Physics Data Processing and Machine Learning in the Cloud

R. Castellotti[1], E. Motesnitsalis[1] , M. Migliorini[1] , L. Canali[1]

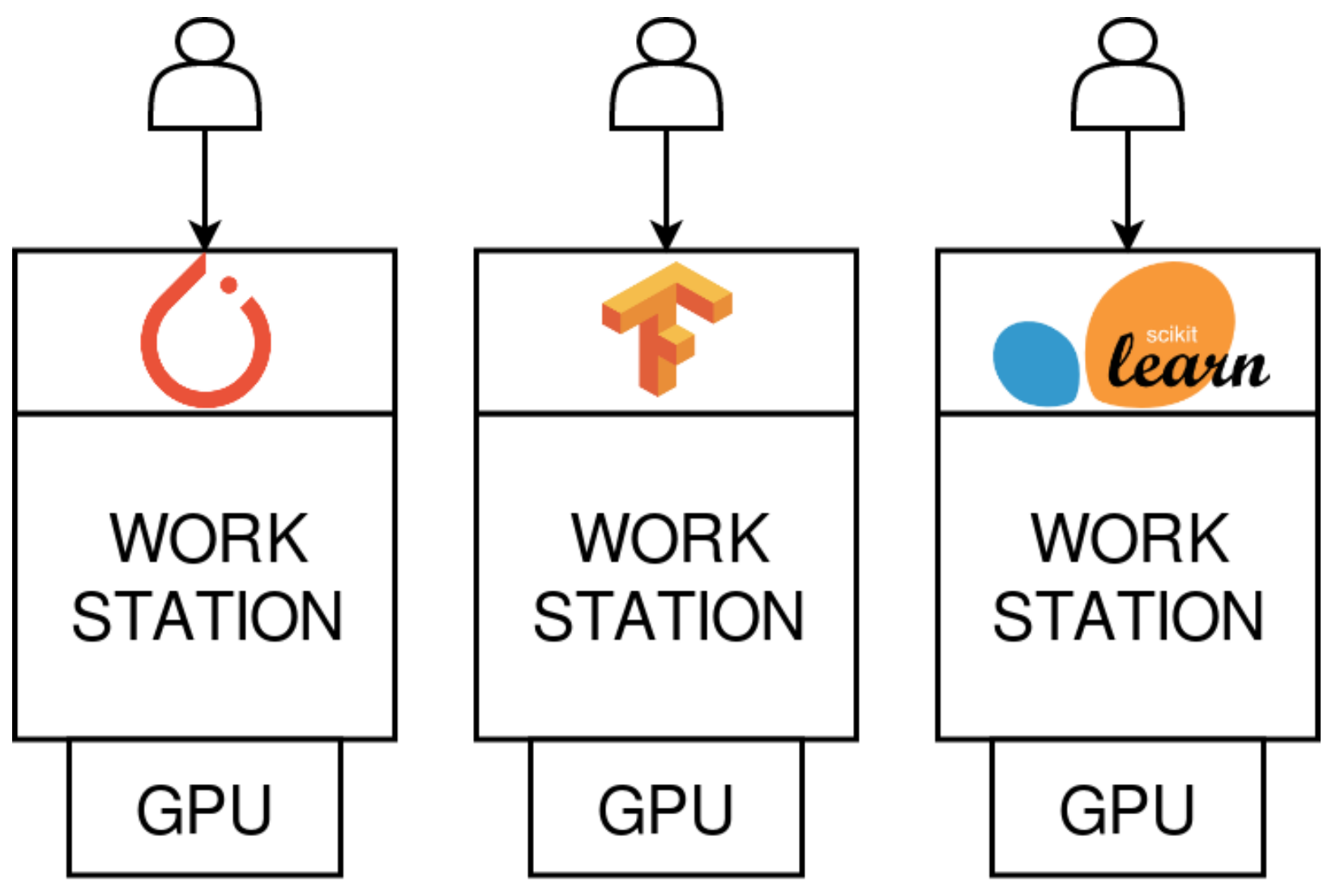
[1]CERN IT; Hadoop, Spark and Streaming service; Geneva; Switzerland

GOALS AND MOTIVATIONS

- The goal of the project is to deploy Physics Data processing and Machine learning on cloud resources, notably using CERN cloud and Oracle Cloud Infrastructure (OCI)
- We want to adopt the latest developments in the Big Data ecosystem in order to profit from the work done in industry and academia
- We have identified, with the user community, real world workloads that we want to run at scale and/or with hardware accelerators

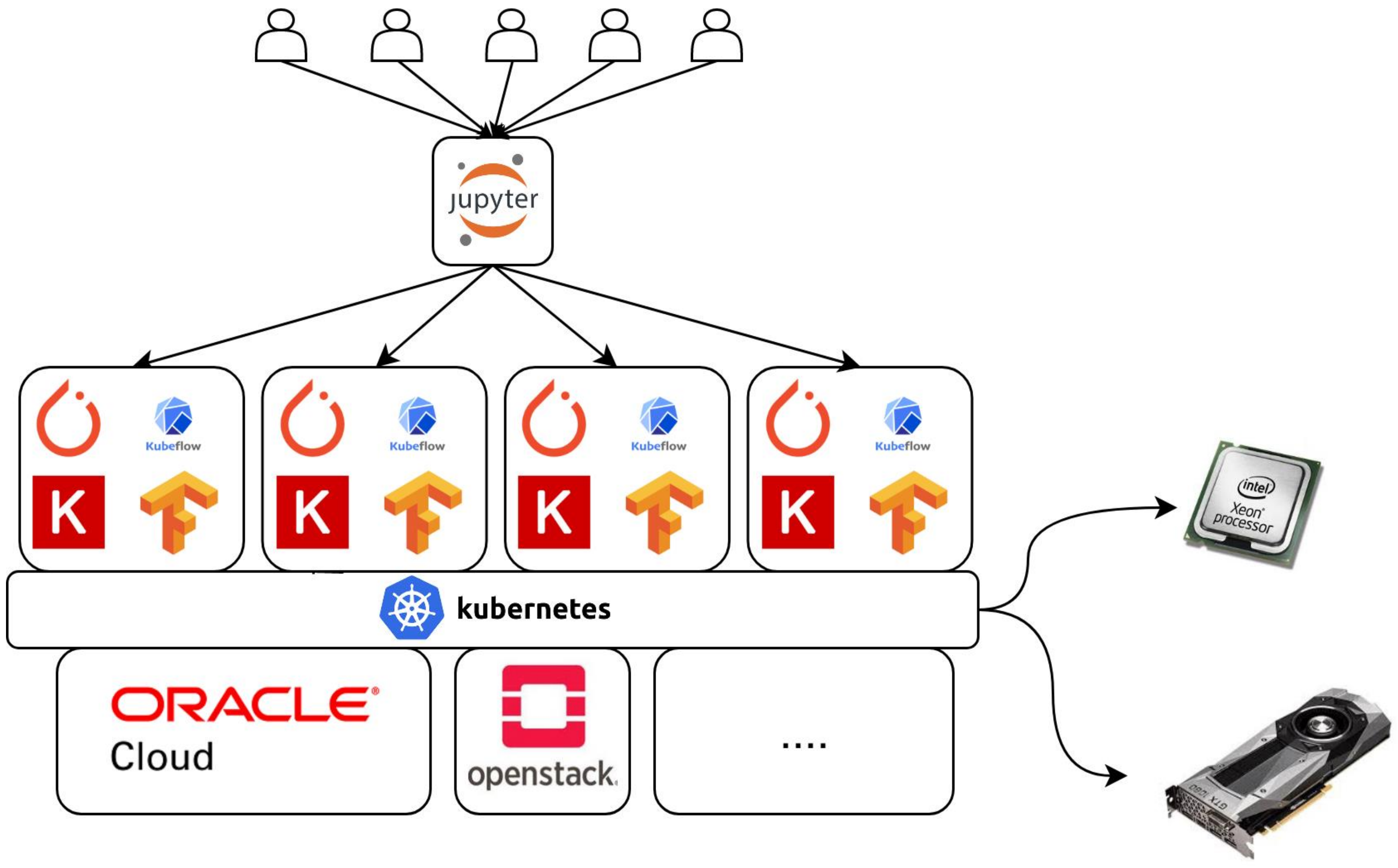
EXISTING PROCESSES

- The vast majority of high energy physics analysis is done in a specialized environment developed over the years by the HEP community
- The ML jobs are often run on workstations managed by every research group, which leads to increased technical overhead



Currently, every user has his own custom software and physical deployment for ML

ANALYTICS PLATFORM IN THE CLOUD

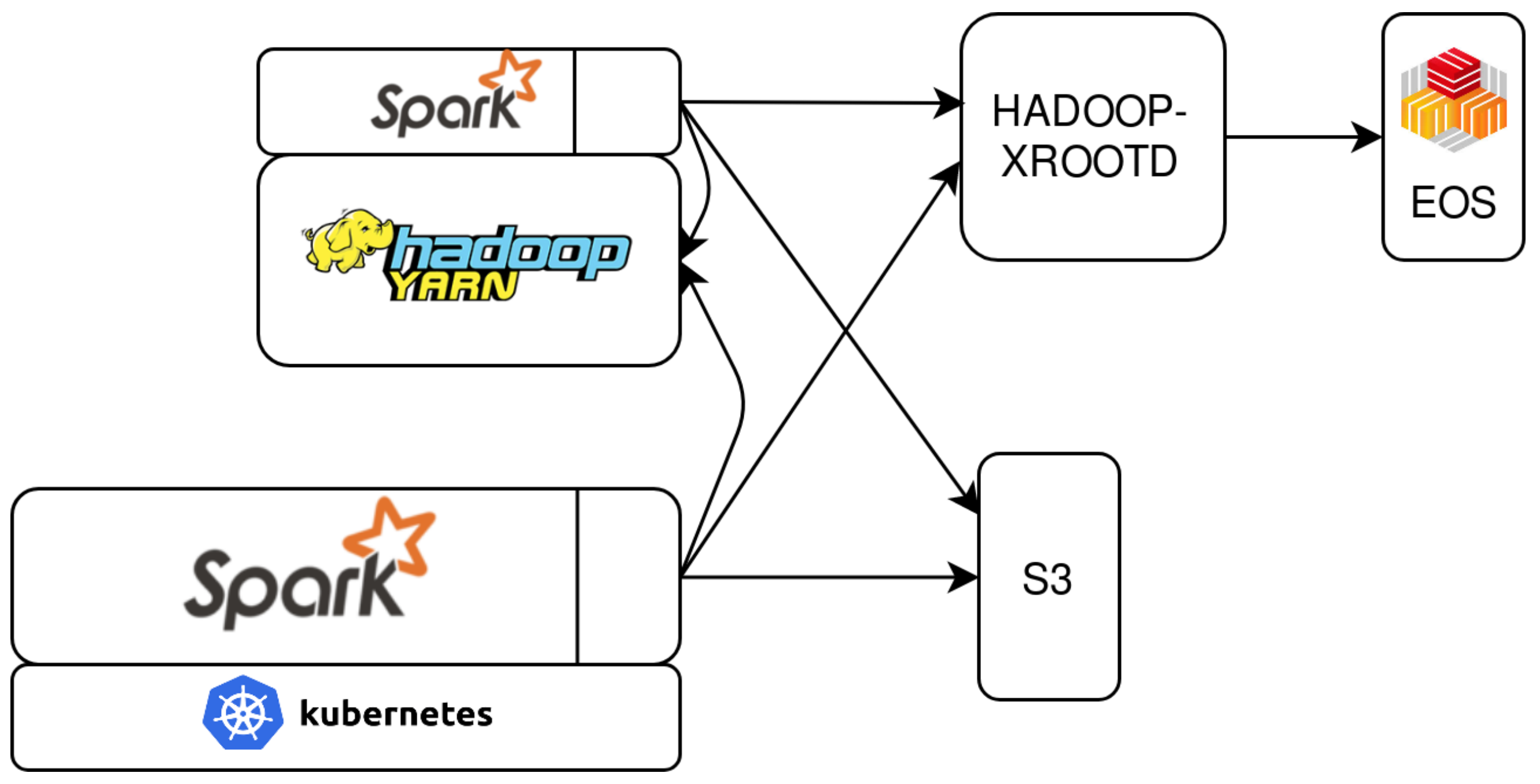


- Scale rapidly from prototype
- Share a cluster regardless of where it is deployed and of the chosen software stack
- Fully featured environment, allows users to focus on analysis rather than data engineering and work collaboratively

SEPARATION OF COMPUTE AND STORAGE

- A cloud-native architecture for data analysis means separating the data storage and computing
- This pattern suits well with CERN, as the physics data is already stored in a dedicated environment, EOS
- Object storage APIs in cloud infrastructure are standardized on Amazon S3 and are implemented both at CERN and in Oracle Cloud as well as in most commercial providers
- CPU resources can be scaled independently from storage using cloud-native solutions like Kubernetes

APACHE SPARK ON-PREMISES AND IN THE CLOUD



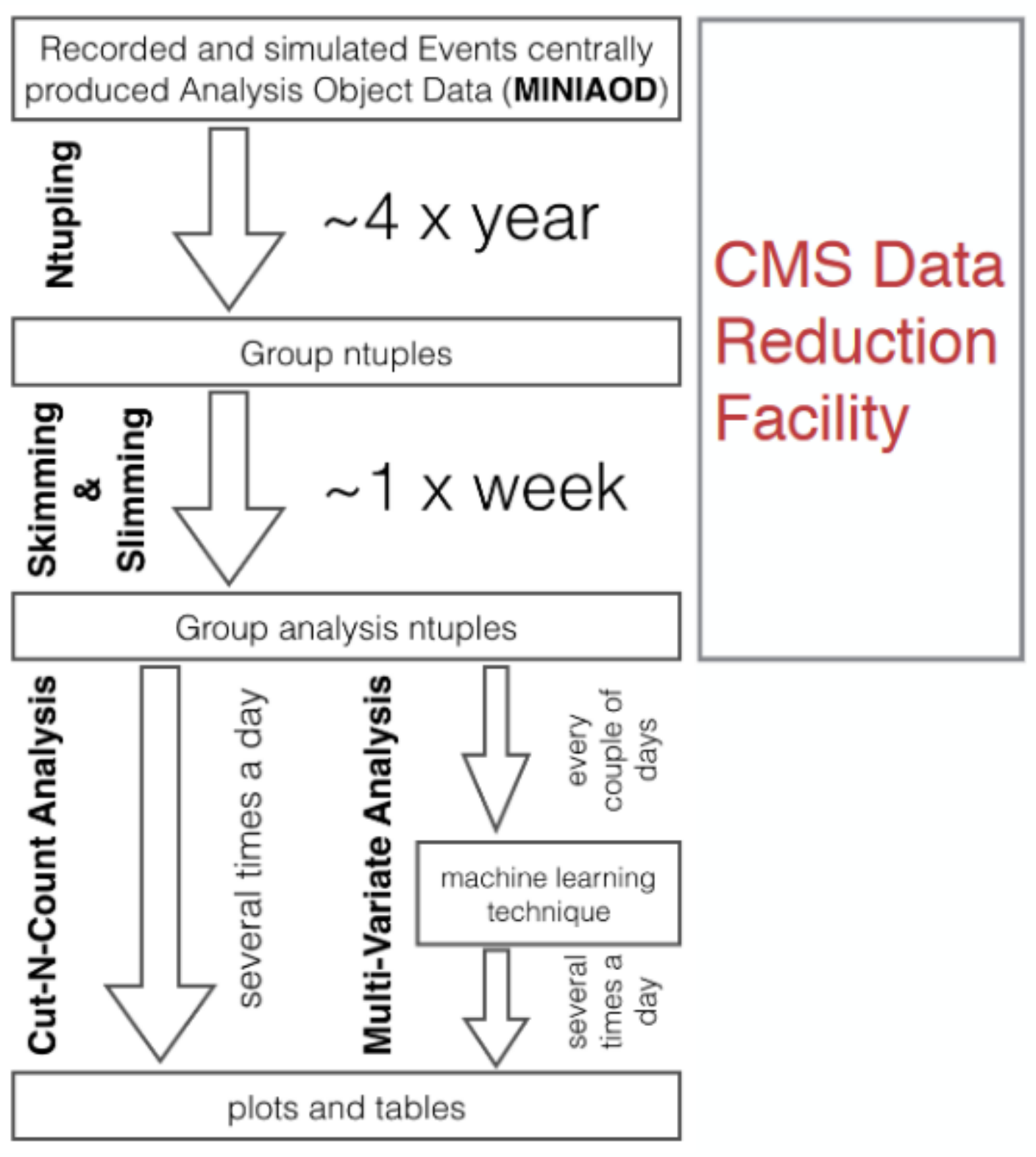
Two different ways to run Spark:

- in a fixed-size Hadoop cluster where data processing happens on the same nodes storing the data
- in a scalable, compute-only Kubernetes cluster with remote data access

PHYSICS USE CASES

- Two existing use cases have been adapted to run on Kubernetes using resources in CERN cloud:
  - Data reduction from CMS Big Data project performing event selection and dimuon invariant mass calculations on 1 PB of data read from EOS
  - An event classifier ML pipeline replicating the work done in “Topology classification with deep learning to improve real-time event selection at the LHC” (https://arxiv.org/abs/1807.00083)

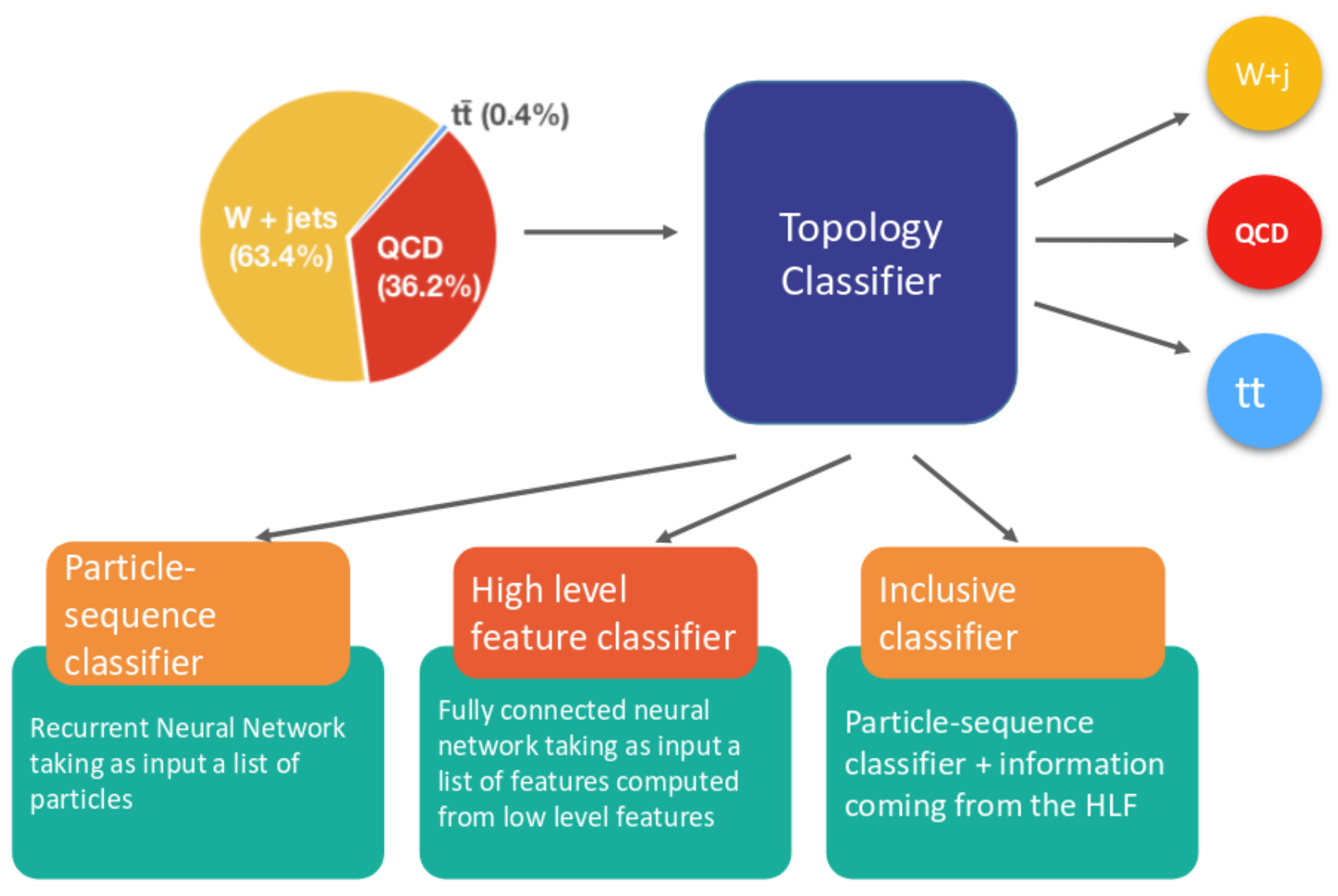
CMS data reduction



# executors	Elapsed time
8	2h52
16	1h11
32	0h37
64	0h26

Above: The architecture of the CMS data reduction facility
Below: execution times of the CMS data reduction workload on a 22TB data sample; every executor process ran on Kubernetes using 4 cores and 4.5 Gb of memory

ML pipeline for an event classifier


tt̄ (0.4%)
W + jets (63.4%)
QCD (36.2%)
Topology Classifier
W+j
QCD
tt
Particle-sequence classifier
Recurrent Neural Network taking as input a list of particles
High level feature classifier
Fully connected neural network taking as input a list of features computed from low level features
Inclusive classifier
Particle-sequence classifier + information coming from the HLF

# executors	Elapsed time
1	4m06
2	2m26
3	2m04
4	1m51

Above: conceptual schematic showing different event classifier implementations
Below: execution times of the training for the HLF classifier on a reduced dataset of 8000 events; every executor had 4 cores and 4.5 Gb of memory

Additional details on both the use cases are available in the posters:

- Physics Data Analysis and Data Reduction with Apache Spark
- Machine Learning Pipelines with Apache Spark and Intel BigDL

FUTURE STEPS

- Run on Oracle Cloud the workloads developed for physics data reduction and ML on physics data highlighted above
- Investigate performance bottlenecks at massive scale and develop optimizations
- Provide Oracle Cloud teams with feedback on functional and performance tests that we will implement
- Include support for GPUs, which are a key requirement for many Deep Learning tasks