O^2 : A novel combined online and offline computing system for ALICE after 2018

Pierre VANDE VYVRE for the O^2 project

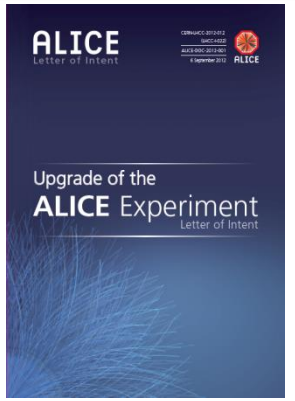
15-Oct-2013 – CHEP – Amsterdam, **Netherlands**

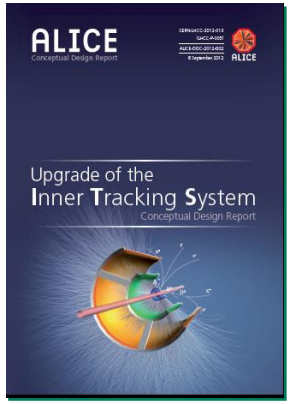
Outline

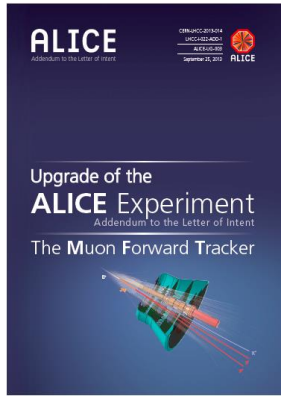
- ALICE apparatus upgrade
- New computing requirements
- O^2 system
  - Detector read-out
  - Data volume reduction
  - Big data
- O^2 project
  - Computing Working Groups
- Next steps

ALICE LS2 Upgrade

- 2018/19 (LHC 2nd Long Shutdown)
- Inner Tracking System (ITS)
  - New, high-resolution, low-material ITS
- Time Project Chamber (TPC)
  - Upgrade of TPC with replacement of MWPCs with GEMs
  - New pipelined continuous readout electronics
- New and common computing system for online and offline computing
- New 5-plane silicon telescope in front of the Muon Spectrometer

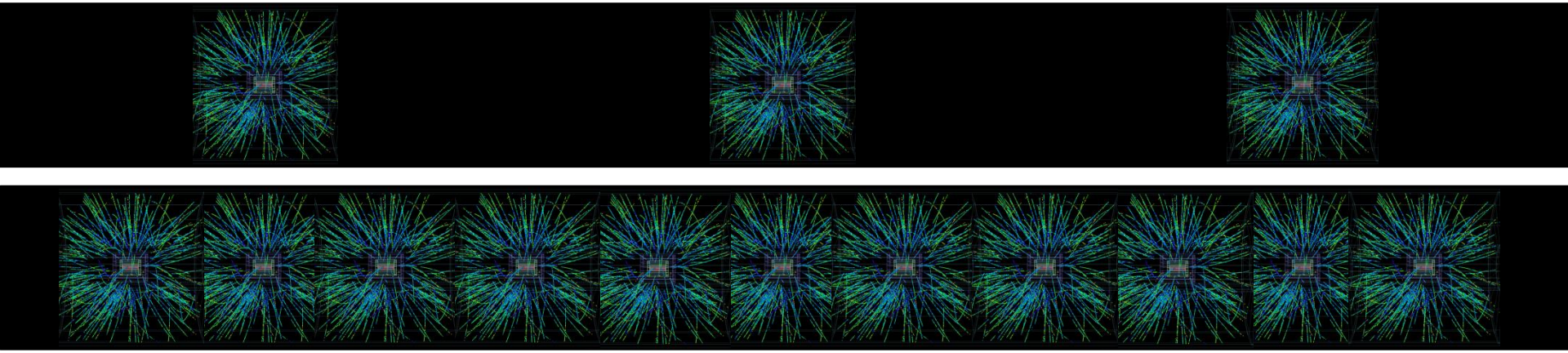






Requirements: Event Rate

- Rate increase: from 500 Hz to 50 kHz
  - Physics topics require measurements characterized by very small signal-over-background ratio → large statistics
  - Large background → traditional triggering or filtering techniques very inefficient for most physics channels.
  - Strategy: read out all particle interactions 50 kHz (anticipated Pb-Pb interaction rate)
- TPC intrinsic rate << 50 kHz
  - In average 5 events overlapping in the detector
  - Continuous read-out

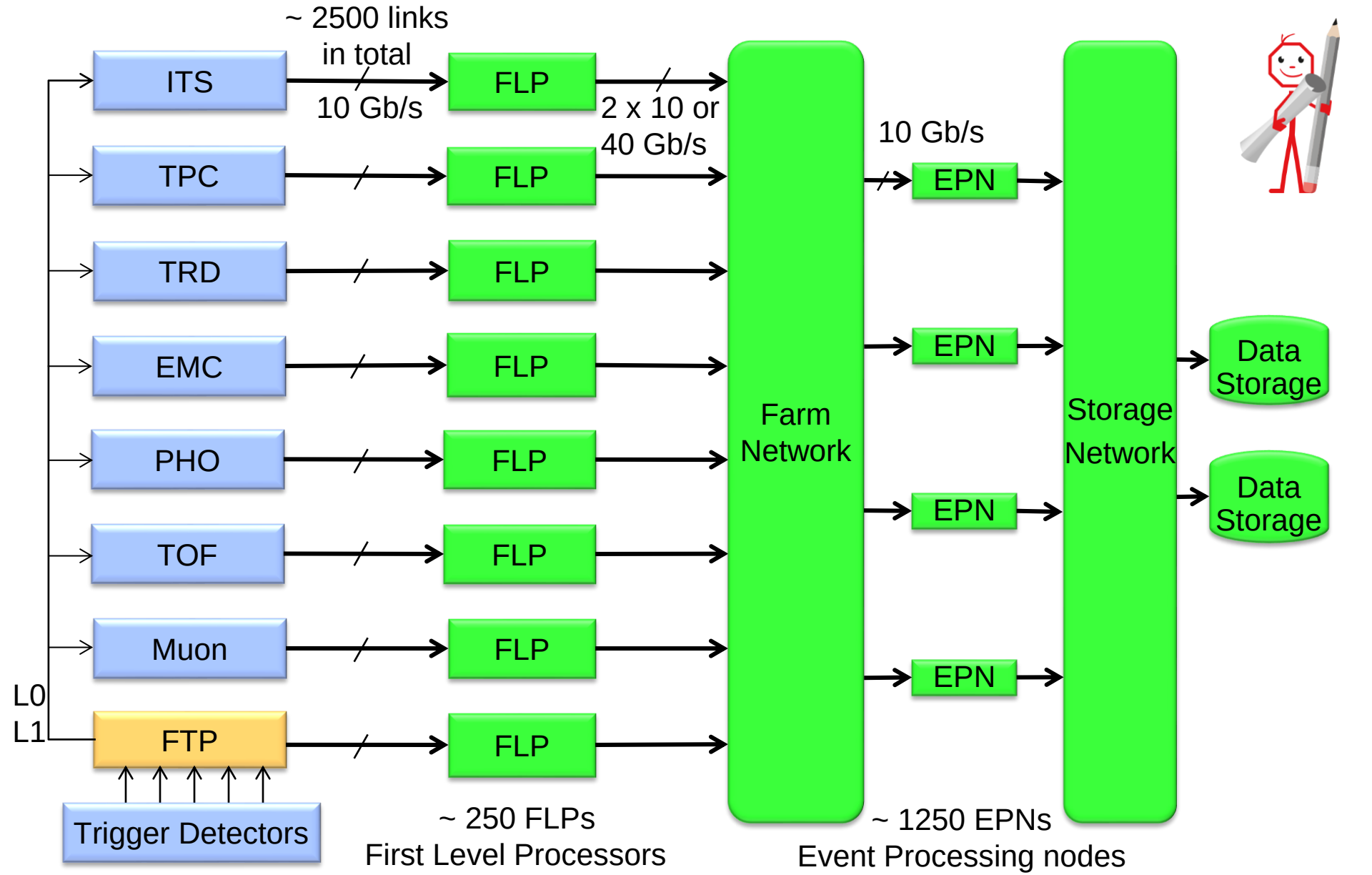


Requirements: Data Volume

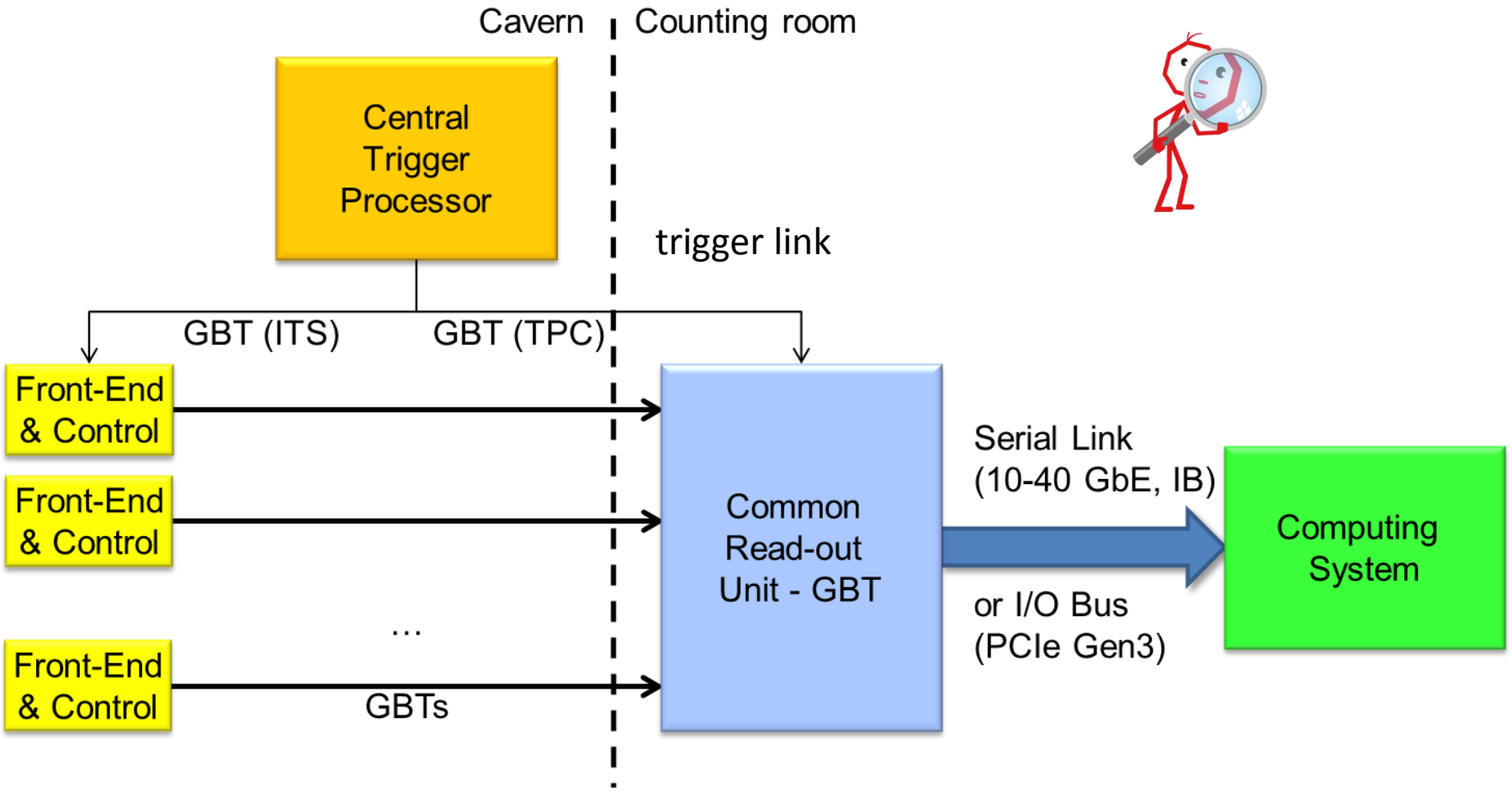
Detector	Event Size After Zero Suppression (MByte)	Bandwidth @50 kHz Pb-Pb (GByte/s)
TPC	20.0	1000
TRD	1.6	81.5
ITS	0.8	40
Others	0.5	25
Total	**22.9**	**1146.5**

- Massive data volume reduction needed
- Only option is by online processing

O² Hardware System



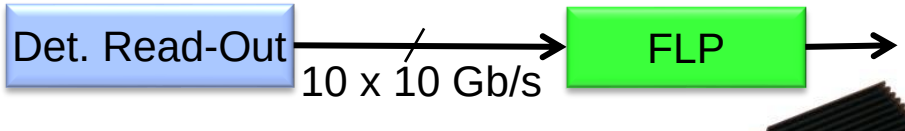
Detector Interface

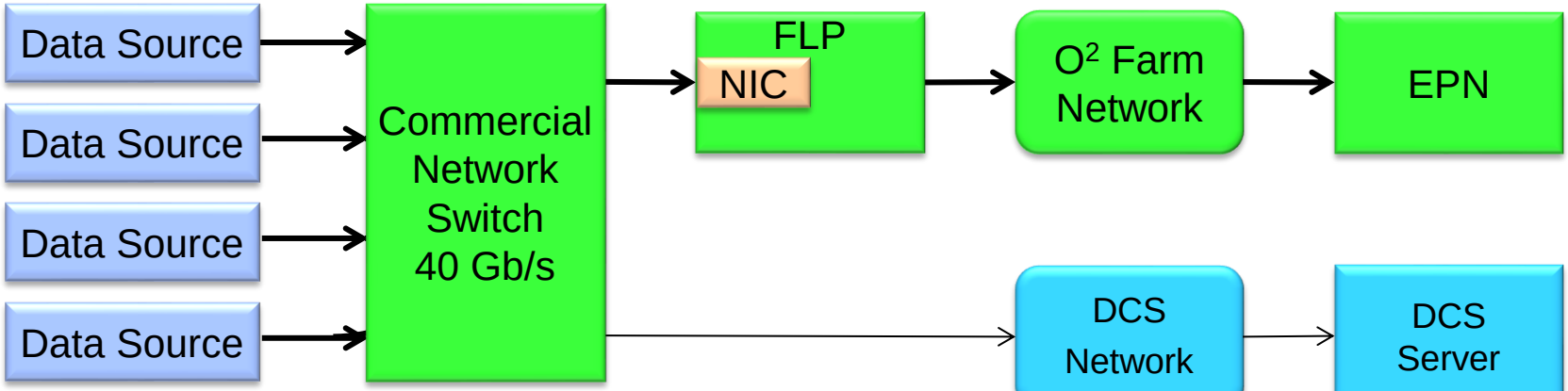


- GBT: custom radiation-hard optical link

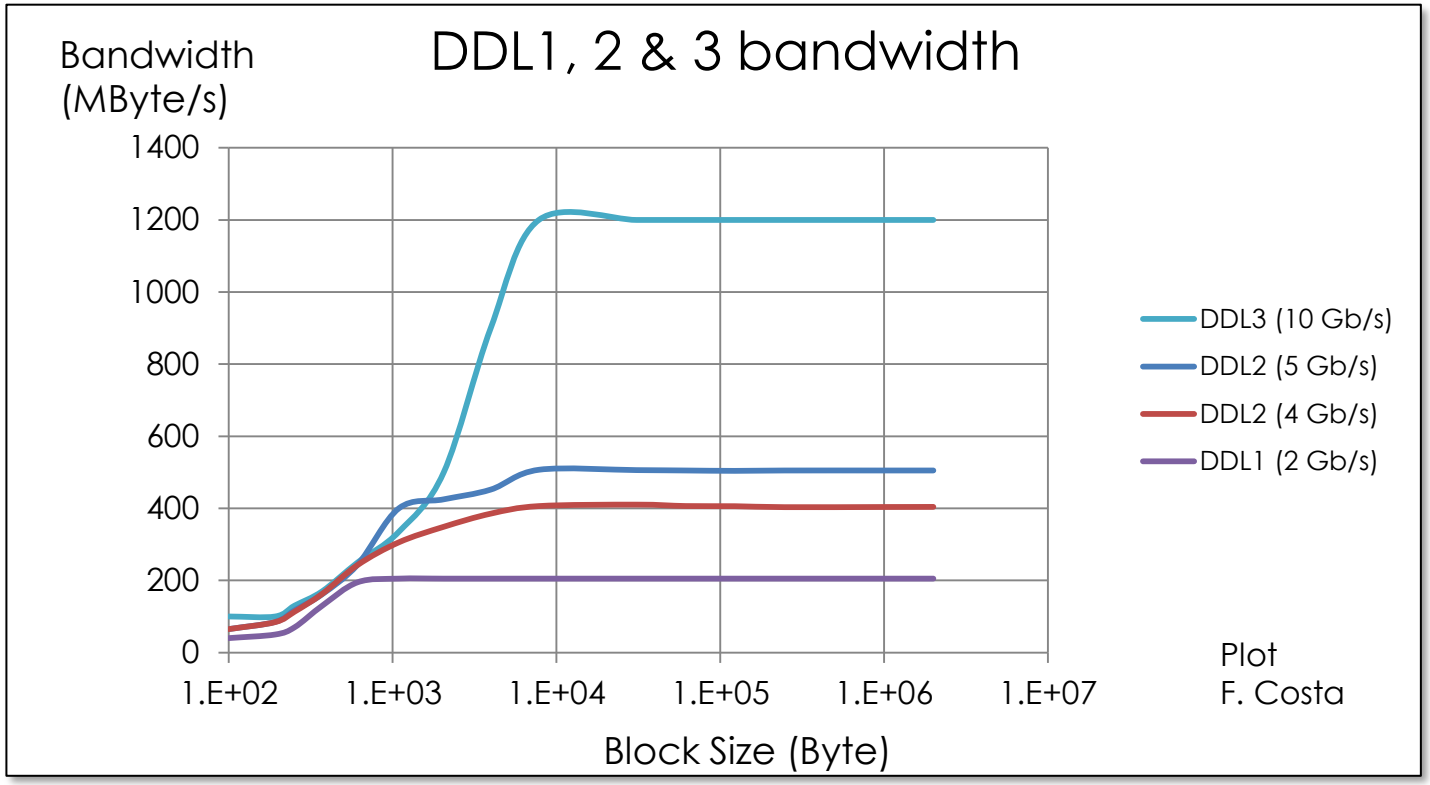
Detector Data Link 3 (DDL3)

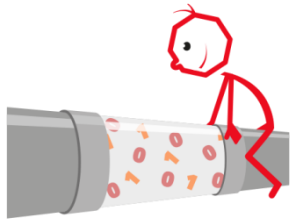
- DDL3: 10 Gb/s (LoI target) using a commercial standard (Ethernet or serial PCIe)
- Commercial products at 40 or 56 Gb/s available now
  - Dual-port 40 GbE Network Interface Card (NIC) (Chelsio) (40 GbE made of four lanes of multi-mode fiber at 10 Gb/s)
  - Dual-port 56 GbIB (QDR) (Mellanox)
  - Multiplex 4 x DDL3 over 1 input port of a commercial NIC
    - Breakout splitter cable (4 x 10 Gbe ↔ 1 x 40 GbE)
    - Commercial network switch (staging, DCS data demultiplex, etc)
  - Both options tested in the lab with equipment on loan giving the expected performance of 4 x 10 Gb/s





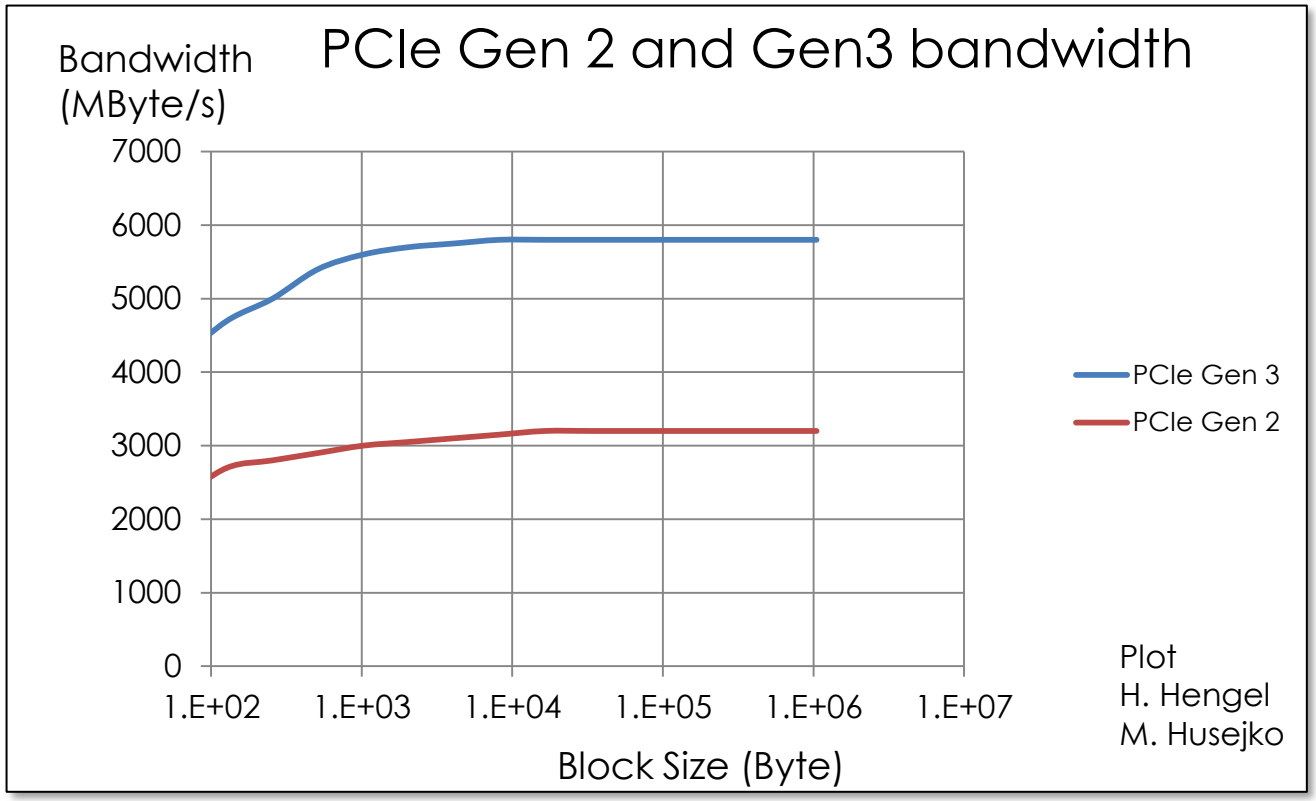
DDL Performance Evolution

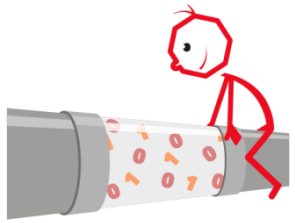




- DDL1 at 2 Gb/s used by all ALICE detectors for Run 1 (radiation tolerant)
- DDL2 at 4 and 5 Gb/s (according to needs) ready for Run 2
- Prototype for one of the DDL3 option considered for Run 3 implemented (Eth. + UDP/IP)
- Expected performance evolution verified

PC Input/Output Bandwidth

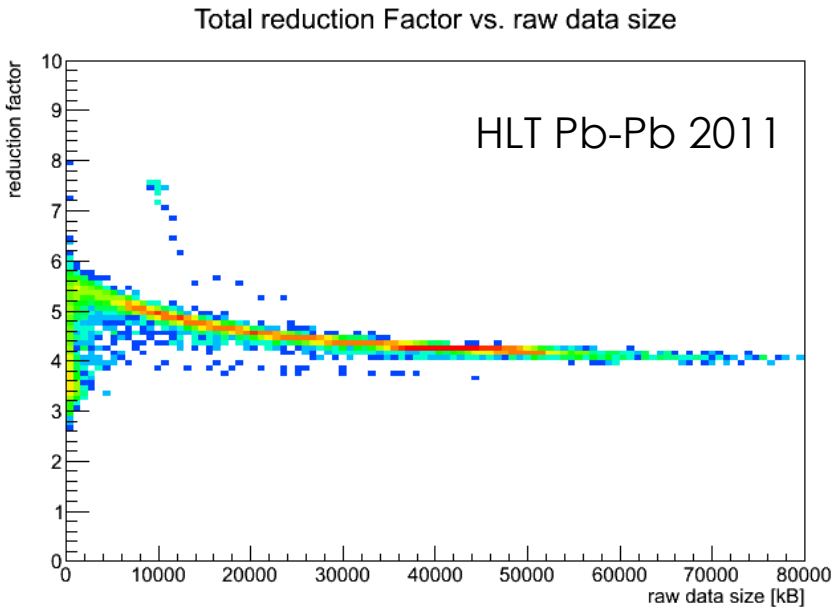




- 1 key element for the O2 system will be the I/O bandwidth of the PCs
- PCIe Gen2 performance measured for the LoI
- PCIe Gen3 measured with a FPGA development board (Xilinx Virex-7 Connectivity Kit VC709)
  - Large data blocks: wire speed 8 GB/s, theoretical max 7.2, measured 5.8
  - FLP I/O capacity needed will at least require: 3 slots PCIe Gen 3 x8 or 2 slot x16

TPC Data Volume Reduction

Data Format		Data Reduction Factor	Event Size (MByte)
	Raw Data	1	700
FEE	Zero Suppression	35	20
HLT	Clustering & Compression	5-7	~3
	Remove clusters not associated to relevant tracks	2	1.5
	Data format optimization	2-3	<1



- TPC data volume reduction by online event reconstruction
- Discarding original raw data
- In production from the 2011 Pb-Pb run

Total Data Volume

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**

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

Heterogeneous Platforms

- Shift from 1 to many platforms
  - Intel Xeon X64
  - Intel Xeon Phi (many cores)
  - GPUs
  - Low cost processors
  - FPGAs

- Benchmarking in progress to assess their relative merits

Dataflow Model (1)

Raw Data Input

Local Processing

Frame dispatch

Global Processing

Event Building

Detectors electronics

TPC ... TRD

Trigger and clock

Data Sample n+1

Data Sample n

Data Sample n-1

Read-Out

Read-Out

FLPs

Buffering

Raw data sample

Raw and Processed data

Data Reduction 0

- e.g. TPC clusters

Tagging

Calibration 0

(on local data, ie. partial detector)

Raw data sample

Time slicing

Time slicing could actually occur before, e.g. on the front-end or receiving part

Sub Time Frame

EPNs

Data aggregation

Full Time Frame

Full Time Frame

Condition & Calibration Databases

- Global reco
- Tracks/Tracklets
- Drift time
- Vertex

Data Reduction 1

Event building Tagging

Calibration 1

(on full detectors)

Events

Raw data

QA/DQM

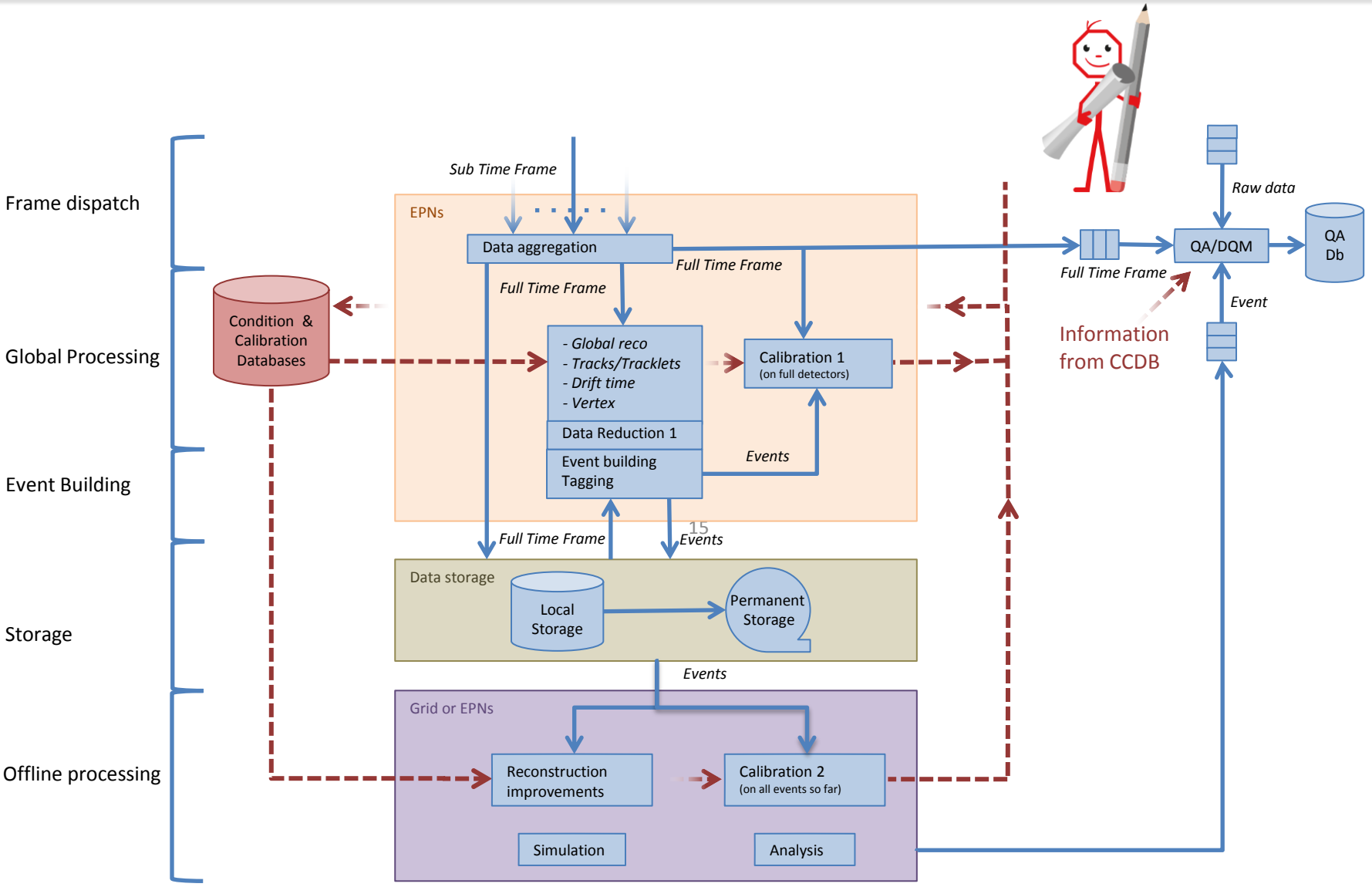
QA Db

Full Time Frame

Event

Information from CCDB

Dataflow Model (2)

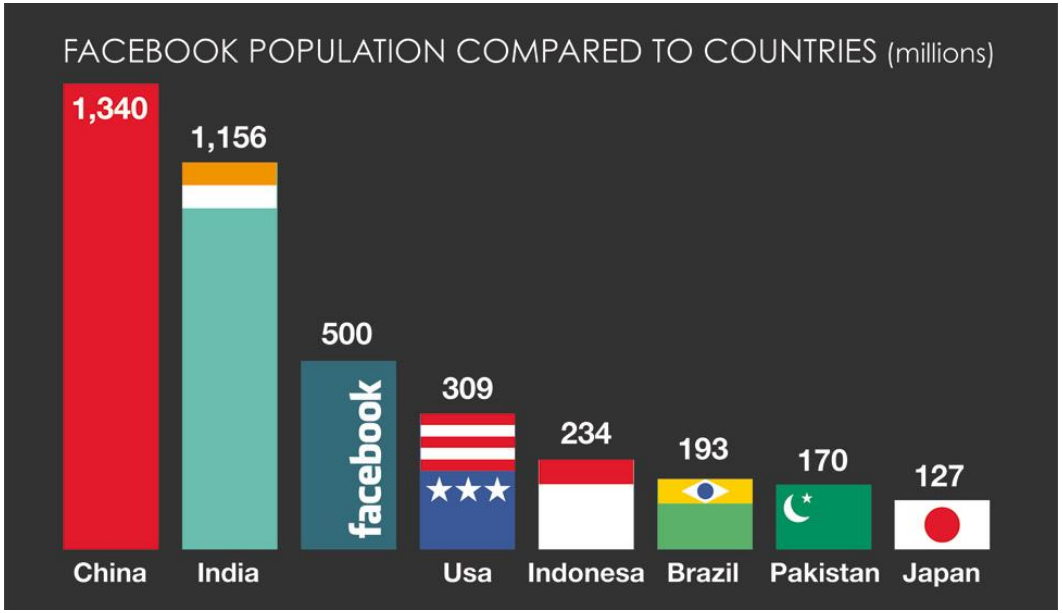


Internet : an inflating universe…

- HEP is not alone
  in the computing universe !

- 1 ZB/year in 2017 (Cisco)
- 35 ZB in 2020 (IBM)
- 1 ZB = 1'000 EB = 1'000'000 PB

- Number of users
  (Kissmetrics)



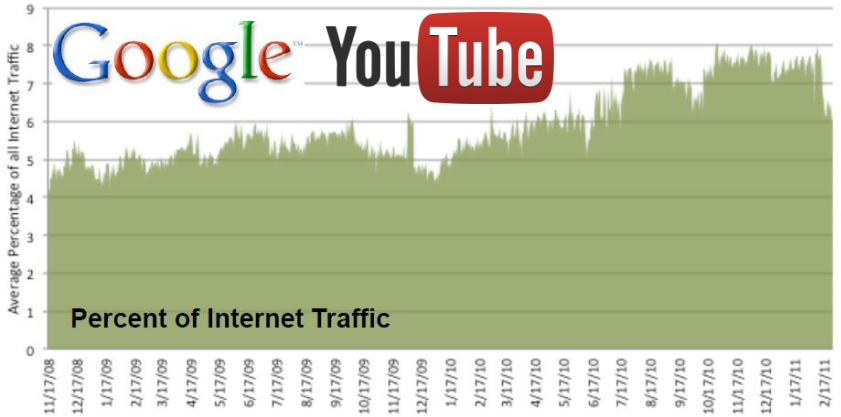
…with a few very large galaxies !

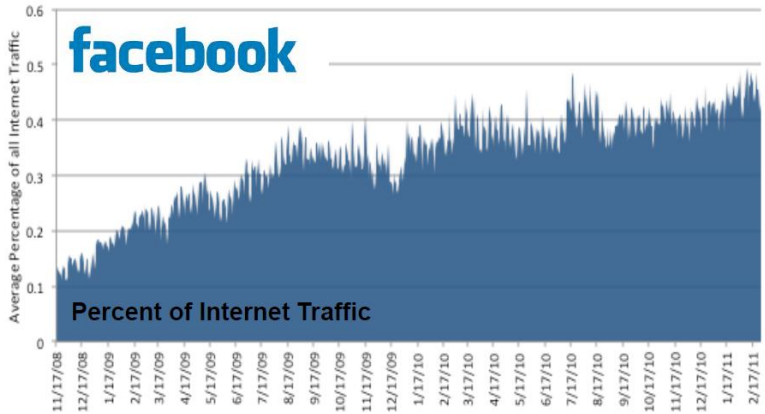
- "Hyper giants": the 150 companies that control 50% of all traffic on the web (Arbor Networks)

- Google :
  100 billion searches/month,
  38'500 searches/second
- YouTube:
  6 billion hours of video are watched each month

- Facebook
  350 millions photos uploaded/day

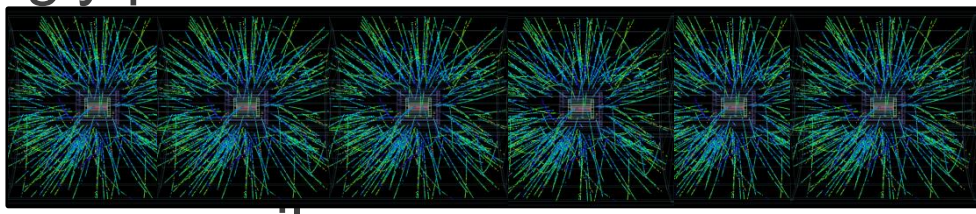
- HEP should definitely try to navigate in the wake of the Big Data hyper giants





Big Data approach

- Very large data sets
  - High Energy Physics data are inherently and embarrassingly parallel… but



  - At the luminosity targeted for the upgrade there will be some pile-up
    → Continuous dataflow → New framework must handle it
- Issues to become a Big Data shop
  - Lots of legacy software not designed for this paradigm
  - Fraction the work into small independent manageable tasks
  - Merge results

O² Framework

Analysis

Data Acquisition

Simulation

For example

ØMQ

Cluster Finding

Analysis

Tracking

ESD, AOD

O^2 Project

O^2 Steering Board

Project Leaders

Projects

DAQ | HLT | Offline

Computing Working Groups

- 50 people active in 1-3 CWGs
- Service tasks

CWG1 Architecture	CWG2 Procedure & Tools	CWG3 DataFlow	CWG4 Data Model	CWG5 Platforms	CWG6 Calibration
CWG7 Reconstruc.	CWG8 Simulation	CWG9 QA, DQM, Vi	CWG10 Control	CWG11 Sw Lifecycle	CWG12 Hardware
CWG13 Sw Framework	CWGnn -----	CWGnn -----	CWGnn -----	CWGnn -----	CWGnn -----

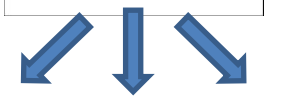
Institution Boards

Online Institution Board

Computing Board

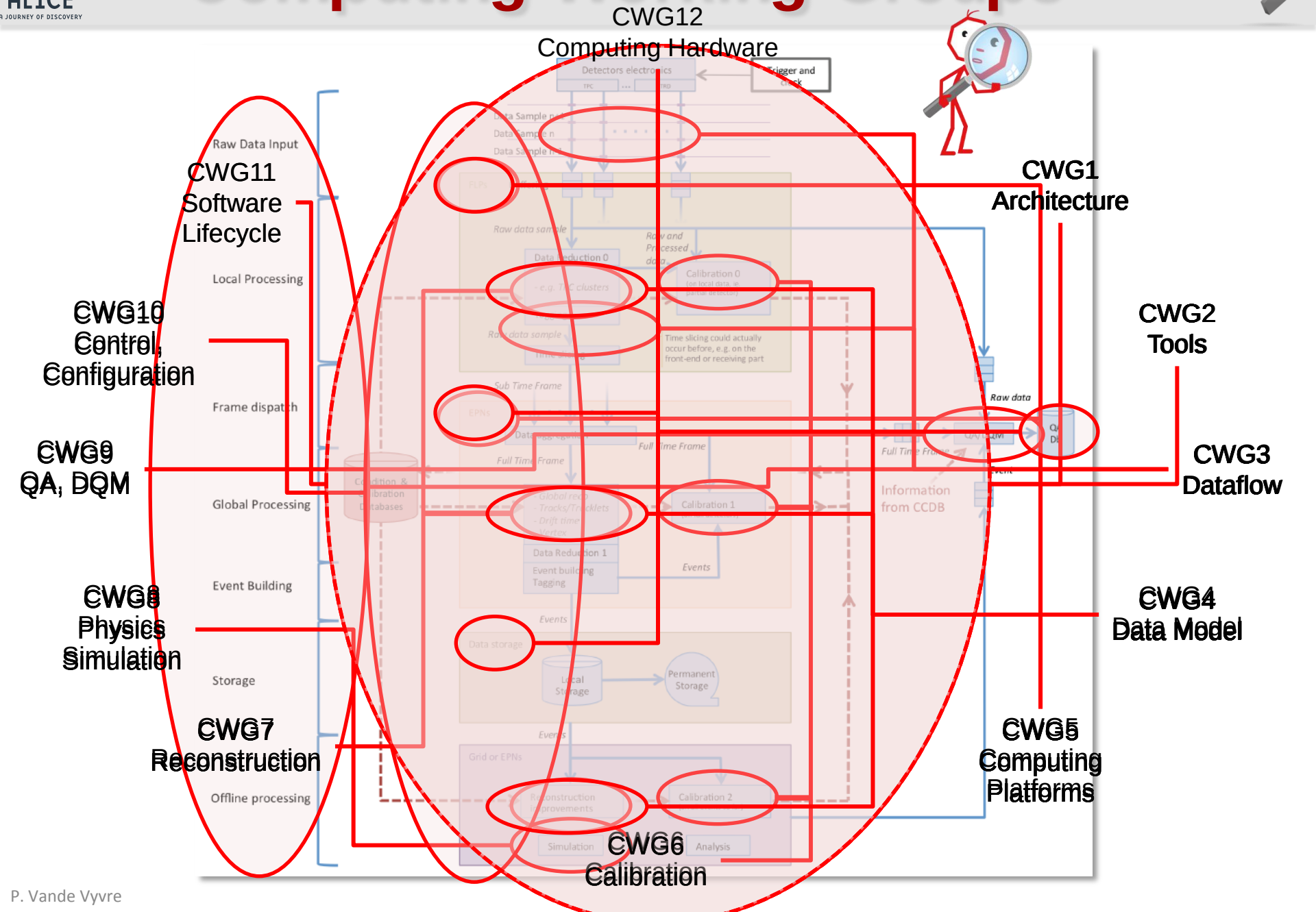
Overall Schedule

- Sep 2012 ALICE Upgrade LoI
- Jan 2013 Report of the DAQ-HLT-Offline software panel on “ALICE Computer software framework for LS2 upgrade”
- Mar 2013 O^2 Computing Working Groups
- Sep 2014 O^2 Technical Design Report



O2 Computing Working Groups

Computing Working Groups



Next steps

- Intensive period of R&D :
  - Collect the requirements: ITS and TPC TDRs
  - System modeling
  - Prototyping and benchmarking
- Technology and time are working with us
  - New options
  - Massive usage of commercial equipment very appealing
- Technical Design Report
  - Sep ’14: submission to the LHCC

Thanks !