ALICE trigger and computing upgrade: design, technology and performance

B. von Haller for the ALICE Collaboration

CERN

6.10.2016

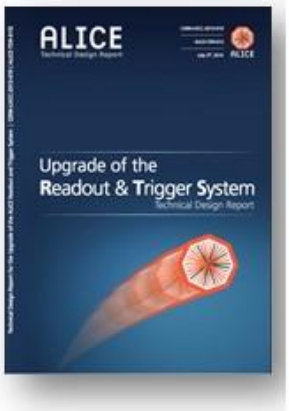
Outline

- ALICE trigger & computing upgrade in a nutshell
- Triggers, Heart Beats and Timeframes
- Hardware and computing model
  - Nodes and network
  - Hardware acceleration
  - O2 Facility and Tiers
- Software
  - Paradigms
  - Development processes and tools
- Project organisation
- Schedule

ALICE Trigger/DAQ Upgrade in a nutshell

Requirements

1. After LS2, LHC min bias PbPb at 50 kHz
   - ~100 x more data than during Run 1
   - ➔ Too much data to be stored
2. Physics topics addressed by ALICE upgrade
   - Rare processes
   - Very small signal over background ratio
   - Needs large statistics of reconstructed events
   - $13 nb^{-1}$ for PbPb
   - ➔ Triggering techniques very inefficient if not impossible in most cases
3. TPC inherent rate (drift time ~100 µs) < 50 kHz
   - ➔ Support for continuous read-out (TPC), as well as triggered read-out

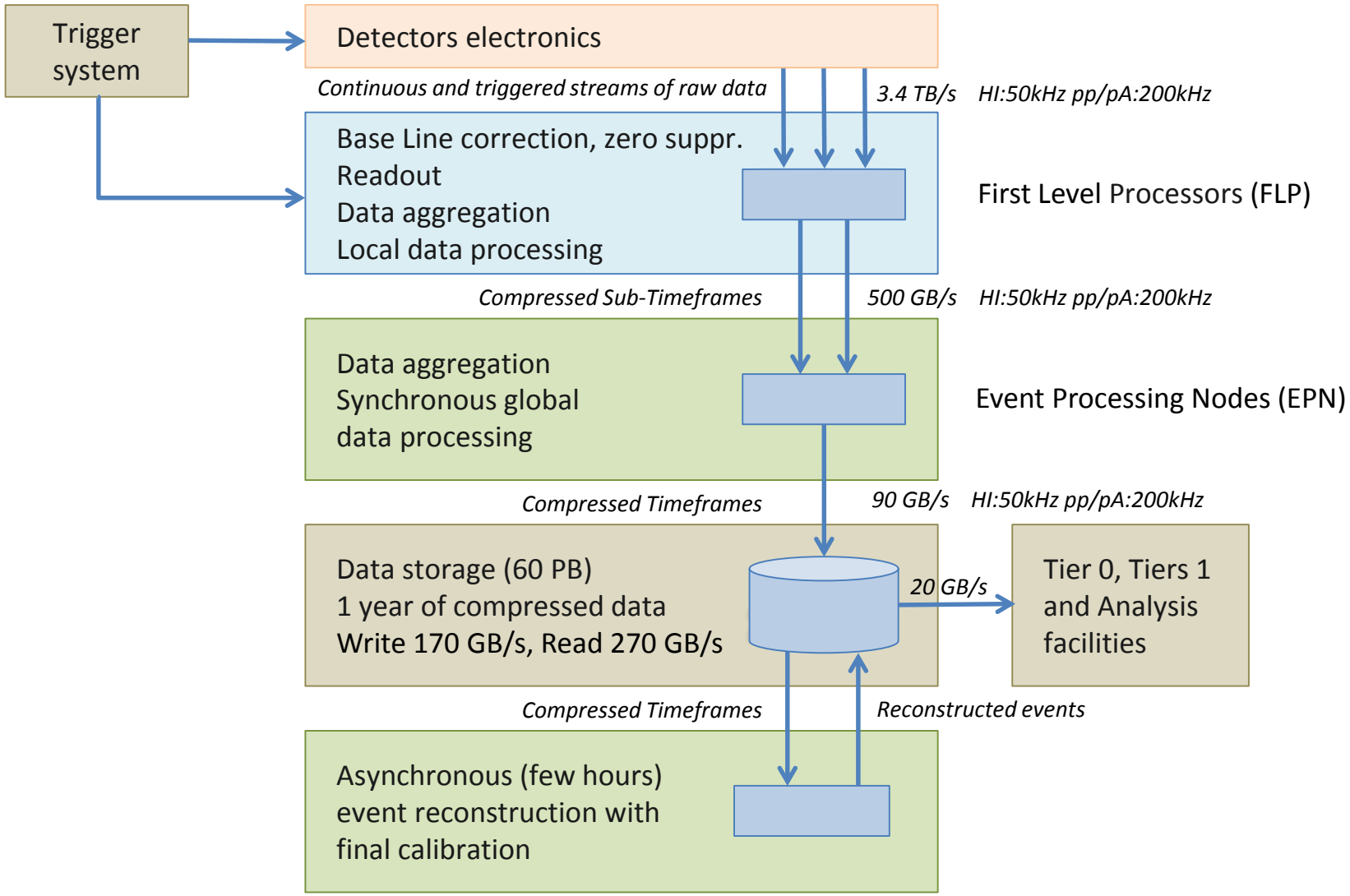


New computing system

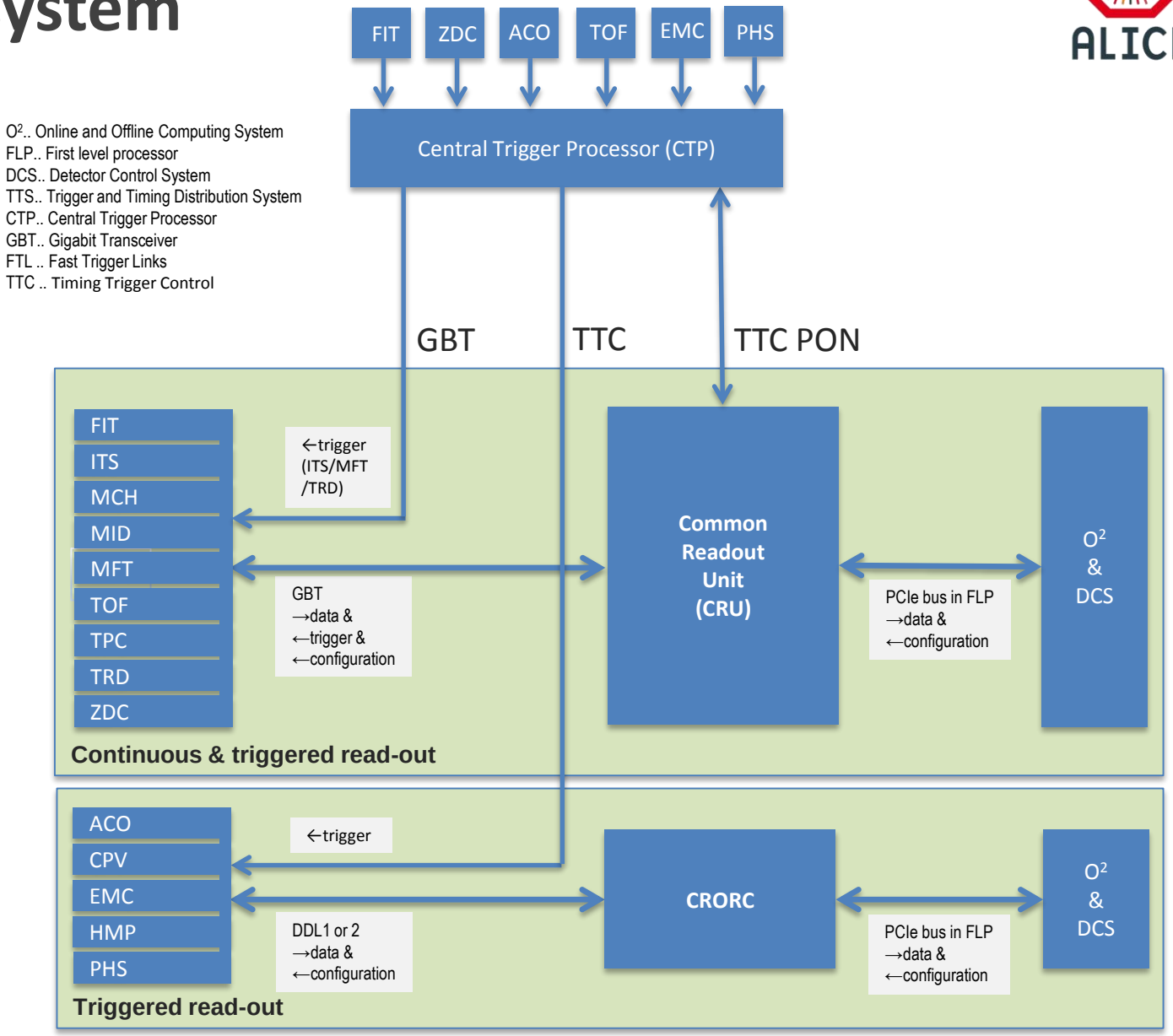
- Read-out the data of all interactions
- Compress these data intelligently by reconstructing them online
- One common online-offline computing system: O^2

➔ Paradigm shift compared to approach for Run 1 and 2

Functional flow



Trigger system



Trigger system

Run 1 and 2 :

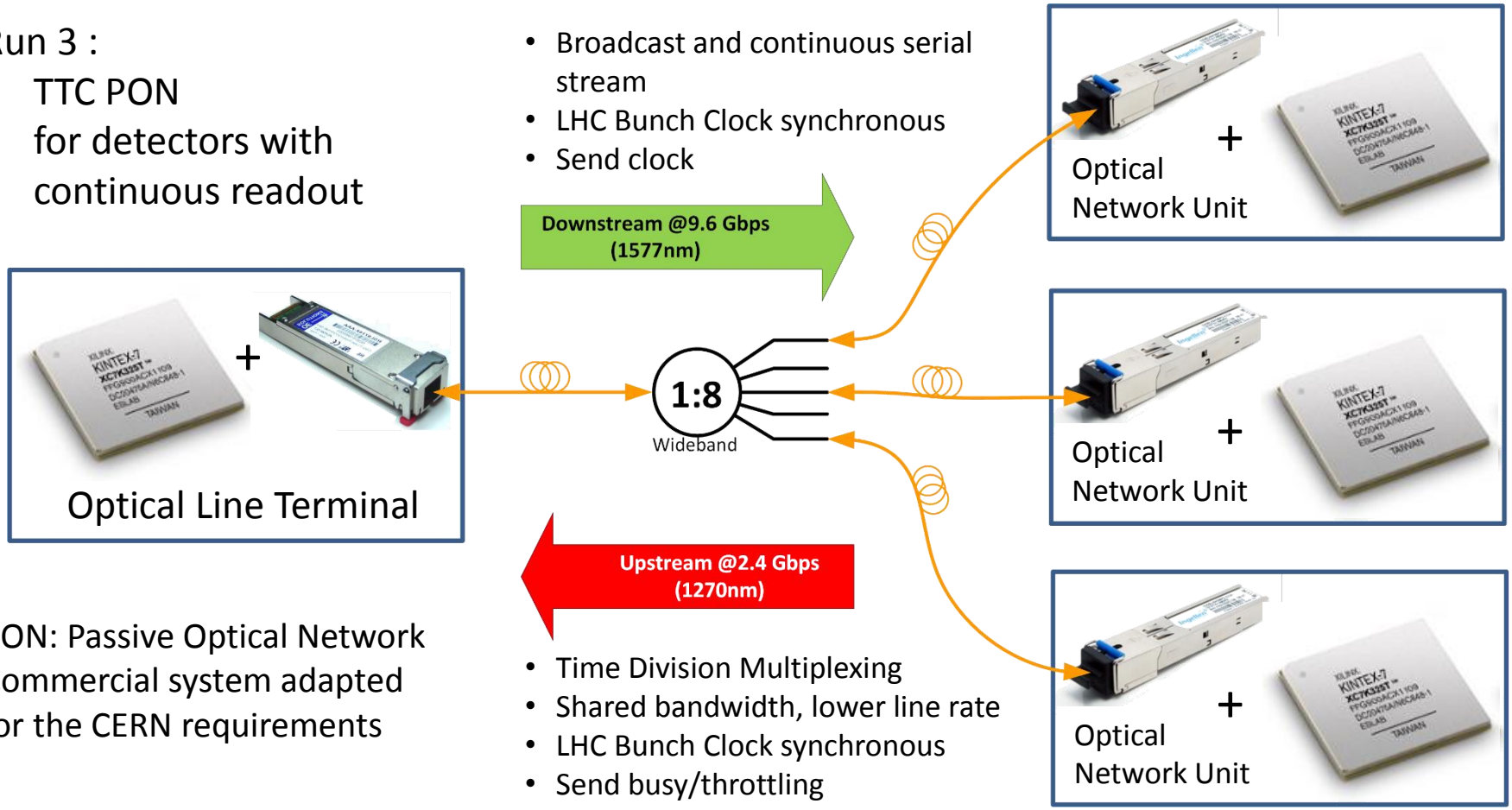
- TTC (Timing Trigger Control) used for trigger distribution (unidirectional)
- Busy/throttling transported electrically on a separate link

S. Baron

Run 3 :

- TTC PON
  for detectors with
  continuous readout

- Broadcast and continuous serial stream
- LHC Bunch Clock synchronous
- Send clock



PON: Passive Optical Network
Commercial system adapted
for the CERN requirements

- Time Division Multiplexing
- Shared bandwidth, lower line rate
- LHC Bunch Clock synchronous
- Send busy/throttling

Triggers, Heart Beats, Timeframe

Heart Beat (HB)

issued in continuous & triggered modes to all detectors

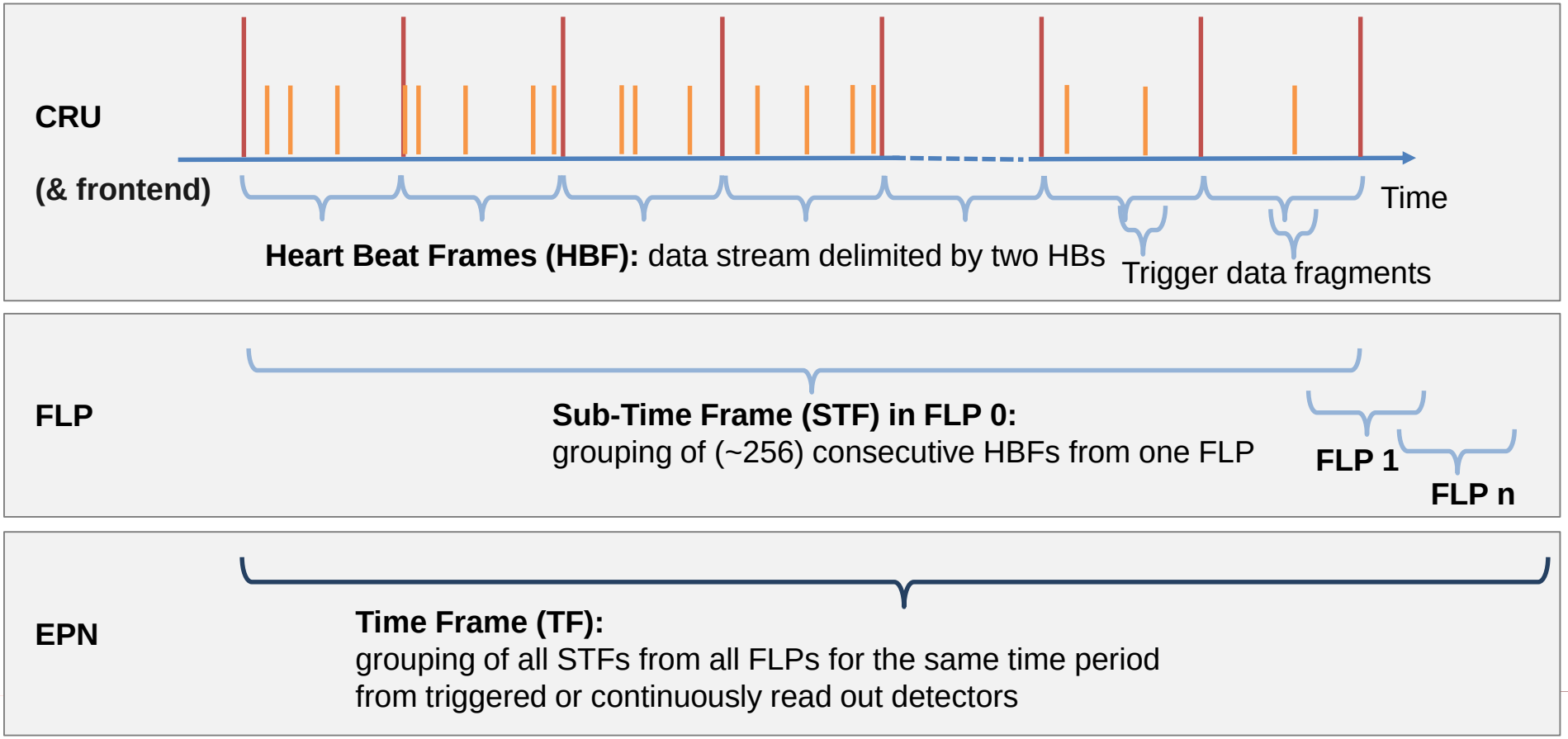
Physics trigger

can be sent to upgraded detectors will be sent to non-upgraded detectors

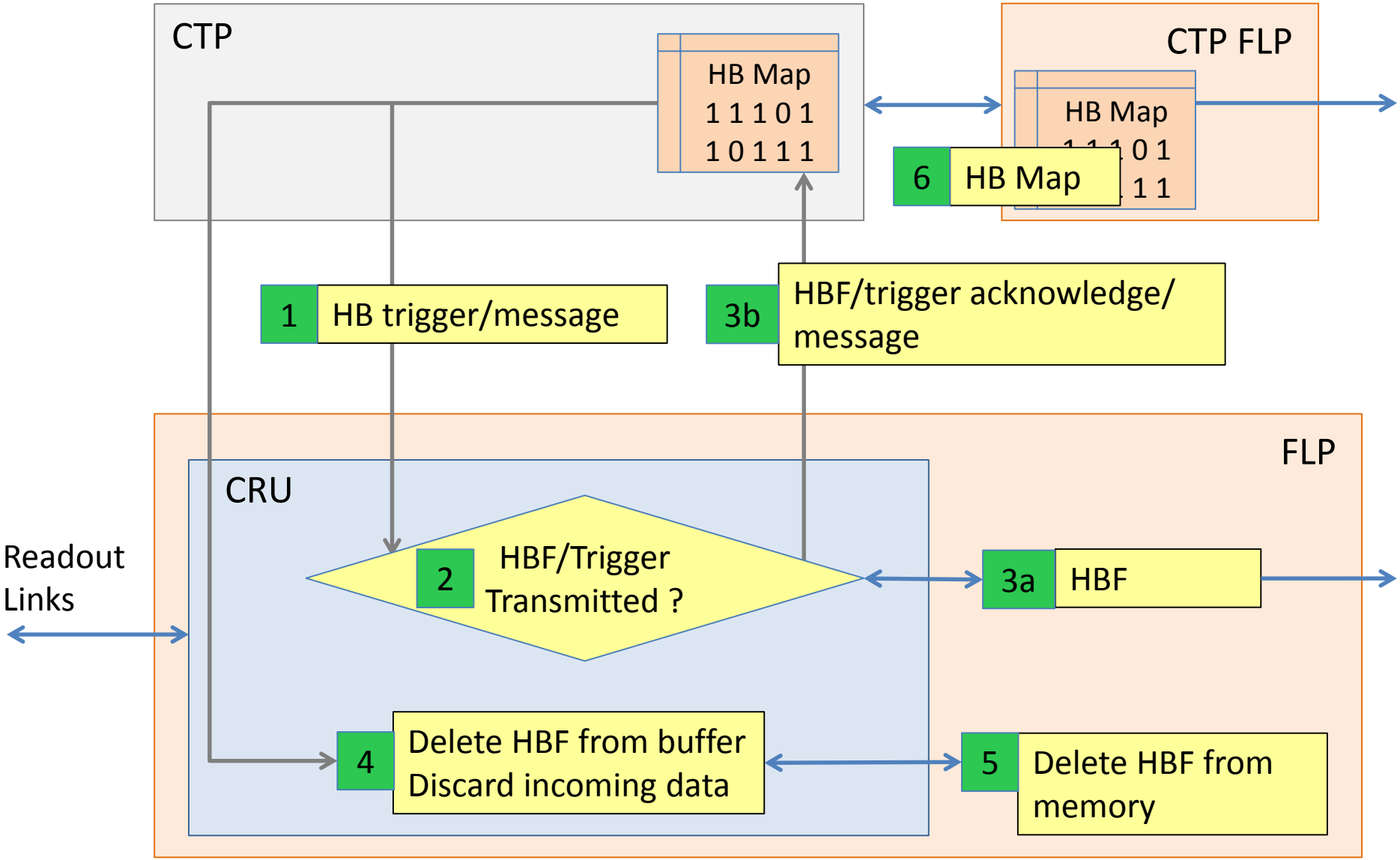
HBF and TF rates programmable
Typical values:

- HB: 1 per orbit, 89.4 µs: ~10 kHz
- TF: 1 every ~20 ms: ~50 Hz
- → 1 TF = ~256 HBF

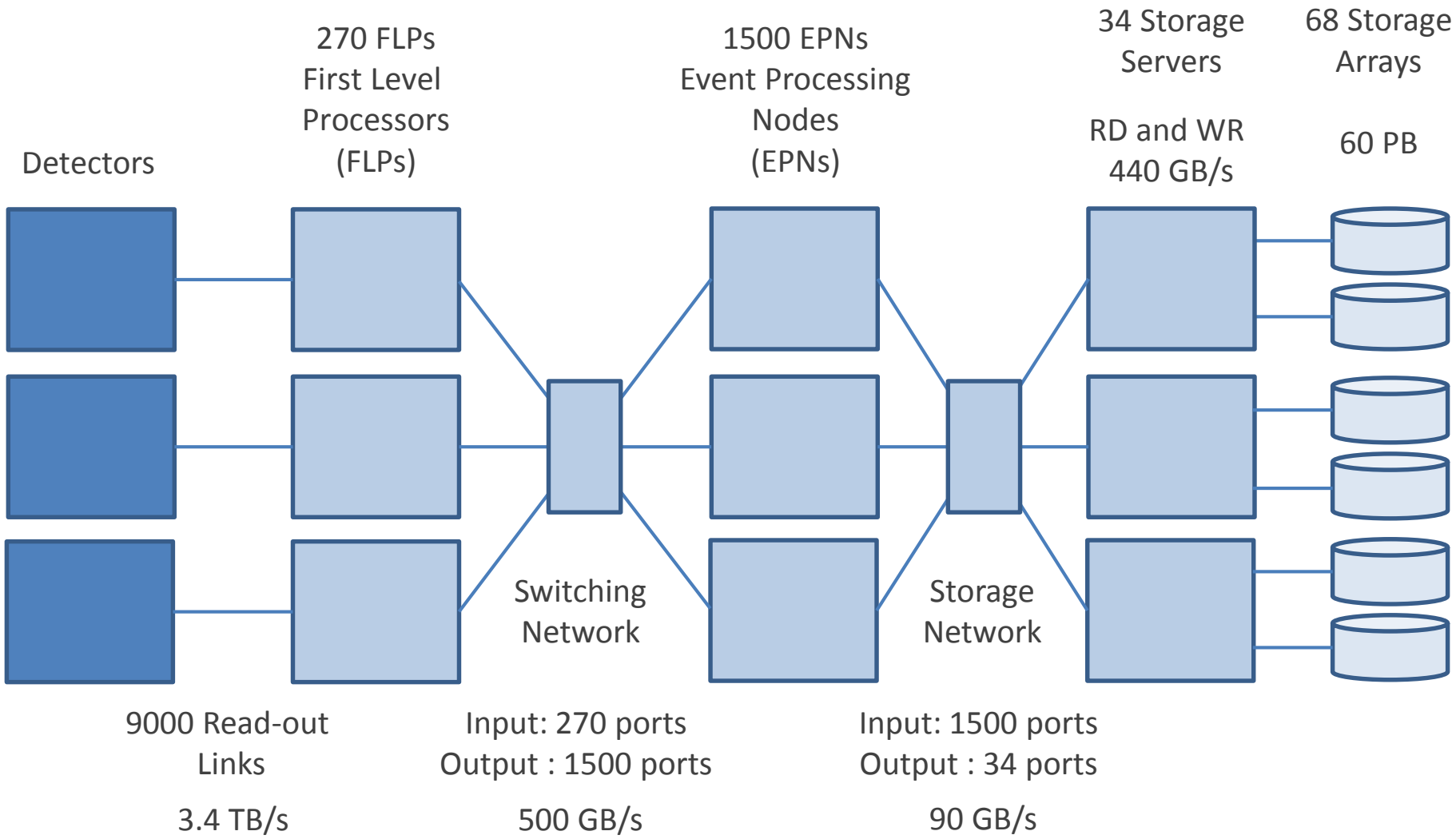
Read-out: Continuous and triggered



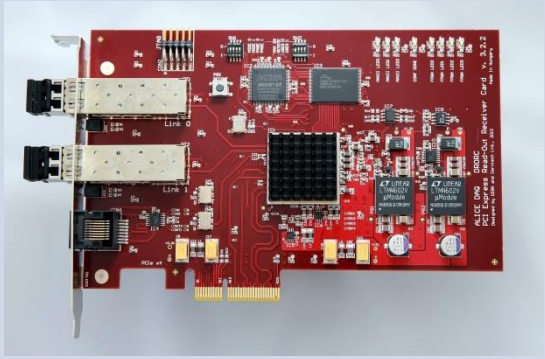
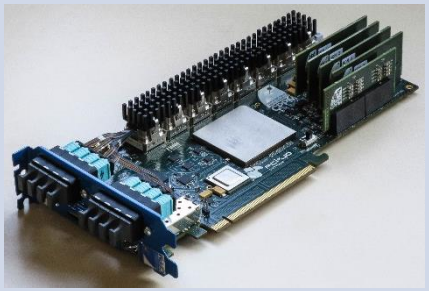
System throttling



O2 Hardware facility

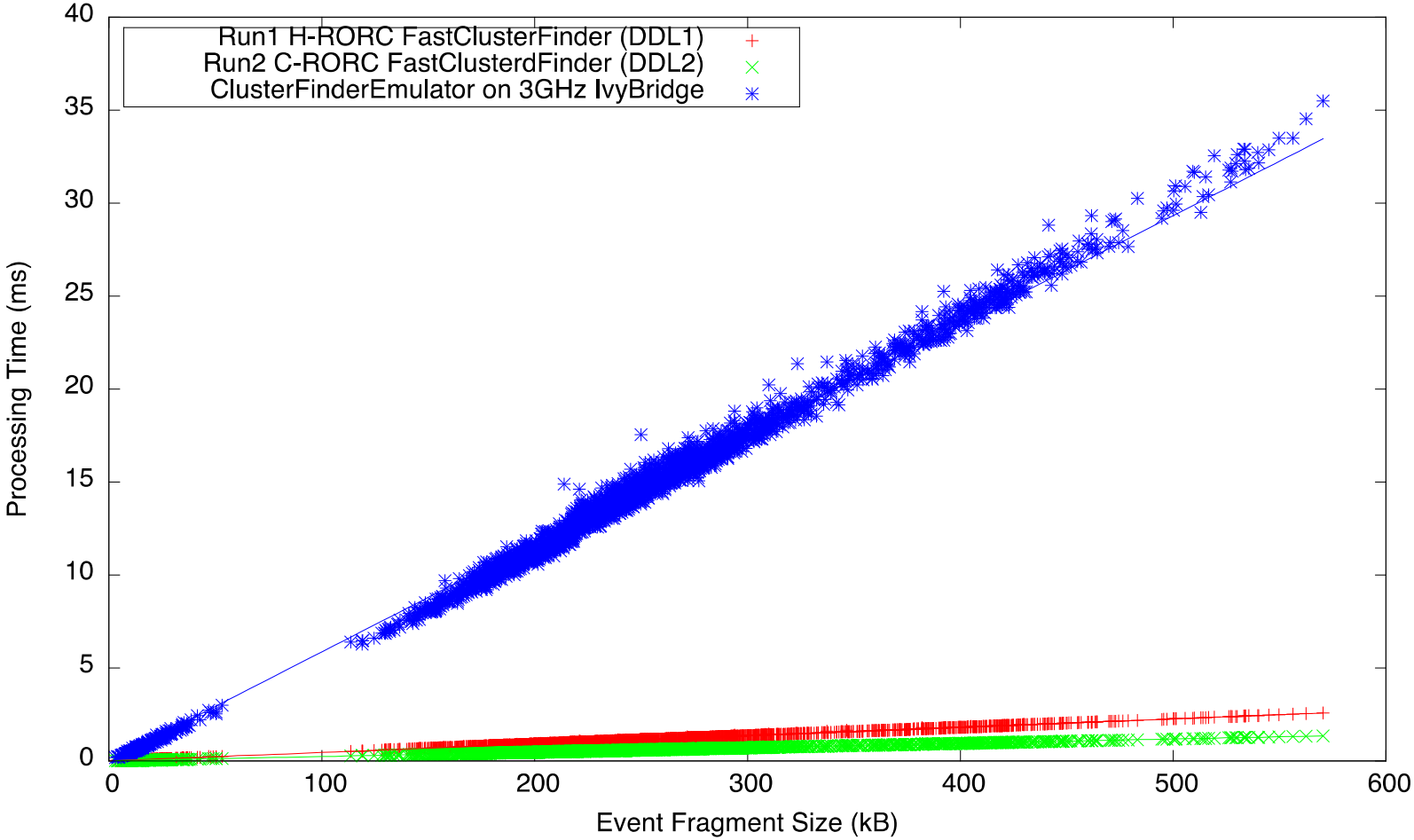


Readout & FPGA Hardware acceleration

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

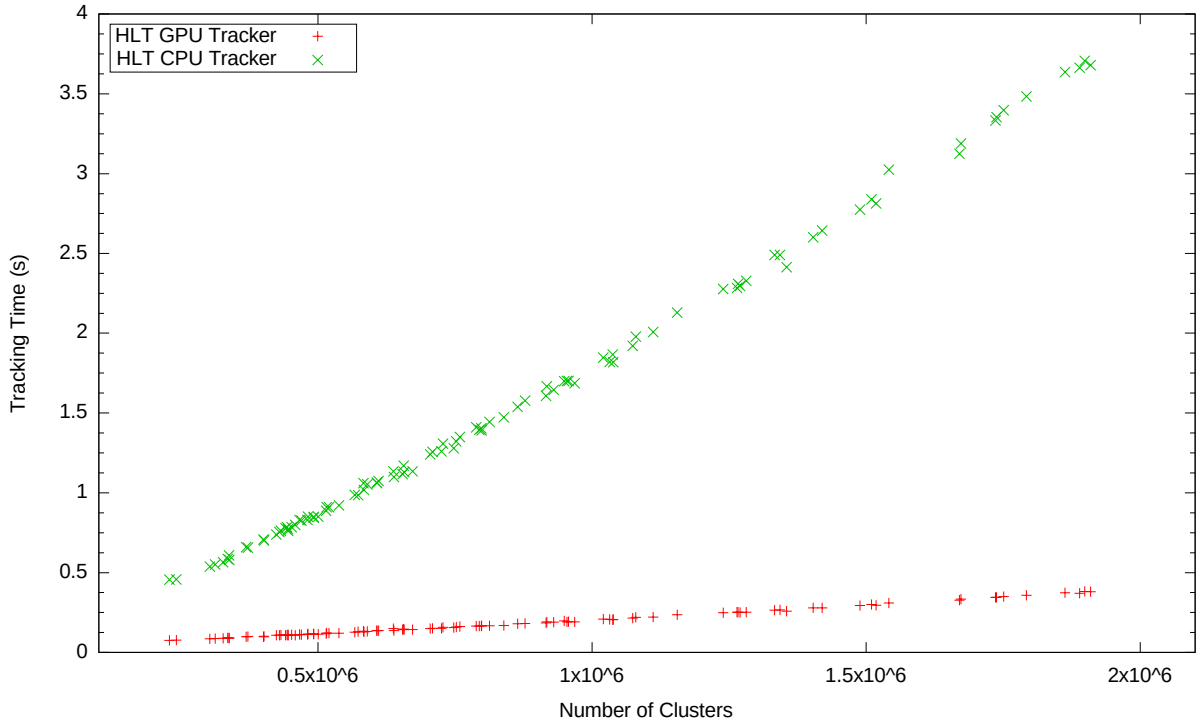
Hardware acceleration (FPGA)



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

Hardware acceleration (GPU)

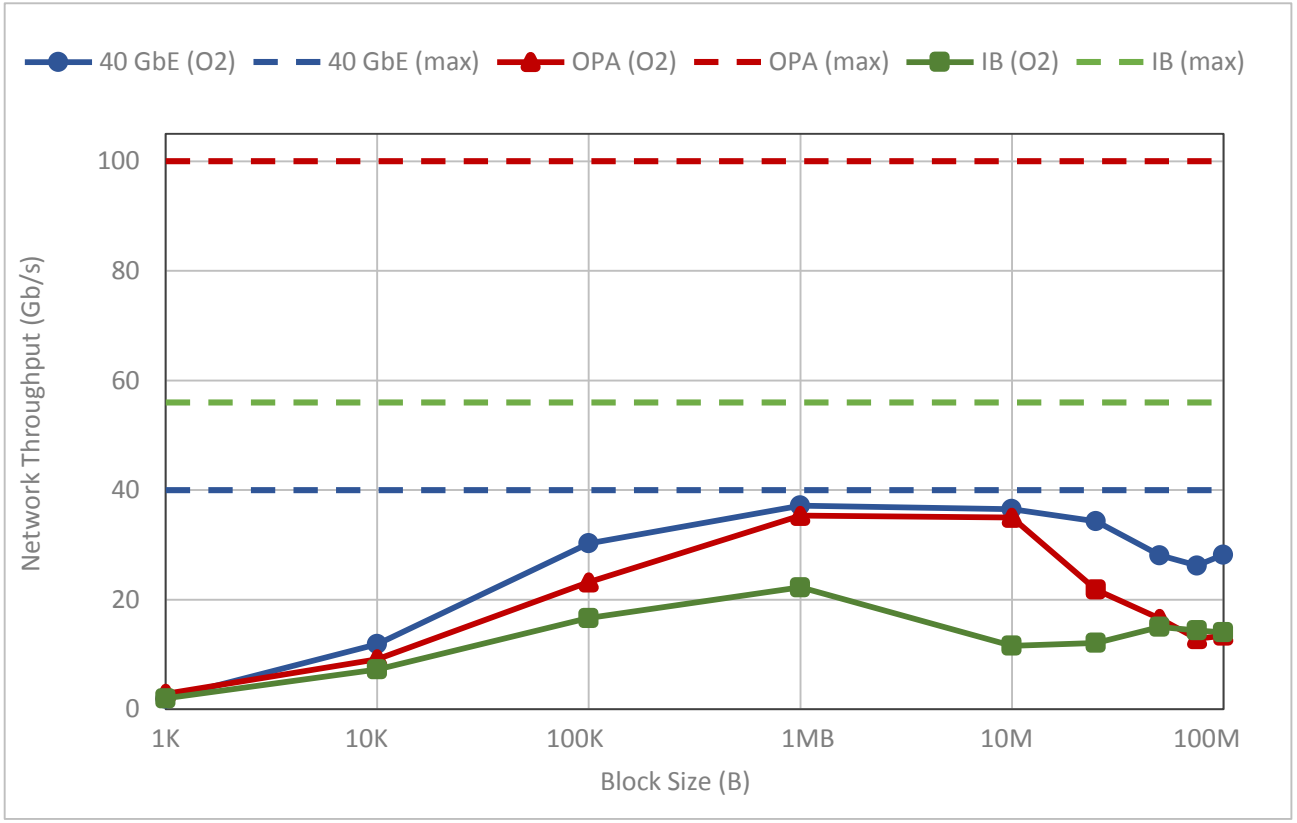
- For TPC track finding on EPNs (as today's HLT)
- Possibly more use cases depending on R&D



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

Network performance tests

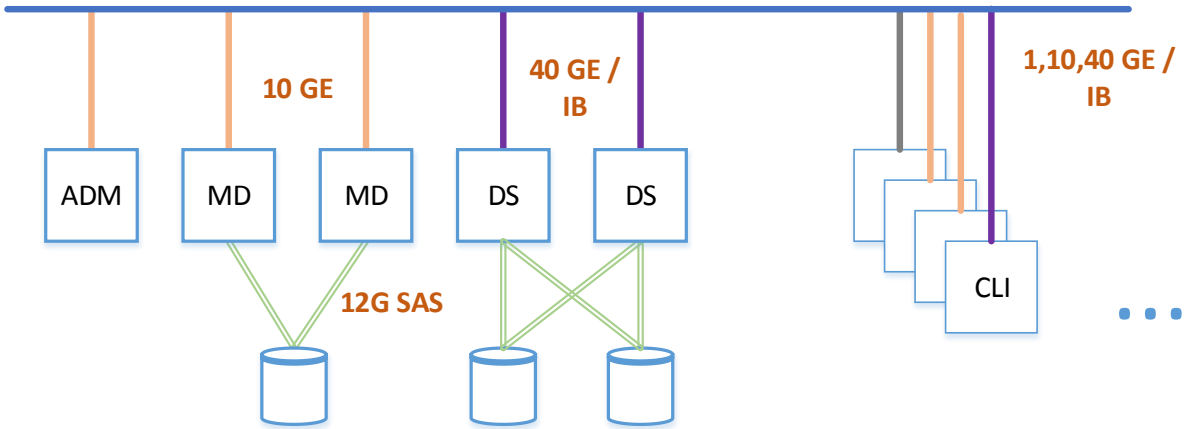
Comparison of Ethernet,
IP over InfiniBand and IP over Omni-Path



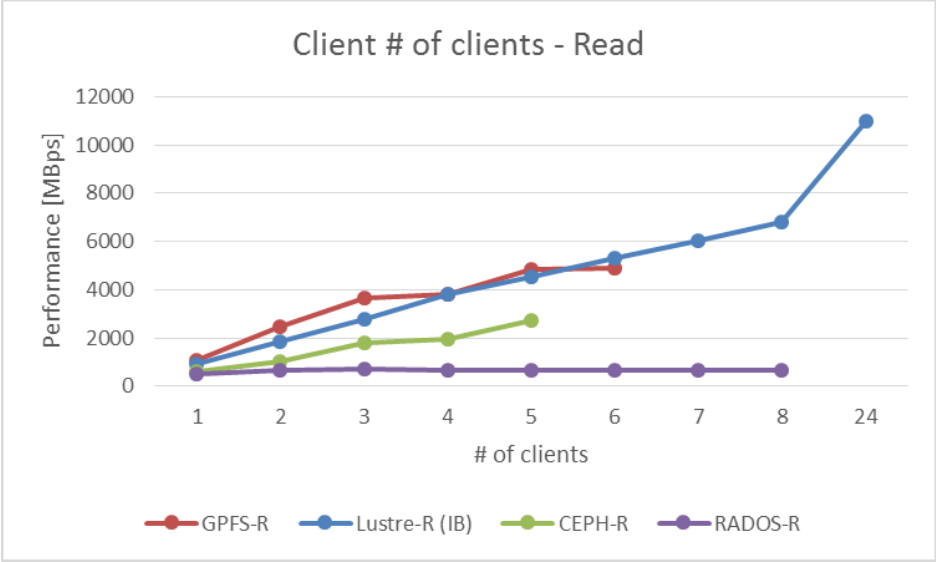
40 GbE: 40 Gigabit Ethernet
OPA: Intel® Omni-Path
IB: InfiniBand
O2: ALICE Online-Offline framework

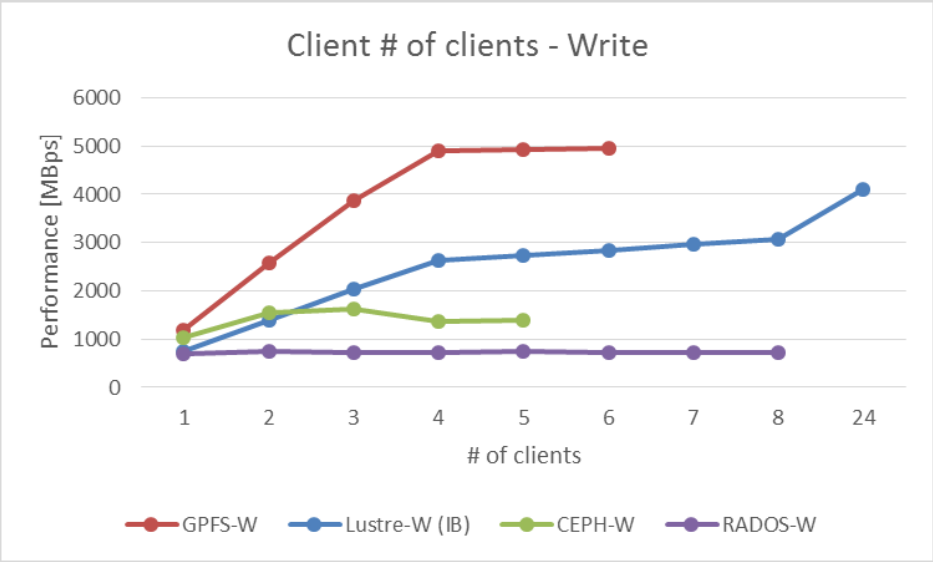
Storage

Client File Systems performance tests



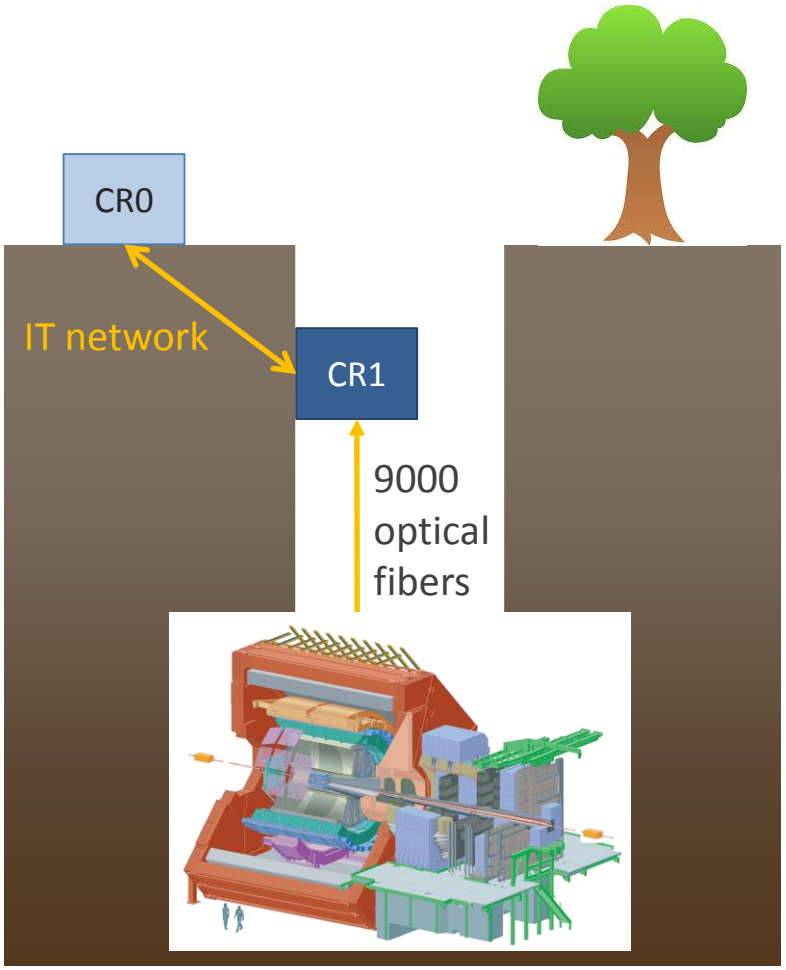
ADM: Administration server
MD: Metadata server
DS: Data server
CLI: Client
GE: Gigabit Ethernet
IB: InfiniBand
GPFS: General Parallel File System
Lustre: open-source parallel file system
Ceph: distributed object store
RADOS: Reliable Autonomic Distributed Object Store



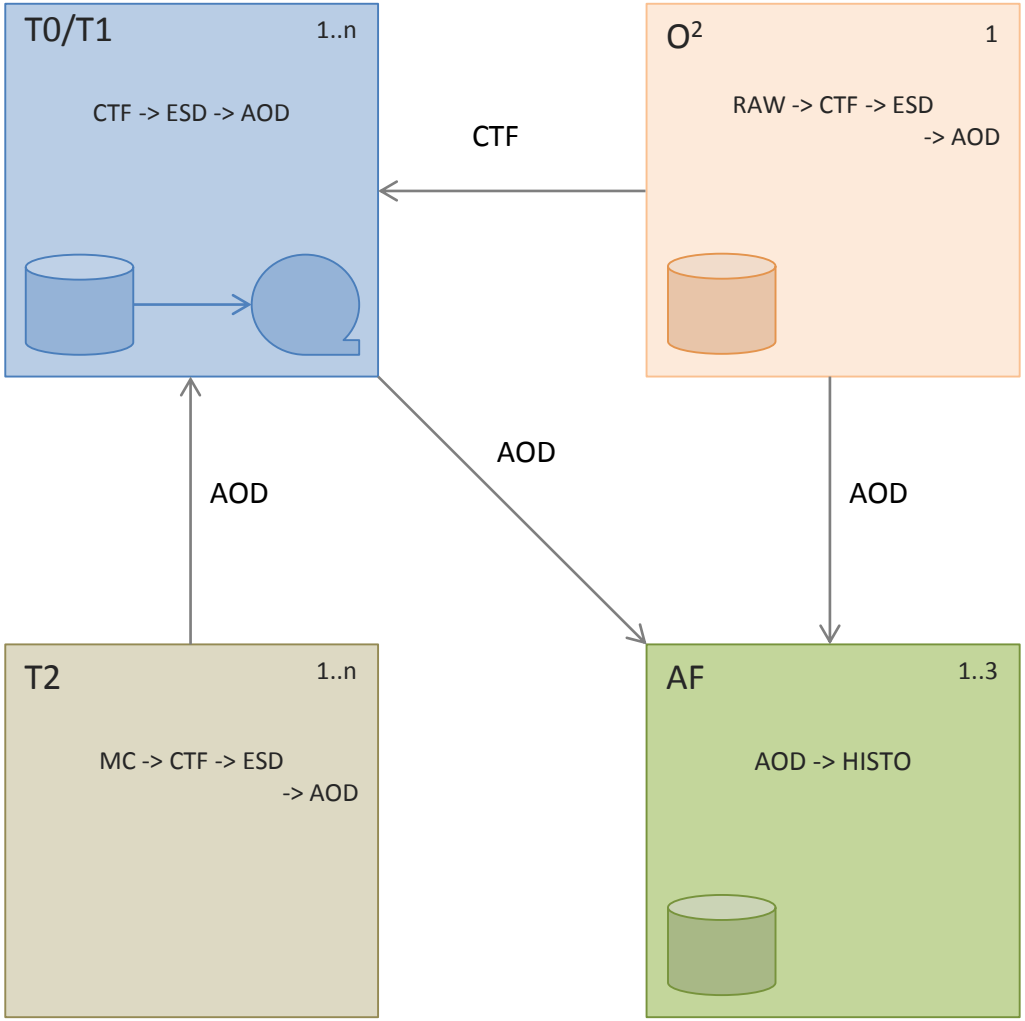


O^2 Farm

- ~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
    - Solution based on the GSI Green Cube
    - Being studied



Computing Model



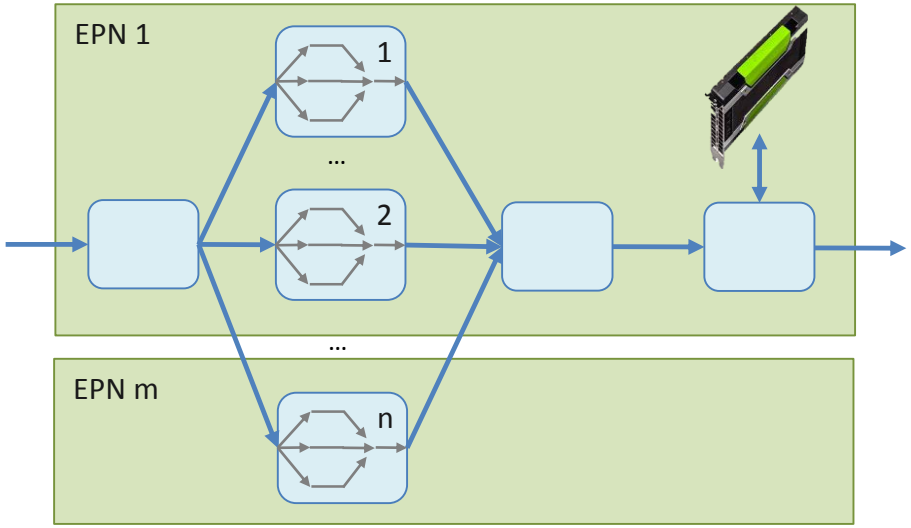
Glossary

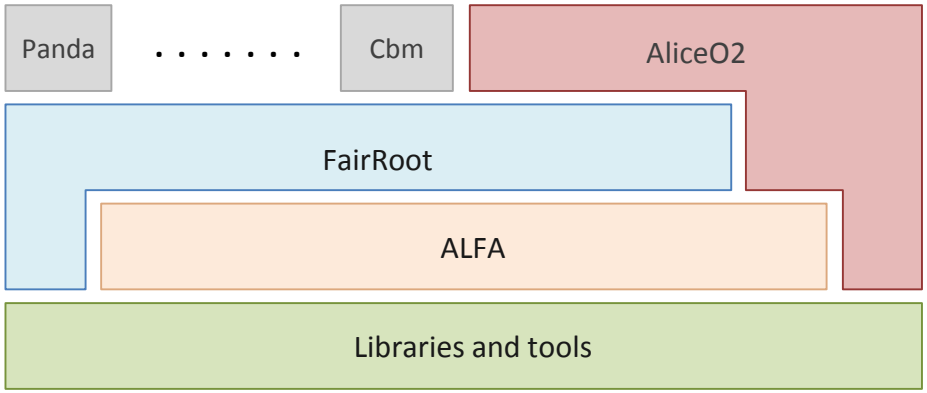
- RAW: raw data
- CTF: Compressed Time Frame
- ESD: Event Summary Data
- AOD: Analysis Data Object
- O2: Online-Offline facility
- T0, T1, T2: Grid Tier 0, 1, 2
- AF: Analysis Facility
- MC: Monte-Carlo
- HISTO: Subset of AOD specific for a given analysis

Software

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
  - Steady development
- AliceO2
  - Prototyping
  - Development started

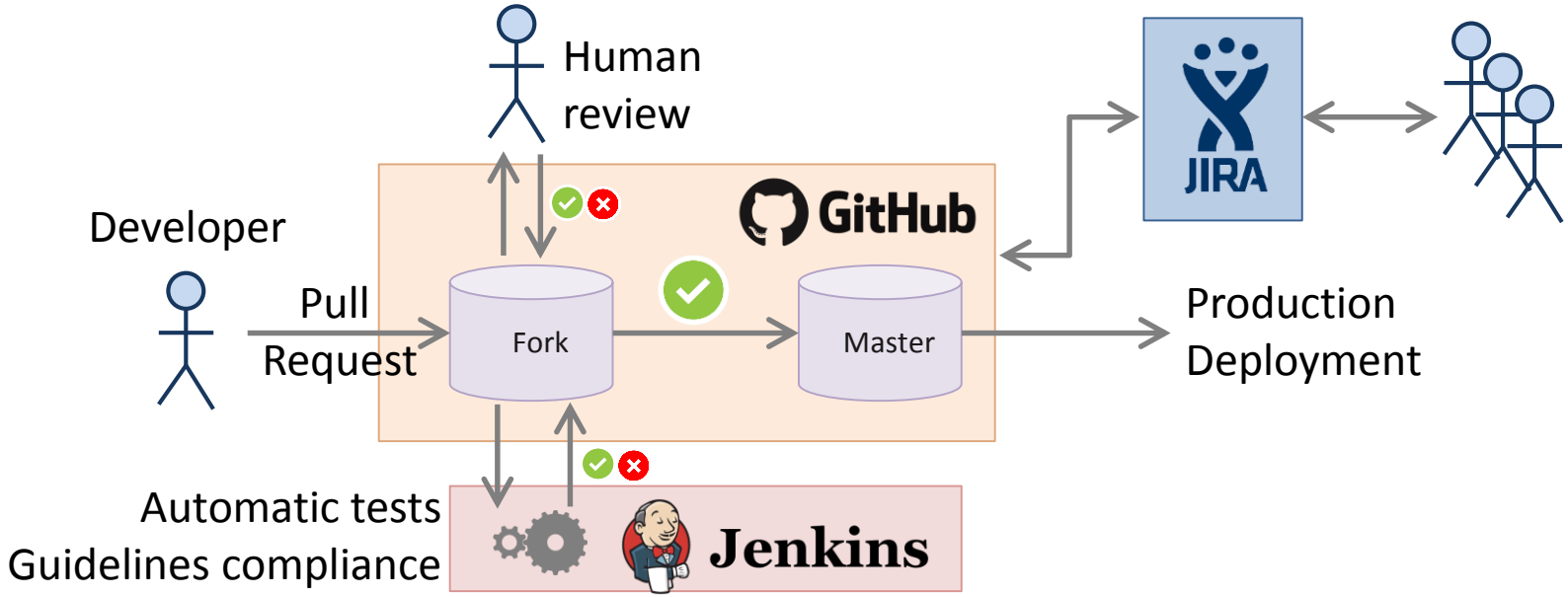




Software

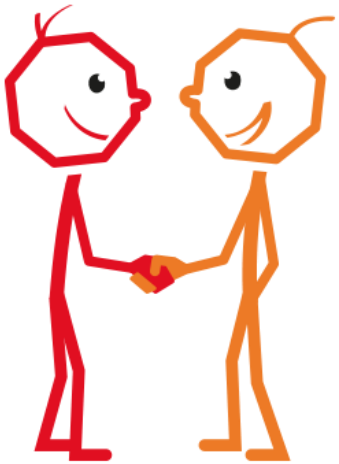
Development processes

- Especially important
  - Most software will run online
  - Large code base
  - Many people involved, from different backgrounds
    - That is also a blessing as we have a large expertise to pick from



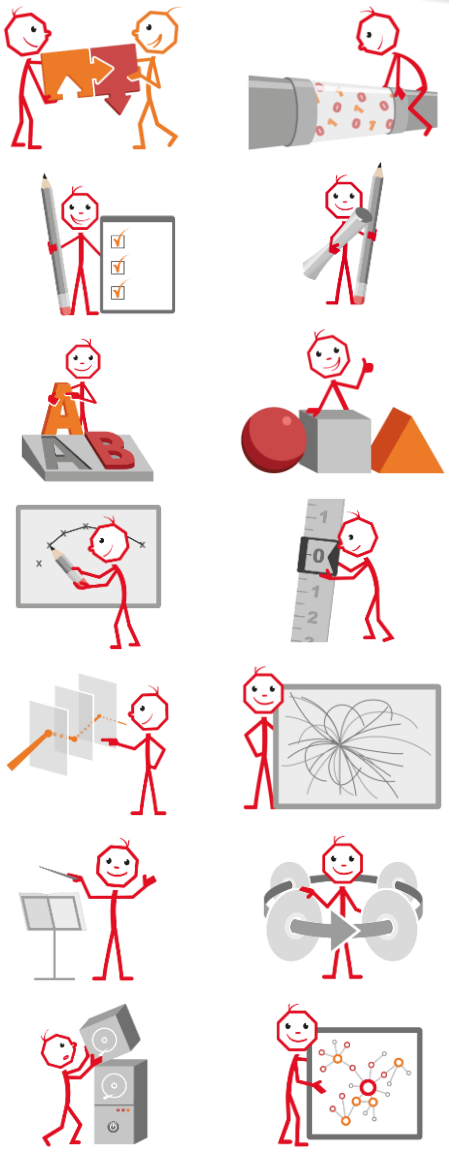
O^2 Organization

- Today
  - 3 distinct projects : DAQ, HLT, Offline
  - 2 different code base : HLT+Offline and DAQ
  - 2 different reconstructions : HLT and Offline
- ALICE O^2
  - 1 project
  - 1 code base
  - 1 reconstruction



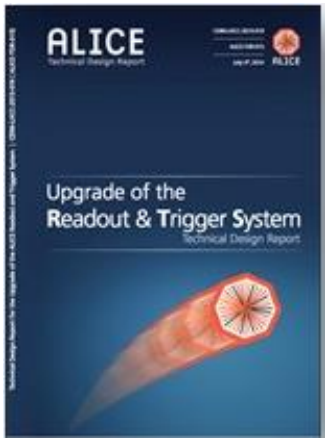
O^2 Organization

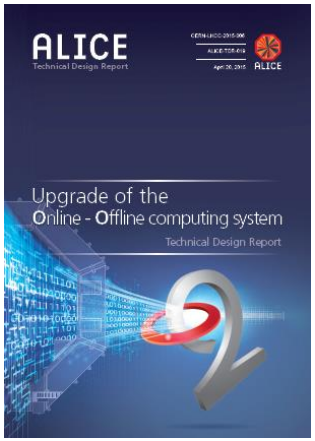
- Steering board
  - Project leaders of the DAQ, HLT and Offline
- 13 Computing working groups (CWGs)
  - Made of people from the DAQ, HLT and Offline projects
  - Plus participants from many detectors
- Plenaries
- Successful transition
  - Cross-fertilization of the different cultures
  - There is a strong and growing sense of belonging



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



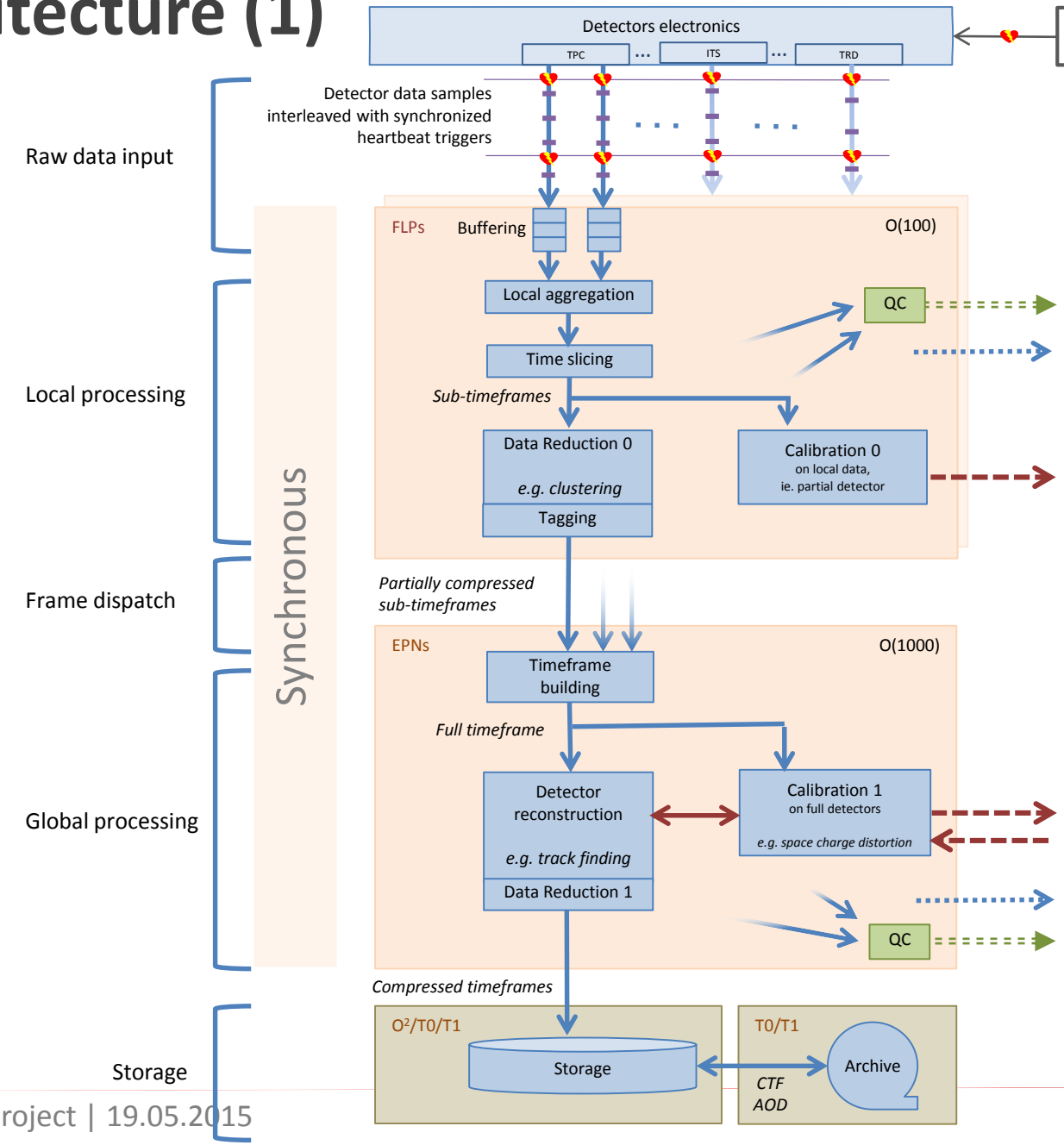


Run 1	LS1	Run 2	LS 2	Run 3

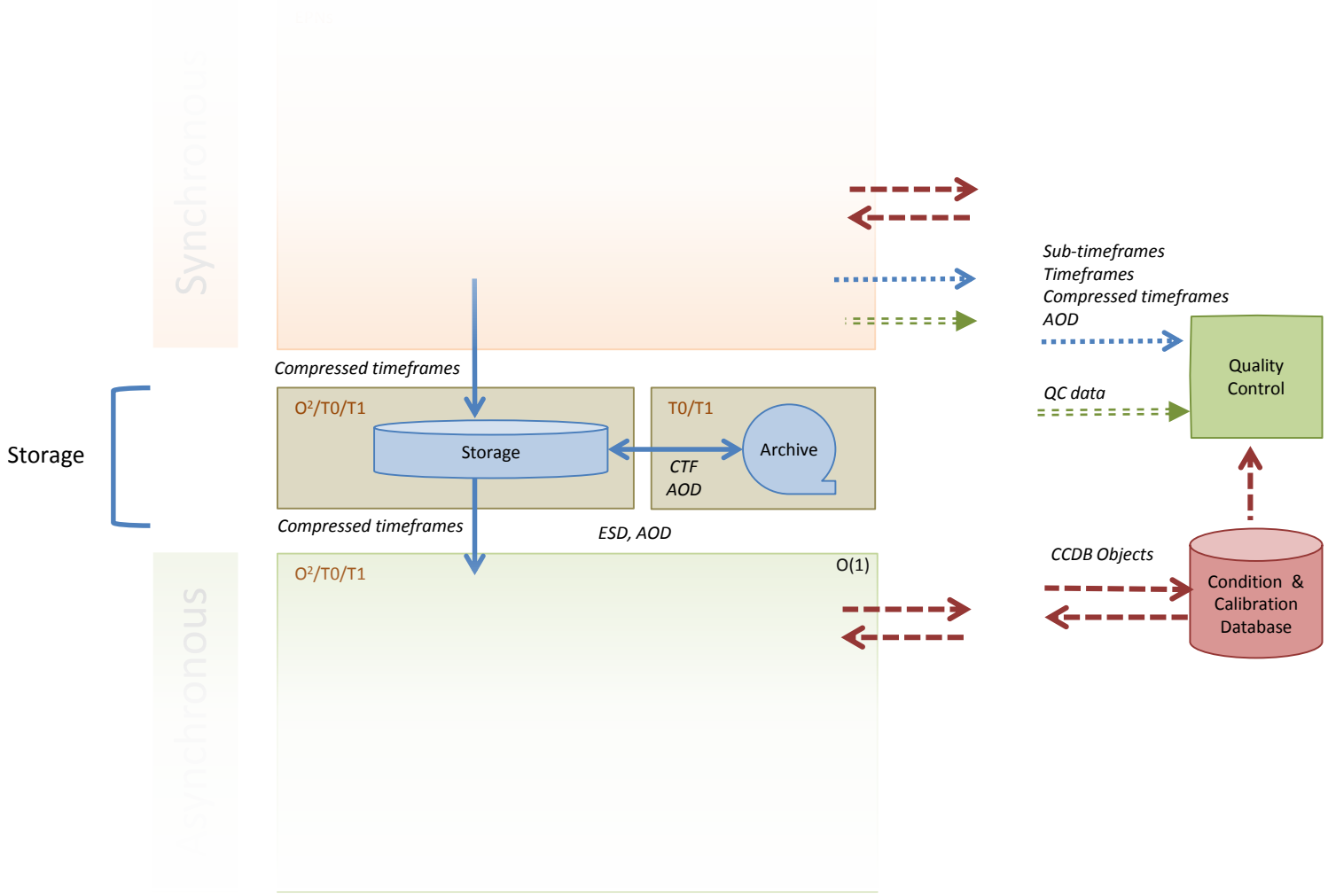
Summary

- ALICE Trigger and computing upgrade is an exciting project with very ambitious requirements
  - $13nb^{-1}$ for PbPb
  - >3TB/s continuous and triggered detector input
- New trigger and throttling strategy to handle the continuous readout
- Major paradigm change with combined offline and online system
- Hardware system
  - HW acceleration (FPGAs, GPUs)
  - O^2 farm with ~100 k CPU cores and ~5000 GPUs
- Software development
  - Collaboration with FAIR team in Germany and JPARC experiments in Japan allows a steady progress for the framework
- People from different groups and backgrounds work together
  - We all improved by doing so !
- On track despite a challenging schedule

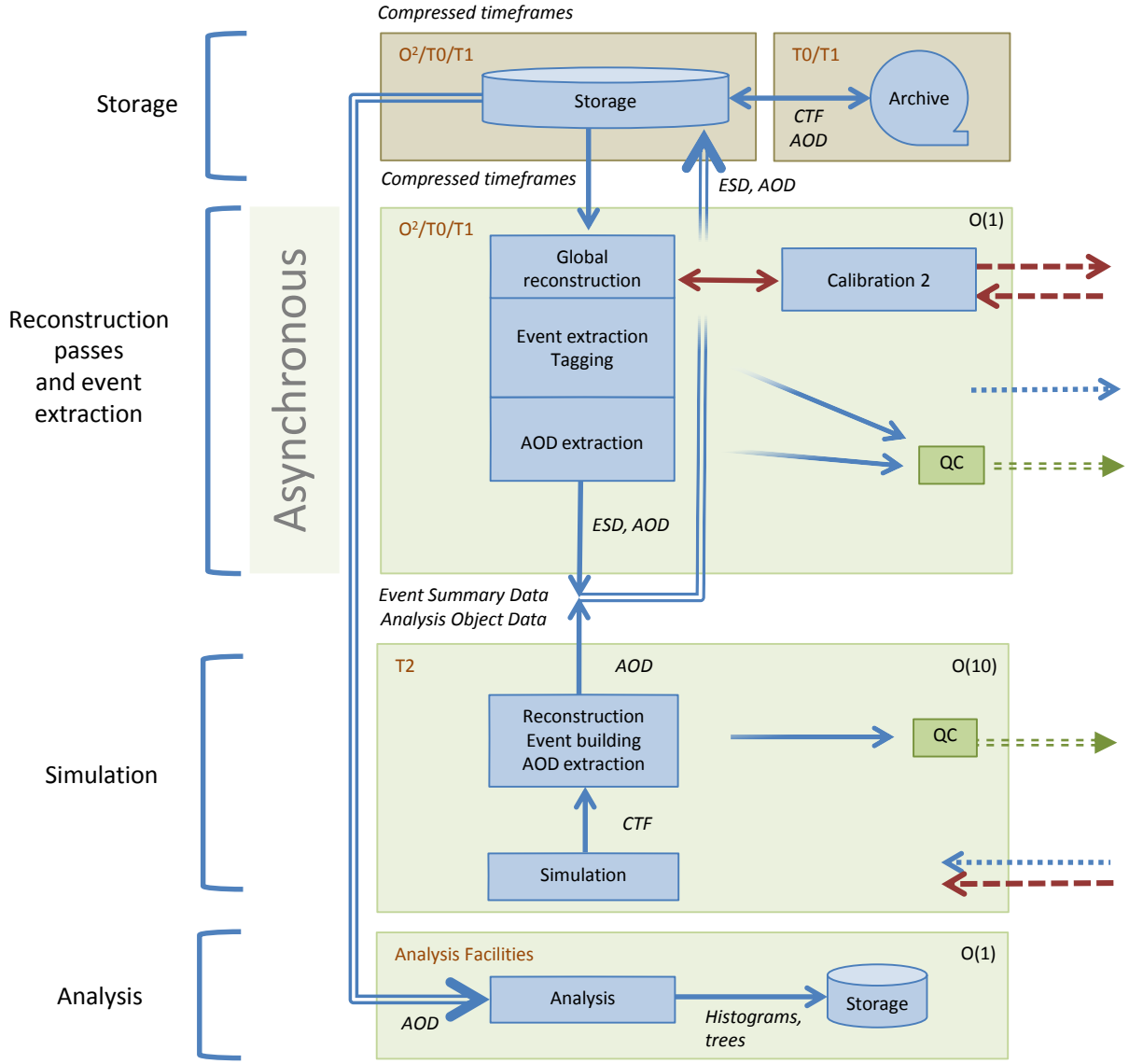
O² architecture (1)



O² architecture (2)

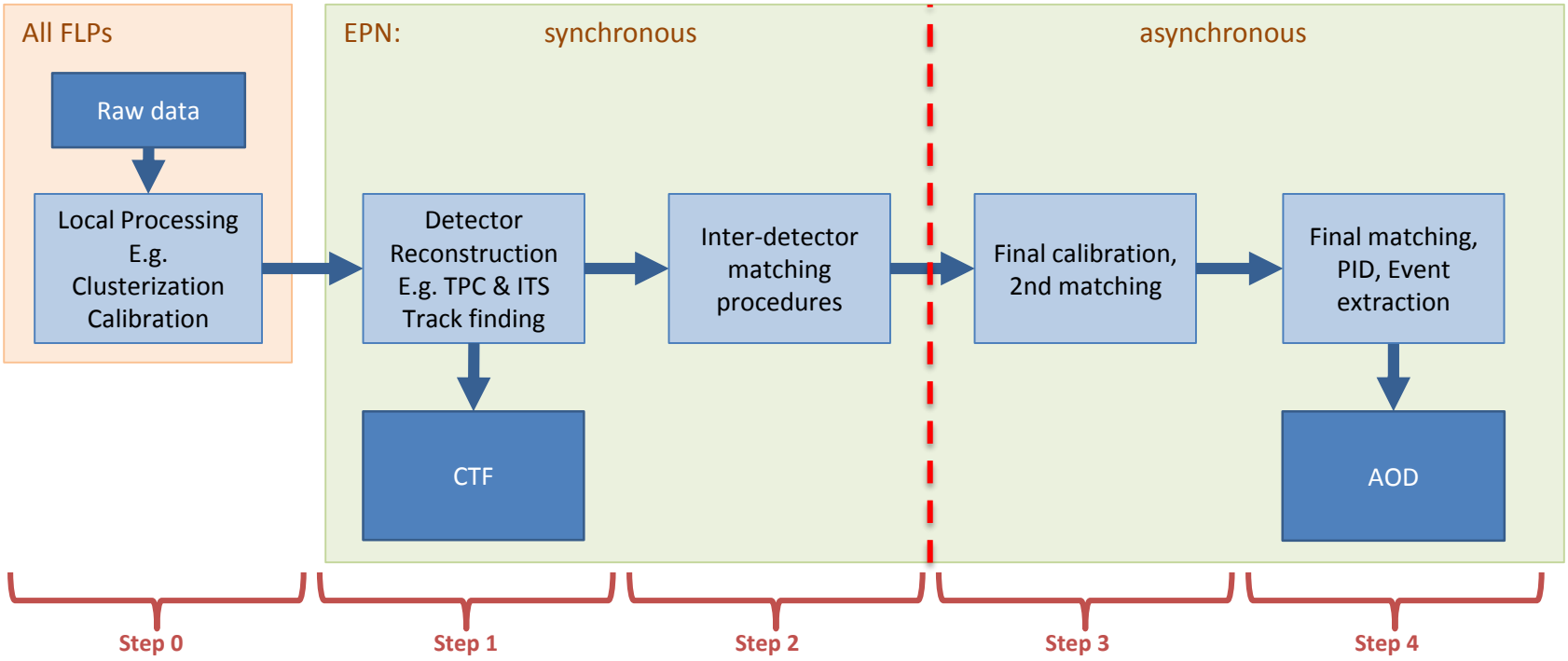


O² architecture (3)



Physics software design

Processing workflow



Computing requirements for processing

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

Goes together, merging and fitting can run on GPUs too

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

Theoretically could run on GPU

Being ported to GPU, conversion factor unknown

Physics programme and data taking scenarios

Chapter 2

Year	System	$\sqrt{s_{NN}}$ (TeV)	L_{int} (pb^{-1})	L_{int} (nb^{-1})	$N_{collisions}$
2020	pp	14	0.4		$2.7 \cdot 10^{10}$
	Pb-Pb	5.5		2.85	$2.3 \cdot 10^{10}$
2021	pp	14	0.4		$2.7 \cdot 10^{10}$
	Pb-Pb	5.5		2.85	$2.3 \cdot 10^{10}$
2022	pp	14	0.4		$2.7 \cdot 10^{10}$
	pp	5.5	6		$4 \cdot 10^{11}$
2025	pp	14	0.4		$2.7 \cdot 10^{10}$
	Pb-Pb	5.5		2.85	$2.3 \cdot 10^{10}$
2026	pp	14	0.4		$2.7 \cdot 10^{10}$
	Pb-Pb	5.5		1.4	$1.1 \cdot 10^{10}$
	p-Pb	8.8		50	10^{11}
2027	pp	14	0.4		$2.7 \cdot 10^{10}$
	Pb-Pb	5.5		2.85	$2.3 \cdot 10^{10}$

Data types characteristics

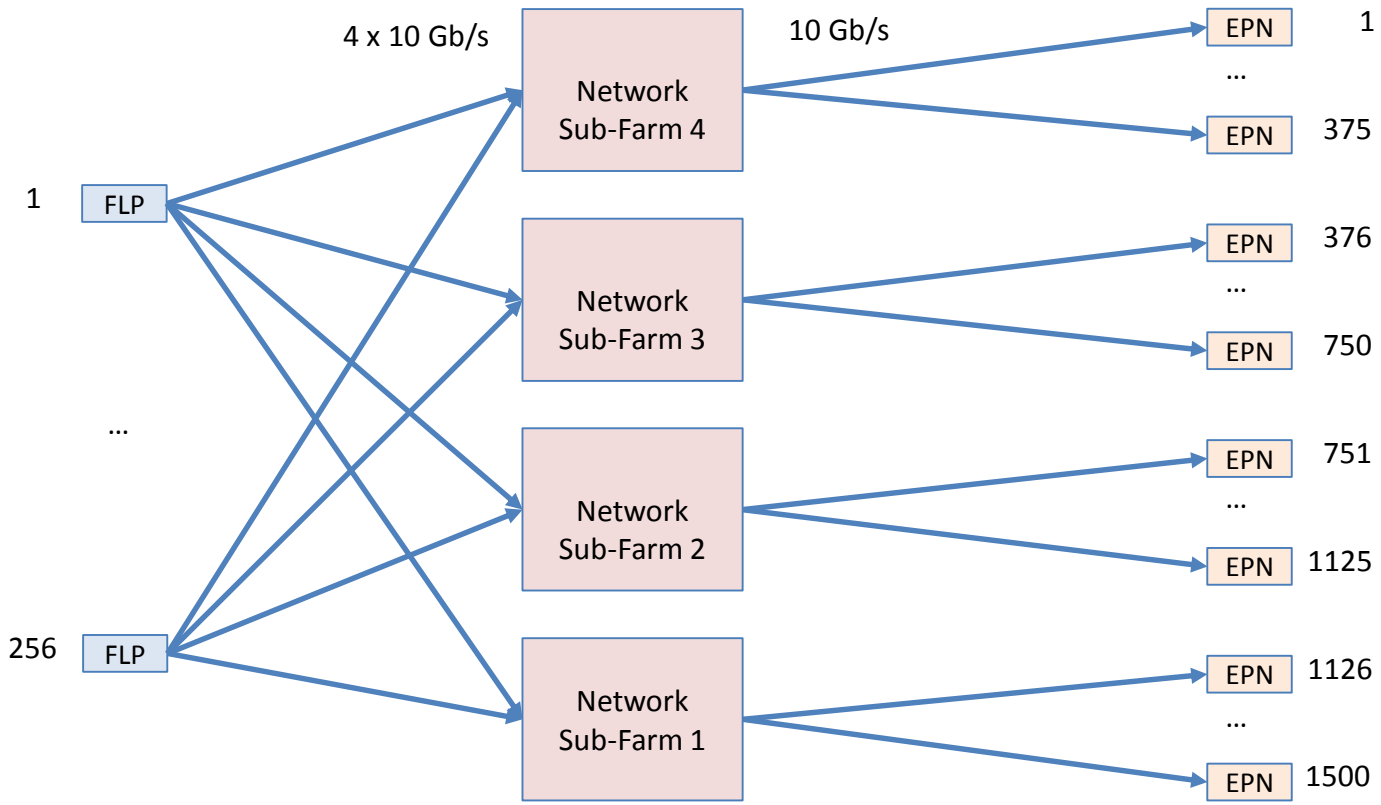
- TF size - Duration of the time window (t_{TF})
  - Data lost at the edges: $0.1/t_{TF}$(ms)
  - For calibration and reconstruction: 20ms - 100ms
  - Shorter is better for buffering and distribution

  → 20ms (1000 interactions in Pb-Pb at 50kHz)

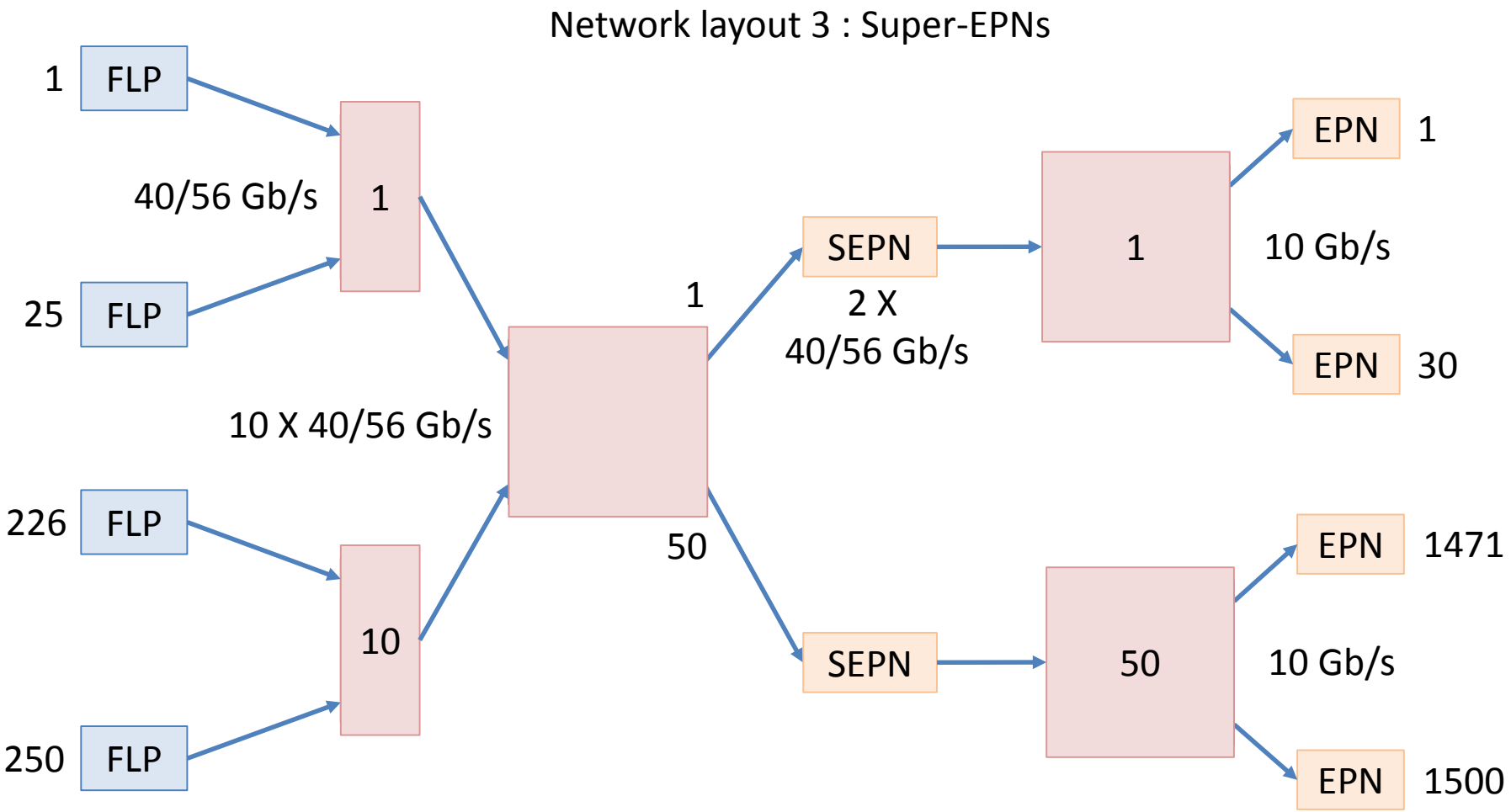
Data type	Size (GB)	Tape copy
TF (Pb-Pb)	10	No
CTF (Pb-Pb)	1.6	Yes
ESD	15% of CTF	No
AOD	10% of CTF	Yes
MC	100% of CTF	No
MCAOD	30% of ESD	Yes
HISTO	1% of ESD	No

O2 facility design (2)

Network layout 2 : 4 independent EPN subfarms

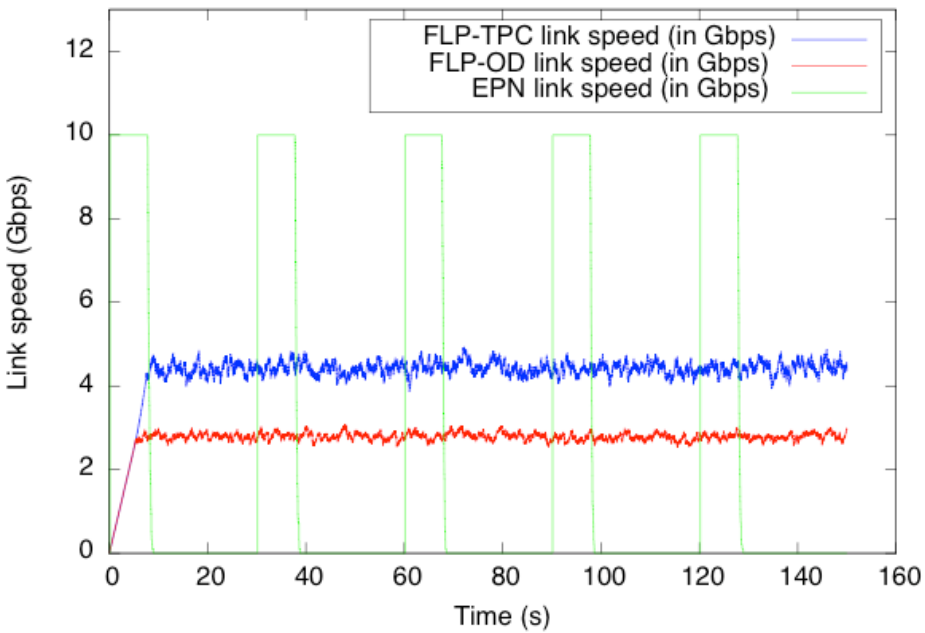


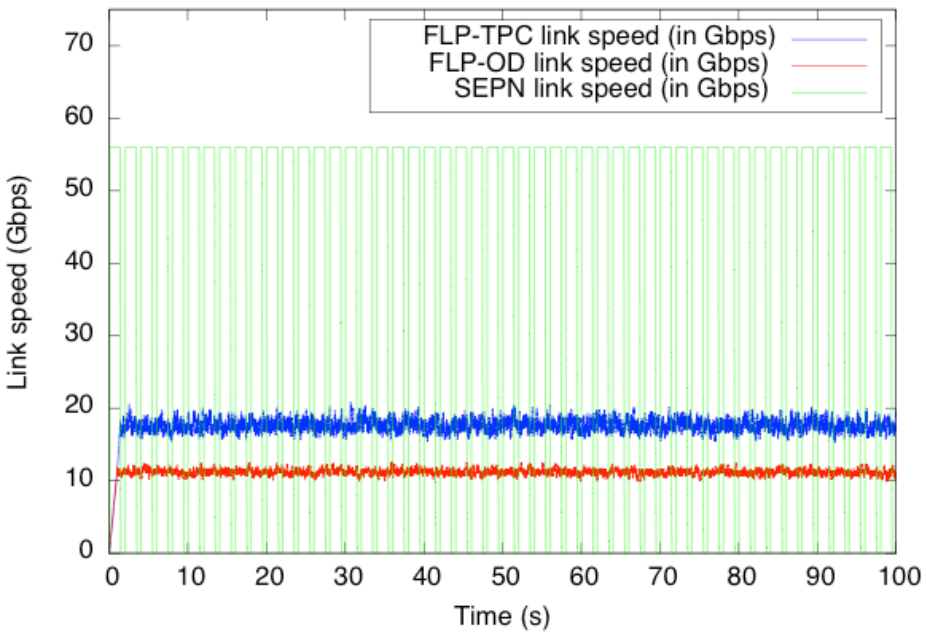
O2 facility design (3)



O2 facility design (4)

Simulation – Link speed



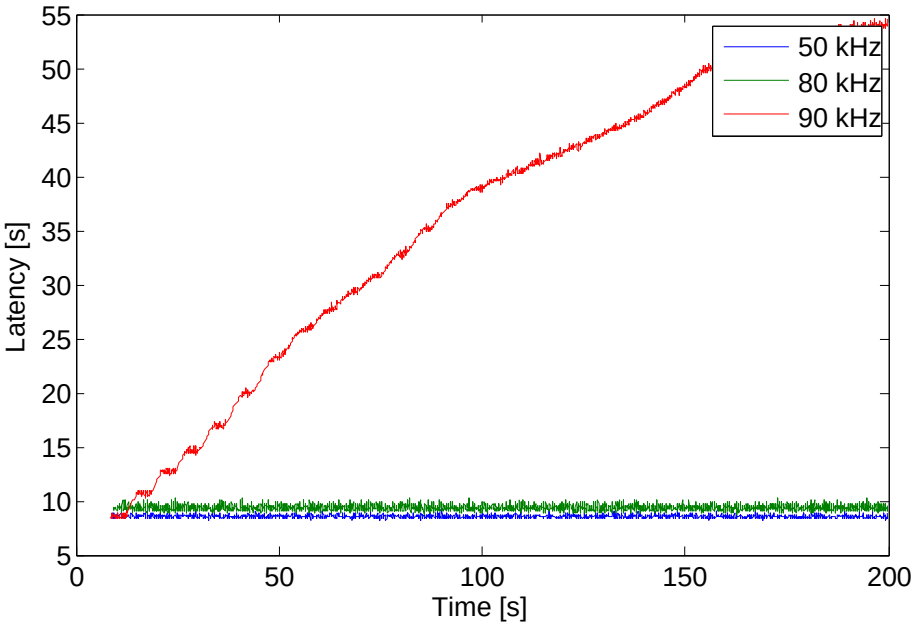


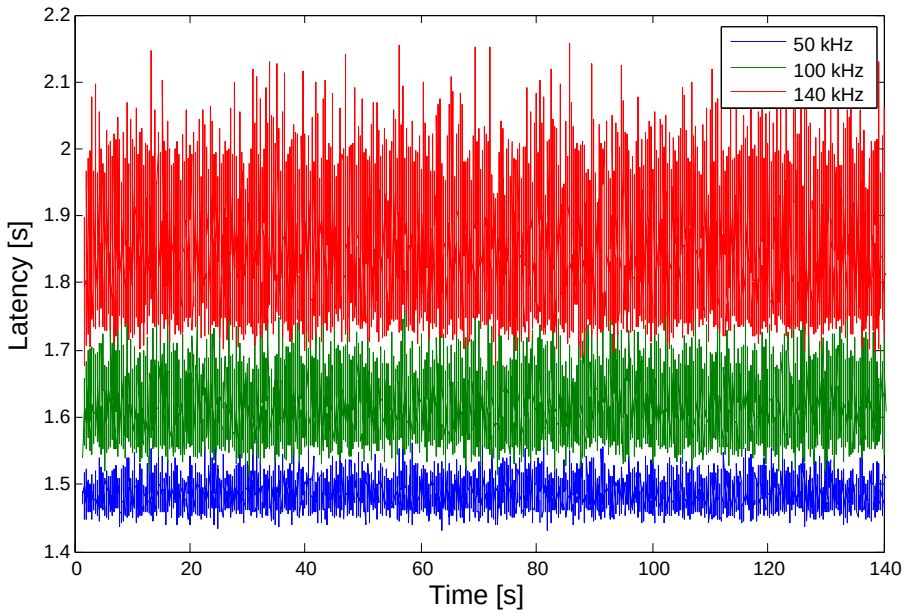
Left : Network Layout 2 :
Link speed on the FLPs and EPNs for a network layout with 4 EPN subfarms for 100 parallel transfers from the FLPs.

Right : Network Layout 3 :
Link speed on the FLPs and Super-EPNs
(configuration based on an Infiniband network at 56 Gb/s)

O2 facility design (6)

Simulation - system scalability





Latency of the timeframes for different interaction rates using layout 2 (left) and layout 3 (right)

→ Layout 2 is cheaper but scales up to 90kHz only.

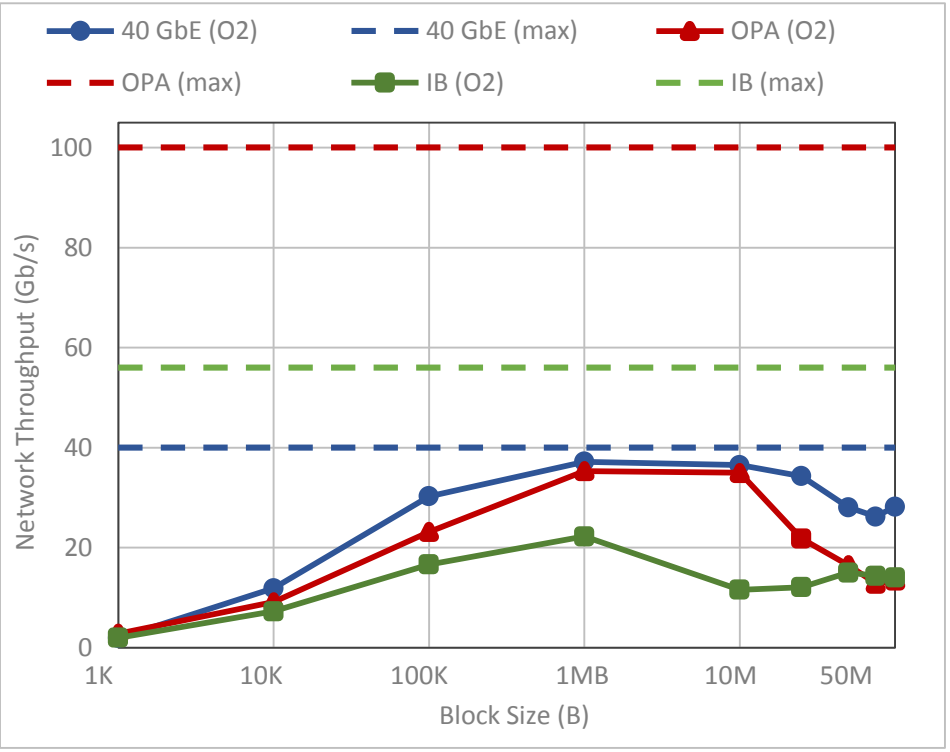
O2 facility – Power and cooling

Table 10.5: Location, rack space, power and cooling needs of the O^2 facility.

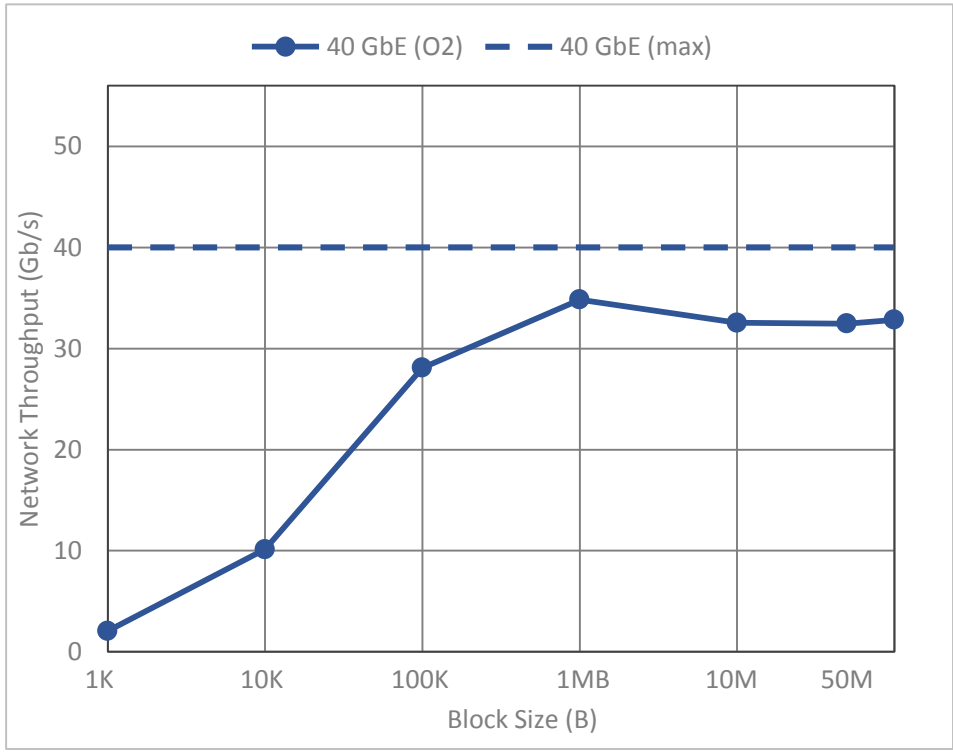
Type	Number of items	Item height (U)	Location-Total height (U)	Number of Racks	Power per rack (kVA)	Total power (kVA)	Cooling per rack (kW)	Total cooling (kW)
FLP	250	2	CR1 - 35	18	12	216	14	252
EPN	1500	1	CR0 - 54	34	50	1700	50	1700
SEPN	50	1	CR0 - 54	1	12	12	12	12
Storage	34	9	CR0 - 54	7	50	350	50	350
Network								
Dataflow	10	3	CR1 - 40	2	12	24	12	24
Dataflow	15	3	CR0 - 54	2	12	24	12	24
Control	4	1	CR1 - 40	1	12	12	12	12
Control	14	1	CR0 - 54	1	12	12	12	12
Services	110	1	CR1 - 40	4	12	48	12	48
Total			CR1	25		300		336
			CR0	45		2086		2086
Grand total				70		2398		2434

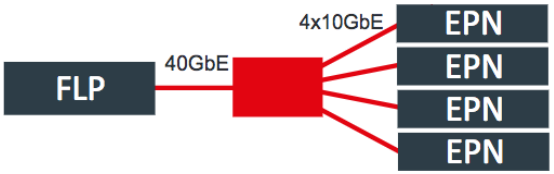
Network performance tests

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



Saturating the Sender
with Ethernet





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