A Large Ion Collider Experiment

ALICE

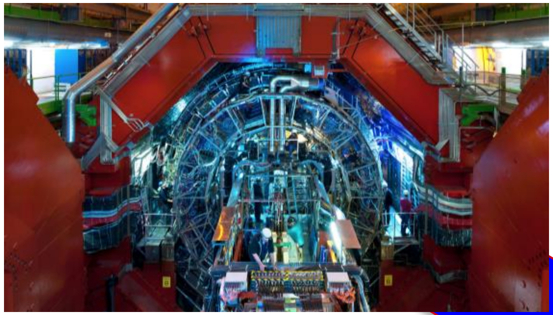
ALICE Upgrade: O^2
Processing Challenges

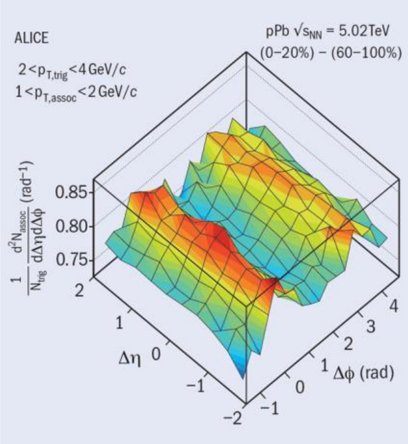
Thorsten Kollegger

FIAS Frankfurt Institute for Advanced Studies

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

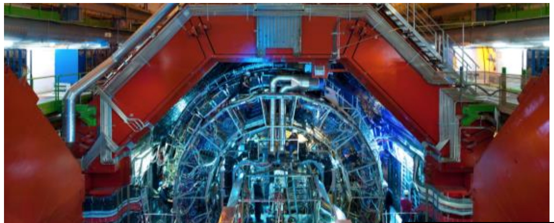
O^2 Project



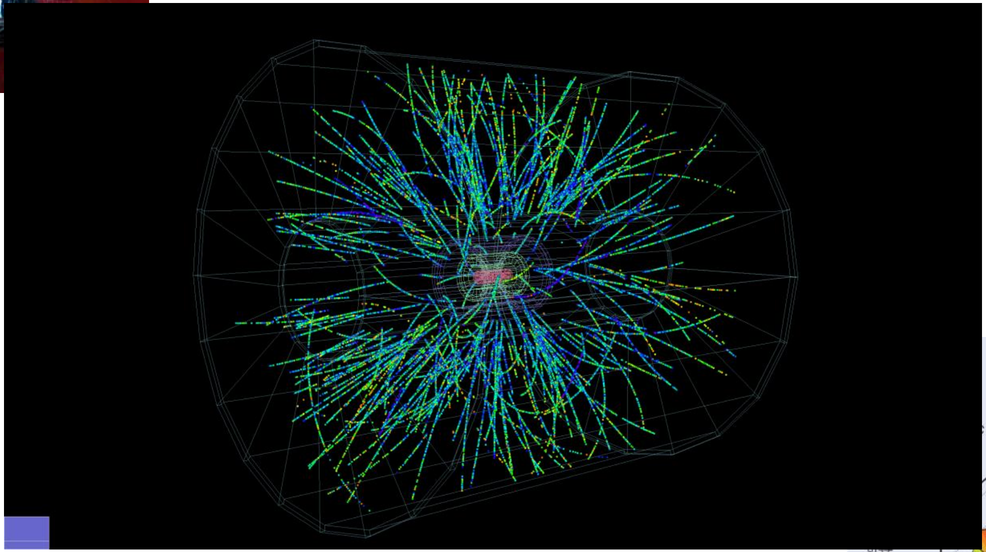


From Detector Readout to Analysis:
What is the “optimal” computing architecture?

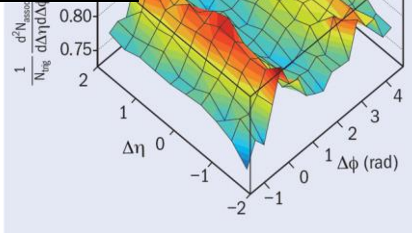
O² Project



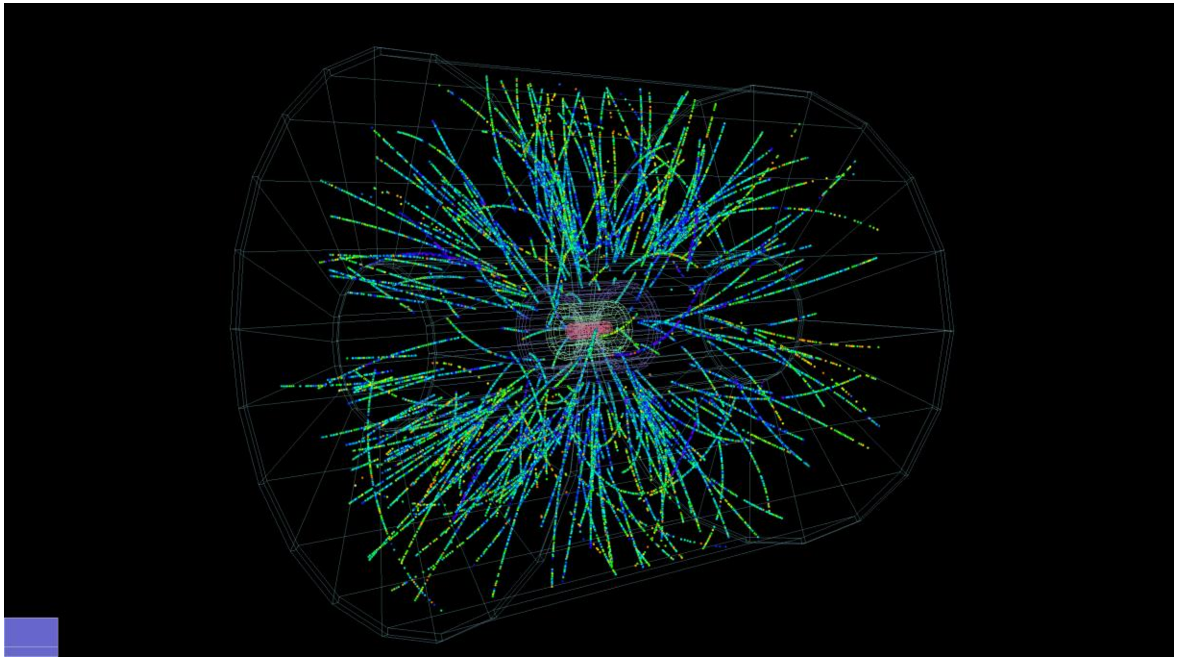
Event Reconstruction
(Online Farm & GRID/Cloud)



Physics Analysis
(Grid/Cloud)



O^2 Project



ALICE in 2018:

50.000 collision events per second, each ~20 MByte

> 1 TByte per second data input

Requirements

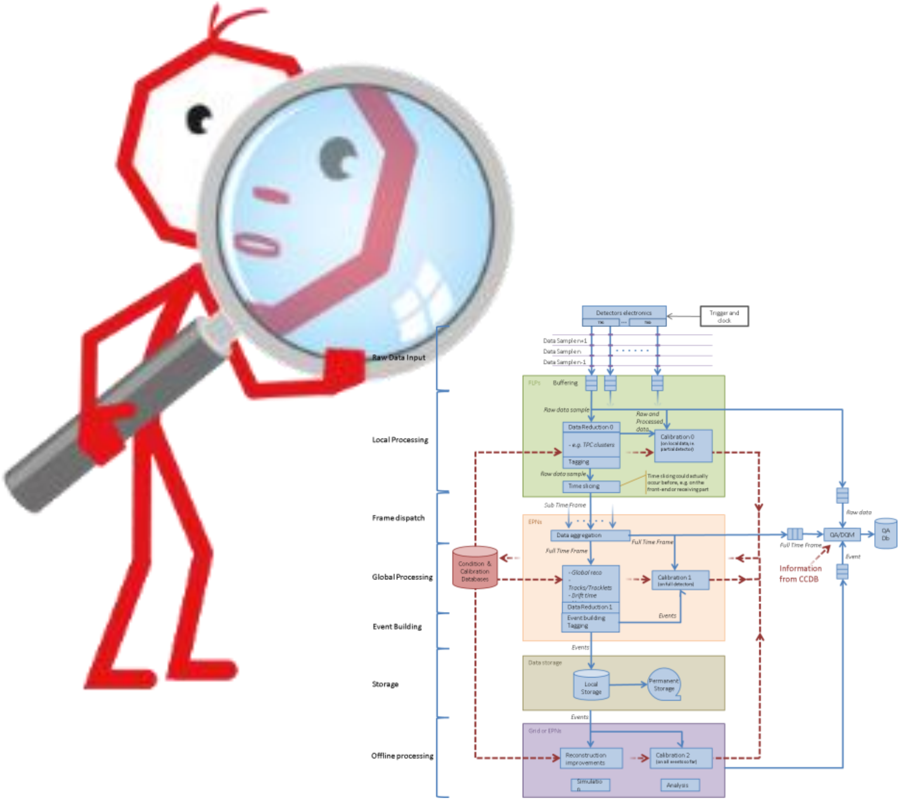
Focus of ALICE upgrade on physics probes requiring high statistics: sample 10 nb^{-1}

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

⇨ ~1.1 TByte/s detector readout
However:
- storage bandwidth limited to ~20 GByte/s
- many physics probes have low S/B:
  classical trigger/event filter approach not efficient
  (N.B. trigger: selecting “interesting” events)

O^2 System

A closer look at selected parts of the system…



Strategy

~1.1 TByte/s detector readout

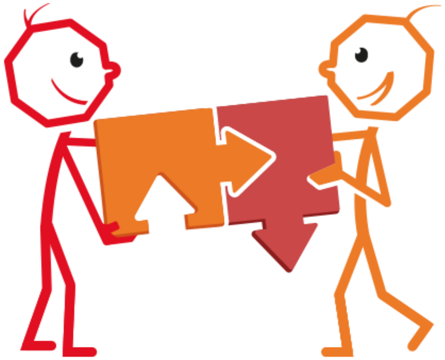
However:

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

Store only reconstruction results, discard raw data

Data reduction by (partial) online reconstruction and compression

⇨ Implies much tighter coupling between online and offline reconstruction software



O^2 System 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

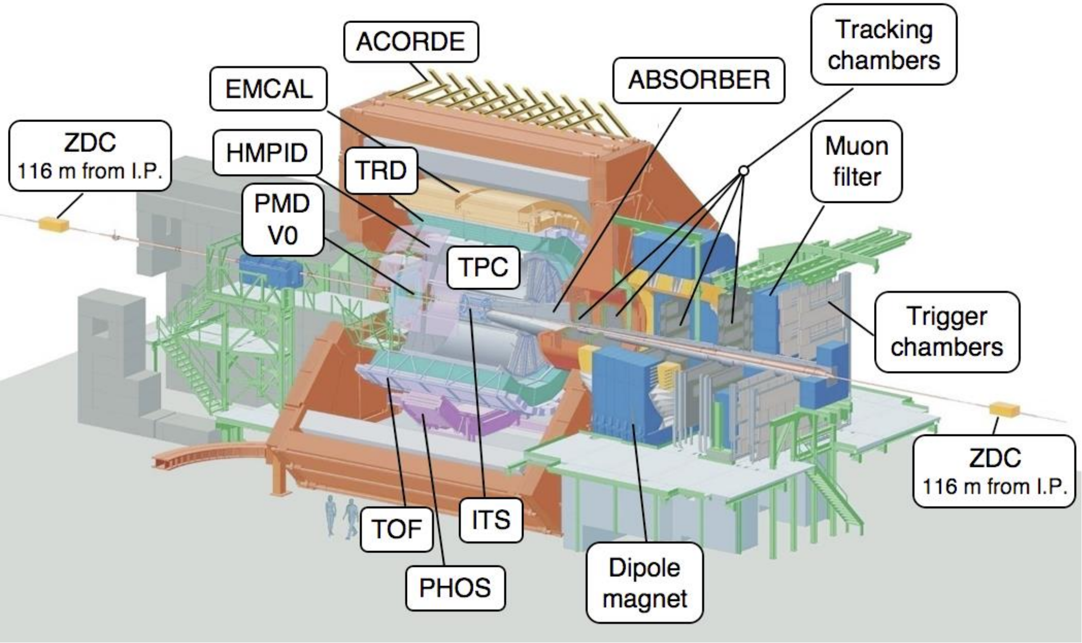
- ➪ Allow for reconstruction with improved calibration, e.g. store clusters associated to tracks instead of tracks
- ➪ Minimize dependence on calibration accuracy
- ➪ Implies "intermediate" storage format

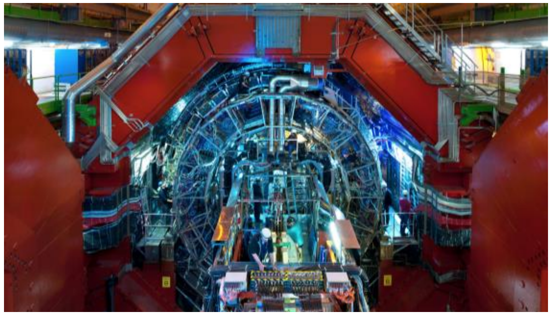
Keep cost "reasonable"

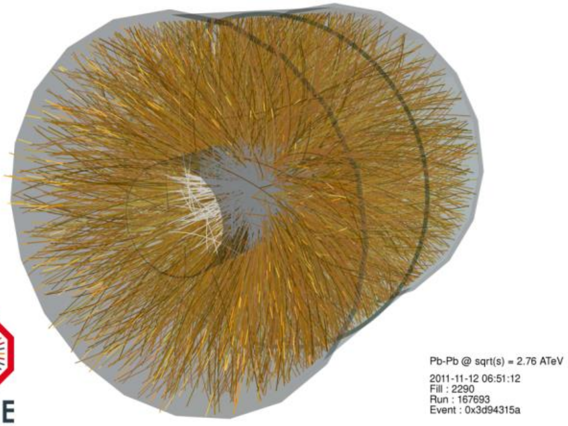
- ➪ Limit final storage system bandwidth
- ➪ Optimize computing capacity

"No" latency requirements & fault-tolerance

A Large Ion Collider Experiment





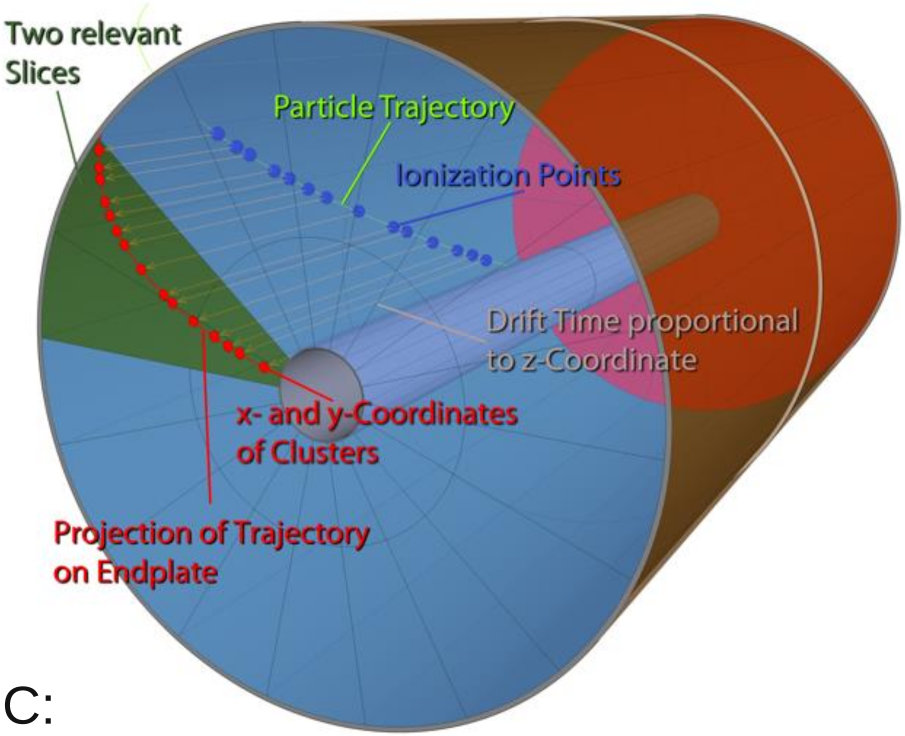


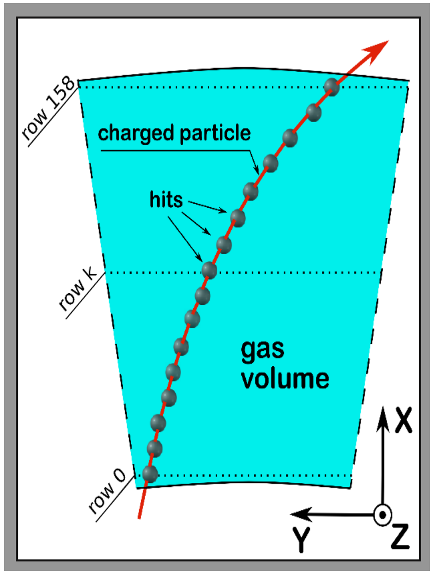
Expected Data Bandwidth

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

Time Projection Chamber

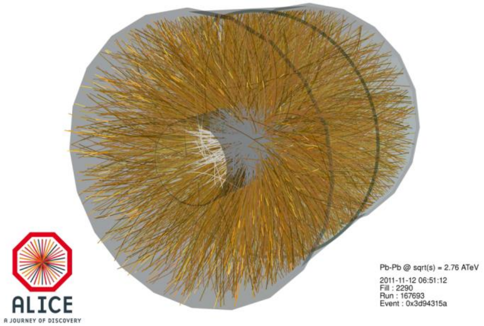




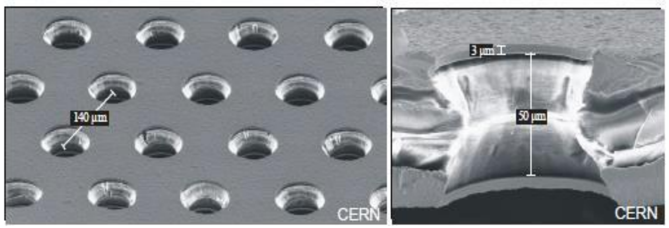
ALICE TPC:

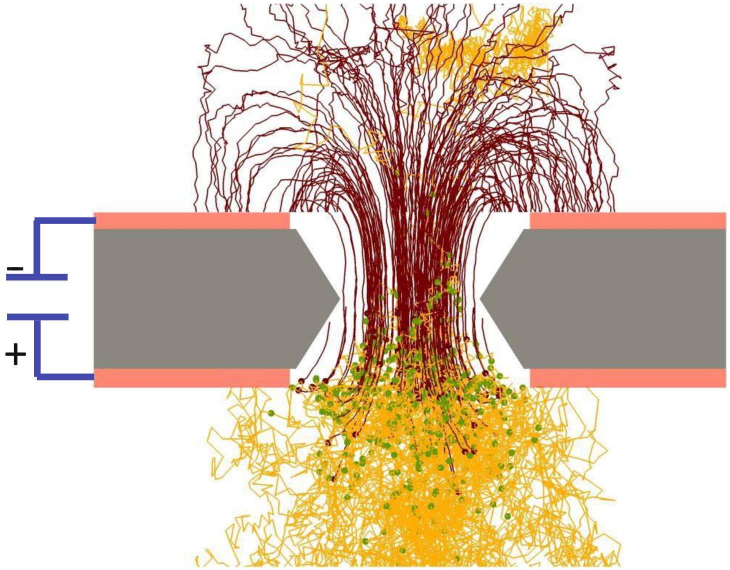
5 m diameter, 5m long

557.568 readout channels * 1000 time samples



ALICE TPC Upgrade





GEM: Gas Electron Multiplier

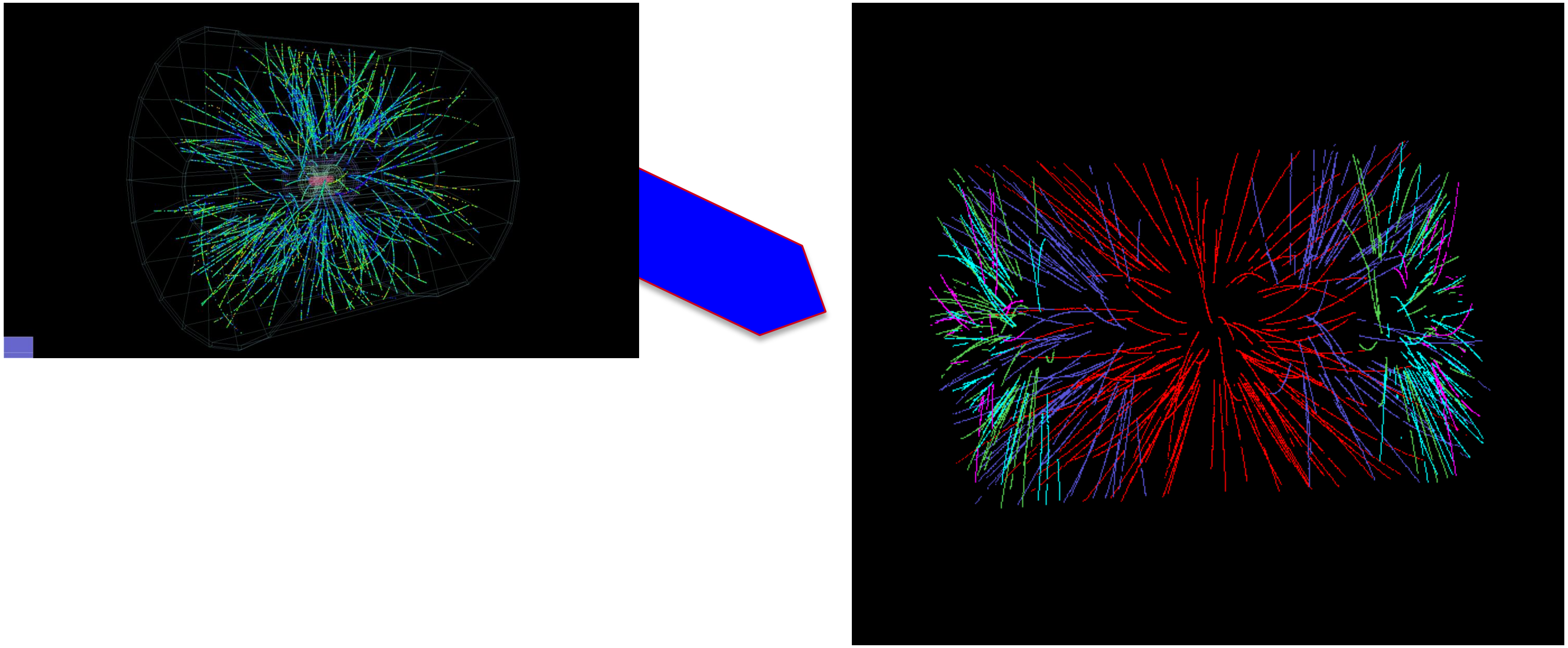
copper – kapton – copper sandwich (~50µm) with holes etched into it

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

fast negative signal (new electronics)

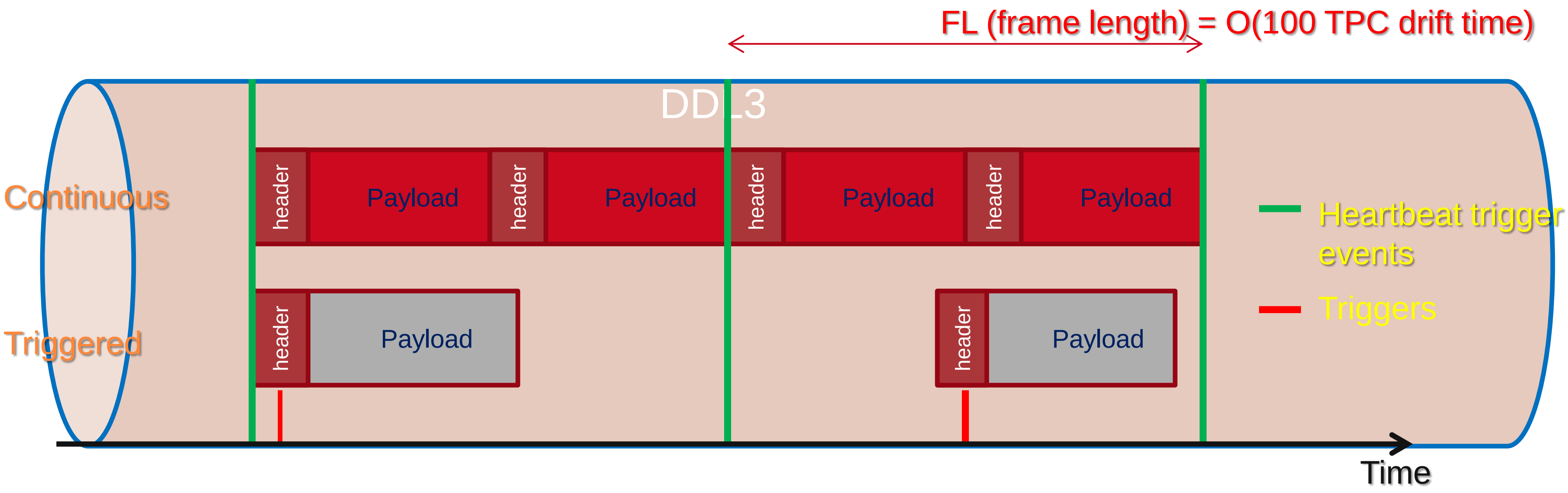
asymmetric field configuration features intrinsic ion blocking

ALICE TPC Upgrade



Operated in continuous mode: self triggered electronic
At 50kHz: on average 5 events in TPC drift time of ~100 µs

Time Frames



Run 1+2 triggered:

event based – one collision to analyze

Run 3+4 continuous readout:

will work with “time frames” – many collision to analyze

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)

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

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

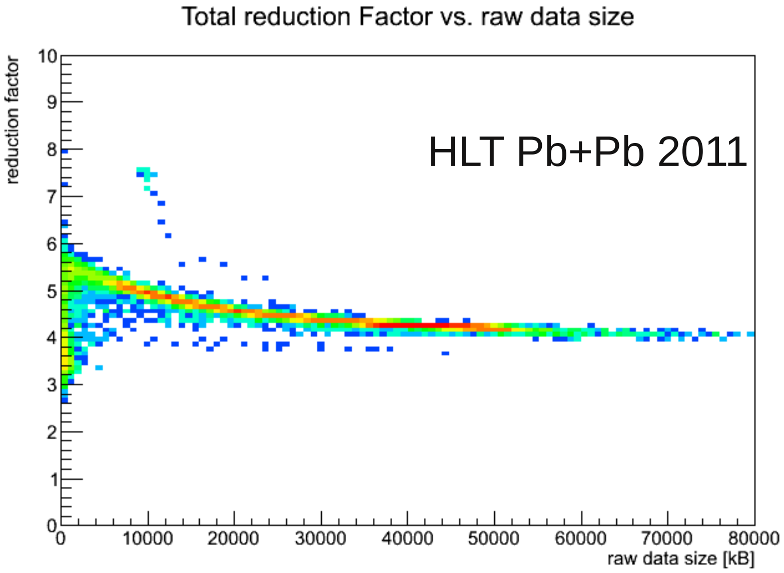
First steps up to clustering on the FPGA of the detector link receiver
Further steps require full event reconstruction, pattern recognition requires only coarse online calibration

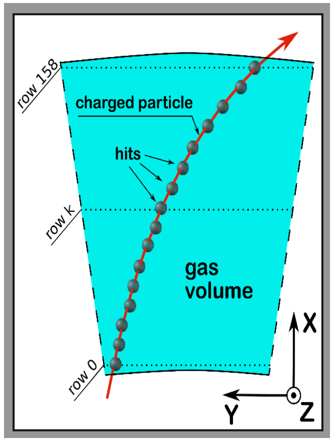
TPC Data Reduction

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



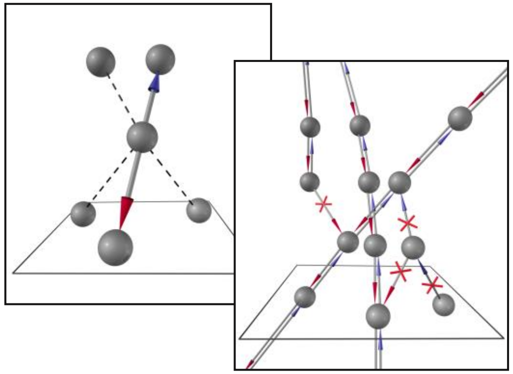


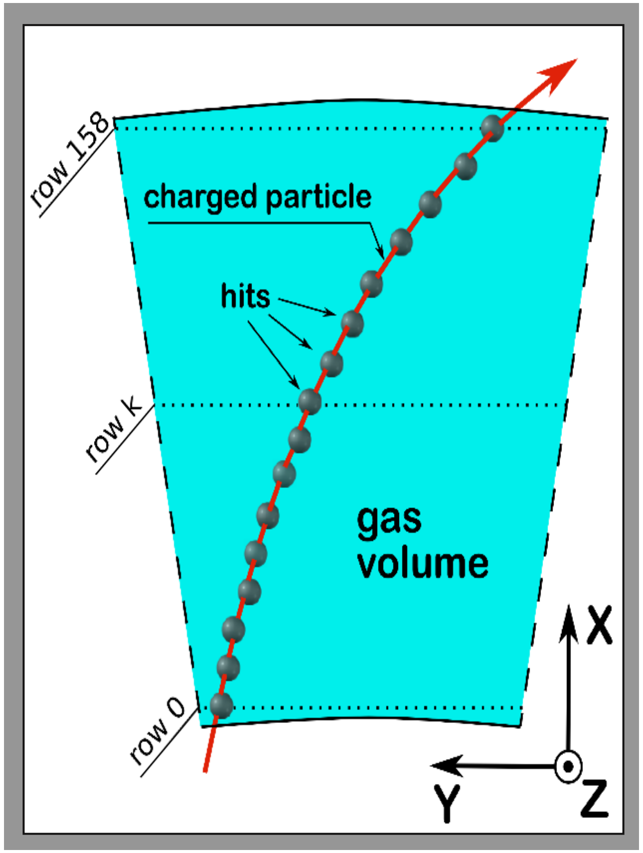
ALICE HLT TPC Tracker

TPC tracking algorithm based on Cellular Automaton approach

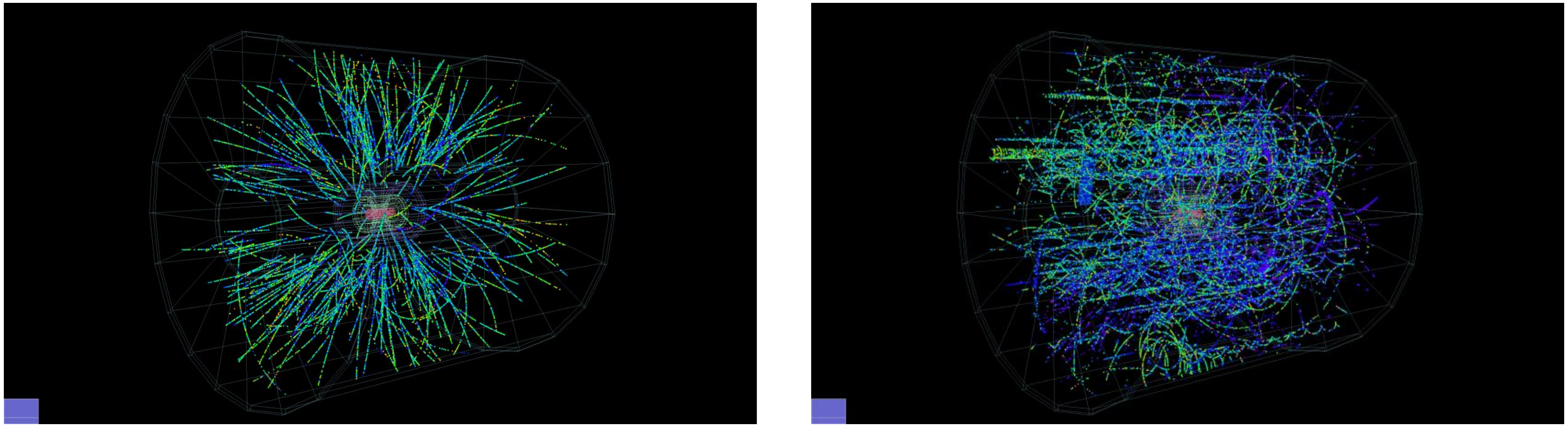
Optimized for multi-core CPUs to fulfill latency requirements

Also available for CUDA/NVIDIA GPUs and currently being ported to OpenCL





Background Rejection



“Background” processes also contribute to the TPC clusters

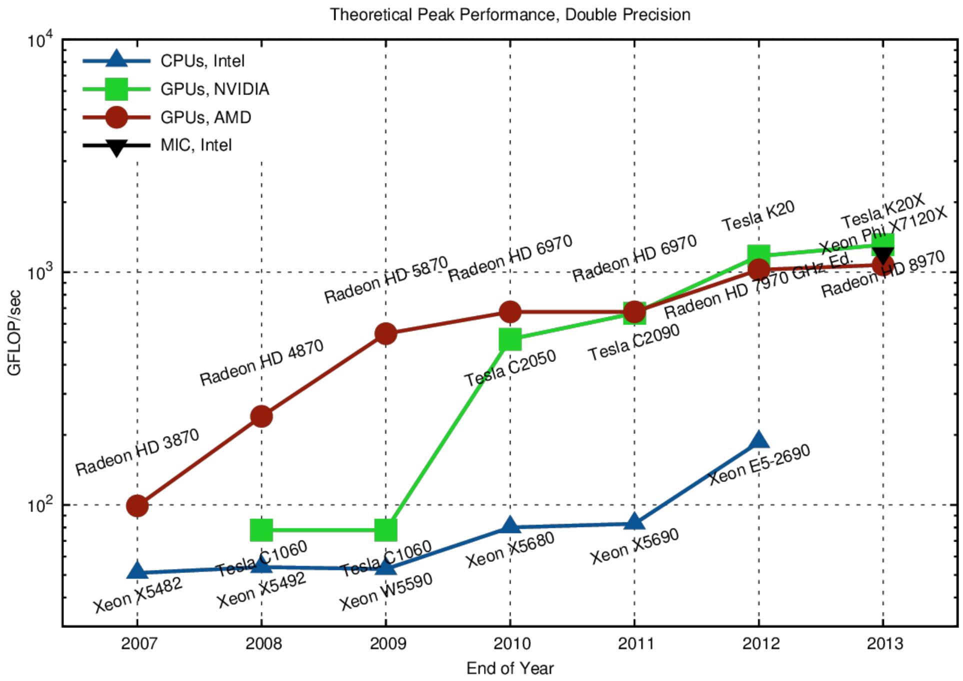
- Number of “background” clusters ~ number of physics clusters

Can we filter this background?

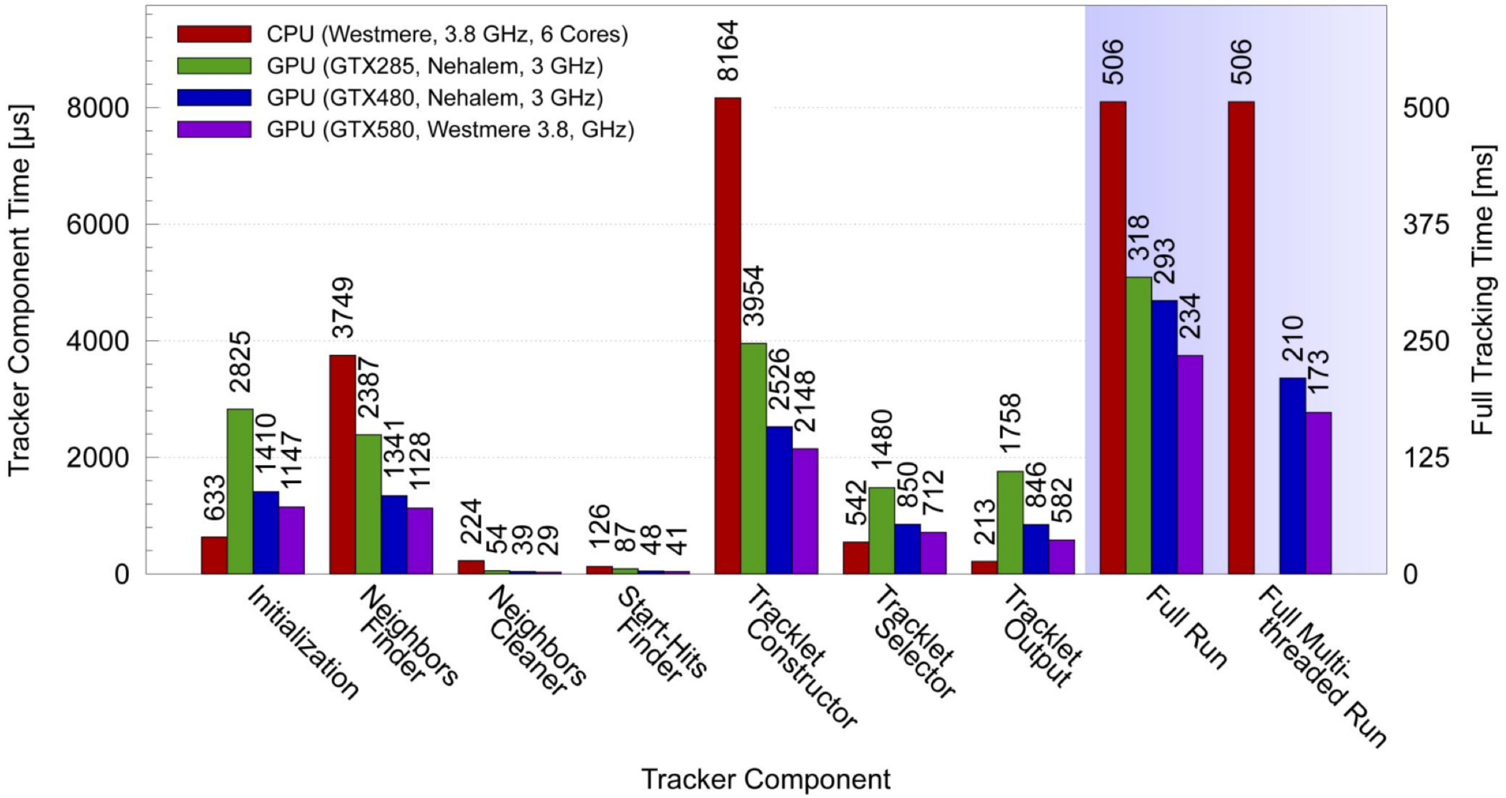
What is the optimal computing algorithm for it?

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)



ALICE HLT TPC Tracker Speedup



4-fold Speedup compared to optimized CPU version

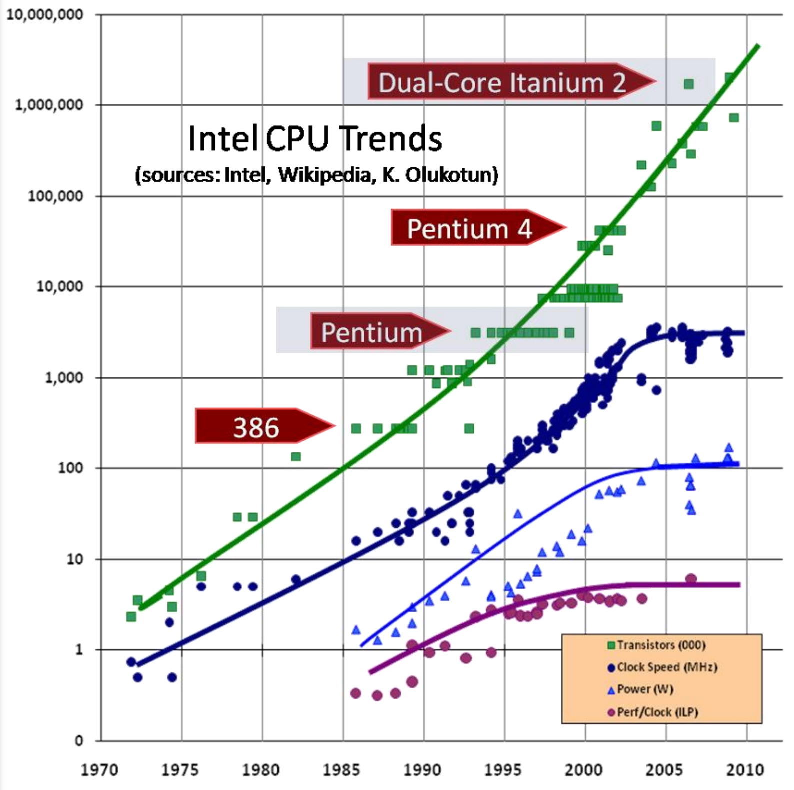
Note: frees CPUs on CN for other operations (tagging/trigger)

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

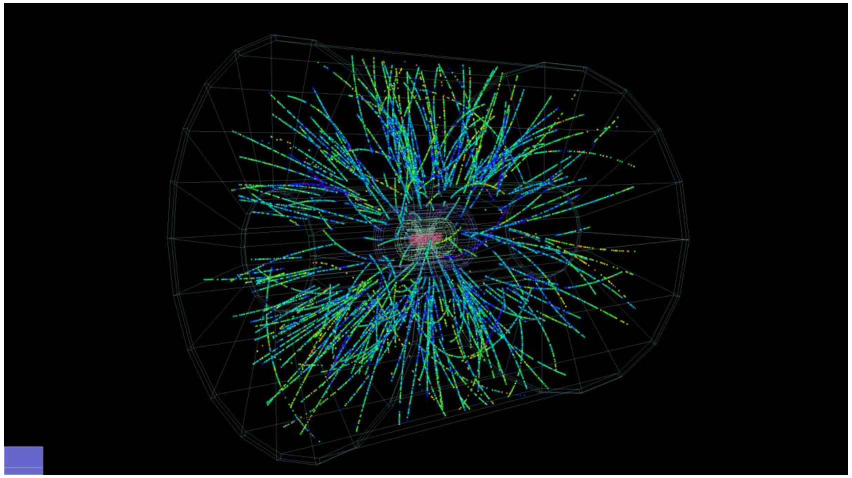
Reconstruction software needs to adapt to full use resources



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)

Processing Requirements

Today



- O(0.1s) online (HLT)
  with accelerator cards
  (FPGA+CPU+GPU)
  limited accuracy
- O(100s) offline
  on the GRID
  “ultimate” performance

Future

Full reconstruction online!

What is the optimal computing architecture?

Processing Power

Estimate for online systems based on current HLT processing power

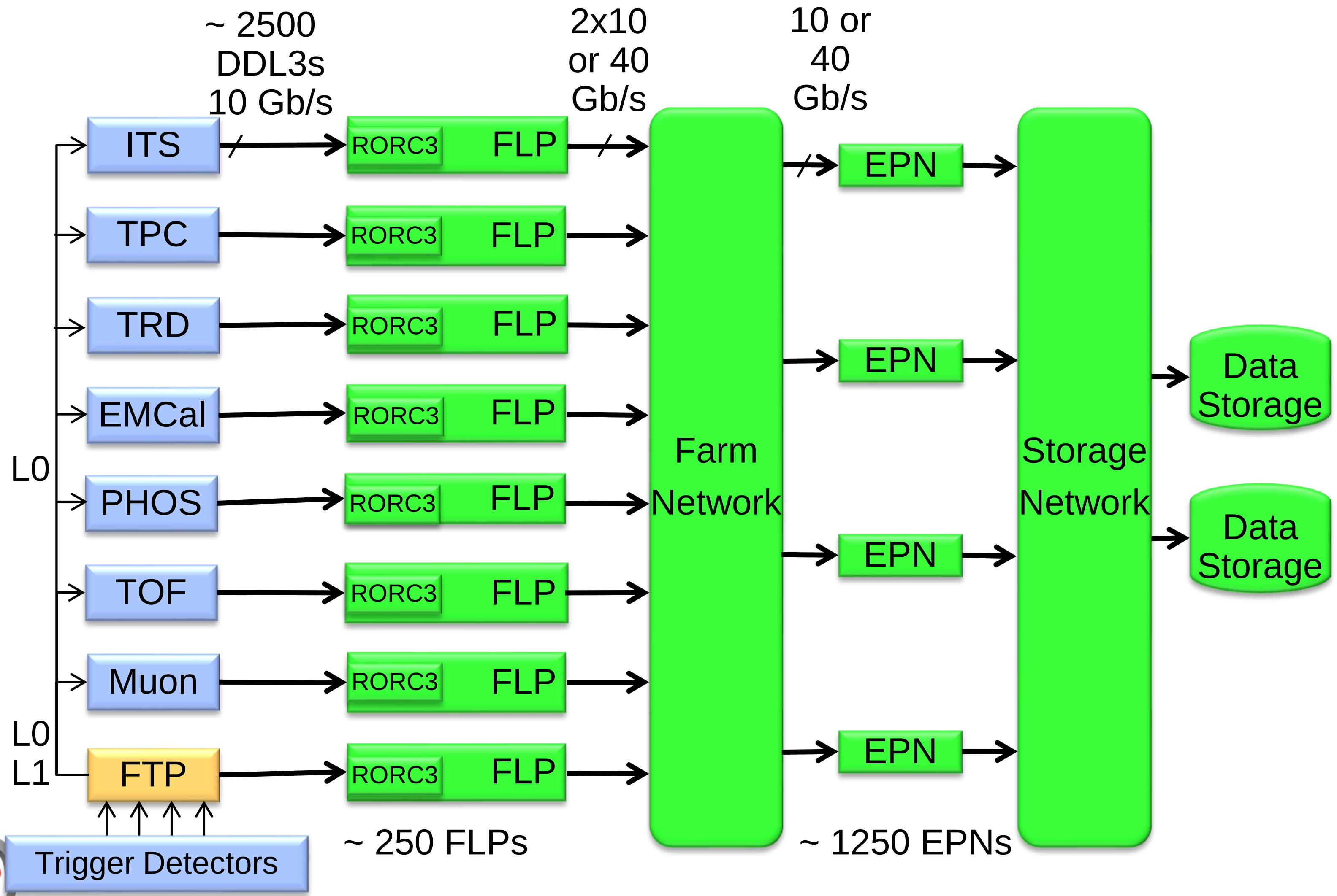
- ~2500 cores distributed over 200 nodes
- 108 FPGAs on H-RORCs for cluster finding
  1 FPGA equivalent to ~80 CPU cores
- 64 GPGPUs for tracking (NVIDIA GTX480 + GTX580)

Scaling to 50 kHz rate to estimate requirements

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

⇨1250-1500 nodes in 2018 with multicores

O² System from the Letter of Intent



~ 2500
DDL3s
10 Gb/s

2x10
or 40
Gb/s

10 or
40
Gb/s

ITS
TPC
TRD
EMCal
PHOS
TOF
Muon
FTP

L0

L0
L1

Trigger Detectors

RORC3 FLP
RORC3 FLP
RORC3 FLP
RORC3 FLP
RORC3 FLP
RORC3 FLP
RORC3 FLP
RORC3 FLP

~ 250 FLPs

Farm
Network

EPN
EPN
EPN
EPN

~ 1250 EPNs

Storage
Network

Data
Storage

Data
Storage

ALICE

Online Reconstruction Mode

Synchronous with data taking

- Need to handle peak load ➪ computing requirements
- Very high-fault tolerance ➪ failure ≙ stop of data taking
- Code stability like online ➪ few updates during run

Asynchronous with data taking

- Need to handle average load
- Faults can be recovered
- More frequent code updates possible

What parts of the Online Reconstruction can be done asynchronously?

Online Reconstruction Mode

Data Input/Data Reduction/Storage synchronously

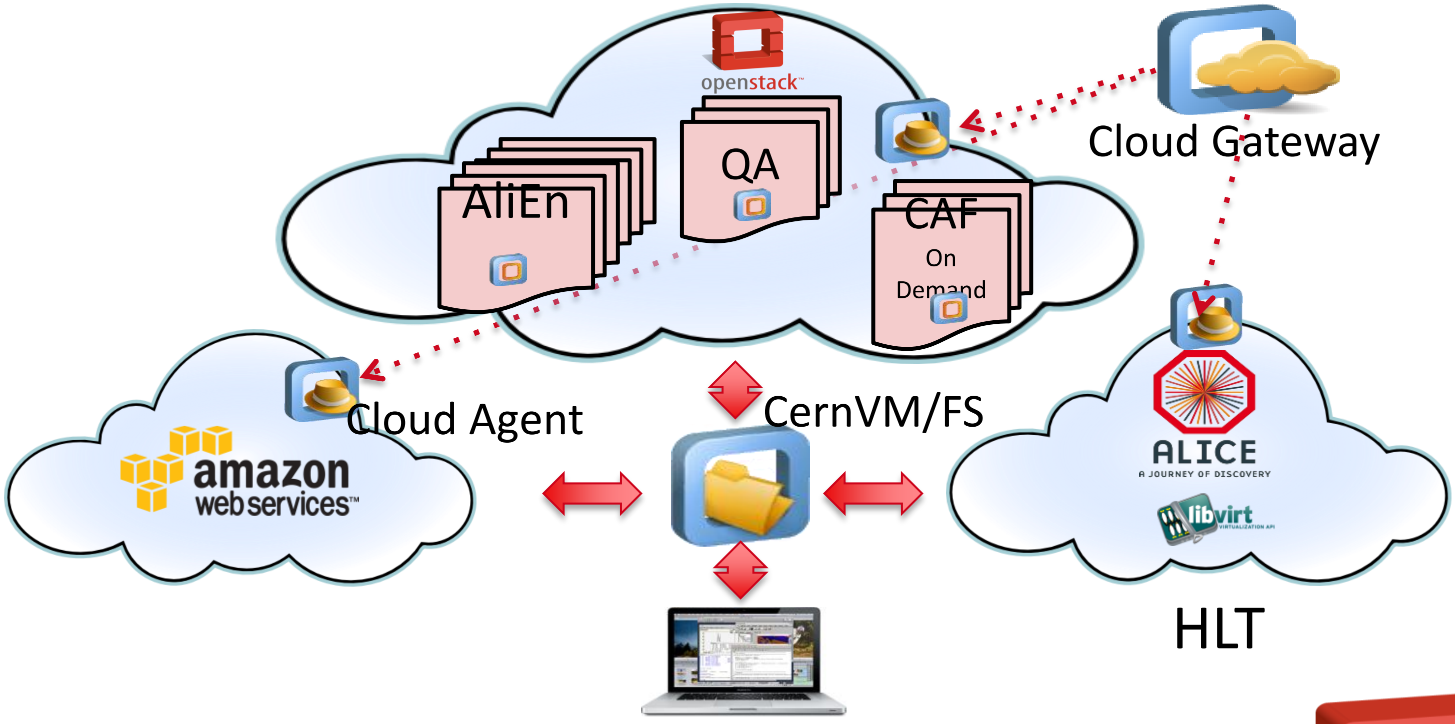
- Designed to handle peak load
- Minimize processing/calibration sensitivity
- Streamed processing – no backloops!
- Feasible to prepare calibrations constants for full reconstruction?
- Monitoring/QA (*can they be asynchronously?*)

Use EPN memory/local storage as buffer

Full reconstruction asynchronously

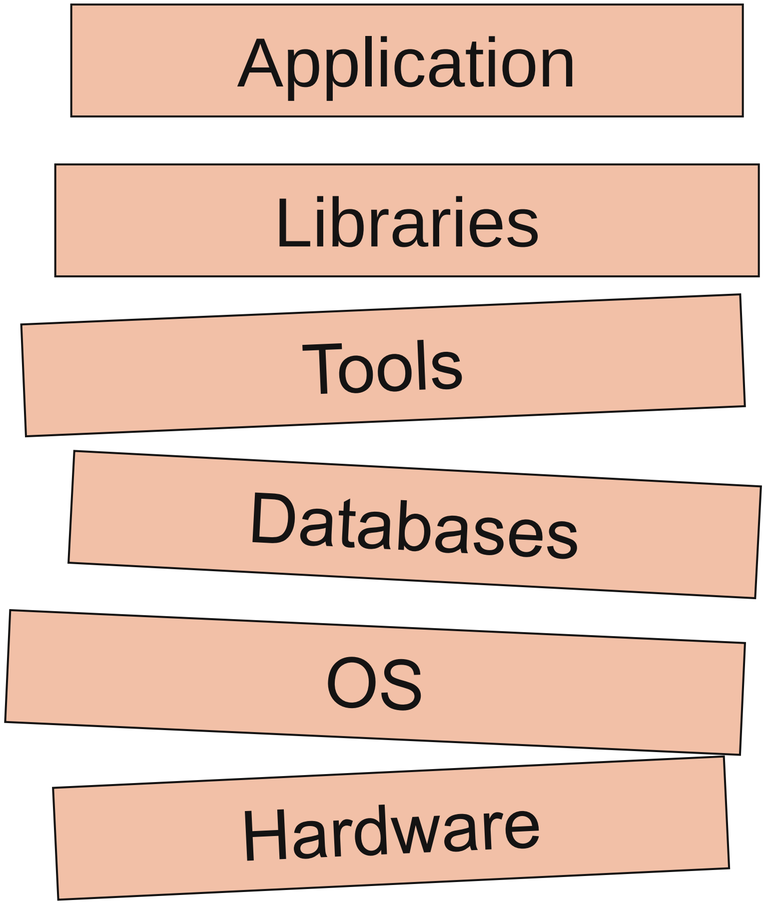
- Designed to handle average load
- Only one pass, avoid backloops…
- AOD output for physics analysis

Clouds…



Based on CernVM family of tools
Prototyped for offline use of HLT farm during Run 2

Software integration problem



Application

Libraries

Tools

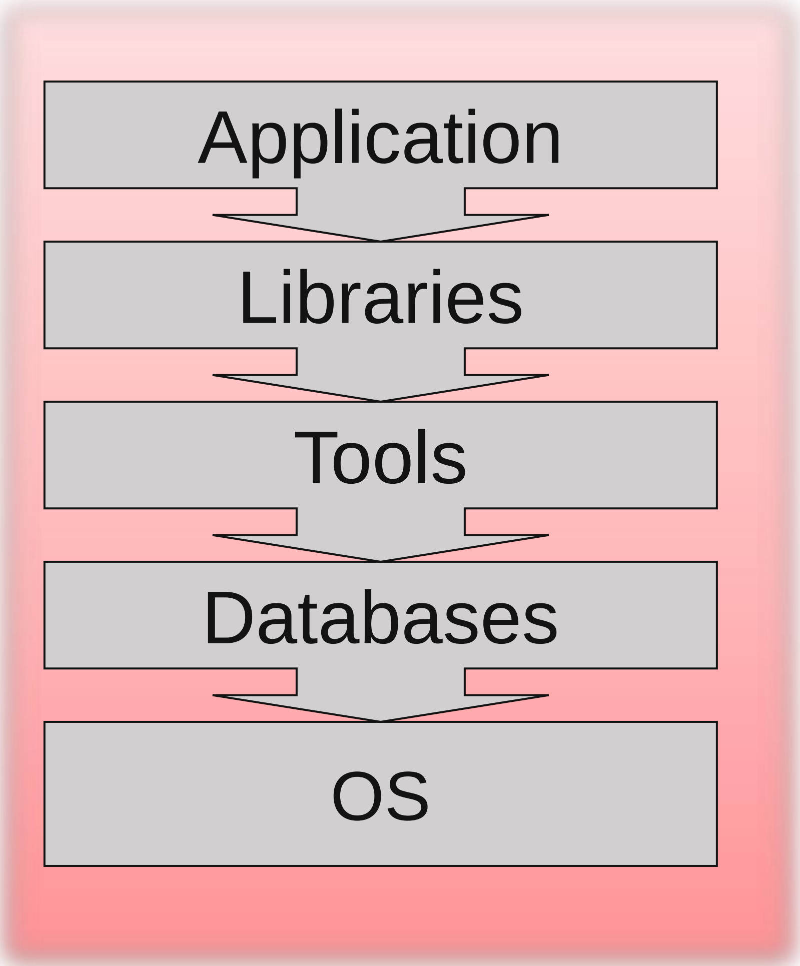
Databases

OS

Hardware

- Traditional model
  - Horizontal layers
  - Independently developed
  - Maintained by the different groups
  - Different lifecycle
- Application is deployed on top of the stack
  - Breaks if any layer changes
  - Needs to be certified every time when something changes
  - Results in deployment and support nightmare
- Difficult to do upgrades
  - Even worse to switch to new OS versions

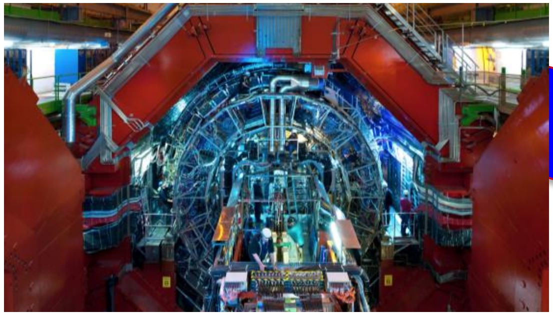
Decoupling Apps and Ops

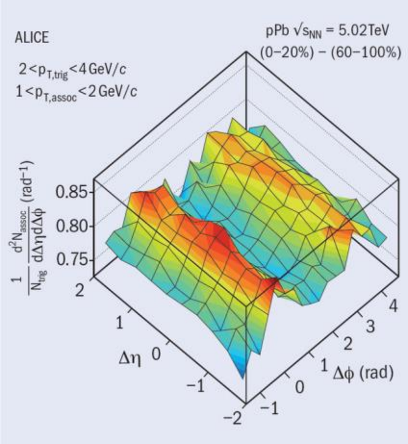


- Application driven approach
  1. Start by analysing the application requirements and dependencies
  2. Add required tools and libraries
- Use virtualization to
  1. Build minimal OS
  2. Bundle all this into Virtual Machine image

•Separates lifecycles of the application and underlying computing infrastructure

Summary



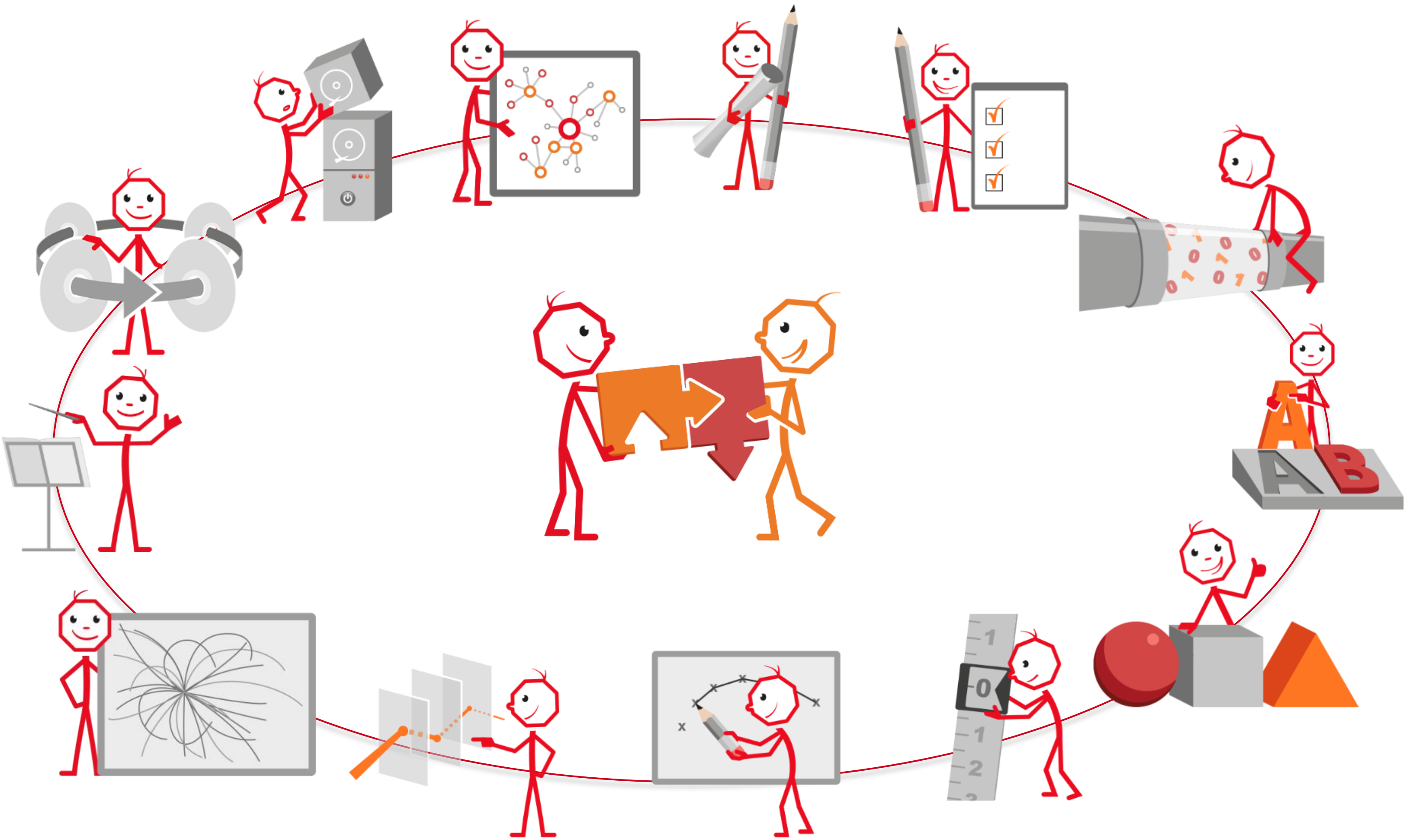


From Detector Readout to Analysis:
What is the “optimal” computing architecture?

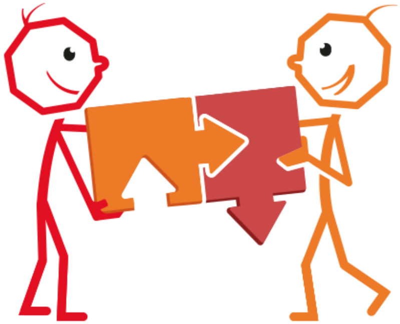
Lots of interesting R&D in the coming years

- Multi-core/accelerator cards
- Data Management
- Clouds
- The online high performance computing farm

O² Project Organization



O² Institutes



Institutes

- FIAS, Frankfurt, Germany
- IIT, Mumbay, India
- Jammu University, Jammu, India
- IPNO, Orsay, France
- IRI, Frankfurt, Germany
- Rudjer Bošković Institute, Zagreb, Croatia
- SUP, Sao Paulo, Brasil
- University Of Technology, Warsaw, Poland
- Wiegner Institute, Budapest, Hungary
- CERN, Geneva, Switzerland

Looking for more people

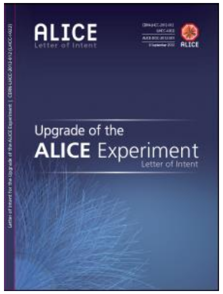
- Need people with computing skills and from detector groups

CWG's membership is neither closed nor rigid:

- New members more than welcome to join

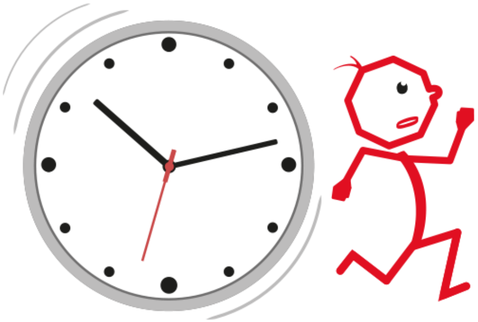
Overall Schedule

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



ALICE
Letter of Intent

Upgrade of the
ALICE Experiment
Letter of Intent



O^2 System from the Letter of Intent

Cost estimate

Item	Cost[MCHF]
DDL fibres	0.9
EPN	4.1
FLP and CRORC	0.9
Infrastructure	1.3
Networks	0.8
Servers	0.5
Storage	0.6
Central DCS	0.2
Total*	**9.3**

Based on extrapolation from existing HLT/DAQ systems

For 50 kHz Pb+Pb interaction rate (scaling to 100 kHz foreseen)

+ 0.5 MCHF for (central) offline investments

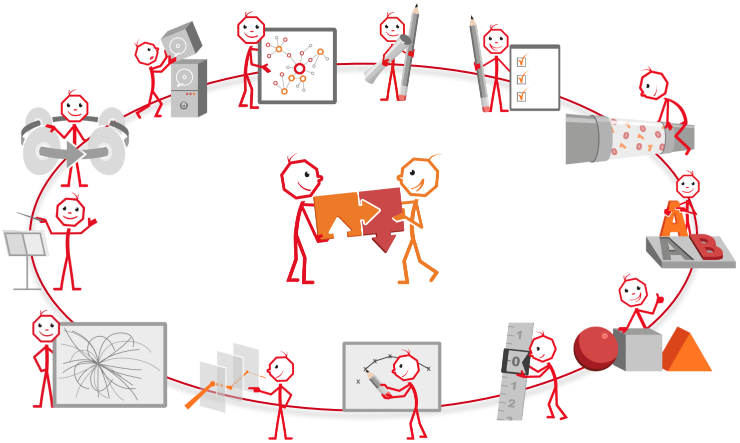
O^2 Project

Computing systems after LS2 have to handle > 1TByte/s input

- Detectors in continuous & triggered read-out mode
- Data reduction by (partial) online reconstruction
- Raw data reconstruction on same farm
- Output: AODs

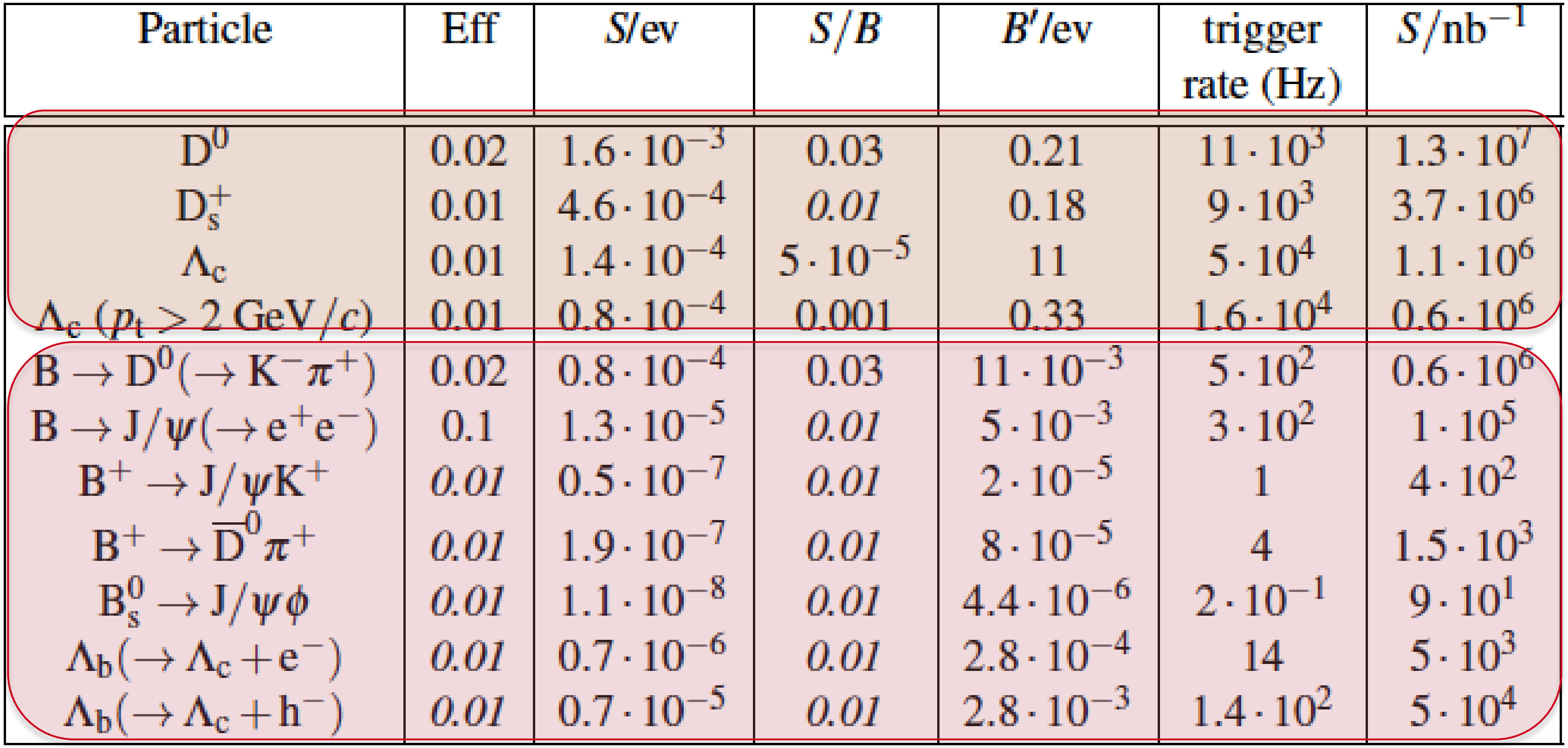
O2 Project organized in CWGs

- Working towards TDR in 2014
- Open for new participants, especially also from detectors



Why not triggering?

Slide from Luciano Musa



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

➡ ~ 36 kHz@50kHz rate...