A Large Ion Collider Experiment

ALICE

O^2 Project : Status Report

4th ALICE ITS, MFT and O2 Asian workshop
Pusan, South Korea, 15-16 December 2014

Pierre Vande Vyvre / CERN-PH

ALICE

PNU-Supercomputing Center

www.kisti.re.kr

Outline

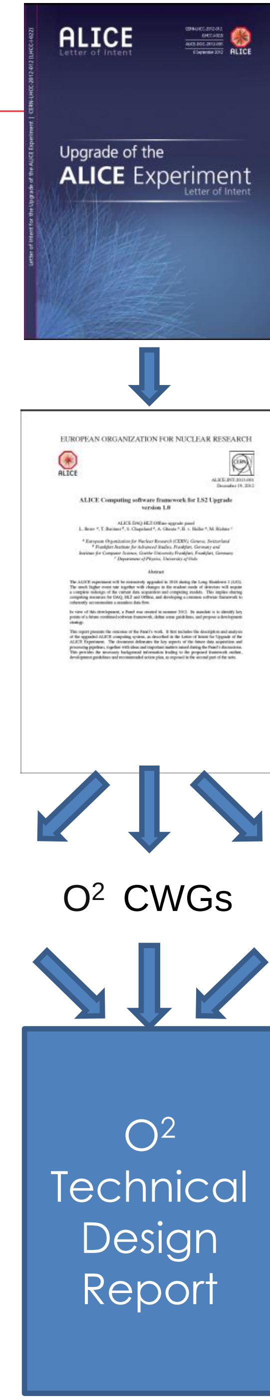
- Project status: CWGs and Institutes
- Design
- Model
- Technology watch and benchmarks
- Prototype
- Milestones, Summary, Outlook

O^2 Project

Project Organization

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

	Computing Working Group(CWG)		**Chair**
1.	Architecture		S. Chapeland
2.	Tools & Procedures		A. Telesca
3.	Dataflow	T. Breitner →	I. Legrand
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 →	D. Berzano
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

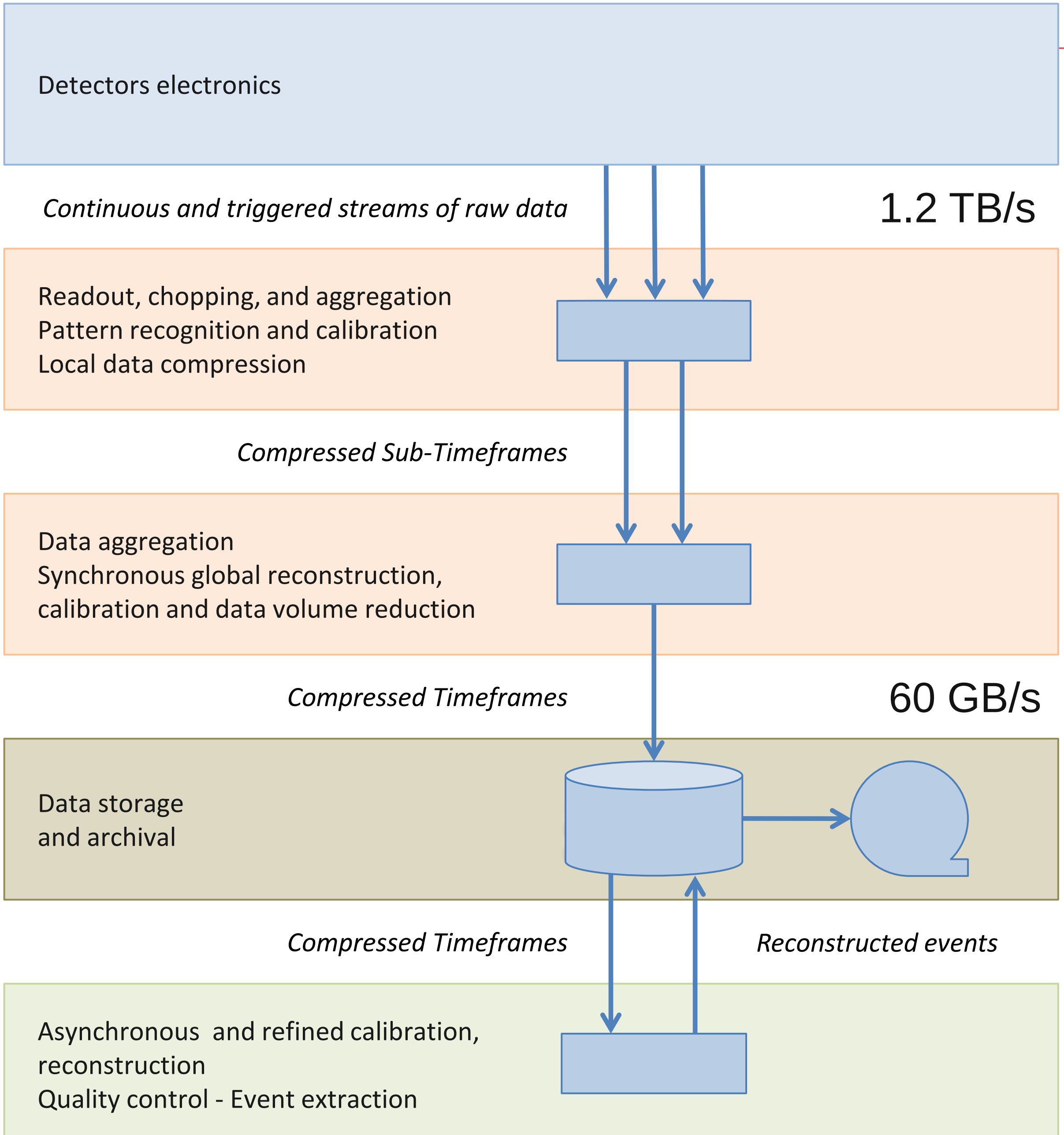
O^2 Project
Institutes

Table 9.2: Institutes participating in the O^2 Project. Based on the institute feedback till 05-Nov. To

	Country	City	Institute
1	Brasil	São Paulo	University of São Paulo
2	CERN	Geneva	European Organization for Nuclear Research
3	Chile	Talca(*)	University of Talca
4	Croatia	Zagreb	Institute Rudjer Boskovic
5	Croatia	Split	Technical University of Split
6	Czech Republic	Rez u Prahy	Nuclear Physics Institute, Academy of Sciences of the Czech Republic
7	France	Clermont -Ferrand	Laboratoire de Physique Corpusculaire (LPC), Clermont Universite, Universite Blaise Pascal, CNRS-IN2P3
8	France	Grenoble	Laboratoire de Physique Subatomique et de Cosmologie (LPSC), Universite Grenoble-Alpes, CNRS-IN2P3
9	France	Nantes	SUBATECH, Ecole des Mines de Nantes, Universite de Nantes, CNRS-IN2P3
10	France	Orsay	Institut de Physique Nucléaire (IPNO), Université Paris-Sud, CNRS-IN2P3
11	France	Strasbourg	Institut Pluridisciplinaire Hubert Curien
12	Germany	Darmstadt	GSI - Helmholtzzentrum fur Schwerionenforschung GmbH
13	Germany	Frankfurt	Frankfurt Institute for Advanced Studies, Johann Wolfgang Goethe-Universität
14	Germany	Frankfurt	Institut für Informatik, Johann Wolfgang Goethe-Universität Frankfurt
15	Hungary	Budapest	Wigner RCP Hungarian Academy of Sciences
16	India	Jammu	University of Jammu
17	India	Mumbai	Indian Institute of Technology
18	Indonesia	Bandung	Indonesian Institute of Sciences
19	Korea	Daejeon	Korea Institute of Science and Technology Information
20	Poland	Warsaw	Warsaw University of Technology
21	Romania	Bucharest	Institute of Space Science
22	South Africa	Cape Town	University of Cape Town
23	Thailand	Bangkok(")	King Mongkut's University of Technology Thonburi
24	Thailand	Bangkok	Thammasat University
25	Turkey	Konya	KTO Karatay University
26	United States	Berkeley, CA	Lawrence Berkelely National Laboratory
27	United States	Detroit, MI	Wayne State University
28	United States	Houston, TX	University of Houston
29	United States	Knoxville, TN	University of Tennessee
30	United States	Oak Ridge, TN	Oak Ridge National Laboratory
31	United States	Omaha, NE (*)	Creighton University
32	United States	Pasadena, CA	California Institute of Technology

Design

Functional Requirements



- Functional requirements of the O2 system
- Data fully compressed before data storage
- Reconstruction with calibrations of better quality
- Grid capacity will evolve much slower than the ALICE data volume
- Data archival of reconstructed events of the current year to keep Grid networking and data storage within ALICE quota
- Needs for local data storage higher than originally anticipated

Physics programme and data-taking scenario

- Scenario (Run 3+4) delivering the same HI integrated luminosity as the LoI scenario as approved by the LHCC:
  - Pb-Pb: 12.8 nb^{-1} at 5.5 TeV
  - p-Pb: 50 nb^{-1} at 8.8 TeV
- Scenario also detailing pp data taking due to the large impact on the O^2 requirements:
  - pp: 26 pb^{-1} at 14 TeV
  - pp: 6 pb^{-1} at 5.5 TeV
- Realistic scenario for 2020

  183 days * 0.6 * 0.3 * 0.7 * 200 kHz = 4.2E+11

  0.6: days for physics

  0.3: LHC efficiency

  0.7: ALICE efficiency

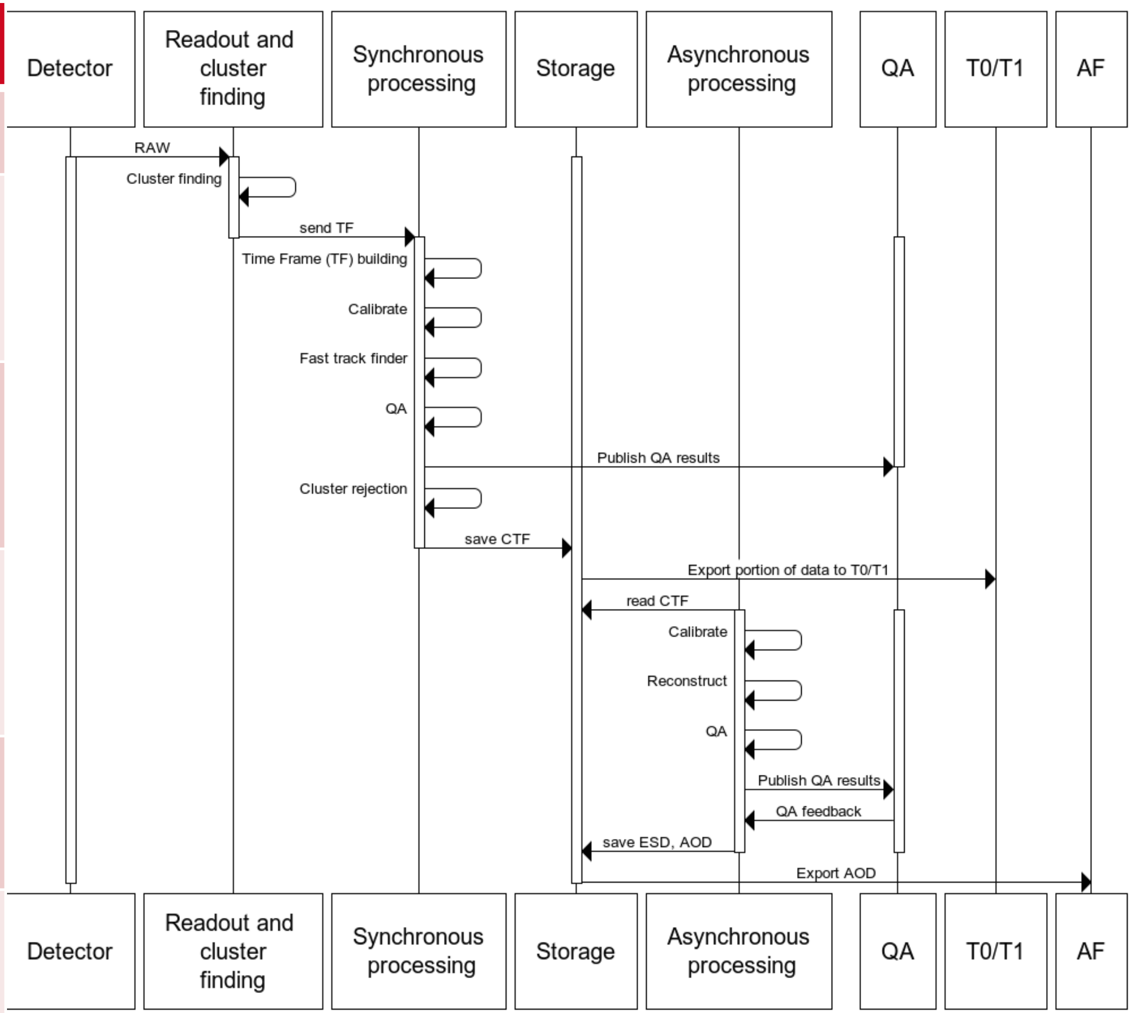
Year	System	$\sqrt{s_{NN}}$	L_{int}	$N_{collisions}$
2020	pp	14 TeV	6 pb^{-1}	$4 \cdot 10^{11}$
	Pb–Pb	5.5 TeV	2.85 nb^{-1}	$2.3 \cdot 10^{10}$
2021	pp	14 TeV	4 pb^{-1}	$2.7 \cdot 10^{11}$
	Pb–Pb	5.5 TeV	2.85 nb^{-1}	$2.3 \cdot 10^{10}$
2022	pp	14 TeV	4 pb^{-1}	$2.7 \cdot 10^{11}$
	pp	5.5 TeV	6 pb^{-1}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

- Data taking scenario defined for O^2 TDR according to the scenario approved by the LHCC
- Sufficient level of details at this stage

Computing model

Data flow

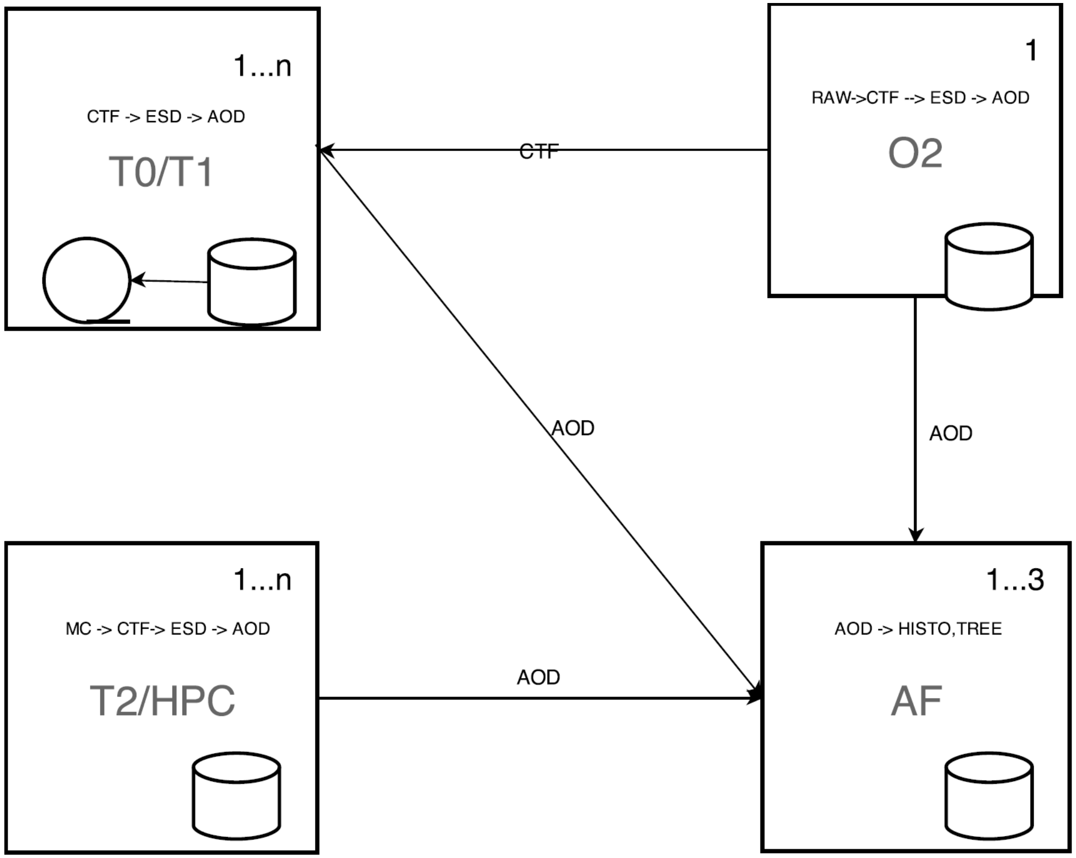
Acronym	Description	Persistency
RAW	Raw data as it comes from the detector.	Trans.
CTF	Compressed Time Frame containing the history of OM(100 ms) of detector readout information in the form of identified clusters that belong to identified tracks.	Persist.
ESD	Event Summary Data. Auxiliary data to CTF containing the output of the reconstruction process that assigns tracks to vertices and identifies the individual collisions.	Trans.
AOD	Analysis Object Data containing the final track parameters in a given vertex and for a given physics event. AODs are collected on dedicated facilities for subsequent analysis.	Persist.
HISTO	The subset of AOD information specific for a given analysis. Can be generated during analysis but needs to be offloaded from the Grid.	Temp.
MC	Simulated energy deposits in sensitive detectors. Removed once the reconstruction of MC data is completed on the Worker Node.	Trans.



Computing model

O² processing flow

Facility	Function
O2	ALICE Online-Offline Facility at LHC Point 2. During data taking: run the online reconstruction in order to achieve maximal data compression. Provides data storage capacity. After data taking: runs the calibration and reconstruction tasks.
T0	CERN Computer Center facility providing CPU, storage and archiving resources. Here reconstruction and calibration tasks are carried out on a portion of the archived CTF data, plus simulation if required.
T1	Grid site connected to T0 with high bandwidth network links (100+ Gb) providing CPU, storage and archiving resources. It runs the reconstruction and calibration tasks on its portion of archived CTF data with simulation if needed.
T2	Regular grid site with good network connectivity (10+ Gb); running simulation jobs.
AF	Dedicated Analysis Facility of HPC type that collects and stores AODs produced elsewhere and runs the organised analysis activity.



- Maintain the advantages of the Grid and the analysis trains
- Make it more open and more effective
- TDR: computing model defined

TPC requirements

Calibration

- 2 meetings in November with the TPC software team to refine the requirements
- Space-charge fluctuations
  - Dominated by event and multiplicity fluctuations
  - Must be taken into account for distortion corrections
  - Sets constraints on the update interval → 5ms
- Efficient methods for SCD calculation based on ion density
  - CPU: Lund group started with optimisations of the current code
  - GPU: Lipi group will work on this.
    More information in the presentation about plans for the TPC sw of Rifki Sadikin/LIPI

TPC requirements

Synchronous reconstruction and data compression

- Cluster finder efficiency
  - 1Dx1D very efficient implementation with a FPGA
  - 2D vs. 1Dx1D for up to 100kHz interaction rate
  - To be verified to validate the choice of computing platform
- HLT tracking performance to low p_T (<150 MeV/c)
  - Values from TPC LOI to be verified
  - Determine cluster association efficiency, fake cluster efficiency, tracking efficiency
  - Overhead due to tracklet merging at time boundaries
- Data compression factor 20 to be verified
  - Cluster format + cluster to track compression
  - Removal of non physics data (low *p*T loopers, noise)
    - Loop detection
  - Maximum compression could be achieved during the synchronous phase

TPC requirements

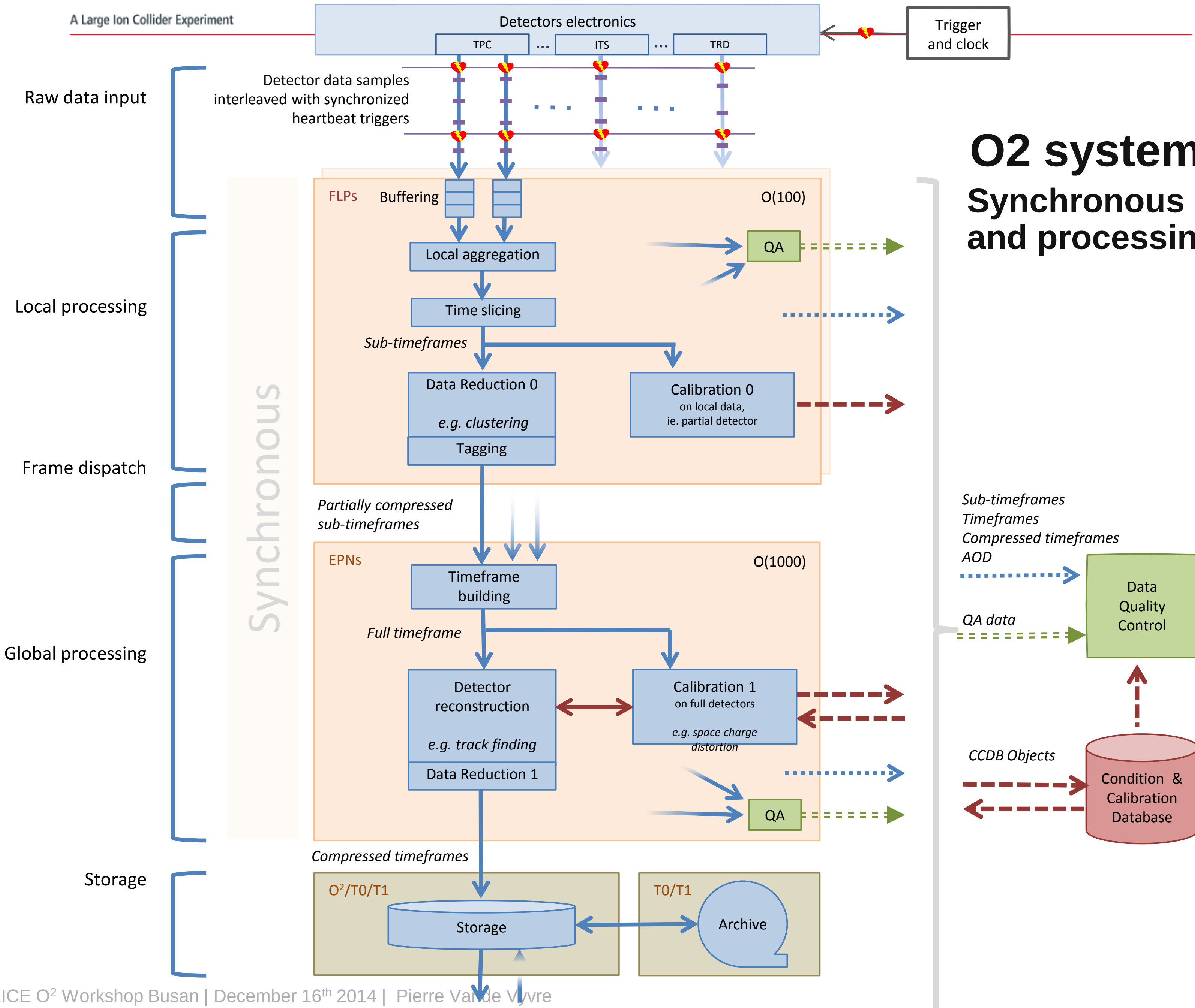
Asynchronous reconstruction

- Additional requirements needed for the asynchronous stage needed to reach physics ready data
  - dE/dx calculation, full B map, material budget –simplified geometry, ...
  - Use estimates from offline code → requires realistic speed-up of the procedures
- Precise computing needs to be estimated

- TPC requirements being refined
- Calibration requirements well defined
- Some issues related to reconstruction and data compression to be verified: compression factor of 20, 1Dx1D cluster efficiency, CPU time required

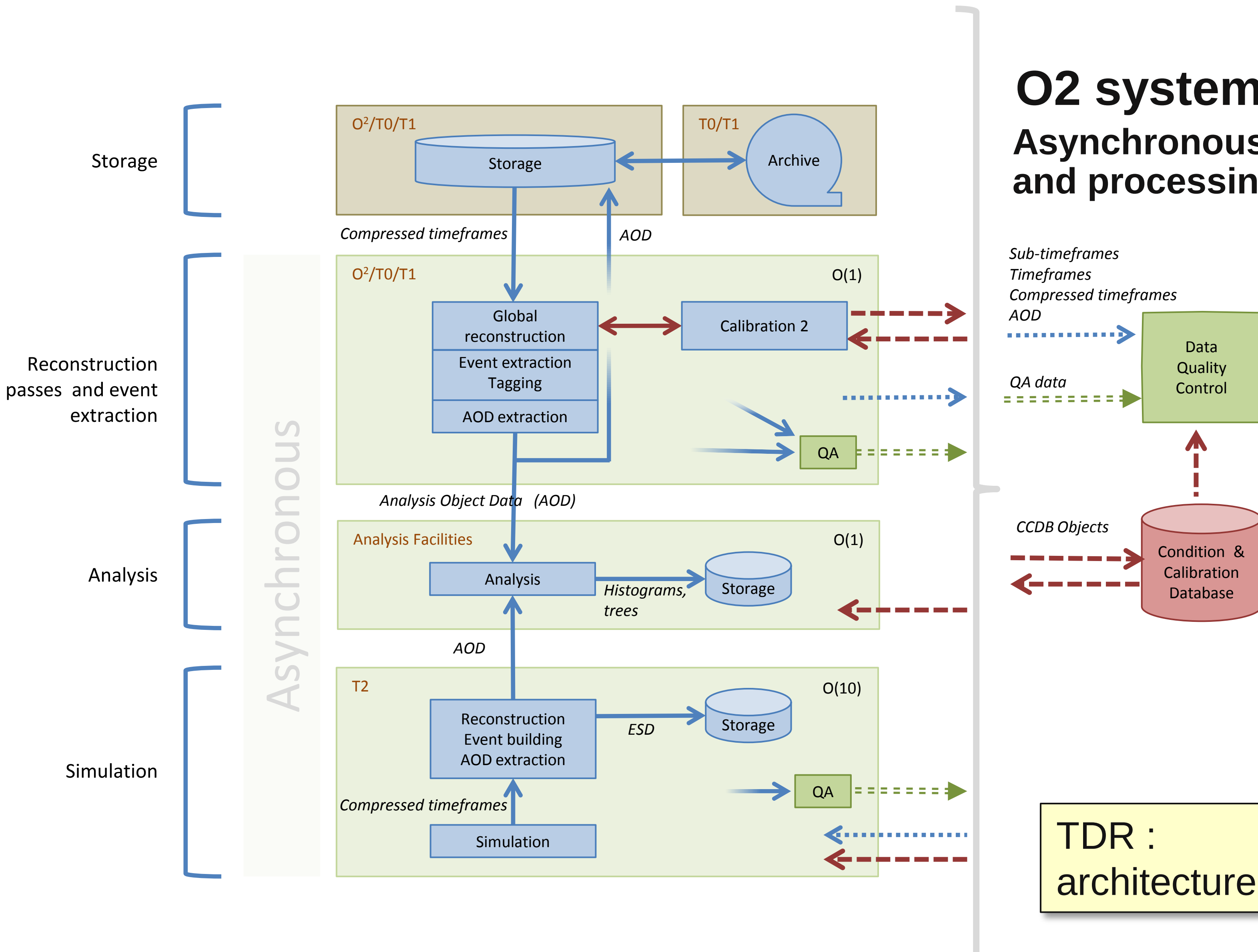
O2 system

Synchronous data flow and processing



O2 system

Asynchronous data flow and processing

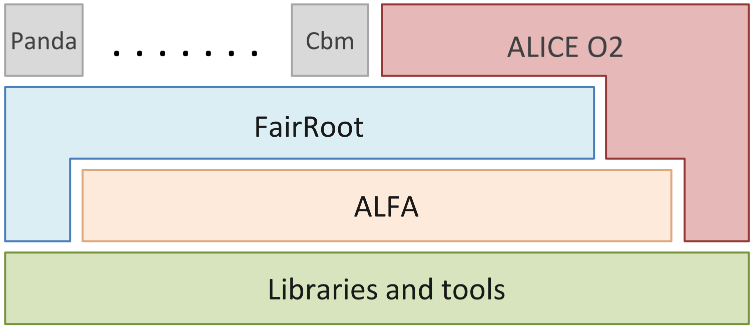


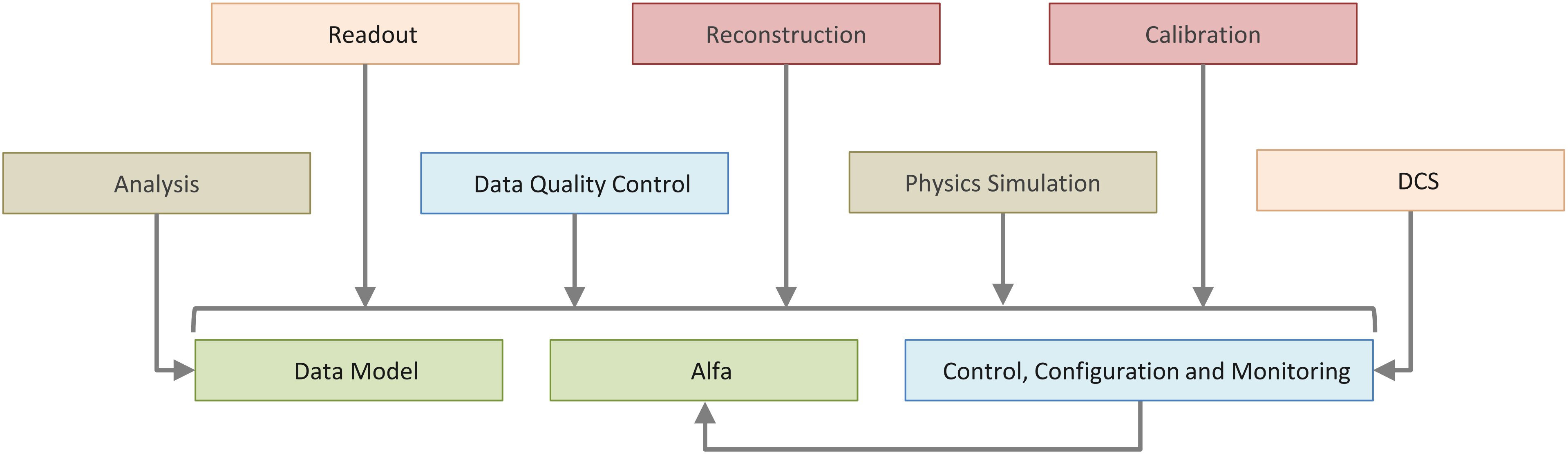
TDR : architecture defined

Software

Modularity

- Structure of ALFA and ALICE O2 software framework decided
- ALICE O2 software modules decomposition defined





Model

System modelling

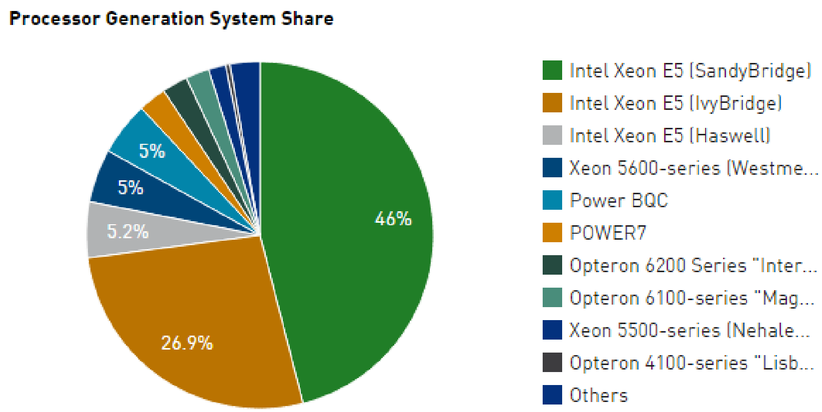
- Full system simulation → system design → hardware architecture, software design
  - A first version of the model exists
  - Used to show e.g. the system scalability shown up to 166 kHz
- Data storage simulation → data storage needs → budget evaluation
  - Data storage needs evaluated
  - To be redone with updated data taking scenarios and refined evaluations of the data sizes
- Network simulation → network layout budget → budget optimization
- More information in the following presentations:
  - Dataflow by Iosif Legrand/Caltech
  - Dataflow simulation of Rifki Sadikin/LIPI

Technology watch and benchmarks

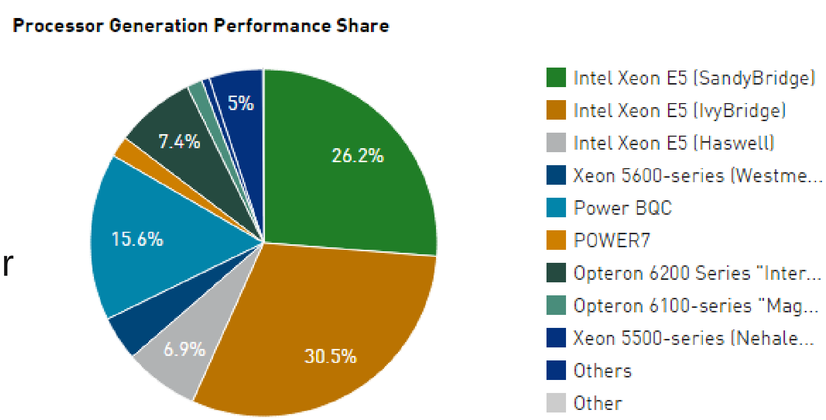
Processing power

CPUs and GPUs

TOP500



- Increasing diversity
  - Intel still the leader by far
  - AMD Unified CPU and GPU on one chip
  - IBM: Power 9 for Titan's successor at ORNL
- Intel
  - "Tick-Tock“: model adopted by Intel Corp. from 2007 to change only one major chip characteristics at each generation: either the process or the microarchitecture
  - “Tick“: shrinking of the process technology of the previous microarchitecture.
  - “Tock“: new microarchitecture
  - Every 12 to 18 months, expected to be one tick or tock.



Processing platforms

- Very ambitious expectation in the LoI for 2018: 50 cores/CPU.
- Evolution: core performance increase and slower increase of their number.

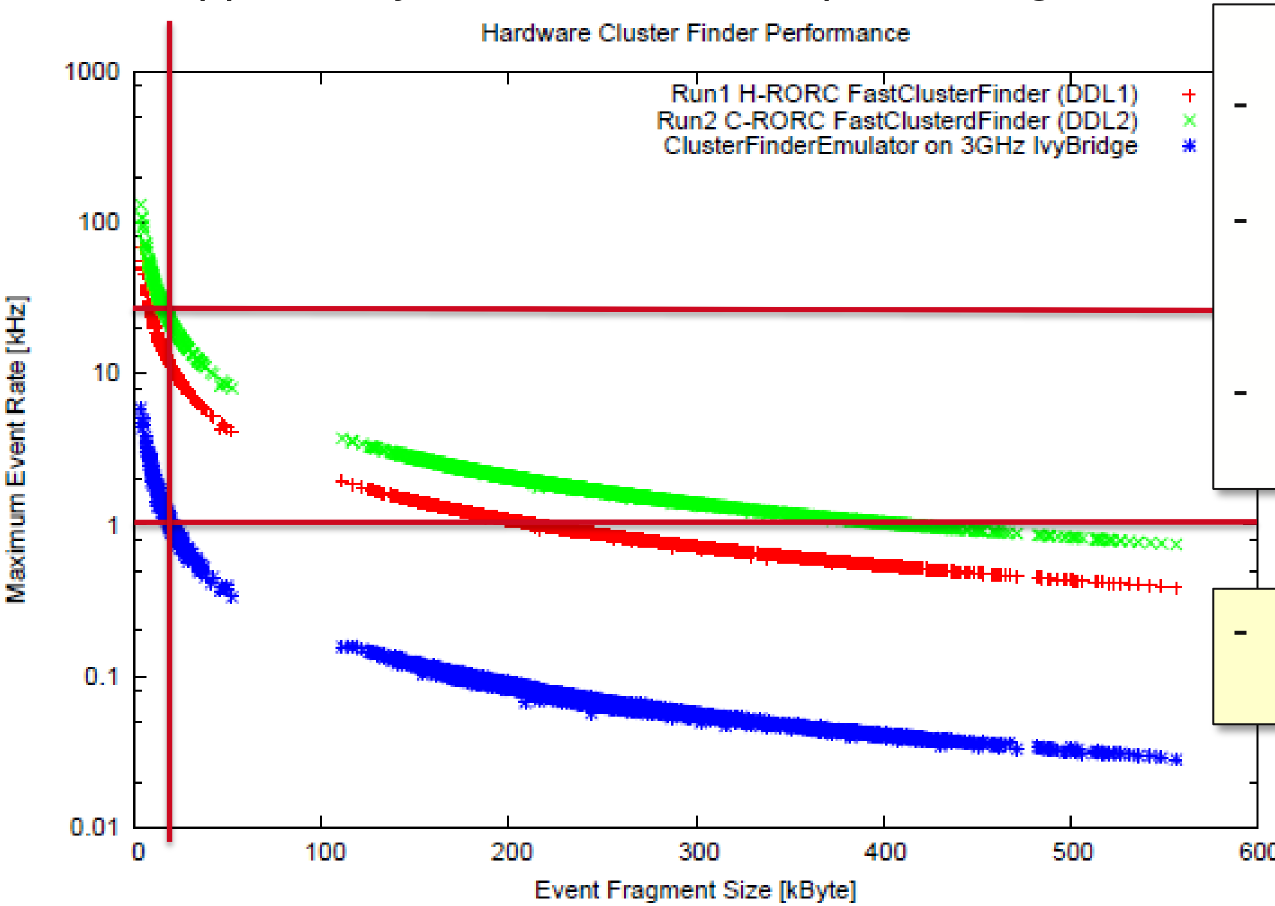
Announc.	Product Name	Code Name	Feature Size	Cores/ CPU	PCIe
2010	Intel Xeon E5 V1	Sandy Bridge	32 nm / Tock	8	Gen3
2011	Intel Xeon E5 V2	Ivy Bridge	22 nm / Tick	12	Gen3
2013	Intel Xeon E5 V3	Haswell	22 nm / Tock	14-18	Gen3
2015	Intel Xeon E5 V4	Broadwell	14 nm / Tick	18	Gen3
		Skylake	14 nm / Tock		Gen4
		Cannonlake	10 nm / Tick		Gen4
2018	LoI previsions			50	Gen3

- Less cores than anticipated per box. PCIe Gen4 available.

- More information in:
  GPU Computing platforms by Joohyung Sun / KU
  Computing Platform Benchmarking by Boonyarit Changaival / KMUTT
  Opportunistic use of CPU cycles from mobile devices Tiranee Achalakul / KMUTT

FPGA hardware accelerator for TPC cluster finder

The performance of hardware accelerators (FPGA, GPU, MIC) keeps increasing. Their applicability to the ALICE data processing is confirmed.



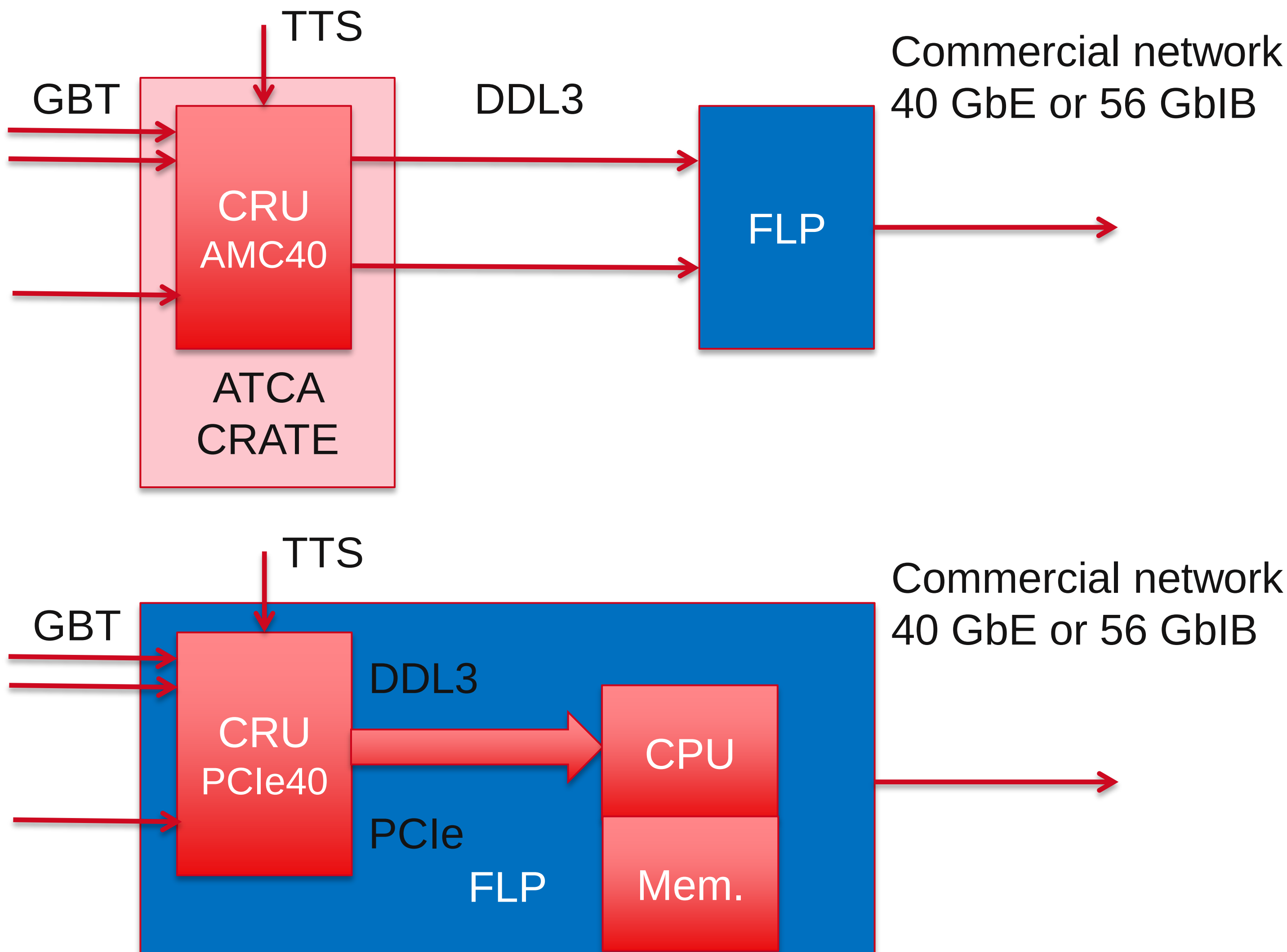
- TPC event fragment 20 MB/2000 DDL = 10 kB
- Process TPC event fragment at ~ 25 kHz with the DDL2.
- 25 x better than the existing best Xeon CPU

- FPGA baseline choice for TPC cluster finder

Plot: T. Alt and H. Engel

O2 hardware architecture

Detectors, CRU, FLPs



CRU form factor and O^2

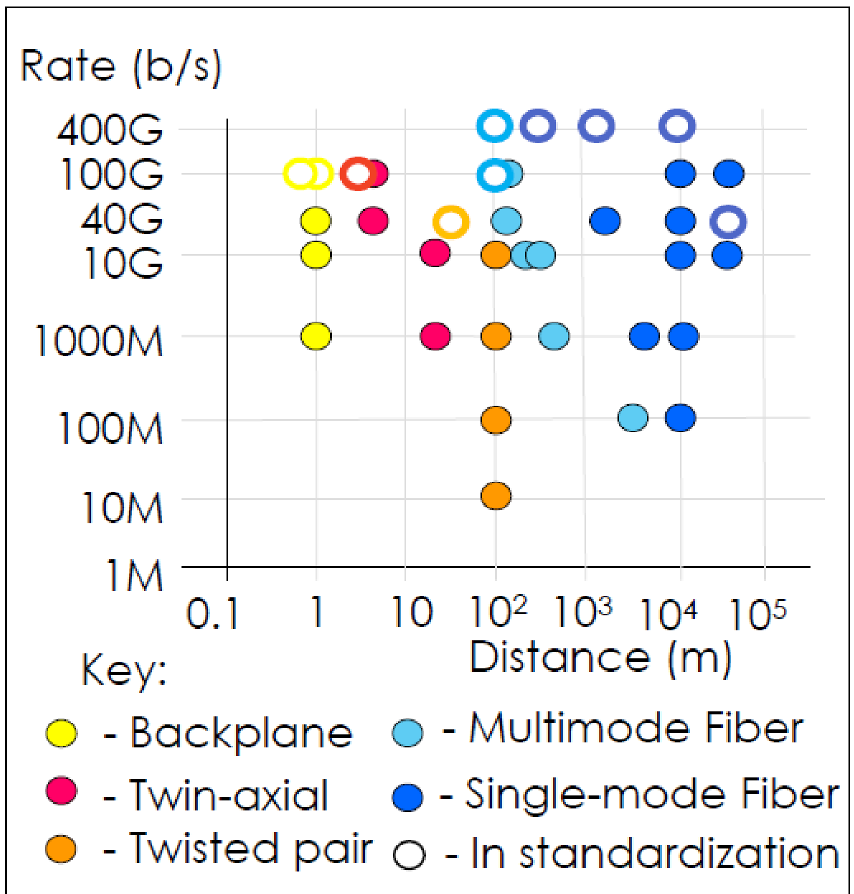
- Two options for the Common Readout Unit (CRU):
  - ATCA card (AMC40) in an ATCA crate with a commercial link (Eth or IB) to the FLP
    +: clear function separation and FLP selected independently of the CRU form factor
    - : additional step in the dataflow chain, additional complexity and cost
  - PCIe card (PCIe40) in the FLP itself
    +: simpler, cheaper
    - : constraint for the FLPs (~20% of the O2 farm): at least one PCIe Gen 3 or 4 slot.
  - No indication that the existence of a PCIe Gen3 or 4 slot could be a problem till Run4.
- CRU includes one FPGA: attractive to use it for the cluster finder as well.
  - Brand of FPGA: not considered as a restriction. Cluster finder code currently used on Xilinx but portable and originally developed for Altera.
  - Cluster finder speed: on the C-RORC processes data from the DDL2 at 4 Gb/s.
    GBT : maximum transfer rate of 5 Gb/s so the processing speed shouldn't be an issue assuming that the TPC data originated from the same padrow.
  - FPGA capacity : the current C-RORC already includes 6 occurrences of the cluster finder. The CRU would need to include at least 24 occurrences because the PCIe40 can multiplex up to 24 or 36 GBTs. To be investigated.

- Both architectures have advantages but the PCIe version is significantly cheaper.
- Dedicated meeting on 13 January with Electronics coordination, detectors and O^2.

Network technologies: Ethernet

- Ethernet: 40 GbE now and probably 100 GbE by 2015

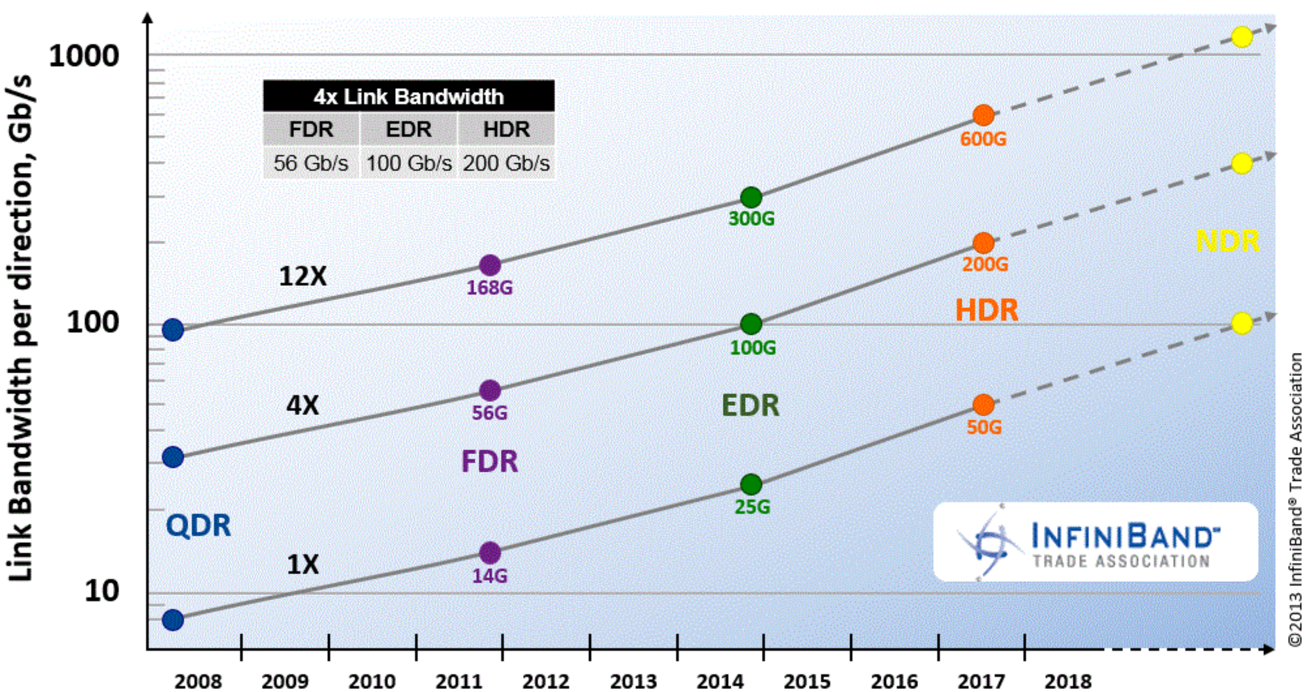
  100GBASE-SR4 over OM3, OM 4 fibers (70/100 m)



ethernet alliance

Network technologies: InfiniBand

- InfiniBand: 56 GbIB now and probably 100 GbIB (EDR) by 2015



Network technologies: Omniscale

- New network technology announced by Intel in June ‘14
  - Intel® Omni Scale Fabric– an end-to-end interconnect optimized for fast data transfers, reduced latencies and higher efficiency – initially available as discreet components in 2015, will also be integrated into next-generation Intel Xeon Phi processor (Knights Landing) and future 14nm Intel® Xeon® processors.
  - Adapter integrated in the CPU chip
  - 100 Gb fabric announced for 2015
  - Probably an impact on the form factors of some systems used by online systems
  - Will be monitored and tested to see if cost effective for O^2
  - Could affect the long-term viability of IB

- TDR
  - Budget according to the most cost-effective available solution (Eth or IB)
  - Keep the two other solutions for the network technologies as possible alternatives. Make the choice later (see milestones).

Prototype

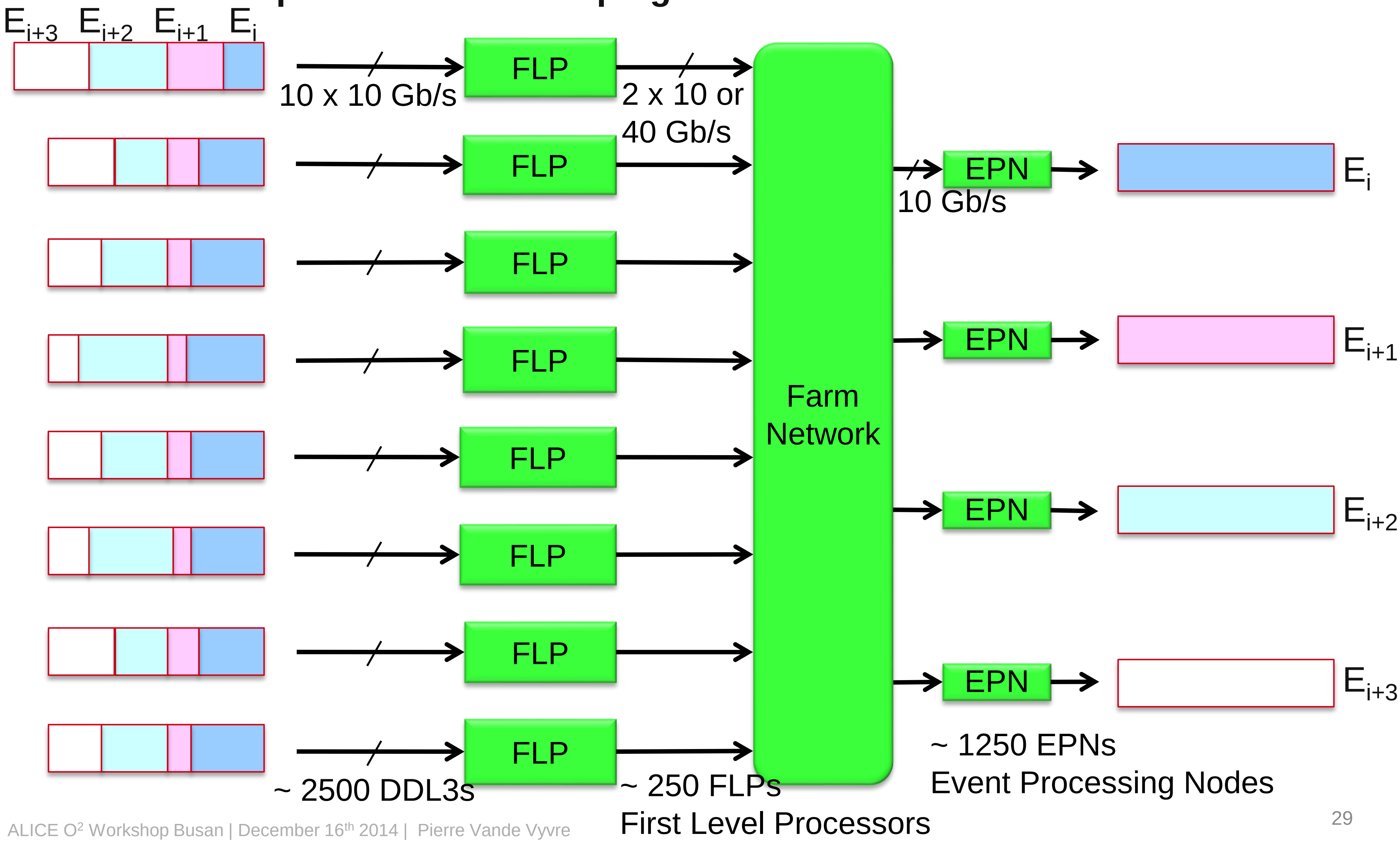
Prototype

Goals

- Assemble a first system with all the existing components
- Involve all the groups (CWGs and detectors)
- Use in production the tools selected and follow the procedures put in place and verify them.
  See presentations of Vasco Barroso / CERN and Rifki Sadikin/LIPI.
- Compare options, measure performances and validate choices made for the software
- Use some hardware with realistic applications
- Currently done on two setups at GSI and CERN
- The request for a new lab at CERN for the O2 project has been accepted: 20 racks in the basement of bld 4 with adequate power and cooling. Available in 2016

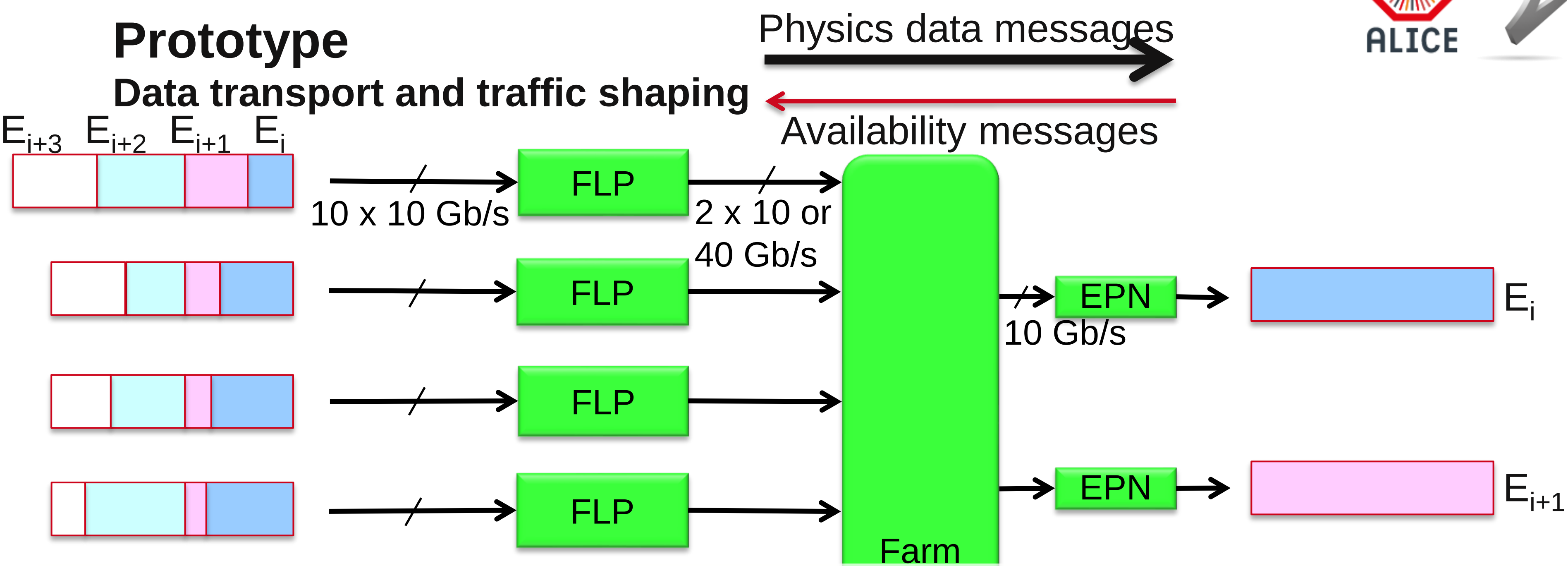
Prototype

Data transport and traffic shaping



Prototype

Data transport and traffic shaping



- FairMQ : proposed transport and load balancing solution
- Underlying protocol with ZeroMQ, nanomsg or other solution
- Traffic shaping: avoid EPN contention, graceful degradation, load balancing
- FLP: decision on EPN used as destination based on the timeframe ID included in the data and the EPN availability messages

  e.g. **EPN# = TimeframeID % numOfAvailableEPNs** or

  EPN# = getEpnIdFromTfId(TimeframeID)
- Zookeeper-based library getEpnIdFromTfId

Prototype

Data and Algorithms

- FairMQ as basis for the framework
- TPC and O^2 working on assembling realistic input data
  - Use current HLT clusters to inject data in the FLPs
- Interfacing the existing HLT online reconstruction modules in AliRoot as a baseline for the O2 prototype development

Prototype

Control, Configuration and Monitoring

- Control, Configuration and monitoring

  See presentations of Vasco Barroso and Khanasin YAMNUAL

- Monitoring with MonaLisa
  - Used by the ALICE offline for the Grid. Also used by the DAQ for Run 2.
  - Reliable, fast and open.
  - Requirements for the prototype
    - # of nodes: OM (10-100) for the tests
    - Processes per machine: OM (100)
    - Parameters per process: OM (10) (including system monitoring such as CPU, memory, etc)
    - Sending frequency: OM (10 Hz) per process
    - Storage policy
      - Frequency: all metrics
      - Storage time: ideally forever, but in practice never look at statistics older than a couple of weeks

Prototype monitoring

Standard GUI

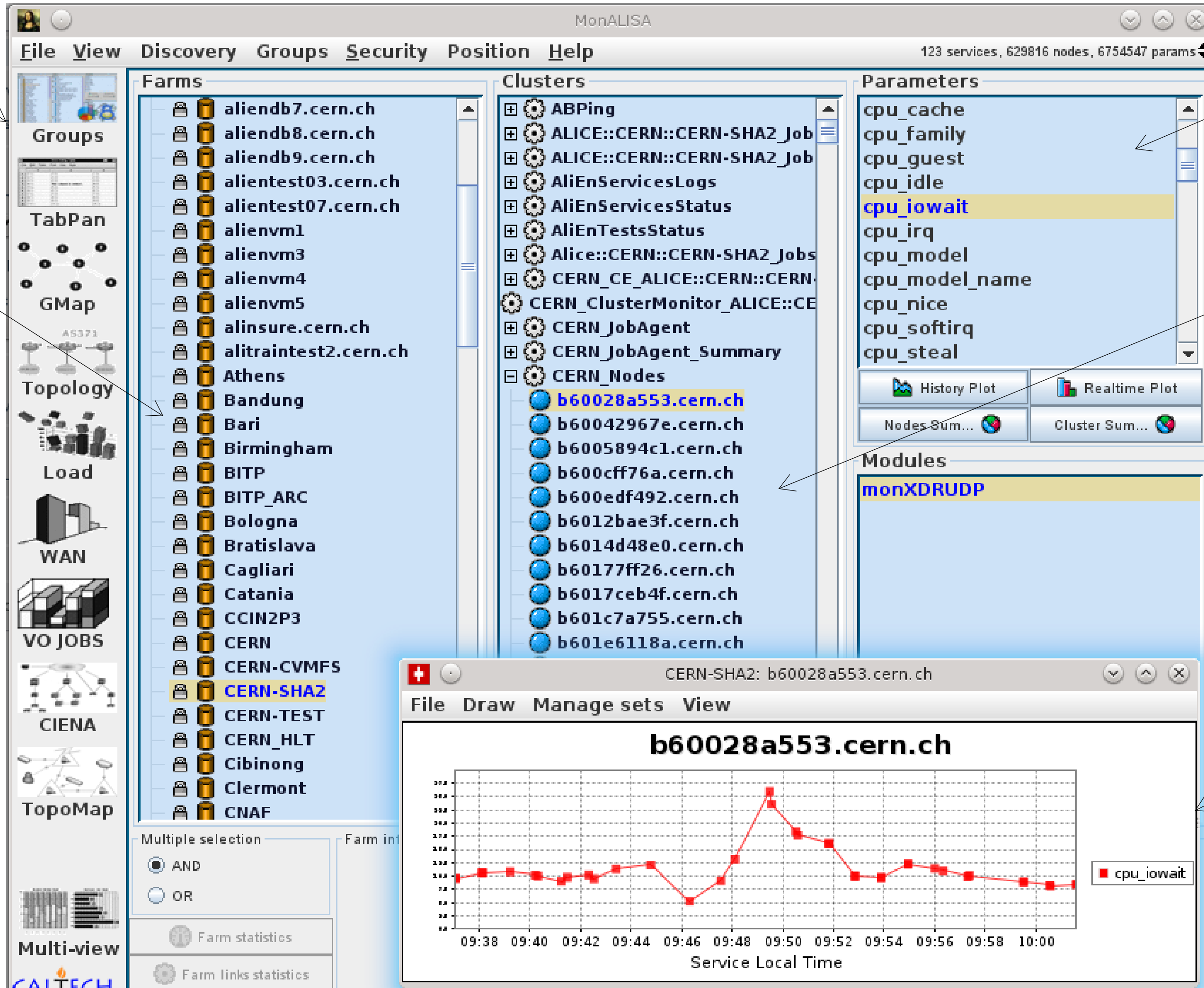
Views

Services in group

Params

Clusters and nodes

Plots



Milestones, Summary, Outlook

O² Technical Design Report

Schedule

- Apr ’14: Draft 0 of the text for review by CWGs
- 5- 7 May ’14 Draft 1: review by O² EC (Editorial Committee)
- 4 July ‘14: Draft 2
- 24 - 26 Sep ’14 Draft 3 review by EC
- 24 Oct ’14: Apply all fixes and general coherence decided by EC
- 28 Nov ‘14: Check by the EC and start of proof-reading
- 12 Dec ‘12: Pre-Draft 4 tag
- 9 Jan ‘15: Draft 4 tag
- Jan ‘15: End of proof-reading for Draft 4
- 19 – 21 Jan Draft 4 review by EC
- 16 Feb - 1 Mar ‘15: ALICE internal review
- Apr ’15: Submission TDR to LHCC (1 month before the meeting)
- Jun ’15: LHCC meeting (3-4 June 2015)

O2 milestones

	Q1	Q2	Q3	Q4
2014	- O^2 dataflow simulation program	- TDR draft 1 - Alfa framework definition	- TDR drafts 2, 3	- TBD: data taking scenarios - TDR draft 4
2015	- Decision CRU form factor - TDR → ALICE - UCG draft 1	- TDR, UCG → LHCC - O^2 lab: racks + PDU order	- O^2 lab: racks installation, hw purchase decision	- O^2 sw: cont. detector readout - O^2 lab cooling comm., hw order
2016	- TPC fully equipped IROC prototype test	- O^2 lab: start hw installation	- ITS elements commissioning (TBD)	- O^2 lab: ≈5-10% O^2 facility
2017	- O^2 lab: soft commissioning	- O^2 lab: large scale tests	- O^2 techno: infrastructure and FLP selection	- O^2 sw: full dataflow
2018	- O^2 system for ITS commissioning - O^2 techno: EPN, storage and network selection - O^2 facility: CR2 racks + PDU order	- ITS commissioning at surface	- O^2 facility: FLP ITS EPN temp - CR2 renovation	- ITS commissioning in ALICE - MFT commissioning at surface (TBD) - O^2 facility: delivery FLP, EPN, network, storage
2019	- O^2 facility: FLP 100% EPN 20 % Storage 20%	- MFT commissioning in ALICE (TBD)	- TPC commissioning on surface - O^2 facility: delivery EPN, network, storage	- TPC commissioning in ALICE
2020	- O^2 facility: EPN 100 % Storage 100 % - ALICE commissioning	- ALICE pp run	- ALICE pp run	- ALICE PbPb run

Summary and Outlook

- More institutes have joined the project, in particular from Asia.

  They have a significant impact.
- Design
  - Physics requirements complete
  - Decision on CRU form factor needed. Probably January/February.
  - Requirements of TPC being refined
    (1Dx1D/2D cluster finder, compression factor, processing times)
  - Computing model and architecture defined
- Model
  - A first version of the model exists. It will be refined and used.
- Technologies
  - Processing platform:
    CPU's in 2018 possibly less performant than anticipated at the LoI time.
    FPGA-based hardware accelerator confirmed for TPC cluster finder.
  - Network: several technologies available would allow to build the system.
    New technology (Intel Omniscale) might be relevant for O2. Keep the choice open.
- Intensive work on the O2 prototype:
  - Code development and integration in progress.
  - O2 lab should be ready in 2016 for larger tests.
- TDR progressing
  - Draft for reviewing inside ALICE in February/March '15.