IV International Workshop ”Data life cycle in physics”, DLC-2020

Report of Contributions

https://indico.kit.edu/e/806

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

A job management system for utilization of idle supercomputer resources

Wednesday, June 10, 2020 10:00 AM (15 minutes)

We propose a system for executing low-priority non-parallel jobs on idle supercomputer resources to increase the effective load of the resources. The jobs are executed inside containers so the checkpoint mechanism can be used to save the state of the jobs during the execution and resume it on a different node. Thanks to splitting the execution of the low-priority jobs into separate shorter intervals, the system can utilize idle computational nodes with little impact on performance with respect to the regular jobs.

The system consists of two components. The first component is a control program that maintains a queue of low-priority non-parallel jobs, assigns both the new jobs and the jobs saved as checkpoints to computational nodes, tracks their status, and manages the checkpoints. It also interacts with the supercomputer scheduler. The second component is an agent program that is executed on computational nodes and interacts with container software by starting the assigned jobs inside containers and saving the progress as checkpoints before the allotted time is over.

Based on our estimates, under varying assumptions 40% to 89% of the idle resources can be effectively utilized with the proposed system.

Authors: FEDOTOVA, Elena (SINP MSU); POLYAKOV, Stanislav (SINP MSU); DUBENSKAYA, Julia (SINP MSU); NGUYEN, Minh-Duc

Presenter: FEDOTOVA, Elena (SINP MSU)

Session Classification: Session 5

Contribution ID: **2** Type: **not specified**

Comparison of container virtualization tools for utilization of idle supercomputer resources

Wednesday, June 10, 2020 10:15 AM (15 minutes)

We propose a system to increase the effective load of supercomputer resources. The key idea of the system is that when idle supercomputer nodes appear, low-priority non-parallel jobs are started occupying these nodes until a regular job from the main queue of the supercomputer arrives. Upon arrival of the regular job, the low-priority jobs temporarily interrupt their execution and wait for the appearance of new idle nodes to be resumed there. This approach can be implemented by running low-priority jobs in containers and using the container migration mechanism to freeze these jobs and then run them from the point they were frozen at. While freezing a job, a stateful checkpoint is created that is a collection of files containing all the information for restoring the job execution (in the general case on another computing node).

Thus, the selection of a specific container virtualization system that is best suited to our goal is an important task. Preliminary analysis allowed us to choose Docker and LXC software products, which were compared in more detail.

When comparing the capabilities of Docker and LXC, it was noted that more lightweight Docker containers start and stop somewhat faster than LXC containers, the launch of which is more like starting a classic virtual machine. At the same time, the LXC project has the best support for the ZFS file system, which significantly speeds up the process of writing checkpoints to disk, as well as restoring containers from checkpoints. However, when testing the LXC, a recurring problem was discovered: the same container was correctly restored from a checkpoint only once. Attempts to checkpoint a container that was previously restored from a checkpoint resulted in an error with the loss of the container state. Since our project assumes a multiple checkpoint and restore of the same container as the main scenario, the above LXC feature prevents us from using this technology for the needs of our project. Thus, we opted for the Docker project, which stably and correctly checkpoints/restores any container multiple times while maintaining the current state of the processes. We implemented a prototype system to increase the effective load of supercomputer resources using Docker containers. Testing of the prototype proved the reliability and stability of the proposed approach.

Authors: DUBENSKAYA, Julia (SINP MSU); POLYAKOV, Stanislav (SINP MSU); NGUYEN, Minh--Duc; FEDOTOVA, Elena (SINP MSU)

Presenter: DUBENSKAYA, Julia (SINP MSU)

Session Classification: Session 5

Contribution ID: **3** Type: **not specified**

Modifications to the EMC algorithm for orientation recovery in Single Particle Imaging experiments on X-ray free electron lasers

Tuesday, June 9, 2020 12:00 PM (15 minutes)

The emergence of super-bright light sources - X-ray free electron lasers(XFELs) combined with Single Particle Imaging(SPI) method, makes it possible to obtain nanometer resolution 3D structure of biological particles such as proteins or viruses without needing to freeze them. SPI relies on the "diffraction before destruction"principle, meaning that each sample only produces a single diffraction image before being destroyed by an X-ray pulse. The orientation of the particle in the beam is random for each shot. This gives rise to the problem of orientation recovery, in which an array of 2D diffraction images has to be combined into a single 3D image, necessary for the reconstruction of 3D structure of the studied particle. The orientation recovery problem is most commonly solved by the EMC algorithm[1], which is the most computationally expensive part of data analysis for SPI experiments. In this work we introduce several modifications to the EMC algorithm aimed at improving the quality of reconstruction and/or increasing the algorithm's speed of convergence. We analyse the effectiveness of these modifications using simulated diffraction data. Loh, N. D. & Elser, V. (2009). Phys. Rev. E, 80, 026705.

Presenter: Mr ZOLOTAREV, Sergei (National Research Center "Kurchatov Institute")

Session Classification: Session 4

Contribution ID: **4** Type: **not specified**

A review of methods of resolution estimation for 3D reconstructions for nanoscale biological objects from experiments data on super-bright X-ray free electron lasers (XFELs)

Tuesday, June 9, 2020 11:45 AM (15 minutes)

The ability to investigate 3D structure of biomolecules, such as proteins and viruses, is essential in biology and medicine. With the invention of super-bright X-ray free electron lasers (XFELs) the Single Particle Imaging (SPI) approach allows to reconstruct 3D structures from many 2D diffraction images produced in the experiment by X-rays scattered on the biomolecule exposed in different orientations. Nowadays the Fourier shell correlation (FSC) [1] is the most common metric for estimating global resolution of the obtained 3D structures in SPI experiments, where the resolution is defined as the spatial frequency at which the correlation between two independently reconstructed structures is equal to some given threshold value. The choice of a threshold value is currently a controversial issue. In addition, this approach can't account fact that the quality of reconstruction can be uneven and depend on the specific area of the biomolecule. Thus, the issue of effective resolution estimation methods remains open. In this way we considered various alternative approaches to the resolution estimation from related scientific fields, such as cryogenic electron microscopy (local resolution estimation methods) and optics (digital camera resolution measurement), and analyzed the applicability of these approaches to resolution estimation in SPI experiments on XFELs. 1. Marin Van Heel and Michael Schatz (2005), Fourier shell correlation threshold criteria. Journal of Structural Biology 151(3): 25-262.

Presenter: Ms IKONNIKOVA, Kseniia (National Research Center «Kurchatov Institute»)

Session Classification: Session 4

Contribution ID: **5** Type: **not specified**

AstroDS - A Distributed Storage for Astrophysics of Cosmic Rays. Current Status.

Monday, June 8, 2020 9:45 AM (30 minutes)

The current state of distributed storage for astrophysics of cosmic ray is considered. The main goal of AstroDS is to unite existing astrophysical data storages of a number of existing experimental collaborations, such as TAIGA, TUNKA, CASKADE and others.

Authors: KRYUKOV, Alexander (MSU); NGUYEN, Minh-Duc; MIKHAILOV, Andrey (IDCST)

Co-authors: POLYAKOV, Stanislav (SINP MSU); KAZARINA, Yulia (API ISU); Mr ZHUROV, Dmitriy (Applied Physics Institute of Irkutsk State University); BYCHKOV, Igor (IDSTU SB RAS); DUBENSKAYA, Julia (SINP MSU)

Presenters: KRYUKOV, Alexander (MSU); POLYAKOV, Stanislav (SINP MSU); Mr ZHUROV, Dmitriy (Applied Physics Institute of Irkutsk State University)

Session Classification: Session 1

Contribution ID: **6** Type: **not specified**

Access Rights Management in Decentralized Distributed Computing Systems

Monday, June 8, 2020 10:15 AM (15 minutes)

The paper presents a solution for decentralized management of data access rights in geographically distributed systems with users from different institutions. This implies possible lack of trust between the user groups. The solution is based on the distributed ledger technology (DLT) together with provenance metadata driven data management.

Authors: Dr DEMICHEV, Andrey (SINP MSU); KRYUKOV, Alexander (MSU); PRIKHODKO, Nikolai

Presenter: Dr DEMICHEV, Andrey (SINP MSU)

Session Classification: Session 1

Contribution ID: **7** Type: **not specified**

Data agregation in astroparticle physics

Monday, June 8, 2020 11:30 AM (15 minutes)

The extension of KCDC (KASCADE Cosmic-ray Data Center, https://kcdc.ikp.kit.edu/) project, which allows the user to access the data of the KASCADE, Tunka-133 and Tunka-Rex astrophysical experiments, has been deployed in the framework of German Russian Astroparticle Data Life Cycle Initiative.
The report will describe the organization of this service, its interaction with storages and aggregation services on Russian side (MSU, ISU), as well as the speed and scalability of this solution.

Author: TOKAREVA, Victoria (KIT)

Presenter: TOKAREVA, Victoria (KIT)

Session Classification: Session 2

Contribution ID: **8** Type: **not specified**

Towards a global analysis and data centre for multi-messenger astroparticle physics

Monday, June 8, 2020 9:15 AM (30 minutes)

Towards a Global Analysis and Data Centre for Multi-Messenger Astroparticle Physics

Presenter: HAUNGS, Andreas (KIT)

Session Classification: Session 1

Contribution ID: **12** Type: **not specified**

KCDC Analysis Extension

Monday, June 8, 2020 11:00 AM (15 minutes)

We will introduce a data analysis extension for the KASCADE Cosmic-ray Data Center (KCDC), based on the Jupyterhub/notebook ecosystem. A user-friendly interface, easy access to data from KCDC, and modern analysis software are of special interest.
This contribution will discuss the service architecture, followed by a brief usage example.

Author: POLGART, Frank (KIT)

Presenter: POLGART, Frank (KIT)

Session Classification: Session 2

Contribution ID: **13** Type: **not specified**

Access Pattern Analysis in the EOS Storage System at CERN

Tuesday, June 9, 2020 11:00 AM (15 minutes)

EOS is a CERN-developed storage system that serves several hundred petabytes of data to the scientific community of the Large Hadron Collider (LHC). In particular, it provides services to the four largest LHC particle detectors: LHCb, CMS, ATLAS and ALICE. Each of these collaborations uses different workflows to process and analyse its data. EOS has a monitoring system that collects detailed information on the file accesses and can give important insights about the specifics of the physics experiments' workflows. In our study, we analyse the monitoring information accumulated over a six months period and amounting to over 1.3 terabytes and have the goal to help the IT department and the experiments' operations teams to better understand the EOS data flows.

In this contribution, we describe a pipeline, mainly developed in R, for processing large volumes of access logs and perform a comparative analysis of the storage usage in scientific workflows. In particular, we calculate aggregated statistics over a six months period and provide a high-level overview of the experiments' data flows. Additionally, we study how the frequency of data accesses changes over time and estimate to what extent different experiments may benefit from an additional caching layer.

Author: CHUCHUK, Olha (CERN, Taras Shevchenko KNU)

Co-author: Dr DUELLMANN, Dirk (CERN)

Presenter: CHUCHUK, Olha (CERN, Taras Shevchenko KNU)

Session Classification: Session 4

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

Educational and outreach resource for astroparticle physics

Wednesday, June 10, 2020 9:45 AM (15 minutes)

The modern astrophysics is moving towards the consolidation and integration of tools aimed at detecting various channels for recording ultrahigh-energy cosmic radiation. In order to obtain reliable data, the experiments should work on the order of several decades, which means that the data will be obtained and analyzed by several generations of physicists. Thus, for the stability of experiments, it is necessary to properly support not only the data life cycle, but also the human aspects, for example, attracting, learning and continuity. To this end, an educational and outreach resource has been deployed in the framework of German-Russian Astroparticle Data Life Cycle Initiative (GRADLCI).

Author: KAZARINA, Yulia (API ISU)

Presenter: KAZARINA, Yulia (API ISU)

Session Classification: Session 5

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

Evaluation of the impact of various local data caching configurations at Tier2/Tier3 WLCG sites

Tuesday, June 9, 2020 10:45 AM (15 minutes)

In this talk we will describe various data caching scenarios and lessons learned. In particular we will talk about local data caches configuration, deployment, and tests. We are using xCache, which is a special type of Xrootd server setup to cache input data for a physics analysis. A relatively large Tier2 storage is used as a primary data source and several geographically distributed smaller WLCG sites configured specifically for this test. All sites are connected to the LHCONE network. The testbed configuration will be evaluated using both synthetic tests and a real ATLAS computational jobs submitted via the HammerCloud toolkit. The impact and realistic applicability of different local cache configurations will be presented, including both the network infrastructure and the configuration of computing nodes.

Authors: KIRYANOV, Andrey (NRC Kurchatov Institute PNPI); ZAROCHENTSEV, Andrey (SPbSU); ALEKSEEV, Alexandr; KORCHUGANOVA, Tatiana; KLIMENTOV, Alexei

Co-authors: OLEYNIK, Danila; SMIRNOV, Serge; MITSYN, Valery

Presenter: KIRYANOV, Andrey (NRC Kurchatov Institute PNPI)

Session Classification: Session 4

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

Big Data Virualization: why and how?

Wednesday, June 10, 2020 9:00 AM (30 minutes)

The fact that over 2000 programs exist for working with various types of data, including Big Data, makes the issue of flexible storage a quintessential one. Storage can be of various types, including portals, archives, showcases, data bases of different varieties, data clouds and networks. They can have synchronous or asynchronous computer connections. Because the type of data is frequently unknown a priori, there is a necessity for a highly flexible storage system, which would allow to easily switch between various sources and systems.
Significant part of the problems can be solved if we use the paradigm of the Virtual Personal Supercomputer, which was developed for computing, but also used to build a framework for distributed ledgers. The idea of this approach is to virtualize not only the processing itself, but also the entire field on which the processing is performed, namely the network, file system and shared memory. This allows you to create a single image of the operating environment, which simplifies the user's work and increases the processing speed.

Authors: Prof. BOGDANOV, Alexander (St.Petersburg State University); Prof. DEGTYAREV, Alexander (St.Petrsburg State University); Prof. SHCHEGOLEVA, Nadezhda (St.Petrsburg State University); Dr KORKHOV, Vladimir (St.Petrsburg State University); Mr KHVATOV, Valery (DGT Technologies AG.)

Presenter: Prof. BOGDANOV, Alexander (St.Petersburg State University)

Session Classification: Session 5

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

The current design and implementation of the AstroDS Data Aggregation Service

Monday, June 8, 2020 11:15 AM (15 minutes)

AstroDS is a distributed storage for Cosmic Ray Astrophysics. The primary goal of Astro DS is to gather data measured by the instruments of various physical experiments such as TAIGA, TUNKA, KASCADE into global storage and provide the users with a standardized user-friendly interface to search for the datasets that match certain conditions. AstroDS consists of a set of distributed microservices components that communicate with each other through the Internet via REST API. The core component of AstroDS is the Data Aggregation Service that orchestrates other components to provide access to data. The development process of AstroDS started in 2019. This paper describes the current design and implementation of the Data Aggregation Service and also the benefits it brings to the astrophysical community in the early state.

Authors: NGUYEN, Minh-Duc; KRYUKOV, Alexander (MSU); MIKHAILOV, Andrey (IDCST)

Presenter: NGUYEN, Minh-Duc

Session Classification: Session 2

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

Development self-trigger algorithms for radio detection of air-showers

Tuesday, June 9, 2020 10:00 AM (15 minutes)

The detection of extensive air-showers with radio method is a relatively young, but promising branch in experimental astrophysics of ultrahigh energies. This method allows one to carry out observations regardless of weather conditions and time of day, and the precision of reconstruction of the properties of primary particles is comparable to the classical methods. The main disadvantage of this method is the complexity of the trigger implementation. Radio signals from extensive air-showers have a duration of few tens nanoseconds and amplitudes comparable to the surrounding background. Moreover, industrial noise, tele- and radio broadcasting signals, as well as noise from the electronic equipment of the experiment, often interfere with measurements. Most of the setups for detecting radio emission from extensive air-showers use an external trigger from optical or particle detectors. Despite numerous attempts to develop autonomous (operating with an internal trigger) cosmic ray radio detectors, there is still no established cost-effective technology for the sparse radio arrays. We give an overview of our progress in this direction, particularly describe noise generator and simulation study using data from Tunka-Rex Virtual Observatory.

Author: FEDOROV, Oleg (ISU)

Presenter: FEDOROV, Oleg (ISU)

Session Classification: Session 3

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

Tunka-Rex Virtual Observatory

Tuesday, June 9, 2020 9:30 AM (15 minutes)

The Tunka Radio Extension (Tunka-Rex) is a cosmic-ray detector operating since 2012. The detection principle of Tunka-Rex is based on the radio technique, which impacts data acquisition and storage. We present the Tunka-Rex Virtual Observatory (TRVO), a framework for open access to the Tunka-Rex data, which currently is prepared for the first release.

Author: KOSTUNIN, Dmitriy (DESY)

Presenter: KOSTUNIN, Dmitriy (DESY)

Session Classification: Session 3

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

Fast Simulation of Electromagnetic Calorimeter using Deep Learning

Tuesday, June 9, 2020 10:15 AM (15 minutes)

The simulation of particle showers in electromagnetic calorimeters with high precision is a computationally expensive and time consuming process. Fast simulation of particle showers using generative models have been suggested to significantly save computational resources. The objective of studies is to perform a fast simulation of particle showers in the Belle II calorimeters using deep learning techniques. In my study, particle showers simulated using the Geant4 simulation toolkit are used to train a generative deep learning model. Once the model is trained, the generative part of the model is used to generate particle shower simulations providing noise vectors as input. The generated particle showers are cross-checked with the Geant4 showers using various observables.

Authors: Mrs IRAKKATHIL JABBAR, Jubna; Prof. BERNLOCHNER, Florian; Dr GOLDENZWEIG, Pablo

Presenter: Mrs IRAKKATHIL JABBAR, Jubna

Session Classification: Session 3

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

TAIGA: status, results and perspectives

Tuesday, June 9, 2020 9:00 AM (30 minutes)

We present the current status of very high-energy cosmic and gamma ray installation TAIGA at the Tunka Astrophysical Center situated at about 50 km from Lake Baikal. The deployment first stage insllation consits of 120 optical staion of HiSCORE array and 3 IACT will be finished in autumn of 2020. The last results of local sources observation during 2019 -2020a and plan for the further modification of the installation will be presented.

Presenter: Prof. KUZMICHEV, Leonid

Session Classification: Session 3

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

Reconstruction radio signals from air-showers with autoencoder

Tuesday, June 9, 2020 9:45 AM (15 minutes)

One of the main challenges related to the measurements of air-shower radio emission is the high background. Plenty of natural and anthropogenic RFI as well as stationary background distort air-shower pulse. Standard methods based on signal-to-noise ratio lead to increasing the threshold of air-shower detection. For extending the energy range towards lower energies we perform data denoising using autoencoder, which is a deep neural network trained in order to decrease the amplitude of the noise, meanwhile keeping useful signal in the trace. We describe our method, present the results of reconstruction and discuss further steps in this direction.

Author: BEZYAZEEKOV, Pavel

Presenter: BEZYAZEEKOV, Pavel

Session Classification: Session 3

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

KCDC: status and future perspectives

Monday, June 8, 2020 10:45 AM (15 minutes)

A few days ago we released the new KCDC version PENTARUS, which contains another DataShop for the first time. A brief outline of the new features will be given here, as well as possible future perspectives.

Author: WOCHELE, Juergen (KIT-IKP)

Co-authors: HAUNGS, Andreas (KIT); KANG, Donghwa (KIT); Dr WOCHELE, Doris (KIT-IKP); SCHOO, Sven (IKP)

Presenter: WOCHELE, Juergen (KIT-IKP)

Session Classification: Session 2

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

Outreach activities in astroparticle physics at KIT

Wednesday, June 10, 2020 9:30 AM (15 minutes)

With various outreach activities, KIT - in particular KCETA - aims to make astroparticle physics more accessible for everyone, not only high-school students and their teachers but also a broader public.
A wide range of activities, from public lectures, internships and practical activities for students to art meets science projects are therefore part of the repertory.
Some of these activities are also embedded in the programme of national (Netzwerk Teilchenwelt) and global institutions (International Particle Physics Outreach Group / Global Cosmic Group). There are also close links at European level with the APPEC Functional Centre being in Karlsruhe.
In this talk we present an overview of our outreach programme.

Author: LINK, Katrin

Presenter: LINK, Katrin

Session Classification: Session 5

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

Publishing multi-purpose data sets from KM3NeT

Tuesday, June 9, 2020 11:30 AM (15 minutes)

The KM3NeT neutrino detector, consisting of several building blocks for the water Cherenkov detection of relativistic charged particles, is currently under construction at various deep-sea locations in the Mediterranean Sea. As inter-domain experiment between neutrino, astroparticle and astrophysics, data processing and data publication from KM3NeT draws on computing paradigms and standardization from both fields. In this contribution, key considerations for the provision of these multi-purpose open data sets, interface options and interoperability requirements are presented.

Author: SCHNABEL, Jutta (Erlangen Centre for Astroparticle Physics (ECAP/FAU))

Presenter: SCHNABEL, Jutta (Erlangen Centre for Astroparticle Physics (ECAP/FAU))

Session Classification: Session 4

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

A model of data processing pipeline for space weather analysis and forecast

Monday, June 8, 2020 12:00 PM (15 minutes)

Space weather is a branch of space physics that studies various factors in the near-Earth space such as solar wind, magnetosphere disturbance, solar proton events, and others, which make a massive impact on the Earth. In practice, data measured by different satellite instruments need to be gathered and appropriately transformed before use in space weather analysis and forecast. The data processing pipeline involves a large number of various programs. It also requires in-depth technical knowledge of both satellite instruments and programming tools so that data will be processed correctly. Building such a data pipeline is time-consuming and error-prone. The correctness of the output data produced by the processing pipeline is one of the critical factors that define the success of an analysis or a forecast model. This work proposes a model that describes how the data processing pipeline might be organized and how to build a distributed data processing system based on the proposed model.

Author: NGUYEN, Minh-Duc

Presenter: NGUYEN, Minh-Duc

Session Classification: Session 2

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

Dynamic Computing Resource Extension using COBalD/TARDIS

Tuesday, June 9, 2020 11:15 AM (15 minutes)

Cloud providers, HPC clusters, and free institute resources can dynamically increase computing power. In order to make theses so-called opportunistic resources transparently available, the services COBalD and TARDIS are developed in collaboration of the Institute of Experimental Particle Physics (ETP) and the Steinbuch Centre for Computing (SCC) at KIT.
The opportunistic resources are integrated into an overlay batch system (OBS), which acts as a single-point-of-entry for the users. Depending on the decisions of the OBS, COBalD/TARDIS adjust the resource allocation at the various resource providers. To supply the necessary software environment for the jobs, required by the scientific communities, virtualization and containerization technologies are used on the heterogeneous resources.

In this contribution, we introduce the general concept of COBalD/TARDIS, present current setups in the HEP context, and show an example application of outside HEP.

Presenter: VON CUBE, Ralf Florian

Session Classification: Session 4

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

Welcome words

Monday, June 8, 2020 9:00 AM (15 minutes)

Presenter: HAUNGS, Andreas (KIT)

Session Classification: Session 1

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

MD Catalog - integration with external data storages. Current state.

Monday, June 8, 2020 11:45 AM (15 minutes)

The current state of the metadata catalog for astrophysics of cosmic rays is considered. The method of integration with external data storages is proposed.

Author: MIKHAILOV, Andrey (IDCST)

Co-authors: KRYUKOV, Alexander (MSU); SHIGAROV, Alexei; NGUYEN, Minh-Duc; TOKAREVA, Victoria (KIT)

Presenter: MIKHAILOV, Andrey (IDCST)

Session Classification: Session 2