PROCEEDINGS OF SCIENCE

Enhancing CMS data analyses using a distributed high throughput platform

Tommaso Diotalevi,[a,b,*] **Carlo Battilana,**[a,b] **Alessandra Fanfani**[a,b] **and Daniele Bonacorsi**[a,b]

[a] *Department of Physics and Astronomy, University of Bologna, Viale C. Berti Pichat 6/2, Bologna, Italy*

[b] *INFN Bologna, Viale C. Berti Pichat 6/2, Bologna, Italy*

E-mail: tommaso.diotalevi@unibo.it

A flexible and dynamic environment capable of accessing distributed data and resources efficiently, is a key aspect for HEP data analysis, especially for the HL-LHC era. A quasi-interactive declarative solution, like ROOT RDataFrame, with scale-up capabilities via open-source standards like Dask, can profit from the "HPC, Big Data and Quantum Computing" Italian Center DataLake model under development. The starting point is a prototypal CMS high throughput analysis platform, offloaded on local Tier-2.

This contribution evaluates the scalability, identifies bottlenecks and explores the interactivity of such platform, on two use-cases: a CMS physics analysis with high-rate triggered events and a study of the CMS muon detector performance in phase-space regions driven by analysis needs, accessing detector datasets. The metrics used to evaluate the scaling and speed-up performance will be reported and results will be discussed, emphasising the differences with the legacy analysis workflows.

42nd International Conference on High Energy Physics (ICHEP2024)
18-24 July 2024
Prague, Czech Republic

*Speaker

PoS(ICHEP2024)1007

1. Introduction

In the coming years, the Large Hadron Collider (LHC) will receive a significant upgrade - the High-Luminosity LHC (HL-LHC) [1] - which is expected to generate approximately 100 PB of data annually from collision events [2]. This will demand an immense amount of computing resources, including CPU and storage, which will become unsustainable after a few years of operation with the current computing infrastructure [3]. It is therefore fundamental to develop new industry-standard analysis paradigms, combining declarative programming solutions and interactive workflows. Behind this, a DataLake-like infrastructure with distributed computing elements capable of running different applications coming from the High Energy Physics (HEP) domain.
Funded by Italy's National Recovery and Resilience Plan (NRRP), the '*High-Performance Computing, Big Data, and Quantum Computing National Centre*' (ICSC) will supply the necessary resources[1] enabling the deployment of a high throughput analysis platform on a national level.
This contribution provides a technical overview of the platform, along with two use cases from the CMS Collaboration [4]: a muon detector performance study and a technical performance overview of a physics data analysis with high-rate triggered events.

2. The high throughput platform

A schematic diagram describing the proposed architecture [5] is shown in Figure 1. By connecting to an entrypoint URL, the user reaches a JupyterHub[2] instance that, after authentication and authorization via INDIGO-IAM[3], allocates the required resources for the user's working area. Afterwards, the user is redirected to a JupyterLab[4] user interface where it is possible to store all the code required to perform the analyses. The working environment is highly customizable using container-based technologies (e.g. Docker[5], Singularity/Apptainer[6]): using centralized services like CVMFS [6], container images can be stored by the users and pulled directly inside the platform. Thanks to this solution, analysts are able to use specific software and tools for their applications, plotting frameworks, statistical tools, etc...). This part of the infrastructure is built on top of a Kubernetes (K8s) cluster[7], hosted at the INFN-CNAF (Italy) cloud service.
On the back-end side, a HTCondor-based overlay connecting to the remote computing resources: for this contribution, the resources hosted at the Italian Tier-2 of Legnaro (CMS pledged) have been used.

2.1 Distributing the workload

As already mentioned, the key technological feature of this platform is the distribution of computational workflows. This is accomplished by parallelizing the computation through two main steps, closely resembling the MapReduce paradigm [7]: first, submitting multiple jobs to process

[1] ICSC main page: https://www.supercomputing-icsc.it
[2] JupyterHub main page: https://jupyter.org/hub
[3] Indigo-IAM main page: https://indigo-iam.github.io/v/current/
[4] JupyterLab main page: https://jupyter.org/lab
[5] Docker main page: https://www.docker.com/
[6] Apptainer/Singularity main page: https://apptainer.org/
[7] Kubernetes main page: https://kubernetes.io/

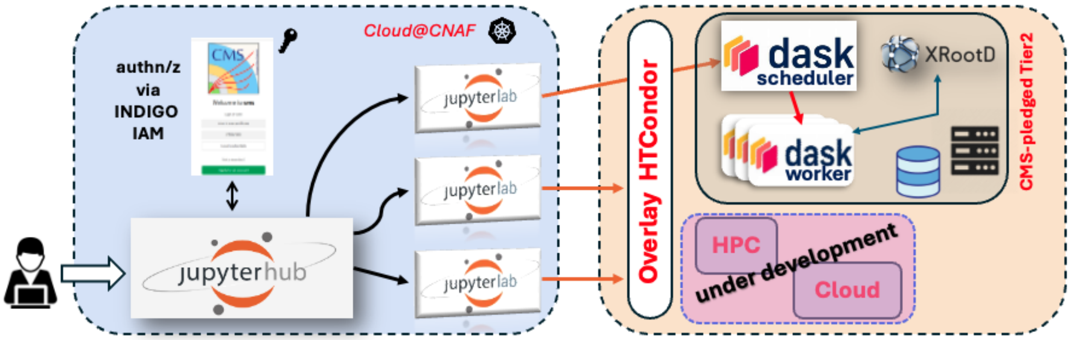


Figure 1: A sketched diagram of the proposed high throughput platform. The offloading on heterogeneous resources is still under development.

different portions of the input dataset; second, combining the partial results to generate the final output, such as histograms and relevant statistics. Among the various open-source software tools available for these tasks, the Dask Python library[8] has been selected. This choice is based on Dask scalability, flexibility, and seamless integration with already existing Python code. It efficiently manages large datasets by distributing tasks across multiple cores and nodes, while offering a user-friendly API compatible with standard data analysis libraries. The datasets can be stored and read within the local filesystem of the user Jupyter instance; alternatively, since more often experimental data are located in WLCG [8] remote storage sites, they can be accessed via XRootD[9] and/or WebDAV [9] protocols.

2.2 A Declarative software framework for HEP

As mentioned above, this platform relies on declarative software solutions. Based on the ROOT toolkit [10], the RDataFrame interface[10] has been used. It provides users with an high-level declarative approach, allowing them to apply filters, define columns, and execute computations efficiently without writing explicit loops. In this way, analysers can concentrate more on the physics itself, freeing them from the repetitive code required for data access, event looping, computation distribution, and result aggregation.

3. Performance evaluation on CMS use cases

In order to evaluate the performance of the platform, a selection of use cases from the CMS Collaboration have been chosen, sharing a common challenge: analysing big amounts of data as quickly and as interactively as possible.

3.1 Muon Detector Performance analysis

Analyses from the Detector Performance Group (DPG) at the CMS experiment, typically, run on a reduced amount of data (e.g. a single run of data taking, or multiple runs in a single beam fill

[8] Dask documentation: http://dask.pydata.org
[9] XRootD protocol main page: https://xrootd.slac.stanford.edu/
[10] ROOT RDataFrame main page: https://root.cern/doc/master/classROOT_1_1RDataFrame.html

at LHC). However, for some specific cases, a full processing of entire years might be needed. For example, to assess or improve any systematic uncertainty of high precision analyses (when they are dominated by the response of a specific detector), or to reprocess data across multiple years, e.g. for detector stability studies (ageing, etc...).
This use case involves a specific detector performance study: the Tag-and-Probe analysis [11] of the Drift Tubes (DT) muon sub-detector [12]. The dataset consists in a skim of $Z \rightarrow \mu\mu$ decay candidates collected by the CMS experiment during 2023, and corresponding to an integrated luminosity of 27fb^{-1}. Unlike the majority of physics analysis datasets, here also low-level detector quantities are included. The total size corresponds to about 224GB.
The legacy code is running mainly on C++, reconstructing DT segments and computing the efficiencies. In order to run on the platform, the code has been ported to Python, and runs on a Jupyter notebook using ROOT RDataFrame. The workflow is the same of the legacy approach: the libraries and functions are stored in a dedicated C++ header file, where the objects are manipulated using ROOT RVec objects. This allows to use a RDataFrame approach with Dask as a back-end, thus enabling the entire analysis flow to be distributed (up to the available resources).

3.1.1 Porting results

The Tag-and-Probe method is used to measure the DT efficiency to reconstruct a local track segment. Events are selected to contain a pair of oppositely charged reconstructed muons with some specific criteria, first for the tag muon and subsequently for the probe muon [11]. A DT chamber crossed by a probe track is considered efficient if a reconstructed segment is near the extrapolated track (within 15cm). In Figure 2a, it is possible to notice how the changes applied to the Tag-and-Probe program - running on the high throughput platform (shown in yellow) - do not affect performance, in agreement with the legacy approach (shown in blue). This agreement also includes high-energy muons, a generally less explored phase-space, as shown in last bins of Figure 2b.

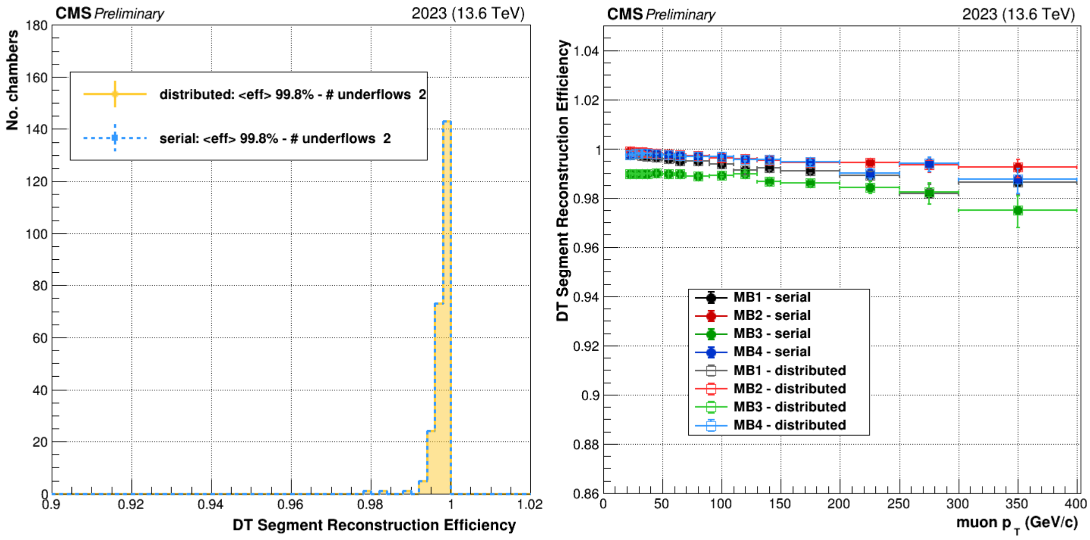
3.1.2 Technical performance

To evaluate the technical performance, the available statistics has been processed three times, mimicking an integrated luminosity of about 82fb^{-1}, reaching about 77M events in total. The total size of the considered dataset reaches now 672GB.
For the legacy execution, the computation has been performed as a single HTCondor job, running on a single CPU of the AMD EPYC 7302 16-Core processor, with 2GB of memory. The computation reached a walltime of about 120 minutes. On the other hand, for the distributed processing on the high throughput platform, the computation has been performed on two AMD EPYC 7413 24-Core processors[11]. The number of CPU used has been gradually incremented up to 92, with 2GB of memory per CPU, reaching a minimum walltime of about 6 minutes.

3.2 Technical performance of a physics data analysis

The performances of the high throughput platform have been also investigated on a CMS physics data analysis[12]. The main challenge of such analysis relies on the large dataset used,

[11]Resources hosted at the Italian CMS Tier-2 of Legnaro, remotely monitored using in-site metrics stored in a database.
[12]For approval reason, results and figures are not shown. Only a technical overview is given.



(a) DT segment reconstruction efficiency (one entry per chamber).

(b) DT segment reconstruction efficiency vs muon p_T, for each DT station (labelled as MB).

Figure 2: Porting results for the Tag-and-Probe analysis. The markers labelled as 'distributed' show the execution running on the high throughput platform, while the markers labelled as 'serial' show the execution running on the legacy software.

coming from a high-rate triggered stream called *b-parking* [13], gathered by CMS during 2018.
The same analysis workflow has been run on a increasing number of Dask workers (sharing the computational load), showing as expected a decrease in the execution time. The computation has been performed in a testbed of resources coming from the Legnaro Tier-2. Unlike the node used for the previous use case, this is equipped with a less powerful network bandwidth (1Gbit/s instead of 10 Gbit/s) while keeping the number of CPUs and memory per CPU unchanged. With a low number of workers, as expected, the CPU usage saturates (highlighting the high computational load); however, for a high number of workers, the worst hardware bandwidth translates in a network access bottleneck (due to the excessive I/O access required from remote storage resources).

4. Conclusions and outlook

A novel and innovative approach has been presented to handle the large volume of data to be processed in the HL-LHC phase: an interactive, high throughput platform built on a parallel and distributed computing system. A use case coming from the CMS detector performance group shows the benefits of such an approach: a significant decrease in the execution time and the quasi-interactivity. Every time a re-execution of the analysis is needed (e.g. tweaking some thresholds or using different selection criteria), now it boils down to run a few Jupyter Notebook cells. This can result in a great improvement for any future detector performance analysis application. On the other hand, the use case from the physics data analysis (in addition to the decrease in execution time) shows a potential bottleneck in terms of network access.
As a future evolution of this work, an effort of extending the entire infrastructure towards a multi-tenant platform is undergoing, intercepting also the needs of data analysts coming from different

scientific collaborations. In this sense, a new platform is under development; entirely deployed on a k8s cluster (thus fully scalable) and offloaded to the new infrastructure under construction by the Italian National Center for HPC, Big Data and Quantum Computing (ICSC). This new center will eventually benefit from cutting-edge resources and network bandwidth: therefore, all these tests will undergo a major scale test, and all the limiting aspects (e.g. network access) will be re-evaluated.

Acknowledgements

This work is supported by ICSC – Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing, funded by European Union – NextGenerationEU.

References

[1] I. Zurbano Fernandez et al., *High-Luminosity Large Hadron Collider (HL-LHC): Technical design report*, Tech. Rep. (12, 2020), 10.23731/CYRM-2020-0010.

[2] E. Elsen, *A Roadmap for HEP Software and Computing R&D for the 2020s*, *Computing and Software for Big Science* **3** (2019) 16.

[3] CMS Offline Software and Computing, *CMS Phase-2 Computing Model: Update Document*, (Geneva), 2022, https://cds.cern.ch/record/2815292.

[4] CMS collaboration, *The CMS experiment at the CERN LHC*, *Journal of Instrumentation* **3** (2008) S08004.

[5] T. Tedeschi et al., *Prototyping a ROOT-based distributed analysis workflow for HL-LHC: The CMS use case*, *Computer Physics Communications* **295** (2024) 108965.

[6] P. Buncic et al., *CernVM: A virtual software appliance for LHC applications*, *J. Phys. Conf. Ser.* **219** (2010) 042003.

[7] J. Dean and S. Ghemawat, *Mapreduce: simplified data processing on large clusters*, *Commun. ACM* **51** (2008) 107–113.

[8] I. Bird, *Computing for the Large Hadron Collider*, *Ann. Rev. Nucl. Part. Sci.* **61** (2011) 99.

[9] J. Whitehead et al., *Web Distributed Authoring and Versioning (WebDAV) Redirect Reference Resources*, No. 4437 in Request for Comments, RFC Editor, Mar., 2006, 10.17487/RFC4437.

[10] R. Brun and F. Rademakers, *ROOT: An object oriented data analysis framework*, *Nucl. Instrum. Meth. A* **389** (1997) 81.

[11] CMS collaboration, *Drift Tube Performance in 2023*, https://cds.cern.ch/record/2868786.

[12] CMS collaboration, *The CMS muon project: Technical Design Report*, Technical design report. CMS, (Geneva), CERN, 1997, https://cds.cern.ch/record/343814.

[13] CMS collaboration, *Enriching the Physics Program of the CMS Experiment via Data Scouting and Data Parking*, 2403.16134.