biodiversity and its Changes – Introduction of Asian Soundscape Monitoring Network

Loss of biodiversity and associated ecosystem services is one of the major challenges facing human-

ity. Monitoring biodiversity status and trends across spatial and temporal scales is necessary to mitigate impacts of human-induced environmental changes and secure human well-being. Soundscape, the collection of all sounds emanating from a landscape, reflects the dynamics of biological, environmental and societal systems in a landscape as well as the interactions among them. Investigating the spatiotemporal patterns of soundscape can provide insights into characteristics of those systems and their responses to environmental changes. Soundscape monitoring is thus a useful tool for tracking biodiversity under global changes. Asian Soundscape Monitoring Network was established in 2014 with the aims to understand soundscape patterns, enhance the capacity of soundscape monitoring, open up soundscape data and promote soundscape research in the Asian region. Since then, long-term monitoring sites have been established in Malaysia, Philippines, Taiwan, Thailand and Vietnam to collect data on terrestrial, marine and ultrasonic soundscapes. In this talk, I will introduce this network, show the achievements so far, describe short- and long-term plans, and provide some example applications of soundscape data collected in the network.

Soundscape Conference / 88

Soundscape phenology in coral reef

Co-authors: Frederic Sinniger [1]; Saki Harii [1]; Tzu-Hao Lin [2]

[1] *University of Ryukyus*

[2] *Japan Fisheries Research and Education Agency*

Coral reef accommodate many symbionts that made it highly biodiverse underwater ecosystem. Soundscape is one of an indicator of biodiversity since phonation of animals represent species or family specific acoustic characteristics. Using unsupervised classification algorism developed by the second author, we found clear change of soundscape of coral reef in different season. Extensive calling of damsel fish was observed in May and June. It lasted until midnight and occurred again very early morning. On the other hand, song of humpback whales dominated soundscape in the winter time. Soundscape analysis is a tool to investigate acoustic phenology based on a long term recordings.

Soundscape Conference / 89

Human disturbance on fish, fisheries and marine soundscape in an intertidal coralline algal reef, Taoyuan

Coralline or coralligenous algal reef, like many other coastal marine ecosystems, are vulnerable to land-source sedimentation and human disturbance. The Taoyuan coralline algal reef is a biodiverse area recently threated by coastal developments and industrial waste runoff. As the reef lies in an area highly disturbed by monsoons, it is difficult to survey the algal reef fish community using traditional netting method, hence knowledge of Taoyuan algal reef is limited and it was long believed to be a barren coast. We employed multiple sampling methods and underwater sound recording (soundscape) to determine fish community response to human disturbance. We selected five different sites ranging from the marine reserve to sites adjacent to the industrial area to represent different levels of human disturbance. To date, we have found both fish sampling methods and soundscape analysis indicated the sites in Datang algal reef are areas of high fish biodiversity and abundance. We also recorded juvenile predatory reef fishes, such as grouper, snapper and hammerhead shark, indicating that these fisheries species use these shallow water reefs as a nursery habitat. During this study, we observed highly turbid effluent discharge from the industrial area as well as a mass fish death incident beside the discharge point. We suggest a marine reserve to protect this extraordinary reef ecosystem from local development.

Soundscape Conference / 90

Listening to underwater noise: impacts of chronic noise exposure on fishes

Co-author: Tzu-Hao Lin [1]

[1] *apan Agency for Marine-Earth Science and Technology*

The soundscape underwater is never silent, but composed of geophonic and biophonic sounds. Anthropogenic sounds, e.g. sounds generated during exploration of oil and gas deposits, shipping, military operations, and development of offshore wind farms (OWF) have altered the underwater soundscape in the past 50 years. The impacts of anthropogenic noise on underwater animals were depended on the exposure level. Strong noise, such as that from pile driving, military sonar, seismic exploration, and shipping, has been known to cause auditory damage, hearing loss, behavioral change, communication masking in underwater animals. On the other hand, recent studies indicate that continuous noise with lower sound intensity may also induce physiological responses. For examples, milkfish (Chanos chanos) and black porgy (Acanthopagrus schlegelii) responded significantly on cortisol metabolism or reactive oxygen species / antioxidants balance, when they were exposed to the turbine noise of OWF for long duration. Furthermore, such long-term noise exposure may affect fish behavior as well. Using auditory brainstem response, a preliminary experiment showed that response thresholds of big-head croakers (Pennahia macrocephalus) to the intra-specific mating calls were found to be higher under long-duration exposure of OWF noise. The results suggested the increment of anthropogenic noise, will not only shape the underwater soundscape, but also affect both soniferous and non-soniferous fish in various levels. On the basis of Asian Soundscape, now it is possible to study the prominence of underwater noise in different aquatic ecosystems, which serve as the essential baseline information for evaluating the potential impacts of sound-generating human activities on underwater animals.

Soundscape Conference / 91

Coral reef soundscapes of Cebu, Philippines: Initial results and future directions

Co-author: Tzu-Hao Lin [1]

[1] *Japan Agency for Marine-Earth Science and Technology*

The coastal waters around the island of Cebu is a complex mix of ecosystem types, substrate composition, and underwater topography that vary spatially and temporally. Levels of protection, human use patterns, and occurrence of natural disturbances also vary across space. Status of the coastal ecosystems is primarily monitored through visual assessments of abundance, biomass, or percent cover of the dominant producers such as seagrass, algae, and mangrove or of the dominant life forms such corals and fishes. While these can provide measures of changes between monitoring periods, they are unable to trace temporal sequence of changes in community composition and are unable to detect nocturnal or cryptic species. The use of soundscapes to investigate the phenology of community structures from soniferous marine animals remains an underutilized tool in the Philippines. To characterize the spatiotemporal soundscape patterns and factors that may influence such patterns, autonomous underwater recorders are initially deployed in the following coastal habitats in Cebu: a seagrass bed and a reef flat in a marine protected area near a proposed coastal development site and a fore reef in a protected area frequented by divers. Unsupervised machine learning techniques are employed to tagged biological choruses, transient calls, anthropogenic noises in long-duration underwater recordings. Results from these initial efforts will be used to demonstrate to conservation managers, local government leaders, and funding agencies the value of integrating soundscape information in assessment and monitoring plans of marine ecosystems. In the future, the soundscape monitoring works will be expanded to mesophotic reef in the northeastern Philippines and the Tubbataha Reef in the western Philippines. Possibilities of combining soundscapes with other emerging tools will also be explored.

IGTF All-Hands Meeting / 92

EUGridPMA Update

Corresponding Author: davidg@nikhef.nl

Soundscape Conference / 93

Collaboration in Soundscape Monitoring: A Use Case for Research Data Management

Co-authors: Cheng-Jen Lee [1]; Chia-Hsun Wang [1]

[1] *Academia Sinica*

We will first provide an overview of depositar which is a research data repository built on top of CKAN, an open source software package originally developed for publishing open (government) data. Several features have been added to depositar to better support the deposit, curation, and exploration of research datasets. The new features include, among others, 1) rich metadata support, 2) spatiotemporal annotation and query, 3) preview and overlay of spatial datasets, 4) Wikidata-powered keywords, and 5) a simplified frontend and registration process.

We will then share our experience working with a collaborative soundscape monitoring project on using the repository to help manage and publish research datasets. We will discuss issues related to project composition and role, data workflow, metadata requirements and practices, and the interfaces to other information systems and tools.

Soundscape Conference / 94

Acoustic Diversity and Activity Patterns of Insectivorous Bats in a Riverine Forest at Gunung Mulu National Park, Malaysian Borneo

Co-author: Faisal Ali Anwarali Khan [1]

[1] *Universiti Malaysia Sarawak*

Bats are keystone species that perform vital ecosystem and economic services and are, therefore, an important fauna to be monitored. Acoustic monitoring with ultrasonic detectors has emerged, in recent years, as an essential tool to quantify the activity of echolocating insectivorous bats and identify key habitats used by them for commuting and foraging. However, only a few acoustic studies have been conducted in the tropical forests of Southeast Asia. In order to monitor bats using acoustic methods, a library of echolocation calls needs to be available for the region or even better from the specific locality. Between May and November 2016, a full spectrum ultrasonic detector was used to record bat activity during 105 nights, at 35 points along three rivers of different width and also adjacent trails and forest interior 50 metres from trails and rivers. Recorded calls were analysed by spectrogram viewer software and compared manually to a collection of references calls recorded from 502 individuals representing 30 species of insectivorous bats, previously captured within the same area. Discriminate function analysis was performed on reference calls to determine which species could be accurately identified from their calls. A total of 273.02 gigabytes of files containing an estimated 104,834 bat passes were recorded. Overall, 93% of passes were recorded at rivers, while 5% were recorded on trails and 2% in forest interior. Based on results from discriminant function analysis, five species of cave roosting Rhinolophidae, seven Hipposideridae, Chaerephon plicatus and Miniopterus australis could be identified and represented 62% of total recorded passes. A further 36% of passes were identified to a Myotis species call group and at least four call types from unknown species were detected. Accurate identification of low intensity, high frequency calls from members of the forest roosting Vespertilionidae subfamilies Kerivoulinae and Murininae proved to be problematic, however this group of species only represented 1% of recorded calls. Most species recorded were detected more often at rivers than in other habitats. Acoustic sampling, although challenging to conduct in a megadiverse tropical environment, is proving to be an effective method in providing information about the ecology of insectivorous bats. This study also highlights the importance of rivers as critical habitats for foraging insectivorous bats in a tropical karst environment.

Soundscape Conference / 95

Bat acoustics in Vietnam: a historical review of research and developing a national database

Vietnam is recognised as an important country in Asia for bat research and conservation. To date, 126 bat species belonging to 36 genera and 8 families are known. However, the acoustic characters of many Vietnamese bats are still poorly studied. Between 2006 and 2018, we conducted a series of bat surveys in Vietnam with an emphasis on bioacoustic research. Echolocation calls of bats were recorded in different situations using PCTape; they were subsequently analysed with Selena software. We have obtained comprehensive data of echolocation calls for over 50 bat species of 7 families: Emballonuridae, Hipposideridae, Megadermatidae, Miniopteridae, Molossidae, Rhinolophidae and Vespertilionidae. Of these, the echolocation calls of the Hipposideridae and Rhinolophidae are sufficiently distinct to enable identification to species level. Our results have been included in a database of Vietnamese bat echolocation call, which will be used for monitoring, ecological studies and conservation. This presentation provides a historical review of research of echolocation of Vietnamese bats; our achievements between 2006 and 2018; and recommendations and potential trends for further research in the future.

Soundscape Conference / 96

Panel Discussion

IGTF All-Hands Meeting / 97

TAGPMA Update

Soundscape Conference / 98

Acoustic Signal Enhancement in a Distributed Microphone Environment

Co-author: Tzu-Hao Lin [1]

[1] *Japan Agency for Marine-Earth Science and Technology*

Recently, information retrieval based on acoustic signals has caught great attention. In real-world scenarios, acoustic signals are easily distorted by additive or convolutional noises or recording de-

vices, which constrain the achievable information retrieval performance. To address this issue, numerous acoustic signal enhancement (ASE) algorithms have been derived in order to improve the distorted acoustic signals and are widely used as a preprocessor in acoustics-related applications. Most of these approaches consider the condition where only one microphone (channel) is available to capture the acoustic signals. With the recent advances of hardware and communication technologies, it is believed that there will be multiple channels available for acoustic information retrieval in the near future. Therefore, ASE methods with considering multiple channels is an emerging research topic in the acoustic signal processing field. In this talk, we present a novel fully convolutional ensemble learning networks (FCEN) for multichannel ASE in the time domain. The proposed FCEN is formed by a dilated convolution and skip-connection structure with three ensemble inputs. Experimental results confirm the outstanding denoising capability of the proposed FCEN model on a ASE task where distributed microphones are available as compared to existing methods.

Soundscape Conference / 99

An application of acoustic and visual information to monitor activities of sika deer

Co-authors: Pei-Jen Lee Shaner [1]; Tzu-Hao Lin [2]

[1] *National Taiwan Normal University*

[2] *Japan Agency for Marine-Earth Science and Technology*

After the extinction of sika deer (Cervus nippon) population in the wild in Taiwan, a restoration program has been in place in Kenting National Park since 1984. To evaluate the effectiveness of restoration program and to address the issue of deer impacts on the environment, the population status and activities of sika deer should be monitored. In addition to conventional methods, such as line transects and camera traps, soundscape monitoring represents an alternative tool for monitoring sika deer population during their breeding season. From November 7 to December 21, 2017, we deployed 9 sound recorders and 9 camera traps to acoustically and visually, respectively, investigate the behavior pattern of sika deer. In the acoustic part, we measured the difference between the mean spectrum and the median spectrum to detect high-intensity transient sounds. A clustering algorithm was employed to further identify rutting vocalizations of sika deer. In the visual part, we calculated a relative activity level, i.e. the number of independent photographs per survey effort, for male sika deer at each camera trap. In addition, we estimated the diel activity pattern of male deer by applying a kernel density estimation. The acoustical analysis shows that 9 acoustic clusters were identified from 3,124 hours of effective recordings. Among them, two clusters were highly associated with deer's rutting vocalizations. The occurrence of deer's rutting vocalizations varied among different recorder locations and recording months. The rutting vocalizations are crepuscular in a diel period at every recording location. In addition, we collected 506 independent photographs of male deer from 392 trap nights of all camera traps. Male deer's relative activity level varied among the traps. Their diel activity pattern was crepuscular. On the basis of camera traps and sound recorders, the dynamic pattern of space use and diel activity of sika deer can be effectively detected. Each of camera trap and sound recorder has pros and cons, such as detection range, efficiency of data process, data precision, restriction of season, cues for animal behavior…etc. Therefore, a combination of acoustic and visual information are potential to be employed to monitor the population status of wild sika deer in the future.

Soundscape Conference / 100

Use of acoustic recorders in monitoring and management of seabird breeding colony in Matsu archipelagos, Taiwan

The Chinese Crested Tern, Thalasseus bernsteini (CCT), is the most critically endangered seabird species in Taiwan. CCT nests sympatrically with the Great Crested Tern, T. bergii (GCT), among seven protected islands within the Matsu Island Tern Refuge (MITR).To minimize disturbances during breeding season, autonomous acoustic recorders were used to monitor the activity of CCTs and GCTs on the protected islands of MITR. Based on records from 2015 to 2017, we found that the average sound pressure level between 1.5-8kHz was associated changes in the number and status of the tern colony. Therefore, it can be considered an important information point for habitat management and ecological protection in MITR. In addition, we have further developed the CCT sound recognition system using machine learning technology. We expect to be able to locate breeding colonies of CCTs through an automatic sound recognition system. Overall, autonomous acoustic recorders provide more accurate and effective information for the monitoring of seabird breeding communities on remote islands such as Matsu.

IGTF All-Hands Meeting / 101

APGridPMA Update

Corresponding Author: eric.yen@twgrid.org

102

eScience Activities in India

eScience Activities in Asia Pacific / 103

eScience Activities in Pakistan

Corresponding Author: arshad.ahmad@ncp.edu.pk

Soundscape Conference / 104

The Soundscape Composition of Mt. Makiling Forest Reserve, an ASEAN Heritage Park in the Philippines

Organisms interact in a complex system that also involves various cues for intra- and interspecific recognition. Sounds are important signals for recognizing kin, prey as well as predators. In the Philippines, very little is known about the characteristics of sounds in the rainforest, how they vary across landscapes and time, and how species interact acoustically. It is aimed that this research will be able to provide an overview and baseline information on the soundscape structure in a typical rainforest in the Philippines, recognizing the astounding diversity of organisms that simultaneously produce sounds at varying frequency levels. The lack of such baseline information about the sound characteristics of different organisms prevents large-scale analysis at the ecosystem level, and as a result, hindering updated policies to be formed in protected areas. This research aims to describe the acoustic dynamics of different species in Mt. Makiling Forest Reserve. Specifically, we hope to (1) establish an extensive collection of voucher calls to permit identification of species through the sounds/calls that the species produce, (2) provide baseline data on the interaction of different species across different landscape and time, and (3) contribute to the effective management of this ASEAN Heritage park and formulate recommendations on the use of bioacoustics analysis in the management and protection of other protected areas in the Philippines and the ASEAN.

Opening Ceremony & Keynote Speech I / 105

Opening Remarks

Corresponding Author: simon.lin@twgrid.org

eScience Activities in Asia Pacific / 106

eScience Activities in Korea (remote presentation)

Soundscape Conference / 107

Using the passive acoustic monitoring to study the calling behaviour of yellow-cheeked gibbon (Nomascus gabriellae) in Bidoup - Nui Ba National Park, Vietnam

Co-author: Yu-Huang Wang [1]

[1] *Indepent Scholar*

Vocalization is an important characteristic of gibbons (family Hylobatidae) in Southeast Asia because it is used to determine and protect group territory and in social behaviour. Even though the important role of calling in gibbon life, understanding the calling behaviour of species often based on short time of study rather than in long-term study because of the limitation of presence surveyor in the field. The one-year data of acoustic that recording by the automatic acoustic recording (Wildlife Acoustic SM2) in the mixed-conifer and broadleaf of Bidoup – Nui Ba National Park was explored to understand the calling behaviour of southern yellow-cheeked gibbon (Nomascus gabriellae). For a year of monitoring, gibbon's song, including solo and duet song, were recorded in 195 days with 264 times and the number days that detected call vary over the month with a maximum in October 2015 and minimum in May 2015. The earliest time of day the gibbon made a call was 5:30 and the last call of the day at 13:00. Regarding the number of groups made a call at the same time, about 18.94% of the total file obtained a call from two or three groups and most of them came from one group. Within the 5-minutes of recording, the gibbon produced average 7.98 calls (range from 1 to 17), of which 3.7 are duet calls (range from 1 to 8). In this study, the species produced more calls in the rainy season than in dry season.

Workshop on Global PBL (Project Based Learning) / 108

Rationale behind the curriculum development for global liberal arts

Corresponding Author: soetosh@gmail.com

Kansai University's Collaborative Online International Learning (COIL) endeavor is elaborated and rationale behind the curriculum development for global liberal arts is elaborated for the benefit of the discussion in the workshop.

Workshop on Global PBL (Project Based Learning) / 109

Report: Stage I: Academic Skills

Corresponding Author: soetosh@gmail.com

The organizer team have conducted pilot studies for curriculum development since 2015. At Stage One, the theme of the learning contents was focused on the global academic skills in the realm of global liberal arts, where proactive active learning (PBL/TBL) in global teams was implemented to nurture critical and creative thinking skills while acquiring research skills in transdisciplinarity region. The pilot studies was conducted between Kansai University (KU) and collaborative universities in Taiwan (National Taiwan University, National University of Tainan, National Central University, Asia University, Soochow University, Hsuan Chuang University, Chia Nan University, National Tsing Hua University, among others). It was found that the approach at Stage One lacked authenticity because the learning motivation had been based on the learners' intellectual curiosity about diversity in global culture. The students at Stage One learned to work in global team to conduct PBL/TBL. However, they were not ready to scenario plan future global society as intended initially for the sense-making of the 21st Century Skills.

Workshop on Global PBL (Project Based Learning) / 110

Learning Environment: Hands-on Session

Corresponding Author: soetosh@gmail.com

ICT-enhanced learning environment is shared in the form of hands-on session.

Global Team Building with Empathy enhanced with Cloud Service
Flipgrid, Padlet, Google Drive
Team Project Management (all members as well as all students and instructors to be on the same page of learning)
Padlet, Trello, Google Drive,

Workshop on Global PBL (Project Based Learning) / 111

Report: Stage II Global Social Entrepreneurship

Corresponding Author: soetosh@gmail.com

At Stage Two, the theme of the learning contents has been extended to the global social entrepreneurship in global liberal arts to have learners more aware of component of scenario planning of future society, where global teams define problems in the society, work on their optimal solutions, design their deployment in the society enhanced with information technology, and then plan for start-up companies projecting to the mezzanine. In this way, students will have more enriched learning opportunity to develop their 21st Century Skills conscious to the current society as well as the future society.

Workshop on Global PBL (Project Based Learning) / 112

Learning Environment for SE & Assessment

Corresponding Author: soetosh@gmail.com

ICT-enhanced learning environment for global teams, leanstack®, is elaborated.
And further, visual assessment strategies are shared.

SE enhanced with IT: leanstack®
SE Assessment and Analytics: Rubrics, Meta Cognitive Reflection

Soundscape Conference / 113

Panel Discussion

Opening Ceremony & Keynote Speech I / 114

Long-range transport of Southeast Asia biomass burning pollutants to Taiwan: Impacts and implications

It happens to be the biomass burning season in spring time from Indochina. Under favor weather conditions, the products of biomass burning pollutants could be transported easily to Taiwan and even East Asia. Actually, the complex interactions of these air pollutants and aerosols features in the boundary layer and aloft have resulted in complex characteristics of air pollutants and aerosols distributions in the lower troposphere. For example, at the Lulin Atmospheric Background Station (LABS) (elevation 2862 m) in central Taiwan, the concentrations of carbon monoxide (CO), ozone (O3) and particulate matter with diameter less than 10 ⊠m (PM10) were found to be 135-200 ppb, 40-56 ppb, and 13-26 µg/m3, respectively in the springtime (February-April) between 2006 and 2009, which are 2-3 times higher than those in other seasons.

The project "Effect of Megacities on the transport and transformation of pollutants on the Regional and Global scales (EMeRGe)" aims to improve our knowledge and prediction of the transport and transformation patterns of European and Asian megacities pollutant outflows. In EMeRGe Asia, the composition of the plumes of pollution entering and leaving Asia measured by the new High Altitude and LOng Range (HALO) aircraft research platform. The HALO aircraft performing optimized transects and vertical profiling in Asia during 12 March and 7 April in 2018. To design the measurement of aircraft flight paths and elevations, a high resolution, 9 km, numerical prediction by Weather Research Forecast (WRF) and WRF-Chem models were joined and performed during the campaigns. The LRT of biomass burning organic aerosol to Taiwan measured by HALO could be more than 2 ug/m3 at the elevation of 2500 m on 20 March, 2018. Model performances and the results will be discussed in this meeting. Overall, this series of studies significantly fill the gap of our understanding on air pollutants transformation and transport to Taiwan and East Asia, and show the potential directions of future studies.

Keynote Speech / 115

Transitioning to the Fourth Paradigm of Discovery

We have now entered the era of big data. Few know what this truly entails but those of us that are confronted with it know that it means we need to do our work differently. Indeed, scholars across disciplines as diverse as ecology, social science, medicine, and physics understand that addressing big data challenges requires entirely new approaches. One visionary that contemplated the impact that big data has on society – Jim Gray of Microsoft – suggested that our approach to science and engineering must require new thinking that is transformative – that we need to enter into a new, fourth paradigm of discovery. Is it possible for us to transition to this this new paradigm? What is necessary for us to achieve this? I outline an approach that I believe that is needed to confront, head on, the enormous challenges of big data. I argue that we cannot retreat either, as the potential that this era has to address grand challenges problems in society are boundless.

Keynote Speech / 116

Is the hype real? An introduction to quantum computing

Quantum computing has been a buzzword for a while. Contradicting information and hype are all over the internet. Many governments, industry leaders, and venture capitalists all over the world are investing heavily into it. I'll introduce the basic concepts in quantum computing and Google's efforts on superconducting quantum computer. I'll also assess what researchers can do with quantum computing now and in the near future, and talk about how to become quantum-ready in this exciting time.

ECAI Workshop / 117

Bibliography and learning

What are the requirements for learning? The implications of a semiotic view of learning will be related to the design and potential uses of bibliographies, extended to all media, and other kinds of reference works.

ECAI Workshop / 118

Bibliography and creativity

How might bibliographical practices of description and analysis be viewed as creative? The talk will attempt to answer this question as it relates to cultural and especially literary studies.

ECAI Workshop / 119

Atlas of Maritime Buddhism Project Updates (remote presentation)

ECAI Workshop / 120

Technical Project Updates

ECAI Workshop / 121

The Building Blocks for Open Ecosystems of Online Resources Serving Buddhist Communities

The presentation gives an overview of the state of the art of the software building blocks for development of online resources serving Buddhist communities and how those are driving new capabilities and broadening access. The central theme described is the huge scale and rapid evolution of the open source movement and modular package management systems that are built on open source. Illustrative examples will be given from the author's experience developing web applications for the study of Buddhist texts, including translation projects for Fo Guang Shan.

ECAI Workshop / 122

Generating East Asia's Bibliographic Past

This talk will suggest that a form of deep learning known as Generative Adversarial Networks (GANs) can be usefully incorporated into bibliographical investigations of older East Asian texts. It will do so by demonstrating that GANs can be used to generate historically accurate representations of Korean, Japanese, and Chinese xylographyic and typographic shapes. To demonstrate their usefulness to historical bibliography, the talk will focus on a description of how GANs have been used to generate nearly all of the unique characters likely to appear in an important facsimile of the Qisha Canon, thereby speeding up its transcription and advancing work underway in the Buddhist community to transcribe all of the documents associated with the Chinese Buddhist Canon into machine-readable text.

ECAI Workshop / 123

Longquan Buddhist Canon punctuation automation process and related areas including new website, API, and punctuation knowledge database

ECAI Workshop / 124

ThakBong goes CIDOC CRM: Motivations, Challenges, and Prospects

ECAI Workshop / 125

Taiwan Baotu Re-imagined in OpenStreetMap

Environmental Computing Workshop / 126

Environmental Computing and Data Management at LRZ

Corresponding Author: kranzlmueller@ifi.lmu.de

Addressing the growing area of Research Data Management (RDM) in today's data-driven science, the Research Department of Leibniz Supercomputing Centre (LRZ, Garching, Germany) has formed a RDM Team. Here, we give an overview about current projects of this team and the whole Department. These projects bring together Environmental Computing with RDM, following FAIR principles.

In fact, RDM as a topic at LRZ has been prominently driven by environmental projects, such as the Alpine Environmental Data Analysis Centre (AlpEnDAC.eu), the climate and hydrology supercomputing projects ClimEx and ViWA (www.climex-project.org, viwa.geographie-muenchen.de), and the projects ePIN and BAYSICS (see later talks). On the other hand, LRZ has been involved in a pure and focused FAIR-RDM effort: the project “Generic Research Data Infrastructure” (www.GeRDI-project.eu), building a common German scientific data search and management system. Recently, the EU Project LEXIS was launched to improve mixed “Supercomputing+Cloud” workflows, including a data-management and Environmental Computing component as well.

To integrate these efforts, the RDM team - collaborating with our Environmental Computing team - is developing a lightweight framework (codenamed “Let the Data Sing”/LTDS) for projects with FAIR RDM at LRZ. LTDS will store standardised metadata for Supercomputing data sets (e.g. ClimEx / ViWA), and make these findable in GeRDI by exposing the metadata via OAI-PMH. A functionality to assign DOIs to data sets at LRZ will be implemented as well as landing pages allowing users to access the actual data.

In the course of our development, we are happy to collect feedback to our ideas at workshops and conferences. This shall allow for an optimum design of the details of LTDS and GeRDI.

Environmental Computing Workshop / 127

European Open Science Cloud - Concept, status and opportunities

Corresponding Author: gergely.sipos@egi.eu

Environmental Computing Workshop / 128

Regional Collaborations on Disaster Mitigation

Corresponding Author: eric.yen@twgrid.org

Environmental Computing Workshop / 129

Progress of development firewatch monitoring sensor in Indonesia

Environmental Computing Workshop / 130

A Case Study of a Flood Producing Heavy Rainfall Event over Northwestern Peninsular Malaysia During 4-5 November 2017

Environmental Computing Workshop / 131

HPC, AI and Machine Learning in Disaster Management

Environmental Computing Workshop / 132

WFSRU's Disasters Related Activities and Thailand's UND Case Studies Updates

Environmental Computing Workshop / 133

Development of Nested-Grid Storm Surge Fast-Calculation System and Case Study of 2019 Tropical Cyclone in Thailand and 2013 Super Typhoon Haiyan in the Philippines

Environmental Computing Workshop / 134

ePIN Pollen Monitoring Project (remote presentation)

Environmental Computing Workshop / 135

Hydrometeorological Simulations including massive data collection (remote presentation)

Environmental Computing Workshop / 136

Building IT Infrastructure for Citizen Science Research on Climate Change (remote presentation)

Due to the far-reaching consequences of climate change, extensive adaptation and climate protection measures are becoming necessary not only on a global, but in particular on more regional scales.

Such measures need to embrace citizens for their success, on an educational as well as a participatory level. For Bavaria, Germany, the BAYSICS project (Bavarian Citizen Science Portal for Climate Research and Science Communication) aims to achieve (1) innovative digital forms of citizens' participation in climate change research, (2) transfer of knowledge on the complexity of climate change and its local consequences, and (3) joint scientific and environmental education goals.

Within the project we develop two main tools for citizen scientists, an interactive web portal and smartphone app. As the project processes data collected by citizen scientists, careful considerations are necessary, especially on legal issues (e.g. copyright), the privacy of citizen scientists, and the credibility of data. Data collected by the citizen scientists are visualized on the interactive web portal. Additionally, the data will be available for download through the web portal. Thus, clear metadata (e.g. on relevant attributes and data usage license) needs to be included. To increase the credibility of data, data collection guideline for citizen scientists and the application of trust metrics to data are considered.

Environmental Computing Workshop / 137

Discussion

Environmental Computing Workshop / 138

Case Study of Flooding in Middle Region of Myanmar

IGTF All-Hands Meeting / 139

Introduction

Corresponding Author: eric.yen@twgrid.org

HTCondor& ARC Workshop / 140

Welcome and Logistics

Corresponding Author: ian.collier@stfc.ac.uk

HTCondor& ARC Workshop / 141

HTCondor User Tutorial

HTCondor& ARC Workshop / 142

Introduction to Workflows with DAGMan

HTCondor& ARC Workshop / 143

HTCondor Administration: Architecture and Deployment

HTCondor& ARC Workshop / 144

HTCondor ClassAds Language

HTCondor& ARC Workshop / 145

Monitoring your HTCondor Pool

HTCondor& ARC Workshop / 146

Job and Machine Policies

HTCondor& ARC Workshop / 147

Negotiator Policy: User and Group Priorities

HTCondor& ARC Workshop / 148

HTCondor Python API Overview

HTCondor& ARC Workshop / 149

Docker and Singularity Container Support in HTCondor

HTCondor& ARC Workshop / 150

Federating HTCondor Pools

HTCondor& ARC Workshop / 151

HTCondor: What Next?

HTCondor& ARC Workshop / 152

Day One Closing Remarks

Corresponding Author: ian.collier@stfc.ac.uk

HTCondor& ARC Workshop / 153

Introduction and Logistics

Corresponding Author: ian.collier@stfc.ac.uk

HTCondor& ARC Workshop / 154

Overview: From Clusters to Grids

HTCondor& ARC Workshop / 155

HTCondor-CE Basics and Architecture

HTCondor& ARC Workshop / 156

HTCondor-CE Configuration Part One

HTCondor& ARC Workshop / 157

HTCondor-CE Configuration Part Two

HTCondor& ARC Workshop / 158

Grid Universe: Interface with Other Batch Systems

HTCondor& ARC Workshop / 159

HTCondor-CE Trouble shooting and Questions / Answers

HTCondor& ARC Workshop / 160

HTCondor-CE: What Next?

HTCondor& ARC Workshop / 161

Workshop Closing Remarks

Corresponding Author: ian.collier@stfc.ac.uk

HTCondor& ARC Workshop / 162

Introduction to ARC

HTCondor& ARC Workshop / 163

ARC for WLCG pilots

HTCondor& ARC Workshop / 164

ARC with Caching

HTCondor& ARC Workshop / 165

aCT

HTCondor& ARC Workshop / 166

Changes from ARC5 to ARC6

HTCondor& ARC Workshop / 167

Hacking ARC

HTCondor& ARC Workshop / 168

Containers and ARC

HTCondor& ARC Workshop / 169

ARCHERY

HTCondor& ARC Workshop / 170

arcctl + the new RTE system

HTCondor& ARC Workshop / 171

Coffee Break

Security Workshop / 172

Forensics hands on training

Security Workshop / 173

Network Monitoring, towards a Security Operations Center

Corresponding Author: david.crooks@stfc.ac.uk

Security Workshop / 174

Developing Incident Response in a federated environment, a practical approach

Corresponding Author: sveng@nikhef.nl

Security Workshop / 175

A Cybersecurity Framework for Open Science: Motivations and Requirements Discussion

Security Workshop / 176

AARC, federated incident response

Corresponding Author: davidg@nikhef.nl

Security Workshop / 177

Wrap-Up

Biomedicine & Life Science Application / 178

Approach to natural product archive and relationships

Environmental Computing Workshop / 179

VN Update (remote presentation)

Environmental Computing Workshop / 180

Event Report of ICT-DM2018 Conference

IGTF All-Hands Meeting / 181

Review of MICS Audit Guideline

Corresponding Author: sakane@nii.ac.jp

IGTF All-Hands Meeting / 182

AARC multi-community BPA and assurance mapping

Corresponding Author: davidg@nikhef.nl

IGTF All-Hands Meeting / 183

Automatic certificate management environment (ACME) protocol

IGTF All-Hands Meeting / 184

CILogon, SciTokens and Grid Community Toolkits

Corresponding Author: jbasney@illinois.edu

IGTF All-Hands Meeting / 185

Combined assurance model

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

IGTF All-Hands Meeting / 186

Self-audit reports and Updates of APGridPMA CAs (CNIC/SDG, HKU, HPCI, IGCA, IHEP, KEK, KISTI, MYIFAM, ASGCCA)

IGTF All-Hands Meeting / 187

Next Meeting (candidates)

IGTF All-Hands Meeting / 188

AoB

Earth & Environmental Sciences & Biodiversity Application / 189

TBA

GDB Meeting / 190

Introduction

Corresponding Author: ian.collier@stfc.ac.uk

GDB Meeting / 191

ASGC Report

GDB Meeting / 192

HTCondor/ARC Workshop Report

GDB Meeting / 193

KISTI – Tapeless archive project

GDB Meeting / 194

Joint HSF-OSG-WLCG Workshop Report

Corresponding Author: ian.collier@stfc.ac.uk

GDB Meeting / 195

Belle II Report

GDB Meeting / 196

LHCONE Development in Asia

Soundscape Conference / 197

Immersive with Sound: City Soundscape Crowdsourcing with Virtual Reality Immersive Experience

GDB Meeting / 198

IPv6 Deployment Report

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