



**ALICE**

# **ALICE approach to streaming readout**

P. Vande Vyvre / CERN-EP

---

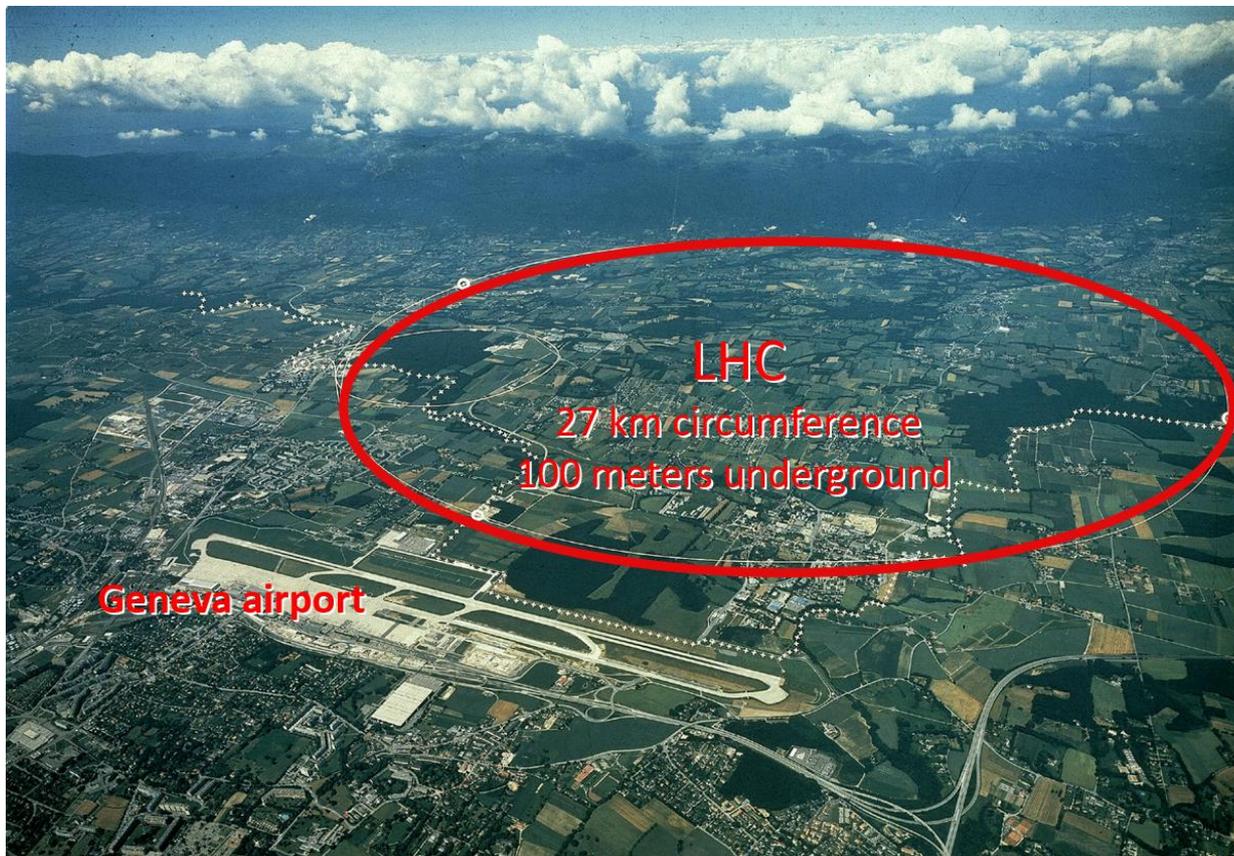
# Outlook

- First, a warm thank you for the invitation in this paradisiac place !
  
- ALICE: current status and upgrade
- New requirements and strategy
- New electronics and computing system
  - Continuous read-out
  - Trigger system and read-out throughput regulation

# CERN LHC



ALICE

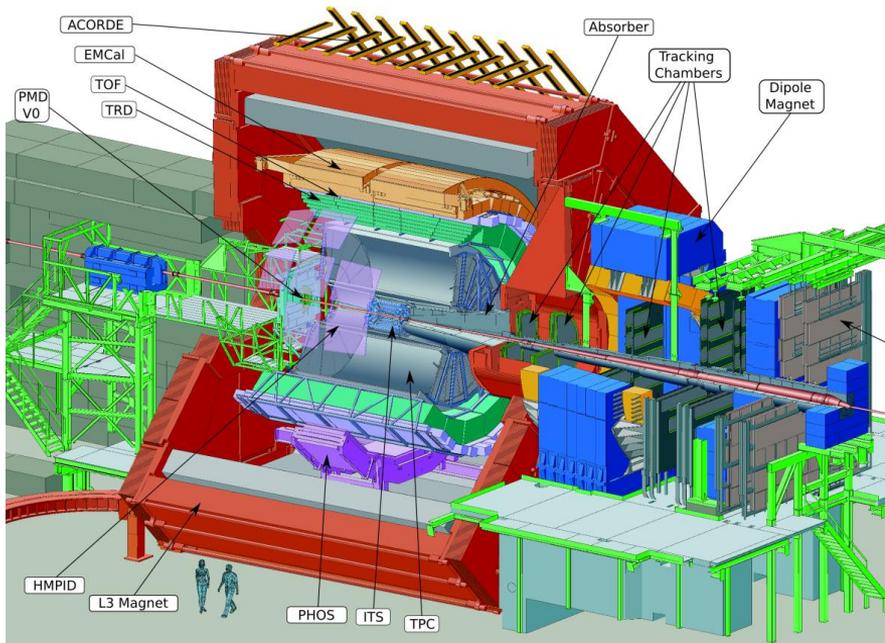


# ALICE experiment till 2019

**Detector: 18 technologies**

**Size: 16 x 26 meters**

**Weight: 10,000 tons**



Collaboration (Jan '19):

**1975 Members**

**175 Institutes**

**40 countries**

## A brief history of ALICE

**1990-1996:** Design

**1993** Official start of ALICE

**1995:** **Technical Proposal**

**1992-2006:** R&D

**2000-2009:** Construction, Installation, Commissioning

**2010-2013:** **Run 1 (Operation)**

**2013-2014:** LS1 (Long Shutdown 1)

**2015-2018:** **Run 2 (Operation)**

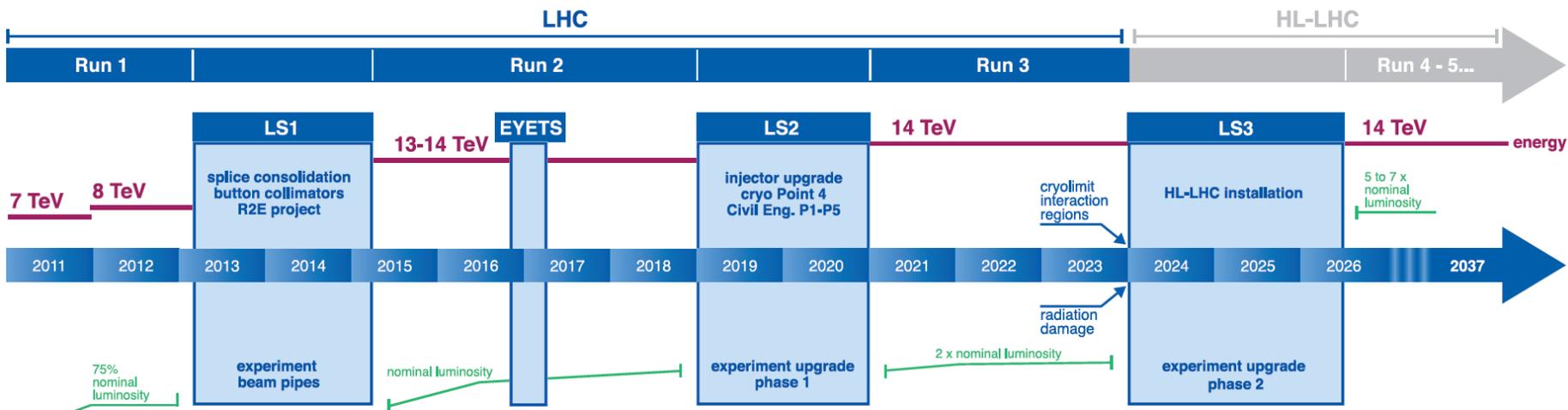
**2019:** LS2 (Long Shutdown 2)



# LHC evolution

Run: exploitation period

LS: Long Shutdown



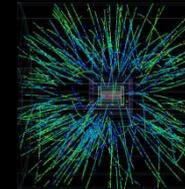
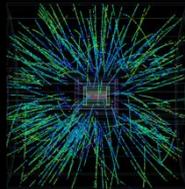
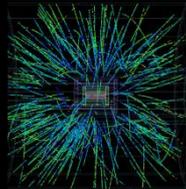
Integrated Luminosity

		Run 1	Run 2	Run 3	Run 4	HL-LHC
ATLAS, CMS	pp	30 fb <sup>-1</sup>	150 fb <sup>-1</sup>	300 fb <sup>-1</sup>		3000 fb <sup>-1</sup>
ALICE	Pb-Pb	1 nb <sup>-1</sup>		13 nb <sup>-1</sup>		

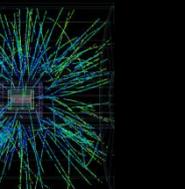
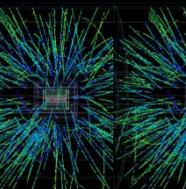
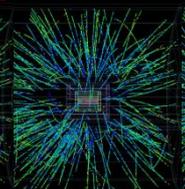
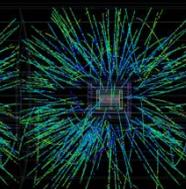
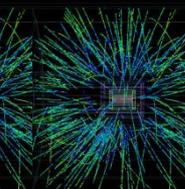
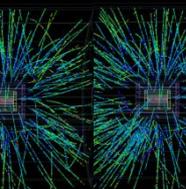
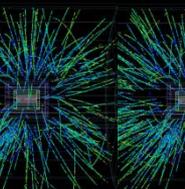
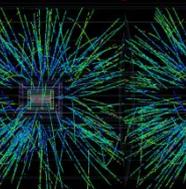
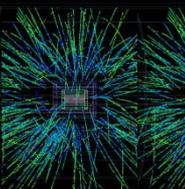
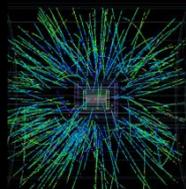


# LHC after 2020

LHC pp now **40 MHz, Luminosity  $7 \times 10^{33} \text{ cm}^{-2} \text{ s}^{-1}$**  - PbPb now **8 MHz, Luminosity  $1 \times 10^{27} \text{ cm}^{-2} \text{ s}^{-1}$**



LHC pp 2021 **40 MHz, Luminosity  $2 \times 10^{34} \text{ cm}^{-2} \text{ s}^{-1}$**  - PbPb 2021 **8 MHz, Luminosity  $6 \times 10^{27} \text{ cm}^{-2} \text{ s}^{-1}$**



# ALICE upgrade

## New Inner Tracking System (ITS)

- improved pointing precision
- less material -> thinnest tracker at the LHC

## Time Projection Chamber (TPC)

- new GEM technology for readout chambers
- faster readout electronics

## TOF, TRD, ZDC

- Faster readout

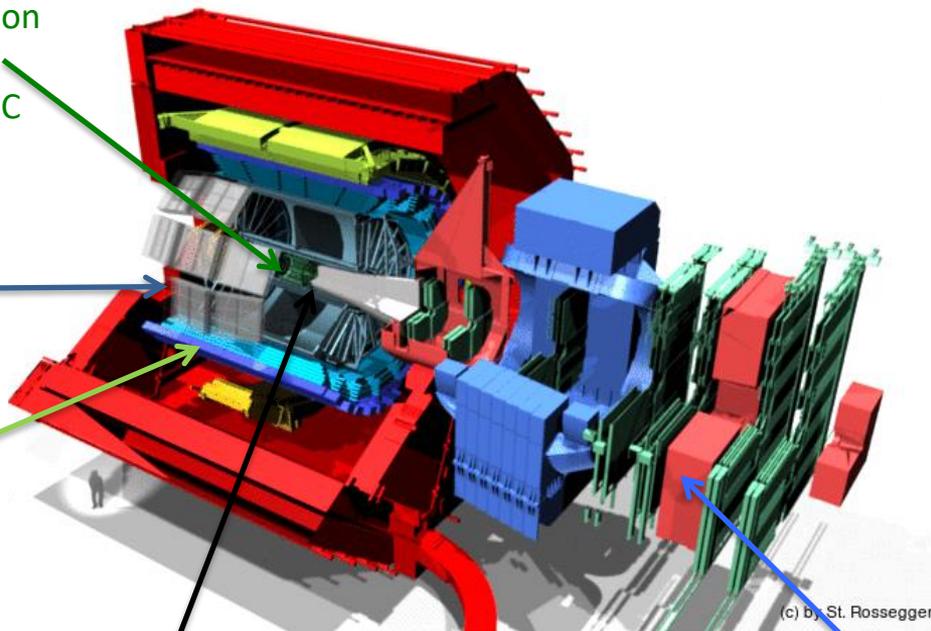
## New Trigger Detectors (FIT)

New Online-Offline computing system

New Central Trigger Processor

## MUON ARM

- continuous readout electronics

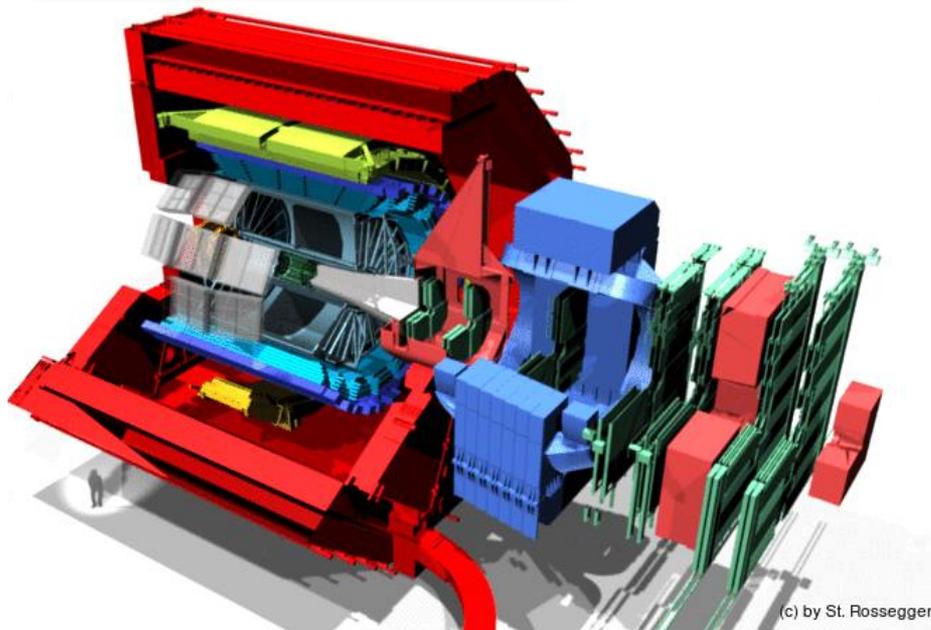


# ALICE experiment till 2029

**Detector:** 18 technologies

**Size:** 16 x 26 meters

**Weight:** 10,000 tons



Collaboration (Jan '19):

**1975** Members

**175** Institutes

**40** countries

## ALICE: plans for the next 10 years

**1990-1996:** Design

**1993** Official start of ALICE

**1995:** Technical Proposal

**1992-2006:** R&D

**2000-2009:** Construction, Installation, Commissioning

**2010-2013:** Run 1 (Operation)

**2013-2014:** LS1 (Long Shutdown 1)

**2015-2018:** Run 2 (Operation)

**2019-2020:** **LS2: major ALICE Upgrade**

**2021-2023:** **Run 3 (Operation)**

**2024-2026:** LS3

**2026-2029:** **Run 4 (Operation)**

# Physics programme and data taking scenarios

Year	System	$\sqrt{s_{NN}}$ (TeV)	$L_{int}$		$N_{collisions}$
			( $pb^{-1}$ )	( $nb^{-1}$ )	
2021	pp	14	0.4		$2.7 \cdot 10^{10}$
	Pb-Pb	5.5		2.85	$2.3 \cdot 10^{10}$
2022	pp	14	0.4		$2.7 \cdot 10^{10}$
	Pb-Pb	5.5		2.85	$2.3 \cdot 10^{10}$
2023	pp	14	0.4		$2.7 \cdot 10^{10}$
	pp	5.5	6		$4 \cdot 10^{11}$
2027	pp	14	0.4		$2.7 \cdot 10^{10}$
	Pb-Pb	5.5		2.85	$2.3 \cdot 10^{10}$
2028	pp	14	0.4		$2.7 \cdot 10^{10}$
	Pb-Pb	5.5		1.4	$1.1 \cdot 10^{10}$
	p-Pb	8.8		50	$10^{11}$
2029	pp	14	0.4		$2.7 \cdot 10^{10}$
	Pb-Pb	5.5		2.85	$2.3 \cdot 10^{10}$

# New requirements

- After LS2, LHC will deliver min bias Pb-Pb collisions at 50 kHz
  - New concepts needed to read out the data and reduce the data volume
- Support for continuous read-out (TPC)
  - Collision rate is faster than intrinsic rate of the slowest detector (TPC) drift time  $\sim 100 \mu\text{s}$
  - Continuous detector read-out introduced (movies rather than pictures)
- Limited selectivity by triggering
  - Very small signal-to-background ratio
  - Triggering (selection) techniques very inefficient if not impossible
  - Needs large statistics
  - Read the data resulting from all interactions
- Software-based reduction of the huge data volume
  - $\sim 100$  x more data than today
  - Reduction of the data volume by online calibration and reconstruction

# Overall strategy

- TPC: unmodified data all the time
  - Zero suppression moved from the front end ASIC to the FPGA-based read-out
  - Data from many TPC read-out pads are available
    - baseline restoration and zero suppression can be performed more effectively
- Continuous read-out with a heart-beat synchronization
  - The LHC clock is still the common reference
- Data fragmented in intervals of continuous data (Time Frame TF) delimited by Heart Beats (HB)
  - Data lost at the extremities of the intervals
  - Duration of intervals selected to minimize the loss while keeping a reasonable memory footprint
- No data retransmission
- Dead-time
  - Created by local decision of discarding data overflow
  - Monitored by the Central Trigger Processor

# Electronics and computing system

## New electronics for new/upgraded detectors

- Common front-end chip: SAMPA
- Common Read-Out Unit (CRU)

## New trigger system

- Heart Beats generation and distribution
- Data throughput monitoring

## New computing system

- Read-out the data of all interactions
- ➔ Compress these data intelligently by online reconstruction
- ➔ One common online-offline computing system: O<sup>2</sup>
- Paradigm shift compared to approach for Run 1 and 2

**Unmodified raw data of all interactions shipped from detector to online farm in triggerless continuous mode**

HI run 3.3 TByte/s

Baseline correction and zero suppression  
Data volume reduction by zero cluster finder. No event discarded.  
Average compression factor 6.6

500 GByte/s

**Data volume reduction by online tracking.**  
**Only reconstructed data to data storage.**  
Average compression factor 5.5

90 GByte/s

Data Storage: 1 year of compressed data

- Bandwidth: Write 90 GB/s Read 90 GB/s
- Capacity: 60 PB

Tier 0, Tiers 1 and  
Analysis Facilities

Asynchronous (few hours) event  
reconstruction with final calibration



ALICE

# O<sup>2</sup> system

## O2/FLP

CR1 – First Level Processors

TPC 3.45 TByte/s  
ITS 40 GByte/s  
TRD 4GByte/s  
Rest 21 Gbyte/s

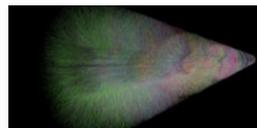
3.5 TB/s

Continuous  
Unmodified  
Raw data

TPC 570 GByte/s  
ITS 40 GByte/s  
TRD 4GByte/s  
Rest 21 Gbyte/s

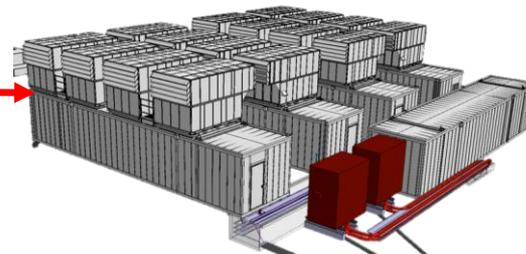
635 GByte/s

Sub-Timeframes (20ms)



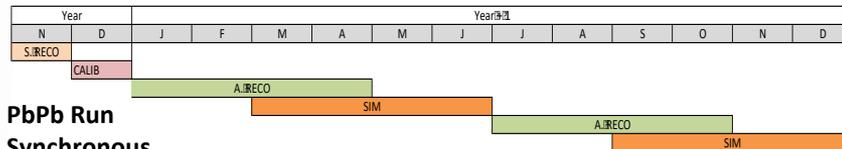
## O2/EPN

CR0 - Event Processing Nodes



## O2/PDP

Physics and Data Processing



PbPb Run  
Synchronous

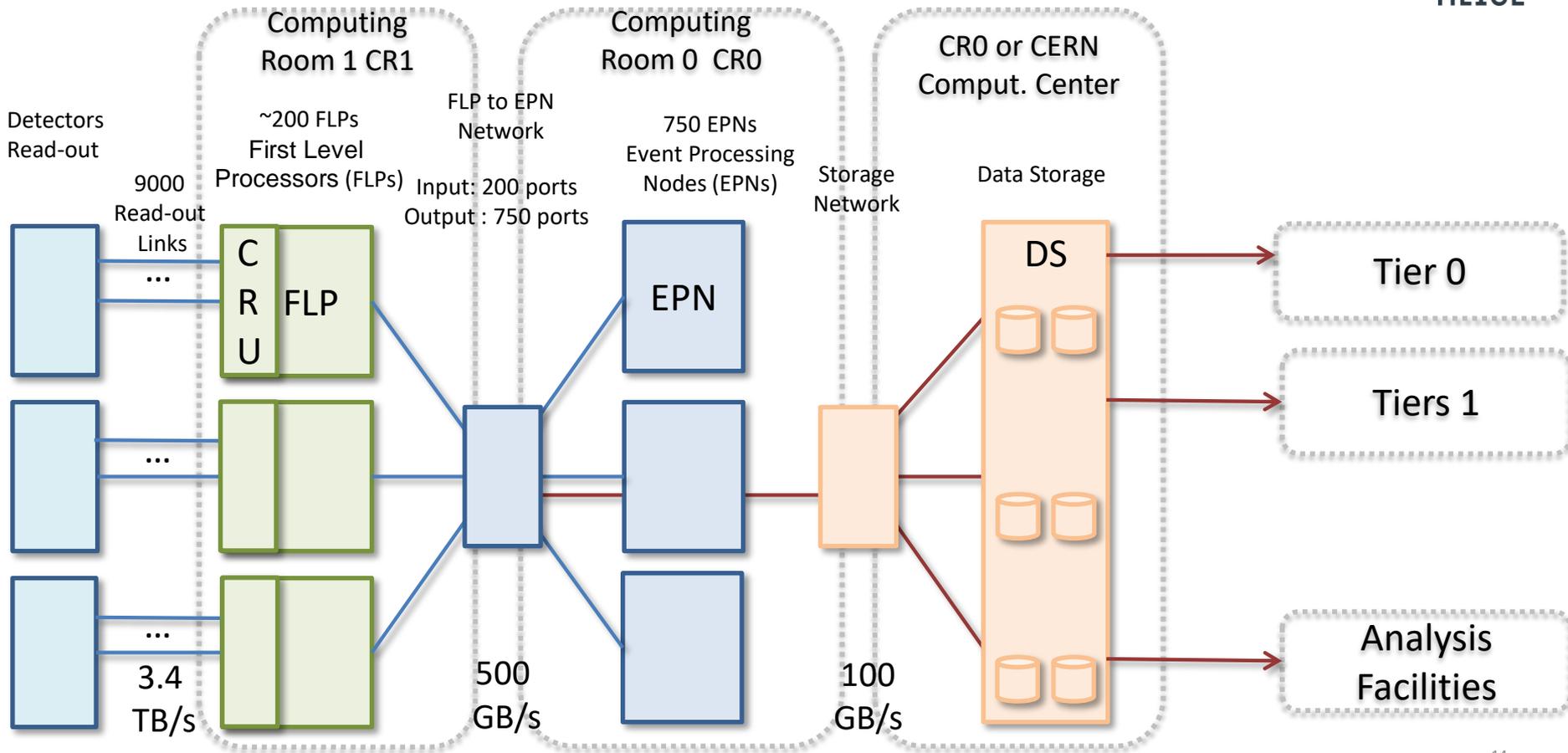
Asynchronous 1

Asynchronous 2

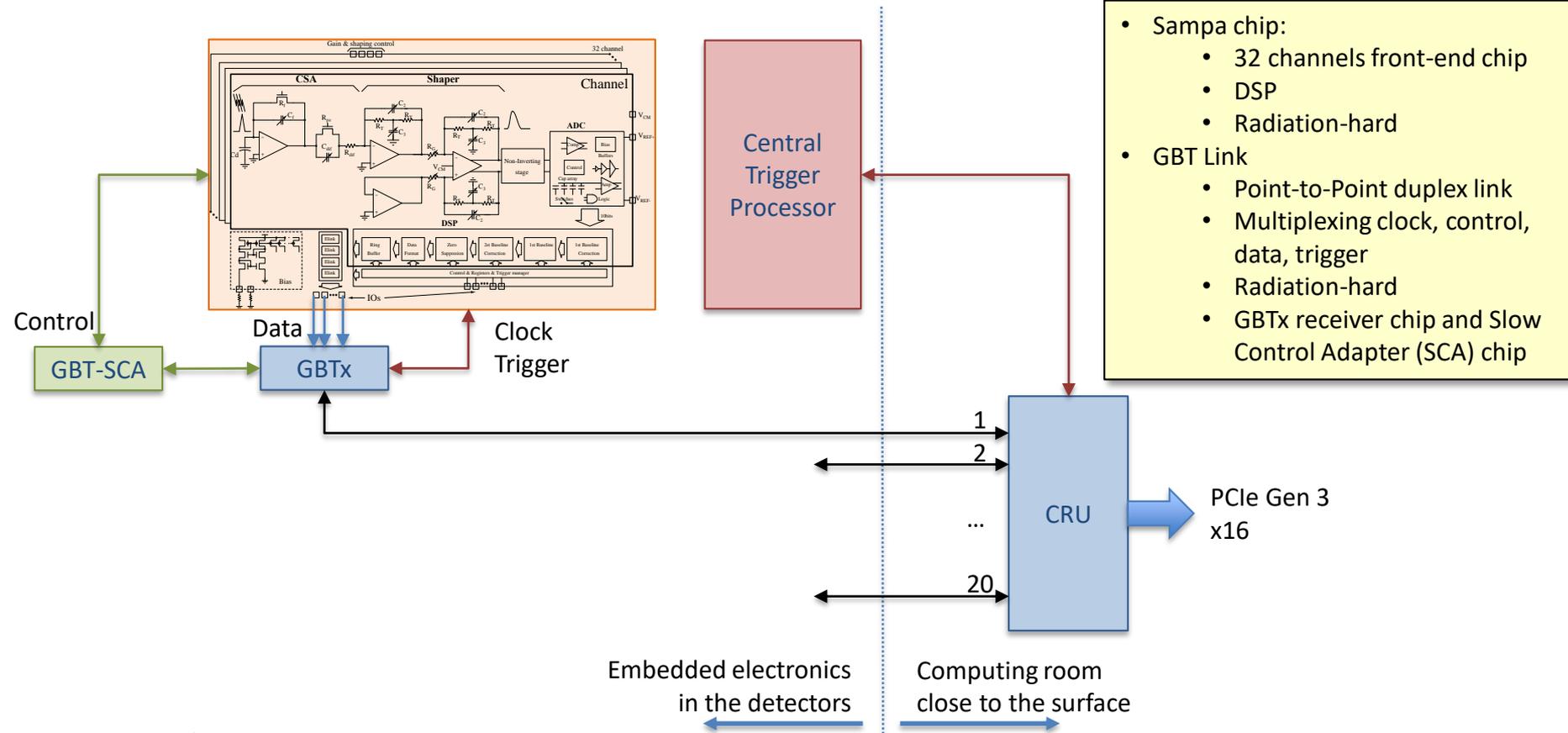
CTP  
Central Trigger Processor

Distribution of timing info, heartbeat trigger, data throughput monitoring

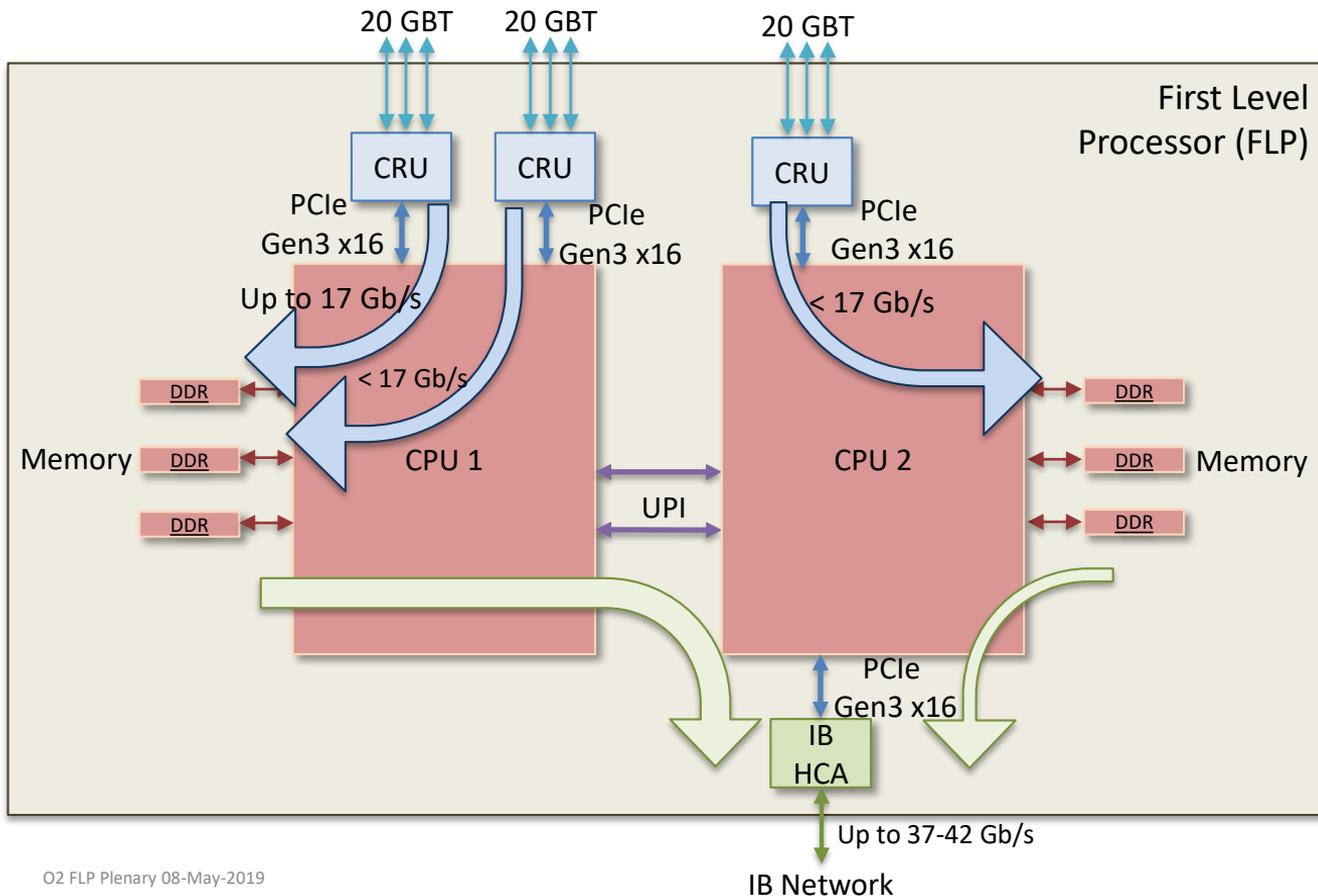
# Hardware architecture



# Electronics : Sampa chip and GBT link



# Data flow inside CRU and FLP



Maximum need for TPC:

- Up to 17 Gb/s per CRU
- Up to 37-42 Gb/s per FLP

O2/FLP read-out sw and fw able to saturate the PCIe Gen3 x16: 110 Gb/s per port

- Up to 87 Gb/s concurrent input and output (factor 2 compared to the maximum need)

# FLP qualification before procurement

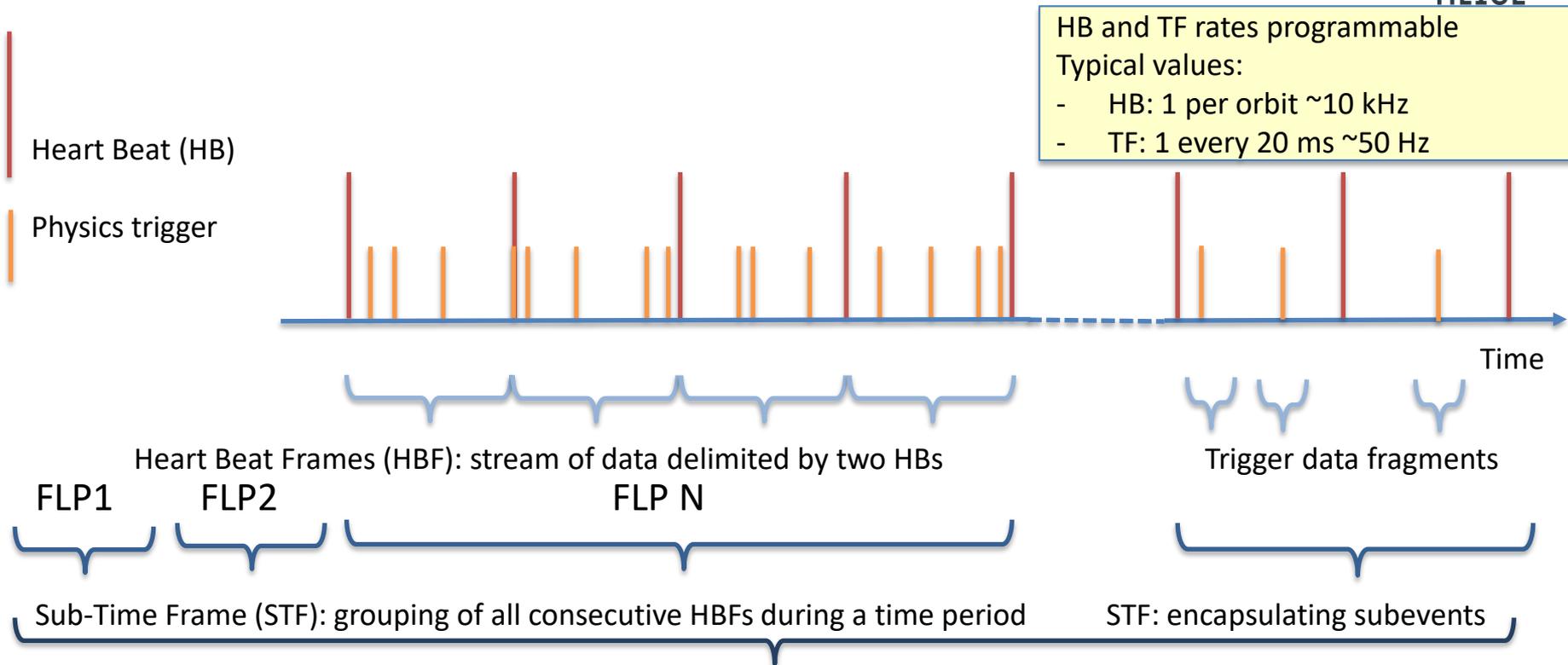
- Results so far (3 CRUs per FLP)

CERN Name	Base Model	Test #1 CRU-FLP Compatibility	Test #2 Input	Test #3 Input + Output	Test #4 Input + Output +Processing
FLP-AG4	ASUS ESC4000-G4		330 Gb/s	87 Gb/s	In progress
FLP-D740	DELL PowerEdge R740		330 Gb/s	87 Gb/s	In progress
FLP-D7425	DELL PowerEdge R7425		220 Gb/s	In progress	
FLP-H380	HP ProLiant DL380	unstable			
FLP-SM	Supermicro X11DPG-QT		330 Gb/s	87 Gb/s	In progress

- Data flow demonstrated in 3 platforms up to x2.4 nominal data rate



# Continuous read-out: Triggers, Heart Beats, Timeframe



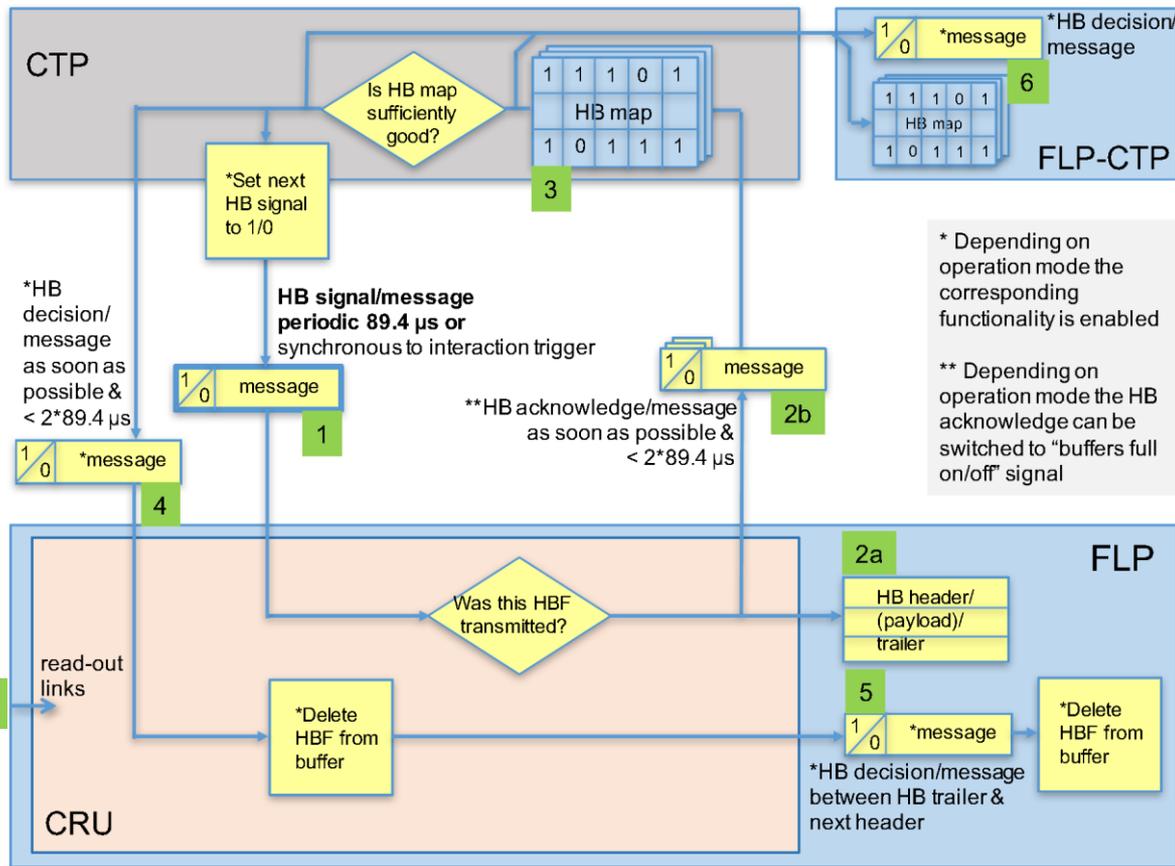
HB and TF rates programmable

Typical values:

- HB: 1 per orbit  $\sim 10$  kHz
- TF: 1 every 20 ms  $\sim 50$  Hz

Time Frame grouping of all STFs from all FLPs for the same time period  
from detectors triggered or read out continuously for the same time period

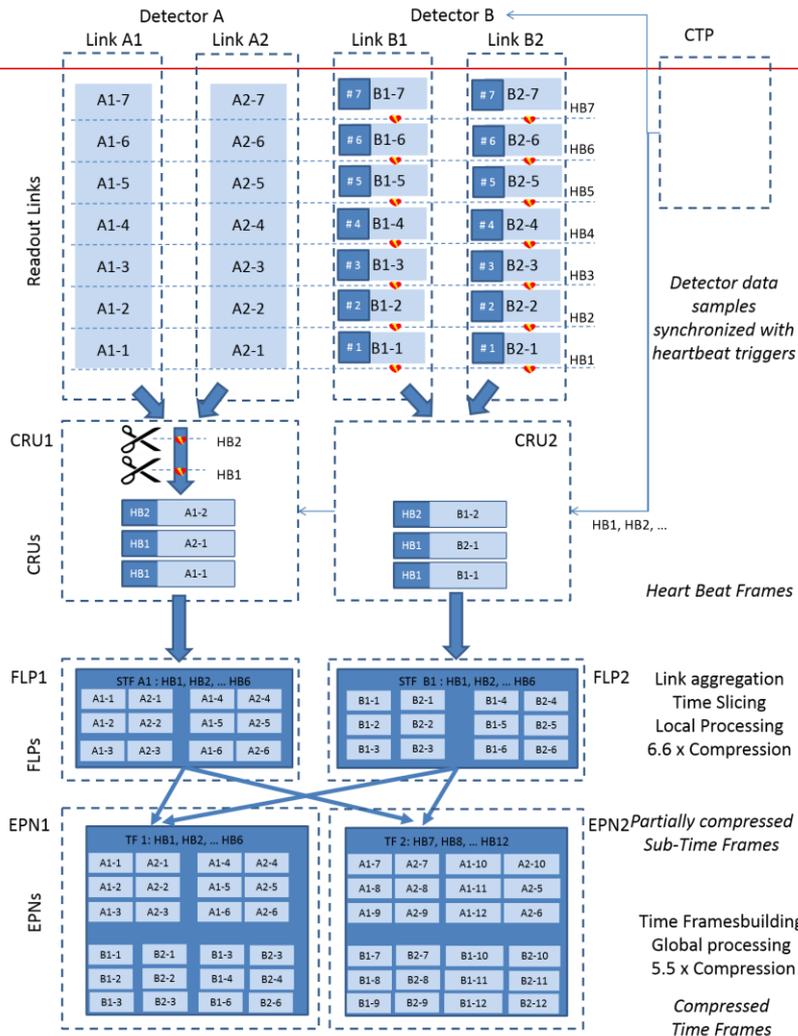
# Triggering and read-out throughput regulation



- 0) Data from front-end electronics
- 1) New Heart Beat (accept or reject)
- 2) Data received successfully for this HB (notify O<sup>2</sup> (a) and CTP(b))
- 3) CTP builds Heart beat map  
→ If Time Frame lost, reject further frames
- 4) Decision message to keep/remove Heart Beat Frame
- 5) Decision message in FLP memory
- 6) From 3) notify FLP-CTP



# Data aggregation

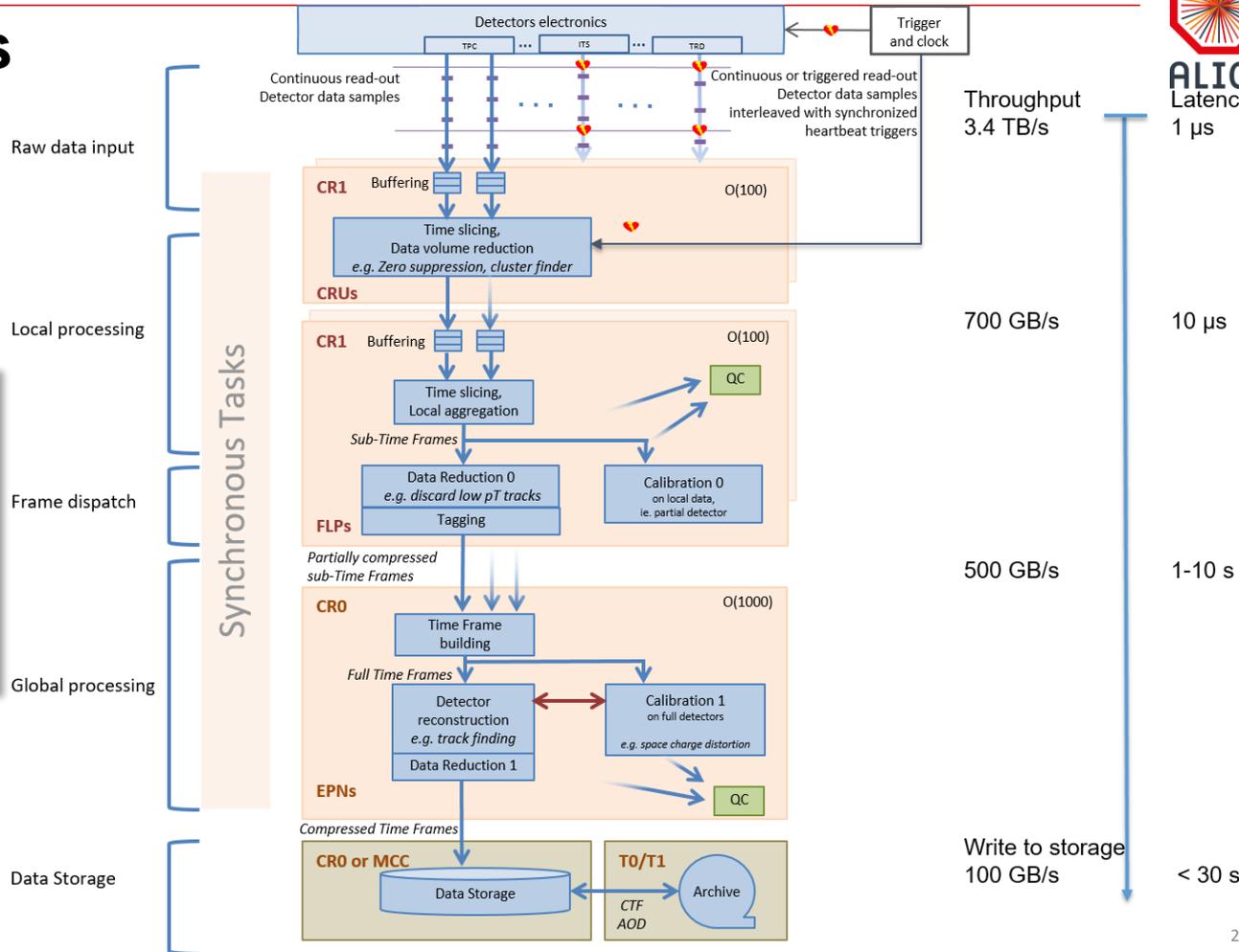




# Synchronous Processing

ALICE  
Latency  
1  $\mu$ s

- Set of operations to be performed before data storage, i.e. all the operations to reduce the data volume
- Fast and simple calibration
- Reconstruction
- Production of the immutable Compressed Time Frames





ALICE

# Asynchronous Processing

Storage

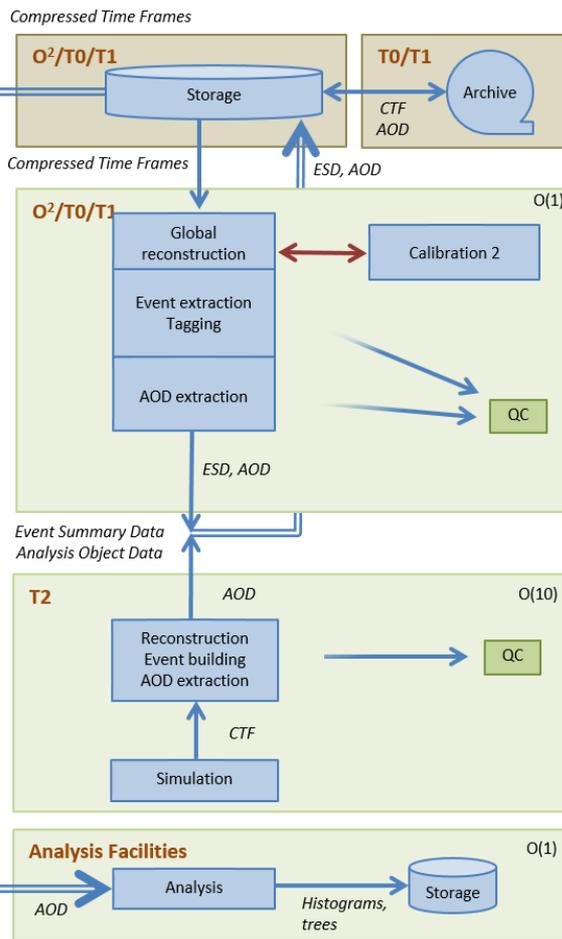
Reconstruction passes and event extraction

- Set of operations to be performed before analysis and publications, i.e. to produce physics quality data
- Refined calibration
- Second pass of reconstruction

Simulation

Analysis

Asynchronous and offline tasks



Throughput

Latency

Read from storage  
 CTF to O<sup>2</sup> 100 GB/s  
 CTF to T0 35 GB/s  
 CTF to T1s 1-3 GB/s  
 AOD to AF 1-3 GB/s  
 Total ~140 GB/s

Write to storage  
 AOD to O<sup>2</sup> 20 GB/s

Hours to  
 2 Months

# O2 Software Framework

## Data Processing Layer (DPL)

Abstracts away the hiccups of a distributed system, presenting the user a familiar "Data Flow" system.

- *Reactive-like design (push data, don't pull)*
- *Declarative Domain Specific Language for implicit workflow definition.*
- *Integration with the rest of the production system, e.g. Monitoring, Logging, Control.*
- *Laptop mode, including graphical debugging tools.*

## Data Layer: O2 Data Model

Message passing aware data model. Support for multiple backends:

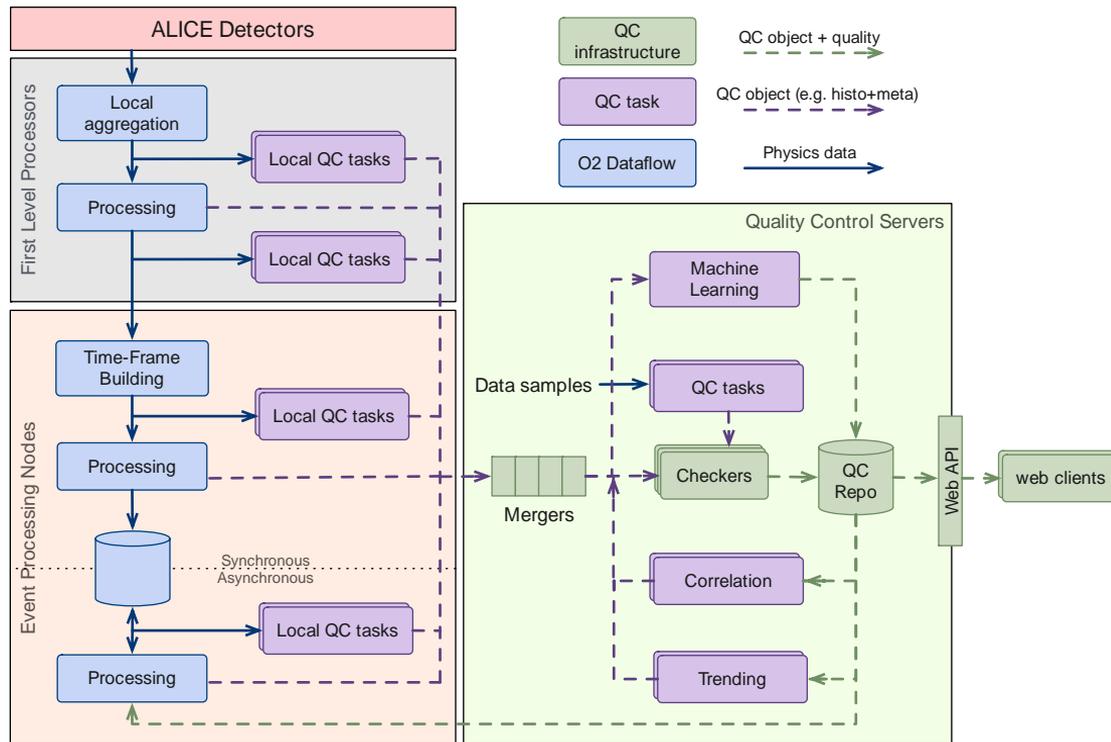
- **Simplified, zero-copy** format optimised for performance and direct GPU usage. Useful e.g. for TPC reconstruction on the GPU.
- **ROOT based serialisation.** Useful for QA and final results.
- **Apache Arrow based.** Useful as backend of the analysis ntuples and for integration with with other tools.

## Transport Layer: ALFA / FairMQ<sup>1</sup>

- **Standalone processes (devices)** for deployment flexibility.
- **Message passing as a parallelism paradigm.**
- **Shared memory backend** for reduced memory usage and improved performance.

# Data Quality Control

- Architecture supporting QC function at each step of the data read-out and processing chain
- Local tasks and offload to remote QC servers whenever needed
- Based on DPL like all the other O2 applications





ALICE

# Hardware acceleration

FPGA used in the FLPs (Long and successful experience of FPGAs in the ALICE HLT)

- Acceleration for TPC cluster finder versus a standard CPU
- 25 times faster than the software implementation
- Use of the CRU FPGA for the TPC cluster finder

GPU used in the EPNs

- TPC Track Finder based on the Cellular Automaton principle to construct track seeds.
- 1 GPU replaces 30 CPU cores and uses 3 for I/O
- Use of GPUs for the TPC track finder

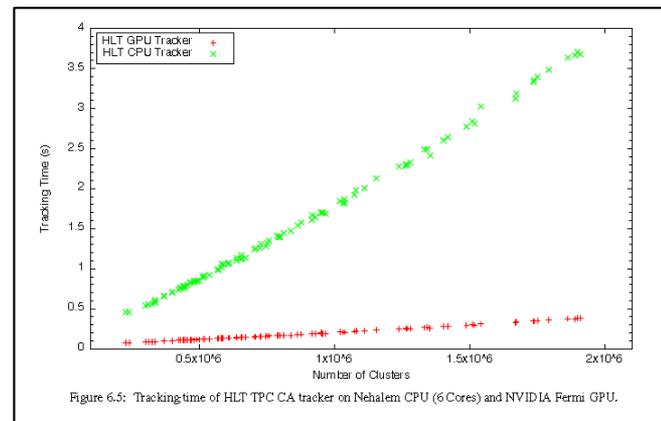
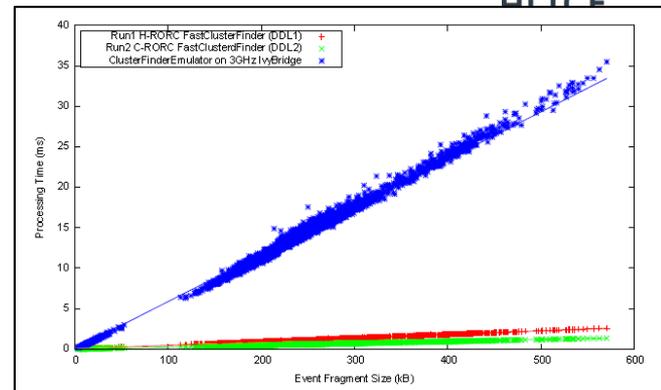
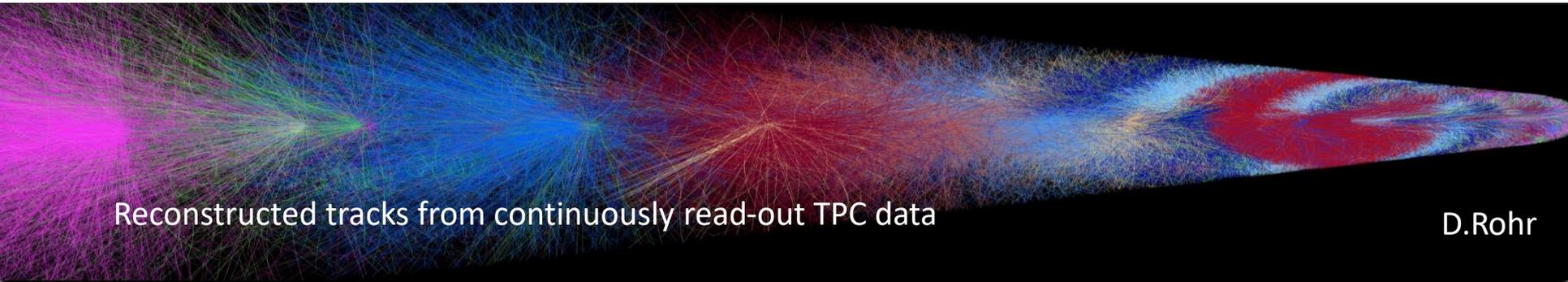


Figure 6.5: Tracking time of HLT TPC CA tracker on Nehalem CPU (6 Cores) and NVIDIA Fermi GPU.

# Online reconstruction

- Tracking of continuously read-out data is operational, performance equivalent or better than the current offline reconstruction.
- Synchronous phase requirement within the envelope
- Demonstrated compression ratio compatible with the target



Reconstructed tracks from continuously read-out TPC data

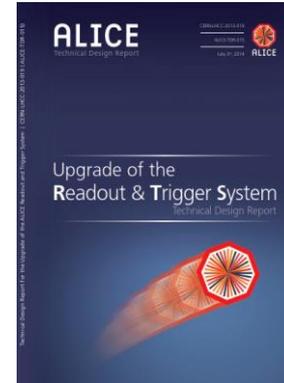
D.Rohr

# Conclusion

- After close to 10 years of successful operation, ALICE is performing a major upgrade to use the increased LHC luminosity after LS2
- This upgrade comes with new requirements for the electronics and the computing system
- Most ALICE detectors will now be continuously read out
- The reduction of the data volume is based on online calibration and reconstruction

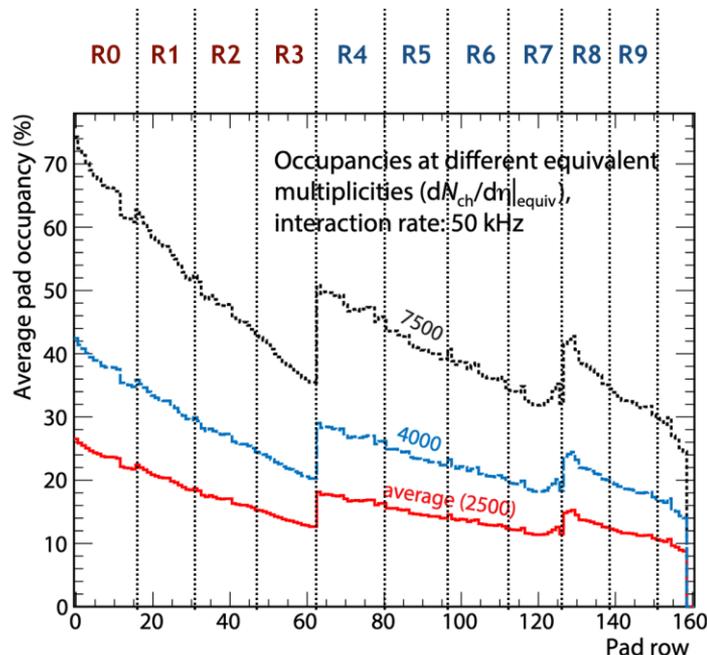
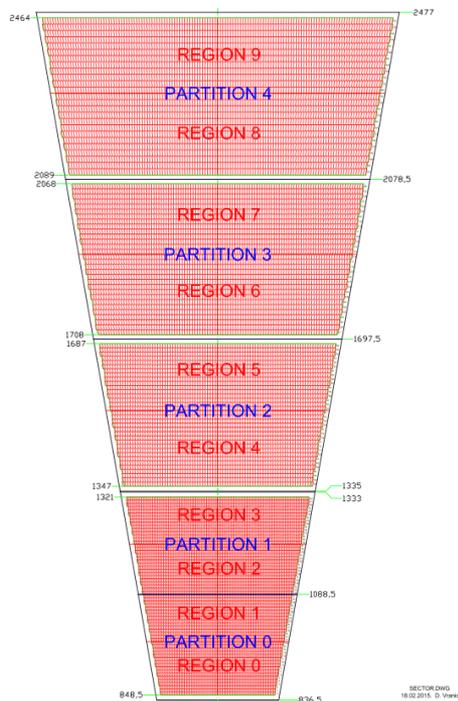
# Supporting documents

- Upgrade of the Readout and Trigger System TDR  
<https://cds.cern.ch/record/1603472/files/ALICE-TDR-015.pdf>
- Upgrade of the Online-Offline computing system TDR  
<https://cds.cern.ch/record/2011297/files/ALICE-TDR-019.pdf>
- The detector read-out in ALICE during Run 3 and 4, July 2016  
[https://svnweb.cern.ch/world/wsvn/alicetdrun3/Notes/Run34SystemNote/detector-read-alice/ALICErun34\\_readout.pdf](https://svnweb.cern.ch/world/wsvn/alicetdrun3/Notes/Run34SystemNote/detector-read-alice/ALICErun34_readout.pdf)



# Update on requirements

## TPC data flow



Thanks to  
Torsten Alt

Figure 6.3: Expected average occupancies within a given time window for equivalent multiplicities of  $dN_{ch}/d\eta|_{equiv} = 2500$ , 4000 and 7500. The data is extrapolated using measured occupancies in isolated (no pileup) events recorded in 2010.

# Update on requirements

## TPC data flow

	Start Row	End Row	Start Pad	End Pad	NumPads	MaxRate [Gb/s]	Occ 2500	DataRate 2500 [Gb/s]	Occ 4000	DataRate 4000 [Gb/s]	Occ 7500	DataRate 7500 [Gb/s]
R0	0	16	0	1199	1200	60	25	15	38	22,8	68	40,8
R1	16	31	1200	2399	1200	60	20	12	32	19,2	56	33,6
R2	32	47	2400	3839	1440	72	18	12,96	27	19,44	47	33,84
R3	48	62	3840	5279	1440	72	15	10,8	23	16,56	39	28,08
R4	63	80	5280	6719	1440	72	17	12,24	27	19,44	48	34,56
R5	81	96	6720	8159	1440	72	15	10,8	24	17,28	42	30,24
R6	97	112	8160	9759	1600	80	14	11,2	21	16,8	38	30,4
R7	113	126	9760	11359	1600	80	12	9,6	19	15,2	34	27,2
R8	127	139	11360	12959	1600	80	14	11,2	22	17,6	38	30,4
R9	140	151	12960	14559	1600	80	12	9,6	19	15,2	33	26,4
Sector								115,4		179,52		315,52
TPC [GB/s]								519,3		807,84		1419,84

Thanks to  
Torsten Alt

Add 10-15%  
for ZS  
overhead



# Update on requirements

## TPC data flow

Two options for the read-out farm (FLPs)  
for the TPC

### Option A (144 FLPs)

4 FLPs per TPC sector  
2+2+3+3 CRUs

TPC CRU @ 12.2 Gb/s

TPC CRU @ 10.8 Gb/s

TPC CRU @ 9.6 Gb/s

**Total: 37.5 Gb/s  
(incl 15% over.)**

### Option B (120 FLPs)

10 FLPs per 3 TPC sectors  
3 CRUs per FLP

TPC CRU @ 12.2 Gb/s

TPC CRU @ 12.2 Gb/s

TPC CRU @ 12.2 Gb/s

**Total: 42.1 Gb/s  
(incl 15% over.)**

# Computing Model

