The ALICE computing upgrade project and network in Asia

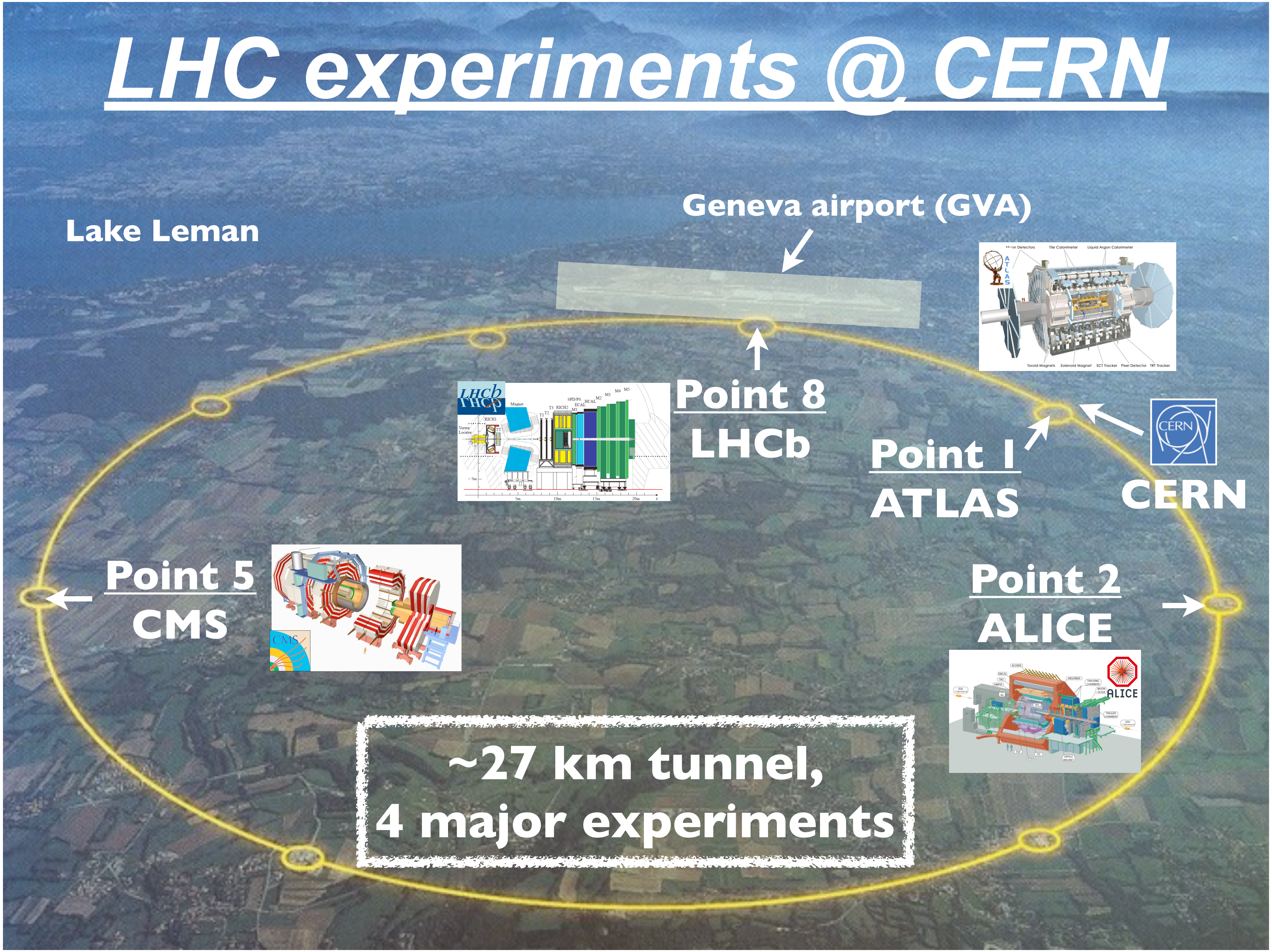
Tatsuya Chujo
(for the ALICE collaboration)

AFAD 2015
6th Asian Forum for Accelerators and Detector
January 26, 2015, NSRRC, Taiwan

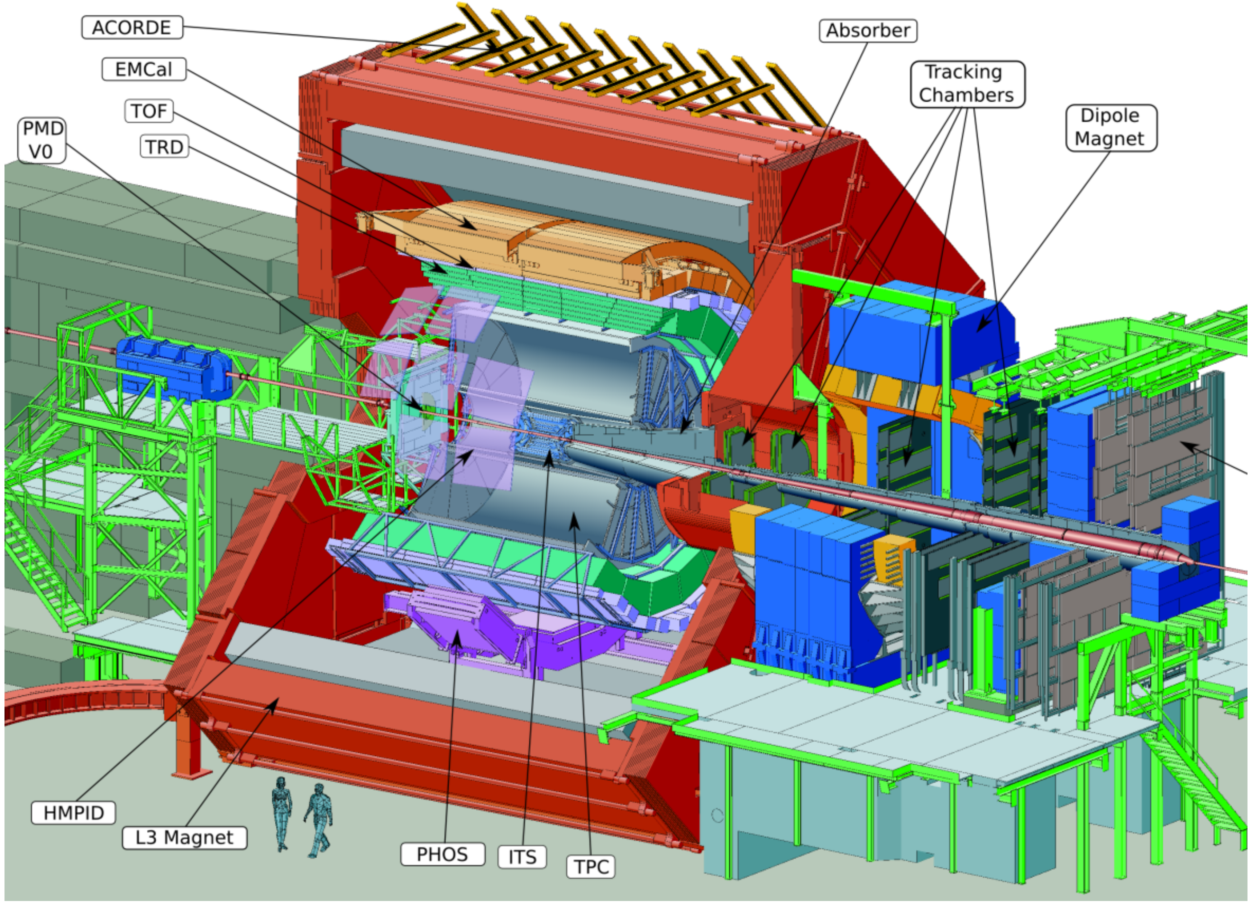
筑波大学
University of Tsukuba

ALICE

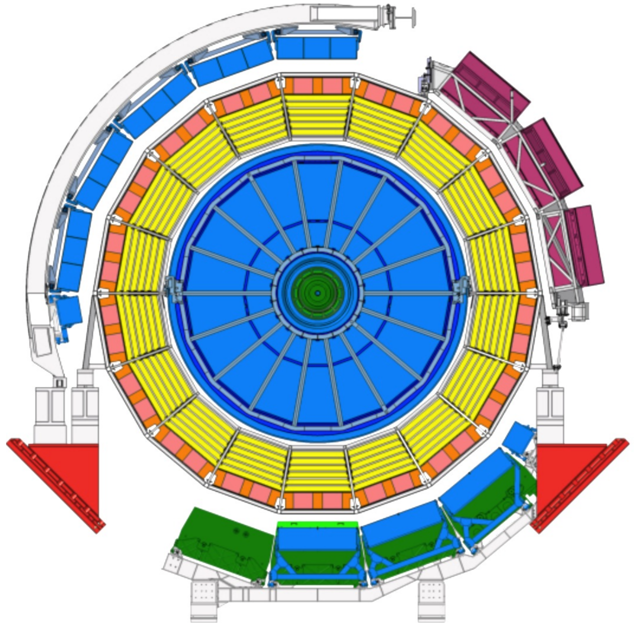
LHC experiments @ CERN



ALICE Experiment



16m x 16m x 26m, 10,000 tons

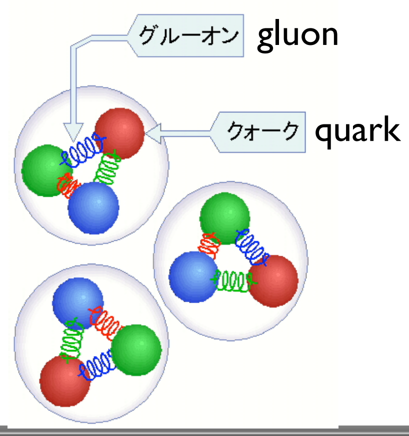


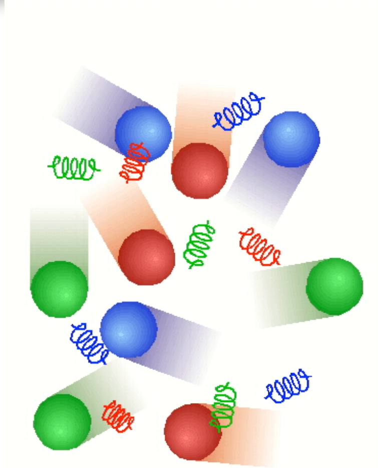
>1400 scientists from 149 Institutes in 40 Countries

The dedicated experiment in LHC experiments to study Quark Gluon Plasma (QGP) by using heavy ion beams.

18 different sub-detectors:
tracking, particle identification, energy measurement, event trigger

What is Quark Gluon Plasma (QGP)?



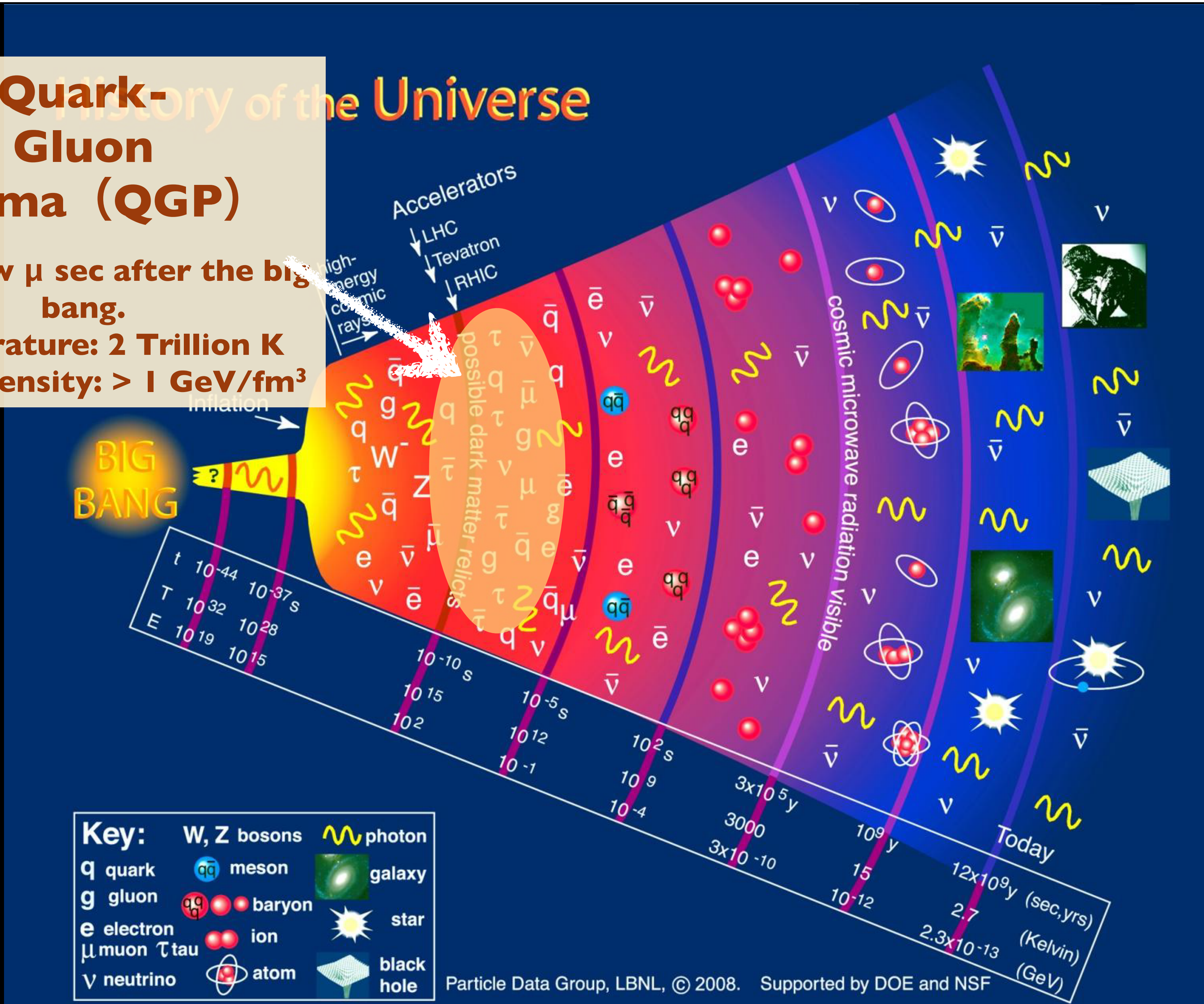


$$\mathcal{L} = -\frac{1}{4} G^a_{\mu\nu} G^{\mu\nu}_a + \bar{q}\gamma^\mu (i\partial_\mu - g t^a A^a_\mu) q - m\bar{q}q$$

- De-confined state of quarks and gluon inside hadrons under the extremely high temperature and energy density
    - New and still unknown state of matter.
- Lattice QCD calculations:
    - Critical temperature: T_c = 150-200 MeV
    - Crossover phase transition from hadronic phase to parton phase.

Quark-Gluon Plasma（QGP）

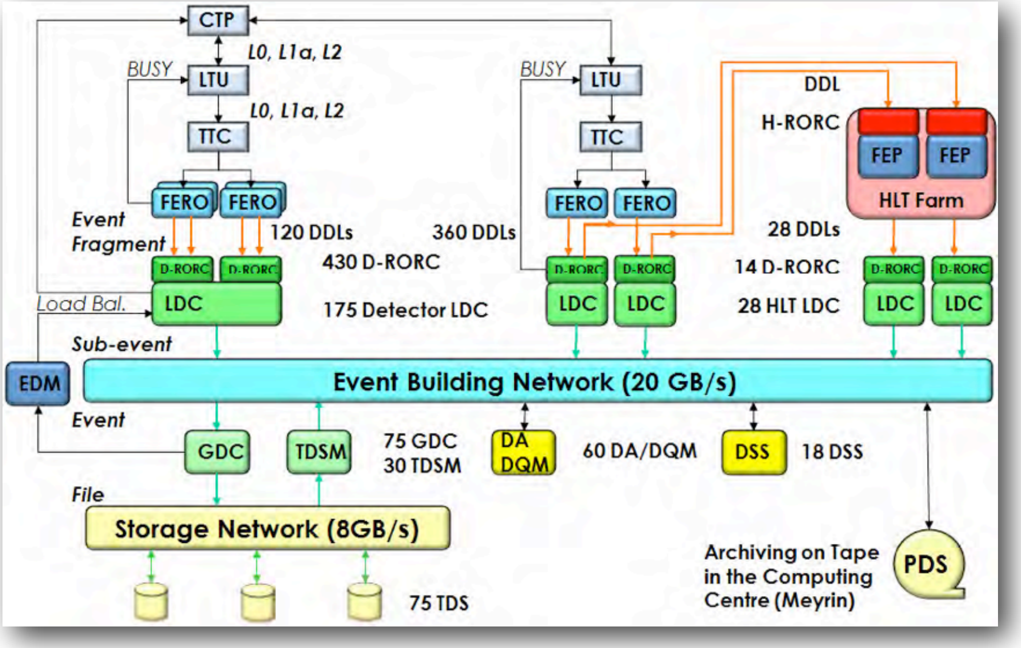
Time: few μ sec after the big bang.
Temperature: 2 Trillion K
Energy density: > 1 GeV/fm³

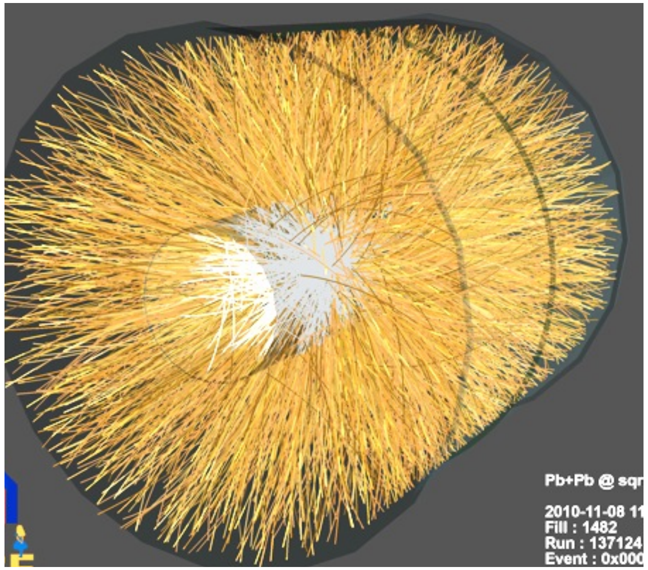


ALICE data collection (current)

Nominal LHC beam crossing at 40 MHz

ALICE:
Multi-level trigger system needed:
40 MHz → a few kHz



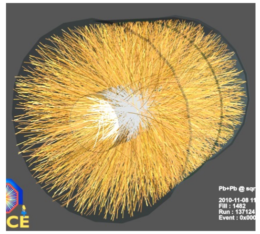
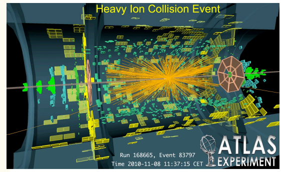


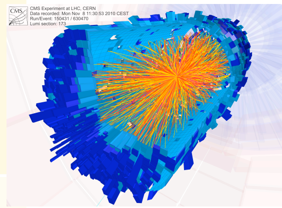
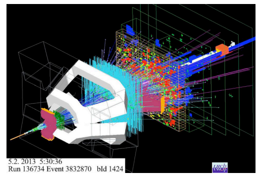
Single Pb-Pb collision events
($\sqrt{s_{NN}}$ = 2.76 TeV)

Online:
1) Reject background
2) Select most interesting interactions
3) Custom computer to reduce the total data volume

Data rate and data volume

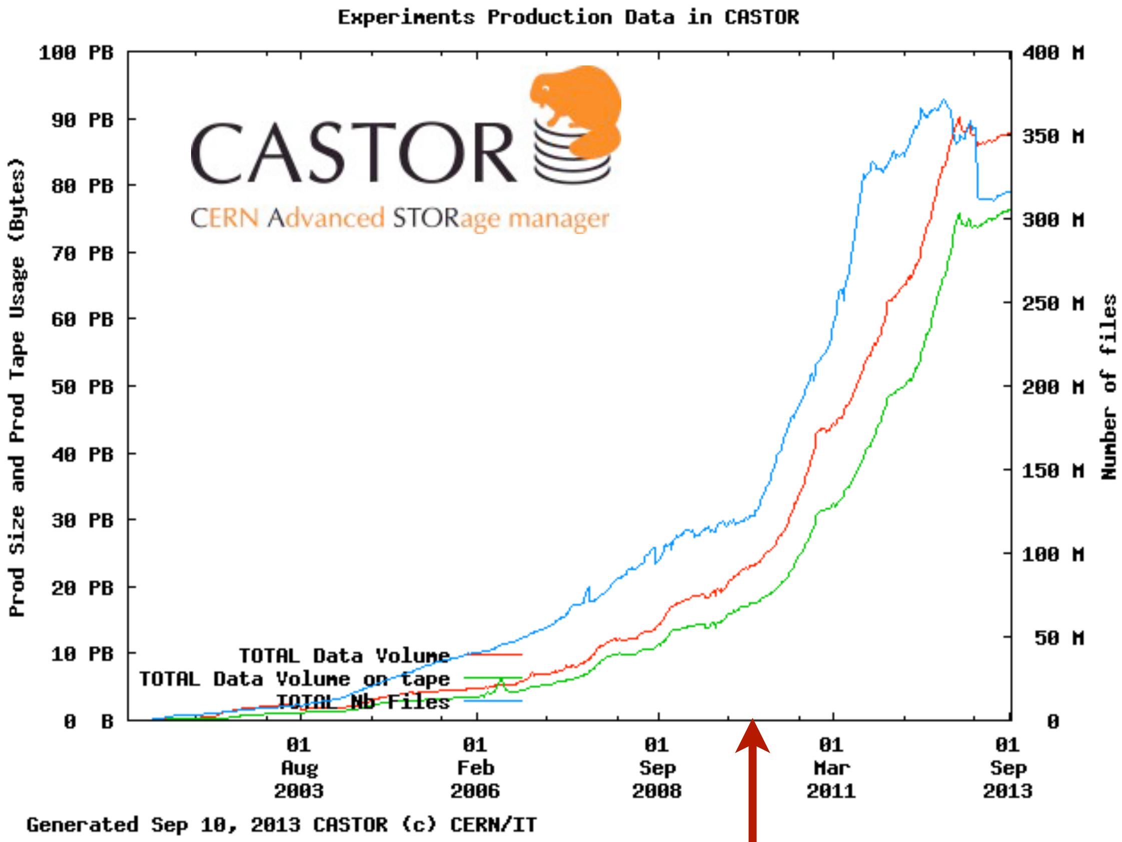
	Beam type	Recording (Events/s)	Recording (MB/s)	Data archived (PB per year)
ALICE	Pb-Pb (one month)	200	1250	2.3
ATLAS	p-p (~8 months)	100	100	6.0
CMS	p-p (~8 months)	100	100	3.0
LHCb	p-p (~8 months)	200	40	1.0

- Pb-Pb events are much heavier than pp ones.
- ONLINE selection + OFFLINE processing →Final data volume archived.
- Each experiment has its own online-offline strategies.

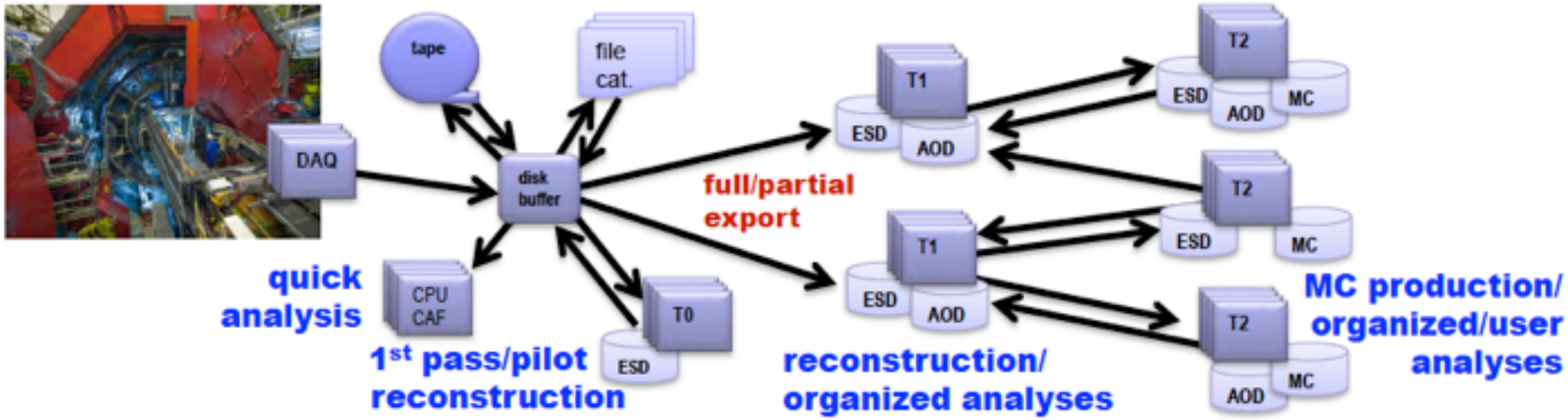
Data archiving



LHC start taking data

- Experiments production data in CASTOR: LHC contribution is the main significant one (data taking starts at the end of 2009)
- LHC contributed with ~12 PB/year between 2010 and 2013
- Nearly 80 PB of experiments data currently stored in CASTOR's tapes

Computing model (Run1 and Run2, -2018)

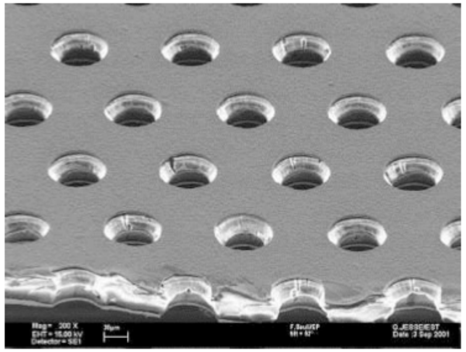


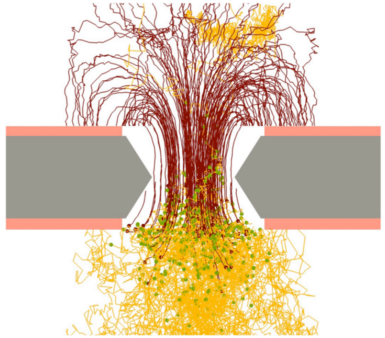
slide by T. Sugitate

ALICE upgrade (2018-)

<u>ALICE upgrade; high rate capability</u>

✓GEM-TPC continuos high rate readout
✓ ITS Silicon high rate readout
✓DAQ (RCU etc.)

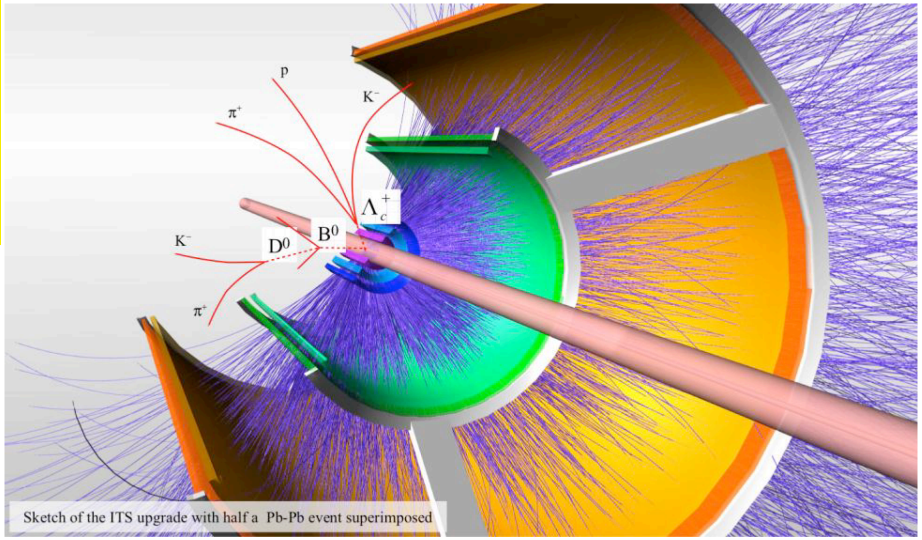




Standard GEM
Pitch=140μm
Hole ϕ=70μm

For LHC high luminosity upgrade, Pb-Pb @50kHz
Record all MB events, x100 statistics
（Unique capability in ALICE）

→**Access to high precision measurements and rare probes**



Sketch of the ITS upgrade with half a Pb-Pb event superimposed

Physics Goals:
Measure
- heavy quarks, photons, lepton pairs
azimuthal anisotropy
- Jet w/ PID hadron simultaneously

ALICE Upgrade Plans: (2018-)

- **Current: reducing the event rate from 40 MHz to ~ 1 kHz**
  - Select the most interesting particle interactions
  - Reduce the data volume to a manageable size
- **After 2018:**
  - Much more data (**X 100**) because:
    - Higher interaction rate
    - More violent collisions → More particles → More data (1 TB/s)
    - Physics topics require measurements characterized by;
      - Very small signal/background ratio → large statistics
      - Large background → traditional triggering or filtering techniques very inefficient for most physics channels
    - Read out all particle interactions (PbPb) at the anticipated interaction rate of **50 kHz**
    - **No more data selection**
      - Continuous detector read-out
      - Read-out and process all interactions with a standard computer farm.
      - ~1,500 nodes with the computing power expected by then

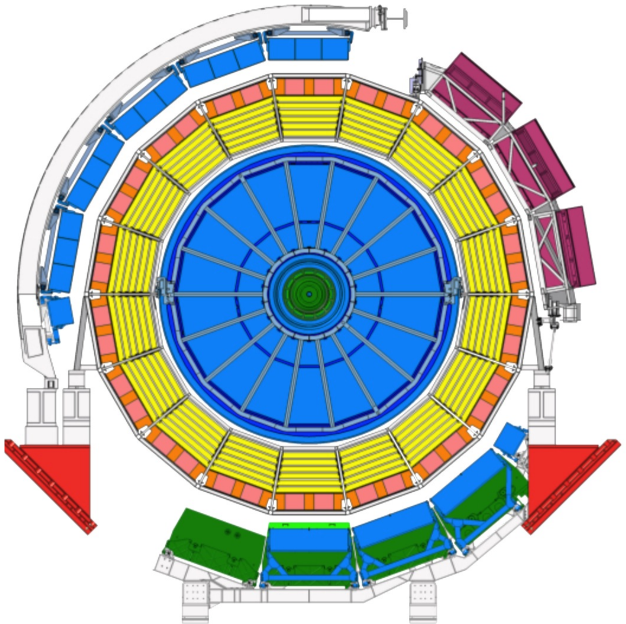
➡ **Total data throughput out of the detectors: 1 TB/s**

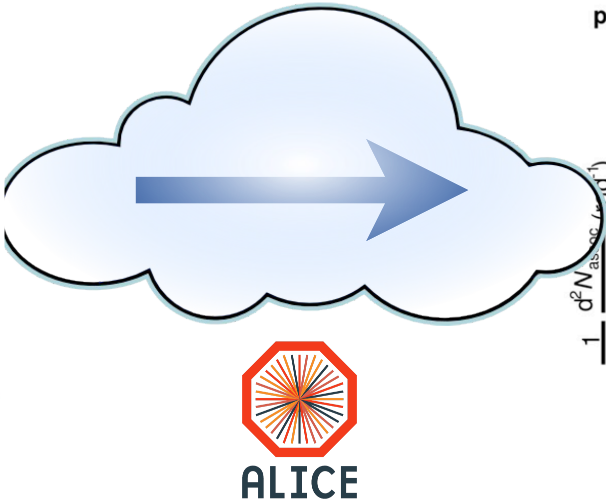
Expected data bandwidth (after 2018-)

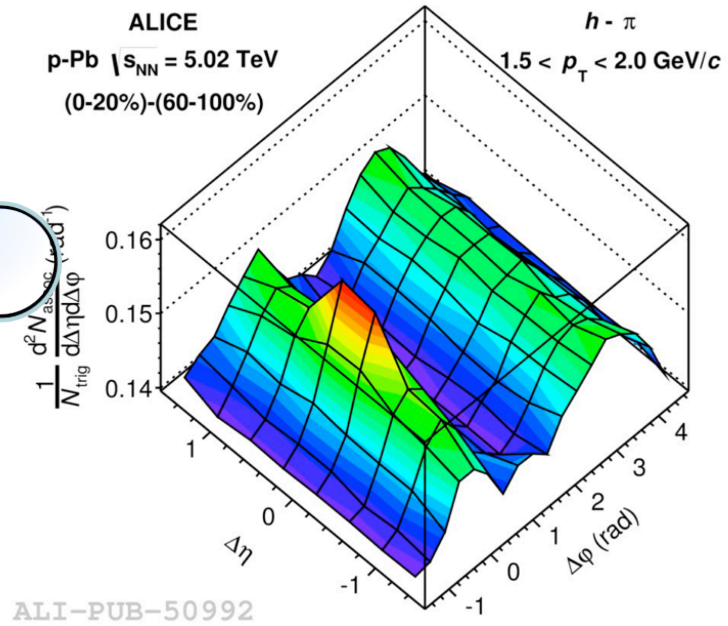
Detector	Input to Online System (GB/s)	Peak Output to Local Data Storage (GB/s)	Average Output to Computing Center (GB/s)
TPC	1,000	50.0	8.0
TRD	81.5	10.0	1.6
ITS	40	10.0	1.6
Others	25	12.5	2.0
TOTAL	**1,146.5**	**82.5**	**13.2**

Note: LHC luminosity variation during fill and efficiency taken into account for average output to computing center

The ALICE Online-Offline (O2) Project







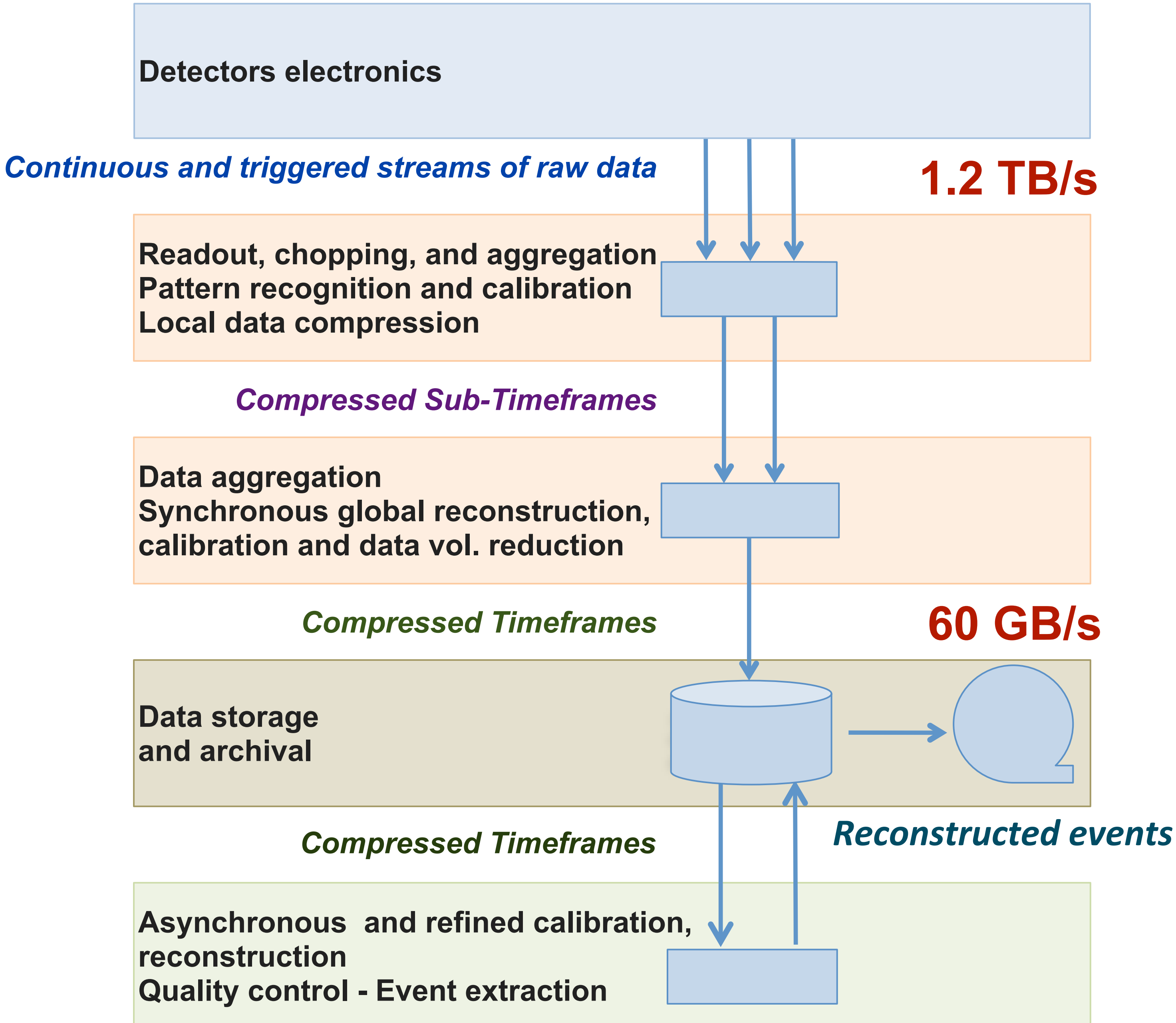
- From Detector Readout to Analysis:
- What is the "optimal" computing architecture?

- Handle **>1 T Byte /s** detector input
- Support for continuous readout
- Online reconstruction to reduce data volume
- Common hardware and software system developed by the DAQ, HLT, Offline teams

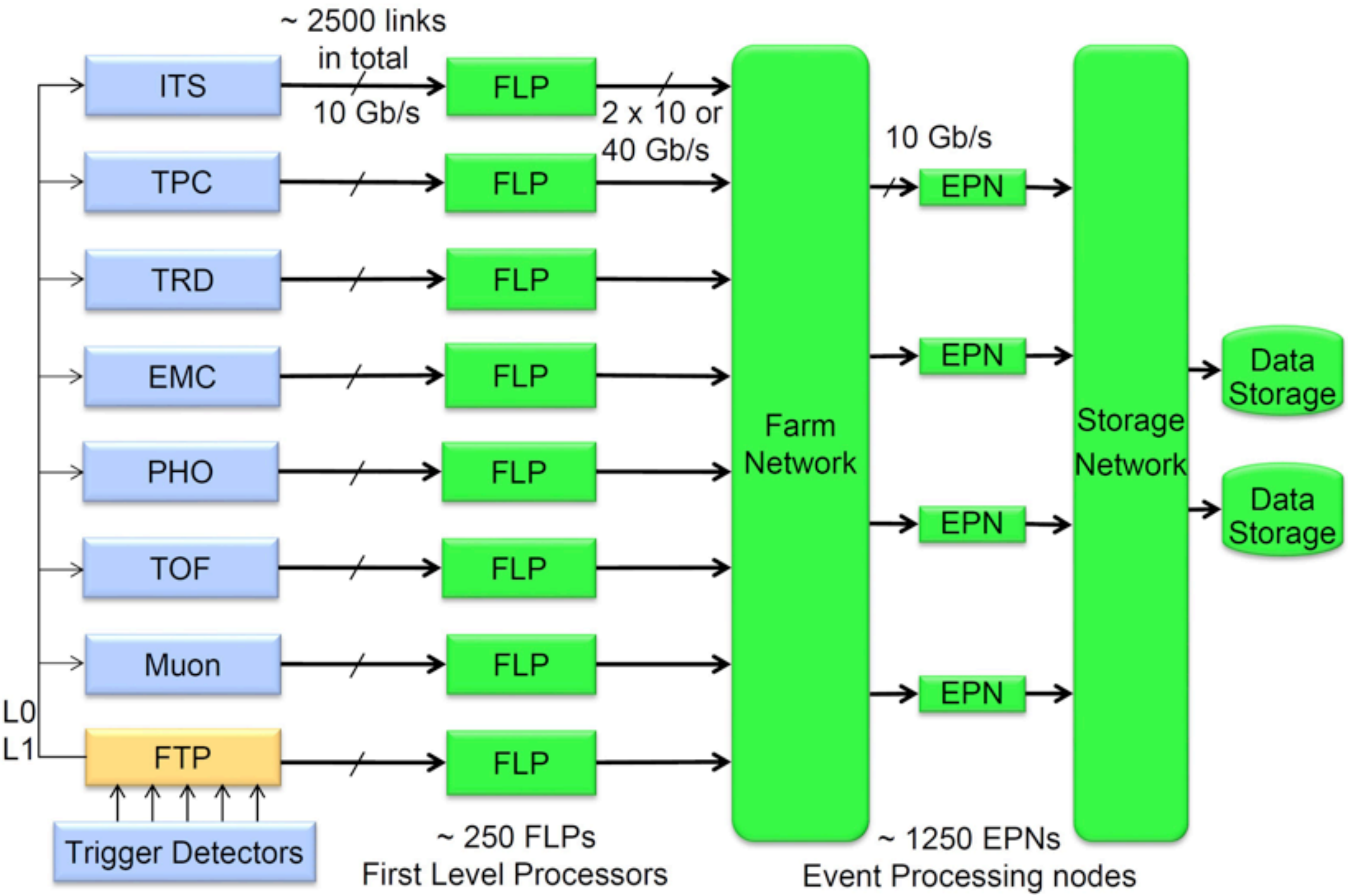
Functional Requirements of the O2 system

- ✓ 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

Basic idea of the O2 system



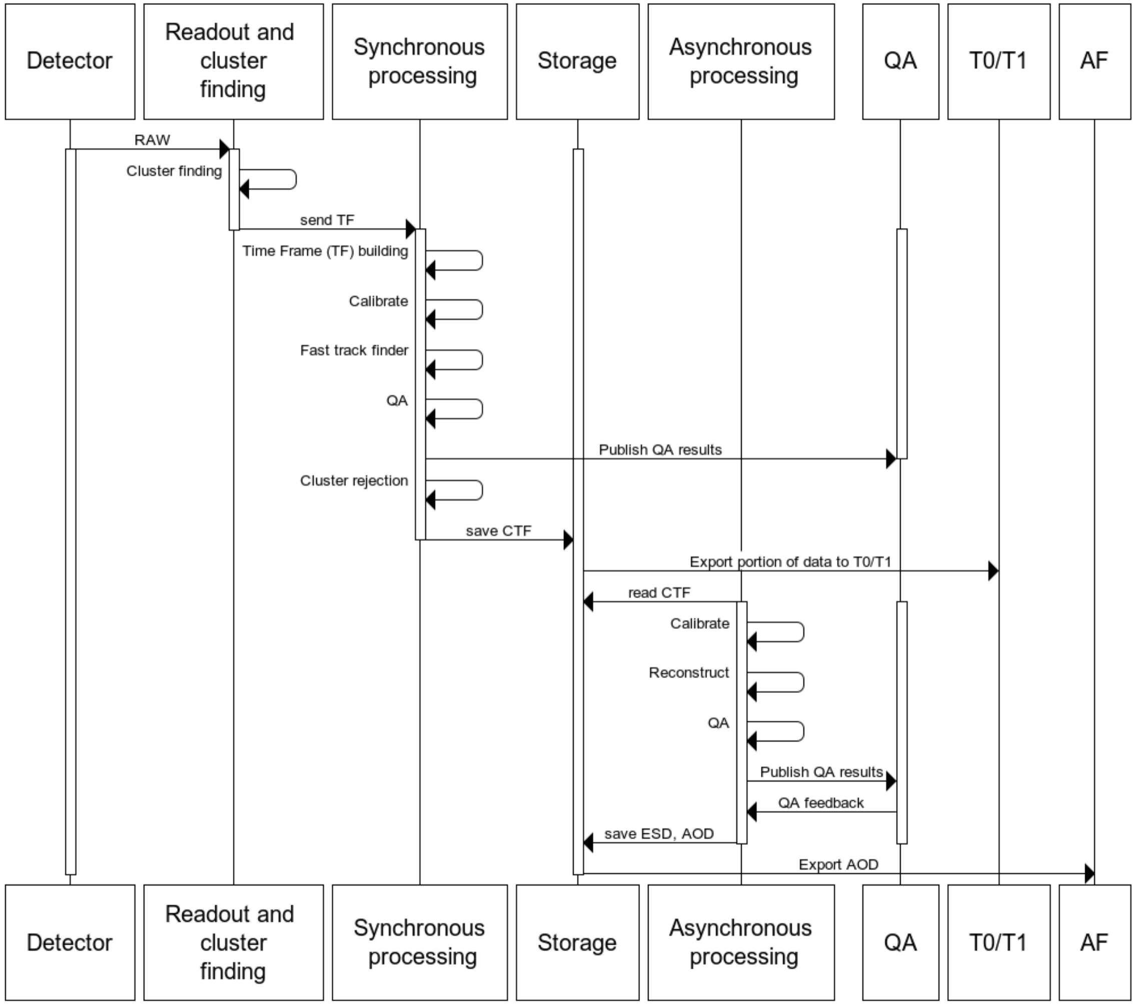
The ALICE O2 Hardware Architecture



Computing model (Data flow)

A Large Ion Collider Experiment

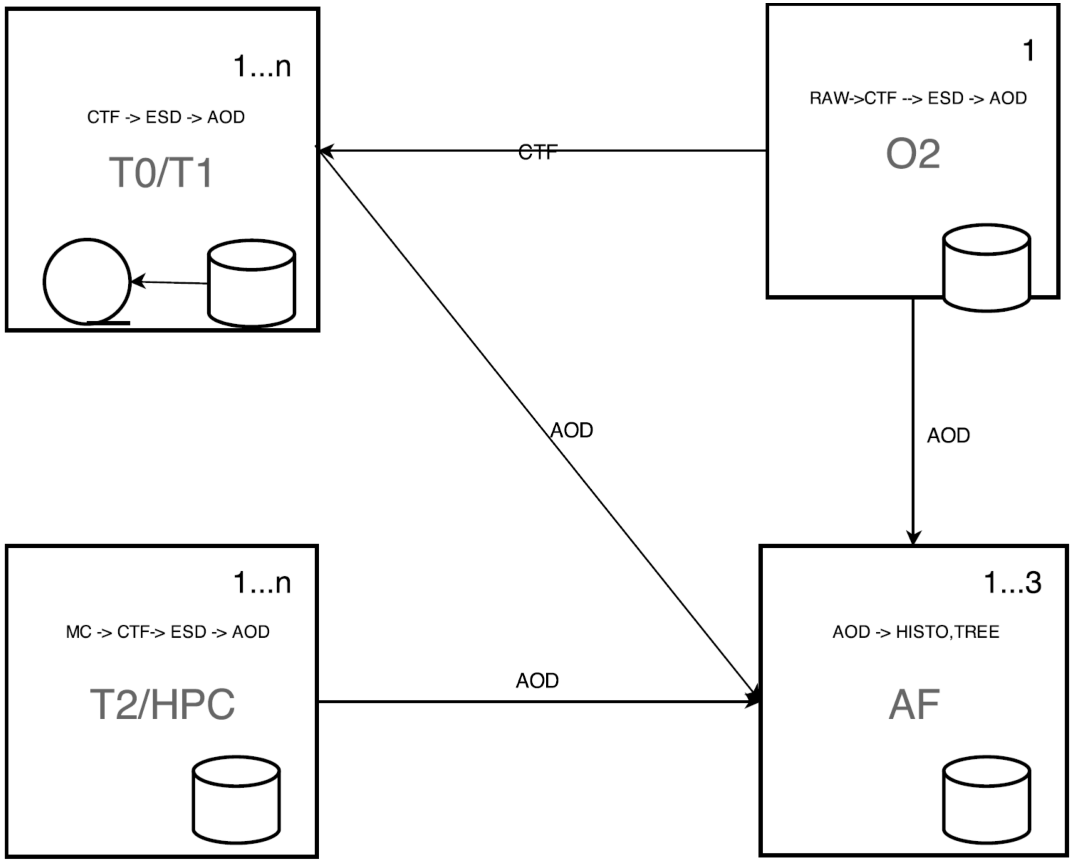
Acronym	Description
RAW	Raw data as it comes from the detector.
CTF	Compressed Time Frame containing the history of OM(100 ms) of detector readout information in the form of identified clusters that belong to identified tracks.
ESD	Event Summary Data.
AOD	Analysis Object Data for physics analysis.
HISTO	The subset of AOD information specific for a given analysis.
MC	Montecarlo simulation



by Pierre Vande Vyvre

Computing model (O2 processing flow)

Facility	Function
O2	**ALICE Online-Offline Facility at LHC Point 2**. Online reconstruction during the run. Provides data storage capacity. After data taking: runs the calibration and reconstruction tasks.
T0	**CERN Computer Center** facility providing CPU, storage and archiving resources.
T1	Grid site connected to T0 with **high bandwidth network links (100+ Gb)** providing CPU, storage and archiving resources. Reconstruction and calibration tasks
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 organized **analysis activity**.

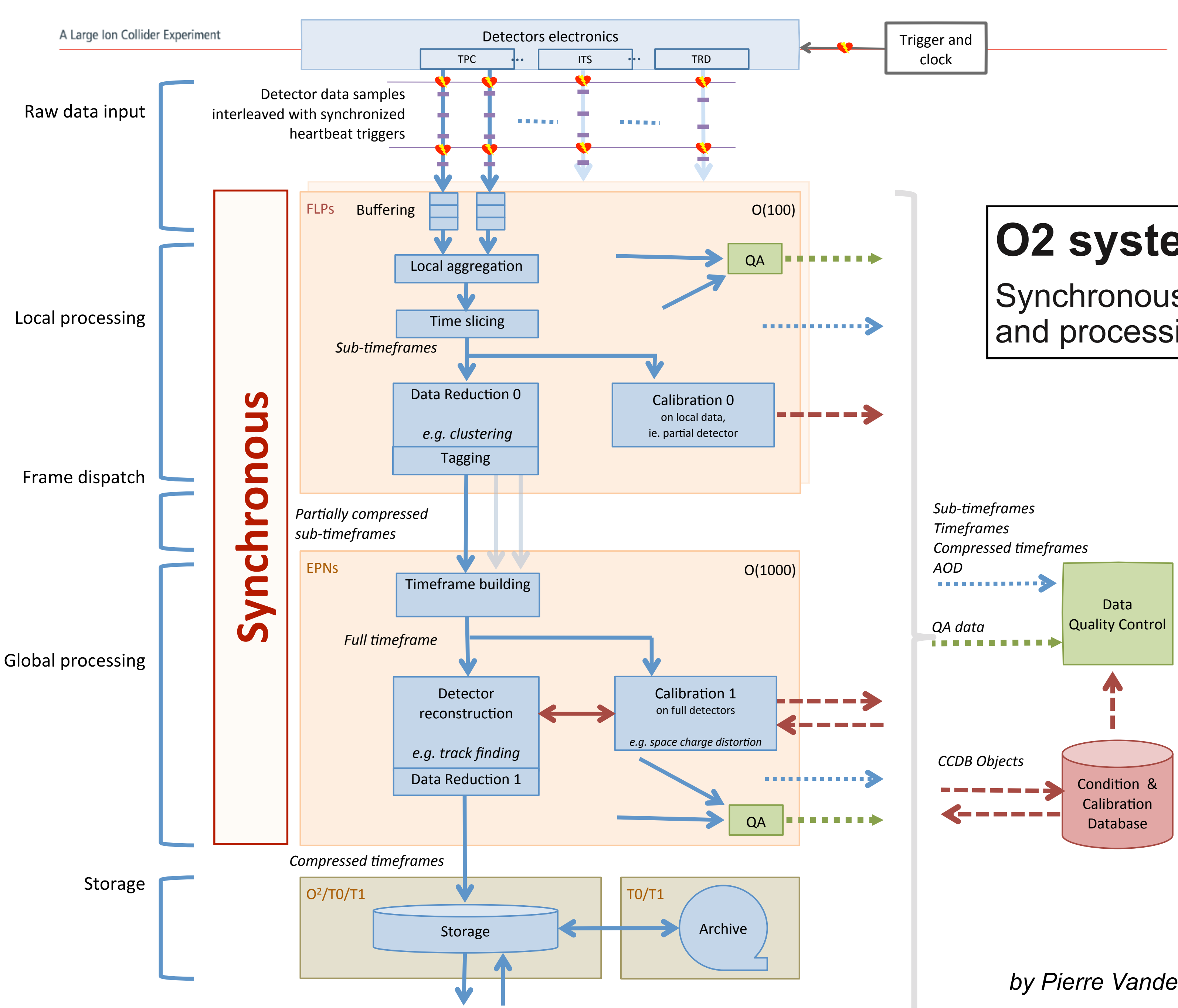


- Maintain the advantages of the Grid and the analysis trains
- Make it more open and more effective

by Pierre Vande Vyvre

O2 system (1)

Synchronous data flow and processing



by Pierre Vande Vyvre

The ALICE O2: Data Inputs

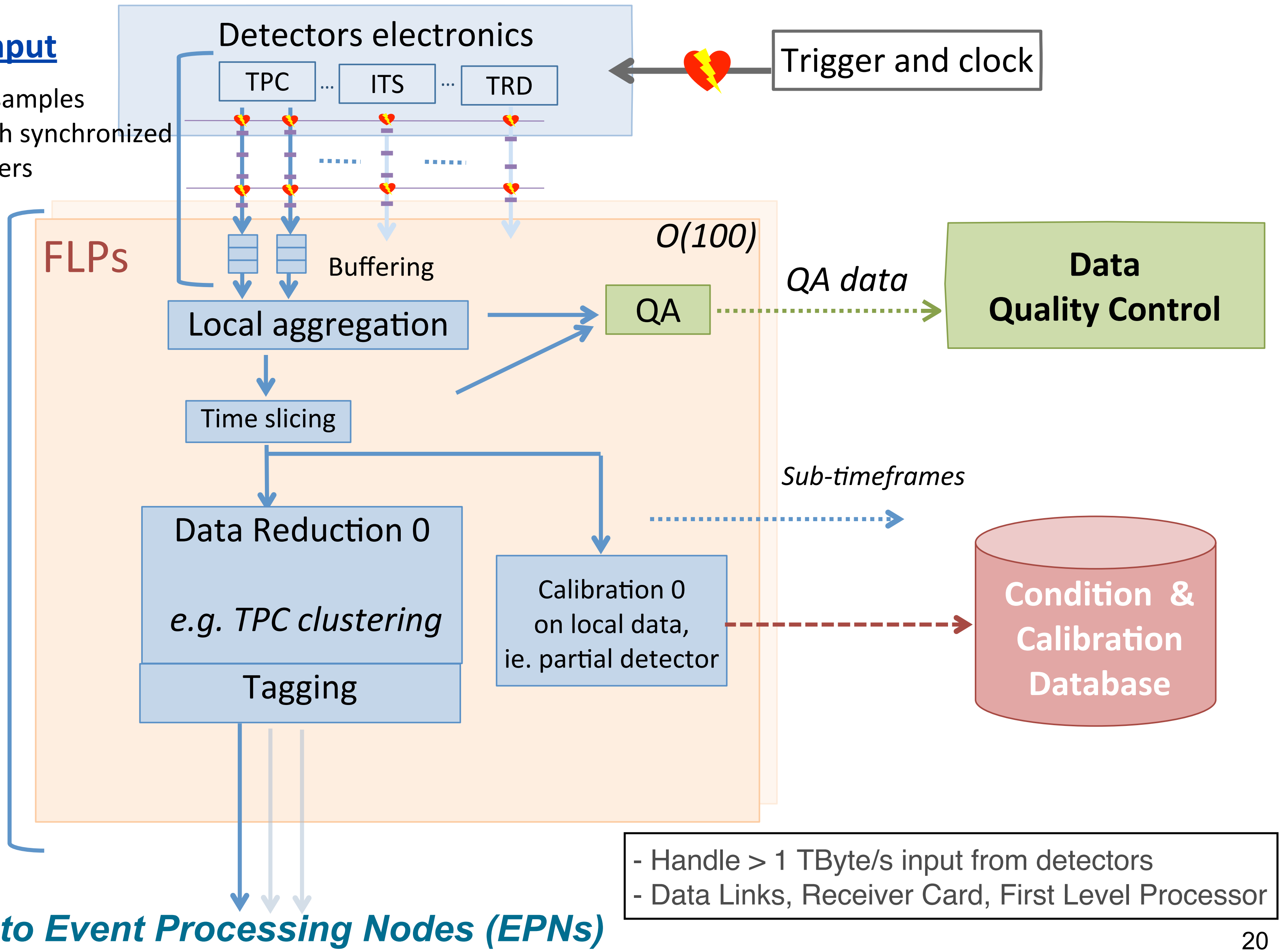
by Pierre Vande Vyvre (modified)

ALICE

Raw data input

Detector data samples interleaved with synchronized heartbeat triggers

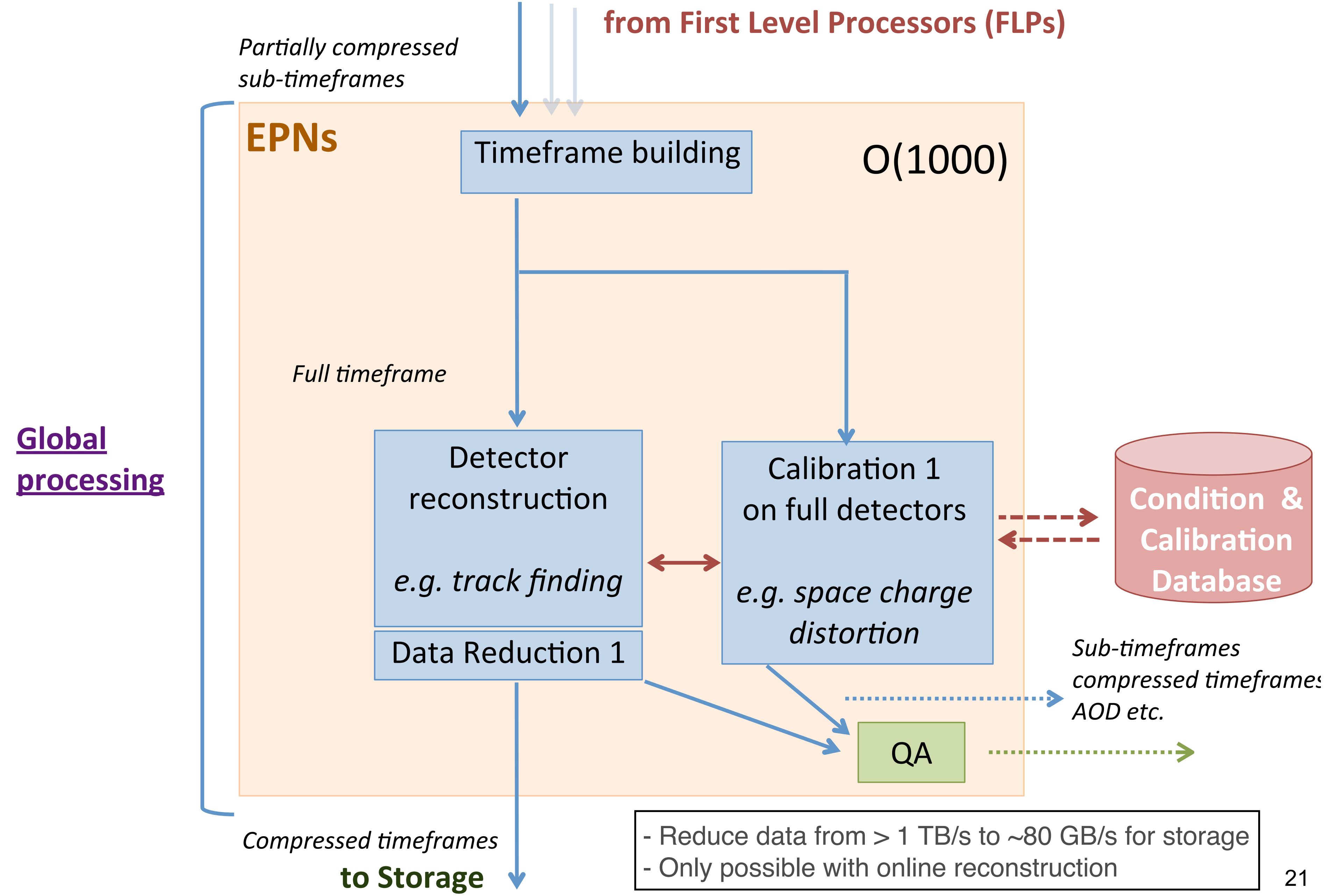
First Level Local Processing



- Handle > 1 TByte/s input from detectors
- Data Links, Receiver Card, First Level Processor

The ALICE O2: Data Reduction (I)

by Pierre Vande Vyvre (modified)



- Reduce data from > 1 TB/s to ~80 GB/s for storage
- Only possible with online reconstruction

The ALICE O2: Data Reduction (II)

	Dataflow Stage	Data Reduction Factor	Event Size (MByte)
	Raw Data	1	700
FEE →	Zero Suppression	35	20
High Level Trigger	Clustering & Compression	5 – 7	~ 3
High Level Trigger	Remove clusters not associated to relevant tracks	2	1.5
High Level Trigger	Data Format Optimization	2 – 3	< 1

slide by A. Uras (IC3INA 2013)

The ALICE O2: Data Reduction (III)

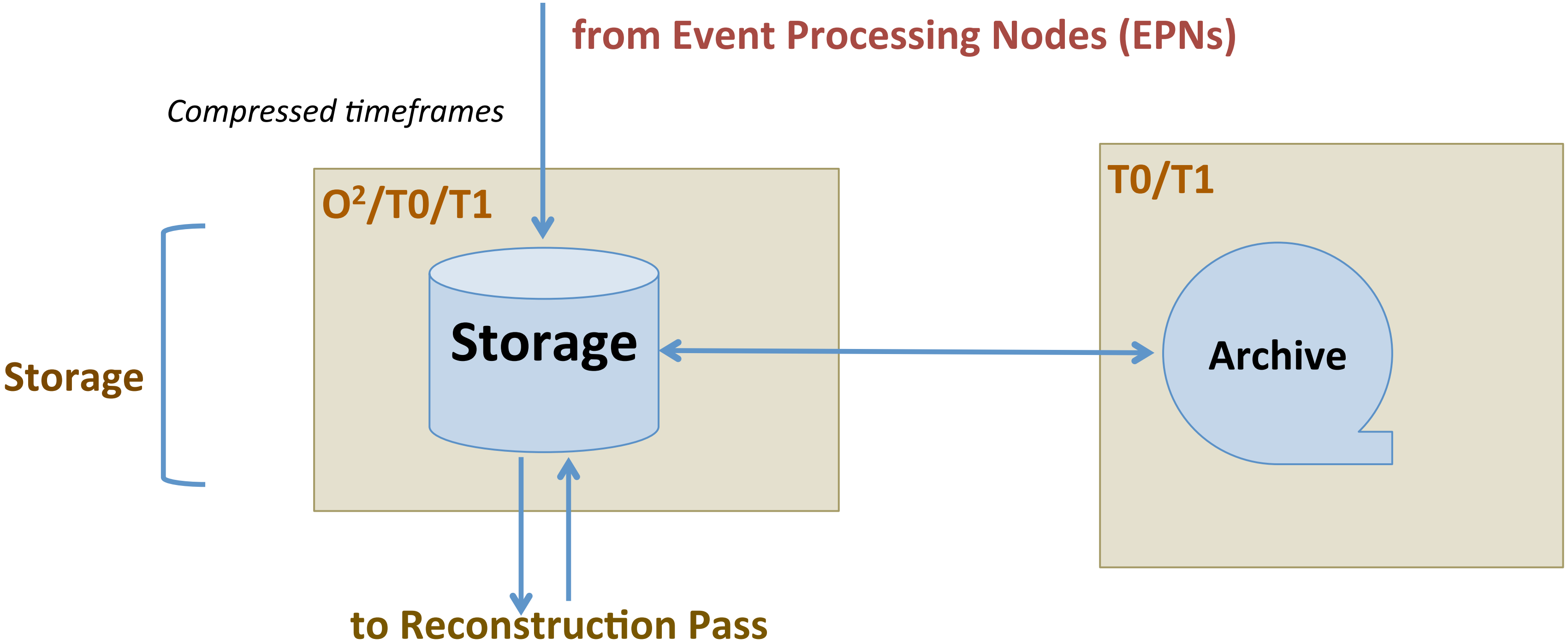
Detector	Event Size (MByte)	
	After Zero Suppression	After Data Compression
TPC	20.0	1.0
TRD	1.6	0.2
ITS	0.8	0.2
Others	0.5	0.25
TOTAL	**22.9**	**1.65**

- Data compression factors ranging from 2 to 20 according to the detector
- TPC still accounts for 60% of the total event size

slide by A. Uras (IC3INA 2013)

The ALICE O2: Data Storage

by Pierre Vande Vyvre (modified)

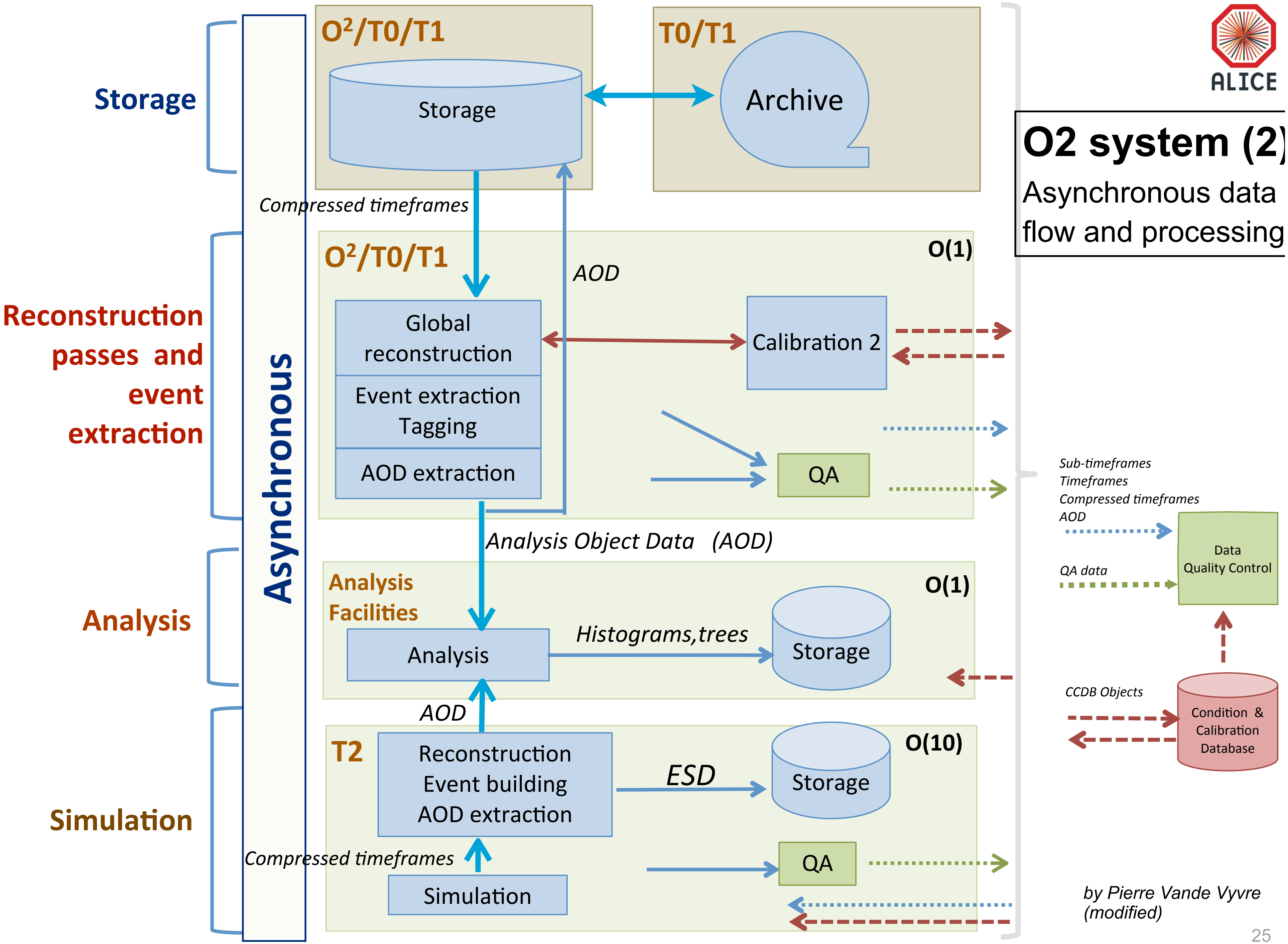


•Data in "intermediate" formats (not directly usable for physics analysis):
- 80 GByte/s peaks to be handled, distributed over ~1,250 nodes
- Average load of 15 GByte/s
- Local storage in O2 system
- Permanent storage in computing center

•Data in "final" formats (usable for physics analysis):
- GRID storage, accessible by experiment's users

O2 system (2)

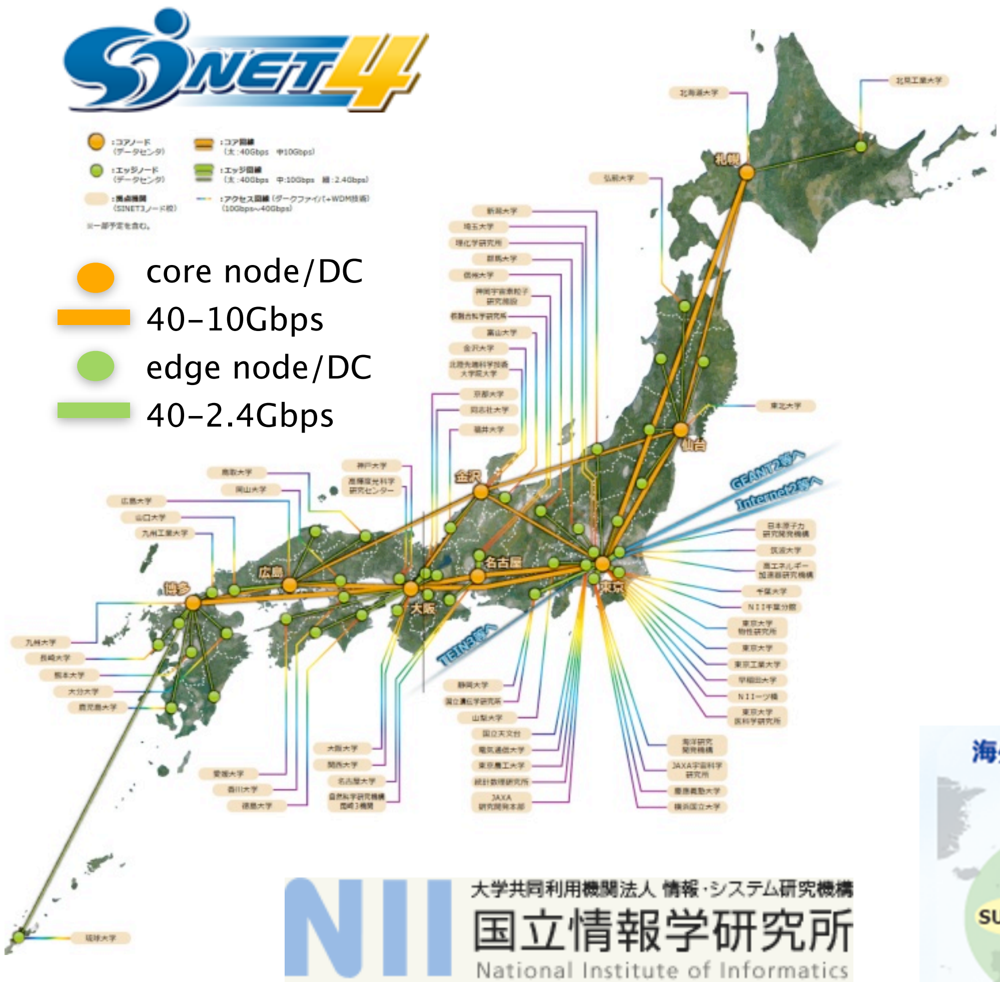
Asynchronous data flow and processing



by Pierre Vande Vyvre (modified)

Network in Asia and Japan for ALICE

Hiroshima T2 Approaching to LHCONE

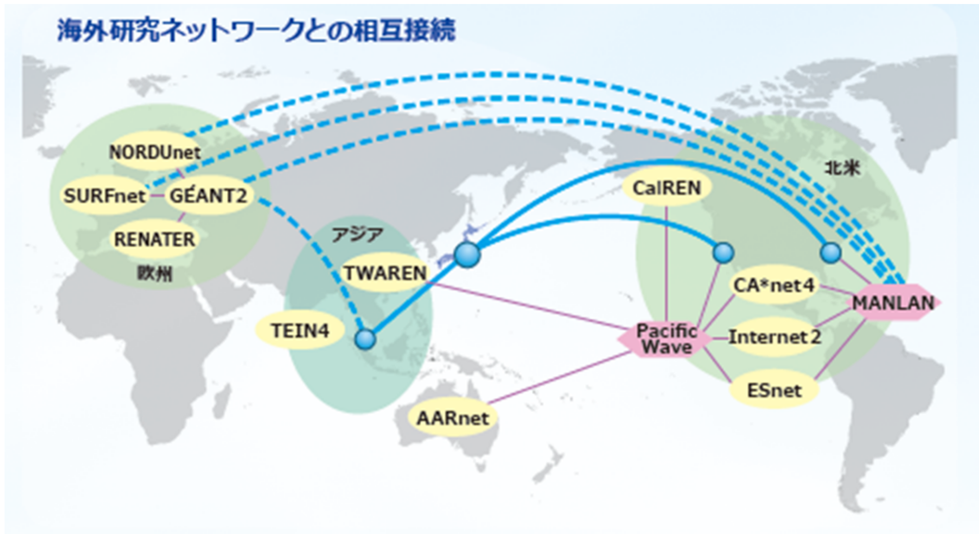


Present in SINET-4 (-2015)

- Hiroshima DC: 40Gbps Core node
- T2 to KEK via Hiroshima DC on 1Gbps-MPLS of HEPNet-J
- Internat'l connection via a default SINET routing

New SINET-5 announced by NII (2016-)

- Domestic nodes at 100Gbps, and upgrade to 400Gbps/1Tbps later
- Direct links to US/EU at 100Gbps
- Approach to LHCONE at 10Gbps



slide by T. Sugitate

Possible scenarios on ALICE network

- **2015-2017:**
  - LHC Run2 data taking, x2 more heavy ion data.
  - Data analysis on Run-2 (+Run-1)
  - Almost no change from Run-1 scheme. Due to data larger data volume, network traffic in Asia will increase, at least x2.
- **2018- beyond:**
  - LHC Run-3 data taking, x100 more data.
  - Architecture change (O2) applied.
    - O2, AF, and T0/T1/T3 scheme.
  - Significant data reductions, reconstruction in O2 mainly, and analysis → reduce data volume.

1. Can keep the similar network traffic as Run-2?
2. Or if have **AF in Asia (using HPC)**, then it will need more network traffic than that in Run-2 → Accelerate local physics analysis in Asia.

➡ **Direct links to US/EU at 100Gbps may be necessary in case of Japan?**

Summary and Outlook

- **ALICE computing upgrades on online-offline for the data taking after 2018 is ongoing.**
  - Continuous minimum bias event readout at **50 kHz in Pb-Pb collisions**.
  - **1 TB/s raw data** from detector, need a significant data reduction down to 80 GB/s to storage, and make a physics outputs timely.
  - O2 Scheme: Online reconstruction and calibration by O2 (near ALICE) & T0/T1, organized analysis at Analysis Farm (AF), and simulation at T2.

- Designing based on Physics requirements is completed.
- Intensive works on, modeling, Technologies (processing platform & network), O2 prototyping.
- Technical Design Repot (TDR) is progressing. It will be submitted to LHCC in April 2015.

ALICE O² Project

A Large Ion Collider Experiment

Project Organization

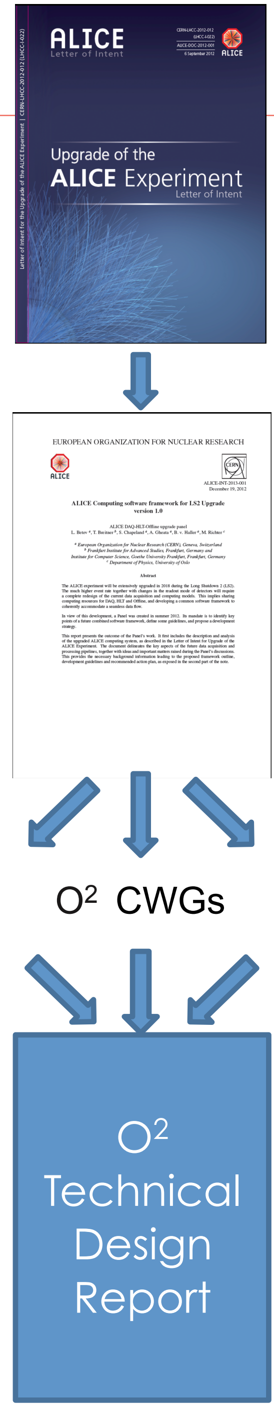
PLs: P. Buncic, T. Kollegger, M. Krzewicki, P. Vande Vyvre

	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

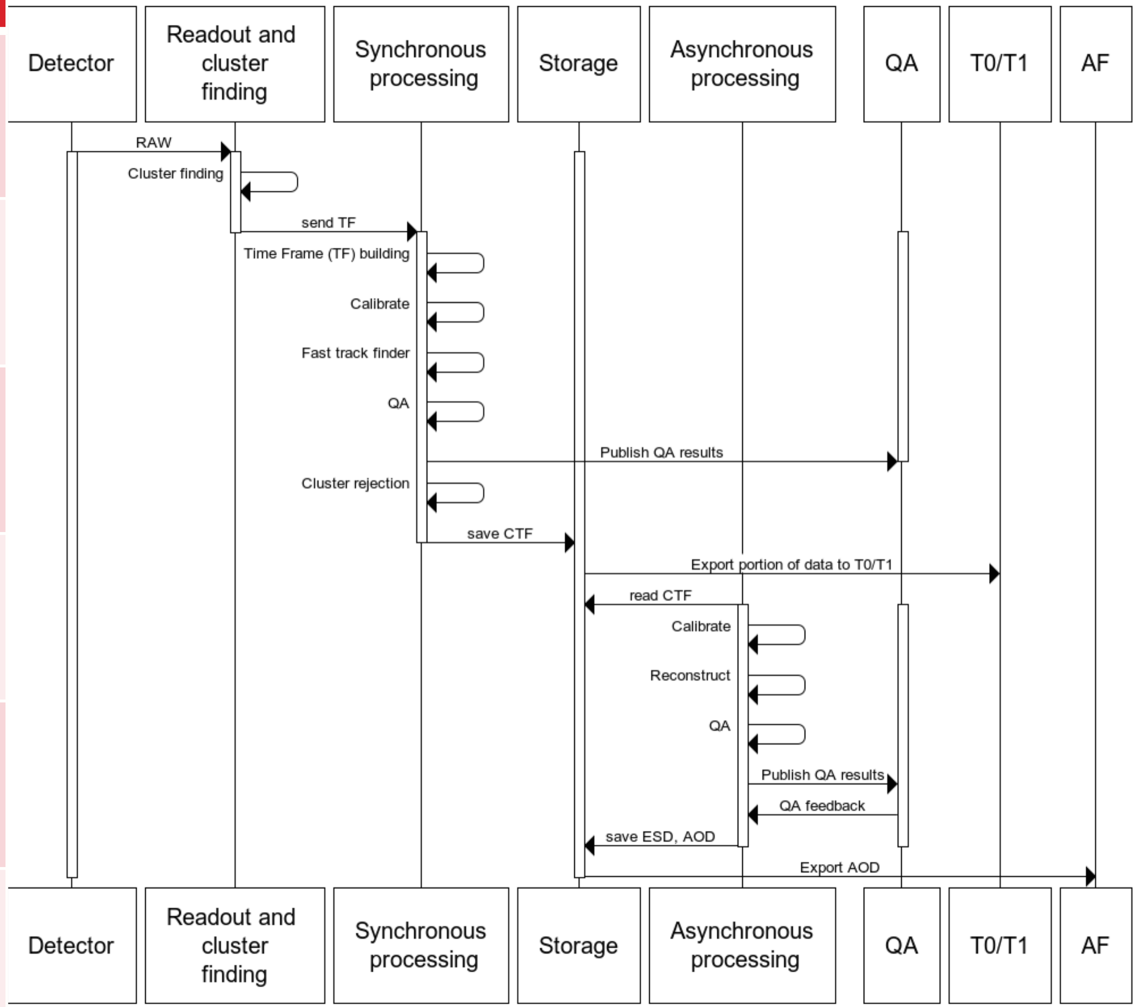


Back up

Computing model (Data flow)

A Large Ion Collider Experiment

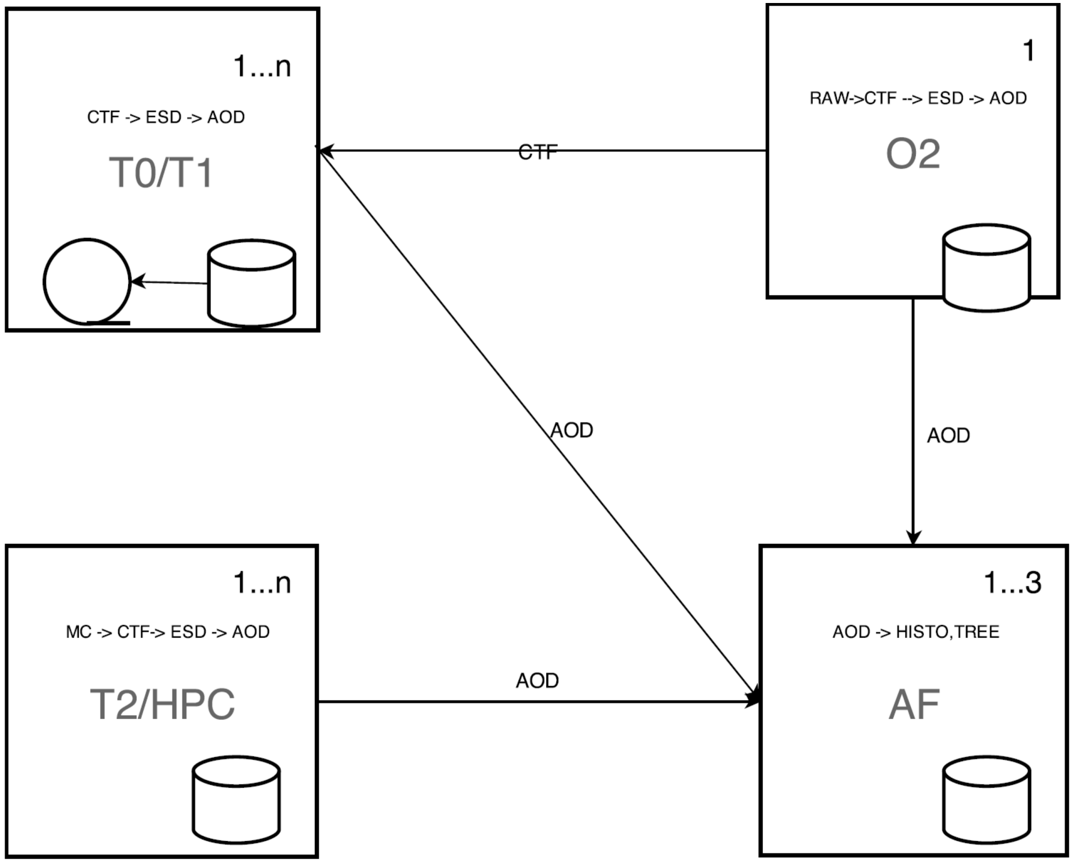
Acronym	Description	Persistency
RAW	Raw data as it comes from the detector.	Trans.
CTF	Compressed Time Frame containing the history of OM(100 ms) of detector readout information in the form of identified clusters that belong to identified tracks.	Persist.
ESD	Event Summary Data. Auxiliary data to CTF 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)

A Large Ion Collider Experiment

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

by Pierre Vande Vyvre