A Large Ion Collider Experiment

O^2 Project :Upgrade of the online and offline computing

Pierre VANDE VYVRE

Requirements

Focus of ALICE upgrade on physics probes requiring high statistics: sample 10 nb^{-1}

Online System Requirements

Sample full 50kHz Pb-Pb interaction rate

- current limit at ~500Hz, factor 100 increase
- system to scale up to 100 kHz

⇨ **~1.1 TByte/s detector readout**

However:

- Storage bandwidth limited to a much lower value (design decision/cost)
- Many physics probes have low S/B:
  classical trigger/event filter approach not efficient

O^2 System from the Letter of Intent

Design Guidelines

Handle >1 TByte/s detector input

Produce (timely) physics result

} Online Reconstruction to reduce data volume
Output of System AODs

Minimize “risk” for physics results

- ➪ Allow for reconstruction with improved calibration, e.g. store clusters associated to tracks instead of tracks
- ➪ Minimize dependence on initial calibration accuracy
- ➪ Implies “intermediate” storage format

Keep cost “reasonable”

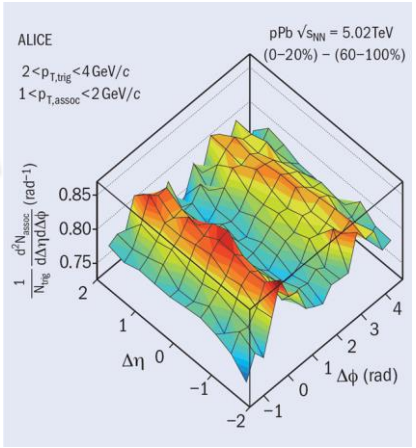
- ➪ Limit storage system bandwidth to ~80 GB/s peak and 20 GByte/s average
- ➪ Optimal usage of compute nodes

Reduce latency requirements & increase fault-tolerance

O^2 Project

Requirements

Detector	Input to Online System (GByte/s)	Peak Output to Local Data Storage (GByte/s)	Avg. Output to Computing Center (GByte/s)
TPC	1000	50.0	8.0
TRD	81.5	10.0	1.6
ITS	40	10.0	1.6
Others	25	12.5	2.0
Total	**1146.5**	**82.5**	**13.2**



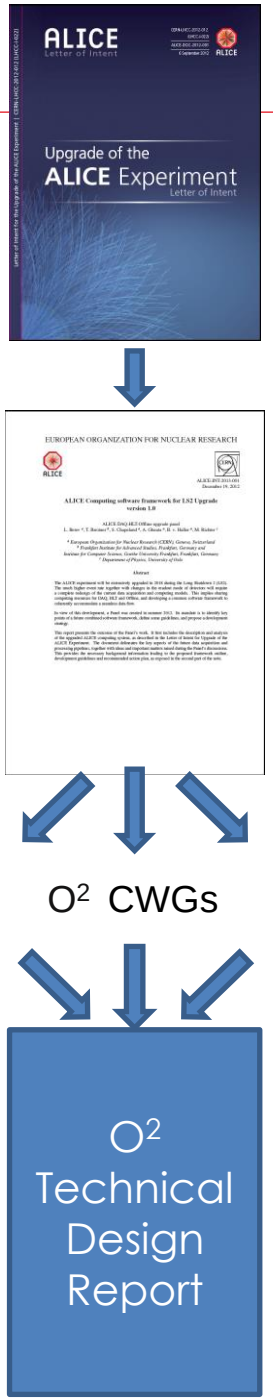
- Handle >1 TByte/s detector input
- Support for continuous read-out
- Online reconstruction to reduce data volume
- Common hw and sw system developed by the DAQ, HLT, Offline teams

O^2 Project

Project Organization

PLs: P. Buncic, T. Kollegger, P. Vande Vyvre

Computing Working Group(CWG)		Chair
1.	Architecture	S. Chapeland
2.	Tools & Procedures	A. Telesca
3.	Dataflow	T. Breitner
4.	Data Model	A. Gheata
5.	Computing Platforms	M. Kretz
6.	Calibration	C. Zampolli
7.	Reconstruction	R. Shahoyan
8.	Physics Simulation	A. Morsch
9.	QA, DQM, Visualization	B. von Haller
10.	Control, Configuration, Monitoring	V. Chibante
11.	Software Lifecycle	A. Grigoras
12.	Hardware	H. Engel
13.	Software framework	P. Hristov

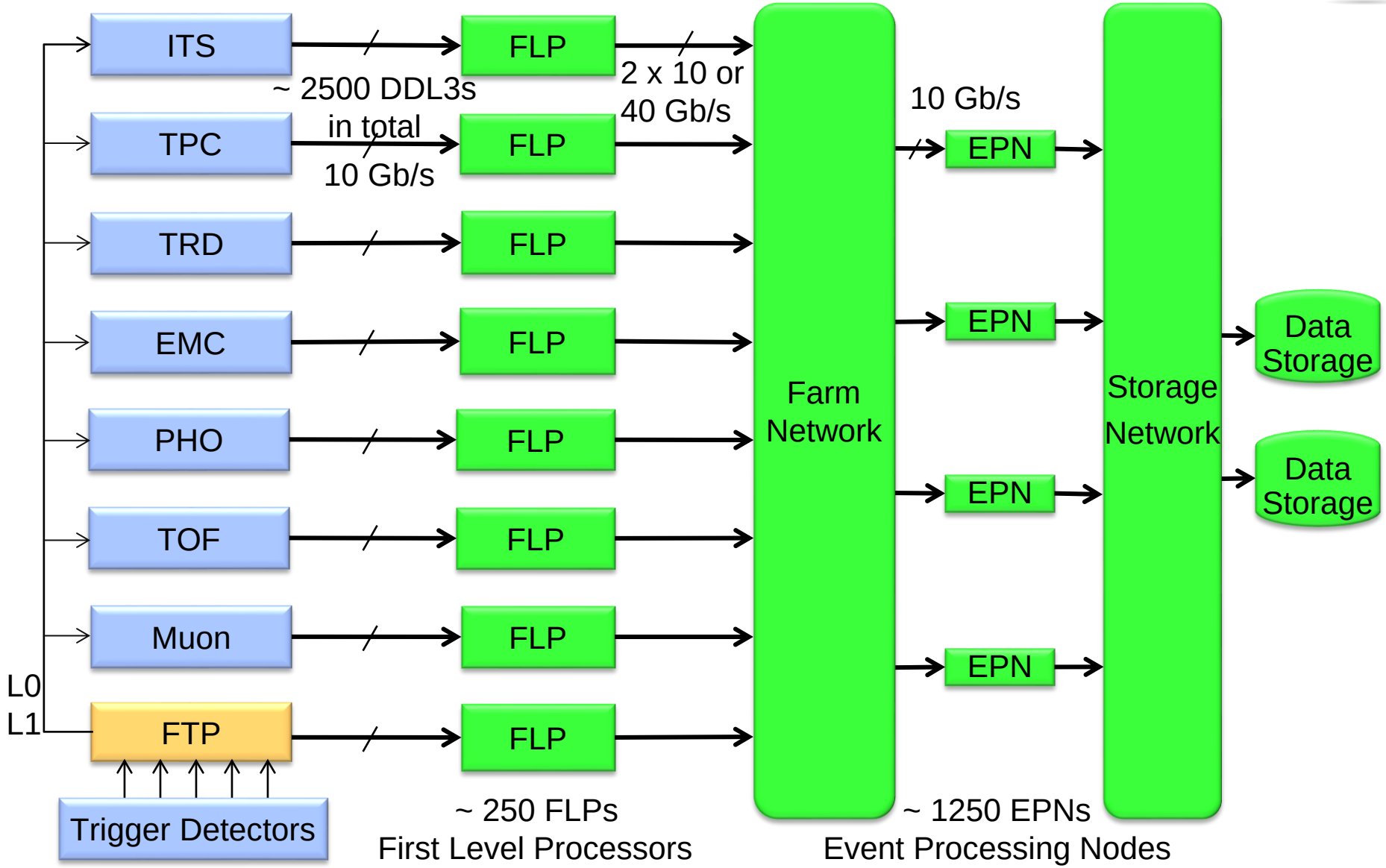


Editorial Committee

L. Betev, P. Buncic, S. Chapeland, F. Cliff, P. Hristov, T. Kollegger, M. Krzewicki, K. Read, J. Thaeder, B. von Haller, P. Vande Vyvre

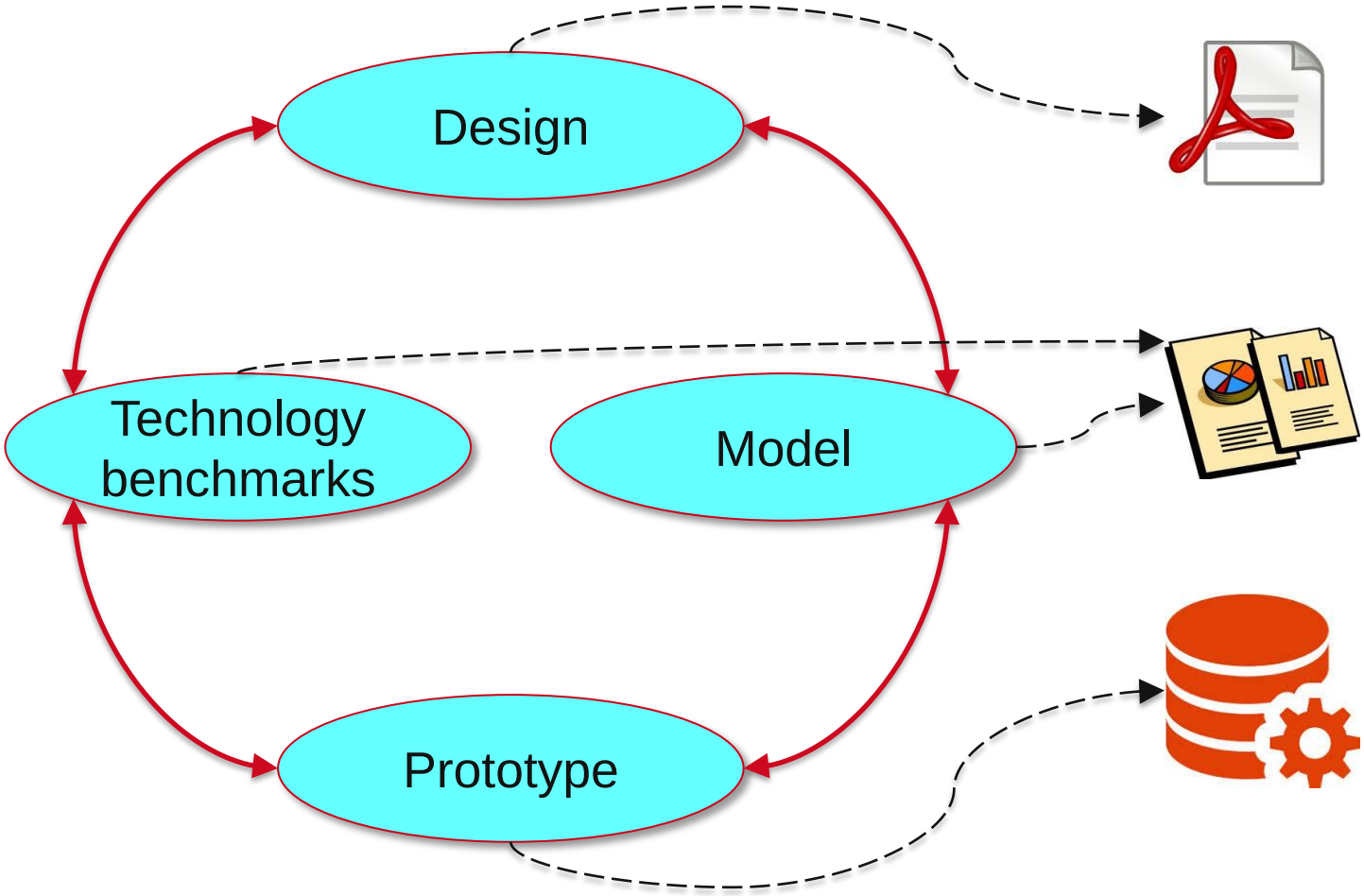
Physics requirement chapter: Andrea Dainese

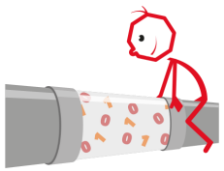
Hardware Architecture



Design strategy

Iterative process: design, benchmark, model, prototype





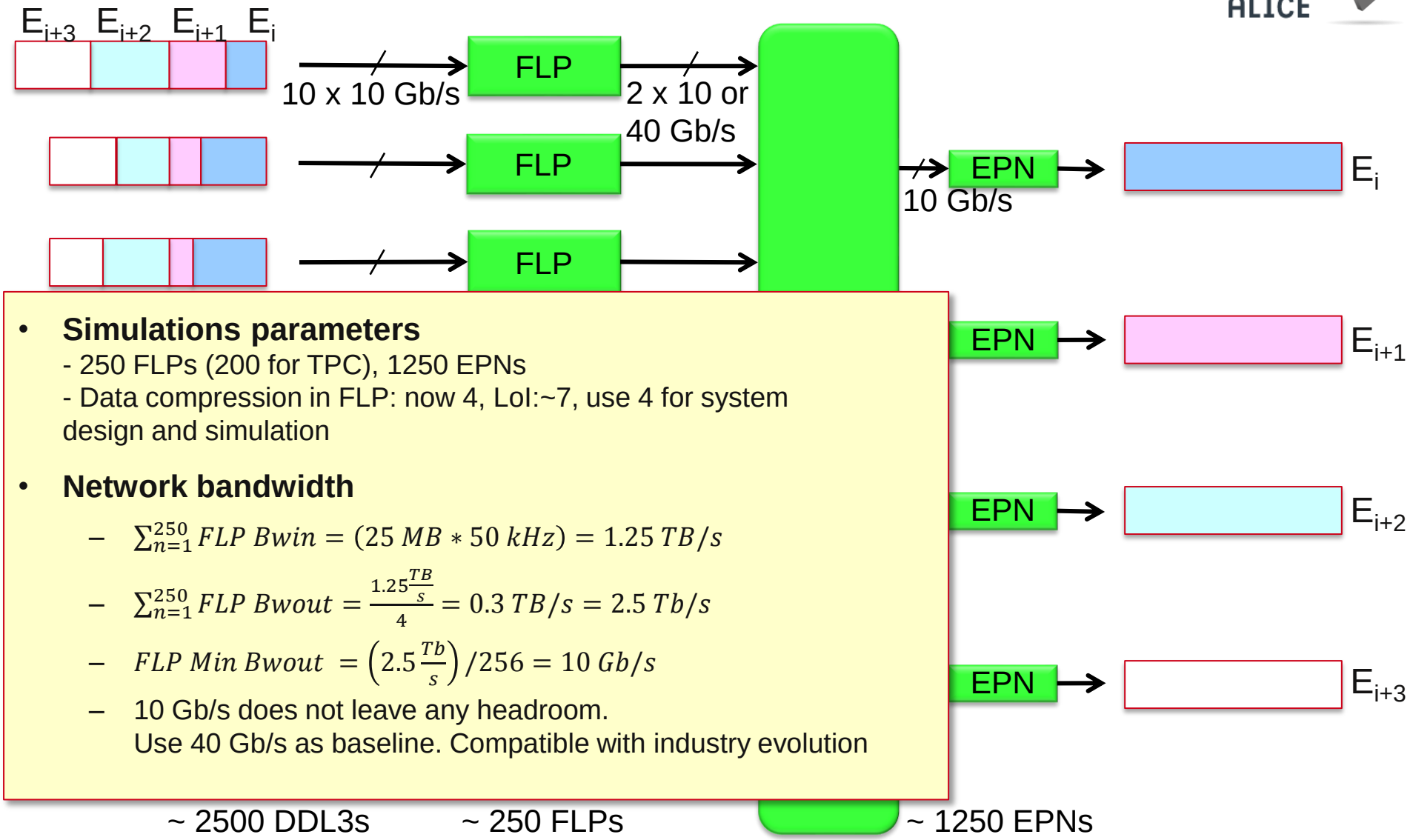
Dataflow

Dataflow modelling

- Dataflow discrete event simulation implemented with OMNET++
  - FLP-EPN data traffic and data buffering
    - Network topologies (central switch; spine-leaf),
    - Data distribution schemes (time frames, parallelism)
    - Buffering needs
    - System dimensions
  - Heavy computing needs

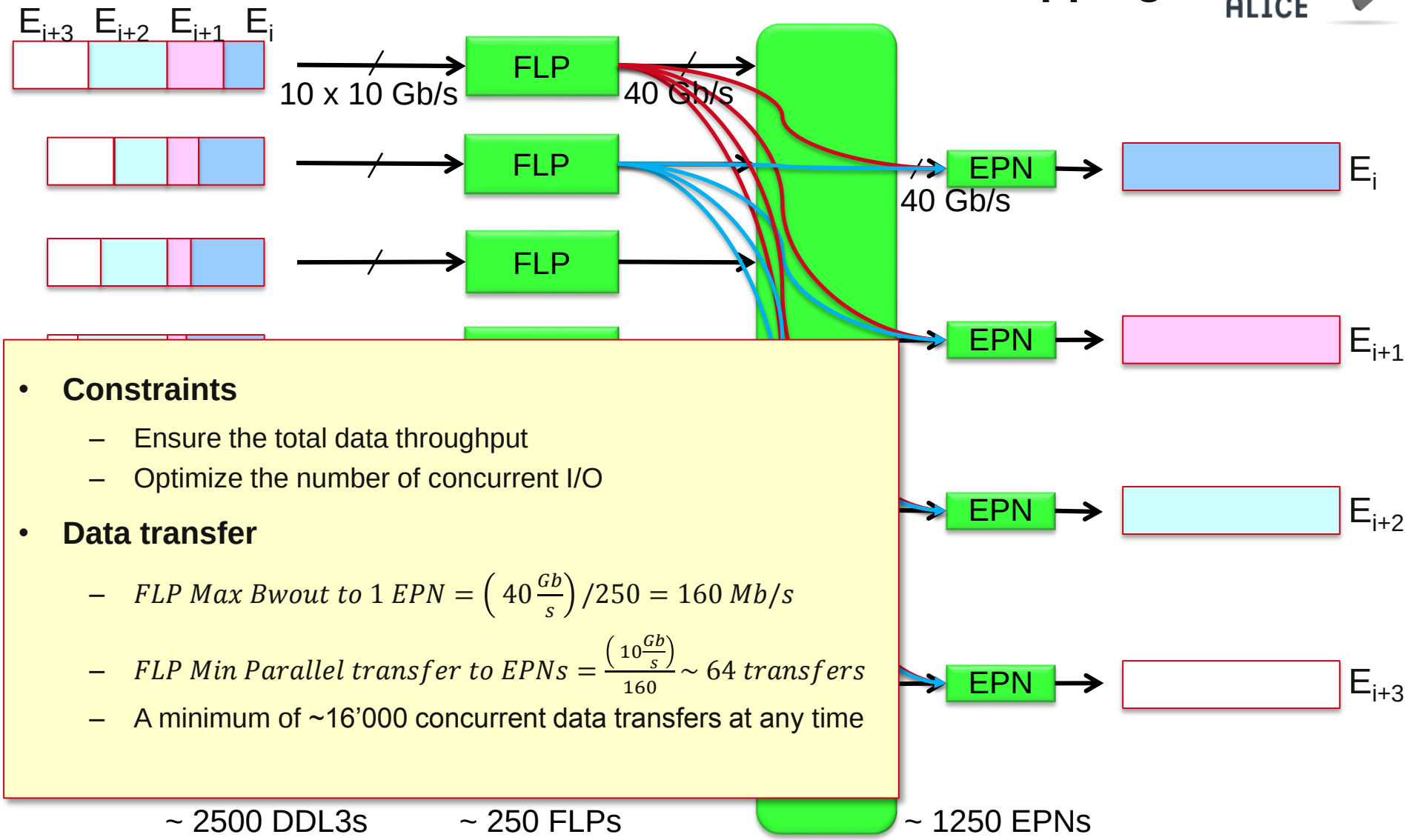
    Downscaling applied for some simulations:
    - Reduce network bandwidth and buffer sizes
    - Simulate a slice of the system
- System global simulation with ad-hoc program

Hardware Architecture: FLP-EPN data transport model



- **Simulations parameters**
  - 250 FLPs (200 for TPC), 1250 EPNs
  - Data compression in FLP: now 4, LoI:~7, use 4 for system design and simulation
- **Network bandwidth**
  - $\sum_{n=1}^{250} FLP\ Bwin = (25\ MB * 50\ kHz) = 1.25\ TB/s$
  - $\sum_{n=1}^{250} FLP\ Bwout = \frac{1.25\frac{TB}{s}}{4} = 0.3\ TB/s = 2.5\ Tb/s$
  - $FLP\ Min\ Bwout\ = \left(2.5\frac{Tb}{s}\right)/256 = 10\ Gb/s$
  - 10 Gb/s does not leave any headroom.
    Use 40 Gb/s as baseline. Compatible with industry evolution

Hardware Architecture: FLP-EPN traffic shapping

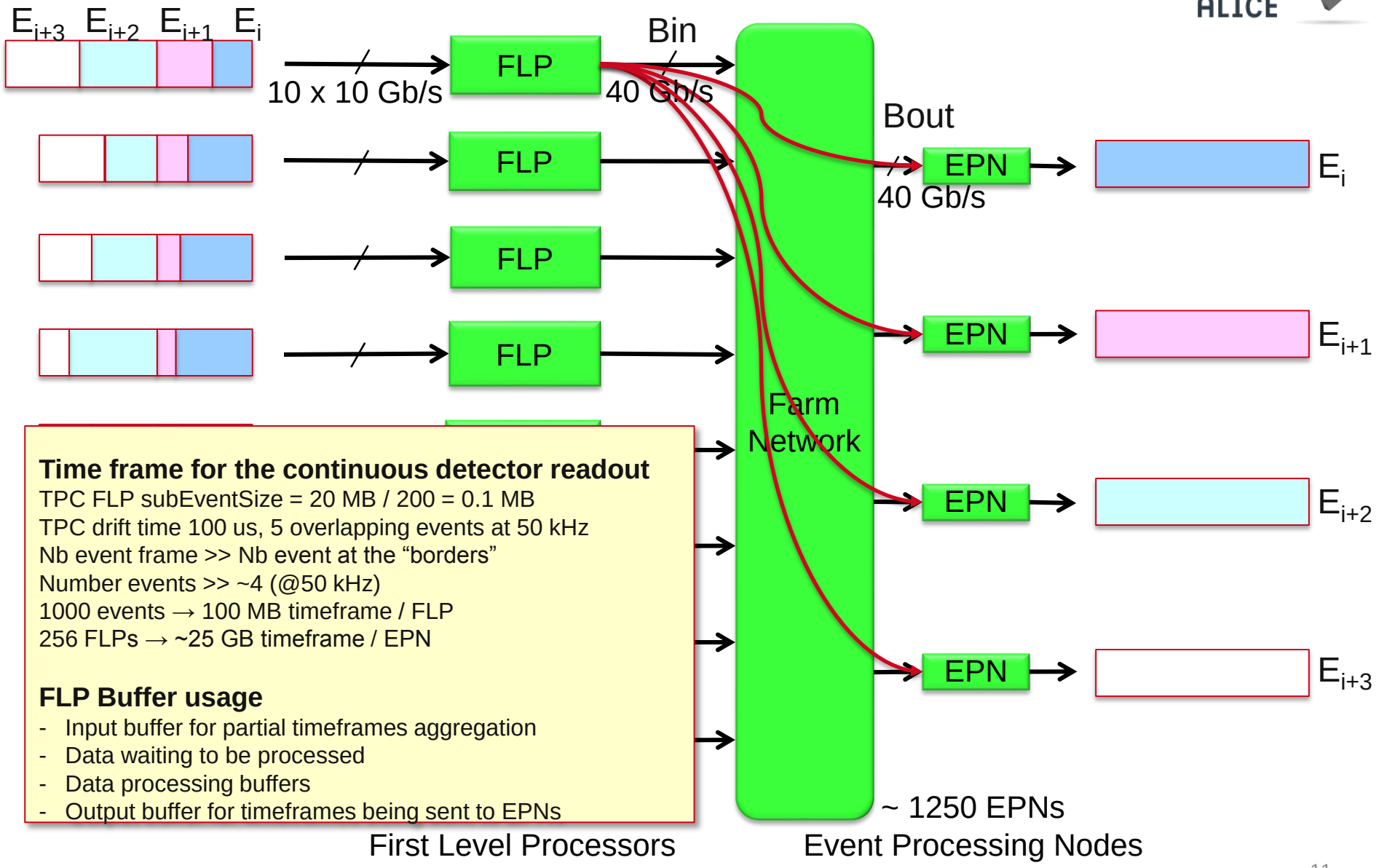


- **Constraints**
  - Ensure the total data throughput
  - Optimize the number of concurrent I/O
- **Data transfer**
  - $FLP\ Max\ Bwout\ to\ 1\ EPN = \left(40\frac{Gb}{s}\right)/250 = 160\ Mb/s$
  - $FLP\ Min\ Parallel\ transfer\ to\ EPNs = \frac{\left(10\frac{Gb}{s}\right)}{160} \sim 64\ transfers$
  - A minimum of ~16'000 concurrent data transfers at any time

~ 2500 DDL3s ~ 250 FLPs ~ 1250 EPNs

First Level Processors Event Processing Nodes

Hardware Architecture: FLP buffer size

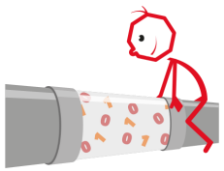


Time frame for the continuous detector readout

TPC FLP subEventSize = 20 MB / 200 = 0.1 MB
TPC drift time 100 us, 5 overlapping events at 50 kHz
Nb event frame >> Nb event at the “borders”
Number events >> ~4 (@50 kHz)
1000 events → 100 MB timeframe / FLP
256 FLPs → ~25 GB timeframe / EPN

FLP Buffer usage

- Input buffer for partial timeframes aggregation
- Data waiting to be processed
- Data processing buffers
- Output buffer for timeframes being sent to EPNs



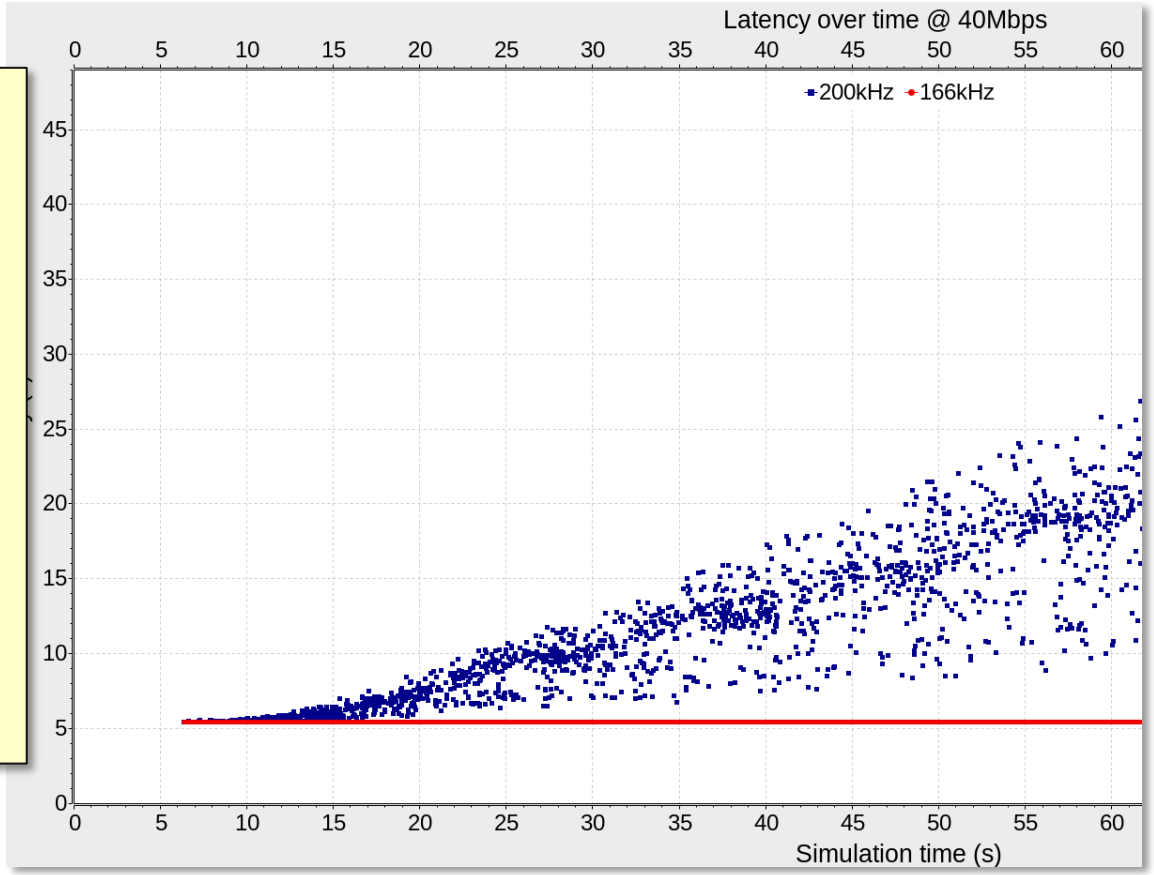
FLP-EPN Dataflow simulation

System scalability study

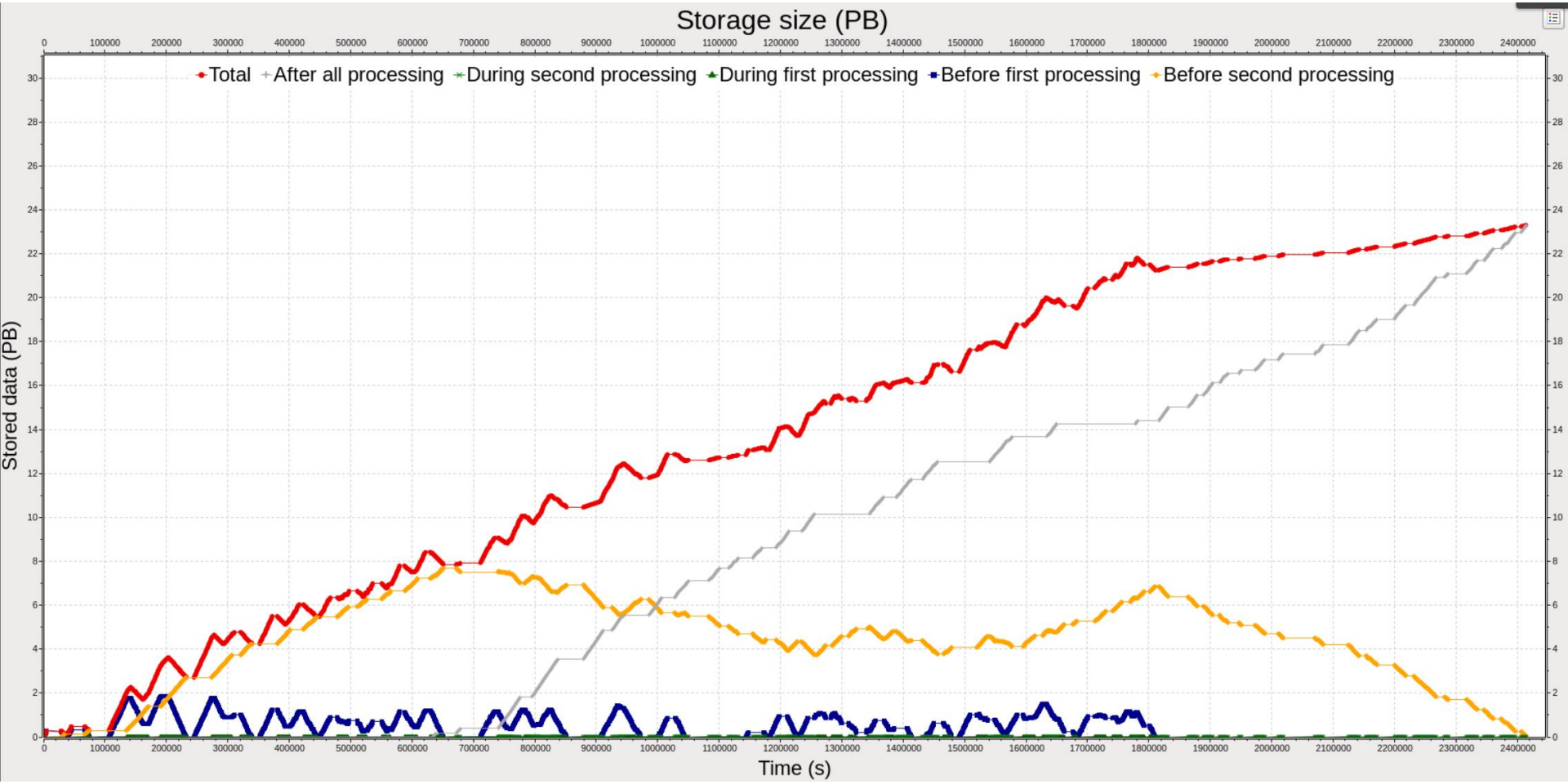
S. Chapeland, C. Delort

- System scalability study
- System studied on a ¼ of the entire system and lower bandwidth to limit the simulation time
- System scales at up to 166 kHz of MB interactions

Configuration 40 Mbps 250x288



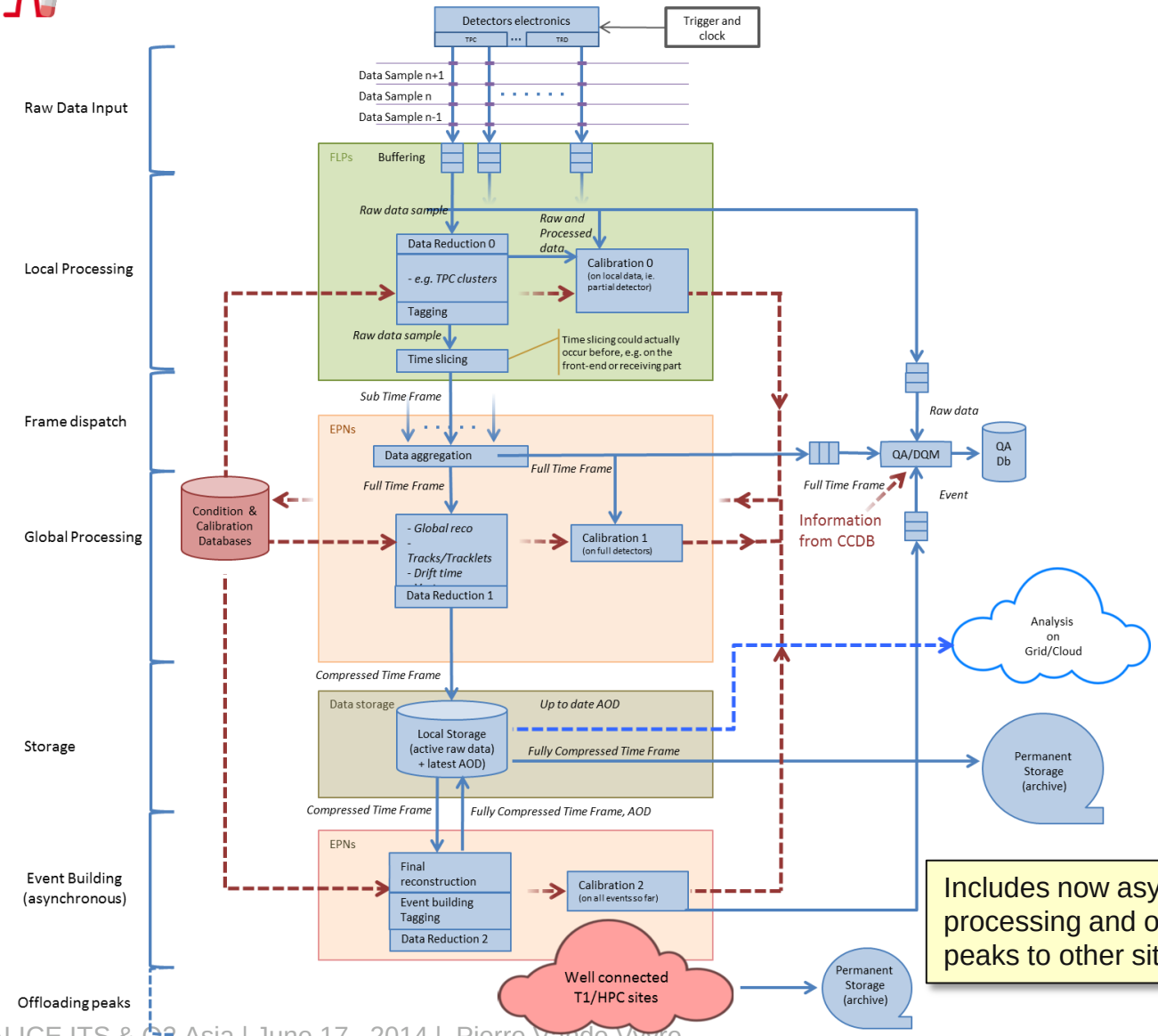
Data storage needs of the O^2 facility



Need of ~25 PB of local data storage for 1 year of data taking

Architecture

O² Architecture and data flow



TPC Data
1 TB/s

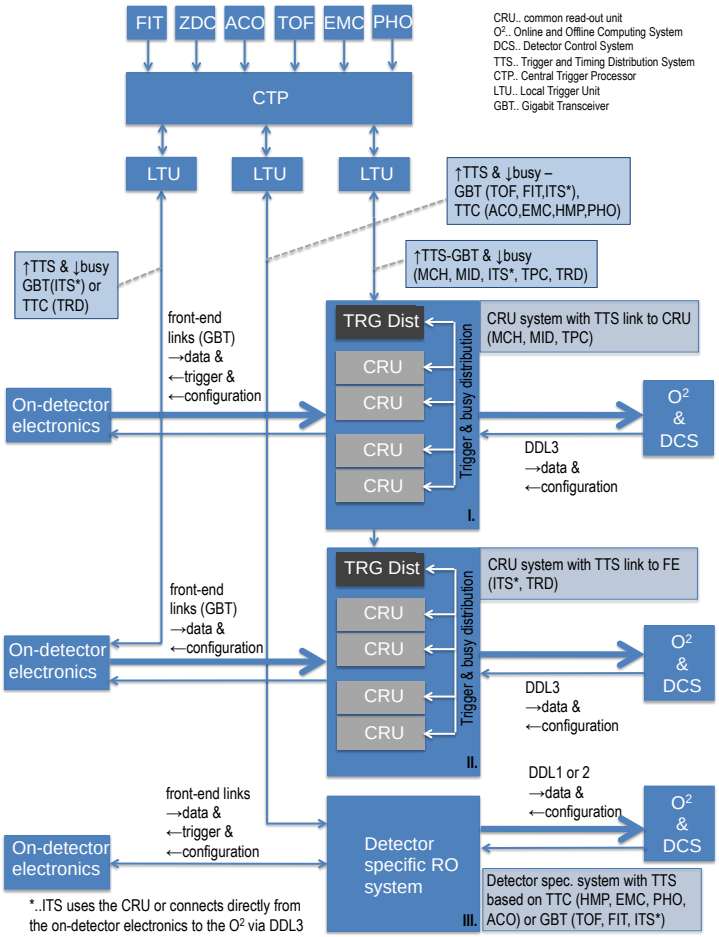
250 GB/s

125 GB/s

80 GB/s

Includes now asynchronous processing and offloading peaks to other sites.

Detector Readout via Detector Data Links (DDLs)



Common Interface to the Detectors:

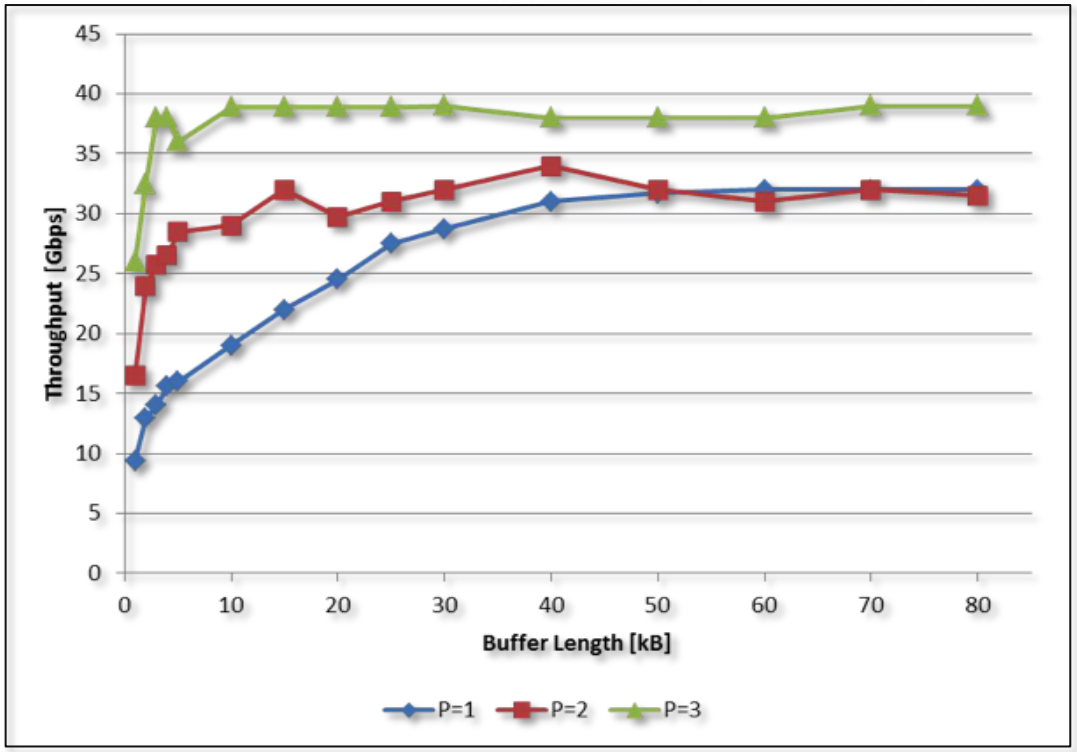
- DDL1 (2.125 Gbit/s)
- DDL2 (5.3125 Gbit/s)
- DDL3 (>=10 Gbit/s)
  - 10 Gbit Ethernet
  - PCIe bus

More development of VHDL code still needed. Looking for more collaborators in this area. See presentation of F. Costa: “Firmware developments for the ALICE Run 2 and Run 3”

FLP and Network prototyping

- FLP requirements
  - Input 100 Gbit/s (10 x 10 Gbit/s)
  - Local processing capability
  - Output with ~20 Gbit/s

- Two network technologies under evaluation
  - 10/40 Gbit/s Ethernet
  - Infiniband FDR (56 Gbit/s)
  - Both used already (DAQ/HLT)



- Benchmark example
- Chelsio T580-LP-CR with TCP/UDP Offload engine 1, 2 and 3 TCP streams, iperf measurements

CWG5: Computing Platforms

The Conversion factors

- Shift from 1 to many platforms
- Speedup of CPU Multithreading:
  - Task takes **n1** seconds on **1** core, **n2** seconds on **x** cores
  - → Speedup is **n1**/**n2** for **x** cores, Factors are **n1**/**n2** and **x**/**1**
  - With Hyperthreading: **n2‘** seconds on **x‘** threads on **x** cores. (**x‘** >= 2**x**)
  - → Will not scale linearly, needed to compare to full CPU performance.
    - Factors are **n1** / **n2‘** and **x** / **1** (Be carefull: Not **x‘** / **1**, we still use only **x** cores.)
- Speedup of GPU v.s. CPU:
  - Should take into account full CPU power (i.e. all cores, hyperthreading).
  - Task on the GPU might also need CPU resources.
    - Assume this occupies **y** CPU cores.
  - Task takes **n3** seconds on GPU.
  - Speedup is **n2‘**/**n3**, Factors are **n2‘**/**n3** and **y**/**x**. (Again **x** not **x‘**.)
- How many CPU cores does the GPU save:
  - Compare to **y** CPU cores, since the GPU needs that much resources.
  - Speedup is **n1** / **n3**, GPU Saves **n1** / **n3** – **y** CPU cores.
  - → Factors are **n1** / **n3**, **y** / **1**, and **n1** / **n3** - **y**.
- Benchmarks: Track Finder, Track Fit, DGEMM (Matrix Multiplication – Synthetic)

CWG5: Computing Platforms

Track finder

Nehalem 4-Core 3,6 GHz (Smaller Event than others)		
1 Thread	3921 ms	**Factors:**
4 Threads	1039 ms	3,77 / 4
12 Threads (x = 4, x‘ = 12)	816 ms	4,80 / 4
Westmere 6-Core 3.6 GHz		
1 Thread	4735 ms	**Factors:**
6 Threads	853 ms	5.55 / 6
12 Threads (x = 4, x‘ = 12)	506 ms	9,36 / 6
Dual Sandy-Bridge 2 * 8-Core 2 GHz		
1 Thread	4526 ms	**Factors:**
16 Threads	403 ms	11,1 / 16
36 Threads (x = 16, x‘ = 36)	320 ms	14,1 / 16
Dual AMD Magny-Cours 2 * 12-Core 2,1 GHz		
36 Threads (x = 24, x‘ = 36)	495 ms	

3 CPU Cores + GPU – All Compared to Sandy Bridge System			
		Factor vs x‘ (Full CPU)	**Factor vs 1 (1 CPU Core)**
GTX580	174 ms	1,8 / 0,19	26 / 3 / 23
GTX780	151 ms	2,11 / 0,19	30 / 3 / 27
Titan	143 ms	2,38 / 0,19	32 / 3 / 29
S9000	160 ms	2 / 0,19	28 / 3 / 25
S10000 (Dual GPU with 6 CPU cores	85 ms	3,79 / 0,38	54 / 6 / 48

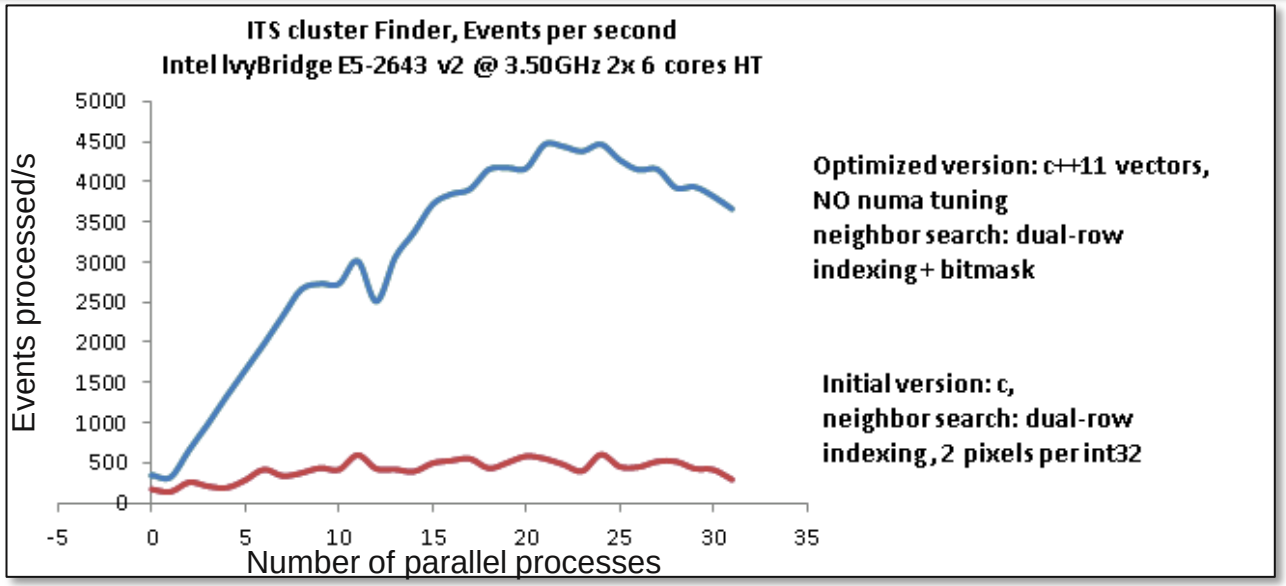
Computing Platforms

ITS Cluster Finder

- Use the ITS cluster finder as optimization use case and as benchmark
- Initial version memory-bound
- Several data structure and algorithms optimizations applied

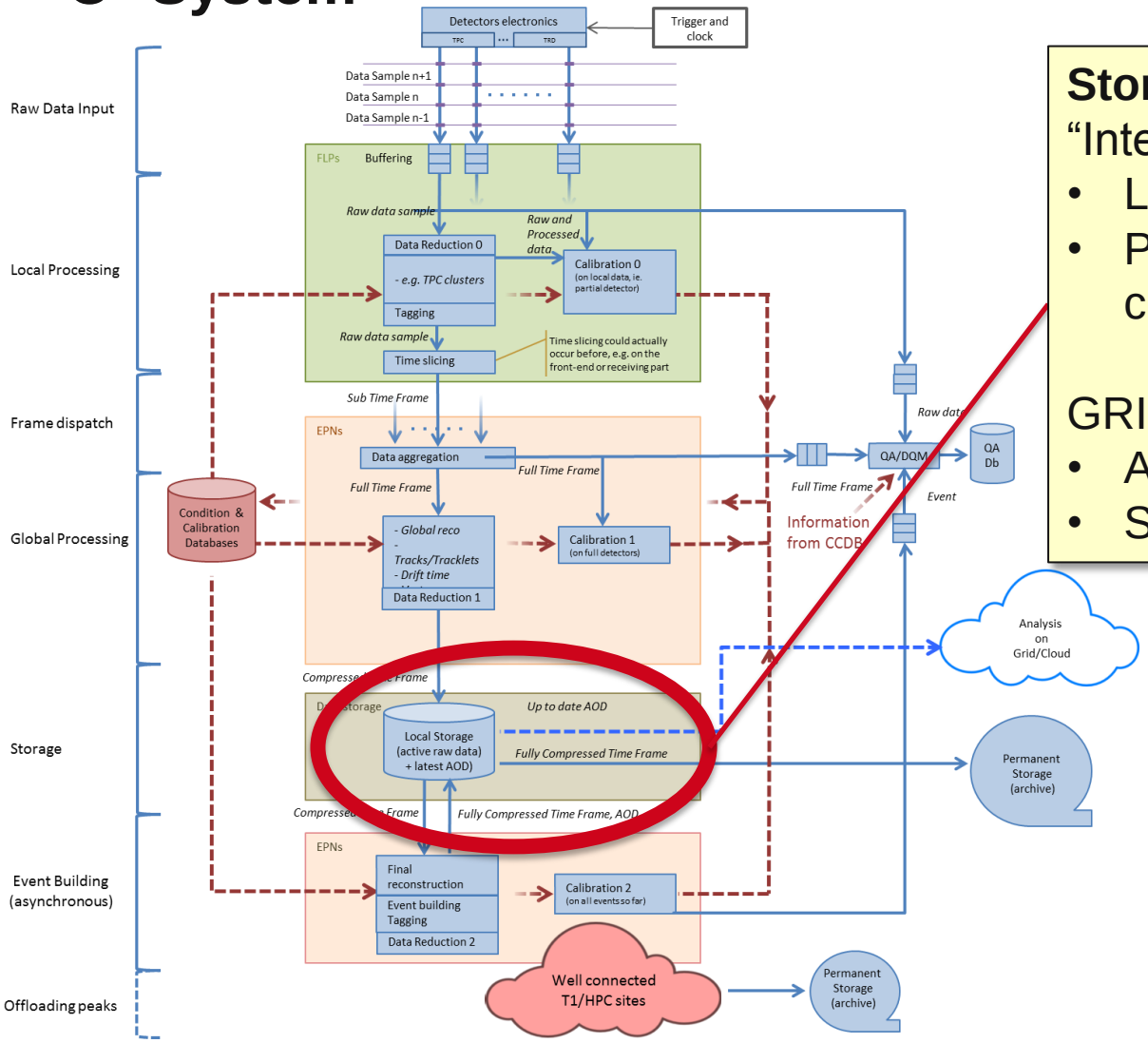
See the presentation of Prof. T. Achalakul about benchmarking

More benchmarking of detector-specific code still needed.
Looking for more collaborators in this area. See presentation of S. Chapeland “Benchmarks for the ITS cluster finder”



S. Chapeland

O^2 System



Storage

"Intermediate" format:

- Local storage in O2 system
- Permanent storage in computing center

GRID storage

- AODs
- Simulations

Data Storage

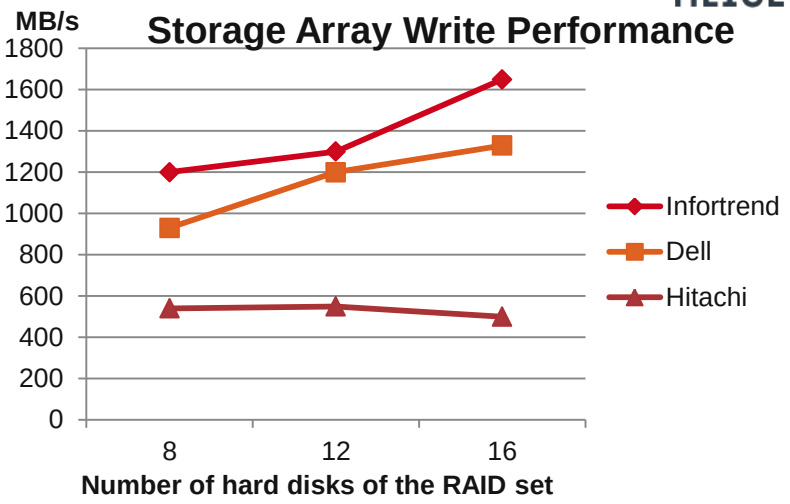
80 GB/s over ~1250 nodes

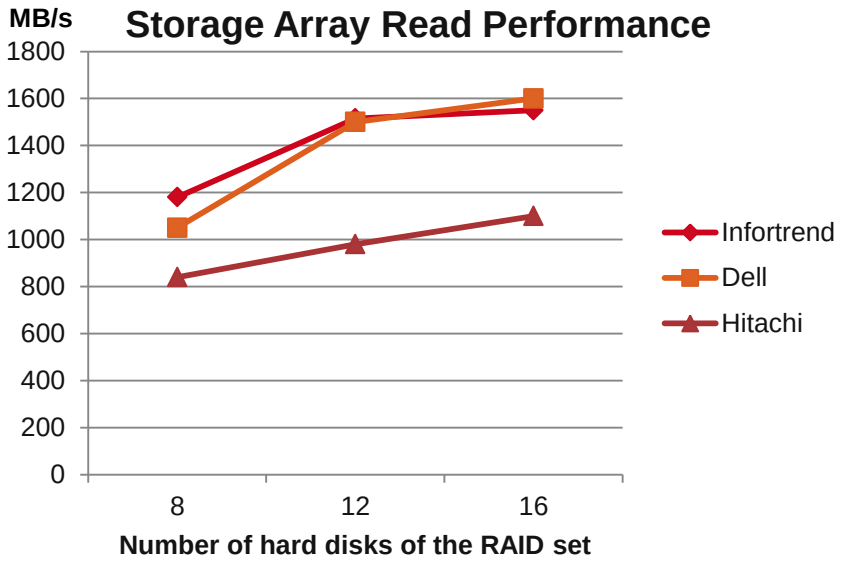
Option 1: SAN
(currently used in the DAQ)

Centralized pool of storage arrays,
Dedicated network

5 racks (same as today)
would provide 40 PB

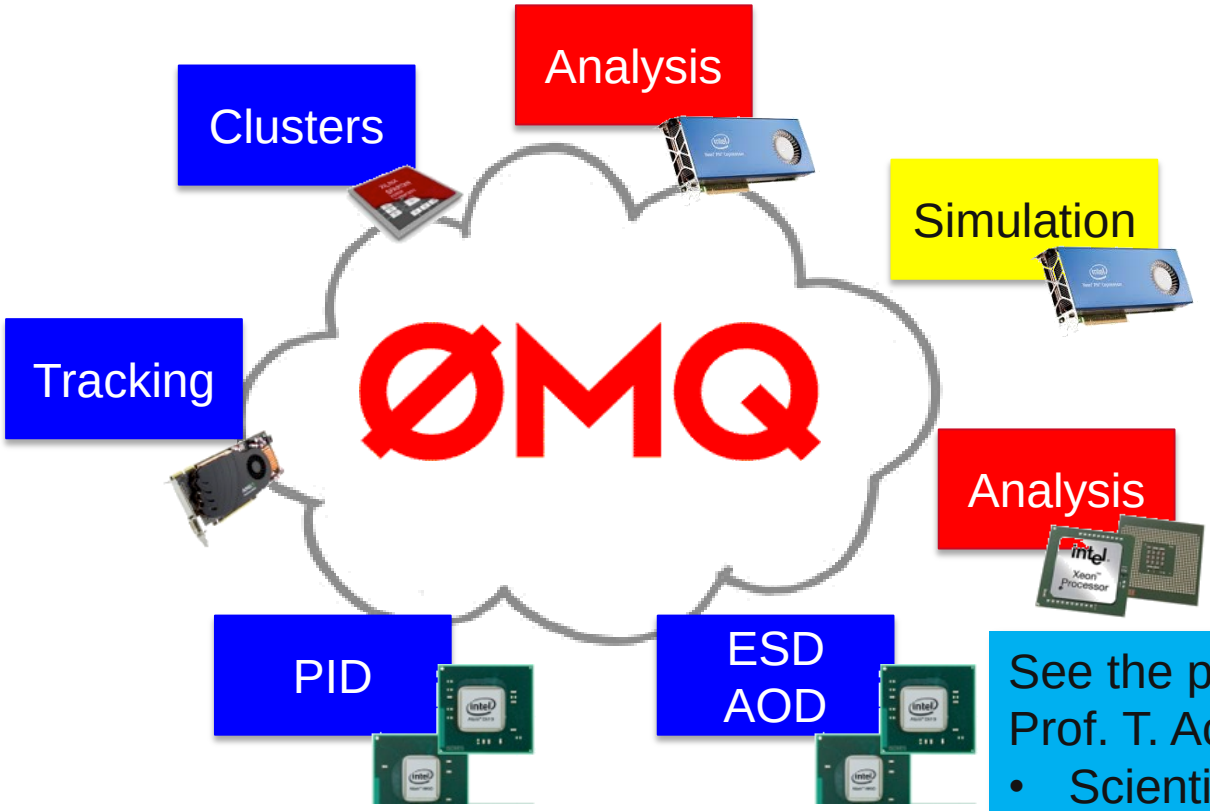
Option 2: DAS
Distributed data storage
1 or a few 10 TB disks in each node





Software Framework

- Multi-platforms
- Multi-applications
- Public-domain software

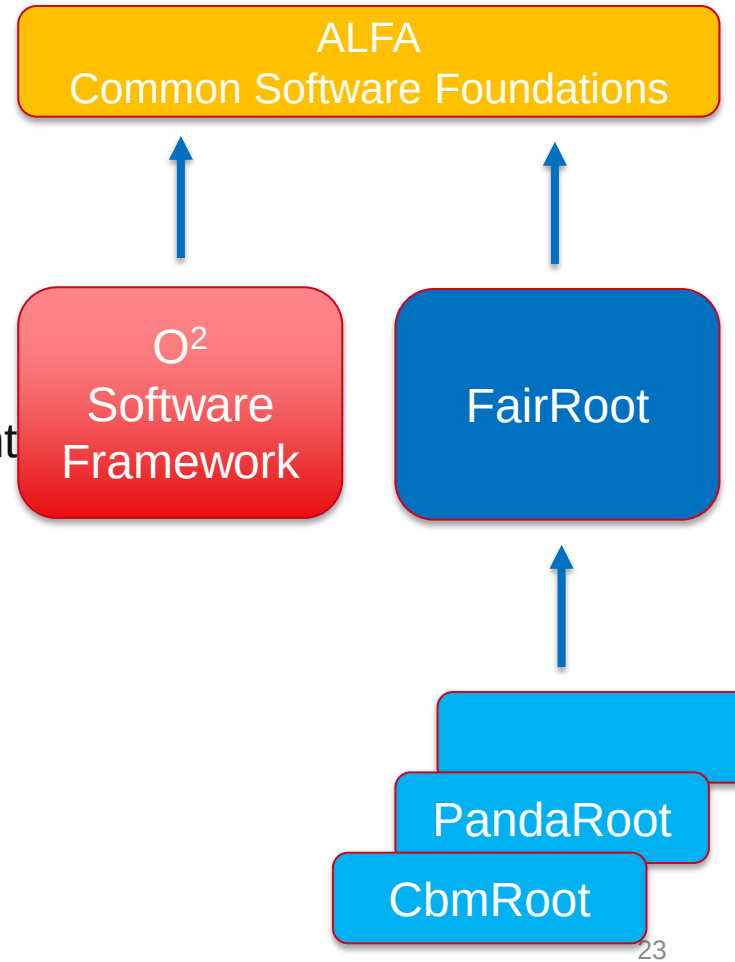


See the presentations of
Prof. T. Achalakul and K. Chanchio

- Scientific computing scheduling
- Live checkpointing and migration

Software Framework Development

- Design and development of a new modern framework targeting Run3
- Should work in Offline and Online environment
  - Has to comply with O^2 requirements and architecture
- Based on new technologies
  - Root 6.x, C++11
- Optimized for I/O
  - New data model
- Capable of utilizing hardware accelerators
  - FPGA, GPU, MIC…
- Support for concurrency and distributed environment
- Based on ALFA - common software foundation developed jointly between ALICE & GSI/FAIR



Large development in progress.
Looking for more collaborators in this area.
See presentation of P. Hristov:
“Software framework development”

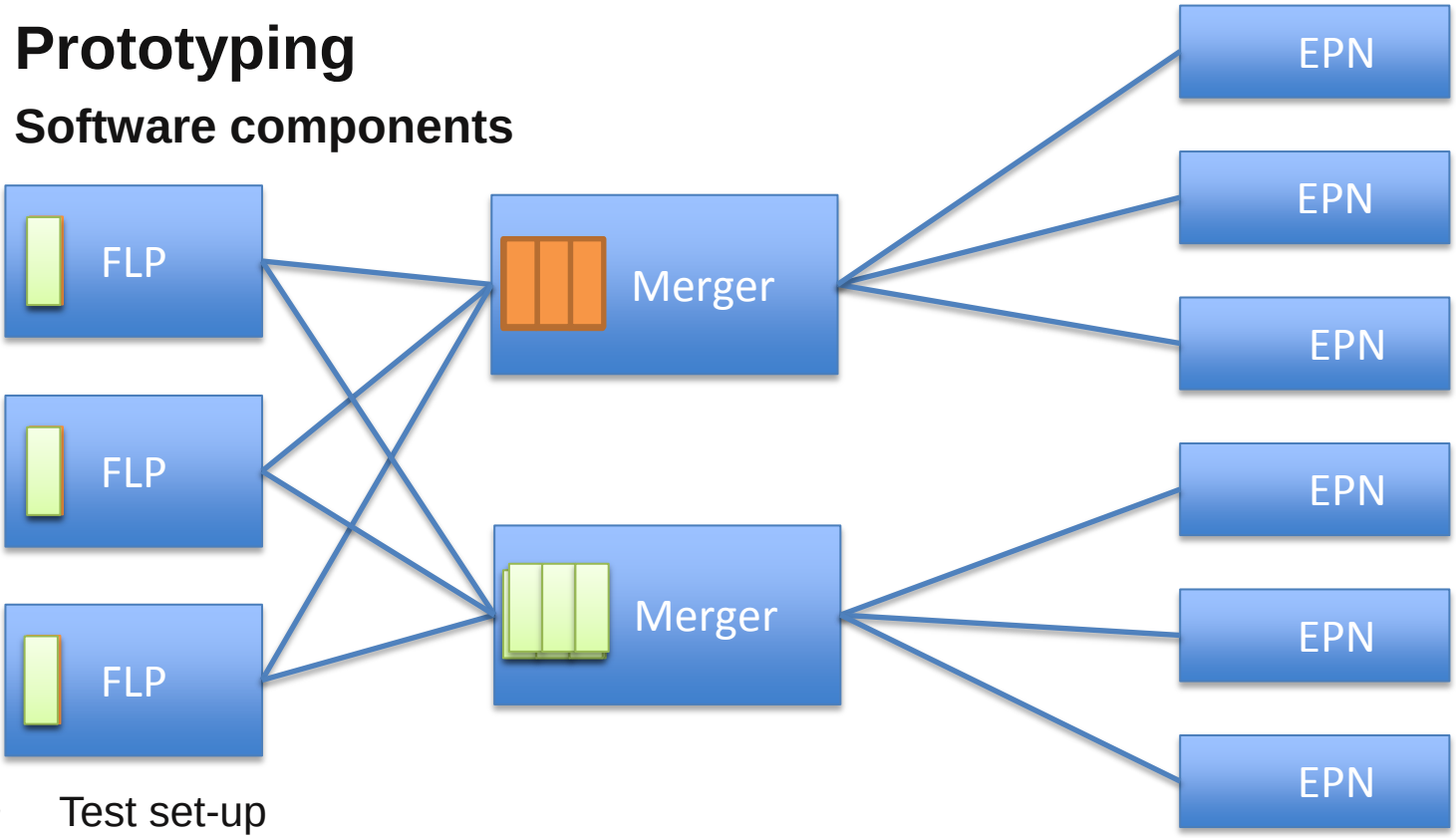
Software Framework Development

ALICE + FAIR = ALFA

- Expected benefits
  - Development cost optimization
  - Better coverage and testing of the code
  - Documentation, training and examples.
  - **ALICE** : work already performed by the FairRoot team concerning features (e.g. the continuous read-out), which are part of the ongoing FairRoot development.
  - **FAIR** experiments : ALFA could be tested with real data and existing detectors before the start of the FAIR facility.
- The proposed architecture will rely:
  - A dataflow based model
  - A process-based paradigm for the parallelism
    - Finer grain than a simple match 1 batch on 1 core
    - Coarser grain than a massively thread-based solution

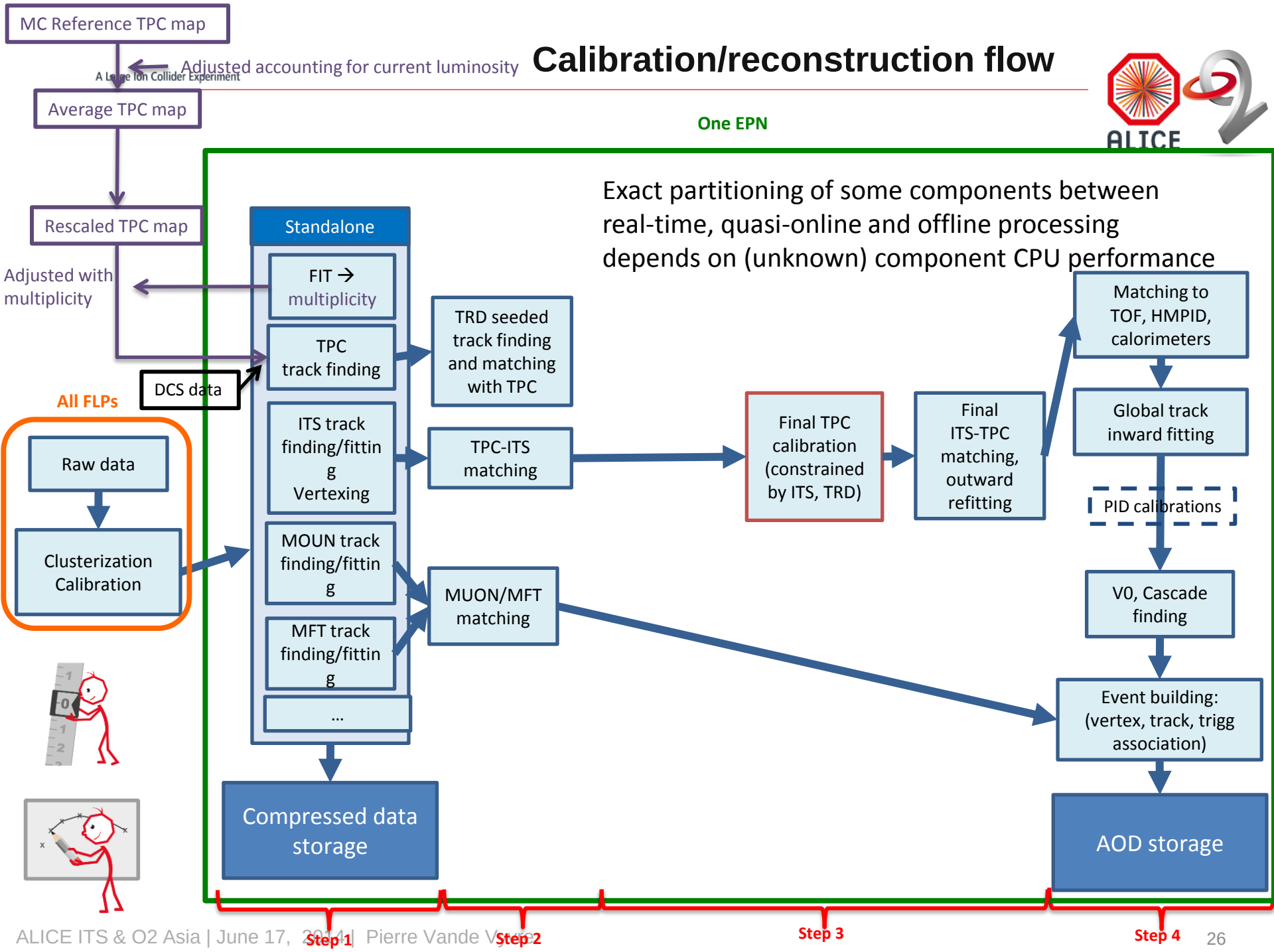
Prototyping

Software components

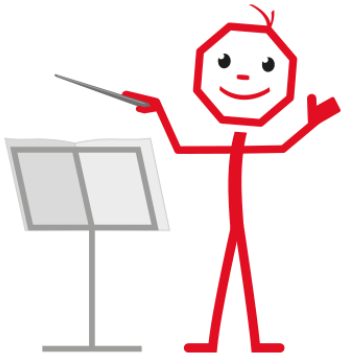


- Test set-up
  - 8 machines
    - Sandy Bridge-EP, dual E5-2690 @ 2.90GHz, 2x8 hw cores - 32 threads, 64GB RAM
  - Network
    - 4 nodes with 40 G Ethernet, 4 nodes with 10 G Ethernet
- Software framework prototype by members of DAQ, HLT, Offline, FairRoot teams
  - Data exchange messaging system
  - Interfaces to existing algorithmic code from offline and HLT

Calibration/reconstruction flow

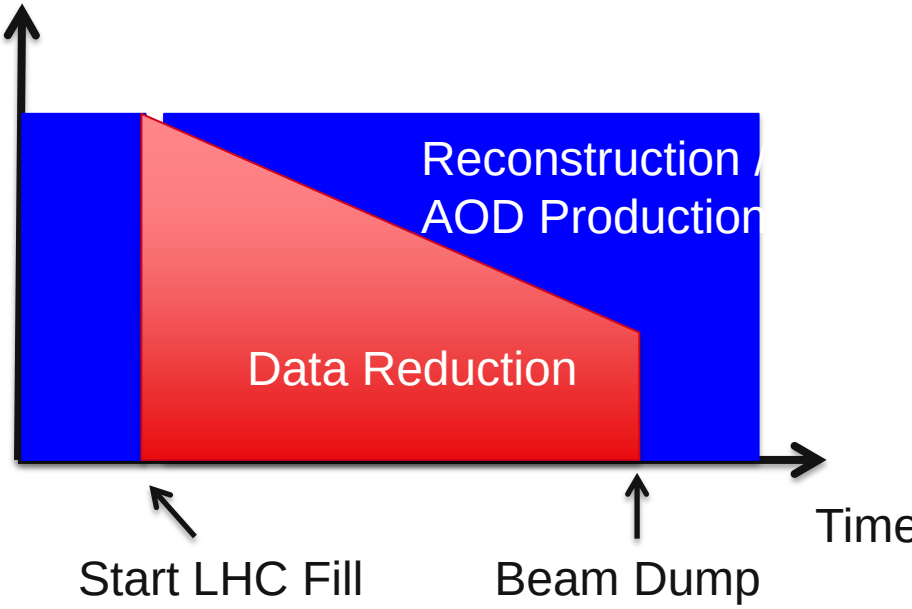


Control, Configuration and Monitoring



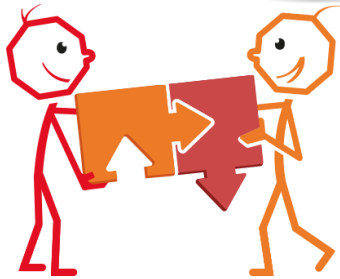
Large computing farm with many concurrent activies
Software Requirements Specifications
Tools survey document
Tools under test

- Monitoring: Mona Lisa, Ganglia, Zabbix
- Configuration: Puppet, Chef



System design and evaluation
of several tools in progress.
Looking for more collaborators in this area.
See presentation of V. Chibante
“Control, Configuration and Monitoring”

O^2 Project
Institutes



- Institutes *(contact person, people involved)*
  - FIAS, Frankfurt, Germany *(V. Lindenstruth, 8 people)*
  - GSI, Darmstadt, Germany *(M. Al-Turany and FairRoot team)*
  - IIT, Mumbay, India *(S. Dash, 6 people)*
  - IPNO, Orsay, France *(I. Hrivnacova)*
  - IRI, Frankfurt, Germany *(Udo Kebschull, 1 PhD student)*
  - Jammu University, Jammu, India *(A. Bhasin, 5 people)*
  - Rudjer Bošković Institute, Zagreb, Croatia *(M. Planicic, 1 postdoc)*
  - SUP, Sao Paulo, Brasil *(M. Gameiro Munhoz, 1 PhD)*
  - University Of Technology, Warsaw, Poland *(J. Pluta, 1 staff, 2 PhD, 3 students)*
  - Wiegner Institute, Budapest, Hungary *(G. Barnafoldi, 2 staffs, 1 PhD)*
  - CERN, Geneva, Switzerland *(P. Buncic, 7 staffs and 5 students or visitors)* *(P. Vande Vyvre, 7 staffs and 2 students)*
- Looking for more groups and people
  - Need people with computing skills and from detector groups
- Active interest from *(contact person, people involved)*
  - Creighton University, Omaha, US *(M. Cherney, 1 staff and 1 postdoc)*
  - KISTI, Daejeon, Korea
  - KMUTT (King Mongkut's University of Technology Thonburi), Bangkok, Thailand *(T. Achalakul, 1 staff and master students)*
  - KTO Karatay University, Turkey
  - Lawrence Berkeley National Lab., US *(R.J. Porter, 1 staff and 1 postdoc)*
  - LIPI, Bandung, Indonesia
  - Oak Ridge National Laboratory, US *(K. Read, 1 staff and 1 postdoc)*
  - Thammasat University, Bangkok, Thailand *(K. Chanchio)*
  - University of Cape Town, South Africa *(T. Dietel)*
  - University of Houston, US *(A. Timmins, 1 staff and 1 postdoc)*
  - University of Talca, Chile *(S. A. Guinez Molinos, 3 staffs)*
  - University of Tennessee, US *(K. Read, 1 staff and 1 postdoc)*
  - University of Texas, US *(C. Markert)*
  - Wayne State University, US *(C. Pruneau)*

Budget

Item	Cost
First Level Processing Nodes (FLP)	800 kCHF
Readout-Receiver Cards (RORC)	900 kCHF
Event Processing Nodes (EPN)	4100 kCHF
Infrastructure	1300 kCHF
Networks	800 kCHF
Servers	500 kCHF
Storage	600 kCHF
Offline	500 kCHF
Total	**9500 kCHF**

- ~80% of budget covered
- Contributions possible by cash or in-kind
- Continuous funding for GRID assumed

Future steps

- A new computing system (O^2) should be ready for the ALICE upgrade during the LHC LS2 (currently scheduled in 2018-19).
- The ALICE O^2 R&D effort has started in 2013 and is progressing well but additional people and expertise are still required in several areas:
  - VHDL code for links and computer I/O interfaces
  - Detector code benchmarking
  - Software framework development
  - Control, configuration and monitoring of the computing farm
- The project funding is not entirely covered.
- Schedule
  - June ‘15 : submission of TDR, finalize the project funding
  - ‘16 – ’17: technology choices and software development
  - June ’18 – June ’20: installation and commissioning