A Large Ion Collider Experiment

Triggering Discoveries in High-Energy Physics

# ALICE $O^2$

## The Upgrade of the ALICE Online and Offline Computing after 2018

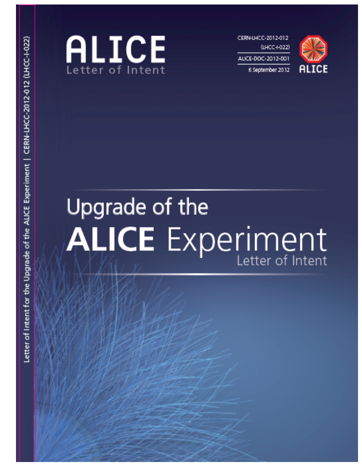

Thorsten Kollegger
for the ALICE Collaboration

FIAS Frankfurt Institute for Advanced Studies

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

# ALICE Upgrade Overview

**Planned for 2018 (LHC 2nd Long Shutdown)**
("Upgrade of the ALICE Experiment", LoI, CERN-LHCC-2012-12)



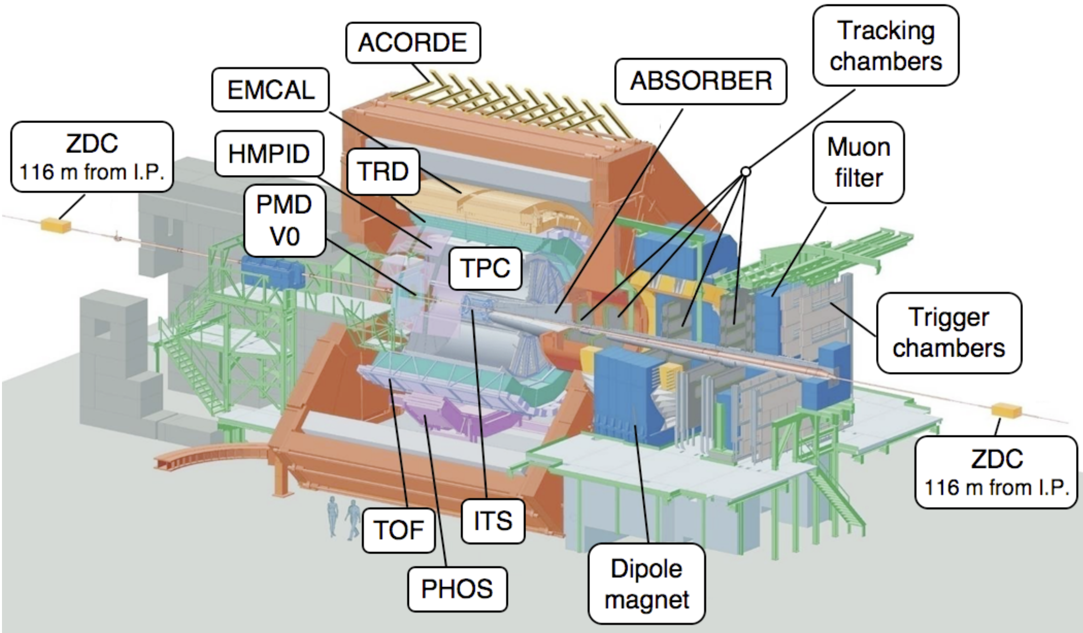

**Physics goals** (Michael Webers Talk on Tuesday)

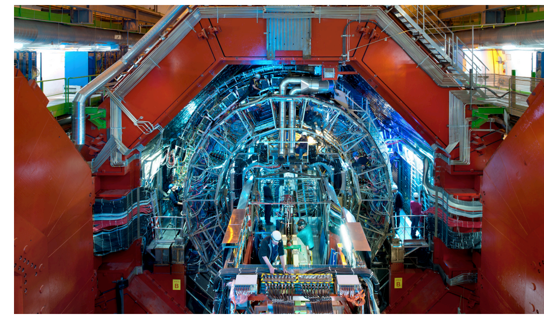
Heavy Flavor
Quarkonia
Low-mass dielectrons
Jets
Anti- and Hypernuclei

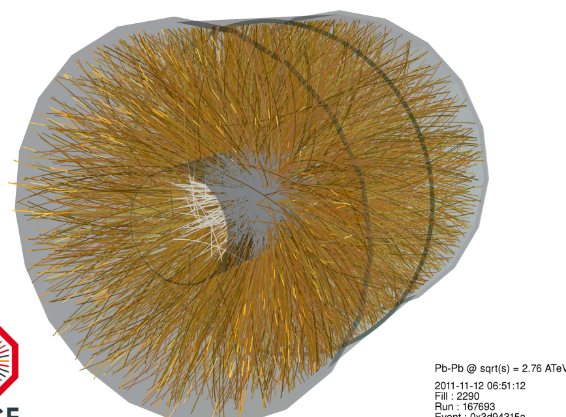
**Target**

- Pb-Pb recorded luminosity $\geq 10\ nb^{-1}$ ➡ $8 \times 10^{10}$ events
- pp (@5.5 Tev) recorded luminosity $\geq 6\ pb^{-1}$ ➡ $1.4 \times 10^{11}$ events

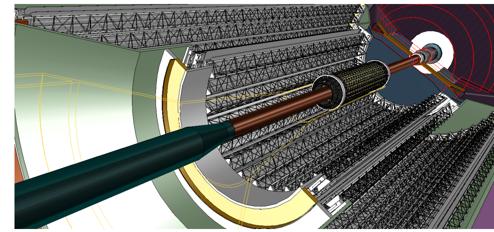
# A Large Ion Collider Experiment



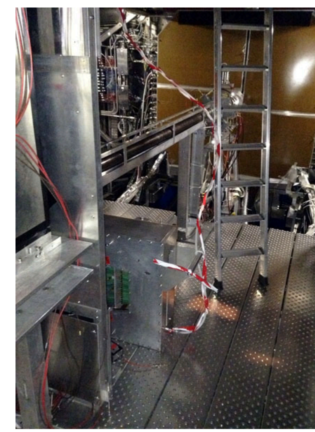


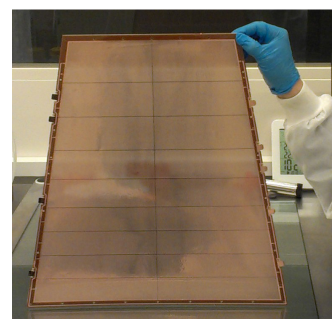
# ALICE Upgrade Overview

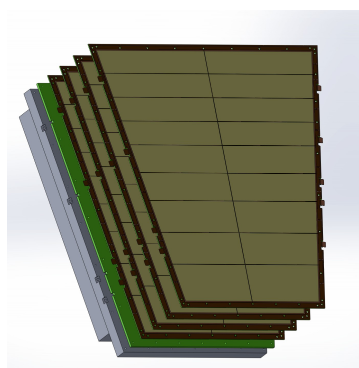


**The upgrade plan entails building**

- New, high-resolution, low-material ITS
- Upgrade of TPC with replacement of MWPCs with GEMs and new pipelined readout electronics
- Upgrade of readout electronics of: TRD, TOF, PHOS







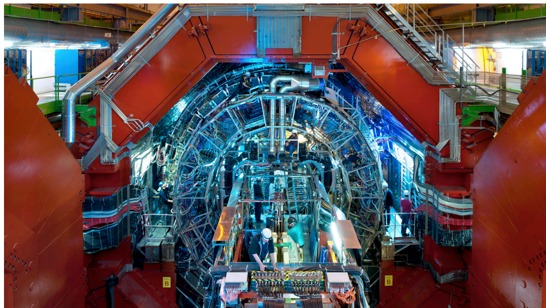
# ALICE Upgrade Overview

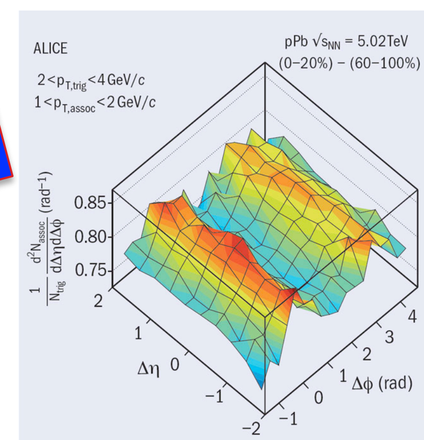
**The upgrade plan entails building**

- Upgrade of readout electronics of the Muon Spectrometer
- Upgrade of the forward trigger detectors and ZDC
- Muon Forward Tracker (MFT)
- **Upgrade of the online systems**
- **Upgrade of the offline reconstruction and analysis framework and code**

ALICE

# O$^2$ Project






From Detector Readout to Analysis:
What is the “optimal” computing architecture?

# Requirements

Sample full 50kHz Pb-Pb interaction rate
(current limit at ~500Hz, factor 100 increase)

Typical event size of PbPb collisions@5.5TeV: 22 MByte
⇨ ~1.1 TByte/s detector readout
⇨ ~500 PByte/HI period (1 month)

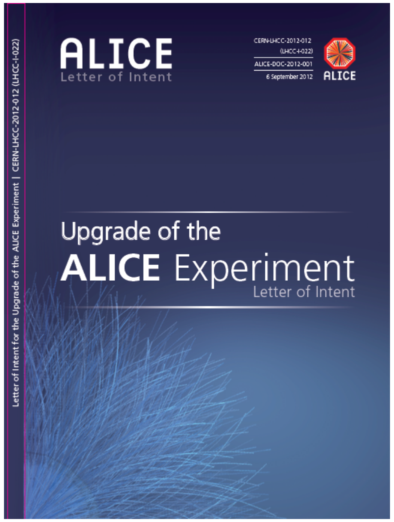
*However:*
storage bandwidth limited to ~20 GByte/s

**How to reduce the data rate? Trigger?!?**

# ALICE Upgrade

**Physics goals** (Michael Webers Talk on Tuesday)

Heavy Flavor
Quarkonia
Low-mass dielectrons
Jets
Anti- and Hypernuclei




**Focus on low $p_T$ probes**
many have low S/B:
➪ **classical trigger/event filter approach not efficient**

ALICE

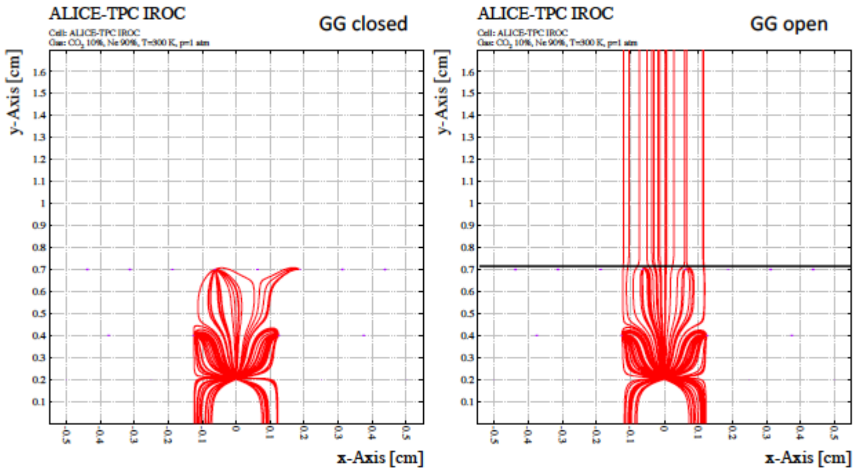
# Why not triggering?

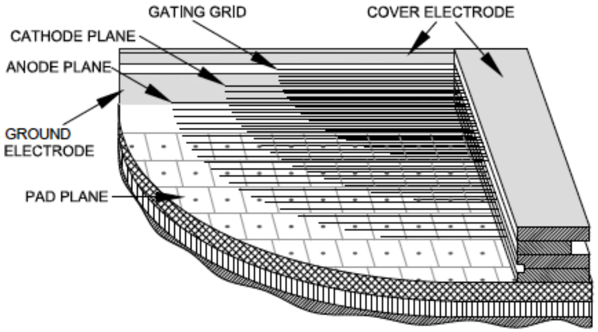
| Particle | Eff | S/ev | $S/B$ | $B'$/ev | trigger rate (Hz) | $S/\text{nb}^{-1}$ |
|---|---|---|---|---|---|---|
| $D^0$ | 0.02 | $1.6\cdot10^{-3}$ | 0.03 | 0.21 | $11\cdot10^3$ | $1.3\cdot10^7$ |
| $D_s^+$ | 0.01 | $4.6\cdot10^{-4}$ | *0.01* | 0.18 | $9\cdot10^3$ | $3.7\cdot10^6$ |
| $\Lambda_c$ | 0.01 | $1.4\cdot10^{-4}$ | $5\cdot10^{-5}$ | 11 | $5\cdot10^4$ | $1.1\cdot10^6$ |
| $\Lambda_c$ ($p_t > 2$ GeV$/c$) | 0.01 | $0.8\cdot10^{-4}$ | 0.001 | 0.33 | $1.6\cdot10^4$ | $0.6\cdot10^6$ |
| $B\to D^0(\to K^-\pi^+)$ | 0.02 | $0.8\cdot10^{-4}$ | 0.03 | $11\cdot10^{-3}$ | $5\cdot10^2$ | $0.6\cdot10^6$ |
| $B\to J/\psi(\to e^+e^-)$ | 0.1 | $1.3\cdot10^{-5}$ | *0.01* | $5\cdot10^{-3}$ | $3\cdot10^2$ | $1\cdot10^5$ |
| $B^+\to J/\psi K^+$ | *0.01* | $0.5\cdot10^{-7}$ | *0.01* | $2\cdot10^{-5}$ | 1 | $4\cdot10^2$ |
| $B^+\to \overline{D}^0\pi^+$ | *0.01* | $1.9\cdot10^{-7}$ | *0.01* | $8\cdot10^{-5}$ | 4 | $1.5\cdot10^3$ |
| $B_s^0\to J/\psi\phi$ | *0.01* | $1.1\cdot10^{-8}$ | *0.01* | $4.4\cdot10^{-6}$ | $2\cdot10^{-1}$ | $9\cdot10^1$ |
| $\Lambda_b(\to\Lambda_c+e^-)$ | *0.01* | $0.7\cdot10^{-6}$ | *0.01* | $2.8\cdot10^{-4}$ | 14 | $5\cdot10^3$ |
| $\Lambda_b(\to\Lambda_c+h^-)$ | *0.01* | $0.7\cdot10^{-5}$ | *0.01* | $2.8\cdot10^{-3}$ | $1.4\cdot10^2$ | $5\cdot10^4$ |

Triggering on $D^0$, $D_s$ and $\Lambda_c$ ($p_T$>2 Gev/c)

➡ **~ 20-25kHz@50kHz rate...**

# ALICE TPC Upgrade




Ions from the amplification region require finite drift time to reach the gating grid

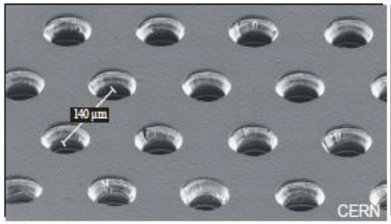
With current ALICE TPC gas:

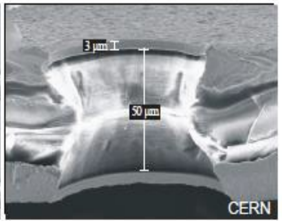
~100 μs drift + 180 ~μs gating grid closing time

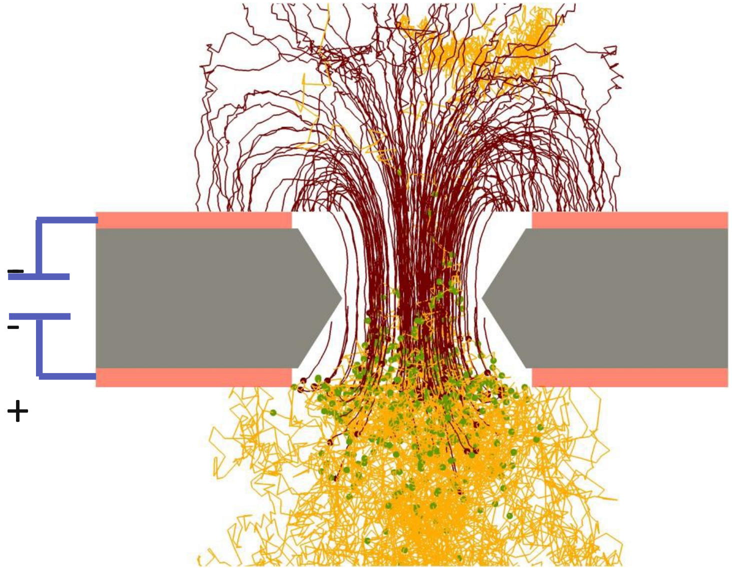
**Rate limited to ~3.5 kHz**

Without gating grid: space-charge distortions O(1m) to be compared with required position resolution O(100μm)

# ALICE TPC Upgrade




GEM: Gas Electron Multiplier

copper – kapton – copper sandwich (~50µm) with holes etched into it

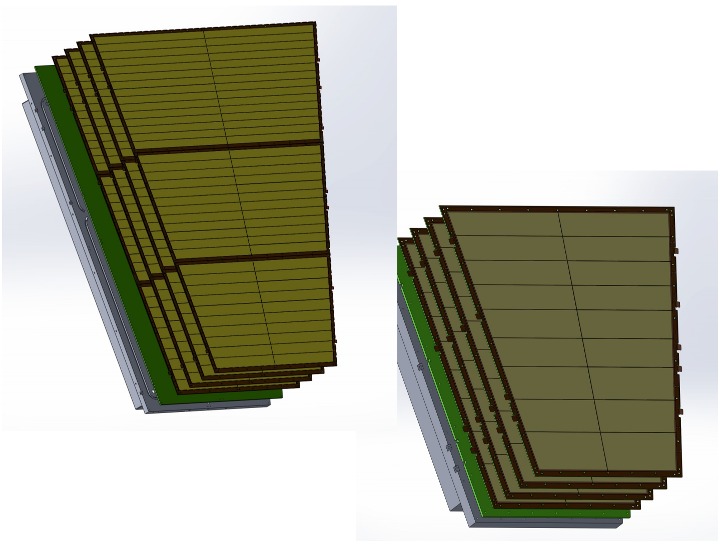
large field strength inside holes, sufficient for avalanche creation (gas amplification)

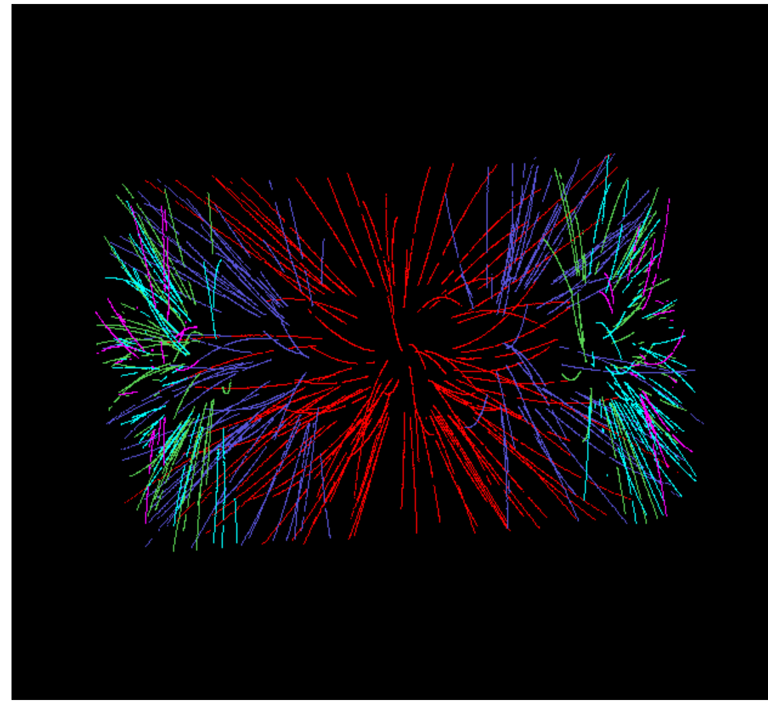
fast negative signal (new electronics)

**asymmetric field configuration features intrinsic ion blocking**

# ALICE TPC Upgrade

New read-out chambers based on 4 GEM layer setup





Operated in continuous mode: self triggered electronic
At 50kHz: on average 5 events in TPC drift time of ~100 µs
-> Factor 5 in data volume for online systems to read-out and process

# Requirements

Sample full 50kHz Pb-Pb interaction rate
(current limit at ~500Hz, factor 100 increase)

Typical event size of PbPb collisions@5.5TeV: 22 MByte
⇨ ~1.1 TByte/s detector readout
⇨ ~500 PByte/HI period (1 month)

*However:*

- storage bandwidth limited to ~20 GByte/s
- many physics probes have low S/B
  and event overlap in TPC:
  *classical trigger/event filter approach not efficient*

… and all this data has to be reconstructed

# Strategy

Data reduction by online reconstruction

Store only reconstruction results, discard raw data

- Demonstrated with TPC data since Pb-Pb 2011
- Optimized data structures for lossless compression
- Algorithms designed to allow for “offline” reconstruction passes with improved calibrations

⇨ Implies much tighter coupling between online and offline computing systems

# TPC Data 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 |

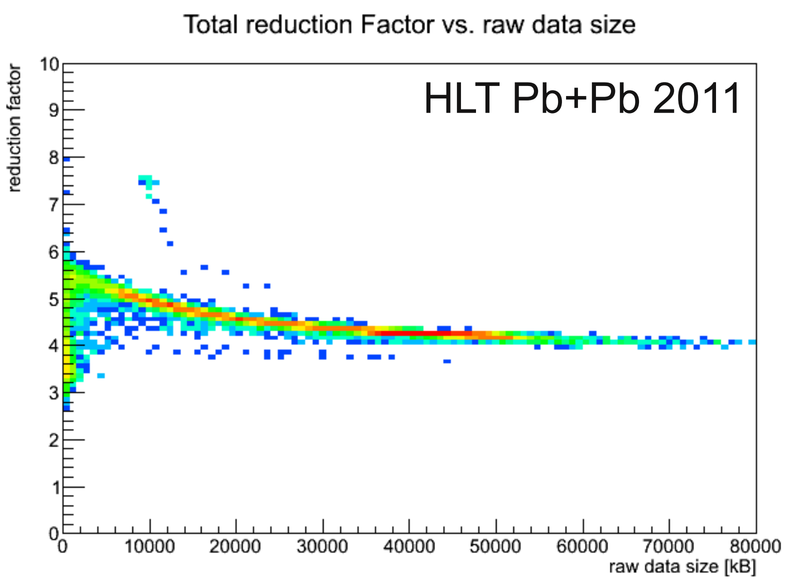
First steps up to clustering on the FPGA of the detector link receiver
Further steps require full event reconstruction, pattern recognition requires only coarse online calibration

# TPC Data Reduction

First compression steps used in production starting with the 2011 Pb+Pb run

Online found TPC clusters are basis for offline reconstruction

Currently R&D towards using online found TPC tracks to complement offline seed finding and online calibration

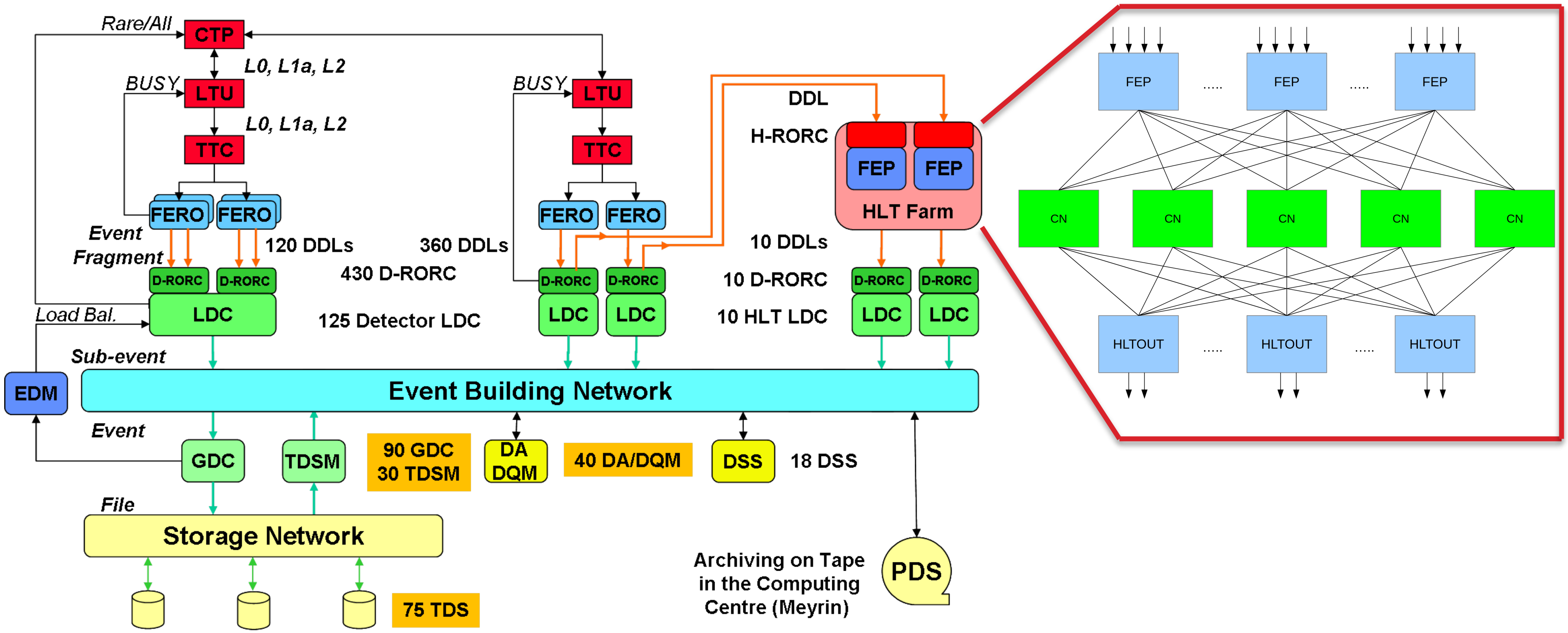


# Data Bandwidth

| 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** |

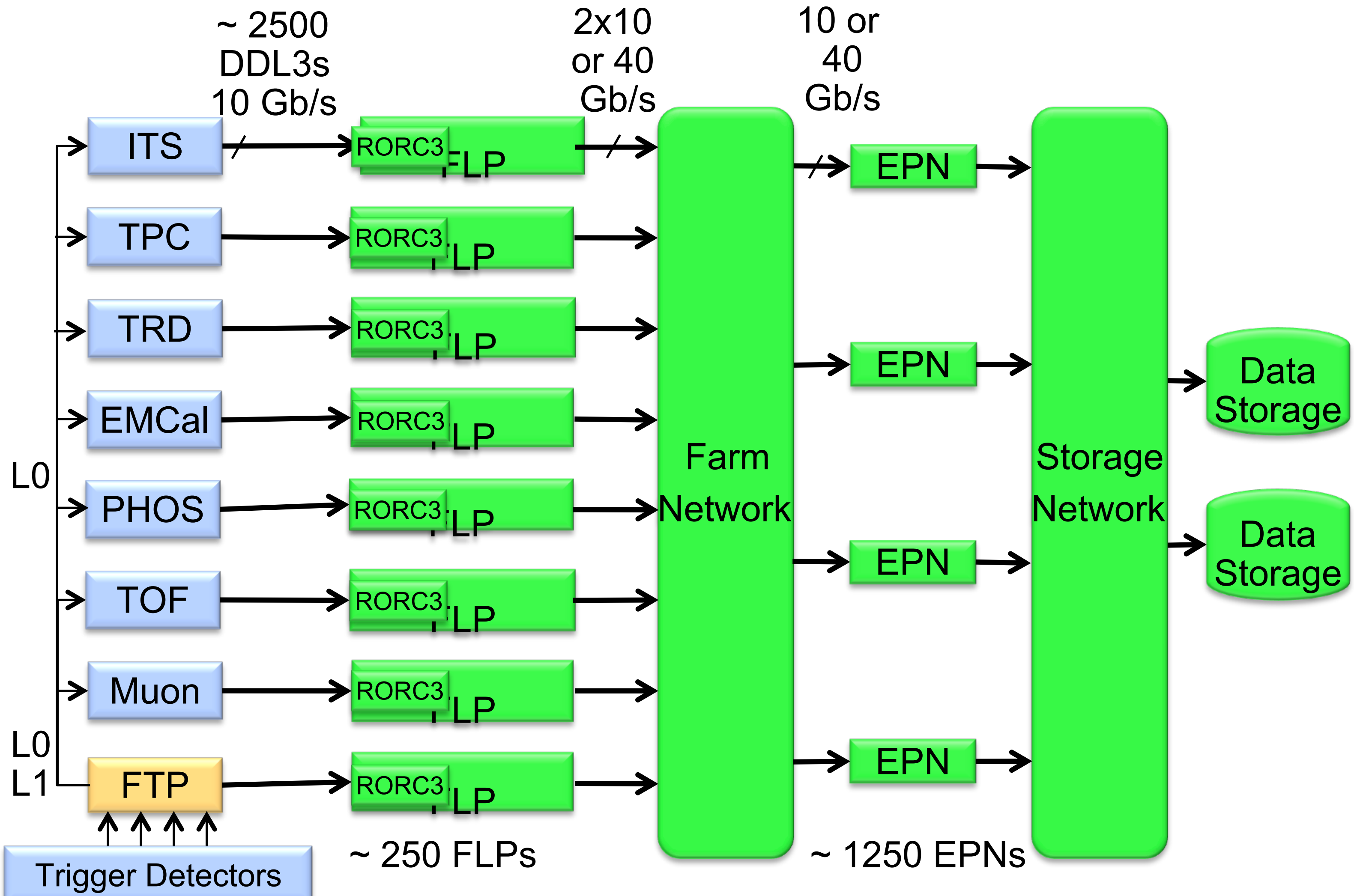
LHC luminosity variation during fill and efficiency taken into account for average output to computing center.

# Current Online Systems

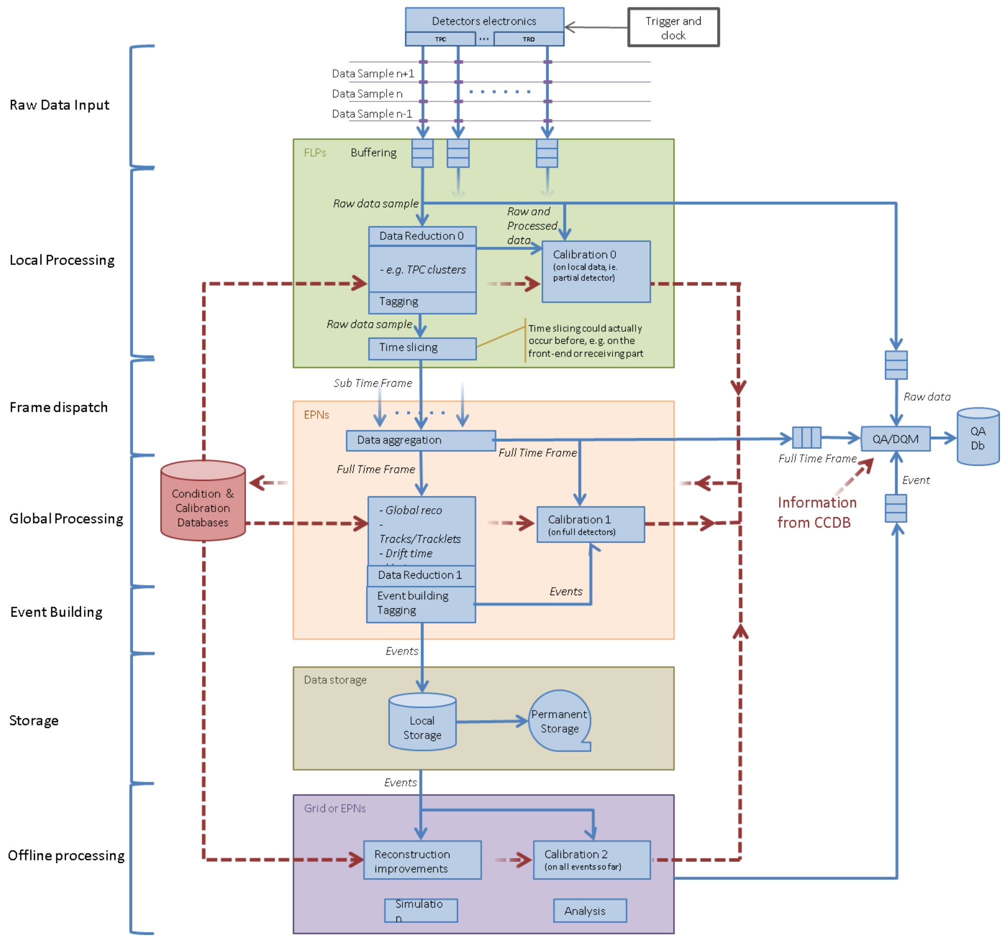



Different technologies/techniques used in DAQ/HLT
e.g. Ethernet <-> Infiniband

# Combined DAQ/HLT System



# O² System Dataflow



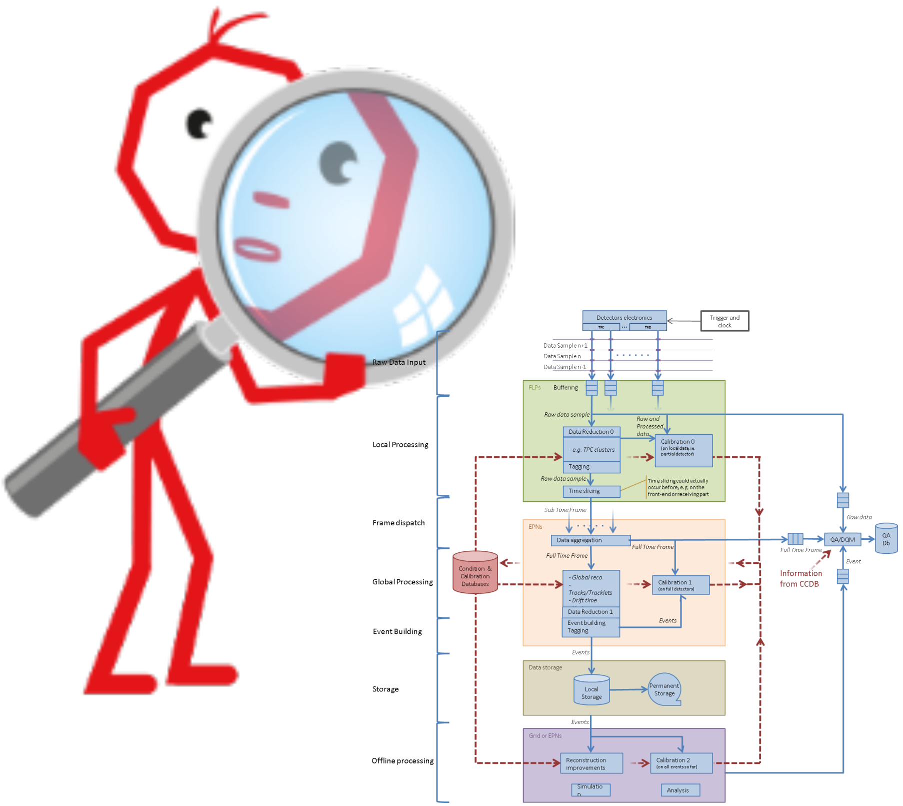

**Main functionality blocks**

- Data input
- Data reduction
- Storage
- Reconstruction
- Simulation
- Analysis

- Networking – Data transport
- Calibration
- Condition and calibration database
- Data monitoring (QA/DQM/etc)
- Software framework

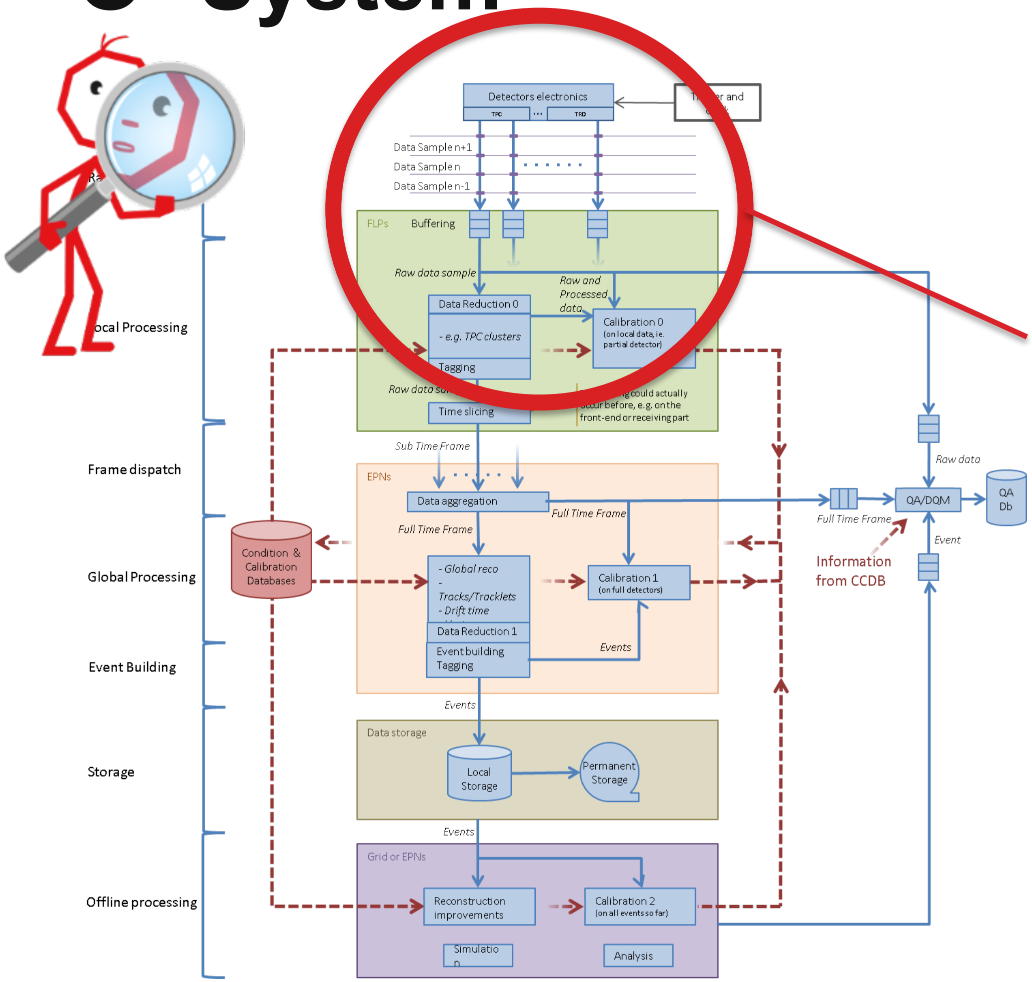
# O² System

A closer look at selected parts of the system…



ALICE

# O² System




**Data Input**

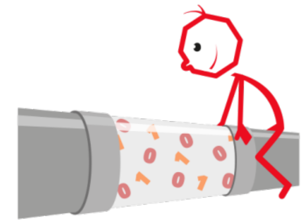
*Detector*
*Trigger*
*CRU*

*Outside of O² project*

Handle > 1TByte/s Input from continuous & triggered detectors

- Data Links
- Receiver Card
- Local Reconstruction

# Detector Readout



Combination of continuous and triggered readout

Continuous readout for TPC (and ITS)

- At 50 kHz, ~5 events in TPC during drift time of ~100 µs
  Continuous readout minimizes needed bandwidth
- Implies change from event granularity in the online systems to time-windows with multiple events
- Implies event building only after partial reconstruction

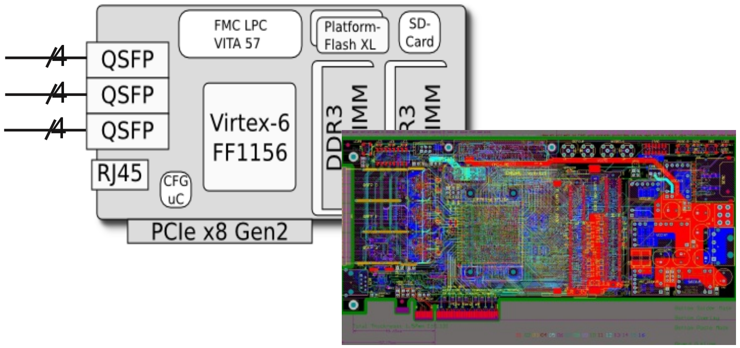
Fast Trigger Processor (FTP) complementing CTP

- Provides clock/L0/L1 to triggered detectors and TPC/ITS for data tagging and test purposes

# DDL/RORC Development

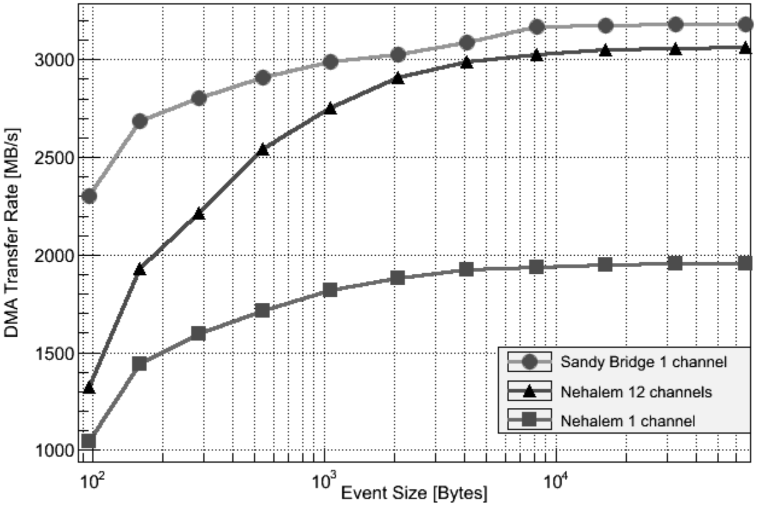
**Data Link**

DDL1 (Run 1): 2Gbit/s
DDL2 (Run 2): 6 Gbit/s
DDL3 (LS2): 10 Gbit/s

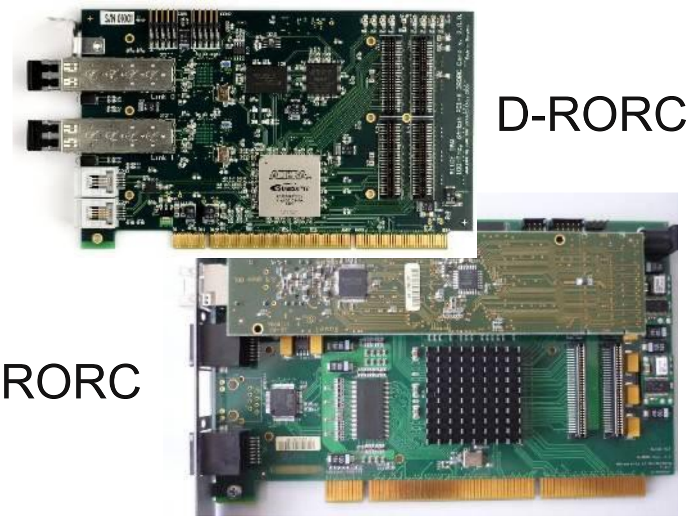



**Receiver Card (FPGA)**

- RORC1 (now)
  2 DDL1, PCI-X&PCIe Gen1x4
- RORC2 (being produced)
  12 DDL2, PCIe Gen2x8
- RORC3 (LS2)
  10-12 DDL3, PCIe Gen3



# Read-Out Receiver Card



D-RORC

H-RORC

C-RORC

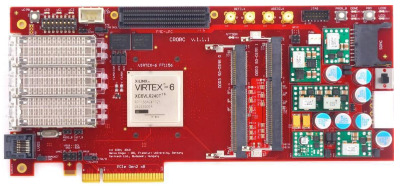
**Common Read-Out Receiver Card**

- mainly developed by IRI Frankfurt&Cerntech for HLT Run 2
- in production now, delivery this year

Increased link speed: **2 Gb/s (DDL1) -> 6 Gb/s (DDL2)**

Increased number of ports: 2 -> 12

# Run 3 Detector Readout



Different link protocols under investigation:
- DDL3 (custom, 10Gb/s)
- Ethernet (10 - 40 Gb/s)
- PCIe over cable (Gen2, Gen3; 16 - 128 Gb/s )
- GBT (3 – 4 Gb/s)

Large variation in link bandwidths

Number of links and FLPs depend upon decision about read-out implementation

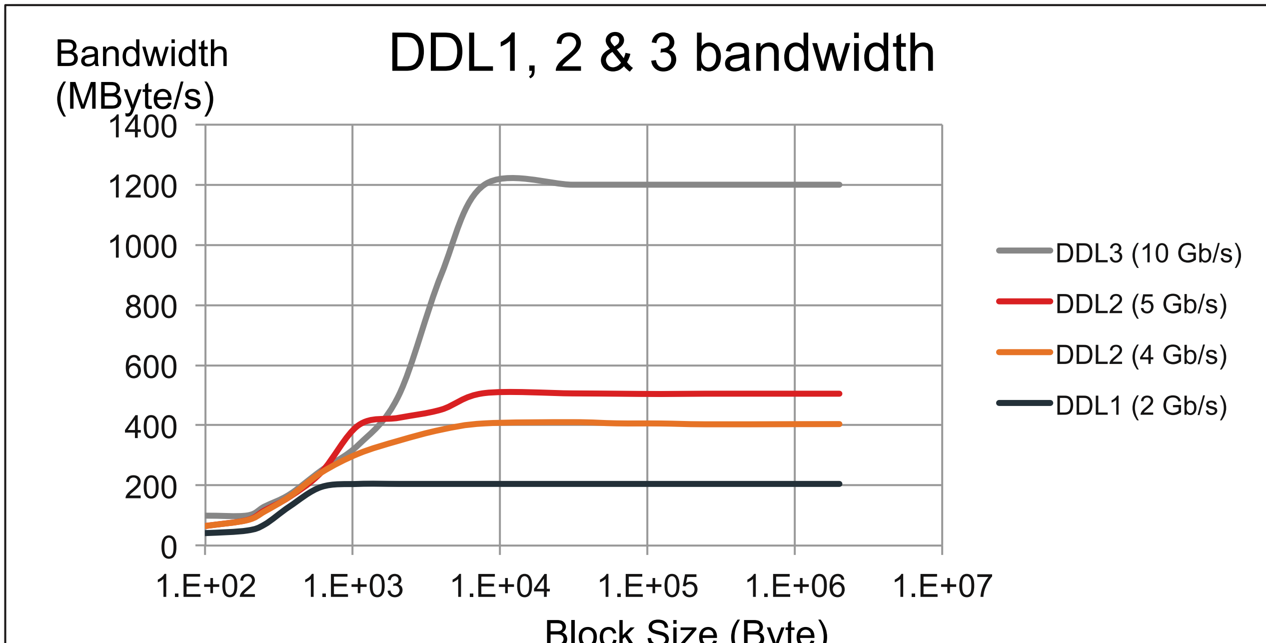
Data compression by co-processing (FPGA or other)

Run 1 and 2: combined in a custom card with DDL receiver

Run 3: could we split the dataflow and the data processing ?
Benchmark of memory bandwidth

S/W compression on FLP?

# DDL Performance Evolution

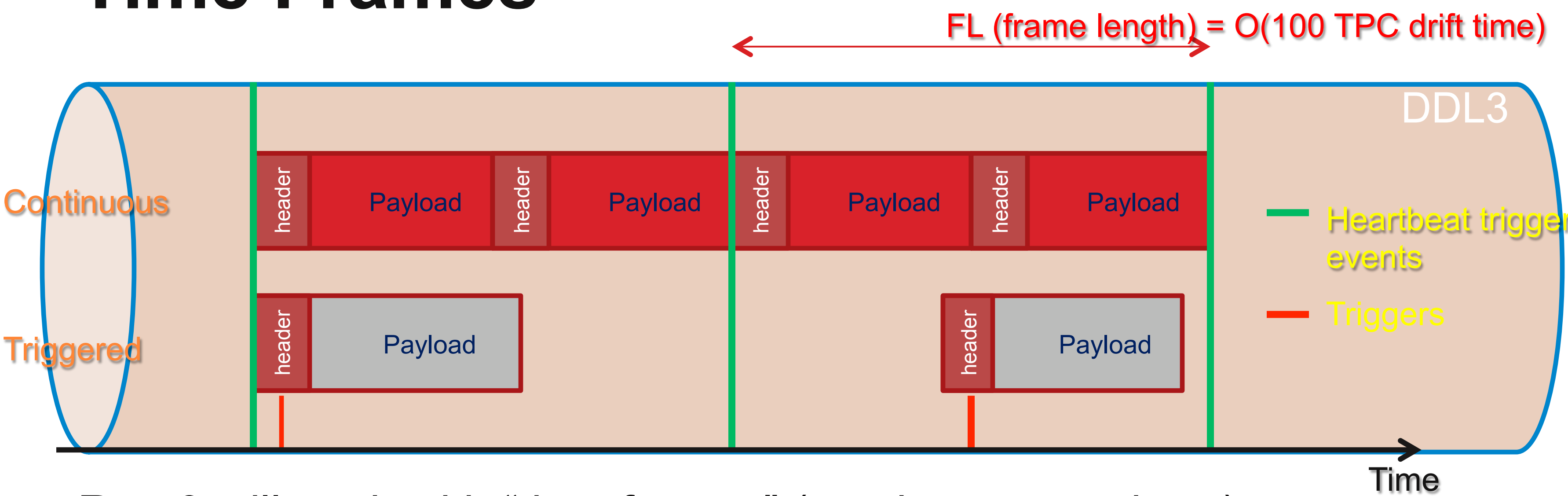



DDL2 at 4 and 5 Gb/s (according to needs) ready for Run 2

Prototype for one of the DDL3 option considered for Run 3 implemented (Eth. + UDP/IP)

Expected performance evolution verified

# Time Frames




Run 3 will work with “time frames” (continuous read-out)

No “events” in the online systems; definition during reconstruction

Defined by Heat-Beat Triggers

- Highest priority trigger in ALICE
- Defines boundary between Time Frames

# Time Frames

**Length of Time Frame/HB Interval**

100 µs TPC drift time determining constant

- Number events >> number events in "border"
  Number events >> 2*5 (@50 kHz)
- 1000 events@50 kHz ≙ 20ms … or even more? 100ms?

Note that Time Frame Rate will be O(1kHz) – *not a high rate system*

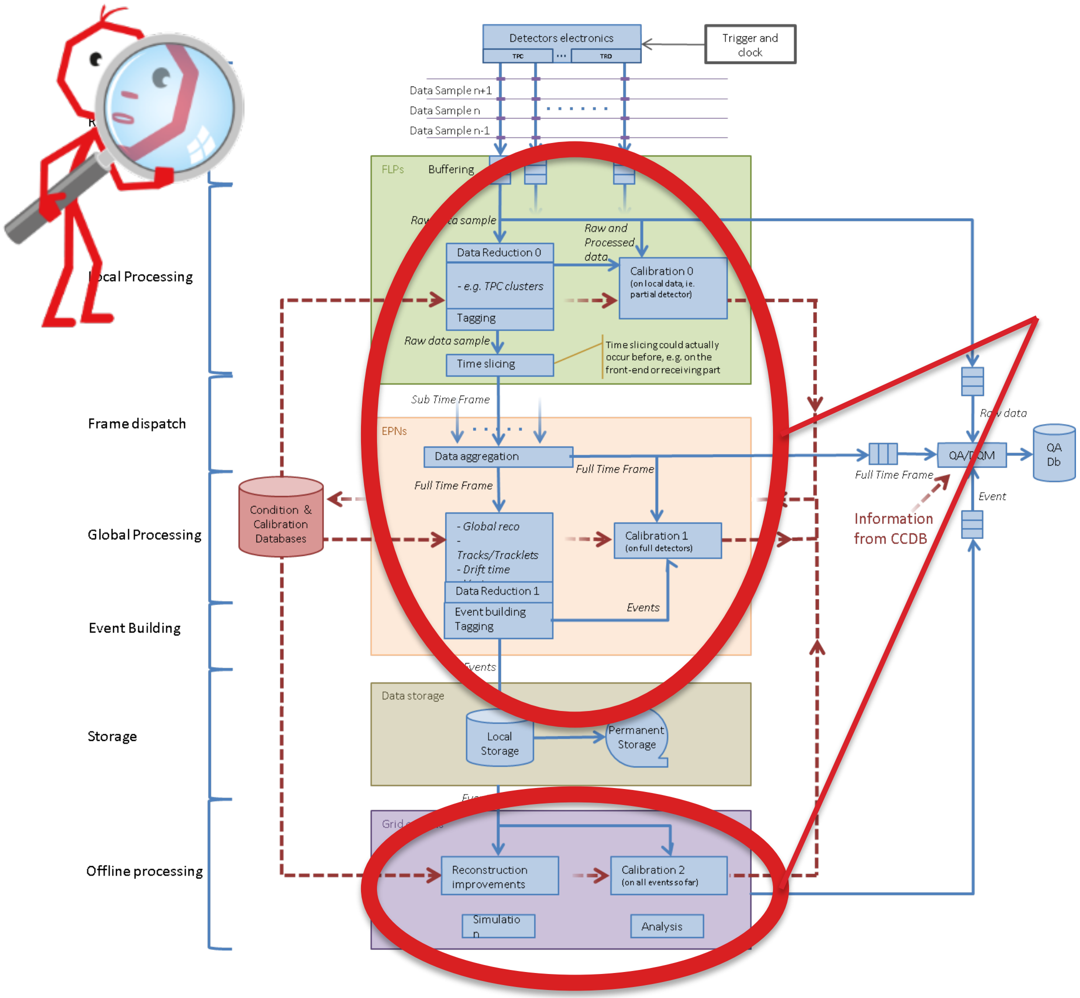
**Limiting factors**

Data size: 1000 events@23 MByte = 23 GByte (w/o FLP comp…)

Data transport: network bandwidth/FLP buffers
avoid cross EPN data transfer/think in streams

ALICE

# $O^2$ System




**Networking and Data Transport**

Connect the different computing layers (FLP/EPN)

Transport of AODs for physics analysis from the O2 farm to the Grid

Data transfer between grid sites

ALICE

# Network

**Requirements**

Total number of nodes: ~1500
FLP Node Output: up to 12 Gbit/s
EPN Node Input: up to 7.2 Gbit/s
EPN Output: up to 0.5 Gbit/s

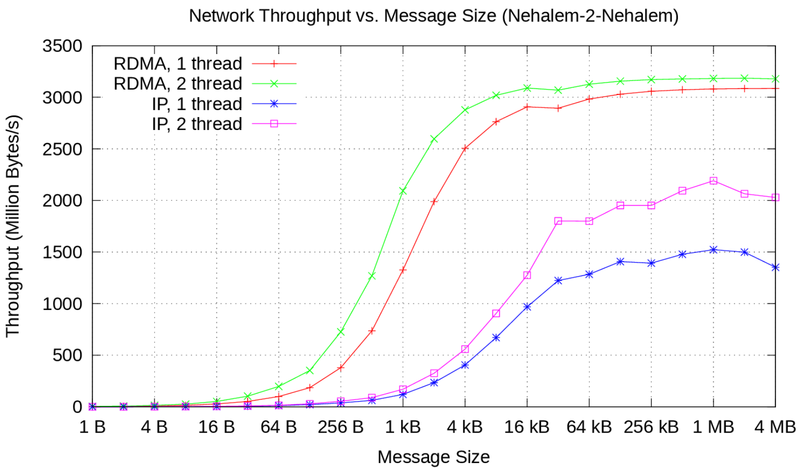
Two technologies available

- 10/100Gbit Ethernet (currently used in DAQ)
- QDR/FDR Infiniband (40/52Gbit, used in HLT)

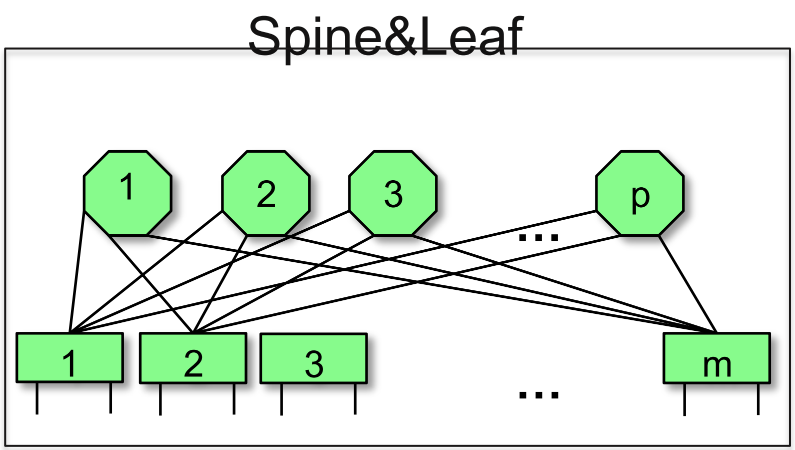
Both would allow to construct a network satisfying the requirements even today

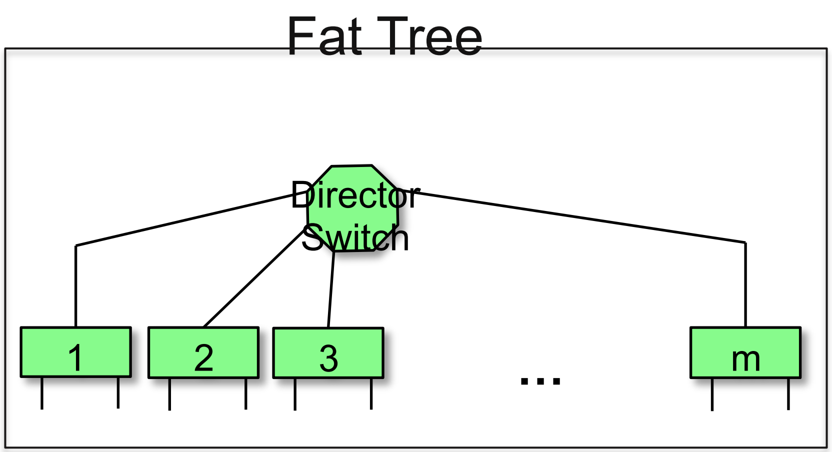
# Network

Throughput tests

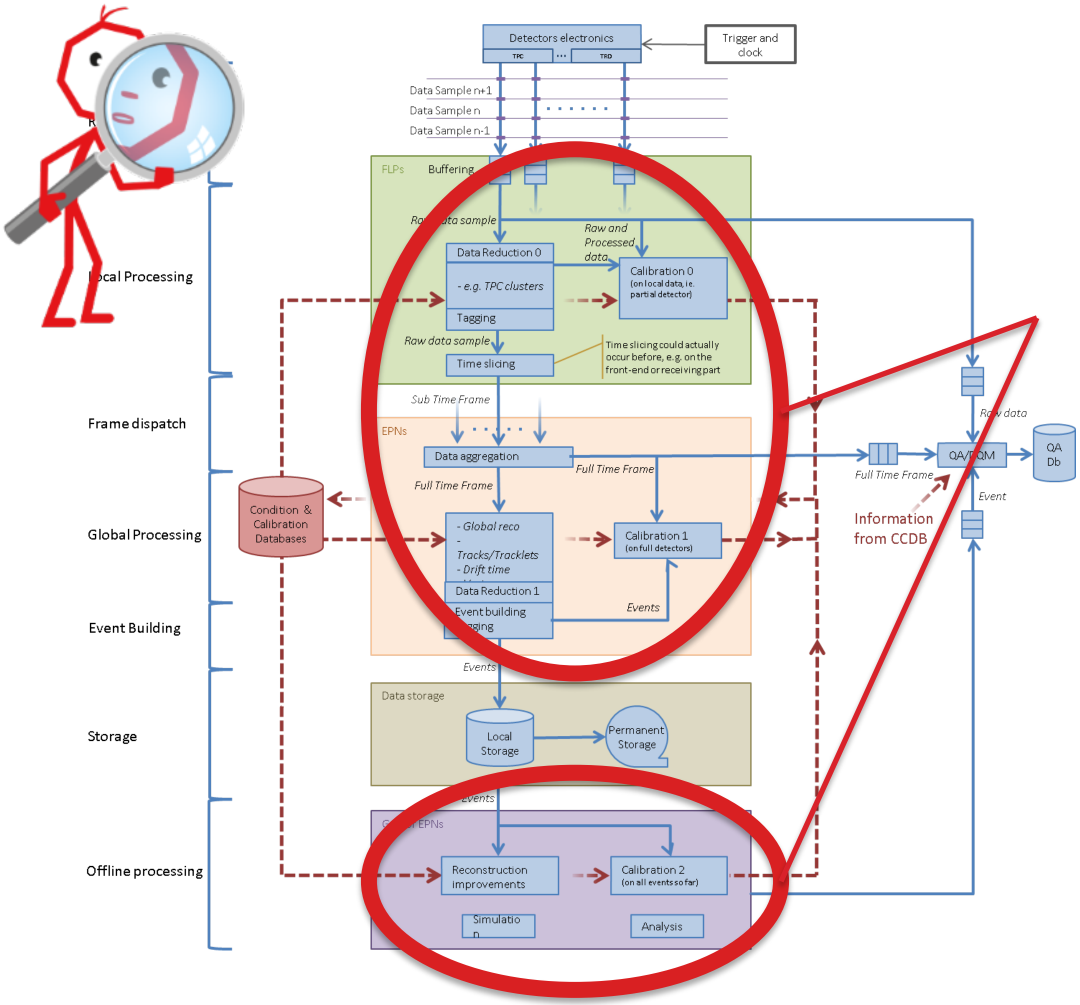



Different topologies under study to
- minimize cost
- optimize failure tolerance
- cabling



# O² System




**Processing Layer**

Full event reconstruction

- for local storage (data reduction)
- for physics analysis

ALICE

# Processing Power

Estimate for online systems based on
current HLT processing power

- ~2500 cores distributed over 200 nodes
- 108 FPGAs on H-RORCs for cluster finding
  1 FPGA equivalent to ~80 CPU cores
- 64 GPGPUs for tracking (NVIDIA GTX480 + GTX580)

Scaling to 50 kHz rate to estimate requirements

- ~ 250.000 cores
- additional processing power by FPGAs + GPGPUs

⇨1250-1500 nodes in 2018 with multicores

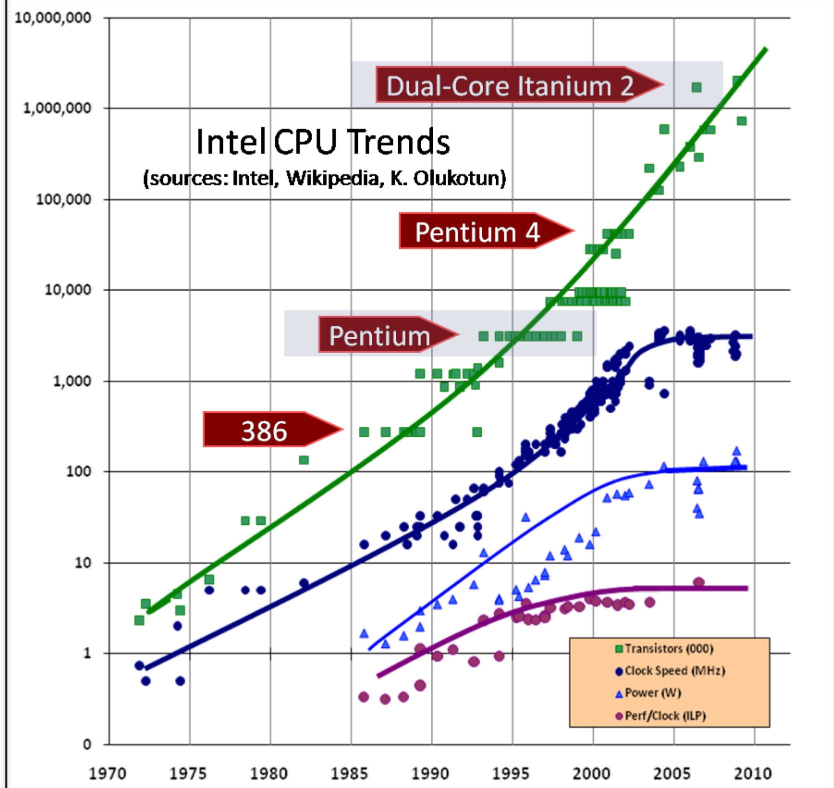
ALICE

# Processing Power

Estimate of processing power based on scaling by Moore's law

However: no increase in single core clock speed, instead multi/many-core

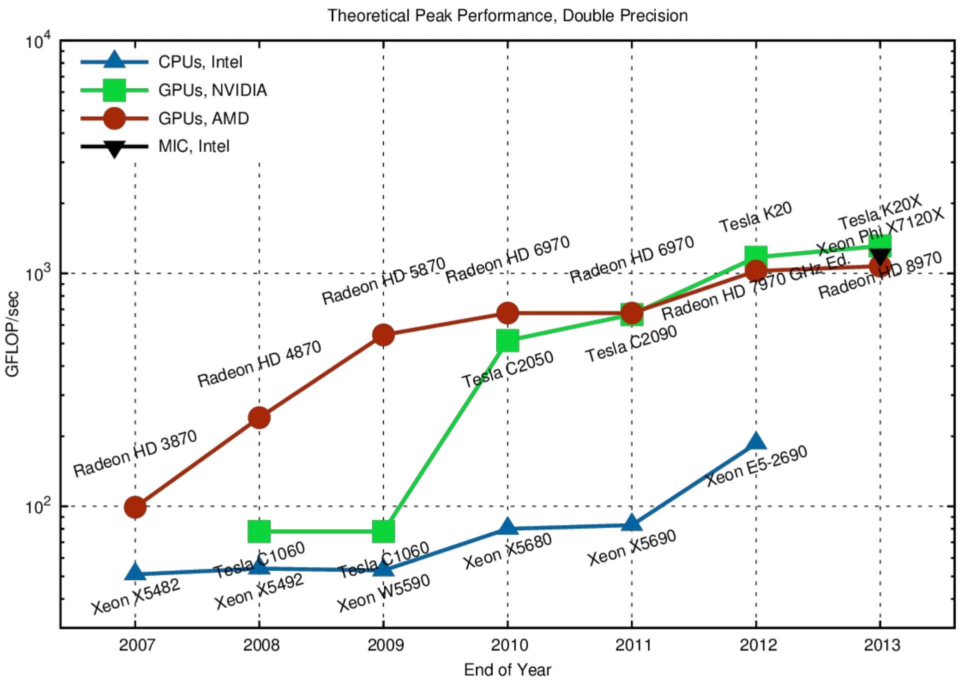
Reconstruction software needs to adapt to full use resources




Picture from Herb Sutte: The Free Lunch Is Over
A Fundamental Turn Toward Concurrency in Software
Dr. Dobb's Journal, 30(3), March 2005 (updated)

# GPUs for General Purpose Computing

Driven by (theoretical) peak performance
GPU: O(1) TFLOP/s (NVIDIA TESLA K20: 3.2 TFLOP/s)
CPU: O(0.1) TFLOP/s (Intel Xeon E5-2690 : 243 GFLOP/s)

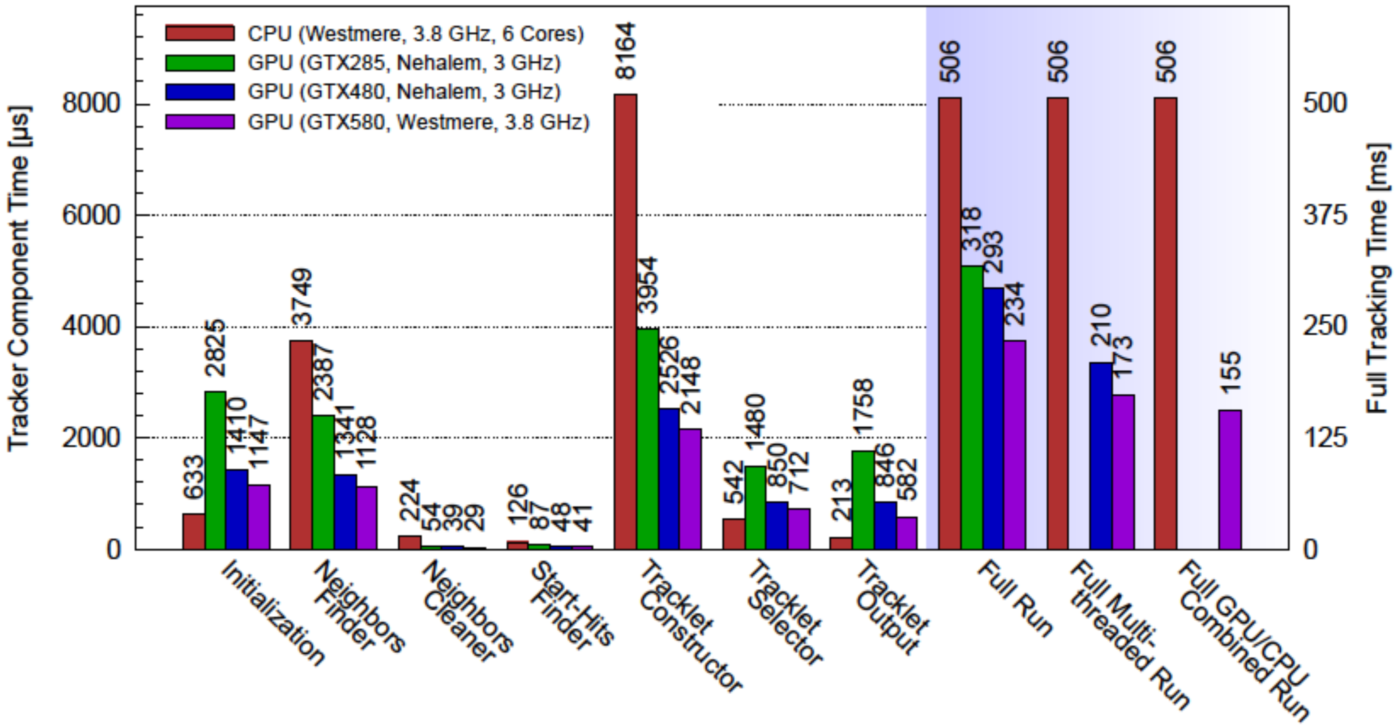



Can this theoretical peak performance be used **efficiently** for the typical HEP workload?

ALICE

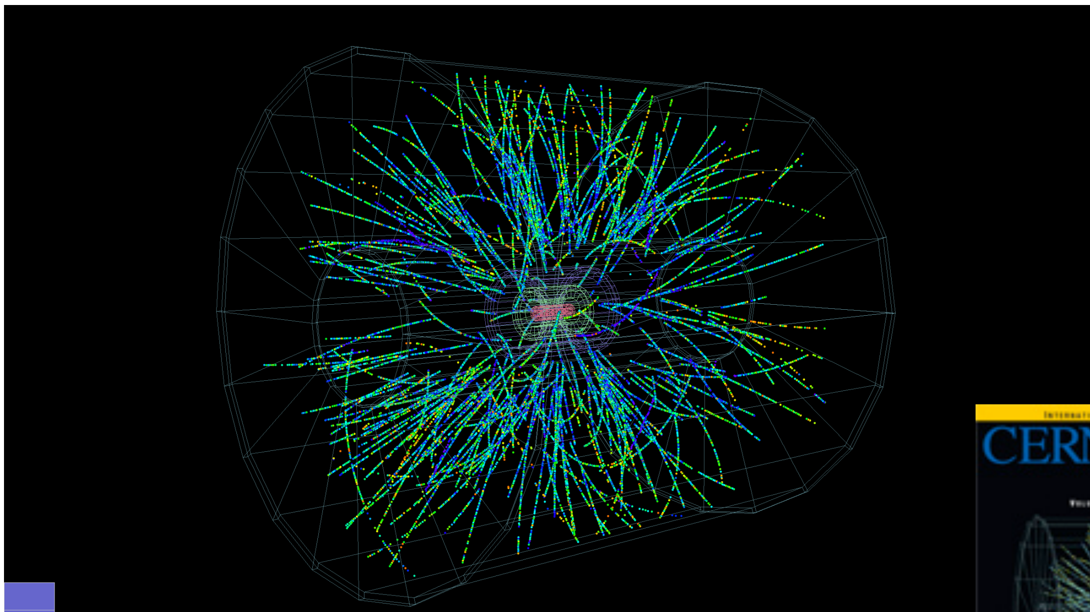
# Parallel Reconstruction

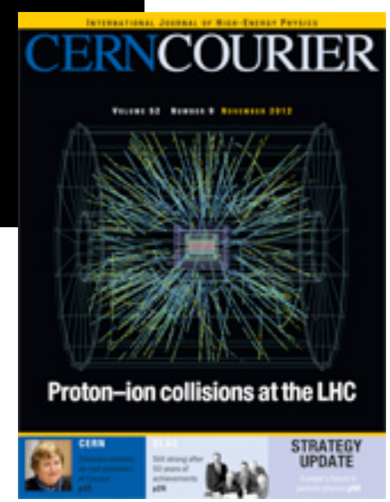
Tracking most time-consuming step in ALICE
Developed multi-threaded tracking algorithm for the HLT
Also adopted to GPUs (NVIDIA Fermi, CUDA)



ALICE

# Online Reconstruction

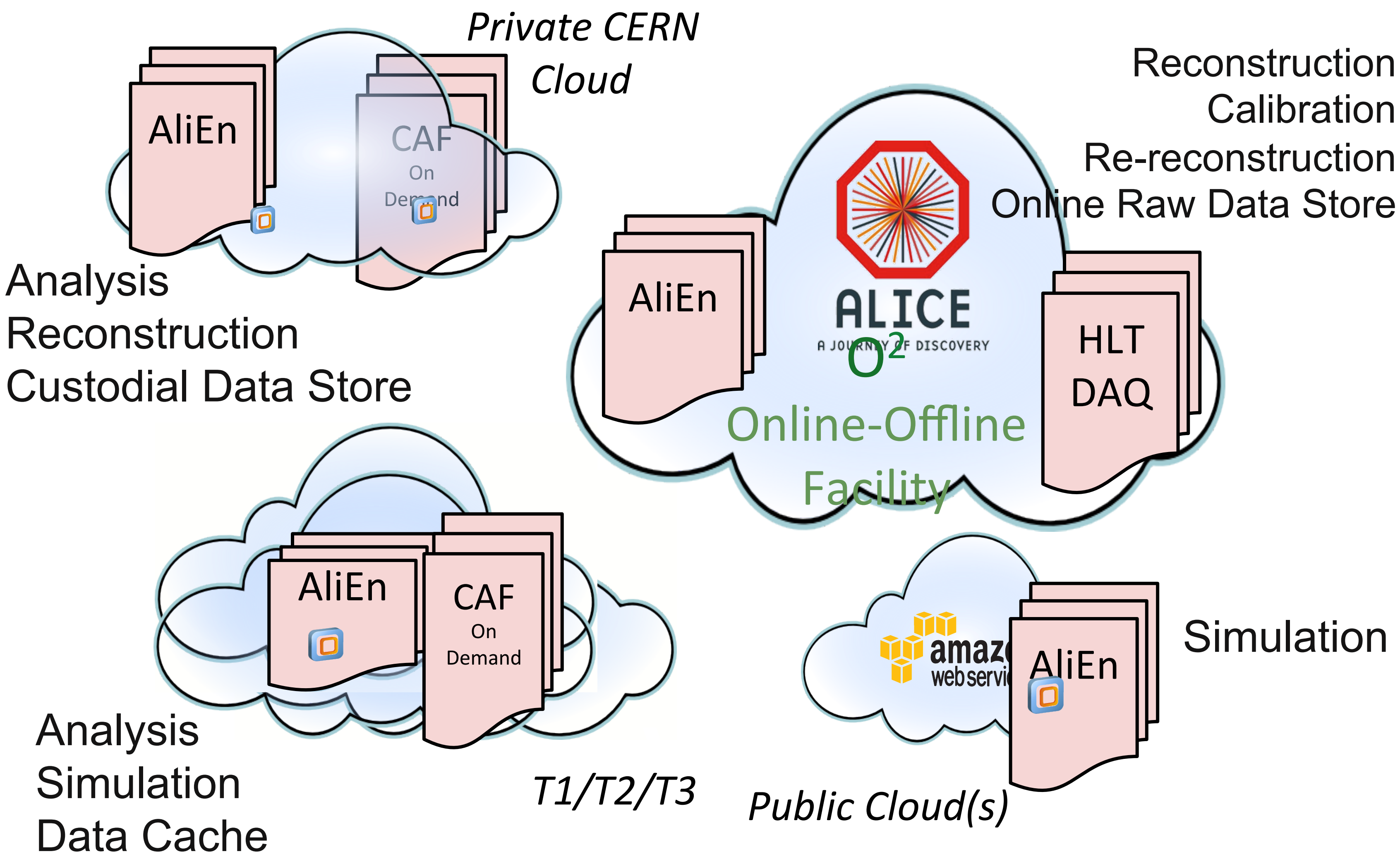





Full GPU tracking available in the HLT since Pb+Pb 2011; based on CA algorithm

Comparable efficiency, fake and clone rates to offline code, order of magnitude faster

# Computing Model



# Summary

ALICE physics program after 2018 requires handling of 50kHz minimum-bias Pb-Pb collisions (1TByte/s) from the online and offline systems

Strategy to handle the load is an ambitious data volume reduction by a first pass online reconstruction & discarding of raw data on a combined DAQ/HLT/offline farm ($O^2$)

Raw data will be stored locally on farm, subsequent reconstruction passes for physics will run there

R&D towards the Online&Computing TDR at end of 2014

**An interesting future ahead…**