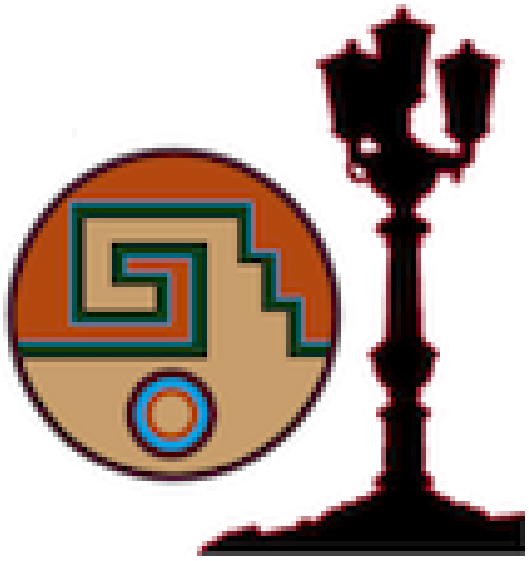
ACAT 2022

Sunday, 23 October 2022 - Friday, 28 October 2022

Villa Romanazzi Carducci, Bari, Italy



Book of Abstracts

Contents

Poster session with coffee break / 2

Machine Learning for Real-Time Processing of ATLAS Liquid Argon Calorimeter Signals with FPGAs

Author: Steffen Stärz[1]

[1] *McGill University, (CA)*

Corresponding Author: steffen.staerz@cern.ch

The Phase-II upgrade of the LHC will increase its instantaneous luminosity by a factor of 7 leading to the High Luminosity LHC (HL-LHC). At the HL-LHC, the number of proton-proton collisions in one bunch crossing (called pileup) increases significantly, putting more stringent requirements on the LHC detectors electronics and real-time data processing capabilities.

The ATLAS Liquid Argon (LAr) calorimeter measures the energy of particles produced in LHC collisions. This calorimeter has also trigger capabilities to identify interesting events. In order to enhance the ATLAS detector physics discovery potential, in the blurred environment created by the pileup, an excellent resolution of the deposited energy and an accurate detection of the deposited time is crucial.

The computation of the deposited energy is performed in real-time using dedicated data acquisition electronic boards based on FPGAs. FPGAs are chosen for their capacity to treat large amount of data with very low latency. The computation of the deposited energy is currently done using optimal filtering algorithms that assume a nominal pulse shape of the electronic signal. These filter algorithms are adapted to the ideal situation with very limited pileup and no overlap of the electronic pulses in the detector. However, with the increased luminosity and pileup, the performance of the optimal filter algorithms decreases significantly and no further extension nor tuning of these algorithms could recover the lost performance.

The back-end electronic boards for the Phase-II upgrade of the LAr calorimeter will use the next high-end generation of INTEL FPGAs with increased processing power and memory. This is a unique opportunity to develop the necessary tools, enabling the use of more complex algorithms on these boards. We developed several neural networks (NNs) with significant performance improvements with respect to the optimal filtering algorithms. The main challenge is to efficiently implement these NNs into the dedicated data acquisition electronics. Special effort was dedicated to minimising the needed computational power while optimising the NNs architectures.

Five NN algorithms based on CNN, RNN, and LSTM architectures will be presented. The improvement of the energy resolution and the accuracy on the deposited time compared to the legacy filter algorithms, especially for overlapping pulses, will be discussed. The implementation of these networks in firmware will be shown. Two implementation categories in VHDL and Quartus HLS code are considered. The implementation results on Stratix 10 INTEL FPGAs, including the resource usage, the latency, and operation frequency will be reported. Approximations in the firmware implementations, including the use of fixed-point precision arithmetic and lookup tables for activation functions, will be discussed. Implementations including time multiplexing to reduce resource usage will be presented. We will show that two of these NNs implementations are viable solutions that fit the stringent data processing requirements on the latency (O(100ns)) and bandwidth (O(1Tb/s) per FPGA) needed for the ATLAS detector operation.

Significance:

References:

Experiment context, if any:

Poster session with coffee break / 4

CMS Tracker Alignment: Legacy results from LHC Run 2 and first results from Run 3

Author: Antonio Vagnerini[1]

[1] *Università di Torino*

Corresponding Author: antonio.vagnerini@cern.ch

The inner tracking system of the CMS experiment, consisting of the silicon pixel and strip detectors, is designed to provide a precise measurement of the momentum of charged particles and to perform the primary and secondary vertex reconstruction. The movements of the individual substructures of the tracker detectors are driven by the change in the operating conditions during data taking. Frequent updates in the detector geometry are therefore needed to describe accurately the position, orientation, and curvature of the tracker modules.

The procedure in which new parameters of the tracker geometry are determined is referred to as the alignment of the tracker. The latter is performed regularly during data taking using reconstructed tracks from both collisions and cosmic rays data, and it is further refined after the end of data-taking. The tracker alignment performance corresponding to the ultimate accuracy of the alignment calibration for the legacy reprocessing of the CMS Run 2 data will be presented. The data-driven methods used to derive the alignment parameters and the set of validations that monitor the performance of the physics observables will be reviewed. The first results obtained with the data taken during the year 2021 and the most recent set of results from LHC Run 3 will be presented.

Significance:

Extensive review of alignment strategies adopted in Run 2 and new developments for Run 3 in the alignment algorithm in both the online & offline reconstruction software, like high granularity automated alignment & new trigger development

References:

Paper submitted to NIMA: https://arxiv.org/pdf/2111.08757.pdf

Experiment context, if any:

CMS

Track 2: Data Analysis - Algorithms and Tools / **5**

Long Short-Term Memory Networks and Bayesian Inference for Time-evolving Systems: an Industrial Case

Author: Davide Pagano[1]

[1] *Universita di Brescia (IT)*

Corresponding Author: davide.pagano@unibs.it

Since the last decade, the so-called *Fourth Industrial Revolution* is ongoing. It is a profound transformation in industry, where new technologies such as smart automation, large-scale machine-to-machine communication, and the internet of things are largely changing traditional manufacturing and industrial practices. The analysis of the huge amount of data, collected in all modern industrial plants, not only has greatly benefited from modern tools of artificial intelligence, but has also spurred the development of new ones. In this context, we present a new approach, based on the combined use of a Long Short-Term Memory (LSTM) neural network and Bayesian inference, for the predictive maintenance of an

industrial plant. *SPE* and *Hotelling* metrics, assessing the degree of compatibility between the time-evolving industrial data and the output of the LSTM, trained on a reference period of good working condition, are used to update the Bayesian probability of a failure of the plant. This method has successfully been applied to a real industrial case and the results are presented and discussed. Finally, it is important to highlight that, although developed to tackle a precise industrial need, the presented approach is general and can be applied to a plethora of other scenarios.

Significance:

References:

Experiment context, if any:

Track 2: Data Analysis - Algorithms and Tools / 6

Neural Estimation of Energy Mover's Distance for Clustering

Author: Ouail Kitouni[1]

Co-authors: J Michael Williams [1]; Jesse Thaler [2]

[1] *Massachusetts Inst. of Technology (US)*

[2] *MIT*

Corresponding Author: ouail.kitouni@cern.ch

We propose a novel neural architecture that enforces an upper bound on the Lipschitz constant of the neural network (by constraining the norm of its gradient with respect to the inputs). This architecture was useful in developing new algorithms for the LHCb trigger which have robustness guarantees as well as powerful inductive biases leveraging the neural network's ability to be monotonic in any subset of features. A new and interesting direction for this architecture is that it can also be used in the estimation of the Wasserstein metric (or the Earth Mover's Distance) in optimal transport using the Kantorovich-Rubinstein duality. In this talk, I will describe how such architectures can be leveraged for developing new clustering algorithms using the Energy Mover's Distance. Clustering using optimal transport generalizes all previous well-known clustering algorithms in HEP (anti-kt, Cambridge-Aachen, etc.) to arbitrary geometries and offers new flexibility in dealing with effects such as pile-up and unconventional topologies. I will also talk in detail about how this flexibility can be used to develop new algorithms which are more suitable for the Electron-Ion Collider setting than conventional ones.

Significance:

This work proposes the use of a novel neural architecture for clustering and jet observable computations in the Energy Mover's Distance framework. It reproduces conventional clustering algorithms and generalizes them to new ones which are more suitable for the Electron-Ion Collider (as an example.)

References:

NeurIPS physical sciences submission for the original architecture and its application to the LHCb trigger: https://ml4physicalsciences.github.io/2021/files/NeurIPS_ML4PS_2021_86.pdf

N.B.: This NeurIPS submission does not refer to any clustering or EMD-related applications.

Experiment context, if any:

Track 3: Computations in Theoretical Physics: Techniques and Methods / 8

Optimally combining BSM searches using graph theory

Authors: Andy Buckley[1]; Benjamin Fuks[None]; Humberto Reyes-González[2]; Jack Araz[3]; James David Yellen[1]; Sophie Williamson[4]; Wolfgang Waltenberger[5]

[1] *University of Glasgow (GB)*

[2] *University of Genoa*

[3] *IPPP - Durham University*

[4] *LPTHE, Sorbonne Université*

[5] *Austrian Academy of Sciences (AT)*

Corresponding Author: james.david.yellen@cern.ch

A decade of data-taking from the LHC has seen great progress in ruling out archetypal new-physics models up to high direct-production energy scales, but few persistent deviations from the SM have been seen. So as we head into the new data-taking era, it is of paramount importance to look beyond such archetypes and consider general BSM models that exhibit multiple phenomenological signatures. But typically each such signature will appear at lower strength than the archetypical simplified models: to significantly constrain them requires a move away from single, "silver-bullet" analyses, to a holistic approach in which many analyses are combined into composite likelihoods. Such combinations require understanding analysis overlaps, and identifying optimal analysis combinations for each point in model space. In this contribution, we present the TACO method, which uses computational statistics in combination with LHC data-reinterpretation tools to estimate analysis correlations, and hence find their optimal combinations. Across several BSM-model scenarios, we show that the TACO approach can significantly increase both exclusion and observation power.

Significance:

This contribution is a new method, to correspond to a paper in late stages of preparation, on novel combination of a bootstrapping method for estimating analysis event-sharing correlations and graph-theory approaches for efficiently & scalably identifying analysis combinations with maximum BSM sensitivity from the combinatoric space of all signal regions.

References:

Prototype correlation analysis in Les Houches 2019

Experiment context, if any:

Poster session with coffee break / 9

Applications of supercomputer Tianhe-II in BESIII

Authors: Biying Hu[1]; Jian Tang[2]; Jingkun Chen[3]; Qiumei Ma[4]; Wei Zheng[5]; Yao Zhang[None]; Ye Yuan[6]; xiaomei zhang[7]

[1] *Sun Yat-sen University*

[2] *Sun Yat-Sen University*

[3] *National Supercomputer Center in Guangzhou*

[4] *IHEP China*

[5] *IHEP*

[6] *Institute of High Energy Physics, Beijing*

[7] *IHEP,Beijing*

Corresponding Author: huby5@mail2.sysu.edu.cn

High energy physics experiments are pushing forward the precision measurements and searching for new physics beyond standard model. It is urgent to simulate and generate mass data to meet

requirements from physics. It is one of the most popular areas to make good use of existing power of supercomputers for high energy physics computing. Taking the BESIII experiment as an illustration, we deploy the offline software BOSS into the top-tier supercomputer "Tianhe-II" with the help of Singularity. With very limited internet connection bandwidth and without root privilege, we synchronize and maintain the simulation software up to date through CVMFS successfully, and an acceleration rate in a comparison of HPC and HTC is realized for the same large-scale task. There are two creative ideas to be shared in the community: on one hand, common users constantly meet problems in the real-time internet connection and the conflict of loading locker. We solve these two problems by deployment a squid server and using fuse in memory in each computing node. On the other hand, we provide a MPI python interface for high throughput parallel computation in TianheII. Meanwhile, the program to deal with data output is also specially aligned so that there is no queue issue in the I/O task. The acceleration rate in simulation reaches 80% so far, as we have done the simulation tests up to 15 K processes in parallel.

Significance:

References:

Experiment context, if any:

Track 2: Data Analysis - Algorithms and Tools / 10

A Machine Learning Method for calorimeter signal processing in sPHENIX

Author: Maxim Potekhin[1]

[1] *Brookhaven National Laboratory (US)*

Corresponding Author: maxim.potekhin@cern.ch

The sPHENIX experiment at RHIC requires substantial computing power for its complex reconstruction algorithms. One class of these algorithms is tasked with processing signals collected from the sPHENIX calorimeter subsystems, in order to extract signal features such as the amplitude, timing of the peak and the pedestal. These values, calculated for each channel, form the basis of event reconstruction in the calorimeter. The baseline technique used for signal feature extraction is fitting the signal waveforms in individual calorimeter channels with a parametrized function which optimally represents the signal shape. Due to the large channel count in the sPHENIX calorimeters, such fitting procedure may consume a non-trivial fraction of the total reconstruction time in a given event. To solve this problem, an alternative technique is being explored, based on a Machine Learning algorithm utilizing a Neural Network, in which the training data sample is produced using the traditional fitting technique. Initial results demonstrate an order of magnitude improvement in speed of signal processing while preserving acceptable level of accuracy. A prototype of a Keras/TensorFlow-based inference application has been created, to be deployed on the worker nodes running sPHENIX event reconstruction software. Comparison with the standard fitting technique has been performed. We present our experience with the design and implementation of the ML-based algorithm for the sPHENIX calorimeter signal processing.

Significance:

The material to be presented describes the first application of ML technology to processing signals produced in a RHIC experiment calorimeter, with a substantial performance gain. The software is optionally packaged as a microservice, which increases modularity and creates flexibility of integration with other applications.

References:

Experiment context, if any:

This research is done in the context of the sPHENIX experiment at RHIC. The abstract has been reviewed and approved by sPHENIX publication board.

Poster session with coffee break / **12**

A comparison of HEPSPEC benchmark performance on ATLAS Grid-Sites versus ideal conditions

Authors: David Cameron[1]; Michael Boehler[2]; David South[3]

[1] *University of Oslo (NO)*

[2] *Albert Ludwigs Universitaet Freiburg (DE)*

[3] *Deutsches Elektronen-Synchrotron (DE)*

Corresponding Author: michael.boehler@cern.ch

The goal of this study is to understand the observed differences in ATLAS software performance, when comparing results measured under ideal laboratory conditions with those from ATLAS computing resources on the Worldwide LHC Computing Grid (WLCG). The laboratory results are based on the full simulation of a single ttbar event and use dedicated, local hardware. In order to have a common and reproducible base to which to compare, thousands of identical ttbar full simulation benchmark jobs were submitted to hundreds of Grid sites using the HammerCloud infrastructure. The impact of the heterogeneous hardware of the Grid sites and the performance difference of different hardware generations is analysed in detail, and a direct, in depth comparison of jobs performed on identical CPU types is also done. The choice of the physics sample used in the benchmark is validated by comparing the performance on each Grid site measured with HammerCloud, weighted by its contribution to the total ATLAS full simulation production output.

Significance:

HEPSPEC06 is still THE metric for the WLCG experiments. In this study we have analysed for the first time centrally via the HammerCloud testing and bench-marking infrastructure how the HEPSPEC06 value varies for a dedicated simulation job over different hardware generations on the entire computing grid. This study can be used as blue print for central evaluation for succeeding benchmark metricizes.

References:

Experiment context, if any:

ATLAS

Track 1: Computing Technology for Physics Research / **13**

The Software Quality Assurance programme of the ASTRI Mini-Array project

Author: Vito Conforti[None]

Co-authors: Andrea Bulgarelli [1]; Nicola La Palombara [1]; Fabrizio Lucarelli [1]; Giorgia Sironi [1]; Lucio Angelo Antonelli [1]; Ciro Bigongiari [1]; Grivel Christine [2]; Stefano Gallozzi [1]; Fulvio Gianotti [1]; Valentina Giordano [1]; Andrea Giuliani [1]; Saverio Lombardi [1]; Rachele Millul [1]; Giovanni Pareschi [1]; Salvatore Scuderi [1]

[1] *INAF*

[2] *TNG IAC*

Corresponding Author: vito.conforti@inaf.it

The ASTRI Mini-Array is a gamma-ray experiment led by Istituto Nazionale di Astrofisica with the partnership of the Instituto de Astrofisica de Canarias, Fundacion Galileo Galilei, Universidade de Sao Paulo (Brazil) and North-West University (South Africa). The ASTRI Mini-Array will consist of nine innovative Imaging Atmospheric Cherenkov Telescopes that are being installed at the Teide Astronomical Observatory (~2400 m a.s.l.) in Tenerife (Canary Islands, Spain). The ASTRI Mini-Array software will cover the entire life cycle of the experiment, including scheduling, operations and data dissemination. The on-site control software will allow the operator to communicate remotely to the array (including automated reaction to critical environmental conditions). Due to the high-speed (10 Gbit/s) networking connection available between Canary Islands and Italy, all data will be delivered every night to the ASTRI dedicated Data Center in Rome for their processing and dissemination. The ASTRI team made experience with ASTRI-Horn, the first Italian dual-mirror Cherenkov telescope, prototype of the ASTRI Mini-Array telescopes. Exploiting lessons learned from ASTRI-Horn, we decided to adopt an iterative incremental model for the software in order to provide more software releases according to the project schedule. Due to this software peculiarity, we have implemented a Quality Assurance (QA) programme specific for the software, which defines the strategy and the organization for the management of the quality control. In this contribution we present the layout and the contents of the ASTRI Mini-Array QA software programme, describing the organization adopted for its management and reporting some examples of how it has been applied so far.

Significance:

References:

Experiment context, if any:

Poster session with coffee break / **14**

Improved Selective Background Monte Carlo Simulation at Belle II with Graph Attention Networks and Weighted Events

Author: Boyang Yu[None]

Co-authors: Nikolai Hartmann [1]; Luca Schinnerl [2]; Thomas Kuhr [1]

[1] *Ludwig Maximilians Universitat (DE)*

[2] *LMU Munich*

Corresponding Author: boyang.yu@physik.uni-muenchen.de

When measuring rare processes at Belle II, a huge luminosity is required, which means a large number of simulations are necessary to determine signal efficiencies and background contributions. However, this process demands high computation costs while most of the simulated data, in particular in case of background, are discarded by the event selection. Thus filters using graph neural networks are introduced at an early stage to save the resources for the detector simulation and reconstruction of events discarded at analysis level. In our work, we improved the performance of the filters using graph attention and invested statistical methods including sampling and reweighting to deal with biases introduced by the filtering.

Significance:

Improved the accuracy of distinguishing between background and expected events while reduced bias. Provided a tool to speedup the generation + skimming process.

References:

DPG Talk 2022:
https://www.dpg-verhandlungen.de/year/2022/conference/heidelberg/part/t/session/53/contribution/1
DPG Talk 2021:

https://www.dpg-verhandlungen.de/year/2021/conference/dortmund/part/t/session/38/contribution/10

Experiment context, if any:

Belle II

Track 1: Computing Technology for Physics Research / **15**

The Virtual Research Environment: towards a complexive analysis platform

Authors: Elena Gazzarrini[1]; Alba Vendrell Moya[1]; Riccardo Di Maria[1]; Rizart Dona[1]; Xavier Espinal[1]; Enrique GARCIA GARCIA[None]

[1] *CERN*

Corresponding Author: elena.gazzarrini@cern.ch

One of the objectives of the EOSC (European Open Science Cloud) Future Project is to integrate diverse analysis workflows from Cosmology, Astrophysics and High Energy Physics in a common framework. The project's development relies on the implementation of the Virtual Research Environment (VRE), a prototype platform supporting the goals of Dark Matter and Extreme Universe Science Projects in the respect of FAIR data policies, making use of a common AAI system, and leveraging experiments data via a reliable and scalable distributed storage infrastructure for multi-science: the Data Lake. The entry point of such a platform is a jupyterhub instance sitting on top of a complex K8s infrastructure, which provides an interactive GUI interface for researchers to access and share data, as well as to run notebooks. The data access and browsability is enabled through API calls to the high level data management and storage orchestration software (Rucio).
The cluster's functionality, currently allowing data injection replication, storage and deletion, is being expanded to include a software repository plug-in enabling researchers to directly select computational environments from Docker images and to host a re-analysis platform (REANA) supporting various distributed computing backends (K8s, HTCondor, Slurm), which allows scientists to spawn and interact with complete re-analysis workflows.
The goal of the VRE project, bringing together data and software access, workflow reproducibility and enhanced user interface, is to facilitate scientific collaboration, ultimately accelerating research in various fields.

Significance:

The VRE will first and foremost provide an easy-to use prototype analysis platform based on some of the most commonly used DevOps technologies (K8s, Helm, Flux, GitLab, DB on-demand), with the nuance of hosting workflows spanning from the field of particle physics to astrophysics.
The novelty of the infrastructure will be its common AAI framework to authenticate with both federated storage services and computing infrastructure, gaining access to the data management software, the software repository and the computational environment necessary for analysis reproduction at the same time.
While similar work has been ongoing in single isolated institutes – at CERN, for example –, the VRE aims at being completely open source, easily reproducible on different clusters, and easily accessible by anyone having an account; the target audience are not only HEP sciences, but also smaller experiments who would hugely benefit from the provisioning of shared computing resources.

References:

https://indico.in2p3.fr/event/26454/
https://indico.cern.ch/event/1151054/

Experiment context, if any:

Various experiments are currently using the VRE platform and providing feedback to develop it further – making it easier to use, enhancing the documentation, improving its deployment – and others are continuously onboarding the project. The postdocs testing the infrastructure are involved in experiments

such as ATLAS, Km3Net, Fermi-LAT, EGO and LOFAR.

Track 2: Data Analysis - Algorithms and Tools / **16**

Evolutionary algorithms for hyperparameter optimization in machine learning for application in high energy physics

Authors: Christian Veelken[1]; Laurits Tani[1]; Diana Rand[1]; Mario Kadastik[1]

[1] *National Institute of Chemical Physics and Biophysics (EE)*

Corresponding Author: laurits.tani@cern.ch

In contemporary high energy physics (HEP) experiments the analysis of vast amounts of data represents a major challenge. In order to overcome this challenge various machine learning (ML) methods are employed. However, in addition to the choice of the ML algorithm a multitude of algorithm-specific parameters, referred to as hyperparameters, need to be specified in practical applications of ML methods. The optimization of these hyperparameters, which is often performed manually, has a significant impact on the performance of the ML algorithm. In this talk we explore several evolutionary algorithms that allow to determine optimal hyperparameters for a given ML task in a fully automated way. Additionally, we study the capability of the two most promising hyperparameter optimisation algorithms, particle swarm optimization and bayesian optimization, for utilising the highly parallel computing architecture that is typical for the field of HEP.

Significance:

ML methods are in common use in HEP data analyses, but very few studies have been performed of the task of finding optimal hyperparameter values. Algorithms that allow to determine optimal hyperparameter values in a fully automated way are presented in this talk. Furthermore, we present results on how well different hyperparameter optimization algorithms parallelise on modern computing architectures, such as computing clusters and the Worldwide LHC Computing Grid (WLCG).

References:

Tani, L., Rand, D., Veelken, C. et al. Evolutionary algorithms for hyperparameter optimization in machine learning for application in high energy physics. Eur. Phys. J. C 81, 170 (2021). https://doi.org/10.1140/epjc/s10052-021-08950-y

Tani, L. & Veelken, C. Comparison of Bayesian and particle swarm algorithms for hyperparameter optimisation in machine learning applications in high energy physics. arXiv preprint arXiv:2201.06809 (2022) https://arxiv.org/pdf/2201.06809.pdf

Experiment context, if any:

None

Poster session with coffee break / **17**

AI Data Quality Monitoring with Hydra

Authors: David Lawrence[None]; Thomas Britton[None]; Kishansingh Rajput[1]; Torri Jeske[None]

[1] *Jefferson Lab*

Corresponding Author: tbritton@jlab.org

Hydra is an AI system employing off-the-shelf computer vision technologies aimed at autonomously monitoring data quality. Data quality monitoring is an essential step in modern experimentation and Nuclear Physics is no exception. Certain failures can be identified through alarms (e.g. electrical heartbeats) while others are more subtle and often require expert knowledge to identify and diagnose. In the GlueX experiment at Jefferson Laboratory data quality monitoring is a multistep, human in the loop process that begins with shift crews looking at a litany of plots (e.g. occupancy plots) which indicate the performance of detector subsystems. With the sheer complexity of the systems and number of plots needing to be monitored subtle issues can be, and are, missed. During its time in production (over 2 years) Hydra has lightened the load of shift takers of GlueX by autonomously monitoring detector systems. This talk will describe the construction, training, and operation of the Hydra system in GlueX as well as the ongoing work to develop and deploy the system with other experiments at Jefferson Laboratory and beyond.

Significance:

This work represents an early deployment of an AI system in production using off-the-shelf technologies. It has been generalized and is being adopted by other experiments and shows a good foundation for the management and deployment of AI monitoring systems and dovetails with ongoing work in AI controls.

References:

@article{ refId0,
author = {{Britton, Thomas} and {Lawrence, David} and {Rajput, Kishansingh}},
title = {AI Enabled Data Quality Monitoring with Hydra},
DOI= “10.1051/epjconf/202125104010”,
url= “https://doi.org/10.1051/epjconf/202125104010”,
journal = {EPJ Web Conf.},
year = 2021,
volume = 251,
pages = “04010”,
}

Experiment context, if any:

GlueX, SBS, CLAS12

Poster session with coffee break / **18**

AtlFast3: Fast Simulation in ATLAS for Run 3 and beyond

Authors: Joshua Falco Beirer[1]; Rui Zhang[2]

[1] *CERN, Georg-August-Universitaet Goettingen (DE)*

[2] *University of Wisconsin Madison (US)*

Corresponding Author: rui.zhang@cern.ch

AtlFast3 is the next generation of high precision fast simulation in ATLAS that is being deployed by the collaboration and was successfully used for the simulation of 7 billion events in Run 2 data taking conditions. AtlFast3 combines a parametrization-based approach known as FastCaloSimV2 and a machine-learning based tool that exploits Generative Adversarial Networks (FastCaloGAN) for the simulation of hadrons.

For the purpose of Run 3, the parametrization of AtlFast3 was fully reworked and many active developments are ongoing to further enhance the quality of fast simulation in ATLAS. This talk will give a brief overview of AtlFast3 with focus on FastCaloSimV2 and outline several improvements with respect to the previous simulator tool AFII. Furthermore, recent advancements in the parametrised simulation, such as the development of a dedicated tune of electromagnetic shower shapes to data are presented.

Significance:

The talk will give an overview of recent developments of fast simulation in ATLAS that play a crucial role in achieving the collaborations computing goals. Novel developments such as the tuning of EM shower shapes to data are presented.

References:

https://link.springer.com/article/10.1007/s41781-021-00079-7

Experiment context, if any:

ATLAS

Poster session with coffee break / **19**

Design and implementation of computational storage system based on EOS for HEP data processing

Authors: Xiaoyu Liu[1]; Xiaoyu Liu[2]; Yaodong Cheng[None]; Yaosong Cheng[None]; minxing zhang[3]

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

[2] *Central China Normal University CCNU (CN)*

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

Corresponding Authors: liuxiaoyu@ihep.ac.cn, xiaoyu.liu@cern.ch

Computing in high energy physics is one kind of typical data-intensive applications, especially some data analysis , which require access to a large amount of data. The traditional computing system adopts the “computing-storage” separation mode, which leads to large data volume move during the computing process, and and also increase transmission delay and network load. Therefore, it can effectively alleviate this situation by pushing down some data-intensive tasks from computing node to storage node. The philosophy is that bringing computing as close to the source of data as possible in order to reduce latency and bandwidth use. Generally, storage nodes have computing resources like CPUs, necessary for deploying distributed file system. However, the computing power in storage node is often ignored. This paper designed and implemented a computational storage system based on CERN Open Storage (EOS). The system presents transparently the computational storage functions through standard POSIX file system interface, such as open, read and write. A plugin implemented in EOS storage node (FST) will execute the specified algorithm or program when it finds the special arguments in filename, for example “&CSS=decode”. The plugin can read and write file locally in FST, then register new-generated file into EOS name node (MGM). The paper finally give some test results showing that the computational storage mode performs faster and supports more parallel computing tasks than the traditional mode in some applications like raw data decode for LHAASO experiment. Computational storage mode reduces computation time by 37% in single task execution and 72% in the case of 40 tasks in parallel compared with traditional mode.

Significance:

References:

Experiment context, if any:

Track 1: Computing Technology for Physics Research / **20**

Optimizing the ATLAS Geant4 detector simulation software

Authors: Evangelos Kourlitis[1]; Marilena Bandieramonte[2]

Co-authors: John Derek Chapman [3]; Tommaso Lari [4]

[1] *Argonne National Laboratory (US)*

[2] *University of Pittsburgh (US)*

[3] *University of Cambridge (GB)*

[4] *University and INFN, Milano*

Corresponding Author: evangelos.kourlitis@cern.ch

The ATLAS experiment at the LHC relies critically on simulated event samples produced by the full Geant4 detector simulation software (FullSim). FullSim was the major CPU consumer during the last data-taking year in 2018 and it is expected to be still significant in the HL-LHC era [1, 2]. In September 2020 ATLAS formed a Geant4 Optimization Task Force to optimize the computational performance of FullSim for the Run 3 Monte Carlo campaign. This contribution summarizes the already implemented and upcoming improvements. These include improved features from the core Geant4 software, optimal options in the simulation configuration, simplifications in geometry and magnetic field description and technical improvements in the way ATLAS simulation code interfaces with Geant4. Overall, more than 50% higher throughput is achieved, compared to the baseline simulation configuration used during Run 2.

[1]: ATLAS Collaboration, "ATLAS HL-LHC Computing Conceptual Design Report", CERN-LHCC-2020-015.
[2]: ATLAS Collaboration, "ATLAS Software and Computing HL-LHC Roadmap", CERN-LHCC-2022-005.

Significance:

This contribution summarizes novel physics and computing optimizations for the ATLAS experiment full detector simulation software. For the first time, these are implemented into the production of the required Monte Carlo simulations for the LHC Run 3. These improvements allow the production of 50% more simulated events, using the same computational resources with the previous baseline software.

References:

1. Geant4 Technical Forum, Geant4 in ATLAS, https://indico.cern.ch/event/1106118/timetable/?view=standard#19-geant4-in-atlas
2. Geant4 Technical Forum, ATLAS Geant4 Simulation Update, https://indico.cern.ch/event/1139613/timetable/?view=atlas-geant4-simulation-upd

Experiment context, if any:

ATLAS Experiment at CERN

Poster session with coffee break / 21

Flexible Event Data Model for track reconstruction in ACTS and integration with ATLAS xAOD

Author: Paul Gessinger[1]

Co-author: Tomasz Bold [2]

[1] *CERN*

[2] *AGH Univ. of Science and Technology, Krakow (PL)*

Corresponding Author: paul.gessinger@cern.ch

ACTS [1] is an experiment independent toolkit for track reconstruction, which is designed from the ground up for thread-safety and high performance. It is built to accommodate different experiment deployment scenarios, and also serves as community platform for research and development of new approaches and algorithms.
The ATLAS experiment [2] plans to use ACTS as the backbone for track reconstruction starting from Run 4, and work is underway to integrate the toolkit into the experiment software.

The Event Data Model (EDM) is a central piece of the tracking library that is visible to clients. Until this point, ACTS was mostly focused on an internal EDM, targeting data interchange betweens various components in the toolkit. This contribution reports on recent development to build on top of this internal EDM to provide a high-level, user-facing client EDM for tracking outputs. It includes a top level track object description and components containing information on the intermediate states originating from a track finding or fitting algorithm. Both are needed for flexible downstream processing and handling of tracking outputs. These outputs generally need to be easily persistable for event data storage.

The new ACTS EDM is designed such that the storage backend can be tailored to an experiment, with minimal overhead. This way, the choice of persistence technology can then be determined by the specific needs of an experiment.

This contribution reports on the integration of the new ACTS track EDM into the existing ATLAS xAOD [3] EDM infrastructure, leveraging the low-overhead abstracted storage mechanism. In addition, the interface of the track EDM with edm4hep [4], a central component of the larger turnkey software stack key4hep [5] for collider studies, will be discussed.

Significance:

The track EDM is central to deployment of ACTS into experiments, as it's the main interface point for downstream clients of tracking. Having this EDM in place is an important milestone for ACTS deployment in ATLAS, but also improving usability in other experiment contexts. The low-overhead abstraction mechanism ensures that the integration into experiment specific contexts does not introduce performance issues.

References:

[1] A Common Tracking Software Project, 2106.13593
[2] The ATLAS Experiment at the CERN Large Hadron Collider, 10.1088/1748-0221/3/08/S08003
[3] Implementation of the ATLAS Run 2 event data model, 10.1088/1742-6596/664/7/072045
[4] https://github.com/key4hep/EDM4hep
[5] https://github.com/key4hep

Experiment context, if any:

ATLAS

Track 2: Data Analysis - Algorithms and Tools / 22

Learning full-likelihoods of LHC results with Normalizing Flows.

Authors: Humberto Reyes-González[1]; Riccardo Torre[2]

[1] *University of Genoa*

[2] *INFN Genoa*

Corresponding Author: harg.zeppelin@gmail.com

The publication of full likelihood functions (LFs) of LHC results is vital for a long-lasting and profitable legacy of the LHC. Although major steps have been put forward in this direction, the systematic publication of LFs remains a big challenge in High Energy Physics (HEP) as such distributions are usually quite complex and high-dimensional. Thus, we propose to describe LFs with Normalizing Flows (NFs); a powerful class of expressive generative networks that provide density estimation by construction. In this talk, we show that NFs are able to accurately model the complex high-dimensional LFs found in HEP, in some cases even with relatively small training samples. This approach opens the possibility of compact and efficient characterisations of the LFs derived from LHC searches, SM measurements, phenomenological studies, etc.

Significance:

The systematic publication of full likelihood functions (LFs) of LHC results is a hot topic in HEP. This would allow for future new statisctical interpretations, for more accurate re-interpretations of the results in the context of different theoretical models, etc. However, there is not a stong concensus on how this LFs should be published, specially since they are often high-dimensional complex distributions. Thus, we present a novel approach for modeling and sharing this LFs using Normalizing Flows. We believe that NFs are very suitable for this task and could be systematically used.

References:

Experiment context, if any:

LHC experiments, specially ATLAS and CMS. Usage could be extended to other experiments.

Track 3: Computations in Theoretical Physics: Techniques and Methods / **23**

Bridge between Classical & Quantum Machine Learning

Authors: Jack Y. Araz[1]; Michael Spannowsky[2]

[1] *IPPP - Durham University*

[2] *IPPP Durham*

Corresponding Author: jack.araz@durham.ac.uk

Tensor Networks (TN) are approximations of high-dimensional tensors designed to represent locally entangled quantum many-body systems efficiently. In this talk, we will discuss how to use TN to connect quantum mechanical concepts to machine learning techniques, thereby facilitating the improved interpretability of neural networks. As an application, we will use top jet classification against QCD jets and compare performance against state-of-the-art machine learning applications. Finally, we will discuss how to convert these models into Quantum Circuits to be compiled on a quantum device and show that classical TNs require exponentially large bond dimensions and higher Hilbert-space mapping to perform comparably to their quantum counterparts.

Significance:

This study shows how to use Quantum inspired algorithms in Machine learning to increase the interpretability of the application and compile such networks in a quantum device to improve the representability of the network.

References:

This talk is based on 2202.10471 [quant-ph] and 2106.08334 [hep-ph].
IRN Terascale: https://indico.in2p3.fr/event/26315/contributions/107811/
at LPSC: https://lpsc-indico.in2p3.fr/event/2873/

Experiment context, if any:

Track 3: Computations in Theoretical Physics: Techniques and Methods / 24

Run Dependent Monte Carlo at Belle II

Authors: Alberto Martini[1]; Stefano Lacaprara[2]

[1] *DESY*

[2] *INFN sezione di Padova*

Corresponding Author: alberto.martini@desy.de

The Belle II is an experiment taking data from 2019 at the asymmetric e+e- SuperKEKB collider, a second generation B-factory, at Tsukuba, Japan. Its goal is to perform high precision measurements of flavor physics observables One of the many challenges of the experiment is to have a Monte Carlo simulation with very accurate modeling of the detector, including any variation occurring during data taking. To this goal, a dedicated "run dependent" Monte Carlo has been developed, using the detector conditions during data taking, as well as using beam induced background collected with random triggers. In this talk, the procedure for setup and processing of run-dependent Monte Carlo at Belle II will be shown.

Significance:

References:

Experiment context, if any:

Belle II

Track 1: Computing Technology for Physics Research / 25

Data Calibration and Processing at Belle II

Author: Stefano Lacaprara[1]

[1] *INFN sezione di Padova*

Corresponding Author: stefano.lacaprara@pd.infn.it

The Belle II experiment has been collecting data since 2019 at the second generation e+/e- B-factory SuperKEKB in Tsukuba, Japan. The goal of the experiment is to explore new physics via high precision measurement in flavor physics. This is achieved by collecting a large amount of data that needs to be calibrated promptly for fast reconstruction and recalibrated thoroughly for the final reprocessing. To fully automate the calibration process a Python plugin package, b2cal, had been developed based on the open-source Apache Airflow package using Directed Acyclic Graphs (DAGs) to describe the ordering of processes and Flask to provide administration and job submission web pages. Prompt processing and reprocessing are performed at different calibration centers (BNL and DESY, respectively). After calibration, the raw data are reconstructed on the GRID to an analysis-oriented format (mDST), also stored on the GRID, and delivered to the collaborations. This talk will describe the whole procedure, from raw data calibration to mDST production.

Significance:

References:

Experiment context, if any:

Belle II

Track 2: Data Analysis - Algorithms and Tools / 26

Simultaneous track finding and track fitting by the Deep Neural Network at BESIII

Authors: Yao Zhang[None]; Ye Yuan[1]; Haiyong Jiang[2]

Co-authors: Wenniu Zhang [2]; Xiao-Rui Lyu [3]; Yangheng Zheng [4]; Jun Xiao [2]

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

[2] *University of Chinese Academy of Sciences*

[3] *UCAS*

[4] *University of Chinese Academy of Sciences (CN)*

Corresponding Author: zhangyao@ihep.ac.cn

Track fitting and track hit classification are highly relevant, hence these two approaches could benefit each other. For example, if we know the underlying parameters of a track, then track hits associated with the track can be easily identified. On the other hand, if we know the hits of a track, then we can get underlying parameters by fitting them. Most existing works take the second scheme by classifying track hits and then estimating track parameters.
Inspired by the above observations and the success of multi-task training, we propose a unified framework to address track fitting and track hit classification simultaneously in an end-to-end fashion. The method takes hits from multiple tracks as inputs, where each hit holds 4-dimensional features, including 2D position, hitting time, and deposit charge. We feed these inputs to a backbone network to extract per-hit features. Then the network is divided into two branches. One branch is a reconstruction branch, which estimates the parameters of each track and its existence. The other branch is a track segmentation branch, which takes learned features of PointNet++ and tracks features to determine a hit-wise track assignment. In essence, we can assign each track hit to its potential track to classify track hits. This method allows us to predict the track parameters of a track candidate while conducting per-track hit classification. This study leverages the simulated multi-track samples of the BESIII drift chamber. Preliminary results indicate our framework is able to categorize hits of different tracks and the candidate track parameters simultaneously.

Significance:

References:

Experiment context, if any:

Track 1: Computing Technology for Physics Research / 27

The journey towards HEPscore, the HEP-specific CPU benchmark for WLCG

Author: Domenico Giordano[1]

[1] *CERN*

Corresponding Author: domenico.giordano@cern.ch

HEPscore is a CPU benchmark, based on HEP applications, that the HEPiX Working Group is proposing as a replacement of the currently used HEPSpec06 benchmark, adopted in WLCG for procurement, computing resource pledges and performance studies.

In 2019, we presented at ACAT the motivations for building a benchmark for the HEP community based on HEP applications. The process from the conception to the implementation and validation of this objective has been inspiring and challenging. In the spirit of the HEP community, it has involved many contributions from software developers, data analysts, experts of the experiments, representatives of several WLCG computing centres, as well as the WLCG HEPscore Deployment Task Force.
In this contribution, we review this long journey and in particular the technological solutions selected, such as containerization of the HEP applications and cvmfs snapshotting. We update the community on the readiness status of HEPscore, the HEP application mix selected to build HEPscore and the deployment plans for 2023. We describe the current campaign of measurements performed on multiple WLCG sites, intended to study the performance of eleven HEP applications on more than 50 different computer systems.
Finally, we also cover how to extend the HEPscore adoption to the benchmarking of heterogeneous resources, and how it can include workloads for physics analysis and Machine Learning algorithms.

Significance:

HEPscore is the candidate benchmark for compute performance. It bridges the experiments' HEP applications with the benchmark paradigms. As replacement of HS06, it will be of major interest for the whole HEP community

References:

https://link.springer.com/article/10.1007/s41781-021-00074-y

Experiment context, if any:

Track 2: Data Analysis - Algorithms and Tools / 28

Flow-Unet for High Dimensional Image Semantic Segmentation

Authors: HaoLai Tian[1]; Yu Hu[1]; Xiaomeng Qiu[2]; Lin Wang[3]

Co-authors: Wei Song [2]; ChunYan Liu [3]

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

[2] *Zhengzhou University*

[3] [redacted]

Corresponding Author: 202022172013231@gs.zzu.edu.cn

Nowadays, medical images play a mainstay role in medical diagnosis, and computer tomography, nuclear magnetic resonance, ultrasound and other imaging technologies have become a powerful means of in vitro imaging. Extracting lesion information from these images can enable doctors to observe and diagnose the lesion more effectively, so as to improve the accuracy of quasi diagnosis. Therefore, the segmentation of medical images has important social value.The achievement of image semantic segmentation shows the potential of the Convolutional Neural Network (CNN) for medical image analysis. However, the application of the existing CNN model to the video neglect the correlation between frames of the video. A video semantic segmentation framework based on U-Net is proposed in this article that the feature map of the pre-frame is propagated to the next frame via an optical flow field. The accuracy of segmentation is boosted with slight performance degradation. The framework includes three parts: 1) a segmentation sub module using UNet to segment the current frame; 2) an optical flow feature extraction module to perform feature extraction on the motion information of the current frame and the previous frame; 3) a correction module, which assigns weights to the segmentation results and optical flow features to achieve the correction effect. The effectiveness of our proposed method is presented on two public datasets (Drosophila melanogaster electron micrographs, Chaos), and private Digital Subtraction Angiography (DSA) video datasets.

Significance:

References:

Experiment context, if any:

Track 3: Computations in Theoretical Physics: Techniques and Methods / 29

First results of Local Unitarity at N3LO

Authors: Ben Ruijl[1]; Valentin Hirschi[2]; Zeno Capatti[1]

[1] *ETH Zürich*

[2] *CERN*

Corresponding Author: zcapatti@phys.ethz.ch

Local Unitarity provides an order-by-order representation of perturbative cross-sections that realises at the local level the cancellation of final-state collinear and soft singularities predicted by the KLN theorem. The representation is obtained by manipulating the real and virtual interference diagrams contributing to transition probabilities using general local identities. As a consequence, the Local Unitarity representation can be directly integrated using Monte Carlo methods and without the need of infrared counter-terms. I will present first results from this new approach with examples up to N3LO accuracy. I will conclude by giving an outlook on future generalisations of the method applicable to hadronic collisions.

Significance:

The new results that I will present from Local Unitarity offer a clear path to go beyond the state-of-the-art collider simulations.

References:

arXiv:1906.06138, arXiv:2010.01068, arXiv:2203.11038

Experiment context, if any:

High-Energy Colliders

Poster session with coffee break / 30

Transparent expansion of a WLCG compute site using HPC resources

Authors: Ralf Florian Von Cube[1]; Alexander Schmidt[2]; Gunter Quast[1]; Manuel Giffels[1]; Andreas Nowack[3]; Thomas Kress[3]; Alexander Jung[3]; Matthias Schnepf[None]

[1] *KIT - Karlsruhe Institute of Technology (DE)*

[2] *RWTH Aachen (DE)*

[3] *Rheinisch Westfaelische Tech. Hoch. (DE)*

Corresponding Author: ralf.florian.von.cube@cern.ch

Restarting the LHC again after more than 3 years of shutdown, unprecedented amounts of data are expected to be recorded. Even with the WLCG providing a tremendous amount of compute resources to process this data, local resources will have to be used for additional compute power. This, however, makes the landscape in which computing takes place more heterogeneous.

In this contribution, we present a solution for dynamically integrating non-HEP resources into existing infrastructures using the COBalD/TARDIS resource manager. By providing all resources through conventional CEs as single point-of-entry, the use of these external resources becomes completely transparent for experiments and users.

In addition, experiences with an existing setup, operated in production since more than a year, extending the German Tier 2 WLCG site operated at RWTH Aachen University with a local HPC cluster will be discussed.

Significance:

References:

Experiment context, if any:

Poster session with coffee break / **31**

Machine Learning Techniques for selecting Forward Electrons $(2.5 < \eta < 3.2)$ with the ATLAS High Level Trigger

Authors: Meinrad Moritz Schefer[1]; Yu Nakahama Higuchi[2]

[1] *Universitaet Bern (CH)*

[2] *High Energy Accelerator Research Organization (JP)*

Corresponding Author: meinrad.moritz.schefer@cern.ch

The ATLAS detector at CERN measures proton proton collisions at the Large Hadron Collider (LHC) which allows us to test the limits of the Standard Model (SM) of particles physics. Forward moving electrons produced at these collisions are promising candidates for finding physics beyond the SM. However, the ATLAS detector is not construed to measure forward leptons with pseudorapidity η of more than 2.5 with high precision. The ATLAS performance for forward leptons can be improved by enhancing the trigger system. This system selects events of interest in order to not overwhelm the data storage with the information of around 1.7 billion collisions per second. First studies using the Neural Ringer algorithm for selecting forward electrons with $2.5 < \eta < 3.2$ show promising results. The Neural Ringer using machine learning to analyse detector information to distinguish electromagnetic from hadronic signatures, is being presented. Additionally, its performance on simulated ATLAS Monte Carlo samples in improving the high level trigger for forward electrons will be shown.

Significance:

This presentation covers the performance evaluation of the NeuralRinger algorithm in selecting electrons in more forward regions (|eta|>2.5) than online electrons are currently triggered within the ATLAS experiment at LHC. A special focus will be laid on the machine learning aspects used for it and not on some general performance of the ATLAS electron trigger.

References:

-

Experiment context, if any:

The ATLAS experiment.

Poster session with coffee break / **32**

A FPGA Implementation of the Hough Transform tracking algorithm for the Phase-II upgrade of ATLAS

Authors: Fabrizio Alfonsi[1]; Yu Nakahama Higuchi[2]

[1] *Universita e INFN, Bologna (IT)*

[2] *High Energy Accelerator Research Organization (JP)*

Corresponding Author: fabrizio.alfonsi@cern.ch

The High Energy Physics world will face challenging trigger requests in the next decade. In particular the luminosity increase to 5-7.5 x 1034 cm-2 s-1 at LHC will push the major experiments as ATLAS to exploit the online tracking for their inner detector to reach 10 kHz of events from 1 MHz of Calorimeter and Muon Spectrometer trigger. The project described here is a proposal for a tuned Hough Transform algorithm implementation on FPGA high-end technology, versatile to adapt different tracking situations. The platform developed allows to study different dataset from a software "emulating" the firmware and consequently to the hardware performance and to generate input dataset from ATLAS simulation. Xilinx FPGA have been destined to this implementation, exploiting up to now the VC709 commercial board and its PCI Express Generation 3 technology. The system provides the features to possibly process a 200 pile up event of ATLAS Run4 in the order of 10 μs averagely, with the possibility to run two events at a time. Best efficiency reached are simulated to be > 95 % for single muon tracking. The project plans to be proposed for the Event Filter TDAQ ATLAS Upgrade of Phase-II.

Significance:

These results are the updates of a FPGA tracking algorithm implementation forwarded
by the INFN Bologna group and stated in the ATLAS TDAQ Phase-II upgrade reports related to the Hardware Tracking for Trigger project.

References:

https://www.mdpi.com/2079-9292/10/20/2546
https://www.mdpi.com/2079-9292/11/4/517

Experiment context, if any:

The ATLAS experiment.

Poster session with coffee break / 34

Fast track seed selection for track following in the Inner Detector Trigger track reconstruction

Authors: Andrius Vaitkus[1]; Yu Nakahama Higuchi[2]

[1] *University of London (GB)*

[2] *High Energy Accelerator Research Organization (JP)*

Corresponding Author: andrius.vaitkus@cern.ch

During ATLAS Run 2, in the online track reconstruction algorithm of the Inner Detector (ID), a large proportion of the CPU time was dedicated to the fast track finding. With the proposed HL-LHC upgrade, where the event pile-up is predicted to reach <Î¼>=200, track finding will see a further large increase in CPU usage. Moreover, only a small subset of Pixel-only seeds is accepted after the fast track finding procedure, essentially discarding the CPU time used on rejected seeds. Therefore, a computationally cheap track candidate seed pre-selection procedure based on approximate track following was designed, which is described in this report. The algorithm uses a parabolic track approximation in the plane perpendicular to the beamline, a combinatorial Kalman filter simplified by a reference-related coordinate system to find the best track candidates. For such candidates,

a set of numerical features are created to classify seeds using machine learning techniques, such as Support Vector Machines (SVM) or kernel-based methods. The algorithm was tuned for high identification and rejection of bad seeds, while ensuring no significant loss of track finding efficiency. Current studies focus on implementing the algorithm into the Athena framework for online seed pre-selection, which could be used during Run 3 or potentially be adapted for the ITk geometry for Run 4 of the HL-LHC.

Significance:

The presentation covers a new approach that could greatly reduce the overall CPU time consumption for the full-scan and large RoI tracking in the ATLAS Inner Detector. A separate algorithm is designed to be used for pre-filtering of track seeds before the combinatorial track following, so it could be used simultaneously with any other optimisation techniques used within the fast track finding algorithm. Reducing the CPU timing of the track finding is especially important for higher pile-up levels, and this algorithm is flexible and can be adapted to Run 4 ITk geometry.

References:

-

Experiment context, if any:

The ATLAS experiment.

Poster session with coffee break / **35**

Parametrized simulation of the micro-RWELL response with PARSIFAL software

Authors: Lia Lavezzi[1]; Riccardo Farinelli[2]; on behalf of CGEM-IT, FCC_RD, EURIZON working groups of Ferrara, Laboratori Nazionali di Frascati and Torino[None]

[1] *Universita e INFN Torino (IT)*

[2] *INFN, Ferrara (IT)*

Corresponding Author: lia.lavezzi@to.infn.it

PARSIFAL (PARametrized SImulation) is a software tool originally implemented to reproduce the complete response of a triple-GEM detector to the passage of a charged particle, taking into account the involved physical processes by their simple parametrization and thus in a very fast way.
Robust and reliable software, such as GARFIELD++, is widely used to simulate the transport of electrons and ions in the gas and all their interactions step by step, but it is CPU-time consuming. The implementation of PARSIFAL code was driven by the need to reduce the processing time, while maintaining the precision of a full simulation.
The software must be initialized with some parameters that can be extracted from the GARFIELD++ simulation, which must be run once-and-for-all. Then it can be run independently to provide a reliable simulation, from the ionization, to diffusion, multiplication, signal induction and electronics, only by sampling from a set of functions which describe the physical effects and depend on the input parameters.
The code has been thoroughly tested on triple-GEM detectors and the simulation was finely tuned to experimental data collected at testbeam.
Recently, PARSIFAL has been extended to another detector in the MPGD family, the micro-RWELL, thanks to the modular structure of the code. The main difference in the treatment of the physical processes is the introduction of the resistive plane and its effect on the formation of the signal. For this purpose, the charge spread on the resistive layer has been described following the work of M. S. Dixit and A. Rankin (NIM A518 (2004) 721-727, NIM A566 (2006) 281-285) and the electronics readout (APV-25) was added to the description.
A fine tuning of the simulation is ongoing to reproduce the experimental data collected during testbeams. A similar strategy already validated for the triple-GEM case is used: the variables of interest for the comparison of the experimental data with simulated results are the cluster charge, cluster size and the position resolution obtained by charge centroid and micro-TPC reconstruction algorithms.

In this case, special attention must be paid to the tuning of the resistivity of the resistive layer.
An illustration of the general code, setting the focus on this latest implementation and the first comparison with experimental data from testbeam are the subject of this contribution.

Significance:

PARSIFAL software has been originally developed for the simulation of the response of a triple-GEM, within the project for the development of a Cylindrical GEM Inner Tracker for the BESIII experiment. The code has been tested on triple-GEM detectors and the simulation was finely tuned to experimental data collected at testbeam.
Recently, PARSIFAL has been extended to micro-RWELL, due to the modular structure of the code, as they also are micro pattern gas detectors.
PARSIFAL important feature is that it allows for reliable simulations of a triple-GEM and of a micro-RWELL reducing significantly the CPU-time with respect to full physics simulators, as GARFIELD++.

References:

- PARSIFAL for triple-GEM was presented by R. Farinelli, "A fast and parametric digitization for triple-GEM detectors" ACAT 2019
- PARSIFAL for micro-RWELL was presented by R. Farinelli, "A parametric simulation of the micro-RWELL detector", RD51 June 2022 Collaboration Meeting, CERN

Experiment context, if any:

FCC-ee/CepC IDEA, EURIZON (EU H2020 project)

Track 1: Computing Technology for Physics Research / **36**

CPU-level resources allocation for optimal execution of multi-process physics code

Author: Marta Bertran Ferrer[1]

[1] *CERN*

Corresponding Author: marta.bertran.ferrer@cern.ch

During the LHC LS2, the ALICE experiment has undergone a major upgrade of the data acquisition model, evolving from a trigger-based model to a continuous readout. The upgrade allows for an increase in the number of recorded events by a factor of 100 and in the volume of generated data by a factor of 10. The entire experiment software stack has been completely redesigned and rewritten to adapt to the new requirements and to make optimal use of storage and CPU resources. The architecture of the new processing software relies on running parallel processes on multiple processor cores and using large shared memory areas for exchanging data between them.

Without mechanisms that guarantee job resource isolation, the deployment of multi-process jobs can result in a usage that exceeds those originally requested and allocated. Internally, jobs may launch as many processes as defined in their workflow, significantly higher than the number of allocated CPU cores. This freedom of execution can be limited by mechanisms like cgroups, already employed by some Grid sites, however these are a minority. If jobs are allowed to run unconstrained, they may interfere with each other in terms of the simultaneous utilization of the resources. Constrainment mechanisms in this context improve the fairness of resource utilization, both between ALICE jobs and towards other users in general.

The efficient use of the worker nodes' cache memory is closely related to the CPU cores executing the job. An important aspect to consider is the host architecture and the cache topology, i.e. cache levels, size and hierarchical connection to individual cores. Memory usage patterns of running tasks,

the memory and cache topologies and the chosen CPU cores to constrain the job to influence the overall efficiency of the execution, in terms of useful work done by unit of time.

This paper presents an analysis of the impact of different CPU pinning strategies on the efficiency of the execution of simulation tasks. The evaluation of the different configurations is performed by extracting a set of metrics tightly related to job turnaround and efficient resource utilization. The results are presented both for the execution of a single job on an idle machine and for whole node saturation, analyzing the interference between jobs. Different host architectures are studied for a global and robust assessment.

Significance:

Efficient use of resources by multi-threaded physics applications on heterogeneous hardware.

References:

Experiment context, if any:

ALICE simulation, processing and analysis code

Track 1: Computing Technology for Physics Research / **37**

Challenges and opportunities in migrating the CNAF datacenter to the Bologna Tecnopolo

Authors: Luca dell'Agnello[1]; Tommaso Boccali[2]

Co-authors: Andrea Chierici [3]; Daniele Cesini [4]; Luigi Scarponi [3]; Pierpaolo Ricci [3]; Stefano Zani ; Vladimir Sapunenko [5]

[1] *INFN*

[2] *INFN Sezione di Pisa*

[3] *INFN-CNAF*

[4] *Universita e INFN, Bologna (IT)*

[5] *INFN-CNAF (IT)*

Corresponding Authors: luca.dellagnello@cern.ch, tommaso.boccali@cern.ch, daniele.cesini@cnaf.infn.it

The INFN Tier1 data center is currently located in the premises of the Physics Department of the University of Bologna, where CNAF is also located. Soon it will be moved to the "Tecnopolo", the new facility for research, innovation, and technological development in the same city area; it will follow the installation of Leonardo, the pre-exascale supercomputing machine managed by CINECA, co-financed as part of the EuroHPC Joint Undertaking.
The construction of the new CNAF data center will consist of two phases, corresponding to the computing requirements of LHC: Phase 1, starting from 2023, will involve an IT power of 3 MW, and Phase 2, starting from 2025, involving an IT power up to 10 MW.
The primary goal of the new data center is to cope with the computing requirements of the data taking of the HL-LHC experiments, in the time spanning from 2026 to 2040, providing, at the same time, computing services for several other INFN experiments, projects, and activities of interest, being they currently in operation, under construction, in advanced design, or even not yet defined.
The co-location with Leonardo will also open new scenarios, with a close integration between the two systems able to share dynamically resources.
In this presentation we will describe the new center design, with a particular focus on the status of the migration, its schedule, and the technical challenges we have to face moving the data center without service interruption. On top of this, we will analyze the opportunities that the new infrastructure will open in the context of the PNRR (National Plan for Resilience and Recovery) funding and strategic plans, within and beyond the High Energy Physics domain.

Significance:

It is a status report, but includes the analysis of the technical challenges we had to face to migrate the data center without service interruption and the integration of the data center itself with a pre-exascale machine.

References:

Experiment context, if any:

Track 1: Computing Technology for Physics Research / 38

Monitoring CMS experiment data and infrastructure for next generation of LHC run

Authors: Benedikt Maier[1]; Brij Kishor Jashal[2]; Ceyhun Uzunoglu[1]; Federica Legger[3]; Felipe Gomez[4]; Garyfallia Paspalaki[5]; Oscar Fernando Garzon Miguez[6]; Valentin Y Kuznetsov[7]

[1] *CERN*

[2] *Tata Inst. of Fundamental Research (IN)*

[3] *Universita e INFN Torino (IT)*

[4] *Universidad de los Andes (CO)*

[5] *Purdue University (US)*

[6] *Fermi National Accelerator Lab. (US)*

[7] *Cornell University (US)*

Corresponding Author: ceyhun.uzunoglu@cern.ch

As CMS starts the Run 3 data taking, the experiment's data management software tools along with the monitoring infrastructure have undergone significant upgrades to cope up with the conditions expected in the coming years. The challenges of an efficient, real-time monitoring for the performance of the computing infrastructure or for data distribution are being met using state-of-the-art technologies that are continuously evolving. In this talk, we describe how we set up monitoring pipelines based on a combination of technologies, such as Kubernetes, Spark/Hadoop and other open-source software stacks. We show how the choice of these components is critical for this new generation of services and infrastructure for CMS data management and monitoring. We also discuss how some of the developed monitoring services such as data management monitoring, CPU efficiency monitoring, data-set access and transfers metrics, have been instrumental for taking strategic decisions and increasing the physics harvest through maximal utilization of computing resources available to us.

Significance:

References:

Experiment context, if any:

CMS

Poster session with coffee break / 39

Transparent extension of INFN-T1 with heterogeneous computing architectures

Authors: Daniele Spiga[None]; Stefano Dal Pra[1]

Co-authors: Lorenzo Rinaldi [1]; Tommaso Boccali [2]

[1] *Universita e INFN, Bologna (IT)*

[2] *INFN Sezione di Pisa*

Corresponding Author: stefano.dal.pra@cern.ch

The INFN-CNAF Tier-1 is engaged for years in a continuous effort to integrate its computing centre with more tipologies of computing resources. In particular, the challenge of providing opportunistic access to nonstandard CPU architectures, such as PowerPC or hardware accelerators (GPUs) has been actively exploited. In this work, we describe a solution to transparently integrate access to ppc64 CPUs as also GPUs. This solution has been tested to transparently extend the INFN-T1 Grid computing centre with Power9 based machines and V100 GPUs from the Marconi 100 HPC cluster managed by CINECA. We also discuss further possible improvements and how this will meet requirements and future plans for the new tecnopolo centre, where the CNAF Tier-1 will be hosted soon.

Significance:

End users can transparently access HPC resources and special resources (non x86 CPUs, GPUs) through the usual and well known methods used to submit payloads to the INFN-T1 batch system. No need for the INFN-T1 users to adapt their submission workflow in case of particular targets. Also no need for them to directly handle specific problems at the resources, who are managed by the INFN-T1 staff.

References:

1) Boccali, T., Dal Pra, S., Spiga, D., Ciangottini, D., Zani, S., Bozzi, C., ... & Bonacorsi, D. (2020). Extension of the INFN Tier-1 on a HPC system. In EPJ Web of Conferences (Vol. 245, p. 09009). EDP Sciences.

2) Enabling CMS Experiment to the utilization of multiple hardware architectures – a Power9 Testbed at CINECA (ACAT 2021)

Experiment context, if any:

The context of this research is provided by several WLCG experiments. The shared use case is the needs to access any available resource, minimizing the effort Operational Wise as well as minimizing the development effort required to integrate new heterogeneous providers.

Track 1: Computing Technology for Physics Research / 40

Speeding up CMS simulations, reconstruction and HLT code using advanced compiler options

Co-authors: Danilo Piparo [1]; Malik Shahzad Muzaffar [1]; Niccolo' Forzano [2]; Vladimir Ivantchenko [1]; Vincenzo Innocente [1]

[1] *CERN*

[2] *Universita & INFN, Milano-Bicocca (IT)*

Corresponding Authors: danilo.piparo@cern.ch, shahzad.malik.muzaffar@cern.ch, niccolo.forzano@cern.ch, civanch@cern.ch, vincenzo.innocente@cern.ch

The CMS simulation, reconstruction, and HLT code have been used to deliver an enormous number of events for analysis during Runs 1 and 2 of the LHC at CERN. In fact, these techniques have been regarded as of fundamental importance for the CMS experiment. In the following arguments presented, several ways to improve efficiency of these procedures will be described and it will be displayed how no particular conceptual or technical blocker has been identified in their implementation.

In this framework, particular attention will be devoted to highlight how CMS simulation, Reco and HLT will gain a considerable increase in speed recompiling several CMS sub-libraries using advanced compiler options. In fact, using this logic, the compiler will be leveraged to obtain a up to 10% speedup. As will be shown, the focus of the reasonings reported will be on the LTO (Link Time Optimization) and PGO (Profile Guided Optimization) approaches: using these advanced tools, several results will be seen about improving the event loop time and event throughput and the differences between the profiles of the processes will be shown. Moreover, an important feature of PGO approach will be considered: profiles obtained running events based on one process will be enough to speedup many other ones (and a profile obtained with the Phase 1 detector configuration will manage to give an improvement for Phase 2 processes too).

Significance:

References:

Experiment context, if any:

CMS Collaboration

Poster session with coffee break / **42**

A graph neural network for B decays reconstruction at Belle II

Authors: Giulio Dujany[1]; Ilias Tsaklidis[None]; Jacopo Cerasoli[2]; James Kahn[3]; Lea Reuter[None]; Markus Götz[4]; Oskar Taubert[None]; Pablo Goldenzweig[5]

[1] *IPHC - CNRS*

[2] *CNRS - IPHC*

[3] *Helmholtz AI, Karlsruhe Institute of Technology (KIT)*

[4] *Karlsruhe Institute of Technology (KIT)*

[5] *KIT - Karlsruhe Institute of Technology (DE)*

Corresponding Author: jacopo.cerasoli@cern.ch

Over the past few years, intriguing deviations from the Standard Model predictions have been reported in measurements of angular observables and branching fractions of B meson decays, suggesting the existence of a new interaction that acts differently on the three lepton families. The Belle II experiment has unique features that allow to study B meson decays with invisible particles in the final state, in particular neutrinos. It is possible to deduce the presence of such particles from the energy-momentum imbalance obtained after reconstructing the companion B meson produced in the event. This task is complicated by the thousands of possible final states B mesons can decay into, and is currently performed at Belle II by the Full Event Interpretation (FEI) software, an algorithm based on Boosted Decision Trees and limited to specific, hard-coded decay processes.
In recent years, graph neural networks have proven to be very effective tools to describe relations in physical systems, with applications in a range of fields. Particle decays can be naturally represented in the form of rooted, acyclic tree graphs, with nodes corresponding to particles and edges representing the parent-child relations between them. In this work, we present a graph neural network approach to generically reconstruct B decays at Belle II by exploiting the information from the detected final state particles, without formulating any prior assumption about the nature of the decay. This task is performed by reconstructing the Lowest Common Ancestor matrix, a novel representation, equivalent to the adjacency matrix, that allows reconstruction of the decay from the final state particles alone. Preliminary results show that the graph neural network approach outperform the FEI by a factor of at least 3.

Significance:

Preliminary results show that this work significantly improves the reconstruction of decays with invisible particles in the final state at Belle II.

References:

Experiment context, if any:

Belle II

Poster session with coffee break / **43**

Custom event sample augmentations for ATLAS analysis data

Author: Peter Van Gemmeren[1]

Co-authors: Alaettin Serhan Mete [1]; Jackson Carl Burzynski [2]; James Catmore [3]; Lukas Alexander Heinrich [4]; Marcin Jerzy Nowak [5]; Nils Erik Krumnack [6]

[1] *Argonne National Laboratory (US)*

[2] *Simon Fraser University (CA)*

[3] *University of Oslo (NO)*

[4] *CERN*

[5] *Brookhaven National Laboratory (US)*

[6] *Iowa State University (US)*

Corresponding Author: lukas.heinrich@cern.ch

High Energy Physics (HEP) has been using column-wise data stored in synchronized containers, such as most prominently ROOT's TTree, for decades. These containers have proven to be very powerful as they combine row-wise association capabilities needed by most HEP event processing frameworks (e.g. Athena) with column-wise storage, which typically results in better compression and more efficient support for many analysis use-cases. The downside, however, is that all events (rows) need to contain the same attributes and therefore extending the list of items to be stored, even if needed only for a subsample of events, can be costly in storage and lead to data duplication.
The ATLAS experiment has developed navigational infrastructure to allow storing custom data extensions for subsample of events in separate, but synchronized containers. These extensions can easily be added to ATLAS standard data products (such as DAOD-PHYS or PHYSLITE) avoiding duplication of those core data products, while limiting their size increase. As a proof of principle, a prototype based on the Long Lived Particle search is implemented. Preliminary results concerning the event-size as well as reading/writing performance implications associated with this prototype will be presented.
Augmented data as described above are stored within the same file as the core data. Storing them in dedicated files will be investigated in future, as this could provide more flexibility to store augmentations separate from core data, e.g. certain sites may only want a subset of several augmentations or augmentations can be archived to disk once their analysis is complete.

Significance:

Derived data is a main consumer of storage resources (for ATLAS in Run 2, derived AOD occupied >30% of disk). The capability of custom augmentation will reduce duplication and reduce storage costs.

References:

Experiment context, if any:

ATLAS

Track 1: Computing Technology for Physics Research / **44**

A cloud-based computing infrastructure for the HERD cosmic-ray experiment

Authors: Matteo Duranti[1]; Nicola Mori[2]; Valerio Formato[3]

Co-authors: Daniele Spiga ; Diego Ciangottini [4]

[1] *Universita e INFN, Perugia (IT)*

[2] *INFN Florence*

[3] *INFN - Sezione di Roma Tor Vergata*

[4] *INFN, Perugia (IT)*

Corresponding Author: nicola.mori@cern.ch

The HERD experiment will perform direct cosmic-ray detection at the highest ever reached energies, thanks to an innovative design that maximizes the acceptance, and its placement on the future Chinese Space Station which will allow for an extended observation period."

Significant computing and storage resources are foreseen to be needed in order to cope with the necessities of a large community driving a big experimental device with an energy reach above PeV for hadrons and multi-TeV for electrons and positrons. For example, at PeV energies Monte Carlo simulations require a massive amount of computing power, and very large simulated data sets are needed for detector performance studies like electron-proton rejection.

The HERD computing infrastructure is currently being investigated and prototyped in order to provide a flexible, robust and easy to use cloud-based computing and storage platform. It is based on technical solutions originally developed by the "Dynamic On Demand Analysis Service" (DODAS) framework in the context of projects such as INDIGO-DataCloud, EOSC-hub and XDC. It allows to seamlessly access both commercial and institutional cloud resources, in order to efficiently make use of opportunistic resources to cope with high-demand periods (like full dataset reprocessings and specialized Monte Carlo productions), as well transparently integrate with with on-premise computing resources managed by an HTCondor batch system. The cloud platform also allows for an easy and efficient deployment of services for the collaboration like calendar, document server, code repository etc. making use of available, free open source solutions. Finally, an Indigo-IAM instance provides a Single-Sign-On service for access control for the whole infrastructure.

An overview of the current status and of the future perspectives will be presented.

Significance:

This contribution is about the first-ever, fully-cloud-based platform for data processing of a space-based cosmic-ray experiment. The presented technical solutions are innovative for the field and are a significant advancement in the definition of the computing model of the HERD experiment.

References:

Experiment context, if any:

HERD

Track 1: Computing Technology for Physics Research / **45**

ML-based tool for RPC currents quality monitoring

Authors: Borislav Pavlov[1]; Elton Shumka[1]; Leandar Litov[1]; Peicho Petkov[1]

[1] *University of Sofia - St. Kliment Ohridski (BG)*

Corresponding Author: elton.shumka@cern.ch

The CMS experiment has 1056 Resistive Plate Chambers (RPCs) in its muon system. Monitoring their currents is the first essential step towards maintaining the stability of the CMS RPC detector performance. An automated monitoring tool to carry out this task has been developed. It utilises the ability of Machine Learning (ML) methods in the modelling of the behavior of the current of these chambers. Two types of ML approaches are used: Generalized Linear Models and Autoencoders. In the GLM case, a set of parameters such as environmental conditions, LHC parameters and working point are used to characterize the behavior of the current. In the autoencoder case, the set of currents for all of the high-voltage channels of the RPC system are used as input and the autoencoder network is trained to reproduce these inputs on the output neurons. Both approaches show very good predictive capabilities, with accuracy of the order of 1-2 μA. These predictive capabilities are the basis for the monitoring tool, which is going to be tested during Run 3. All the developed tools are integrated in a framework that can be easily accessed and controlled by a specially developed Web User Interface that allows the end user to work with the monitoring tool in a simple manner.

Significance:

The tool provides an automatized approach to current monitoring, making it possible to monitor the whole system and detect changes in behavior that would otherwise be difficult to detect.

References:

Experiment context, if any:

This tool will be integrated in the CMS RPC automation tools for Run 3

Poster session with coffee break / **46**

Enabling continuous speedup of CMS Event Reconstruction through continuous benchmarking

Author: Claudio Caputo[1]

[1] *Universite Catholique de Louvain (UCL) (BE)*

Corresponding Author: claudio.caputo@cern.ch

The outstanding performances obtained by the CMS experiment during Run1 and Run2 represent a great achievement of seamless hardware and software integration. Among the different software parts, the CMS offline reconstruction software is essential for translating the data acquired by the detectors into concrete objects that can be easily handled by the analyzers. The CMS offline reconstruction software needs to be reliable and fast. The long shutdown 2 (LS2) elapsed between LHC Run2 and Run3 has been instrumental in the optimization of the CMS offline reconstruction software and for the introduction of new algorithms reaching a continuous CPU speedup. In order to reach these goals, a continuous benchmarking pipeline has been implemented; CPU timing and memory profiling, using the igprof tool, are performed on a regular basis to monitor the footprint of the new developments and identify the possible areas of performance improvement. The current status and achievement obtained by a continuous benchmarking of CMS experiment offline reconstruction software are described here.

Significance:

References:

Experiment context, if any:

CMS Experiment

Poster session with coffee break / 47

Secrets Management for CMSWEB

Author: Muhammad Imran[1]

Co-authors: Valentin Y Kuznetsov [2]; Panos Paparrigopoulos [3]; Spyridon Trigazis [3]; Andreas Pfeiffer [3]

[1] *National Centre for Physics (PK)*

[2] *Cornell University (US)*

[3] *CERN*

Corresponding Author: muhammad.imran@cern.ch

Secrets Management is a process where we manage secrets, like certificates, database credentials, tokens, and API keys in a secure and centralized way. In the present CMSWEB (the portfolio of CMS internal IT services) infrastructure, only the operators maintain all services and cluster secrets in a secure place. However, if all relevant persons with secrets are away, then we are left with no choice but to contact them to get secrets in case of emergency needs.

In order to overcome this issue, we performed an R&D study for the management of secrets and explored various strategies such as Hashicorp Vault, Github credential manager, and SOPS/age. In this talk, we'll discuss the process by which CMS investigated these strategies and perform a feasibility analysis of them. We will also underline why CMS chose SOPS as a solution, reviewing how the features of SOPS with age satisfy our needs. We will also discuss how other experiments could adopt our solution.

Significance:

In this talk, we'll discuss the process by which CMS investigated the strategies and perform a feasibility analysis for selecting a best solution for secrets management. We will also discuss how other experiments could adopt our solution.

References:

Experiment context, if any:

CMS Experiment at CERN

Poster session with coffee break / 49

A distributed infrastructure for interactive analysis: the experience at INFN

Authors: Daniele Spiga[1]; Diego Ciangottini[2]; Mirco Tracolli[3]; Piergiulio Lenzi[4]; Tommaso Tedeschi[1]

[1] *Universita e INFN, Perugia (IT)*

[2] *INFN, Perugia (IT)*

[3] *INFN Perugia*

[4] *Universita e INFN, Firenze (IT)*

Corresponding Author: diego.ciangottini@cern.ch

The challenges expected for the HL-LHC era, both in terms of storage and computing resources, provide LHC experiments with a strong motivation for evaluating ways of re-thinking their computing models at many levels. In fact a big chunk of the R&D efforts of the CMS experiment have been focused on optimizing the computing and storage resource utilization for the data analysis, and Run3 could provide a perfect benchmark to make studies on new solutions in a realistic scenario. The work that will be shown is focused on the integration and validation phase of an interactive environment for data analysis with the peculiarity of providing a seamless scaling over grid resources at Italian T2s, and possibly opportunistic providers such as HPC. In this approach the integration of new resources has been proved to be exceptionally easy in terms of requirements, thus computing power can be included dynamically in a very effective way. The presentation will firstly focus on an overview of the architectural pillars and the integration challenges. Then the results of a first set of performance measurements will be presented, thanks to a first real user CMS analysis built on top of Root RDataFrame ecosystem that has been successfully executed over such an infrastructure.

Significance:

References:

Experiment context, if any:

Poster session with coffee break / 50

Enhanced Data Analytics capabilities in the ELK Stack - a review of the premium features and their benefit to a Scientific Compute Facility

Author: Michael Poat[None]

Co-author: Jerome LAURET [1]

[1] *Brookhaven National Laboratory*

Corresponding Author: mpoat@bnl.gov

In real-time computing facilities - system, network, and security monitoring are core components to run efficiently and effectively. As there are many diverse functions that can go awry, such as load, network, processes, and power issues, having a well-functioning monitoring system is imperative. In many facilities you will see the standard set of tools such as Ganglia, Grafana, Nagios, etc. While these are noteworthy, the diversity of tools used clearly points to an adequacy gap (none is self-sufficient) and furthermore, they lack in their alerting and anomaly detection capabilities beyond the binary events.

The ELK stack (Elasticsearch, Logstash, & Kibana) is the combination of three open-source projects to ingest, search, and visualize logs and data. The basic free license of ELK enables these features but overall is limited for use in a real-time facility. Instead, by leveraging the full capabilities of ELK, the gained features are significant. ELK offerings provide many enhancements from single sign-on and means to control Authorization for security, including alerting for unusual events, Machine Learning capabilities, and many other tools that are useful for advanced data analytics.

With the advanced set of Machine Learning techniques, the ELK toolbox adds features such as clustering, time series decomposition, and correlation analysis. For example, these Machine Learning techniques can be applied to alerts, providing you with the details of events for an unusual uptick in resource usage, if there is rare or high process activity, or unusual port activity. A standard monitoring tool would typically not have such capability.

In this report, will discuss the details and features of how a facility could benefit from the open source and premium versions of the ELK stack. We will provide procedures and details for configuring these tools, and how it benefits compute facility monitoring postures within a scientific based environment.

Significance:

Focus on the ELK offerings with use of their Machine Learning techniques to provide scientific compute facilities specialized alerting and monitoring.

References:

Flexible visualization of a 3rd party Intrusion Prevention (Security) tool: A use case with the ELK stack - https://indico.cern.ch/event/855454/contributions/4604980/

Experiment context, if any:

Track 2: Data Analysis - Algorithms and Tools / **52**

Hierarchical Graph Neural Networks for Particle Track Reconstruction

Authors: Daniel Thomas Murnane[1]; Ryan Liu[2]

Co-authors: Paolo Calafiura [1]; Steven Farrell [3]; Xiangyang Ju [1]

[1] *Lawrence Berkeley National Lab. (US)*

[2] *University of California, Berkeley*

[3] *Lawrence Berkeley National Laboratory*

Corresponding Author: liuryan30@berkeley.edu

Graph Neural Networks (GNN) have recently attained competitive particle track reconstruction performance compared to traditional approaches such as combinatorial Kalman filters. In this work, we implement a version of Hierarchical Graph Neural Networks (HGNN) for track reconstruction, which creates the hierarchy dynamically. The HGNN creates "supernodes" by pooling nodes into clusters, and builds a "supergraph" which enables message passing among supernodes. A new differentiable pooling algorithm that can maintain the sparsity and produce variable number of supernodes is proposed to facilitate the hierarchy construction. We perform an apples-to-apples comparison between the Interaction Network (IN) and HGNN on track finding performance using node embedding metric learning, which shows that in general HGNNs are more robust against imperfectly constructed input graphs, and more powerful in recognizing long-distance patterns. Equipped with soft assignment, HGNN also allows assigning a given hit to multiple track candidates. The HGNN model can be used as a node-supernode pair classifier, where supernodes are considered to be track candidates. Under this regime, the pair-classifying HGNN is even more powerful than the node embedding HGNN. We show that the HGNN can not only improve upon the performance of common GNN architectures on embedding and clustering problems but also opens up other approaches for GNNs in high energy physics.

Significance:

We explore the behavior of hierarchical GNNs in the context of high energy physics, proposing HEP-specific ways they can be applied, and demonstrating superior performance to traditional GNNs

References:

Experiment context, if any:

Poster session with coffee break / 53

Recent Developments in the FullSimLight Simulation Tool from ATLAS

Authors: Andrea Dell'Acqua[1]; Joseph Boudreau[2]; Marilena Bandieramonte[3]; Raees Ahmad Khan[3]; Riccardo Maria Bianchi[3]; Vakho Tsulaia[4]

[1] *CERN*

[2] *University of Pittsburgh*

[3] *University of Pittsburgh (US)*

[4] *Lawrence Berkeley National Lab. (US)*

Corresponding Author: raees.ahmad.khan@cern.ch

FullSimLight is a lightweight, Geant4-based command line simulation utility intended for studies of simulation performance. It is part of the GeoModel toolkit (geomodel.web.cern.ch) which has been stable for more than one year. Currently, the FullSimLight component is undergoing renewed development aimed at extending its functionality. It has been endowed with a GUI for fast, transparent, and foolproof configuration and with a plugin mechanism allowing users and developers with diverse goals to extend and customize the simulation. Geometry and event input can be easily specified on the fly, allowing rapid evaluation of different geometry options and their effect on simulation performance. User actions and sensitive detectors can also be loaded through the new plugin mechanism, allowing for customization of Geant4 processing and hit production. Simulation of radiation backgrounds in ATLAS, which is well adapted to lightweight simulation, is expected to run within FullSimLight in the near future. FullSimLight, brought to you by the ATLAS collaboration, is an experiment independent software tool.

Significance:

References:

Experiment context, if any:

Track 3: Computations in Theoretical Physics: Techniques and Methods / 54

Quantum computing of the 6Li nucleus via ordered unitary coupled cluster

Author: Oriel Orphee Moira Kiss[1]

Co-authors: Federico Sanchez [1]; Michele Grossi [2]; Pavel Lougovski [3]; Sofia Vallecorsa [2]; Thomas Papenbrock [4]

[1] *Universite de Geneve (CH)*

[2] *CERN*

[3] *AWS center for quantum computing*

[4] *The University of Tennessee, Knoxville*

Corresponding Author: oriel.kiss@cern.ch

The variational quantum eigensolver (VQE) is an algorithm to compute ground and excited state energy of quantum many-body systems. A key component of the algorithm and an active research area is the construction of a parametrized trial wavefunction – a so called variational ansatz. The wavefunction parametrization should be expressive enough, i.e. represent the true eigenstate of a quantum system for some choice of parameter values. On the other hand, it should be trainable, i.e. the number of parameters should not grow exponentially with the size of the system. Here, we

apply VQE to the problem of finding ground and excited state energies of the odd-odd nucleus 6Li. We study the effects of ordering fermionic excitation operators in the unitary coupled clusters ansatz on the VQE algorithm convergence by using only operators preserving the Jz quantum number. The accuracy is improved by two order of magnitude in the case of descending order. We first compute optimal ansatz parameter values using a classical state-vector simulator with arbitrary measurement accuracy and then use those values to evaluate energy eigenstates of 6Li on a superconducting quantum chip from IBM. We post-process the results by using error mitigation techniques and are able to reproduce the exact energy with an error of 3.8% and 0.1% for the ground state and for the first excited state of 6Li, respectively.

Significance:

We compute the ground and first excited state energy of the 6Li nuclei in the shell model using a quantum processor, which has not been done before.

References:

https://arxiv.org/abs/2205.00864

Experiment context, if any:

Track 3: Computations in Theoretical Physics: Techniques and Methods / **55**

Quantum neural networks force fields generation

Author: Oriel Orphee Moira Kiss[1]

Co-authors: Francesco Tacchino [2]; Ivano Tavernelli [3]; Sofia Vallecorsa [4]

[1] *Universite de Geneve (CH)*

[2] *IBM Research Zürich*

[3] *IBM Research - Zurich*

[4] *CERN*

Corresponding Author: oriel.kiss@cern.ch

Accurate molecular force fields are of paramount importance for the efficient implementation of molecular dynamics techniques at large scales. In the last decade, machine learning methods have demonstrated impressive performances in predicting accurate values for energy and forces when trained on finite size ensembles generated with ab initio techniques. At the same time, quantum computers have recently started to offer new viable computational paradigms to tackle such problems. On the one hand, quantum algorithms may notably be used to extend the reach of electronic structure calculations. On the other hand, quantum machine learning is also emerging as an alternative and promising path to quantum advantage. Here we follow this second route and establish a direct connection between classical and quantum solutions for learning neural network potentials. To this end, we design a quantum neural network architecture and apply it successfully to different molecules of growing complexity. The quantum models exhibit larger effective dimension with respect to classical counterparts and can reach competitive performances, thus pointing towards potential quantum advantages in natural science applications via quantum machine learning.

Significance:

We propose a quantum neural network (QNN) for to computation of molecular energy and forces and use it to drive molecular dynamic simulations. While QNNs have already been proposed in quantum chemistry, our reaches better accuracy and computes the forces as well.

References:

https://arxiv.org/abs/2203.04666

Experiment context, if any:

Track 1: Computing Technology for Physics Research / **56**

Performance of Run 3 Software of the ATLAS Experiment

Author: Alaettin Serhan Mete[1]

Co-authors: Alessandro Di Girolamo [2]; Federico Meloni [3]; John Kenneth Anders [2]; Marilena Bandieramonte [4]; Zach Marshall [5]

[1] *Argonne National Laboratory (US)*

[2] *CERN*

[3] *Deutsches Elektronen-Synchrotron (DE)*

[4] *University of Pittsburgh (US)*

[5] *Lawrence Berkeley National Lab. (US)*

Corresponding Author: alaettin.serhan.mete@cern.ch

In preparation for Run 3, ATLAS has upgraded its data-processing software, Athena, to support multi-threading. Although writing and maintaining multi-threaded software is complicated, it was a necessary undertaking in order to maximize the event throughput by better utilizing the available computing resources. Athena is used in all official ATLAS workflows, including but not limited to simulation, digitization, trigger, and reconstruction. In this work, the preliminary performance of this new software in various workflows, including the processing of both simulated and real data from the latest data-taking run, will be presented.

Significance:

This work will be the first public presentation of the performance of ATLAS' Athena framework after its migration to multi-threading using early Run 3 data.

References:

Experiment context, if any:

The ATLAS Experiment

Poster session with coffee break / **57**

Evaluating Generative Adversarial Networks for particle hit generation in a cylindrical drift chamber using Fréchet Inception Distance

Authors: Irene Andreou[None]; Noam Mouelle[1]

[1] *Imperial College London*

Corresponding Authors: irene.andreou18@imperial.ac.uk, noam.mouelle18@imperial.ac.uk

We evaluate two Generative Adversarial Network (GAN) models developed by the COherent Muon to Electron Transition (COMET) collaboration to generate sequences of particle hits in a Cylindrical Drift Chamber (CDC). The models are first evaluated by measuring the similarity between distributions of particle-level, physical features. We then measure the Effectively Unbiased Fréchet Inception Distance (FID) between distributions of high-dimensional representations obtained with: InceptionV3; then a version of InceptionV3 fine-tuned for event classification; and a 3D Convolutional Neural Network that has been specifically designed for event classification. We also normalize the obtained FID values by the FID for two sets of real samples, setting the scores for different representations on the same scale. This novel relative FID metric is used to compare our GAN models to state-of-the-art natural image generative models.

Significance:

Fréchet Inception Distance (FID) is the standard metric, in academia and industry alike, used to evaluate natural image generative models. However, to our knowledge, it has only been used once to evaluate models developed by the HEP community. This is because FID was specifically designed for natural image generation applications.

In previous works, HEP-specific neural networks were substituted to InceptionV3 to measure FID. In this work we take the same approach, and measure FID using 2 neural networks trained on HEP data and bring two contributions:

```
- Rather than the FID for a finite number of samples, we measure the effectively unbiased FID

- FID values obtained with different neural networks are not directly comparable. We introduce
```

The evaluation method we present could allow physicists to evaluate generative models more reliably and to directly compare them to the current state-of-the-art natural image generative models, allowing the progress of HEP ML applications to be monitored with respect to that of the industry.

References:

The authors intend to publish this work as a journal publication too.

Experiment context, if any:

This work was carried out in the context of the COherent Muon to Electron Transition (COMET) experiment and supervised by Professor Yoshi Uchida. The project was an integral part of the integrated master's in science (MSci) degree at Imperial College London. It builds on the work of the COMET collaboration and the Imperial COMET group but is independent of the collaboration.

Track 1: Computing Technology for Physics Research / **58**

Next generation task scheduler for ATLAS software framework

Authors: Beojan Stanislaus[1]; Charles Leggett[2]; Julien Esseiva[1]; Paolo Calafiura[1]; Vakho Tsulaia[1]; Xiangyang Ju[1]

[1] *Lawrence Berkeley National Lab. (US)*

[2] *Lawrence Berkeley National Lab (US)*

Corresponding Author: beojan.stanislaus@cern.ch

Experiments at the CERN High-Luminosity Large Hadron Collider (HL-LHC) will produce hundreds of Petabytes of data per year. Efficient processing of this dataset represents a significant human resource and technical challenge. Today, ATLAS data processing applications run in multi-threaded mode, using Intel TBB for thread management, which allows efficient utilization of all available CPU cores on the computing resources. However, modern HPC systems and high-end computing clusters are increasingly based on heterogeneous architectures, usually a combination of CPU and accelerators (e.g., GPU, FPGA). To run ATLAS software on these machines efficiently, we started developing a distributed, fine-grained, vertically integrated task scheduling software system. A first simplified implementation of such a system called Raythena was developed in late 2019. It is based on Ray - a high-performance distributed execution platform developed by Riselab at UC Berkeley. Raythena leverages the ATLAS event-service architecture for efficient utilization of CPU resources on HPC systems by dynamically assigning fine-grained workloads (individual events or event ranges) to ATLAS data-processing applications running simultaneously on multiple HPC compute nodes.

The main purpose of the Raythena project was to gain the experience of developing real-life applications with the Ray platform. However, in order to achieve our main objective, we need to design a new system capable of utilizing heterogeneous computing resources in a distributed environment. To accomplish this, we have started to evaluate HPX as an alternative to TBB/Ray. HPX

is a C++ library for concurrency and parallelism developed by the Stellar group, which exposes a uniform, standards-oriented API for programming parallel, distributed, and heterogeneous applications.

This presentation will describe the preliminary results of the evaluation of HPX for implementation of the task scheduler for ATLAS data-processing applications aimed to enable cross-node scheduling in heterogeneous systems that offer a mixture of CPU and GPU architectures. We present the prototype applications implemented using HPX and the preliminary results of performance studies of these applications.

Significance:

This presentation describes design ideas and first simple prototype implementations of the distributed and heterogeneous task scheduling system for the ATLAS experiment. Given the increased data volumes expected to be recorded in the era of HL LHC, it becomes critical for the experiments to efficiently utilize all available computing resources, including the new generation of supercomputers, most of which will be based on heterogeneous architectures.

References:

Experiment context, if any:

ATLAS

Poster session with coffee break / **59**

Faster simulated track reconstruction in the ATLAS Fast Chain

Authors: Debajyoti Sengupta[1]; Fang-Ying Tsai[2]; William Axel Leight[3]

[1] *Universite de Geneve (CH)*

[2] *Stony Brook University (US)*

[3] *University of Massachusetts Amherst*

Corresponding Author: william.axel.leight@cern.ch

The production of simulated datasets for use by physics analyses consumes a large fraction of ATLAS computing resources, a problem that will only get worse as increases in the instantaneous luminosity provided by the LHC lead to more collisions per bunch crossing (pile-up). One of the more resource-intensive steps in the Monte Carlo production is reconstructing the tracks in the ATLAS Inner Detector (ID), which takes up about 60% of the total detector reconstruction time [1]. This talk discusses a novel technique called track overlay, which substantially speeds up the ID reconstruction. In track overlay the pile-up ID tracks are reconstructed ahead of time and overlaid onto the ID tracks from the simulated hard-scatter event. We present our implementation of this track overlay approach as part of the ATLAS Fast Chain simulation, as well as a method for deciding in which cases it is possible to use track overlay in the reconstruction of simulated data without performance degradation.

[1] ATL-PHYS-PUB-2021-012 (60% refers to Run3, mu=50, including large-radius tracking, p11)

Significance:

This presentation covers a new method of speeding up the reconstruction of simulated data in ATLAS that will become necessary with the high luminosities of the HL-LHC.

References:

Experiment context, if any:

ATLAS

Track 1: Computing Technology for Physics Research / 60

The Level 1 Scouting system of the CMS experiment

Author: Thomas Owen James[1]

[1] *CERN*

Corresponding Author: thomas.owen.james@cern.ch

A novel data collection system, known as Level-1 (L1) Scouting, is being introduced as part of the L1 trigger of the CMS experiment at the CERN Large Hadron Collider. The L1 trigger of CMS, implemented in FPGA-based hardware, selects events at 100 kHz for full read-out, within a short 3 microsecond latency window. The L1 Scouting system collects and stores the reconstructed particle primitives and intermediate information of the L1 trigger processing chain, at the full 40 MHz bunch crossing rate. This system will provide vast amounts of data for detector diagnostics, luminosity measurements, and the study of otherwise inaccessible signatures, either too common to fit in the L1 accept budget, or with requirements orthogonal to the standard physics triggers. Demonstrator systems consisting of PCIe-based FPGA stream-processing boards and associated host PCs have ben deployed at CMS to capture data from both the Global Muon Trigger (GMT), and Calorimeter Trigger sub-systems. In addition, a neural-network based re-calibration and fake identification engine has been developed using the Micron Deep Learning Accelerator (MDLA) FPGA framework. An overview of the new system, and the first results from 2022 data taking will be shown. Plans and development progress towards the continued expansion of the L1 Scouting system throughout LHC Run 3, and for Phase II of CMS at the High Luminosity LHC, will also be presented.

Significance:

First results with LHC Run 3 data taking, with a new and novel data collection system (L1 scouting).

References:

Experiment context, if any:

CMS

Track 1: Computing Technology for Physics Research / 61

The new GPU-based HPC cluster at ReCaS-Bari

Author: Gioacchino Vino[1]

Co-authors: Alessandro Italiano [2]; Domenico Elia [3]; Giacinto Donvito [4]; Marica Antonacci [5]

[1] *INFN Bari (IT)*

[2] *INFN - National Institute for Nuclear Physics*

[3] *INFN Bari*

[4] *Universita e INFN, Bari (IT)*

[5] *INFN*

Corresponding Author: gioacchino.vino@cern.ch

The ReCaS-Bari datacenter enriches its service portfolio providing a new HPC/GPU cluster for Bari University and INFN users. This new service is the best solution for complex applications requiring a massively parallel processing architecture. The cluster is equipped with cutting edge Nvidia GPUs, like V100 and A100, suitable for those applications able to use all the available parallel hardware. Artificial intelligence, complex model simulation (weather and earthquake forecasts, molecular dynamics and galaxy formation) and all high precision floating-point based applications are possible candidates to be executed on the new service. The cluster is composed of 10 machines with a total computing resource equals to 1755 cores, 13.7 TB RAM, 55 TB local disk and 38 high performance GPUs (18 Nvidia A100 and 20 Nvidia V100). Each node can access the ReCaS-Bari distributed storage based on GPFS equals to 8.3 PB. Applications are executed only within Docker containers, conferring to the HPC/GPU cluster features like easy application configuration and execution, reliability, flexibility and security. Currently, users are able to choose among different ready-to-use services like remote IDEs (Jupyter Notebook and RStudio), by which execute GPU based applications, or a job orchestration to whom submit complex workflow represented as DAG (Directed Acyclic Graphs). The user service portfolio is in evolution. If the provided user services do not cover the user needs, user-defined Docker containers can be executed on the Cluster. Long running services and job submission are managed with Marathon and Chronos respectively, two frameworks running along with Apache Mesos. These three tools add high availability, fault tolerant and security additional to the native capacity to manage all compute resources and user requests. The implemented technological solution allows users to continue to access their own data both from HTC cluster (based on HTCondor) and from HPC/GPU Cluster, based on Mesos.
The first phase, where local beta-testers used the cluster, concluded successfully. The service is now ready to join the national INFN-Cloud federation. Leveraging the INDIGO PaaS orchestrator, provides multiple ready-to-used frameworks and services (ML_INFN, Apache Spark, JupyterLab, ...), a stable and secure authentication layer, a simple web dashboard that can be used to deploy services on top of and an heterogeneous set of resources. The evolution of the service, where a performance evaluation of Kubernetes as replacement of Apache Mesos, is in the pipeline.
In this contribution will be presented and discussed resources and technological solutions related to the HPC/GPU Cluster in the ReCaS-Bari data center and the most important applications running on the cluster.

Significance:

References:

Experiment context, if any:

Track 3: Computations in Theoretical Physics: Techniques and Methods / **62**

DMG4: a fully GEANT4-compatible package for the simulation of Dark Matter

Author: Henri Hugo Sieber[1]

[1] *ETH Zurich (CH)*

Corresponding Author: henri.hugo.sieber@cern.ch

The search of New Physics through Dark Sectors is an exciting possibility to explain, among others, the origin of Dark Matter (DM). Within this context, the sensitivity study of a given experiment is a key point in estimating its potential for discovery. In this contribution we present the fully GEANT4-compatible Monte Carlo simulation package for production and propagation of DM particles, DMG4. In particular, we discuss the implementation of production cross-sections in its GEANT4-independent sub-package, DarkMatter, and DMG4 latest release, including a finer application programming interface (API) to GEANT4. We also cover its recent developments with faster and more accurate cross-sections computations, sampling methods, extended energy range, as well as the expansion of the package to $B-L$ and semi-visible models. We finally discuss the improvements in the simulations of New Physics processes specific to muon beams.

Significance:

References:

https://www.sciencedirect.com/science/article/abs/pii/S0010465521002411?via%3Dihub

Experiment context, if any:

NA64

Poster session with coffee break / 63

The adaptation of a deep learning model to locating primary vertices in the CMS and ATLAS experiments

Authors: Elliott Kauffman[1]; Henry Fredrick Schreiner[2]; Lauren Alexandra Tompkins[3]; Michael David Sokoloff[4]; Michael Peters[5]; Rida Shahid[None]; Rocky Bala Garg[3]; Simon Akar[4]; Will Tepe[None]

[1] *Duke University (US)*

[2] *Princeton University*

[3] *Stanford University (US)*

[4] *University of Cincinnati (US)*

[5] *University of Cincinatti*

Corresponding Author: elliott.kauffman@duke.edu

Over the past several years, a deep learning model based on convolutional neural networks has been developed to find proton-proton collision points (also known as primary vertices, or PVs) in Run 3 LHCb data. By converting the three-dimensional space of particle hits and tracks into a one-dimensional kernel density estimator (KDE) along the direction of the beamline and using the KDE as an input feature into a neural network, the model has achieved an efficiency of 98% with a low false positive rate. The success of this method motivates its extension to other experiments, including ATLAS and CMS. Although LHCb is a forward spectrometer and ATLAS and CMS are central detectors, both ATLAS and CMS have the necessary characteristics to compute KDEs analogous to the LHCb detector. While the ATLAS and CMS detectors will benefit from higher precision, the expected number of visible PVs per event will be approximately 10 times that for LHCb, resulting in only slightly altered KDEs. The KDE and a few related input features are fed into the same neural network architectures used to achieve the results for LHCb. We present the development of the input feature and initial results across different network architectures. The results serve as a proof-of-principle that a deep neural network can achieve high efficiency and low false positive rates for finding vertices in ATLAS and CMS data.

Significance:

The work presented will demonstrate that deep neural network architecture designed to find primary vertices in LHCb data also works for ATLAS and CMS data, which come from central detectors rather than a forward detector.

References:

https://arxiv.org/abs/2103.04962

Experiment context, if any:

ATLAS, CMS, LHCb

Poster session with coffee break / 64

Evolution of the CMS Submission Infrastructure to support heterogeneous resources in the LHC Run 3

Author: Antonio Perez-Calero Yzquierdo[1]

Co-authors: Edita Kizinevic [2]; Farrukh Aftab Khan [3]; Hyunwoo Kim [3]; Marco Mascheroni [4]; Maria Acosta Flechas [3]; Saqib Haleem [5]

[1] *Centro de Investigaciones Energéticas Medioambientales y Tecnológicas*

[2] *CERN*

[3] *Fermi National Accelerator Lab. (US)*

[4] *Univ. of California San Diego (US)*

[5] *National Centre for Physics (PK)*

Corresponding Author: antonio.perez.calero.yzquierdo@cern.ch

The landscape of computing power available for the CMS experiment is rapidly evolving, from a scenario dominated by x86 processors deployed at WLCG sites, towards a more diverse mixture of Grid, HPC, and Cloud facilities incorporating a higher fraction of non-CPU components, such as GPUs. Using these facilities' heterogeneous resources efficiently to process the vast amounts of data to be collected in the LHC Run3 and beyond, in the HL-LHC era, is key to CMS's achieving its scientific goals.

The CMS Submission Infrastructure is the main computing resource provisioning system for CMS workflows, including data processing, simulation and analysis. It currently aggregates nearly 400k CPU cores distributed worldwide from Grid, HPC and cloud providers. The Submission Infrastructure, together with other elements in the CMS workload management, has been modified in its strategies and enlarged in its scope to make use of these new resources.

In this evolution, key questions such as the optimal level of granularity in the description of the resources, or how to prioritize workflows in this new resource mix must be taken into consideration. In addition, access to many of these resources is considered opportunistic by CMS, thus each resource provider may also play a key role in defining particular allocation policies, diverse from the up-to-now dominant system of pledges. All these matters must be addressed in order to ensure the efficient allocation of resources and matchmaking to tasks to maximize their use by CMS.

This contribution will describe the evolution of the CMS Submission Infrastructure towards a full integration and support of heterogeneous resources according to CMS needs. In addition, a study of the pool of GPUs already available to CMS Offline Computing will be presented, including a survey of their diversity in relation to CMS workloads, and the scalability reach of the infrastructure to support them.

References:

Experiment context, if any:

The CMS experiment at the LHC at CERN

Significance:

The Submission Infrastructure is the main component of the resource acquisition and workload to resource matchmaking systems in CMS Offline Computing. It is therefore mandatory to adapt it to be able to send GPU allocation requests to resource providers, to integrate those GPUs into the CMS HTCondor infrastructure, and finally to optimize workload to heterogeneous resource assignment in order for CMS to succeed in this future vasts amount of computing power available in the form of GPUs. This contribution will present how this has been achieved, and indeed the already existing pool of GPUs ready for CMS use.

Poster session with coffee break / 65

Stability of the CMS Submission Infrastructure for the LHC Run 3

Author: Antonio Perez-Calero Yzquierdo[1]

Co-authors: Edita Kizinevic [2]; Farrukh Aftab Khan [3]; Hyunwoo Kim [3]; Marco Mascheroni [4]; Maria Acosta Flechas [3]; Saqib Haleem [5]

[1] *Centro de Investigaciones Energéticas Medioambientales y Tecnológicas*

[2] *CERN*

[3] *Fermi National Accelerator Lab. (US)*

[4] *Univ. of California San Diego (US)*

[5] *National Centre for Physics (PK)*

Corresponding Author: antonio.perez.calero.yzquierdo@cern.ch

The CMS Submission Infrastructure is the main computing resource provisioning system for CMS workflows, including data processing, simulation and analysis. It currently aggregates nearly 400k CPU cores distributed worldwide from Grid, HPC and cloud providers. CMS Tier-0 tasks, such as data repacking and prompt reconstruction, critical for data-taking operations, are executed on a collection of computing resources at CERN, also managed by the CMS Submission Infrastructure.

All this computing power is harnessed via a number of federated resource pools, supervised by HTCondor and GlideinWMS services. Elements such as pilot factories, job schedulers and connection brokers are deployed in HA mode across several "availability zones", providing stability to our services via hardware redundancy and numerous failover mechanisms.

Given the upcoming start of the LHC Run 3, the Submission Infrastructure stability has been recently tested in a series of controlled exercises, performed without interruption of our services. These tests have demonstrated the resilience of our systems, and additionally provided useful information in order to further refine our monitoring and alarming system.

This contribution will describe the main elements in the CMS Submission Infrastructure design and deployment, along with the performed failover exercises, proving that our systems are ready to serve their critical role in support of CMS activities.

References:

Experiment context, if any:

The CMS experiment at the LHC at CERN

Significance:

This presentation will cover how the CMS Submission Infrastructure (SI) has been designed and set up to avoid single points of failure, along with the tests performed in order to verify its resilience and stability, considering that the SI plays a critical role in the capability of the CMS experiment's Tier-0 node to take and process collisions data.

Track 1: Computing Technology for Physics Research / **66**

The Awkward World of Python and C++

Authors: Manasvi Goyal[1]; Ianna Osborne[2]; Jim Pivarski[2]

[1] *Delhi Technological University*

[2] *Princeton University*

Corresponding Author: manasvigoyal_2k19pe034@dtu.ac.in

There are undeniable benefits of binding Python and C++ to take advantage of the best features of both languages. This is especially relevant to the HEP and other scientific communities that have invested heavily in the C++ frameworks and are rapidly moving their data analyses to Python.

The version 2 of Awkward Array, a Scikit-HEP Python library, introduces a set of header-only C++ libraries that do not depend on any application binary interface. The users can directly include these libraries in their compilation, rather than linking against platform-specific libraries. This new development makes the integration of Awkward Arrays into other projects easier and more portable as the implementation is easily separable from the rest of the Awkward Array codebase.

The code is minimal, it does not include all of the code needed to use Awkward Arrays in Python, nor does it include references to Python or pybind11. The C++ users can use it to make arrays and then copy them to Python without any specialised data types - only raw buffers, strings, and integers. This C++ code also simplifies the process of JIT-compilation in ROOT. This implementation approach solves some of the drawbacks like packaging projects where native dependencies can be challenging.

In this talk, we will demonstrate the techniques of exposing C++ classes and their methods to Python and vice versa. We will also describe the implementation of a new LayoutBuilder and a GrowableBuffer that are more performant in building the Awkward Arrays as compared to the previous approach. Furthermore, examples of wrapping the C++ data into Awkward Arrays and exposing Awkward Arrays to C++ without copying them will be discussed.

Significance:

This submission represents our evolving view of best practices for creating Awkward Arrays in C++. Previously, the main codebase was written in C++ with the idea that downstream code would link to libawkward.so, but as Jim Pivarski described in his talk at the last ACAT, that route is full of hidden gotchas. This method of a small, header-only library that only fills array buffers for downstream code to pass from C++ to Python using C types (integers and raw pointers) has considerably more promise. Already, two applications are ready to use it: Awkward ←→ RDataFrame (which needs the header-only library to JIT-compile) and ctapipe in gamma ray astronomy (which has array types that are known at compile-time).

References:

Experiment context, if any:

IRIS-HEP

Poster session with coffee break / **67**

A Deep Learning based algorithm for PID study with cluster counting

Author: Guang Zhao[1]

[1] *Institute of High Energy Physics*

Corresponding Author: zhaog@ihep.ac.cn

Ionization of matters by charged particles are the main mechanism for particle identification in gaseous detectors. Traditionally, the ionization is measured by the total energy loss (dE/dx). The concept of cluster counting, which measures the number of clusters per track length (dN/dx), was proposed in the 1970s. The dN/dx measurement can avoid many sources of fluctuations from the dE/dx measurement, which in the end can potentially have a resolution two times better than the dE/dx.

The dN/dx measurement requires highly efficient reconstruction algorithm. One need to determine the number of peaks associated with the primary electrons in the induced current waveform in a single detection unit. The main challenge of the algorithm is to handle the highly pileup situations of the single peaks and to discriminate the primary peaks from the secondary electrons and noises. A machine learning based algorithm is developed for the cluster counting problem. The algorithm consists of a peak finding algorithm, which aims to find all peaks in the waveform, based on the Recurrent Neural Network (RNN). And a clustering algorithm, which is to determine the number of primary peaks, based on the Convolutional Neural Network (CNN).

In the talk, the basic idea of cluster counting and the reconstruction algorithm based on machine learning will be presented.

Significance:

The peak finding is essential for the cluster counting technique. The new developed algorithm based on machine learning overcomes the traditional algorithm such as derivatives for the peak finding.

References:

Experiment context, if any:

The study is applied for the drift chamber design in Circular Electron Positron Collider (CEPC).

Track 2: Data Analysis - Algorithms and Tools / 68

End-to-end multi-particle reconstruction in high occupancy imaging calorimeters with graph neural networks

Authors: Jan Kieseler[1]; Kenneth Long[2]; Maurizio Pierini[1]; Nadya Chernyavskaya[1]; Oleksandr Viazlo[3]; Philipp Zehetner[4]; Raheel Nawaz[5]; Shah Rukh Qasim[6]

[1] *CERN*

[2] *Massachusetts Inst. of Technology (US)*

[3] *Florida State University (US)*

[4] *Ludwig Maximilians Universitat (DE)*

[5] *Staffordshire University*

[6] *Manchester Metropolitan University (GB)*

Corresponding Author: philipp.zehetner@cern.ch

We present an end-to-end reconstruction algorithm to build particle candidates from detector hits in next-generation granular calorimeters similar to that foreseen for the high-luminosity upgrade of the CMS detector. The algorithm exploits a distance-weighted graph neural network, trained with object condensation, a graph segmentation technique. Through a single-shot approach, the reconstruction task is paired with energy regression. We describe the reconstruction performance in terms of efficiency as well as in terms of energy resolution. In addition, we show the jet reconstruction performance of our method and discuss its inference computational cost. To our knowledge, this work is the first-ever example of single-shot calorimetric reconstruction of 𝒪(1000) particles in high-luminosity conditions with 200 pileup.

Significance:

To our knowledge, this work is the first-ever example of single-shot calorimetric reconstruction with machine learning of O(1000) particles in high-luminosity conditions with up to 200 pileup.

References:

arXiv:2204.01681
arXiv:1902.07987

arXiv:2002.03605

Experiment context, if any:

Loosely related to CMS HGCAL

Track 2: Data Analysis - Algorithms and Tools / 69

Automatic data processing for prompt calibration of the CMS ECAL

Author: Simone Pigazzini[1]

[1] *ETH Zurich (CH)*

Corresponding Author: simone.pigazzini@cern.ch

The CMS ECAL has achieved an impressive performance during the LHC Run1 and Run2. In both runs, the ultimate performance has been reached after a lengthy calibration procedure required to correct ageing-induced changes in the response of the channels. The CMS ECAL will continue its operation far beyond the ongoing LHC Run3: its barrel section will be upgraded for the LHC Phase-2 and it will be operated for the entire duration of the High Luminosity HLC program. With the increase of instantaneous luminosity, the ageing effects will increase, and so will the required frequency of calibrations: it is therefore crucial for the CMS ECAL community to reduce the time and resources needed for this task, in order to ensure with limited personpower a smooth operation and excellent performance on the long term. A new system has been developed during the LHC second long shut down to automatically execute the calibration workflows on a daily basis during the data taking. The new system is based on industry standard tools (Openshift, Jenkins, Influxdb, and Grafana) and provides a general interface to orchestrate standalone workflows written in different programming languages. It also provides interfaces to other existing CMS systems to steer the processing of selected data streams and to upload newly computed calibration into the database used for the data processing for physics analyses. The new system is designed with the ambitious goal of cutting the time needed to provide the best possible performance for physics analyses by one order of magnitude. The system offers an extensive suite of diagnostic tools that provide a constant monitoring of its status as well as the option to send alerts in case of problems. In this talk, the general structure of the system will be presented, along with the results from the first year of operation. The detail of the monitoring and alert system will also be discussed.

Significance:

The presentation will discuss a profound change in how detector calibration is performed (on the computational side) in the CMS ECAL. Despite continuos and autonomous calibration system are wide spread in HEP experiments, with this work we pushed the automation boundaries to include the high level refinements that are usually computed only after data-taking. Another crucial aspect is the versatility of the system, which provide simple interface to integrate existing calibration workflows rather than requiring a complete re-writing of existing code. As such the system can be easily ported to other experiments.
The system is currently in the commissioning phase, the experience we have been gaining during this first year of operation will pave the way for the design of a more general system for the calibration of the CMS detector during HL-LHC.

References:

Experiment context, if any:

CMS, LHC

Track 3: Computations in Theoretical Physics: Techniques and Methods / **70**

Conditional Born machine for Monte Carlo events generation

Authors: Enrique Kajomovitz Must[1]; Michele Grossi[2]; Oriel Orphee Moira Kiss[3]; Sofia Vallecorsa[2]

[1] *Technion, Israel Institute of Technology*

[2] *CERN*

[3] *Universite de Geneve (CH)*

Corresponding Author: michele.grossi@cern.ch

The potential exponential speed-up of quantum computing compared to classical computing makes it to a promising method for High Energy Physics (HEP) simulations at the LHC at CERN.
Generative modeling is a promising task for near-term quantum devices, the probabilistic nature of quantum mechanics allows us to exploit a new class of generative models: quantum circuit Born machine (QCBM).
These models use the stochastic nature of quantum measurement as random-like sources and have no classical analog.
More specifically, they produce samples from the underlying distribution of a pure quantum state by measuring a parametrized quantum circuit with probability given by the Born rule
This work presents an application of Born machines to Monte Carlo simulations and extends their reach to multivariate and conditional distributions.
Even if generating multivariate distributions with Born machines has already been explored, we propose an alternative circuit design with a reduced connectivity, better suited for NISQ devices.
Indeed, models are run on (noisy) simulators and IBM Quantum superconducting devices.
More specifically, Born machines are used to generate muonic force carriers (MFC) events resulting from scattering processes between muons and the detector material in high-energy-physics colliders experiments. MFCs are bosons appearing in beyond the standard model theoretical frameworks, which are candidates for dark matter. Empirical evidences suggest that Born machines can reproduce the underlying distribution of datasets coming from Monte Carlo simulations, and are competitive with classical machine learning-based generative models of similar complexity.

Significance:

The submitted idea represents an extension of a prior work that incorporate important updated coming from external review and multidisciplinary discussion.

References:

https://arxiv.org/abs/2205.07674

Experiment context, if any:

Track 2: Data Analysis - Algorithms and Tools / **71**

Generative Models for Fast Simulation of Electromagnetic and Hadronic Showers in Highly Granular Calorimeters

Authors: Anatolii Korol[1]; Daniel Hundhausen[2]; Engin Eren[3]; Erik Buhmann[4]; Frank-Dieter Gaede[3]; Gregor Kasieczka[4]; Katja Kruger[3]; Lennart Rustige[5]; Peter McKeown[6]; Sascha Daniel Diefenbacher[4]; William Korcari[4]

[1] *Centre National de la Recherche Scientifique (FR)*

[2] *Deutsches Elektronen-Synchrotron DESY*

[3] *Deutsches Elektronen-Synchrotron (DE)*

[4] *Hamburg University (DE)*

[5] *CDCS / DESY*

[6] *DESY*

Corresponding Authors: anatolii.korol@cern.ch, sascha.diefenbacher@desy.de, engin.eren@cern.ch, erik.buhmann@desy.de

Simulation in High Energy Physics (HEP) places a heavy burden on the available computing resources and is expected to become a major bottleneck for the upcoming high luminosity phase of the LHC and for future Higgs factories, motivating a concerted effort to develop computationally efficient solutions. Methods based on generative machine learning methods hold promise to alleviate the computational strain produced by simulation while providing the physical accuracy required of a surrogate simulator.

In this contribution, an overview of a growing body of work focused on simulating showers in highly granular calorimeters will be reported, which is making significant steps towards realistic fast simulation tools based on deep generative models. Progress on the simulation of both electromagnetic and hadronic showers will be presented, with a focus on the high degree of physical fidelity and computational performance achieved. Additional steps taken to address the challenges faced when broadening the scope of these simulators, such as those posed by multi-parameter conditioning, will also be discussed.

Significance:

Generative modeling of hadron shower simulation and high fidelity angular simulation of photons

References:

https://iopscience.iop.org/article/10.1088/2632-2153/ac7848

Experiment context, if any:

Track 2: Data Analysis - Algorithms and Tools / 72

Particle Tracking with Noisy Intermediate-Scale Quantum Computers

Authors: Cenk Tuysuz[1]; Cigdem Issever[1]; Hannsjorg Weber[2]; Karl Jansen[3]; Stefan Kühn[4]; Teng Jian Khoo[2]; Tim Schwägerl[1]

[1] *Humboldt University of Berlin and DESY (DE)*

[2] *Humboldt University of Berlin (DE)*

[3] *DESY*

[4] *Computation-based Science and Technology Research Center, The Cyprus Institute*

Corresponding Author: tim.schwaegerl@desy.de

Particle track reconstruction poses a key computing challenge for future collider experiments. Quantum computing carries the potential for exponential speedups and the rapid progress in quantum hardware might make it possible to address the problem of particle tracking in the near future. The solution of the tracking problem can be encoded in the ground state of a Quadratic Unconstrained Binary Optimization. In our study, sets of three hits in the detector are grouped into triplets. True triplets are part of trajectories of particles, while false triplets are random combinations of three hits. By approximating the ground state, the Variational Quantum Eigensolver algorithm aims at identifying true triplets. Different circuits and optimizers are tested for small instances of the tracking problem with up to 23 triplets. Precision and recall are determined in a noiseless simulation and the effects of readout errors are studied. It is planned to repeat the experiments on real hardware and to combine the solutions of small instances to address the full-scale tracking problem.

Significance:

We present a comprehensive study to tackle small instances of the particle tracking problem at an ATLAS-like detector using the Variational Quantum Eigensolver algorithm

References:

Experiment context, if any:

ATLAS, LHC

Track 2: Data Analysis - Algorithms and Tools / **73**

Standalone track reconstruction in LHCb's SciFi detector for the GPU-based High Level Trigger

Authors: Arantza De Oyanguren Campos[1]; Arthur Hennequin[2]; Brij Kishor Jashal[3]; Christina Agapopoulou[4]; Jiahui Zhuo[1]; Louis Henry[5]; Lukas Calefice[6]

[1] *Univ. of Valencia and CSIC (ES)*

[2] *Massachusetts Inst. of Technology (US)*

[3] *Tata Inst. of Fundamental Research (IN)*

[4] *Centre National de la Recherche Scientifique (FR)*

[5] *CERN*

[6] *Technische Universität Dortmund (DE), LPNHE/Sorbonne Université (FR)*

Corresponding Author: arthur.hennequin@cern.ch

As part of the Run 3 upgrade, the LHCb experiment has switched to a two stage event trigger, fully implemented in software. The first stage of this trigger, running in real time at the collision rate of 30MHz, is entirely implemented on commercial off-the-shelf GPUs and performs a partial reconstruction of the events.
We developed a novel strategy for this reconstruction, starting with two independent tracking algorithms, in the VELO and SciFi detectors, forming track segments which are then matched and merged to form full tracks, suitable for selecting events at LHCb. A key point enabling this sequence is the SciFi tracking algorithm, which was implemented for GPU with special care in order to meet the throughput requirements of a real time trigger.
Developing such algorithm is challenging due to the high number of track hypothesis that needs to be tested. We discuss how this challenge was overcome by using the GPU architecture efficiently and how the efficiency of the new sequence is compared to the current baseline reconstruction.

Significance:

References:

https://cds.cern.ch/record/2811214

Experiment context, if any:

LHCb

Poster session with coffee break / **74**

Commissioning CMS online reconstruction with GPUs

Author: CMS collaboration[None]

Corresponding Author: borislav.pavlov@cern.ch

Building on top of the multithreading functionality that was introduced in Run-2, the CMS software framework (CMSSW) has been extended in Run-3 to offload part of the physics reconstruction to

NVIDIA GPUs. The first application of this new feature is the High Level Trigger (HLT): the new computing farm installed at the beginning of Run-3 is composed of 200 nodes, and for the first time each one is equipped with two AMD Milan CPUs and two NVIDIA T4 GPUs. In order to guarantee that the HLT can run on machines without any GPU accelerators - for example as part of the large scale Monte Carlo production running on the grid - the HLT reconstruction has been implemented both for NVIDIA GPUs and for traditional CPUs.

CMS has undertaken a comprehensive validation and commissioning activity to ensure the successful operations of the new HLT farm and the reproducibility of the physics results while using either of the two implementations: some have taken place offline, on dedicated Tier-2 centres equipped with NVIDIA GPUs; other activities ran online during the LHC commissioning period, after installing GPUs on few of the nodes from the Run-2 HLT farm. The final steps were the optimisation of the HLT configuration, after the installation of the new HLT farm.

This contribution will describe the steps taken to validate the GPU-based reconstruction and commission the new HLT farm, leading to the successful data taking activities after the LHC Run-3 start up.

Significance:

References:

Experiment context, if any:

CMS Collaboration

Poster session with coffee break / **75**

Progress towards an improved particle flow algorithm at CMS with machine learning

Author: CMS Collaboration[None]

Corresponding Author: borislav.pavlov@cern.ch

The particle-flow (PF) algorithm is of central importance to event reconstruction at the CMS detector, and has been a focus of developments in light of planned Phase-2 running conditions with an increased pileup and detector granularity. Current rule-based implementations rely on extrapolating tracks to the calorimeters, correlating them with calorimeter clusters, subtracting charged energy and creating neutral particles from significant energy deposits. Such rule-based algorithms can be difficult to extend and may be computationally inefficient under high detector occupancy, while also being challenging to port to heterogeneous architectures in full detail.

In recent years, end-to-end machine learning approaches for event reconstruction have been proposed, including for PF at CMS, with the possible advantage of directly optimising for the physical quantities of interest, being highly reconfigurable to new conditions, while also being a natural fit for deployment on heterogeneous accelerators.

One of the proposed approaches for machine-learned particle-flow (MLPF) reconstruction relies on graph neural networks to infer the full particle content of an event from the tracks and calorimeter clusters based on a training on simulated samples, and has been recently implemented in CMS as a possible future reconstruction R&D direction to fully map out the characteristics of such an approach in a realistic setting.

We discuss progress in CMS towards an improved implementation of the MLPF reconstruction, now optimised on generator-level particle information for the first time to our knowledge, thus paving the way to potentially improving the detector response in terms of physical quantities of interest. We show detailed physics validation with respect to the current PF algorithm in terms of high-level physical quantities such as jet and MET resolution. Furthermore, we discuss progress towards deploying

the MLPF algorithm in the CMS software framework on heterogeneous platforms, performing large-scale hyperparameter optimization using HPC systems, as well as the possibilities of making use of explainable artificial intelligence (XAI) to interpret the output.

Significance:

References:

Experiment context, if any:

CMS Collaboration

Track 3: Computations in Theoretical Physics: Techniques and Methods / **77**

Integrations with a neural network

Author: Daniel Maitre[1]

[1] *IPPP, Durham University*

Corresponding Author: daniel.maitre@durham.ac.uk

In this presentation I will show how one can perform parametric integrations using a neural network. This could be applied for example to perform the integration over the auxiliary parameters in the integrals that result from the sector decomposition of multi-loop integrals.

Significance:

This method would allow for a much faster evaluation of parametric integrals (at the cost of precison)

References:

Experiment context, if any:

Poster session with coffee break / **78**

An Autoencoder-based Online Data Quality Monitoring for CMS ECAL

Authors: Abhirami Harilal[1]; Kyungmin Park[1]; Manfred Paulini[1]; Michael Andrews[1]

[1] *Carnegie-Mellon University (US)*

Corresponding Author: abhirami.harilal@cern.ch

The online Data Quality Monitoring (DQM) system of the CMS electromagnetic calorimeter (ECAL) is a vital operations tool that allows ECAL experts to quickly identify, localize, and diagnose a broad range of detector issues that would otherwise hinder physics-quality data taking. Although the existing ECAL DQM system has been continuously updated to respond to new problems, it remains one step behind new and never-before-seen issues. As the ECAL electronics continue to age, previously rare and obscure failure modes have become more common, emphasizing the need for a more robust anomaly detection system. Using unsupervised deep learning, a real-time autoencoder-based anomaly detection system is developed that is able to detect ECAL anomalies unseen in past data. After accounting for spatiotemporal variations in the response of the ECAL, the new system is able to efficiently detect anomalies while maintaining an estimated false discovery rate between 10^{-2}

to 10^{-4}, besting existing benchmarks by several orders of magnitude. The real-world performance of the system is validated using anomalies found in 2018 data taking and with early data taken from 2022 collisions.

Significance:

This presentation will show results from the Ecal Endcaps not previously presented and new results from early Run 3 data which shows the live performance of the auto encoder DQM system which was trained on Run 2 data.

References:

Previously presented at APS April 2022 meeting: https://meetings.aps.org/Meeting/APR22/Session/X09.5

Experiment context, if any:

CMS

Poster session with coffee break / **79**

Event Display Development for Mu2e using Eve-7

Authors: Namitha Chithirasreemadam[1]; Sophie Middleton[2]

Co-author: Simone Donati

[1] *University of Pisa*

[2] *Caltech*

Corresponding Author: n.chithirasreemad@studenti.unipi.it

The Mu2e experiment will search for the CLFV neutrinoless coherent conversion of muon to electron, in the field of an Aluminium nucleus. A custom offline event display has been developed for Mu2e using TEve, a ROOT based 3-D event visualisation framework. Event displays are crucial for monitoring and debugging during live data taking as well as for public outreach. A custom GUI allows event selection and navigation. Reconstructed data like the tracks, hits and clusters can be displayed within the detector geometries upon GUI request. True Monte Carlo trajectory of particles traversing the muon beam line, obtained directly from Geant4 can also be displayed. Tracks are coloured according to their particle ID and users can select the trajectories to be displayed. Reconstructed tracks are refined using a Kalman filter. The resulting tracks can be displayed alongside truth information, allowing visualisation of the track resolution. The user can remove/add data based on energy deposited in a detector or arrival time. This is a prototype and an online event display, is currently under-development using Eve-7 which allows remote access for live data taking and lets multiple users to simultaneously view and interact with the display.

References:

Experiment context, if any:

Mu2e is an upcoming experiment at the Fermilab. It will search for the Charged Lepton Flavour Violating process of neutrinoless, coherent conversion of muon to electron in the field of an Al nucleus. We develop a custom, sophisticated yet user friendly event display for the experiment that would be useful to the experts and amateurs.

Significance:

An Event Display is the top layer of a robust framework, helping to visualise the physics in each event. They are crucial in the early planning stages of the experiment, for debugging of simulation and reconstruction codes, in detector calibration, physics analysis, online monitoring as well as for public outreach. We have developed a custom display for Mu2e using Eve-7. Art (Mu2e software framework) and Eve are both ROOT based frameworks which integrates the display well with the Mu2e environment, with

full access to all Mu2e data products. A custom GUI has been developed for the display making it user friendly. It is also useful during analysis as the user can access all the event details through the ROOT Browser button available on the main window. Therefore, this is a sophisticated yet easy to use, bespoke display of Mu2e.

Poster session with coffee break / 80

Machine learning techniques for data quality monitoring at the CMS detector

Authors: Luka Lambrecht[1]; Rosamaria Venditti[2]

[1] *Ghent University (BE)*

[2] *Universita e INFN, Bari (IT)*

Corresponding Author: rosamaria.venditti@cern.ch

The CMS experiment employs an extensive data quality monitoring (DQM) and data certification (DC) procedure. Currently, this approach consists mainly of the visual inspection of reference histograms which summarize the status and performance of the detector. Recent developments in several of the CMS subsystems have shown the potential of computer-assisted DQM and DC using autoencoders, spotting detector anomalies with high accuracy and a much finer time granularity than previously accessible. We will discuss a case study for the CMS pixel tracker, as well as the development of a common infrastructure to host computer-assisted DQM and DC workflows. This infrastructure facilitates accessing the input histograms, provides tools for preprocessing, training and validating, and generates an overview of potential detector anomalies.

Significance:

We show how machine learning or other computer-assisted techniques can be applied for data quality monitoring (DQM). A case study for the CMS pixel tracker has been presented recently in a poster at PM2021. Here, the intention is to focus more on the general infrastructure and strategy to enable machine learning assisted DQM across all CMS subdetectors.

References:

Poster and proceedings at PM2021: https://cds.cern.ch/record/2815415?ln=en
Public CMS note: https://cds.cern.ch/record/2812026?ln=en

Experiment context, if any:

CMS (CERN)

Poster session with coffee break / 81

Trigger Rate Monitoring Tools at CMS

Author: John Lawrence[1]

[1] *University of Notre Dame (US)*

Corresponding Author: jlawren6@nd.edu

With the start of run 3 in 2022, the LHC has entered a new period, now delivering higher energy and luminosity proton beams to the Compact Muon Solenoid (CMS) experiment. These increases make it

critical to maintain and upgrade the tools and methods used to monitor the rate at which data is collected (the trigger rate). Software tools have been developed to allow for automated rate monitoring, and we present several upgrades to these software tools, which maintain and expand on their functionality. These trigger rate monitoring tools allow for real-time monitoring including alerts which go out to on-call experts in the case of abnormalities. Fits are produced from previously collected data and extrapolate the behaviors of the triggers as a function of pile-up (the average number of particle interactions per bunch-crossing). These fits allow for visualization and statistical analysis of the behavior of the triggers and are displayed on the online monitoring system (OMS). The rate monitoring code can also be used for offline data certification and more complex trigger analysis. This presentation will show some of the upgrades to this software with an emphasis on the automation for easier and consistent upgrades and fixes to the software, and the increased interactivity with the users.

Significance:

References:

Experiment context, if any:

High Level Trigger at the Compact Muon Solenoid

Poster session with coffee break / 84

Particle Flow Reconstruction on Heterogeneous Architecture for CMS

Author: CMS Collaboration[None]

Corresponding Author: borislav.pavlov@cern.ch

The Particle Flow (PF) algorithm, used for a majority of CMS data analyses for event reconstruction, provides a comprehensive list of final-state state particle candidates and enables efficient identification and mitigation methods for simultaneous proton-proton collisions (pileup). The higher instantaneous luminosity expected during the upcoming LHC Run 3 will impose challenges for CMS event reconstruction. This will be amplified in the HL-LHC era, where luminosity and pileup rates are expected to be significantly higher. One of the approaches CMS is investigating to cope with this challenge is to adopt the heterogeneous computing architectures and accelerate event reconstruction. In this talk, we will discuss the effort to adopt the PF reconstruction to take advantage of GPU accelerators.

We will discuss the design and implementation of PF clustering for the CMS Electromagnetic and Hadronic Calorimeters using Cuda, including optimizations of the PF algorithm. The physics validation and performance of the GPU-accelerated algorithms will be demonstrated by comparing these to the CPU-based implementation.

Significance:

References:

Experiment context, if any:

CMS Collaboration

Poster session with coffee break / 85

Comparing and improving hybrid deep learning algorithms for identifying and locating primary vertices

Author: Michael Peters[None]

Co-authors: Michael David Sokoloff [1]; William Tepe [2]

[1] *University of Cincinnati (US)*

[2] *University of Cincinnati*

Corresponding Author: peter2mj@mail.uc.edu

Identifying and locating proton-proton collisions in LHC experiments (known as primary vertices or PVs) has been the topic of numerous conference talks in the past few years (2019-2021). Efforts to search for a variety of potential architectures have yielded potential candidates for PV-finder. The UNet model, for example, has achieved an efficiency of 98% with a low false-positive rate. These results can be obtained with numerous other neural network architectures. It also converges faster than any previous model. While this does not answer the question of how the algorithm learns, it does provide some useful insights into the open question. We present the results from this architectural study of different algorithms and their performance in locating PVs for LHCb data. The goal is to demonstrate progress in developing a performant architecture and evaluate different algorithms' learning.

Significance:

Provides analysis and comparison of different machine learning algorithms with respect to LHCb primary vertex finding. These results also show a near-upper limit with currently available data.

References:

https://arxiv.org/pdf/1906.08306.pdf
https://arxiv.org/abs/2103.04962

Experiment context, if any:

LHCb pv-finder

Track 1: Computing Technology for Physics Research / 86

Design and implementation of zstd compression algorithm for high energy physics experiment data processing based on FPGA

Author: Xuyang Zhou[None]

Co-authors: Haibo li ; Yaodong Cheng ; Yaosong Cheng ; Yu Gao ; Yujiang Bi [1]

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

Corresponding Author: zhouxuyang@ihep.ac.cn

With the continuous increase in the amount of large data generated and stored in various scientific fields ,such as cosmic ray detection, compression technology becomes more and more important in reducing the requirements for communication bandwidth and storage capacity. Zstandard, abbreviated as zstd, is a fast lossless compression algorithm. For zlib-level real-time compression scenarios, it can have a good compression ratio and a faster speed than similar algorithms. In this paper, we introduce the architecture of a new zstd compression kernel, and combine it with the root framework (an open-source data analysis framework used by high energy physics and others), and optimize the proposed architecture for the specific use case of lhaaso km2a data decode. The optimized kernel is implemented on Xilinx alveo U200 board.

Significance:

References:

Experiment context, if any:

Poster session with coffee break / **87**

Machine learning-based vertex reconstruction for reactor neutrinos in JUNO

Author: Wuming Luo[1]

[1] *Institute of High Energy Physics, Chinese Academy of Science*

Corresponding Author: luowm@ihep.ac.cn

Jiangmen Underground Neutrino Observatory (JUNO), located at the southern part of China, will be the world's largest liquid scintillator(LS) detector. Equipped with 20 kton LS, 17623 20-inch PMTs and 25600 3-inch PMTs in the central detector, JUNO will provide a unique apparatus to probe the mysteries of neutrinos, particularly the neutrino mass ordering puzzle. One of the challenges for JUNO is the high precision vertex reconstruction for reactor neutrino events. This talk will present machine learning-based vertex reconstruction in JUNO, particularly the comparison of different machine learning models as well as the optimization of the model inputs for better reconstruction performance.

Significance:

the content of this talk mainly comes from two papers, one has already been published and the other has been received by the Journal.

References:

Vertex and energy reconstruction in JUNO with machine learning methods (https://doi.org/10.1016/j.nima.2021.165527)
Improving the machine learning based vertex reconstruction for large liquid scintillator detectors with multiple types of PMTs (https://doi.org/10.48550/arXiv.2205.04039)

Experiment context, if any:

JUNO

Poster session with coffee break / **88**

Awkward Arrays to RDataFrame and back

Authors: Ianna Osborne[1]; Jim Pivarski[1]

[1] *Princeton University*

Corresponding Author: ianna.osborne@cern.ch

Awkward Arrays and RDataFrame provide two very different ways of performing calculations at scale. By adding the ability to zero-copy convert between them, users get the best of both. It gives users a better flexibility in mixing different packages and languages in their analysis.

In Awkward Array version 2, the ak.to_rdataframe function presents a view of an Awkward Array as an RDataFrame source. This view is generated on demand and the data is not copied. The column

readers are generated based on the run-time type of the views. The readers are passed to a generated source derived from ROOT::RDF::RDataSource.

The ak.from_rdataframe function converts the selected columns as native Awkward Arrays.

We discuss the details of the implementation exploiting JIT techniques. We present examples of analysis of data stored in Awkward Arrays via a high-level interface of an RDataFrame.

We show a few examples of the column definition, applying user-defined filters written in C++, and plotting or extracting the columnar data as Awkward Arrays.

We discuss current limitations and future plans.

Significance:

References:

Experiment context, if any:

CMS

Track 3: Computations in Theoretical Physics: Techniques and Methods / **89**

Quantum-Inspired Machine Learning

Author: Domenico Pomarico[1]

Co-authors: Albino Biafora [2]; Alfredo Zito [3]; Annarita Fanizzi [3]; Daniele La Forgia [3]; Maria Irene Pastena [3]; Nicola Amoroso [4]; Pasquale Tamborra [3]; Raffaella Massafra [3]; Roberto Bellotti [5]; Samantha Bove [3]; Vito Lorusso [3]; Vittorio Didonna [3]

[1] *INFN Sezione di Bari*

[2] *Dipartimento di Economia e Finanza, Università degli Studi di Bari*

[3] *Istituto tumori "Giovanni Paolo II" IRCCS*

[4] *Dipartimento di Farmacia-Scienze del Farmaco, Università degli Studi di Bari*

[5] *Dipartimento di Fisica, Università degli Studi di Bari*

Corresponding Author: domenico.pomarico@ba.infn.it

Learning tasks are implemented via mappings of the sampled data set, including both the classical and the quantum framework. The quantum-inspired approach mimics the support vector machine mapping in a high-dimensional feature space, yielded by the qubit encoding. In our application such scheme is framed in the formulation of a least-squares problem for the minimization of the mean squared error cost function, implemented by means of measurements. The ability of quantum algorithms to manage a high number of parameters will characterize their analysis capability for complex systems, like the targeted biomedical framework.

Significance:

References:

https://www.mdpi.com/2227-7390/9/4/410

Experiment context, if any:

Poster session with coffee break / **91**

Experience in SYCL/oneAPI for event reconstruction at the CMS experiment

Authors: Andrea Bocci[1]; Aurora Perego[2]; Felice Pantaleo[1]; Tony Di Pilato[3]; Wahid Redjeb[4]

[1] *CERN*

[2] *Universita & INFN, Milano-Bicocca (IT)*

[3] *CASUS - Center for Advanced Systems Understanding (DE)*

[4] *Rheinisch Westfaelische Tech. Hoch. (DE)*

Corresponding Author: aurora.perego@cern.ch

The CMS software framework (CMSSW) has been recently extended to perform part of the physics reconstruction with NVIDIA GPUs. To avoid writing a different implementations of the code for each back-end the decision was to use a performance portability library and so Alpaka has been chosen as the solution for Run-3.
In the meantime different studies have been performed to test the track reconstruction and clustering algorithms on different back-ends like CUDA and Alpaka.
With the idea of exploring new solutions, INTEL GPUs have been considered as a new possible back-end and their implementation is currently under development.
This is achieved using SYCL, that is a cross-platform abstraction C++ programming model for heterogeneous computing. It allows developers to reuse code across different hardware and also perform custom tuning for a specific accelerator. The SYCL implementation used is the Data Parallel C++ library (DPC++) in the Intel oneAPI Toolkit.

In this work, we will present the performance of physics reconstruction algorithms on different hardware. Strengths and weaknesses of this heterogeneous programming model will also be presented.

Significance:

Heterogeneous Computing is becoming more and more popular, so is very important to find solutions to avoid code-duplication, like the use of abstraction layers or performance portability libraries.
For this reason is also essential to explore new solutions and one, among the others, is the use of Intel GPUs and the Intel oneAPI Toolkit that provides the portability library DPC++.
This work extends a study performed on other portability libraries and aims at testing these new solutions in order to compare their performance and to be able to converge on the best choice in the end.

References:

Experiment context, if any:

This is one of the studies within the Patatrack project that aim at exploring innovative software and hardware technologies for the CMS experiment at CERN

Poster session with coffee break / 92

Extending ADL/CutLang with a new dynamic multipurpose protocol

Authors: Burak Sen[1]; Gokhan Unel[2]; Sezen Sekmen[3]

[1] *Middle East Technical University*

[2] *University of California Irvine (US)*

[3] *Kyungpook National University (KR)*

Corresponding Author: gokhan.unel@cern.ch

Use of declarative languages for HEP data analysis is an emerging, promising approach. One highly developed example is ADL (Analysis Description Language), an external domain specific language that expresses the analysis physics algorithm in a standard and unambiguous way, independent of frameworks. The most advanced infrastructure that executes an analysis written in the formal ADL syntax is the CutLang (CL) runtime interpreter based on traditional parsing tools. CL which was previously presented in this conference, has been further developed in the last years to cope with most LHC analyses. The new additions include full fledged histogramming and data-MC comparison facilities alongside an interface to a number of well known limit setting tools.

The ADL/CL architecture was thus far prepared and built with a general-purpose programming language, without formal computing expertise and has grown into a complex monolithic structure. To facilitate maintenance and further development of CL, while making it reusable in other (non-scientific) domains, we designed a protocol called Dynamic Domain Specific eXtensible Language (DDSXL) that modularizes its monolithic structure. The DDSXL protocol provides a set of strict rules that allow each researcher to work in their area of expertise and understand the work done without any expertise in other areas, completely independent of the programming languages and frameworks used.
DDSXL integrates a domain ecosystem (such as CL) into the development environment with a completely abstract structure using various OOP design patterns and with a set of rules determined through communication over the network. This protocol also integrates numerous programming languages and frameworks, allowing each developer to integrate it into their own module without the need for expertise in technologies from other modules.

Here, we introduce the latest developments in ADL/CL focusing on the working principles of the DDSXL protocol and integration.

Significance:

ADL/CutLang aims to solve the complexity of HEP analyses by using a domain specific language adapted to collider physics. This presentation focuses on a new protocol proposal which bring a fundamental change to the interpreter infrastructure design. The protocol would convert the interpreter's monolithic structure into a more generic setup where hexagonal architecture design principles are applied. This means independent blocks such as parsers, interpretation engines etc are communicating over the network and can dynamically be replaced or extended. Therefore we expect the resulting protocol to be able to handle not only HEP analyses but other data management tasks as well. A completely out of HEP context example would be the analysis of insurence data to detect frauds.

References:

ADL/CutLang have been published many times, as listed below. However the recent work on the DDSXL protocol proposed for presentation above has not been yet published.
– Project website: cern.ch/adl (includes all references)
Publicatons
– B. Gokturk, A. M. Toon, A. Paul, B. Orgen, N. Ravel, J. Setpal, G. Unel, S. Sekmen, "CutLang V2: towards a unified Analysis Description Language", Frontiers in Science, Big Data, 2021, doi:10.3389/fdata.2021.659986, arXiv:2101.09031
– G. Unel, S. Sekmen and A.M. Toon, "CutLang: a cut-based HEP analysis description language and runtime interpreter," J. Phys. Conf. Ser. 1525 (2020) no.1, 012025 doi:10.1088/1742-6596/1525/1/012025, arXiv:1909.10621.
– S. Sekmen and G. Unel, "CutLang: A Particle Physics Analysis Description Language and Runtime Interpreter," Comput. Phys. Commun. 233 (2018), 215-236, doi:10.1016/j.cpc.2018.06.023, arXiv:1801.05727.
Proceedings: 12 proceedings including the following 2 for ACAT:
– ACAT 2021: "Declarative interfaces for HEP data analysis: FuncADL and ADL/CutLang", 29 Nov - 3 Dec 2021, Daejeon, South Korea
– ACAT 2019: "CutLang analysis description language and runtime interpreter" (poster), 10-15 March 2019, Saas Fe, Switzerland, G. Unel et al 2020 J. Phys.: Conf. Ser. 1525 012025 , https://arxiv.org/abs/1909.10621

Experiment context, if any:

Poster session with coffee break / **93**

JETFLOW: Generating jets with Normalizing Flows using the jet mass as condition and constraint

Authors: Benno Kach[1]; Dirk Krucker[1]; Isabell Melzer-Pellmann[1]; Moritz Scham[1]; Simon Schnake[1]

[1] *Deutsches Elektronen-Synchrotron (DE)*

Corresponding Author: benno.kaech@desy.de

In this study, jets with up to 30 particles are modelled using Normalizing Flows with Rational Quadratic Spline coupling layers. The invariant mass of the jet is a powerful global feature to control whether the flow-generated data contains the same high-level correlations as the training data. The use of normalizing flows without conditioning shows that they lack the expressive power to do this. Using the mass as a condition for the coupling transformation enhances the model's performance on all tracked metrics. In addition, we demonstrate how to sample the original mass distribution with the use of the empirical cumulative distribution function and we
study the usefulness of including an additional mass constraint in the loss term. On the JetNet dataset, our model shows state-of-the-art performance combined with a general model and stable training.

Significance:

Significance: The contribution demonstrates that Normalising Flows with Rational Quadratic Splines can model high-dimensional data efficiently (i.e. stable training and state-of-the-art performance) when global features (mass) are used for conditioning the transformation.

References:

Reference: The study uses the public JetNet dataset: https://zenodo.org/record/4834876 and arXiv:2106.11535

Experiment context, if any:

None

Poster session with coffee break / **94**

XRootD caching for Belle II

Authors: Gunter Quast[1]; Manuel Giffels[1]; Matthias Schnepf[None]; Max Fischer[2]; Moritz David Bauer[None]

[1] *KIT - Karlsruhe Institute of Technology (DE)*

[2] *Karlsruhe Institute of Technology*

Corresponding Author: moritz.bauer@kit.edu

The Belle II experiment at the second generation e+/e- B-factory SuperKEKB has been collecting data since 2019 and aims to accumulate 50 times more data than the first generation experiment, Belle.
To efficiently process these steadily growing datasets of recorded and
simulated data that end up on the order of 100 PB and to support
Grid-based analysis workflows using the DIRAC Workload Management
System, an XRootD-based caching architecture is presented.
The presented mechanism decreases job waiting time for often-used datasets by transparently adding copies of these files at smaller sites without managed storage.
The described architecture seamlessly integrates local storage services and supports the use of dynamic computing resources with minimal deployment effort.

This is especially useful in environments with many institutions providing comparatively small numbers of cores and limited personpower.

This talk will describe the implemented cache at GridKa, a main computing centre for Belle II, as well as its performance and upcoming opportunities for caching for Belle II.

Significance:

This is the first application of the XRootD caching technology for the DIRAC Workload Management System and a novel application of caching for the Belle II experiment. Owing to the unique challenges of Belle II, solutions like these may prove to be essential to analyze the large dataset we aim to collect and increase the efficiency of the resources available to the experiment.

References:

Experiment context, if any:

Belle II

Poster session with coffee break / 95

Implementation of generic SoA data structure in the CMS software

Authors: Andrea Bocci[1]; Eric Cano[1]

[1] *CERN*

Corresponding Author: eric.cano@cern.ch

GPU applications require a structure of array (SoA) layout for the data to achieve good memory access performance. During the development of the CMS Pixel reconstruction for GPUs, the Patatrack developers crafted various techniques to optimise the data placement in memory and its access inside GPU kernels. The work presented here gathers, automates and extends those patterns, and offers a simplified and consistent programming interface.

The work automates the creation of SoA structures, fulfilling technical requirements like cache line alignment, while optionally providing alignment and cache hinting to the compiler and range checking. Protection of read-only products of the CMS software framework (CMSSW) is also ensured with constant versions of the SoA. A compact description of the SoA is provided to minimize the size of data passed to GPU kernels. Finally, the user interface is designed to be as simple as possible, providing an AoS-like semantic allowing compact and readable notation in the code.

The result of porting of CMSSW to SoA will be presented, along with performance measurements.

Significance:

This generic SoA software automatically implements techniques previously used to various degrees in the different modules of the CMS software.

References:

Experiment context, if any:

CMS

Track 1: Computing Technology for Physics Research / 96

Power Efficiency in HEP (x86 vs. arm)

Author: Emanuele Simili[None]

Co-authors: David Britton [1]; Gordon Stewart ; Samuel Cadellin Skipsey

[1] *University of Glasgow (GB)*

Corresponding Author: pepposub@gmail.com

The power consumption of computing is coming under intense scrutiny worldwide, driven both by concerns about the carbon footprint, and by rapidly rising energy costs.
ARM chips, widely used in mobile devices due to their power efficiency, are not currently in widespread use as capacity hardware on the Worldwide LHC Computing Grid.
However, the LHC experiments are increasingly able to compile their workloads on the ARM architecture to take advantage of various HPC facilities (e.g., ATLAS, CMS).

The work described in this paper attempts to compare the energy consumption of various workloads on two almost identical machines, one with an arm64 CPU and the other with a standard AMD x86_64 CPU, operating in identical conditions.
This builds on our initial study of two rather dissimilar machines, located at different UK Universities, which produced some interesting, but at times contradictory, results, showing the need to control the comparison more closely.

The set of benchmarks used include CPU intensive, memory intensive, and I/O bound tasks, ranging from simple scripts, through compiled C programs, to typical HEP workloads (full ATLAS simulations).
We also plan to test the most recent HEPscore containerized jobs, which are actively being developed to match LHC Run3 conditions and can already target different architectures.

The results compare both the power consumption and execution time of the same workload on the two different architectures (arm64 and x86_64).
This will help inform Grid sites whether there are any scenarios where power efficiency can be improved for LHC computing by deploying ARM-based hardware.

Significance:

Power efficiency is an important, often overlooked, feature of computing in High-Energy Physics.
A close comparison in efficiency and throughput of HEP workloads among different architectures provides actual data to guide Grid sites in the choice of new hardware.

This builds and largely improves on our previous study by taking advantage of brand new hardware and an extended set of benchmarks.

References:

https://indico.cern.ch/event/1128343/contributions/4787174/

Experiment context, if any:

LHC, WLCG, ATLAS

Track 1: Computing Technology for Physics Research / 97

Challenges and opportunities integrating LLAMA into AdePT

Author: Bernhard Manfred Gruber[1]

[1] *Technische Universitaet Dresden (DE)*

Corresponding Author: bernhard.manfred.gruber@cern.ch

Particle transport simulations are a cornerstone of high-energy physics (HEP), constituting almost half of the entire computing workload performed in HEP. To boost the simulation throughput and energy efficiency, GPUs as accelerators have been explored in recent years, further driven by the increasing use of GPUs on HPCs. The Accelerated demonstrator of electromagnetic Particle Transport (AdePT) is an advanced prototype for offloading the simulation of electromagnetic showers in Geant4 to GPUs, and still undergoes continuous development and optimization. Improving memory layout and data access is vital to use modern, massively parallel GPU hardware efficiently, contributing to the challenge of migrating traditional CPU based data structures to GPUs in AdePT. The low-level abstraction of memory access (LLAMA) is a C++ library that provides a zero-runtime-overhead data structure abstraction layer, focusing on multidimensional arrays of nested, structured data. It provides a framework for defining and switching custom memory mappings at compile time to define data layouts and instrument data access, making LLAMA an ideal tool to tackle the memory-related optimization challenges in AdePT. Our contribution shares insights gained with LLAMA when instrumenting data access inside AdePT, complementing traditional GPU profiler outputs. We demonstrate traces of read/write counts to data structure elements as well as memory heatmaps. The acquired knowledge allowed for subsequent data layout optimizations.

Significance:

AdePT is central to the current strategy for improving simulation throughput in Geant4. We contribute further optimizations to the project. By coupling these optimizations with LLAMA, a general-purpose library, the demonstrated strategies, insights and optimizations will be transferable to other projects targeting GPUs and heterogeneous systems as well.

References:

AdePT at ACAT2021: https://indico.cern.ch/event/855454/contributions/4605037/
LLAMA paper: https://doi.org/10.1002/spe.3077

Experiment context, if any:

Track 3: Computations in Theoretical Physics: Techniques and Methods / **98**

Accelerating LHC event generation with simplified pilot runs and fast PDFs

Authors: Andy Buckley[1]; Christian Gutschow[2]; Enrico Bothmann[3]; Marek Schoenherr[4]; Max Knobbe[None]; Stefan Hoeche[5]

[1] *University of Glasgow (GB)*

[2] *UCL (UK)*

[3] *University of Göttingen*

[4] *University of Durham*

[5] *Fermilab*

Corresponding Author: chris.g@cern.ch

High-precision calculations are an indispensable ingredient to the success of the LHC physics programme, yet their poor computing efficiency has been a growing cause for concern, threatening to become a paralysing bottleneck in the coming years. We present solutions to eliminate the apprehension by focussing on two major components of general purpose Monte Carlo event generators: The evaluation of parton-distribution functions along with the generation of perturbative matrix elements. We show that for the cost-driving event samples employed by the ATLAS experiment to model omnipresent irreducible Standard Model backgrounds, such as weak boson+jets as well as top-quark-pair production, these components dominate the overall run time by up to 80%. We demonstrate that a reduction of the computing footprint of LHAPDF and SHERPA by factors of

around 50 can be achieved for multi-leg NLO event generation, thereby smashing one of the major milestones set by the HSF event generator working group whilst paving the way towards affordable state-of-the-art event simulation in the HL-LHC era.

Significance:

This presentation covers a new targeted effort enabled by the SWIFT-HEP project, bringing together experimentalists and MC developers to greatly improve the computational efficiency of multi-leg NLO calculations, following a dedicated CPU profiling of these setups - typically the most expensive ones produced by the LHC experiments. The resulting improvements achieve a significant milestone set by the HSF generators working group and will help the experiments stay within the projected budget in the coming years by making high-precision calculations more affordable as we head into the high-luminosity phase of the LHC.

References:

in preparation

Experiment context, if any:

relevant for the LHC experiments, mainly ATLAS and CMS (abstract does not require involvement of the experiments' publication boards)

Track 2: Data Analysis - Algorithms and Tools / **99**

Performance study of the CLUE algorithm with the alpaka library

Authors: Felice Pantaleo[1]; Marco Rovere[1]; Tony Di Pilato[2]; Wahid Redjeb[3]

[1] *CERN*

[2] *CASUS - Center for Advanced Systems Understanding (DE)*

[3] *Rheinisch Westfaelische Tech. Hoch. (DE)*

Corresponding Author: tony.dipilato@cern.ch

CLUE (CLUsters of Energy) is a fast, fully-parallelizable clustering algorithm developed to optimize such a crucial step in the event reconstruction chain of future high granularity calorimeters. The main drawback of having an unprecedentedly high segmentation in this kind of detectors is a huge computation load that, in case of the CMS, must be reduced to fit the harsh requirements of the Phase-2 High Level Trigger.

With the adoption of alpaka as performance portability library in CMSSW, the CLUE algorithm has been tested on multiple accelerators and hybrid platforms. This work presents the latest results obtained with the alpaka implementation of CLUE, which can fully exploit the available hardware on each machine and fulfill the task with high performance.

Significance:

This talk will show how the alpaka performance portability library, the software technology chosen by CMS for hardware accelerators, can mantain the high performance for novel algorithms, developed with throughput and efficiency in mind for the new generation of detectors and Phase-2 upgrades at the experiment.

References:

https://doi.org/10.3389/fdata.2020.591315

Experiment context, if any:

CMS

Poster session with coffee break / 100

Supporting multiple hardware architectures at CMS: the integration and validation of Power9

Authors: Christoph Wissing[1]; Daniele Spiga[2]

Co-authors: Alan Malta Rodrigues [3]; Antonio Perez-Calero Yzquierdo [4]; Dirk Hufnagel [5]; Hasan Ozturk [6]; Jordan Martins [7]; Kirill Skovpen [8]; Marco Mascheroni [9]; Saqib Haleem [10]; Todor Trendafilov Ivanov [11]; Tommaso Boccali [12]

[1] *Deutsches Elektronen-Synchrotron (DE)*

[2] *Universita e INFN, Perugia (IT)*

[3] *University of Notre Dame (US)*

[4] *Centro de Investigaciones Energéticas Medioambientales y Tecnológicas*

[5] *Fermi National Accelerator Lab. (US)*

[6] *CERN*

[7] *Universidade do Estado do Rio de Janeiro (BR)*

[8] *Ghent University (BE)*

[9] *Univ. of California San Diego (US)*

[10] *National Centre for Physics (PK)*

[11] *University of Sofia - St. Kliment Ohridski (BG)*

[12] *INFN Sezione di Pisa*

Corresponding Author: daniele.spiga@cern.ch

Computing resources in the Worldwide LHC Computing Grid (WLCG) have been based entirely on the x86 architecture for more than two decades. In the near future, however, heterogeneous non-x86 resources, such as ARM, POWER and Risc-V, will become a substantial fraction of the resources that will be provided to the LHC experiments, due to their presence in existing and planned world-class HPC installations. The CMS experiment, one of the four large detectors at the LHC, has started to prepare for this situation, with the CMS software stack (CMSSW) already compiled for multiple architectures. In order to allow for a production use, the tools for workload management and job distribution need to be extended to be able to exploit heterogeneous architectures.

Profiting from the opportunity to exploit the first sizable IBM Power9 allocation available on Marconi100 HPC system at CINECA, CMS developed all the needed modifications to the CMS workload management system. After a successful proof of concept, a full physics validation has been performed in order to bring the system in production. The experiences are of very high value, when it comes to commissioning of the similar (even larger) Summit HPC system at Oak Ridge, where CMS is also expecting a resource allocation. Moreover the compute power of those systems is being provided also via GPUs and this represents an extremely valuable opportunity to exploit the offloading capability already implemented in CMSSW.

The status of the current integration including the exploitation of the GPUs, the results of the validation as well as the future plans will be shown and discussed.

Significance:

The presentation shows how CMS experiment is preparing to transparently integrate at large scale heterogeneous non-x86 resources, including the strategy for physics validation

References:

Experiment context, if any:

CMS experiment

Poster session with coffee break / **101**

Updates on the Low-Level Abstraction of Memory Access

Author: Bernhard Manfred Gruber[1]

[1] *Technische Universitaet Dresden (DE)*

Corresponding Author: bernhard.manfred.gruber@cern.ch

Choosing the best memory layout for each hardware architecture is increasingly important as more and more programs become memory bound. For portable codes that run across heterogeneous hardware architectures, the choice of the memory layout for data structures is ideally decoupled from the rest of a program.
The low-level abstraction of memory access (LLAMA) is a C++ library that provides a zero-runtime-overhead abstraction layer, underneath which memory layouts can be freely exchanged, focusing on multidimensional arrays of nested, structured data.
It provides a framework for defining and switching custom memory mappings at compile time to define data layouts, data access and access instrumentation, making LLAMA an ideal tool to tackle memory-related optimization challenges in heterogeneous computing.
After its scientific debut, several improvements and extensions have been added to LLAMA. This includes compile-time array extents for zero memory overhead, support for computations during memory access, new mappings (e.g. int/float bit-packing or byte-swapping) and more. This contribution provides an overview of the LLAMA library, its recent development and an outlook of future activities.

Significance:

LLAMA provides a general C++ library solution for memory layout and memory access abstractions which are crucial to exploit the heterogeneous landscape of modern hardware architectures, including CPUs, GPUs and FPGAs. LLAMA is an orthogonal framework to compute kernel abstractions such as alpaka and Kokkos and of similar importance.

References:

Journal paper: https://doi.org/10.1002/spe.3077

Experiment context, if any:

Poster session with coffee break / **102**

Distributed data processing pipelines in ALFA

Author: Alexey Rybalchenko[1]

Co-authors: Dennis Klein [1]; Mohammad Al-Turany [2]

[1] *GSI - Helmholtzzentrum fur Schwerionenforschung GmbH (DE)*

[2] *CERN*

Corresponding Author: alexey.rybalchenko@cern.ch

The common ALICE-FAIR software framework ALFA offers a platform for simulation, reconstruction and analysis of particle physics experiments. FairMQ is a module of ALFA that provides building blocks for distributed data processing pipelines, composed out of components communicating via

message passing. FairMQ integrates and efficiently utilizes standard industry data transport technologies, while hiding the transport details behind an abstract interface. In this work we present the latest developments in FairMQ, focusing on the new and improved features of the transport layer, primarily the shared memory transport and the generic interface features. Furthermore, we present the new control and configuration facilities, that allow programmatically controlling a group of FairMQ components. Additionally, new debugging and monitoring tools are highlighted. Finally, we outline how these tools are used by the ALICE experiment.

Significance:

References:

Experiment context, if any:

ALICE, FAIR

Poster session with coffee break / 103

The CMS Roadmap towards HL-LHC Software and Computing

Authors: Danilo Piparo[1]; James Robert Letts[2]

[1] *CERN*

[2] *Univ. of California San Diego (US)*

Corresponding Author: danilo.piparo@cern.ch

The Phase-2 upgrade of CMS, coupled with the projected performance of the HL-LHC, shows great promise in terms of discovery potential. However, the increased granularity of the CMS detector and the higher complexity of the collision events generated by the accelerator pose challenges in the areas of data acquisition, processing, simulation, and analysis. These challenges cannot be solved solely by increments in the computing resources available to CMS, but must be accompanied by major improvements of the computing model and computing software tools, as well as data processing software and common software tools. We present aspects of our roadmap for those improvements, focusing on the plans to reduce storage and CPU needs as well as take advantage of heterogeneous platforms, such as the ones equipped with GPUs, and High Performance Computing Centers. We describe the most prominent research and development activities being carried out in the experiment, demonstrating their potential effectiveness in either mitigating risks or quantitatively reducing computing resource needs on the road to the HL-LHC.

Significance:

This presentation would be based on the documentation submitted to the LHCC for the November 2021 review of HL-LHC computing models, referenced below.

References:

https://cds.cern.ch/record/2815292?ln=en

Experiment context, if any:

Submitted on behalf of the CMS Collaboration. Abstract has been approved by the CMS Conference Committee.

Poster session with coffee break / 104

The Level-1 Global Trigger for Phase-2: Algorithms, configuration and integration in the CMS offline framework

Authors: Benjamin Huber[1]; Dinyar Rabady[2]; Elias Leutgeb[1]; Gabriele Bortolato[3]; Hannes Sakulin[2]

[1] *Technische Universitaet Wien (AT)*

[2] *CERN*

[3] *Universita e INFN, Padova (IT)*

Corresponding Author: benjamin.huber@cern.ch

The CMS Level-1 Trigger, for its operation during Phase-2 of LHC, will undergo a significant upgrade and redesign. The new trigger system, based on multiple families of custom boards, equipped with Xilinx Ultrascale Plus FPGAs and interconnected with high speed optical links at 25 Gb/s, will exploit more detailed information from the detector subsystems (calorimeter, muon systems, tracker). In contrast to its implementation during Phase-1, information from the CMS tracker is now also available at the Level-1 Trigger and can be used for particle flow algorithms. The final stage of the Level-1 Trigger, called Global Trigger (GT), will receive more than 20 different trigger object collections from upstream systems and will be able to evaluate a menu of more than 1000 cut-based algorithms distributed over 12 boards. These algorithms may not only apply conditions on parameters such as momentum or angle of a particle, but can also do arithmetic calculations, like the invariant mass of a suspected mother particle of interest or the angle between two particles. The Global Trigger is designed as a modular system, with an easily re-configurable algorithm unit, to meet the demand of high flexibility required for shifting trigger strategies during Phase-2 operation of the LHC. The algorithms themselves are kept highly configurable and tools are provided to allow their study from within the CMS offline software framework (CMSSW) without the need for knowledge of the underlying firmware implementation. To allow the reproducible translation of the physicist-designed trigger menu to VHDL for use in the hardware trigger, a tool has been developed that converts the Python-based configuration used by CMSSW to VHDL. In addition to cut-based algorithms, neural net algorithms are being developed and integrated into the Global Trigger framework. To make use of these algorithms in hardware, the HLS4ML framework is used, which transpiles pre-trained neural nets, generated in the most commonly used software frameworks, into firmware code. A prototype firmware for a single Global Trigger board has been developed, which includes the de-multiplexing logic, conversion to an internal common object format and distribution of the data over all Super Logic Regions. In this framework 312 algorithms are implemented at a clock speed of 480MHz. The prototype has been thoroughly tested and verified with the bit-wise compatible C++ emulator. In this contribution we present the Phase-2 Global Trigger with an emphasis on the Global Trigger algorithms, their implementation in hardware, configuration with Python and the novel integration within the CMS offline software framework (CMSSW).

Significance:

n this contribution we present the Phase 2 Global Trigger with an emphasis on the Global Trigger algorithms, their implementation in hardware, configuration with Python and the novel integration within the CMS offline software framework (CMSSW).

References:

Experiment context, if any:

Phase 2 upgrade, Level 1 Trigger, CMS

Poster session with coffee break / 105

CERNLIB status

Authors: Andrii Verbytskyi[1]; Dirk Duellmann[2]; Frank Berghaus[3]; Gerardo Ganis[2]; Marcello Maggi[4]; Matthias Schroeder[2]; Ulrich Schwickerath[2]

[1] *Max Planck Society (DE)*

[2] *CERN*

[3] *Argonne National Laboratory (US)*

[4] *Universita e INFN, Bari (IT)*

Corresponding Author: andrii.verbytskyi@cern.ch

We present a revived version of the CERNLIB, the basis for software ecosystems of most of the pre-LHC HEP experiments. The efforts to consolidate the CERNLIB are part of the activities of the Data Preservation for High Energy Physics collaboration to preserve data and software of the past HEP experiments.

The presented version is based on the CERNLIB version 2006 with numerous patches made for the compatibility with modern compilers and operating systems. The code is available publicly in the CERN GitLab repository with all the development history starting from the early 1990s. The updates also include a re-implementation of the build system in cmake to make CERNLIB compliant with the current best practices and to increase the chances of preserving the code in a compilable state for the decades to come.

The revived CERNLIB project also includes an updated documentation, which we believe is a cornerstone for any preserved software depending on it.

Significance:

CERNLIB has top importance for the Data Preservation in High Energy Physics as it is the basis software for the most of the pre-LHC HEP experiments.

The revival of the CERNLIB after more than 15 years of absence of maintenance is an example of the scientific software preservation and a source of lessons to learn for the benefits of ongoing software development and related physics experiments.

References:

Experiment context, if any:

Data Preservation for High Energy Physics (DPHEP), ALEPH, OPAL, L3, DELPHI, JADE, H1

Poster session with coffee break / **106**

Variational AutoEncoders for Anomaly Detection in VBS events within an EFT framework

Authors: Giacomo Boldrini[1]; Giulia Lavizzari[None]; Pietro Govoni[1]; Simone Gennai[1]

[1] *Universita & INFN, Milano-Bicocca (IT)*

Corresponding Author: giulia.lavizzari@cern.ch

We present a machine-learning based method to detect deviations from a reference model, in an almost independent way with respect to the theory assumed to describe the new physics responsible for the discrepancies.

The analysis is based on an Effective Field Theory (EFT) approach: under this hypothesis the Lagrangian of the system can be written as an infinite expansion of terms, where the first ones are those from the Standard Model (SM) Lagrangian and the following terms are higher dimension operators. The presence of the EFT operators impacts the distributions of the observables by producing deviations from the shapes expected when the SM Lagrangian alone is considered .

We use a Variational AutoEncoder (VAE) trained on SM processes to identify EFT contributions as anomalies. While SM events are expected to be reconstructed properly, events generated taking into account EFT contributions are expected to be poorly reconstructed, thus accumulating in the tails of the loss function distribution. Since the training of the model does not depend on any specific new physics signature, the proposed strategy does not make specific assumptions on its nature. In order to improve the discrimination performances, we introduced a DNN classifier that distinguishes between EFT and SM events based on the values of the reconstruction and regularization losses of the model. In this second model a cross entropy term is added to the usual loss of the VAE, optimizing at the same time the reconstruction of the input variables and the classification. This procedure ensures that the model is optimized for discrimination, with a small price in terms of model independency due to the use of one of the 15 operators from the EFT model in the training.

In this talk we will discuss in detail the above-mentioned methods using generator level VBS events produced at LHC and assuming, in order to compute the significance of possible new physics contributions, an integrated luminosity of $350 fb^{-1}$.

Significance:

For the first time an Anomaly Detection strategy is applied to VBS events within an EFT framework: this new approach could deeply improve the strategy employed to address those kinds of analyses. Such an algorithm would deliver a list of anomalous events for further analysis and the recurring event topologies in this dataset could inspire novel new-physics models and new experimental searches.

References:

https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.120.081801
https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.123.161801

Experiment context, if any:

LHC

Poster session with coffee break / **107**

HDTFS□Cost-effective Hadoop Distributed & Tiered File System for High Energy Physics

Authors: Xiaoyu Liu[1]; Libin Xia[1]; Xiaowei Jiang[2]; Gongxing Sun[1]

[1] *IHEP*

[2] *IHEP□□□□□□□□□□□□□□□*

Corresponding Author: liuxiaoyu@ihep.ac.cn

With the scale and complexity of High Energy Physics(HEP) experiments increase, researchers are facing the challenge of large-scale data processing. In terms of storage, HDFS, a distributed file system that supports the “data-centric” processing model, has been widely used in academia and industry. This file system can support Spark and other distributed data localization calculations, researching the application of Hadoop Distributed File System(HDFS) in the field of HEP is the basis for ensuring the application of upper-layer computing in this field. However, HDFS expand the cluster capacity by adding cluster nodes, this way cannot meet the high cost-effective system requirements for the persistence and backup process of massive HEP experimental data. In response to the above problems, researching Hadoop Distributed & Tiered File System(HDTFS) that supports disk-tape storage, taking full advantage of the fast disk access speed and the advantages of large tape storage capacity, low price, and long storage period, to solve the high cost of horizontal expansion of HDFS clusters. The system provides users with a single global namespace, and avoids dependence on external metadata servers to access the data stored on tape. In addition, tape layer resources are managed internally so that users do not have to deal with complex tape storage. The experimental results show that this method can effectively solve the massive data storage of HEP Hadoop cluster.

Significance:

References:

Experiment context, if any:

Poster session with coffee break / 108

Application of Unity for detector modeling in BESIII

Authors: Kaixuan Huang[1]; Shengsen Sun[2]; Yumei Zhang[1]; Zhengyun You[1]; Zhijun Li[3]

[1] *Sun Yat-Sen University(CN)*

[2] *Chinese Academy of Sciences(CN)*

[3] *Sun Yat-Sen University (CN)*

Corresponding Author: zhijun.li@cern.ch

Detector modeling and visualization are essential in the life cycle of a High Energy Physics (HEP) experiment. Unity is a professional multi-media creation software that has the advantages of rich visualization effects and easy deployment on various platforms. In this work, we applied the method of detector transformation to convert the BESIII detector description from the offline software framework into the 3D detector modeling in Unity. By matching the geometric units with detector identifiers, the new event display system based on Unity can be developed for BESIII. The potential for further application development into virtual reality will also be introduced.

Significance:

References:

Experiment context, if any:

Track 3: Computations in Theoretical Physics: Techniques and Methods / 109

Speeding up Madgraph5_aMC@NLO through CPU vectorization and GPU offloading: towards a first alpha release

Authors: Andrea Valassi[1]; Carl Vuosalo[2]; David Smith[1]; Laurence Field[1]; Nathan Nichols[3]; Olivier Mattelaer[4]; Stefan Roiser[1]; Stephan Hageboeck[1]; Taylor Childers[3]; Walter Hopkins[3]

[1] *CERN*

[2] *University of Wisconsin Madison (US)*

[3] *Argonne National Laboratory (US)*

[4] *UCLouvain*

Corresponding Author: andrea.valassi@cern.ch

The matrix element (ME) calculation in any Monte Carlo physics event generator is an ideal fit for implementing data parallelism with lockstep processing on GPUs and on CPU vector registers. For complex physics processes where the ME calculation is the computational bottleneck of event generation workflows, this can lead to very large overall speedups by efficiently exploiting these hardware architectures, which are now largely underutilized in HEP. In this contribution, we will present the latest status of our work on the reengineering of the Madgraph5_aMC@NLO event

generator for these architectures. The new implementations of the ME calculation in vectorized C++, in CUDA and in the ALPAKA, KOKKOS and SYCL portability frameworks will be described in detail, as well as their integration into the existing MadEvent framework to keep the same overall look-and-feel of the user interface. Performance numbers will be reported both for the ME calculation alone and for the overall production workflow for unweighted event generation. First experience with an alpha release of the software supporting LHC LO processes, which is expected by the time of the ACAT2022 conference, will also be discussed.

Significance:

- We plan to present the first functional release of the software usable by LHC experiments (or at least a clear timeline towards that).
- This contribution is relevant to both ACAT track1 and track3. It is relevant to track1 because it discusses approaches to exploiting heterogeneous resources which may be reused also by other HEP workloads, such as GPU/CPU data parallelism through compiler vector extensions, AOSOAs, portability frameworks and various threading implementations. It is relevant to track3 because we believe that similar large speedups on GPUs and vector CPUs are within reach also for any other MC matrix element event generator. For these reasons we kindly suggest to the organizers to also consider it for a plenary talk, which would cover the topic of speeding up Monte Carlo event generators also in more generic terms.

References:

- HSFWS2020: https://indico.cern.ch/event/941278/contributions/4101793/
- vCHEP2021: https://doi.org/10.1051/epjconf/202125103045
- ICHEP2022: https://agenda.infn.it/event/28874/abstracts/20368/

Experiment context, if any:

Madgraph5_aMC@NLO is routinely used, amongst others, by ATLAS and CMS

Track 2: Data Analysis - Algorithms and Tools / 110

Efficient search for new physics using Active Learning in the ATLAS Experiment

Authors: Irina Espejo Morales[1]; Janik Von Ahnen[2]; Kyle Stuart Cranmer[3]; Lukas Alexander Heinrich[4]; Patrick Rieck[3]; Philipp Gadow[2]; Zubair Bhatti[3]

[1] *New York University*

[2] *Deutsches Elektronen-Synchrotron (DE)*

[3] *New York University (US)*

[4] *CERN*

Corresponding Author: patrick.rieck@cern.ch

Searches for new physics set exclusion limits in parameter spaces of typically up to 2 dimensions. However, the relevant theory parameter space is usually of a higher dimension but only a subspace is covered due to the computing time requirements of signal process simulations. An Active Learning approach is presented to address this limitation. Compared to the usual grid sampling, it reduces the number of parameter space points for which exclusion limits need to be determined. Hence it allows to extend interpretations of searches to higher dimensional parameter spaces and therefore to raise their value, e.g. via the identification of barely excluded subspaces which motivate dedicated new searches.

In an iterative procedure, a Gaussian Process is fit to excluded signal cross-sections. Within the region close to the exclusion contour predicted by the Gaussian Process, Poisson disc sampling is used to determine further parameter space points for
which the cross-section limits are determined. The procedure is aided by a warm-start phase based on computationally inexpensive, approximate limit estimates such as total signal cross-sections. A python package, excursion 1, provides the Gaussian Process routine. The procedure is applied to a Dark Matter search performed by the ATLAS experiment, extending its interpretation from a 2 to a 4-dimensional parameter space while keeping the computational effort at a low level.

1 https://github.com/diana-hep/excursion

Significance:

Follow-up on ACAT 2019 contribution 479, now applying Active Learning to a full-scale ATLAS physics analysis

References:

https://indico.cern.ch/event/708041/contributions/3269754/

Experiment context, if any:

ATLAS Experiment

Track 2: Data Analysis - Algorithms and Tools / 111

Full Quantum GAN Model for High Energy Physics Simulations

Author: Florian Rehm[1]

Co-authors: Alexis Harilaos Verney Provatas [2]; Dirk Krucker [2]; Kerstin Borras [3]; Michele Grossi [4]; Simon Schnake [3]; Sofia Vallecorsa [4]

[1] *CERN / RWTH Aachen University*

[2] *DESY*

[3] *DESY / RWTH Aachen University*

[4] *CERN*

Corresponding Author: florian.matthias.rehm@cern.ch

The prospect of possibly exponential speed-up of quantum computing compared to classical computing marks it as a promising method when searching for alternative future High Energy Physics (HEP) simulation approaches. HEP simulations like at the LHC at CERN are extraordinarily complex and, therefore, require immense amounts of computing hardware resources and computing time. For some HEP simulations classical machine learning models are already successfully tested leading to speed-ups in the order of magnitudes. In this research we proceed to the next step and test if quantum computing can further improve HEP machine learning simulations.

With a small prototype model we showcase a full quantum Generative Adversarial Network (GAN) model for successfully generating real calorimeter shower images with high precision. The advantage compared to previous other quantum models is, that with employing angle encoding the pixel to qubit ratio scales linear and the model generates real images with pixel energy values instead of simple probability distributions. The model is constructed and evaluated for images with eight pixels and requires only eight qubits for the generator and discriminator quantum circuit. The quantum circuits make use of the properties of entanglement and superposition to learn and reproduce the correlations in the images.

To complete the picture, the results of the full quantum GAN model are compared to other quantum and hybrid quantum-classical models.

Significance:

Exploring future ideas of employing quantum computing to a HEP use case. We are the first ones who test a full quantum generative adversarial network (qGAN) model for generating calorimeter detector images. Furthermore, we employ a different encoding strategy than in a previous research which allows us to generate real images with pixel energies instead of only probability distributions. With quantum computing we hope to combat computing hardware restraints for future HEP experiments and start, therefore, already with initial tests. Lastly, we compare our results to other state of the art quantum models.

References:

We were working already on a hybrid qGAN which, however, only was able to generate probability distributions in those two papers:
- http://ceur-ws.org/Vol-3041/363-368-paper-67.pdf
- https://arxiv.org/abs/2203.01007

The new model is full quantum (compared to the previous hybrid model) and uses another encoding allowing to generate real images.

Experiment context, if any:

Study is carried out general for HEP detectors at CERN, no specific experiment.

Poster session with coffee break / **113**

ML-based discrimination of same sign WW VBS processes at CMS with hadronic tau in final state

Author: Tommaso Tedeschi[1]

[1] *Universita e INFN, Perugia (IT)*

Corresponding Author: tommaso.tedeschi@cern.ch

Vector Boson scattering (VBS) processes are a significant testing ground for the Standard Model and especially for the Higgs sector since their longitudinal component is sensitive to the couplings between Higgs and Vector Bosons. Here an analysis with CMS detector targeting same-sign WW channel is performed considering the final state with a light lepton and a hadronic tau lepton, which could be an important probe of BSM processes thanks to its large mass. In order to maximize the discrimination of such events, Machine Learning applications have been put in place: BDT/DNN-based discriminators are trained and optimized to recognize SM and BSM signals (considering an EFT approach) using kinematic information of the selected object. In this contribution, we will present the details of this ML-based approach and the results achieved, comparing them with the ones obtained with a non-ML-based approach.

References:

Experiment context, if any:

CMS experiment

Significance:

This presentation covers the application of ML algorithms, exploiting modern techniques, in the analysis of Vector Boson Scattering processes with a final state that has never been studied before.

Poster session with coffee break / **114**

Improving robustness of jet tagging algorithms with adversarial training

Author: Annika Stein[1]

[1] *Rheinisch Westfaelische Tech. Hoch. (DE)*

Corresponding Author: annika.stein@cern.ch

In the field of high-energy physics, deep learning algorithms continue to gain in relevance and provide performance improvements over traditional methods, for example when identifying rare signals or finding complex patterns. From an analyst's perspective, obtaining highest possible performance is desirable, but recently, some focus has been laid on studying robustness of models to investigate how well these perform under slight distortions of input features. Especially for tasks that involve many (low-level) inputs, the application of deep neural networks brings new challenges. In the context of jet flavor tagging, adversarial attacks are used to probe a typical classifier‘s vulnerability and can be understood as a model for systematic uncertainties. A corresponding defense strategy, adversarial training, improves robustness, while maintaining high performance. This contribution presents different approaches using a set of attacks with varying complexity. Investigating the loss surface corresponding to the inputs and models in question reveals geometric interpretations of robustness, taking correlations into account. Additional cross-checks against other, physics-inspired mismodeling scenarios are performed and give rise to the presumption that adversarially trained models can cope better with simulation artifacts or subtle detector effects.

Significance:

Such studies are crucial to understand if potential mismodelings in simulation could lead to differences in performance in data compared to simulation. Sophisticated calibration techniques are applied which at times might still leave residual disagreement. Therefore, any technique that evades that problem during training and that probes the trade-off between performance and robustness is of importance for identification of physics objects with deep learning algorithms, especially with large numbers of (low-level) input features. In this contribution, a successful application of defense strategies for deep-learned flavor tagging algorithms is shown and is accompanied by novel insights into the neural network‘s properties that help explaining the observed behavior.

References:

https://arxiv.org/abs/2203.13890

Experiment context, if any:

Context: ATLAS, CMS

Track 1: Computing Technology for Physics Research / **115**

Precision Cascade: A novel algorithm for multi-precision extreme compression

Authors: Gene Van Buren[None]; Jerome LAURET[1]; Juan Gonzalez[2]; Philippe Canal[3]; Yueyang Ying[4]

[1] *Brookhaven National Laboratory*

[2] *Accelogic, Inc.*

[3] *Fermi National Accelerator Lab. (US)*

[4] *Massachusetts Inst. of Technology (US)*

Corresponding Author: kying@mit.edu

Lossy compression algorithms are incredibly useful due to powerful compression results. However, lossy compression has historically presented a trade-off between the retained precision and

the resulting size of data compressed with a lossy algorithm. Previously, we introduced BLAST, a state-of-the-art compression algorithm developed by Accelogic. We presented results that demonstrated BLAST can achieve a compression factor that undeniably surpasses compression algorithms currently available in the ROOT framework. However, the leading concern of utilizing the lossy compression technique is the delayed realization that more precision is necessary. This precision may have been irretrievably lost in an effort to decrease storage size. Thus, there is immense value in retaining higher precision data in reserve. Though, in the era of exabyte computing, it becomes extremely inefficient and costly to duplicate data stored at different compressive precision values. A tiered cascade of stored precision optimizes data storage and resolves these fundamental concerns.

Accelogic has developed a game-changing compression technique, known as "Precision Cascade", which enables higher precision to be stored separately without duplicating information. With this novel method, varying levels of precision can be retrieved, potentially minimizing live storage space. Preliminary results from STAR and CMS demonstrate that multiple layers of precision can be stored and retrieved without significant penalty to the compression ratios and (de)compression speeds, when compared to the single-precision BLAST baseline.

In this contribution, we will present the integration of Accelogic's "Precision Cascade" into the ROOT framework, with the principal purpose of enabling high-energy physics experiments to leverage this state-of-the-art algorithm with minimal friction. We also present our progress in exploring storage reduction and speed performance with this new compression tool in realistic examples from both STAR and CMS experiments and feel we are ready to deliver the compression algorithm to the wider community.

Significance:

Lossy compression algorithms are incredibly useful due to powerful compression results and represents the next evolution in the HEP/NP IO workflows for extreme space saving. Precision cascade is a novel technique that targets the primary concern of lossy compression (that too much precision is lost), by enabling higher precision to be stored and retrieved later on, without duplicating data.

Track 1 encompassing Computing Technology for Physics Research and architecture seem suited for our abstract.

References:

G. V. Buren, J. Lauret, J. Gonzales. R. Nunez, P. Canal, A. Naumann – "Extreme compression for Large Scale Data store" – CHEP 2019 proceedings, EPJ Web of Conferences 245, 06024 (2020), doi 10.1051/epjconf/202024506024

P. Canal, G.V. Buren, J. Lauret, I.A. Cali. J. Gonzales, , P. Canal, R. Nunez, Y. Ying – "ROOT Files Improved with Extreme Compression", ACAT 2021 proceedings, accepted for publication.

Experiment context, if any:

STAR, CMS and the ROOT team

Track 1: Computing Technology for Physics Research / 116

GPU acceleration of Monte Carlo simulations: particle physics methods applied to medicine

Authors: Marco Barbone[None]; Rafael Brandt[1]

Co-authors: Alexander Howard [2]; Mihaly Novak [3]; Alex Tapper [4]; Wayne Luk ; Georgi Gaydadjiev [1]

[1] *University of Groningen*

[2] *Imperial College (GB)*

[3] *CERN*

[4] *Imperial College London*

Corresponding Author: m.barbone19@imperial.ac.uk

GPU acceleration has been successfully utilised in particle physics for real time analysis and simulation, in this study, we investigate the potential benefits for medical physics applications by analysing performance, development effort, and availability. We selected a software developer with no high performance computing experience to parallelise and accelerate a stand-alone Monte Carlo simulation consisting of electron single coulomb scattering. Such simulations contribute to real-time dose estimation for real-time adaptive radiotherapy, a new and emerging cancer treatment that heavily relies on high performance computing. As a proof of principle, we implement a single scattering process of electrons in a homogeneous material with pencil beam at constant initial energy. We compared performance gain offered by GPU acceleration against an optimised CPU implementation and evaluated it by computing 100M histories of a 128 keV electron interacting in water. We also evaluated 1B histories to measure the scalability. The results show that when comparing the multicore CPU implementation running with 24 cores, a speedup of 808x (100M) and 1727x (1B), which corresponds to a 320x and 648x cost-equivalent speedup. The results on both architectures were statistically equivalent.The successful implementation and measured acceleration combined with the low level of expertise needed for obtaining such speedup is a promising first step for the use of GPU acceleration in a context such as real-time adaptive radiotherapy where there are strict performance and time requirements.

Significance:

This study analyzes the possibility of applying HEP methods for GPU acceleration of Monte Carlo simulation to the medical context. The results show that it is possible to achieve multiple orders of magnitude speedup compared to equivalent multicore implementations

References:

Experiment context, if any:

Poster session with coffee break / **117**

Analysis of spectroscopy data with Machine Learning

Authors: Aamod Kulkarni[1]; Jan Ebert[2]; Stefan Kesselheim[2]; Sören Möller[2]

[1] *Technische Universität Dresden*

[2] *Forschungszentrum Jülich*

Corresponding Author: s.kesselheim@fz-juelich.de

Different spectroscopy techniques lead to data with a similar structure: One-dimensional frequency resolved measurements with characteristic features such as peaks or shoulders. In many cases, classical fitting methods are sufficient for the analysis of such data, but it is desirable to have a robust alternative based on Machine Learning (ML). ML models can, for example, learn that only a part of the measured signal is relevant for the scientist, and another part is to be ignored ("background"). In this work we systematically study how to generate such ML models. We formulate a simplified data generation procedure where spectra with peak-like features are generated on the fly during the training process. With access to unlimited synthetic data, we can systematically study the role of dataset size, model architecture and capacity, different losses and all relevant hyperparameters. With this approach, we study if transformer-based models are able to capture imposed correlations in peak positions better than convolutional models, as their attention mechanism should have the inductive bias to also capture long-range correlation. We apply the approach to Nuclear Reaction Analysis (NRA), a spectroscopy method based on MeV ions. With our contribution, we would like to open the discussion with domain experts about what other classes of data analysis can be facilitated with such ML techniques.

Significance:

We want to make our novel approach public and are interested in getting feedback from scientists from different fields in order to refine our ML approach.

References:

Experiment context, if any:

Track 3: Computations in Theoretical Physics: Techniques and Methods / 118

A differentiable simulation approach for Solar Power Plants

Authors: Max Pargmann[1]; Jan Ebert[2]; Daniel Maldonado Quinto[1]; Stefan Kesselheim[2]

[1] *German Aerospace Center*

[2] *Forschungszentrum Jülich*

Corresponding Author: s.kesselheim@fz-juelich.de

In Solar Power Plants, temperatures sufficient for chemical processes or the generation of electrical power are created by reflecting sunlight with thousands of mirrors ("heliostats") to a surface ("the receiver"). In operation, the temperature distribution on the receiver is critical for the performance and must be optimized. The heliostats are never perfectly flat as due to budget constraints, the construction is not optimal. We have devised a method to infer the heliostat surface from the reflection of the sun. The technique is based on an implementation of a simulation in PyTorch, where the automatic differentiation engine is used to optimize the surface. The surface is modeled as by a Non-Uniform Rational B-Spline (NURBS) and the NURBS parameters are subject to optimization. Furthermore we employ a regularization technique to mitigate the appearing challenge of ambiguous solutions. Our approach makes efficient use of GPUs based on PyTorch's linear algebra engine. We believe our approach poses an interesting example of a fruitful interaction of techniques originating from Machine Learning and simulation.

Significance:

This is a novel approach where we bring together tools from Machine Learning with "classical" simulation. We hope to spark discussion where similar approaches could be useful.

References:

Experiment context, if any:

Poster session with coffee break / 119

Automatic differentiation of binned likelihoods with RooFit and Clad

Authors: Garima Singh[1]; Jonas Rembser[2]; Lorenzo Moneta[2]; Vassil Vasilev[1]

[1] *Princeton University (US)*

[2] *CERN*

Corresponding Author: garima.singh@cern.ch

RooFit is a toolkit for statistical modeling and fitting used by most experiments in particle physics. Just as data sets from next-generation experiments grow, processing requirements for physics analysis become more computationally demanding, necessitating performance optimizations for RooFit. One possibility to speed-up minimization and add stability is the use of automatic differentiation

(AD). Unlike for numerical differentiation, the computation cost scales linearly with the number of parameters, making AD particularly appealing for statistical models with many parameters. In this talk, we report on one possible way to implement AD in RooFit. Our approach is to add a facility to generate C++ code for a full RooFit model automatically. Unlike the original RooFit model, this generated code is free of virtual function calls and other RooFit-specific overhead. In particular, this code is then used to produce the gradient automatically with Clad. Clad is a source transformation AD tool implemented as a plugin to the clang compiler, which automatically generates the derivative code for input C++ functions. We show results demonstrating the improvements observed when applying this code generation strategy to HistFactory and other commonly used RooFit models. HistFactory is the subcomponent of RooFit that implements binned likelihood models with probability densities based on histogram templates. These models frequently have a very large number of free parameters, and are thus an interesting first target for AD support in RooFit.

References:

Experiment context, if any:

Significance:

This contribution will demonstrate significant advancements in state-of-the-art capabilities for both High-Energy Physics (HEP) and computer science domains. The widespread use of automatic differentiation of statistical models in HEP will significantly improve the performance and numeric stability of statistical analysis. On the computer science side, this work demonstrates that source-transformation based automatic differentiation can be added to complex libraries like RooFit. It also showcases an application where different AD strategies can be compared, as other research groups are experimenting with other AD implementations for differentiable likelihoods (usually the ones available in the Python ecosystem, such as TensorFlow).

Track 2: Data Analysis - Algorithms and Tools / **120**

Accurate dE/dx simulation and prediction using ML method in the BESIII experiment

Author: Wenxing Fang[None]

Co-authors: Fang Liu [1]; Jinfa Qiu [1]; Kai Zhu [2]; Shengsen Sun ; Tao Lin ; Tong Chen [1]; Weidong Li [3]; Xiaobin Ji [4]; Xiaoling Li [5]

[1] *IHEP*

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

[3] *IHEP, Beijing*

[4] *IHEP, CAS*

[5] *SDU*

Corresponding Author: fangwx@ihep.ac.cn

The Beijing Spectrometer III (BESIII) 1 is a particle physics experiment at the Beijing Electron–Positron Collider II (BEPC II) 2 which aims to study physics in the tau-charm region precisely. Currently, the BESIII has collected an unprecedented number of data and the statistical uncertainty is reduced significantly. Therefore, systematic uncertainty is key for getting more precise results. In the BESIII, the measurement of energy deposition per unit length (so-called dE/dx) from the drift chamber is used for charged particles identification (PID) which is quite important for most analyses 3. Due to the Geant4 can not simulate the energy loss of charged particles in thin gas precisely, a sampling method using experimental data is adopted for dE/dx simulation and it works smoothly 3. In order to reduce the systematic uncertainty from dE/dx PID, advanced machine learning techniques can be tried for accurate dE/dx simulation.

This contribution will present the dE/dx simulation model based on normalizing flows 4 which are stable in training and easy to convergent. Plenty of dE/dx measurements from the experiment are used for training. The metrics for judging the quality of the simulation include the comparison of dE/dx distribution and the dE/dx PID performance between data and simulation. Performance studies show that the simulation has very high fidelity and the dE/dx PID systematic can be reduced to within 1%.

Besides, due to the lack of understanding about dE/dx measurements at a very low beta *gamma region, the expected dE/dx value and resolution can not be fitted well using the traditional method which decreases the dE/dx PID efficiency, especially for protons(anti-protons). To overcome the barrier, fully-connected neural networks are trained to predict the expected dE/dx value and resolution accurately. With this method, the efficiency of dE/dx PID at a very low beta*gamma region can be restored to ~100%.

Reference:
1: BESIII Collaboration, Design and Construction of the BESIII Detector. Nucl.Instrum.Meth.A614:345-399,2010
2: For BEPC II Team, BEPC II: construction and commissioning, Chinese Phys. C 33 60, 2009
3: Cao Xue-Xiang et al. Studies of dE/dx measurements with the BESIII. Chinese Phys. C 34 1852,2010
4: I. Kobyzev, S. J. D. Prince and M. A. Brubaker, "Normalizing Flows: An Introduction and Review of Current Methods," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 11, pp. 3964-3979, 1 Nov. 2021, doi: 10.1109/TPAMI.2020.2992934.

References:

Experiment context, if any:

Significance:

Ideas/methods from this talk could benefit other experiments which use drift chamber for dE/dx PID.

Poster session with coffee break / **121**

Preliminary Results of Vectorization of Density Functional Theory calculations in Geant4/V for amino acids

Author: Oscar Roberto Chaparro Amaro[1]

Co-author: Jesús Alberto Martínez Castro [2]

[1] *Instituto Politécnico Nacional. Centro de Investigación en Computación*

[2] *Instituto Politécnico Nacional. Centro de Investigación en Computación.*

Corresponding Author: ochaparroa2019@cic.ipn.mx

Density Functional Theory (DFT) is an extended ab initio method used for calculating the electronic properties of molecules. Considering Hartree Fock methods, the DFT offers appropriate approximations regarding the time calculations. Recently, the DFT method has been used for discovering and analyzing protein interactions by means of calculating the free energies of these macro-molecules from short to large scales. However, calculating the ground-state energy by DFT for many-body systems of molecules as proteins, in a reasonable time with enough accuracy, is still a very challenging and intensive task for the CPU's resources.
On the other hand, Geant4 is a toolkit for simulating the effects of energy through matter and the nature of materials with a wide range of specialized methods that include DNA and protein exploration. Unfortunately, the execution time to obtain an effective protein analysis is still a strong restriction for CPU processors. In this sense, the GeantV project searches to exploit the vectorization of CPUs, designed to tackle the problem of intensive charge of calculus at the cores of CPUs. In this work, we present the preliminary results of the partial implementation of the DFT in the Geant4 framework

and the vectorized GeantV project. We show the advantages and the partial methods used for vectorizing several sub-routines in the calculus of ground-state energy for some amino acids and some molecules.

Significance:

The novelty of this project consists of the addition of the DFT method to the workflow implemented in Geant4 with an extended analysis dedicated to proteins. Besides, with the incorporation of new strategies to vectorize several applications on modern CPU proposed by the GeantV project in HEP, we aim to include these advances to reduce the execution time of complex processes as the DFT calculations, following the philosophy of modern parallelism.

Finally, We would like to innovate with this work, expanding the aims and scopes of the GeantV project over the calculation of molecular structures at CPU with enough speedup and efficiency, contributing with part of libraries for fast simulation techniques. Therefore, we hope the final results will provide a straightforward approach to methods in software for parallelization.

References:

• Collaboration at the project: GeantV. Results from the Prototype of Concurrent Vector Particle Transport Simulation in HEP: https://doi.org/10.1007/s41781-020-00048-6
• Vectorization techniques for probability distribution functions using VecCore: https://doi.org/10.1088/1742-6596/1525/1/012106
• Vectorization techniques for probability distribution function using VecCore, the 19th-2019 ACAT conference: https://indico.cern.ch/event/708041/contributions/3272109/attachments/1810083/2955720/posterf.pdf
• Hot Spots & Hot Regions Detection Using Classification Algorithms in BMPs Complexes at the Protein-Protein Interface with the Ground-State Energy Feature: https://doi.org/10.1007/978-3-031-07750-0_1

Experiment context, if any:

This work is inspired in Geant4-DNA project: http://geant4-dna.in2p3.fr/index.html Furthermore, this work is intended to be an expansion of the GeantV project: https://geant.web.cern.ch/geant/

Poster session with coffee break / 122

Application of Machine Learning to Particle Identification at the BESIII experiment

Authors: Zhengyuan Chen[1]; Shengsen Sun[1]

Co-authors: Huainmin Liu [1]; Gang Li [1]; Xiaobin Ji [1]; Guang Zhao [1]; XingTao Huang [2]; Tong Chen [1]; Fang Liu [1]; XiaoLing Li [2]; Teng Li [2]; Shuopin Wen [1]; Chunxiu Liu [1]; Wenxing Fang [1]; LiangLiang Wang [1]

[1] *Institute of High Energy Physics Chinese Academy of Sciences*

[2] *Shandong university*

Corresponding Author: chenzhengyuan18@mails.ucas.ac.cn

Particle identification is an important ingredient to particle physics experiments. Distinguishing the charged hadrons (pions, kaons, protons and their antiparticles) is often crucial, in particular for hadronic decays which could be studied with an efficient particle identification to obtain a desirable signal-to-background ratio. An optimal performance of particle identification in a large momentum range requires an effective combination of various relevant variables provided by almost all subsystems of a general-purpose detector. Since the particle identification capability of each variable has a complicated dependence on particle momentum and potential correlations between the other

variables, it is intricate to obtain a perfect performance utilizing a predetermined equation as a model. Machine learning algorithms have been developing to use computational methods to "learn" information directly from data. Particle identification is a typical application of classification techniques of machine learning which involves a large amount of data and lots of features. The BESIII detector is used for studies of hadron physics and $\tau-$charm physics, which is composed of a helium-gas based drift chamber, a time-of-flight system, a CsI(Tl) crystal electromagnetic calorimeter and a resistive plate chamber based muon counter. High statistics and high purity data samples of pion, kaon and (anti-)proton are selected using real data accumulated in BESIII experiment, and the detector responses of these hadron samples have been investigated and summarized. These hadron samples and their characteristics of detector responses offer an unique opportunity to take advantages of machine learning techniques to make classifications or predictions. With application of gradient boosted decision trees (BDT) and usage of various features provided by all sub-detectors, the inherent potentialities of BESIII detector for particle identification is exploited and an enhancement of hadron identification capability has been achieved, especially for high momentum particles.

Significance:

References:

Experiment context, if any:

BESIII

Poster session with coffee break / 123

Data Quality Monitoring for the JUNO Experiment

Authors: Kaixuan Huang[None]; Zhengyun You[1]

[1] *Sun Yat-Sen University (CN)*

Corresponding Author: huangkx28@mail2.sysu.edu.cn

In High Energy Physics (HEP) experiment, Data Quality Monitoring (DQM) system is crucial to ensure the correct and smooth operation of the experimental apparatus during the data taking. DQM at Jiangmen Underground Neutrino Observatory (JUNO) will reconstruct raw data directly from JUNO Data Acquisition (DAQ) system and use event visualization tools to show the detector performance for high quality data taking. The strategy of the JUNO DQM, as well as its design and performance will be presented.

Significance:

References:

Experiment context, if any:

Track 1: Computing Technology for Physics Research / 124

Development of the Topological Trigger for LHCb Run 3

Authors: Blaise Raheem Delaney[1]; J Michael Williams[1]; Johannes Albrecht[2]; Nicole Schulte[2]; Niklas Nolte[3]

[1] *Massachusetts Inst. of Technology (US)*

[2] *Technische Universitaet Dortmund (DE)*

[3] *Massachusetts Institute of Technology (US)*

Corresponding Author: nicole.schulte@tu-dortmund.de

The data-taking conditions expected in Run 3 of the LHCb experiment will be unprecedented and challenging for the software and computing systems. Accordingly, the LHCb collaboration will pioneer the use of a software-only trigger system to cope with the increased event rate efficiently. The beauty physics programme of LHCb is heavily reliant on topological triggers. These are devoted to selecting beauty-hadron candidates inclusively, based on the characteristic decay topology and kinematic properties expected from beauty decays. We present the Run 3 implementation of the topological triggers using Lipschitz monotonic neural networks. This architecture offers robustness under varying detector conditions and sensitivity to long-lived candidates, opening the possibility of discovering New Physics at LHCb.

Significance:

Topological triggers play a fundamental role in the LHCb b-physics program. However, while the selections were based on boosted decision trees in prior years of data taking
1, the former selection algorithms are no longer usable due to an increased luminosity and varying detector conditions during LHC Run 3. For this reason, the Run 3 implementation of the topological triggers uses an entirely new architecture, the so-called Monotonic Lipschitz neural networks 2, which provide robustness against these deviations. Presented is one of the first applications of this architecture. It also provides high efficiency in selecting long-lived beauty candidates due to the introduction of monotonic behaviour in certain selection variables. As a result, a significantly increased efficiency is achieved, which will be crucial to maintaining LHCb's outstanding role in b-physics.

References:

1 https://inspirehep.net/literature/1711636, Design and performance of the LHCb trigger and full real-time reconstruction in Run 2 of the LHC, LHCb Collaboration
2 https://inspirehep.net/literature/1981931, Robust and Provably Monotonic Networks, Ouail Kitouni, Niklas Nolte, Mike Williams

Experiment context, if any:

LHCb Experiment, Flavor Physics

Track 2: Data Analysis - Algorithms and Tools / **125**

Navigation, field integration and track parameter transport through detectors using GPUs and CPUs within the ACTS R&D project

Authors: Andreas Salzburger[1]; Attila Krasznahorkay[1]; Beomki Yeo[None]; Joana Niermann[2]; Stephen Nicholas Swatman[3]

[1] *CERN*

[2] *Georg August Universitaet Goettingen (DE)*

[3] *University of Amsterdam (NL)*

Corresponding Authors: beom.ki.yeo@cern.ch, joana.niermann@cern.ch, andreas.salzburger@cern.ch

The use of hardware acceleration, particularly of GPGPUs is one promising strategy for coping with the computing demands in the upcoming high luminosity era of the LHC and beyond. Track reconstruction, in particular, suffers from exploding combinatorics and thus could greatly profit from the massively parallel nature of GPGPUs and other accelerators. However, classical pattern recognition algorithms and their current implementations, albeit very successfully deployed in the CPU based software of current LHC experiments, show several shortcomings when adapted to modern accelerator architectures; the geometry, for example, is often characterized by runtime-polymorphic shapes,

which are incompatible with common heterogeneous programming platforms. In addition, field integration modules need efficient access to the magnetic field on a variety of devices, and adaptive Runge-Kutta methods may cause thread divergence.

In order to investigate whether state-of-the-art CPU based track reconstruction software can be adapted to run efficiently on GPUs, the ACTS project has launched a dedicated R&D program aiming to develop a demonstrator that mirrors the current track reconstruction chain based on seed finding followed by a combinatorial Kalman filter available in the ACTS suite. We demonstrate the implementation and performance of a core component of this chain: the propagation of track parameters and their associated covariances through a non-homogenous magnetic field including the navigation through a highly complex geometry with different shapes together with the application of material effects when passing through detector material. This demonstrator showcases the usage of the detray library for geometry description and navigation, the covfie library for an efficient description and interpolation of a complex magnetic field on different hardware backends, a dedicated algebra plugin that allows using different math implemenations, and is based on the vecmem library, which has been developed to handle memory resources on host and device. We demonstrate that it is possible to perform this task using single-source code across multiple devices, and we compare the performance of this heterogeneous reconstruction chain to existing CPU-based code in the ACTS project.

Significance:

This is a major step in our R&D program to bring a realistic track reconstruction chain based on Combinatorial Kalman Filtering onto GPUs - it showcases the components that have been developed and partly presented at ACAT and other conferences in a real-world scenario.

References:

The detray library has been presented in ACAT 2021:
https://indico.cern.ch/event/855454/contributions/4605075/

The vecmem library has been presented at ACAT 2021:
https://indico.cern.ch/event/855454/contributions/4605054

The newly developed covfie library will be submitted as a dedicated abstract to this conference, as it could serve a more broader scope (e.g. simulation).

Experiment context, if any:

Track 1: Computing Technology for Physics Research / **126**

Real-time tracking on FPGAs at LHCb

Author: Giulia Tuci[1]

Co-authors: Andrea Contu [2]; Federico Lazzari [3]; Giovanni Bassi [4]; Giovanni Punzi [5]; Michael J. Morello [6]; Riccardo Fantechi [7]; Sofia Kotriakhova [8]; Wander Baldini [8]

[1] *University of Chinese Academy of Sciences (CN)*

[2] *INFN*

[3] *Universita di Siena & INFN Pisa (IT)*

[4] *SNS & INFN Pisa (IT)*

[5] *Universita & INFN Pisa (IT)*

[6] *SNS and INFN-Pisa (IT)*

[7] *INFN - Sezione di Pisa*

[8] *Universita e INFN, Ferrara (IT)*

Corresponding Author: giulia.tuci@cern.ch

In the past four years, the LHCb experiment has been extensively upgraded, and it is now ready to start Run 3 performing a full real-time reconstruction of all collision events, at the LHC average

rate of 30 MHz. At the same time, an even more ambitious upgrade is already being planned (LHCb "Upgrade-II"), and intense R&D is ongoing to boost the real-time processing capability of the experiment. The instantaneous luminosity will significantly increase (x5÷x10), and the trigger system should deal with data coming from more granular and complex detectors. In an effort of moving reconstruction and data reduction to the earliest possible stages of processing, heterogeneous computing solutions are being explored. Specialized coprocessors (computing accelerators) will take responsibility for the most intensive and parallelizable tasks, freeing the more flexible general-purpose processors for higher-level functions. In this talk we describe the results obtained with a life-size demonstrator for the reconstruction of pixel tracking detectors, implemented in commercial, PCIe hosted, FPGA cards. They are interconnected by fast optical links and they operate parasitically on live LHCb data from Run 3. This demonstrator is based on a extremely parallel, 'artificial retina' architecture, and is intended as a first life-size test of the technology, to explore its potential for future larger-scale applications in Real-Time reconstruction at LHCb at high luminosity.

Significance:

For the first time, results obtained with a life-size demonstrator system running on live LHCb data will be shown.

References:

Experiment context, if any:

LHCb

Track 1: Computing Technology for Physics Research / **127**

covfie: a compositional library for heterogeneous vector fields

Author: Stephen Nicholas Swatman[1]

Co-authors: Andreas Salzburger [2]; Attila Krasznahorkay [2]

[1] *University of Amsterdam (NL)*

[2] *CERN*

Corresponding Author: stephen.nicholas.swatman@cern.ch

Vector fields are ubiquitous mathematical structures in many scientific domains including high-energy physics where mdash; among other things mdash; they are used to represent magnetic fields. Computational methods in these domains require methods for storing and accessing vector fields which are both highly performant and usable in heterogeneous environments. In this paper we present covfie, a co-processor-aware vector field library developed by the ACTS community which aims to flexibly and performantly represent vector fields for a wide variety of scientific domains and across a range of programming platforms. To this end, we employ a compositional design philosophy which enables us to meet domain requirements through the composition of simple structures we refer to as vector field *transformers*. In this work, we detail the design and implementation of our library, and enumerate the different kinds of vector fields that our library supports. Furthermore, we evaluate the performance of our library using a mini-application that renders vector magnitudes of a slice of the ATLAS magnetic field on both an x86-based CPU platform and a CUDA-compatible GPGPU platform; through this mini-application, we demonstrate that different storage methods mdash; all of which can be implemented using our library mdash; can have a significant impact on the performance of client applications.

Significance:

This talk is on a novel library with a novel design that has not been presented at ACAT or any other conference; it has only been presented internally. The library is of interest to a wide variety of experiments and applications, including track reconstruction, simulation, and fields other than HEP.

References:

Not applicable.

Experiment context, if any:

Developed in the context of the Acts project, and uses data from the ATLAS detector

Poster session with coffee break / **128**

Data Management interfaces for CMS experiment: building an improved user experience

Author: Rahul Chauhan[1]

[1] *CERN*

Corresponding Author: rahul.chauhan@cern.ch

After a successful adoption of Rucio following its inception in 2018 as the new data management system, a subsequent step is to advertise this to the users among other stakeholders. In this perspective, one of the objectives is to keep improving the tooling around Rucio. As Rucio introduces a new data management paradigm w.r.t the previous model, we begin by tackling the challenges arising from such a shift in the data model, while trying to alleviate the impact on users. Thus we focus on building a monitoring system capable of answering questions that do not naturally fit the current paradigm while also providing new features and services for the users to naturally push further the adoption and the benefits of the new implementation. In this regard, we present the process of development and evolution path of a set of new interfaces dedicated to the extension of the current monitoring infrastructure and the integration of a user-dedicated CLI capable of granting users an almost seamless transition and enhancement for their daily data management activity. We try to maintain minimum dependencies and ensure decoupling to these tools making them of potential use for other experiments. These will form a set of extensions to the Rucio API that is intended at automating a series of most frequent use cases. Eventually enhancing the user experience and lowering the barriers for newcomers.

Significance:

References:

Experiment context, if any:

CMS

Track 1: Computing Technology for Physics Research / **130**

Adoption of the alpaka performance portability library in the CMS software

Author: Andrea Bocci[1]

[1] *CERN*

Corresponding Author: andrea.bocci@cern.ch

To achieve better computational efficiency and exploit a wider range of computing resources, the CMS software framework (CMSSW) has been extended to offload part of the physics reconstruction to NVIDIA GPUs, while the support for AMD and Intel GPUs is under development. To avoid the need to write, validate and maintain a separate implementation of the reconstruction algorithms for each back-end, CMS decided to adopt a performance portability framework. After evaluating different alternative, it was decided to adopt Alpaka as the solution for Run-3.

Alpaka (Abstraction Library for Parallel Kernel Acceleration) is a header-only C++ library that provides performance portability across different back-ends, abstracting the underlying levels of parallelism. It supports serial and parallel execution on CPUs, and extremely parallel execution on GPUs.

This contribution will show how Alpaka is used inside CMSSW to write a single code base; to use different toolchains to build the code for each supported back-end, and link them into a single application; and to select the best back-end at runtime. It will highlight how the alpaka-based implementation achieves near-native performance, and will conclude discussing the plans to support additional back-ends.

Significance:

To the best of our knowledge, this is the first time that a High Energy Physics experimental software makes use of a "performance portability" framework, and builds a *single binary* distribution that can leverage GPUs from different vendors.

References:

https://indico.cern.ch/event/1073640/#2-patatrack
https://indico.cern.ch/event/855454/contributions/4604992/

Experiment context, if any:

CMS

Poster session with coffee break / 131

Exploring the use of accelerators for lossless data compression in CMS

Authors: Andrea Bocci[1]; Stefan Rua[2]

[1] *CERN*

[2] *Aalto University*

Corresponding Author: stefu.rua@gmail.com

The CMS collaboration has a growing interest in the use of heterogeneous computing and accelerators to reduce the costs and improve the efficiency of the online and offline data processing: online, the High Level Trigger is fully equipped with NVIDIA GPUs; offline, a growing fraction of the computing power is coming from GPU-equipped HPC centres. One of the topics where accelerators could be used for both online and offline processing is data compression.

In the past decade a number of research papers exploring the use of GPUs for lossless data compression have appeared in academic literature, but very few practical application have emerged. In the industry, NVIDIA has recently published the nvcomp GPU-accelerated data compression library, based on closed-source implementations of standard and dedicated algorithms. Other platforms, like the IBM Power 9 processors, offer dedicated hardware for the acceleration of data compression tasks.

In this work we review the recent developments on the use of accelerators for data compression. After summarising the recent academic research, we will measure the performance of representative

open- and closed-source algorithms over CMS data, and compare it with the CPU-only algorithms currently used by ROOT and CMS (lz4, zlib, zstd).

Significance:

References:

Experiment context, if any:

CMS

Track 2: Data Analysis - Algorithms and Tools / 132

Particle Transformer for Jet Tagging

Authors: Congqiao Li[1]; Huilin Qu[2]; Sitian Qian[1]

[1] *Peking University (CN)*

[2] *CERN*

Corresponding Author: stqian@pku.edu.cn

Jet tagging is a critical yet challenging classification task in particle physics. While deep learning has transformed jet tagging and significantly improved performance, the lack of a large-scale public dataset impedes further enhancement. In this work, we present JetClass, a new comprehensive dataset for jet tagging. The JetClass dataset consists of 100 M jets, about two orders of magnitude larger than existing public datasets. A total of 10 types of jets are simulated, including several types unexplored for tagging so far. Based on the large dataset, we propose a new Transformer-based architecture for jet tagging, called Particle Transformer (ParT). By incorporating pairwise particle interactions in the attention mechanism, ParT achieves higher tagging performance than a plain Transformer and surpasses the previous state-of-the-art, ParticleNet, by a large margin. The pre-trained ParT models, once fine-tuned, also substantially enhance the performance on two widely adopted jet tagging benchmarks.
https://arxiv.org/abs/2202.03772

Significance:

Jet tagging is a widely adopted analysis technique in high energy physics experiment. In this work, we propose a large and comprehensive public dataset, JetClass. We also propose a transformer based machine learning model for jet tagging, Particle Transformer (ParT). Leveraging novel architect and special pairwise particle interaction information, ParT achieves state-of-the-art performance in jet tagging. Moreover, powered by the comprehensiveness and largeness, model pre-trained with JetClass performed better after fine-tuning with downstream tasks compared with the directly trained ones.

References:

https://icml.cc/virtual/2022/poster/17989
https://indico.cern.ch/event/1144064/abstracts/144880/
https://indico.cern.ch/event/1078970/timetable/?view=standard#29-particle-transformer-for-je

Experiment context, if any:

Track 2: Data Analysis - Algorithms and Tools / 133

Boost-Invariant Polynomials: an efficient and interpretable approach to jet tagging

Authors: Christoph Ortner[1]; Ilyes Batatia[2]; Jose M Munoz[3]

[1] *UBC math Department*

[2] *Engineering Laboratory, University of Cambridge*

[3] *EIA University*

Corresponding Authors: jose.miguel.munoz.arias@cern.ch, ilyes.batatia@ens-paris-saclay.fr

Besides modern architectures designed via geometric deep learning achieving high accuracies via Lorentz group invariance, this process involves high amounts of computation. Moreover, the framework is restricted to a particular classification scheme and lacks interpretability.
To tackle this issue, we present BIP, an efficient and computationally cheap framework to build rotational, permutation, and boost in the jet mean axis invariances. Moreover, we show the versatility of our approach to obtaining state-of-the-art range accuracies in both supervised and unsupervised jet tagging by using several out-of-the-box classifiers.

Significance:

We show that a mathematically inspired multi-body representation of jets can generate a fast, accurate, and interpretable classification of jets at Hadron Colliders via sub-group invariances. Thus obtaining state-of.the-art results within a fraction of the computational cost and speed-ups of orders of magnitude in both training and inference.

References:

Experiment context, if any:

We benchmark our method in both of the following datasets: https://zenodo.org/record/2603256 and https://zenodo.org/record/3164691

Poster session with coffee break / 134

Continuous Integration for the FairRoot Software Stack

Authors: Dennis Klein[1]; Christian Tacke[1]

Co-authors: Florian Uhlig [1]; Mohammad Al-Turany [1]

[1] *GSI - Helmholtzzentrum fur Schwerionenforschung GmbH (DE)*

Corresponding Authors: c.tacke@gsi.de, d.klein@gsi.de

The FairRoot software stack is a toolset for the simulation, reconstruction, and analysis of high energy particle physics experiments (currently used i.e. at FAIR/GSI, and CERN). In this work we give insight into recent improvements of Continuous Integration (CI) for this software stack. CI is a modern software engineering method to efficiently assure software quality. We discuss relevant development workflows and how they were improved through automation. Furthermore, we present our infrastructure detailing its hardware and software design choices. The entire toolchain is composed of free and open source software. Finally, this work concludes with lessons learned from an operational as well as a user perspective and outlines ideas for future improvements.

Significance:

References:

Experiment context, if any:

Track 3: Computations in Theoretical Physics: Techniques and Methods / 135

Two-loop five-point amplitudes in massless QCD with finite fields

Authors: Christian Broennum-Hansen[None]; Dmitry Chicherin[1]; Heribertus Bayu Hartanto[2]; Johannes Henn[3]; Matteo Marcoli[None]; Ryan Moodie[4]; Simon David Badger[5]; Simone Zoia[None]; Thomas Gehrmann[None]; Tiziano Peraro[6]

[1] *Max Planck Institute for Physics*

[2] *Cambridge University*

[3] *Humboldt University Berlin*

[4] *Turin University*

[5] *Universita e INFN Torino (IT)*

[6] *Max Planck Institute for Physics - Munich*

Corresponding Author: ryanmoodie@gmail.com

I will discuss the analytic calculation of two-loop five-point helicity amplitudes in massless QCD. In our workflow, we perform the bulk of the computation using finite field arithmetic, avoiding the precision-loss problems of floating-point representation. The integrals are provided by the pentagon functions. We use numerical reconstruction techniques to bypass intermediate complexity and obtain compact forms for the rational coefficients. I will present results for NLO gluon-initiated diphoton-plus-jet production and NNLO trijet production.

Significance:

References:

arxiv:2106.08664

Experiment context, if any:

Track 1: Computing Technology for Physics Research / 138

Portable Programming Model Exploration for LArTPC Simulation in a Heterogeneous Computing Environment: OpenMP vs. SYCL

Authors: Zhihua Dong[1]; Kyle Knoepfel[2]; Meifeng Lin[3]; Vincent Pascuzzi[1]; Brett Viren[1]; Tianle Wang[1]; Haiwang Yu[1]

[1] *Brookhaven National Laboratory*

[2] *Fermi National Accelerator Laboartory*

[3] *Brookhaven National Laboratory (US)*

Corresponding Author: mlin@bnl.gov

The evolution of the computing landscape has resulted in the proliferation of diverse hardware architectures, with different flavors of GPUs and other compute accelerators becoming more widely available. To facilitate the efficient use of these architectures in a heterogeneous computing environment, several programming models are available to enable portability and performance across different computing systems, such as Kokkos, SYCL, OpenMP and others. As part of the High Energy Physics Center for Computational Excellence (HEP-CCE) project, we investigate if and how these different programming models may be suitable for experimental HEP workflows through a few 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 up on our previous investigations [1, 2] of using Kokkos to port LArTPC simulation in the Wire-Cell Toolkit (WCT) to GPUs, we have explored OpenMP and SYCL as potential portable programming models for WCT, with the goal to make diverse computing resources accessible to the LArTPC simulations. In this presentation, we will describe how we utilize relevant features of OpenMP and SYCL for the LArTPC simulation module in WCT. We will also show performance benchmark results on multi-core CPUs, NVIDIA and AMD GPUs for both the OpenMP and the SYCL implementations. Comparisons with different compilers will be given. Advantages and disadvantages of using OpenMP, SYCL and Kokkos in this particular use case will also be discussed.

Significance:

OpenMP and SYCL are two very different programing models, with OpenMP being compiler directive-based and SYCL a C++-based framework. OpenMP is easy to add to existing codes, while using SYCL will require more code changes. This is the first time OpenMP has been implemented in the context of HEP-CCE. In this presentation we intend to show the feasibility of using OpenMP in C++ code bases to achieve performance portability, while contrasting it with C++-based frameworks, Kokkos and SYCL. We believe this will be of great interest to the broader HEP community.

The experimental high energy physics community has traditionally relied on homogeneous CPU resources, which have been the main target of many HEP software suites. However, it is expected that the CPU resources alone will not be able to meet the computational requirements of the next-generation HEP experiments, such as the Deep Underground Neutrino Experiment (DUNE). We have to adapt our software to utilize compute-accelerator-based heterogeneous computing resources that are provided in large-scale high performance computing facilities. Our exploration of different portable programming models will help guide the software adaptation strategies for WCT, and also inform other HEP software projects of the pros and cons of these programming models.

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

Experiment context, if any:

The Liquid Argon Time Projection Chamber (LArTPC) is a key detector technology that is widely used in current and next generation experiments for neutrino physics, e.g., DUNE and the SBN program. Neutrino events in the LArTPC manifest a large number of disparate patterns, which raises the opportunity for deep-learning algorithms. However, training such algorithms requires very large data sets to achieve accurate performance. But generating such large data sets remains challenging for traditional x86 CPU centric computing facilities. Accelerating the LArTPC simulation with heterogeneous architectures could significantly boost the efficiency of algorithm developments and further the accuracy of physics analyses.

Poster session with coffee break / **139**

General shower simulation MetaHEP in key4hep framework

Authors: Anna Zaborowska[1]; Dalila Salamani[1]; Witold Pokorski[1]

[1] *CERN*

Corresponding Author: dalila.salamani@cern.ch

Description of development of cascades of particles in a calorimeter of a high energy physics experiment relies on precise simulation of particle interactions with matter. It is inherently slow and constitutes a challenge for HEP experiments. Furthermore, with the upcoming high luminosity upgrade of the Large Hadron Collider and a much increased data production rate, the amount of required simulated events will increase accordingly. Several research directions investigated the use of Machine

Learning (ML) based models to accelerate particular calorimeter response simulation. These models typically require a large amount of data and time for training, and the result is a specifically tuned simulation. Meanwhile, meta-learning has emerged in ML community as a fast learning algorithm using small training datasets. In this contribution, we present MetaHEP, a meta-learning approach to accelerate shower simulation in different calorimeters using very high granular data. We show its application using a calorimeter proposed for the Future Circular Collider (FCC-ee) and integration into key4hep framework.

Significance:

References:

Experiment context, if any:

Poster session with coffee break / 140

CMS tracking performance in Run 2 and early Run 3 data using the tag-and-probe technique

Authors: Brunella D'Anzi[1]; CMS Collaboration[None]

[1] *Universita e INFN, Bari (IT)*

Corresponding Author: brunella.d'anzi@cern.ch

Accurate reconstruction of charged particle trajectories and measurement of their parameters (tracking) is one of the major challenges of the CMS experiment. A precise and efficient tracking is one of the critical components of the CMS physics program as it impacts the ability to reconstruct the physics objects needed to understand proton-proton collisions at the LHC. In this work, we present the tracking performance measured in data where the tag and-probe technique was applied to $Z \longrightarrow \mu^+\mu^-$ di-muon resonances for all reconstructed muon trajectories and the subset of trajectories in which the CMS Tracker is used to seed the measurement. The performance is assessed using LHC Run 2 at $\sqrt{s}$ = 13 TeV and early LHC Run 3 data at $\sqrt{s}$ = 13.6 TeV.

Significance:

References:

Experiment context, if any:

CMS experiment

Track 2: Data Analysis - Algorithms and Tools / 141

Gaussian process for calibration and control of GlueX Central Drift Chamber

Author: Torri Jeske[None]

Co-authors: Thomas Britton [1]; Nikhil Kalra ; Naomi Jarvis [2]; Diana McSpadden [1]; David Lawrence [1]

[1] *Jefferson Lab*

[2] *Carnegie Mellon University*

Corresponding Author: dianam@jlab.org

We have developed and implemented a machine learning based system to calibrate and control the GlueX Central Drift Chamber at Jefferson Lab, VA, in near real-time. The system monitors environmental and experimental conditions during data taking and uses those as inputs to a Gaussian process (GP) with learned prior. The GP predicts calibration constants in order to recommend a high voltage (HV) setting for the detector that maintains consistent detector performance (gain and resolution) throughout data taking. This approach is in stark contrast to traditional detector operations in which the detector operates at fixed HV and its calibration parameters vary quite considerably with time. Additionally, the ML based system utilizes uncertainty quantification to correct the recommended control parameters when appropriate. We will present results from the ML system autonomously during the Charged Pion Polarizability (CPP) experiment conducted in Hall D at Jefferson Lab.

Significance:

First instance of utilizing an ML based system to autonomously calibrate and control the GlueX Central Drift Chamber, with uncertainty quantification. Using this system eliminates the need to calibrate the experimental data after the experiment has completed.

References:

Experiment context, if any:

Charged Pion Polarizability, GlueX

Track 2: Data Analysis - Algorithms and Tools / 144

Optimization and deployment of ML fast simulation models

Authors: Anna Zaborowska[1]; Dalila Salamani[1]; Guneet Singh Kohli[None]; Maciej Dragula[None]; Piyush Raikwar[None]; Priyam Mehta[None]; Witold Pokorski[1]

[1] *CERN*

Corresponding Author: piyushraikwar555@gmail.com

In high energy physics experiments, the calorimeter is a key detector measuring the energy of particles. These particles interact with the material of the calorimeter, creating cascades of secondary particles, the so-called showers. Describing development of cascades of particles relies on precise simulation methods, which is inherently slow and constitutes a challenge for HEP experiments. Furthermore, with the upcoming high luminosity upgrade of the LHC with more complex events and a much increased trigger rate, the amount of required simulated events will increase. Machine Learning (ML) techniques such as generative models are currently widely explored for faster simulation alternatives. The pipeline of a ML fast simulation solution consists of multiple components starting from data generation and preprocessing to model training, optimization, validation and deployment within C++ framework. In this contribution, we will present our latest developements: to build a portable and a scalable pipeline with Kubeflow, to automate hyperparameter search with Optuna and NAS and to optimize the inference memory footprint in C++ by leveraging quantization and graph optimization strategies for different hardware acceleretors.

Significance:

References:

Experiment context, if any:

Track 3: Computations in Theoretical Physics: Techniques and Methods / 145

Loop integral computation in the Euclidean or physical kinematical region using numerical integration and extrapolation

Authors: Elise de Doncker[1]; Fukuko Yuasa[2]; Tadashi Ishikawa[2]; Kiyoshi Kato[3]

[1] *Western Michigan University*

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

[3] *Kogakuin University, Japan*

Corresponding Author: elise.dedoncker@wmich.edu

The computation of loop integrals is required in high energy physics to account for higher-order corrections of the interaction cross section in perturbative quantum field theory. Depending on internal masses and external momenta, loop integrals may suffer from singularities where the integrand denominator vanishes at the boundaries, and/or in the interior of the integration domain (for physical kinematics).

In previous work we implemented iterated integration numerically using one- or low-dimensional adaptive integration algorithms in subsequent coordinate directions, enabling intensive subdivision in the vicinity of singularities. To handle a threshold singularity originating from a vanishing denominator in the interior of the domain, we add a term (for example, $i\delta$) in the denominator, and perform a nonlinear extrapolation to a sequence of integrals obtained for a (geometrically) decreasing sequence of δ.

In addition this may give rise to UV singularities, treated by dimensional regularization, where the space-time dimension $n = 4$ is replaced by $n = 4 - 2\varepsilon$ for a sequence of ε values, and a linear extrapolation is applied as ε tends to zero. Presence of both types of singularities may warrant a double extrapolation. In this paper we will devise and apply a strategy for loop integral computations by combining these methods as needed for a set of Feynman diagrams. In view of the compute-intensive nature, the code is further multi-threaded to run in a shared memory environment.

Significance:

In view of the improvements in the technology of high energy physics experiments, accurate theoretical predictions with higher-order corrections are required for an accurate theoretical prediction of the cross-section for particle interactions.
Whereas symbolic or symbolic/numerical calculations are performed for some challenging problems using existing software packages, we focus on the development of fully numerical methods for the evaluation of Feynman loop integrals. The integration strategies adhere to automatic integration, which is a black-box approach for generating an approximation, assuming little or no knowledge of the problem, apart from the specification of the integrand function. In this paper we apply and evaluate integration and extrapolation strategies for a set of Feynman diagrams.

References:

1 "Regularization with Numerical Extrapolation for Finite and UV-Divergent Multi-loop Integrals", E de Doncker, F Yuasa, K Kato, T Ishikawa, J Kapenga, O Olagbemi, Computer Physics Communications 224 (2018), pp. 164-185,
https://doi.org/10.1016/j.cpc.2017.11.001

2 "Numerical Computation of Two-loop Box Diagrams with Masses", Y. Yuasa, E. de Doncker, N. Hamaguchi, T. Ishikawa, K. Kato, Y. Kurihara, Y. Shimizu, Computer Physics Communications 183 (2012), 2136-2144,
http://www.sciencedirect.com/science/article/pii/S0010465512001877

3 "Quadpack Computation of Feynman Loop Integrals", E. de Doncker, J. Fujimoto, N. Hamaguchi, T. Ishikawa, Y. Kurihara, Y. Shimizu, F. Yuasa, Journal of Computational Science (JoCS) 3 (3), (2012), 102-112, doi:10.1016/j.jocs.2011.06.003 , http://www.sciencedirect.com/science/article/pii/S1877750311000573

4 "Extrapolation Algorithms for Infrared Divergent Integrals", E. de Doncker, J. Fujimoto, N. Hamaguchi, T. Ishikawa, Y. Kurihara, M. Ljucovic, Y. Shimizu, F. Yuasa, Proceedings of Science PoS (CPP2010)011,

https://arxiv.org/pdf/1110.3587.pdf

Experiment context, if any:

Track 3: Computations in Theoretical Physics: Techniques and Methods / 146

Physics-informed neural networks: The tug-of-war between knowledge and noise

Authors: Ben Moews[1]; Ricardo Vilalta[2]; Zhenyu Dai[2]

[1] *Carnegie Mellon University & Pittsburgh Supercomputing Center*

[2] *University of Houston*

Corresponding Author: bmoews@andrew.cmu.edu

Physics-informed neural networks (PINNs) have emerged as a coherent framework to build predictive models that combine statistical patterns with domain knowledge. The underlying notion is to enrich the optimization loss function with known equations to constrain the space of possible model solutions. Successful applications cover a variety of areas, including physics, astronomy and bioinformatics. This study investigates the effect of two components within PINNs. First, knowledge-encapsulating physical laws, purposely injected into the learning mechanism and independent of the training data, are used to constrain the model bias and improve generalization performance. Secondly, noisy data has a substantial impact on the search for accurate models when dealing with limited sensors and systematic errors, which can lead to poor generalizations. We use the Bateman–Burgers equation, commonly used in fluid mechanics and nonlinear acoustics, to understand the tug-of-war between knowledge injection and noise perturbation when building models using PINNs.

Significance:

Current state-of-the-art developments in physics-informed neural networks fail to provide a clear understanding of the effect that noise and knowledge exert on the final model's generalization performance. As understanding the role of knowledge as a regularization factor, to narrow the space of possible models, is critical to a number of applications in physics, our study shines a light on these topics and provides new insights. The medium-term direction of this investigation focuses on galaxy evolution in theoretical astrophysics, as well as comparisons to observational telescope data in the same domain, to establish an investigative tool for not yet ascertained variable relationships.

References:

Experiment context, if any:

Poster session with coffee break / 147

AI/ML for PID in the Charged Pion Polarizability Experiment at Jefferson Lab}

Authors: Andrew Schick[None]; David Lawrence[None]; Nikhil Kalra[None]

Corresponding Author: aschick@umass.edu

A precise measurement of the polarizability of the charged pion provides an important experimental test of our understanding of low-energy QCD. The goal of the Charged Pion Polarizability (CPP) experiment in Hall D at JLab, currently underway, is to make a precision measurement of this quantity through a high statistics study of the $\gamma\gamma \rightarrow \pi^+\pi^-$ reaction near 2π threshold. The production

of Bethe-Heitler electron and muon pairs present significant backgrounds, which demand high discrimination between e/π and μ/π to select a clean pion-pair signal. Two independent AI/ML projects were developed to classify μ/π and e/π respectively: a tensorflow-lite model (training in python, inference in C++) for μ/π, and the TMVA package from ROOT for e/π. A new detector, consisting of iron absorbers interspersed with multi-wire proportional chambers, was constructed to enhance the discrimination between muons and pions. Both models were deployed in real time data monitoring to verify good experimental conditions.

Significance:

References:

Experiment context, if any:

Charged Pion Polarizability Experiment (JLAB E12-13-008), The GlueX Collaboration

Track 2: Data Analysis - Algorithms and Tools / 148

A multi-purposed reconstruction method based on machine learning for atmospheric neutrino at JUNO

Authors: Teng LI[1]; Hongyue Duyang[1]; Zhen Liu[2]; Jiaxi Liu[2]

[1] *Shandong University, CN*

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

Corresponding Author: tengli@sdu.edu.cn

The Jiangmen Underground Neutrino Observation (JUNO) experiment is designed to measure the neutrino mass order (NMO) using a 20-kton liquid scintillator detector to solve one of the biggest remaining puzzles in neutrino physics. Regarding the sensitivity of JUNO's NMO measurement, besides the precise measurement of reactor neutrinos, the independent measurement of the atmospheric neutrino oscillation has great potential to enhance the sensitivity in the combined analysis. This heavily relies on the event reconstruction performance at high energy (GeV) level, including the angular resolution of the incident neutrino, the energy resolution, as well as the accuracy of the flavor identification etc.
In this contribution, we present a multi-purposed reconstruction algorithm for high energy particles in JUNO based on machine learning method. This includes extracting effective features from tens of thousands of PMT waveforms, as well as the development of two types of machine learning models (spherical GNN and planar CNN/Transformer). Novel techniques, such as improving the model convergence speed and eliminating reconstruction bias by maintaining the rotation-invariance are also discussed. Preliminary results based on JUNO simulation present reconstruction precision at an unprecedented level, showing great application potential for other large liquid scintillator detectors as well.

Significance:

This contribution covers the novel reconstruction method based on machine learning and results that are un-reported in other conferences.

References:

Experiment context, if any:

Jiangmen Underground Neutrino Observation (JUNO)

Poster session with coffee break / 149

Pyrate: a novel system for data transformations, reconstruction and analysis for the SABRE experiment

Author: Federico Scutti[1]

[1] *Swinburne University of Technology*

Corresponding Author: fscutti@swin.edu.au

The pyrate framework provides a dynamic, versatile, and memory-efficient approach to data format transformations, object reconstruction and data analysis in particle physics. Developed within the context of the SABRE experiment for dark matter direct detection, pyrate relies on a blackboard design pattern where algorithms are dynamically evaluated throughout a run and scheduled by a central control unit. The system intends to improve the user experience, portability and scalability of offline software systems currently available in the particle physics community, with particular attention to medium to small-scale experiments. Pyrate is implemented with the python programming language, allowing easy access to the scientific python ecosystem and commodity big data technologies. This presentation addresses the pyrate design and implementation.

Significance:

The presentation will introduce a new tool designed for data format transformations, object reconstruction and data analysis which improves upon the versatility of similar systems used for offline data processing and analysis at particle physics experiments, with particular advantage to medium to small-scale experiments.

References:

https://indico.cern.ch/event/855454/contributions/4605006/

Experiment context, if any:

The system has been developed in the context of the SABRE experiment for dark matter direct detection in construction at the SUPL laboratory in Victoria Australia and at the LNGS in Italy. The system aims to be an efficient tool for experiments of the same scale.

Poster session with coffee break / 150

A web based graphical user interface for X-ray computed tomography imaging

Author: Yu Hu[None]

Co-authors: Haolai Tian [1]; Kai Zhang [2]; Qi Fazhi [3]; Qingbao Hu [3]; Yan Wang [2]

[1] *Institute of High Energy Physics*

[2] *IHEP, CAS*

[3] *IHEP*

Corresponding Author: huyu@ihep.ac.cn

The high-performance fourth-generation synchrotron radiation light source, e.g., the High Energy Photon Source (HEPS) has been proposed and built successively. The advent of beamlines at fourth-generation synchrotron sources and the advanced detector has made significant progress that push the demand for computing resource at the edge of current workstation capabilities. On the other hand, the vast data volume produced by specific experiments makes it difficult for users to take data away. In this case, on-site data analysis services are necessary both during and after experiments.

On top of this, most synchrotron light source has shifted to prolonged remote operation because of the outbreak of a global pandemic, with the need for remote access to the online instrumental system during the experiments.
A data analysis platform with a graphical user interface (GUI) accessible via the browser-based Jupyter notebook framework was developed to address the above requirements. It aims to provide an interactive and user-friendly tool for the analysis of X-ray synchrotron radiation CT data collected during experiments. This platform allows remote access and quick reconstruction of large datasets from synchrotron radiation CT experiments. Various techniques to subtract background, normalize signal, reconstruct slice, and post-process the image have been made available. Through containerization and container orchestration techniques, it allows the platform to operate on heterogeneous and different scale computing resources.
This presentation will describe the design and status of the web-based data analysis platform for the CT imaging beamline of HEPS, as well as the future plan for this platform.

Significance:

This paper details the work behind the new solution for data analysis of X-ray synchrotron radiation CT imaging. This platform is the first web-based CT data analysis software according to the existing literature.

References:

1.Hu, Y., Li, L., Tian, H.L., Liu, Z.B., Huang, Q.L., Zhang, Y., Hu, H. and Qi, F.Z., 2021b. Daisy:Data analysis integrated software system for x-ray experiments. Epj web conf., 251, p.04020

Experiment context, if any:

High Energy Photon Source (HEPS)

Track 1: Computing Technology for Physics Research / **151**

Using a DSL to read ROOT TTrees faster in Uproot

Authors: Aryan Roy[None]; Jim Pivarski[1]

[1] *Princeton University*

Corresponding Author: aryanroy5678@gmail.com

Uproot reads ROOT TTrees using pure Python. For numerical and (singly) jagged arrays, this is fast because a whole block of data can be interpreted as an array without modifying the data. For other cases, such as arrays of std::vector<std::vector<float», numerical data are interleaved with structure, and the only way to deserialize them is with a sequential algorithm. When written in Python, such algorithms are very slow.

We solve this problem by writing the same logic in a language that can be executed quickly. AwkwardForth is a Domain Specific Language (DSL), based on Standard Forth with I/O extensions for making Awkward Arrays, and it JIT-compiles to a fast virtual machine without requiring LLVM as a dependency. We generate code as late as possible to take advantage of optimization opportunities. All ROOT types previously implemented with Python are being converted to AwkwardForth.

Double and triple-jagged arrays have already been implemented and are 400× faster in AwkwardForth than in Python, with multithreaded scaling up to 1 second/GB because AwkwardForth releases the Python GIL. In this talk, we describe design aspects, performance studies, and future directions in accelerating Uproot with AwkwardForth.

Significance:

This talk presents an implementation of the acceleration anticipated in the talk and paper referenced below. Previously, the I/O speed was measured in a mocked-up (but realistic) test, now it is implemented in Uproot in a way that will be used in production, in Uproot version 5. The observe the same magnitude

of speed-up (with respect to the state-of-the-art Uproot 4) as in the previous talk and paper.

References:

https://indico.cern.ch/event/948465/contributions/4324131/ (vCHEP 2021)
https://inspirehep.net/literature/1849024

Experiment context, if any:

IRIS-HEP

Track 2: Data Analysis - Algorithms and Tools / **152**

Hunting for signals using Gaussian Process regression

Author: Abhijith Gandrakota[1]

Co-authors: Alexandre Morozov [2]; Amitabh Lath [3]

[1] *Fermi National Accelerator Lab. (US)*

[2] *Rutger, The state University of New Jersey*

[3] *Rutgers, The State University of New Jersey*

Corresponding Author: abhijith.gandrakota@cern.ch

We present a novel computational approach for extracting weak signals, whose exact location and width may be unknown, from complex background distributions with an arbitrary functional form. We focus on datasets that can be naturally presented as binned integer counts, demonstrating our approach on the datasets from the Large Hadron Collider. Our approach is based on Gaussian Process (GP) regression - a powerful and flexible machine learning technique that allowed us to model the background without specifying its functional form explicitly, and to separate the background and signal contributions in a robust and reproducible manner. Unlike functional fits, our GP-regression-based approach does not need to be constantly updated as more data becomes available. We discuss how to select the GP kernel type, considering trade-offs between kernel complexity and its ability to capture the features of the background distribution. We show that our GP framework can be used to detect the Higgs boson resonance in the data with more statistical significance than a polynomial fit specifically tailored to the dataset. Finally, we use Markov Chain Monte Carlo (MCMC) sampling to confirm the statistical significance of the extracted Higgs signature.

Significance:

Efficient and accurate background estimation is one of the most crucial aspects of searching for new physics or studying rare physics. Non-resonant backgrounds involving Quantum Chromodynamic processes are often very challenging to estimate. Conventionally these backgrounds are estimated using ad-hoc function fitting, which often fails in the presence of signals with smaller magnitudes or larger data.

Here we present a novel approach using machine learning techniques such as Gaussian Process regression and Bayesian framework to hunt for new physics. This approach is shown to be sensitive and well functioning in the ever-increasing dataset of experimental particle physics.

References:

https://arxiv.org/abs/2202.05856

Experiment context, if any:

This presentation is regarding the background estimation procedure used in experimental high energy physics.

Poster session with coffee break / **153**

CernVM 5: a versatile container-based platform to run HEP applications

Authors: Jakob Blomer[1]; Jakob Karl Eberhardt[2]

[1] *CERN*

[2] *University of Applied Sciences (DE)*

Corresponding Author: jakob.karl.eberhardt@cern.ch

Since its inception, the minimal Linux image CernVM provides a portable and reproducible runtime environment for developing and running scientific software. Its key ingredient is the tight coupling with the CernVM-FS client to provide access to the base platform (operating system and tools) as well as the experiment application software. Up to now, CernVM images are designed to use full virtualization. The goal of CernVM 5 is to deliver all the benefits of the CernVM appliance and to be equally practical as a container and as a full VM. To this end, the CernVM 5 container image consists of a "Just Enough Operating System (JeOS)", with its contents defined by the HEP_OSlibs meta-package commonly used as a base platform in HEP. CernVM 5 further aims at smooth integration of the CernVM-FS client in various container environments (such as Docker, kubernetes, podman, apptainer). Lastly, CernVM 5 uses special build tools and post-build processing to ensure that experiment software stacks using their custom compilers and build chains can coexist with standard system application stacks. As a result, CernVM 5 aims at providing a single, minimal container image that can be used as a virtual appliance for mounting the CernVM-FS client and for running and developing HEP application software.

Significance:

Unlike previous versions, the CernVM 5 appliance works equally well as a container and as a virtual machine. To achieve this novelty, special build methods had to be evaluated and implemented. The CernVM 5 container image can be deployed in various container runtimes such as Docker, kubernetes, podman or apptainer. In addition, the image can be used as a base layer for custom images built with standard tools such as Docker build or buildah.

References:

Experiment context, if any:

Poster session with coffee break / **154**

Mock Data Challenge for the JUNO experiment

Author: Alessandra Carlotta Re[1]

Co-authors: Cailian Jiang [2]; Jilei Xu [3]; Rui Li [4]; Tao Lin [3]; Zhengyun You [5]

[1] *Universita' degli Studi & INFN of Milano (Italy)*

[2] *NJU*

[3] *IHEP*

[4] *SJTU*

[5] *SYSU*

Corresponding Author: alessandra.re@mi.infn.it

The Jiangmen Underground Neutrino Observatory (JUNO) is under construction in South China at a depth of about 700~m underground: the data taking is expected to start in late 2023. JUNO has a very rich physics program which primarily aims to the determination of the neutrino mass ordering

and to the precisely measurement of oscillation parameters.
The JUNO average raw data volume is expected to be about 2~PB/year and
will be transferred from the experimental site to the main computing center (IHEP, Beijing, China) using a dedicated link. When raw data arrive to IHEP, a Data Quality Monitoring (DQM) system will be used to monitor their quality. A so called Keep-Up-Production (KUP) will reconstruct the data and these processed data will be used for detector status studies and for some prompt physics analysis. In order to validate the complete data processing chain, a Mock Data Challenge is being performed and will produce a large scale Monte Carlo data-set for the JUNO experiment.
Due to the rare signals, most of the JUNO expected events are backgrounds, coming from natural radioactivity of rocks, cosmic muons and from the detector itself. There are 17 different components considered in this Mock Data Challenge, and the simulation of each component is performed using the JUNO Distributed Computing Infrastructure (JUNO-DCI). The Monte Carlo output can then be used for the electronics and digitization simulation. However, the electronics simulation needs to simultaneously read a huge amount of data for each background component, and that makes the production on JUNO-DCI really challenging. A pre-mixing method is implemented to mix the radioactivity events beforehand so that the number of required input files can be significantly reduced: a radioactivity background event is picked from the existing data files according to the event rates and then saved into a pre-mixed data file.

In this contribution, details on the Mock Data Challenge, on the JUNO data processing logic-flow and on the practical challenges to be faced for a successful production, will be reported.

Significance:

This would be the first presentation focussed on the data processing chain of the JUNO experiment.

References:

Experiment context, if any:

JUNO

Poster session with coffee break / 155

Primary Vertex Reconstruction for Heterogeneous Architecture at CMS

Authors: Adriano Di Florio[1]; CMS Collaboration[None]; Carlos Francisco Erice Cid[2]; David Sperka[2]; Giorgio Pizzati[3]

[1] *Politecnico e INFN, Bari*

[2] *Boston University (US)*

[3] *Universita & INFN, Milano-Bicocca (IT)*

Corresponding Authors: giorgio.pizzati@cern.ch, carlos.francisco.erice.cid@cern.ch

The future development projects for the Large Hadron Collider will constantly bring nominal luminosity increase, with the ultimate goal of reaching a peak luminosity of $5 \times 10^{34} cm^{-2} s^{-1}$. This would result in up to 200 simultaneous proton collisions (pileup), posing significant challenges for the CMS detector reconstruction.

The CMS primary vertex (PV) reconstruction is a two-step procedure consisting of vertex finding and fitting. First, the Deterministic Annealing algorithm clusters tracks coming from the same interaction vertex. Secondly, an Adaptive Vertex Fit computes the best estimate of the vertex position. In High Luminosity LHC (HL-LHC) conditions, due to the high track density, the reconstruction of PVs is expected to be particularly time expensive (up to 6\% of reconstruction time).

This work presents a complete study about adapting the CMS primary vertex reconstruction algorithms in order to be run on heterogeneous architectures that allows us to exploit parallelization

techniques to significantly reduce the processing time, while retaining similar physics performance. Results obtained for both Run3 and HL-LHC conditions will be discussed.

Significance:

This is the first time a 3D overtaxing algorithm running on GPU is presented by the CMS experiment. No previous result, no previous report.

References:

Experiment context, if any:

CMS Experiment at LHC. Abstract approved by the collaboration.

Poster session with coffee break / **156**

Auto-tuning capabilities of the ACTS track reconstruction suite

Authors: Andreas Salzburger[1]; Corentin Allaire[2]; Elyssa Frances Hofgard[3]; Hadrien Benjamin Grasland[2]; Lauren Alexandra Tompkins[3]; Rocky Bala Garg[3]

[1] *CERN*

[2] *Université Paris-Saclay (FR)*

[3] *Stanford University (US)*

Corresponding Authors: rocky.bala.garg@cern.ch, corentin.allaire@cern.ch

The reconstruction of particle trajectories is a key challenge of particle physics experiments as it directly impacts particle reconstruction and physics performances. To reconstruct these trajectories, different reconstruction algorithms are used sequentially. Each of these algorithms use many configuration parameters that need to be fine-tuned to properly account for the detector/experimental setup, the available CPU budget and the desired physics performance. Examples for such parameters are cut values limiting the search space of the algorithm, approximations accounting for complex phenomenons or parameters controlling algorithm performance. Until now, these parameters had to be optimised by human experts which is inefficient and raises issues for the long term maintainability of such algorithms. Previous experiences with using machine learning for particle reconstruction (such as the TrackML challenge) have shown that they can be easily adapted to different experiments by learning directly from the data. We propose to bring the same approach to the classic track reconstruction algorithms by connecting them to an agent driven optimiser which will allow us to find the best set of input parameters using an iterative tuning approach. We have so far demonstrated this method on different track reconstruction algorithms within A Common Tracking Software (ACTS) framework using the Open Data Detector (ODD). These algorithms include the trajectory seed reconstruction and selection, the particle vertex reconstruction and the generation of simplified material map used for trajectory reconstruction. Finally, we present a development plan for a flexible integration of tunable parameters within the ACTS framework to bring this approach to all aspects of trajectory reconstruction.

Significance:

One of the main lesson from the TrackML challenge and recent machine learning research have shown that using large numbers of tuneable parameters is ideal to create tracking algorithms adapted to different experiments. We present the application of this approach to different tracking algorithms and a plan to generalise it in the ACTS track reconstruction suite.

References:

Auto-tuning with Acts was previously presented in CTD : https://indico.cern.ch/event/1103637/contributions/4821875/

Experiment context, if any:

Poster session with coffee break / 157

k4Clue: Having CLUE at future colliders experiments

Author: Erica Brondolin[1]

Co-authors: Felice Pantaleo [1]; Marco Rovere [1]

[1] *CERN*

Corresponding Author: erica.brondolin@cern.ch

CLUE is a fast and innovative density-based clustering algorithm to group digitized energy deposits (hits) left by a particle traversing the active sensors of a high-granularity calorimeter in clusters with a well-defined seed hit. Outliers, i.e. hits which do not belong to any clusters, are also identified. Its outstanding performance has been proven in the context of the CMS Phase-2 upgrade using both simulated and test beam data.

Initially CLUE was developed in a standalone repository to allow performance benchmarking with respect to its CPU and GPU implementations, demonstrating the power of algorithmic parallelization in the coming era of heterogeneous computing. In this contribution we will outline CLUE's capabilities outside CMS and more specifically, at experiments at future colliders. In order to do so, CLUE was adapted to run in the key4hep framework (k4Clue): it was integrated in the Gaudi software framework and it now supports EDM4hep data format for inputs and outputs.

Implementation details and physics performance will be shown not only for several options of highly granular calorimeters for e+e- linear and circular future colliders, but also for the new Open Data Calorimeter detector, a recent extension to the Open Data Tracking detector, whose aim is to build a simulation-on-the-fly testbed for future algorithm R&D.

Significance:

At the latest ACAT in 2021, the performance of the CLUE algorithm for future experiments was not yet presented. On top of this, the work done to include the calorimeter detector in the Open Data Detector was also never presented before.

References:

- CLUE: A Fast Parallel Clustering Algorithm for High Granularity Calorimeters in High Energy Physics, https://arxiv.org/abs/2001.09761
- CLUE: a clustering algorithm for current and future experiments, E. Brondolin, ACAT 2021, https://indico.cern.ch/event/855454/contributions/4596547/

Experiment context, if any:

Track 3: Computations in Theoretical Physics: Techniques and Methods / 158

Loop Amplitudes from Precision Networks

Authors: Anja Butter[None]; Michel Luchmann[1]; Sebastian Pitz[2]; Simon David Badger[3]; Tilman Plehn[None]

[1] *Universität Heidelberg*

[2] *ITP, Universität Heidelberg*

[3] *Universita e INFN Torino (IT)*

Corresponding Author: butter@thphys.uni-heidelberg.de

Evaluating loop amplitudes is a time-consuming part of LHC event generation. For di-photon production with jets we show that simple, Bayesian networks can learn such amplitudes and model their uncertainties reliably. A boosted training of the Bayesian network further improves the uncertainty estimate and the network precision in critical phase space regions. In general, boosted network training of Bayesian networks allows us to move between fit-like and interpolation-like regimes of network training.

Significance:

For the first time we integrate uncertainties in the training process of Bayesian neural networks for the prediction of amplitudes. This allows a boosting for performance and reliability of the predicted amplitudes.

References:

https://arxiv.org/abs/2206.14831

Previous publications:
https://arxiv.org/abs/2110.13632
https://arxiv.org/abs/2106.09474

Experiment context, if any:

Poster session with coffee break / **159**

The Key4hep Turnkey Software Stack: Beyond Future Higgs Factories

Authors: Andre Sailer[1]; Benedikt Hegner[1]; Clement Helsens[2]; Erica Brondolin[1]; Frank-Dieter Gaede[3]; Gerardo Ganis[1]; Graeme A Stewart[1]; Jiaheng Zou[None]; Placido Fernandez Declara[1]; Sang Hyun Ko[4]; Sylvester Joosten[None]; Tao Lin[None]; Teng LI[5]; Thomas Madlener[6]; Valentin Volkl[1]; Wenxing Fang[None]; Wouter Deconinck[None]; Xingtao Huang[7]; xiaomei zhang[8]

[1] *CERN*

[2] *KIT - Karlsruhe Institute of Technology (DE)*

[3] *Deutsches Elektronen-Synchrotron (DE)*

[4] *Seoul National University (KR)*

[5] *Shandong University, CN*

[6] *Deutsches Elektronen-Synchrotron (DESY)*

[7] *Shandong University*

[8] *IHEP,Beijing*

Corresponding Author: valentin.volkl@cern.ch

The Key4hep project aims to provide a turnkey software solution for the full experiment life-cycle, based on established community tools. Several future collider communities (CEPC, CLIC, EIC, FCC, and ILC) have joined to develop and adapt their workflows to use the common data model EDM4hep and common framework. Besides sharing of existing experiment workflows, one focus of the Key4hep project is the development and integration of new experiment independent software libraries. Ongoing collaborations with projects such as ACTS, CLUE, PandoraPFA and the OpenDataDector show the potential of Key4hep as an experiment-independent testbed and development platform. In this talk, we present the challenges of an experiment-independent framework along with the lessons learned from discussions of interested communities (such as LUXE) and recent adopters of Key4hep in order to discuss how Key4hep could be of interest to the wider HEP community while staying true to its goal of supporting future collider designs studies.

Significance:

The Key4hep software project has recently seen expressions of interests and contributions from outside the initial project stakeholders (EIC, Muon Collider, LUXE and the Open Data Detector). While previous presentations have focused on the progress with regard to the declared goals of the project, this

presentation will focus on the potential and challenges of the use of Key4hep for other experiments and as a common software ecosystem for the wider HEP community.

References:

- ACAT 2021 (poster): https://indico.cern.ch/event/855454/contributions/4604989/
- Ganis, G., Helsens, C. & Völkl, V. Key4hep, a framework for future HEP experiments and its use in FCC. Eur. Phys. J. Plus 137, 149 (2022). https://doi.org/10.1140
- Key4hep Status and Plans at CHEP 2021 https://www.epj-conferences.org/articles/epjconf/abs/2021/05/epjconf_che
- Podio/EDM4hep at CHEP 2021 https://www.epj-conferences.org/articles/epjconf/abs/2021/05/epjconf_chep2021_0
- EPS/HEP Presentation about key4hep: https://indico.desy.de/event/28202/contributions/105603/

Experiment context, if any:

FCC, CLIC, ILC, CEPC, EIC, LUXE

Track 3: Computations in Theoretical Physics: Techniques and Methods / **160**

Invertible Networks for the Matrix Element Method

Authors: Anja Butter[None]; Sascha Peitzsch[1]; Theo Heimel[2]; Till Martini[3]; Tilman Plehn[None]

[1] *Humboldt-Universität zu Berlin*

[2] *Heidelberg University*

[3] *HU Berlin*

Corresponding Author: heimel@thphys.uni-heidelberg.de

For many years, the matrix element method has been considered the perfect approach to LHC inference. We show how conditional invertible neural networks can be used to unfold detector effects and initial-state QCD radiation, to provide the hard-scattering information for this method. We illustrate our approach for the CP-violating phase of the top Yukawa coupling in associated Higgs and single-top production.

Significance:

Our work describes a novel analysis method that combines our precise theory predictions of scattering cross sections at colliders with a machine-learning based method to understand parton shower and detector effects that would be otherwise analytically intractable. In contrast to many conventional analysis methods, this allows us to work with high-dimensional data and to extract as much information as possible from it.

References:

Experiment context, if any:

Track 2: Data Analysis - Algorithms and Tools / **161**

Advances in parallelization of particle showers simulations in CORSIKA 8

Author: Antonio Augusto Alves Junior[1]

[1] *KIT - IAP*

Corresponding Author: augusto.alvesjunior@kit.edu

CORSIKA 8 is a Monte Carlo simulation framework to model ultra-high energy secondary particle cascades in astroparticle physics. This presentation is devoted to the advances in the parallelization of CORSIKA 8, which is being developed in modern C++ and is designed to run on multi-thread modern processors and accelerators, are discussed.

Aspects such as out-of-the-order particle shower calculations, generation of high quality random numbers in multi-thread machines and fast task scheduling and submission on massively parallel platforms are discussed, followed by presentation of CORSIKA 8 approaches, including preliminary performance measurements.

Finally, the design choices and status of integration into CORSIKA 8 are presented, together with some basic examples.

Significance:

This presentation summarizes the efforts for the parallelization of the main software package (CORSIKA) used by the artroparticle physics community for the simulation of extensive high energy particle showers.

References:

Experiment context, if any:

CORSIKA 8, IceCube, Pierre Auger Collaboration

Poster session with coffee break / **163**

CaloPointFlow - Generating Calorimeter Showers as Point Clouds

Author: Simon Schnake[1]

Co-authors: Benno Kach [2]; Dirk Krucker [2]; Kerstin Borras [1]; Moritz Scham [2]; Sofia Vallecorsa [3]

[1] *DESY / RWTH Aachen University*

[2] *Deutsches Elektronen-Synchrotron (DE)*

[3] *CERN*

Corresponding Author: simon.schnake@desy.de

In particle physics, precise simulations are necessary to enable scientific progress. However, accurate simulations of the interaction processes in calorimeters are complex and computationally very expensive, demanding a large fraction of the available computing resources in particle physics at present. Various generative models have been proposed to reduce this computational cost. Usually, these models interpret calorimeter showers as 3D images in which each active cell of the detector is represented as a voxel. This approach becomes difficult for high-granularity calorimeters due to the larger sparsity of the data.

In this study, we use this sparseness to our advantage and interpret the calorimeter showers as point clouds. More precisely, we consider each hit as part of a hit distribution depending on a global latent calorimeter shower distribution.

Our model is based on PointFlow (Yang et al. 2019) and consists of a permutation invariant encoder and two normalizing flows. One flow models the global latent calorimeter shower distribution. The other flow models the distribution of individual hits conditioned on the calorimeter shower distribution.

We present first results, they are shown and compared with state-of-the-art voxel methods.

Significance:

First model to generate calorimeter showers as point clouds.

References:

Experiment context, if any:

Track 1: Computing Technology for Physics Research / 164

Application of Portable Parallelization Strategies for GPUs on track reconstruction kernels

Authors: Alexei Strelchenko[1]; Giuseppe Cerati[1]; Ka Hei Martin Kwok[1]; Matti Kortelainen[1]; Oliver Gutsche[1]

[1] *Fermi National Accelerator Lab. (US)*

Corresponding Authors: ka.hei.martin.kwok@cern.ch, astrel@fnal.gov

Utilizing the computational power of GPUs is one of the key ingredients to meet the computing challenges presented to the next generation of High-Energy Physics (HEP) experiments. Unlike CPUs, developing software for GPUs often involves using architecture-specific programming languages promoted by the GPU vendors and hence limits the platform that the code can run on. Various portability solutions have been developed to achieve portable, performant software across different GPU vendors. Given the rapid evolution of these portability solutions, an early adoption of them in simple HEP testbed applications will help us understand the strengths and weaknesses of respective approaches.

We apply several portability solutions, such as Kokkos, SYCL, std::execution::par and Alpaka, on kernels for track propagation extracted from the mkFit project. We report on the development experience of the same application with different portability solutions, as well as their performance on GPUs, measured as the throughput of the kernels, from different manufacturers such as NVIDIA, AMD and Intel.

Significance:

This is a novel result covering application of portability technologies to HEP-centric kernels on the major GPU vendors.

References:

Experiment context, if any:

Poster session with coffee break / 165

BESIII track reconstruction algorithm based on machine learning

Authors: Xiaoqian Jia[1]; Xiaoshuai Qin[1]; Teng Li[1]; Xingtao Huang[1]; Xueyao Zhang[1]; Yao Zhang[2]; Ye Yuan[2]

[1] *Shandong University*

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

Corresponding Author: jiaxq@mail.sdu.edu.cn

Track reconstruction (or tracking) plays an essential role in the offline data processing of collider experiments. For the BESIII detector working in the tau-charm energy region, plenty of efforts were made previously to improve the tracking performance with traditional methods, such as pattern recognition and Hough transform etc. However, for challenging tasks, such as the tracking of low momentum tracks, tracks from secondary vertices and tracks with high noise level, there is still large room for improvement.

In this contribution, we demonstrate a novel tracking algorithm based on machine learning method. In this method, a hit pattern map representing the connectivity between drift cells is established using an enormous MC sample, based on which we design an optimal method of graph construction, then an edge-classifying Graph Neural Network is trained to distinguish the hit-on-track from noise hits. Finally, a clustering method based on DBSCAN is developed to cluster hits from multiple tracks. Track fitting algorithm based on GENFIT is also studied to obtain the track parameters, where deterministic annealing filter are implemented to deal with ambiguities and potential noises.

The preliminary results on BESIII MC sample presents promising performance, showing potential to apply this method to other drift chamber based trackers as well, such as the CEPC and STCF detectors under pre-study.

Keywords: machine learning, tracking, drift chamber, GNN

Reference:
1. Steven Farrell et al, Novel deep learning methods for track reconstruction. arxiv: 1810.06111
2. A Generic Track-Fitting Toolkit. https://github.com/GenFit/GenFit

Significance:

This contribution covers novel tracking method based on machine learning dealing with drift chamber based trackers. The results present promising performance.

References:

https://indico.cern.ch/event/1128328/contributions/4900740/

Experiment context, if any:

BESIII experiment: http://bes3.ihep.ac.cn/

Track 3: Computations in Theoretical Physics: Techniques and Methods / **166**

Theory prediction in PDF fitting

Author: Felix Hekhorn[None]

Corresponding Author: felix.hekhorn@mi.infn.it

Continuously comparing theory predictions to experimental data is a common task in analysis of particle physics such as fitting parton distribution functions (PDFs). However, typically, both the computation of scattering amplitudes and the evolution of candidate PDFs from the fitting scale to the process scale are non-trivial, computing intesive tasks. We develop a new stack of software tools that aim to facilitate the theory predictions by computing FastKernel (FK) tables that reduce the theory computation to a linear algebra operation. Specifically, I present PineAPPL, our workhorse for grid operations, EKO, a new DGLAP solver, and yadism, a new DIS library. Alongside, I review several projects that become available with the new tools.

Significance:

The tools presented in this talk are all open-source and although developed in the context of the NNPDF PDF fitting collaboration they are completely general purpose and can be leveraged to benefit other common tasks in particle physics. We apply modern best-practice software development tools to ensure

a flexible and extensible framework.

References:

https://arxiv.org/abs/2202.02338
https://arxiv.org/abs/2008.12789

Experiment context, if any:

Track 3: Computations in Theoretical Physics: Techniques and Methods / **167**

Towards an automatized framework to perform quantum calibration

Authors: Andrea Pasquale[1]; Stefano Carrazza[2]

[1] *University of Milan*

[2] *CERN*

Corresponding Author: andrea.pasquale@unimi.it

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 refereed to as noisy intermediate-scale quantum (NISQ) computers and they require calibration routines in order to obtain consistent results.

In this context, we present the latest developments of Qibo, an open-source framework for quantum computing.
Qibo was initially born as a simulation tool in order to simulate quantum circuits.
Through its modular layout for backend abstraction it is possible to change effortlessly between different backends, including a high-performance simulator based on just-in-time compilation able to simulate circuit with large number of qubits (greater than 35).

The latest addiction has been the possibility to employ the language developed by Qibo to execute quantum circuit on real quantum hardware.

Given the necessity to apply calibration routines to characterize the experimental setup, we've also developed a plugin for Qibo, which implements both basic and more advanced calibration routines, including
randomized benchmarking and gate set tomography.

Significance:

The latest updates on Qibo regards the possibility of executing quantum circuits directly on quantum hardware instead of simulating them. We also show how to perform calibration routines and quantum protocols in an automatized way.

References:

https://inspirehep.net/literature/1986757
https://inspirehep.net/literature/2032781
https://inspirehep.net/literature/2054070

Experiment context, if any:

Poster session with coffee break / **168**

Accelerating ROOT compression with Intel ISA-L library

Author: Yu Gao[None]

Co-authors: Yaodong Cheng ; Yaosong Cheng

Corresponding Author: gaoyu94@ihep.ac.cn

ROOT TTree has been widely used in the analysis and storage of various high-energy physical experiment data. The event data generated by the experiment is stored in TTree's bunch and further compressed and archived into a standard ROOT format file. At present, ROOT supports the compression storage of TBasket, the buffer of TBranch, using compression algorithms such as zlib, lzma, lz4, zstd, etc., and maximizes performance by using different compression algorithms in different scenarios, which is of great significance for the increasing amount of high-energy physical data. With the continuous improvement of hardware technology, it is possible to accelerate specific commonly used algorithms from the underlying hardware layer. In this article, by using ISA-L(The Intel Intelligent Storage Acceleration Library), the compression algorithm of ROOT is extended on the Intel X86 machine, enriching the options for ROOT data compression and further improving the comprehensive performance of TTree data compression. Performance tests on intel Xeon Silver 4215R CPUs indicate that the compression time using the ISA-L library is 25% higher than that of the ZSTD algorithm, and the compression rate is slightly better than ZSTD, but the decompression speed is slower than ZSTD. Adding ISA-L support to root allows users to choose more compression methods and effectively reduces compression time.

Significance:

Adding ISA-L support to root allows users to choose more compression methods and effectively reduces compression time.

References:

Experiment context, if any:

Track 2: Data Analysis - Algorithms and Tools / 169

A method for inferring signal strength modifiers by conditional invertible neural networks

Authors: Martin Erdmann[1]; Mate Zoltan Farkas[1]; Niclas Steve Eich[1]; Svenja Diekmann[1]

[1] *Rheinisch Westfaelische Tech. Hoch. (DE)*

Corresponding Author: mate.zoltan.farkas@cern.ch

The continuous growth in model complexity in high-energy physics (HEP) collider experiments demands increasingly time-consuming model fits. We show first results on the application of conditional invertible networks (cINNs) to this challenge. Specifically, we construct and train a cINN to learn the mapping from signal strength modifiers to observables and its inverse. The resulting network infers the posterior distribution of the signal strength modifiers rapidly and for low computational cost. We present performance indicators of such a setup including the treatment of systematic uncertainties and highlight the features of cINNs estimating a signal strength for HEP-data on simulations.

Significance:

Evaluating trained cINNs is ultrafast and gives remarkable precision in reproducing arbitrary probability distributions from Gaussian distributions. Thus cINNs can be used for inference (ref. below). As a new approach to inference in HEP we construct a cINN, include systematic uncertainties, and present benchmark results using a typical simulated Higgs-plus-backgrounds physics case.

References:

Inference of cosmic-ray source properties by conditional invertible neural networks, EPJC 82 (2022) 171

Experiment context, if any:

Simulations of CMS Experiment

Poster session with coffee break / 170

Hyperparameter optimization, multi-node distributed training and benchmarking of AI-based HEP workloads using HPC

Authors: David Southwick[1]; Eduard Cuba[2]; Eric Wulff[1]; Lars Soerlie[3]; Maria Girone[1]

[1] *CERN*

[2] *University of Zurich (CH)*

[3] *Norwegian University of Science and Technology (NTNU) (NO)*

Corresponding Author: eric.wulff@cern.ch

In the European Center of Excellence in Exascale Computing "Research on AI- and Simulation-Based Engineering at Exascale" (CoE RAISE), researchers from science and industry develop novel, scalable Artificial Intelligence technologies towards Exascale. In this work, we leverage European High performance Computing (HPC) resources to perform large-scale hyperparameter optimization (HPO), multi-node distributed data-parallel training as well as benchmarking, using multiple compute nodes, each equipped with multiple GPUs.

Training and HPO of deep learning-based AI models is often compute resource intensive and calls for the use of large-scale distributed resources as well as scalable and resource efficient hyperparameter search algorithms. We evaluate the benefits of HPC for HPO by comparing different search algorithms and approaches, as well as performing scaling studies. Furthermore, the scaling and benefits of multi-node distributed data-parallel training using Horovod are presented, showing significant speed-up in model training. In addition, we present results from the development of a containerized benchmark based on an AI-model for event reconstruction that allows us to compare and assess the suitability of different hardware accelerators for training deep neural networks. A graph neural network (GNN) model known as MLPF, which has been developed for the task of Machine Learned Particle-Flow reconstruction in High Energy Physics (HEP), acts as the base model for which studies are performed.

Further developments of AI models in CoE RAISE have the potential to greatly impact the field of High Energy Physics by efficiently processing the very large amounts of data that will be produced by particle detectors in the coming decades. In order to do this efficiently, techniques that leverage modern HPC systems like multi-node training, large-scale distributed HPO as well as standardized benchmarking will be of great use.

Significance:

We present the latest work in hyperparameter optimization (HPO) of MLPF and the first HPO results using a generator-level ground-truth definition for training a machine-learned algorithm for Particle-Flow reconstruction. In addition, we show multi-node scaling of MLPF training for the first time as well as an AI benchmark based on the MLPF training workload.

References:

https://indico.cern.ch/event/855454/contributions/4598499/
https://arxiv.org/abs/2203.00330
https://arxiv.org/abs/2101.08578

Experiment context, if any:

CMS

Track 2: Data Analysis - Algorithms and Tools / 171

Affine Parametric Neural Networks for High-Energy Physics

Author: Luca Anzalone[1]

Co-authors: Daniele Bonacorsi [2]; Tommaso Diotalevi [1]

[1] *Universita e INFN, Bologna (IT)*

[2] *University of Bologna / INFN*

Corresponding Author: luca.anzalone@cern.ch

Signal-background classification is a central problem in High-Energy Physics (HEP), that plays a major role for the discovery of new fundamental particles. The recent Parametric Neural Network (pNN) is able to leverage multiple signal mass hypotheses as an additional input feature to effectively replace a whole set of individual neural classifiers, each providing (in principle) the best response for the corresponding mass hypothesis. In this work we aim at deepening the understanding of pNNs in light of real-world usage. We discovered several peculiarities of parametric networks, providing intuition, metrics, and guidelines to them. We further propose the affine parametrization scheme, resulting in a new parameterized architecture: the affine parametric neural network (AffinePNN); along with many other generally applicable improvements, like the balanced training procedure, and the background's mass distribution. Finally, we extensively and empirically evaluate our models on the HEPMASS dataset, along its imbalanced version (HEPMASS-IMB) provided by us, to further validate our approach. Presented results are in terms of the impact of the proposed design decisions, classification performance, and interpolation capability.

Significance:

Parametric neural networks can be generally applied to any signal-background classification task, they are experiment agnostic. This means that such method can be applied, in principle, to any analysis and any particle decay. The important aspect of our work is that with the improvements we propose, pNNs are now ready for being employed in a real analysis. In general, we address several aspects of the original pNN method (http://arxiv.org/abs/1601.07913 - which were not addressed by the authors) that are usually a source of confusion across the community.

References:

CMS ML Forum (last talk): https://indico.cern.ch/event/1099308/

Experiment context, if any:

Poster session with coffee break / 172

Deploying a cache content delivery network for CMS experiment in Spain

Authors: Carlos Perez Dengra[1]; Jose Flix Molina[2]; Anna Sikora[3]

[1] *PIC-CIEMAT*

[2] *CIEMAT - Centro de Investigaciones Energéticas Medioambientales y Tec. (ES)*

[3] *Universitat Autònoma de Barcelona (UAB)*

Corresponding Author: carlos.perez.dengra@cern.ch

The Xrootd protocol is used by CMS experiment of LHC to access, transfer, and store data within Worldwide LHC Computing Grid (WLCG) sites running different kinds of jobs on their compute nodes. Its redirector system allows some execution tasks to run by accessing input data that is stored on any WLCG site. In 2029 the Large Hadron Collider (LHC) will start the High-Luminosity LHC (HL-LHC) program, when the luminosity will increase in a factor 10 as compared to the current

values. This scenario will also imply an unprecedented increase of simulation and collision data to transfer, process and store in disk and tape systems. The Spanish WLCG sites that support CMS, the PIC Tier-1 and the CIEMAT Tier-2 have explored content delivery network type solutions in the Spanish region. One of the possible solutions under development has been the deployment of caches between the two sites that store the data requested by the jobs remotely, so that they get closer to the nodes to improve their job efficiency and input data transfer latency. In this contribution, we analyze the impact of deploying physical caches in production in the CMS region between PIC and CIEMAT, as well as the impact they have on job efficiency, latency and bandwidth gains, and potential storage savings.

Significance:

The relevance of this contribution is to discuss the results of the new data management system in the Spanish CMS region in convergence with the WLCG-DOMA data challenges.

References:

1 J. Albrecht, et al, "A Roadmap for HEP Software and Computing RD for the 2020s", Computing and Software for Big Science volume 3, Article number: 7 (2019) https://doi.org/10.1007/s41781-018-0018-8.
2 CMS data access and usage studies at PIC Tier-1 and CIEMAT Tier-2,
Antonio Delgado Peris, José Flix Molina, José M. Hernández, Antonio Pérez-Calero Yzquierdo, Carlos Pérez Dengra, Elena Planas, Francisco Javier Rodríguez Calonge, Anna Sikora, EPJ Web Conf. 245 04028 (2020).
3 Pérez Dengra, C., 2022 "Simulating a network delivery content solution for the CMS experiment in the Spanish" WLCG Tiers, International Symposium on Grids & Clouds (ISGC) 2022 Virtual Conference, 21-25 March 2022, last access on 25th of May of 2022: https://indico4.twgrid.org/event/20/contributions/1116/

Experiment context, if any:

CMS Collaboration

Track 3: Computations in Theoretical Physics: Techniques and Methods / **173**

Product Jacobi-Theta Boltzmann machines with score matching

Authors: Andrea Pasquale[1]; Stefano Carrazza[2]

[1] *University of Milan*

[2] *CERN*

Corresponding Author: andrea.pasquale@unimi.it

We introduce a restricted version of the Riemann-Theta Boltzmann machine, a generalization of the Boltzmann machine with continuous visible and discrete integer valued hidden states. Though the normalizing higher dimensional Riemann-Theta function does not factorize, the restricted version can be trained efficiently with the method of score matching, which is based on the Fisher divergence. At hand of several common two dimensional datasets, we show that the quality of the fits obtained are comparable to state-of-the-art density estimation techniques such as normalizing flows or kernel density estimation. We also discuss how some of these methods can converge to an overfitted solution and we try to quantify this overfitting behavior.
Furthermore, we show that our model is less likely to converge to such non ideal solutions.
We also prove that the recursive calculation of the one dimensional Riemann-Theta function can be extended to the calculation of the first and second order gradients.
We also hint at the possibility of using the density estimated by this model
to perform multi-dimensional integration using Monte Carlo methods with a particular focus on High Energy Physics applications.

Significance:

The major updates on the Riemann-Theta Boltzmann machine are the possibility to train efficiently systems with more than 2 hidden layer thanks to the method of score matching. We also present a novel way to
quantify the overfitting through the surface fractal dimension.

References:

https://inspirehep.net/literature/1644620
https://inspirehep.net/literature/1694236
https://inspirehep.net/literature/1737266

Experiment context, if any:

Poster session with coffee break / **174**

Optimizing electron and photon reconstruction using deep learning: application to the CMS electromagnetic calorimeter

Author: Davide Valsecchi[1]

[1] *ETH Zurich (CH)*

Corresponding Author: davide.valsecchi@cern.ch

The reconstruction of electrons and photons in CMS depends on topological clustering of the energy deposited by an incident particle in different crystals of the electromagnetic calorimeter (ECAL). These clusters are formed by aggregating neighbouring crystals according to the expected topology of an electromagnetic shower in the ECAL. The presence of upstream material (beampipe, tracker and support structures) causes electrons and photons to start showering before reaching the calorimeter. This effect, combined with the 3.8T CMS magnetic field, leads to energy being spread in several clusters around the primary one. It is essential to recover the energy contained in these satellite clusters in order to achieve the best possible energy resolution for physics analyses.

Historically satellite clusters have been associated to the primary cluster using a purely topological algorithm which does not attempt to remove spurious energy deposits from additional pileup interactions (PU). The performance of this algorithm is expected to degrade during LHC Run 3 (2022+) because of the larger average PU levels and the increasing levels of noise due to the ageing of the ECAL detector. New methods are being investigated that exploit state-of-the-art deep learning architectures like Graph Neural Networks (GNN) and self-attention algorithms. These more sophisticated models improve the energy collection and are more resilient to PU and noise.

This contribution covers the model optimization results and the steps to put it in production inside the realistic CMS reconstruction sequence. The impact on the electron and photon energy resolution and tests of the resiliency of the algorithm to the changing detector conditions are shown.

Significance:

The topic of this talk has been shown for the first time at ACAT2021. This new contribution
will cover the process to put the ML algorithm into production in the CMS software and the optimization of the model to obtain the final physics performance for electron and photon reconstruction and the best computational efficiency.

References:

https://arxiv.org/pdf/2204.10277.pdf

Experiment context, if any:

CMS

Poster session with coffee break / 175

Speeding up the CMS track reconstruction with a parallelized and vectorized Kalman-filter-based algorithm during the LHC Run 3

Authors: Allison Reinsvold Hall[1]; Avi Yagil[2]; Bei Wang[3]; Boyana Norris[None]; Brian Gravelle[4]; CMS Collaboration[None]; Daniel Sherman Riley[5]; Frank Wurthwein[6]; Giuseppe Cerati[7]; Kevin McDermott[5]; Leonardo Giannini[2]; Manos Vourliotis[2]; Mario Masciovecchio[2]; Matevz Tadel[2]; Matti Kortelainen[7]; Patrick Gartung[8]; Peter Elmer[3]; Peter Wittich[5]; Slava Krutelyov[2]; Sophie Berkman[None]; Steven R Lantz[5]; Tres Reid[5]

[1] *Fermilab*

[2] *Univ. of California San Diego (US)*

[3] *Princeton University (US)*

[4] *Oregon U. (US)*

[5] *Cornell University (US)*

[6] *UCSD*

[7] *Fermi National Accelerator Lab. (US)*

[8] *Fermilab (US)*

Corresponding Author: emmanouil.vourliotis@cern.ch

One of the most challenging computational problems in the Run 3 of the Large Hadron Collider (LHC) and more so in the High-Luminosity LHC (HL-LHC) is expected to be finding and fitting charged-particle tracks during event reconstruction. The methods used so far at the LHC and in particular at the CMS experiment are based on the Kalman filter technique. Such methods have shown to be robust and to provide good physics performance, both in the trigger and offline. In order to improve computational performance, we explored Kalman-filter-based methods for track finding and fitting, adapted for many-core SIMD architectures. This adapted Kalman-filter-based software, called "mkFit", was shown to provide a significant speedup compared to the traditional algorithm, thanks to its parallelized and vectorized implementation. The mkFit software was recently integrated into the offline CMS software framework, in view of its exploitation during the Run 3 of the LHC. At the start of the LHC Run 3, mkFit will be used for track finding in a subset of the CMS offline track reconstruction iterations, allowing for significant improvements over the existing framework in terms of computational performance, while retaining comparable physics performance. The performance of the CMS track reconstruction using mkFit at the start of the LHC Run 3 is presented, together with prospects of further improvement in the upcoming years of data taking.

Significance:

The deployment of a novel parallel Kalman-filter-based algorithm (called "mkFit") for the charged track reconstruction of the CMS experiment for the LHC Run 3 allows for a very significant improvement of the CMS reconstruction computational performance, while retaining comparable physics performance with respect to the traditional tracking algorithm in use during the LHC Run 2, with clear prospects of further improvement.

References:

Speeding up particle track reconstruction using a parallel Kalman filter algorithm, Steven Lantz (Cornell U.), Kevin McDermott (Cornell U.), Michael Reid (Cornell U.), Daniel Riley (Cornell U.), Peter Wittich (Cornell U.) et al., e-Print: 2006.00071 [physics.ins-det], DOI: 10.1088/1748-0221/15/09/P09030, Published in: JINST 15 (2020) 09, P09030.

Experiment context, if any:

CMS

Poster session with coffee break / 176

Of Frames and schema evolution - The newest features of podio

Authors: Andre Sailer[1]; Benedikt Hegner[1]; Clement Helsens[2]; Frank-Dieter Gaede[3]; Gerardo Ganis[1]; Graeme A Stewart[1]; Placido Fernandez Declara[1]; Thomas Madlener[4]; Valentin Volkl[1]

[1] *CERN*

[2] *KIT - Karlsruhe Institute of Technology (DE)*

[3] *Deutsches Elektronen-Synchrotron (DE)*

[4] *Deutsches Elektronen-Synchrotron (DESY)*

Corresponding Author: thomas.madlener@cern.ch

The podio event data model (EDM) toolkit provides an easy way to generate a performant implementation of an EDM from a high level description in yaml format. We present the most recent developments in podio, most importantly the inclusion of a schema evolution mechanism for generated EDMs as well as the "Frame", a thread safe, generalized event data container. For the former we discuss some of the technical aspects in relation with supporting different I/O backends and leveraging potentially existing schema evolution mechanisms provided by them. Regarding the Frame we introduce the basic concept and highlight some of the functionality as well as important aspects of its implementation. We also present some other, smaller new features, which have been inspired by the usage of podio for generating different EDMs for future collider projects, most importantly EDM4hep, the common EDM for the Key4hep project. We end with a brief overview on current developments towards a first stable version as well as an outlook on future developments beyond that.

Significance:

Missing schema evolution has become a significant issue for podio. Putting together a working solution without re-inventing the wheel is a major milestone. Additionally, we think the Frame concept is worthy of being presented as it addresses many issues in the context of multithreading.

References:

- Podio/EDM4hep at CHEP 2021 https://www.epj-conferences.org/articles/epjconf/abs/2021/05/epjconf_chep2021_03026/epjc

Experiment context, if any:

Future collider projects (FCC, CLIC, ILC, CEPC, EIC, LUXE)

Track 1: Computing Technology for Physics Research / **177**

Investigation of HPC friendly data storage for HEP experiments in the HPC Era

Authors: Amit Bashyal[1]; Peter Van Gemmeren[1]; Saba Sehrish[2]; Kyle Knoepfel[2]; Shane Snyder[1]; Suren Byna[3]

[1] *ANL*

[2] *FNAL*

[3] *LBL*

Corresponding Author: abashyal@anl.gov

With the start of the Run III LHC, computing and storage requirements of the energy and intensity frontiers will grow significantly. In the intensity frontier, with large trigger readouts during supernovae explosions, the DUNE will have unique computing challenges that could be addressed by the

use of parallel and accelerated data-processing capabilities. Most of the requirements of the energy and intensity frontier experiments rely on increasing the role of HPCs in the HEP community. In this presentation, we will describe our ongoing projects contributed by the HEP-Computational Center of Excellence's I/O and Storage working group. The group has been focusing on three topics that will help the experiments to utilize HPC resources for their current and future needs. The first topic is I/O monitoring with Darshan for the computational jobs processed by HPC systems. This effort has led to a number of enhancements to the Darshan tool to enable comprehensive instrumentation of multithread and multiprocess HEP event processing frameworks like ATLAS' Athena and CMS's CMSSW. The second topic is investigating the use of HDF5 as an HPC supported I/O library for intermediate HEP data. Our approach is to rely on ROOT serialization of the often rather complex data model required by HEP experiments and use HDF5 to store the serialized data buffer. We will discuss our methodology of mapping serialized buffers to HDF5, considering both serial and collective IO approaches and present performance scaling studies. The last topic involves demonstration of studies related to the HPC friendly data models that can be stored directly in the HDF5 file format (w/o rely on ROOT serialization) and studies related to the offloading of these data into GPUs.

Significance:

This presentation gives an overview of ongoing efforts to increasing the role of HPCs in the HEP experiments. We demonstrate how HPC resources can be utilized for the HEP experiment requirements and show some of the results from our performance studies.

References:

https://arxiv.org/abs/2203.07885
https://indico.fnal.gov/event/53251/contributions/238753/attachments/153877/199794/Snowmass_compf4_plenary_ver1.p

Experiment context, if any:

This work will be presented on behalf of the HEP-CCE Group. Further information on the group can be found out at: https://www.anl.gov/hep-cce

Poster session with coffee break / **178**

Differentiating through Awkward Arrays using JAX and a new CUDA backend for Awkward Arrays

Author: Anish Biswas[1]

Co-authors: Jim Pivarski [2]; Lukas Alexander Heinrich [3]

[1] *Princeton University (US)*

[2] *Princeton University*

[3] *CERN*

Corresponding Author: anish.biswas@cern.ch

Awkward Array is a library for nested, variable-sized data, including arbitrary-length lists, records, mixed types, and missing data, using NumPy-like idioms. Auto-differentiation (also known as "autograd" and "autodiff") is a technique for computing the derivative of a function defined by an algorithm, which requires the derivative of all operations used in that algorithm to be known.

The grad-hep group is primarily focused on end-to-end analysis, and they use JAX as their primary library for auto-differentiation. As part of such an effort, we developed an interoperability layer between JAX and Awkward Arrays using JAX's pytrees API. JAX now differentiates most of the Awkward Array functions including reducers algorithms. This allows investigators to differentiate through their functions if they are using Uproot with Awkward Arrays. However, extending JAX's vectorized mapping APIs is not possible currently, because of the fundamental differences between the two libraries.

Future work on this might involve testing for a large subset of most commonly used differentiable cases. Currently, testing is carried out on a relatively small number of cases which were developed to catch edge cases.

We also developed a GPU backend for Awkward Arrays by leveraging CuPy's CUDA capabilities. Awkward Arrays now has the entire infrastructure to support operations on a GPU. However, many low-level "C" Kernels (115/204) are yet to be translated to CUDA. After implementing this, Awkward Arrays will have full GPU support and this would indirectly help in making auto-differentiation fully deployable on the GPUs too.

Significance:

References:

Experiment context, if any:

Track 3: Computations in Theoretical Physics: Techniques and Methods / **179**

Performance of modern color decompositions for standard candle LHC tree amplitudes

Authors: Enrico Bothmann[1]; Joshua Isaacson[None]; Max Knobbe[None]; Rui Wang[2]; Stefan Hoeche[3]; Taylor Childers[2]; Walter Giele[None]

[1] *University of Göttingen*

[2] *Argonne National Laboratory (US)*

[3] *Fermilab*

Corresponding Author: max.knobbe@uni-goettingen.de

For more than a decade the current generation of fully automated, matrix element generators has provided hard scattering events with excellent flexibility and good efficiency.
However, as recent studies have shown, they are a major bottleneck in the established Monte Carlo event generator toolchains. With the advent of the HL-LHC and ever rising precision requirements, future developments will need to focus on computational performance, especially at intermediate to large jet multiplicities.
We present the novel BlockGen family of fast matrix element algorithms that are amenable for GPU acceleration, making use of modern, minimal color decompositions. Moreover, we discuss the performance achieved for standard candle processes such as V+jets and $t\bar{t}$+jets production.

Significance:

This presentation will cover first implementation of novel QCD amplitude methods relevant for next-generation matrix element generators. After previous pathfinder studies, we will present the finished implementation and the resulting significant computational improvements. The resulting algorithms are especially suited for the deployment in modern architectures, but already achieve great performance in the established frameworks.

References:

Publication/Talk for pathfinder-study:
https://inspirehep.net/literature/1868130

Experiment context, if any:

Poster session with coffee break / **180**

ROCm on Gentoo: efficient and portable GPGPU software management

Author: Yiyang Wu[1]

[1] *Tsinghua University*

Corresponding Author: xgreenlandforwyy@gmail.com

Gentoo is a GNU/Linux distribution with a vast collection of scientific software. Its package manager, Portage, can resolve complex dependencies and build software from source automatically. Packages are highly customizable, and can be installed on a prefixed path called Gentoo Prefix, providing a feature-rich and portable deployment solution.
ROCm™ open software platform is a collection of free software focusing on general-purpose computing on graphics processing units (GPGPU), high performance computing (HPC) and heterogeneous computing. It is under actively development, and releases the great performance of AMD GPUs and accelerators.
Deploying ROCm may not be easy, so there are various efforts going on to package ROCm, including major Linux distributions. Among all distributions, Gentoo has the most complete support of ROCm in its main repository, and have popular frameworks like PyTorch in its overlays. With Gentoo Prefix, ROCm along with other physics software can be installed anywhere without privilege. Gentoo can even support AMD GPUs beyond official ones. With the powerful functions from Portage and supports from community, Gentoo ROCm packages are quality-assured, highly maintainable and customizable, providing an elegant solution for deploying ROCm in most scenarios of physics research.

Significance:

References:

Experiment context, if any:

Track 1: Computing Technology for Physics Research / **181**

Computing for Gravitational-wave Research towards O4

Author: Stefano Bagnasco[1]

[1] *Istituto Nazionale di Fisica Nucleare, Torino*

Corresponding Author: bagnasco@to.infn.it

The LIGO, VIRGO and KAGRA Gravitational-wave interferometers are getting ready for their fourth observational period, scheduled to begin in March 2023, with improved sensitivities and higher event rates.

Data from the interferometers are exchanged between the three collaborations and processed by running search pipelines for a range of expected signals, from coalescing compact binaries to continuous waves and burst events, along with sky localisation and parameter estimation pipelines. One of the most important peculiarities of GW computing (and, more generally, of time-domain astrophysics) is that data processing happens both offline and on special low-latency infrastructures, in order to provide timely "event candidate alerts" to other observatories and make multi-messenger astronomy possible.

Significant efforts have been made in recent years to design and build a common computing infrastructure, both in terms of a common architecture and shared resources, to prepare for growing computing demand and increasingly exploit distributed computing resources. Many custom tools, difficult to maintain, have been replaced by more mainstream tools, more widely adopted in the

physics community, in order to streamline workflows and reduce the burden of maintenance and operations.

We report on this activities, the status of the infrastructure and the plans for the upcoming observation period.

Significance:

Computing infrastructure for GW is coming of age, and will need to be upgraded also in view of future upgrades and of the proposed third generation observatories

References:

See, for example, https://indico.egi.eu/event/5464/contributions/15714/

Experiment context, if any:

Virgo, LIGO, KAGRA

Poster session with coffee break / **182**

Fast Named Data Networking Based Open Storage System Plugin For XRootD

Authors: Catalin Iordache[1]; Justas Balcas[1]; Yuanhao Wu[2]

Co-authors: Davide Pesavento [3]; Edmund Yeh [2]; Harvey Newman [1]; Jason Cong [4]; Lixia Zhang [4]; Michael Lo [4]; Raimondas Sirvinskas [1]; Sankalpa Timilsina [5]; Sichen Song [4]; Susmit Shannigrahi [5]

[1] *California Institute of Technology (US)*

[2] *Northeastern University (US)*

[3] *National Institute of Standards and Technology (US)*

[4] *University of California, Los Angeles (US)*

[5] *Tennessee Tech University (US)*

Corresponding Authors: justas.balcas@cern.ch, catalinn.iordache@gmail.com

We present an NDN-based Open Storage System (OSS) plugin for XRootD instrumented with an accelerated packet forwarder, built for data access in the CMS and other experiments at the LHC, together with its current status, performance as compared to other tools and applications, and plans for ongoing developments.

Named Data Networking (NDN) is a leading Future Internet Architecture where data in the network is accessed directly by its name rather than the location of the data containers (hosts). NDN enables the joint design of multipath forwarding and caching to achieve superior latency and failover performance. The Caltech team, together with Northeastern University, UCLA, Tennessee Tech and other collaborators in the NDN for Data Intensive Science Experiments (N-DISE) project, has implemented (1) a small C++ NDN library (NDNc) to bridge the existing NDN libraries with the newly developed high-throughput NDN-DPDK forwarder by NIST, (2) a corresponding NDN naming scheme in order to access datasets in the network, (3) two fundamental entities for transferring data in NDN: a consumer and a producer, and (4) an NDN-based OSS plugin for XRootD.

The XRootD plugin offers implementation for all filesystem related calls (e.g., open, read, close) and it embeds the NDN consumer that is able to translate these calls to NDN Interest packets: the Interest addressing a read operation for the third segment from a file at */path/to/foo* location on disk has the corresponding Name */ndnc/ft/path/to/foo/v=1/seg=3*. Once Interest packets are assembled they are passed to a proxy object that runs different congestion window algorithms (e.g., fixed window, congestion-aware AIMD) to send them to the local interface accordingly. The window also handles retransmission and timeouts. The local interface implements a memif shared memory packet interface providing high-performance packet transmission to and from the local NDN-DPDK forwarder.

The forwarder broadcasts the packets on the NDN network to find data from either in-network caches or data producers. At the beginning of the runtime session of the XRootD service, the OSS plugin configures the local forwarder via a C++ GraphQL client available in the NDNc library by creating interfaces and advertising Name prefixes. Alongside this plugin, a corresponding producer has been implemented, which can communicate with multiple file systems (CEPH, HDFS); upon receiving NDN Interests, the producer responds with data encapsulating byte ranges at an offset from an existing file, indicated by the segment number of the arriving packet.

In this paper we present the architecture of: the NDNc library, the consumer application and the NDN-based XRootD plugin. We will also present the throughput performance of the plugin over a continental-scale wide area network testbed and comparisons with other tools and applications used for accessing data at the CMS experiment.

Significance:

This work presents a novel file system plugin that allows the CMS community to take advantage of NDN through in-network caching, built-in multicast, and location independent data retrieval.

References:

https://www.nist.gov/news-events/events/2021/10/ndn-community-meeting-2021
https://www.epj-conferences.org/articles/epjconf/abs/2020/21/epjconf_chep2020_04018/epjconf_chep2020_04018.html
https://ieeexplore.ieee.org/document/9444084
https://vast.cs.ucla.edu/sites/default/files/publications/atc19-final97.pdf
https://nsf.gov/awardsearch/showAward?AWD_ID=2019012

Experiment context, if any:

Compact Muon Solenoid experiment at the LHC

Poster session with coffee break / **183**

Bayesian method for waveform analysis with GPU acceleration

Author: Yuyi Wang[1]

[1] *Tsinghua University*

Corresponding Author: strawberry_str@hotmail.com

One way to improve the position and energy resolution in neutrino experiments, is to give parameters with high resolution to the reconstruction method. These parameters, the photon electron(PE) hit time and the expectation of PE count, can be analyzed from the waveforms. We developed a new waveform analysis method called Fast Scholastic Matching Pursuit(FSMP). It is based on Bayesian principles, and the possible solutions are sampled with Markov Chain Monte Carlo(MCMC). To accelerate the method, we ported it to GPU, and could analysis the waveforms with 0.01s per waveform. This method extracts all the information in the waveforms, and will benefit event reconstruction with high resolution. With the improved resolution, we can make our way to our final physics goal.

Significance:

References:

Experiment context, if any:

Poster session with coffee break / **184**

Reconstructing Particle Decay Trees with Quantum Graph Neural Networks for High Energy Physics

Author: Melvin Strobl[None]

Co-authors: Eileen Kuehn [1]; Max Fischer [1]; Achim Streit [2]

[1] *Karlsruhe Institute of Technology*

[2] *Karlsruhe Institute of Technology (KIT)*

Corresponding Author: melvin.strobl@kit.edu

Quantum Computing and Machine Learning are both significant and appealing research fields. In particular, the combination of both has led to the emergence of the research field of quantum machine learning which has recently taken enormous popularity. We investigate in the potential advantages of this synergy for the application in high energy physics, more precisely in the reconstruction of particle decay trees in particle collision experiments. Due to the larger computational space of quantum computers, this highly complex combinatorical problem is well suited for investigating in a potential quantum advantage compared to the classical scenario. However, current quantum devices are subject to noise and provide only a limited number of qubits. We therefore propose the utilization of a variational quantum circuit within a classical graph neural network which has been shown to be feasible for reconstruction of particle decay trees before. We evaluate our approach on artificially generated decay trees on a quantum simulator and a real quantum computer by IBM Quantum and compare our results to the purely classical approach. Our proposed approach does not only enable the effective utilization of nowadays quantum devices, but also shows competitive results even in the presence of noise.

Significance:

The presentation will show the results of our investigations of using variational quantum circuits in graph neural networks under noisy conditions for reconstructing particle decay trees.

References:

Experiment context, if any:

Poster session with coffee break / **185**

Optimized GPU usage in High Energy Physics applications

Author: Tim Voigtlaender[1]

Co-authors: Gunter Quast [1]; Manuel Giffels [1]; Matthias Schnepf ; Roger Wolf [1]

[1] *KIT - Karlsruhe Institute of Technology (DE)*

Corresponding Author: tim.voigtlaender@cern.ch

Machine Learning (ML) applications, which have become quite common tools for many High Energy Physics (HEP) analyses, benefit significantly from GPU resources. GPU clusters are important to fulfill the rapidly increasing demand for GPU resources in HEP. Therefore, the Karlsruhe Institute of Technology (KIT) provides a GPU cluster for HEP accessible from the physics institute via its batch system and the Grid. As the exact hardware needs of such applications heavily depend on the ML hyperparameters, a flexible resource setup is necessary to utilize the available resources as efficient as possible. Therefore, the multi-instance GPU feature of the Nvidia A100 GPUs was studied. Several neural network training scenarios performed on the GPU cluster at KIT are discussed to illustrate possible performance gains and the setup that has been used.

Significance:

The basics we use are HTCondor and the MIG feature from NVIDIA and are described in other publications. However, we provide the resources, as one of a handful of Grid sites, to the Grid. Furthermore, the resources are shared with end-users with a more complex set of resource requirements than Grid jobs. Our experience and ideas on how to use GPUs efficiently in such an environment seem unique.

References:

Experiment context, if any:

CMS

Poster session with coffee break / **186**

Advancing Opportunistic Resource Management via Simulation

Author: Max Fischer[1]

Co-author: Eileen Kuehn [1]

[1] *Karlsruhe Institute of Technology*

Corresponding Author: max.fischer@kit.edu

Modern high energy physics experiments and similar compute intensive fields are pushing the limits of dedicated grid and cloud infrastructure. In the past years research into augmenting this dedicated infrastructure by integrating opportunistic resources, i.e. compute resources temporarily acquired from third party resource providers, has yielded various strategies to approach this challenge. However, work on this topic is usually driven by practical needs to use specific resource providers for production workflows; in this context, research is ad hoc and relies on impressions gained during unique situations of resource providers, resource demand and opportunistic resource management. Replicating or even preparing a specific situation to investigate opportunistic resource management is extremely challenging or even impossible. More importantly research in the field of opportunistic resource management is therefore extremely limited.

We propose to tackle this challenge using simulation and to this end present the simulation framework LAPIS, a general purpose scheduling simulator offering programmatic control of resources. We demonstrate this approach by integrating LAPIS with the COBalD/TARDIS resource manager to investigate the behaviour of this resource manager in a simulated environment.

Significance:

Our work is integral for advancing opportunistic resource management in a principled way. To our knowledge, this is the first successful attempt to run a production resource manager in a completely simulated environment.

References:

"Transparent Integration of Opportunistic Resources into the WLCG Compute Infrastructure" Böhler, M.; Caspart, R.; Fischer, M.; Freyermuth, O.; Giffels, M.; Kroboth, S.; Kuehn, E.; Schnepf, M.; Cube, F. von; Wienemann, P. 2021. The European physical journal / Web of Conferences, 251, Art.-Nr.: 02039. doi:10.1051/epjconf/202125102039

"Effective Dynamic Integration and Utilization of Heterogenous Compute Resources" Fischer, M.; Giffels, M.; Heiss, A.; Kuehn, E.; Schnepf, M.; Cube, R. F. von; Petzold, A.; Quast, G. 2020. The European physical journal / Web of Conferences, 245, Article no: 07038. doi:10.1051/epjconf/202024507038

Experiment context, if any:

Poster session with coffee break / 187

Equivariant Graph Neural Networks for Charged Particle Tracking

Authors: Ameya Thete[1]; Daniel Thomas Murnane[2]; Savannah Jennifer Thais[3]

[1] *Birla Institute of Technology and Science, Pilani - KK Birla Goa Campus (IN)*

[2] *Lawrence Berkeley National Lab. (US)*

[3] *Princeton University (US)*

Corresponding Author: ameyathete11@gmail.com

A broad range of particle physics data can be naturally represented as graphs. As a result, Graph Neural Networks (GNNs) have gained prominence in HEP and have increasingly been adopted for a wide array of particle physics tasks, including particle track reconstruction. Most problems in physics involve data that have some underlying compatibility with symmetries. These problems may either require, or at the very least, benefit from models that perform computations and construct representations that reflect these symmetries. In this work, we explore the application of symmetry group equivariance to GNNs within the context of charged particle tracking in pileup conditions similar to those expected at the high-luminosity Large Hadron Collider. In particular, we investigate whether rotationally-equivariant GNNs can perform competitively and yield models that either contain fewer, more expressive learned parameters or are more efficient vis-à-vis data and computational requirements. To our knowledge, this is the first study exploring equivariant GNNs for a track reconstruction use case. Additionally, we perform a side-by-side comparison of equivariant and non-equivariant architectures over evaluation metrics that capture both outright tracking performance as well as the track-building power-to-weight ratio of physics-constrained GNNs.

Significance:

While there has been recent progress in developing symmetry equivariant GNNs for particle physics, they have mostly been limited to tasks like jet tagging. The presented work is perhaps the first study that discusses the application of equivariant graph neural network architectures to a particle track reconstruction task. It represents an essential contribution toward future architectures that are potentially more robust under newer computational paradigms.

References:

Experiment context, if any:

Track 3: Computations in Theoretical Physics: Techniques and Methods / 188

Developments in Performance and Portability of BlockGen

Authors: Enrico Bothmann[1]; Joshua Isaacson[None]; Max Knobbe[None]; Rui Wang[2]; Stefan Hoeche[3]; Taylor Childers[2]; Walter Giele[None]

[1] *University of Göttingen*

[2] *Argonne National Laboratory (US)*

[3] *Fermilab*

Corresponding Authors: rui.wang@cern.ch, jchilders@anl.gov

For more than a decade Monte Carlo (MC) event generators with the current matrix element algorithms have been used for generating hard scattering events on CPU platforms, with excellent flexibility and good efficiency.
While the HL-LHC is approaching and precision requirements are becoming more demanding, many

studies have been made to solve the bottleneck in the current MC event generator toolchains. The novel family of fast matrix element algorithms (BlockGen) shown in this report, is one of the new developments that are more suitable for GPU acceleration.
We report the development experience of porting Blockgen using Kokkos. Moreover, we discuss the performance of the Kokkos version in comparison with the dedicated GPU version in CUDA.

Significance:

References:

Experiment context, if any:

Poster session with coffee break / 189

RDataFrame: a flexible and scalable analysis experience

Authors: Enrico Guiraud[1]; Vincenzo Eduardo Padulano[2]; Enric Tejedor Saavedra[3]; Ivan Kabadzhov[4]; Pawan Pawan[None]

[1] *EP-SFT, CERN*

[2] *Valencia Polytechnic University (ES)*

[3] *CERN*

[4] *Albert Ludwig University of Freiburg*

Corresponding Author: vincenzo.eduardo.padulano@cern.ch

The growing amount of data generated by the LHC requires a shift in how HEP analysis tasks are approached. Usually, the workflow involves opening a dataset, selecting events, and computing relevant physics quantities to aggregate into histograms and summary statistics. The required processing power is often so high that the work needs to be distributed over multiple cores and multiple nodes. This contribution establishes ROOT RDataFrame as the single entry point for virtually all HEP data analysis use cases. In fact, the typical steps of an analysis workflow can be easily and flexibly written with RDataFrame. Data ingestion from multiple sources is streamlined through a single interface. Relevant metadata can be made available to the dataframe and used during analysis execution. A declarative API offers the most common operations to the users, while transparently taking care of data processing optimisations. For example, it is possible to inject user-defined code to compute complex quantities, gather them into histograms or other relevant statistics, include large sets of systematic variations and use machine-learning inference kernels. A Pythonic layer allows dynamic injection of Python functions in the main C++ event loop. Finally, any RDataFrame application can seamlessly scale out to hundreds of cores on the same machine or multiple distributed nodes by changing a single line of code. The latest performance validation studies are also included in this contribution to demonstrate the efficiency of the tool on both the computation complexity and the scalability spectra.

Significance:

This contribution demonstrates how a physics analysis can be written from begin to end with a single interface. All users can benefit from having a coherent interface that removes the burden of thinking about the programming implementation and just focus on the desired results. Virtually all analysis workflows can be written with RDataFrame, which is now more flexible than ever. Most relevant new features of the tool are: the possibility of writing a Python-only application that still exploits a fast C++ core; the inclusion of machine learning kernels in the event loop; handling metadata of large data samples directly within RDataFrame itself, thus enabling usage of such information directly within the event loop on a per-sample basis. The attendees will see a HEP analysis being written step by step with this tool, from easier tasks with the aim of exploring the dataset to complex operations that represent a full analysis in production.

References:

Experiment context, if any:

Poster session with coffee break / 190

Evaluating Portable Parallelization Strategies for Heterogeneous Architectures

Authors: Alexei Strelchenko[1]; Beomki Yeo[None]; Brett Viren[2]; Charles Leggett[3]; Haiwang Yu[None]; Ka Hei Martin Kwok[1]; Kyle Knoepfel[None]; Mark Dewing[None]; Matti Kortelainen[1]; Meghna Battacharya[None]; Meifeng Lin[4]; Mohhamad Atif[5]; Oliver Gutsche[1]; Paolo Calafiura[6]; Salman Habib[7]; Taylor Childers[8]; Tianle Wang[2]; Vakho Tsulaia[6]; Vincent Pascuzzi[2]; Zhihua Dong[None]

[1] *Fermi National Accelerator Lab. (US)*

[2] *Brookhaven National Laboratory*

[3] *Lawrence Berkeley National Lab (US)*

[4] *Brookhaven National Laboratory (US)*

[5] *BNL*

[6] *Lawrence Berkeley National Lab. (US)*

[7] *Argonne National Laboratory*

[8] *Argonne National Laboratory (US)*

Corresponding Authors: charles.g.leggett@gmail.com, astrel@fnal.gov

High-energy physics (HEP) experiments have developed millions of lines of code over decades that are optimized to run on traditional x86 CPU systems. However we are seeing a rapidly increasing fraction of floating point computing power in leadership-class computing facilities and traditional data centers coming from new accelerator architectures, such as GPUs. HEP experiments are now faced with the untenable prospect of rewriting millions of lines of x86 CPU code, for the increasingly dominant architectures found in these computational accelerators. This task is made more challenging by the architecture specific languages and APIs promoted by manufacturers such as NVIDIA, Intel and AMD. Producing multiple, architecture specific implementations is not a viable scenario, given the available person power and code maintenance issues.

The Portable Parallelization Strategies team of the HEP Center for Computational Excellence is investigating the use of Kokkos, SYCL, OpenMP, std::execution::parallel and Alpaka as potential portability solutions that promise to execute on multiple architectures from the same source code, using an assortment of representative use cases from DUNE, LHC ATLAS and CMS experiments. Central to the project is to develop a list of metrics that evaluate the suitability of each portability layer for the various testbeds. This list includes both subjective ratings, such as the ease of learning the language, and objective criteria such as performance.

We report on the status of these projects, the development and evaluation of the metrics, as well as the current benchmarks and evaluations of the portability layers for the testbeds under study and recommendations for HEP experiments seeking forward looking portability solutions.

Significance:

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 problem addressing all currently available heterogeneous architectures.

The HEP-CCE/PPS effort is the only cross experiment investigation that is tackling the issue of software portability on heterogeneous architectures with a very broad selection of portability solutions. We are addressing the needs of both large and small experiments with representative testbeds taken from a broad variety of sources. We are working in close proximity with the various experiments, with core developers from the experiments being part of the CCE/PPS team, which facilitates the cross pollination

of knowledge and experiences, and feedback cycles with the experiments.

References:

- Childers, Taylor, et al. "Porting CMS Heterogeneous Pixel Reconstruction to Kokkos." vCHEP 2021. arXiv:2104.06573v1.
- Dong, Zhihua, et al. "Porting HEP Parameterized Calorimeter Simulation Code to GPUs." Frontiers in Big Data. arXiv:2103.14737v2.
- Kortelainen, Matti J., et al. "Performance of CUDA Unified Memory in CMS Heterogeneous Pixel Reconstruction." vCHEP 2021.
- Pascuzzi, Vincent R., Goli, Mehdi. "Achieving Near Native Runtime Performance and Cross-Platform Performance Portability for Random Number Generation Through SYCL Interoperability." arXiv:2109.01329
- Yu, Haiwang, et al. "Evaluation of Portable Acceleration Solutions for LArTPC Simulation Using Wire-Cell Toolkit." vCHEP 2021. arXiv:2104.08265v1.

Experiment context, if any:

We have taken representative testbeds to evaluate the various portability layers from a number of current HEP experiments, namely Patatrack (pixel tracking) from CMS, p2r (central "propagate to R" tracker from CMS), Wirecell (Liquid Argon Time Projection Chamber toolkit) from DUNE, FastCaloSim (parametrized Liquid Argon Calorimeter Simulation) from ATLAS, and ACTS (detector agnostic tracking toolkit) with developers from ATLAS, LHCb, and sPHENIX. All these experiments are investigating software solutions for heterogeneous architectures for current and future runs. While some have made interim decisions for their current runs that are focussed on NVIDIA hardware, all are facing the prospect of diversifying heterogeneous architectures for future experimental phases, and are seeking portable solutions to address them.

Poster session with coffee break / 191

Lamarr: LHCb ultra-fast simulation based on machine learning models

Authors: Adam Davis[1]; Artem Maevskiy[2]; Denis Derkach[2]; Gloria Corti[3]; Lucio Anderlini[4]; Matteo Barbetti[4]; Nikita Kazeev[5]; Sergei Mokhnenko[2]

[1] *University of Manchester (GB)*

[2] *National Research University Higher School of Economics (RU)*

[3] *CERN*

[4] *Universita e INFN, Firenze (IT)*

[5] *HSE University*

Corresponding Author: lucio.anderlini@cern.ch

About 90% of the computing resources available to the LHCb experiment has been spent to produce simulated data samples for Run 2 of the Large Hadron Collider. The upgraded LHCb detector will operate at much-increased luminosity, requiring many more simulated events for the Run 3. Simulation is a key necessity of analysis to interpret data in terms of signal and background and estimate relevant efficiencies. The amount of simulation required will far exceed the pledged resources, requiring an evolution in technologies and techniques to produce simulated data samples. In this conference contribution, we discuss Lamarr, a Gaudi-based framework to speed-up the simulation production parametrizing both the detector response and the reconstruction algorithms of the LHCb experiment.
Deep Generative Models powered by several algorithms and strategies are employed to effectively parameterize the high-level response of the single components of the LHCb detector, encoding

within neural networks the experimental errors and uncertainties introduced in the detection and reconstruction phases. Where possible, models are trained directly on real data, statistically subtracting any background components through the application of weights.
Embedding Lamarr in the general LHCb simulation framework (Gauss) allows to combine its execution with any of the available generators in a seamless way. The resulting software package enables a simulation process completely independent of the detailed simulation used to date.

Significance:

Emerging use of deep generative models in LHC simulation.

References:

Experiment context, if any:

LHCb

Poster session with coffee break / **192**

Development of a lightweight database interface for accessing JUNO conditions and parameters data

Authors: Jiaheng Zou[None]; Qiumei Ma[1]; Tao Lin[2]; Teng LI[3]; Weidong Li[2]; Wenhao Huang[4]; Xiaomei Zhang[2]; Xingtao Huang[4]; Ziyan Deng[None]

[1] *IHEP China*

[2] *Chinese Academy of Sciences (CN)*

[3] *Shandong University, CN*

[4] *Shandong University*

Corresponding Author: tao.lin@cern.ch

The Jiangmen Underground Neutrino Observatory (JUNO) has a very rich physics program which primarily aims to the determination of the neutrino mass ordering and to the precisely measurement of oscillation parameters. It is under construction in South China at a depth of about 700~m underground. As data taking will start in 2023, a complete data processing chain is developed before the data taking. Conditions and parameters data, as non-event data, are one of important parts in the data processing chain, which are used by reconstruction and simulation. These data could be accessed via Frontier on JUNO-DCI (Distributed Computing Infrastructure), or via databases, such as MySQL and SQLite in local clusters.

In this contribution, the latest development of a lightweight database interface (DBI) for JUNO conditions and parameters data management system will be shown. This interface provides a unified method to access data from different backends, such as Frontier, MySQL and SQLite: production jobs could run on JUNO-DCI with Frontier; testing jobs could run in a local cluster with MySQL to validate the conditions and parameters data; fast reconstruction could run in a DAQ environment onsite using SQLite without any connections to remote database. Modern C++ template techniques are used in DBI: extension of a new backend is defined by a simple \texttt{struct} with two methods \texttt{doConnect} and \texttt{doQuery}; result sets are binding to \texttt{std::tuple} and the types of all the elements are known at compile-time. Finally, DBI is used by high-level user interfaces: data models in the database are mapping to normal C++ classes, so that users could access these objects without knowing DBI.

Significance:

In CHEP 2019, our group give an oral on JUNO Conditions Database Management System. In the recent years, we add the support of the parameters data. Therefore, a unified database interface is necessary to support both conditions and parameters data. We also support the multiple database backends, including Frontier, MySQL and SQLite. These backends could be used in difference cases. Even though it is

developed for JUNO, this database interface could be also used by other experiments.

References:

1. Xingtao Huang (JUNO Collaboration), EPJ Web Conf. 245 (2020) 04030

Experiment context, if any:

JUNO

Poster session with coffee break / **193**

Real-time alignment procedure at the LHCb experiment for Run3

Author: Florian Reiss[1]

[1] *University of Manchester (GB)*

Corresponding Author: florian.reiss@cern.ch

The LHCb detector at the LHC is a general purpose detector in the forward region with a focus on studying decays of c- and b-hadrons. For Run 3 of the LHC (data taking from 2022), LHCb will take data at an instantaneous luminosity of 2 × 10^{33} cm−2 s−1, five times higher than in Run 2 (2015-2018). To cope with the harsher data taking conditions, LHCb will deploy a purely software based trigger with a 30 MHz input rate.
The software trigger at LHCb is composed of two stages: in the first stage the selection is based on a fast and simplified event reconstruction, while in the second stage a full event reconstruction is used. This gives room to perform a real-time alignment and calibration after the first trigger stage, which provides an offline-quality detector alignment in the second stage of the trigger. The detector alignment is an essential ingredient to have the best detector performance in the full event reconstruction. The alignment of the whole tracking system of LHCb is evaluated in real-time by an automatic iterative procedure. This is particularly important for the vertex detector, which is retracted for LHC beam injection and centered around the primary vertex position with stable beam conditions in each fill. Hence it is sensitive to position changes on fill-by-fill basis.
The real-time alignment procedure is fully automatic procedure in the online framework that uses a multi-core farm. It is executed as soon as the required data sample is collected. The alignment tasks are split in two parts to allow the parallelization of the event reconstruction via a multi-threads process, while the the evaluation of the alignment parameters is performed on a single thread after collecting all the needed information from all the reconstruction processes in the first part. The execution of the alignment tasks is under the control of the LHCb Experiment Control System, and it is implemented as a finite state machine. The procedure is run at the beginning of each LHC fill and for the alignment of the full tracking system (about 300 elements and about 1000 dofs) takes few minutes. The parameters are updated immediately in the software trigger. This in turn allows to achieve the optimal performance in the trigger output data that can be used for physics analysis without a further offline event reconstruction.
The framework and the procedure for a real-time alignment of the LHCb detector developed for Run 3 data taking are discussed from both the technical and operational point of view. Specific challenges of this procedure and its performance are presented.

Significance:

The real-time alignment and calibration was pioneer in Run2 at LHCb. In Run 3, LHCb runs with a full software trigger, this increases further the importance of this real-time alignment to be used in the trigger without affecting the physics performance.
A new implementation of the real-time alignment was needed to be adapted and optimised to the new computing framework and to the new detectors.
The data taking just started and we plan to present the implementation and the performance of the real-time alignment within novel fully software LHCb trigger in Run 3 for the first time at ACAT.

References:

Experiment context, if any:

LHCb collaboration

Poster session with coffee break / **194**

Integration of machine learning-trained models into JUNO's offline software

Authors: Jiaheng Zou[None]; Tao Lin[1]; Teng LI[2]; Weidong Li[1]; Xingtao Huang[3]

[1] *Chinese Academy of Sciences (CN)*

[2] *Shandong University, CN*

[3] *Shandong University*

Corresponding Author: tao.lin@cern.ch

The Jiangmen Underground Neutrino Observatory (JUNO) is under construction in South China and will start data taking in 2023. It has a central detector with a 20-kt liquid scintillator, equipped with 17,612 20-inch PMTs (photo-multiplier tubes) and 25,600 3-inch PMTs. The requirement on energy resolution of 3\%@1MeV makes the offline data processing challenging, so several machine learning based methods have been developed for reconstruction, particle identification, simulation etc. These methods are implemented with machine learning libraries in Python, however, the offline software is based on a C++ framework called SNiPER. Therefore, how to integrate them and run the inference in offline software is important.

In this contribution, integration of machine learning-trained models into JUNO's offline software will be presented. Three methods are explored: using SNiPER's Python binding to share data between C++ and Python; using native C/C++ APIs of the machine learning libraries, such as TensorFlow and PyTorch; using ONNX runtime. Even though SNiPER is implemented in C++, it provides Python binding via Boost Python. In recent updates of SNiPER, a special data buffer is implemented to share data between C++ and Python, which makes it possible to run machine learning methods in following way: a C++ algorithm reads event data and converts them to \texttt{numpy} arrays; a Python algorithm then accesses these \texttt{numpy} arrays and invokes machine learning libraries in Python; finally, the C++ algorithm puts the results into event data. For the native C/C++ APIs of machine learning libraries and ONNX runtime, a C++ algorithm is used to convert the event data to the corresponding formats and invoke the C/C++ APIs. The deployments of the three methods are also studied: using SNiPER's Python binding is the most flexible method for users, as users could install any Python libraries using \texttt{pip} by themselves; using native C/C++ APIs requires the users to use the same versions in JUNO official software release; using ONNX runtime only requires users to convert their own models to ONNX format. By comparing the three methods, ONNX is recommended for most of users in JUNO. For developing and testing of machine learning-models in offline software, developers could choose the other two methods.

Significance:

As a JUNO L3 manager for software release, I try to deploy the C/C++ API of the native machine learning libraries into JUNO offline in several years ago. But the requirements from developers are different, such as one uses TF 1.x, while another one uses TF 2.x. So that drives our core software group to find some more generic solutions. As mentioned in the abstract, SNiPER provides a data buffer, which could be accessed from both C++ and Python side. This feature lets us do the data conversion in C++ and do the inference in Python. Recently, we also notice that the ONNX is popular, such as Geant4 includes an example using ONNX. So we also apply it in our software.

References:

1. An example to show the data interoperability in C++ and Python: https://github.com/JUNO-collaboration/offline-example-pyalg

Experiment context, if any:

JUNO

Track 3: Computations in Theoretical Physics: Techniques and Methods / **195**

Conditional Normalizing Flow for Markov Chain Monte Carlo Sampling in the Critical Region of Lattice Field Theory

Authors: Dipankar chakraborti[None]; Vipul Arora[None]; ankur singha[None]

Corresponding Author: anksing@iitk.ac.in

In Lattice Field Theory, one of the key drawbacks of the Markov Chain Monte Carlo(MCMC) simulation is the critical slowing down problem. Generative machine learning methods, such as normalizing flows, offer a promising solution to speed up MCMC simulations, especially in the critical region. However, training these models for different parameter values of the lattice theory is inefficient. We address this issue by interpolating or extrapolating the flow model in the critical region. We demonstrate the effectiveness of the proposed method for MCMC sampling in critical regions for multiple parameter values of phi4 scalar theory and U(1) gauge theory in 1+1 dimensions and compare its performance against HMC and flow-based methods.

Significance:

Our flow-based approach for MCMC simulation in the critical region of lattice theory outperforms the existing flow-based method and algorithm like HMC.

References:

http://arxiv.org/abs/2207.00980

Experiment context, if any:

Poster session with coffee break / **196**

Hyperparameter Optimization as a Service on INFN Cloud

Authors: Lucio Anderlini[1]; Matteo Barbetti[1]

[1] *Universita e INFN, Firenze (IT)*

Corresponding Author: matteo.barbetti@cern.ch

The simplest and often most effective way of parallelizing the training of complex Machine Learning models is to execute several training instances on multiple machines, possibly scanning the hyperparameter space to optimize the underlying statistical model and the learning procedure.
Often, such a meta learning procedure is limited by the ability of accessing securely a common database organizing the knowledge of the previous and ongoing trials. Exploiting opportunistic GPUs provided in different environments represents a further challenge when designing such optimization campaigns.
In this contribution we discuss how a set of REST APIs can be used to access a dedicated service based on INFN Cloud to monitor and possibly coordinate multiple training instances, with gradientless optimization techniques, via simple HTTP requests. The service, named *Hopaas* (Hyperparameter OPtimization As A Service), is made of web interface and sets of APIs implemented with a

FastAPI back-end running through Uvicorn and NGINX in a virtual instance of INFN Cloud. The optimization algorithms are currently based on Bayesian techniques as provided by Optuna. A Python front-end is also made available for quick prototyping.
We present applications to hyperparameter optimization campaigns performed combining private, INFN Cloud and CINECA resources.

Significance:

API to distribute hyperparameter optimization campaigns through HTTP requests

References:

Experiment context, if any:

Track 1: Computing Technology for Physics Research / **197**

APEIRON: composing smart TDAQ systems for high energy physics experiments

Authors: Alessandro Lonardo[1]; Andrea Biagioni[2]; Andrea Ciardiello[3]; Francesca Lo Cicero[1]; Francesco Simula[1]; Luca Pontisso[4]; Matteo Turisini[1]; Michele Martinelli[2]; Ottorino Frezza[1]; Paolo Cretaro[2]; Piero Vicini[1]; Roberto Ammendola[5]

[1] *Sapienza Universita e INFN, Roma I (IT)*

[2] *INFN*

[3] *Istituto Nazionale di Fisica Nucleare Sezione di Roma*

[4] *Universita e INFN sezione di Napoli (IT)*

[5] *INFN e Universita Roma Tor Vergata (IT)*

Corresponding Author: alessandro.lonardo@cern.ch

APEIRON is a framework encompassing the general architecture of a distributed heterogeneous processing platform and the corresponding software stack, from the low level device drivers up to the high level programming model.
The framework is designed to be efficiently used for studying, prototyping and deploying smart trigger and data acquisition (TDAQ) systems for high energy physics experiments.
The general architecture of such a distributed processing platform includes *m* data sources, corresponding to the detectors or sub-detectors, feeding a sequence of *n* stream processing layers, making up the whole data path from readout to trigger processor (or storage server).
The processing platform features a modular and scalable low-latency network infrastructure with configurable topology. This network system represents the key element of the architecture, enabling the low-latency recombination of the data streams arriving from the different input channels through the various processing layers.
Developers can define scalable applications using a dataflow programming model (inspired by Kahn Process Networks) that can be efficiently deployed on a multi-FPGAs system: the APEIRON communication IPs allow low-latency communication between processing tasks deployed on FPGAs, even if hosted on different computing nodes.
Thanks to the use of High Level Synthesis tools in the workflow, tasks are described in high level language (C/C++) while communication between tasks is expressed through a lightweight API based on non-blocking *send()* and blocking *receive()* operations.
The mapping between the computational data flow graph and the underlying network of FPGAs is defined by the designer with a configuration tool, by which the framework will produce all project files required for the FPGAs bitstream generation. The interconnection logic is therefore automatically built according to the application needs (in terms of input/output data channels), allowing the designer to focus on the processing tasks expressed in C/C++ .
The aim of the APEIRON project was to develop a flexible framework that could be adopted in the design and implementation of both “traditional” low level trigger systems and of data reduction stages in trigger-less or streaming readout experimental setups characterized by high event rates.
For this purpose we studied and implemented algorithms capable of boosting the efficiency of these

classes of online systems based on Neural Networks (NN), trained offline and leveraging the HLS4ML software package for deployment on FPGA.
We have validated the framework on the physics use case represented by the partial particle identification system for the low-level trigger of the NA62 experiment, working on data from its Ring Imaging Cherenkov detector to pick out electrons and number of charged particles.

Significance:

This is the first presentation of the results of the INFN APEIRON project in a workshop.

References:

Experiment context, if any:

NA62, APEIRON

Track 3: Computations in Theoretical Physics: Techniques and Methods / 198

lips: complex phase space goes singular and p-adic

Author: Giuseppe De Laurentis[1]

[1] *Freiburg University*

Corresponding Author: giuseppe.de.laurentis@physik.uni-freiburg.de

High-multiplicity loop-level amplitude computations involve significant algebraic complexity, which is usually sidestepped by employing numerical routines. Yet, when available, final analytical expressions can display improved numerical stability and reduced evaluation times. It has been shown that significant insights into the analytic structure of the results can be obtained by tailored numerical evaluations. I present new developments on the object-oriented python package lips (Lorentz invariant phase space) for the generation and manipulation of complex massless kinematics. Phase-space points can be defined at the spinor level over complex numbers ($\mathbb{C}$), finite fields ($\mathbb{F}_p$), and p-adic numbers ($\mathbb{Q}_p$). Facilities are also available for the evaluation of arbitrary spinor-helicity expressions in any of these fields. Through the algebraic-geometry submodule, which relies on Singular through the python interface syngular, one can define and manipulate ideals in spinor variables (either covariant components or invariant brackets). These allow to identify irreducible varieties, where amplitudes have well-defined zeros and poles, and to fine-tune numerical phase-space points to be on or close to such varieties. Explicit precision tracking in the p-adic implementation allows one to perform numerical computations in singular configurations while keeping track of the numerical uncertainty as an $\mathcal{O}(p^k)$ term. As an example application, I will show how to infer valid partial-fraction decompositions from p-adic evaluations.

Significance:

1) Finite fields have become a staple of modern amplitude computations, but have their limitations. p-adic numbers, while retaining many features of finite fields, address two of these limitations: they enable non-trivial scale separations and the evaluation of transcendental functions.
2) Likewise, partial-fraction decompositions play an important role in managing the complexity of the analytical expressions. However, these decompositions are generally obtained only after analytical expressions have already been obtained. The proposed approach allows determining the validity of a decomposition purely from numerical evaluations.

References:

The theoretical/mathematical background can be found at:
arXiv:1904.04067, arXiv:2203.04269, arXiv:2203.17170 (Appendix C mainly)

Experiment context, if any:

Poster session with coffee break / **199**

Track reconstruction using quantum algorithms at LUXE

Authors: Lena Funcke[1]; Tobias Hartung[2]; Beate Heinemann[3]; Karl Jansen[4]; Annabel Kropf[5]; Stefan Kühn[6]; Federico Meloni[7]; David Spataro[4]; Cenk Tuysuz[None]; Yee Chinn Yap[7]

[1] *MIT*

[2] *University of Bath and The Cyprus Institute*

[3] *DESY and University of Freiburg (Germany)*

[4] *DESY*

[5] *DESY Hamburg*

[6] *The Cyprus Institute*

[7] *Deutsches Elektronen-Synchrotron (DE)*

Corresponding Author: annabel.kropf@desy.de

LUXE (Laser Und XFEL Experiment) is a proposed experiment at DESY using the electron beam of the European XFEL and a high-intensity laser. LUXE will study Quantum Electrodynamics (QED) in the strong-field regime, where QED becomes non-perturbative. One of the key measurements is the positron rate from electron-positron pair creation, which is enabled by the use of a silicon tracking detector. Precision tracking of positrons becomes very challenging at high laser intensities due to the high rates, which can be computationally expensive for classical computers. The talk will present the latest progress of quantum algorithm-based tracking, which relies on Variational Quantum Eigensolver (VQE) or Quantum Approximate Optimisation Algorithm (QAOA) to reconstruct tracks, and compare the results with classical methods using Graph Neural Networks or a Combinatorial Kalman Filter.

Significance:

This talk will present the first results reconstructing tracks from a full LUXE bunch-crossing (instead of just a test subset) and a set of studies on candidate algorithms (e.g. QUBO splitting) to solve a the quadratic unconstrained binary optimisation problem.

References:

https://indico.cern.ch/event/1103637/contributions/4821835/ (CTD 2022)
https://indico.cern.ch/event/855454/contributions/4597417/ (ACAT 2021)

Experiment context, if any:

LUXE

Poster session with coffee break / **200**

A calibrated particle identification for Belle II

Authors: Connor Hainje[1]; Jan Strube[None]; Marcel Hohmann[None]

[1] *PNNL*

Corresponding Author: mhohmann@student.unimelb.edu.au

The Belle II experiment has been taking data at the SuperKEKB collider since 2018. Particle identification is a key component of the reconstruction, and several detector upgrades from Belle to Belle II were designed to maintain performance with the higher background rates.
We present a method for a data-driven calibration that improves the overall particle identification performance and is resilient against imperfections in the calibration of individual detectors. Our

framework also defines a “blame” metric that identifies the detectors with largest contributions to correctly and incorrectly assigned particle hypotheses.

Significance:

References:

Experiment context, if any:

Belle II

Poster session with coffee break / 201

Accelerating the DBSCAN clustering algorithm for low-latency primary vertex reconstruction

Authors: Alex Tapper[1]; Andrew Rose[None]; Lucas Santiago Borgna[2]; Marco Barbone[None]; Robert John Bainbridge[2]; Wayne Luk[None]

[1] *Imperial College London*

[2] *Imperial College (GB)*

Corresponding Authors: lucas.santiago.borgna@cern.ch, m.barbone19@imperial.ac.uk

In this work we present the adaptation of the popular clustering algorithm DBSCAN to reconstruct the primary vertex (PV) at the hardware trigger level in collisions at the High-Luminosity LHC. Nominally, PV reconstruction is performed by a simple histogram-based algorithm. The main challenge in PV reconstruction is that the particle tracks need to be processed in a low-latency environment $\mathcal{O}(1\ \mu s)$. To achieve this an accelerated version of the DBSCAN algorithm was developed to run in a Field Programmable Gate Array (FPGA). A CPU-optimized version of DBSCAN was implemented in C++ to serve as a benchmark for comparison. The CPU version of DBSCAN resulted in an average PV reconstruction latency of 93 µs, while the FPGA firmware only had a latency of 0.73 µs resulting in a 127x speedup. The speedup is a result of running all the input tracks in parallel, which ultimately results in high resource consumption, of up to 48.6 % of the available logic. Most of the logic was attributed to the use of sorting networks that allows for the parallel processing of the input tracks. To tune the firmware for a specific latency and resource usage constraints, the firmware has been parametrized by the number of input tracks to consider at a time. The accelerated DBSCAN method yielded a higher PV reconstruction efficiency when compared to the simpler histogram-based method. As clustering applications are prominent in High Energy Physics, we modified the accelerated DBSCAN algorithm for higher-dimensional datasets.

Significance:

In general the DBSCAN clustering algorithm is one of the most flexibility and accurate clustering algorithms available. This work demonstrates that DBSCAN can be utilized in a low-latency environment by using FPGA acceleration. The accelerated algorithm was used to reconstruct primary vertices in collisions at the LHC, however it can be generalized to any clustering application.

References:

Experiment context, if any:

CMS

Poster session with coffee break / 202

High performance analysis with RDataFrame and the python ecosystem: Scaling and Interoperability

Authors: Josh Bendavid[1]; Kenneth Long[2]

[1] *CERN*

[2] *Massachusetts Inst. of Technology (US)*

Corresponding Authors: joshbendavid@gmail.com, kenneth.long@cern.ch

The unprecedented volume of data and Monte Carlo simulations at the HL-LHC will pose increasing challenges for data analysis both in terms of computing resource requirements as well as "time to insight". Precision measurements with present LHC data already face many of these challenges today. We will discuss performance scaling and optimization of RDataFrame for complex physics analyses, including interoperability with Eigen, Boost Histograms, and the python ecosystem to enable this.

Significance:

Performance optimizations in this work are critical to enable higher complexity analyses while maintaining fast turnaround time. Identification of issues and bottlenecks are driving important and ongoing improvements in Root, Numba, and other libraries. Progress and impact on performance are being tracked and incorporated into benchmarking and implementation.

References:

https://indico.fnal.gov/event/23628/contributions/237985/attachments/154987/201732/highPerfAnalysis-May11-2022.pdf

Experiment context, if any:

CMS

Track 2: Data Analysis - Algorithms and Tools / 203

First performance measurements with the Analysis Grand Challenge

Authors: Alexander Held[1]; Oksana Shadura[2]

[1] *University of Wisconsin Madison (US)*

[2] *University of Nebraska Lincoln (US)*

Corresponding Author: oksana.shadura@cern.ch

The IRIS-HEP Analysis Grand Challenge (AGC) is designed to be a realistic environment for investigating how analysis methods scale to the demands of the HL-LHC. The analysis task is based on publicly available Open Data and allows for comparing usability and performance of different approaches and implementations. It includes all relevant workflow aspects from data delivery to statistical inference.

The reference implementation for the AGC analysis task is heavily based on tools from the HEP Python ecosystem. It makes use of novel pieces of cyberinfrastructure and modern analysis facilities in order to address the data processing challenges of the HL-LHC.

This contribution compares multiple different analysis implementations and studies their performance. Differences between the implementations include the use of multiple data delivery mechanisms and caching setups for the analysis facilities under investigation.

Significance:

This presentation shows for the first time quantitative performance results for various workflows envisioned at the HL-LHC in a realistic analysis environment. Novel aspects are the evaluation of a full analysis pipeline and the use of an analysis task that captures the relevant workflow aspects and scale physicists require in practice.

References:

ICHEP presentation, targeted at introducing the overall project at a general level: https://agenda.infn.it/event/28874/contri
demonstration of analysis task and several implementations at a dedicated workshop: https://indico.cern.ch/event/112610
Both of these presentations did not target quantitative performance evaluation.

Experiment context, if any:

ATLAS / CMS in particular for the analysis task, to a lesser extent LHCb and other experiments with similar data processing pipelines

Poster session with coffee break / 204

Control of cryogenic dark matter detectors through deep reinforcement learning

Author: Felix Wagner[1]

[1] *HEPHY Vienna*

Corresponding Author: felix.wagner@oeaw.ac.at

Cryogenic phonon detectors are used by direct detection dark matter experiments to achieve sensitivity to light dark matter particle interactions. Such detectors consist of a target crystal equipped with a superconducting thermometer. The temperature of the thermometer and the bias current in its readout circuit need careful optimization to achieve optimal sensitivity of the detector. This task is not trivial and has to be done manually by an expert. In our work, we created a simulation of the detector response as an OpenAI Gym reinforcement learning environment. In the simulation, we test the capability of a soft actor critic agent to perform the task. We accomplish the optimization of a standard detector in the equivalent of 30 minutes of real measurement time, which is faster than most human experts. Our method can improve the scalability of multi-detector setups.

References:

We did not publish our method before. Experiments, that could benefit from our method are e.g. CRESST (https://arxiv.org/abs/1904.00498) and COSINUS (https://link.springer.com/article/10.1007/s10909-020-02464-9).

Experiment context, if any:

CRESST, COSINUS

Significance:

Multiple dark matter experiments plan to scale up the number of detectors used, up to the simultaneous operation of 100 detectors. Our method is valuable for them to avoid the repetitive, manual task of detector optimization.

Track 1: Computing Technology for Physics Research / 205

EJFAT: Towards Intelligent Compute Destination Load Balancing

Author: michael goodrich[None]

Corresponding Author: goodrich@jlab.org

To increase the science rate for high data rates/volumes, JLab is partnering with ESnet for development of an AI/ML directed dynamic Compute Work Load Balancer (CWLB) of UDP streamed data. The CWLB is an FPGA featuring dynamically configurable, low fixed latency, destination switching and high throughput. The CLWB effectively provides seamless integration of edge / core computing to support direct experimental data processing for immediate use by JLab science programs and others such as the EIC as well as data centers of the future. The ESnet/JLaB FPGA Accelerated Transport (EJFAT) project is targeting near future projects requiring high throughput and low latency for both hot and cooled data for both running experiment data acquisition systems and data center use cases.

The essential function of the CWLB data plane is to redirect so designated data channel streams sharing a common data event designation to selectable destination hosts as a function of data event id, and target host ports as a function of data channel id. Thus is effected a form of hierarchical horizontal scaling at two levels; the first across compute host machines data event by data event for a type of pipe-lined processing for a series of events and secondly across ports on a compute host so that different data id channels may be assigned to different processors for parallelized further processing, e.g., reassembly, event reconstruction, physics harvesting, etc.

An EJFAT control plane running external to the CLWB and using both network and compute farm telemetry, effects AI directed and predictive resource allocation, capacity assessment, and scheduling of compute farm resources in order to dynamically reconfigure the CLWB in-situ as the operating context and conditions require.

Significance:

innovation for hierarchical horizontal scaling / load balancing

References:

Experiment context, if any:

Track 3: Computations in Theoretical Physics: Techniques and Methods / 206

Preliminary Lattice Boltzmann Method Simulation using Intel® Quantum SDK

Author: Tejas Shinde[1]

Co-authors: Helena Liebelt [1]; Rui Li [1]

[1] *Deggendorf Institute of Technology*

Corresponding Author: tejas.shinde@stud.th-deg.de

The present work is based on the research within the framework of cooperation between Intel Labs and Deggendorf Institute of Technology, since the Intel® Quantum SDK (Software Development Kit) has recently released. Transport phenomena e.g. heat transfer and mass transfer are nowadays the most challenging unsolved problems in computational physics due to the inherent nature of fluid complexity. As the revolutionary technology, quantum computing opens a grand new perspective for numerical simulation including the computational fluid dynamics (CFD). It is true that the current CFD algorithms based on the different scales (e.g. macroscopic or microscopic) need to be translated into quantum system. In the current work the quantum algorithms have been preliminarily implemented for fluid dynamics using the Intel Quantum SDK, one mesoscopic approach has been applied i.e. to solve the lattice Boltzmann equation. Taking the simplest transport phenomena as a starting point, the preliminary quantum simulation results have been validated with the analytical solution and the classical numerical simulation. The potential of quantum in simulating fluid will be discussed.

Significance:

This is a highly innovative topic, that has not been researched in Europe yet but could have immeasurable impact on a range of subjects from our daily life like meteorology, materials, energy, pharmacology and others.

References:

Experiment context, if any:

Track 1: Computing Technology for Physics Research / **207**

The LHCb simulation software: Gauss and its Gaussino core framework

Authors: Adam Davis[1]; Adam Morris[2]; Gloria Corti[2]; Michal Kreps[3]; Michal Mazurek[2]

[1] *University of Manchester (GB)*

[2] *CERN*

[3] *University of Warwick (GB)*

Corresponding Author: gloria.corti@cern.ch

The LHCb experiment underwent a major upgrade for data taking with higher luminosity in Run 3 of the LHC. New software that exploits modern technologies in the underlying LHCb core software framework, is part of this upgrade. The LHCb simulation framework, Gauss, is adapted accordingly to cope with the increase in the amount of simulated data required for Run 3 analyses. An additional constraint rises from the fact that Gauss also relies on external simulation libraries.
The new version of Gauss, based on a newly-developed, experiment-agnostic core framework where the generic simulation components have been encapsulated, is called Gaussino. This simulation framework allows easier prototyping and testing of new technologies where only the core elements are affected. Gaussino provides a plug&play mechanism for modelling collisions and interfacing generators like Pythia and EvtGen. It relies on Gaudi for general functionalities and the Geant4 toolkit for particle transport, combining their specific multi-threaded approaches. A fast simulation interface to replace the Geant4 physics processes with a palette of fast simulation models for a given sub-detector, including new deep learning based options, is the most recent addition. Geometry layouts can be provided through DD4Hep or experiment-specific software. A new, built-in mechanism to define simple volumes at configuration time can ease the development cycle.
In this contribution, will describe the structure and functionality of Gaussino, as well as its more recent developments and performance. We will also show how the new version of Gauss exploits the Gaussino infrastructure to match the requirements of the simulation(s) of the LHCb experiment.

Significance:

Exposition of a major experiment agnostic simulation framework.

References:

Experiment context, if any:

LHCb

Track 2: Data Analysis - Algorithms and Tools / **208**

The Federation - A novel machine learning technique applied on data from the Higgs Boson Machine Learning Challenge

Author: Maximilian Mucha[1]

Co-author: Eckhard Von Torne [1]

[1] *University of Bonn (DE)*

Corresponding Author: maximilian.mucha@cern.ch

The Federation is a new machine learning technique for handling large amounts of data in a typical high-energy physics analysis. It utilizes Uniform Manifold Approximation and Projection (UMAP) to create an initial low-dimensional representation of a given data set, which is clustered by using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). These clusters can then be used for a federated learning approach, in which we separately train a classifier on the data of each individual cluster. As a requirement for this approach, we need to apply an imbalanced learning method to the data in the found clusters before the training. By using a Dynamic Classifier Selection method, the Federation can then make predictions for the whole data set. As a proof of concept for this novel technique, open data from the Higgs Boson Machine Learning Challenge is used and comparisons to results from established methods will be presented. We also investigated the issue of handling missing values and the jet-count feature for this data.

Significance:

First application of UMAP and imbalanced learning methods in high-energy particle physics.

References:

Experiment context, if any:

Track 1: Computing Technology for Physics Research / **209**

Quality assurance of the LHCb simulation

Authors: Adam Davis[1]; Adam Morris[2]; Dmitry Popov[3]; Gloria Corti[2]; Michal Kreps[4]; Michal Mazurek[2]

[1] *University of Manchester (GB)*

[2] *CERN*

[3] *University of Chinese Academy of Sciences (CN)*

[4] *University of Warwick (GB)*

Corresponding Author: dmitry.popov@cern.ch

Monte Carlo simulation is a vital tool for all physics programmes of particle physics experiments. Their accuracy and reliability in reproducing detector response is of the utmost importance. For the LHCb experiment, which is embarking on a new data-take era with an upgraded detector, a full suite of verifications has been put in place for its simulation software to ensure the quality of the samples produced. The chain of tests exploits the LHCb infrastructure for software quality control.
In this contribution we will describe the procedure and the tests that have been put in place. First-level verifications are performed as soon as new software is submitted for integration in the LHCb GitLab repository. They range from Continous Integration (CI) tests to, so called, 'nightlies': short jobs run overnight to verify the integrity of the software. More in-depth performance and regression tests are carried with dedicated infrastructure (LHCbPR), which compares samples of O(1000) events. Simulation data quality shifters look for anomalies and alert the authors in the case of unexpected changes. Work is also in progress to enable the automatic verification of important variable distributions from a small number of simulated events before the whole production is launched.

Significance:

Announcement of newly developed automatic quality assurance tools

References:

Experiment context, if any:

LHCb

Poster session with coffee break / 210

Equivariant Neural Networks for Particle Physics: PELICAN

Authors: Alexander Bogatskiy[1]; Jan Tuzlic Offermann[2]; Timothy Hoffman[3]; Xiaoyang Liu[3]; David Miller[2]

[1] *Flatiron Institute, Simons Foundation*

[2] *University of Chicago (US)*

[3] *University of Chicago*

Corresponding Author: abogatskiy@flatironinstitute.org

We hold these truths to be self-evident: that all physics problems are created unequal, that they are endowed with their unique data structures and symmetries, that among these are tensor transformation laws, Lorentz symmetry, and permutation equivariance. A lot of attention has been paid to the applications of common machine learning methods in physics experiments and theory. However, much less attention is paid to the methods themselves and their viability as physics modeling tools. One of the most fundamental aspects of modeling physical phenomena is the identification of the symmetries that govern them. Incorporating symmetries into a model can reduce the risk of over-parameterization, and consequently improve a model's robustness and predictive power. As usage of neural networks continues to grow in the field of particle physics, more effort will need to be invested in narrowing the gap between the black-box models of ML and the analytic models of physics.

Building off of previous work, we demonstrate how careful choices in the details of network design – creating a model both simpler and more grounded in physics than the traditional approaches – can yield state-of-the-art performance within the context of problems including jet tagging and particle four-momentum reconstruction. We present the Permutation-Equivariant and Lorentz-Invariant or Covariant Aggregator Network (PELICAN), which is based on three key ideas: symmetry under permutations of particles, Lorentz symmetry, and the ambiguity of the aggregation process in Graph Neural Networks. For the first, we use the most general permutation-equivariant layer acting on rank 2 tensors, which can be viewed as a maximal generalization of Message Passing. For the second, we use classical theorems of Invariants Theory to reduce the 4-vector inputs to a tensor of Lorentz-invariant latent quantities. Finally, the flexibility of the aggregation process commonly used in Graph Networks can be leveraged for improved accuracy, in particular to allow variable scaling with the size of the input.

Significance:

This is one of the first applications of group equivariant neural architectures in particle physics. In particular, the novel permutation-equivariant layer allows for efficient weight-sharing, state-of-the-art performance, and better generalization with lower model complexity. This is also a unique architecture in particle physics that is able to tackle manifestly Lorentz-equivariant tasks such as momentum reconstruction without breaking core physical symmetries. In the setting of classification tasks, the Top Tagging dataset has been widely used as a benchmark for various architectures, and PELICAN is now the state of the art, proving the importance of further research into symmetry-based network design.

References:

[Previous work on this problem] A. Bogatskiy, B. Anderson, J. T. Offermann, M. Roussi, D. W. Miller, and R. Kondor (ICML, 2020) arXiv:2006.04780

Poster on PELICAN presented at SNOWMASS 2022 in Seattle, WA, USA (not yet available online as of this writing).

Experiment context, if any:

Poster session with coffee break / **212**

Scaling MadMiner with a deployment on REANA

Author: Irina Espejo Morales[1]

Co-authors: Kenyi Hurtado ; Kyle Stuart Cranmer [1]; Sinclert Perez [2]

[1] *New York University (US)*

[2] *NYU*

Corresponding Author: irina.espejo.morales@cern.ch

MadMiner is a python module that implements a powerful family of multivariate inference techniques that leverage both matrix element information and machine learning.

This multivariate approach neither requires the reduction of high-dimensional data to summary statistics nor any simplifications to the under-lying physics or detector response.

In this paper, we address some of the challenges arising from deploying MadMiner in a real scale HEP analysis with the goal of offering a new tool in HEP that is easily accessible.

The proposed approach streamlines a typical MadMiner pipeline into a parametrized yadage workflow in yaml files. The general workflow is split in two yadage subworkflows, one dealing with the physics dependencies and the other with the ML ones. After that, the worfklow is deployed using REANA, a reproducible research data analysis platform that takes care of flexibility, scalability, reusability and reproducibility features.

To test the performane of our method, we performed scaling experiments for a MadMiner workflow on the National Energy Research Sscientific Computer luster (NERSC) cluster with an HTCondor backend.
All the stages of the physics subworkfow had a linear dependency between resources & walltime and number of event generated. This trend has allowed us to run a typical MadMiner workflow consiting of 1M events and the generation step just used 2930 MB of memory and walltime of 2919s.

Significance:

Our work is the first effort to scale MadMiner in order to be a reliable and widespread tool in the HEP community. We proved in the past that phyics analyisinvolving a multivariate method are more.

References:

Experiment context, if any:

ATLAS

Poster session with coffee break / **213**

Binning high-dimensional classifier output for HEP analyses through a clustering algorithm

Authors: Martin Erdmann[1]; Niclas Steve Eich[1]; Svenja Diekmann[1]

[1] *Rheinisch Westfaelische Tech. Hoch. (DE)*

Corresponding Author: svenja.diekmann@cern.ch

The usage of Deep Neural Networks (DNNs) as multi-classifiers is widespread in modern HEP analyses. In standard categorisation methods, the high-dimensional output of the DNN is often reduced to a one-dimensional distribution by exclusively passing the information about the highest class score to the statistical inference method. Correlations to other classes are hereby omitted.
Moreover, in common statistical inference tools, the classification values need to be binned, which relies on the researcher's expertise and is often non-trivial. To overcome the challenge of binning multiple dimensions and preserving the correlations of the event-related classification information, we perform K-means clustering on the high-dimensional DNN output to create bins without marginalising any axes.
We evaluate our method in the context of a simulated cross section measurement at the CMS experiment, showing an increased expected sensitivity over the standard binning approach.

Significance:

DNNs have shown to be an indispensable tool in searches for rare processes at the LHC. The sensitivity of an analysis, however, can suffer under the search for an optimal binning and often results in a substantial reduction of the high-dimensional DNN output. This study utilises the full DNN prediction in order to increase the overall sensitivity of a HEP analysis, which would be applicable in a wide range of use cases of multi-classifiers.

References:

Experiment context, if any:

Simulations of CMS Experiment

Track 1: Computing Technology for Physics Research / 214

PHASM: A toolkit for creating AI surrogate models within legacy codebases

Author: Nathan Brei[1]

Co-authors: Xinxin Mei [1]; David Lawrence [1]

[1] *Jefferson Lab*

Corresponding Author: nbrei@jlab.org

PHASM is a software toolkit, currently under development, for creating AI-based surrogate models of scientific code. AI-based surrogate models are widely used for creating fast and inverse simulations. The project anticipates an additional, future use case: adapting legacy code to modern hardware. Data centers are investing in heterogeneous hardware such as GPUs and FPGAs; meanwhile, many important codebases are unable to take advantage of this hardware's superior parallellism without undergoing a costly rewrite. An alternative is to train a neural net surrogate model to mimic the computationally intensive functions in the code, and deploy the surrogate on the exotic hardware instead. PHASM addresses three specific challenges: (1) systematically discovering which functions can be effectively replaced with a surrogate, (2) automatically identifying, for a given function, the true space of inputs and outputs including those not apparent from the type signature, and (3) integrating a machine learning model into a legacy codebase cleanly and with a high level of abstraction. In the first year of development, a proof of concept has been developed for each challenge. A surrogate API makes it easy to bring PyTorch models into the C++ ecosystem and uses profunctor optics to establish a two-way data binding between C++ datatypes and tensors. A model variable discovery tool performs a dynamic binary analysis using Intel PIN in order to identify a target function's model variable space, including types, shapes, and ranges, and generate the optics code necessary to bind the model to the function. Future work may include exploring the limits of surrogate models

for functions of increasing size and complexity, and adaptively generating synthetic training data based on uncertainty estimates.

Significance:

This is the first time presenting this project outside of Jefferson Lab. It already explores several novel technical approaches, specifically (1) using profunctor optics to create a two-way binding between arbitrary C++ datatypes and tensors, and (2) using dynamic binary analysis to discover the space of model variables within an unknown codebase.

References:

No publications yet!

Experiment context, if any:

Not used by any experiments yet!

Poster session with coffee break / **215**

New RooFit Developments on Performance Optimization

Authors: Carsten Burgard[1]; Jonas Rembser[2]; Lorenzo Moneta[2]; Patrick Bos[3]; Wouter Verkerke[3]; Zef Wolffs[3]

[1] *Technische Universitaet Dortmund (DE)*

[2] *CERN*

[3] *Nikhef National institute for subatomic physics (NL)*

Corresponding Author: zwolffs@nikhef.nl

RooFit is a toolkit for statistical modeling and fitting, and together with RooStats it is used for measurements and statistical tests by most experiments in particle physics, particularly the LHC experiments. As the LHC program progresses, physics analyses become more computationally demanding. Therefore, recent RooFit developments were focused on performance optimization, in particular to speed up the minimization of the negative log likelihood when fitting a model to a dataset.

Two such improvements will be discussed in this session: gradient-based CPU parallelization and batched computations. The former strategy parallelizes the calculation of the gradient in the line search approach (MIGRAD) used for minimum likelihood estimation in RooFit. Here, the parallelization approach and computational tools used will be discussed. The second strategy comprises a restructuring of the computational graph associated with a model and dataset in order to allow for batched computations. With batched computations RooFit can evaluate batches of events simultaneously per computational graph node, rather than event by event. This simultaneous computation can be either supported by vectorization or GPU parallelization.

Throughout this session, there will be an emphasis on detailed benchmarking and how it was used to optimize various parts of the developed performance improvements, including load balancing and the reduction of communication overhead. Benchmarks are primarily shown for cutting-edge Higgs combination fits, where the developed improvements were intended to achieve order-of-magnitude improvements in execution wall time.

Significance:

RooFit is a library that is used in a large number of physics analyses, especially those from the LHC experiments. Physicists that use RooFit in any of their analyses can benefit from these new performance improvements, especially in those that are limited by computational runtime. From a computational perspective, the benchmarks and optimizations tailored to cutting-edge Higgs combination fits provide interesting insights into the computational challenges provided by such problems and how they could be tackled.

References:

Experiment context, if any:

Poster session with coffee break / **217**

The Linear Template Fit

Author: Daniel Britzger[1]

[1] *Max-Planck-Institut für Physik München*

Corresponding Author: daniel.britzger@cern.ch

In this contribution the Linear Template Fit is presented, which provides an analytic expressions for a maximum likelihood estimator in a uni- or multi-variate parameter estimation problem. The convenient algorithm and its implementation provides an outstandingly simple and statistically profound methodology for statistical inference that is ideally suited for precision phenomenology. It can also be applied in performance critical applications.
The contribution will discuss the derivation from basic statistical laws and first applications for the determination of the strong coupling constant from jet data, or the top-quark mass from top-quark pair production cross sections.

Significance:

A new analytic expression for parameter estimation is presented, which is already used by the LHC collaborations for precision phenomenology.

References:

arXiv:2112.01548, accepted by EPJ C

Experiment context, if any:

Poster session with coffee break / **218**

RNTuple: Towards First-Class Support for HPC data centers

Authors: Giovanna Lazzari Miotto[1]; Javier Lopez Gomez[2]

[1] *Universidade Federál Do Rio Grande Do Sul (BR)*

[2] *CERN*

Corresponding Author: giovanna.lazzari.miotto@cern.ch

Compared to LHC Run 1 and Run 2, future HEP experiments, e.g. at the HL-LHC, will increase the volume of generated data by an order of magnitude. In order to sustain the expected analysis throughput, ROOT's RNTuple I/O subsystem has been engineered to overcome the bottlenecks of the TTree I/O subsystem, focusing also on a compact data format, asynchronous and parallel requests, and a layered architecture that allows supporting distributed filesystem-less storage systems, e.g. HPC-oriented object stores.
In a previous publication, we introduced and evaluated the RNTuple's native backend for Intel DAOS. Since its first prototype, we carried out a number of improvements both on RNTuple and its DAOS backend aiming to saturate the physical link, such as support for vector writes and an improved RNTuple-to-DAOS mapping, only to name a few. In parallel, the latest developments allow for better integration between RNTuple and ROOT's storage-agnostic, declarative interface to write HEP analyses, RDataFrame.

In this work, we contribute with the following: (i) a redesign and evaluation of the RNTuple DAOS backend, including a mechanism for efficient population of the object store based on existing data; and (ii) an experimental evaluation of single-node and distributed analyses using RDataFrame as a proxy between the user and RNTuple, showing a significant increase in the analysis throughput for typical HEP workflows.

Significance:

Our contribution lies at the intersection between High Energy Physics and High Performance Computing. In this contribution, we provide key updates to RNTuple, the designated successor of the ROOT TTree I/O subsystem. RNTuple comes with a user-friendly API and aims at higher throughput and smaller files. This work describes the latest developments on RNTuple and its integration with RDataFrame, focusing on their use on HPC data centers that leverage Intel DAOS as a distributed object store.

References:

1 https://www.epj-conferences.org/articles/epjconf/abs/2021/05/epjconf_chep2021_02066/epjconf_chep2021_02066.html
2 https://arxiv.org/abs/2204.09043
3 https://www.researchgate.net/publication/346917416_Evolution_of_the_ROOT_Tree_IO

Experiment context, if any:

Poster session with coffee break / **219**

Uncertainty estimation in deep learning based-classifiers of High Energy Physics events using Monte Carlo Dropout

Author: Raquel Pezoa Rivera[1]

Co-authors: Luis Salinas [2]; Sebastián Bórquez [2]; William Brooks [2]; Claudio Torres [2]

[1] *Universidad de Valparaíso*

[2] *Universidad Técnica Federico Santa Maria*

Corresponding Author: raquel.pezoa@cern.ch

The classification of HEP events, or separating *signal* events from the *background*, is one of the most important analysis tasks in High Energy Physics (HEP), and a foundational task in the search for new phenomena. Complex deep learning-based models have been fundamental for achieving accurate and outstanding performance in this classification task. However, the quantification of the uncertainty has traditionally been neglected when deep learning-based methods are used, despite its critical importance in scientific applications 1, 2.

In this work, we propose a Bayesian deep learning-based method for measuring uncertainty when classification of HEP events is performed using a deep neural network classifier. The work is focused on the use of the Monte Carlo Dropout (MC-Dropout) method, a variational inference technique proposed in 3 that is based on Dropout 4, the well-known regularization technique used to overcome overfitting. The Monte Carlo Dropout method allows production of the posterior distribution of the network weights by training a dropout network that approximates Bayesian inference. Thus, a Bayesian deep neural network considers a distribution over network parameters instead of a single point. The traditional dropout method randomly toggles off some neurons, with probability D_{rate} during the training stage. However, the MC-Dropout method toggles off neurons both during the training stage and also during the inference stage.

In this work, we use the publicly available Higgs dataset described in [5]. This is simulated data, and the problem is to distinguish the signal from the background, where the signal corresponds to a Higgs boson decaying to a pair of bottom quarks according to the process: $gg \rightarrow H^0 \rightarrow W^{\mp}H^{\pm} \rightarrow W^{\mp}W^{\pm}h^0 \rightarrow W^{\mp}W^{\pm}b\bar{b}$. Furthermore, we plan to apply the proposed method using simulated data of the ω *meson production off nuclear targets*. Here, the problem is that the ω meson decays

into four final-state particles: $\pi^+ \pi^- \gamma \gamma$, and the pions can also decay into muons and neutrinos, especially at low momentum [6].

The methodology of this work includes (i) training of Bayesian deep learning-based classifiers for the identification of signal and background (binary classification), using the Monte Carlo Dropout method, (ii) evaluate different D_{rate}; (iii) evaluate the classification performance; and (iv) compute three uncertainty measures including variance, mutual information, and predictive entropy. Preliminary results show on average 0.66 accuracy, 0.68 precision, 0.72 recall, and 0.70 F1 score, when a Monte Carlo Dropout model-based is used, with three hidden layers with 300 neurons each, and $D_{rate} = 0.5$. We expect to increase the classification performance using hyper-parameters optimization, evaluating different network architectures, and varying the D_{rate} parameter.

1 Aishik Ghosh, Benjamin Nachman, and Daniel Whiteson. Uncertainty-aware machine learning for high energy physics. Phys. Rev. D, 104:056026, Sep 2021.

2 Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion, 76:243–297, 2021.

3 Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In international conference on machine learning, pages 1050–1059. PMLR, 2016.

4 Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1):1929–1958, 2014.

[5] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. Nature communications, 5(1):1–9, 2014.

[6] Andrés Bórquez. The ω hadronization studies in the nuclear medium with the CLAS spectrometer. Master's thesis, UTFSM, Valparaíso, Chile, 2021.

Significance:

Machine learning-based systems perform an essential role in the analysis of huge volumes of data, in diverse scientific fields, including physics. More precisely, deep learning-based models have demonstrated outstanding results, however, accurate models are as important as the uncertainty quantification of the model prediction, and deep learning-based models only produce a point estimate which does not allow quantifying the uncertainty of each prediction.

Thus, in recent years, there has been increased interest within the scientific community to quantify the uncertainty related to machine learning models increasing the publications with the keywords *machine learning, uncertainty quantification*. In this work, we aim to measure uncertainty when event classification is performed using deep neural networks. We will apply the Monte Carlo Dropout method to measure epistemic uncertainty, including mutual information, variance, and predictive entropy measures, and hence to understand how the model makes decisions, an essential task in the scientific domain.

References:

Experiment context, if any:

Poster session with coffee break / 220

Binned histogram fitting for Bayesian inference via Automatic Differentiation in JuliaLang

Authors: Jerry ⊠ Ling[1]; Lukas Alexander Heinrich[2]; Oliver Schulz[3]

[1] *Harvard University (US)*

[2] *CERN*

[3] *Max Planck Society (DE)*

Corresponding Author: jerry.ling@cern.ch

Template Bayesian inference via Automatic Differentiation in JuliaLang

Binned template-fitting is one of the most important tools in the High-Energy physics (HEP) statistics toolbox. Statistical models based on combinations of histograms are often the last step in a HEP physics analysis. Both model and data can be represented in a standardized format - HistFactory (C++/XML) and more recently pyHF (Python/JSON), have taken advantage of that fact to make template fits both easy and reproducible.

We present a port of pyHF to the Julia programming language much like the way pyHF started out as a port of the C++ HistFactory. The new package, LiteHF.jl, provides an independent, fully compatible implementation of the pyHF JSON specification. Since Julia compiles to native code via LLVM and has a lower function-call overhead than Python, LiteHF.jl can outperform the original pyHF. We utilize Julia's meta-programming capabilities to keep the implementation simple and flexible, and the likelihood gradient is obtained for free via automatic differentiation. LiteHF.jl also makes it easy for the user to add custom template modifiers.

Models generated by LiteHF.jl can be used directly in BAT.jl (Bayesian Analysis Toolkit) in Julia and other Julia inference packages. This enables full Bayesian inference with a few simple commands. BAT.jl provides a full suite of analysis tools including MCMC, nested sampling, automatic re-parametrization, Bayesian evidence calculation, and plotting. A user-friendly likelihoodist inference path for LiteHF.jl is available as well.

Significance:

References:

https://github.com/JuliaHEP/
https://github.com/JuliaHEP/LiteHF.jl
https://github.com/bat/BAT.jl

Experiment context, if any:

Poster session with coffee break / **221**

High Performance Computing Workflow for Liquid Argon Time Projection Chamber Neutrino Experiments

Authors: Allison Reinsvold Hall[1]; Boyana Norris[2]; Giuseppe Cerati[3]; Jim Kowalkowski[1]; Kyle Knoepfel[None]; Marc Paterno[None]; Marianette Wospakrik[4]; Orcun Yildiz[5]; Patrick Gartung[6]; Robert Ross[5]; Saba Sehrish[1]; Sajid Syed[None]; Sophie Berkman[None]; Thomas Peterka[5]; Wesley Ketchum[4]

[1] *Fermilab*

[2] *University of Oregon*

[3] *Fermi National Accelerator Lab. (US)*

[4] *Fermi National Accelerator Laboratory*

[5] *Argonne National Lab*

[6] *Fermilab (US)*

Corresponding Author: sberkman@fnal.gov

Neutrino experiments that use liquid argon time projection chamber (LArTPC) detectors are growing bigger and expect to see more neutrinos with next generation beams, and therefore will require

more computing resources to reach their physics goals of measuring CP violation in the neutrino sector and exploring anomalies. These resources can be used to their full capacity by incorporating parallelism through multi-threading and vectorization within algorithms, and by running these algorithms on High Performance Computers (HPCs). A HPC workflow is being developed for LArTPC experiments to take advantage of all of levels of parallelism, within and across nodes. It will be used to enhance the statistics available for use in physics analysis and will also make it possible to efficiently incorporate AI algorithms. Additional opportunities to incorporate parallelism within LArTPC algorithms is also being explored.

Significance:

References:

Experiment context, if any:

Track 2: Data Analysis - Algorithms and Tools / 222

Constraining Cosmological Parameters from Dark Matter Halo Abundance using Simulation-Based Inference

Authors: Moonzarin Reza[1]; Yuanyuan Zhang[None]

Co-authors: Aleksandra Ćiprijanović ; Brian Nord

[1] *Texas A&M University*

Corresponding Author: moonzarin@tamu.edu

Constraining cosmological parameters, such as the amount of dark matter and dark energy, to high precision requires very large quantities of data. Modern survey experiments like DES, LSST, and JWST, are acquiring these data sets. However, the volumes and complexities of these data – variety, systematics, etc. – show that traditional analysis methods are insufficient to exhaust the information contained in these survey data. Specifically, explicit likelihood-based inference as performed with MCMC likelihood fitting is prone to biases because the likelihoods are written as analytic expressions. This calls for a method that can simultaneously process large volumes of data and handle biases in an efficient manner. Simulation-based inference (SBI or likelihood-free inference) is rapidly gaining popularity for addressing diverse cosmological problems because of its ability to incorporate complex physical processes (statistical fluctuations of cluster properties) and observational effects (non-linear measurement errors) while generating the observables by forward simulations. In this work, we train a normalizing-flow-based machine learning algorithm embedded in the SBI framework on two datasets - generated by analytical forward models (via CosmoSIS) and N-body simulations (Quijote simulations suite). We use number counts and mean masses of dark matter halos to estimate posteriors of multiple cosmological parameters (e.g., Ωm, Ωb, h, ns, σ8). Our results show that the SBI method constrains the cosmological parameters within 2σ, which is comparable to the state-of-the-art MCMC-based inference methods, and results in a smaller bias for some parameters (h and ns) than MCMC. Furthermore, SBI trained on the Quijote simulations data permits a much shorter computational time when dealing with large datasets, compared to MCMC method.

Significance:

References:

Experiment context, if any:

Poster session with coffee break / 223

Cluster counting algorithms for particle identification at future colliders

Authors: Brunella D'Anzi[1]; Federica Cuna[None]; Francesco Grancagnolo[2]; Nicola De Filippis[3]; Walaa Elmetenawee[1]

Co-authors: Alessandro Corvaglia [4]; Alessandro Miccoli [5]; Andrea Ventura [5]; Angela Taliercio [6]; Claudio Caputo [6]; Cosimo Pastore [1]; Edoardo Gorini [5]; Gianluigi Chiarello ; Giovanni F. Tassielli [7]; Kurtis Johnson [8]; Marco Panareo ; Margherita Primavera [5]; Matteo Greco [5]; Maurizio Mongelli [9]

[1] *Universita e INFN, Bari (IT)*

[2] *INFN - Lecce*

[3] *Politecnico/INFN Bari (IT)*

[4] *INFN Lecce*

[5] *INFN Lecce e Universita del Salento (IT)*

[6] *Universite Catholique de Louvain (UCL) (BE)*

[7] *INFN Lecce / Università del Salento*

[8] *Florida State University*

[9] *INFN Bari*

Corresponding Author: brunella.d'anzi@cern.ch

The large statistical fluctuations in the ionization energy loss high energy physics process by charged particles in gaseous detectors implies that many measurements are needed along the particle track to get a precise mean, and this represent a limit to the particle separation capabilities that should be overcome in the design of future colliders. The cluster counting technique (dN/dx) represents a valid alternative which takes advantage of the Poisson nature of the primary ionization process and offers a more statistically robust method to infer mass information. Simulation studies by using Garfield++ and Geant4 prove that the cluster counting allows to reach a resolution two times better than traditional dE/dx method over a wide momentum range in the use-case of a helium-based drift chamber. It consists in singling out, in ever recorded detector signal, the electron peak structures related to the arrival of the electrons belonging to a single primary ionization act (cluster) on the anode wire. However, the search for hundreds of electron peaks and the cluster recognition in real data-driven waveform signals is extremely challenge because of their superimposition in the time scale. The state-of-the-art open-source algorithms fail in finding the expected number even in low-noise conditions. In this talk, we present cutting-edge algorithms to search for electrons peaks and identify ionization clusters in experimental data using the latest available computing tools and physics knowledge. To validate the algorithms and show the advantages of the cluster counting technique, two beam tests has been performed at CERN/H8 facility collecting data with different helium based gas mixtures at different gas gains and angles between the wire direction and the ionizing tracks using a muon beam ranging from 40 GeV/c to 180 GeV/c on a setup made of different size drift tubes, equipped with different diameter sense wires. We show the data analysis results concerning the ascertainment of the Poisson nature of the cluster counting technique, the establishment of the most efficient cluster counting and electrons clustering algorithms among the various ones proposed, and the definition of the limiting effects for a fully efficient cluster counting, like the cluster dimensions, the space charge density around the sense wire and the dependence of the counting efficiency versus the beam particle impact parameter.

Significance:

The most recent test beam in July 2022 gave us the possibility to demonstrate that the number of counted clusters follows the Poisson statistics, as expected, indicating that a particle identification at the 2% level is at reach.

References:

https://agenda.infn.it/event/28874/contributions/169554/
https://agenda.infn.it/event/22092/contributions/166630/

Experiment context, if any:

The beam tests were performed on drift tubes which could be the elementary units of the IDEA drift chamber. The IDEA (Innovative Detector for an Electron-positron Accelerator) general-purpose detector concept has been designed to study electron-positron collisions in a wide energy range provided by

a very large circular leptonic collider.

Poster session with coffee break / 225

Performances studies for a real time HEP data analysis

Author: Dung Hoang[1]

Co-authors: Adriano Di Florio [2]; Alexis Pompili [3]; Umit Sozbilir [3]; Vincenzo Mastrapasqua [3]

[1] *Rhodes College*

[2] *Politecnico e INFN, Bari*

[3] *Universita e INFN, Bari (IT)*

Corresponding Author: hoadh-23@rhodes.edu

In recent years, new technologies and new approaches have been developed in academia and industry to face the necessity to both handle and easily visualize huge amounts of data, the so-called "big data". The increasing volume and complexity of HEP data challenge the HEP community to develop simpler and yet powerful interfaces based on parallel computing on heterogeneous platforms. Good examples are 1) the pandas framework, which is an open source set of data analysis tools allowing the configuration and fast manipulation of data structures, and 2) the Jupyter Notebook, which is a web application that allows users to create and share documents that contain live executable code. Similarly to the python-based pandas, ROOT::RDataFrame offers another parallel data analysis tool also providing a C++ interface as well as Python bindings (thus compatible with the Jupyter Notebook).

In this contribution we aim to document our experience and performance studies in deploying an HEP analysis workflow, in a realtime analysis fashion, being developed within a Jupyter environment (from the selection criteria to extract the physical signal to the fitting tasks). For this purpose we exploit CMS Run1 Open Data to extract the signal associated with the decay of a beauty meson particle.
We will discuss how the combination of HEP specific tools and technologies coming from the much wider data analysis world may result in a powerful and easy-to-use tool for a HEP data analyst. Among these tools we will test the advantage of offloading some of the most compute intensive tasks on heterogeneous architectures through GooFit, a tool that exploits the computational capabilities of GPUs to perform maximum likelihood fits.

Significance:

This contribution provides a performance study of cutting-edge HEP data analysis tools by comparing different approaches to the problem of speeding-up a standard analysis task on an heterogeneous computing platform, thus providing useful advice to the HEP analysts.

References:

Experiment context, if any:

Four out of five co-authors are CMS members; CMS open data are used.

Poster session with coffee break / 226

Implementation of the Cluster Counting and Timing realtime algorithm on FPGA to improve the impact parameter estimates of the Drift Chamber and particle identification.

Authors: Federica Cuna[None]; Francesco Grancagnolo[1]; gianluigi chiarello[None]

Co-authors: Alessandro Crovaglia ; Alessandro Miccoli [2]; Andrea Ventura [2]; Angela Taliercio [3]; BRUNELLA D'ANZI [4]; Claudio Caputo [3]; Cosimo Pastore [5]; Giovanni F. Tassielli [6]; Marco Panareo ; Margherita Primavera [2]; Maurizio Mongelli [7]; Nicola De Fillipis ; Walaa Elmetenawee [5]

[1] *INFN - Lecce*

[2] *INFN Lecce e Universita del Salento (IT)*

[3] *Universite Catholique de Louvain (UCL) (BE)*

[4] *University of Bari Aldo Moro*

[5] *Universita e INFN, Bari (IT)*

[6] *INFN Lecce / Università del Salento*

[7] *INFN Bari*

Corresponding Author: gianluigi.chiarello@le.infn.it

Ultra-low mass and high granularity Drift Chambers fulfill the requirements for tracking systems of modern High Energy Physics experiments at the future high luminosity facilities (FCC-ee or CEPC). \indent We present how, in Helium based gas mixtures, by measuring the arrival times of each individual ionization cluster and by using proper statistical tools, it is possible to perform a bias free estimate of the impact parameter and a precise PID. Typically, in a helium-based drift chamber, consecutive ionization clusters are separated in time by a few ns, at small impact parameters up to a few tens of ns, at large impact parameters. For an efficient application of the cluster timing technique, consisting in isolating pulses due to different ionization cluster, it is, therefore, necessary to have read-out interfaces capable of processing high speed signals. We present a full front-end chain, able to treat the low amplitude sense wire signals (a $\sim$few mV), converted from analog to digital with the use of FADCs, with a high bandwidth ($\sim$1 GHz). The requirement of high sampling frequency, together with long drift times, usually of the order of several hundreds of ns, and large number of readout channels, typically of the order of tens of thousand, impose a sizable data reduction, meanwhile preserving all relevant information. Measuring both the amplitude and the arrival time of each peak in the signal associated to each ionization cluster is the minimum requirement on the data transfer for storage to prevent any significant data loss. An electronic board including a Fast ADC and an FPGA for a real-time processing of the drift chamber signals is presented. Various peak finding algorithms, implemented and tested in real time with VHDL code, are also compared.

Significance:

This project, by immediately digitizing the signals of the drift chamber, respecting the performance requirements, imposes conversions at high sampling rates with high resolution. These constraints, together with maximum drift times and with a large number of readout channels, impose some sizable data reduction, preserving all relevant information. Measuring both the amplitude and the arrival time of each peak in the signal associated to each ionization cluster is the minimum requirement on the data transfer for storage to prevent any data loss.

References:

-The use of FPGA in drift chambers for data transfer rate reduction", Journal of Instrumentation 15 (2020) C09058 , doi:10.1088/1748-0221/15/09/C09058;
-https://agenda.infn.it/event/22092/contributions/166648/

Experiment context, if any:

it will be used in the IDEA drift chamber and it is WP of AidaINNOVA

Poster session with coffee break / **227**

A Checker-Board Sky: Automating Telescope Scheduling with Reinforcement Learning

Authors: Maggie Voetberg[None]; Sophia Zhou[None]

Co-authors: Ben Cohen ; Brian Nord ; Eric Neilsen

Corresponding Authors: sophia.zhou@yale.edu, maggiev@fnal.gov

The size, complexity, and duration of telescope surveys are growing beyond the capacity of traditional methods for scheduling observations. Scheduling algorithms must have the capacity to balance multiple (often competing) observational and scientific goals, address both short-term and long-term considerations, and adapt to rapidly changing stochastic elements (e.g., weather). Reinforcement learning (RL) methods have the potential to significantly automate the scheduling and operation of telescope campaigns and greatly reduce the amount of human effort needed to vet schedules produced via costly simulation work.
In this work, we present the application of an RL-based scheduler, which uses a Markov decision process framework to construct scheduling policies in a way that is scalable, recoverable in the case of interruptions during observation, and computationally efficient for surveys that can include over a hundred observations.
We simulate surveys of objects in the Galactic equator, assuming the location and optics of Stone Edge Observatory. We present schedules generated by our RL technique. While initial results are not comparable to human-tuned schedules, we are encouraged by the technique's scalable, automated approach. We examine how well an RL agent's produced schedules compare to human-designed schedules by comparing different formulations of cumulative reward for these schedules. We also investigate the success of our model as we vary the complexity of the telescope environment and as we vary the reward function. We present this work as a motivation to explore more complex situations and surveys.

Significance:

Showing the potential use of reinforcement learning as a novel technique for automating small-scale telescope surveys.

References:

Experiment context, if any:

Poster session with coffee break / 228

Deep learning based event reconstruction for the HEPD-02 detector on board the China Seismo-Electromagnetic Satellite

Author: Andrea Di Luca[1]

[1] *Universita degli Studi di Trento and INFN (IT)*

Corresponding Author: andrea.di.luca@cern.ch

HEPD-02 is a new, upgraded version of the High Energy Particle Detector as part of a suite of instruments for the second mission of the China Seismo-Electromagnetic Satellite (CSES-02) to be launched in 2023. Designed and realized by the Italian Collaboration LIMADOU of the CSES program, it is optimized to identify fluxes of charged particles (mostly electrons and protons) and determine their energy and incoming direction, providing new measurements of cosmic rays at low energies (up to 200 MeV for protons and up to 100 MeV for electrons). As already experienced in the previous version of the detector, i.e. HEPD-01 on board CSES-01, the reconstruction of the collected events will be performed using a strategy based entirely on deep learning~(DL). This choice is motivated by the fact that deep learning models are very effective when working with particle detectors, in which a variety of electrical signals are produced and may be treated as low-level features. The new HEPD-02 DL-based event reconstruction will be trained on dedicated Monte Carlo simulation and tested on both simulated and test-beam data. Moreover, the collaboration is working on new deep-learning approaches to increase the robustness of the performance assessments, especially when passing from simulated samples to real data, and the interpretability of these algorithms to be used in future analysis.

In this contribution, the entire event reconstruction of the HEPD-02 detector will be described and the performance will be reported.

Significance:

References:

Experiment context, if any:

Contribution submitted on behalf of the CSES-Limadou Collaboration

Track 2: Data Analysis - Algorithms and Tools / **229**

Temporal Variational Autoencoders and Simulation-based inference for interpolation of light curves of Gravitationally Lensed Quasars

Author: Egor Danilov[1]

Co-authors: Aleksandra Ćiprijanović ; Brian Nord

[1] *Fermilab and EPFL*

Corresponding Author: egor.danilov@epfl.ch

The Hubble Tension presents a crisis for the canonical LCDM model of modern cosmology: it may originate in systematics in data processing pipelines or it may come from new physics related to dark matter and dark energy. The aforementioned crisis can be addressed by studies of time-delayed light curves of gravitationally lensed quasars, which have the capacity to constrain the Hubble constant (H_0). A critical task in this analysis is the interpolation of time series with varying duration and irregular time sampling. In this problem, the baseline approach is Gaussian processes (GPs), which have issues in converging on the maximum likelihood.
In this work, we compare the interpolation performance of multiple models: GPs inferred with maximum likelihood optimization, GPs inferred with neural density estimation (NDE), and heteroscedastic temporal neural networks. For the NDE approach, a normalizing flow infers the posteriors of GP's parameters from time series' encodings independent of duration or time sampling. Of the neural networks, we use spline-based convolutional variational autoencoders (VAEs) and multi-time attention VAEs.
We validate our methods on simulations of Gaussian processes, on the observed lensed quasar light curves as well as on real-world datasets that are baselines for irregularly sampled time series interpolation. Our analysis shows that the Gaussian processes inferred with neural density estimators outperform the other approaches in interpolation quality.

Significance:

We introduce a modification to the stochastic model of the quasar light curve that lifts the necessity for numeric convolutions and facilitates maximum likelihood inference.

Moreover, we enhance the inference of quasar parameters using a combination of the latest advancements in temporal generative networks, NLP, and simulation-based inference.

As a result, the work makes a comparative analysis of state of the art stochastic and deep learning approaches in time series interpolation problems.

References:

https://indico.fnal.gov/event/53945/contributions/243362/

Experiment context, if any:

Lensed quasars observations by LSST, DES, COSMOGRAIL

Track 2: Data Analysis - Algorithms and Tools / 230

Galaxy survey data reduction with deep learning

Authors: Laura Cabayol-Garcia[1]; Martin Eriksen[None]

[1] *IFAE*

Corresponding Author: eriksen@pic.es

PAUS is a 40 narrow-band imaging survey using the PAUCam instrument installed at the William Herschel Telescope (WHT). Since the survey started in 2015, this instrument has acquired a unique dataset, performing a relatively deep and wide survey, but with a simultaneous excelled redshift accuracy. The survey is a compromise in performance between deep spectroscopic survey and wide field imaging, showing an order of magnitude better redshift resolution than typical broad band surveys.

The survey data reduction was designed based on classical data reduction techniques. For example the redshift template fitting needed a different algorithm to properly handle the PAUS data (Eriksen 2019). While the data reduction and redshift estimation worked, it had room for improvements. In this talk, we detail the different efforts of replacing steps in the PAUS data reduction with deep learning algorithms. First, deep learning techniques obtain a 50 per.cent reduction in the photo-z scatter for the fainted galaxies. This is achieved through various techniques, including using transfer learning from simulations to handle a small data set.

Furthermore, we have constructed multiple algorithms to improve the data reduction stage. Noise estimation from background estimation from a non-uniform background was handled in BKGNet (Cabayol-Garcia 2019), the galaxy photometry (light measure) was introduced with Lumus (Cabayol-Garcia 2021). Recent work includes the effort of directly estimating the galaxy distance from images. In this talk we also discuss the challenges encountered by differences between the survey fields and recent advances in applying unsupervised denoising techniques.

Significance:

Introduced the techniques for training on simulations for improving the redshift estimations, image background subtraction and flux estimation. This can also be relevant for large broad band surveys, like LSST and Euclid.

References:

- Eriksen 2019, "The PAU Survey: early demonstration of photometric redshift performance in the COSMOS field", MNRAS, Volume 484, Issue 3, April 2019, Pages 4200–4215
- Eriksen 2020, "The PAU Survey: Photometric redshifts using transfer learning from simulations", MNRAS, Volume 497, Issue 4, October 2020, Pages 4565–4579
- Cabayol-Garcia 2020, "The PAU Survey: background light estimation with deep learning techniques",
  MNRAS, Volume 491, Issue 4, February 2020, Pages 5392–5405,
- Cabayol-Garcia 2021, "The PAU survey: estimating galaxy photometry with deep learning", MNRAS, Volume 506, Issue 3, September 2021, Pages 4048–4069

Experiment context, if any:

The PAU Survey (PAUS)

Track 1: Computing Technology for Physics Research / 231

Implementing Machine Learning inference on FPGAs: from software to hardware using hls4ml

Author: Marco Lorusso[1]

Co-authors: Daniele Bonacorsi [2]; Riccardo Travaglini [3]

[1] *Universita e INFN, Bologna (IT)*

[2] *University of Bologna / INFN*

[3] *INFN, Bologna (IT)*

Corresponding Author: marco.lorusso@bo.infn.it

In the past few years, using Machine and Deep Learning techniques has become more and more viable, thanks to the availability of tools which allow people without specific knowledge in the realm of data science and complex networks to build AIs for a variety of research fields. This process has encouraged the adoption of such techniques: in the context of High Energy Physics, new algorithms based on ML are being tested for event selection in trigger operations, end-user physics analysis, computing metadata based optimizations, and more. Time critical applications can benefit from implementing algorithms on low-latency hardware like specifically designed ASICs and programmable micro-electronics devices known as FPGAs. The latter offers a unique blend of the benefits of both hardware and software. Indeed, they implement circuits just like hardware, providing power, area and performance benefits over software, yet they can be reprogrammed cheaply and easily to implement a wide range of tasks, at the expense of performance with respect to ASICs.

In order to facilitate the translation of ML models to fit in the usual workflow for programming FPGAs, a variety of tools have been developed. One example is the HLS4ML toolkit, developed by the HEP community, which allows the translation of Neural Networks built using tools like TensorFlow to a High-Level Synthesis description (e.g. C++) in order to implement this kind of ML algorithms on FPGAs.

This paper presents and discusses the activity started at the Physics and Astronomy department of University of Bologna and INFN-Bologna devoted to preliminary studies for the trigger systems of the Compact Muon Solenoid (CMS) experiment at the CERN LHC accelerator. A broader-purpose open-source project from Xilinx (a major FPGA producer) called PYNQ is being tested combined with the HLS4ML toolkit. The PYNQ purpose is to grant designers the possibility to exploit the benefits of programmable logic and microprocessors using the Python language. This software environment can be deployed on a variety of Xilinx platforms, from IOT devices like the ZYNQ-Z1 board, to the high performance ones, like Alveo accelerator cards and on the cloud AWS EC2 F1 instances.

Even though a rich documentation can be found on how to use hls4ml, a comprehensive description of the entire workflow from Python to FPGA is still hard to find. This work tries to fill this gap, presenting hardware and software set-up, together with performance tests on various baseline models used as benchmarks. The presence or not of some overhead causing an increase in latency will be investigated. Eventually, the consistency in the predictions of the NN, with respect to a more traditional way of interacting with the FPGA using C++ code, will be verified.

Significance:

This talk would present, through examples and actual lines of code, for the first time the entire workflow needed to go from a purely software Neural Network in Python to the hardware implementation on a generic FPGA, together with the possibility of using PYNQ to run the inference on compatible boards.

References:

https://pos.sissa.it/378/005/
https://indico4.twgrid.org/event/20/contributions/1119/attachments/672/775/ISGC2022_slides.pdf
https://agenda.infn.it/event/28874/contributions/169219/attachments/94335/129059/ICHEP2022_slides_Lorusso.pdf

Experiment context, if any:

Compact Muon Solenoid at CERN

Poster session with coffee break / **232**

The TICL reconstruction at the CMS Phase-2 High Granularity Calorimeter Endcap

Author: Felice Pantaleo[1]

[1] *CERN*

Corresponding Author: felice.pantaleo@cern.ch

To sustain the harsher conditions of the high-luminosity LHC, the CMS Collaboration is designing a novel endcap calorimeter system. The new calorimeter will predominantly use silicon sensors to achieve sufficient radiation tolerance and will maintain highly granular information in the readout to help mitigate the effects of the pile up. In regions characterized by lower radiation levels, small scintillator tiles with individual SiPM on-tile readout are employed. A unique reconstruction framework (TICL: The Iterative CLustering) is being developed within the CMS Software CMSSW to fully exploit the granularity and other significant detector features, such as particle identification and precision timing, with a view to mitigating pile up in the very dense environment of HL-LHC. The TICL framework has been thought of with heterogeneous computing in mind: the algorithms and their data structures are designed to be executed on GPUs. In addition, geometry agnostic data structures have been designed to provide fast navigation and searching capabilities. Seeding capabilities (also exploiting information coming from other detectors), dynamic cluster masking, energy calibration, and particle identification are the main components of the framework. To allow for maximal flexibility, TICL allows the composition of different combinations of modules that can be chained together in an iterative fashion. The presenter will describe the design of TICL pattern recognition algorithms and advanced neural networks under development, as well as future plans.

Significance:

TICL is the CMS Phase-2 reconstruction framework for the CMS HGCAL.
In 2022, its version 4 is coming out with new pattern recognition algorithms like CLUE3D, FastJet and new Graph Neural Networks

References:

Experiment context, if any:

CMS phase 2 upgrade

Track 1: Computing Technology for Physics Research / **233**

Efficient and Accurate Automatic Python Bindings with Cppyy and Cling

Authors: Baidyanath Kundu[1]; Vassil Vasilev[1]; Wim Lavrijsen[2]

[1] *Princeton University (US)*

[2] *Lawrence Berkeley National Lab. (US)*

Corresponding Author: baidyanath.kundu@cern.ch

The simplicity of Python and the power of C++ provide a hard choice for a scientific software stack. There have been multiple developments to mitigate the hard language boundaries by implementing language bindings. The static nature of C++ and the dynamic nature of Python are problematic for bindings provided by library authors and in particular features such as template instantiations with user-defined types or more advanced memory management.

The development of the C++ interpreter Cling has changed the way we can think of language bindings as it provides an incremental compilation infrastructure available at runtime. That is, Python can interrogate C++ on demand and fetch only the necessary information. This way of automatic binding provision requires no binding support by the library authors and offers better performance than Pybind11. This approach pioneered in ROOT with PyROOT and later was enhanced with its successor Cppyy. However, until now, Cppyy relied on the reflection layer of ROOT which is limited in terms of provided features and performance.

In this talk we show how basing Cppyy purely on Cling yields better correctness, performance and installation simplicity. We illustrate more advanced language interoperability of Numba-accelerated Python code capable of calling C++ functionality via Cppyy. We outline a path forward for integrating the reflection layer in LLVM upstream which will contribute to the project sustainability and will foster greater user adoption. We demonstrate usage of Cppyy through Cling's LLVM mainline version Clang-Repl.

Significance:

The automatic Python bindings based on the C++ interpretation is unique, cutting-edge technology worldwide. The Cppyy project is a cornerstone for the ROOT user base and bridges the C++ and Python ecosystems and analyses. The Cppyy project has users beyond HEP. It outperforms Pybind11 and is a key motivator for starting the nanobind project by the Pybind11 community. In this talk we talk about how to bring the technology on the next level in terms of performance, robustness and sustainability. This will have a direct effect on HEP and its community should know about the project plans.

References:

Experiment context, if any:

Poster session with coffee break / **234**

ROOT Machine Learning Ecosystem for Data Analysis

Author: Lorenzo Moneta[1]

Co-authors: Sitong An [2]; Omar Andres Zapata Mesa [3]; Sanjiban Sengupta ; Ahmat Hamdan

[1] *CERN*

[2] *CERN, Carnegie Mellon University (US)*

[3] *University of Antioquia & Metropolitan Institute of Technology*

Corresponding Author: lorenzo.moneta@cern.ch

Through its TMVA package, ROOT provides and connects to machine learning tools for data analysis at HEP experiments and beyond. In addition, ROOT provides through its powerful I/O system and RDataFrame analysis tools the capability to efficiently select and query input data from large data sets as typically used in HEP analysis. At the same time, several existing Machine Learning tools exist in a diversified landscape outside of ROOT.

In this talk, we present new developments in ROOT that bridge the gap between external tools and ROOT, by providing better interoperability in a common software ecosystem for Machine Learning in data analysis.
We present recently included features in TMVA allowing for generating batches of events for ROOT I/O and RDataFrame to train efficiently machine learning models using Python tools such as Tensorflow and PyTorch. This will facilitate direct access to the ROOT input data when training using external tools. Another focus is put on fast machine learning inference, which enables analysts to deploy their machine learning models rapidly on large scale datasets. A new tool has been recently developed in ROOT, SOFIE, allowing for generating C++ code for evaluation of deep learning models, which are trained from external tools. This provides the capability to better integrate Machine Learning model evaluation in HEP data analysis.
The new developments are paired with newly designed C++ and Python interfaces for TMVA supporting modern C++ paradigms and providing full interoperability in the Python ecosystem.

Significance:

This presentation covers some novel results, teh development of a batch generator for better integration of ROOT RDataFrame with external Machine tools for training models.
Furthermore it will contain an update on other TMVA developments such as SOFIE which has greatly developed since the last ACAT presentation, being able to parse complex ML models used by LHC experiments.

References:

Experiment context, if any:

Poster session with coffee break / **235**

Quantum anomaly detection in the latent spaces of high energy physics events

Authors: Vasilis Belis[1]; Guenther Dissertori[1]; Kinga Anna Wozniak[2]; Ema Puljak[3]; Michele Grossi[4]; Sofia Vallecorsa[4]; Maurizio Pierini[4]

[1] *ETH Zurich (CH)*

[2] *University of Vienna*

[3] *Universitat Autonoma de Barcelona (ES)*

[4] *CERN*

Corresponding Author: vasileios.belis@cern.ch

We developed supervised and unsupervised quantum machine learning models for anomaly detection tasks at the Large Hadron Collider at CERN. Current Noisy Intermediate Scale Quantum (NISQ) devices have a limited number of qubits and qubit coherence. We designed dimensionality reduction models based on Autoencoders to accommodate the constraints dictated by the quantum hardware. Different designs were investigated, such as convolutional and Sinkhorn Autoencoder architectures, that can compress HEP data while preserving the class structure of the original dataset. The quantum algorithms are trained to identify anomalies in the latent spaces generated by the Autoencoders. A collection of results for a quantum classifier and a set of quantum anomaly detection algorithms is presented. Our study is supported by a performance comparison to the corresponding classical models.

Significance:

In our work, as well as in other studies addressing classification tasks in HEP, no significant difference in performance between quantum and classical ML models has been observed. Classical ML and deep learning approaches typically outperform quantum algorithms when one allows the size of the training dataset or the number of model parameters to increase beyond the limits of both current quantum hardware and quantum simulation software on classical devices. We believe that these results can stimulate fundamental research towards quantum machine learning or hybrid quantum-classical algorithm

design that would manifest interesting behaviour that cannot be replicated by classical models.

References:

Experiment context, if any:

CMS

Poster session with coffee break / **236**

Ceph S3 Object Storage for CMS data

Authors: Bo Jayatilaka[1]; David Alexander Mason[1]; Nick Smith[1]; Oliver Gutsche[1]

[1] *Fermi National Accelerator Lab. (US)*

Corresponding Author: nick.smith@cern.ch

To support the needs of novel collider analyses such as long-lived particle searches, considerable computing resources are spent forward-copying data products from low-level data tiers like CMS AOD and MiniAOD to reduced data formats for end-user analysis tasks. In the HL-LHC era, it will be increasingly difficult to ensure online access to low-level data formats. In this talk, we present a novel online data storage mechanism that obviates the need for data tiers by storing individual data products in column objects using RadosGW, a Ceph object store technology. Benchmarks of the performance of storage and retrieval of the event data through the S3 protocol for a prototype of typical analysis workflows will be presented, and compared with traditional xrootd ROOT file access protocols.

Significance:

The use of Ceph object stores and S3 protocol to access experiment data is novel within HEP. Our experience will help guide evaluation and possible adoption of these technologies.

References:

https://indico.cern.ch/event/1125222/timetable/?view=standard#32-object-store-rd
https://uscms-software-and-computing.github.io/postdocs/nsmith-.html

Experiment context, if any:

CMS

Poster session with coffee break / **237**

Federated Learning Strategies of Generative Adversarial Networks for High Energy Physics Calorimeter Simulation

Author: Mohamed Hemdan[None]

Co-authors: Jose Cabrero Holgueras [1]; Sofia Vallecorsa [2]

[1] *University Carlos III (ES)*

[2] *CERN*

Corresponding Author: mohamed.hemdan@cern.ch

Particle physics experiments spend large amounts of computational effort on Monte Carlo simulations. Due to the computational expense of simulations, they are often executed and stored in large distributed computing clusters. To lessen the computational cost, physicists have introduced alternatives to speed up the simulation. Generative Adversarial Networks (GANs) are an excellent Deep-Learning-based alternative due to their ability to imitate probability distributions. Concretely, one of the more tackled problems is calorimeter simulations since they involve a large portion of the computing power. GANs simulate calorimeter particle showers with good accuracy and reduced computational resources. Previous works have already explored the generation of calorimeter simulation data with GANs, but in most cases as a centralized perspective (i.e., where the dataset is present on the training node).
This separation creates a disparity between the training data generation (i.e., in distributed clusters) and training (i.e., centralized), introducing a limiting factor to the amount of data the centralized node can use to train. Federated Learning has arisen as a successful decentralized training solution where data is non-necessarily balanced, independent, and identically distributed (IID). Federated Learning is a training method where a group of \textit{collaborators} trains a model by sharing training updates with an \textit{aggregator}. The sparsity and distributed nature of the simulated data pairs favorably with the features of Federated Learning. In this work, we introduce new federated learning-based approaches for GAN training and test them on the 2DGAN model*. This work covers different training schemes for GANs with FL (e.g., centralized discriminator or centralized generator). Our work provides insights into the various architectures by performing model training and extracting performance metrics. The results permit the evaluation of the effectiveness of the different strategies.

- Rehm, F., Vallecorsa, S., Borras, K., & Krücker, D. (2021). Validation of Deep Convolutional Generative Adversarial Networks for High Energy Physics Calorimeter Simulations. doi:10.48550/ARXIV.2103.13698

Significance:

Monte-Carlo Simulations are classically used in High-Energy Physics to simulate particle interactions in detectors. These simulations are often performed in distributed computing grids, creating an inherent dispersion of resulting data. Generative Adversarial Networks have emerged as a potential solution to these expensive simulations. However, they still enforce the centralization of data. Federated Learning is a novel decentralized deep learning model training approach. We combine Federated Learning with GANs for calorimeter simulation. We aim to provide different data architectures where data training is not uniformly distributed or balanced. Furthermore, we want to test diverse network configurations to understand the advantages, disadvantages, and differences between the classical centralized approaches and federated learning.

References:

Experiment context, if any:

Poster session with coffee break / 238

Fast analysis facility for HEP experiments

Author: Gabor Biro[1]

Co-author: Gergely Gabor Barnafoldi [1]

[1] *Wigner Research Centre for Physics (Wigner RCP) (HU)*

Corresponding Author: gabor.biro@cern.ch

The ever growing increase of computing power necessary for the storage and data analysis of the high-energy physics experiments at CERN requires performance optimization of the existing and planned IT resources.

One of the main computing capacity consumers in the HEP software workflow is the data analysis. To optimize the resource usage, the concept of Analysis Facility (AF) for Run 3 has been introduced. The AFs are special computing centres with a combination of CPU and fast interconnected disk storage resources, allowing for rapid turnaround of analysis tasks on a subset of data. This in turn allows for optimization of the analysis process and the codes before the analysis is performed on the large data samples on the WLCG Grid.

In this paper, the structure and the first benchmark tests of the Wigner AF are presented.

Significance:

References:

Experiment context, if any:

Track 3: Computations in Theoretical Physics: Techniques and Methods / **239**

Studying Hadronization by Machine Learning Techniques

Author: Gabor Biro[1]

Co-authors: Bence Tankó-Bartalis [2]; Gergely Gabor Barnafoldi [1]

[1] *Wigner Research Centre for Physics (Wigner RCP) (HU)*

[2] *Wigner RCP*

Corresponding Author: gabor.biro@cern.ch

Hadronization is a non-perturbative process, which theoretical description can not be deduced from first principles. Modeling hadron formation requires several assumptions and various phenomenological approaches. Utilizing state-of-the-art Computer Vision and Deep Learning algorithms, it is eventually possible to train neural networks to learn non-linear and non-perturbative features of the physical processes.

Here, I would like to present the latest results of two deep neural networks, by investigating global and kinematical quantities, indeed jet- and event-shape variables. The widely used Lund string fragmentation model is applied as a baseline in $\sqrt{s}$=7 TeV proton-proton collisions to predict the most relevant observables at further LHC energies. Non-liear QCD scaling properties were also identified and validated by experimental data.

1 G. Bíró, B. Tankó-Bartalis, G.G. Barnaföldi; arXiv:2111.15655

Significance:

References:

https://arxiv.org/abs/2111.15655
https://agenda.infn.it/event/28874/contributions/170292/
https://indico.wigner.hu/event/1393/contributions/3160/
https://indico.cern.ch/event/1097820/contributions/4623938/

Experiment context, if any:

Poster session with coffee break / **240**

Law: End-to-End Analysis Automation over Distributed Resources

Author: Marcel Rieger[1]

[1] *Hamburg University (DE)*

Corresponding Author: marcel.rieger@cern.ch

In particle physics, workflow management systems are primarily used as tailored solutions in dedicated areas such as Monte Carlo production. However, physicists performing data analyses are usually required to steer their individual, complex workflows manually, frequently involving job submission in several stages and interaction with distributed storage systems by hand. This process is not only time-consuming and error-prone, but also leads to undocumented relations between particular workloads, rendering the steering of an analysis a serious challenge.
This contribution presents the Luigi Analysis Workflow (law) Python package which is based on the open-source pipelining tool luigi, originally developed by Spotify. It establishes a generic design pattern for analyses of arbitrary scale and complexity, and shifts the focus from executing to defining the analysis logic. Law provides the building blocks to seamlessly integrate with interchangeable remote resources without, however, limiting itself to a specific choice of infrastructure.
In particular, it introduces the concept of complete separation between analysis algorithms on the one hand, and run locations, storage locations, and software environments on the other hand. To cope with the sophisticated demands of end-to-end HEP analyses, law supports job execution on WLCG infrastructure (ARC, gLite) as well as on local computing clusters (HTCondor, Slurm, LSF), remote file access via various protocols using the Grid File Access Library (GFAL2), and an environment sandboxing mechanism with support for sub-shells and virtual environments, as well as Docker and Singularity containers. Moreover, the novel approach ultimately aims for analysis preservation out-of-the-box.
Law is developed open-source and independent of any experiment or the language of executed code. Over the past years, its user-base increased steadily with applications now ranging from (pre-)processing workflows in CMS physics objects groups, to pipelines performing the statistical inference in most CMS di-Higgs searches, and it serves as the underlying core software for large scale physics analyses across various research groups.

Significance:

I've presented this topic in an earlier stage at ACAT 2019, but over the past years the user-base multiplied, making it the most widely used workflow management tool (apart from central production pipelines) at CMS. Law will be used for many Run 3 analyses, and apart from a gentle introduction into the key concepts and main features, I intend to show a demonstrator for a full-blown and not necessarily CMS-related physics analysis with law.

References:

Experiment context, if any:

I am with CMS. The presented software is experiment-independent.

Track 2: Data Analysis - Algorithms and Tools / 241

Hybrid Quantum-Classical Networks for Reconstruction and Classification of Earth Observation Images

Authors: Su Yeon Chang[1]; Bertrand Le Saux[2]; Michele Grossi[3]; Sofia Vallecorsa[3]

[1] *EPFL - Ecole Polytechnique Federale Lausanne (CH)*

[2] *ESA*

[3] *CERN*

Corresponding Author: su.yeon.chang@cern.ch

Earth Observation (EO) has experienced promising progress in the modern era via an impressive amount of research on establishing a state-of-the-art Machine Learning (ML) technique to learn a large dataset. Meanwhile, the scientific community has also extended the boundary of ML to the quantum system and exploited a new research area, so-called Quantum Machine Learning (QML), to integrate advantages from both ML and Quantum Computing (QC). Recent papers investigated the application of QML in the EO domain mainly based on Parameterized Quantum Circuits (PQCs), which are regarded as suitable architecture for quantum neural networks (QNNs) due to their potential to be efficiently simulated on near-term quantum hardware. But more contributions are still required in-depth, and various challenges should be tackled, such as large EO image size for the current quantum simulators, trainability of the quantum circuit, etc.
This work introduces a hybrid Quantum-Classical model performing reconstruction and classification simultaneously and explores its application for EO image multi-class classification. Moreover, we investigate for the first time the correlation between different PQC descriptors and the training results in the realistic EO use case. The results demonstrate that the hybrid model successfully achieves up to 10 class classification suggesting a potential usage of QNNs for a realistic context, and also hint at generic approaches for choosing the suitable PQC architecture for a given problem.

Significance:

References:

Experiment context, if any:

Track 2: Data Analysis - Algorithms and Tools / 242

Pruning and resizing deep neural networks for FPGA implementation in trigger systems at collider experiments

Authors: Andrea Di Luca[1]; Daniela Mascione[1]; Francesco Maria Follega[1]; Marco Cristoforetti[2]; Roberto Iuppa[1]

[1] *Universita degli Studi di Trento and INFN (IT)*

[2] *Universita degli Studi di Trento e INFN (IT)*

Corresponding Author: daniela.mascione@cern.ch

Deep Learning algorithms are widely used among the experimental high energy physics communities and have proved to be extremely useful in addressing a variety of tasks. One field of application for which Deep Neural Networks can give a significant improvement is event selection at trigger level in collider experiments. In particular, trigger systems benefit from the implementation of Deep Learning models on FPGAs. However, this task poses specific challenges to Deep Learning algorithm design, due to the microsecond latency requirements and limited resources of FPGA-based trigger systems. Before being implemented on an FPGA, Neural Networks may need to be appropriately compressed in order to reduce the number of neurons and synapses. A widespread technique to reduce the size of Deep Neural Networks is pruning. Numerous approaches have been developed to create a pruned model from an untrained one. Nearly all of them use a similar procedure, according to which the network is first trained to convergence, then single weights are removed on the basis of a particular ranking. To recover from accuracy loss, pruned networks are finally retrained. The pruning and retraining process is repeated iteratively, shrinking the network's size. This procedure however can be quite long and resource demanding. Moreover, the relative importance of parameters changes along iterations and this may lead to converging to sub-optimal configurations.
Here we propose a different pruning strategy, which proved to be a mathematically rigorous and faster method for optimizing Neural Networks under size constraints. Our approach works by overlaying a shadow network on the one that has to be optimized. The shadow network is very simple to incorporate into already developed Deep Neural Networks and can be used to prune the whole network or just a portion. Through the training process, the combined optimization of the shadow and standard networks takes place. As a result, the pruning procedure occurs along with the training, and not in two different phases. The proposed method performs a pruning of the nodes, rather

than of the single connections, allowing for a determination of an ideal network layout, with the number of total nodes determined by the user so to match the FPGA resources available. After finding the optimal network layout, the reduced network can be retrained as a new independent model. Preliminary results will be presented, along with new developments and applications.

Significance:

Here we are presenting a novel pruning strategy for compressing Deep Neural Networks for FPGA implementation. Our method is mathematically sound and time-saving with respect to standard pruning strategies. It allows to prune during the training stage, and to prune nodes rather than single connections. It provides network layouts as effective as the optimal one.

References:

Experiment context, if any:

The presenter and the research team are members of the ATLAS collaboration, actively working on jet flavor-tagging.

Track 2: Data Analysis - Algorithms and Tools / **243**

Automated Lens Parameter Estimation using Simulation-Based Inference

Author: Jason Poh[1]

Co-authors: Ashwin Samudre ; Aleksandra Ćiprijanović ; Brian Nord [2]

[1] *University of Chicago*

[2] *Fermi National Accelerator Laboratory*

Corresponding Author: jasonpoh@uchicago.edu

Modern cosmology surveys are producing data at rates that are soon to surpass our capacity for exhaustive analysis – in particular for the case of strong gravitational lenses. While the Dark Energy Survey may discover thousands of galaxy-scale strong lenses, the upcoming Legacy Survey of Space and Time (LSST) will find hundreds of thousands more. These large numbers of objects will make strong lensing a highly competitive and complementary cosmic probe of dark energy and dark matter. Unfortunately, the traditional analysis of a single lens is highly computationally expensive, requiring up to a day of human-intensive work. Being able to accurately estimate the lens parameters of a large sample of lenses will enable us to study the dark matter distribution across populations of lenses, as well as potentially constrain dark energy models. To leverage the increased statistical power from these surveys, we will need highly automated lens analysis techniques.

We present work in which we automate and accelerate parameter estimation of galaxy-galaxy lenses using Simulation-Based Inference (SBI). In particular, we demonstrate the successful application of Neural Posterior Estimators (NPE), based on masked autoregressive flows, to efficiently infer a 5-parameter lens mass model. We also present preliminary results on a 12-parameter model including lens mass, source light and external shear. We compare our NPE constraints to a Bayesian Neural Network (BNN) and find that it outperforms the BNN, often producing posterior distributions that are both more accurate and more precise. In some scenarios, the NPE method predicts constraints on lens parameters that are several times smaller than that from the BNN.

Significance:

Neural posterior estimators in simulation-based inference methods are still quite new in many fields of astronomy and we present a novel application of it to astronomical data.

References:

Presented at Fermilab's New Perspective conference: https://indico.fnal.gov/event/53945/contributions/243359/

Experiment context, if any:

We used simulated Dark Energy Survey data, but did not use any proprietary data from that experiment..

Poster session with coffee break / **244**

Ultra-low latency recurrent neural network inference on FPGAs for physics applications with hls4ml

Authors: Aaron Wang[None]; Caterina Vernieri[1]; Chaitanya Paikara[2]; Dylan Sheldon Rankin[3]; Elham E Khoda[4]; Michael Aaron Kagan[1]; Philip Coleman Harris[3]; Rafael Teixeira De Lima[1]; Richa Rao[2]; Scott Hauck[None]; Shih-Chieh Hsu[5]; Sioni Paris Summers[6]; Vladimir Loncar[6]

[1] *SLAC National Accelerator Laboratory (US)*

[2] *University of Washington*

[3] *Massachusetts Inst. of Technology (US)*

[4] *University of Washington (US)*

[5] *University of Washington Seattle (US)*

[6] *CERN*

Corresponding Author: elham.e.khoda@cern.ch

Recurrent neural networks have been shown to be effective architectures for many tasks in high energy physics, and thus have been widely adopted. Their use in low-latency environments has, however, been limited as a result of the difficulties of implementing recurrent architectures on field-programmable gate arrays (FPGAs). In this paper we present an implementation of two types of recurrent neural network layers- long short-term memory and gated recurrent unit- within the hls4ml 1 framework. We demonstrate that our implementation is capable of producing effective designs for both small and large models, and can be customized to meet specific design requirements for inference latencies and FPGA resources. We show the performance and synthesized designs for multiple neural networks, many of which are trained specifically for jet identification tasks at the CERN Large Hadron Collider.

1 J. Duarte et al., "Fast inference of deep neural networks in FPGAs for particle physics", JINST 13 (2018) P07027, arXiv:1804.06913

Significance:

RNNs have shown substantial success for many tasks in particle physics. They are particularly well-suited to those problems involving sequences of particle or detector signals, outperforming densely connected deep neural networks (DNNs) and convolution neural networks (CNNs) on certain jet classification tasks. In spite of this success, RNNs have not seen widespread adoption in ultra-low latency environments in physics when compared to DNNs and CNNs. This difference is owed in part to tools such as hls4ml that simplify the adaptation of the latter models from Keras to HLS. The support for GRUs and LSTMs in hls4ml that we present in this work represents the removal of a major barrier to the use of RNNs in ultra-low latency environments. This has ramifications not only for high energy physics but also other research areas where RNNs have become popular. While we have focused on the usage of hls4ml with FPGAs, it is important to note that hls4ml can also be used to create ASIC designs, and thus this work also allows for the possibility of RNN usage on ASICs as well. The recurrent or repeating nature of many modern algorithms, such as RNNs, transformers and graph neural networks, make them very difficult to be run, particularly at low latency, on FPGAs. In this work, we present the successful deployment of RNNs in models with number of trainable parameters ranging from O(1 k) to O(100 k) achieving latencies of O(1 s) to O(100s). This represents an important step in enabling support in hls4ml for more complex architectures with recursive computations.

References:

https://arxiv.org/abs/2207.00559

Experiment context, if any:

Poster session with coffee break / 245

Noise removal of the events at main drift chamber of BESIII with deep learning techniques

Author: Hosein Karimi Khozani[1]

Co-authors: Yao Zhang ; Ye Yuan [2]

[1] *IHEP*

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

Corresponding Author: karimi@ihep.ac.cn

There are established classical methods to reconstruct particle tracks from recorded hits on the particle detectors. Current algorithms do this either by cut in some features, like recorded time of the hits, or by the fitting process. This is potentially error prone and resource consuming. For high noise events, these issues are more critical and this method might even fail. We have been developing artificial neural networks which can learn to separate noise from signal in the simulated data. The data sample we use for this purpose is Monte-Carlo simulated Bhabha events generated by BESIII offline software system. We study different types of deep neural networks and their effectiveness to remove the noise which happens in the main drift chamber of BESIII from various origins.

The fully connected networks that we first try find sophisticated cuts in hit features of each cell of the detector. These features include raw time of a hit and the recorded charge associated to it. This leads to about 85 percent efficiency and purity of the signal separation. This sets up a lower limit for us since such a network judges every hit only by its own features. Next, we develop a CNN network and show that with information of only four neighboring cells, the noise removal happens with 99 percent purity and efficiency at the same time. We discuss the effectiveness of the network for events with different noise levels.

The main drift chamber is consisted of 6796 sense wires arranged in 43 layers. The structure of the wire system is known and therefore we also examine the idea of looking at the main drift chamber structure as a graph. We make a model based on graph convolutional layers and chose node classification approach. We include a message passing process in three of the hidden layers and get 95 percent efficiency and purity for the noise removal. We then describe the results of our network for other events such as j/psi to p+ p_ pi+ pi-. In the end, we compare all of this with the classical methods.

Significance:

As a proof of concept, we show how different machine learning techniques can be applied to remove noise from events (FNN, CNN, GCN) and find tracks of particles at the same time. Also, compared to classical algorithms, our neural networks can be used to much more efficiently remove noise, especially for events with a high level of it.
For the graph approach, two aspects of our work are different from the literature. Firstly, we model the real structure of the main drift chamber as a graph and secondly, we have a node classification approach to find tracks of the particles and remove noise hits.

References:

Experiment context, if any:

Poster session with coffee break / **246**

Compiling Awkward Lorentz Vectors with Numba

Authors: Henry Fredrick Schreiner[1]; Jim Pivarski[1]; Saransh Chopra[2]

[1] *Princeton University*

[2] *Cluster Innovation Centre, University of Delhi*

Corresponding Author: saransh0701@gmail.com

Due to the massive nature of HEP data, performance has always been a factor in its analysis and processing. Languages like C++ would be fast enough but are often challenging to grasp for beginners, and can be difficult to iterate quickly in an interactive environment . On the other hand, the ease of writing code and extensive library ecosystem make Python an enticing choice for data analysis. Increasing interoperability between Python and C++, as well as the introduction of libraries such as Numba, had been accelerating Python's traction in the HEP community.

Vector is a Python library for 2D, 3D, and Lorentz vectors, especially arrays of vectors, designed to solve common physics problems in a NumPy-like way. Vector currently supports pure Python Object, NumPy, Awkward, and Numba-based (Numba-Object, Numba-Awkward) backends.

We are introducing the library, with a focus on the Numba-based Awkward Lorentz vectors to perform operations on HEP data without compromising on the speed and the ease of writing code. Awkward is one of the core libraries of the Scikit-HEP ecosystem that allows data analysis with jagged arrays. Numba, on the other hand, allows Python codebases to harness the power of Just-In-Time compilation, enabling the Python code to be compiled before executing.

The library seamlessly integrates with the existing Scikit-HEP libraries, especially with Awkward. Our talk will start with an introduction to this library, with the main agenda of compiling Awkward Lorentz vectors with Numba. Furthermore, Vector is still under active development and preparing for a 1.0 release; hence, we will also take in user feedback while discussing the overall development roadmap.

Significance:

The Vector library is relatively new and has not been independently presented at any HEP-focused conference. Additionally, its interoperability with the Scikit-HEP ecosystem makes it a crucial part of most HEP analyses. Furthermore, the seamless Numba and Awkward support allows it to integrate with any existing HEP pipeline, making the pipeline faster, simpler, and more effective.

References:

All the links are available here - https://iris-hep.org/projects/vector.html

Experiment context, if any:

Track 3: Computations in Theoretical Physics: Techniques and Methods / **247**

Quantum annealing applications in high-energy phenomenology

Author: Juan Carlos Criado[1]

[1] *Durham University*

Corresponding Author: criadoalamo@gmail.com

Quantum annealing provides an optimization framework with the potential to outperform classical algorithms in finding the global minimum of non-convex functions. The availability of quantum

annealers with thousands of qubits makes it possible today to tackle real-world problems using this technology. In this talk, I will review the quantum annealing paradigm and its use in the minimization of general functions. I will then discuss some of the applications of this method in high-energy physics, including training neural networks for classification, and fitting effective field theories to experimental data.

Significance:

References:

Experiment context, if any:

Poster session with coffee break / 248

Performance portability with alpaka

Authors: Bernhard Manfred Gruber[1]; Jan Stephan[None]; Jiří Vyskočil[2]; Michael Bussmann[None]; René Widera[3]; Sergei Bastrakov[3]; Simeon Ehrig[None]; Tony Di Pilato[4]

[1] *Technische Universitaet Dresden (DE)*

[2] *CASUS - Center for Advanced Systems Understanding*

[3] *Helmholtz-Zentrum Dresden-Rossendorf*

[4] *CASUS - Center for Advanced Systems Understanding (DE)*

Corresponding Author: j.stephan@hzdr.de

The alpaka library is a header-only C++17 abstraction library for development across hardware accelerators (CPUs, GPUs, FPGAs). Its aim is to provide performance portability across accelerators through the abstraction (not hiding!) of the underlying levels of parallelism. In this talk we will show the concepts behind alpaka, how it is mapped to the various underlying hardware models, and show the features introduced over the last year. In addition, we will also (shortly) present the software ecosystem surrounding alpaka.

Significance:

The alpaka library has been adopted by the CMS experiment at CERN to be integrated in CMSSW as a performance portability library. It is also used in other projects, for example the PIConGPU project for particle-in-cell simulations.

References:

https://doi.org/10.1007/978-3-319-67630-2_36 - latest peer-reviewed paper about alpaka
https://github.com/alpaka-group/alpaka - Main software repository
https://www.hzdr.de/publications/Publ-33634 - Introductory lecture given at the ESC21 school in Bertinoro, Italy

Experiment context, if any:

Track 3: Computations in Theoretical Physics: Techniques and Methods / 249

Unweighted event generation for multi-jet production processes based on matrix element emulation

Authors: Daniel Maitre[None]; Frank Siegert[1]; HENRY TRUONG[None]; Steffen Schumann[None]; Timo Janssen[None]

[1] *Technische Universitaet Dresden (DE)*

Corresponding Author: timo.janssen@theorie.physik.uni-goettingen.de

The generation of unit-weight events for complex scattering processes presents a severe challenge to modern Monte Carlo event generators. Even when using sophisticated phase-space sampling techniques adapted to the underlying transition matrix elements, the efficiency for generating unit-weight events from weighted samples can become a limiting factor in practical applications. Here we present the combination of a two-staged unweighting procedure with a factorisation-aware matrix element emulator using neural networks which we make accessible in the Sherpa event generation framework. The algorithm can significantly accelerate the unweighting process, while it still guarantees unbiased sampling from the correct target distribution. We apply, validate and benchmark the approach in high-multiplicity LHC production processes, including Z/W+4 jets and $t\bar{t}$+3 jets, where we find speed-up factors up to 60.

Significance:

References:

Experiment context, if any:

Track 3: Computations in Theoretical Physics: Techniques and Methods / 250

Emulation of high multiplicity NLO k-factors

Author: HENRY TRUONG[None]

Co-author: Daniel Maitre

Corresponding Author: henry.truong@durham.ac.uk

Evaluation of one-loop matrix elements is computationally expensive and makes up a large proportion of time during event generation. We present a neural network emulator that builds in the factorisation properties of matrix elements which accurately reproduces the NLO k-factors for electron-position annihilation into up to 5 jets.

We show that our emulator retains good performance for high multiplicities and that there is a significant speed advantage over more traditional loop provider tools.

Significance:

Previous studies have shown that emulation of one-loop matrix elements is possible but the accuracy drops for higher multiplicities. We show that using the factorisation properties of matrix elements we are able to retain good performance even for 2->5 processes at one-loop.

References:

Experiment context, if any:

Track 3: Computations in Theoretical Physics: Techniques and Methods / 251

Anomaly searches for new physics at the LHC

Author: Barry Dillon[1]

[1] *University of Heidelberg*

Corresponding Author: dillon@thphys.uni-heidelberg.de

In this talk I will give an overview of our recent progress in developing anomaly detection methods for finding new physics at the LHC. I will discuss how we define anomalies in this context, and the deep learning tools that we can use to find them. I will also discuss how self-supervised representation learning techniques can be used to enhance anomaly detection methods.

Significance:

This talk will provide details that go beyond what has currently been published in the literature, and will cover important updates on the status of our work on anomaly detection for new physics searches at the LHC.

References:

https://arxiv.org/abs/2108.04253
https://arxiv.org/abs/2202.00686
https://arxiv.org/abs/2205.10380
https://arxiv.org/abs/2206.14225

Experiment context, if any:

Poster session with coffee break / **252**

Data transfer to remote GPUs over high performance networks

Author: Ali Marafi[1]

Co-author: Andrea Bocci [2]

[1] *Kuwait University (KW)*

[2] *CERN*

Corresponding Authors: andrea.bocci@cern.ch, ali.abdullah.marafi@cern.ch

In the past years the CMS software framework (CMSSW) has been extended to offload part of the physics reconstruction to NVIDIA GPUs. This can achieve a higher computational efficiency, but it adds extra complexity to the design of dedicated data centres and the use of opportunistic resources, like HPC centres. A possible solution to increase the flexibility of heterogeneous clusters is to offload part of the computations to GPUs installed in external, dedicated nodes.

Our studies on this topic have been able to achieve high-throughput, low-latency data transfers to and from a remote NVIDIA GPU across Mellanox NICs, using the Remote Direct Memory Access (RDMA) technology to access the GPU memory without involving either nodes' operating system.

In this work we present our approach based on the Open MPI framework, and compare the performance of data transfers of local and remote GPUs from different generations, using different communication libraries and network protocols.

Significance:

References:

Experiment context, if any:

CMS

Poster session with coffee break / **253**

SCD: an open, realistic calorimeter for ML studies in HEP

Authors: Nathalie Soybelman[1]; Nilotpal Kakati[2]

[1] *Weizmann Institute of Science (IL)*

[2] *Weizmann Institute of Science*

Corresponding Authors: nilotpal.kakati@cern.ch, nathalie.soybelman@cern.ch

The feature complexity of data recorded by particle detectors combined with the availability of large simulated datasets presents a unique environment for applying state-of-the-art machine learning (ML) architectures to physics problems. We present the Simplified Cylindrical Detector (SCD): a fully configurable GEANT4 calorimeter simulation which mimics the granularity and response characteristics of general purpose detectors at the LHC. The SCD will be released as a public software to accelerate development of ML-based reconstruction and calorimeter models. Two use-cases based on data from the SCD are presented: first, an ML-based global particle reconstruction which shows potential to outperform traditional approaches. Second, a fast simulation model transforming a set of truth particles into a set of reconstructed particles.

Significance:

References:

Experiment context, if any:

Planary / **254**

Graph Neural Networks and their application in IceCube

Corresponding Author: martin.haminh@icecube.wisc.edu

The interpretation of detector data to observables that we can use to perform our physics analyses is an essential part in modern day experimental physics. It is also a field among the biggest profiteers in the recent advances of machine learning. In this contribution we want to highlight our event reconstruction efforts using Graph Neural Networks in the IceCube experiment. Using a pulse-based approach our network can adapt to the irregular architecture of our detector. We can show not only speed-ups on the order of magnitudes but also increases in reconstruction resolution of up to 20% compared to our current baseline algorithms. Our goal is to provide an easy-to-use but effective entry into machine learning-based event reconstruction for any physics
purpose: from neutrino oscillations, over beyond-the-standard-model searches, to neutrino astronomy. In addition, our software package is not just compatible with the current IceCube experiment, but also for future extensions, like the IceCube Upgrade or Gen2, as well as any neutrino detector.

Experiment context, if any:

References:

Significance:

Planary / **255**

Practical Quantum Computing for Scientific Applications

Corresponding Author: jungsang@duke.edu

Trapped ion is the leading candidate for realizing practically useful quantum computers, as the system features highest performance quantum computational operations. Introduction of advanced integration technologies has provided an opportunity to convert a complex atomic physics experiment into a stand-alone programmable quantum computer. In this talk, I will discuss recent technological progress that changed the perception of a trapped ion system as a scalable quantum computer and enabled commercially viable quantum computer. I will also discuss several application areas where quantum computers can make a practical contribution to the computational frontier in scientific applications.

Experiment context, if any:

References:

Significance:

Planary / **256**

Adapting C++ for Data Science

Corresponding Author: vasil.georgiev.vasilev@cern.ch

Over the last decade the C++ programming language has evolved significantly into safer, easier to learn and better supported by tools general purpose programming language capable of extracting the last bit of performance from bare metal. The emergence of technologies such as LLVM and Clang have advanced tooling support for C++ and its ecosystem grew qualitatively. C++ has an important role in the field of scientific computing as the language design principles promote efficiency, reliability and backward compatibility - a vital tripod for any long-lived codebase. Other ecosystems such as Python have prioritized better usability and safety while making some tradeoffs on efficiency and backward compatibility. That has led developers to believe that there is a binary choice between performance and usability.

In this talk we would like to present the advancements in the C++ ecosystem; its relevance for scientific computing and beyond; and foreseen challenges. The talk introduces three major components for data science – interpreted C++; automatic language bindings; and differentiable programming. We outline how these components help Python and C++ ecosystems interoperate making a little compromise on either performance or usability. We elaborate on a future hybrid Python/C++ differentiable programming analysis framework which might accelerate science discovery in HEP by amplifying the power and physics sensitivity of data analyses into end-to-end differentiable pipelines.

Experiment context, if any:

References:

Significance:

Planary / **257**

Scientific Software and Computing in the HL-LHC, EIC, and Future Collider Era

Corresponding Author: danilo.piparo@cern.ch

A bright future awaits particle physics. The LHC Run 3 just started, characterised by the most energetic beams ever created by humankind and the most sophisticated detectors. In the next few years we will accomplish the most precise measurements to challenge our present understanding of nature that will, potentially, lead us to prestigious discoveries. However, Run 3 is just the beginning. A rich programme is ahead of us at the HL-LHC, the EIC, and at future colliders, like the FCC. These programs imply a large effort and substantial funding, for example to develop future detector and accelerator technologies, to construct new experiments and facilities, or expanding the scope of the existing ones. This contribution is about the software and computing that will lead us to the full exploitation of such infrastructure, the software and computing that will empower us to make important strides in humanity's understanding of the universe. The HL-LHC, EIC and FCC eras will be taken in consideration in this contribution. We will discuss the role of education, innovation and technology in our preparation for the future. We will also review the current state of the art, discuss ongoing technology evolutions, for instance in hardware and programming languages, and extrapolate most relevant trends into the next decades. Moreover, we'll identify the areas where our efforts could be focussed to boost the progression of particle physics software and computing, as well as the steps we can take to take advantage of veritable revolutions.

Experiment context, if any:

References:

Significance:

Planary / **258**

Lattice QCD on supercomputers with Chinese CPU

Corresponding Author: cheny@ihep.ac.cn

Lattice QCD is ab initio approach for QCD and plays an indispensable role in understanding the low energy properties of the strong interaction. Last four decades have witnessed the rapid development of the lattice QCD numerical calculation along with the progress of the high performance computing (HPC) techniques. Lattice QCD becomes one of the most resource-consuming HPC fields. China has built several native supercomputers with different hardware architectures,
such as Sunway series, Tianhe series and Sunrising-1 etc., which provide potentially massive HPC resources for lattice QCD studies.
This talk will give a brief introduction to the code developing and the performance of lattice QCD software on these strikingly different computing systems.

Experiment context, if any:

References:

Significance:

Planary / **259**

Quantum computing: a grand era for simulating fluid

Corresponding Author: rui.li@th-deg.de

Transport phenomena remains nowadays the most challenging unsolved problems in computational physics due to the inherent nature of Navier-Stokes equations. As the revolutionary technology, quantum computing opens a grand new perspective for numerical simulations for instance the computational fluid dynamics (CFD). In this plenary talk, starting with an overview of quantum computing including basic conceptions for instance qubits, quantum gates and circuit, more focus are then put on how to translate the algorithms from the classical computation system to quantum system. The possible quantum algorithms (e.g. partial different equation solver, eigenvalue solvers, etc.) for fluid dynamics are overviewed. Two concrete typical examples are presented with details namely: first one based on lattice Boltzmann method, the second one based on quantum Navier-Stokes algorithm. In the latter method the key process of reducing partial different equations to ordinary differential equations is explained. In the end the advantages of quantum computing are compared with the classical computation, indicating that a large application area for simulating fluid using quantum system is yet coming.

Experiment context, if any:

References:

Significance:

Planary / **260**

Loop amplitudes at the precision frontier

Corresponding Author: simon.badger@cern.ch

Precision simulations for collider phenomenology require intensive evaluations of complicated scattering amplitudes. Uncovering hidden simplicity in these basic building blocks of quantum field theory can lead us to new, efficient methods to obtain the necessary theoretical predictions. In this talk I will explore some new approaches to multi-scale loop amplitudes that can overcome conventional bottlenecks in their evaluation. Computational techniques based on evaluations over finite fields are now being used to obtain analytic information from numerical evaluations and can lead to fast and efficient implementations that can be used directly in Monte Carlo simulations. In some cases even the most compact representations of amplitudes can still mean prohibitive evaluation times. Approximating these complicated functions with Machine Learning technology has the potential to provide an order of magnitude improvement in evaluation times yet it remains a challenge to keep deviations from the complete amplitude under quantitative control. I will present some advances in the use of Neural Networks to provide reliable amplitude evaluations.

Experiment context, if any:

References:

Significance:

Planary / **261**

Simpler, faster and bigger: HEP analysis in the LHC Run 3 era

Corresponding Author: enrico.guiraud@cern.ch

The production, validation and revision of data analysis applications is an iterative process that occupies a large fraction of a researcher's time-to-publication.
Providing interfaces that are simpler to use correctly and more performant out-of-the-box not only reduces the community's average time-to-insight but it also unlocks completely novel approaches that were previously impractically slow or complex.
All of the above becomes especially true at the unprecedented integrated luminosity that will be achieved during LHC Run 3 and beyond, which further motivates the fast-paced evolution that has been taking place in the HEP analysis software ecosystem in recent years.
This talk analyzes the trends and challenges that characterize this evolution.
In particular we focus on the emerging pattern of strongly decoupling end-user analysis logic from low-level I/O and work scheduling by interposing high-level interfaces that gather semantic information on the particular analysis application.
We show how this pattern brings benefits to analysis ergonomics and reproducibility, as well as opportunities for performance optimizations.
We highlight potential issues in terms of extensibility and debugging experience, together with possible mitigations.
Finally, we explore the consequences of this convergent evolution towards smart, HEP-aware "middleman analysis software" in the context of future analysis facilities and data formats:
both will have to support a bazaar of high-level solutions while optimizing for typical low-level data structures and access patterns.
Our goal is to provide novel insights useful to boost the ever-ongoing, stimulating conversation that, since always, characterizes the HEP software community.

Experiment context, if any:

References:

Significance:

Planary / 262

TBC

Corresponding Author: jusovitsch@googlemail.com

Experiment context, if any:

References:

Significance:

Planary / 263

How Good is the Standard Model?

Corresponding Author: wulzer@cern.ch

Strategies to detect data departures from a given reference model, with no prior bias on the nature of the new physical model responsible for the discrepancy might play a vital role in experimental programs where, like at the LHC, increasingly rich experimental data are accompanied by an increasingly blurred theoretical guidance in their interpretation. I will describe one such strategy that employs neural networks, leveraging their virtues as flexible function approximants, but builds its foundations directly on the canonical likelihood-ratio approach to hypothesis testing. The algorithm compares observations with an auxiliary set of reference-distributed events, possibly obtained with a Monte Carlo event generator. It returns a p-value, which measures the compatibility of the reference model with the data. It also identifies the most discrepant phase-space region of the dataset, to be selected for further investigation. Imperfections due to mismodelling in the reference dataset can be taken into account straightforwardly as nuisance parameters.

Experiment context, if any:

References:

Significance:

Planary / **264**

Machine learning for phase space integration with SHERPA

Corresponding Author: enrico.bothmann@uni-goettingen.de

Simulated event samples from Monte-Carlo event generators (MCEGs) are a backbone of the LHC physics programme.
However, for Run III, and in particular for the HL-LHC era, computing budgets are becoming increasingly constrained, while at the same time the push to higher accuracies
is making event generation significantly more expensive.
Modern ML techniques can help with the effort of creating such costly samples in two ways.

One way is to use inference models to try to learn the event distribution of the entire MCEG toolchain, or parts of it, such that events can then be generated with those \emph{replacement models}
in a fraction of the time a full MCEG would require.
This ansatz is however intrinsically constrained by the available training data.
Another way, and this is the one discussed in this talk, is to keep the MCEG,
and to use ML \emph{assistant models} to increase the efficiency of certain performance bottlenecks.

One of those bottlenecks is the sampling of the high-dimensional phase space of complex processes, for which a given distribution must be approximated as closely as possible.
This is indeed a very generic problem, such that methods can be explored that have been developed in entirely different fields of physics or even outside of physics.

In this talk I will discuss the potential to increase the phase space sampling efficiency
using the methods of Neural Importance Sampling and Nested Sampling,
and of neural network surrogates of the integrand to increase the efficiency of event unweighting.
The application of these methods within the \textsc{Sherpa} generator framework is then reviewed.

Experiment context, if any:

References:

Significance:

Planary / **265**

The European Processor Initiative (EPI), an status update

Corresponding Author: e.suarez@fz-juelich.de

TBC

Experiment context, if any:

References:

Significance:

Planary / **266**

Machine Learning for Beyond the Standard Model Physics

Corresponding Author: sven.krippendorf@physik.uni-muenchen.de

In this talk I discuss how machine learning can be used for identifying underlying mathematical structures in physical systems. Geared towards relevant structures in Beyond the Standard Model Physics I will focus on how we can use ML to discover symmetries. I discuss how standard ML pipelines have to be adopted to enable such discoveries and comment on further applications of these methods in physics beyond symmetries.

Experiment context, if any:

References:

Significance:

Planary / **267**

TBC

Corresponding Author: appuswam@eurecom.fr

Experiment context, if any:

References:

Significance:

Planary / **268**

Machine Learning in the Search for New Fundamental Physics

Corresponding Author: gregor.kasieczka@cern.ch

As the search for new fundamental phenomena at modern particle colliders is a complex and multifaceted task dealing with high-dimensional data, it is not surprising that machine learning based techniques are quickly becoming a widely used tool for many aspects of searches. On the one hand, classical strategies are being supercharged by ever more sophisticated tagging algorithms; on the other hand, new paradigms — such as searching for anomalies in a data-driven way — are being proposed. This talk will review some key developments and consider which steps might be needed to maximise the discovery potential of particle physics experiments.

Experiment context, if any:

References:

Significance:

Planary / **269**

TBC

Corresponding Author: a.l.varbanescu@uva.nl

Experiment context, if any:

References:

Significance:

Planary / **270**

Lattice QCD with the Supercomputer Fugaku - progress and prospects

Corresponding Author: yasumichi.aoki@riken.jp

The Japanese flagship supercomputer Fugaku started its operation in early 2021.
After one and half years of production runs it is producing some initial results in Lattice QCD applications, such as thermodynamics, heavy and light quark flavor physics, and hadron structures and interactions.
In this talk, we first touch on the basis of Fugaku and its software status.
Discussion is given on the ongoing projects highlighting some initial results, mainly focusing on those using domain wall fermions, a practical chiral fermion formulation on the lattice.

Experiment context, if any:

References:

Significance:

Planary / 271

TBC

Corresponding Author: helena.liebelt@th-deg.de

Experiment context, if any:

References:

Significance:

Planary / 272

Lightning Talk 1

Planary / 273

Lightning Talk 2

Planary / 274

Lightning Talk 3

Planary / 275

Track 1 Summary

Planary / 276

Track 2 Summary

Planary / 277

Track 3 Summary

Planary / **278**

ACAT 2022 Summary

Corresponding Author: david.britton@cern.ch

Planary / **279**

Welcome

Planary / **280**

Foundation Models for Accelerated Discovery

Corresponding Author: jsmith@us.ibm.com

AI is making an enormous impact on scientific discovery. Growing volumes of data across scientific domains are enabling the use of machine learning at ever increasing scale to accelerate discovery. Examples include using knowledge extraction and reasoning over large repositories of scientific publications to quickly study scientific questions or even come up with new questions, applying AI surrogate models to speed up simulation campaigns and generate critical new data and knowledge, leveraging generative models to construct new hypotheses and make predictions about them, and automating experimentation through robotic labs to enable tighter loops of hypothesis-test cycles. At the same time, new machine learning techniques based on “foundation models” are gaining focus in AI. Foundation models aim to learn “universal representations” from enormous amounts of data, typically using self-supervised or unsupervised training, with the goal to effectively enable subsequent downstream tasks. Prominent examples are large-language models, which have been driving state-of-the-art performance for natural language processing tasks. In this talk, we review how foundation models work by learning representations at scale and show examples of how they can further accelerate scientific discovery. By targeting bottlenecks in the scientific method, we discuss the potential of foundational models to impact a broad set of scientific challenges.

Planary / **281**

Updates from the organizers

Corresponding Authors: axel.naumann@cern.ch, lucia.silvestris@cern.ch

Planary / **282**

Updates from the organizers

Corresponding Authors: axel.naumann@cern.ch, lucia.silvestris@cern.ch

Planary / 283

Updates from the organizers

Corresponding Authors: axel.naumann@cern.ch, lucia.silvestris@cern.ch