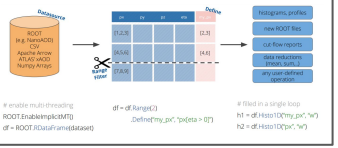
Benchmarking distributed-RDataFrame with CMS analysis workflows on the INFN analysis infrastructure

Daniele Spiga - INFN
On behalf of … (see next slide)

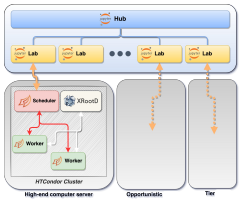
This Talk

A combined R&D Project

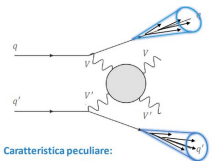
Adoption of modern tools: RDF



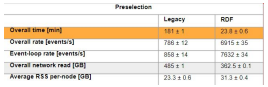
Data Analysis: a new approach



Benchmark definition: VBS ssWW analysis Porting to RDF



Legacy vs new comparison: benchmark measurements and results



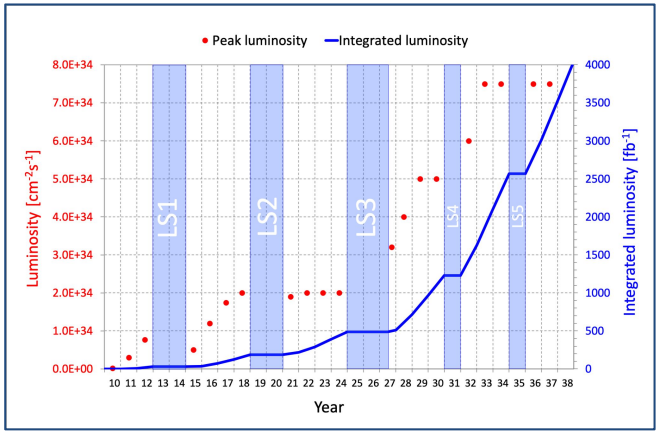
Preselection	Legacy	RDF
Overall time [min]	181 ± 1	23.8 ± 0.6
Overall rate [events/s]	786 ± 12	6915 ± 35
Event-loop rate [events/s]	858 ± 14	7632 ± 34
Overall network read [GB]	485 ± 1	362.5 ± 0.1
Average RSS per-node [GB]	23.3 ± 0.6	31.3 ± 0.4

The (growing) Team

- Tommaso Tedeschi (**INFN**)
- Vincenzo Eduardo Padulano (**CERN**)
- Daniele Spiga (**INFN**)
- Diego Ciangottini (**INFN**)
- Enric Tejedor Saavedra (**CERN**)
- Enrico Guiraud (**Princeton University, CERN**)
- Massimo Biasotto (**INFN**)
- Tommaso Diotalevi (**University of Bologna**)
- Alessandra Fanfani (**University of Bologna**)

(Main) Motivations

NOTE2022_008



R&D on analysis at High Luminosity LHC (HL-LHC)

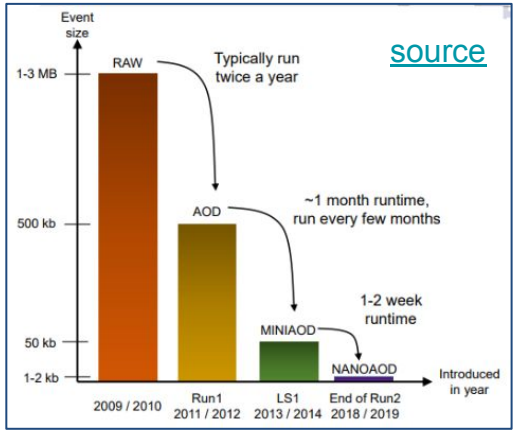
- Promote adoption of NanoAOD
- optimizing the computing and storage resource utilization

Testing software featuring a declarative programming model and interactive workflows

- Increasing data processing throughput is crucial
- Ergonomic interfaces remove the lower-level programming burden from analysts
- Fast Turnaround Reducing analysis “time to insight”

Prototype resources integration models to efficiently leverage computing capacity

- Integrate already deployed (grid) infrastructure
- Transparently access specialized HW
- Scale toward opportunistic (cloud/HPC)



source

The Foundations

Toward the end of 2020 we started a Proof of Concept on (very) high throughput analysis at INFN

User Perspective

- **A single entrypoint (HUB) for the data analysis**
  - Web based (not necessarily)
  - High level analysis framework agnostic
- **Bring in user runtime environment**
  - Allow the usage of user tailored images both locally and over all the distributed resources.
- **Scale seamlessly from 1 to 1000+ cores**
  - Transparently distributing the user payload on dispersed resources
  - “Get a jupyter session as big as a Tier2”

Computing Perspective

- **Implement the continuum (HTC/HPC/Cloud)**
  - Integrate heterogeneous resources under the same pool
  - Lower the bar for integrating distributed facilities
- **Use batch-systems (also) for interactive processing**
  - Distributing payloads from remote (cloud-native) services
- **Seamlessly exploiting the existing WLCG infrastructure for interactive use**
  - No dedicated Hardware, except for a seed of resources at INFN Cloud
  - Looking forward DataLake

How it is Made

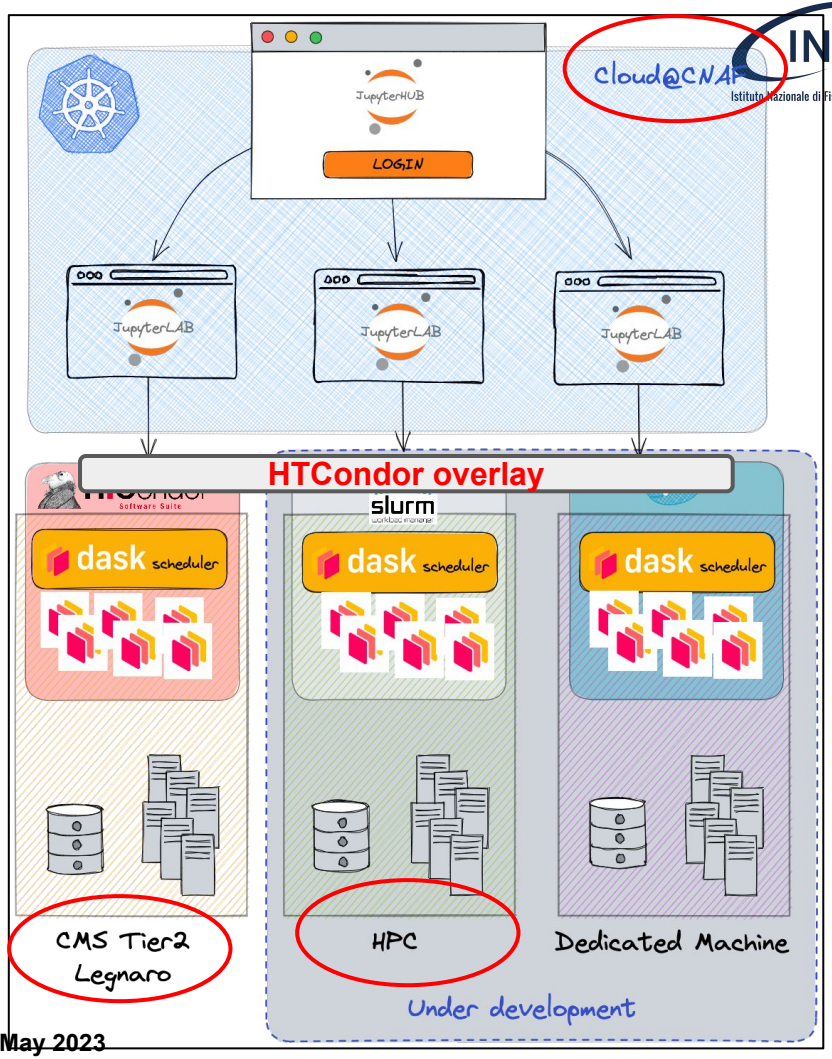
We developed a production ready system running at INFN

- Based on industry standard (plus very few customization)
- HTCondor as overlay technology
- DASK
- Completely exportable and replicable
- Token based AuthN/Z (via INDIGO-IAM)

The challenge:
benchmark this new facility using a full scale CMS analysis

What users see

What the offloading hides to the user



Benchmarking Strategy

A real analysis re-coded using RDataFrame and run over the very same Hardware setup and compare the two distinct approaches (Legacy processing vs RDataFrame) over pre-defined metrics

Selected analysis was scattering (VBS) of two same-sign W bosons decaying to a hadronic tau and a light lepton

"medium" size of the analysis:
- Preselection keeps 2% of initial O(1B) MC events. O(100k) MC events make it to the final histograms

Data format already used (**NanoAOD**)

Physical importance for Run 3 and beyond
- Using Run3 as playground, looking forward for HL-LHC

Defined Metrics

Overall execution time	Time elapsed from the start of the execution (legacy: first job submitted, RDF: execution triggered) to the end of execution.
Rate (events/s). **Job (initialization time) and event-loop-only**	The ratio between the total number of events processed and sum of processing times obtained from single job logs.
Network read	Per node information about total bytes read from the network during the execution. This value is summed across all nodes.
Absolute memory occupancy (RSS)	Per node information averaged across executions time and across all the available nodes.

A key element of the project

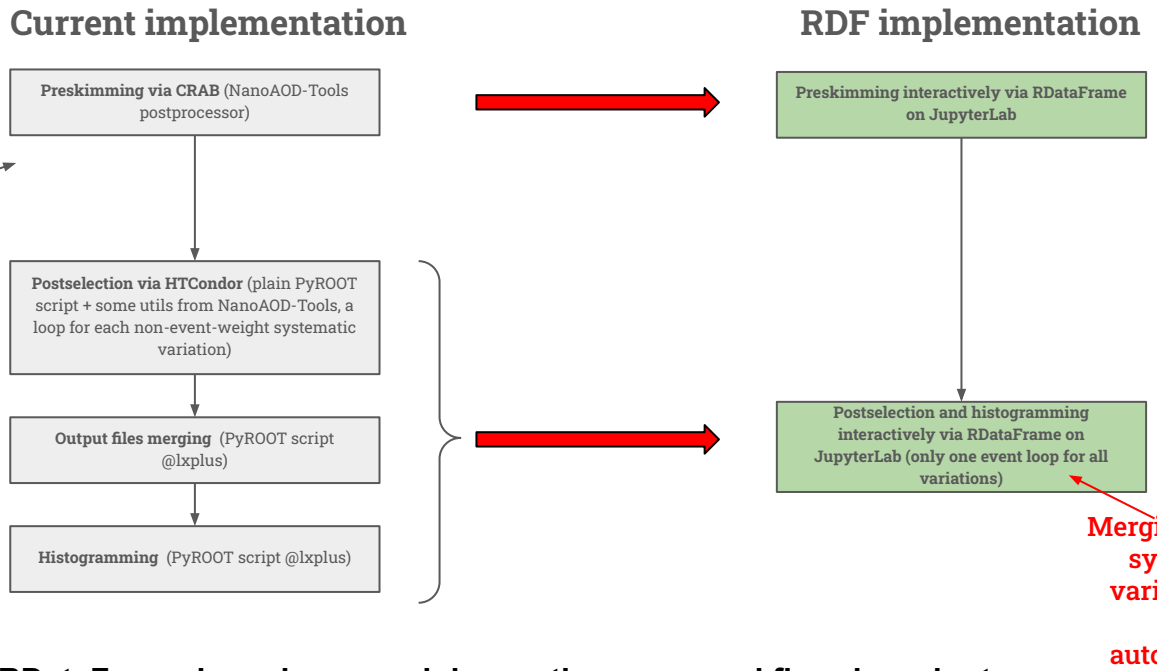
The physics analysis is converted from a legacy iterative approach to the modern declarative approach offered by RDataFrame

The legacy approach of this **analysis is based on a two-step procedure**:

A preselection step, where the original files are skimmed producing reduced flat ROOT-files;

Postselection step, where the proper analysis is run.

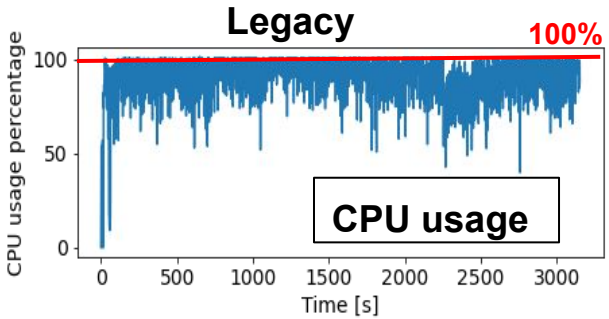
- production of histograms, for each systematic variation,

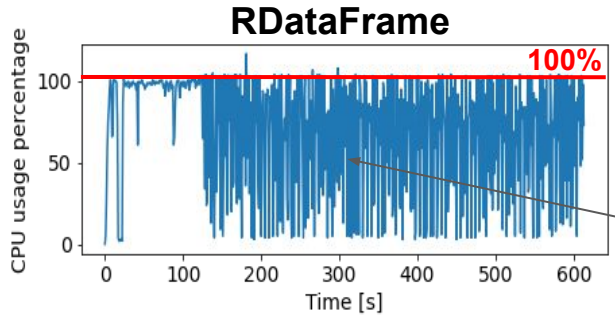


RDataFrame-based approach keeps the same workflow, in order to achieve a one-to-one mapping to the legacy approach

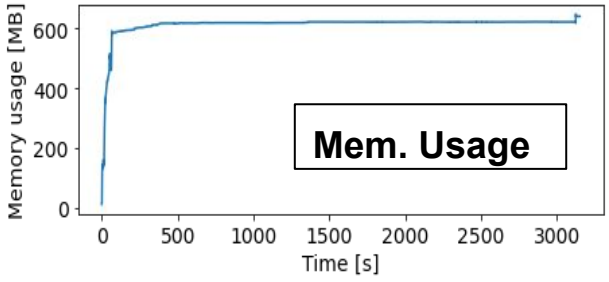
Performance checks: per job

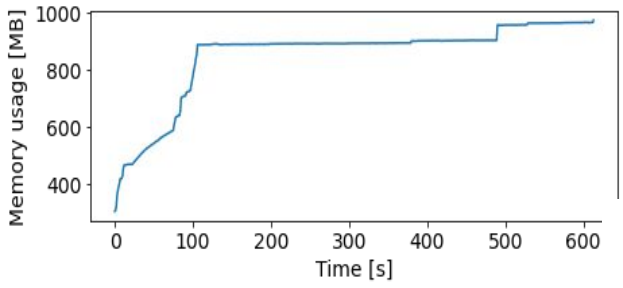
Montecarlo samples, simulating 2017 data-taking operating conditions, for a total of 657M events, into 1274 nanoAOD files (1.1 TB)





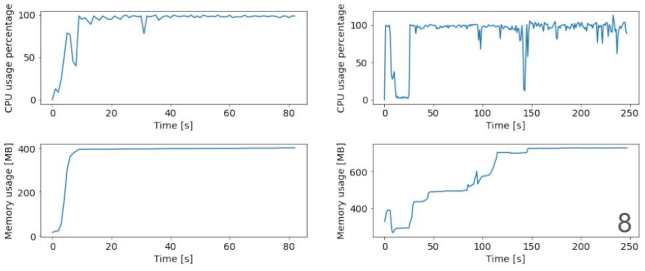
Oscillation in the second part of the execution due to the **network saturation**





A higher throughput add more stress on network (and storage I/O)

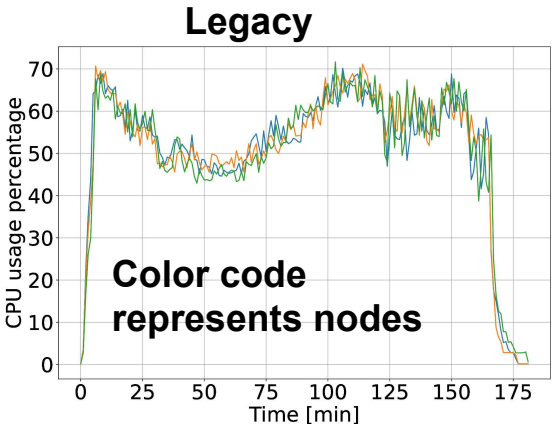
Checks made also post selection

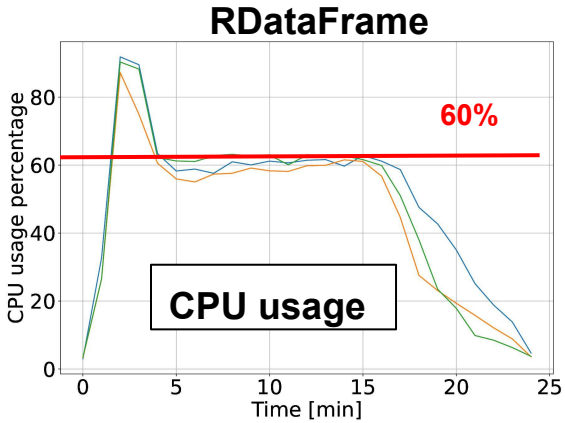


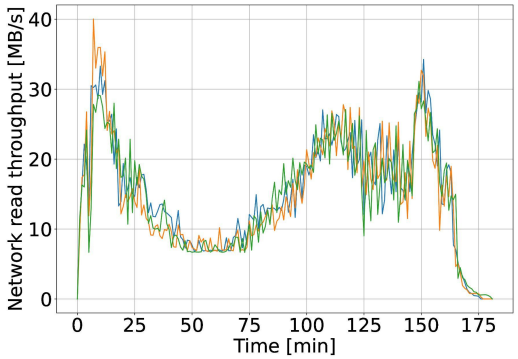
Legacy on the same HTCondor pool Dask is deployed on:

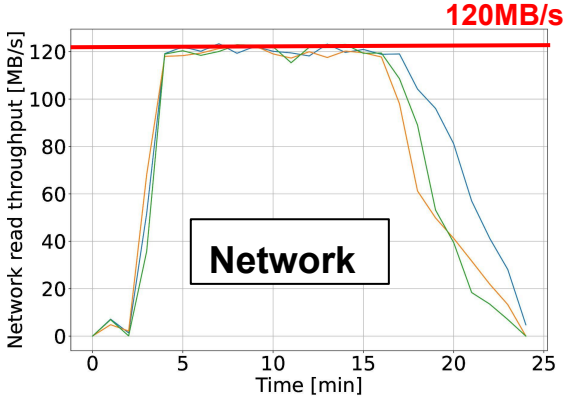
- 3 nodes, each one with 32 logical CPU (16 physical) - 128 GB RAM - 1 Gb/s @ T2_LNL_PD

Performance checks: per node









CPU Usage @RDF limited by network saturation

The network read throughput, which reaches a plateau at 120 MB/s corresponding to the throughput of the network interface on the node, namely 1 Gb/s.

Results & Comparison

Our case study shows a **factor 8 speedup** (a lower limit)

- About 84% reduction of overall execution time
- opening to the possibility of running the analysis in just 1 step?

Overall network read reduction of about 33%.

RDataFrame-based approach outperforms the legacy one in terms of time and event rate in both scenario

Preselection		
	Legacy	**RDF**
Overall time [min]	181 ± 1	23.8 ± 0.6
Overall rate [events/s]	60.5k ± 0.3k	465k ± 11k
Job rate [events/s]	786 ± 12	6915 ± 35
Job event-loop rate [events/s]	858 ± 14	7632 ± 34
Overall network read [GB]	485 ± 1	362.5 ± 0.1
Average RSS per-node [GB]	23.3 ± 0.6	31.3 ± 0.4

Postselection		
	Legacy	**RDF**
Overall time [min]	48.3 ± 0.5	12.6 ± 0.3
Overall rate [events/s]	4.56k ± 0.05k	17.5k ± 0.4k
Job rate [events/s]	62.9 ± 0.1	288 ± 1
Job event-loop rate [events/s]	65.69 ± 0.05	355 ± 3
Overall network read [GB]	84.46 ± 0.08	17.46 ± 0.08
Average RSS per-node [GB]	5.5 ± 0.2	26.7 ± 0.5

spiga@pg.infn.it

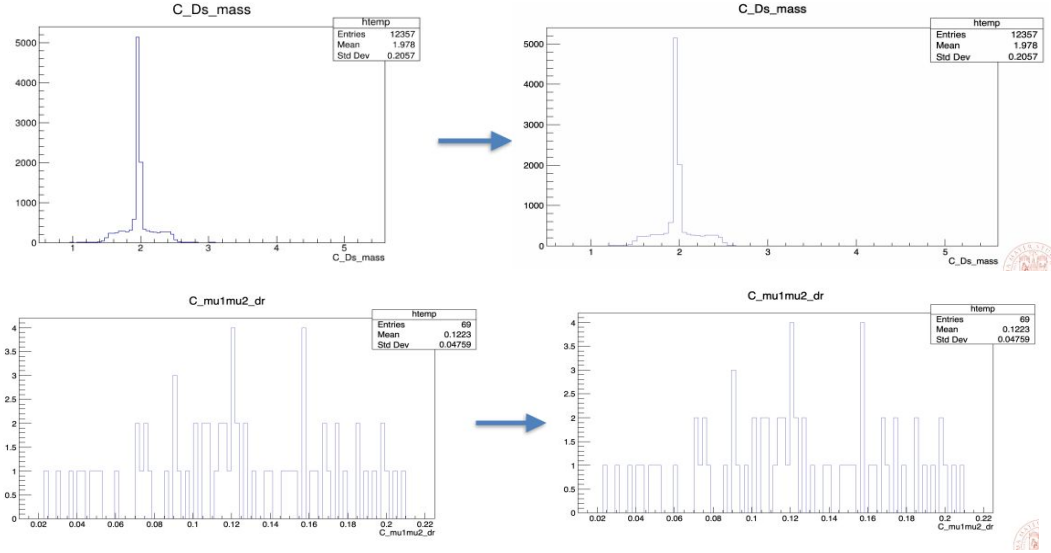
Not only benchmarking

Interests from users is growing and more use cases are coming.

- Recently started an activity for validating

“Heavy Neutral Lepton (HNL) search in D decays”

- Originally developed using RDataFrame, a porting has been performed for its usage with the Dask environment and the usage of RDataFrame distributed



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

Summary and Future

INFN deployed a model for high throughput analysis integrating **Cloud-Native services and offloading to regular WLCG resources** (possibly HPC and other providers).

- Evolution process! Not a Revolution.
- A successful benchmark allowed for the first time interesting comparison of the very same CMS analysis both legacy and an RDataFrame-based approach

Extremely fruitful collaboration between experts with distinct backgrounds and skills! A key to deliver

A playground for CMS for further activities

- Study the **impact on Network/ Storage I/O** (study **Datalake** models)
- Benchmark same systems with **multiple Frameworks (i.e Coffea)**
- Further stimulate NanoAOD adoptions

A R&D platform where to test new technologies

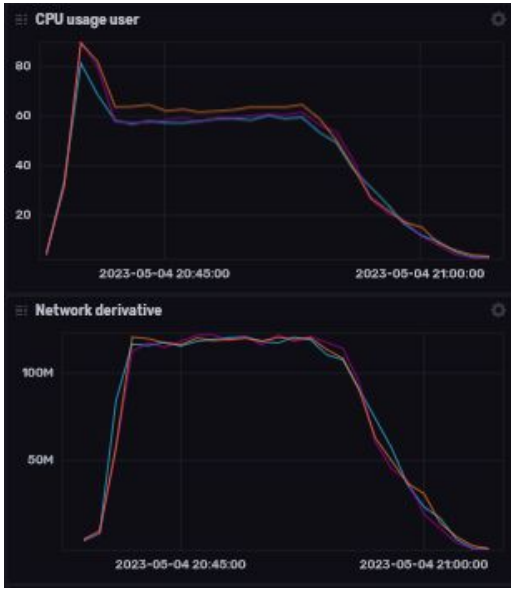
- Evolving the **offloading model toward a Virtual Kublet based solution**
  - **the interTwin EU project (GA. 101058386)**
- Extend the model to other discipline Talk at CHEP: ID 544
  - **ICSC funded by European Union – Next Generation EU".** Talks at CHEP: 114, 497
- **Enhance the National resource federation approach**

spiga@pg.infn.it

Backup

Studies on CPU usage vs network (**preliminary**)

92 Dask workers



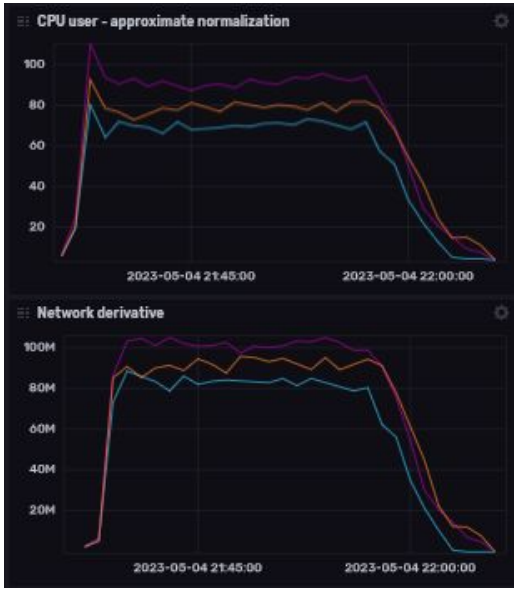
Overall time: 23min 50s

Job rate: 6879.23 Hz

Job Event-loop rate: 7794.12 Hz

Average CPU usage single task: 64 %

46 Dask workers



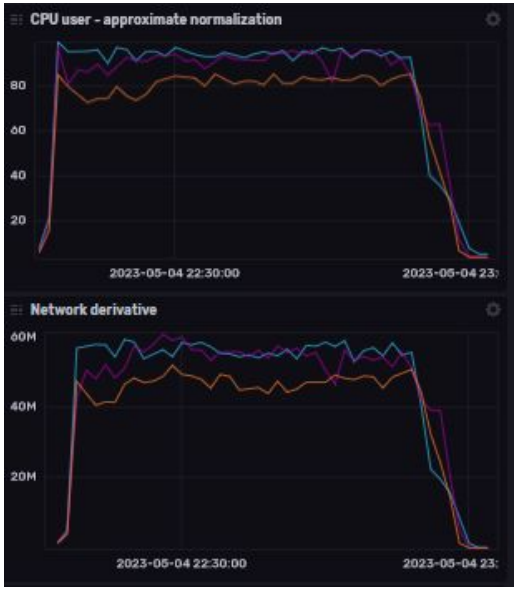
Overall time: 28min 13s

Job rate: 10385.10 Hz

Job Event-loop rate: 11228.41 Hz

Average CPU usage single task: 76 %

23 Dask workers



Overall time: 42min 29s

Job rate: 12171.10 Hz

Job Event-loop rate: 12922.41 Hz

Average CPU usage single task: 83 %

spiga@pg.infn.it