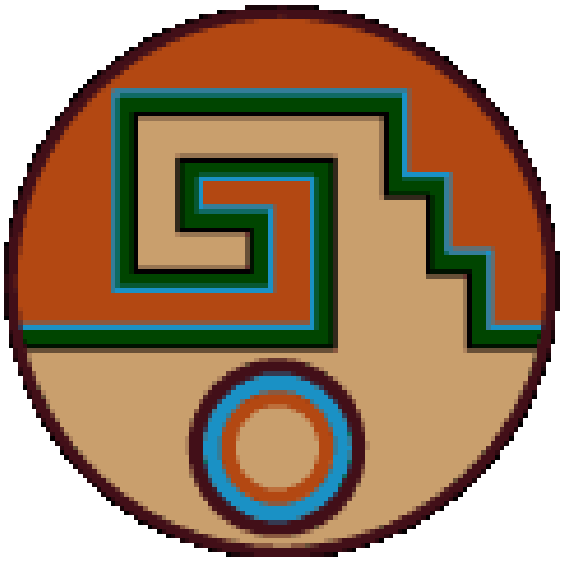
# ACAT 2024



# Report of Contributions

https://indico.cern.ch/e/acat2024

Contribution ID: **1** Type: **Oral**

# Deep Learning-Based C14 Pile-Up Identification in the JUNO Experiment

*Thursday, March 14, 2024 2:50 PM (20 minutes)*

Measuring neutrino mass ordering (NMO) poses a fundamental challenge in neutrino physics. To address this, the Jiangmen Underground Neutrino Observatory (JUNO) experiment is scheduled to commence data collection in late 2024, with the ambitious goal of determining the NMO at a 3-sigma confidence level within a span of 6 years. A key factor in achieving this is ensuring a high-quality energy resolution of positrons. However, the presence of residual C14 isotopes in the liquid scintillator introduces pile-up effects that can impact the positron energy resolution. Mitigating these pile-up effects requires the identification of pile-up events, which presents a significant challenge. The signal from C14 is considerably smaller compared to the positron signal, making its identification difficult. Additionally, the close event time and vertex between a positron and a C14 further compound the identification challenge.

This contribution focuses on the application of deep learning models for the identification of C14 pile-up events. It encompasses a range of models, including convolutional-based models and advanced transformer models. Through performance evaluation, the study showcases the robust capabilities of deep learning models in accurately and effectively identifying pile-up events.

## Significance

Considering that pile-up event identification is a common issue across various experiments, the methods proposed in this contribution hold the potential for wider adoption and utilization.

## References

## Experiment context, if any

**Author:** FANG, Wenxing

**Co-authors:** Dr LI, Weidong (IHEP, Beijing); LUO, Wuming (Institute of High Energy Physics, Chinese Academy of Science)

**Presenter:** FANG, Wenxing

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **2** Type: **Oral**

# JUNO raw data management system

*Monday, March 11, 2024 2:30 PM (20 minutes)*

The Jiangmen Underground Neutrino Observatory (JUNO) is a multipurpose neutrino experiment. JUNO will start to take data in the fall of 2024 with 2PB data each year. It is important that raw data is copied to permanent storage and distributed to multiple data center storage system in time for backup. To make available for re-reconstruction among these data centers, raw data also need to be registered into metadata and replicas catalogs of the JUNO distributed computing system. The raw data management system will take care of distributing raw data and running data processing activities in JUNO data centers. An automatic system based on JUNO distributed computing system has been designed and developed to do registering, replicating, archiving and data reconstruction in a data-driven chain. The monitoring dashboard has been designed and developed to ensure the quality of data transfer and processing. The prototype of the system has been tested with commissioning data in 2023 and the system will continue to join JUNO data challenge in early 2024.

## Significance

## References

## Experiment context, if any

JUNO

**Author:** ZHANG, Xiaomei (Chinese Academy of Sciences (CN))

**Presenter:** ZHANG, Xiaomei (Chinese Academy of Sciences (CN))

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **3** Type: **Oral**

# Leveraging Language Models for Particle Reconstruction

*Tuesday, March 12, 2024 11:50 AM (20 minutes)*

Particle detectors play a pivotal role in the field of high-energy physics. Traditionally, detectors are characterized by their responses to various particle types, gauged through metrics such as energy or momentum resolutions. While these characteristics are instrumental in determining particle properties, they fall short of addressing the initial challenge of reconstructing particles.

We propose an innovative approach to enhance particle reconstruction by harnessing the power of language models. The idea is to tokenize the detector readout signals and train a language model to embed these detector readouts to a latent space as new detector data representations that capture the essence of particle interactions. The talk will show our preliminary results, providing a first proof-of-concept demonstration of solving the challenging particle tracking problem with language models. By leveraging language models, we aim to revolutionize particle reconstruction methodologies, opening new avenues for understanding particle detectors.

## Significance

## References

## Experiment context, if any

**Author:** JU, Xiangyang (Lawrence Berkeley National Lab. (US))

**Presenter:** JU, Xiangyang (Lawrence Berkeley National Lab. (US))

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **4** Type: **Oral**

# Towards a framework for GPU event generation

*Wednesday, March 13, 2024 2:30 PM (20 minutes)*

We demonstrate some advantages of a top-bottom approach in the development of hardware-accelerated code by presenting the PDFFlow-VegasFlow-MadFlow software suite. We start with an autogenerated hardware-agnostic Monte Carlo generator, which is parallelized in the event axis. This allow us to take advantage of the parallelizable nature of Monte Carlo integrals even if we do not have control of the hardware in which the computation will run (i.e., an external cluster). The generic nature of such an implementation can introduce spurious bottlenecks or overheads. Fortunately, said bottlenecks are usually restricted to a subset of operations and not to the whole vectorized program. By identifying the more critical parts of the calculation one can get very efficient code and at the same time minimize the amount of hardware-specific code that needs to be written. We show benchmarks demonstrating how simply reducing the memory footprint of the calculation can increase the performance of a $2 \longrightarrow 4$ process. Finally, we present summary results about the performance achieved so far for PDF query on GPU and Monte Carlo integration.

## Significance

In view of recent interested from the theory community in hardware accelerators, our goal is to present a development paradigm which could accelerate in the introduction of GPUs for MC simulation.

## References

https://arxiv.org/abs/2211.14056
https://arxiv.org/abs/2106.10279
https://arxiv.org/abs/2012.08221
https://arxiv.org/abs/2010.09341

## Experiment context, if any

**Authors:** Dr CRUZ MARTINEZ, Juan M. (CERN); CARRAZZA, Stefano (CERN)

**Presenter:** CARRAZZA, Stefano (CERN)

**Session Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

**Track Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

Contribution ID: **5** Type: **Oral**

# AI-based Data Popularity, Placement Optimization for a Novel Multi-tiered Storage System at BNL/SDCC Facility

*Thursday, March 14, 2024 5:10 PM (20 minutes)*

Scientific experiments and computations, particularly in Nuclear Physics (NP) and High Energy Physics (HEP) programs, are generating and accumulating data at an unprecedented rate. Big data presents opportunities for groundbreaking scientific discoveries. However, managing this vast amount of data cost-effectively while facilitating efficient data analysis within a large-scale, multi-tiered storage architecture poses a significant challenge for the Scientific Data and Computing Center (SDCC).

The storage team is currently addressing optimization challenges related to data classification, placement, and migration in the existing multi-tier storage system. While users and administrators manually optimize storage by migrating data based on simple rules derived from human knowledge, decisions, and basic usage statistics, evaluating the placement of data in different storage classes with I/O-intensive workloads remains a complex task.

To overcome the aforementioned challenge and address existing limitations, we have developed a precise data popularity prediction model utilizing state-of-the-art AI/ML techniques. Additionally, we have designed a data placement policy engine based on data popularity, allowing us to migrate infrequently accessed data to more economical storage media, such as tape drives, while storing frequently accessed data on faster yet costlier storage media like HDD or SSD. This strategy optimally places data into the proper storage classes, maximizing storage capacity while minimizing data access latency for end users. This paper delves into the analysis of the data, demonstration patterns, tag files. Specifically, we detail the design and development of an accurate AI/ML prediction model to forecast future data popularity, based on an analysis of access patterns, facilitating optimal data movement and placement. Additionally, we provide insights into the implementation of a policy engine and data placement tool to execute automated migration actions. Finally, the evaluation of different strategies is illustrated, including those involving AI/ML models,etc.

## Significance

## References

## Experiment context, if any

**Author:** HUANG, Qiulan (Brookhaven National Laboratory (US))

**Co-authors:** Mr LEONARDI, James (Brookhaven National Laboratory); Dr GARONNE, Vincent (Brookhaven National Laboratory); Dr YOO, Shinjae (Brookhaven National Laboratory)

**Presenter:** HUANG, Qiulan (Brookhaven National Laboratory (US))

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **6** Type: **Oral**

# Quantum simulation with just-in-time compilation

*Monday, March 11, 2024 2:30 PM (20 minutes)*

Quantum technologies are moving towards the development of novel hardware devices based on quantum bits (qubits). In parallel to the development of quantum devices, efficient simulation tools are needed in order to design and benchmark quantum algorithms and applications before deployment on quantum hardware.
In this context, we present a first attempt to perform circuit-based quantum simulation using the just-in-time (JIT) compilation technique on multiple hardware architectures and configurations based on single-node central processing units (CPUs) and graphics processing units (GPUs).
One of the major challenges in scientific code development is to balance the level of complexity between algorithms and programming techniques without losing performance or degrading code readability. In this context, we have developed `qibojit`: a new module for the `Qibo` quantum computing framework, which uses a just-in-time compilation approach through Python.
We also present recent results within the `Qibo` framework concerning different simulation methods such as tensor networks and multi-node deployment.
We perform systematic performance benchmarks between `Qibo` and a subset of relevant publicly available libraries for quantum computing.

## Significance

This talk will present the latest enhancements in the Qibo framework concerning full state vector simulation and novel results regarding tensor networks and multi-node implementations.

## References

https://iopscience.iop.org/article/10.1088/2058-9565/ac39f5
https://quantum-journal.org/papers/q-2022-09-22-814/

## Experiment context, if any

**Authors:** PASQUALE, Andrea (University of Milan); ROBBIATI, Matteo (Università degli Studi e INFN Milano (IT)); PEDICILLO, edoardo (Università degli Studi di Milano)

**Co-author:** CARRAZZA, Stefano (CERN)

**Presenter:** PASQUALE, Andrea (University of Milan)

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 3: Computations in Theoretical Physics: Techniques and Meth-

ods

Contribution ID: **7** Type: **Oral**

# Real-time error mitigation for variational optimization on quantum hardware

*Monday, March 11, 2024 3:30 PM (20 minutes)*

The development of quantum computers as tools for computation and data analysis is continually increasing, even in the field of machine learning, where numerous routines and algorithms have been defined, leveraging the high expressiveness of quantum systems to process information. In this context, one of the most stringent limitations is represented by noise. In fact, the devices currently available are not clean enough to implement complex routines. One of the strategies that can be adopted to face this problem is called quantum error mitigation: the noise configuration of a device is learned as a noise map, which is then used to mitigate the results. In this talk, we present a real-time error mitigation algorithm applicable in the context of optimizing a quantum variational model. In particular, we use the Importance Clifford Sampling method to mitigate both predictions and gradients in a gradient-based optimization procedure. This process is carried out by monitoring the device's noise, so that the noise map is re-learned when the previous one becomes unreliable.
The routine we describe can easily be extended to any training context and, being problem-agnostic, can be applied to any type of problem addressable with optimization techniques. We present the algorithm and then show promising results obtained by training noisy quantum circuits up to eight qubits.

## Significance

We put forward the inclusion of error mitigation routines in the context of training quantum variational models. We have concretized the theoretical findings from the Los Alamos National Laboratory (https://arxiv.org/pdf/2109.01051.pdf) implementing an efficient and computationally lightweight algorithm.

## References

Pre-print (under review): https://arxiv.org/abs/2311.05680
Code: https://github.com/qiboteam/rtqem

## Experiment context, if any

**Author:** ROBBIATI, Matteo (Università degli Studi e INFN Milano (IT))

**Co-authors:** Mr SOPENA, Alejandro (Instituto de Fisica Téorica, UAM-CSIC, Universidad Autònoma de Madrid, Cantoblanco, Madrid, Spain); Mr PAPALUCA, Andrea (School of Computing, The Australian National University, Canberra, ACT, Australia); CARRAZZA, Stefano (CERN)

**Presenter:** ROBBIATI, Matteo (Università degli Studi e INFN Milano (IT))

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **8** Type: **Oral**

# Total 10-th order QED electon anomalous magnetic moment calculation

*Thursday, March 14, 2024 3:10 PM (20 minutes)*

Total 5-loop quantum electrodynamics calculation results for the electron anomalous magnetic moment will be presented. These results provide the first check of the previously known value obtained by T. Aoyama, M. Hayakawa, T. Kinoshita, M. Nio. A comparison will be provided. The results for the Feynman diagrams without lepton loops were presented by the author in 2018-2019. The remaining part of the diagrams will be presented here.

The difficulty is that known universal methods require enormous amount of computer resources to obtain the value. Author's method of reduction to finite integrals will be briefly explained as well as a specially developed Monte Carlo integration method.

The results are split into 95 gauge-invariant classes. Such a detailzation is provided for the first time and is useful for independent checking and theoretical investigations.

## Significance

1. After emerging a new experimental value in 2022 the tension between the theory and experiment for the electron g-2 became over 3.5 sigma (depending of the fine structure constant used). The 10-th order QED coefficient had not been double-checked yet; the coefficient is sensitive in experiments and will be more sensitive in the future.
2. The high-order calculation methods in quantum field theory are important themselves. Author's method does not use dimensional regularization or other limit-like regularizations and is based on a deep understanding of the structure of divergences in Feynman diagrams; this allows us to spare computer resources significantly.

## References

https://indico.cern.ch/event/1164804/contributions/5384597/
https://indico.cern.ch/event/855454/contributions/4606407/
https://arxiv.org/abs/2308.11560

## Experiment context, if any

**Author:** VOLKOV, Sergey

**Presenter:** VOLKOV, Sergey

**Session Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

**Track Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

Contribution ID: **9** Type: **Poster**

# Introduction of dynamic job matching optimization for Grid middleware using Site Sonar infrastructure monitoring

*Monday, March 11, 2024 4:15 PM (30 minutes)*

In the realm of Grid middleware, efficient job matching is paramount, ensuring that tasks are seamlessly assigned to the most compatible worker nodes. This process hinges on meticulously evaluating a worker node's suitability for the given task, necessitating a thorough assessment of its infrastructure characteristics. However, adjusting job matching parameters poses a significant challenge due to the involvement of both central and site services within the Grid middleware. This necessitates deploying a new middleware version across the entire Grid, introducing potential bugs and raising the risk of a single point of failure.

Furthermore, the inherent limitations in the number of available job matching parameters, stemming from insufficient infrastructure monitoring in pilot jobs, further complicate the task for Grid middleware developers.

This paper introduces an entirely new approach for dynamically adding and modifying job matching parameters in Grid middleware, leveraging the Site Sonar Grid Infrastructure monitoring framework. This solution empowers Grid administrators to seamlessly add or modify job matching parameters without altering the core middleware code. This flexibility enables dynamic job matching based on diverse infrastructure properties of worker nodes. By decoupling job matching parameters from the Grid middleware, the proposed approach enhances flexibility, mitigates complexities, and reduces risks associated with introducing and changing job matching parameters.

This transformative approach bolsters the adaptability of Grid middleware for heterogeneous systems, fostering optimized resource allocation.

## Significance

Currently, the job matching in Grid middleware is done by using a set of hardcoded parameters in all the popular Grid middleware systems. This causes a major problem when introducing or changing and existing matching parameters because it requires updating both central and site services to accommodate this change. Hence the Grid administrators limit the number of matching parameters to a limited number and avoid changing them often.

Another reason for this is that a pilot job submitted to a site does not have a lot of capabilities in terms of collecting infrastructure information and hence most of the information to be collected is hardcoded. Therefore these parameters cannot be changed dynamically when necessary.

But, if we can make this approach more flexible and allow the Grid administrators to dynamically define job matching parameters, we can have an improved job matching process that will lead to a more efficient use of available worker nodes in a Grid, that could also help in reducing job failure rates.

We have previously introduced Site Sonar - a new Grid infrastructure monitoring system that can collect a lot of information from a Grid worker node. In this project we have integrated Site Sonar with the Grid middleware to facilitate the use of any infrastructure information collected through Site Sonar in the job matching process. Further, we have changed how central services handle the job matching parameters to allow dynamically changing parameters. This approach can be

used by any Grid middleware to improve their job matching process and allow a more optimized resource allocation.

## References

https://indico.jlab.org/event/459/contributions/11495/ (Previous presentation on introduction of Site Sonar)

## Experiment context, if any

The project was conducted on ALICE Computing Grid in ALICE experiment at CERN (https://alice-collaboration.web.cern.ch)

**Author:** WIJETHUNGA, Kalana (University of Moratuwa (LK))

**Co-authors:** GRIGORAS, Costin (CERN); Prof. PERERA, Indika (University of Moratuwa(LK)); BETEV, Latchezar (CERN)

**Presenter:** WIJETHUNGA, Kalana (University of Moratuwa (LK))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **10** Type: **Poster**

# Visualizing BESIII Events with Unity

*Monday, March 11, 2024 4:15 PM (30 minutes)*

In high-energy physics experiments, the software's visualization capabilities are crucial, aiding in detector design, assisting with offline data processing, offering potential for improving physics analysis, among other benefits. Detailed detector geometries and architectures, formatted in GDML or ROOT, are integrated into platforms like Unity for three-dimensional modeling. In this study, based on the BESIII spectrometer, Unity is utilized to display BESIII events in three-dimensional and even animated formats. This method of event display vividly illustrates the collision and tracks of particles within the detector. Utilizing this event display system instances through software facilitates improved analysis, fosters interdisciplinary applications, and expands into the realm of education.

## Significance

## References

## Experiment context, if any

**Authors:** LI, Jingshu (Sun Yat-Sen University (CN)); YOU, Zhengyun (Sun Yat-Sen University (CN))

**Presenter:** LI, Jingshu (Sun Yat-Sen University (CN))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **11** Type: **Oral**

# Persistifying the complex Event Data Model of the ATLAS Experiment in RNTuple

*Monday, March 11, 2024 4:50 PM (20 minutes)*

The ATLAS experiment at CERN's Large Hadron Collider has been using ROOT TTree for over two decades to store all of its processed data. The ROOT team has developed a new I/O subsystem, called RNTuple, that will replace TTree in the near future. RNTuple is designed to adopt various technological advancements that happened in the last decade and be more performant from both the computational and storage perspectives. On the other hand, RNTuple has limited/streamlined data model support compared to TTree.

The ATLAS Event Data Model (EDM) must support functionality arising from the vast complexity of the underlying detector and the constraints of the computing model. It takes advantage of C++ (object oriented) language features that allow efficient processing of highly complex algorithms that produce physics objects from various different sub-detectors. To encapsulate this complexity needed for transient processing, ATLAS had introduced a separation between the transient and the persistent (T/P) representations of the EDM. This approach simplified the adoption of TTree as the main event data format at the time. It also allows us to embrace different technologies and storage backends more easily while keeping the reconstruction and simulation software stack as complex as it needs to be.

In this presentation, we will discuss all the foundational work that allowed ATLAS to persistify all its processed event data, including complex simulation and reconstruction data, in the RNTuple format. We will discuss the key elements of ATLAS'core EDM and I/O software and how encapsulation via T/P separation can guide other (future) experiments in designing their own models and future-proofing their I/O and storage infrastructure.

## Significance

This will be the first public presentation on the foundational work that allowed ATLAS to persistify all its upstream event data, including complex simulation and reconstruction data, in the RNTuple format.

## References

## Experiment context, if any

The ATLAS Experiment

**Authors:** METE, Alaettin Serhan (Argonne National Laboratory (US)); NOWAK, Marcin (Brookhaven National Laboratory (US)); VAN GEMMEREN, Peter (Argonne National Laboratory (US))

**Presenter:** METE, Alaettin Serhan (Argonne National Laboratory (US))

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **12** Type: **Poster**

# Web based HXMT data analysis platform

*Monday, March 11, 2024 4:15 PM (30 minutes)*

The HXMT satellite is China's first space astronomy satellite. It is a space-based X-ray telescope capable of broadband and large-field X-ray sky surveys, as well as the study of high-energy celestial objects such as black holes and neutron stars, focusing on short-term temporal variations and broadband energy spectra. It also serves as a highly sensitive all-sky monitor for gamma-ray bursts. The HXMT User Data Analysis Software (HXMTDAS) is primarily designed to analyze pointed observational data from the HXMT satellite, including on-axis and off-axis observations, to produce energy spectra, light curves, and energy response files.
This report presents an interactive data analysis platform based on web technology and HXMTDAS. Using containers, virtualisation technology and JupyterLab, this platform allows users to perform interactive data analysis via a web browser. The platform is out-of-the-box and operating system independent, eliminating the need for complex software installations and environment configuration steps. It is particularly user-friendly and can be used for educational purposes or for training new users.

## Significance

## References

## Experiment context, if any

**Author:** HU, Yu

**Co-authors:** QI, Fazhi (IHEP, CAS); ZHANG, Hongmei (IHEP, CAS); LIU, Jianli (IHEP, CAS); HU, Qingbao (IHEP, CAS); WANG, Shuang (IHEP, CAS)

**Presenter:** WANG, lei (Institute of High Energy Physics)

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **13** Type: **Poster**

# Supervised job preemption methodology for controlled memory consumption of jobs running in the ALICE Grid

*Wednesday, March 13, 2024 4:15 PM (30 minutes)*

The ALICE experiment's Grid resources vary significantly in terms of memory capacity, CPU cores, and resource management. Memory allocation for scheduled jobs depends on the hardware constraints of the executing machines, system configurations, and batch queuing policies. The O2 software framework introduces multi-core tasks where deployed processes share resources. To accommodate these new use cases, most Grid sites provide ALICE with multi-core slots of a customizable amount of cores. The Grid middleware manages the resources within a slot, sub-partitioning and distributing them among allocated jobs. This allows for parallel execution of jobs with different natures and usage patterns within the same resource-sharing slot. From the scheduling system's perspective, this job set is treated as a single unit for resource usage accounting. Overconsumption by any job can lead to the entire slot being killed, terminating all co-executing ALICE jobs. To prevent this and promote job completion with reasonable resource usage, the Grid middleware should implement targeted preemption of top-consuming jobs when overall consumption approaches the system's killing threshold.

This paper analyzes site resource limiting procedures, including killing policies and memory thresholds, and the design of the ALICE Grid middleware framework's methodology for targeted preemption. Preemption decisions are made in real time, considering various factors of running payloads, weighted according to experiment priorities, to maximize efficiency and successful task completion.

## Significance

The contribution presents an analysis of Grid site resource allocation limiting procedures, including killing policies and memory thresholds, and the design of the ALICE Grid middleware framework's methodology for targeted preemption of over-consuming jobs. Preemption decisions are made in real time, considering various factors of running payloads, weighted according to experiment priorities, to maximize efficiency and successful task completion.

## References

Paper related to multi-core job support in the ALICE Grid - https://dx.doi.org/10.1088/1742-6596/2438/1/012009

## Experiment context, if any

LHC ALICE Experiment

**Authors:** WIJETHUNGA, Kalana (University of Moratuwa (LK)); BERTRAN FERRER, Marta (CERN)

**Presenter:** WIJETHUNGA, Kalana (University of Moratuwa (LK))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **14** Type: **Poster**

# Interface to Unity for High Energy Physics detector visualization

*Monday, March 11, 2024 4:15 PM (30 minutes)*

The visualization process of detector is one of the important problems in high energy physics (HEP) software. At present, the description of detectors in HEP is complicated. Industry professional visualization platforms such as Unity, have the most advanced visualization capabilities and technologies, which can help us to achieve the visualization of detectors. The work is to find an automated interface to efficiently convert all detector descriptions from HEP experiments in formats such as GDML, DD4hep, root, Geant4, directly to 3D models in Unity. Such an interface has been successfully applied to several detectors, converted them into 3D models and imported into unity. This work has great potential to play an auxiliary role in detector design, HEP offline software development, physical analysis and other parts HEP experiments, and it also provides a good foundation for future research such as event display.

## Significance

## References

## Experiment context, if any

**Author:** SONG, Tianzi (Sun Yat-Sen University (CN))

**Presenter:** SONG, Tianzi (Sun Yat-Sen University (CN))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **15** Type: **Oral**

# Efficient precision simulation of processes with many-jet final states at the LHC

*Wednesday, March 13, 2024 3:10 PM (20 minutes)*

The success of the LHC physics programme relies heavily on high-precision calculations. However, the increased computational complexity for high-multiplicity final states has been a growing cause for concern, with the potential to evolve into a debilitating bottleneck in the foreseeable future. We present a flexible and efficient approach for the simulation of collider events with multi-jet final states for both leading and next-to-leading order QCD calculations. The technique is based on an improved parton-level event file format with efficient scalable data handling. We validate the proposed framework using a range of processes, including Higgs boson plus multi-jet production with up to seven jets, and demonstrate its use in both the Sherpa and Pythia event generators, paving the way towards economically and ecologically sustainable event generation in the high-luminosity era.

## Significance

We propose a new event file format akin to the Les Houches Event format but based on the HDF5 library, lending itself very well to HPC applications, including GPU-accelerated Monte Carlo event generators.

## References

https://arxiv.org/abs/2309.13154

## Experiment context, if any

The work is relevant for ATLAS and CMS but was done outside of the experiments and hence doesn't require involvement the respective publication boards.

**Authors:** GUTSCHOW, Christian (UCL (UK)); BOTHMANN, Enrico (University of Göttingen); ISAACSON, Joshua; KNOBBE, Max (University of Göttingen); HOVLAND, Paul (Argonne National Laboratory); LATHAM, Robert (Argonne National Laboratory); HOECHE, Stefan (Fermilab); CHILDERS, Taylor

**Presenter:** GUTSCHOW, Christian (UCL (UK))

**Session Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

**Track Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

Contribution ID: **16** Type: **Oral**

# Consistent multi-differential histogramming and summary statistics with YODA2

*Thursday, March 14, 2024 3:30 PM (20 minutes)*

In the contemporary landscape of advanced statistical analysis toolkits, ranging from Bayesian inference to machine learning, the seemingly straightforward concept of a histogram often goes unnoticed. However, the power and compactness of partially aggregated, multi-dimensional summary statistics with a fundamental connection to differential and integral calculus make them formidable statistical objects. Expressing these concepts robustly and efficiently in high-dimensional parameter spaces is a non-trivial challenge, especially when the resulting library is meant to remain usable by scientists rather than software engineers.

A decade after its initial release, the YODA statistical library has been redesigned from the ground, aiming to generalise its principles while addressing real-world usage requirements in the era of expanding computational power and vast datasets. We will summarise the core principles required for consistent generalised histogramming and outline some of the C++ metaprogramming techniques adopted to handle dimensionality relationships in the revamped YODA histogramming library. Used both in Rivet and Contur, YODA is a key component of physics data–model comparison and statistical interpretation in collider physics.

## Significance

The YODA library is a key component in the Rivet and Contur packages. 10 years after its initial release, YODA has been redesigned using modern C++ techniques to provide generalised histogramming in arbitrary dimensions and addressing various other shortcomings of the initial release series.

## Experiment context, if any

## References

https://arxiv.org/abs/2312.15070

**Authors:** BUCKLEY, Andy (University of Glasgow (GB)); GUTSCHOW, Christian (UCL (UK)); YELLEN, Jamie (University of Glasgow); YEH, Yoran (University College London (UK))

**Presenter:** GUTSCHOW, Christian (UCL (UK))

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **18** Type: **Oral**

# RNTupleInspector: A storage information utility for RNTuple

*Monday, March 11, 2024 5:10 PM (20 minutes)*

Inspired by over 25 years of experience with the ROOT TTree I/O subsystem and motivated by modern hard- and software developments as well as an expected tenfold data volume increase with the HL-LHC, RNTuple is currently being developed as ROOT's new I/O subsystem. Its first production release is foreseen for late 2024, and various experiments have begun working on the integration of RNTuple with their existing software frameworks and data models. To aid developers in this integration process, and to help them further understand and monitor the storage patterns of their data with RNTuple, we have developed the RNTupleInspector utility interface, which will be available with every ROOT installation that includes RNTuple. The RNTupleInspector provides storage information for full RNTuples as well as specific fields or columns, and is designed in such a way that it can be used as part of a larger monitoring tool as well as in an exploratory manner, for example through the ROOT interpreter. In this contribution, we will discuss the motivation and design considerations behind the RNTupleInspector and demonstrate its use through example use cases.

## Significance

As RNTuple becomes more mature in preparation for its first production release, it becomes essential for software developers of experiments as well as RNTuple to understand how different data models and I/O parameters affect the storage efficiency of RNTuple. The RNTupleInspector utility is meant to provide insights into these aspects and can aid in designing an optimal RNTuple I/O parameter configuration and monitoring the storage behaviour of different data sets.

## References

## Experiment context, if any

**Author:** DE GEUS, Florine (CERN)

**Co-authors:** BLOMER, Jakob (CERN); CANAL, Philippe (Fermi National Accelerator Lab. (US)); Dr PADULANO, Vincenzo Eduardo (CERN)

**Presenter:** DE GEUS, Florine (CERN)

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **19** Type: **Oral**

# Seamless transition from TTree to RNTuple analysis with RDataFrame

*Monday, March 11, 2024 5:30 PM (20 minutes)*

As the High-Luminosity LHC era is approaching, the work on the next-generation ROOT I/O subsystem, embodied by the RNTuple, is advancing fast with demonstrated implementations of the LHC experiments' data models and clear performance improvements over the TTree. Part of the RNTuple development is to guarantee no change in the RDataFrame analysis flow despite the change in the underlying data format.
In this talk, we present integration of RNTuple and RDataFrame. The engine can process RNTuple datasets on a local machine, sequentially with one core or using implicit multithreading with multiple cores. Furthermore, RNTuple processing is also introduced in the distributed RDataFrame layer and benchmarked using SWAN, a web-based platform, to transparently offload analysis tasks to the CERN HTCondor pools. The new workflow is demonstrated using existing RDataFrame analyses on one or multiple nodes with no change in the API. One notable example is the t-tbar Analysis Grand Challenge benchmark, which is also used as a blueprint to showcase differences in performance of (distributed) execution with the two data formats.

## Significance

LHC experiments are already involved in the process of testing and validating the next-generation ROOT I/O. ROOT will progressively fade out support for writing new datasets with TTree, so RNTuple will have a clear impact on future HEP computing workflows at many levels, from infrastructures to final analyses. This presentation demonstrates how the ROOT efforts go in the direction of making the transition as effortless as possible for the HEP users, while aligning with the experiments' expected computing challenges.

## References

CHEP 2023 https://indico.jlab.org/event/459/contributions/11582/
ACAT 2022 https://indico.cern.ch/event/1106990/contributions/4998129/

## Experiment context, if any

**Author:** CZURYLO, Marta (CERN)

**Co-authors:** FALKO, Andrii; PIPARO, Danilo (CERN); TEJEDOR SAAVEDRA, Enric (CERN); GUIRAUD, Enrico (Princeton University, CERN); BLOMER, Jakob (CERN); CANAL, Philippe (Fermi National Accelerator Lab. (US)); Dr PADULANO, Vincenzo Eduardo (CERN)

**Presenter:** CZURYLO, Marta (CERN)

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **20** Type: **Oral**

# An empirical performance-portability evaluation for Lorentz Vectors computations via SYCL

*Wednesday, March 13, 2024 5:50 PM (20 minutes)*

In recent years, we have seen a rapid increase in the variety of computational architectures, featuring GPUs from multiple vendors, a trend that will likely continue in the future with the rise of possibly new accelerators. The High Energy Physics (HEP) community employs a wide variety of algorithms for accelerators which are mostly vendor-specific, but there is a compelling demand to expand the target capabilities of these tools via single-source cross-platform performance-portable abstraction layers, such as SYCL.
In this talk, we present GenVectorX, a SYCL-based multi-platform extension of the GenVector package of ROOT, that provides classes and functionalities to represent and manipulate particle events. This tool is intended for general usage, but it specifically targets HEP experiments data processing. Moreover, we discuss results showing that the SYCL-based implementation exhibits comparable performance and scalability as the CUDA implementation when targeting NVIDIA GPUs.

## Significance

In this presentation, we detail the migration of a large, complex, C++ code base to both SYCL and CUDA, providing guidance and insights regarding the analogies and differences between the two frameworks for other developers interested in migrating their own codes. We provide a detailed performance analysis of the migrated SYCL code on different platforms and architectures. Moreover, we compute and compare the performance portability and code divergence of GenVectorX, to explore the trade-offs between maintaining a single source code and specializing small regions of code for specific targets.

## References

## Experiment context, if any

**Author:** DESSOLE, Monica (EP SFT)

**Co-authors:** NAUMANN, Axel (CERN); CHEN, Jolly (CERN)

**Presenter:** DESSOLE, Monica (EP SFT)

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **21** Type: **Oral**

# A Deep Generative Model for Hadronization

*Wednesday, March 13, 2024 2:30 PM (20 minutes)*

Hadronization is a critical step in the simulation of high-energy particle and nuclear physics experiments. As there is no first principles understanding of this process, physically-inspired hadronization models have a large number of parameters that are fit to data. We propose an alternative approach that uses deep generative models, which are a natural replacement for classical techniques, since they are more flexible and may be able to improve the overall precision. We first demonstrate using neural networks to emulate specific hadronization when trained using the inputs and outputs of classical methods. A protocol is then developed to fit a deep generative hadronization model in a realistic setting, where we only have access to a set of hadrons in data. Finally, we build a deep generative hadronization model that includes both kinematic (continuous) and flavor (discrete) degrees of freedom. Our approach is based on Generative Adversarial Networks and we show the performance within the context of the cluster model within the Herwig event generator.

## Significance

This presentation shows results that demonstrate our proposed new methods for simulating hadronization in particle physics. It provides better flexibility and can fit to data, which can potentially replace the current standard hadronization models in the future.

## References

This presentation will be mainly based on the following papers:

1. https://arxiv.org/abs/2203.12660
2. https://arxiv.org/abs/2305.17169
3. https://arxiv.org/abs/2312.08453

## Experiment context, if any

**Authors:** CHAN, Jay (Lawrence Berkeley National Lab. (US)); JU, Xiangyang (Lawrence Berkeley National Lab. (US)); KANIA, Adam; NACHMAN, Ben (Lawrence Berkeley National Lab. (US)); SIODMOK, Andrzej Konrad (Jagiellonian University (PL))

**Presenter:** CHAN, Jay (Lawrence Berkeley National Lab. (US))

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **22** Type: **Oral**

# A Function-As-Task Workflow Management Approach with PanDA and iDDS

*Monday, March 11, 2024 2:50 PM (20 minutes)*

The growing complexity of high energy physics analysis often involves running a large number of different tools. This demands a multi-step data processing approach, with each step requiring different resources and carrying dependencies on preceding steps. It's important and useful to have a tool to automate these diverse steps efficiently.
With the Production and Distributed Analysis (PanDA) system and the intelligent Data Delivery Service (iDDS), we provide a platform for coordinating sequences of tasks with a workflow, orchestrating the seamless execution of tasks in a specified order and under predefined conditions, in order to automate the task sequence. In this presentation, we will present our efforts, beginning with an overview of the platform's architecture. We'll then describe a user-friendly interface with workflows described in python and tasks described by python functions. Next, we detail the flow to transform python functions into tasks and schedule tasks to distributed heterogeneous resources, coupled with a messaging-based asynchronous result-processing mechanism. Finally, we'll showcase a practical example illustrating how this platform effectively converts a machine learning hyperparameter optimization processing on an ATLAS ttH analysis to a distributed workflow.

## Significance

## References

## Experiment context, if any

**Authors:** WEBER, Christian (Brookhaven National Laboratory (US)); Dr KARAVAKIS, Edward (Brookhaven National Laboratory (US)); LIN, Fa-Hui (University of Texas at Arlington (US)); BARREIRO MEGINO, Fernando Harald (University of Texas at Arlington); DE, Kaushik (University of Texas at Arlington (US)); NILSSON, Paul (Brookhaven National Laboratory (US)); ZHANG, Rui (University of Wisconsin Madison (US)); MAENO, Tadashi (Brookhaven National Laboratory (US)); WENAUS, Torre (Brookhaven National Laboratory (US)); GUAN, Wen (Brookhaven National Laboratory (US)); ZHAO, Xin (Brookhaven National Laboratory (US)); YANG, Zhaoyu (Brookhaven National Laboratory (US))

**Presenter:** GUAN, Wen (Brookhaven National Laboratory (US))

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **23** Type: **Poster**

# interTwin - an Interdisciplinary Digital Twin Engine for Science

*Monday, March 11, 2024 4:15 PM (30 minutes)*

The interTwin project, funded by the European Commission, is at the forefront of leveraging 'Digital Twins' across various scientific domains, with a particular emphasis on physics and earth observation. Two of the most advanced use-cases of interTwin are event generation for particle detector simulation at CERN as well as the climate-based Environmental Modelling and Prediction Platform (EMP2) jointly developed at CERN and the Julich Supercomputing Center (JSC) using foundation models. interTwin enables those use-cases to leverage AI methodologies on cloud to high-performance computing (HPC) resources by using itwinai - the AI workflow and method lifecycle module of interTwin.

The itwinai module is developed collaboratively by CERN and JSC and is a pivotal contribution within the interTwin project. Its role is advancing interdisciplinary scientific research through the synthesis of learning and computing paradigms. This framework stands as a testament to the commitment of the interTwin project towards co-designing and implementing an interdisciplinary Digital Twin Engine. Its main functionalities and contributions are:

**Distributed Training:** itwinai offers a streamlined approach to distributing existing code across multiple GPUs and nodes, automating the training workflow. Leveraging industry-standard backends, including PyTorch Distributed Data Parallel (DDP), TensorFlow distributed strategies, and Horovod, it provides researchers with a robust foundation for efficient and scalable distributed training. The successful deployment and testing of itwinai on JSC's HDFML cluster underscore its practical applicability in real-world scenarios.

**Hyperparameter Optimization:** One of the core functionalities of itwinai is its hyperparameter optimization, which plays a crucial role in enhancing model accuracy. By intelligently exploring hyperparameter spaces, itwinai eliminates the need for manual parameter tuning. The functionality, empowered by RayTune, contributes significantly to the development of more robust and accurate scientific models.

**Model Registry:** A key aspect of itwinai is its provision of a robust model registry. This feature allows researchers to log and store models along with associated performance metrics, thereby enabling comprehensive analyses in a convenient manner. The backend, leveraging MLFlow, ensures seamless model management, enhancing collaboration and reproducibility.

In line with the theme of the 2024 ACAT workshop, "Foundation Models for Physics," interTwin and its use-cases empowered by itwinai are positioned at the convergence of computation and physics and showcase the significant potential of foundation models, supported by HPC resources. Together, they contribute to a narrative of interconnected scientific frontiers, where the integration of digital twins, AI frameworks, and foundation models broadens possibilities for exploration and discovery through itwinai's user-friendly interface and powerful functionalities.

## Significance

The frameworks developed within interTwin enable the integration of foundation models for physics and earth observation within a Digital Twin Engine and alleviate their development by a seamless use of advanced AI workflows powered by HPC resources.

## References

## Experiment context, if any

**Authors:** ZOECHBAUER, Alexander (CERN); LUISE, Ilaria (CERN); TSOLAKI, Kalliopi (CERN); Dr GIRONE, Maria (CERN); Mr BUNINO, Matteo (CERN); Dr VALLECORSA, Sofia (CERN)

**Presenter:** ZOECHBAUER, Alexander (CERN)

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **24** Type: **Oral**

# ”Accelerating Particle Physics Simulations with Machine Learning using Normalizing Flows and Flow Matching”

*Wednesday, March 13, 2024 5:50 PM (20 minutes)*

The simulation of high-energy physics collision events is a key element for data analysis at present and future particle accelerators. The comparison of simulation predictions to data allows us to look for rare deviations that can be due to new phenomena not previously observed. We show that novel machine learning algorithms, specifically Normalizing Flows and Flow Matching, can be effectively used to perform accurate simulations with several orders of magnitude of speed-up compared to traditional approaches. The classical simulation chain starts from a physics process of interest, computes energy deposits of particles and electronics response; and finally employs the same reconstruction algorithms used for data. Eventually, the data is reduced to some high-level analysis format. Instead, we propose an end-to-end approach, simulating the final data format directly from physical generator inputs, skipping any intermediate steps. We use particle jets simulation as a benchmark for comparing both *discrete* and *continuous* Normalizing Flows models. The models are validated across a variety of metrics to select the best ones. We discuss the scaling of performance with the increase in training data, as well as the generalization power of these models on physical processes different from the training one. We investigate sampling multiple times from the same inputs, a procedure we name *oversampling*, and we show that it can effectively reduce the statistical uncertainties of a sample. This class of ML algorithms is found to be highly expressive and useful for the task of HEP simulation. Their speed and accuracy, coupled with the stability of the training procedure, make them a compelling tool for the needs of current and future experiments.

## Significance

Application of novel machine learning algorithms and training routines to the end-to-end simulation problem in high energy physics. Demonstrated major improvements in simulation speed while retaining a high level of accuracy and fidelity.

## References

## Experiment context, if any

**Author:** VASELLI, Francesco (Scuola Normale Superiore & INFN Pisa (IT))

**Co-authors:** RIZZI, Andrea (Universita & INFN Pisa (IT)); CATTAFESTA, Filippo (Universita & INFN Pisa (IT)); ASENOV, Patrick (Universita & INFN Pisa (IT))

**Presenter:** VASELLI, Francesco (Scuola Normale Superiore & INFN Pisa (IT))

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **25** Type: **Oral**

# HEP Benchmark Suite: Enhancing Efficiency and Sustainability in Worldwide LHC Computing Infrastructures

*Monday, March 11, 2024 3:10 PM (20 minutes)*

As the scientific community continues to push the boundaries of computing capabilities, there is a growing responsibility to address the associated energy consumption and carbon footprint. This responsibility extends to the Worldwide LHC Computing Grid (WLCG), encompassing over 170 sites in 40 countries, supporting vital computing, disk, and tape storage for LHC experiments. Ensuring efficient operational practices across these diverse sites is crucial beyond mere performance metrics.

This paper introduces the HEP Benchmark suite, an enhanced suite designed to measure computing resource performance uniformly across all WLCG sites, using HEPScore23 as performance unit. The suite expands beyond assessing only the execution speed via HEPScore23. In fact the suite incorporates metrics such as machine load, memory usage, memory swap, and notably, power consumption. Its adaptability and user-friendly interface enable comprehensive acquisition of system-related data alongside benchmarking.

Throughout 2023, this tool underwent rigorous testing across numerous WLCG sites. The focus was on studying compute job slot performance and correlating these with fabric metrics. Initial analysis unveiled the tool's efficacy in establishing a standardized model for compute resource utilization while pinpointing anomalies, often stemming from site misconfigurations.

This paper aims to elucidate the tool's functionality and present the results obtained from extensive testing. By disseminating this information, the objective is to raise awareness within the community about this probing model, fostering broader adoption and encouraging responsible computing practices that prioritize both performance and environmental impact mitigation.

## Significance

High relevant for the validation and improvement of the compute performance of WLCG sites, via a centralized probing and analytic system.

## References

## Experiment context, if any

**Authors:** SZCZEPANEK, Natalia Diana (CERN); GIORDANO, Domenico (CERN)

**Co-authors:** ONDRIS, Ladislav (Brno University of Technology (CZ)); KETELE, Ewoud (CERN); MENENDEZ BORGE, Gonzalo (CERN); DI GIROLAMO, Alessandro (CERN); GLUSHKOV, Ivan (University of Texas at Arlington (US))

**Presenter:** SZCZEPANEK, Natalia Diana (CERN)

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **26** Type: **Oral**

# Evaluating Application Characteristics for GPU Portability Layer Selection

*Wednesday, March 13, 2024 2:30 PM (20 minutes)*

GPUs have become the dominant source of computing power for HPCs and are increasingly being used across the High Energy Physics computing landscape for a wide variety of tasks. Though NVIDIA is currently the main provider of GPUs, AMD and Intel are rapidly increasing their market share. As a result, programming using a vendor-specific language such as CUDA can significantly reduce deployment choices. There are a number of portability layers such as Kokkos, Alpaka, SYCL, OpenMP and std::par that permit execution on a broad range of GPU and CPU architectures, significantly increasing the flexibility of application programmers. However, each of these portability layers has its own characteristics, performing better at some tasks and worse at others, or placing limitations on aspects of the application. In this presentation, we report on a study of application and kernel characteristics that can influence the choice of a portability layer and show how each layer handles these characteristics. We have analyzed representative heterogeneous applications from CMS (patatrack and p2r), DUNE (Wire-Cell Toolkit), and ATLAS (FastCaloSim) to identify key application characteristics that have different behaviors for the various portability technologies. Using these results, developers can make more informed decisions on which GPU portability technology is best suited to their application.

## Significance

Flexibly porting code originally written for CPUs to diverse heterogeneous architectures is currently an unsolved problem in the HEP community. While some experiments have ported some code bases to a single or a small number of platforms as they have already purchased their selected hardware backends, there has not been a systematic study of the problem addressing all currently available heterogeneous architectures. Some experiments have selected technologies such as Alpaka or HIP, simply because it functioned for their code bases. This does not help other experiments make a portability layer selection, as their use cases are likely different. This study is cross-cutting in nature, identifying application characteristics that result in different performance for the various layers. By using this information, application developers can more easily select a portability technology without having to try each one.

## References

## Experiment context, if any

**Author:** Dr LEGGETT, Charles (Lawrence Berkeley National Lab (US))

**Co-authors:** VIREN, Brett (Brookhaven National Laboratory); MOHAMMAD ATIF, FNU (Brookhaven National Laboratory); YU, Haiwang; ESSEIVA, Julien (Lawrence Berkeley National Lab. (US)); KWOK, Ka Hei Martin (Fermi National Accelerator Lab. (US)); DEWING, Mark; KORTELAINEN, Matti (Fermi National Accelerator Lab. (US)); BHATTACHARYA, Meghna (Fermilab); LIN, Meifeng; STRELCHENKO, Oleksii (Fermi National Accelerator Lab. (US)); GUTSCHE, Oliver (Fermi National Accelerator Lab. (US)); WANG, Tianle (Brookhaven National Lab); TSULAIA, Vakho (Lawrence Berkeley National Lab. (US)); DONG, Zhihua

**Presenter:** Dr LEGGETT, Charles (Lawrence Berkeley National Lab (US))

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **27** Type: **Poster**

# Athena MPI: A Multi-Node Version of ATLAS's Athena Framework, Using Message Passing Interface

*Monday, March 11, 2024 4:15 PM (30 minutes)*

With the coming luminosity increase at the High Luminosity LHC, the ATLAS experiment will find itself facing a significant challenge in processing the hundreds of petabytes of data that will be produced by the detector.

The computing tasks faced by the LHC experiments such as ATLAS are primarily throughput limited, and our frameworks are optimized to run these on High Throughput Computing resources. However the focus of funding agencies is increasingly shifting towards High Performance Computing resources. As a result we urgently require the capability to efficiently run our computing tasks on these High Performance Computing resources.

One of the first tasks towards implementing this capability is to enable a single instance of the ATLAS software framework, Athena, to run over multiple (tens, or hundreds) nodes. Here we present a multi-node version of Athena that uses the industry standard Message Passing Interface (MPI) to assign work to the various worker nodes.

## Significance

This work presents a preliminary multi-node version of the Athena software framework used for data processing by the ATLAS experiment. This will enable the experiment to make more efficient use of the HPC resources available to it.

## References

https://indico.jlab.org/event/459/contributions/11444/

https://indico.cern.ch/event/1106990/contributions/4991224/

## Experiment context, if any

ATLAS

**Authors:** STANISLAUS, Beojan (Lawrence Berkeley National Lab. (US)); Dr LEGGETT, Charles (Lawrence Berkeley National Lab (US)); ESSEIVA, Julien (Lawrence Berkeley National Lab. (US)); CALAFIURA, Paolo (Lawrence Berkeley National Lab. (US)); TSULAIA, Vakho (Lawrence Berkeley National Lab. (US)); JU, Xiangyang (Lawrence Berkeley National Lab. (US))

**Presenter:** STANISLAUS, Beojan (Lawrence Berkeley National Lab. (US))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **28** Type: **Poster**

# FATRAS integration for ATLAS fast simulation at HL-LHC

*Monday, March 11, 2024 4:15 PM (30 minutes)*

The computing challenges in collecting, storing, reconstructing, and analyzing the colossal volume of data produced by the ATLAS experiment and producing similar numbers of simulated Monte Carlo (MC) events put formidable requirements on the computing resources of the ATLAS collaboration. ATLAS currently expends around 40% of its CPU resources on detector simulation, in which half of the events are produced with full simulation using GEANT4. Fast Chain provides a quicker alternative to the standard ATLAS MC production chain (full simulation).

The Fast ATLAS Track Simulation (FATRAS), which simulates charged particles passing through complex magnetic and calorimetric systems, has been seamlessly integrated into the ATLAS fast simulation pipeline. This integration accelerates the simulation process and maintains high precision in reproducing particle interactions within the ATLAS inner detector using a simplified detector geometry and a parameterization of particle interactions with the detector material. Recent updates to the FATRAS have focused on improving its modeling of particle interactions, extending its applicability to a broader range of physics scenarios, and enhancing its efficiency for large-scale simulations. For High Luminosity LHC (HL-LHC), the ATLAS experiment aims to use mostly fast simulation and plans to migrate the current FATRAS to a multithread-compatible version, ACTS-FATRAS. We will discuss specific features and performance benchmarks of the updated FATRAS-integrated ATLAS fast simulation, showcasing its capabilities in accurately reproducing the physics processes of interest and the impact on reducing computational resources required for large-scale simulation campaigns.

## Significance

## References

## Experiment context, if any

ATLAS

**Authors:** SHEMYAKIN, Dmitry (Weizmann Institute of Science (IL)); CHAPMAN, John Derek (University of Cambridge (GB)); MIJOVIC, Liza (University of Edinburgh); JAVURKOVA, Martina (University of Massachusetts (US)); WANG, Rui (Argonne National Laboratory (US))

**Presenter:** WANG, Rui (Argonne National Laboratory (US))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **29** Type: **Poster**

# AtlFast3: Fast Simulation in ATLAS for LHC Run 3 and beyond

*Monday, March 11, 2024 4:15 PM (30 minutes)*

As we are approaching the high-luminosity era of the LHC, the computational requirements of the ATLAS experiment are expected to increase significantly in the coming years. In particular, the simulation of MC events is immensely computationally demanding, and their limited availability is one of the major sources of systematic uncertainties in many physics analyses. The main bottleneck in the detector simulation is the detailed simulation of electromagnetic and hadronic showers in the ATLAS calorimeter system using Geant4. In order to increase the MC statistics and to leverage the available CPU resources for LHC Run 3, the ATLAS collaboration has recently put into production a refined and significantly improved version of its state-of-the-art fast simulation tool AtlFast3. AtlFast3 uses classical parametric and machine learning based approaches such as Generative Adversarial Networks (GANs) for the fast simulation of LHC events in the ATLAS detector.

This talk will present the newly improved version of AtlFast3 that is currently in production for the simulation of Run 3 samples. In addition, ideas and plans for the future of fast simulation in ATLAS will also be discussed.

## Significance

The talk will give an overview of the completely revised configuration of the Run 3 fast simulation used in ATLAS, which is currently in production for the simulation of physics samples. ATLAS aims to use >90% of fast simulation samples in the coming years, such that improvements in fast simulation accuracy are crucial for the success of the ATLAS physics programme.

## References

## Experiment context, if any

ATLAS

**Author:** BEIRER, Joshua Falco (CERN)

**Presenter:** BEIRER, Joshua Falco (CERN)

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **30** Type: **Oral**

# Towards a Simplified (Fast) Simulation Infrastructure in ATLAS

*Wednesday, March 13, 2024 4:50 PM (20 minutes)*

To increase the number of Monte Carlo simulated events that can be produced with the limited CPU resources available, the ATLAS experiment at CERN uses a variety of fast simulation tools in addition to the detailed simulation of the detector response with Geant4. The tools are deployed in a heterogeneous simulation infrastructure known as the Integrated Simulation Framework (ISF), which was originally developed over a decade ago. While ISF allows for a flexible combination of simulation tools, it has accumulated a significant level of complexity over the last few years and is becoming increasingly difficult to maintain by the collaboration. In addition, the complex particle routing algorithms implemented by ISF have been found to cause a measurable overhead in the simulation time. At the same time, recent advances in Geant4 may allow a complete replacement of ISF by outsourcing its entire functionality as a particle stack dispatcher to the Geant4 toolkit.

This talk presents a first implementation of FastCaloSimV2 as a Geant4 fast simulation model. FastCaloSimV2 provides a fast parametric simulation of the ATLAS calorimeter and is part of the state-of-the-art fast simulation tool AtlFast3. Its integration into Geant4 will serve as a reference for the integration of other simulators, which is expected to significantly streamline the simulation infrastructure of the ATLAS experiment in the coming years.

## Significance

ATLAS aims to use >90% of fast simulation samples in the coming years, such that improvements in fast simulation accuracy are crucial for the success of the ATLAS physics programme. In order to maintain the growing fast and full simulation infrastructure of the experiment, a major refactoring of the infrastructure will be required in the coming years. This talk presents the integration of FastCaloSimV2, one of the major fast simulation tools used in the collaboration, as a Geant4 fast simulation model, which will serve as a reference for the integration of other simulators.

## References

## Experiment context, if any

ATLAS

**Author:** BEIRER, Joshua Falco (CERN)

**Presenter:** BEIRER, Joshua Falco (CERN)

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **31** Type: **Poster**

# Describe Data to get Science-Data-Ready Tooling: Awkward as a Target for Kaitai Struct YAML

*Monday, March 11, 2024 4:15 PM (30 minutes)*

In some fields, scientific data formats differ across experiments due to specialized hardware and data acquisition systems. Researchers need to develop, document, and maintain specific analysis software to interact with these data formats. These software are often tightly coupled with a particular data format. This proliferation of custom data formats has been a prominent challenge for small to mid-scale experiments. The widespread adoption of ROOT has largely mitigated this problem for the Large Hadron Collider (LHC) experiments. However, not all experiments use ROOT for their data formats. Experiments such as Cryogenic Dark Matter Search (CDMS) continue to use custom data formats to meet specific research needs. Therefore, simplifying the process of converting a unique data format to analysis code still holds immense value for scientific communities even beyond HEP. We have added Awkward Arrays, a Scikit-HEP library for storing nested and variable data into Numpy-like arrays, as a target language for Kaitai Struct for this purpose.

Kaitai Struct is a declarative language that uses a YAML-like description of a binary data structure to generate the code to read a raw data file in any of the supported languages. Researchers can simply describe their custom data format in the Kaitai Struct YAML (KSY) language only once. The Kaitai Struct Compiler generates C++ code to fill the LayoutBuilder buffers using the KSY format. In a few simple steps, the Kaitai Struct Awkward Runtime API can convert the generated C++ code into a compiled Python module using ctypes. Finally, the raw data file can be passed to the module to produce Awkward Arrays.

This talk will introduce the Awkward Target for the Kaitai Struct Compiler and the Kaitai Struct Awkward Runtime API. It will demonstrate the conversion of a given KSY for a specific custom file format to Awkward Arrays.

## Significance

Collaborations that use a custom data format spend many hours writing their own tools to read and analyse their data. It is difficult to fund such tools because they are not usable outside the collaboration, and as a result, the analysis can be restricted to what the original author envisioned. Even if the tool continues to have a maintainer, it is usually poorly documented and tested, making it difficult to continue maintaining it. It also poses a significant barrier for scientists entering the collaboration.

Switching to a supported standard data format sounds like the most obvious solution. However, in the case of the Cryogenic Dark Matter Search Collaboration, for example, this would require substantial rewriting of the data-acquisition system or switching to a different one. This would require additional personnel and substantial time investment. In addition, there are legacy datasets that still have the potential for new science. This project provides a simple and effective solution to this problem of reading and analysing custom data formats for small and mid-scale collaborations across the sciences. Instead of developing their own tools, the collaborations only need to describe their custom data formats in KSY language just once and then directly use the Kaitai Struct Awkward Runtime API to convert their data into Awkward Arrays.

## References

## Experiment context, if any

Cryogenic Dark Matter Search (CDMS)

**Authors:** GOYAL, Manasvi (Princeton University (US)); OSBORNE, Ianna (Princeton University); PIVARSKI, Jim (Princeton University); Dr ROBERTS, Amy (University of Colorado Denver); ZONCA, Andrea

**Presenter:** GOYAL, Manasvi (Princeton University (US))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **32** Type: **Oral**

# Reconstructing Particle Tracks in One Go with a Recursive Graph Attention Network

*Monday, March 11, 2024 5:30 PM (20 minutes)*

Track reconstruction is a crucial task in particle experiments and is traditionally very computationally expensive due to its combinatorial nature. Many recent developments have explored new tracking algorithms in order to improve scalability in preparation of the HL-LHC. In particular, Graph neural networks (GNNs) have emerged as a promising approach due to the graph nature of particle tracks. Most of these GNN-based methods implement a three-step algorithm, including graph construction, edge classification, and graph segmentation. Others perform object condensation (OC) after the graph construction stage followed by a clustering of the detector hits. In this presentation, we consider a one-shot OC approach which reconstructs particle tracks directly from a set of hits (point cloud) by recursively applying Graph Attention Networks with an evolutional graph structure. This approach simplifies the procedure compared to the three-step approaches and also allows to further regress the hit properties. Preliminary studies on the trackML dataset show physics and computing performance comparable to current production algorithms for track reconstruction.

## Significance

This presentation presents a novel approach for track finding and provides several advantages to the standard tracking algorithm as well as the currently most common ML-based approach. We demonstrate the idea and show very promising results. This novel approach can potentially replace the current method in the future.

## References

This is our very first result.

## Experiment context, if any

**Author:** CHAN, Jay (Lawrence Berkeley National Lab. (US))

**Presenter:** CHAN, Jay (Lawrence Berkeley National Lab. (US))

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **34** Type: **Oral**

# FASER Tracking and Emulsion Station Alignment

*Thursday, March 14, 2024 4:50 PM (20 minutes)*

FASER, the ForwArd Search ExpeRiment, is an LHC experiment located 480 m downstream of the ATLAS interaction point along the beam collision axis. FASER has been taking collision data since the start of LHC Run3 in July 2022. The first physics results were presented in March 2023 [1,2], including the first direct observation of collider neutrinos. FASER includes four identical tracker stations constructed from silicon microstrip detectors, which play a key role in the physics analysis. Specifically the tracker stations are designed to precisely reconstruct the charged particles arising from a new particle decay, as well as high-energy muons from neutrino interactions. For the current analyses the three upstream tracking stations, the tracking spectrometer, have been used. To take full advantage of the detector in neutrino analyses we need to include the neutrino detector, an emulsion detector, and the InterFace Tracker (IFT). The unique geometry of FASER requires a 2-prong alignment procedure for the tracking stations. The tracking spectrometer, which is supported by a precision aluminium beam, is aligned first then the more challenging alignment of the IFT is performed to correct for its relatively large misalignments. After the alignment of the tracking stations we perform the emulsion-IFT alignment. This talk will present updated results for the spectrometer and first result on the IFT and emulsion-IFT alignment using the 2022 collision data.

## Significance

In addition to the tacking station alignment of the FASER detector we will also present the alignment between the emulsion and IFT detectors, this is a unique combination of detectors that need to be aligned, and is not often done in the LHC. The combination of the non-electronic and high density tracks of the emulsion tracks and the much lower density tracks of the IFT provides a challenging alignment effort.

## References

[1] FASER Collaboration, "First direct observation of collider neutrinos with FASER at the LHC." Physics Review Letters, 2023, 031801.
[2] FASER Collaboration, "Search for dark photons with the FASER detector at the LHC", Physics Letters B, 848, 2024, 138378.

## Experiment context, if any

FASER

**Authors:** GARABAGLU, Ali (University of Washington (US)); LI, Ke (University of Washington (US))

**Presenter:** GARABAGLU, Ali (University of Washington (US))

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **35** Type: **Poster**

# Boosting CPU Efficiency in ATLAS Inner Detector Reconstruction with Track Overlay

*Monday, March 11, 2024 4:15 PM (30 minutes)*

In response to the rising CPU consumption and storage demands, as we enter a new phase in particle physics with the High-Luminosity Large Hadron Collider (HL-LHC), our efforts are centered around enhancing the CPU processing efficiency of reconstruction within the ATLAS inner detector. The track overlay approach involves pre-reconstructing pileup tracks and subsequently running reconstruction exclusively on hard-scatter tracks. This allows us to conserve valuable CPU resources by concentrating on events of interest. Integral to track overlay is the incorporation of machine learning (ML)-based decision processes. ML decisions guide the selection of events suitable for track overlay, while events in denser environments continue to use the standard overlay. This strategy ensures judicious use of resources, balancing efficiency and precision in inner detector reconstruction. This presentation focuses on constructing the ML model and verifying the workflow with ML decisions. The improvement of the track overlay approach on CPU usage and the reduction in the size of standard data format files in the Run 3 detector setup are also demonstrated. Preliminary results in the context of the forthcoming ITk inner detector at HL-LHC will be presented as well.

## Significance

This presentation goes beyond traditional status reports by showcasing the results of the track overlay approach, coupled with machine learning-based decision processes and significant advancements in the workflow. We will also present a preliminary result with Run 4 setup, which is a totally different inner detector at HL-LHC from Run 3.

## References

Faster simulated track reconstruction in the ATLAS Fast Chain
https://indico.jlab.org/event/459/papers/11440/

## Experiment context, if any

This abstract is conducted in the context of the ATLAS experiment at the LHC.

**Authors:** DUDA, Dominik (The University of Edinburgh (GB)); TSAI, Fang-Ying (Stony Brook University (US)); CHAPMAN, John Derek (University of Cambridge (GB)); JAVURKOVA, Martina (University of Massachusetts (US)); JIGGINS, Stephen (Deutsches Elektronen-Synchrotron (DE))

**Presenter:** TSAI, Fang-Ying (Stony Brook University (US))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **36** Type: **Oral**

# Optimizing the CMS Offline Software Infrastrucure for Run 3

*Wednesday, March 13, 2024 2:50 PM (20 minutes)*

The CMSSW framework has been instrumental in data processing, simulation, and analysis for the CMS detector at CERN. It is expected to remain a key component of the CMS Offline Software for the foreseeable future. Consequently, CMSSW is under continuous development, with its integration system evolving to incorporate modern tools and keep pace with the latest software improvements in the High Energy Physics (HEP) community. This contribution presents an in-depth examination of the recent enhancements made to the CMSSW infrastructure. Technical improvements, such as advanced compiler techniques like Link Time Optimization (LTO) and Profile-Guided Optimization (PGO), have been successfully integrated into the CMSSW infrastructure. Additionally, the adoption of heterogeneous resources and multi-vectorization architectures has contributed to a variety of software flavors and architectures, providing different approaches to identify bugs and legacy code at an early stage. To efficiently accommodate the increasing workloads of such techniques, the migration of the CernVM File System to a parallel publishing setup has also been engineered according to the experiment's needs. We will finally discuss the enhancement of the CMS Continuous Integration infrastructure, focusing on the adoption of new methods for monitoring and scheduling, testing, and integrating the software stack. Overall, these advancements in CMSSW have not only prepared it for the ongoing Run 3 but also underscore our commitment to continuously optimizing the CMS software infrastructure.

## Significance

This contribution includes all the improvements made to the CMS Offline Software infrastructure to prepare it for the challenges of Run 3 and builds upon the last progress report at CHEP 2019. We believe that the novel techniques adopted will be of significant interest to the HEP community.

## References

Last reporting at CHEP 2019: "Modernizing the CMS software stack"(https://indico.cern.ch/event/773049/contributions/347
Preliminary results of the compiler techniques Link Time Optimization (LTO) and Profile-Guided Optimization (PGO) were presented at ACAT 2022: "Speeding up CMS simulations, reconstruction and HLT code using advanced compiler options"(https://indico.cern.ch/event/1106990/contributions/4991214/)

## Experiment context, if any

This contribution is set in the context of the CMS experiment at CERN. It has been submitted on behalf of the CMS Collaboration. The abstract has been approved by the CMS Conference Committee.

**Author:** VALENZUELA RAMIREZ, Andrea (CERN)

**Co-authors:** Dr RAZUMOV, Ivan (Princeton University (US)); MUZAFFAR, Malik Shahzad (CERN)

**Presenter:** VALENZUELA RAMIREZ, Andrea (CERN)

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **37** Type: **Oral**

# Implementing an emissions model for dual phase xenon TPCs with probabilistic programming

*Thursday, March 14, 2024 4:50 PM (20 minutes)*

Traditionally, analysis of data from experiments such as LZ and XENONnT have relied on summary statistics of large sets of simulated data, generated using emissions models for particle interactions in liquid xenon such as NEST. As these emissions models are probabilistic in nature, they are a natural candidate to be implemented in a probabilistic programming framework. This would also allow for direct inference of latent variables that we are interested in, such as energy. In this work, I will describe the challenges faced in creating such an implementation, and the possible applications, such as probabilistic energy reconstruction.

## Significance

This presentation will cover a new attempt to implement a liquid xenon emissions model in probabilistic programming, so that a model that was previously a black-box model for inference can have explicit likelihoods without the need for summary statistics. The benefits of this approach would also be discussed.

## References

## Experiment context, if any

**Author:** Dr QIN, Juehang (Rice University)

**Co-author:** Prof. TUNNELL, Christopher (Rice University)

**Presenter:** Dr QIN, Juehang (Rice University)

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **39** Type: **Oral**

# Pepper –A Portable Parton-Level Event Generator for the High-Luminosity LHC

*Wednesday, March 13, 2024 3:30 PM (20 minutes)*

Parton-level event generators are one of the most computationally demanding parts of the simulation chain for the Large Hadron Collider. The rapid deployment of computing hardware different from the traditional CPU+RAM model in data centers around the world mandates a change in event generator design. These changes are required in order to provide economically and ecologically sustainable simulations for the high-luminosity era of the LHC. We present the first complete leading-order parton-level event generation framework capable of utilizing most modern hardware, and discuss its performance in standard-candle processes at the LHC.

## Significance

This is the first time we present Pepper at a conference. It is the first production-ready portable parton-level event generator framework.

## References

## Experiment context, if any

**Authors:** Dr BOTHMANN, Enrico (U Goettingen); ISAACSON, Joshua; KNOBBE, Max (University of Göttingen); HOECHE, Stefan (Fermilab); CHILDERS, Taylor; GIELE, Walter

**Presenter:** Dr BOTHMANN, Enrico (U Goettingen)

**Session Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

**Track Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

Contribution ID: **40** Type: **Oral**

# QUnfold: Quantum Annealing for Distributions Unfolding in High-Energy Physics

*Monday, March 11, 2024 3:10 PM (20 minutes)*

In High-Energy Physics (HEP) experiments, each measurement apparatus exhibit a unique signature in terms of detection efficiency, resolution, and geometric acceptance. The overall effect is that the distribution of each observable measured in a given physical process could be smeared and biased. Unfolding is the statistical technique employed to correct for this distortion and restore the original distribution. This process is essential to make effective comparisons between the outcomes obtained from different experiments and the theoretical predictions.
The emerging technology of Quantum Computing represents an enticing opportunity to enhance the unfolding performance and potentially yield more accurate results.
This work introduces QUnfold, a simple Python module designed to address the unfolding challenge by harnessing the capabilities of quantum annealing. In particular, the regularized log-likelihood minimization formulation of the unfolding problem is translated to a Quantum Unconstrained Binary Optimization (QUBO) problem, solvable by using quantum annealing systems.
The algorithm is validated on a simulated sample of particles collisions data generated combining the Madgraph Monte Carlo event generator and the Delphes simulation software to model the detector response. A variety of fundamental kinematic distributions are unfolded and the results are compared with conventional unfolding algorithms commonly adopted in precision measurements at the Large Hadron Collider (LHC) at CERN.
The implementation of the quantum unfolding model relies on the D-Wave Ocean software and the algorithm is run by heuristic classical solvers as well as the physical D-Wave Advantage quantum annealer boasting 5000+ qubits.

## Significance

## References

GitHub repo: https://github.com/JustWhit3/QUnfold

## Experiment context, if any

**Authors:** Dr BIANCO, Gianluca (Universita e INFN, Bologna (IT)); GASPERINI, Simone (Universita e INFN, Bologna (IT))

**Co-author:** LORUSSO, Marco (Universita Di Bologna (IT))

**Presenters:** Dr BIANCO, Gianluca (Universita e INFN, Bologna (IT)); GASPERINI, Simone (Universita e INFN, Bologna (IT))

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **41** Type: **Poster**

# Offline data processing in the First JUNO Data Challenge

*Monday, March 11, 2024 4:15 PM (30 minutes)*

The Jiangmen Underground Neutrino Observatory (JUNO) is currently under construction in southern China, with the primary goals of the determining the neutrino mass ordering and the precisely measurement of oscillation parameters. The data processing in JUNO is challenging. When JUNO starts data taking in the late 2024, the expected event rate is about 1 kHz, which is about 31.5 billions of events per year. About 60 MB of byte-stream raw data is produced every second, which is about 2 PB per year. The raw data is transferred from the JUNO onsite to the IHEP data center via a dedicated network. At IHEP data center, the raw data is preprocessed and converted to the ROOT-based raw data format (RTRAW). Then both raw and RTRAW data are replicated to the other data centers, including CC-IN2P3, INFN-CNAF and JINR. There are several critical components in the data processing, such as data quality monitoring (DQM), keep up reconstruction (KUP) and physics production (PP).

A series of JUNO Data Challenges are proposed to evaluate and validate the complete data processing chain in advance. In this contribution, the offline data processing in the first JUNO Data Challenge (DC-1) will be presented. The major goal of DC-1 is processing 1-week of RTRAW data with conditions database and multi-threaded reconstruction. The workflow consists of production of the 45 TB of simulated RTRAW data and reconstruction of the RTRAW data in DQM, KUP and PP. In order to test the conditions database, 7 batches of conditions data are prepared and added in the simulation, then these conditions data are loaded from database in the reconstruction. The tests show that the conditions data are loaded correctly when new events are loaded. A JUNO-Hackathon is organized to let the core software experts and algorithm developers work together in order to support the multi-threaded reconstruction algorithms in the DC-1. The multi-threaded reconstructions are performed well within local computing resources and distributed computing resources respectively.

## Significance

This contribution will summarize the current status of the JUNO Data Challege before data taking.

## References

## Experiment context, if any

JUNO

**Author:** LIN, Tao (Chinese Academy of Sciences (CN))

**Presenter:** LIN, Tao (Chinese Academy of Sciences (CN))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **42** Type: **Poster**

# Energy consumption characterization of Subnuclear Physics computing workloads

*Monday, March 11, 2024 4:15 PM (30 minutes)*

Among human activities that contribute to the environmental footprint of our species, computational footprint, i.e. the environmental impact that results from the employment of computing resources, might be one of the most underappreciated ones. While many modern scientific discoveries have been obtained thanks to the availability of more and more performing computers and algorithms, the energy and carbon footprint related to powering and exploiting such technologies is often underacknowledged. Since hardware-related improvements (i.e. Moore's Law and Dennard Scaling) are considered close to reach a *plateau*, this underestimation might lead, in the near future, the computing sector to become unsustainable for the environment and economically unaccessible for researchers.
Investigations in computing sustainability mainly took into account the hardware lifecycle, the focus being curbing the footprint related to resource provisioning and de-commisioning phases. Today, due to the pervasiveness of computing, it is becoming evident that resource exploitation must be taken into account too. Some computing niches, AI as an example, are indeed starting to shed some light on the usage of computing resources in their sector in order to understand how can their research be kept as energy efficient as reasonably possible. Nonetheless, the energy impact of many other scientific applications is typically only vaguely outlined.
With the goal of encouraging a better footprint description of computing activities in physics, in this work we show the footprint of computing workloads belonging to the branch of Subnuclear physics. We focus on the performance-related interplay between workload type, CPUs and energy usage. This data is expected to characterize the resource usage and offer actionable insights for curbing their energy eagerness.

## Significance

This work should give a frame of reference for the energy efficiency of benchmark subnuclear physics workloads which, to the best of my knowledge, is not available in literature yet.

## References

## Experiment context, if any

**Authors:** MINARINI, Francesco; LORUSSO, Marco; LORUSSO, Marco; LORUSSO, Marco (Universita Di Bologna (IT))

**Presenters:** LORUSSO, Marco; LORUSSO, Marco; LORUSSO, Marco (Universita Di Bologna (IT))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **43** Type: **Oral**

# Line Segment Tracking: Improving the Phase 2 CMS High Level Trigger Tracking with a Novel, Hardware-Agnostic Pattern Recognition Algorithm

*Monday, March 11, 2024 2:50 PM (20 minutes)*

Charged particle reconstruction is one the most computationally heavy components of the full event reconstruction of Large Hadron Collider (LHC) experiments. Looking to the future, projections for the High Luminosity LHC (HL-LHC) indicate a superlinear growth for required computing resources for single-threaded CPU algorithms that surpass the computing resources that are expected to be available. The combination of these facts creates the need for efficient and computationally performant pattern recognition algorithms that will be able to run in parallel and possibly on other hardware, such as GPUs, given that these become more and more available in LHC experiment and high-performance computing centres. Line Segment Tracking (LST) is a novel such algorithm which has been developed to be fully parallelizable and hardware agnostic. The latter is achieved through the usage of the Alpaka library. The LST algorithm has been tested with the CMS central software as an external package and has been used in the context of the CMS HL-LHC High Level Trigger (HLT). When employing LST for pattern recognition in the HLT tracking, the physics and timing performances are shown to improve with respect to the ones utilizing the current pattern recognition algorithms. The latest results on the usage of the LST algorithm within the CMS HL-LHC HLT are presented, along with prospects for further improvements of the algorithm and its CMS central software integration.

## Significance

This presentation covers a new pattern recognition algorithm, Line Segment Tracking, shown for the first time in this conference. The algorithm is developed in the context of the CMS experiment and its application and performance in the High Level Trigger of CMS will be presented in this talk for the first time ever in a conference.

## References

## Experiment context, if any

CMS

**Authors:** YAGIL, Avi (Univ. of California San Diego (US)); SATHIA NARAYANAN, Balaji Venkat (Univ. of California San Diego (US)); NIENDORF, Gavin (Cornell University (US)); GUIANG, Jonathan (Univ. of California San Diego (US)); VOURLIOTIS, Manos (Univ. of California San Diego (US)); TADEL, Matevz (Univ. of California San Diego (US)); SILVA, Mayra (University of Florida); ELMER, Peter (Princeton University (US)); WITTICH, Peter (Cornell University (US)); CHANG, Philip (University of

Florida (US)); KRUTELYOV, Slava (Univ. of California San Diego (US)); REID, Tres (Cornell University (US)); GU, Yanxi (Univ. of California San Diego (US))

**Presenter:** VOURLIOTIS, Manos (Univ. of California San Diego (US))

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **44** Type: **Oral**

# Deep learning methods for noise filtering in the NA61/SHINE experiment.

*Thursday, March 14, 2024 3:30 PM (20 minutes)*

The NA61/SHINE experiment is a prominent venture in high-energy physics, located at the SPS accelerator within CERN. Recently, the experiment's physics program has been extended, which necessitated the upgrade of detector hardware and software for new physics purposes.

The upgrade included a fundamental modification of the readout electronics (front-end) in the detecting system core of the NA61/SHINE, namely the time projection chambers (TPCs). This improvement increased data flow rates, raising them from 80 Hz to 1.7 kHz.

In light of the significant increase in the amount of data collected, it has become necessary to implement an online noise filtering tool.

Traditionally, this task has relied on the reconstruction of particle tracks and the subsequent removal of clusters that lack association with any discernible particle trajectory. However, it's important to acknowledge that this method consumes a noteworthy amount of time and computational resources.

In the year 2022, the initial dataset was collected through the utilization of the upgraded detector system. In relation to this data, a collection of machine learning models was developed, employing two distinct categories of neural networks: dense and convolutional networks (DNN, CNN). Of utmost significance is the seamless integration of these trained models into the existing NA61/SHINE C++ software framework, utilizing the capabilities of the TensorFlow C++ library. Furthermore, to facilitate easier deployment, containerization using Docker was applied. It is production ready.

This presentation aims to unveil the results attained through the application of these algorithms for noise reduction, encompassing training times for both CNN and DNN models, post-filtering data reconstruction duration, and the Receiver Operating Characteristic (ROC) analysis of the CNN-filtered data. During this presentation, we intend to unveil the outcomes yielded by the application of these algorithms for noise filtration. Additionally, we will delve into the time performance of these models, offering insights into various pertinent metrics.

We will compare the results of reference invariant mass spectra generated with and without the ML noise rejection and draw the conclusions on the influence of employed ML methods on the physical results.

## Significance

The incremental updates of an important project are: integration of these trained models into the existing NA61/SHINE C++ software framework, utilizing the capabilities of the TensorFlow C++ library, test of the methods on new, 2022 data, comparison of the results of reference invariant mass spectra generated with and without the ML noise rejection.

## References

J.Phys.Conf.Ser. 2438 (2023) 1, 012104 • Contribution to: ACAT 2021

## Experiment context, if any

NA61/SHINE CERN

**Authors:** SCHMIDT, Katarzyna (University of Silesia (PL)); BRYLINSKI, Wojciech (Warsaw University of Technology (PL)); Mr MAKULSKI, Kordian (Warsaw University of Technology, Warsaw, Poland); SLODKOWSKI, Marcin (Warsaw University of Technology (PL)); WYSZYNSKI, Oskar (Jan Kochanowski University (PL))

**Presenter:** SCHMIDT, Katarzyna (University of Silesia (PL))

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **45** Type: **Oral**

# A Mechanism for Asynchronous Offloading in the Multithreaded Gaudi Event Processing Framework

*Wednesday, March 13, 2024 3:10 PM (20 minutes)*

High Performance Computing resources are increasingly prominent in the plans of funding agencies, and the tendency of these resources is now to rely primarily on accelerators such as GPUs for the majority of their FLOPS. As a result, High Energy Physics experiments must make maximum use of these accelerators in our pipelines to ensure efficient use of the resources available to us.

The ATLAS and LHCb experiments share a common data processing architecture called Gaudi. In Gaudi, data processing workloads are ultimately split into units called Algorithms, and Gaudi uses a smart scheduler (the Avalanche scheduler) to schedule these Algorithms on a fixed pool of CPU threads managed by Intel's TBB.

This is an architecture that efficiently fills the available CPU capacity provided the algorithms are primarily CPU-limited. However when the algorithms offload a large portion of their computational work to GPUs they can be left blocking a CPU thread, wasting precious core-time.

Here we present a prototype of an addition to this scheduler, which places such GPU-accelerated algorithms on a separate pool of dedicated threads. By making use of lightweight Boost Fibers, and the ability to suspend these fibers without suspending the underlying OS thread, we can run the GPU workload asynchronously, without blocking the thread. This allows more efficient use of the CPU resources, and where the work offloaded by a single Algorithm doesn't fill the GPU resources available can also improve GPU-efficiency by making use of separate CUDA streams.

## Significance

This work presents an addition to the Gaudi Avalanche scheduler which enables it to deal with GPU-accelerated algorithms in a CPU efficient manner.

## References

## Experiment context, if any

ATLAS, LHCb

**Authors:** STANISLAUS, Beojan (Lawrence Berkeley National Lab. (US)); Dr LEGGETT, Charles (Lawrence Berkeley National Lab (US)); ESSEIVA, Julien (Lawrence Berkeley National Lab. (US)); CALAFIURA, Paolo (Lawrence Berkeley National Lab. (US)); TSULAIA, Vakho (Lawrence Berkeley National Lab. (US)); JU, Xiangyang (Lawrence Berkeley National Lab. (US))

**Presenter:** STANISLAUS, Beojan (Lawrence Berkeley National Lab. (US))

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **46** Type: **Oral**

# Finetuning Foundation Models for Joint Analysis Optimization

*Tuesday, March 12, 2024 12:10 PM (20 minutes)*

In this work we demonstrate that significant gains in performance and data efficiency can be achieved moving beyond the standard paradigm of sequential optimization in High Energy Physics (HEP). We conceptually connect HEP reconstruction and analysis to modern machine learning workflows such as pretraining, finetuning, domain adaptation and high-dimensional embedding spaces and quantify the gains in the example usecase of searches of heavy resonances decaying via an intermediate di-Higgs to four b-jets.

## Significance

We demonstrate a finetuning workflow in the hierarchical setting of per-object representation and event-level inference within particle physics, quantifying the significant gains due to end-to-end optimization with respect to data efficiency and performance at fixed sample size. We also provide evidence of successful domain adaptation in a hierarchical setting of HEP foundation models finetuned on datasets other than the one they are pretrained with.

## References

Paper: https://arxiv.org/abs/2401.13536
ML4Jets 2023: https://indico.cern.ch/event/1253794/contributions/5588562/
2023 ATLAS Flavour Tagging Workshop: https://indico.cern.ch/event/1311519/contributions/5582015/

## Experiment context, if any

CMS open data

**Authors:** HEINRICH, Lukas Alexander (Technische Universitat Munchen (DE)); VIGL, Matthias (Technische Universitat Munchen (DE)); HARTMAN, Nicole Michelle (TUM (DE))

**Presenter:** VIGL, Matthias (Technische Universitat Munchen (DE))

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **47** Type: **Poster**

# The performance profiling of ptycho-W1Net AI algorithm on DCU and HUAWEI NPU 910

*Thursday, March 14, 2024 4:10 PM (30 minutes)*

APS at USA completed the high-precision training and inference in Nvidia GPU clusters taking the ptychoNN algorithm combined with ePIE Conjugate Gradient method. By the reference of that idea, we came up with a new model called W1-Net whose training speed was faster with higher precision of inference. After this development, we implemented the model onto DCU cluster. However, the performance was only 1/6 of Nvidia GPU A100. Profiling action was done to the training process and the low speed was caused by the atom operation during the training function. After tuning the code, the training time was reduced by half then the previous model. Apart from DCU, we also trained on HUAWEI NPU card. This paper will show the profiling result of HUAWEI NPU 910*8 cluster.

## Significance

The training process is on the heterogeneous computing card on HUAWEI Ascend 910 which is different from Nvidia and training speed can be comparable to Nvidia GPU A100.

## References

Title: 'W1-Net:A fast training and highly scalable ptychography convolutional neural network'. This paper is underreview by 'The European Physical Journal Plus'.

## Experiment context, if any

The data comes from this website: https://github.com/mcherukara/PtychoNN

**Authors:** Dr WANG, Lei (Institute of High Energy Physics); Dr MU, Yangyang (Institute of High Energy Physics)

**Co-authors:** Dr XING, Chengye (Institute of High Energy Physics); Dr CHANG, Guangcai (Institute of High Energy Physics); Dr SHI, Jingyan (Institute of High Energy Physics); Dr HU, Jiarui (Institute of High Energy Physics); Dr HU, Yu (Institute of High Energy Physics); Dr LIU, Jianli (Institute of High Energy Physics); Dr SUN, Haokai (Institute of High Energy Physics); Dr LIU, Rui (Institute of High Energy Physics); Dr FU, Shiyuan (Institute of High Energy Physics); Dr QI, Fazhi (Institute of High Energy Physics); Dr CHENG, Yaodong (Institute of High Energy Physics)

**Presenter:** Dr WANG, Lei (Institute of High Energy Physics)

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **49** Type: **Poster**

# The ATLAS Web Run Control system

*Monday, March 11, 2024 4:15 PM (30 minutes)*

The ATLAS experiment at the Large Hadron Collider (LHC) operated very successfully in the years 2008 to 2023. ATLAS Control and Configuration (CC) software is the core part of the ATLAS Trigger and DAQ system, it comprises all the software required to configure and control the ATLAS data taking. It provides essentially the glue that holds the various ATLAS sub-systems together. During recent years, more and more applications in CC software were made available as web applications, thus making them easily available to experts for remote operations. The important missing part, however, was the main data taking control and monitoring application known as 'Igui' (Integrated Graphical User Interface), the front-end GUI tool used by operators in ATLAS Control Room to steer the data taking sessions.

This paper presents the new web application called 'WebRC' (Web Run Control), which provides Igui-like functionality of the data taking monitoring and control from a web browser, including: presenting the Run Control tree of all applications, dynamically updating their states, browsing their log files and messages stream, monitoring different system and trigger rates and detector busy information, thus allowing experts to promptly assess the state of the data taking and to investigate possible issues.

WebRC is built using Apache Wicket framework, and it is java-only back-end application. Important requirement which led to this choice was the necessity to closely integrate with CC services at the back-end side, with high performance and high scalability in mind. Another aspect is the long-term maintainability of the code, where in case of Wicket we benefit from not having to maintain any front-end JS library: on the browser side, a Wicket application is a plain HTML markup page, and all graphical elements management is fully done in Java on the server side. Wicket leverages the standard HTTP and AJAX technologies for achieving dynamic behavior of the application.

WebRC can operate in two modes, Control and Display, where the former allows to fully control user data taking session by sending control commands to the applications. Recent development includes integration of WebRC with CERN IT Open-ID authentication infrastructure, allowing the actions to be performed on behalf of the user authenticated in CERN SSO page in a web browser, thus enabling full user control over TDAQ data taking sessions, including changes in the DAQ configuration, sending of RC commands and starting and stopping the DAQ session processes. However this mode is disabled for operations in ATLAS, where WebRC is running in Display mode.

WebRC is widely used for monitoring of data taking sessions during ongoing Run 3 period, with dozens of users connecting daily.

## Significance

This paper presents the new web application called 'WebRC' (Web Run Control), which provides Igui-like functionality of the data taking monitoring and control from a web browser, including: presenting the Run Control tree of all applications, dynamically updating their states, browsing their log files and messages stream, monitoring different system and trigger rates and detector busy information, thus allowing experts to promptly assess the state of the data taking and to investigate possible issues.

## References

## Experiment context, if any

ATLAS, CERN

**Authors:** KOULOURIS, Aimilianos (CERN); OH, Alexander (University of Manchester (GB)); KAZAROV, Andrei (University of Johannesburg (SA))

**Presenter:** KOULOURIS, Aimilianos (CERN)

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **50** Type: **Oral**

# Phase-2 Upgrade of the ATLAS L1 Central Trigger

*Thursday, March 14, 2024 5:30 PM (20 minutes)*

The ATLAS trigger system will be upgraded for the Phase 2 period of LHC operation. This system will include a Level-0 (L0) trigger based on custom electronics and firmware, and a high-level software trigger running on off-the-shelf hardware. The upgraded L0 trigger system uses information from the calorimeters and the muon trigger detectors. Once information from all muon trigger sectors has been received, trigger candidate multiplicities are calculated by the Muon-to-Central-Trigger-Processor Interface (MUCTPI). Muon multiplicity information is sent to the Central-Trigger-Processor (CTP) and trigger objects are sent to the L0 Global Trigger Processor (L0Global). In Phase 2, the CTP will be a newly designed and custom-built electronics system, based on the ATCA standard. It will employ a System-on-Chip (SoC) and optical serial inputs to receive trigger information from the Global Trigger and the MUCTPI system. The control and monitoring software run directly on the SoC, while the trigger logic runs on an FPGA. The CTP will need to allow a set of 1024 trigger items based on 1024 usable single-bit inputs, requiring updates in the trigger logic implementation, as well as the software for compiling the trigger conditions into FPGA configuration files. New features will also be introduced, such as delayed triggers. We will present the design and status of the Phase 2 L0CT system and its new features, including a view of the pilot Phase 1 upgrade, which paves the way for the upcoming upgrades.

## Significance

Comprehensive overview on the Phase-2 upgrade of the Level 1 Central Trigger of ATLAS.

## References

## Experiment context, if any

ATLAS, CERN

**Authors:** KOULOURIS, Aimilianos (CERN); OH, Alexander (University of Manchester (GB))

**Presenter:** KOULOURIS, Aimilianos (CERN)

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **51** Type: **Oral**

# ATLAS TDAQ Phase-2

*Thursday, March 14, 2024 5:10 PM (20 minutes)*

The ATLAS experiment at CERN will be upgraded for the "High Luminosity LHC", with collisions due to start in 2029. In order to deliver an order of magnitude more data than previous LHC runs, 14 TeV protons will collide with an instantaneous luminosity of up to 7.5 x 10e34 cmˆ-2sˆ-1, resulting in higher pileup and data rates. This increase brings new requirements and challenges for the trigger and data acquisition system (TDAQ), as well as for the detector and computing systems.

The design of the TDAQ upgrade comprises:
- a hardware-based low-latency real-time Trigger operating at 40 MHz,
- data acquisition which combines custom readout with commodity hardware and networking to deal with 4.6 TB/s input, and
- an Event Filter running at 1 MHz which combines offline-like algorithms on a large commodity computing service with the potential to be augmented by commercial accelerators.

Commodity servers and networks are used as far as possible, with custom ATCA boards, high speed links and powerful FPGAs deployed in the low-latency parts of the system. Offline-style clustering and jet-finding in FPGAs, as well as accelerated track reconstruction are designed to combat pileup in the Trigger and Event Filter respectively.

This contribution will report recent progress on the design, technology and construction of the system. The physics motivation and expected performance will be shown for key physics processes.

## Significance

This contribution will report recent progress on the design, technology and construction of the ATLAS Trigger and Data Acquisition system for LHC Phase-2. The physics motivation and expected performance will be shown for key physics processes.

## References

## Experiment context, if any

ATLAS, CERN

**Authors:** OH, Alexander (University of Manchester (GB)); PASTORE, Francesca (Royal Holloway, University of London)

**Presenter:** PASTORE, Francesca (Royal Holloway, University of London)

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **52** Type: **Oral**

# Machine learning-based particle identification of atmospheric neutrinos in JUNO

*Wednesday, March 13, 2024 3:30 PM (20 minutes)*

The Jiangmen Underground Neutrino Observatory (JUNO) is a next-generation large (20 kton) liquid-scintillator neutrino detector, which is designed to determine the neutrino mass ordering from its precise reactor neutrino spectrum measurement. Moreover, high-energy (GeV-level) atmospheric neutrino measurements could also improve its sensitivity to mass ordering via matter effects on oscillations, which depend on the capability to identify electron (anti-)neutrinos and muon (anti-)neutrinos against each other and against neutral current background, as well as to identify neutrinos against antineutrinos. However, this particle identification task has never been attempted in large homogeneous liquid scintillator detectors like JUNO.
This contribution presents a machine learning approach for the particle identification of atmospheric neutrinos in JUNO. In this method, several features relevant to event topology are extracted from PMT waveforms and used as inputs to the machine learning models. Moreover, the features from captured neutrons could also provide the capability of neutrinos versus anti-neutrinos identification. Two independent strategies are developed to utilize neutron information and to combine these two types of inputs information in different machine learning models. Preliminary results based on Monte Carlo simulations show promising potential for this approach.

## Significance

In this contribution, we provides a multi-purpose machine learning method for atmospheric neutrino reconstruction and particle identification in large unsegmented liquid scintillator detectors like JUNO. Specifically, for this particle identification task, our models' architecture are updated to utilize the extra neutron information that could help classifying neutrinos and anti-neutrinos.

## References

https://indico.cern.ch/event/1264216/contributions/5548526/attachments/2702032/4731931/ML_FlavorIdent_IPRD_Fanrui.p

## Experiment context, if any

Jiangmen Underground Neutrino Observatory (JUNO)

**Authors:** ZENG, Fanrui (China, Shandong University); DUYANG, Hongyue (Shandong University); LIU, Jiaxi; Dr LI, Teng (Shandong University, CN); MA, Wing Yan (SDU); LUO, Wuming (Institute of High Energy Physics, Chinese Academy of Science); HE, Xinhai (The Institute of High Energy Physics of the Chinese Academy of Sciences)

**Presenter:** LIU, Jiaxi

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **53** Type: **Poster**

# Deployment of ATLAS Calorimeter Fast Simulation Training Through Container Technology

*Monday, March 11, 2024 4:15 PM (30 minutes)*

Simulation of the detector response is a major computational challenge in modern High Energy Physics experiments, as for example it accounts for about two fifths of the total ATLAS computing resources. Among simulation tasks, calorimeter simulation is the most demanding, taking up about 80% of resource use for simulation and expected to increase in the future. Solutions have been developed to cope with this demand, notably fast simulation tools based on Machine Learning (ML) techniques, which are faster than Geant4 when simulating calorimeter response and maintain a high level of accuracy. However, these ML-based models require a lot of computing resources to train.
Moreover, computational resources can also be saved by deploying their training on other resources than the CERN HTCondor batch system or the Worldwide LHC Computing Grid, with the opportunity to have an additional boost in computing performance.
In this work we introduce FastCaloGANtainer, a containerized version of FastCaloGAN, a fast simulation tool developed by the ATLAS Collaboration. FastCaloGANtainer allows the training of this tool on more powerful devices such as High Performance Computing clusters and reduces software dependencies on local or distributed file systems (such as CVMFS). We describe the testing methodology and the results obtained on different resources with different operating systems and installed software, with or without GPUs.

## Significance

Fast simulation is one way to address the issue of calorimeter simulation activities being a major source of computing resource consumption and demand, but its training is still resource demanding. Containerization of fast simulation training can help reduce this burden by enabling the use of more powerful devices for these tasks. This can free up the commonly used computing resources and create more space for additional processing requests.

## References

## Experiment context, if any

ATLAS

**Authors:** CORCHIA, Federico Andrea (Universita e INFN, Bologna (IT)); BEIRER, Joshua Falco (Georg August Universitaet Goettingen (DE)); BEIRER, Joshua Falco (CERN); RINALDI, Lorenzo (Universita e INFN, Bologna (IT)); FAUCCI GIANNELLI, Michele (INFN e Universita Roma Tor Vergata (IT)); ZHANG, Rui (University of Wisconsin Madison (US))

**Presenters:** BEIRER, Joshua Falco (Georg August Universitaet Goettingen (DE)); BEIRER, Joshua Falco (CERN)

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **54** Type: **Poster**

# The Workflow Management System for Data Processing towards Photon Sources

*Monday, March 11, 2024 4:15 PM (30 minutes)*

The new-generation light sources, such as the High Energy Photon Source (HEPS) under construction, are one of the advanced experimental platforms that facilitate breakthroughs in fundamental scientific research. These large scientific installations are characterized by numerous experimental beam lines (more than 90 at HEPS), rich research areas, and complex experimental analysis methods, leading to many data processing challenges: high-throughput multi-modal data, flexible and flexible and diverse scientific methodology, and highly differentiated experimental analytical processes. For a long time, there has been a lack of a general, user-friendly, and fully functional experimental data processing full-process management system in China and abroad. This project will use the idea of "workflow" to independently design and implement a set of graphical general-purpose management systems to solve the following key problems: how to quickly share and apply data processing methods to experiments by beam-line scientists, experiment users, and methodology developers during analysis; how researchers can flexibly customize and monitor complex and diverse data processing processes; how the whole process of experimental analysis can be applied in batches to similar experiments and the results can be reproduced. Two typical photon source experiments, the acquisition of the pair distribution function (PDF) of diffraction scatterings and the structural analysis of biological macro-molecules, will be used as application examples to show us how the workflow management system facilitates scientific research.

## Significance

For a long time, there has been a lack of a universal, user-friendly, and comprehensive experimental data processing end-to-end management system, both domestically and internationally. Employing the workflow as the central tool offers an excellent solution by establishing connections between algorithm developers and experimental users. Moreover, it serves as a bridge linking hardware facilities and various modules within data processing software.

## References

## Experiment context, if any

**Authors:** Dr SUN, Hao-Kai (IHEP, CAS); HU, Yu

**Co-authors:** FUSY, FU Shiyuan; LIU, Jianli (Institute of High Energy Physics); FAZHI, Qi (IHEP); LIU, Rui (Institute of High Energy Physics); WANG, lei (Institute of High Energy Physics)

**Presenter:** WANG, lei (Institute of High Energy Physics)

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **56** Type: **Oral**

# Towards an open-source hybrid quantum operating system

*Monday, March 11, 2024 5:30 PM (20 minutes)*

Over the last 20 years, thanks to the development of quantum technologies, it has been possible to deploy quantum algorithms and applications that before were only accessible through simulation on real quantum hardware.
The current devices available are often referred to as noisy intermediate-scale quantum (NISQ) computers, and they require calibration routines in order to obtain consistent results.

In this context, we present Qibo, an open-source framework for quantum computing. Qibo was initially born as a tool for simulating quantum circuits.
Through its modular layout for backend abstraction, it is possible to change effortlessly between different backends, including simulator based on just-in-time compilation, Qibojit.

In order to enable the execution and calibration of self-hosted quantum hardware we have developed two open-source libraries integrated with the Qibo framework: Qibolab and Qibocal.
Qibolab provides the software layer required to automatically execute circuit-based algorithms on custom self-hosted quantum hardware platforms.
It enables experimentalists and developers to delegate all complex aspects of hardware implementation to the library so they can standardize the deployment of quantum computing algorithms in a hardware-agnostic way.

Qibocal is based on a modular QPU (Quantum Processing Unit) platform agnostic approach and introduces tools
that support the calibration and characterization of QPUs on three different levels: development, deployment and distribution. Qibocal provides a code library to rapidly develop protocols for different hardware abstraction layers.
The integration with Qibo allows one to easily switch between hardware execution and high-perfomance simulation.

## Significance

This talk will focus mainly on Qibocal and Qibolab as tools for calibrating and characterizing quantum devices.

## References

https://arxiv.org/pdf/2303.10397.pdf
https://arxiv.org/pdf/2308.06313.pdf

## Experiment context, if any

**Authors:** PASQUALE, Andrea (University of Milan); PEDICILLO, Edoardo; ROBBIATI, Matteo (Università degli Studi e INFN Milano (IT)); CARRAZZA, Stefano (CERN)

**Presenter:** PEDICILLO, Edoardo

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **58** Type: **Oral**

# Generalized Parton Distribution Functions via Quantum Computing

Quantum simulation of quantum field theories offers a new way to investigate properties of the fundamental constituents of matter. We develop quantum simulation algorithms based on the light-front formulation of relativistic field theories. The process of quantizing the system in light-cone coordinates will be explained for a Hamiltonian formulation, which becomes block diagonal, each block approximating the Fock space with a certain harmonic resolution K. We analyze a QCD theory in 2+1D. We compute the analogue of parton distribution functions, the generalized parton distribution functions for hadrons in these theories. In particular, we look at the generalized parton distribution functions for a $\pi^0$ meson as well as a baryon in a quark-diquark model.

## Significance

This work builds upon the groundwork produced by the Tufts quantum information group to develop a formalism to simulate quantum field theories on a quantum computer via light cone coordinates. This work to be presented is the first attempt at calculating GPDs on a quantum computer. GPDs are important non-perturbative distribution functions that relay information about the 3D structure of hadrons.

## References

https://arxiv.org/abs/2211.07826

## Experiment context, if any

**Author:** GUSTIN, Carter M. (Tufts University)

**Co-author:** GOLDSTEIN, Gary R

**Presenter:** GUSTIN, Carter M. (Tufts University)

**Session Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

**Track Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

Contribution ID: **59** Type: **Oral**

# quantum GAN for fast shower simulation

*Monday, March 11, 2024 2:50 PM (20 minutes)*

High-energy physics relies on large and accurate samples of simulated events, but generating these samples with GEANT4 is CPU intensive. The ATLAS experiment has employed generative adversarial networks (GANs) for fast shower simulation, which is an important approach to solving the problem. Quantum GANs, leveraging the advantages of quantum computing, have the potential to outperform standard GANs.

Considering the limitations of the current quantum hardware, we conducted preliminary studies utilizing a hybrid quantum-classical GAN model to produce downsampled 1D(8 pixels) and 2D(64 pixels) calorimeter average shower shapes on quantum simulators. The impact of quantum noise is also investigated on the noisy simulator, and the performance is checked on the real quantum hardware.

After producing the average shower shape, we implemented a new generator model to produce the actual shower image with event fluctuation.

## References

https://indico.ihep.ac.cn/event/19316/contributions/143669/

## Experiment context, if any

no specific experiment.

## Significance

Concerning the study of average shower shape generation, we have fixed the training instability shown in this study (https://ceur-ws.org/Vol-3041/363-368-paper-67.pdf).

Concerning the study of actual shower image generation, we have improved the pixel energy distribution compared to this study (https://doi.org/10.22323/1.449.0573).

**Authors:** Prof. LI, Weidong (IHEP); Dr HUANG, Xiaozhong (IHEP)

**Presenter:** Dr HUANG, Xiaozhong (IHEP)

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **61** Type: **Poster**

# Enabling Computing Resources to Support Grid Jobs and Cluster Jobs Simultaneously

*Monday, March 11, 2024 4:15 PM (30 minutes)*

The Institute of High Energy Physics' computing platform includes isolated grid sites and local clusters. Grid sites manage grid jobs from international experiments, including ATLAS, CMS, LHCb, BELLEII, JUNO, while the local cluster concurrently processes data from experiments leading by IHEP like BES, JUNO, LHAASO. These resources have distinct configurations, such as network segments, file systems, and user namespaces etc.
The local cluster operates at a high job slot utilization rate, exceeding 95%, and still with the significant queuing. In contrast, grid site utilization is below 80%. To optimize resource use, we developed a model enabling worker nodes to handle both grid and local cluster jobs.
This involves preconfiguring the local cluster with container technology and initiating the local cluster's startd on grid nodes through glidein. Dynamically monitoring the grid site job queue, we schedule suitable local cluster jobs to idle grid job slots. This flexible model efficiently provides additional computing resources for experiments

## Significance

## References

## Experiment context, if any

**Authors:** SHIJY, 石京燕; WANG, lei (Institute of High Energy Physics)

**Co-authors:** Mr JIANG, Xiaowei (IHEP（中国科学院高能物理研究所）); Mr GUO, Chaoqi (IHEP); ZHENG, Wei (IHEP); YAN, Xiaofei (Institute of High Energy Physics); Ms YAO, Qiuling

**Presenter:** WANG, lei (Institute of High Energy Physics)

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **63** Type: **Oral**

# CaloDiT: Diffusion with transformers for fast shower simulation

*Wednesday, March 13, 2024 5:10 PM (20 minutes)*

Recently, transformers have proven to be a generalised architecture for various data modalities, i.e., ranging from text (BERT, GPT3), time series (PatchTST) to images (ViT) and even a combination of them (Dall-E 2, OpenAI Whisper). Additionally, when given enough data, transformers can learn better representations than other deep learning models thanks to the absence of inductive bias, better modelling of long-range dependencies, and interpolation and extrapolation capabilities. On the other hand, diffusion models are the state-of-the-art approach for image generation, which still use conventional U-net models for generation, mostly consisting of convolution layers making little use of the advantages of transformers. While these models show good generation performance it lacks the generalisation capabilities obtained from the transformer model. Standard diffusion models with an Unet architecture have already proven to be able to generate calorimeter showers, while transformer-based models, like those based on a VQ-VAE architecture, also show promising results. A combination of a diffusion model with a transformer architecture should bridge the quality of the generation sample obtained from diffusion with the generalisation capabilities of the transformer architecture. In this paper, we propose CaloDiT, to model our problem as a diffusion process with transformer blocks. Furthermore, we show the ability of the model to generalise to different calorimeter geometries, bringing us closer to a foundation model for calorimeter shower generation.

## Significance

This contribution presents a novel transformer-based machine learning model for general fast shower simulation, fitting perfectly the focus theme of ACAT 2024 - the foundation models.

## References

## Experiment context, if any

**Authors:** DA COSTA CARDOSO, Renato Paulo (CERN); RAIKWAR, Piyush (CERN); ZABOROWSKA, Anna (CERN); SALAMANI, Dalila (CERN); JARUSKOVA, Kristina (CERN); Dr VALLECORSA, Sofia (CERN); YEO, Kyongmin (IBM Research); EKAMBARAM, Vijay (IBM Research); NGUYEN, Nam (IBM Research); KALAGNANAM, Jayant (IBM Research); SRIVATSA, Mudhakar (IBM Research)

**Presenter:** RAIKWAR, Piyush (CERN)

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **64** Type: **Oral**

# Bridging Worlds: Achieving Language Interoperability between Julia and Python in Scientific Computing

*Wednesday, March 13, 2024 3:30 PM (20 minutes)*

In the realm of scientific computing, both Julia and Python have established themselves as powerful tools. Within the context of High Energy Physics (HEP) data analysis, Python has been traditionally favored, yet there exists a compelling case for migrating legacy software to Julia.
This talk focuses on language interoperability, specifically exploring how Awkward Array data structures can seamlessly bridge the gap between Julia and Python. The talk offers insights into key considerations such as memory management, data buffer copies, and dependency handling. It delves into the performance enhancements achieved by invoking Julia from Python and vice versa, particularly for intensive array-oriented calculations involving large-scale, though not excessively dimensional, arrays of HEP data.
Join us for this talk to gain a deeper understanding of the advantages and challenges inherent in achieving interoperability between Julia and Python in the domain of scientific computing.

## Significance

## References

## Experiment context, if any

**Authors:** OSBORNE, Ianna (Princeton University); LING, Jerry  (Harvard University (US)); PIVARSKI, Jim (Princeton University)

**Presenter:** OSBORNE, Ianna (Princeton University)

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **66** Type: **Poster**

# Generative Modeling for Fast Shower Simulation

*Monday, March 11, 2024 4:15 PM (30 minutes)*

Here, we present deep generative models for the fast simulation of calorimeter shower events. Using a three-dimensional, cylindrical scoring mesh, a shower event is parameterized by the total energy deposited on each cell of the scoring mesh. Due to the three-dimensional geometry, to simulate a shower event, it is required to learn a complex probability distribution of $O(10^3) \sim O(10^4)$ dimensional random variable. It should be noted that, on top of the high dimensionality, the sparse nature of a shower event, where energy is randomly deposited only on less than than 20\% of the cells, and the intermittency, which requires to capture large magnitudes but low probability events, make the development of a fast shower simulator challenging. To overcome these challenges, we develop a deep-learning framework, which can facilitate a model development by combining different neural network architectures with a range of generative models, e.g., variational auto-encoder, generative adversarial network, Wasserstein generative adversarial network, and denoising diffusion probabilistic model. In this study, we compare various generative models for the shower simulation. Also, the effects of the data scaling and regularizations are discussed. A distributed computing strategy for the deep generative model is discussed to develop a foundation model for the fast shower simulation.

## Significance

In this study, we aim to build a foundation model for the fast shower simulation. To achieve the goal, we first develop a deep learning framework to build a range of deep generative models to learn the high-dimensional probability density function and compare their strengths and weaknesses for the fast shower simulation. We also discuss a parallel computing strategy to digest the large data set to build a foundation model.

## References

## Experiment context, if any

**Authors:** YEO, Kyongmin (IBM Research); EKAMBARAM, Vijay (IBM Research); NGUYEN, Nam (IBM Research); RAIKWAR, Piyush (CERN); DA COSTA CARDOSO, Renato Paulo (CERN); ZABOROWSKA, Anna (CERN); SALAMANI, Dalila (CERN); JARUSKOVA, Kristina (CERN); Dr VALLECORSA, Sofia (CERN); KALAGNANAM, Jayant (IBM Research)

**Presenter:** YEO, Kyongmin (IBM Research)

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **67** Type: **Poster**

# AdaptivePerf: a portable, low-overhead, and comprehensive code profiler for single- and multi-threaded applications

*Wednesday, March 13, 2024 4:15 PM (30 minutes)*

Considering the slowdown of Moore's Law, an increasing need for energy efficiency, and larger and larger amounts of data to be processed in real-time and non-real-time in physics research, new computing paradigms are emerging in hardware and software. This requires bigger awareness from developers and researchers to squeeze as much performance as possible out of applications. However, this is a difficult task in case of complex and large codebases, where pinpointing a specific bottleneck is hard without profiling. At the same time, finding a suitable low-overhead and extensive profiler which works for a wide range of homogeneous and heterogeneous architectures proves to be challenging, especially when multi-threaded application support, user-friendliness, and reliability are important.

To address this, we introduce AdaptivePerf, developed in the context of the SYCLOPS EU project: an open-source comprehensive profiling tool based on sampling and system call tracing through patched Linux perf. It improves shortcomings sometimes observed in Linux perf such as broken stacks, no off-CPU data, and reports unclear to developers. The tool also further builds on its functionality by profiling how threads and processes are spawned within a program and what code parts are most time-consuming on- and off-CPU for each thread/process. Additionally, AdaptivePerf presents results in a user-friendly way by showing an interactive timeline with stack traces of functions starting new threads/processes and the browser of per-thread/per-process non-time-ordered flame graphs and time-ordered flame charts.

AdaptivePerf is designed with architecture portability in mind: the main features are based on hardware-independent metrics like wall time, while hardware-specific extensions such as profiling non-CPU devices (GPUs, FPGAs etc.) or capturing CPU-dependent perf events are possible. Thanks to our tool, detecting bottlenecks and addressing them within software and/or hardware is significantly easier for developers and researchers.

In this presentation, we will discuss the status of the profiler and its future prospects.

## Significance

## References

## Experiment context, if any

**Author:** GRACZYK, Maksymilian (CERN)

**Co-author:** ROISER, Stefan (CERN)

**Presenter:** GRACZYK, Maksymilian (CERN)

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **68** Type: **Poster**

# Preservation of the Direct Photons and Neutral Pions Analysis in the PHENIX Experiment at RHIC

*Wednesday, March 13, 2024 4:15 PM (30 minutes)*

The PHENIX Collaboration has actively pursued a Data and Analysis Preservation program since 2019, the first such dedicated effort at RHIC. A particularly challenging aspect of this endeavor is preservation of complex physics analyses, selected for their scientific importance and the value of the specific techniques developed as a part of the research. For this, we have chosen one of the most impactful PHENIX results, the joint study of direct photons and neutral pions in high-energy d+Au collisions. To ensure reproducibility of this analysis going forward, we partitioned it into self-contained tasks and used a combination of containerization techniques, code management, and robust documentation. We then leveraged REANA (the platform for reproducible analysis developed at CERN) to run the required software. We present our experience based on this example, and outline our future plans for analysis preservation.

## References

## Experiment context, if any

Heavy-Ion experiment at RHIC, leveraging the Electromagnetic Calorimeter capabilities.

## Significance

The Data and Analysis Preservation effort in the PHENIX Experiment at RHIC has expanded in the past year, with many additional elements of the direct photon and neutral pion analysis added to the preservation framework. This is the first such effort at RHIC and experience gained in this process will be useful for the Nuclear Physics community.

**Authors:** SMIRNOV, Dmitri (BNL); DAVID, Gabor (Stony Brook University); POTEKHIN, Maxim (Brookhaven National Laboratory (US))

**Presenters:** DAVID, Gabor (Stony Brook University); POTEKHIN, Maxim (Brookhaven National Laboratory (US))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **69** Type: **Poster**

# Quasi interactive analysis of High Energy Physics big data with high throughput

*Wednesday, March 13, 2024 4:15 PM (30 minutes)*

The need to interject, process and analyze large datasets in an as-short-as-possible amount of time is typical of big data use cases. The data analysis in High Energy Physics at CERN in particular will require, ahead of the next phase of high-luminosity at LHC, access to big amounts of data (order of 100 PB/year). However, thanks to continuous developments on resource handling and software, it is possible to offer users a more flexible and dynamic data access as well as access to open-source industry standards like Jupyter, Dask and HTCondor. This paves the way for innovative approaches: from a batch-based approach to an interactive high throughput platform, based on a parallel and geographically distributed back-end and leveraging the "High-Performance Computing, Big Data e Quantum Computing Research Centre"Italian National Center (ICSC) DataLake model.
This contribution will report the effort of porting multiple data analysis applications - from different collaborations and covering a wide range of physics processes - from a legacy approach to an interactive approach based on declarative solutions, like ROOT RDataFrame. These applications are then executed on the above-mentioned cloud infrastructure, splitting the workflow on multiple worker nodes and outputting the results on a single interface. A performance evaluation, therefore, will also be provided: tentative metrics will be identified, with speed-up benchmarks by upscaling to distributed resources. This will allow to find bottlenecks and/or drawbacks of the proposed high-throughput interactive approach, and eventually help developers committed to its deployment in the Italian National Center.

## Significance

This presentation inherits the state of the art of the INFN high throughput infrastructure (presented in the past, although in a more experiment-specific context, see References [1]), but then will cover the novel upscaling on a national level (within the Spoke-2 of the ICSC Italian National Center: https://www.supercomputing-icsc.it/en/spoke-2-fundamental-research-space-economy-en/), benchmarking multiple physics applications which cover different HEP experiments and different demands in terms of computing resources.

## References

[1] https://indico.jlab.org/event/459/contributions/11593/ (CHEP 2023)
What is ICSC: https://indico.jlab.org/event/459/contributions/11805/ (CHEP 2023)

## Experiment context, if any

Not a specific experiment per se, the physics applications considered are coming from different collaborations (e.g. CMS, ATLAS, FCC,...)

**Authors:** TARASIO, Alessandro (Universita della Calabria e INFN (IT)); CAGNOTTA, Antimo (Universita Federico II e INFN Sezione di Napoli (IT)); SPISSO, Bernardino (Universita Federico II e INFN Sezione di Napoli (IT)); SIMONE, Federica Maria (Universita e INFN, Bari (IT)); GRAVILI, Francesco Giuseppe (INFN Lecce e Universita del Salento (IT)); SABELLA, Gianluca; BARTOLINI, Matteo (Universita e INFN, Firenze (IT)); ANWAR, Muhammad Numan (Universita e INFN, Bari (IT)); MASTRANDREA, Paolo (Universita & INFN Pisa (IT)); DIOTALEVI, Tommaso (Universita e INFN, Bologna (IT)); TEDESCHI, Tommaso (Universita e INFN, Perugia (IT))

**Presenter:** TEDESCHI, Tommaso (Universita e INFN, Perugia (IT))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **71** Type: **Poster**

# Declarative paradigms for analysis description and implementation

*Wednesday, March 13, 2024 4:15 PM (30 minutes)*

The software toolbox used for "big data" analysis in the last few years is changing fast. The adoption of approaches able to exploit the new hardware architectures plays a pivotal role in boosting data processing speed, resources optimisation, analysis portability and analysis preservation.
The scientific collaborations in the field of High Energy Physics (e.g. the LHC experiments, the next-generation neutrino experiments, and many more) devote increasing resources to the development and implementation of bleeding-edge software technologies, pushing the reach of the single experiment and the whole HEP community.

The introduction of declarative paradigms in the analysis description and implementation is gaining interest and support in the main collaborations. This approach can simplify and speed-up the analysis description phase, support the portability of an analysis among different datasets/experiments, and strengthen the preservation of the results.
Furthermore, this approach - providing a deep decoupling between the analysis algorithm and back-end implementation - is a key element for present and future processing speed.

In the panorama of the approaches currently under study, an activity is ongoing in the ICSC (Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing, Italy) which focuses on the development of a framework characterized by the use of a declarative paradigm for the analysis description and the ability to operate on datasets from different experiments.
Using as a building base for a demonstrator the NAIL (Natural Analysis Implementation Language) Python package (developed in the context of the CMS data analysis for the event processing), the activity focuses both on the development of a general and effective interface able to support the data format of different experiments, and on the extension of the declarative approach to the full analysis chain.

## Significance

The application of declarative paradigms to data analysis has been an active field of development in the last decade. The presented developments focus on aspects not yet fully explored by previous studies/demonstrators: configurable input interface (i.e. access to multiple experiments) and the definition of the full analysis chain (not only the "event-loop").

## References

https://indico.cern.ch/event/769263/contributions/3413006/attachments/1840145/3016759/NAIL_Project_Natural_Analysis

## Experiment context, if any

Authors: CMS, ATLAS - target application: HEP main collaborations (e.g. LHC experiments)

**Authors:** ANNOVI, Alberto (INFN Sezione di Pisa); RIZZI, Andrea (Universita & INFN Pisa (IT)); VASELLI, Francesco (Scuola Normale Superiore & INFN Pisa (IT)); MASTRANDREA, Paolo (Universita & INFN Pisa (IT)); Dr BOCCALI, Tommaso (INFN Sezione di Pisa)

**Presenter:** VASELLI, Francesco (Scuola Normale Superiore & INFN Pisa (IT))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **72** Type: **Poster**

# LHC beam monitoring via real-time hit reconstruction in the LHCb VELO pixel detector

*Wednesday, March 13, 2024 4:15 PM (30 minutes)*

The increasing computing power and bandwidth of programmable digital devices opens new possibilities in the field of real-time processing of HEP data. LHCb is exploiting this technology advancements in various ways to enhance its capability for complex data reconstruction in real time. Amongst them is the real-time reconstruction of hits in the VELO pixel detector, by means of cluster-finding "on-the-fly" embedded in the readout board firmware. This reconstruction, in addition to savings of DAQ bandwidth and HLT computing resources, also enables further useful applications in precision monitoring and diagnostics of LHC beam conditions. In fact, clusters of pixels, while being more reliable and robust indications of physical particle hits than raw pixel counts, are exempt from the complications associated to the reconstruction of tracks, that involve alignment issues and are sensitive to multi-layer efficiency products. In this talk, we describe the design and implementation of a flexible system embedded in the readout firmware of the VELO detector, allowing real-time counting of cluster density in many parts of the detector simultaneously, and separately for every bunch ID, for every single LHC collision, without any slowdown of data acquisition. Quantitative applications of this system to luminosity measurement and beam monitoring are demonstrated.

## Significance

This is the first public presentation of a successful implementation of on-the-fly hit-statistics evaluation, transparently embedded in the readout in a complex detector at the full LHC average collision rate of 30 MHz. It also demonstrates its practical application to perform useful functions never before achieved at this high data rate.

## References

This work is a spinoff of the following published work:
G. Bassi et al., "A FPGA-Based Architecture for Real-Time Cluster Finding in the LHCb Silicon Pixel Detector", IEEE Trans. Nucl. Sci. 70 (2023) 1189, arXiv:2302.03972

## Experiment context, if any

LHCb (Real Time Analysis project)

**Authors:** PASSARO, Daniele (SNS & INFN Pisa (IT)); LAZZARI, Federico (Universita di Pisa & INFN Pisa (IT)); PUNZI, Giovanni (Universita & INFN Pisa (IT)); CORDOVA, Giulio (Universita & INFN Pisa (IT)); BASSI, Giovanni (SNS & INFN Pisa (IT)); MORELLO, Michael J. (SNS and INFN-Pisa (IT)); GRAVERINI, Elena (EPFL - Ecole Polytechnique Federale Lausanne (CH))

**Presenter:** PASSARO, Daniele (SNS & INFN Pisa (IT))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **73** Type: **Poster**

# Is Quantum Computing energy efficient? An Investigation on a quantum annealer.

*Thursday, March 14, 2024 4:10 PM (30 minutes)*

The environmental impact of computing activities is starting to be acknowledged as relevant and several scientific initiatives and research lines are gathering momentum in the scientific community to identify and curb it. Governments, industries, and commercial businesses are now holding high expectations for quantum technologies as they have the potential to create greener and faster methods for information processing. The energy perspective of such technologies, however, has remained rather outside the scopes of current deployment strategies, which might limit future adoptions.
In order to shed some light upon the interplay between classical/quantum computing and energy efficiency, we perform a comparison between these two paradigms over selected benchmark activities. In particular, we will compare traditional HPC technologies with the D-Wave Advantage quantum annealer and analyze the outcome of the experiment.

## Significance

## References

## Experiment context, if any

**Authors:** MINARINI, Francesco; Dr BIANCO, Gianluca (Universita e INFN, Bologna (IT)); GASPERINI, Simone (Universita e INFN, Bologna (IT))

**Presenters:** MINARINI, Francesco; Dr BIANCO, Gianluca (Universita e INFN, Bologna (IT)); GASPERINI, Simone (Universita e INFN, Bologna (IT))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **74** Type: **Oral**

# HPC Friendly HEP data model and RNTuple in HEP-CCE

*Monday, March 11, 2024 5:50 PM (20 minutes)*

As the role of High Performance Computers (HPC) increases in the High Energy Physics (HEP) experiments, the experiments will have to adopt HPC friendly storage format and data models to efficiently utilize these resources. In its first phase, the HEP-Center for Computational Excellence (HEP-CCE) has demonstrated that the complex HEP data products can be stored in the HPC native storage backends, such as HDF5, after converting them into byte stream serialization buffers. To efficiently leverage the HPC resources including compute accelerators such as GPUs, the storage format has to allow efficient I/O on parallel file systems used on HPC and the data models have to be capable of being offloaded to the GPUs for processing without conversions. In its second phase, HEP-CCE is studying the design and development of the HEP data models that will be HPC friendly and relevant for the future HEP experiments. At the same time, ROOT, an open data analysis framework, widely used by the HEP community, has been developing a new I/O subsystem called ROOT::RNTuple. RNTuple optimizes performance and minimizes storage, which requires a more streamlined design than the current I/O subsystem (ROOT::TTree) and hence has limited support on data model complexity. When designing data models suitable for offloading to compute accelerators, we also consider their storage in both HPC native backends (such as HDF5) and the more typical HEP persistence in ROOT::RNTuple. Both offloading and storage technologies have different restrictions to construct HEP data models. Only those data models that can take these restrictions into account can be truly HPC friendly and fulfill the requirements of future HEP experiments (including processing using grid resources). In this paper, we will show our results and ongoing works related to data model design and persistence of future HEP experimental data.

## Significance

Implementation and scaling test of I/O of HEP data in HPC friendly storage like HDF5.
Design of HEP data models that are HPC friendly, investigation of persistence in both HPC friendly format and RNTuple

## References

https://indico.jlab.org/event/459/contributions/11807/attachments/9286/13474/CHEP2023%20Parallel%20IO.pdf
Amit Bashyal et. al., “Data Storage for HEP Experiments in the Era of High-Performance Computing”, 2022 Snowmass Summer Study, arXiv:2203.07885.

## Experiment context, if any

Study targeted for HL-LHC and DUNE era experiments where the role of HPCs will further grow.

**Author:** BASHYAL, Amit

**Co-authors:** KNOEPFEL, Kyle (Fermi National Accelerator Laboratory); BHATTACHARYA, Meghna (Fermilab); VAN GEMMEREN, Peter (Argonne National Laboratory (US)); SEHRISH, Saba (Fermilab)

**Presenter:** BASHYAL, Amit

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **75** Type: **Oral**

# Real-time track reconstruction with FPGAs in the LHCb Scintillating Fibre Tracker beyond Run 3

*Thursday, March 14, 2024 6:10 PM (20 minutes)*

Finding track segments downstream of the magnet is an important and computationally expensive task, that LHCb has recently ported to the first stage of its new GPU-based trigger of the LHCb Upgrade I. These segments are essential to form all good physics tracks with precision momentum measurement, when combined with those reconstructed in the vertex track detector, and to reconstruct long-lived particles, such as K-short and strange baryons, decaying after the vertex track detector.
LHCb is currently developing a project for a new real-time tracking device based on distributed system of FPGAs, dedicated to the reconstruction of track primitives in the forward Scintillating Fibre tracker detector at the full LHC collision rate. The aim is to accelerate reconstruction in Run 4, and to develop this new technology in view of the higher instantaneous luminosity conditions foreseen for Run 5 (Upgrade II). In this talk we report the first detailed study of the reconstruction performance expected from this device, based on an accurate simulation of its architecture at the bit level.

## Significance

This is the first complete public report of the performance expected from a new ambitious project of FPGA-based real-time reconstruction aimed at LHCb Upgrade-2, that is obtained from a realistic, detailed simulation of the envisioned device at the bit-level. It is a significant advancement over the initial study on the subject, that was based on a behavioral simulation (see ref. below), and in its current form is part of an official LHCb enhancement proposal, due for submission to the LHCC committee in Feb.2024. The device object of this study is the first tracking device envisioned to have a native throughtput matching the full event rate of LHC collision (averaging 30 MHz) without any time-multiplexing.

## References

An initial study on this subject was presented at ACAT19:
https://arxiv.org/abs/2006.11067
A preliminary, incomplete version of the current work was shown at Connecting the Dots 2023:
https://indico.cern.ch/event/1252748/contributions/5521497/
while the current talk covers the final, full result.

## Experiment context, if any

LHCb (Real Time Analysis project)

**Authors:** XU, Ao (Universita & INFN Pisa (IT)); LAZZARI, Federico (Universita di Pisa & INFN Pisa (IT)); TERZUOLI, Francesco (Università di Siena & INFN Pisa (IT)); PUNZI, Giovanni (Univer-

sita & INFN Pisa (IT)); TUCI, Giulia (Heidelberg University (DE)); ZHUO, Jiahui (Univ. of Valencia and CSIC (ES)); PICA, Lorenzo (SNS & INFN Pisa (IT)); MARTINELLI, Maurizio (Universita & INFN, Milano-Bicocca (IT)); MORELLO, Michael J. (SNS and INFN-Pisa (IT))

**Co-authors:** CONTU, Andrea (INFN); DE OYANGUREN CAMPOS, Arantza (Univ. of Valencia and CSIC (ES)); Dr JASHAL, Brij Kishor (IFIC, Univ. of Valencia, CSIC (ES) and Tata Institute of Fundamental Research (TIFR)); MENDOZA, Diego (Instituto de Física Corpuscular - CERN); BASSI, Giovanni (SNS & INFN Pisa (IT)); HE, Jibo (University of Chinese Academy of Sciences (CN)); SHI, Qi (University of Chinese Academy of Sciences (CN)); FANTECHI, Riccardo (Universita & INFN Pisa (IT)); KOTRIAKHOVA, Sofia (Universita e INFN, Ferrara (IT)); BALDINI, Wander (Universita e INFN, Ferrara (IT))

**Presenter:** XU, Ao (Universita & INFN Pisa (IT))

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **76** Type: **Oral**

# ACTS as a Service

*Thursday, March 14, 2024 2:30 PM (20 minutes)*

Recent advancements in track finding within the challenging environments expected in the High-Luminosity Large Hadron Collider (HL-LHC) have showcased the potential of Graph Neural Network (GNN)-based algorithms. These algorithms exhibit high track efficiency and reasonable resolutions, yet their computational burden on CPUs hinders real-time processing, necessitating the integration of accelerators like GPUs. However, the substantial size of the involved graphs, with approximately 300k nodes and 1M edges, demands significant GPU memory, posing a challenge for facilities lacking high-end GPUs such as NVIDIA A100s or V100s. These computing challenges must be addressed to deploy GNN-based track finding or any algorithm that requires coprocessors, into production.

To overcome these challenges, we propose the as-a-service approach to deploy the GNN-based track-finding algorithm in the cloud or high-performance computing centers such as the NERSC Perlmutter system with over 7000 A100 GPUs. In addressing this, we have developed a tracking-as-a-service prototype within A Common Tracking Software (ACTS), an experiment-independent toolkit for charged particle track reconstruction.

The GNN-based track finding is implemented as a service within ACTS, showcasing its versatility as a demonstrator. Moreover, this approach is algorithm-agnostic, allowing the incorporation of other algorithms as new backends through interactions with the client interface implemented in ACTS.

In this contribution, we present the implementation of the GNN-based track-finding workflow as a service using the Nvidia Triton Inference Server within ACTS. The GNN pipeline comprises three distinct deep-learning models and two CUDA-based algorithms, enabling full tracking reconstruction within ACTS. We explore different server configurations to assess track-finding throughput and GPU utilization, exploring the scalability of the inference server across the NERSC Perlmutter supercomputer and cloud resources such as AWS and Google Cloud.

## Significance

This contribution describes the work that uses the as-a-service computing model to accelerate track finding in dense environments for HL-LHC. The as-a-service method provides more flexibility to scale and balance computing resources using coprocessors for state-of-art tracking reconstruction.

Furthermore, this approach is independent of the underlying tracking algorithm. The GNN-based tracking finding is implemented in ACTS to provide the first demonstrator for such an approach to showcase the potential and explore scalability using supercomputers

## References

## Experiment context, if any

**Authors:** ZHAO, Haoran (University of Washington (US)); JU, Xiangyang (Lawrence Berkeley National Lab. (US)); YAO, Yao (Purdue University (US)); FENG, Yongbin (Fermi National Accelerator Lab. (US)); CHOU, Yuan-Tang (University of Washington (US))

**Co-authors:** NAYLOR, Andrew (Lawrence Berkeley National Lab); RANKIN, Dylan Sheldon (University of Pennsylvania (US)); KHODA, Elham E (University of Washington (US)); PEDRO, Kevin (Fermi National Accelerator Lab. (US)); CALAFIURA, Paolo (Lawrence Berkeley National Lab. (US)); HARRIS, Philip Coleman (Massachusetts Inst. of Technology (US)); HSU, Shih-Chieh (University of Washington Seattle (US)); MCCORMACK, William Patrick (Massachusetts Inst. of Technology (US))

**Presenter:** CHOU, Yuan-Tang (University of Washington (US))

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **77** Type: **Oral**

# Optimizing ANN-Based Triggering for BSM events with Knowledge Distillation

*Tuesday, March 12, 2024 11:30 AM (20 minutes)*

In recent years, the scope of applications for Machine Learning, particularly Artificial Neural Network algorithms, has experienced an exponential expansion. This surge in versatility has uncovered new and promising avenues for enhancing data analysis in experiments conducted at the Large Hadron Collider at CERN. The integration of these advanced techniques has demonstrated considerable potential for elevating the efficiency and efficacy of data processing in this experimental setting.
Nevertheless, one frequently overlooked aspect of utilizing Artificial Neural Networks (ANNs) revolves around the imperative of efficiently processing data for online applications. This becomes particularly crucial when exploring innovative methods for selecting intriguing events at the trigger level, as seen in the pursuit of Beyond Standard Model (BSM) events. The study delves into the potential of Autoencoders (AEs), an unbiased algorithm capable of event selection based on abnormality without relying on theoretical priors. However, the distinctive latency and energy constraints within the Level-1 Trigger domain necessitate tailored software development and deployment strategies. These strategies aim to optimize the utilization of on-site hardware, with a specific focus on Field-Programmable Gate Arrays (FPGAs).
This is why a technique called Knowledge Distillation (KD) is studied in this work. It consists in using a large and well trained "teacher", like the aforementioned AE, to train a much smaller student model which can be easily implemented on an FPGA. The optimization of this distillation process involves exploring different aspects, such as the architecture of the student and the quantization of weights and biases, with a strategic approach that includes hyperparameter searches to find the best compromise between accuracy, latency and hardware footprint.
The strategy followed to distill the teacher model will be presented, together with consideration on the difference in performance of applying the quantization before or after the best student model has been found. Finally, a second way to perform KD will be introduced called co-training distillation which sees the teacher and the student models trained at the same time.

## Significance

## References

## Experiment context, if any

CMS experiment

**Author:** LORUSSO, Marco (Universita Di Bologna (IT))

**Presenter:** LORUSSO, Marco (Universita Di Bologna (IT))

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **78** Type: **Oral**

# Fair Universe: HiggsML Uncertainty Challenge

*Tuesday, March 12, 2024 12:50 PM (20 minutes)*

The Fair Universe project is building a large-compute-scale AI ecosystem for sharing datasets, training large models and hosting challenges and benchmarks. Furthermore, the project is exploiting this ecosystem for an AI challenge series focused on minimizing the effects of systematic uncertainties in High-Energy Physics (HEP), and on predicting accurate confidence intervals. This talk will describe the challenge platform we have developed that builds on the open-source benchmark ecosystem Codabench to interface it to the NERSC HPC center and its Perlmutter system with over 7000 A100 GPUs. This presentation will also launch the first of our Fair Universe public challenges hosted on this platform, the Fair Universe: HiggsML Uncertainty Challenge, the a pilot phase of which will be run concurrently with ACAT so that attendees will be able to enter the competition; interact with organizers; and have their uncertainty-aware ML methods evaluated on large datasets.

This challenge will present participants with a much larger training dataset than previous competitions corresponding measurement of the Higgs decay to tau leptons at the Large Hadron Collider. They should design an advanced analysis technique able to not just measure the signal strength but also to provide a confidence interval, from which correct coverage will be evaluated automatically from pseudo-experiments. The confidence interval should include statistical uncertainty and also systematic uncertainties (including, for example, detector calibration,background levels among others). It is expected that advanced analysis techniques that are able to control the impact of systematics will perform best, thereby pushing the field of uncertainty aware AI techniques for HEP and beyond.

The Codabench/NERSC platform also allows for hosting challenges from other communities, and we also intend to make our benchmark designs available as templates so similar efforts can be easily launched in other domains.

## Significance

This contribution describes work that pushes the state of the art in the ML challenge platform; the ML challenge itself; and in the evaluation of uncertainty-aware methods.
For the platform we describe a system capable of operating at much larger scale than other approaches, including on large datasets and trained and evaluated on multiple GPUs in parallel. The platform also provides a leaderboard and ecosystem for long-lived benchmarks, as well as capabilities to not only evaluate different models but also test models against new datasets.
For the "Fair Universe: HiggsML Uncertainty Challenge"we provide larger datasets, with multiple systematic uncertainties applied, as well as evaluation of uncertainties as part of the challenge, performed on multiple pseudo-experiments. All of these aspects are novel to HEP ML challenges as far as we are aware.
Furthermore we will present methodological innovations including novel metrics for evaluation of uncertainty aware methods as well as improvements in uncertainty aware methods themselves.

## References

## Experiment context, if any

**Authors:** GHOSH, Aishik (University of California Irvine (US)); NACHMAN, Ben (Lawrence Berkeley National Lab. (US)); HARRIS, Chris (NERSC, Lawrence Berkeley National Laboratory); WHITESON, Daniel (University of California Irvine (US)); ROUSSEAU, David (IJCLab-Orsay); KHODA, Elham E (University of Washington (US)); ULLAH, Ihsan (ChaLearn); GUYON, Isabelle (ChaLearn/Google); CALAFIURA, Paolo (Lawrence Berkeley National Lab. (US)); NUGENT, Peter (Lawrence Berkeley National Laboratory); CHAKKAPPAI, Ragansu (Université Paris-Saclay (FR)); DIEFENBACHER, Sascha (Lawrence Berkeley National Lab. (US)); HSU, Shih-Chieh (University of Washington Seattle (US)); FARRELL, Steven (Lawrence Berkeley National Laboratory); BHIMJI, Wahid; CHOU, Yuan-Tang (University of Washington (US))

**Presenter:** BHIMJI, Wahid

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **79** Type: **Oral**

# Generic representations of jets at detector-level with self-supervised learning

*Wednesday, March 13, 2024 3:50 PM (20 minutes)*

Supervised learning has been used successfully for jet classification and to predict a range of jet properties, such as mass and energy. Each model learns to encode jet features, resulting in a representation that is tailored to its specific task. But could the common elements underlying such tasks be combined in a single foundation model to extract features generically? To address this question, we explore self-supervised learning (SSL), inspired by its applications in the domains of computer vision and natural language processing. Besides offering a simpler and more resource-effective route when learning multiple tasks, SSL can be trained on unlabeled data, e.g. large sets of collision data. We demonstrate that a jet representation obtained through SSL can be readily fine-tuned for downstream tasks of jet kinematics prediction and jet classification. Compared to existing studies in this direction, we use a realistic full-coverage calorimeter simulation, leading to results that more faithfully reflect the prospects at real collider experiments.

## Significance

Going beyond previous work, we present novel approaches to the training of our foundation model, with the aim of leveraging large unlabeled datasets (opening the door to novel data-driven analysis techniques) and learning transferable jet representations that are invariant to detector properties.

## References

- Presentation at the ML4Jets Workshop, Hamburg, 2023: https://indico.cern.ch/event/1253794/contributions/5588641/
- Paper presenting the collider detector simulation used in this work: "Configurable calorimeter simulation for AI applications"

## Experiment context, if any

**Authors:** KOBYLIANSKII, Dmitrii (Weizmann Institute of Science (IL)); GROSS, Eilam (Weizmann Institute of Science (IL)); DREYER, Etienne (Weizmann Institute of Science (IL)); MERZ, Garrett; CRANMER, Kyle Stuart (University of Wisconsin Madison (US)); SOYBELMAN, Nathalie (Weizmann Institute of Science (IL)); KAKATI, Nilotpal (Weizmann Institute of Science (IL)); RIECK, Patrick (New York University (US))

**Presenter:** RIECK, Patrick (New York University (US))

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **80** Type: **Oral**

# Boosting statistical anomaly detection via multiple test with NPLM

*Wednesday, March 13, 2024 3:10 PM (20 minutes)*

Statistical anomaly detection empowered by AI is a subject of growing interest at collider experiments, as it provides multidimensional and highly automatized solutions for signal-agnostic data quality monitoring, data validation and new physics searches.
AI-based anomaly detection techniques mainly rely on unsupervised or semi-supervised machine learning tasks. One of the most crucial and still unaddressed challenges of these applications is how to optimize the chances of detecting unexpected anomalies when prior knowledge about the nature of the latter is not available.
In this presentation we show how to exploit multiple tests to improve sensitivity to rare anomalies of different nature. We focus on a kernel methods based implementation of the NPLM algorithm, a signal-agnostic goodness of fit test based on a ML approximation of the likelihood ratio test [1, 2].
First, we show how performing multiple tests with different model configurations on the same data allows us to work around the problem of hyperparameters tuning, improving the algorithm' s chance of discovery at the same time. Second, we show how multiple samples of streamed data can be optimally exploited to increase sensitivity to rare signals.
The presented findings offer the ability to perform fast, efficient, and sensitivity-enhanced applications of the NPLM algorithm to a larger and potentially more inclusive set of data, both offline and quasi-online.
With low-dimensional problems, we show this tool acts as a powerful diagnostic and compression algorithm. Furthermore, we find the agnostic nature of the strategy becomes especially relevant when the input data representation results from unsupervised ML algorithms, whose response to anomalies cannot be predicted.

## Significance

The proposed strategies are new developments of the algorithm that have not been published yet. The tests carried out for this work show improved results over a set of benchmarks with respect to the previous implementation of the algorithm.

## References

Previous work related to the topic:
https://link.springer.com/article/10.1140/epjc/s10052-022-10830-y
https://arxiv.org/abs/2305.14137
https://iopscience.iop.org/article/10.1088/2632-2153/acebb7

## Experiment context, if any

CMS

**Authors:** Dr GROSSO, Gaia (IAIFI, MIT); Dr LETIZIA, Marco; HARRIS, Philip Coleman (Massachusetts Inst. of Technology (US))

**Presenter:** Dr GROSSO, Gaia (IAIFI, MIT)

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **81** Type: **Poster**

# Implementation of zero trust security strategy in HEPS scientific computing system

*Thursday, March 14, 2024 4:10 PM (30 minutes)*

Traditionally, data centers provide computing services to the outside world, and their security policies are usually separated from the outside world based on firewalls and other security defense boundaries. Users access the data center intranet through VPN, and individuals or endpoints connected through remote methods receive a higher level of trust to use computing services than individuals or endpoints outside the perimeter. But this approach to security design is never ideal. Zero Trust security is based on de-peripheralization and least-privilege access, which protects intranet assets and services from vulnerabilities inherent in the network perimeter and implicit trust architecture. In order to meet the diverse data analysis needs of light source users, the HEPS scientific computing system provides an interactive computing service model for external network users. Users can directly access intranet computing resources through web pages. In this service model, how do we refer to the zero-trust security idea? It has become very urgent to realize the minimum permission access between various services of the computing system and improve the security level of the system environment. Based on the zero-trust security strategy, this paper designs an inter-service communication mechanism based on user identity tokens. During the function call process between different services, service permissions are allocated based on the token user identity to achieve fine-grained management of service permissions and ensure the Cybersecurity of HEPS scientific computing systems.

## Significance

## References

## Experiment context, if any

**Authors:** Mr XU, Jiping (IHEP); HU, Qingbao (IHEP); CHENG, Yaosong (Institute of High Energy Physics Chinese Academy of Sciences, IHEP); LUO, qi (中科院高能物理所计算中心)

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

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **82** Type: **Poster**

# Design and Implementation of a Container-based Public Service Cloud Platform for HEPS

*Thursday, March 14, 2024 4:10 PM (30 minutes)*

High Energy Photon Source (HEPS) is a crucial scientific research facility that necessitates efficient, reliable, and secure services to support a wide range of experiments and applications. However, traditional physical server-based deployment methods suffer from issues such as low resource utilization, limited scalability, and high maintenance costs.Therefore, the objective of this study is to design and develop a container-based public service cloud platform that caters to the experimental and application needs of synchrotron radiation sources. By leveraging Kubernetes as the container orchestration technology, the platform achieves elastic scalability, multi-tenancy support, and dynamic resource allocation, thereby enhancing resource utilization and system scalability. Furthermore, incorporating robust security measures such as access control, authentication, and data encryption ensures the safety and integrity of users' applications and data.This research also focuses on the design, application, and deployment of Continuous Integration and Continuous Delivery (CICD). By implementing CICD workflows, the platform automates the build, testing, and deployment processes of applications, resulting in improved efficiency and quality throughout the development and deployment lifecycle.HEPS Container Public Service Cloud offers a comprehensive range of services including ganglia and nagios monitoring, puppet, cluster login nodes, nginx proxy, user service system, LADP and AD domain authentication nodes, KRB5 slave nodes, and more.The research findings demonstrate that the container-based public service cloud design and application deliver high-performance, stable, and secure services, effectively meeting the demands of synchrotron radiation source experiments and applications. Additionally, the utilization of CICD further enhances the efficiency and quality of development and deployment processes. Future work should focus on optimizing and expanding the capabilities of the container-based public service cloud to accommodate diverse user requirements and scenarios.

## Significance

## References

## Experiment context, if any

**Authors:** QI, Fazhi (IHEP, CAS); XU, Jiping (IHEP); HU, Qingbao (IHEP, CAS); ZHENG, Wei (IHEP)

**Presenter:** WANG, lei (Institute of High Energy Physics)

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **83** Type: **Oral**

# Reconstruction of atmospheric neutrinos and muons using Machine Learning-based methods in JUNO

*Monday, March 11, 2024 5:10 PM (20 minutes)*

The Jiangmen Underground Neutrino Observatory (JUNO), located in Southern China, is a multi-purpose neutrino experiment that consists of a 20-kton liquid scintillator detector. The primary goal of the experiment is to determine the neutrino mass ordering (NMO) and measure other neutrino oscillation parameters to sub-percent precision. Atmospheric neutrinos are sensitive to NMO via matter effects and can improve JUNO's total sensitivity in a joint analysis with reactor neutrinos; Atmospheric muons contribute to one of the most important background sources to neutrino signals. Good capability of reconstructing atmospheric neutrinos and muons in JUNO is crucial for its physics goal.

In this contribution, we present a novel multi-purpose reconstruction method for atmospheric neutrinos, muons and other physics events at similar energies (few GeV to tens of GeV) by combining PMT waveform analysis and machine learning techniques. Multiple machine learning approaches, including planer, spherical, and 3-dimensional models, as well as other novel techniques in improving reconstruction precision, are discussed and compared. We show the performance of reconstructing atmospheric neutrino's directionality and energy using Monte-Carlo simulations, and demonstrate that this method can achieve unprecedented reconstruction precision for multiple physics quantities and fulfils the needs of JUNO. This method also has the potential to be applied to other liquid scintillator detectors.

## Significance

## References

## Experiment context, if any

Jiangmen Underground Neutrino Observatory (JUNO)

**Author:** MA, Wing Yan (SDU)

**Co-authors:** ZENG, Fanrui (China, Shandong University); DUYANG, Hongyue (Shandong University); LIU, Jiaxi; Dr LI, Teng (Shandong University, CN); LUO, Wuming (Institute of High Energy Physics, Chinese Academy of Science); HE, Xinhai (The Institute of High Energy Physics of the Chinese Academy of Sciences)

**Presenter:** MA, Wing Yan (SDU)

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **84** Type: **Poster**

# ServiceX, the novel data delivery system, for physics analysis

*Wednesday, March 13, 2024 4:15 PM (30 minutes)*

Effective data extraction has been one of major challenges in physics analysis and will be more important in the High-Luminosity LHC era. ServiceX provides a novel data access and delivery by exploiting industry-driven software and recent high-energy physics software in the python ecosystem. An experiment-agnostic nature of ServiceX will be described by introducing various types of transformer containers that run on Kubernetes cluster. Latest updates in the backend will be also discussed. The newly designed python client library, communicates with REST API of ServiceX, will be introduced with practical use cases within physics analysis pipelines. The future of ServiceX also will be briefly described.

## References

https://iris-hep.org/projects/servicex.html

## Experiment context, if any

## Significance

ServiceX is now ready for users to come and try. Potential to change current and future physics analysis workflow. Possibility to extend its scope outside of ATLAS and CMS

**Authors:** GALEWSKY, Benjamin (Univ. Illinois at Urbana Champaign (US)); WATTS, Gordon (University of Washington (US)); VUKOTIC, Ilija (University of Chicago (US)); CHOI, Kyungeon (University of Texas at Austin (US)); ONYISI, Peter (University of Texas at Austin (US)); GARDNER JR, Robert William (University of Chicago (US))

**Presenter:** CHOI, Kyungeon (University of Texas at Austin (US))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **85** Type: **Oral**

# Key4hep

*Wednesday, March 13, 2024 3:50 PM (20 minutes)*

Detector studies for future experiments rely on advanced software tools to estimate performance and optimize their design and technology choices. Similarly, machine learning techniques require realistic data sets that allow estimating their performance beyond simplistic toy-models. The Key4hep software stack provides tools to perform detailed full simulation studies for a number of different detector models for future Higgs factories, including a palette of different detector technologies for tracking or calorimeter subsystems. The Key4hep stack includes generic tools for full and fast simulation, reconstruction, such as tracking and particle flow clustering, and analysis. The presentation will detail how Key4hep can be used for full simulation studies for generic or specific detector models and give examples of the available detector models and reconstruction tools, and how the results can be converted into formats most convenient for machine learning tools.

## Significance

There have been major improvements in the Key4hep stack. Full simulation based on DD4hep is available for more detector models for the FCC. Tracking and Particle flow reconstruction can be used beyond what was available previously. Several full simulation studies using Key4heo are now the way.

## References

ACAT 2022: https://indico.cern.ch/event/1106990/contributions/4991332/
CHEP 2023: https://indico.jlab.org/event/459/contributions/11535/

## Experiment context, if any

**Author:** SAILER, Andre (CERN)

**Co-authors:** TOLOSA-DELGADO, Alvaro (CERN); HEGNER, Benedikt (CERN); FRANCOIS, Brieuc (CERN); GAEDE, Frank-Dieter (Deutsches Elektronen-Synchrotron (DE)); GANIS, Gerardo (CERN); STEWART, Graeme A (CERN); Mr ZOU, Jiaheng; CARCELLER, Juan Miguel (CERN); SMIESKO, Juraj (CERN); REICHENBACH, Leonhard (University of Bonn (DE)); KO, Sang Hyun (Seoul National University (KR)); SASIKUMAR, Swathi (CERN); JOOSTEN, Sylvester; LIN, Tao (Chinese Academy of Sciences (CN)); Dr LI, Teng (Shandong University, CN); MADLENER, Thomas (Deutsches Elektronen-Synchrotron (DESY)); VOLKL, Valentin (CERN); LI, Weidong (IHEP); FANG, Wenxing; DECONINCK, Wouter; ZHANG, Xiaomei (Chinese Academy of Sciences (CN))

**Presenter:** HEGNER, Benedikt (CERN)

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **86** Type: **Oral**

# Track reconstruction for future colliders with quantum algorithms

*Monday, March 11, 2024 5:50 PM (20 minutes)*

Tracking is one of the most crucial components of reconstruction in the collider experiments. It is known for high consumption of computing resources, and various innovations have been being introduced until now. Future colliders such as the High-Luminosity Large Hadron Collider (HL-LHC) will face further enormously increasing demand of the computing resources. Usage of cutting-edge artificial intelligence will likely be the baseline at the HL-LHC, but the rapid development of quantum algorithms and hardware could bring in further paradigm-shifting improvement to this challenge. The track reconstruction can be considered as a quadratic unconstrained binary optimization (QUBO) problem. The Quantum Approximate Optimization Algorithm (QAOA) is one of the most promising algorithms to solve such combinatorial problems and to seek for a quantum advantage in the era of the Noisy Intermediate-Scale Quantum computers. It is found that the QAOA shows promising performance both in simulator and hardware from Origin Quantum. It demonstrated itself as one of the candidates for the track reconstruction using quantum computers. Ongoing studies with other quantum algorithms will also be presented.

## Significance

QAOA had not been successfully implemented in the track reconstruction in previous studies in our field. Another important implementation is a theoretically robust method considered for the first time regarding the sub-QUBO method (an approach to split the QUBO into small subsets to match with the available number of qubits). Other sub-QUBO methods (e.g. qbsolv used in D-Wave) are empirical and do not have a theoretical foundation to guarantee quasi-optimal solutions. Lastly, the work utilizes a quantum hardware from Origin Quantum, the first practical quantum computer in China.

## References

Previous results: https://arxiv.org/abs/2310.10255 (accepted as a peer-reviewed conference paper at IC2023, published through Springer CCIS)

## Experiment context, if any

This work uses a public dataset from the TrackML Challenge intended for the HL-LHC.

**Author:** OKAWA, Hideki (Chinese Academy of Sciences (CN))

**Presenter:** OKAWA, Hideki (Chinese Academy of Sciences (CN))

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **87** Type: **Oral**

# Towards the construction of Foundational Models at the LHC

*Tuesday, March 12, 2024 11:30 AM (20 minutes)*

The emergence of models pre-trained on simple tasks and then fine-tuned to solve many downstream tasks has become a mainstay for the application of deep learning within a large variety of domains. The models, often referred to as foundation models, aim, through self-supervision, to simplify complex tasks by extracting the most salient features of the data through a careful choice of pre-training strategy. When done effectively, these models can lead to world-leading algorithm performance on a large number of tasks. We present a re-simulation strategy for a model pretraining (R3SL) and show that this strategy applied to quark and gluon jets at the Large Hadron Collider can be used to create a foundation model for hadronically decaying objects. We show R3SL creates a feature space insensitive to the parton shower model uncertainties while retaining the core features of quark and gluon jets. On downstream tasks utilizing the pre-trained feature space, we demonstrate our method achieves comparable, if not better, tagging performance to established benchmarks for jet tagging in Higgs to bottom quarks, but with greatly reduced uncertainties. The algorithm presents a crucial step towards more robust searches for new physics involving particle jets and paves the way for the development of foundation models at the Large Hadron Collider.

## Significance

This study is one of the first examples of a foundation model for applications in high-energy physics.

## References

## Experiment context, if any

LHC

**Authors:** MAIER, Benedikt (KIT - Karlsruhe Institute of Technology (DE)); KRUPA, Jeffrey (Massachusetts Institute of Technology); PIERINI, Maurizio (CERN); KAGAN, Michael (SLAC National Accelerator Laboratory (US)); HARRIS, Philip Coleman (Massachusetts Inst. of Technology (US))

**Presenter:** HARRIS, Philip Coleman (Massachusetts Inst. of Technology (US))

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **88** Type: **Oral**

# Application of ACTS for gaseous tracking detectors

*Monday, March 11, 2024 2:30 PM (20 minutes)*

Based on the tracking experience at LHC, the project, A Common Tracking Software (ACTS), aims to provide an open-source experiment-independent and framework-independent software designed for modern computing architectures. It provides a set of high-level performant track reconstruction tools which are agnostic to the details of the detection technologies and magnetic field configuration, and tested for strict thread-safety to support multi-threaded event processing.

ACTS has been used as a tracking toolkit at experiments such as ATLAS, sPHENIX, FASER, ALICE etc. and has shown very promising tracking performance in terms of both physics performance and time performance. So far, the applications of ACTS are mainly focusing on silicon-based tracking systems. However, its application for gaseous tracking detectors, for example, drift chamber, is very limited. Therefore, an example gaseous tracking detector is still lacking in ACTS.

In this contribution, we will introduce the progress recently we have made in extending the current version of ACTS to support tracking with the gaseous detector like uRWell-based detector and drift chamber. The detailed implementation will be described. The application of this implementation to two future electron-positron collision experiments i.e. Circular Electron Positron Collider (CEPC) and Super Tau Charm Factory (STCF) will be presented. In addition, the efforts of adding an open-access drift chamber to the Open Data Detector will also be introduced to facilitate the development of common tracking algorithms in the future.

## Significance

Part of the research is already published here as listed in the “References”.

## References

https://iopscience.iop.org/article/10.1088/1748-0221/18/07/P07026
https://indico.cern.ch/event/1252748/contributions/5521504/
https://indico.cern.ch/event/1295479/contributions/5635040/

## Experiment context, if any

CEPC, STCF

**Authors:** AI, Xiaocong (Zhengzhou University); HUANG, Xingtao; Dr LI, Weidong (IHEP, Beijing); Dr LIN, Tao

**Presenter:** Dr LIN, Tao

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **90** Type: **Oral**

# 3-loop Feynman integrals in the Euclidean or physical kinematical region

*Thursday, March 14, 2024 3:30 PM (20 minutes)*

We recently explored methods for 2-loop Feynman integrals in the Euclidean or physical kinematical region, using numerical extrapolation and adaptive iterated integration. Our current goal is to address 3-loop two-point integrals with up to 6 internal lines.

Using double extrapolation, the integral $\mathcal{I}$ is approximated numerically by the limit of a sequence of integrals $\mathcal{I}(\varepsilon)$ as $\varepsilon \to 0$, where $\varepsilon$ enters in the space-time dimension $\nu = 4 - 2\varepsilon$. For a fixed value of $\varepsilon = \varepsilon_\ell$, the integral $\mathcal{I}(\varepsilon_\ell)$ is approximated by the limit of a sequence $I(\varepsilon_\ell, \varrho)$ as $\varrho \to 0$. Here, $\varrho$ enters in the modification of a factor $V$ to $V - i\varrho$ in the integrand denominator, applied since $V$ may vanish in the integration domain. Alternatively, we can integrate after expanding with respect to $\varepsilon$, followed by a single extrapolation in $\varrho$ only.

In this work, we will give an analysis with applications to sample diagrams.

## Significance

Accurate theoretical predictions are needed in view of improvements in the technology of high energy physics experiments. Higher-order corrections are required for accurate theoretical predictions of the cross-section for particle interactions. The Feynman diagrammatic approach is commonly used to address higher-order corrections. We use numerical integration and extrapolation methods to handle integrand singularities in Feynman loop integrals.

## References

Paper on 2-loop integrals in ACAT 2022: “Loop integral computation in the Euclidean or physical kinematical region using numerical integration and extrapolation”, E. de Doncker, F Yuasa, T. Ishikawa and K. Kato

## Experiment context, if any

**Authors:** Dr DE DONCKER, Elise (Western Michigan University); Dr YUASA, Fukuko (High Energy Accelerator Research Organization (KEK), Oho 1-1, Tsukuba, Ibaraki, 305-0801, Japan); Dr ISHIKAWA, Tadashi (High Energy Accelerator Research Organization (KEK), Oho 1-1, Tsukuba, Ibaraki, 305-0801, Japan); Dr KATO, K. (Department of Physics, Kogakuin University, Shinjuku, Tokyo 163-8677, Japan)

**Presenter:** Dr DE DONCKER, Elise (Western Michigan University)

**Session Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

**Track Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

Contribution ID: **91** Type: **Oral**

# Common Analysis Tools in CMS

*Thursday, March 14, 2024 3:10 PM (20 minutes)*

The CMS experiment has recently established a new Common Analysis Tools (CAT) group. The CAT group implements a forum for the discussion, dissemination, organization and development of analysis tools, broadly bridging the gap between the CMS data and simulation datasets and the publication-grade plots and results. In this talk we discuss some of the recent developments carried out in the group, including the structure of the group, the facilities and services provided, the communication channels, the ongoing developments in the context of frameworks for data processing, strategies for the management of analysis workflows and their preservation and tools for the statistical interpretation of analysis results.

## Significance

This is the first public presentation of the CMS Common Analysis Tools group

## References

## Experiment context, if any

CMS

**Author:** TEDESCHI, Tommaso (Universita e INFN, Perugia (IT))

**Presenter:** TEDESCHI, Tommaso (Universita e INFN, Perugia (IT))

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **93** Type: **Poster**

# Retrieval Augmented Generation for Particle Physics: A Case Study with the Snowmass White Papers and Reports

*Wednesday, March 13, 2024 4:15 PM (30 minutes)*

Particle physics faces many challenges and opportunities in the coming decades, as reflected by the Snowmass Community Planning Process, which produced about 650 reports on various topics. These reports are a valuable source of information, but they are also difficult to access and query. In this work, we explore the use of Large Language Models (LLMs) and Retrieval Augmented Generation (RAG) to answer questions based on the Snowmass corpus. RAG is a technique that combines LLMs with document retrieval, allowing the model to select relevant passages from the corpus and generate answers. We describe how we indexed the Snowmass reports for RAG, how we compared different LLMs for this task, and how we evaluated the quality and usefulness of the answers. We discuss the potential applications and limitations of this approach for particle physics and beyond.

## Significance

LLM's are new - and we are figuring out how to apply them in our field in ways that leverage their power. Search and reasoning over local document collections is one such approach.

This is new work, and hasn't been presented before.

## References

## Experiment context, if any

None, though both of us are doing this with IRIS-HEP in mind, which isn't actually an experiment…

**Authors:** GALEWSKY, Benjamin (Univ. Illinois at Urbana Champaign (US)); WATTS, Gordon (University of Washington (US))

**Presenter:** WATTS, Gordon (University of Washington (US))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **94** Type: **Oral**

# Beyond Language: Foundation Models for Collider Physics Data

*Tuesday, March 12, 2024 12:30 PM (20 minutes)*

Foundation models have revolutionized natural language processing, demonstrating exceptional capabilities in handling sequential data. Their ability to generalize across tasks and datasets offers promising applications in high energy physics (HEP). However, collider physics data, unlike language, involves both continuous and discrete data types, including four-vectors, particle IDs, charges, etc. Additionally, the particles are permutation invariant, which is fundamentally different from natural language. To address these challenges, we investigate various embedding schemes and techniques that introduce physical biases into the framework. Our findings provide valuable insights into the incorporation of foundation models into the HEP domain.

## Significance

Although foundation models are already widely used in NLP, there is still more research to be done on their application in HEP. Currently, the HEP community is primarily investigating ways to encode collider physics data such that it can serve as a basis for a variety of tasks. We provide studies and insights at this frontier with our work on jet physics.

## References

## Experiment context, if any

**Authors:** HALLIN, Anna (University of Hamburg); KASIECZKA, Gregor (Hamburg University (DE)); BIRK, Joschka Valentin Maria

**Presenter:** HALLIN, Anna (University of Hamburg)

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **95** Type: **Oral**

# Accelerating Machine Learning Inference on GPUs with SYCL using SOFIE

*Tuesday, March 12, 2024 11:50 AM (20 minutes)*

Recently, machine learning has established itself as a valuable tool for researchers to analyze their data and draw conclusions in vari- ous scientific fields, such as High Energy Physics (HEP). Commonly used machine learning libraries, such as Keras and PyTorch, might provide functionality for inference, but they only support their own models, are constrained by heavy dependencies and often provide only a Python API and not a C++ one. SOFIE [13], which stands for System for Optimized Fast Inference code Emit, a part of the ROOT project developed at CERN, creates standalone C++ inference code from an input model in one of the popular machine learning formats. This code is directly invokable from other C++ projects and has minimal dependencies. We will present the new developments of SOFIE extending the functionality to generate SYCL code for machine learning model inference that can run on various GPU platforms and is only dependent on Intel MKL BLAS and portBLAS libraries, achieving a speedup of up to x258 over plain C++ code for large convolutional models.

## Significance

This presentation covers new results coming from new developments that happened last year

## References

## Experiment context, if any

Work happening within the ROOT project (CERN EP/SFT) and in collaboration with CERN Openlab

**Authors:** PANAGOU, Ioanna Maria; MONETA, Lorenzo (CERN); SENGUPTA, SANJIBAN; Dr PADULANO, Vincenzo (CERN)

**Presenter:** Dr PADULANO, Vincenzo (CERN)

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **96** Type: **Oral**

# Fast and Precise Amplitude Surrogates with Bayesian and Symmetry Preserving Networks

*Tuesday, March 12, 2024 12:30 PM (20 minutes)*

One of the biggest obstacles for machine learning algorithms that predict amplitudes from phase space points is the scaling with the number of interacting particles. The more particles there are in a given process, the more challenging it is for the model to provide accurate predictions for the matrix elements. We present a deep learning framework that is built to reduce the impact of this issue, based on the implementation of permutation invariance and Lorentz equivariance within the network architecture. We demonstrate how the use of both of these symmetries grants the model the necessary structure to reproduce LO and NLO amplitude distributions in a competent way for processes with multiple QCD jets in the final state. Additionally, we use Bayesian networks as a main ingredient for all studied amplitude surrogates. That way, we can perform a Bayesian analysis to understand and optimize the uncertainty on model predictions.

## Significance

## References

## Experiment context, if any

**Authors:** PLEHN, Tilman; BRESÓ PLA, Víctor (University of Heidelberg)

**Presenter:** BRESÓ PLA, Víctor (University of Heidelberg)

**Session Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

**Track Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

Contribution ID: **97** Type: **Oral**

# McMule – a Monte Carlo generator for low energy processes

*Thursday, March 14, 2024 2:50 PM (20 minutes)*

McMule, a Monte Carlo for MUons and other LEptons, implements many major QED processes at NNLO (eg. $ee \to ee$, $e\mu \to e\mu$, $ee \to \mu\mu$, $\ell p \to \ell p$, $\mu \to \nu\bar{\nu}e$) including effects from the lepton masses. This makes McMule suitable for predictions for low-energy experiments such as MUonE, CMD-III, PRad, or MUSE.

Recently, McMule gained the ability to generate events at NNLO directly rather than just differential distributions. To avoid negative event weights it employs cellular resampling (2109.07851 & 2303.15246) directly as part of the generation step which further reduces the fraction of negative weights.

## Significance

1) McMule caters to many low-energy experiments whose accuracy requires NNLO-QED predictions as part of the full simulation. This requires an NNLO event generator.
2) McMule offers the first demonstration of cellular resampling at NNLO as well as the benefits it allows when part of the event generation rather than used as a postprocessing step.

## References

https://arxiv.org/abs/2007.01654
https://mcmule.readthedocs.io/

## Experiment context, if any

**Author:** ULRICH, Yannick (Universitaet Bern (CH))

**Presenter:** ULRICH, Yannick (Universitaet Bern (CH))

**Session Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

**Track Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

Contribution ID: **98** Type: **Poster**

# Lamarr: implementing a flash-simulation paradigm at LHCb

*Wednesday, March 13, 2024 4:15 PM (30 minutes)*

In the LHCb experiment, during Run2, more than 90% of the computing resources available to the Collaboration were used for detector simulation. The detector and trigger upgrades introduced for Run3 allow to collect larger datasets that, in turn, will require larger simulated samples. Despite the use of a variety of fast simulation options, the demands for simulations will far exceed the pledged resources.
To face upcoming and future requests for simulated samples, we propose Lamarr, a novel framework implementing a flash-simulation paradigm via parametric functions and deep generative models.
Integrated within the general LHCb Simulation software framework, Lamarr provides analysis-level variables taking as input particles from physics generators, and parameterizing the detector response and the reconstruction algorithms. Lamarr consists of a pipeline of machine-learning-based modules that allow, for selected sets of particles, to introduce reconstruction errors or infer high-level quantities via (non-)parametric functions.
Good agreement is observed by comparing key reconstructed quantities obtained with Lamarr against those from the existing detailed Geant4-based simulation. A reduction of at least two orders of magnitude in the computational cost for the detector modeling phase of the LHCb simulation is expected when adopting Lamarr.

## Significance

In this contribution we will provide an update on the models used to parametrize the tracking reconstruction, now entirely based on deep neural networks, and on a pioneering research on modeling particle-to-particle correlation effects, focusing on the electromagnetic calorimeter reconstruction as an application. An update on the software infrastructure to define low-latency pipelines of machine-learning models will also be discussed.

## References

1 L. Anderlini et al., "Lamarr: the ultra-fast simulation option for the LHCb experiment", PoS ICHEP2022 (2022) 233

2 M. Barbetti, "Lamarr: LHCb ultra-fast simulation based on machine learning models deployed within Gauss", in 21th International Workshop on Advanced Computing and Analysis Techniques in Physics Research: AI meets Reality, 2023, arXiv:2303.11428

[3] L. Anderlini et al., "The LHCb ultra-fast simulation option, Lamarr: design and validation", in 26th International Conference on Computing in High Energy & Nuclear Physics, 2023, arXiv:2309.13213

## Experiment context, if any

LHCb

**Authors:** DAVIS, Adam (University of Manchester (GB)); DERKACH, Denis (National Research University Higher School of Economics (RU)); CORTI, Gloria (CERN); ANDERLINI, Lucio (Universita e INFN, Firenze (IT)); BARBETTI, Matteo (Universita e INFN, Firenze (IT)); MARTINELLI, Maurizio (Universita & INFN, Milano-Bicocca (IT)); MAZUREK, Michal (CERN); MAZUREK, Michał; CAPELLI, Simone (Universita & INFN, Milano-Bicocca (IT))

**Presenters:** MAZUREK, Michal (CERN); MAZUREK, Michał

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **100** Type: **Poster**

# Using Legacy ATLAS C++ Calibration Tools in Modern Columnar Analysis Environments

*Thursday, March 14, 2024 4:10 PM (30 minutes)*

The ATLAS experiment at the LHC relies on crucial tools written in C++ to calibrate physics objects and estimate systematic uncertainties in the event-loop analysis environment. However, these tools face compatibility challenges with the columnar analysis paradigm that operates on many events at once in Python/Awkward or RDataFrame environments. Those challenges arise due to the intricate nature of certain tools, as a result of years of continuous development, and the necessity to support a diverse range of compute environments. In this contribution, we present the ATLAS R&D efforts to adapt these legacy tools to be used in both event-loop and columnar environments with minimal code modifications. This approach enables on-the-fly calibration and uncertainties calculations, minimizing the reliance on intermediate data storage. We demonstrate the functionality and performance of this approach in a Python Jupyter notebook that reproduces a toy Z-boson-peak analysis.

## Significance

This presentation introduces an innovative strategy for incorporating legacy C++ code into the modern data science ecosystem. This approach enables the on-the-fly computation of corrections and uncertainties, consequently diminishing the requirement for intermediary data files. This initiative aligns with the broader goals of the HEP community to curtail reliance on disk storage, especially in preparation for the HL-LHC era and beyond. While columnar analysis using ATLAS lightweight data formats (PHYSLITE) has been demonstrated previously, this marks the first instance where corrections and uncertainties can be computed during data reading. Traditionally, these values were pre-calculated and stored in intermediary data files. ATLAS aims to extend such methodologies to a majority of tools essential for physics analysis, indicating a transformative shift in the HEP data analysis workflow.

## References

1. Columnar analysis and on-the-fly analysis corrections at ATLAS https://indico.jlab.org/event/459/contributions/1158
2. Columnar data analysis with ATLAS analysis formats http://dx.doi.org/10.1051/epjconf/202125103001
3. PHYSLITE - A new reduced common data format for ATLAS https://cds.cern.ch/record/2870350/

## Experiment context, if any

ATLAS

**Authors:** Dr STARK, Giordon Holtsberg (University of California,Santa Cruz (US)); HEINRICH, Lukas Alexander (Technische Universitat Munchen (DE)); FEICKERT, Matthew (University of Wisconsin Madison (US)); VIGL, Matthias (Technische Universitat Munchen (DE)); KRUMNACK, Nils Erik (Iowa State University (US)); KOURLITIS, Vangelis (Technische Universitat Munchen (DE))

**Co-authors:** HELD, Alexander (University of Wisconsin Madison (US)); WATTS, Gordon (University of Washington (US)); HARTMANN, Nikolai (Ludwig Maximilians Universitat (DE))

**Presenter:** VIGL, Matthias (Technische Universitat Munchen (DE))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **101** Type: **Poster**

# Ahead-of-time (AOT) compilation of Tensorflow models for deployment

*Wednesday, March 13, 2024 4:15 PM (30 minutes)*

In a wide range of high-energy particle physics applications, machine learning methods have proven as powerful tools to enhance various aspects of physics data analysis. In the past years, various ML models were also integrated in central workflows of the CMS experiment, leading to great improvements in reconstruction and object identification efficiencies. However, the continuation of successful deployments might be limited in the future due to memory and processing time constraints of more advanced models evaluated on central infrastructure.

A novel inference approach for models trained with TensorFlow, based on Ahead-of-time (AOT) compilation is presented. This approach offers a substantial reduction in memory footprint while preserving or even improving computational performance. This talk outlines strategies and limitations of this novel approach, and presents integration workflow for deploying AOT models in production.

## Significance

The continuation of successful ML model deployments might be limited in the future due to memory and processing time constraints, and this contribution presents a novel approach for inference on central infrastructure that can drastically reduce resource consumption.

## References

## Experiment context, if any

CMS

**Authors:** WIEDERSPAN, Bogdan (Hamburg University (DE)); RIEGER, Marcel (Hamburg University (DE))

**Presenter:** WIEDERSPAN, Bogdan (Hamburg University (DE))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **102** Type: **Poster**

# The Good, The Bad, and the Ugly: A Tale of Physics, Software, and ML

*Wednesday, March 13, 2024 4:15 PM (30 minutes)*

The search for long lived particles, a common extension to the Standard Model, requires a sophisticated neural network design, one that is able to accurately discriminate between signal and background. At the LHC's ALTAS experiment, beam induced background (BIB) and QCD jets are the two significant sources of background. For this purpose, a recurrent neural network (RNN) with an adversary was used to distinguish long lived particles from BIB and QCD as well as control systematic errors. We are presenting the modernization of this neural network through the use of software tools and techniques including testing, continuous integration, and current software design techniques. This modernization has improved the sustainability, functionality, versatility, and performance of the network which will be used in future analyses.

## Significance

This work discusses the improvement of a recurrent neural network used in the LHC ATLAS experiment. We improved the sustainability and performance of the network through the use of modern software techniques such as testing and continuous integration. This network will be used in future analyses which will benefit from its improvement.

## References

## Experiment context, if any

ATLAS, all open source

**Authors:** GOLUB, Alexandra (University of Washington (US)); WATTS, Gordon (University of Washington (US))

**Presenter:** GOLUB, Alexandra (University of Washington (US))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **103** Type: **Oral**

# From Amsterdam to ACAT 2024: The Evolution and Convergence of Declarative Analysis Language Tools and Imperative Analysis Tools

*Thursday, March 14, 2024 3:50 PM (20 minutes)*

Declarative Analysis Languages (DALs) are a paradigm for high-energy physics analysis that separates the desired results from the implementation details. DALs enable physicists to use the same software to work with different experiment's data formats, without worrying about the low-level details or the software infrastructure available. DALs have gained popularity since the HEP Analysis Ecosystem Retreat in Amsterdam in 2017, where they were first seriously discussed as in a community setting. Since then, several languages and tools have adopted features from DALs, such as ADL, awkward, RDF, and func_adl. These languages and tools vary in how much they adhere to the declarative principle, and how they interact with imperative code. In this presentation, we will review this convergence and some of the common features and challenges of DALs, using examples from various languages and tools. We will also discuss recent developments allowing different languages access to custom experiment data formats. We will conclude by reflecting on the future directions of DALs and imperative analysis tools, and how they can improve their convergence, usefulness, and efficiency.

## Significance

Declarative Languages have captured the imagination of many physicists. As they have started to gain some maturity the imperative tools and languages have started to adopt some of their features. This talk will talk about some recent developments and work demonstrating this and predict (or push) for a future for more of this.

## References

## Experiment context, if any

IRIS-HEP

**Author:** WATTS, Gordon (University of Washington (US))

**Presenter:** WATTS, Gordon (University of Washington (US))

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **104** Type: **Oral**

# dilax: Differentiable Binned Likelihoods in JAX

*Thursday, March 14, 2024 5:50 PM (20 minutes)*

dilax is a software package for statistical inference using likelihood functions of binned data. It fulfils three key concepts: performance, differentiability, and object-oriented statistical model building.
dilax is build on JAX - a powerful autodifferentiation Python framework. By making every component in dilax a "PyTree", each component can be jit-compiled (jax.jit), vectorized (jax.vmap) and differentiated (jax.grad). This enables additionally novel computational concepts, such as running thousands of fits simultaneously on a GPU or differentiating through measurements of physical observables.
We present the key concepts of dilax, show its features, and discuss performance benchmarks with toy datasets.

## Significance

This project is a new statistics tool suited for typical measurements in LHC analyses. It focusses on performance, usability, and novel computing techniques such as autodifferentiation and vectorization of full fits. This project has not been presented so far. It introduces new concepts that have not been covered by other statistics libraries.

## References

## Experiment context, if any

Use case for CMS, ATLAS, LHCb analyses

**Authors:** FACKELDEY, Manfred Peter (RWTH Aachen University (DE)); FISCHER, Benjamin (RWTH Aachen University (DE)); ZINN, Felix Philipp (Rheinisch Westfaelische Tech. Hoch. (DE)); ERDMANN, Martin (Rheinisch Westfaelische Tech. Hoch. (DE))

**Presenter:** FACKELDEY, Manfred Peter (RWTH Aachen University (DE))

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **105** Type: **Poster**

# Optimizing Resource Provisioning Across Diverse Computing Facilities with Virtual Kubelet Integration

*Wednesday, March 13, 2024 4:15 PM (30 minutes)*

The integration of geographically diverse computing facilities involves dynamically allocating unused resources, relocating workflows, and addressing challenges in heterogeneous, distributed, and opportunistic compute provisioning. Key hurdles include effective resource management, scheduling, data transfer optimization, latency reduction, and ensuring security and privacy. Our proposed solution, part of the "JLAB Integrating Research Infrastructure Across Facilities (JIRIAF)" project, leverages the Kubernetes framework within userspace. It utilizes Virtual Kubelet implementation to overcome high-level permission limitations on worker nodes, connecting Kubernetes with arbitrary APIs. This implementation enables Virtual Kubelet deployment in userspace for executing shell commands, resulting in an elastic and cross-site Kubernetes cluster that offers enhanced flexibility and resource utilization.

## Significance

This solution leverages the widely recognized Kubernetes (K8s) framework. By integrating Virtual Kubelet (VK), it effectively addresses challenges in diverse, geographically distributed computing environments. This approach enhances resource allocation and scheduling, enabling the deployment of an elastic, cross-site Kubernetes cluster. With its utilization of the well-known K8s framework, this advancement is poised to significantly optimize workflows and efficiently utilize resources across varied computing facilities.

## References

https://indico.jlab.org/event/459/contributions/11501/

## Experiment context, if any

The conducted experiment involved implementing a streaming data workflow from ESnet to NERSC, with the deployment of Virtual Kubelet (VK) at NERSC. The key observation from this experiment is that users can submit jobs to the local control-plane, and Kubernetes (K8s) efficiently distributes these jobs to remote sites based on available resources. This practical application demonstrates the effectiveness of the proposed solution in optimizing resource utilization and streamlining job distribution across geographically distributed computing facilities.

**Author:** TSAI, Jeng-Yuan

**Co-authors:** LARRIEU, Christopher (Thomas Jefferson National Accelerator Facility); LAWRENCE, David; HEYES, Graham (Jefferson Lab); Dr GYURJYAN, Vardan

**Presenter:** TSAI, Jeng-Yuan

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **106** Type: **Poster**

# Paving the Way for HPC: An XRootD-Based Approach for Efficiency and Workflow Optimizations for HEP Jobs on HPC Centers

*Wednesday, March 13, 2024 4:15 PM (30 minutes)*

Today, the Worldwide LHC Computing Grid (WLCG) provides the majority of compute resources for the High Energy Physics (HEP) community. With its homogeneous Grid centers all around the world trimmed to a high throughput of data, it is tailored to support typical HEP workflows, offering an optimal environment for efficient job execution.

With the future German HEP computing strategy, however, there will be a shift away from dedicated resources to official shares on national HPC centers.
This bears many challenges, since these more heterogeneous resources are designed for performance and security rather than transfering and processing large amounts of data. The different focus and certain limitations can lead to higher failure rates and worse efficiency of HEP jobs running at such centers, i.e. because of tendentially slower WAN connections.
Monitoring data collected at the HoreKa HPC Center at KIT further confirmed that assumption. Lower CPU efficiency and an increased failure rate compared to the KIT Tier-1 center indicated a bandwidth limitation, in particular for data intensive workflows.

An efficient resource utilization, however, is the main objective for the success of the German HEP strategy in the future, not only in terms of sustainability, but also to cope with the anticipated data rates of the HL-LHC era. To tackle these challenges, we developed an XRootD/XCache based solution - in close contact with the XRootD/XCache developers - that aims to maximize computational throughput and mitigates the limitations of contemporary HPC centers. The currently operational setup at HoreKa leverages the parallel filesystem and a transfer node of the cluster as some sort of XRootD caching 'buffer', leading to a more stable performance and better utilization of the cluster.

In this contribution, the prerequisites and challenges associated with HEP workflows on HPC centers are pointed out and our solution for a more efficient utilization, backed by a fully operational proof of concept on HoreKa, is presented.

## Significance

Since HPC centers gain importance in the HEP computing environment in the future, especially in Germany in the next year, a well prepared and fluent transition to such resources is important for the efficent operation of the WLCG. We are ensuring this with our work and other sites can benefit from our experience of incorporating HPC efficiently.

## References

## Experiment context, if any

The current P.o.C is running with CMS jobs, but the presented concepts are not specifically aimed at CMS and can in principle be generalized to other experiments.

**Author:** HOFSAESS, Robin (KIT - Karlsruhe Institute of Technology (DE))

**Co-authors:** STREIT, Achim (KIT - Karlsruhe Institute of Technology (DE)); PETZOLD, Andreas (KIT - Karlsruhe Institute of Technology (DE)); GOTTMANN, Artur Il Darovic (KIT - Karlsruhe Institute of Technology (DE)); QUAST, Gunter (KIT - Karlsruhe Institute of Technology (DE)); GIFFELS, Manuel (KIT - Karlsruhe Institute of Technology (DE)); SCHNEPF, Matthias Jochen

**Presenter:** HOFSAESS, Robin (KIT - Karlsruhe Institute of Technology (DE))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **107** Type: **Poster**

# Optimal XCache deployment for the CMS experiment in Spain

*Wednesday, March 13, 2024 4:15 PM (30 minutes)*

The Large Hadron Collider at CERN in Geneva is poised for a transformative upgrade, preparing to enhance both its accelerator and particle detectors. This strategic initiative is driven by the tenfold increase in proton-proton collisions anticipated for the forthcoming high-luminosity phase scheduled to start by 2029. The vital role played by the underlying computational infrastructure, the World-Wide LHC Computing Grid, in processing the data generated during these collisions underlines the need for its expansion and adaptation to meet the demands of the new accelerator phase. The provision of these computational resources by the worldwide community remains essential, all within a constant budgetary framework. While technological advancements offer some relief for the expected increase, numerous research and development projects are underway. Their aim is to bring future resources to manageable levels and provide cost-effective solutions to effectively handle the expanding volume of generated data. In the quest for optimised data access and resource utilisation, the LHC community is actively investigating Content Delivery Network (CDN) techniques. These techniques serve as a mechanism for the cost-effective deployment of lightweight storage systems that support both, traditional and opportunistic compute resources. Furthermore, they aim to enhance the performance of executing tasks by facilitating the efficient reading of input data via caching content near the end user. A comprehensive study is presented to assess the benefits of implementing data cache solutions for the Compact Muon Solenoid (CMS) experiment. This in-depth examination serves as a use-case study specifically conducted for the Spanish compute facilities, playing a crucial role in supporting CMS activities. Data access patterns and popularity studies suggest that user analysis tasks benefit the most from CDN techniques. Consequently, a data cache has been introduced in the region to acquire a deeper understanding of these effects. In this contribution we will focus on the remote data accesses from users that execute tasks in the Spanish CMS sites, in order to simulate and discern the most optimal requirements in both size and network connectivity for a data cache serving the whole Spanish region. This is a mandatory step towards a better understanding of the service to be deployed in a federated fashion in the region.

## Significance

This study is pioneer, never done and presented before, and a similar approach can be used by other countries who would like to explore CDNs/XCache solutions.

## References

## Experiment context, if any

CMS

**Author:** FLIX MOLINA, Jose (CIEMAT - Centro de Investigaciones Energéticas Medioambientales y Tec. (ES))

**Co-authors:** Dr SIKORA, Anna (UAB); PEREZ DENGRA, Carlos (PIC); Ms SERRANO, Paula (UAB)

**Presenter:** FLIX MOLINA, Jose (CIEMAT - Centro de Investigaciones Energéticas Medioambientales y Tec. (ES))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **108** Type: **Poster**

# Awkward Family: expanding functionality through interrelated Python packages

*Wednesday, March 13, 2024 4:15 PM (30 minutes)*

In the 5+ years since their inception, Uproot and Awkward Array have become cornerstones for particle physics analysis in Python, both as direct user interfaces and as base layers for physicist-facing frameworks. Although this means that the software is achieving its mission, it also puts the need for stability in conflict with new, experimental developments. Boundaries must be drawn between the code that must stay robust and the code that implements new ideas.

In this poster, I'll describe how we leverage Python's packaging infrastructures to separate stable components from experimental components at the package boundaries. The uproot and awkward packages were chosen to be long-term maintenance components, while new capabilities are provided in:

- dask-awkward: distributed computing
- awkward-pandas: DataFrame interface
- AwkwardArray.jl: Julia interface and reinterpretation
- kaitai_struct_awkward_runtime: arbitrary file format ⟶ Awkward Array generator
- odapt: high-level file operations: copying, converting, concatenating, skimming and slimming
- uproot-browser: high-level TUI interface to Uproot
- ragged: just the ragged arrays, but satisfying the Python Array API
- vector: Lorentz vector manipulation (in and out of Awkward Arrays)

I'll also describe some best practices (learned through mistakes!) in coordinating versions, managing deprecations, public/private API boundaries, and cross-package testing.

## Significance

## References

- Uproot: https://uproot.readthedocs.io/
- Awkward Array: https://awkward-array.org/
- dask-awkward: https://github.com/dask-contrib/dask-awkward
- awkward-pandas: https://github.com/intake/awkward-pandas
- AwkwardArray.jl: https://github.com/JuliaHEP/AwkwardArray.jl
- kaitai_struct_awkward_runtime: https://github.com/ManasviGoyal/kaitai_struct_awkward_runtime
- odapt: https://github.com/zbilodea/odapt
- uproot-browser: https://github.com/scikit-hep/uproot-browser
- https://github.com/jpivarski/ragged

- Vector: https://github.com/scikit-hep/vector
- Python Array API: https://data-apis.org/array-api/latest/index.html

## Experiment context, if any

**Author:** PIVARSKI, Jim (Princeton University)

**Presenter:** PIVARSKI, Jim (Princeton University)

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **109** Type: **Oral**

# Fully containerised approach for the HPC cluster at FAIR

*Thursday, March 14, 2024 2:50 PM (20 minutes)*

The scientific program of the future FAIR accelerator covers a broad spectrum of topics in modern nuclear and atomic physics. This diversity leads to a multitude of use cases and workflows for the analysis of experimental data and simulations. To meet the needs of such a diverse user group, a flexible and transparent High-Performance Computing (HPC) system is required to accommodate all FAIR experiments and users.

In this presentation, we present an operational approach for the computing cluster at GSI/FAIR that is characterized by an exceptionally minimal host system. This is achieved by installing and running all user applications in containers. Our successful implementation of this approach on a production system hosting approximately 700 users, 80,000 CPUs and 400 GPUs demonstrates its feasibility and scalability.

We present a transparent solution for interactive work in a containerized environment that addresses different levels of user experience. In addition, users have the opportunity to construct and submit their own containers. The presentation will cover also how usage of Spack and CVMFS contributes to the overall efficiency and adaptability of the computing cluster at GSI/FAIR.

## Significance

## References

## Experiment context, if any

**Authors:** PREUSS, Carsten (Unknown); BERTINI, Denis (GSI Darmstadt); KRESAN, Dmytro (GSI - Helmholtzzentrum fur Schwerionenforschung GmbH (DE)); DESSALVI, Matteo (GSI); AL-TURANY, Mohammad (CERN); GROSSO, Raffaele (GSI - Helmholtzzentrum fur Schwerionenforschung GmbH (DE)); FLEISCHER, Soren Lars Gerald (GSI - Helmholtzzentrum fur Schwerionenforschung GmbH (DE)); KOLLEGGER, Thorsten (GSI - Helmholtzzentrum fur Schwerionenforschung GmbH (DE)); PENSO HOF, Victor Manuel (GSI - Helmholtzzentrum fur Schwerionenforschung GmbH (DE))

**Presenter:** KRESAN, Dmytro (GSI - Helmholtzzentrum fur Schwerionenforschung GmbH (DE))

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **110** Type: **Oral**

# Wire-Cell: A High Quality Automated LArTPC Reconstruction for Neutrino Experiments

*Monday, March 11, 2024 4:50 PM (20 minutes)*

Liquid Argon Time Projection Chamber, or LArTPC, is a scalable tracking calorimeter featuring rich event topology information. It provides the core detector technology for many current and next-generation large-scale neutrino experiments, such as DUNE and the SBN program. In neutrino experiments, LArTPC faces numerous challenges in both hardware and software to achieve optimum performance. On the software side, the main challenge is two-fold. First, there is a need for further accumulation of deep domain knowledge. Second, the event's degree of freedom is high due to its large scale and the uncertainties in initial neutrino-argon interactions.

To address the reconstruction challenge from LArTPC detectors, we developed a comprehensive software suite and a set of algorithms, collectively termed 'Wire-Cell.' This innovative system is founded on the concept of topographical 3D reconstruction using multiple 2D LArTPC images. Key to enabling this 3D reconstruction was an extensive study of detector signal formation and signal processing, providing crucial insights. Building on the outcomes of the 3D reconstruction, we crafted a high-quality automated neutrino reconstruction chain. This chain integrates both traditional and machine-learning algorithms. The effectiveness of the Wire-Cell reconstruction approach is validated with real experiment data, specifically in the context of MicroBooNE physics analyses.

In this presentation, we delve into the Wire-Cell LArTPC reconstruction paradigm, with a particular emphasis on the underlying algorithms. Our focus is to demonstrate how Wire-Cell not only addresses but also advances the field of LArTPC detector data interpretation.

## Significance

The Wire-Cell development is a multi-year project that aims to address fundamental issues in LArTPC reconstruction. Now, the full set of algorithms is ready and has been tested in real data analyses. We think it is time to present the big picture of this project, along with its algorithms and toolkits, to a broader community to exchange ideas.

## References

## Experiment context, if any

Neutrino Experiments, DUNE, MicroBooNE

**Author:** YU, Haiwang

**Presenter:** YU, Haiwang

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **111** Type: **Oral**

# High Pileup Particle Tracking with Learned Clustering

*Monday, March 11, 2024 3:30 PM (20 minutes)*

The sub-optimal scaling of traditional tracking algorithms based on combinatorial Kalman filters causes performance concerns for future high-pileup experiments like the High Luminosity Large Hadron Collider. Graph Neural Network-based tracking approaches have been shown to significantly improve scaling at similar tracking performance levels. Rather than employing the popular edge classification approach, we use learned clustering to reconstruct track candidates. This talk presents our first results on the full-detector trackML dataset. We also show that standard embedding strategies deliver similar results to the more complicated object condensation approach and how base models trained with simplified approaches can be fine-tuned for optimal performance. Finally, we show results using a node filtering first stage that reduces the point cloud size before graphs are built, improving inference speeds.

## Significance

First Object Condensation GNN tracking results that include all detector layers (not just pixel detector) of the trackML dataset without truth cuts or other simplifications. New approach with simplified loss functions. New hit filter approach that increases the inference speed.

## References

Connecting the dots 23: https://indico.cern.ch/event/1252748/contributions/5521458/ and http://arxiv.org/abs/2312.03823
Chep 23: https://indico.jlab.org/event/459/contributions/11741/ and https://arxiv.org/abs/2309.16754

## Experiment context, if any

**Authors:** DEZOORT, Gage (Princeton University (US)); LIERET, Kilian (Princeton University)

**Presenter:** LIERET, Kilian (Princeton University)

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **112** Type: **Poster**

# Scalable GNN Training for Track Finding

*Wednesday, March 13, 2024 4:15 PM (30 minutes)*

Graph Neural Networks (GNNs) have demonstrated significant performance in addressing the particle track-finding problem in High-Energy Physics (HEP). Traditional algorithms exhibit high computational complexity in this domain as the number of particles increases. This poster addresses the challenges of training GNN models on large, rapidly evolving datasets, a common scenario given the advancements in data generation, collection, and increase in storage capabilities. The computational and GPU memory requirements present significant roadblocks in efficiently training GNNs on large graph structures. One effective strategy to reduce training time is distributed data parallelism on multi-GPUs, which involves averaging gradients across the devices used for training.

This poster will report the speed-up of GNN training time when using distributed data parallelism with different numbers of GPUs and computing nodes. Running GNN training with distributed data parallelism leads to a decrease in accuracy. We are investigating the relationship between the number of devices and model accuracy degradation and strategies to mitigate it. Preliminary results on the TrackML dataset will be reported. GPU nodes from Perlmutter at NERSC will be used to run the experiments.

## Significance

As the availability of HPC platforms with multi-GPUs increases, distributed deep learning training becomes an essential tool for exploring and experimenting with cutting-edge deep learning architectures and methodologies. By handling larger datasets and complex models, researchers and HEP scientists can push the boundaries of AI capabilities to improve the physics performance of track-finding experiments.

## References

Ju, X., Murnane, D., Calafiura, P., Choma, N., Conlon, S., Farrell, S., ... & Lazar, A. (2021). Performance of a geometric deep learning pipeline for HL-LHC particle tracking. The European Physical Journal C, 81, 1-14.

Lazar, A., Ju, X., Murnane, D., Calafiura, P., Farrell, S., Xu, Y., ... & Lucas, A. (2023, February). Accelerating the Inference of the Exa. TrkX Pipeline. In Journal of Physics: Conference Series (Vol. 2438, No. 1, p. 012008). IOP Publishing.

## Experiment context, if any

We report the results of training GNN models on the TrackML dataset. Even if this dataset is based on a simulation of a generic HL-LHC experiment tracker, the results could be extended to design and evaluate particle tracking algorithms for any of the experiments.

**Authors:** LADUSKA, Ivan (Youngstown State University); REEVES, Brenden (Youngstown State Uni-

versity); MANJEROVIC, Caroline (Youngstown State University); LAZAR, Alina (Youngstown State University); PHAM, Minh-Tuan (University of Wisconsin Madison (US)); CHAN, Jay (Lawrence Berkeley National Lab. (US)); MURNANE, Daniel Thomas (Lawrence Berkeley National Lab. (US)); JU, Xiangyang (Lawrence Berkeley National Lab. (US)); CALAFIURA, Paolo (Lawrence Berkeley National Lab. (US))

**Presenter:** LAZAR, Alina (Youngstown State University)

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **113** Type: **Poster**

# Hydra: Computer Vision for Data Quality Monitoring

*Wednesday, March 13, 2024 4:15 PM (30 minutes)*

Hydra is an extensible framework for training, managing, and deploying machine learning models for near real-time data quality monitoring. It is designed to take some of the burden off of shift crews by providing 'round-the-clock'monitoring of plots representing the data being collected. The Hydra system is backed by a database which is leveraged for near push button training and is primarily controlled and viewed through a set of web interfaces. This web interface contains a simple-to-use web based GUI for labeling of the datasets used in training; making it possible to label thousands of images quickly and efficiently. To aid in the analysis of Hydra inferences gradCAM, a method of interpretability, is performed and overlaid on the target image, highlighting regions of interest, making diagnosis of problems much faster.
Development began in 2019 for the GlueX Experiment in Hall-D, the Hydra system has grown to encompass all of the experimental halls at Jefferson Laboratory. This talk will focus on the features of Hydra as well as provide details of the challenges present with deploying to disparate experimental halls. With a roadmap of development established, Hydra aims to continue to grow in richness of feature set and expand to encompass other monitoring tasks in differing communities.

## Significance

Hydra represents a successful integration of AI/ML into in-situ data acquisition at Jefferson Laboratory. It is deployed in all of the experimental halls and has quickly become part of the standard operating procedures for shift crews, routinely outperforming its human counterparts in the detection of problems with the acquisition of physics quality data.

## References

Chep 2021 https://www.epj-conferences.org/articles/epjconf/pdf/2021/05/epjconf_chep2021_04010.pdf

Acat 2022 https://indico.cern.ch/event/1106990/contributions/4991255/

## Experiment context, if any

All experimental halls at Jefferson Laboratory

**Authors:** ROY, Ayan; LAWRENCE, David; BRITTON, Thomas; JESKE, Torri

**Co-author:** MATSIUK, Nataliia (Jefferson Laboratory)

**Presenter:** ROY, Ayan

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **114** Type: **Poster**

# RTDP: Streaming Readout Real-Time Development and Testing Platform

*Wednesday, March 13, 2024 4:15 PM (30 minutes)*

The Thomas Jefferson National Accelerator Facility (JLab) has created and is currently working on various tools to facilitate streaming readout (SRO) for upcoming experiments. These include reconstruction frameworks with support for Artificial Intelligence/Machine Learning, distributed High Throughput Computing (HTC), and heterogeneous computing which all contribute significantly to swift data processing and analysis. Designing SRO systems that combine such components for new experiments would benefit from a platform that would combine both simulation and execution components for simulation, testing, and validation before large investments are made. The Real-Time Development Platform (RTDP) is being developed as part of an LDRD funded project at JLab. RTDP aims to establish a seamless connection between algorithms, facilitating the seamless processing of data from SRO to analysis, as well as enabling the execution of these algorithms in various configurations on compute and data centers. Individual software components simulating specific hardware can be replaced with actual hardware when it is available.

## Significance

In the data acquisition phase, we have directly captured network packets from the high-speed Network Interface Card, that included hardware timestamps. This functionality allows us to replay the data offline as input to the platform when the beam is unavailable. Furthermore, this process also facilitates the observation and examination of potential interference between hardware components, such as buffer oscillations.

## References

https://wiki.jlab.org/epsciwiki/images/4/41/FY24-LDRD_Proposal_SRO.pdf
https://link.springer.com/article/10.1140/epjp/s13360-022-03146-z

## Experiment context, if any

The RTDP tool will be generically useful to develop and validate streaming systems for multiple experiments. These will include SoLID(Hall-A), CLAS12(Hall-B), GlueX(Hall-D), and ePIC (EIC).

**Authors:** ROY, Ayan; LAWRENCE, David; TSAI, Jeng-Yuan; BATTAGLIERI, Marco (INFN); DIEFENTHALER, Markus; GYURJYAN, Vardan (Jefferson Lab); Dr GYURJYAN, Vardan; MEI, Xinxin (Cissie)

**Presenters:** GYURJYAN, Vardan (Jefferson Lab); Dr GYURJYAN, Vardan

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **115** Type: **Oral**

# Leveraging Large-Scale Pretraining for Efficient Jet Classification: An Evaluation of Transfer Learning, Model Architectures, Dataset Scaling, and Domain Adaptation in Particle Physics

*Monday, March 11, 2024 3:10 PM (20 minutes)*

In particle physics, machine learning algorithms traditionally face a limitation due to the lack of truth labels in real data, restricting training to only simulated samples. This study addresses this challenge by employing self-supervised learning, which enables the utilization of vast amounts of unlabeled real data, thereby facilitating more effective training.

Our project is particularly motivated by the need for improved data-Monte Carlo (MC) agreement in CMS analyses, seeking to bridge the gap between simulation and real-world data. We employ contrastive learning to leverage the JetClass dataset for large-scale pretraining, aiming to capture generalizable features about jets. These features can then be fine-tuned for downstream classification tasks such as Top Tagging and $H \to b\bar{b}$ vs QCD, with minimal additional effort.

The research explores several key questions: the scalability of dataset size in pretraining, the comparative analysis of contrastive learning techniques like SimCLR and VICReg, the effectiveness of the ParticleTransformer architecture over conventional transformer models, and whether self-supervised pretraining on unlabeled data combined with fine-tuning on labeled simulation aids the model in adapting to the data domain.

By investigating these aspects, we aim to provide insights into the impact of dataset size on pretraining, evaluate the strengths and weaknesses of various contrastive learning methods, assess the architectural advantages of ParticleTransformer in jet classification, and facilitate the domain adaptation of machine learning algorithms for enhanced applicability in particle physics.

This study significantly contributes to the field of machine learning in particle physics, demonstrating the immense potential of self-supervised learning in utilizing real, unlabeled data for more efficient and accurate jet classification.

## Significance

This research introduces a groundbreaking approach in experimental particle physics by utilizing self-supervised learning, particularly contrastive learning, to exploit large datasets of unlabeled real data. It explores novel aspects such as the scalability of pre-training dataset size, the comparative effectiveness of contrastive learning techniques, and the potential of the ParticleTransformer architecture over conventional models. These contributions represent substantial progress beyond standard practices, providing practical advancements in machine learning applications within experimental particle physics.

## References

## Experiment context, if any

**Authors:** PAREJA, Carlos (University of California, San Diego); MOKHTAR, Farouk (Univ. of California San Diego (US)); LI, Haoyang (Univ. of California San Diego (US)); DUARTE, Javier Mauricio (Univ. of California San Diego (US)); KANSAL, Raghav (Univ. of California San Diego (US)); ZHAO, Zihan (Univ. of California San Diego (US))

**Presenter:** ZHAO, Zihan (Univ. of California San Diego (US))

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **117** Type: **Oral**

# Denoising Graph Super-Resolution with Diffusion Models and Transformers for Improved Particle Reconstruction

*Tuesday, March 12, 2024 12:50 PM (20 minutes)*

Accurately reconstructing particles from detector data is a critical challenge in experimental particle physics. The detector's spatial resolution, specifically the calorimeter's granularity, plays a crucial role in determining the quality of the particle reconstruction. It also sets the upper limit for the algorithm's theoretical capabilities. Super-resolution techniques can be explored as a promising solution to address the limitations imposed by the detector's spatial resolution. Super-resolution refers to enhancing the resolution of low-resolution images to obtain higher-resolution versions. In the specific case of calorimeter data, which is characterized by sparsity and non-homogeneity, representing it using graphs provides the most faithful representation. Building upon this idea, we propose a diffusion model for graph super-resolution that uses a transformer-based de-noising network to enhance the resolution of calorimeter data. Notably, this study represents the first instance of applying graph super-resolution with diffusion. The low-resolution image, corresponding to recorded detector data, is also subject to noise from various sources. As an added benefit, the proposed model aims to remove these noise artifacts, further contributing to improved particle reconstruction.

## Significance

Increasing resolutions of detector data with Diffusion powered graph super-resolution that can help improve reconstruction, and push the theoretical limits of reconstruction

## References

1. Presentation at ML4Jets, Hamburg, Germany https://indico.cern.ch/event/1253794/contributions/5588579/

## Experiment context, if any

LHC experiments in general

**Author:** KAKATI, Nilotpal (Weizmann Institute of Science (IL))

**Co-authors:** GROSS, Eilam (Weizmann Institute of Science (IL)); DREYER, Etienne (Weizmann Institute of Science (IL))

**Presenter:** KAKATI, Nilotpal (Weizmann Institute of Science (IL))

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **118** Type: **Poster**

# HPC, HTC and Cloud: converging toward a seamless computing federation with interLink

*Wednesday, March 13, 2024 4:15 PM (30 minutes)*

In the era of digital twins a federated system capable of integrating High-Performance Computing (HPC), High-Throughput Computing (HTC), and Cloud computing can provide a robust and versatile platform for creating, managing, and optimizing Digital Twin applications. One of the most critical problems involve the logistics of wide-area with multi stage workflows that move back and forth across multiple resource providers. We envision a model where such a challenge can be addressed enabling a "transparent offloading"of containerized payloads using the Kubernetes API primitives enabling a transparent access to any number of external hardware machines and type of backends. Thus we created the interLink project, an open source extension to the concept of Virtual-Kubelet with a design that aims for a common abstraction over heterogeneous and distributed backends. The primary goal is to have HPC centers exploitable with native Kubernetes APIs with an effort close to zero from all the stakeholders' standpoint.
interLink is developed by INFN in the context of interTwin, an EU funded project that aims to build a digital-twin platform ( Digital Twin Engine) for sciences, and the ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing in Italy. In this talk we will walk through the key features and the early use cases of a Kubernetes-based computing platform capable of extending its computational capabilities over heterogeneous providers: among others, the integration of a world-class supercomputer such as EuroHPC Vega will be showcased.

## Significance

The presented solution enhances the Virtual-Kubelet technology with a design that aims for a common abstraction over heterogeneous and distributed Kubelet backends. The primary goal is to have HPC centers exploitable with native Kubernetes APIs with an effort close to zero for all the stakeholder.
The characterising feature of the interLink project is the definition of a common API spec between the Virtual-Kubelet and the remote host runtime implementation, so that any provider can be completely free deciding how a container execution request can be satisfied.

## References

## Experiment context, if any

**Authors:** SPIGA, Daniele (Universita e INFN, Perugia (IT)); CIANGOTTINI, Diego (INFN, Perugia (IT))

**Co-authors:** MEMON, Ahmed Shiraz; MANZI, Andrea; FILIPCIC, Andrej (Jozef Stefan Institute (SI)); PRICA, Teo (IZUM); Dr BOCCALI, Tommaso (INFN Sezione di Pisa); TEDESCHI, Tommaso (Universita e INFN, Perugia (IT)); SURACE, giacomo (infn)

**Presenter:** TEDESCHI, Tommaso (Universita e INFN, Perugia (IT))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **119** Type: **Oral**

# Advancing Image Classification using Intel SDK: Integrating NAQSS Encoding with Hybrid Quantum-Classical PQC Models

*Monday, March 11, 2024 4:50 PM (20 minutes)*

Artificial intelligence has been used for the real and fake art identification and different machine learning models are being trained then employed with acceptable accuracy in classifying artworks. As the future revolutionary technology, quantum computing opens a grand new perspective in the art area. Using Quantum Machine Learning (QML), the current work explores the utilization of Normal Arbitrary Quantum Superposition State (NAQSS) for encoding images into a quantum circuit. The learning of trainable parameters for image classification is achieved through the use of layers of Parameterized Quantum Circuit (PQC) with a hybrid optimizer. Starting with the simplest example i.e. 2x2-colored images, the accuracy has been improved with the increasing size of the images, as the circuit depth increases linearly with the image size namely quantum gates. The potential of QML and parameters influencing accuracy are extensively investigated. The implementations have been carried out using the Intel Quantum SDK (Software Development Kit), based on the research within the framework of cooperation between Intel Labs and Deggendorf Institute of Technology.

## Significance

A real innovative case of Quantum Machine Learning in the art identification.

## References

## Experiment context, if any

**Authors:** Mr AJAREKAR, Digvijaysinh (Deggendorf Institute of Technology); Mr AL-ROUSAN, Suhaib (Deggendorf Institute of Technology); Prof. LIEBELT, Helena (Deggendorf Institute of Technology); LI, Rui (Deggendorf Institute of Technology)

**Presenter:** Mr AJAREKAR, Digvijaysinh (Deggendorf Institute of Technology)

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **121** Type: **Oral**

# The Neural Network First-Level Hardware Track Trigger of the Belle II Experiment

*Wednesday, March 13, 2024 4:50 PM (20 minutes)*

We describe the principles and performance of the first-level (”L1”) hardware track trigger of Belle II, based on neural networks. The networks use as input the results from the standard \belleii trigger, which provides “2D” track candidates in the plane transverse to the electron-positron beams. The networks then provide estimates for the origin of the 2D track candidates in direction of the colliding beams (z-vertex), as well as their polar emission angles theta. Given the z-vertices of the neural tracks allows identifying events coming from the collision region (z ˜ 0), and suppressing the overwhelming background from outside by a suitable cut d. Requiring |z| < d for at least one neural track in an event with two or more 2D candidates will set an L1 trigger. The networks also enable a minimum bias trigger, requiring a single 2D track candidate validated by a neural track with a momentum larger than 0.7 GeV in addition to the |z| condition. The momentum of the neural track is derived with the help of the polar angle theta.

## Significance

The Level 1 Neural Network Track Trigger is the first of its kind operating in a high energy physics experiment. It provides even a minimum bias single track trigger, also the first of its kind in an electron-positron experiment.

## References

Talk given by collaborator on last year’s ACAT conference, giving the status of the hardware development. A publication for NIMA is in preparation.

## Experiment context, if any

The neural trigger is operating at the Belle II experiment at KEK, Japan

**Author:** KIESLING, Christian (Max Planck Institut für Physik (DE))

**Co-authors:** LENZ, Alex (TUM (Inf)); KNOLL, Alois (TUM (Inf)); MEGGENDORFER, Felix (Max Planck Institut für Physik (DE)); BECKER, Jürgen (KIT (ITIV)); UNGER, Kai (KIT (ITIV); BÄHR, Steffen (KIT (ITIV)); JÜLG, Tobias (TUM (Inf))

**Presenter:** KIESLING, Christian (Max Planck Institut für Physik (DE))

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **122** Type: **Poster**

# Easy columnar file conversions with ”odapt”

*Thursday, March 14, 2024 4:10 PM (30 minutes)*

When working with columnar data file formats, it is easy for users to devote too much time to file manipulation. With Python, each file conversion requires multiple lines of code and the use of multiple I/O packages. Some conversions are a bit tricky if the user isn’t very familiar with certain formats, or if they need to work with data in smaller batches for memory management. To try and address this issue, we are developing Python package ‘odapt.’This package allows users to convert files with just one function call, with automatic memory management, compression settings, and other features added based on user feedback. Some such features include merging ROOT files (hadd-like), adding and dropping branches or TTrees from ROOT files. Odapt uses reliable columnar I/O packages h5py, Uproot, Awkward, and dask-awkward.

## Significance

Though the project is still in development, we have gotten a lot of interest and feature-requests from users who frequently need to do columnar file conversions.

## References

## Experiment context, if any

Converting large files between different columnar formats.

**Author:** BILODEAU, Zoë (Princeton University (US))

**Presenter:** BILODEAU, Zoë (Princeton University (US))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **123** Type: **Oral**

# To be or not to be Equivariant?

*Wednesday, March 13, 2024 5:30 PM (20 minutes)*

Equivariant models have provided state-of-the-art performance in many ML applications, from image recognition to chemistry and beyond. In particle physics, the relevant symmetries are permutations and the Lorentz group, and the best-performing networks are either custom-built Lorentz-equivariant architectures or more generic large transformer models. A major unanswered question is whether the high performance of Lorentz-equivariant architectures is in fact due to their equivariance. Here we report a study designed to isolate and investigate effects of equivariance on network performance. A particular equivariant model, PELICAN, has its symmetry broken down with no to minimal architectural changes via both explicit and implicit methods. Equivariance is broken explicitly by supplying model inputs that are equivariant under strict Lorentz subgroups, while it is broken implicitly by adding spurious particles which imply laboratory-frame geometry. We compare its performance on common benchmark tasks in the equivariant and non-equivariant regimes.

## Significance

This talk presents a thorough investigation of the usefulness of equivariance in the context of a particular state-of-the-art neural network architecture. The topic of equivariance is a pressing issue as particle physics experiments continue to adopt machine learning methods for both data analysis and detector operation. to which this presentation will be a novel addition.

## References

Main pedagogical paper: https://arxiv.org/abs/2307.16506

## Experiment context, if any

**Authors:** BOGATSKIY, Alexander (Flatiron Institute, Simons Foundation); OFFERMANN, Jan Tuzlic (University of Chicago (US)); HOFFMAN, Timothy

**Presenters:** BOGATSKIY, Alexander (Flatiron Institute, Simons Foundation); HOFFMAN, Timothy

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **124** Type: **Poster**

# A Microbenchmark Framework for Performance Evaluation of OpenMP Target Offloading

*Thursday, March 14, 2024 4:10 PM (30 minutes)*

We present a framework based on Catch2 to evaluate performance of OpenMP's target offload model via micro-benchmarks. The compilers supporting OpenMP's target offload model for heterogeneous architectures are currently undergoing rapid development. These developments influence performance of various physics applications in different ways. This framework can be employed to track the impact of compiler upgrades and compare their performance with the native programming models. We use the framework to benchmark performance of a few commonly used operations on leadership class supercomputers such as Perlmutter at National Energy Research Scientific Computing (NERSC) Center and Frontier at Oak Ridge Leadership Computing Facility (OLCF). Such a framework will be useful for compiler developers to gain insights into the overall impact of many small changes, as well as for users to decide which compilers and versions are expected to yield best performance for their applications.

## Significance

Several portable programming models have been developed in the last decade as a solution to avoid code duplication and diversion for different GPU backends. This work focuses on OpenMP's target offload model which has undergone rapid development in the past few years and gained increasing vendor support. Simplified algorithmic performance benchmarks provide a good overview of the current compiler support for OpenMP target offload, however, they often lack the granularity to evaluate the performance of specific, often specialized, operations, such as atomic, memset or scan. We have encountered several issues related to these operations when porting application codes to OpenMP, which motivated us to develop necessary tools for the
better evaluation of specific operations in OpenMP target offload.

## References

## Experiment context, if any

**Authors:** LIN, Meifeng; ATIF, Mohammad (Brookhaven National Laboratory); WANG, Tianle (Brookhaven National Lab); DONG, Zhihua

**Presenter:** ATIF, Mohammad (Brookhaven National Laboratory)

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **125** Type: **Poster**

# Porting and optimizing the performance of LArTPC Detector Simulations with C++ standard parallelism

*Thursday, March 14, 2024 4:10 PM (30 minutes)*

There is a significant expansion in the variety of hardware architectures these years, including different GPUs and other specialized computing accelerators. For better performance portability, various programming models are developed across those computing systems, including Kokkos, SYCL, OpenMP, and others. Among these programming models, the C++ standard parallelism (std::par) has gained considerable attention within the community. Its inclusion as a part of the C++ standard library underscores its significance and potential impact, and it is also supported on AMD and Intel GPU recently.
As part of the High Energy Physics Center for Computational Excellence (HEP-CCE) project, we investigate if and how std::par may be suitable for experimental HEP workflows with some representative use cases. One of such use cases is the Liquid Argon Time Projection Chamber (LArTPC) simulation which is essential for LArTPC detector design, validation and data analysis. Following our earlier work of using Kokkos, OpenMP, and SYCL to port LArTPC simulations module, we are going to present the following topics: 1). How std::par is currently supported on different architectures and compiler, and comparison with other programming models; 2). Lesson learned from optimizing kernels with std::par; 3). Advantages and disadvantages of using std::par in porting LArTPC simulation and other HEP programs.

## Significance

In this presentation, we intend to show the feasibility of porting HEP applications using C++ standard parallelism. As part of the C++ standard, we believe this will greatly interest the broader HEP community.
std::par was previously only supported by limited types of GPU architecture and compiler. However, recently, std::par has been supported on AMD GPU by llvm and intel GPU by onedpl. Because of that we would like to present how it performs on a wider range of architecture and the experience we learned to improve the performance using std::par.

## References

1 Yu, Haiwang; Dong, Zhihua; Knoepfel, Kyle; Lin, Meifeng; Viren, Brett; Yu, Kwangmin; Evaluation of Portable Acceleration Solutions for LArTPC Simulation Using Wire-Cell Toolkit,EPJ Web of Conferences,251,03032,2021,EDP Sciences

2 Dong, Zhihua; Knoepfel, Kyle; Lin, Meifeng; Viren, Brett; Yu, Haiwang; Evaluation of Portable Programming Models to Accelerate LArTPC Detector Simulations, arXiv preprint arXiv:2203.02479,2022

[3] M Lin; Z Dong; T Wang; M Atif; M Battacharya; K Knoepfel; C Leggett; B Viren; H Yu; Portable Programming Model Exploration for LArTPC Simulation in a Heterogeneous Computing Environment: OpenMP vs. SYCL, arXiv preprint arXiv:2304.01841,2023

## Experiment context, if any

The Liquid Argon Time Projection Chamber (LArTPC) technology is widely used in high energy physics experiments, including the upcoming Deep Underground Neutrino Experiment (DUNE).

**Authors:** Dr LEGGETT, Charles (Lawrence Berkeley National Lab (US)); LIN, Meifeng; ATIF, Mohammad (Brookhaven National Laboratory); WANG, Tianle (Brookhaven National Lab); DONG, Zhihua

**Presenter:** WANG, Tianle (Brookhaven National Lab)

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **126** Type: **Oral**

# Reducing Systematic Differences between Data and Simulation with Generative Models

*Wednesday, March 13, 2024 2:50 PM (20 minutes)*

High Energy Physics (HEP) experiments rely on scientific simulation to develop reconstruction algorithms. Despite the remarkable fidelity of modern simulation frameworks, residual discrepancies between simulated and real data introduce a challenging domain shift problem. The existence of this issue raises significant concerns regarding the feasibility of implementing Deep Learning (DL) methods in detector analysis, impeding the adoption of such techniques.

We present our ongoing research in developing new DL strategies to mitigate the differences between simulation and real data. Our approach is based on a combination of Generative Adversarial Networks (GANs) for the translation of simulated samples into the real data domain, reducing the magnitude of domain shift effects.

We discuss our progress made in applying this approach specifically to LArTPC-based particle detectors. We demonstrate the effectiveness of our method on a simplified and cropped LArTPC benchmark dataset. Then we highlight various performance and computational challenges encountered in the process of adapting the method to realistic LArTPC datasets of high-resolution images ($\sim 6000 \times 960$ pixels). By systematically addressing these challenges, our research aims to bridge the gap between the simulated and real data and advance the applicability of DL methods in HEP experiments.

## Significance

We successfully applied the proposed method on a toy problem (https://arxiv.org/abs/2304.12858). The novelty here is that we managed to make the method work on a realistic dataset of high-resolution images and made the method computationally efficient.

## References

https://arxiv.org/abs/2304.12858

## Experiment context, if any

ProtoDUNE, DUNE

**Author:** TORBUNOV, Dmitrii

**Presenter:** TORBUNOV, Dmitrii

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **127** Type: **Poster**

# columnflow: Fully automated analysis through flow of columns over arbitrary, distributed resources

*Thursday, March 14, 2024 4:10 PM (30 minutes)*

To study and search for increasingly rare physics processes at the LHC, a staggering amount of data needs to be analyzed with progressively complex methods. Analyses involving tens of billions of recorded and simulated events, multiple machine learning algorithms for different purposes, and an amount of 100 or more systematic variations are no longer uncommon. These conditions impose a complex data flow on an analysis workflow and render its steering and bookkeeping a serious challenge.
For this purpose, a toolkit for columnar HEP analysis, called *columnflow*, has been developed. It is written in Python, experiment agnostic in its core, and supports any flat file format, such as ROOT-based trees or Parquet files. Leveraging on the vast Python ecosystem, vectorization and convenient physics objects representation can be achieved through NumPy, awkward arrays and other libraries. Based upon the Luigi Analysis Workflow (*law*) package, *columnflow* provides full analysis automation over arbitrary, distributed computing resources. Despite the end-to-end nature, this approach allows for persistent, intermediate outputs for purposes of debugging, caching, and exchange with collaborators. Job submission to various batch systems, such as HTCondor, Slurm, or CMS-CRAB, is natively supported. Remote files can be seamlessly accessed via various protocols using either the Grid File Access Library (GFAL2) or the fsspec file system interface. In addition, a sandboxing mechanism can encapsulate the execution of parts of a workflow into dedicated environments, supporting subshells, virtual environments, and containers.
This contribution introduces the key components of *columnflow* and highlights the benefits of a fully automated workflow for complex and large-scale HEP analyses, showcasing an implementation of the Analysis Grand Challenge.

## Significance

## References

## Experiment context, if any

CMS

**Authors:** WIEDERSPAN, Bogdan (Hamburg University (DE)); RIEGER, Marcel (Hamburg University (DE))

**Presenter:** WIEDERSPAN, Bogdan (Hamburg University (DE))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **128** Type: **Oral**

# Precision-Machine Learning for the Matrix Element Method

*Tuesday, March 12, 2024 12:10 PM (20 minutes)*

The matrix element method is the LHC inference method of choice for limited statistics. We present a dedicated machine learning framework, based on efficient phase-space integration, a learned acceptance and transfer function. It is based on a choice of INN and diffusion networks, and a transformer to solve jet combinatorics. Bayesian networks allow us to capture network uncertainties, bootstrapping allows us to estimate integration uncertainties. We showcase this setup for the CP-phase of the top Yukawa coupling in associated Higgs and single-top production.

## Significance

## Experiment context, if any

## References

Paper: arXiv: 2310.07752 ;
Paper from 2022 that we are building on: arXiv: 2210.00019 ;

Slides: https://indico.cern.ch/event/1311972/contributions/5705529/attachments/2773167/4832338/Wien2023.pdf

**Authors:** BUTTER, Anja (Centre National de la Recherche Scientifique (FR)); HUETSCH, Nathan (Heidelberg University, ITP Heidelberg); WINTERHALDER, Ramon (UCLouvain); HEIMEL, Theo (Heidelberg University); PLEHN, Tilman (Heidelberg University)

**Presenter:** HUETSCH, Nathan (Heidelberg University, ITP Heidelberg)

**Session Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

**Track Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

Contribution ID: **129** Type: **Poster**

# Performance of the Gaussino CaloChallenge-compatible infrastucture for ML-based fast simulation in the LHCb Experiment

*Thursday, March 14, 2024 4:10 PM (30 minutes)*

Efficient fast simulation techniques in high energy physics experiments are crucial in order to produce the necessary amount of simulated samples. We present a new component in the Gaussino core simulation framework that facilitates the integration of fast simulation hooks in Geant4 with machine learning serving, based on Gaudi's scheduling and data processing tools. The implementation supports both PyTorch and ONNXRuntime.

We will also show how this new component can be used to integrate generic ML models developed within the scope of CaloChallenge, a collaborative community-wide initiative aiming to develop and benchmark ML models for modeling of calorimeter shower. A few simple examples within the Gaussino framework, including full observability of the inferred variables, as well as conversion mechanisms to the experiment's event model, will be shown.

Finally, we will present the very first, production-ready implementation of a ML-based fast simulation model for electromagnetic showers in the calorimeter of the LHCb experiment. It is a Vartiational Autoencoder (VAE) with a custom sampling head that increases throughput and improves energy precision. Performance and results of the model and infrastructure, along with insights gained from its utilization, will be presented.

## Significance

In this talk, the very first, production-ready ML model for fast simulation of calorimeter showers in the electromagnetic calorimeter in the LHCb Experiment will be presented. The inference/training of that model is based on the adaptation of the CaloChallenge setup and generic ML infrastructure in Gaussino, the experiment-agnostic core simulation framework, that can be used by any other HEP experiment.

## References

https://indico.jlab.org/event/459/contributions/11528/
https://indico.jlab.org/event/459/contributions/11549/
https://iopscience.iop.org/article/10.1088/1742-6596/2438/1/012108
https://doi.org/10.22323/1.414.0225
https://www.ncbj.gov.pl/seminaria/prototypes-large-scale-detectors-monte-carlo-simulations
https://doi.org/10.1109/NSS/MIC42101.2019.9060074
https://doi.org/10.31577/cai_2021_4_815

## Experiment context, if any

LHCb Experiment

**Author:** MAZUREK, Michal (CERN)

**Co-authors:** CORTI, Gloria (CERN); KMIEC, Mateusz (National Centre for Nuclear Research (PL))

**Presenter:** MAZUREK, Michal (CERN)

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **130** Type: **Plenary**

# Welcome to ACAT 2024 in Stony Brook

*Monday, March 11, 2024 9:00 AM (15 minutes)*

## Experiment context, if any

## References

## Significance

**Presenter:** Dr LAURET, Jerome (Brookhaven National Laboratory)

**Session Classification:** Plenary

Contribution ID: **131** Type: **Plenary**

# Recent advances in neuromorfic computing

*Monday, March 11, 2024 10:10 AM (30 minutes)*

**Presenter:** AKOPYAN, Filipp (IBM)

**Session Classification:** Plenary

Contribution ID: **132** Type: **Plenary**

# AI and Ethics

*Monday, March 11, 2024 11:20 AM (1 hour)*

**Presenter:** CONITZER, Vincent (University of Oxford)

**Session Classification:** Plenary

Contribution ID: **133** Type: **Plenary**

# Long-term high-volume data storage in ceramic

*Thursday, March 14, 2024 9:00 AM (30 minutes)*

## Experiment context, if any

## References

## Significance

**Presenters:** PFLAUM, Christian (Cerabyte); HELLMOLD, Steffen (Cerabyte, Inc.)

**Session Classification:** Plenary

Contribution ID: **134** Type: **Plenary**

# Intersection of machine learning and quantum physics

*Friday, March 15, 2024 9:15 AM (30 minutes)*

**Presenter:** YOU, Yi-Zhuang (UCSD)

**Session Classification:** Plenary

Contribution ID: **135** Type: **Plenary**

# Q&A Panel

*Monday, March 11, 2024 12:20 PM (40 minutes)*

## Experiment context, if any

## References

## Significance

**Presenters:** CIPRIJANOVIC, Aleksandra (Fermi National Accelerator Laboratory); SEXTON-KENNEDY, Elizabeth (Fermi National Accelerator Lab. (US)); WATTS, Gordon (University of Washington (US)); CONITZER, Vincent (CMU/Oxford)

**Session Classification:** Plenary

Contribution ID: **136** Type: **Plenary**

# Digital Twins from Industry to Science

*Wednesday, March 13, 2024 9:45 AM (30 minutes)*

**Presenter:** GIBBS, Tom (Nvidia)

**Session Classification:** Plenary

Contribution ID: **137** Type: **Plenary**

# What will it take to do a HL-LHC analysis in 15'?

*Thursday, March 14, 2024 12:00 PM (30 minutes)*

**Presenter:** GRAY, Lindsey (Fermi National Accelerator Lab. (US))

**Session Classification:** Plenary

Contribution ID: **138** Type: **Plenary**

# Interfaces, discovery, and intelligence in Digital Earth Observation

*Thursday, March 14, 2024 9:30 AM (30 minutes)*

**Presenter:** LUMNITZ, Stefanie (ESA)

**Session Classification:** Plenary

Contribution ID: **139** Type: **Plenary**

# Quantum computers for particle theory

*Wednesday, March 13, 2024 11:30 AM (30 minutes)*

## Experiment context, if any

## References

## Significance

**Presenter:** CARRAZZA, Stefano (CERN)

**Session Classification:** Plenary

Contribution ID: **140** Type: **Plenary**

# Updates from the organizers

*Tuesday, March 12, 2024 9:00 AM (15 minutes)*

## Experiment context, if any

## References

## Significance

**Session Classification:** Plenary

Contribution ID: **141** Type: **Plenary**

# Astronomical foundation models & the Polymathic AI Initiative

*Tuesday, March 12, 2024 9:15 AM (30 minutes)*

**Presenter:** GOLKAR, Siavash (Flat Iron Institute)

**Session Classification:** Plenary

Contribution ID: **142** Type: **Plenary**

# The co-evolution of HPC computing and LQCD

*Thursday, March 14, 2024 10:30 AM (30 minutes)*

**Presenter:** CHRIST, Norman (Columbia University)

**Session Classification:** Plenary

Contribution ID: **143** Type: **Plenary**

# Updates from the organizers

*Wednesday, March 13, 2024 9:00 AM (15 minutes)*

## Experiment context, if any

## References

## Significance

**Session Classification:** Plenary

Contribution ID: **144** Type: **Plenary**

# Quantum algorithms for the simulation of QCD processes in the perturbative regime

*Friday, March 15, 2024 9:45 AM (30 minutes)*

## Experiment context, if any

## References

## Significance

**Presenter:** CHAWDHRY, Herschel (Florida State University)

**Session Classification:** Plenary

Contribution ID: **145** Type: **Plenary**

# AI and microelectronics for science

*Tuesday, March 12, 2024 9:45 AM (30 minutes)*

**Presenter:** TRAN, Nhan (Fermi National Accelerator Lab. (US))

**Session Classification:** Plenary

Contribution ID: **146** Type: **Plenary**

# The Open Compute Project and its application at Meta

*Wednesday, March 13, 2024 10:15 AM (30 minutes)*

**Presenter:** HELVIE, Steve (Open Compute Project)

**Session Classification:** Plenary

Contribution ID: **147** Type: **Plenary**

# Differentiable Programming in HEP

*Thursday, March 14, 2024 12:30 PM (30 minutes)*

**Presenter:** HEINRICH, Lukas Alexander (Technische Universitat Munchen (DE))

**Session Classification:** Plenary

Contribution ID: **148** Type: **Plenary**

# Making ML Robust for Physics Discovery

*Wednesday, March 13, 2024 12:00 PM (30 minutes)*

**Presenter:** WILLIAMS, J Michael (Massachusetts Inst. of Technology (US))

**Session Classification:** Plenary

Contribution ID: **149** Type: **Plenary**

# Mojo: A novel programming language for AI

*Friday, March 15, 2024 10:15 AM (30 minutes)*

**Presenter:** CLAYTON, Jack (Modular)

**Session Classification:** Plenary

Contribution ID: **150** Type: **Plenary**

# Updates from the organizers

### Experiment context, if any

### References

### Significance

**Session Classification:** Plenary

Contribution ID: **151** Type: **Plenary**

# Exascale infrastructures for Science

*Friday, March 15, 2024 8:45 AM (30 minutes)*

**Presenter:** BARD, Deborah (NERSC)

**Session Classification:** Plenary

Contribution ID: **152** Type: **Plenary**

# Sustainable computing and the MLPerf project

*Tuesday, March 12, 2024 10:15 AM (30 minutes)*

**Presenter:** WU, Carole-Jean (MLPerf)

**Session Classification:** Plenary

Contribution ID: **153** Type: **Plenary**

# Scalable neural networks for event reconstruction at current and future colliders

*Thursday, March 14, 2024 10:00 AM (30 minutes)*

**Presenter:** PATA, Joosep (National Institute of Chemical Physics and Biophysics (EE))

**Session Classification:** Plenary

Contribution ID: **154** Type: **Plenary**

# Updates from the organizers

### Experiment context, if any

### References

### Significance

**Session Classification:** Plenary

Contribution ID: **156** Type: **Plenary**

# Track 1 Summary

*Friday, March 15, 2024 11:30 AM (20 minutes)*

## Experiment context, if any

## References

## Significance

**Presenter:** HEGNER, Benedikt (CERN)

**Session Classification:** Plenary

Contribution ID: **157** Type: **Plenary**

# Track 2 Summary

*Friday, March 15, 2024 11:50 AM (20 minutes)*

## Experiment context, if any

## References

## Significance

**Presenter:** HEINRICH, Lukas Alexander (Technische Universitat Munchen (DE))

**Session Classification:** Plenary

Contribution ID: **158** Type: **Plenary**

# Track 3 Summary

*Friday, March 15, 2024 12:10 PM (20 minutes)*

## Experiment context, if any

## References

## Significance

**Presenter:** SIGNORILE , Chiara

**Session Classification:** Plenary

Contribution ID: **159** Type: **Plenary**

# ACAT 2024 Conclusion and Outlook

*Friday, March 15, 2024 12:30 PM (30 minutes)*

## Experiment context, if any

## References

## Significance

**Presenters:** BRITTON, David (University of Glasgow (GB)); NGADIUBA, Jennifer (FNAL); Dr LAURET, Jerome (Brookhaven National Laboratory)

**Session Classification:** Plenary

Contribution ID: **160** Type: **Plenary**

# Detecting rare events using artificial intelligence

*Thursday, March 14, 2024 11:30 AM (30 minutes)*

**Presenter:** LI, Aobo (UCSD)

**Session Classification:** Plenary

Contribution ID: **161** Type: **Plenary**

# Plenary

**Session Classification:** Plenary

Contribution ID: **163** Type: **Oral**

# Tracking and vertexing downstream the LHCb magnet at the first stage of the trigger.

*Monday, March 11, 2024 3:50 PM (20 minutes)*

A new algorithm, called "Downstream", has been developed at LHCb which is able to reconstruct and select very displaced vertices in real time at the first level of the trigger (HLT1). It makes use of the Upstream Tracker (UT) and the Scintillator Fiber detector (SciFi) of LHCb and it is executed on GPUs inside the Allen framework. In addition to an optimized strategy, it utilizes a Neural Network (NN) implementation to increase the track efficiency and reduce the ghost rates, with very high throughput and limited time budget. Besides serving to reconstruct Ks and Lambda vertices to calibrate and align the detectors, the Downstream algorithm and the associated two-track vertexing will largely increase the LHCb physics potential for detecting long-lived particles during the Run3.

## Significance

The algorithms developed under this work hugely increases the LHCb potential to discover new physics.

## References

https://arxiv.org/abs/2312.14016

## Experiment context, if any

LHCb

**Authors:** DE OYANGUREN CAMPOS, Arantza (Univ. of Valencia and CSIC (ES)); Dr JASHAL, Brij Kishor (IFIC, Univ. of Valencia, CSIC (ES) and Tata Institute of Fundamental Research (TIFR)); ZHUO, Jiahui (Univ. of Valencia and CSIC (ES)); KHOLOIMOV, Valerii (IFIC - Valencia); SVINTOZELSKYI, Volodymyr (IFIC - Valencia)

**Presenter:** SVINTOZELSKYI, Volodymyr (IFIC - Valencia)

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **164** Type: **Plenary**

# Welcome address by SBU

*Monday, March 11, 2024 9:15 AM (15 minutes)*

## Experiment context, if any

## References

## Significance

**Presenter:** Prof. LEJUEZ, Carl (Stony Brook University)

**Session Classification:** Plenary

Contribution ID: **165** Type: **Plenary**

# Science at SBU/BNL

*Monday, March 11, 2024 9:30 AM (40 minutes)*

## Experiment context, if any

## References

## Significance

**Presenter:** DESHPANDE, Abhay

**Session Classification:** Plenary

Contribution ID: **166** Type: **Plenary**

# Building Foundational Models for Environmental Modelling and Prediction

*Wednesday, March 13, 2024 9:15 AM (30 minutes)*

**Presenter:** LUISE, Ilaria (CERN)

**Session Classification:** Plenary

Contribution ID: **167** Type: **Plenary**

# Computing the wave: Where the Gravitational Wave Community benefits from HEP, and where it differs?

*Wednesday, March 13, 2024 12:30 PM (30 minutes)*

## Experiment context, if any

## References

## Significance

**Presenter:** MEYER-CONDE, Marco (Osaka Metropolitan University (JP), Tokyo City University (JP))

**Session Classification:** Plenary

Contribution ID: **168** Type: **Oral**

# CLAS12 remote data-stream processing using ERSAP framework

*Monday, March 11, 2024 3:30 PM (20 minutes)*

Implementing a physics data processing application is relatively straightforward with the use of current containerization technologies and container image runtime services, which are prevalent in most high-performance computing (HPC) environments. However, the process is complicated by the challenges associated with data provisioning and migration, impacting the ease of workflow migration and deployment. Transitioning from traditional file-based batch processing to data-stream processing workflows is suggested as a method to streamline these workflows. This transition not only simplifies file provisioning and migration but also significantly reduces the necessity for extensive disk space. Data-stream processing is particularly effective for real-time processing during data acquisition, thereby enhancing data quality assurance. This paper introduces the integration of the JLAB CLAS12 event reconstruction application within the ERSAP data-stream processing framework that facilitates the execution of streaming event reconstruction at a remote data center and enables the return streaming of reconstructed events to JLAB while circumventing the need for temporary data storage throughout the process.

## Significance

## References

## Experiment context, if any

CLAS12

**Author:** Dr GYURJYAN, Vardan

**Co-authors:** TIMMER, Carl; LAWRENCE, David; Dr HOWARD, Derek; HEYES, Graham (Jefferson Lab); TSAI, Jeng-Yuan; Dr GOODRICH, Michael; Dr TYLER, Nicholas; Dr SHELDON, Stacey; Dr KUMAR, Yatish

**Presenter:** Dr GYURJYAN, Vardan

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **169** Type: **Oral**

# Effective denoising diffusion probabilistic models for fast and high fidelity whole-event simulation in high-energy heavy-ion experiment

*Thursday, March 14, 2024 3:50 PM (20 minutes)*

AI generative models, such as generative adversarial networks (GANs), variational auto-encoders, and normalizing flows, have been widely used and studied as efficient alternatives for traditional scientific simulations, such as Geant4. However, they have several drawbacks such as training instability and unable to cover the entire data distribution especially for the region where data are rare. This is particularly challenging for whole-event full-detector simulations in the high-energy heavy-ion experiments, such as sPHENIX at RHIC and LHC experiments, where thousands of particles are produced per event and interact through the detector. AI-based surrogate models need to reproduce short-range and long-range patterns, and their energy spread, all of which stems from the evolution of the quark-gluon plasma produced in the collisions and its detector response.

Here, we investigate the effectiveness of denoising diffusion probabilistic models (DDPM) as an AI-based generative surrogate model for sPHENIX experiment that include both the heavy ion event generation and its response in the whole calorimeter stack. DDPM, a new type of AI generative model, underpins the recent generative AI success such as Stable Diffusion and Mid-Journey. For photographic images and arts, it can outperform previous generative models in image quality, data diversity and training stability. We study its performance in sPHENIX data compared with a popular rival –GANs. Our results show that both DDPM and GAN can reproduce the data distribution where the examples are abundant (low to medium tower energies). But DDPM significantly outperforms GANs in overall distribution distance and high-tower energy regions, where events are rare. The results are consistent between both central and peripheral centrality heavy ion collision events.

## Significance

This work represents the first use of the diffusion model in the full detector full event simulation of heavy ion experiments. Our approach is self-supervising and data-driven that results in faithful reproduction of the full calorimeter detector data. Compared with traditional Geant4-based simulation, this approach leads to two orders of magnitude faster speed gain.

## References

## Experiment context, if any

This exploratory work uses sPHENIX simulation software and simulation data for demonstration. However, this work is developed independently of the sPHENIX collaboration.

**Authors:** TORBUNOV, Dmitrii; Dr HUANG, Jin (Brookhaven National Lab); RINN, Timothy

Thomas (Brookhaven National Laboratory); GO, Yeonju (Brookhaven National Laboratory (US)); REN, Yihui

**Presenter:** GO, Yeonju (Brookhaven National Laboratory (US))

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **170** Type: **Poster**

# Celeritas: evaluating performance of HEP detector simulation on GPUs

*Thursday, March 14, 2024 4:10 PM (30 minutes)*

Celeritas is a Monte Carlo (MC) detector simulation library that exploits current and future heterogeneous leadership computing facilities (LCFs). It is specifically designed for, but not limited to, High-Luminosity Large Hadron Collider (HL-LHC) simulations. Celeritas implements full electromagnetic (EM) physics, supports complex detector geometries, and runs on CPUs and Nvidia or AMD GPUs. Celeritas provides a simple interface to integrate seamlessly with Geant4 applications such as CMSSW and ATLAS FullSimLight.
Using EM-only benchmark problems, we show that one A100 GPU is equivalent to 32-240 EPYC CPU cores on the Perlmutter supercomputer. In a test beam application using the ATLAS tile calorimeter geometry and full hadronic physics simulated by Geant4, offloading EM particles to Celeritas results in a 3x overall speedup on GPU and 1.2x on CPU.
We will present the current capabilities, focusing on performance results including recent optimization work, power efficiency, and throughput improvement.

## Significance

Heterogeneous architectures are increasingly more common, particularly within the TOP500 systems. LHC experiments such as ATLAS and CMS spend a significant amount of their computing budget on detector simulation traditionally done on CPUs. With the upcoming HL-LHC, the data complexity and quantity will significantly increase, challenging the current simulation software. This work will enable experiments to use GPUs for detector simulations.

## References

https://indico.jlab.org/event/459/contributions/11818/

## Experiment context, if any

ATLAS,CMS

**Authors:** LUND, Amanda; MORGAN, Benjamin (University of Warwick); ESSEIVA, Julien (Lawrence Berkeley National Lab. (US)); JOHNSON, Seth (Oak Ridge National Laboratory (US)); JUN, Soon Yung (Fermi National Accelerator Lab. (US))

**Presenter:** ESSEIVA, Julien (Lawrence Berkeley National Lab. (US))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **172** Type: **Oral**

# Improving Computational Performance of a GNN Track Reconstruction Pipeline for ATLAS

*Tuesday, March 12, 2024 12:10 PM (20 minutes)*

Track reconstruction is an essential element of modern and future collider experiments, including within the ATLAS detector. The HL-LHC upgrade of the ATLAS detector brings an unprecedented tracking challenge, both in terms of number of silicon hit cluster readouts, and throughput required for both high level trigger and offline track reconstruction. Traditional track reconstruction techniques often contain steps that scale combinatorically, which could be ameliorated with deep learning approaches. The GNN4ITk project has been shown to apply geometric deep learning algorithms for tracking to a similar level of physics performance with traditional techniques, while scaling sub-quadratically. In this contribution, we provide comparisons of physics and computational performance across a variety of model configurations, as well as optimizations that reduce computational cost without significantly affecting physics performance. These include the use of structured pruning, knowledge distillation, simplified and customized convolutional kernels, regional tracking approaches, and GPU-optimized graph segmentation techniques.

## Significance

This represents the first set of computational performance results for the novel GNN-based tracking pipeline for the upgraded ATLAS ITk subdetector. This proves that the approach is realistic for both physics requirements and compute budgets.

## References

https://cds.cern.ch/record/2882507/files/ATL-SOFT-PROC-2023-047.pdf
https://arxiv.org/abs/2103.06995
https://arxiv.org/abs/2103.00916

## Experiment context, if any

ATLAS

**Authors:** SHMAKOV, Alexander (University of California Irvine (US)); VALLIER, Alexis (L2I Toulouse, CNRS/IN2P3, UT3); Dr COLLARD, Christophe (Laboratoire des 2 Infinis - Toulouse, CNRS / Univ. Paul Sabatier); MURNANE, Daniel Thomas (Lawrence Berkeley National Lab. (US)); TORRES, Heberth (L2I Toulouse, CNRS/IN2P3, UT3); STARK, Jan (Laboratoire des 2 Infinis - Toulouse, CNRS / Univ. Paul Sabatier (FR)); BURLESON, Jared (University of Illinois at Urbana-Champaign); CHAN, Jay (Lawrence Berkeley National Lab. (US)); NEUBAUER, Mark (Univ. Illinois at Urbana Champaign (US)); PHAM, Minh-Tuan (University of Wisconsin Madison (US)); CALAFIURA, Paolo (Lawrence Berkeley National Lab. (US)); CAILLOU, Sylvain (Centre National de la Recherche Scientifique (FR)); JU, Xiangyang (Lawrence Berkeley National Lab. (US))

**Presenter:** MURNANE, Daniel Thomas (Lawrence Berkeley National Lab. (US))

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **173** Type: **Oral**

# Machine learning based surrogates for particle-resolved direct numerical simulation

*Tuesday, March 12, 2024 12:50 PM (20 minutes)*

In atmospheric physics, particle-resolved direct numerical simulation (PR-DNS) models constitute an important tool to study aerosol-cloud-turbulence interactions which are central to the prediction of weather and climate . They resolve the smallest turbulent eddies as well as track the development and motion of individual particles [1,2]. PR-DNS is expected to complement experimental and observational facilities to further our understanding of the cloud dynamics. With a sufficient computational speed and scale, it can also be part of the digital twin to the experimental facilities such as the cloud chambers [2,3]. The original version of PR-DNS does not scale well on multi-core CPUs, thereby limiting the scale it can simulate. There is therefore a great need to accelerate the PR-DNS solver with tools from traditional high performance computing (HPC) as well as replacing computationally expensive modules with machine learning models. In this study, we exploit the potential of Fourier Neural Operators (FNO) [4] learning to yield fast and accurate surrogate models for the velocity and vorticity fields. We have investigated two classes of FNO –2D FNO which has two spatial dimensions with a recurrent structure in time and 3D FNO with two spatial and one temporal dimensions. We discuss results from numerical experiments designed to assess the performance of these architectures as well as their suitability for capturing the behavior of relevant dynamical systems.

1 Zheng Gao et al. Investigation of turbulent entrainment-mixing processes with a new particle-resolved direct numerical simulation model. J. Geophys. Res. Atmos. 123(4), 2018.
2 Lulin Xue et al. Progress and challenges in modeling dynamics–microphysics interactions: From the Pi chamber to monsoon convection. Bull Am Meteorol Soc, 103(5), 2022.
[3] K. Chang et al. A laboratory facility to study gas–aerosol–cloud interactions in a turbulent environment: The  chamber. Bull Am Meteorol Soc, 97(12), 2016.
[4] Z. Li et al. Fourier neural operator for parametric partial differential equations, 2021. https://arxiv.org/abs/2010.08895

## Significance

This presentation discusses the speedup gained by using a machine learning (ML) model as a surrogate for a traditional partial differential equation solver. The corresponding errors from two different ML architectures in the context of atmospheric physics, climate science, and turbulence will also be discussed.

## References

https://arxiv.org/pdf/2312.12412.pdf
https://d197for5662m48.cloudfront.net/documents/publicationstatus/143045/preprint_pdf/6e1cbcdcef459f235515059f10113

## Experiment context, if any

**Authors:** AL MUTI SHARFUDDIN, Abdullah (Stony Brook University); Dr YANG, Fan (Brookhaven National Lab); Prof. LADEINDE, Foluso (Stony Brook University); Dr YU, Kwangmin (Brookhaven National Lab); Dr LI, Lingda (Brookhaven National Lab); Dr LIN, Meifeng (Brookhaven National Lab); ATIF, Mohammad (Brookhaven National Laboratory); Dr ZHANG, Tao (Brookhaven National Lab); Dr LÓPEZ-MARRERO, Vanessa (Brookhaven National Lab); Dr LIU, Yangang (Brookhaven National Lab)

**Presenter:** ATIF, Mohammad (Brookhaven National Laboratory)

**Session Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

**Track Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

Contribution ID: **174** Type: **Poster**

# Portable acceleration of CMS computing workflow with coprocessors as a service

*Thursday, March 14, 2024 4:10 PM (30 minutes)*

Computing demands for large scientific experiments, such as the CMS experiment at the CERN LHC, will increase dramatically in the next decades. To complement the future performance increases of software running on central processing units (CPUs), explorations of coprocessor usage in data processing hold great potential and interest. Coprocessors are a class of computer processors that supplement CPUs, often improving the execution of certain functions due to architectural design choices. In this talk, I will introduce the approach of Services for Optimized Network Inference on Coprocessors (SONIC) and discuss the study of the deployment of this as-a-service approach in large-scale data processing.

In the studies, we take a data processing workflow of the CMS experiment and run the main workflow on CPUs, while offloading several machine learning (ML)inference tasks onto either remote or local coprocessors, specifically graphics processing units (GPUs). With experiments performed at Google Cloud, the Purdue Tier-2 computing center, and combinations of the two, we demonstrate the acceleration of these ML algorithms individually on coprocessors and the corresponding throughput improvement for the entire workflow. We will also show this approach can be easily generalized to different types of coprocessors and deployed on local CPUs without decreasing the throughput performance.

## Significance

## References

## Experiment context, if any

The CMS Experiment

**Author:** FENG, Yongbin (Fermi National Accelerator Lab. (US))

**Presenter:** FENG, Yongbin (Fermi National Accelerator Lab. (US))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **175** Type: **Oral**

# Pinpoint resource allocation for GPU batch applications

*Thursday, March 14, 2024 3:10 PM (20 minutes)*

With the increasing usage of Machine Learning (ML) in High Energy Physics (HEP), the breadth of new analyses with a large spread in compute resource requirements, especially when it comes to GPU resources. For institutes, like the Karlsruhe Institute of Technology (KIT), that provide GPU compute resources to HEP via their batch systems or the Grid, a high throughput, as well as energy efficient usage of their systems is of the essence. With low intensity GPU analyses specifically, inefficiencies are created by the standard scheduling, as resources are over-assigned to such workflows. An approach that is flexible enough to cover the entire spectrum, from multi-process per GPU, to multi-GPU per process, is necessary. As a follow-up to the techniques presented at the 2022 ACAT, this time we study Nvidia's multi-process service (MPS), its ability to securely distribute device memory and its interplay with the KIT HTCondor batch system. A number of ML applications were benchmarked using this less demanding and more flexible approach to illustrate the performance implications regarding throughput and energy efficiency.

## Significance

Batch systems are crucial for the efficient and high-throughput computing that is required in modern high energy physics. Often, these batch systems are limited by their coarse granularity. Especially for GPU resources, the safe sharing of high performance datacenter GPUs is necessary to avoid gross over-allocation of costly hardware, while still allowing for workflows that require multiple GPUs at once. Nvidia's multi-process service (MPS) enables this kind of flexibility, and we therefore consider it a valuable tool for our goal of high throughput and high energy efficiency.

## References

https://indico.cern.ch/event/1106990/contributions/4991345/

## Experiment context, if any

CMS

**Author:** VOIGTLAENDER, Tim (KIT - Karlsruhe Institute of Technology (DE))

**Co-authors:** QUAST, Gunter (KIT - Karlsruhe Institute of Technology (DE)); GIFFELS, Manuel (KIT - Karlsruhe Institute of Technology (DE)); SCHNEPF, Matthias Jochen; WOLF, Roger (KIT - Karlsruhe Institute of Technology (DE))

**Presenter:** VOIGTLAENDER, Tim (KIT - Karlsruhe Institute of Technology (DE))

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **176** Type: **Oral**

# New developments and applications of a Deep-learning-based Full Event Interpretation (DFEI) in proton-proton collisions

*Tuesday, March 12, 2024 12:30 PM (20 minutes)*

The LHCb experiment at the Large Hadron Collider (LHC) is designed to perform high-precision measurements of heavy-hadron decays, which requires the collection of large data samples and a good understanding and suppression of multiple background sources. Both factors are challenged by a five-fold increase in the average number of proton-proton collisions per bunch crossing, corresponding to a change in the detector operation conditions for the recently started LHC Run 3. The limits in the storage capacity of the trigger have brought an inverse relation between the number of particles selected to be stored per event and the number of events that can be recorded, and the background levels have risen due to the enlarged combinatorics. To tackle both challenges, we have proposed a novel approach, never attempted before in a hadronic collider: a Deep-learning based Full Event Interpretation (DFEI), to perform the simultaneous identification, isolation, and hierarchical reconstruction of all the heavy-hadron decay chains in each event. We have developed a prototype for such an algorithm based on Graph Neural Networks. The construction of the algorithm and its current performance have recently been described in a publication [Comput.Softw.Big Sci. 7 (2023) 1, 12]. This contribution will summarise the main findings in that paper. In addition, new developments towards speeding up the inference of the algorithm will be presented, as well as novel applications of DFEI for data analysis. The applications, showcased using simulated datasets, focus on decay-mode-inclusive studies and automated methods for background suppression/characterization.

## Significance

This work presents a novel approach for the trigger in hadronic colliders that shall reduce the average event size, maximizing the number of events the experiments can store. Furthermore, this new approach has several applications for physics analysis so far disregarded in hadronic machines.

## References

Comput.Softw.Big Sci. 7 (2023) 1, 12 (https://rdcu.be/dxee4)

## Experiment context, if any

LHCb

**Authors:** MATHAD, Abhijit (CERN); MAURI, Andrea (Imperial College (GB)); UZUKI, Azusa (University of Zurich (CH)); SOUZA DE ALMEIDA, Felipe Luan (Syracuse University (US)); ESCHLE, Jonas (University of Zurich (CH)); GARCIA PARDINAS, Julian (CERN); CALVI, Marta (Univ. degli Studi Milano-Bicocca); SERRA, Nicola (University of Zurich (CH)); SILVA COUTINHO, Rafael (Syracuse

University (US)); CAPELLI, Simone (Universita & INFN, Milano-Bicocca (IT)); SUTCLIFFE, William (University of Zurich (CH))

**Presenter:** SOUZA DE ALMEIDA, Felipe Luan (Syracuse University (US))

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **177** Type: **Oral**

# Quantum-centric Supercomputing for Physics Research

*Monday, March 11, 2024 5:10 PM (20 minutes)*

The rise of parallel computing, in particular graphics processing units (GPU), and machine learning and artificial intelligence has led to unprecedented computational power and analysis techniques. Such technologies have been especially fruitful for theoretical and experimental physics research where the embarrassingly parallel nature of certain workloads —e.g., Monte Carlo event generation, detector simulations, workflows, and data analysis —are exploited to attain significant performance improvements. Despite these capabilities, there still exist an array of problems that are manifestly intractable with classical computation alone, or for which classical computation provides only approximate or inefficient solutions.

Quantum computing is able to give exponential gains in both time and space for certain classes of problems. However, quantum workloads require significant classical computing support for preprocessing, including optimization and compilation, and postprocessing. This naturally leads to the concept of Quantum-centric Supercomputing (QCSC): the integration of quantum computational devices and high performance computing (HPC) resources. QCSC will leverage quantum and classical computing devices to enable execution of parallel and asynchronous hybrid workloads, unlocking the potential for computations beyond what is currently possible. IBM Quantum is engaging with the scientific and HPC communities to deliver them unrivaled quantum computing capabilities that will play a central role in the most powerful supercomputing systems in the world.

In this talk, we will give a brief overview of IBM Quantum's development roadmap and show how QCSC naturally fits this vision. We will explore ways in which the physics community and computational scientists can benefit from QCSC, and detail use cases suitable for these integrated systems.

## Significance

We will detail new technological developments from IBM Quantum which we believe physics research can benefit from. We hope to attract attention such that new collaborations can be built to expand the QCSC ecosystem; e.g., workflow management systems, programming models, use cases etc. Also, we want to engage with the physics research community, learn about their specific problems and what requirements they have from quantum, and how we can work together to bring useful quantum computing technologies to the community.

## References

Quantum-centric Supercomputing for Materials Science: A Perspective on Challenges and Future Directions (arXiv:2312.09733)

Building Towards QCSC (https://www.youtube.com/watch?v=L5PwmFnHCBI)

Quantum-centric Supercomputing (https://www.simonsfoundation.org/event/quantum-centric-supercomputing/)

## Experiment context, if any

ATLAS, CMS, DESI, DUNE, sPHENIX

**Authors:** CÓRCOLES, Antonio (IBM Quantum); PASCUZZI, Vincent R. (IBM Quantum)

**Presenter:** PASCUZZI, Vincent R. (IBM Quantum)

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **178** Type: **Oral**

# Offline filter of data with abnormal high voltage at BESIII drift chamber

*Thursday, March 14, 2024 2:30 PM (20 minutes)*

Stable operation of the detector is essential for high quality data taking in high energy physics experiment. But it is not easy to keep the detector always running stably during data taking period in environment with high beam induced background. In the BESIII experiment, serious beam related background may cause instability of the high voltages in the drift chamber which is the innermost sub detector. This could result in the decrease of gain and wrong dE/dx measurement. The relationship between the dE/dx measurement and the changes in high voltages has been studied. To guarantee the data quality for the physics study, an offline filter algorithm has been developed to remove the events with abnormal high voltages of the drift chamber. After applying the event filter on the data set with serious high voltage instability, the events with wrong dE/dx measurement were removed effectively.

## Significance

## References

## Experiment context, if any

**Authors:** WU, Linghui (Chinese Academy of Sciences (CN)); Mr ZHANG, Zeheng (Institute of High Energy Physics, Chinese Academy of Sciences)

**Co-authors:** Prof. LIU, Huaimin (Institute of High Energy Physics, Chinese Academy of Sciences); Dr WANG, Liangliang (Institute of High Energy Physics, Chinese Academy of Sciences)

**Presenter:** WU, Linghui (Chinese Academy of Sciences (CN))

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **179** Type: **Poster**

# Acceleration of the ML based fast simulation in high energy physics

*Thursday, March 14, 2024 4:10 PM (30 minutes)*

The diffusion model has demonstrated promising results in image generation, recently becoming mainstream and representing a notable advancement for many generative modeling tasks. Prior applications of the diffusion model for both fast event and detector simulation in high energy physics have shown exceptional performance, providing a viable solution to generate sufficient statistics within a constrained computational budget in preparation for the High Luminosity LHC. However, many of these applications suffer from slow generation with large sampling steps and face challenges in finding the optimal balance between sample quality and speed. The study focuses on the latest benchmark developments in efficient ODE/SDE-based samplers, schedulers, and fast convergence training techniques. We test on the public CaloChallenge and JetNet datasets with the designs implemented on the existing architecture, the performance of the generated classes surpass previous models, achieving significant speedup via various evaluation metrics.

## Significance

## References

## Experiment context, if any

**Author:** JIANG, Cheng (The University of Edinburgh (GB))

**Co-authors:** QU, Huilin (CERN); QIAN, Sitian (Peking University (CN))

**Presenter:** QIAN, Sitian (Peking University (CN))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **181** Type: **Oral**

# First experiences with the LHCb heterogeneous software trigger

*Thursday, March 14, 2024 5:50 PM (20 minutes)*

Since 2022, the LHCb detector is taking data with a full software trigger at the LHC proton-proton collision rate, implemented in GPUs in the first stage and CPUs in the second stage. This setup allows to perform the alignment & calibration online and to perform physics analyses directly on the output of the online reconstruction, following the real-time analysis paradigm.
This talk will give a detailed overview of the LHCb trigger implementation and its underlying computing infrastructure, discuss challenges of using a heterogeneous architecture and report on the experience from the first running periods in 2022 and 2023.

## Significance

This is the first full overview of the 2023 running period of the purely software-based trigger of LHCb, which is also the first year where a large amount of data was processed.

## References

## Experiment context, if any

LHCb

**Authors:** DE CIAN, Michel (Heidelberg University (DE)); BOETTCHER, Thomas (University of Cincinnati (US))

**Presenter:** BOETTCHER, Thomas (University of Cincinnati (US))

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **182** Type: **Poster**

# Study of columnar data analysis methods to complete an ATLAS analysis

*Thursday, March 14, 2024 4:10 PM (30 minutes)*

As the LHC continues to collect larger amounts of data, and in light of the upcoming HL-LHC, using tools that allow efficient and effective analysis of HEP data becomes more and more important. We present a test of the applicability and user-friendliness of several columnar analysis tools, most notably ServiceX and Coffea, by completing a full Run-2 ATLAS analysis. Working collaboratively with a group using traditional methods, we show that our columnar workflow can be used to achieve publishable results. Additionally, we will discuss the difficulties in adapting the workflow to ATLAS procedures, and our experience deploying this workflow at a supercomputer center.

## Significance

This study is to our knowledge the first use of serviceX + coffea to complete a full, publishable analysis at ATLAS. Our methods allow a look into how future physicists might approach analyses and provide valuable insight into the needs of users looking for more modern tools to complete the research goals of the collaboration.

## References

## Experiment context, if any

The physics analysis this poster is based on is an ATLAS study. We are anticipating that our paper will be on the arXiv before the start of ACAT 2024.

**Author:** TOST, Marc (University of Texas at Austin (US))

**Presenter:** TOST, Marc (University of Texas at Austin (US))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **183** Type: **Poster**

# AI-driven HPC Workflows Execution with Adaptivity and Asynchronicity in Mind

*Thursday, March 14, 2024 4:10 PM (30 minutes)*

With the increased integration of machine learning and the need for the scale of high-performance computing infrastructures, scientific workflows are undergoing a transformation toward greater heterogeneity. In this evolving landscape, adaptability has emerged as a pivotal factor in accelerating scientific discoveries through efficient execution of workflows. To increase resource utilization, reduce makespan, and minimize costs, it is essential to enable adaptive and asynchronous execution of heterogeneous tasks within scientific workflows. Consequently, middleware capable of scheduling and executing heterogeneous workflows must incorporate support for adaptive and asynchronous execution. We conduct an investigation into the advantages, prerequisites, and characteristics of a novel workflow execution middleware. Our proposed middleware dynamically adjusts the allocated resources for various task types based on historical execution data and executes them asynchronously. Through a comprehensive analysis, we elucidate how different degrees of asynchronicity impact workflow performance. Furthermore, we demonstrate the benefits in terms of performance and resource utilization by executing a real-world workflow (XYZ) at scale, using our execution middleware.

## Significance

This work will enable more suitable execution models for AI/ML-coupled scientific workflows. With adaptive and asynchronous execution, highly heterogeneous physics workflows can increase resource utilization and reduce the cost of execution. Hence, this will enable faster and cheaper scientific discoveries.

## References

https://link.springer.com/chapter/10.1007/978-3-031-43943-8_2

## Experiment context, if any

**Author:** KILIC, Ozgur Ozan (Brookhaven National Laboratory)

**Co-authors:** TURILLI, Matteo (Rutgers University); Prof. JHA, Shantenu (Rutgers University)

**Presenter:** KILIC, Ozgur Ozan (Brookhaven National Laboratory)

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **184** Type: **Oral**

# Fast and Robust ML for uncovering BSM physics

*Wednesday, March 13, 2024 5:10 PM (20 minutes)*

Navigating the demanding landscapes of real-time and offline data processing at the Large Hadron Collider (LHC) requires the deployment of fast and robust machine learning (ML) models for advancements in Beyond Standard Model (SM) discovery. This presentation explores recent breakthroughs in this realm, focusing on the use of knowledge distillation to imbue efficient model architectures with essential inductive bias. Additionally, novel techniques in robust multi-background representation learning for detecting out-of-distribution BSM signatures will be discussed, emphasizing the potential of these approaches in propelling discoveries within the challenging LHC environment.

## Significance

Fast and robust ML will be required when analyzing very high data rates in the ear of HL-LHC. These techniques will go beyond the conventional tools to address these issues

## References

arXiv:2401.08777, arXiv:2311.17162, and arXiv:2311.14160

## Experiment context, if any

Related to (HL)-LHC experiments

**Authors:** GANDRAKOTA, Abhijith (Fermi National Accelerator Lab. (US)); NGADIUBA, Jennifer (FNAL); LIU, Ryan

**Co-author:** TRAN, Nhan (Fermi National Accelerator Lab. (US))

**Presenter:** GANDRAKOTA, Abhijith (Fermi National Accelerator Lab. (US))

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **185** Type: **Oral**

# Multiscale Lattice Gauge Theory Algorithms and Software for Exascale hardware.

*Thursday, March 14, 2024 2:30 PM (20 minutes)*

I discuss software and algorithm development work in the lattice gauge theory community to develop performance portable software across a range of GPU architectures (Nvidia, AMD and Intel) and corresponding multi scale aware algorithm research to accelerate computation.
An example is given of a large effort to calculate the hadronic vacuum polarisation contribution to the anomalous magnetic moment of the muon, where bespoke multigrid algorithms are being developed and run on six different supercomputers in the USA and the EU.

## Significance

Cutting edge lattice gauge theory performance with transformative multigrid algorithms using GPU hardware that accelerates muon g-2 theory calculations by a factor of around 15x.

## References

arXiv:2401.16620, 2203.17119, 2203.06777, 2103.05034, 1512.03487
Recent algorithms plenary talk at Lattice 2023 and talk at Algorithms ‘23.

## Experiment context, if any

Theory prediction of HVP is critical to muon g-2 at FNAL.

**Author:** BOYLE, Peter

**Presenter:** BOYLE, Peter

**Session Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

**Track Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

Contribution ID: **186** Type: **Oral**

# Low Latency, High Bandwidth Streaming of Experimental Data with EJFAT

*Monday, March 11, 2024 3:50 PM (20 minutes)*

Thomas Jefferson National Accelerator Facility (JLab) has partnered with Energy Sciences Network (ESnet) to define and implement an edge to compute cluster data processing computational load balancing architecture. The ESnet-JLab FPGA Accelerated Transport (EJFAT) architecture focuses on FPGA acceleration to address compression, fragmentation, UDP packet destination redirection (Network Address Translation (NAT)) and decompression and reassembly.

EJFAT seamlessly integrates edge and cluster computing to support direct processing of streamed experimental data. This will directly benefit the JLab science program as well as data centers of the future that require high throughput and low latency for both time-critical data acquisition systems and data center workflows.

The principle benefits of the EJFAT architecture include (a) reduction in latency of experimental data processing by allowing it to be processed in real time or near real time (b) redirect streamed data dynamically without needed to reconfigure or restart the data source (c) decoupling of dynamic cluster resource management from the data source

The EJFAT project will be presented along with how it is synergistic with other DOE activities such as an Integrated Research Infrastructure (IRI), and recent results using data sources at JLab, an EJFAT LB at ESnet, and computational cluster resources at Lawrence Berkeley National Laboratory (LBNL).

## Significance

## References

## Experiment context, if any

**Author:** GOODRICH, michael

**Co-authors:** TIMMER, Carl; LAWRENCE, David; Mr HOWARD, Derek (ESnet); HEYES, Graham (Jefferson Lab); SHELDON, Stacey; Dr GYURJYAN, Vardan; KUMAR, Yatish

**Presenter:** GOODRICH, michael

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **187** Type: **Poster**

# Monitoring the OSDF - Open Science Data Federation

*Thursday, March 14, 2024 4:10 PM (30 minutes)*

Extensive data processing is becoming commonplace in many fields of science, especially in computational physics. Distributing data to processing sites and providing methods to share the data with others efficiently has become essential. The Open Science Data Federation (OSDF) builds upon the successful StashCache project to create a global data distribution network. The OSDF expands the StashCache project to add new data origins and caches (14 origins and 32 caches), new access methods, and more monitoring and accounting mechanisms. Additionally, the OSDF has become an integral part of the U.S. national cyberinfrastructure landscape due to the sharing requirements of recent NSF solicitations, which the OSDF is uniquely positioned to enable. To monitor all the OSDF services were created, and improved scripts, data collectors, and data visualizations. This system makes it possible to check the OSDF's health during all operations.

## Significance

## References

## Experiment context, if any

**Author:** ANDRIJAUSKAS, Fabio (Univ. of California San Diego (US))

**Co-author:** WURTHWEIN, Frank (UCSD)

**Presenter:** ANDRIJAUSKAS, Fabio (Univ. of California San Diego (US))

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **188** Type: **Oral**

# The SciDAC QuantOM Framework: A Composable Workflow

*Thursday, March 14, 2024 5:30 PM (20 minutes)*

As part of the Scientific Discovery through Advanced Computing (SciDAC) program, the Quantum Chromodynamics Nuclear Tomography (QuantOM) project aims to analyze data from Deep Inelastic Scattering (DIS) experiments conducted at Jefferson Lab and the upcoming Electron Ion Collider. The DIS data analysis is performed on an event-level by leveraging nuclear theory models and accounting for experimental conditions. In order to efficiently run multiple analyses under varying conditions, a composable workflow was designed where each section (theory, experiment, objective minimization, etc.) has its own dedicated module.The optimization, i.e. the fit of theory to experimental data is carried out by deep learning techniques, such as Generative Adversarial Networks (GANs) or Reinforcement Learning (RL).

This presentation highlights the details and novelties of this workflow, discusses present and future challenges, and highlights possible extensions to other projects with similar requirements.

## Significance

This presentation discusses a framework that not only combines theoretical and experimental nuclear physics into one single workflow, but also leverages state of the art deep learning techniques.

## References

## Experiment context, if any

Deep Inelastic Experiments conducted at Jefferson Lab and the furure EIC.

**Author:** LERSCH, Daniel (Jefferson Lab)

**Co-authors:** RAJPUT, Kishansingh (Jefferson Lab); SCHRAM, Malachi; GOLDENBERG, Steven (Jefferson Lab)

**Presenter:** LERSCH, Daniel (Jefferson Lab)

**Session Classification:** Track 2: Data Analysis - Algorithms and Tools

**Track Classification:** Track 2: Data Analysis - Algorithms and Tools

Contribution ID: **189** Type: **Poster**

# Scaling the SciDAC QuantOM Workflow

*Thursday, March 14, 2024 4:10 PM (30 minutes)*

As part of the Scientific Discovery through Advanced Computing (SciDAC) program, the Quantum Chromodynamics Nuclear Tomography (QuantOM) project aims to analyze data from Deep Inelastic Scattering (DIS) experiments conducted at Jefferson Lab and the upcoming Electron Ion Collider. The DIS data analysis is performed on an event-level by combining the input from theoretical and experimental nuclear physics into a single, composable workflow. The optimization itself (I.e. fitting the experimental data with theoretical predictions) is carried out by a machine / deep learning algorithm. The size of the acquired DIS data as well as the complexity of the workflow itself require that the analysis is performed across multiple GPUs on high performance computing systems, such as Polaris at Argonne National Laboratory.

This presentation discusses the novelties and challenges that came along with parallelizing this workflow. Recent results are compared to common distributed training techniques.

## Significance

This presentation shows novel techniques that were developed to train a GAN based workflow across multiple GPUs

## References

## Experiment context, if any

Deep Inelastic Experiments conducted at Jefferson Lab and the furure EIC.

**Author:** LERSCH, Daniel (Jefferson Lab)

**Co-authors:** SCHRAM, Malachi; DAI, Zhenyu (Jefferson Lab)

**Presenter:** LERSCH, Daniel (Jefferson Lab)

**Session Classification:** Poster session with coffee break

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **190** Type: **Oral**

# Optimizing the ATLAS Geant4 detector simulation

*Wednesday, March 13, 2024 5:30 PM (20 minutes)*

The ATLAS experiment at the LHC heavily depends on simulated event samples produced by a full Geant4 detector simulation. This Monte Carlo (MC) simulation based on Geant4 was a major consumer of computing resources during the 2018 data-taking year and is anticipated to remain one of the dominant resource users in the HL-LHC era. ATLAS has continuously been working to improve the computational performance of this simulation for the Run 3 Monte Carlo campaign. This report highlights the recent implementation of Woodcock tracking in the Electromagnetic Endcap Calorimeter and provides an overview of other implemented and upcoming optimizations. These improvements include enhancements to the core Geant4 software, strategic choices in simulation configuration, simplifications in geometry and magnetic field descriptions, and technical refinements in the interface between ATLAS simulation code and Geant4. Overall, these improvements have resulted in a more than 100% increase in throughput compared to the baseline simulation configuration utilized during Run 2.

## Significance

## References

## Experiment context, if any

ATLAS

**Authors:** VISHWAKARMA, Akanksha (The University of Edinburgh (GB)); SUKHAREV, Andrei (Budker Institute of Nuclear Physics (RU)); WYNNE, Benjamin Michael (The University of Edinburgh (GB)); MORGAN, Benjamin (University of Warwick (GB)); MARCON, Caterina (Università degli Studi e INFN Milano (IT)); KIM, Dongwon (Stockholm University (SE)); TCHERNIAEV, Evgueni (University of Pittsburgh (US)); AMADIO, Guilherme (CERN); APOSTOLAKIS, John (CERN); CHAPMAN, John Derek (University of Cambridge (GB)); BANDIERAMONTE, Marilena (University of Pittsburgh (US)); MUSKINJA, Miha (Jozef Stefan Institute (SI)); NOVAK, Mihaly (CERN); SCHMIDT, Mustafa Andre (Bergische Universitaet Wuppertal (DE)); LARI, Tommaso (University and INFN, Milano); KOURLITIS, Vangelis (Technische Universitat Munchen (DE)); HOPKINS, Walter (Argonne National Laboratory (US))

**Presenter:** SCHMIDT, Mustafa Andre (Bergische Universitaet Wuppertal (DE))

**Session Classification:** Track 1: Computing Technology for Physics Research

**Track Classification:** Track 1: Computing Technology for Physics Research

Contribution ID: **191** Type: **Oral**

# The MadNIS Reloaded

*Tuesday, March 12, 2024 11:30 AM (20 minutes)*

Theory predictions for the LHC require precise numerical phase-space integration and generation of unweighted events. We combine machine-learned multi-channel weights with a normalizing flow for importance sampling to improve classical methods for numerical integration. By integrating buffered training for potentially expensive integrands, VEGAS initialization, symmetry-aware channels, and stratified training, we elevate the performance in both efficiency and accuracy. We empirically validate these enhancements through rigorous tests on diverse LHC processes, including VBS and W+jets.

## Significance

## References

## Experiment context, if any

**Authors:** Prof. MALTONI, Fabio (Universite Catholique de Louvain (UCL) (BE) and Università di Bologna); HUETSCH, Nathan (Heidelberg University, ITP Heidelberg); MATTELAER, Olivier (UCLouvain); WINTERHALDER, Ramon (UCLouvain); HEIMEL, Theo (Heidelberg University); PLEHN, Tilman

**Presenter:** HEIMEL, Theo (Heidelberg University)

**Session Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

**Track Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

Contribution ID: **192** Type: **Oral**

# Modern Machine Learning Tools for Unfolding

*Tuesday, March 12, 2024 11:50 AM (20 minutes)*

Unfolding is a transformative method that is key to analyze LHC data. More recently, modern machine learning tools enable its implementation in an unbinned and high-dimensional manner. The basic techniques to perform unfolding include event reweighting, direct mapping between distributions and conditional phase space sampling, each of them providing a way to unfold LHC data accounting for all correlations in many dimensions. We describe a set of known and new unfolding methods and tools and discuss their respective advantages. Their combination allows for a systematic comparison and performance control for a given unfolding problem.

## Significance

## References

## Experiment context, if any

**Authors:** BUTTER, Anja (Centre National de la Recherche Scientifique (FR)); NACHMAN, Ben (Lawrence Berkeley National Lab. (US)); MARIÑO VILLADAMIGO, Javier; HUETSCH, Nathan (Heidelberg University, ITP Heidelberg); DIEFENBACHER, Sascha (Lawrence Berkeley National Lab. (US)); HEIMEL, Theo (Heidelberg University); PLEHN, Tilman; MIKUNI, Vinicius Massami (Lawrence Berkeley National Lab. (US))

**Presenter:** MARIÑO VILLADAMIGO, Javier

**Session Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

**Track Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

Contribution ID: **193** Type: **Oral**

# A fresh look at the nested soft-collinear subtraction scheme: NNLO QCD corrections to N-gluon final states in quark-anti-quark annihilation

*Wednesday, March 13, 2024 2:50 PM (20 minutes)*

In this talk, I describe how the nested soft-collinear subtraction scheme can be used to compute NNLO QCD corrections to the production of an arbitrary number of gluonic jets in hadron collisions. In particular, I show how to identify NLO-like recurring structures of infrared subtraction terms that in principle can be applied to any partonic process. As an example, I demonstrate the cancellation of all singularities in the fully-differential cross section for the quark-anti-quark annihilation into an arbitrary number of final state gluons at NNLO in QCD.

## Significance

## References

## Experiment context, if any

**Author:** TAGLIABUE, DAVIDE MARIA

**Presenter:** TAGLIABUE, DAVIDE MARIA

**Session Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

**Track Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

Contribution ID: **194** Type: **Oral**

# Two-loop five-points QCD amplitudes in full colour

*Thursday, March 14, 2024 3:50 PM (20 minutes)*

In this talk I will present recent developments on the calculation of five-point scattering amplitudes in massless QCD beyond the leading-colour approximation.
I will discuss the methodology that we pursued to compute these highly non-trivial amplitudes. In this respect, I will argue that it is possible to tackle and tame the seemingly intractable algebraic complexity at each step of the calculation.
I will then illustrate the salient features of the final results and discuss their relevance in view of current and future phenomenological studies at hadron colliders.

## Significance

## References

## Experiment context, if any

**Author:** DEVOTO, Federica

**Presenter:** DEVOTO, Federica

**Session Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

**Track Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

Contribution ID: **195** Type: **Oral**

# Rational-function interpolation from p-adic evaluations in scattering amplitude calculations

*Wednesday, March 13, 2024 3:50 PM (20 minutes)*

Computer algebra plays an important role in particle physics calculations. In particular, the calculation and manipulation of large multi-variable polynomials and rational functions are key bottlenecks when calculating multi-loop scattering amplitudes. Recent years have seen the widespread adoption of interpolation techniques to target these bottlenecks. This talk will present new techniques using p-adic numbers to interpolate such rational functions in a compact form. The techniques are demonstrated on large rational functions at the edge of current capabilities, taken from 2-loop 5-point amplitude calculations. The number of required numerical (p-adic) samples is found to be around 25 times smaller than in conventional (finite-field-based) techniques, and the obtained result is 130 times more compact.

## Experiment context, if any

## References

## Significance

**Presenter:** CHAWDHRY, Herschel (Florida State University)

**Session Classification:** Track 3: Computations in Theoretical Physics: Techniques and Methods

Contribution ID: **196** Type: **not specified**

# Best poster 1st place: Interface to Unity for High Energy Physics detector visualization

*Friday, March 15, 2024 10:45 AM (5 minutes)*

**Presenter:** SONG, Tianzi (Sun Yat-Sen University (CN))

**Session Classification:** Plenary

Contribution ID: **197** Type: **not specified**

# Best poster 2nd place: columnflow: Fully automated analysis through flow of columns over arbitrary, distributed resources

*Friday, March 15, 2024 10:50 AM (5 minutes)*

## Experiment context, if any

## References

## Significance

**Presenters:** WIEDERSPAN, Bogdan (Hamburg University (DE)); RIEGER, Marcel (Hamburg University (DE))

**Session Classification:** Plenary

Contribution ID: **198** Type: **not specified**

# Best poster 3rd place: Introduction of dynamic job matching optimization for Grid middleware using Site Sonar infrastructure monitoring

*Friday, March 15, 2024 10:55 AM (5 minutes)*

## Experiment context, if any

## References

## Significance

**Presenter:** WIJETHUNGA, Kalana (University of Moratuwa (LK))

**Session Classification:** Plenary