Student Projects within CERN IT-SDC

The Worldwide LHC Computing GRID (WLCG) is a global collaboration of more than 170 computing centers in 40 countries linking up national and international GRID infrastructures. WLCG provides global computing resources to store, distribute and analyse ~30 Petabytes of data annually generated by the Large Hadron Collider (LHC) at CERN.
With more than 350.000 cores WLCG is a highly complex and powerful infrastructure for scientific computing.
The projects we offer students to collaborate in as part of their studies are linked to the evolution of the infrastructure and the software components used to operate and monitor it. The students will be integrated in the CERN WLCG team and visits to CERN are foreseen as an integral part of the activity.

Automation and Integration of WLCG Accounting Information

Project description:
The use of the WLCG distributed resources and services is being constantly monitored. Comparison of the pledged resources and the real utilization is an important indicator of site performance.
At the end of every month information covering the activities of thousand of users at hundreds of sites is aggregated and summarised in the monthly WLCG accounting reports, which provide a complete site overview of the sites performance. This data has to be highly precise, because it is also presented to funding agencies.
Since the different computing centers are based on a variety of technologies the combination and validation of the data is non trivial.

The goal of the project is to enable automatic generation of WLCG Accounting Reports [1] using the existing tool chain as a starting point for the development of the required components that will be integrated into the existing framework.

Learning experience:
You will work with the biggest science GRID infrastructure in the world, learning how production quality in a globally distributed infrastructure can be ensured, how resources are accounted and quality is evaluated. You will acquire experience in using ORACLE databases, learn about mature software lifecycle management procedures in a large-scale Python project and gain experience in the development of user interfaces.

Skills: Python

Project duration: 2 months

Contact for further details: Julia.Andreeva@cern.ch

[1] https://espace2013.cern.ch/WLCG-document-repository/Accounting/Tier-1/2014/august-14/Master_accounting_summaries_August2014.pdf

Extending the functionality of the WLCG Site Status Board

Project description:
The WLCG Site Status Board (SSB) is a framework to collect various monitoring metrics, store them and to provide an advanced user interface to visualize the data. The system is widely used by the LHC experiments to assist in the distributed operations and site commissioning activities. A continuous stream of data from 170 grid sites is providing the information that has to be analysed and displayed. Extensive use of the SSB brings new requirements from the LHC computing community.
The SSB provides vast data storage and complex data processing functionality. For example, it is used to calculate the availability and reliability of the distributed infrastructure based on the results of the remote tests. From time to time it is necessary to correct spurious measurements to maintain a consistent assessment of the quality of sites and services. The challenge is to provide an editing functionality that is easy to use with security enabled access control and logging mechanisms.

Learning experience:
The student will gain practical experience in the development of the large-scale flexible monitoring system and modern web technologies. He/she will learn how to handle the security implications in the web UIs.

Skills required: Python, SQL, Javascript
Project duration: 4 months

Contact for further details: Julia.Andreeva@cern.ch

Cyber Security for the Experiment Dashboard Framework

Project description:

The Experiment Dashboard provides a wide range of applications for monitoring of the LHC computing activities on the WLCG infrastructure. Certain policies define privileges for data access, recording and modifications. The goal of the project is to improve the authentication and authorisation system for the Experiment Dashboard Applications.

Currently the authorisation in the Dashboard applications is based on the User's credentials recorded in the GRID certificate. Most of applications use the X509 certificates to grant different permissions to various categories of users (like admin rights, or creation of the new metrics).

Potential improvements in this area:

1. Introduce Single Sign On for authentication, and make it easy to use for any Dashboard application.
2. Use e-groups for authorisation enabling rules like 'all people belonging to a particular e-group will have admin rights'. E-group is an interface to manage groups at CERN. Authorization based on e-group would provide an easy way to delegate authorization/authentication policy implementation to the group managers rather than to the support team of the monitoring services.

Learning experience :
The student will learn about various options for the implementation of the authentication/authorization for the web applications and will gain experience in choosing and implementing the most appropriate authentication/authorization technique, as well as experience in testing, deployment and validation of the authentication/authorization components.

Skills required: Python, interest in cyber security
Project duration: 3 months

Contact for further details: Julia.Andreeva@cern.ch

Design of a highly functional Web presence for the WLCG Experiment Dashboard Portal using Drupal

Project description:

The WLCG Experiment Dashboard portal provides an entry point for the WLCG monitoring data for thousands of users around the world and brings together the distributed development community. The goal of the project is to redesign the existing portal [1] in accordance with modern web standards, improving its' content, layout, user experience and usability. The new portal should be developed using Drupal [2], the Content Management Framework supported by the CERN IT department. Possible integration with modern distributed revision control and project management systems GIT[3] and JIRA[4] should be investigated.

Learning experience:

The student will acquire practical experience of building web applications using a state of the art Content Management Framework integrated with modern distributed revision control and project management systems (GIT and JIRA)

Skills required or build during the project: HTML, CSS, Javascript
Project duration: 1-3 months (based on prior experience)

[1] http://dashboard.cern.ch
[2] www.drupal.org
[3] git-scm.com
[4] www.atlassian.com

Contact for further details: Julia.Andreeva@cern.ch

Proactive monitoring of dynamic Virtual Machine Clusters

Project description:
The LHC experiments are progressively moving towards computing resources which are provided dynamically by Cloud services. It is important to monitor the health and performance of the virtual machines of these dynamic cluster and to provide early warnings in order to prevent the problems of degraded service and interruptions due to eventual failures of the cluster nodes.The goal of the project is to develop a system that will digest monitoring information coming from the cluster, analyze it almost in real time and provide necessary input for the control engine of the workload management systems of the experiments.
The system should be generic and not coupled to any experiment frameworks, so that it can be used by any LHC experiment. The Esper event processing technology is considered for the implementation, though other alternatives should be evaluated.

Learning experience:
The project offers an opportunity to take part in the construction of the sophisticated self-regulated work-load management systems aimed to use the provided heterogeneous computing resources in a most efficient way. The main challenge consists of creating a generic performant decision-taking unit integrated with the computing systems of the LHC experiments.

Technologies used: Ganglia [1], Esper[2]
Project duration: 3 months

[1] ganglia.sourceforge.net
[2] esper.codehaus.org

Contact for further details: Julia.Andreeva@cern.ch

WLCG Storage Space Monitor

Project description

The Storage Space Monitor should gain at-a-glance insight into storage occupancy and capacity for the globally distributed WLCG storage resources. Reliable Storage Space monitoring plays a key role for optimizing the data distribution and storage capacity planning. The goal of the project is to prototype the storage space monitor for the CMS experiment based on information which is already being collected from the CMS computing centers. In the scope of the project, the ATLAS Grid Information System has to be evaluated as a generic topology descriptor of the storage resources. The monitoring data should be exposed via an informative, responsive and intuitive user interface. The Storage Space Monitor prototype has to be sufficient generic so that it can be adapted to all LHC experiment.

Learning experience:

The student will take part in the development of a large-scale distributed monitoring system integrated with various data storage technologies. Good understanding of the topology of the WLCG infrastructure described in an ORACLE database will be required. The student will gain practical experience in the development of the responsive and intuitive UI using modern web technologies.

Skills and technologies involved: ORACLE, Python, Django, JavaScript
Project duration: 3 months

Contact for further details: Julia.Andreeva@cern.ch

Design an online data serving layer for WLCG Monitoring Data based on modern analytics technologies

Project description:

The current WLCG monitoring system has proven to be a solid and reliable solution to support WLCG during LHC data-taking years. A variety of data coming from different services and experiment-specific frameworks is gathered, processed and archived and a generic web-based dashboard provides a uniform and customisable monitoring interface for scientists and sites. The WLCG monitoring applications handle large volumes of data collected from more than 170 computing centers. In the near future, the WLCG monitoring infrastructure has to cope with an extension of the volume (steadily growing rate of job processing , data access and data transfers) and the variety (e.g. new data-transfer protocols and new resource-types, as cloud-computing) of the monitoring data. Tens of TBytes are stored to allow users to analyse historical data in detail.
Traditional architectures in monitoring, where relational database systems are used to store, to process and to serve monitoring events, face at the given scale and complexity limitations.

The goal of the project is to evaluate NoSQL technologies, in particular Elasticsearch, as a data **serving layer** for the WLCG monitoring data using indexing techniques to make it efficiently query-able. This includes integration of the existing User Interface, which is based on ORACLE database technology, with an Elasticsearch backend.

Learning experience:
The project offers an opportunity to participate in the partial redesign of the data flow of the large scale monitoring system for a fully distributed production quality GRID infrastructure. The main challenge of the project is to offer a scalable and performant solution based on the Elasticsearch technology for serving data to the responsive UI and capable to cope with steadily growing amount and complexity of the monitoring data. In addition the student will be given the opportunity to gain hands on experience with modern analytics techniques like Elasticsearch

Skills required or obtained during the project: SQL, Python, Elasticsearch
Project duration: 6 months

Contact for further details: Julia.Andreeva@cern.ch

Applying Lambda architecture for the WLCG monitoring data processing

Project description:
The current WLCG monitoring system has proven to be a solid and reliable solution to support WLCG functions and operations during LHC data-taking years. A variety of data coming from different services and experiment-specific frameworks is gathered, processed and archived and a generic web-based dashboard provides a uniform and customisable monitoring interface for scientists and sites. The WLCG monitoring applications handle large volumes of data collected from more than 170 computing centers. In the near future, the WLCG monitoring infrastructure has to cope with an extension of the volume (steadily growing rate of job processing , data access and data transfers) and the variety (e.g. new data-transfer protocols and new resource-types, as cloud-computing) of the monitoring data. However traditional architectures, where relational database systems are used to store, to process and to serve monitoring events, have limitations.
Lambda architecture is a data-processing architecture designed to handle massive quantities of data by taking advantage of both batch- and stream- processing data. The goal of the project is to apply Lambda architecture for the processing of the WLCG monitoring data. This implies evaluation of the Hadoop/MapReduce technology as a **batch layer** for processing of the constantly growing WLCG monitoring data providing the ability to compute arbitrary functions on it.

Learning experience:
The project offers an opportunity to participate in the partial redesign of the data flow of the large scale monitoring system for a fully distributed production quality GRID infrastructure. The redesign will follow the Lambda architecture paradigm. The main challenge of the project is to offer a scalable and performant solution for data processing based on map-reduce technology capable to cope with steadily growing amount and complexity of the monitoring data. Hands on experience with a real world application of advanced data analytics technologies.

Skills related to the project: SQL, Python, Java, Hadoop/MapReduce (to be learned during the project)
Project duration: 6 months

Contact for further details: Julia.Andreeva@cern.ch

Using data analytics for WLCG data transfer optimization

Project description:
The overall success of LHC data processing depends heavily on the stable, reliable and fast data distribution performed by the WLCG File Transfer Service (FTS). FTS transfers around 15 PB of data each month representing millions of files per day. The efficient functioning of this service is crucial for successful exploitation of the LHC data. The large scale of the transfer activity and the shared nature of the LHC computing infrastructure, which is used by several virtual organizations, create a challenge for the FTS service.
The project proposes the exploration of the FTS historical monitoring data with the aim of improving the service efficiency. Data analysis should consider all kinds of transfer routes, protocols, and experiments' data transfer workflows with various FTS configurations. The goal of the project is to assist the FTS3 infrastructure to sustain higher traffic while optimizing the resource usage and reducing data transfer latencies. This includes creating a data analytics platform for the FTS performance analysis and predictions.

Learning experience:
The project offers an opportunity to contribute to the evolution of the WLCG data transfer service by taking part in the design and implementation of the new analytics platform to support big data analysis using large-scale data processing and storage technologies (such as MapReduce, HDFS, NoSQL...).The student will get understanding of the Grid, data transfer tools and workflows of the LHC experiments.
Skills required: Python, SQL and basic knowledge of TCP/IP protocol.
Project duration: 9 months
Contact for further details: oliver.keeble@cern.ch

Exposing network information via distributed storage services

The LHC experiments have advanced data distribution systems which manage the global transfer of their multi Petabyte data sets. These transfers can be optimised using information about the wide area network connecting the two endpoints involved. The task is to design and implement a simple service which can be installed with any grid storage element and can return to the client the observed round trip times and packet loss rates to a second storage system. By making this information available to the file transfer service, we aim to increase the efficiency of the LHC data distribution.

Learning and Experience: This project will offer an insight into the LHC's global data distribution network and will give the chance to become familiar with specialised systems employed for storage and transfer. Problems with wide area networking will have to be understood and overcome, with the chance of helping to get LHC data delivered faster than ever.

Skills: Python, REST, TCP/IP

Project Duration: 3 months

Reference: http://fts3-service.web.cern.ch/

Contact for further details: oliver.keeble@cern.ch

The CERN volunteer computing platform

CERN-IT is developing a volunteer computing solution intended to be a common platform for the LHC experiments' activities in this area and which should help to maximise the number of cycles they can acquire. The task is to accompany this project through its initial prototyping, work on all problems discovered and help to guide it to the level of maturity required for production. A major component of the system is based on the storage federation technology of the group ("dynafed") which mediates data transfer between the trusted grid infrastructure and the untrusted volunteer domain.

Learning and Experience: The project will give a chance to work on enabling a potentially large computing resource for the LHC experiments, and will give insights into many of the challenges involved in taking a sophisticated service into production. It will offer an opportunity to get to know how distributed computing solutions are built and managed.

Skills: C++/Linux, Virtualisation technology, HTTP

Project Duration: 3 months

Contact for further details: oliver.keeble@cern.ch

The potential of HTTP proxy caches for LHC computing

Managing storage is one of the major contributors to operational costs on the LHC's grid infrastructure (WLCG). The task is to design and prototype an HTTP proxy cache system, built using standard components, intended to allow pure unmanaged cache storage at a grid site or to accelerate data access in cloud environments.This project could reduce the costs of running the LHC's grid infrastructure by removing storage management overheads at smaller sites and by improving the efficiency of cloud computing resources.

Learning and Experience: This project offers the opportunity to understand how advanced, peta-scale storage systems work and to get to grips with the technicalities of how caching can be used to improve efficiency.

Skills Required: System integration, HTTP, Linux, Python

Project Duration: 6 months

Contact for further details: oliver.keeble@cern.ch

Dynamic storage federations

The group runs a project whose goal is the dynamic federation of HTTP based storage systems, allowing a set of globally distributed resources to be integrated and appear via a single entry point. The task is to work on the development of this project ("dynafed"), implementing functional and performance extensions, in particular

- Redirection monitoring, to allow the logging of federator behaviour for real-time monitorng and subsequent analytics
- Metadata integration, beginning with the incorporation of space usage information, allowing the federator to expose grid-wide storage metrics

Learning and Experience: this project offers experience in how advanced, distributed storage systems are being used to handle the peta-scale data requirements of the LHC experiments. It will offer the chance to become familiar with HTTP technologies and federation concepts.

Skills Required: C++/Linux

Project Duration: From 3 months, depending on task selected

Reference: https://svnweb.cern.ch/trac/lcgdm/wiki/Dynafeds

Contact for further details: oliver.keeble@cern.ch

System modelling

Investigate the potential of analytical models and discrete event simulation of the group's storage and transfer systems, in order to optimise them in their present use cases and extrapolate their scaling behaviour.

Learning and experience: The project will offer an insight into both the selected subject of the study (an advanced data management system) and into the use of modelling in order to understand behaviour at scales inaccessible to practical experimentation.

Skills Required: Familiarity with some of the following would be advantageous: system modelling, networking and i/o, comparative analysis of performance related data

Project Duration: 12 months

Contact for further details: markus.schulz@cern.ch

Implementing 3rd party copy support in HTTP

The group is pursuing ways in which HEP computing workflows can be achieved using standard components and protocols and part of this work focuses on the use of the HTTP protocol. One area where the protocol lacks support is for 3rd party copies, where a client initiates a direct transfer between two storage systems. The task is to complete a formal description of an extension to HTTP which will allow the necessary methods, including delegation of credentials, to enable push or pull 3rd party copy. The extensions should be described as a standards document and implemented for the DPM storage system.

Learning and Experience: Acquire a deep understanding of the protocol underpinning the web, HTTP. Gain exposure to distributed storage for big data and an appreciation of how the standards process works.

Skills Required: HTTP, C++/Linux

Project Duration: 6 months

Contact for further details: oliver.keeble@cern.ch

Distributed storage systems for big data

The group maintains a framework called dmlite which is used to integrate various types of storage with different protocol frontends. It is the basis of a number of the group's products such as the Disk Pool Manager (DPM), a grid storage system which holds over 50PB of storage in the global infrastructure.

DPM/dmlite extensions

The task is to contribute to the dmlite project by working on functional extensions to the framework. Example projects include

- Exposing system data through a "procfs" style plugin
- Incorporation of new AA mechanisms, eg outh
- Creation of a web admin interface
- Work on draining and file placement within the system

dmliteSE

Help to realise the group's vision of a "dmliteSE" by working on the gradual retirement of legacy daemons within the DPM system. In this context, tackle the modernisation of pool management and file placement, and the incorporation of new resource types (eg cluster file systems) into the system. Complete the functional development required to allow operation of a disk storage system purely through standard protocols.

For both sub projects above:

Skills Required: C++/Linux

Reference: https://svnweb.cern.ch/trac/lcgdm/wiki/Dpm

Learning and Experience: This project offers the chance to become involved with one of the storage systems used in computing for LHC and will give an opportunity to become familiar with big data storage and system programming.

Project Duration: From 3 months, depending on task selected

Contact for further details: oliver.keeble@cern.ch

Code quality and software lifecycle tools for reducing maintenance costs

Improve the group's code quality assurance and lifecycle management by understanding the advantages of relevant technologies and implementing solutions where appropriate:

- Collaborative development tools, eg
    - code review
    - github & co.
- Use of static analysis tools

- Release automation, jira/bamboo/drupal workflow
- Comparison of existing solutions (eg Bamboo & Jenkins)

Learning and Experience: This task will offer valuable experience in understanding how software quality assurance works and its place in the lifecycle management of a project. It will offer exposure to a number of standard tools which are widely used in large software projects.

Skills Required: C++/Linux, Some familiarity with the concepts of Software lifecycle management

Project Duration: 3 months

Contact for further details: oliver.keeble@cern.ch

File Transfer Service (FTS) extensions

The File Transfer Service (FTS) manages the global distribution of LHC data, moving multiple petabytes per month during a run and underpinning the whole data lifecycle. Join the FTS team in their development of this critical service. Possible projects include

- authorised proxy sharing: allowing a production service to delegate a proxy and authorising others to use it
- incorporation of support for new types of endpoint, for example cloud or archival storage

Learning and Experience: This project offers the chance to become involved with one of the critical data management systems used in computing for LHC and will give an opportunity to become familiar with big data storage, wide area networking and system programming.

Skills Required: C++/Linux, Python

Reference: http://fts3-service.web.cern.ch/

Project Duration: From 3 months, depending on task selected

Contact for further details: oliver.keeble@cern.ch

Augmenting Sync and Share services

Sync and share services (such as dropbox) are increasing rapidly in popularity. They offer a number of opportunities, through incorporating support for existing grid systems or exploiting sync logic in novel environments. The task is to investigate how the group's solutions can complement efforts to create sync and share services. Aspects include the following investigations;

- the role of the group’s federation technology to accelerate access and to supply placement logic
- how institutional services can be federated to create a single 'sync & share' universe

- global distribution strategies; the role of existing solutions for supplying the workflow and the use of FTS as an asynchronous file mover; relevance for dataset replication/synchronisation between cloud data centres for other sciences.
- integration of davix into sync clients to enable grid storage to be accessed

Learning and Experience: This task offers exposure to the rapidly evolving area of sync and share services and, through considering integration with the existing LHC distributed computing infrastructure, will provide experience in current HEP computing models and the data services which support them.

Skills Required: C++/Linux, System integration

Project Duration: From 3 months, depending on task selected

Contact for further details: oliver.keeble@cern.ch