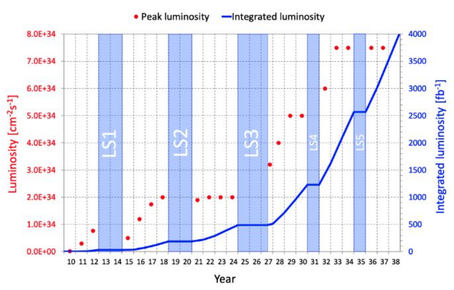
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

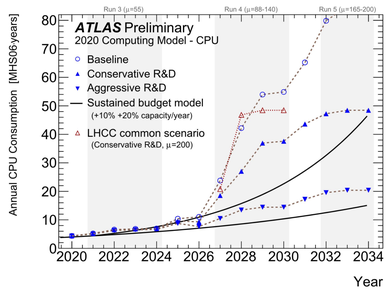
Benchmarking distributed-interactive HEP analysis workflows on the new Italian National Centre analysis infrastructure

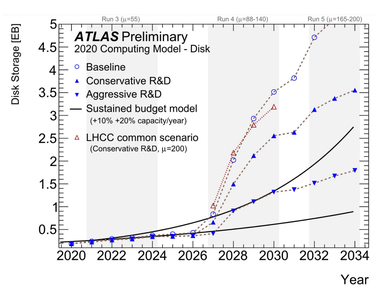
Adelina D'Onofrio[1], Muhammad Anwar[2], Matteo Bartolini[1,5], Antimo Cagnotta[1,6], Antonio D'Avanzo[1,6], Tommaso Diotalevi[1,3], Francesco Giuseppe Gravili[1,7], Paolo Mastrandrea[1], Elvira Rossi[1,6], Gianluca Sabella[1,6], Federica Maria Simone[1,2], Bernardino Spisso[1], Alessandro Tarasio[4], Tommaso Tedeschi[1]

Motivation

The upcoming high-luminosity phase at the **CERN Large Hadron Collider (LHC)** and at future accelerator facilities will require an increasing amount of computing resources [1]







Higher rates of collision events → Higher demand for computing and storage resources

To better analyse this increasing amount of Big Data:

- Optimize the usage of CPU and storage;
- Promote the usage of better data formats;
- **Develop new analysis paradigms**!

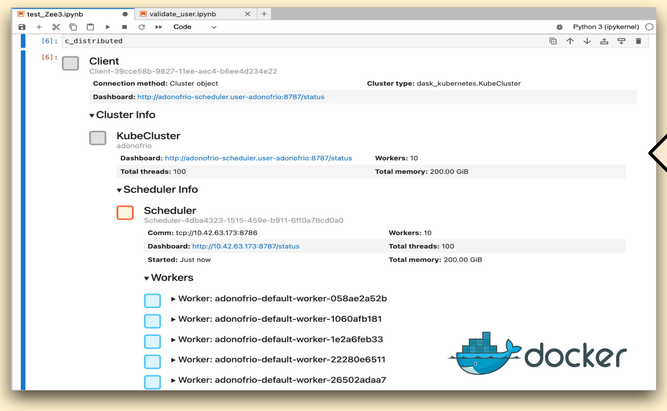
→
- New software based on declarative programming and interactive workflows;
- Distributed computing on geographically separated resources

Access and security

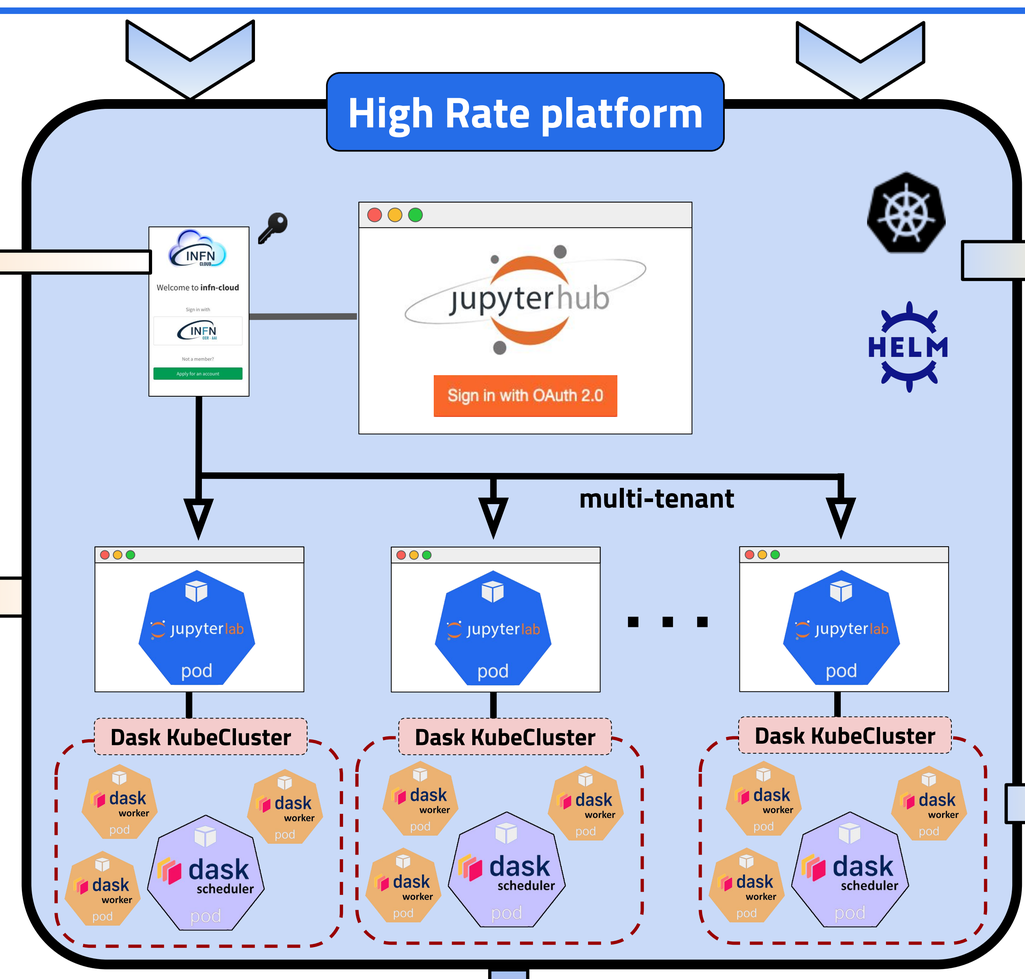
After connecting to an entrypoint URL, the user reaches a **Jupyterhub** [2] instance that, after authentication and authorization via INDIGO-IAM [3], allocates the required resources for the user's working area

User Interface

The user interface is based on **Jupyterlab**, customised with specific plugins for specific purposes (e.g. Dask).



The working environment is highly customizable, using tailored Docker containers. This is important when analyses require specific software (collaboration-wise)



Deployment

The deployment of the **Kubernetes** [4] resources needed for the spawning of this platform, is handled via **HELM** [5] **charts** available in the GitHub organization [6].

Check the docs!

This allows a seamless, flexible, scalable and fault-tolerant deployment on the available resources, with a limited impact on the admin's work time

Software

From the software perspective, interactive/quasi interactive analysis is a promising paradigm

- User-friendly environment
- Adopting open-source industry standards: *Dask, Jupyter Notebooks* and *HTCondor*
- Validating new frameworks (e.g. *ROOT RDataFrame* [7] with multi-threading)

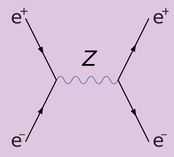
HEP analysis performance evaluation

Evaluating the performance of several High Energy Physics analyses from different experiments, using an approach based on interactive/quasi interactive analysis and parallel computing

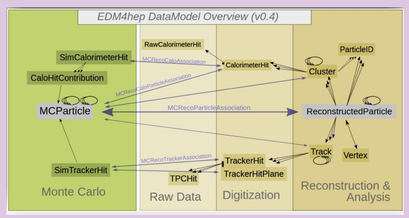
Z*ee* simulations at Future Colliders (FCCee)

Feasibility study: Mimic systematic variations applying a gaussian smearing to e^+e^- energies many times.

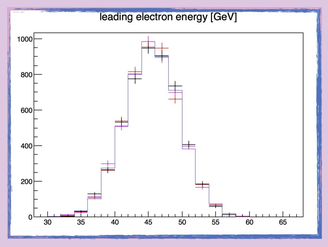
Physics process



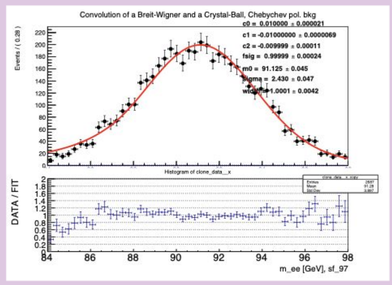
Simulations → flat dataformat



Electron energy variations



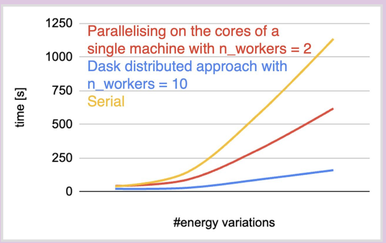
Invariant mass fit



Events selection and histogramming: interactively with **ROOT RDataFrame** and **Jupyterlab** [8]

Dask [9] used as backend

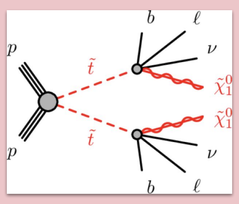
Preliminary Results



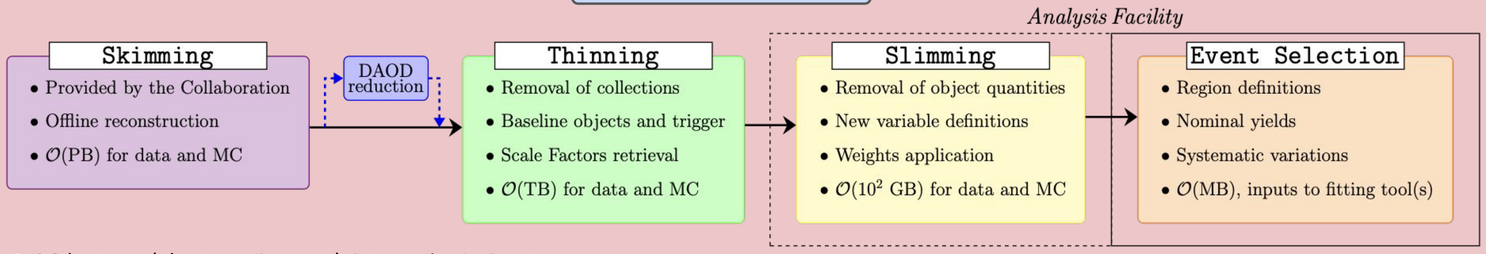
Beyond Standard Model Physics search at LHC

Use case: *Search for new phenomena in events with two opposite-charge leptons, jets and missing transverse momentum in pp collisions at* $\sqrt{s}$ *= 13 TeV with the ATLAS detector* [10]

Physics process



Workflow

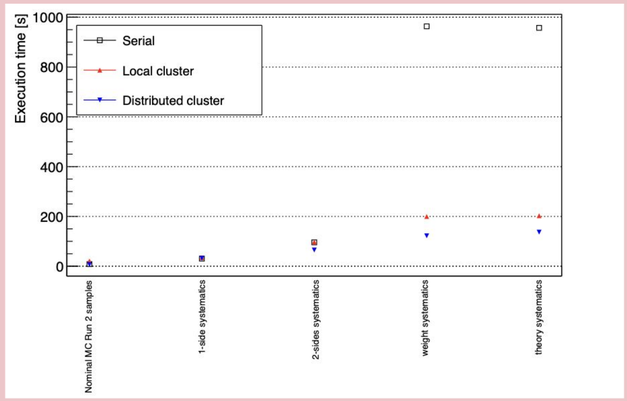


- Final state signature: 2 OS leptons (electrons/muons), jets and missing transverse energy
- Cut & Count based approach
- The analysis workflow was performed over the nominal case and several systematic variations

Events selection and histogramming: interactively with **ROOT RDataFrame** and **Jupyterlab** [8]

Dask [9] used as backend

Preliminary Results



- Both use cases show similar preliminary results: the high rate platform works properly and it is transparent to the experiment tested
- Considering the overall execution time as metric and running the same workflow, there is a performance improvement in the distributed approach wrt the standard/serial approach;
- Moreover, it was tested that scaling resources, the performance further improves.

References

1. https://cds.cern.ch/record/2729668/files/LHCC-G-178.pdf
2. https://jupyterhub.readthedocs.io/en/stable/
3. https://github.com/indigo-iam/iam
4. https://kubernetes.dask.org/en/latest/operator.html
5. https://helm.sh/
6. https://github.com/ICSC-Spoke2-repo/HighRateAnalysis-WP5
7. https://root.cern/doc/master/classROOT__1__1RDataFrame.html
8. https://jupyterlab.readthedocs.io/en/latest
9. https://docs.dask.org/en/stable/
10. The ATLAS Collaboration JHEP 04 (2021) 165

1: INFN - Istituto Nazionale di Fisica Nucleare
2: Polytechnic of Bari
3: University of Bologna
4: University of Calabria
5: University of Firenze
6: University of Naples
7: University of Salento

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