Online data processing in ALICE

B. von Haller

CERN

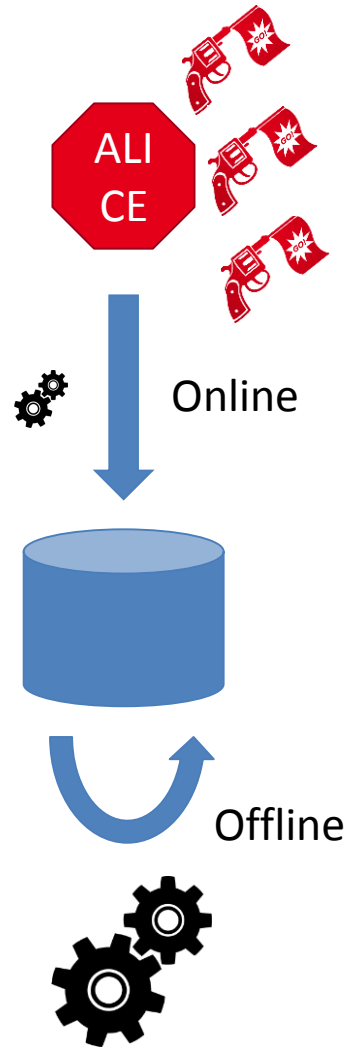
23.03.2017

Overview

- Context
  - Major upgrade of ALICE for Run 3 and 4 (i.e. post 2020)
  - O^2 project : a common online/offline computing system
- Outline
  - Rationales for O2 and online data processing
  - Requirements, architecture and design
  - Data quality control

Today's system

- Triggered
  - only a small amount of events is actually recorded
- Limited online processing
  - the High Level Trigger processes the data of the largest detector
  - a few calibration are done online without human intervention
- Full reconstruction and calibration is done offline, in the days, weeks and months after the data taking



Rationales for a new computing system

- After LS2, LHC min bias PbPb at 50 kHz
  - ~100 x higher event rate than during Run 1
  - ➔ Too much data to be stored
- Physics topics addressed by ALICE upgrade
  - Rare processes, very small signal over background ratio
  - Needs large statistics of reconstructed events, 13nb-1 for PbPb
  - ➔ Triggering techniques very inefficient or impossible in most cases
- TPC inherent rate (drift time ~100 µs) < 50 kHz
  - ➔ Support for continuous read-out, as well as triggered read-out

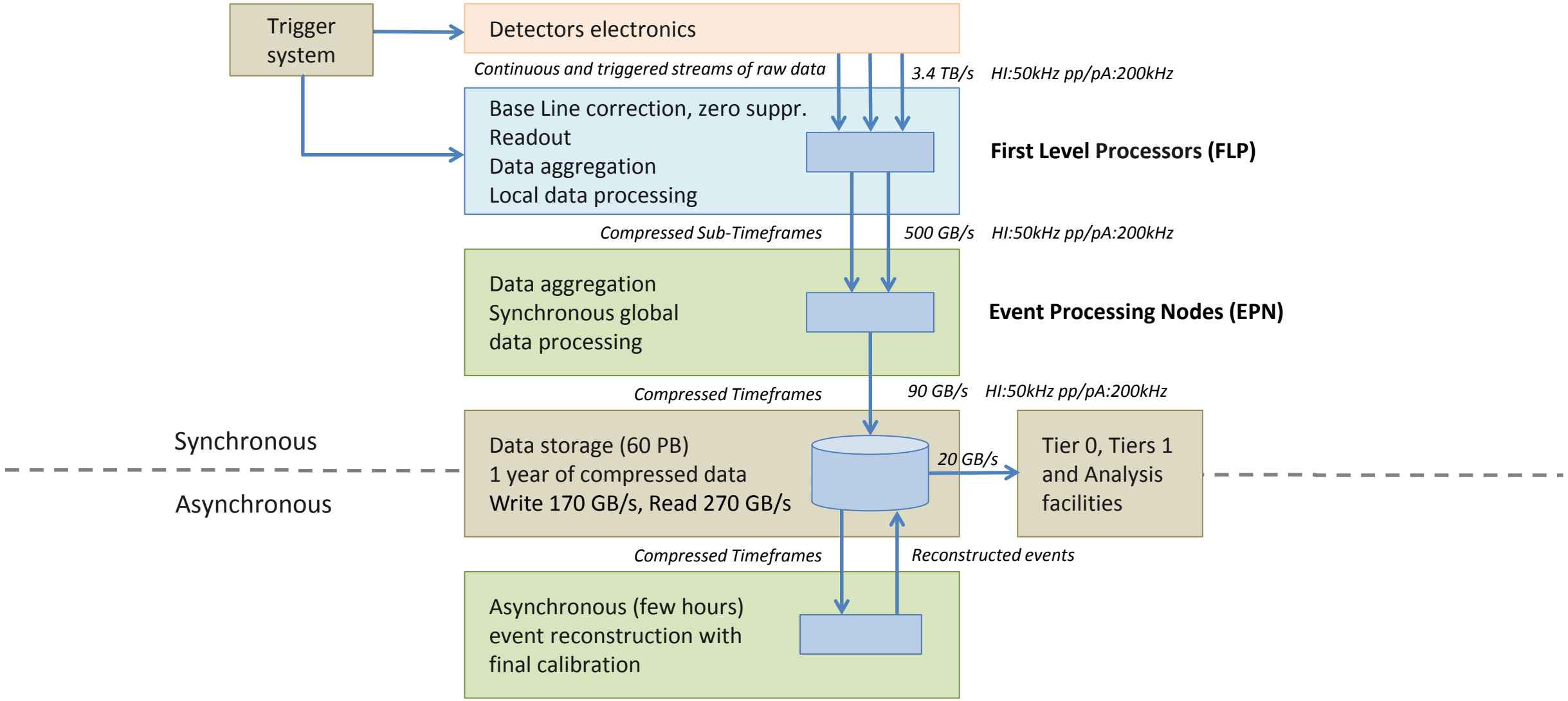
Tomorrow's system

- Read-out the data of all collisions
- Compress these data intelligently by reconstructing and calibrating them online
- One common online-offline computing system: O^2

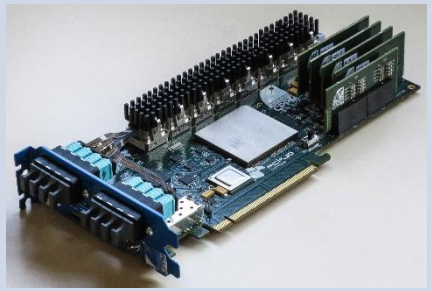
➔Paradigm shift compared to today's approach

➔But built on our experience with the HLT

Functional flow

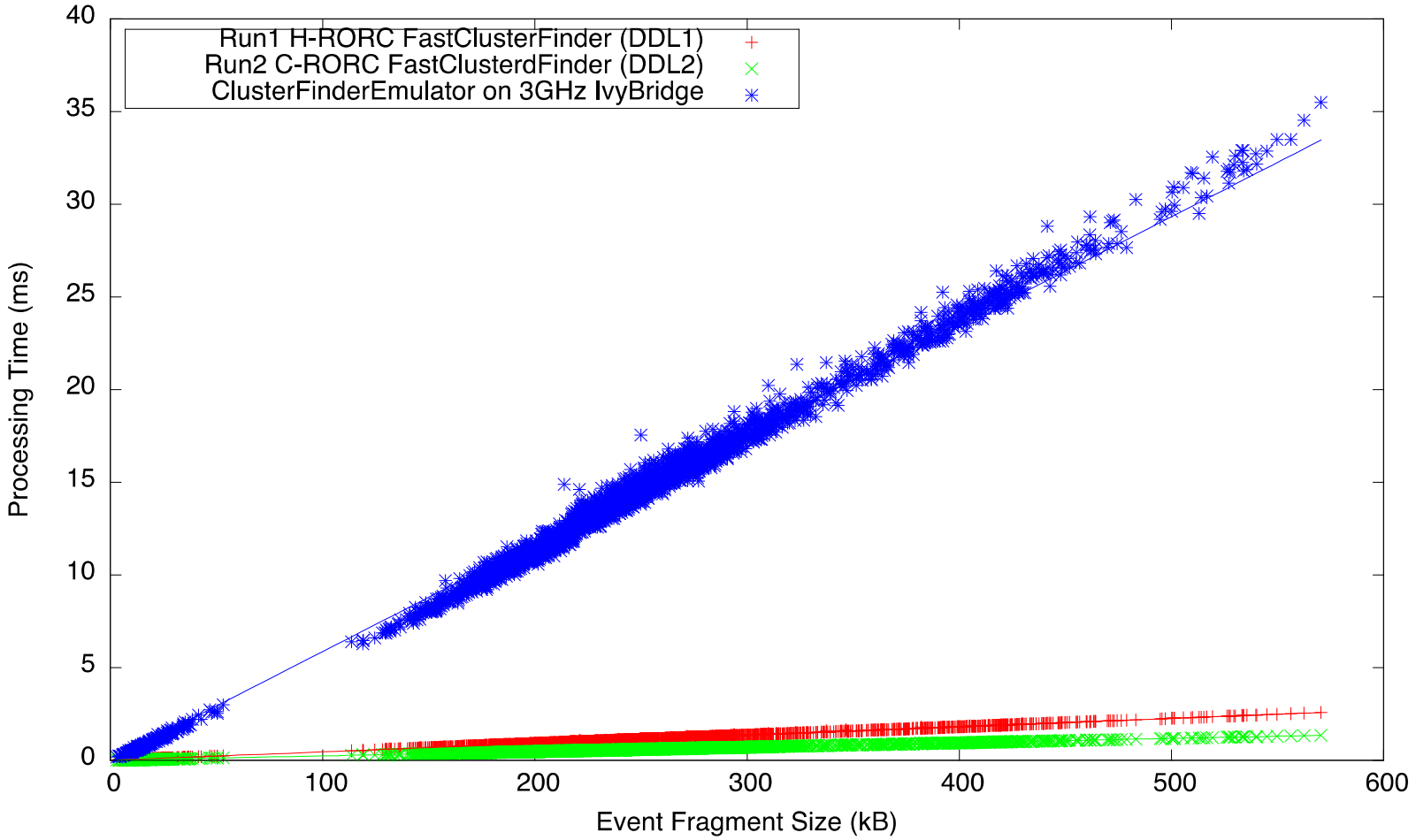


Readout & FPGA Hardware acceleration

RORC 1	C-RORC	CRU
		
2 ch @ 2 Gb/s PCIe gen.1 x4 (1 GB/s)	12 ch @ up to 6 Gb/s PCIe gen.2 x 8 (4 GB/s)	24 ch @ 5 Gb/s PCIe gen.3 X 16 (16 GB/s)
Custom DDL protocol	Custom DDL protocol (same protocol but faster)	GBT
Protocol handling TPC Cluster Finder	Protocol handling TPC Cluster Finder	Protocol handling TPC Cluster Finder Common-Mode correction Zero suppression

Run 1 | LS1 | Run 2 | LS 2 | Run 3

Hardware acceleration (FPGA)



Performance of the FPGA-based FastClusterFinder algorithm for DDL1 (Run1) and DDL2 (Run2) compared to the software implementation on a recent server PC.

Computing requirements for online processing

Computing requirements -> Total : ~ 100000 CPU cores 5000 GPU chips

Detector	Process	Processing requirement [CPU cores or GPUs.]	Processing Platform	System reference
TPC	Calibration	1000	CPU	Intel I7-4600U 2.70 GHz
TPC	Track seeding, following	5000	GPU	AMD S9000
TPC	Track merging, fitting	15000	CPU	Intel I7-980X 3.60 GHz
ITS	Tracking	75000	CPU	Intel I7-2720QM 2.20 GHz
MCH	Preclustering	200	CPU	Intel I7 2.30 GHz
MCH	Clustering	5000	CPU	Intel I7 2.20 GHz

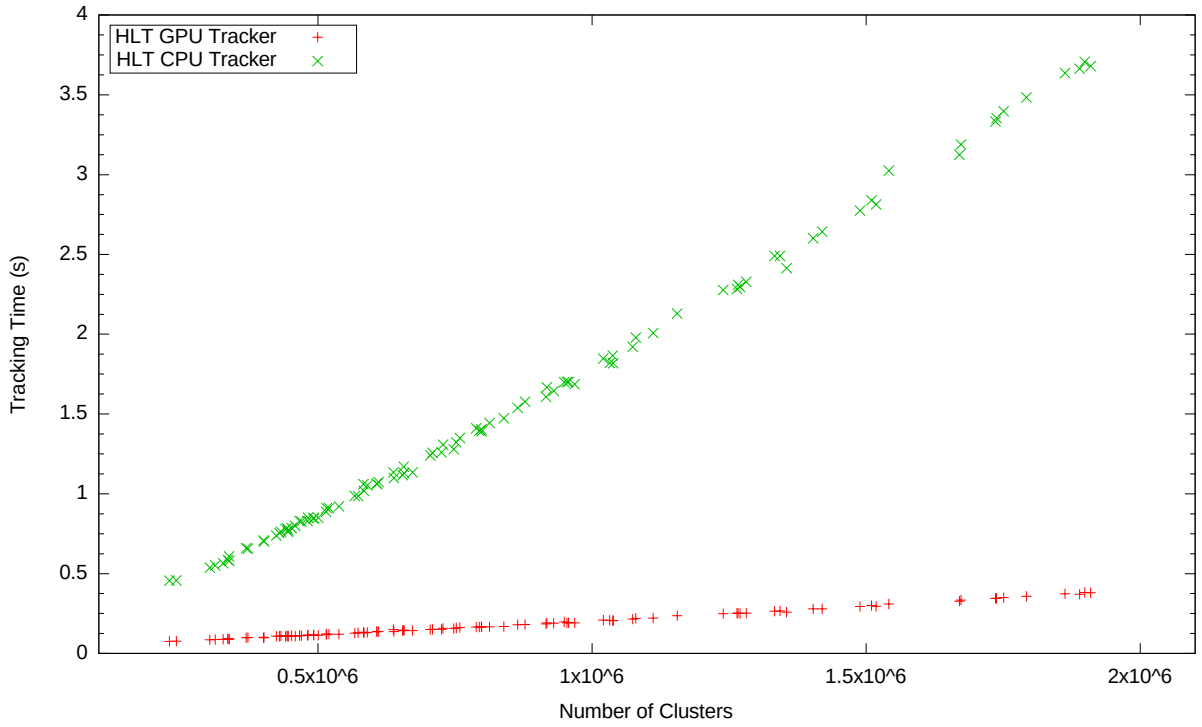
Goes together, merging and fitting can run on GPUs too

Theoretically could run on GPU

Being ported to GPU, conversion factor unknown

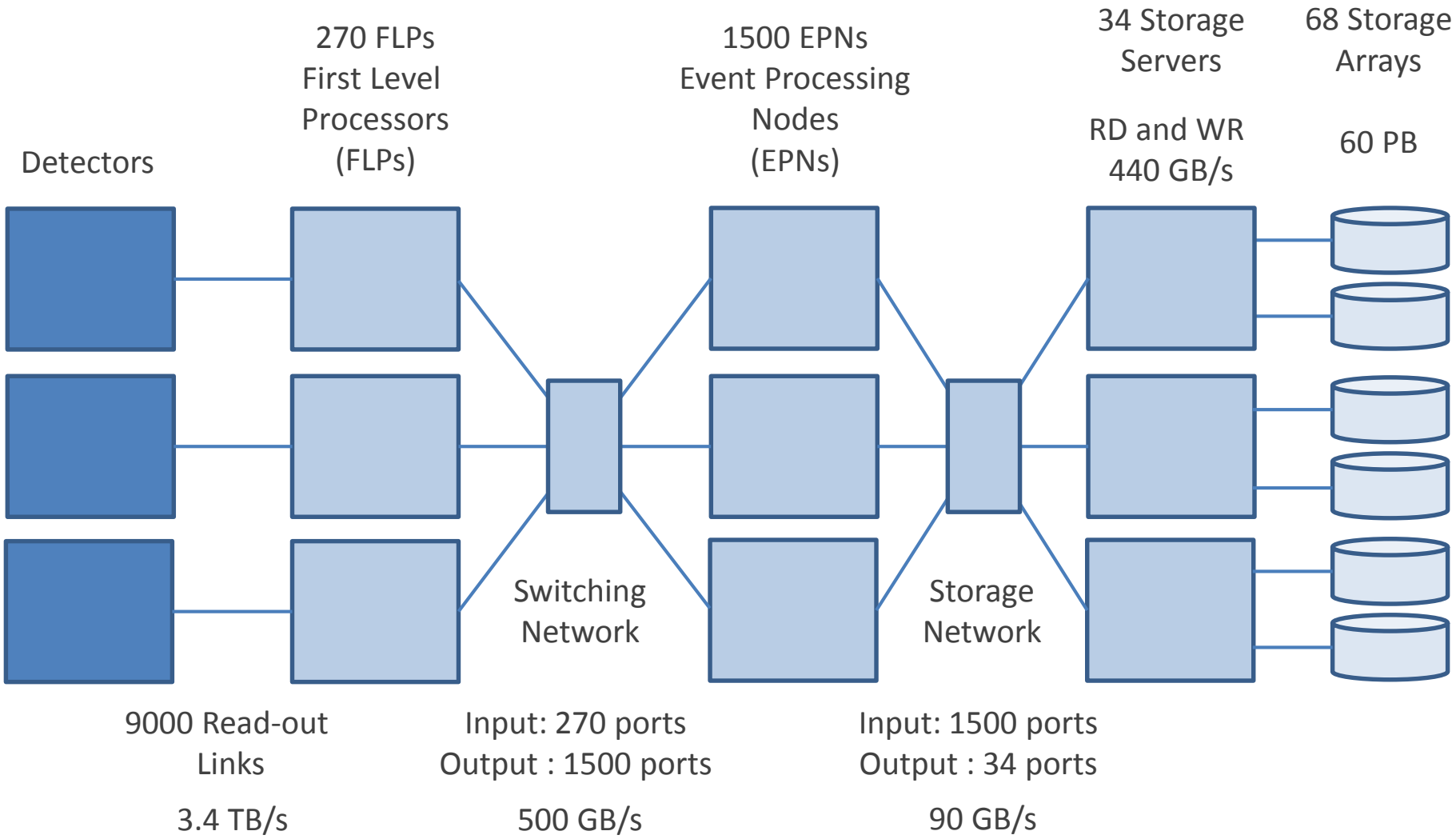
Hardware acceleration (GPU)

- For TPC track finding on EPNs (as today's HLT)
- Possibly more use cases depending on R&D



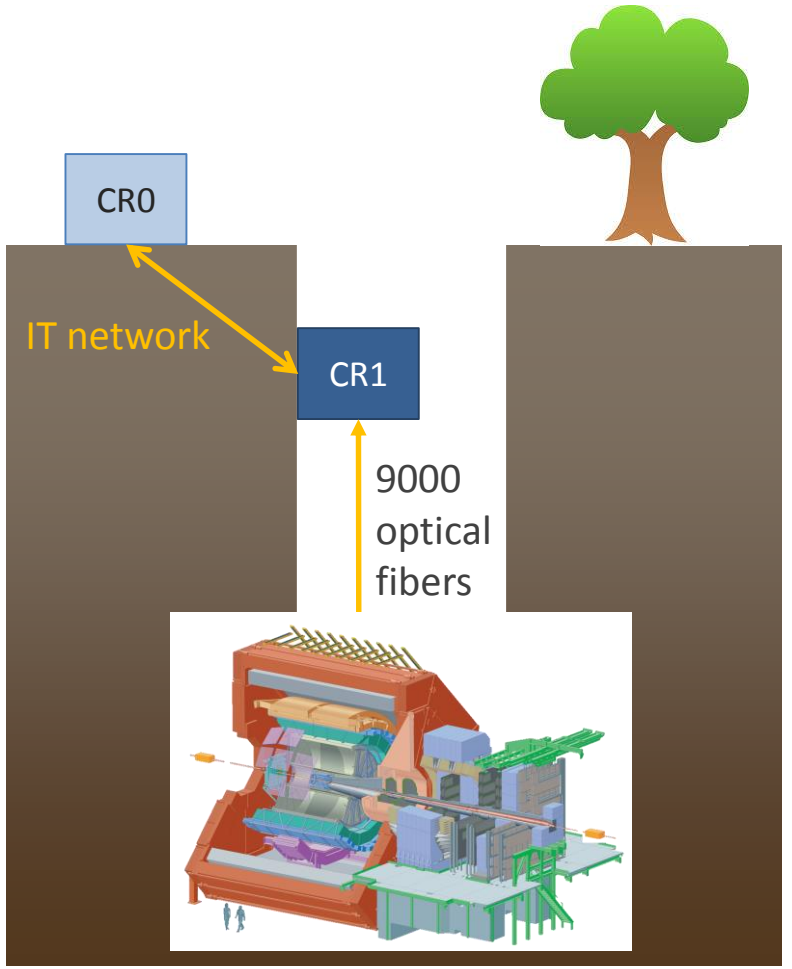
Tracking time of HLT TPC Cellular Automata tracker on Nehalem CPU (6Cores) and NVIDIA Fermi GPU.

O2 Hardware facility



O^2 Farm

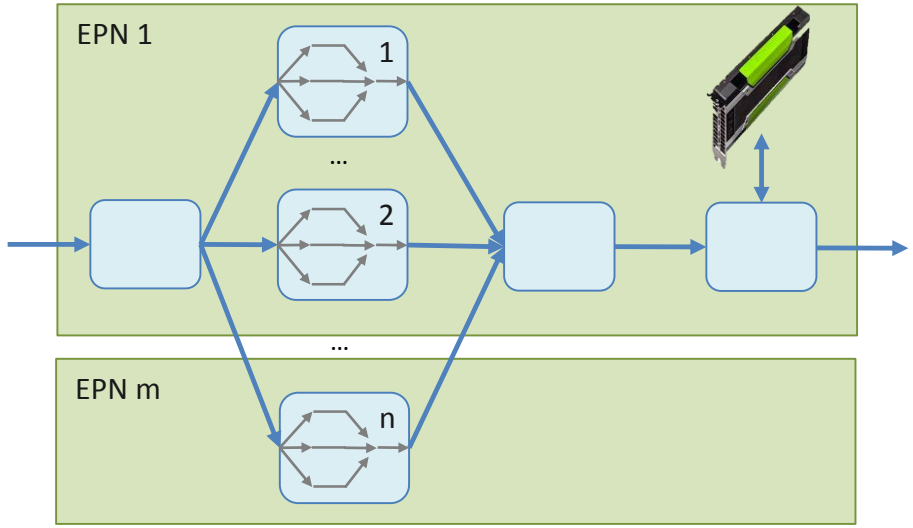
- ~100k CPUs, ~5k GPUs, ~500 FPGAs
- FLPs at P2 in existing CR1
- EPNs and storage need a new dedicated room
  - Space and weight limitations
- Two scenarios
  - CR0
    - Container(s)
    - Call for tender (common with LHCb and neutrino platform)
  - Common Data Center in Prevessin
    - An alternative to the CR0 at P2 has been proposed by CERN
    - New common data center in Prevessin
    - Being studied

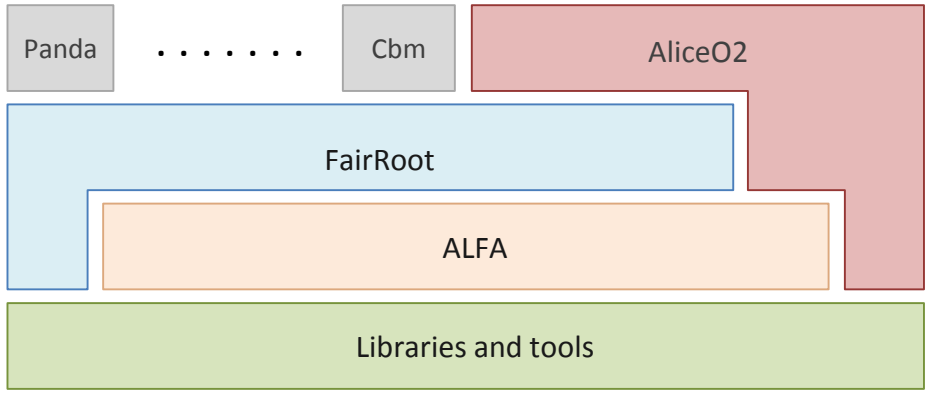


Software for online processing

Design

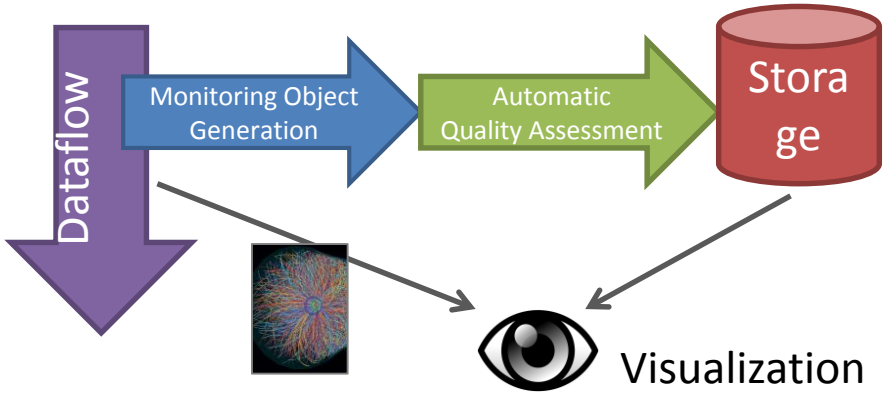
- Message-based multi-processing
  - Ease of development
  - Ease to scale horizontally
  - Possibility to extend with different hardware
  - Multi-threading possible within processes
- ALFA : ALICE-FAIR concurrency framework
  - Data transport layer
  - ZeroMQ
  - Multi-process
  - First version available, development ongoing
- AliceO2
  - Prototyping
  - Steady started



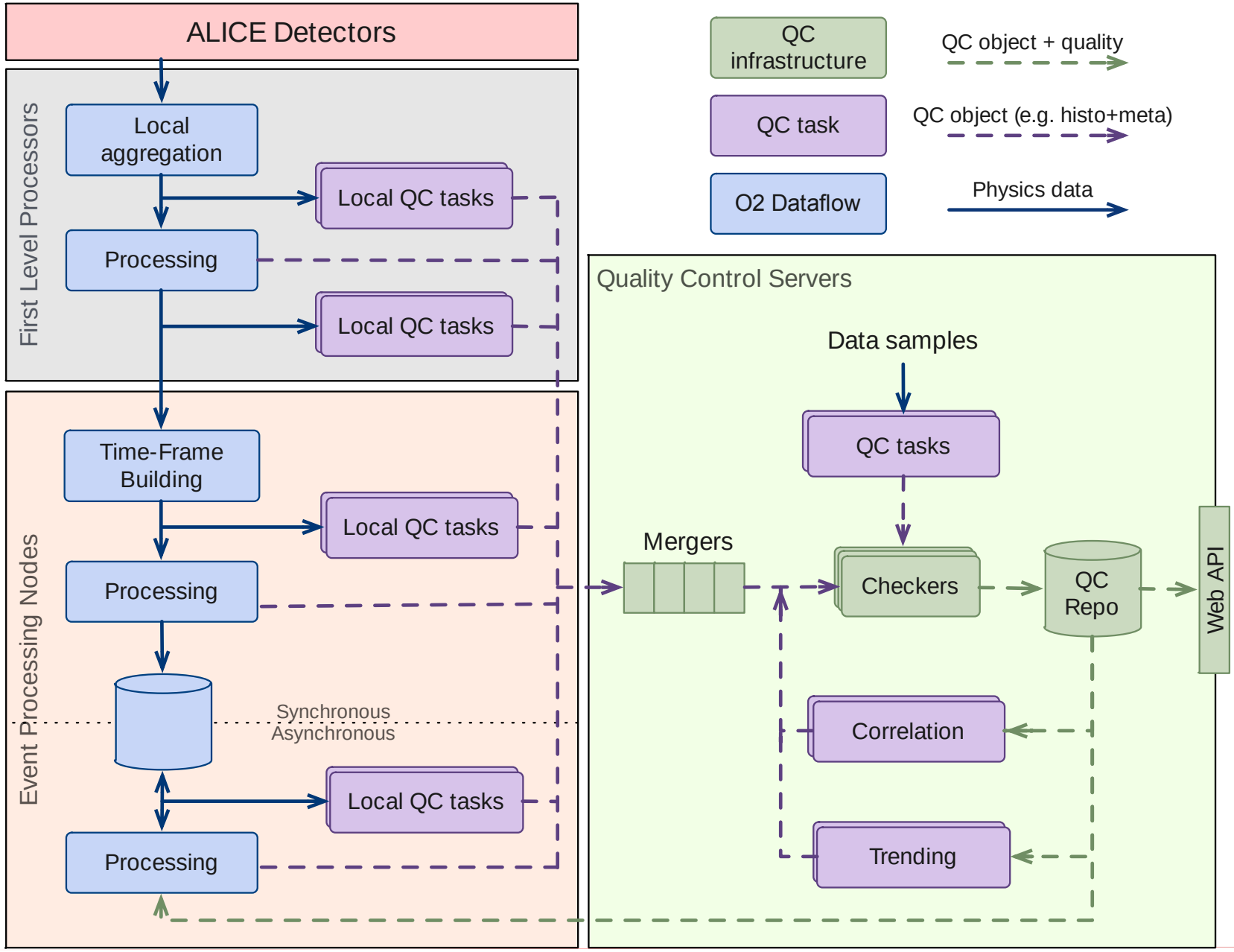


Data Quality Control

- Run 3 Data Quality Control (QC) combines
  - Data Quality Monitoring (online)
    - Make sure to **record** and **reconstruct** high quality data
  - Quality Assurance (offline)
    - Make sure to **reconstruct** and **analyze** high quality data



- Crucial because we do a lossy compression
  - Get feedback on the quality of the data and the processing
  - Identify and solve issues early
    - With or without human interventions
  - Help streamline the processing by identifying ”good” data
- Without QC we are blind


ALICE Detectors
First Level Processors
Local aggregation
Local QC tasks
Processing
Local QC tasks
Event Processing Nodes
Time-Frame Building
Local QC tasks
Processing
Synchronous
Asynchronous
Local QC tasks
Processing
QC infrastructure
QC task
O2 Dataflow
QC object + quality
QC object (e.g. histo+meta)
Physics data
Quality Control Servers
Data samples
QC tasks
Mergers
Checkers
QC Repo
Web API
Correlation
Trending

QC requirements

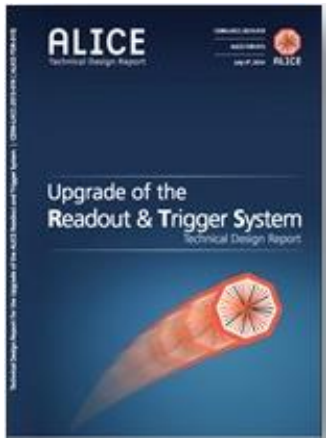
- Based on 2 detailed surveys and on our experience
- 100 tasks (most are parallelized over 100s of machines)
- Analysing 1-100% of the data
  - Possibly in stage, not everything synchronously with data
- 10’000 objects (mostly histos) after merging, updated every minute
  - We actually expect 25’000 objects and plan for peaks of 50’000
- 5% of objects to be kept forever, all the rest kept for days or weeks
- Short feedback loop (seconds) with initial setup within minutes
- Automatic as much as possible, machine learning ideally

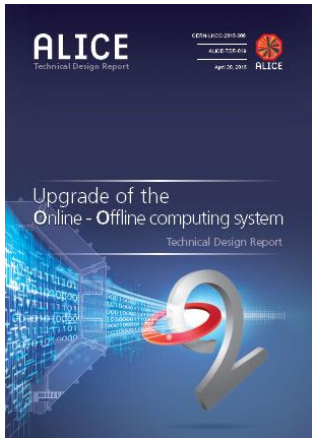
Summary

- O2 is a project with ambitious requirements
  - > 3.4TB/s detector input, ~100x more than today
  - **Online synchronous compression** factor of >30
  - Major paradigm change with **combined offline and online** system
- Hardware
  - HW acceleration (FPGAs, GPUs) for online processing
  - O2 farm with ~100 k CPU cores and ~5000 GPUs
- Software
  - Multi-processes + multi-threaded
  - Data Quality Control is both crucial and challenging

O^2 Schedule

2012-2013	2014-2015	2016-2017	2018-2019	2020-2021
High-Level Design R&D Trigger TDR Project organization	Design R&D Demonstrators O^2 TDR Products selection	Detailed design R&D Prototyping Development Products selection Prototypes	Detailed design R&D Prototyping Development Products selection Prototypes Final components Deployment Commissioning	Development Products selection Final components Deployment Commissioning Production

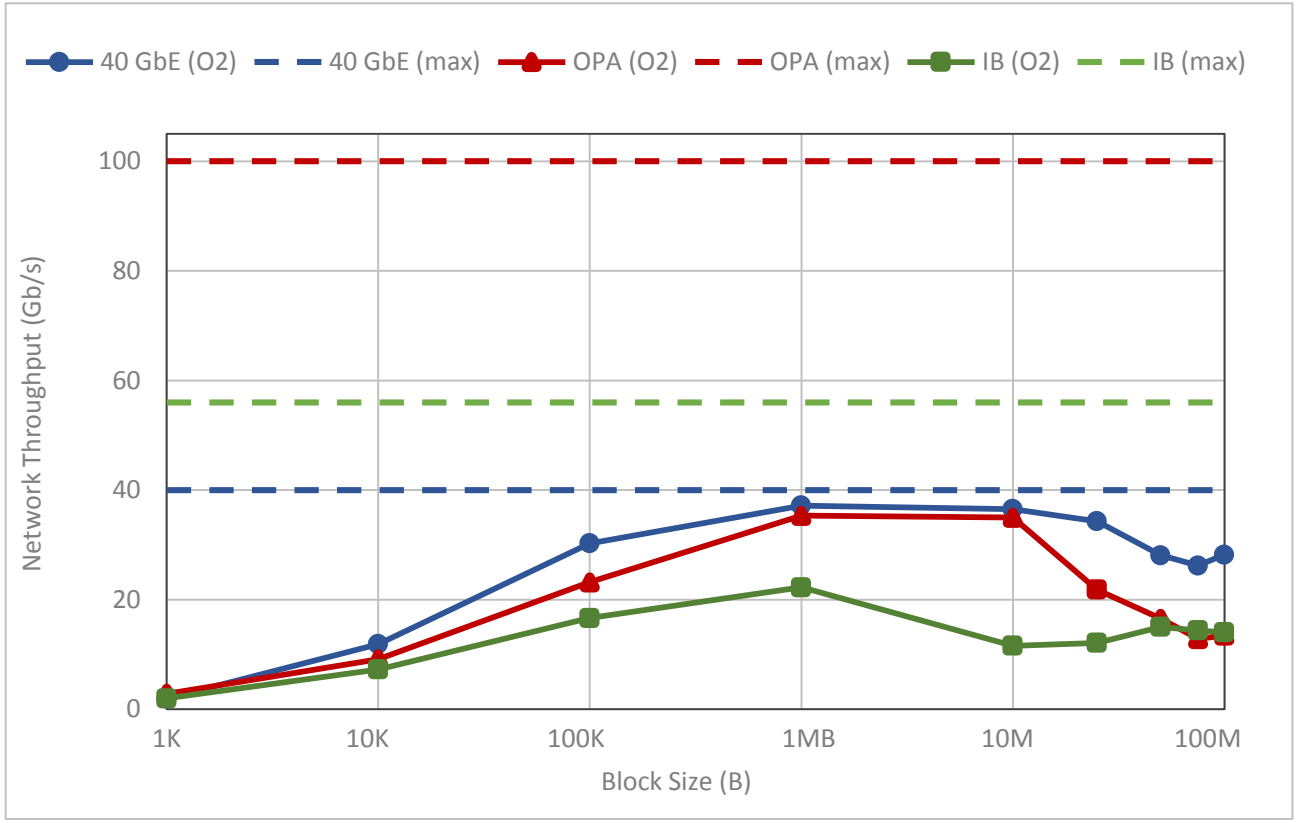




Run 1	LS1	Run 2	LS 2	Run 3

Network performance tests

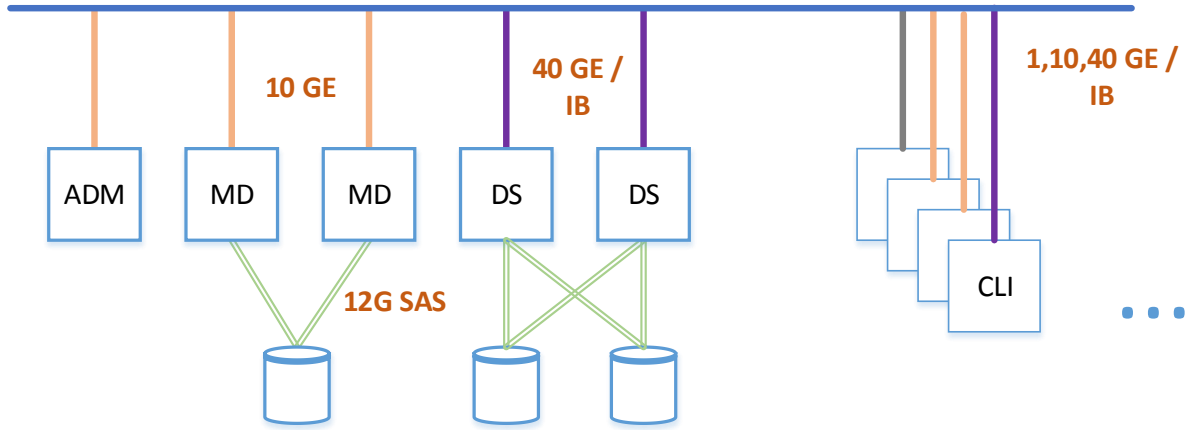
Comparison of Ethernet,
IP over InfiniBand and IP over Omni-Path



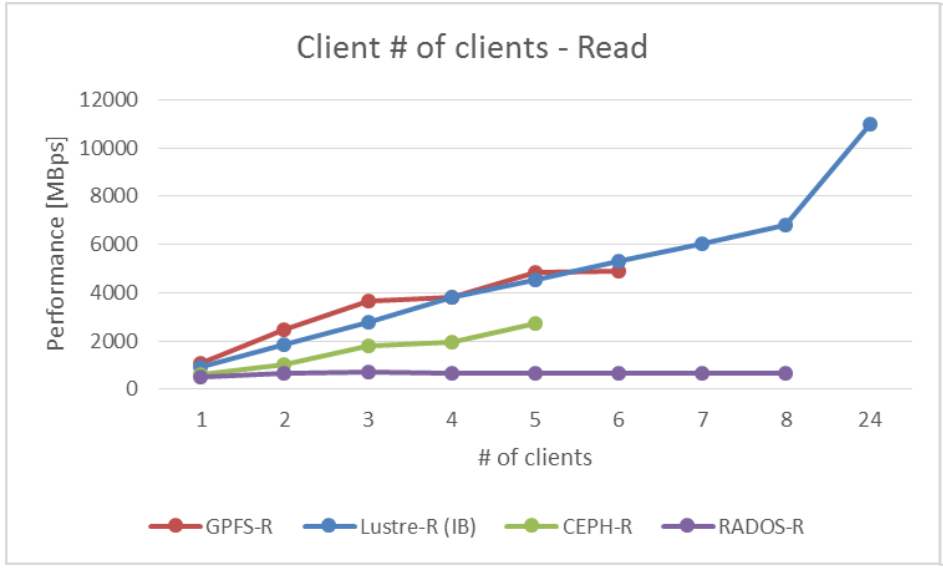
40 GbE: 40 Gigabit Ethernet
OPA: Intel® Omni-Path
IB: InfiniBand
O2: ALICE Online-Offline framework

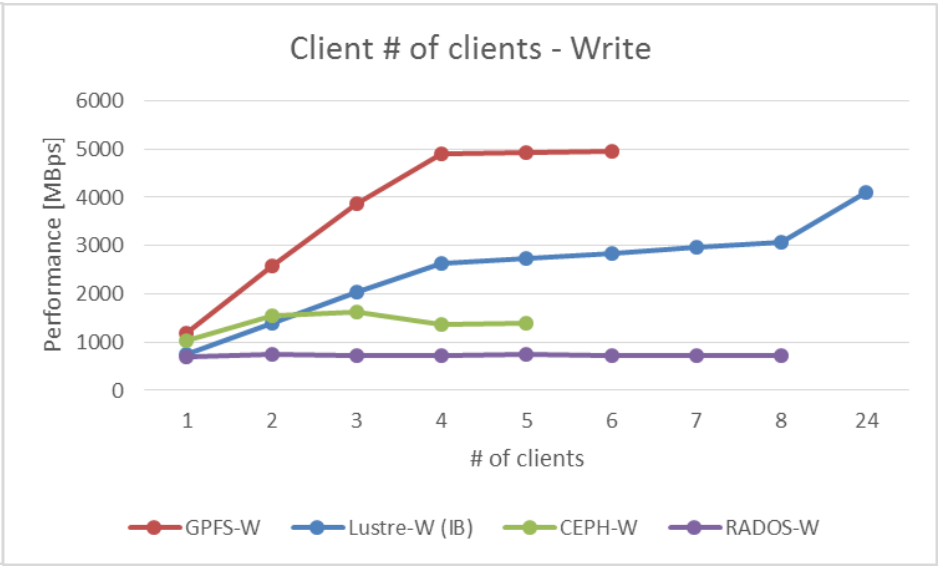
Storage

Client File Systems performance tests



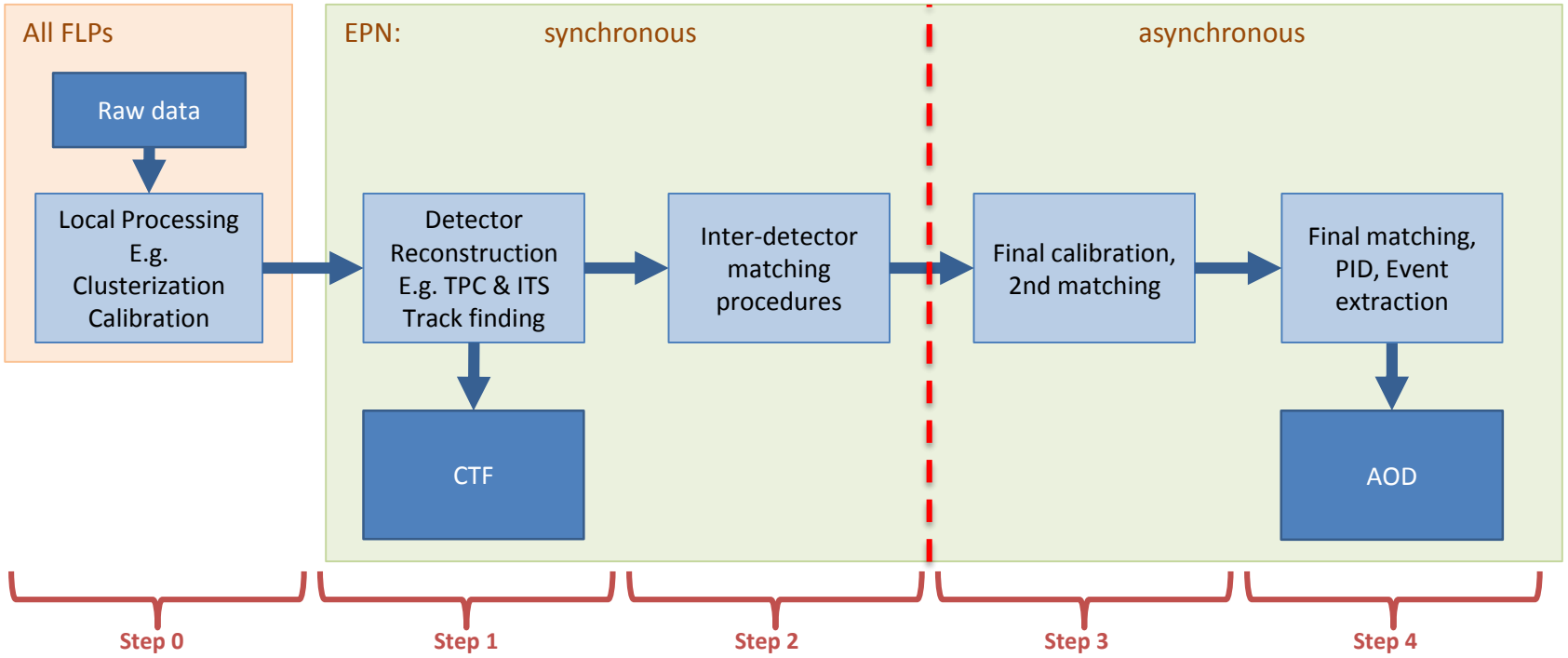
ADM: Administration server
MD: Metadata server
DS: Data server
CLI: Client
GE: Gigabit Ethernet
IB: InfiniBand
GPFS: General Parallel File System
Lustre: open-source parallel file system
Ceph: distributed object store
RADOS: Reliable Autonomic Distributed Object Store



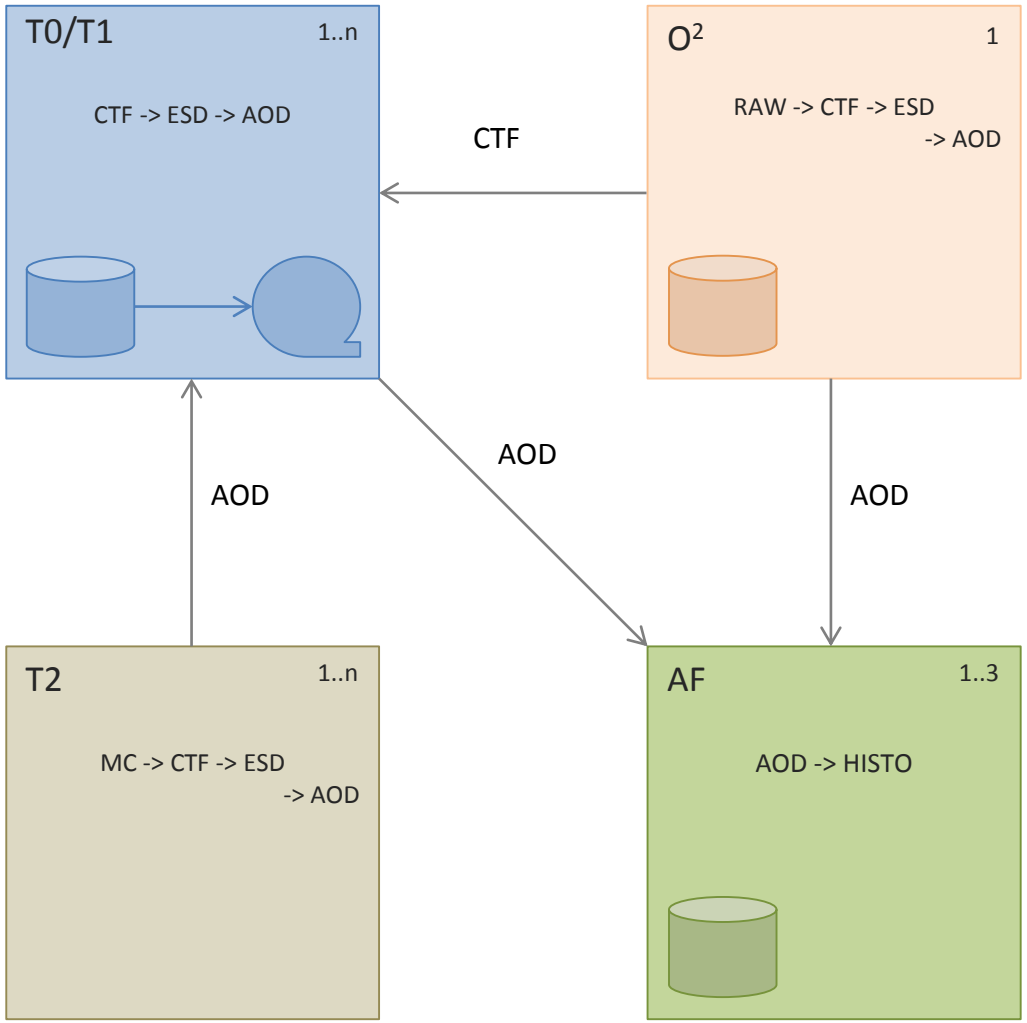


Physics software design

Online processing workflow



Computing Model



Glossary

- RAW: raw data
- CTF: Compressed Time Frame
- ESD: Event Summary Data
- AOD: Analysis Data Object
- O2: Online-Offline facility
- T0, T1, T2: Grid Tier 0, 1, 2
- AF: Analysis Facility
- MC: Monte-Carlo
- HISTO: Subset of AOD specific for a given analysis