3rd Rucio Community Workshop

Report of Contributions

https://indico.cern.ch/e/RUCIO2020

Contribution ID: **1** Type: **not specified**

Keynote: ESnet: DOE's data circulatory system

Wednesday 11 March 2020 09:00 (45 minutes)

Presenter: MONGA, Inder (ESNet)

Contribution ID: **14** Type: **not specified**

Operational Intelligence - General Introduction

Wednesday 11 March 2020 13:30 (20 minutes)

In the near future, large scientific collaborations will face unprecedented computing challenges. Processing and storing exabyte datasets require a federated infrastructure of distributed computing resources. The current systems have proven to be mature and capable of meeting the experiment goals, by allowing timely delivery of scientific results. However, a substantial amount of interventions from software developers, shifters and operational teams is needed to efficiently manage such heterogeneous infrastructures. On the other hand, logging information from computing services and systems is being archived on ElasticSearch, Hadoop, and NoSQL data stores. Such a wealth of information can be exploited to increase the level of automation in computing operations by using adequate techniques, such as machine learning (ML), tailored to solve specific problems. The Operational Intelligence project is a joint effort from various WLCG communities aimed at increasing the level of automation in computing operations. We discuss how state-of-the-art technologies can be used to build general solutions to common problems and to reduce the operational cost of the experiment computing infrastructure.

Presenter: LEGGER, Federica (Universita e INFN Torino (IT))

Session Classification: Operational Intelligence meeting

Contribution ID: **15** Type: **not specified**

Framework Design

Wednesday 11 March 2020 14:15 (25 minutes)

This contribution describes how and why we decided to create the "OpInt Framework", what it offers and how we architected it. Last year we began the development of the "Rucio OpInt" project in order to optimise the operational effort and minimize human interventions in the distributed data management.
When we brought "Rucio OpInt" to the Operational intelligence forum we realized that there were a lot of shared requirements with other projects and there was a need for the creation of a framework that hosts all those shared components. After researching the open source market and realizing there was not an out of the box solution we decided to architect our own solution which offers APIs, authentication, authorization, source data fetching mechanisms and machine learning pipelines to the whole OpInt community.

Presenter: PAPARRIGOPOULOS, Panos (CERN)

Session Classification: Operational Intelligence meeting

Contribution ID: **16** Type: **not specified**

Automation of Rucio operations

Wednesday 11 March 2020 13:50 (25 minutes)

Presenter: CHRISTIDIS, Dimitrios (University of Texas at Arlington (US))

Session Classification: Operational Intelligence meeting

Contribution ID: **17** Type: **not specified**

JobsBuster

Wednesday 11 March 2020 14:40 (25 minutes)

Reliable automatization of the root cause analysis procedure is an essential prerequisite for the Operational Intelligence deployment. That kind of data processing is important as an input for the automatic decision making and has its own value as an instrument for offloading shifters operations. The order of magnitude of failing rate in distributed computing, for instance in ATLAS experiment, is the tenth thousand jobs a day. This is why manual problem identification requires sufficient efforts. We created a prototype of the system, which finds the least common denominator for the computational jobs failures called Jobs Buster. In this talk, we provide an overview of this system, its current status and development plans.

Presenter: PADOLSKI, Siarhei (BNL)

Session Classification: Operational Intelligence meeting

Contribution ID: **18** Type: **not specified**

Job outcome prediction with Google's AutoML Tables

Wednesday 11 March 2020 15:05 (25 minutes)

Presenters: RETZKE, Kevin Michael (Fermi National Accelerator Lab. (US)); BHAT, Shreyas

Session Classification: Operational Intelligence meeting

Contribution ID: **19** Type: **not specified**

Welcome to Fermilab

Tuesday 10 March 2020 09:00 (20 minutes)

Presenter: SEXTON-KENNEDY, Elizabeth (Fermi National Accelerator Lab. (US))

Session Classification: Welcome & Introduction

Contribution ID: **20** Type: **not specified**

Logistics

Tuesday 10 March 2020 09:20 (10 minutes)

Presenter: BENELLI, Gabriele (Brown University (US))

Session Classification: Welcome & Introduction

Contribution ID: **21** Type: **not specified**

Introduction

Tuesday 10 March 2020 09:30 (15 minutes)

Presenter: BARISITS, Martin (CERN)

Session Classification: Welcome & Introduction

Contribution ID: **22** Type: **not specified**

Keynote I: Quantum Computing at Fermilab

Tuesday 10 March 2020 09:45 (45 minutes)

Presenter: Dr LYON, Adam (Fermilab)

Session Classification: Welcome & Introduction

Contribution ID: **23** Type: **not specified**

ATLAS (Remote)

Tuesday 10 March 2020 11:00 (20 minutes)

Presenters: SOUTH, David Michael (Deutsches Elektronen-Synchrotron (DE)); LASSNIG, Mario (CERN)

Session Classification: Community reports

Contribution ID: **24** Type: **not specified**

ESCAPE Project (Remote)

Tuesday 10 March 2020 11:20 (20 minutes)

The ESCAPE European Union funded project aims at integrating facilities of astronomy, astroparticle and particle physics into a single collaborative cluster or data lake. The data requirements of such data lake are in the exabyte scale and the data should follow the FAIR principles (Findable, Accessible, Interoperable, Re-usable). To fulfill those requirements significant RnD is foreseen with regards to data orchestration, management and access. To set up the ESCAPE data lake, Rucio will be used as a reference implementation. We are therefore contributing to the Rucio development, integration and commissioning effort, particularly for the functionalities needed by the ESCAPE partners.

Presenter: FKIARAS, Aristeidis (CERN)

Session Classification: Community reports

Contribution ID: **25** Type: **not specified**

DUNE Data Management Experience with Rucio

Tuesday 10 March 2020 11:40 (20 minutes)

The DUNE collaboration has been using Rucio since 2018 to transport data to our many European remote storage elements. We currently have 13.8 PB of data under Rucio management at 13 remote storage elements. We present our experience thus far, as well as our future plans to make Rucio our sole file location catalog.
We will present our planned data discovery system, and the role of Rucio in the data ingest system and data delivery of files to jobs. We will describe the associated metadata service which is in development. Finally we will describe some of the unique challenges of configuring Rucio to the tape-backed dCache/Enstore disk store at Fermilab.

Presenter: TIMM, Steven (Fermi National Accelerator Lab. (US))

Session Classification: Community reports

Contribution ID: **26** Type: **not specified**

Using Rucio for LCLS (Remote)

Tuesday 10 March 2020 12:00 (20 minutes)

We will describe our plans for using RUCIO within the data management system at the Linac Coherent Light Source (LCLS) at SLAC. An overview of the LCLS data management system will be presented and what role RUCIO will play for cataloging, distributing and archiving of the data files. We are still in the testing phase but plan to use RUCIO in production within the next few month.

Presenter: KROEGER, Wilko (SLAC National Accelerator Laboratory)

Session Classification: Community reports

Contribution ID: **27** Type: **not specified**

Kubernetes & Rucio (Remote)

Tuesday 10 March 2020 13:30 (15 minutes)

Presenter: BEERMANN, Thomas (Bergische Universitaet Wuppertal (DE))

Session Classification: Technical Discussions

Contribution ID: **28** Type: **not specified**

Tales From ATLAS DDM Operations (Remote)

Tuesday 10 March 2020 13:45 (15 minutes)

Presenter: CHRISTIDIS, Dimitrios (University of Texas at Arlington (US))

Session Classification: Technical Discussions

Contribution ID: **29** Type: **not specified**

CRIC: Computing Resource Information Catalogue as a topology system for computing infrastructures and an interface for effortless Rucio configuration (Remote)

Tuesday 10 March 2020 14:00 (20 minutes)

CRIC is a high-level information system which provides flexible, reliable and complete topology and configuration description for a large scale distributed heterogeneous computing infrastructure. CRIC aims to facilitate distributed computing operations for HEP experiments and consolidate WLCG topology information. Being a topology framework, CRIC offers a generic solution with out of the box interfaces, APIs, authentication and authorisation mechanisms, advanced logging and much more. Every community, small or big, can take advantage of CRIC's capabilities. In close collaboration with the Rucio team, CRIC can provide interfaces to configure Rucio and tie this configuration with the actual topology of the computing infrastructure of any Rucio user. Configuring RSEs, running on top of the same physical storage, through CRIC can drastically minimise the number of attributes that need to be filled by Rucio operators. The complex transfer matrix between all the RSEs can be bootstrapped and maintained through a simple table and all the information regarding Users and permissions can be organised through CRIC's A&A system and propagated into Rucio.
The contribution describes the overall CRIC architecture, the new lightweight-CRIC standalone service that can be easily installed and how with minimum effort one can fully exploit Rucio's capabilities using the CRIC framework.

Presenter: PAPARRIGOPOULOS, Panos (CERN)

Session Classification: Technical Discussions

Contribution ID: **30** Type: **not specified**

Discussion

Tuesday 10 March 2020 14:20 (20 minutes)

Session Classification: Technical Discussions

Contribution ID: **31** Type: **not specified**

Belle II (Remote)

Tuesday 10 March 2020 14:40 (20 minutes)

Presenter: SERFON, Cedric (Brookhaven National Laboratory (US))

Session Classification: Community reports

Contribution ID: **32** Type: **not specified**

Data Management Needs in Cancer Research (Remote)

Tuesday 10 March 2020 15:00 (20 minutes)

MSKCC's Computational Oncology group performs prospective and retrospective studies on a number of cancer types with a focus on cancer evolution. The data being collected and managed for research comes from many sources. Broadly, the data may be categorized into molecular, imaging and clinical data types. The studies tend to be cross-sectional and longitudinal. Users require heterogenous permissions to the data with varying levels of control. The storage and compute infrastructure is expected to span on-premise clusters, public and private clouds. This presentation will elaborate on the group's data management needs in comparison to Rucio's current feature set.

Presenter: PASHA, Arfath (MSKCC)

Session Classification: Community reports

Contribution ID: **33** Type: **not specified**

dCache QoS and Storage Events

Wednesday 11 March 2020 13:30 (20 minutes)

dCache is highly scalable distributed storage system that is used to
implement storage elements with and without tape back-ends.
dCache is offering a comprehensive RESTFul data management interface
that uses language of QoS states and transitions to steer the data
life-cycle. This interface provides functionality inspired by the
experiences of the LHC and other data intensive experiments. Additionaly,
dCache provides storage events - a publish-subscribe notification
subsystem which lends itself to greater scalability compared to
polling dCache for data states. A data management system like Rucio
can take advantage of these features to provide a more robust,
efficient and scalable data delivery solution.

Presenter: LITVINTSEV, Dmitry (Fermi National Accelerator Lab. (US))

Session Classification: Technical Discussions

Contribution ID: **34** Type: **not specified**

CERN Tape Archive status and plans (Remote)

Wednesday 11 March 2020 13:50 (20 minutes)

CTA is designed to replace CASTOR as the CERN Tape Archive solution, in order to face scalability and performance challenges arriving with LHC Run-3.

This presentation will focus on the current CTA deployment and will provide an up-to-date snapshot of CTA achievements.

It will also cover the final Run3 CTA Service architecture and underlying hardware that have been deployed at the end of 2019.

Presenter: LEDUC, Julien (CERN)

Session Classification: Technical Discussions

Contribution ID: **35** Type: **not specified**

Connecting Xcache and RUCIO for User Analysis (Remote)

Wednesday 11 March 2020 14:10 (20 minutes)

Presenter: YANG, Wei (SLAC National Accelerator Laboratory (US))

Session Classification: Technical Discussions

Contribution ID: **36** Type: **not specified**

Discussion

Wednesday 11 March 2020 14:30 (20 minutes)

Session Classification: Technical Discussions

Contribution ID: **37** Type: **not specified**

Rucio at RAL and the UK

Wednesday 11 March 2020 09:45 (20 minutes)

Presenter: ELLIS, Katy (Science and Technology Facilities Council STFC (GB))

Session Classification: Community reports

Contribution ID: **38** Type: **not specified**

CMS Transition to Rucio

Wednesday 11 March 2020 10:05 (20 minutes)

An update on the CMS transition to Rucio, expected to be completed this year, will be given.

Results of scale tests, data consistency work, and improvements in the kubernetes infrastructure will be the focus of this talk.

Presenter: VAANDERING, Eric (Fermi National Accelerator Lab. (US))

Session Classification: Community reports

Contribution ID: **39** Type: **not specified**

EGI data management requirements and plans (Remote)

Wednesday 11 March 2020 11:00 (20 minutes)

The Data Management requirements coming from the EGI and EOSC-Hub user communities have pictured Rucio (together with a Data transfer engine) as one of the possible solutions for their needs. Since the 2nd Rucio workshop a number of enhancements and new developments (in primis the support for OIDC and the kubernetes deployment improvements) have been implemented and they are going towards the direction of an easier integration of Rucio in the EGI and EOSC environment. We would like in this talk to highlight the progress done and the desired missing functionalities/integration activities.

Presenter: MANZI, Andrea

Session Classification: Community reports

Contribution ID: **40** Type: **not specified**

iDDS: A New Service with Intelligent Orchestration and Data Transformation and Delivery (Remote)

Wednesday 11 March 2020 11:20 (20 minutes)

The Production and Analysis system (PanDA system) has continuously been evolving in order to cope with rapidly changing computing infrastructure and paradigm. The system is required to be more dynamic and proactive to integrate emerging workflows such as data carousel and active learning, in contrast to conventional HEP workflows such as Monte-Carlo simulation and data reprocessing.

Intelligent Data Delivery Service (iDDS) is an experiment agnostic service to orchestrate workload management and data management systems, in order to transform and deliver data and let clients consume data in near real-time. iDDS has been actively developed by ATLAS and IRIS-HEP. iDDS has a modular structure to separate core functions and workflow-specific plugins to meet a diversity of requirements in various workflows, simplify the development and operation of new workflows, and provide a uniform monitoring view. The goal of iDDS is the seamless integration of new workflows as well as to address performance issues and suboptimal resource usage in existing workflows.

This talk will report architecture overview of iDDS, orchestration of PanDA and Rucio for optimal storage usage in data carousel, dynamic task chaining in ATLAS production system with instant decision making for active learning, data streaming with on-demand marshaling to minimize data delivery from data ocean to analysis facilities and users, integration of iDDS with other workload management systems, and plans for the future.

Presenter: GUAN, Wen (University of Wisconsin (US))

Session Classification: Technical Discussions

Contribution ID: **41** Type: **not specified**

Rucio-DIRAC Integration at Belle II (Remote)

Wednesday 11 March 2020 11:40 (10 minutes)

Presenter: SERFON, Cedric (Brookhaven National Laboratory (US))

Session Classification: Technical Discussions

Contribution ID: **42** Type: **not specified**

Discussion

Wednesday 11 March 2020 11:50 (20 minutes)

Session Classification: Technical Discussions

Contribution ID: **43** Type: **not specified**

IceCube

Thursday 12 March 2020 09:00 (20 minutes)

Presenter: SCHULTZ, David (University of Wisconsin-Madison)

Session Classification: Community reports

Contribution ID: **44** Type: **not specified**

Data management for the XENON Collaboration with Rucio

Thursday 12 March 2020 09:20 (20 minutes)

The search for Dark Matter in the XENON experiment at the LNGS laboratory in Italy enters a new phase, XENONnT in 2020. Managed by the University of Chicago, Xenon's Rucio deployment plays a central role in the data management between the collaboration's end points. In preparation for the new phase, there have been notable upgrades in components of the production and analysis pipeline and they way they interface with Rucio services as such the inclusion of processed data for distribution between the different RSEs. We will describe some of changes in the pipeline and focus in discussing the aDMiX wrapper which calls the rucio API directly to ingest data into Rucio from the DAQ.

Presenter: PASCHOS, Paschalis (University of Chicago)

Session Classification: Community reports

Contribution ID: **45** Type: **not specified**

Rucio at BNL and NSLS2 (Remote)

Thursday 12 March 2020 09:40 (20 minutes)

Rucio has evolved as a distributed data management system to be used by scientific communities beyond High Energy Physics. This includes disengaging its core code from a specific file transfer tool. In this talk I will discuss using Globus Online as a file transfer tool with Rucio, the current state of testing and the possibilities for the future in light of NSLSII's data ecosystem

Presenter: SNYDER, Matthew (Brookhaven National Laboratory)

Session Classification: Community reports

Contribution ID: **46** Type: **not specified**

Rucio @ LIGO-Virgo-KAGRA (Remote)

Thursday 12 March 2020 10:00 (25 minutes)

Presenter: Mr FRONZE', Gabriele Gaetano (University e INFN Torino (IT), Subatech Nantes (FR))

Session Classification: Community reports

Contribution ID: **47** Type: **not specified**

FTS news and plans (Remote)

Thursday 12 March 2020 11:00 (20 minutes)

The File Transfer Service (FTS) is distributing the majority of the LHC data across the WLCG infrastructure and, in 2019, it has transferred more than 800 million files and a total of 0.95 exabyte of data. It is used by more than 28 experiments at CERN and in other data-intensive sciences outside of the LHC and even the High Energy Physics domain.

The FTS team has been very active in performing several significant performance improvements to its core to prepare for the LHC Run-3 data challenges, supporting the new CERN Tape Archival (CTA) system which has been stress tested by the ATLAS Data Carousel activity, supporting a more user-friendly authentication and delegation method using tokens and supporting the Third Party Copy WLCG DOMA activity. This talk will cover all these developments.

Presenter: KARAVAKIS, Edward (CERN)

Session Classification: Technical Discussions

Contribution ID: **48** Type: **not specified**

Multi-VO Rucio

Thursday 12 March 2020 11:20 (20 minutes)

Presenter: CHADWICK, Eli Benjamin (Science and Technology Facilities Council STFC (GB))

Session Classification: Technical Discussions

Contribution ID: **49** Type: **not specified**

Rucio FUSE Interface (Remote)

Presenter: FRONZE', Gabriele Gaetano (University e INFN Torino (IT), Subatech Nantes (FR))

Session Classification: Technical Discussions

Contribution ID: **50** Type: **not specified**

Rucio Roadmap (Remote)

Thursday 12 March 2020 11:40 (20 minutes)

Presenter: BARISITS, Martin (CERN)

Session Classification: Technical Discussions

Contribution ID: **51** Type: **not specified**

Discussion

Thursday 12 March 2020 12:00 (20 minutes)

Session Classification: Technical Discussions

Contribution ID: **52** Type: **not specified**

CANCELLED: Fermilab Colloquium: The Science of LSST and the big data it will produce

Wednesday 11 March 2020 16:00 (1 hour)

https://events.fnal.gov/colloquium/events/event/open-17/