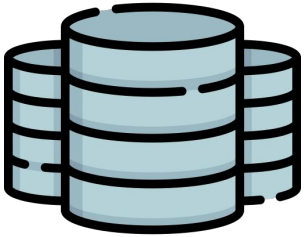
Benchmarking Distributed Analysis at the Jülich HPC Center

Joseph Boulis

Axel Naumann, Maria Girone

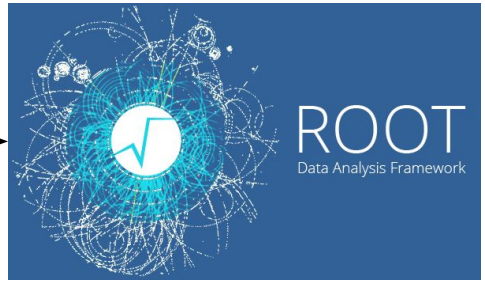
15/08/2023

Introduction & Background



HEP huge amount of data

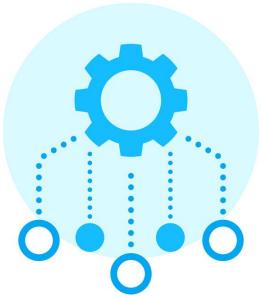
Advancement of grid computing & European HPCs



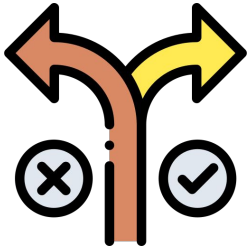
RDataFrame

Why benchmarking?

Areas of improvement

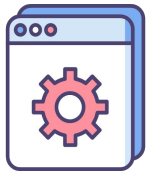


Better Resource utilization



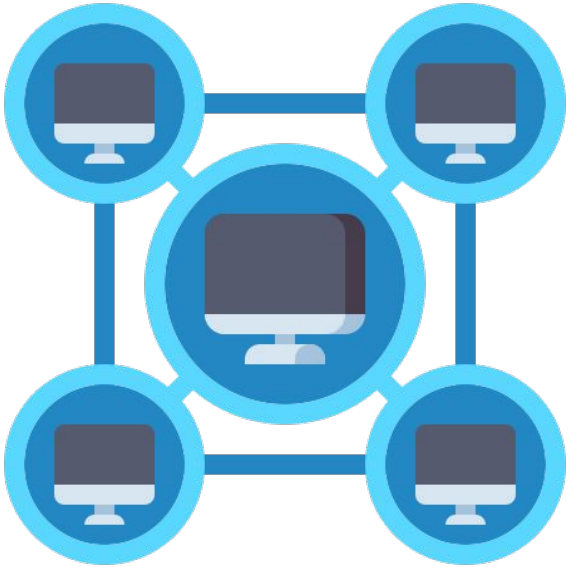
Take informed Decision

The Concept



program.py

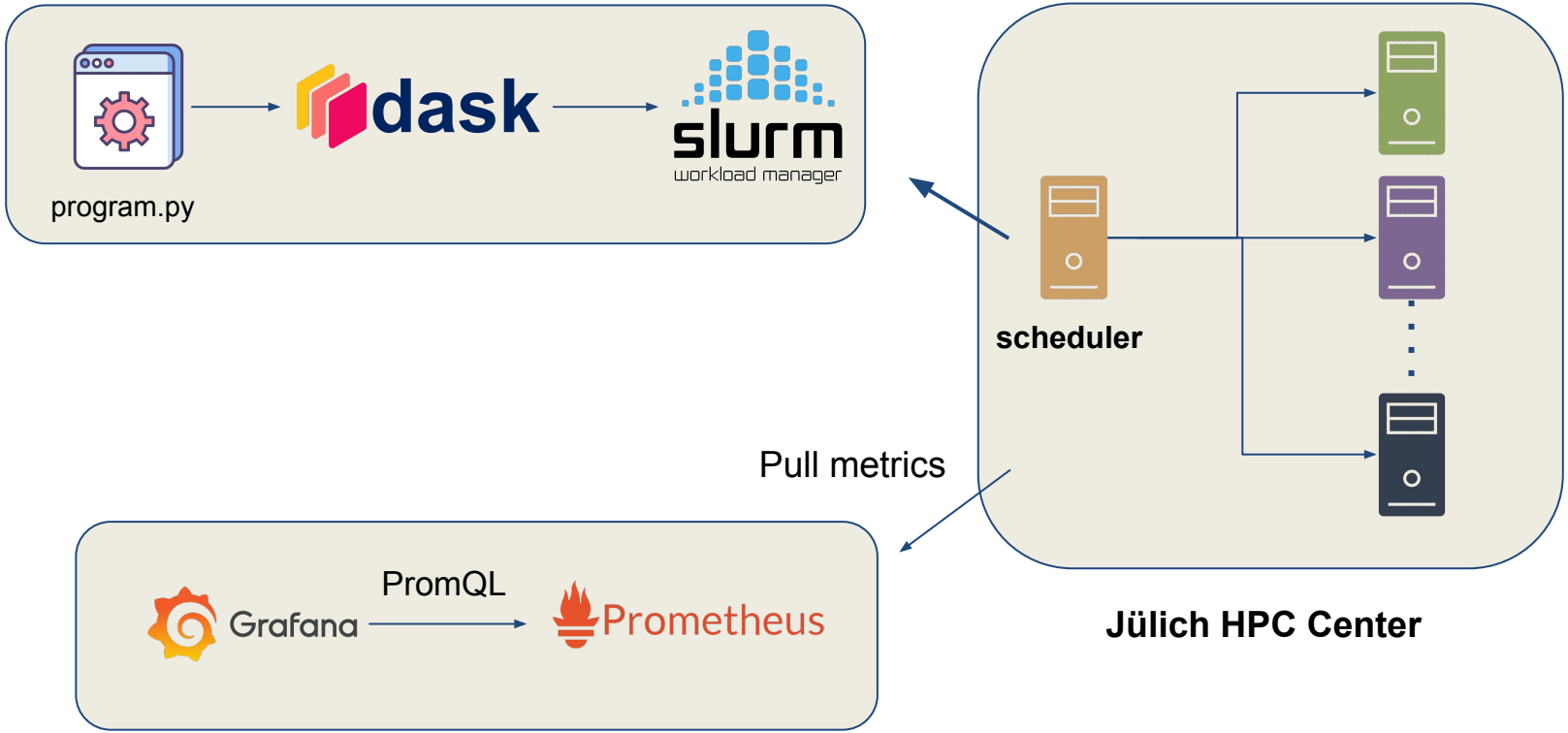
```
df = RDataFrame(...)
```



High Performance Computing Center

Performance Metrics

Architecture



Challenges

Lack of Documentation

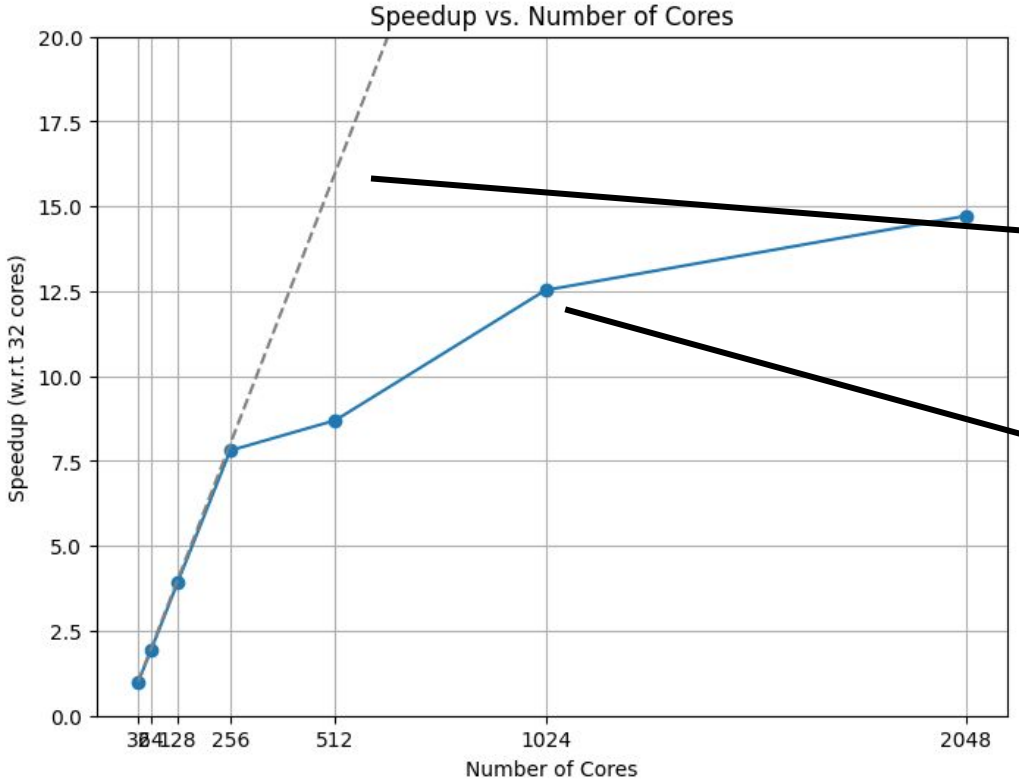
Different HPC Policies

- Limits for job submissions
- Limits for internet access

Time consuming

Installing ROOT Distributedly

Benchmark



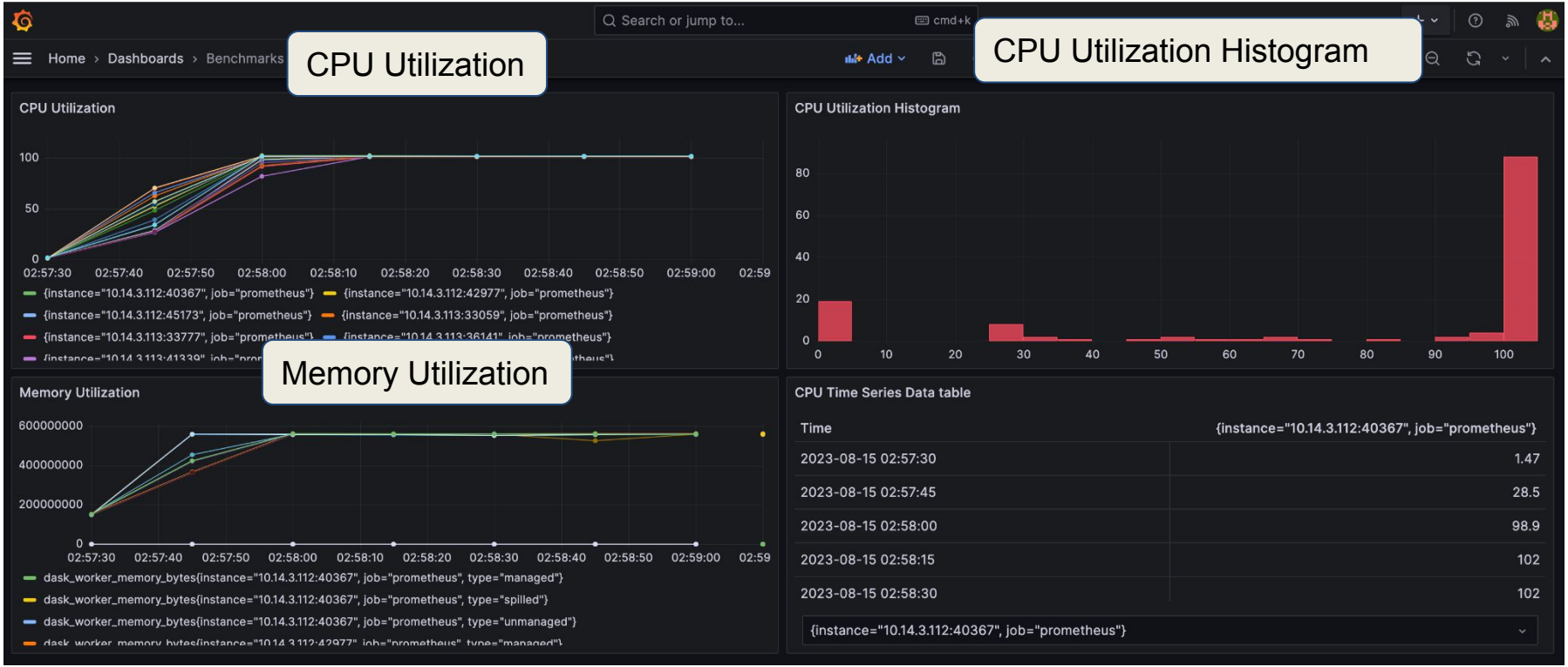
Ideal Case:

- Different physics events are statistically independent.
- Parallelising the computation on different data chunks is a valid approach.

Benchmark:

- 8 TB data on SSD local storage
- Increasing nodes from 2 to 64
- Julich Computing Center
- 1 node = 64 cores

Grafana Dashboard



Future Work

Fetching data remotely

Testing with different data formats (Ex: RNTuple)

Testing on heterogeneous computing resources

Testing on other HPCs (Ex: LUMI)

Thank You!

Questions?

joseph.boulis@cern.ch
josephhany78@gmail.com