



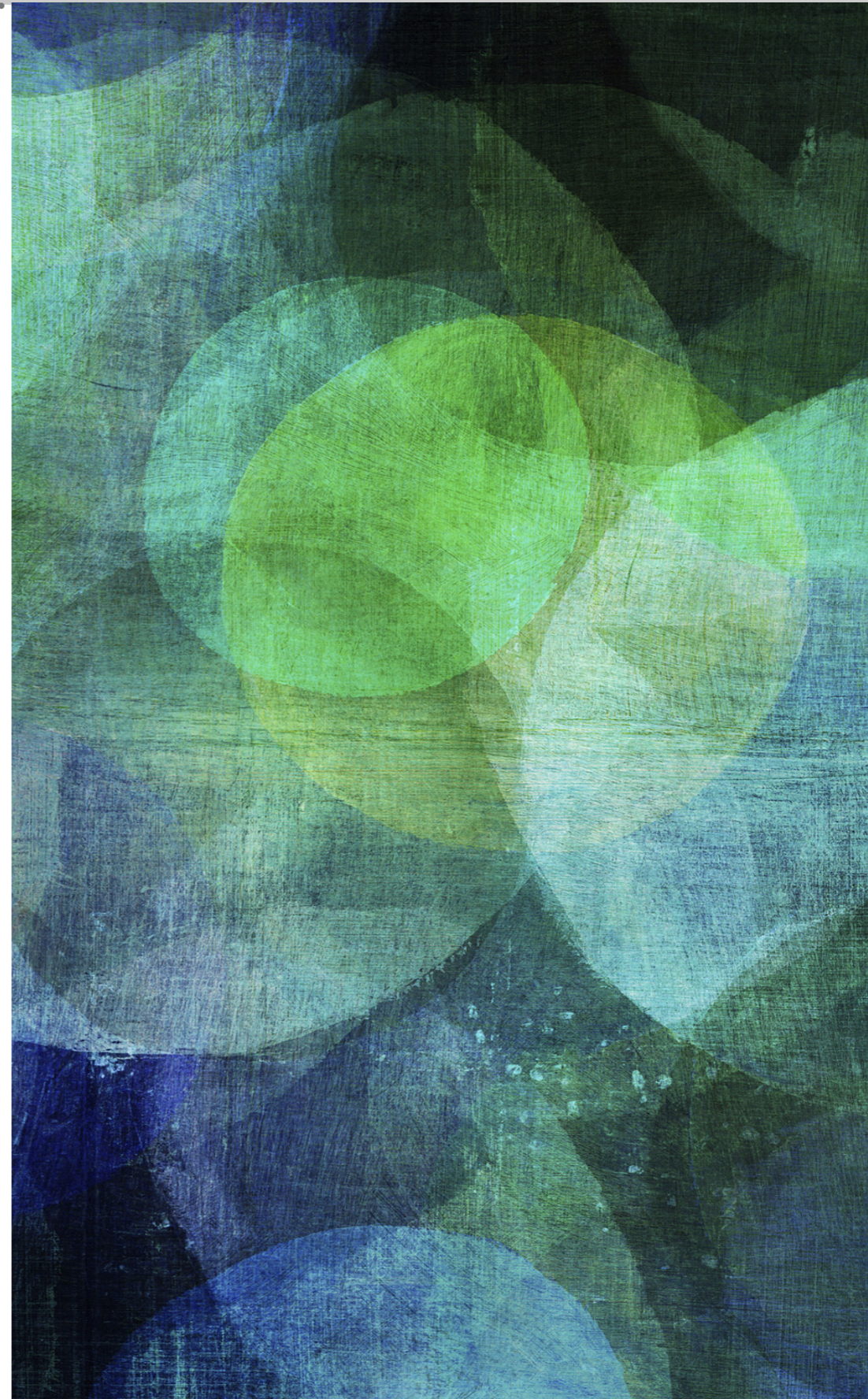
# STRATEGIE E UPGRADE DEL TDAQ NEGLI ESPERIMENTI LHC

*F. Pastore (Royal Holloway Un. of London)*  
*francesca.pastore@cern.ch*



# THE CONTENTS OF THIS LECTURE

- ➔ **Triggering e data taking a LHC**
- ➔ **Strategie per il futuro High-Lumi LHC**
- ➔ **Quattro esperimenti, quattro differenti approcci e sviluppi di architetture TDAQ**
- ➔ **E qualche esempio interessante**

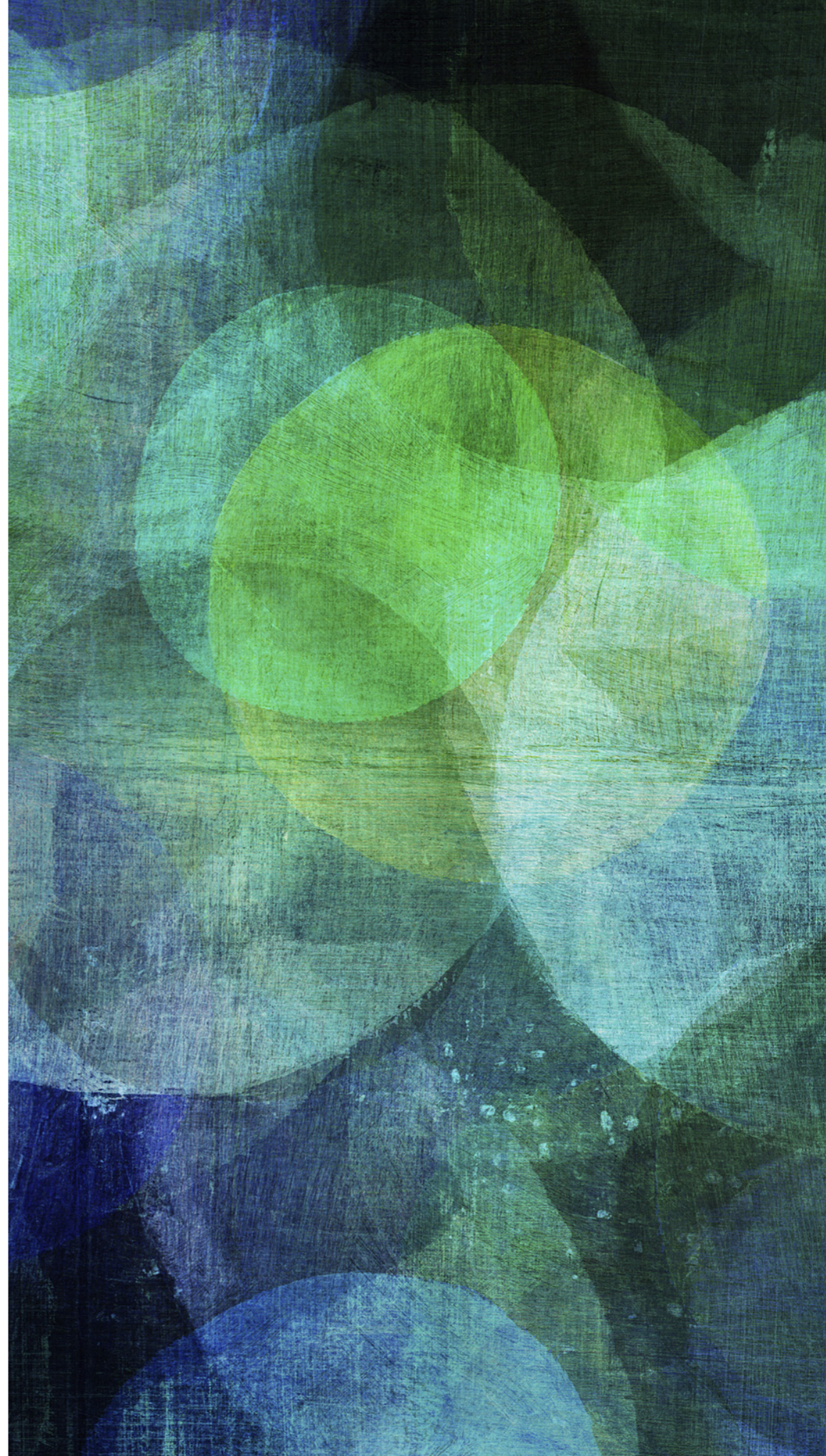




# IL TRIGGER E LA PRESA DATI AD LHC

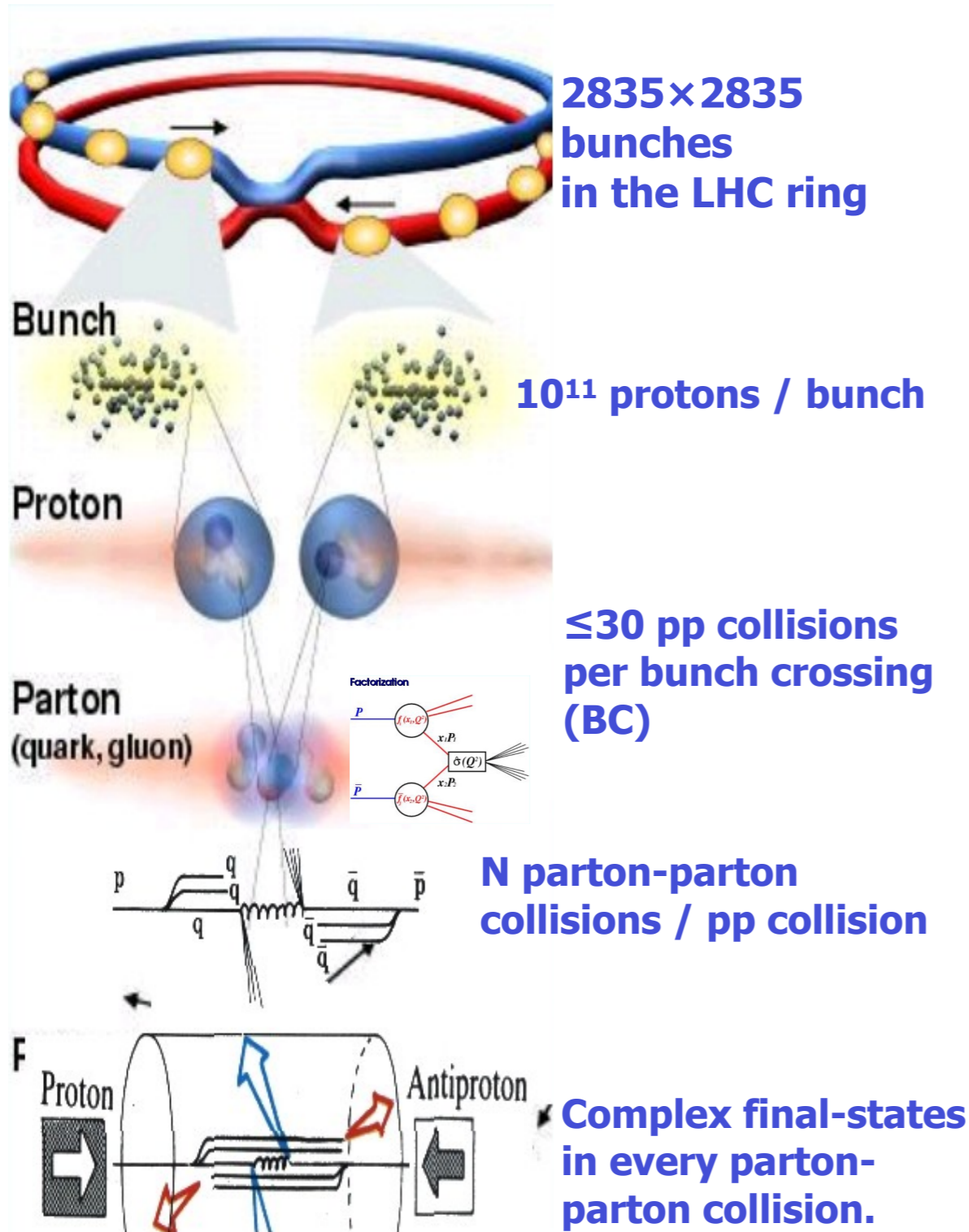
---

*TDAQ for large discovery  
experiments*





# LHC ENGINE AND ITS CHALLENGES



$$E_{\text{cms}} = 14 \text{ TeV}$$

$$L = 10^{34} / \text{cm}^2 \text{ s}$$

$$\text{BC clock} = 40 \text{ MHz}$$

Search for rare events overwhelmed in abundant low-energy particles

Three major challenges for T/DAQ

→ Face High Luminosity:

- fast electronics, to resolve in time
- fine granularity detector, to resolve in space ⇒ high data volume

→ Search for rare physics:

- high rejection or large data collection

→ Be radiation resistant:

- very costly for electronics ⇒ survive up to 100 Mrad = 1 MGy



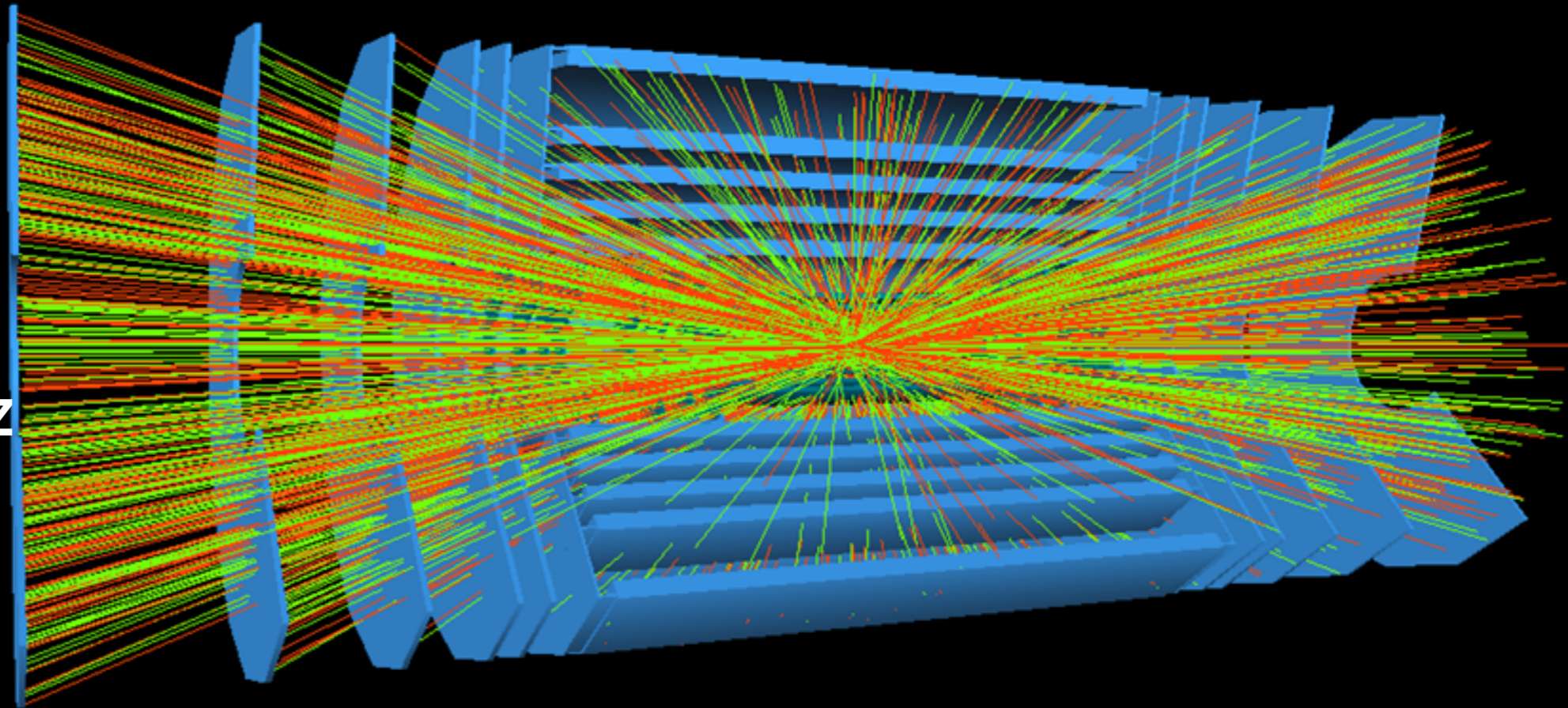
# LHC DATA DELUGE

p-p collisions

$E_{\text{cms}} = 13\text{-}14 \text{ TeV}$

$L = 10^{34} / \text{cm}^2 \text{ s}$

BC clock = 40 MHz



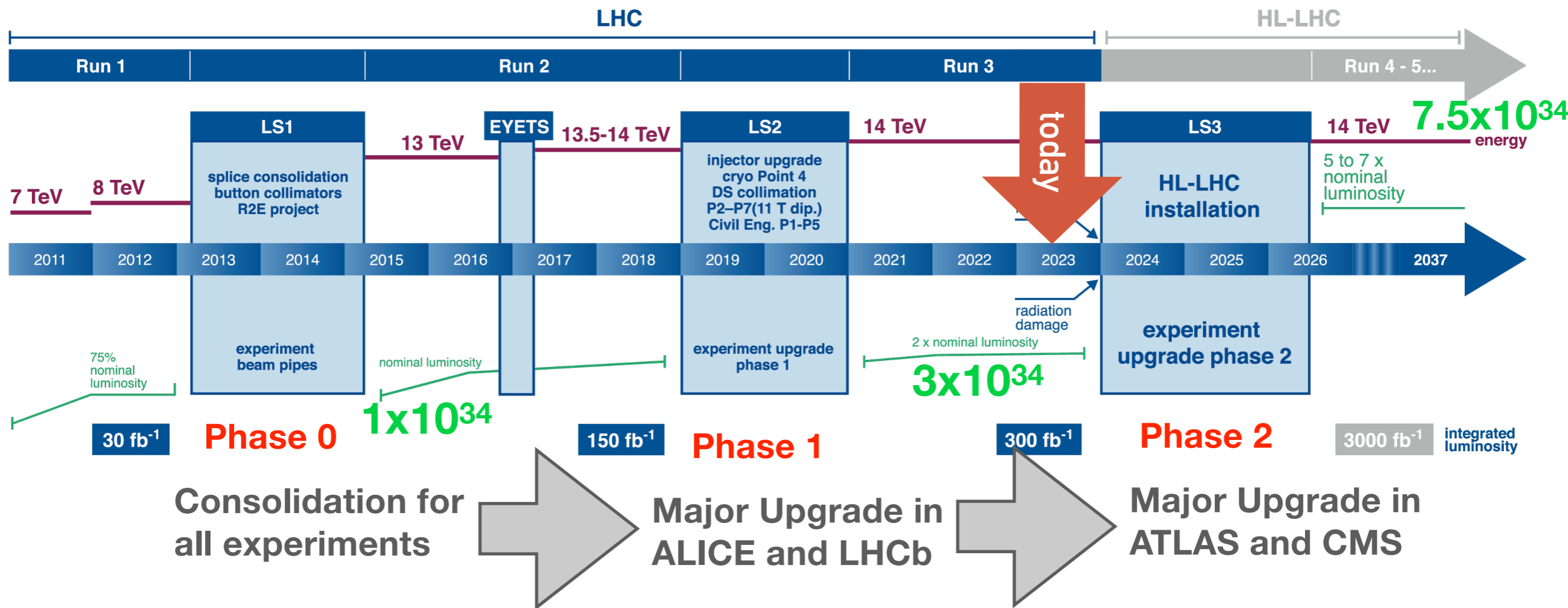
- High Luminosity with collisions close in time and space (1 collision/25ns)
  - abundant data in time and space
- Search for rare physics from hadronic collisions:
  - to store all the possibly relevant data is UNREALISTIC and often UNDESIRABLE
- Three approaches are possible:
  - Reduce the amount of data (packing and/or filtering)
  - Have faster data transmission and processing
  - Both!



# LHC BECOMING IMPRESSIVELY LUMINOUS

European Council (2014): "CERN is the strong European focal point for particle physics in next 20 years"

## LHC / HL-LHC Plan



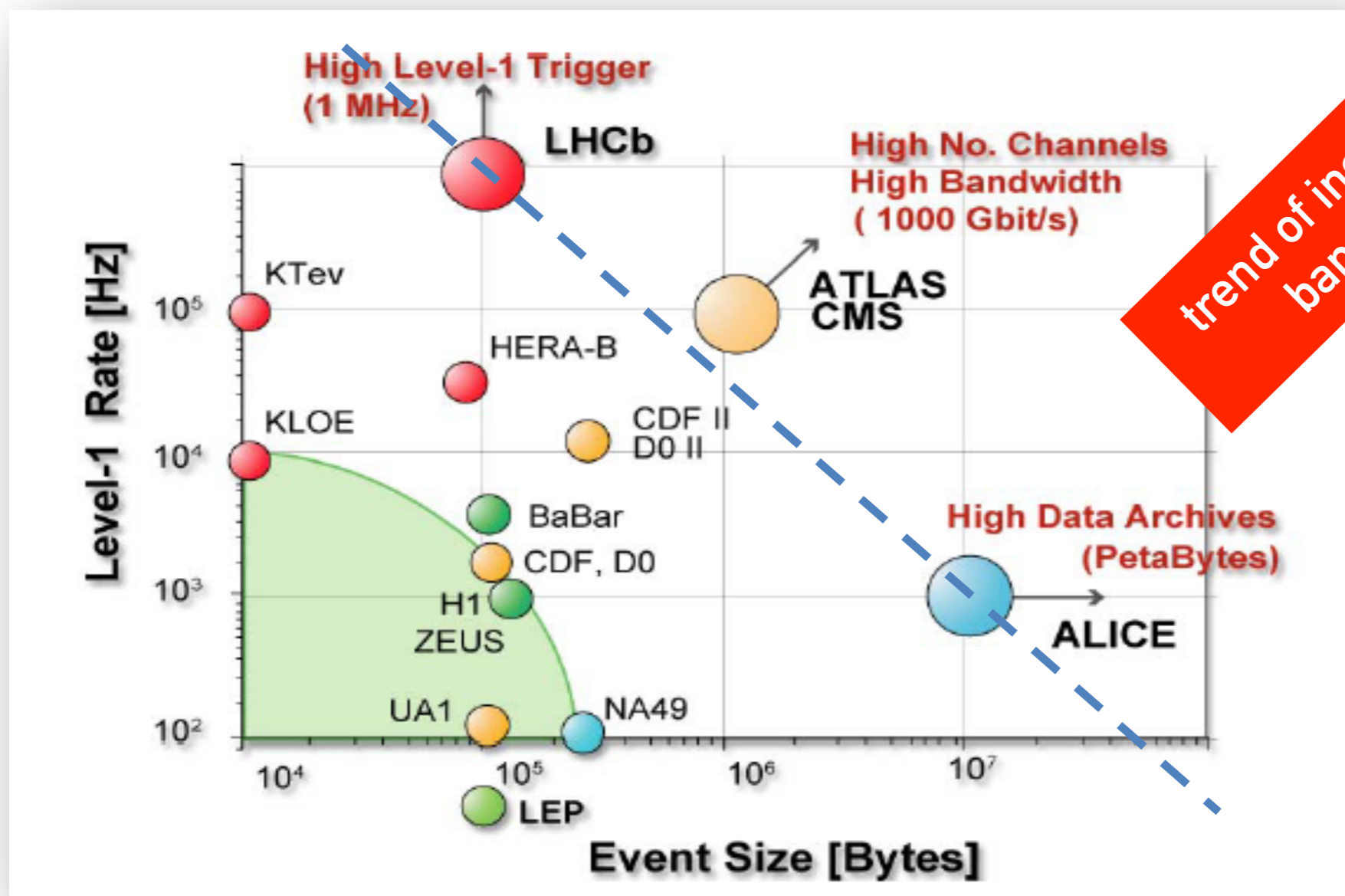
→ Experiments go beyond design specifications ( $1 \times 10^{34} / \text{cm}^2 \text{s}$ ) and need upgrade as well, to improve or at least maintain the design performance



# READOUT AND DAQ THROUGHPUTS

$$R_{DAQ} = R_T^{max} \times S_E$$

faster L1 electronics



trend of increasing bandwidth

## ATLAS/CMS

Data to Process:

100 kHz \* 1 MB = 100 GB/s

Data to Store:

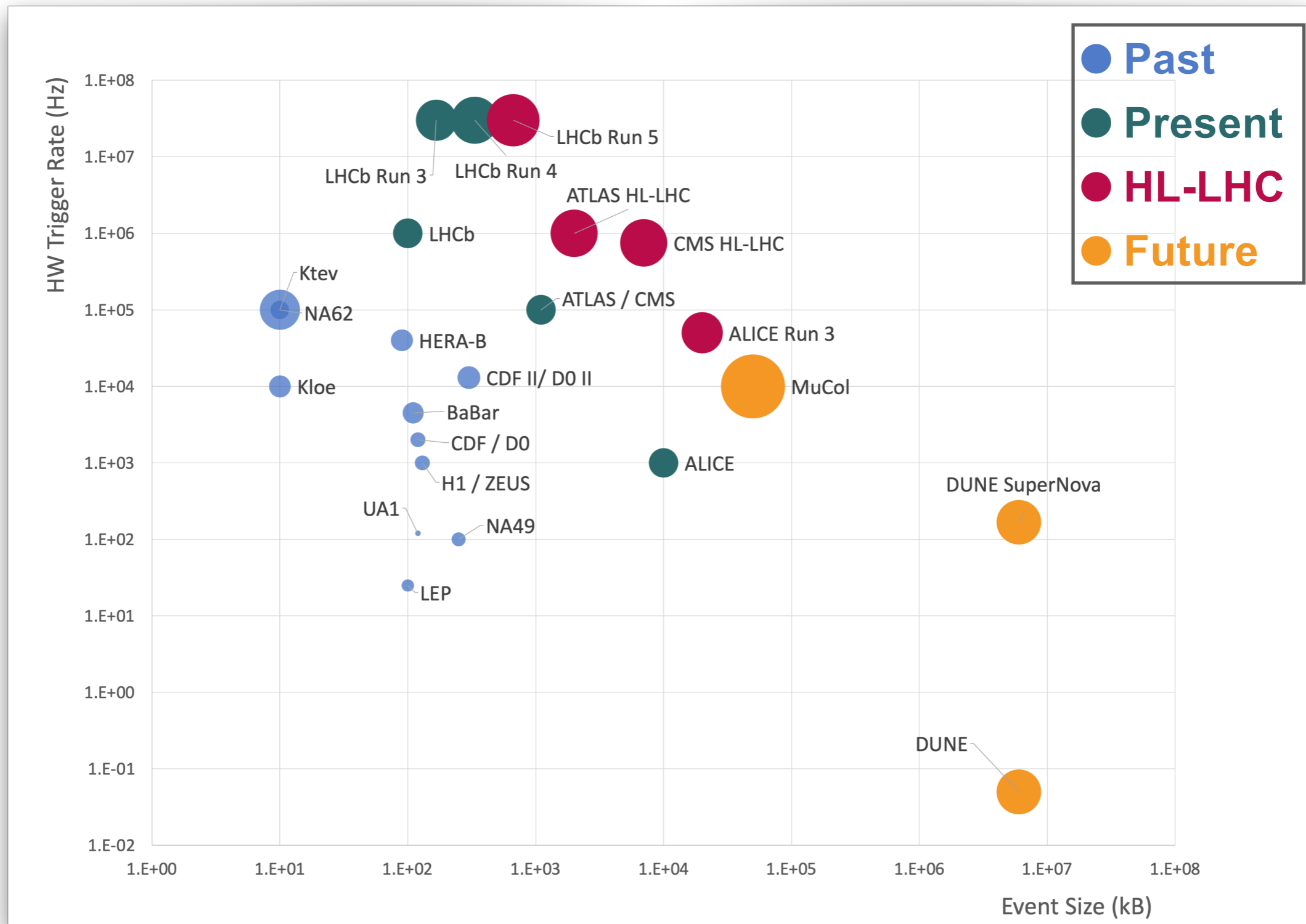
~ 1 PB / year / experiment

more channels, more complex events

As the data volumes and rates increase, new architectures need to be developed



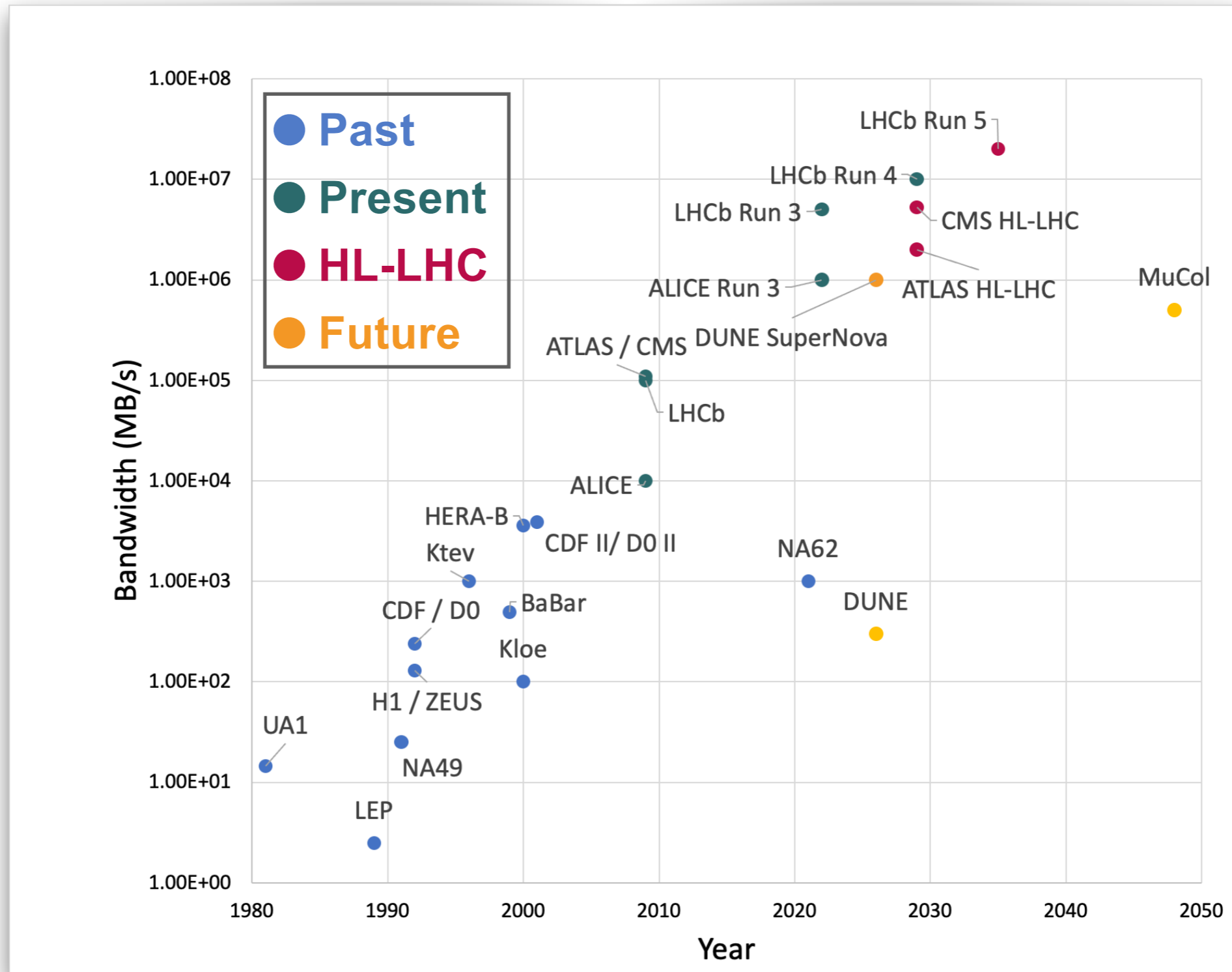
# UPDATED FIGURE!



Courtesy of A. Cerri



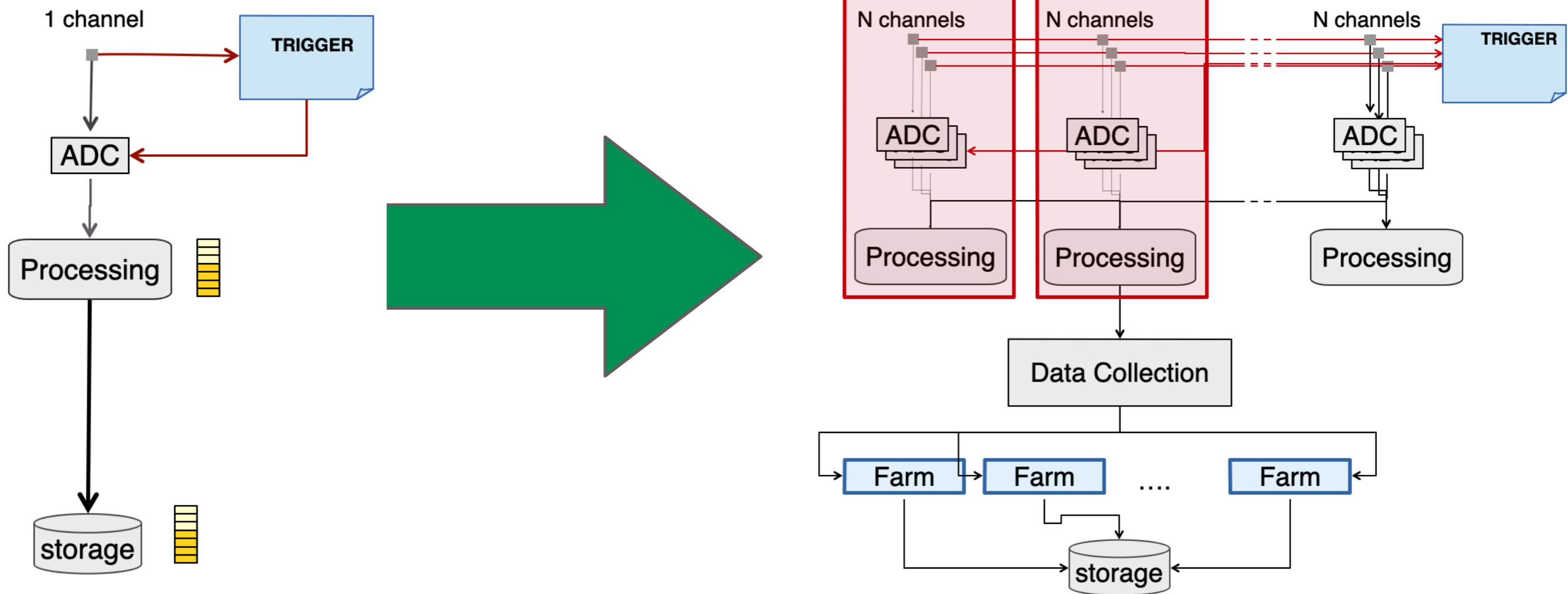
# LOOKING FOR MORE DATA IN THE FUTURE



Courtesy of A. Cerri



# RECAP ON T/DAQ SYSTEMS AND SCALING



- ➔ More Rate  $\implies$  **More buffers**
- ➔ More channels  $\implies$  Parallelism  $\implies$  **Segmented Readout (and trigger)**
- ➔ More Front-end elements  $\implies$  **Multiple processing units (local data)**
  - ➔ Decouple storage from processing unit (PU)  $\implies$  **Data collection**
- ➔ Extend trigger latency  $\implies$  **Multi-level trigger**
- ➔ Avoid dead-time and back-pressure  $\implies$  **Dataflow control**

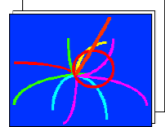


# MANY PLAYERS, COMPLEX TDAQ ARCHITECTURES

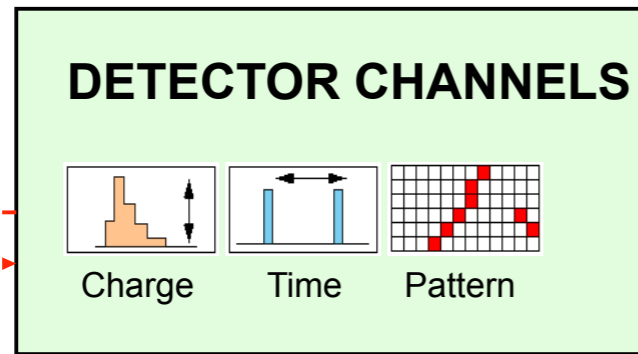
## Buffering and parallelism

Maximum 1-2% deadtime

40 MHz  
COLLISION RATE



Level-1

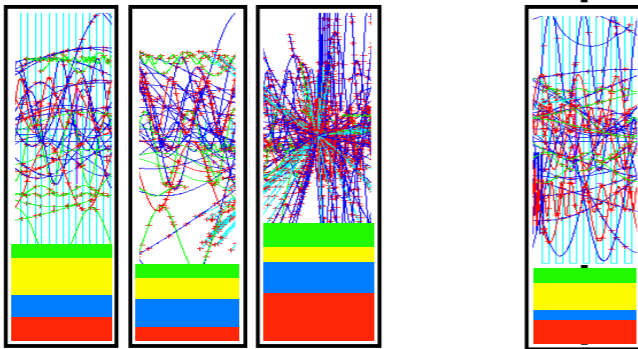


High speed electronics



- Level-1 triggers**
- ➔ Set max Readout rate
  - ➔ Hardware, synchronous
  - ➔ Readout parallelism
  - ➔ Latency ~  $\mu\text{sec/event}$

Readout Buffers

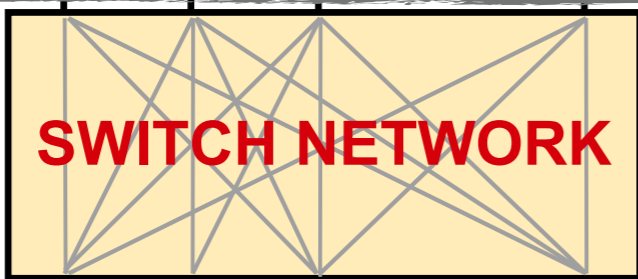


Readout links and buffering

Readout

L1/Readout

Event building

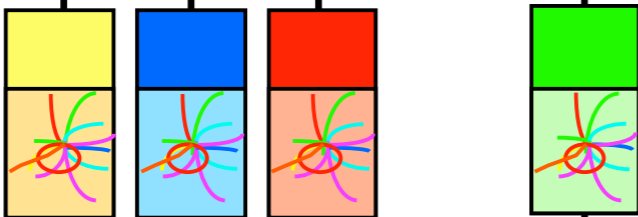


Large data network with dedicated technology

DAQ

HLT/DAQ

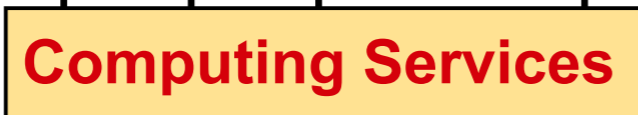
Event filtering



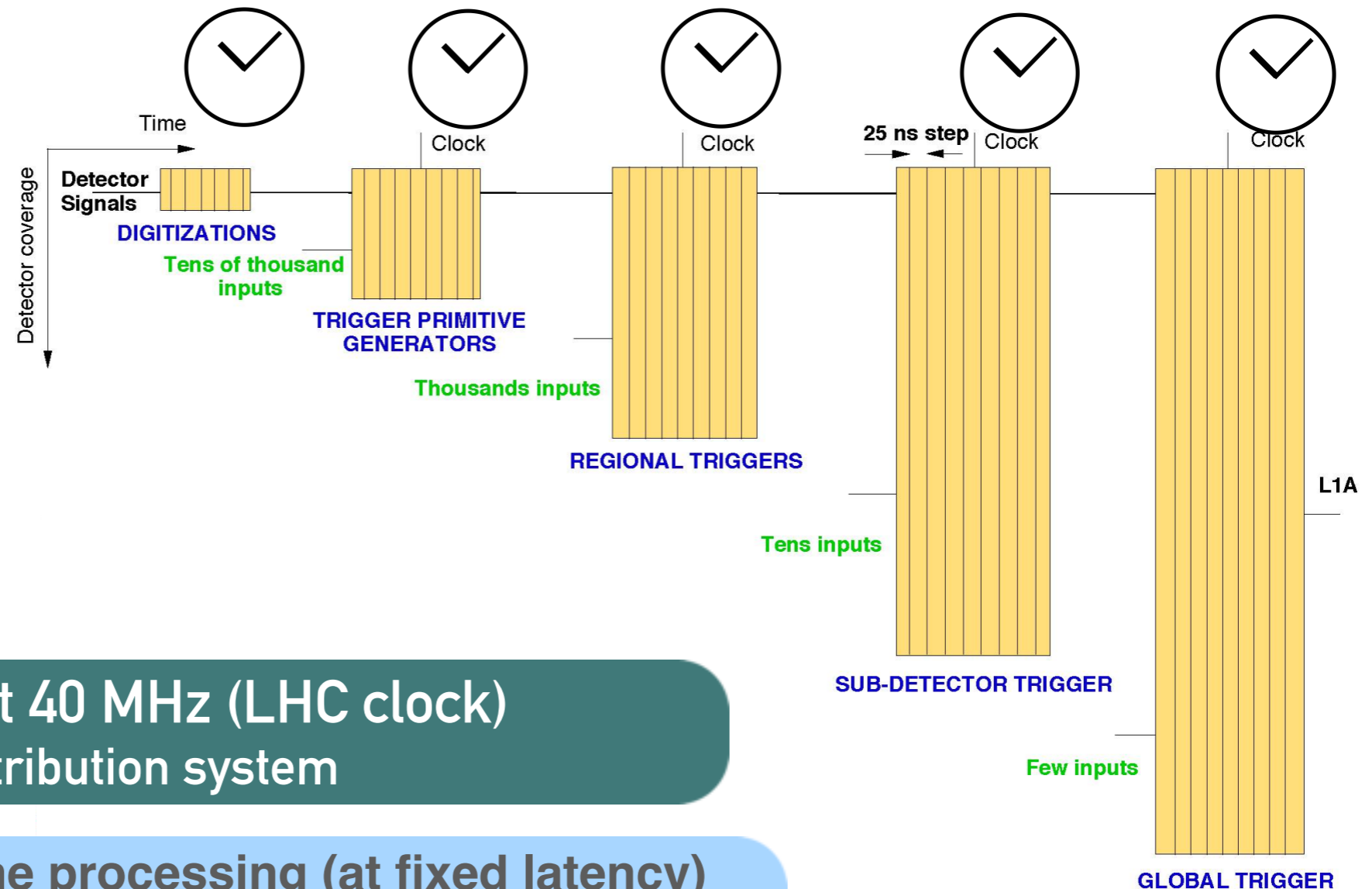
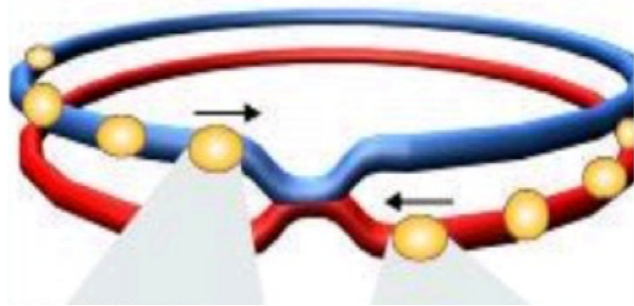
Dedicated PC farms

- Higher level triggers**
- ➔ Set max storage rate
  - ➔ Software, asynchronous
  - ➔ Event parallelism
  - ➔ Latency < 1 sec/event

Petabyte archive



# LEVEL-1 TRIGGER REQUIREMENTS



Full synchronisation at 40 MHz (LHC clock)

➤ large optical time distribution system

- ➔ Synchronous: pipeline processing (at fixed latency)
- ➔ Low latency (fast processing and high speed links)
- ➔ Scalable
- ➔ Massively parallel
- ➔ Bunch Crossing identification capability

ALICE	No pipeline
ATLAS	2.5 $\mu$ s
CMS	3 $\mu$ s
LHCb	4 $\mu$ s

## Fast, robust electronics

Latency dominated by cable/transmission delay



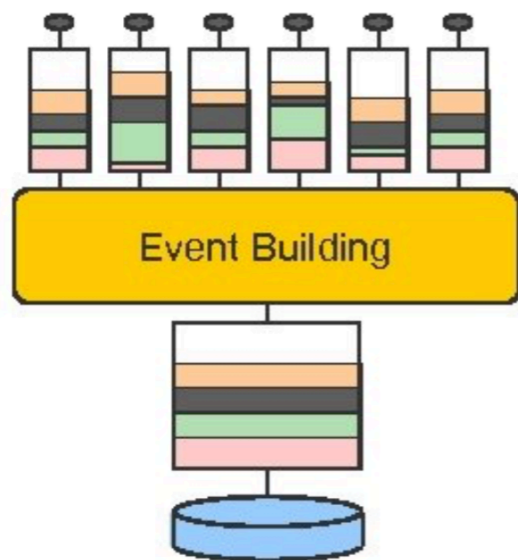
# HLT/DAQ REQUIREMENTS

Data sources

Event Fragments

Full Events

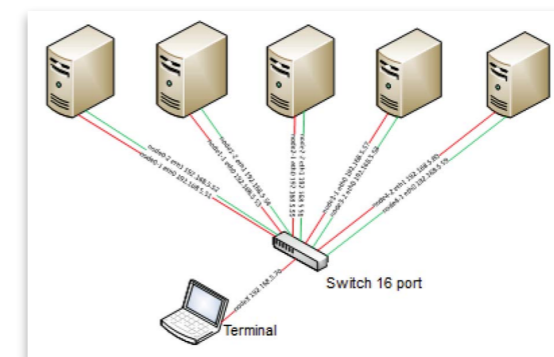
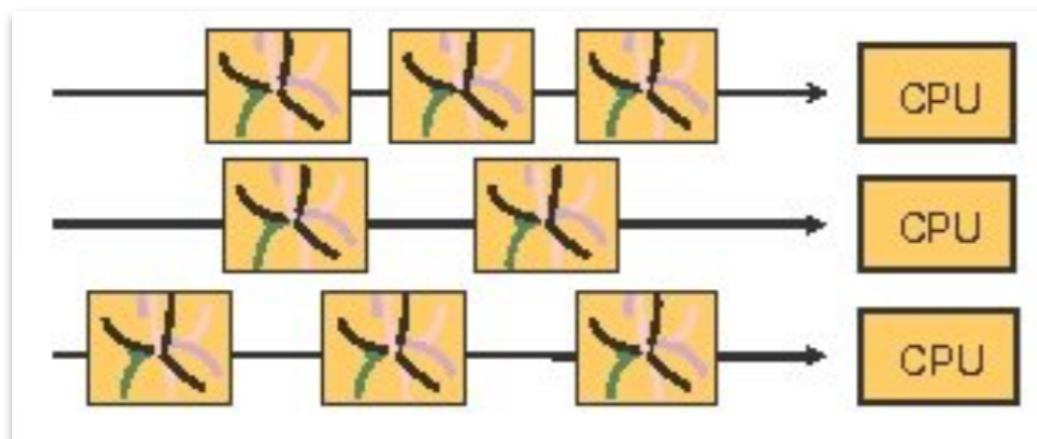
Data storage



- ➔ Robustness and redundancy
- ➔ Scalability to adapt to Luminosity, detectors,...
- ➔ Flexibility (10-years experiments)
- ➔ Based on commercial products
- ➔ Limited cost

**Prefer use of PCs (linux based), Ethernet protocols, standard LAN, configurable devices**

- PC farms on networks for Event Building and Event Filtering (HLT)
  - farm processing: one event per processor (larger latency, but scalable)
  - additional networks regulates the CPU assignment and traffic

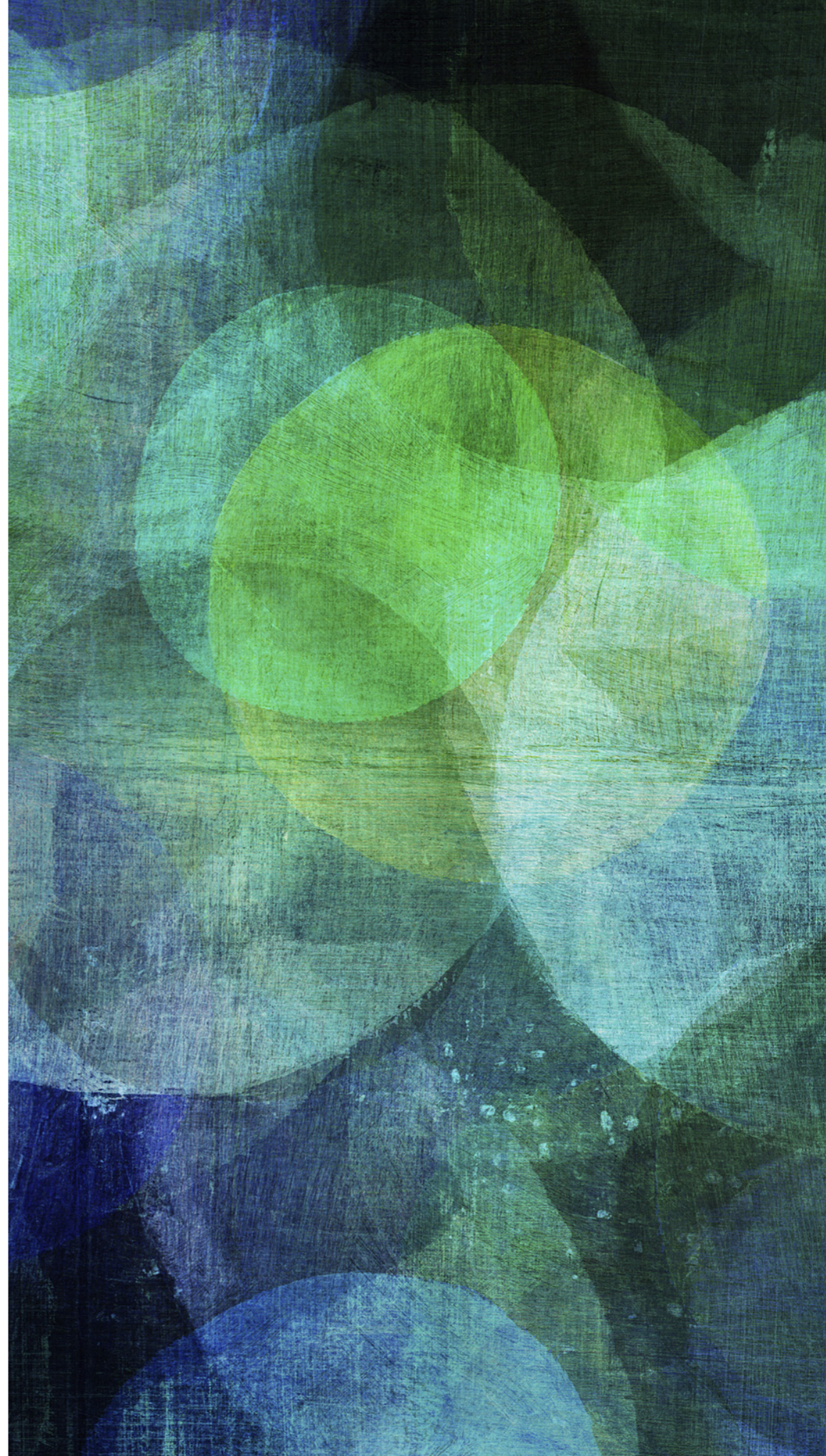


See S.Cittolin, DOI: 10.1098/rsta.2011.0464



**COME  
PREPARIAMO  
L'HIGH  
LUMINOSITY**

.....  
*What about ... tomorrow?*





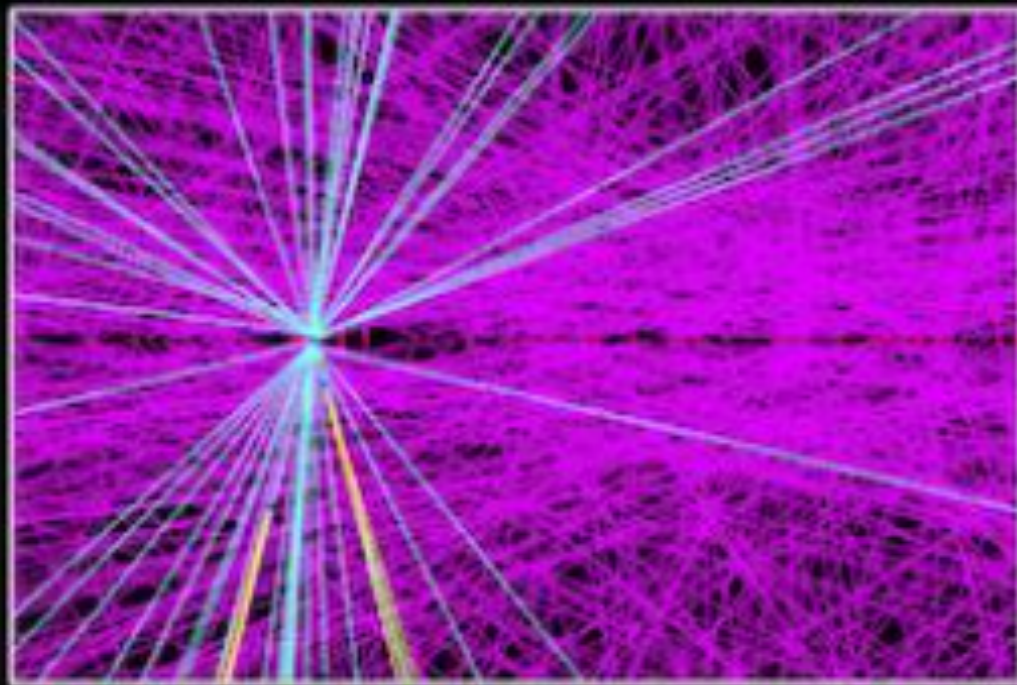
# ONE EVENT AT HIGH-LUMINOSITY ( $L=7.5 \times 10^{34}$ /CM<sup>2</sup>/S)

## Design Luminosity x7.5

- 200 collisions per bunch crossing (any 25 ns)
- ~ 10 000 particles per event
- Mostly low  $p_T$  particles due to low transfer energy interactions

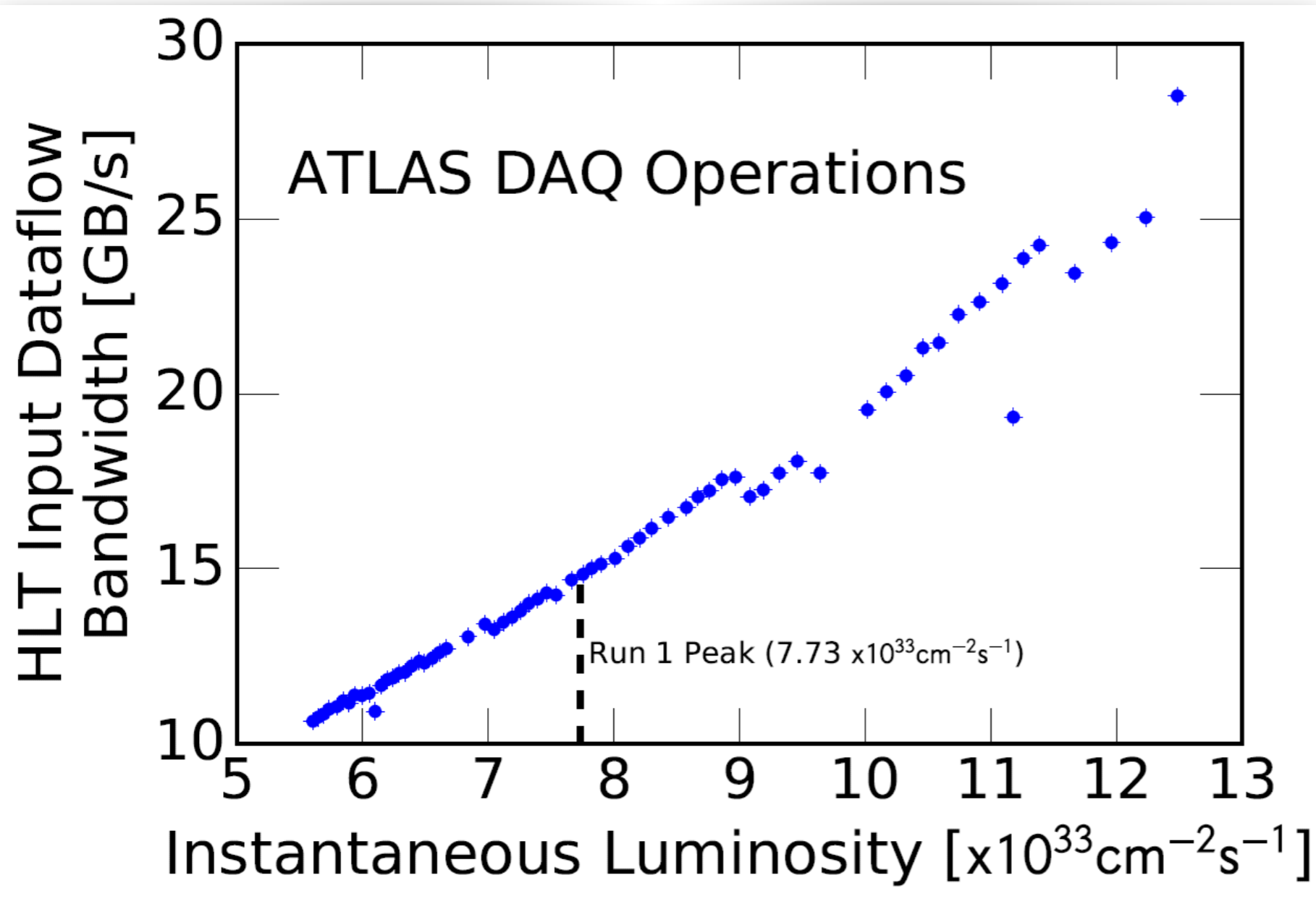


HL-LHC  $t\bar{t}$  event in ATLAS ITK  
at  $\langle\mu\rangle=200$



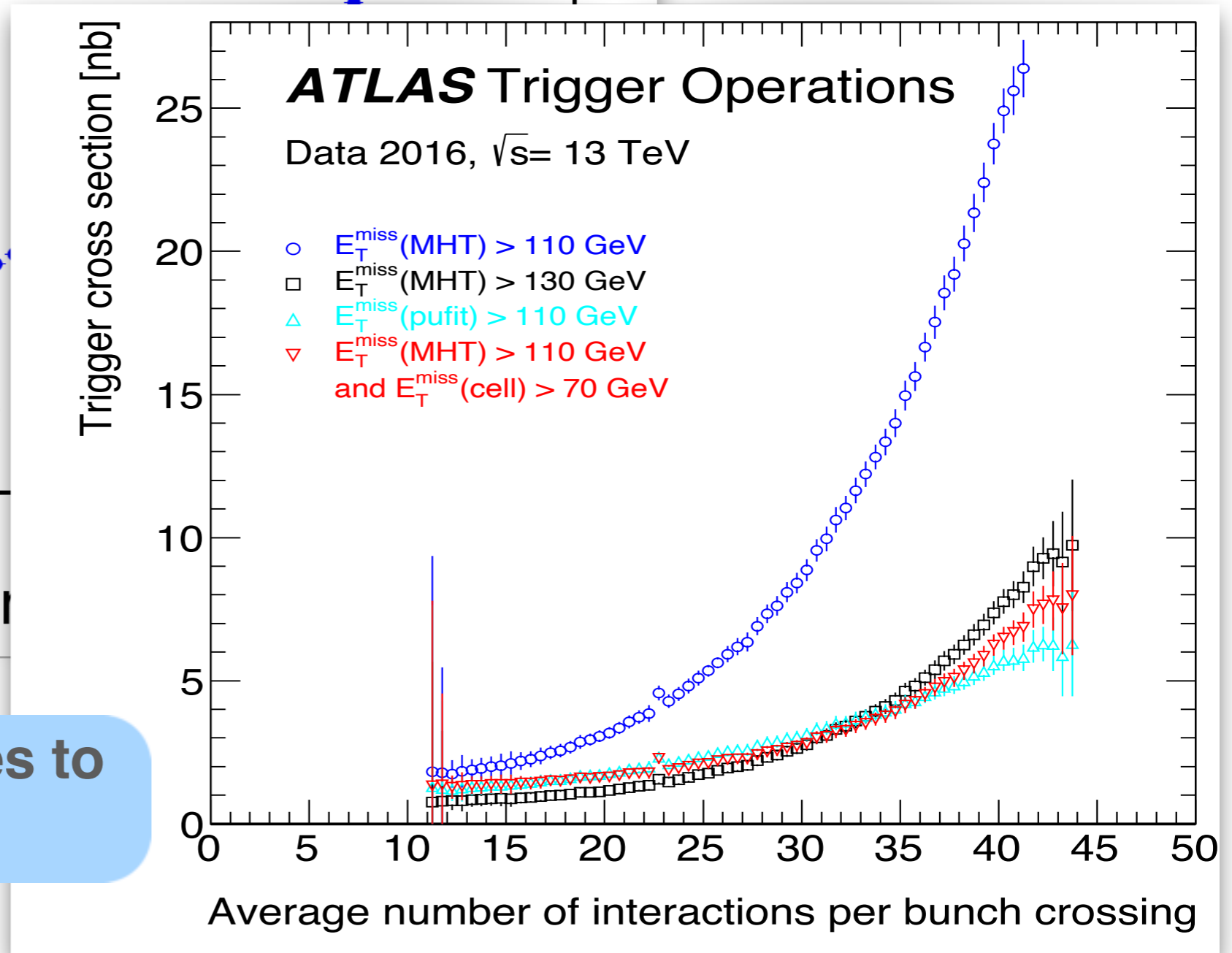
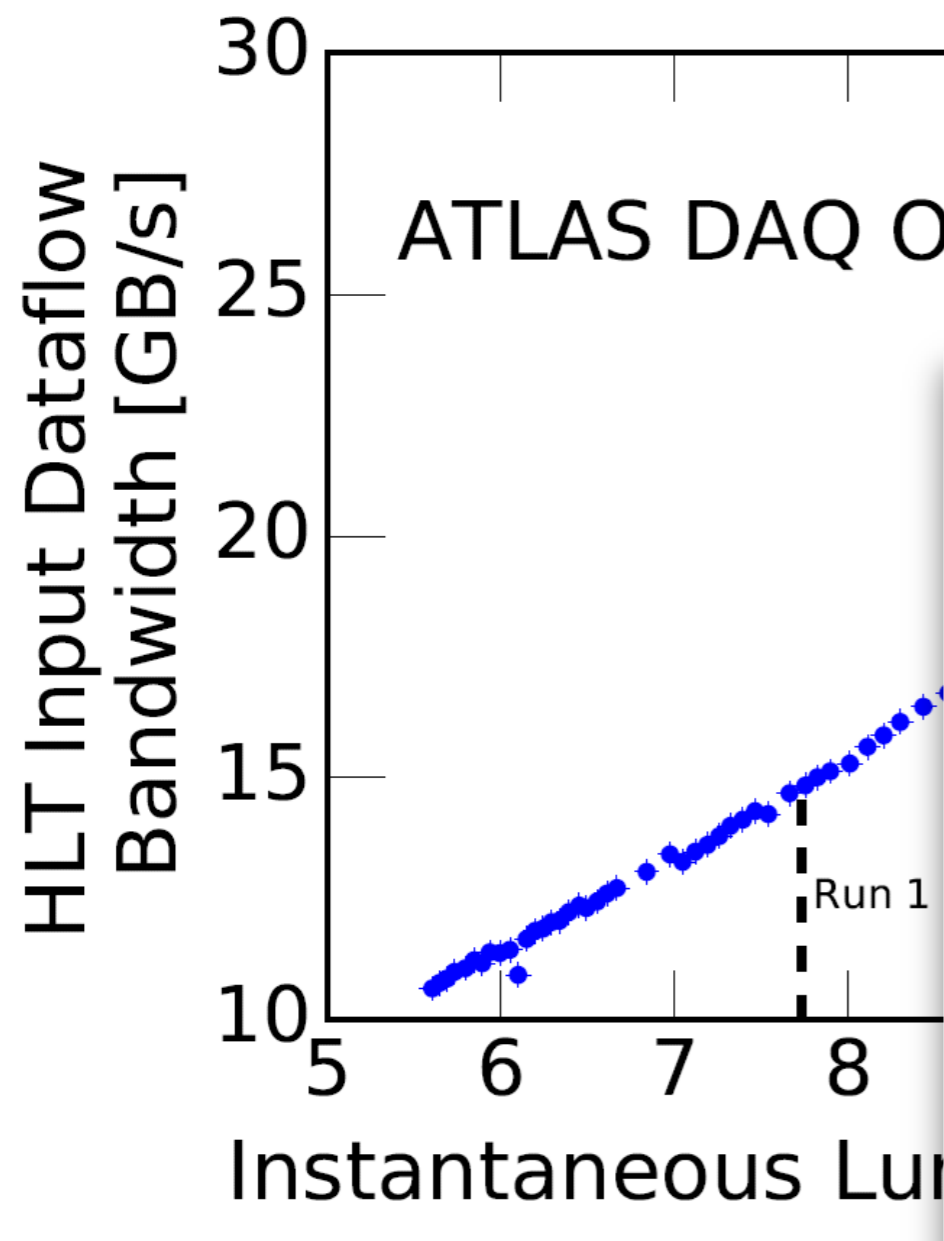
**Physics program for the future  
is towards more rare processes  
at the same energy scale**

# WHAT DO YOU EXPECT FOR THE FUTURE?





# WHAT DO YOU EXPECT FOR THE FUTURE?



Very large uncertainties to take into account!

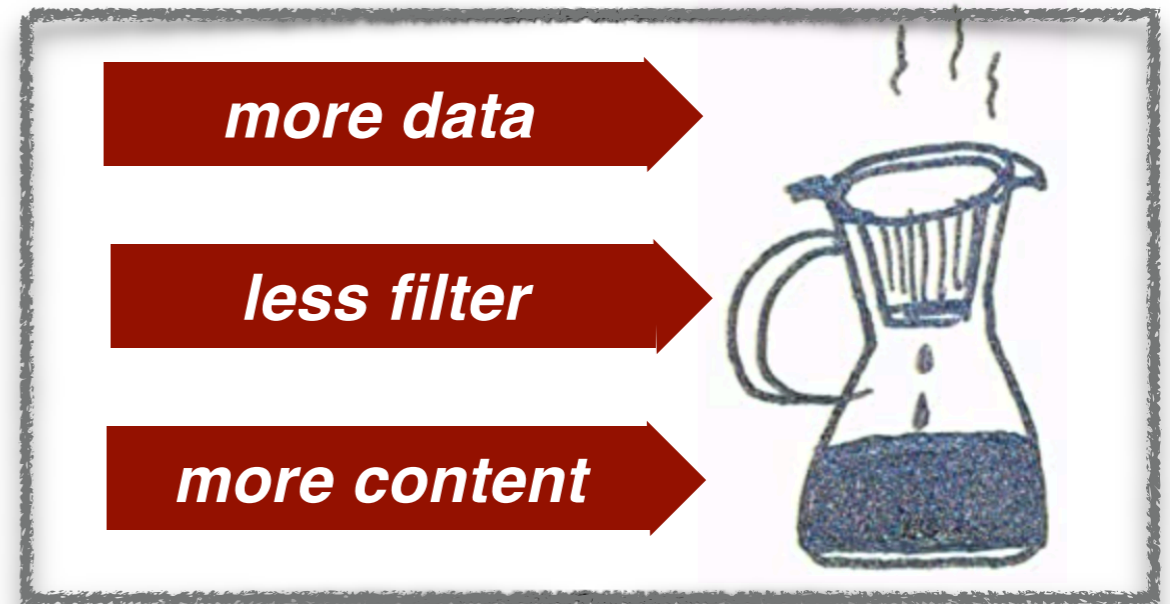
# ADDITIONAL COMPLICATION AT HL-LHC

**Luminosity x10, complexity x100: we cannot simply scale current approach**

## x10 higher Luminosity means...

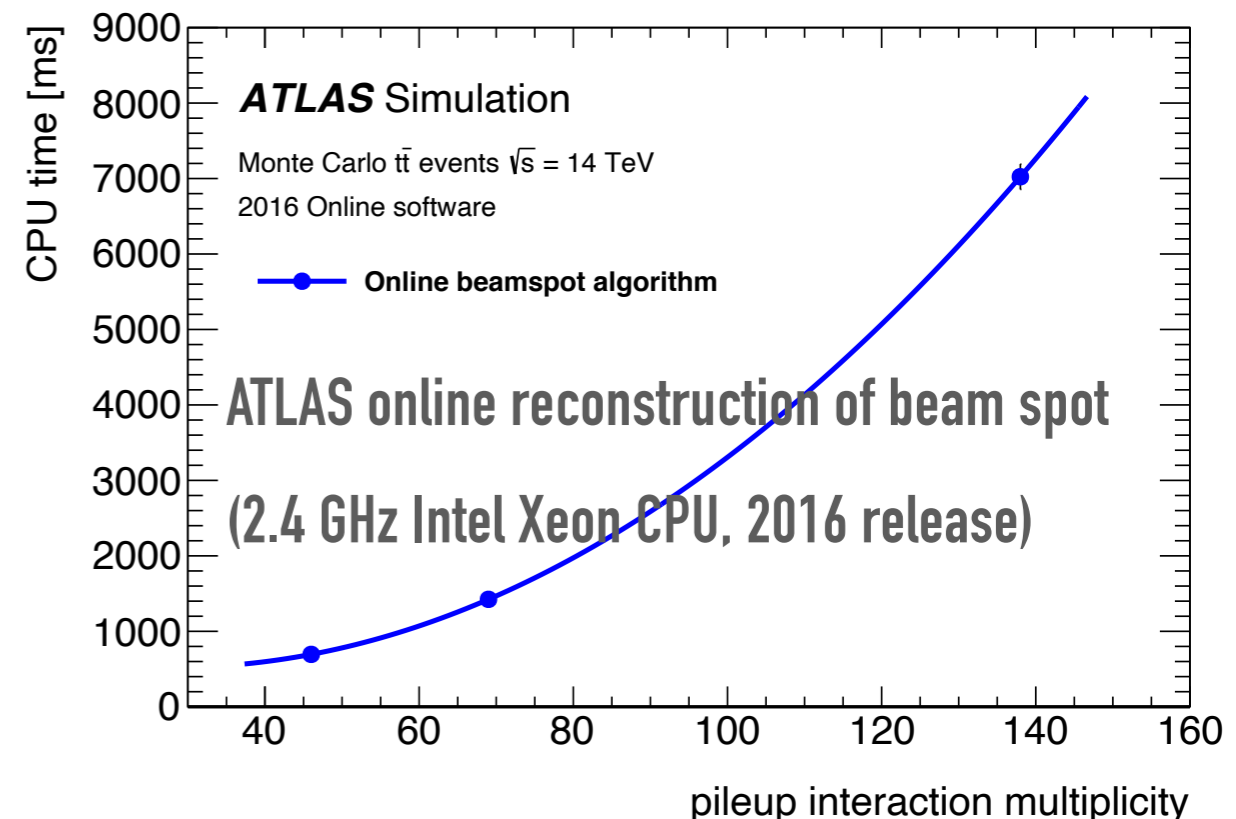
- ➔ **More interactions per BC (pile-up)**
  - ➔ Less rejection power (worse pattern recognition and resolution)
  - ➔ Larger event size
- ➔ **Larger data rates:**
  - ➔ FE readout rate @L1: 0.1  $\Rightarrow$  1 MHz
  - ➔ DAQ throughput: 1  $\Rightarrow$  50 Tbps

*ATLAS/CMS numbers*



## But cannot...

- ➔ **Increase trigger thresholds**
  - ➔ Need to maintain physics acceptance
- ➔ **Scale dataflow with Luminosity**
  - ➔ **H/W:** more parallelism  $\Rightarrow$  more links  $\Rightarrow$  more material and cost
  - ➔ **S/W:** processing time not linear  $\sim L$

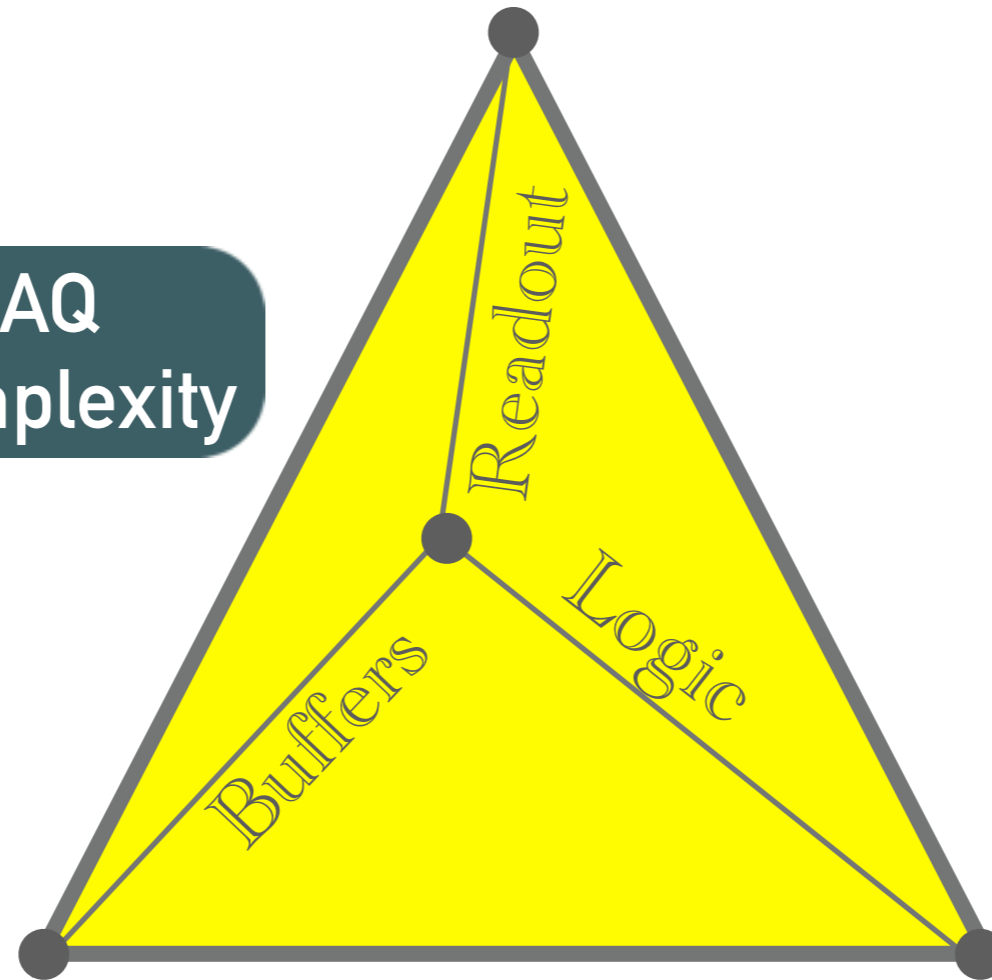




# BE SMARTER! INCREASE RESOLUTION FOR BETTER S/B

## Trigger-less DAQ

Tension between TDAQ architecture and FE complexity



High performance farms

Triggering detectors

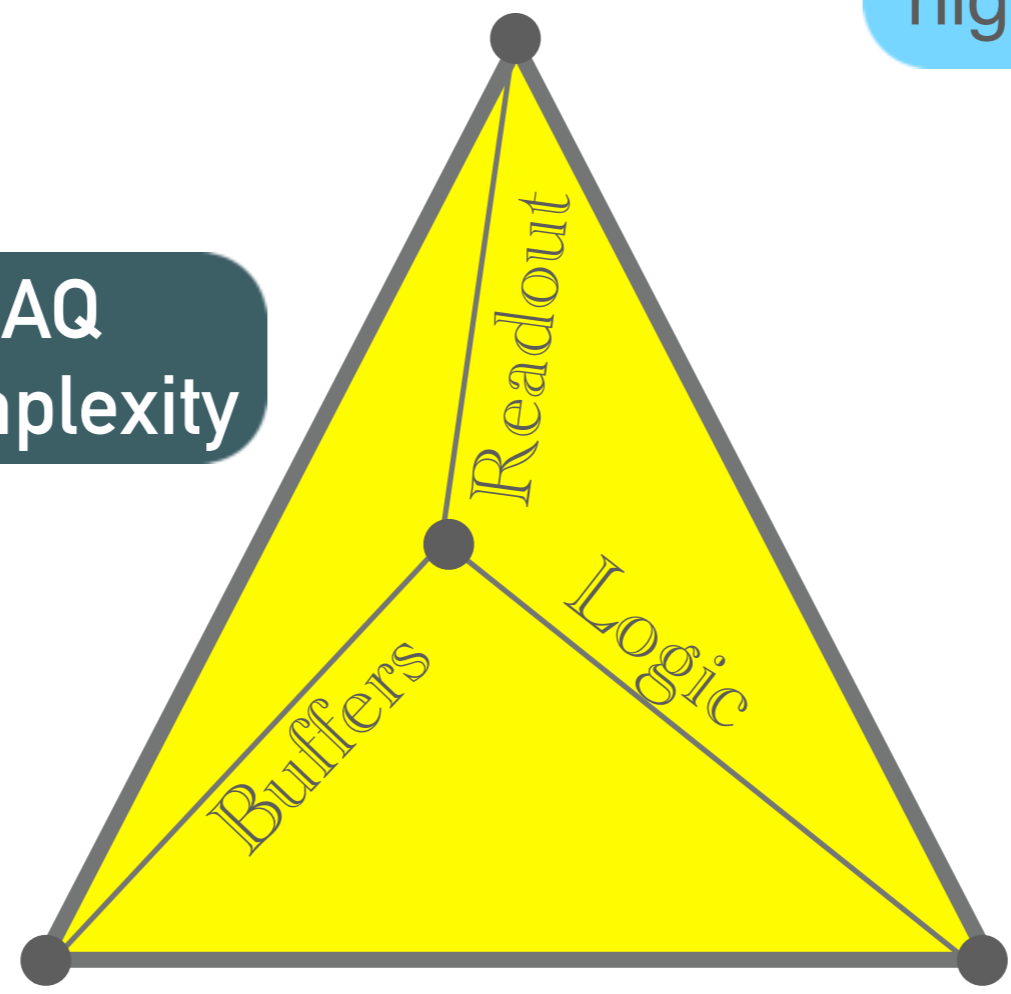
# BE SMARTER! INCREASE RESOLUTION FOR BETTER S/B

## What we do?

### Trigger-less DAQ

high detector granularity

Tension between TDAQ architecture and FE complexity



High performance farms

refine calibrations, as offline

Triggering detectors

complex ASIC logic



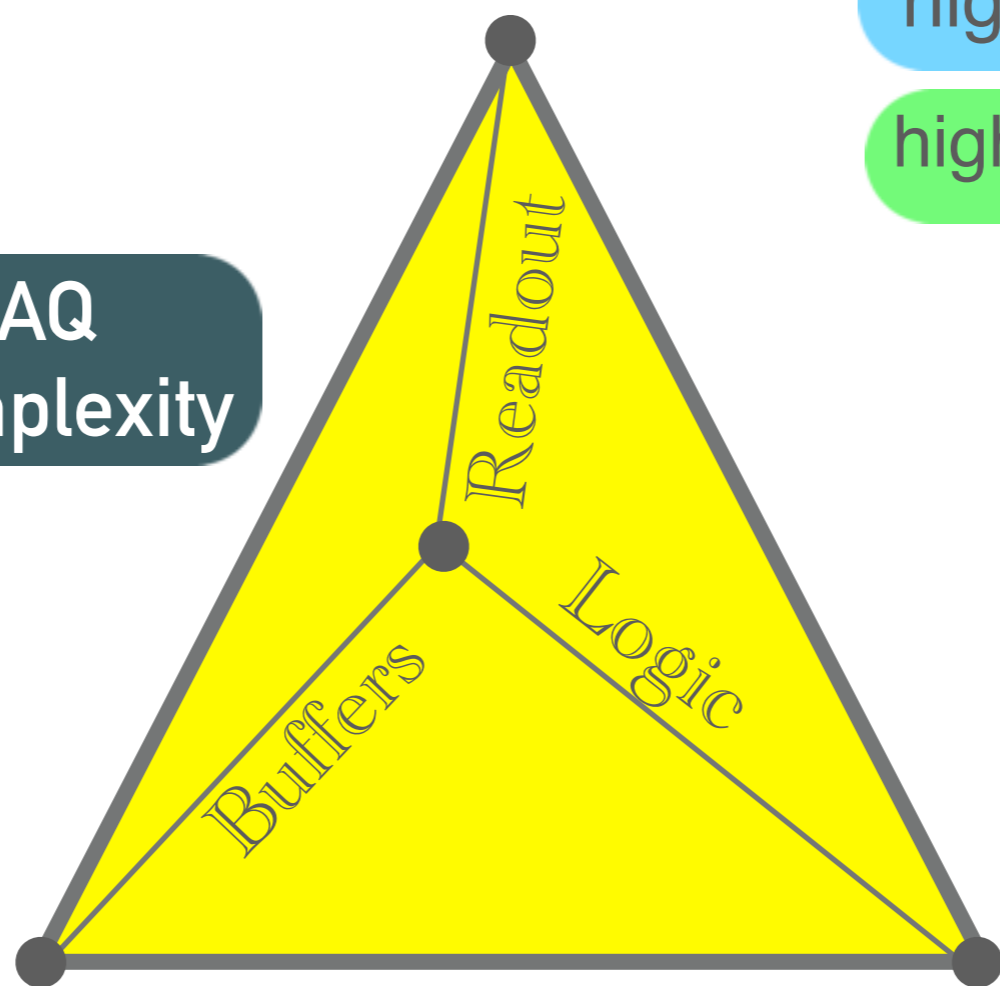
# BE SMARTER! INCREASE RESOLUTION FOR BETTER S/B

What we do?

How?

Tension between TDAQ architecture and FE complexity

## Trigger-less DAQ



high detector granularity

high speed electronics/links

## High performance farms

refine calibrations, as offline

large buffers, long latency

## Triggering detectors

complex ASIC logic

trigger-driven design

# BE SMARTER! INCREASE RESOLUTION FOR BETTER S/B

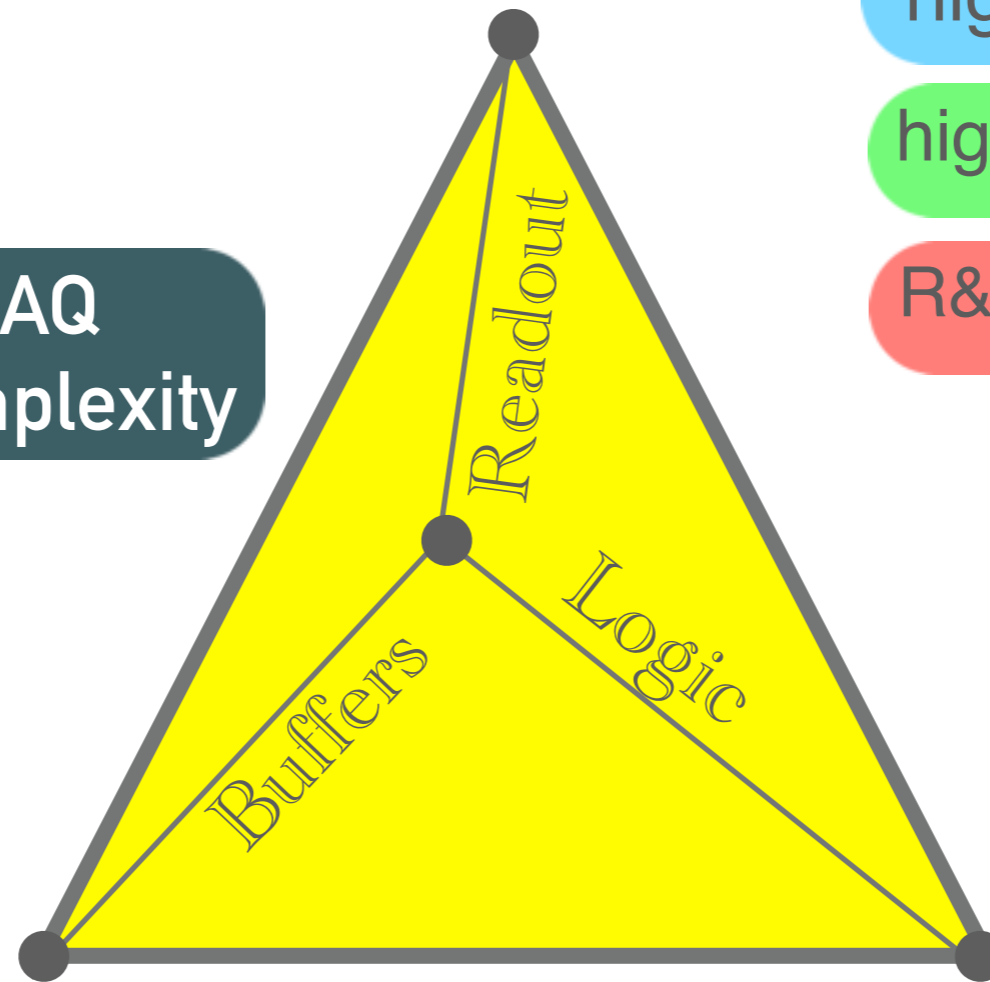
What we do?

How?

Example

Tension between TDAQ architecture and FE complexity

## Trigger-less DAQ



high detector granularity

high speed electronics/links

R&D on detectors Front-End

## High performance farms

refine calibrations, as offline

large buffers, long latency

tight: offline=online (LHCb, ALICE)

soft: decouple trigger/DAQ (ATLAS, CMS)

## Triggering detectors

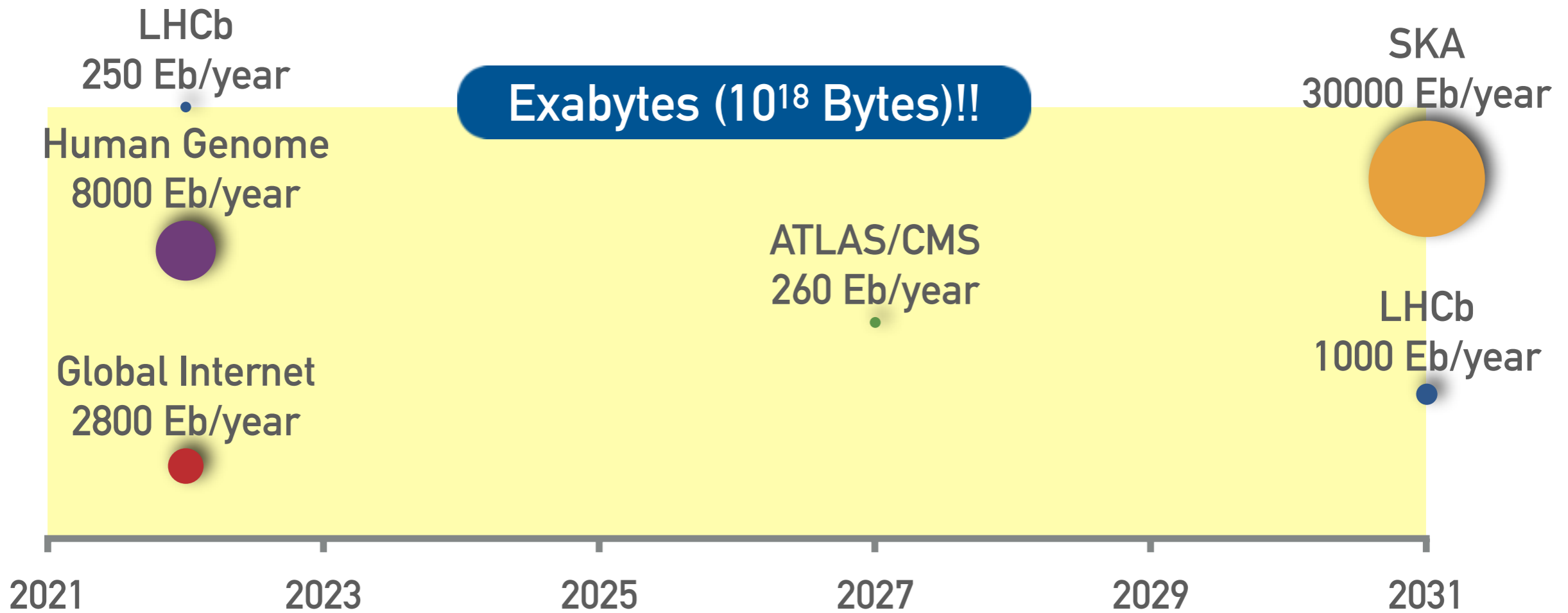
complex ASIC logic

trigger-driven design

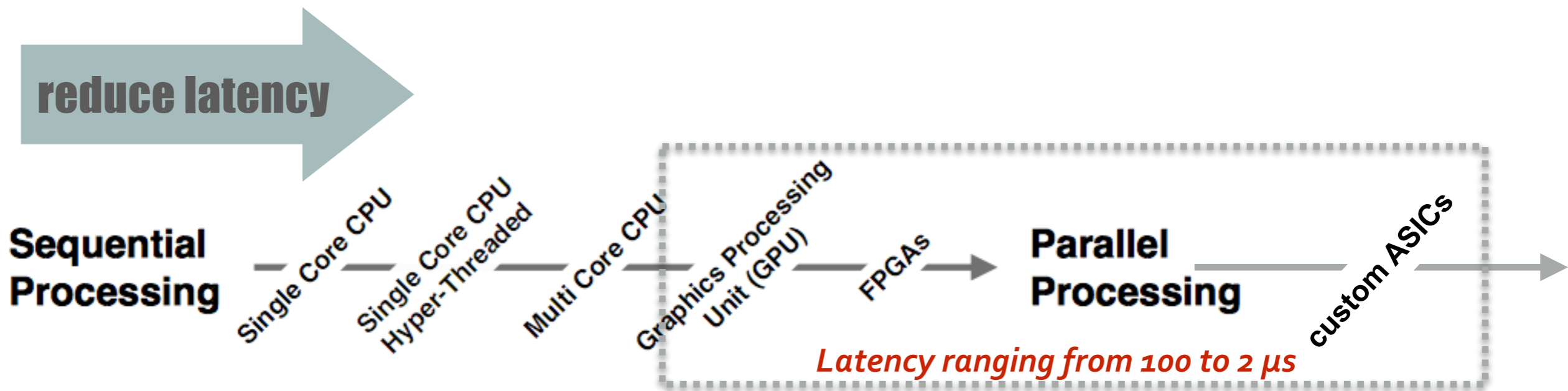
hardware track trigger (CMS)



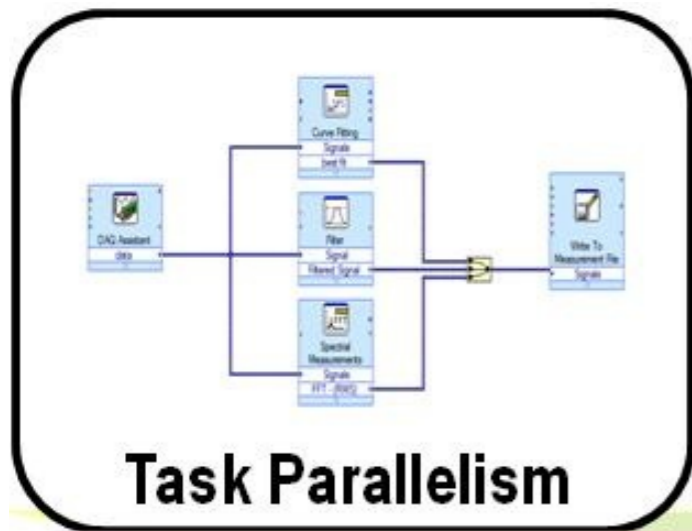
# THE REAL-TIME ADVENTURE



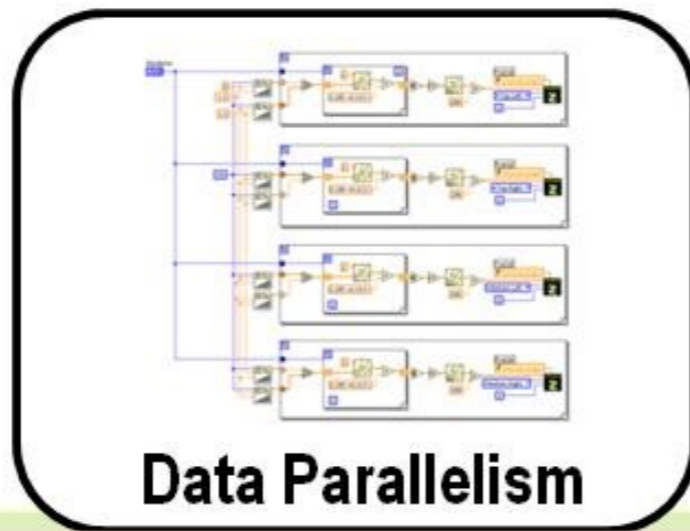
*See Openlab workshop*



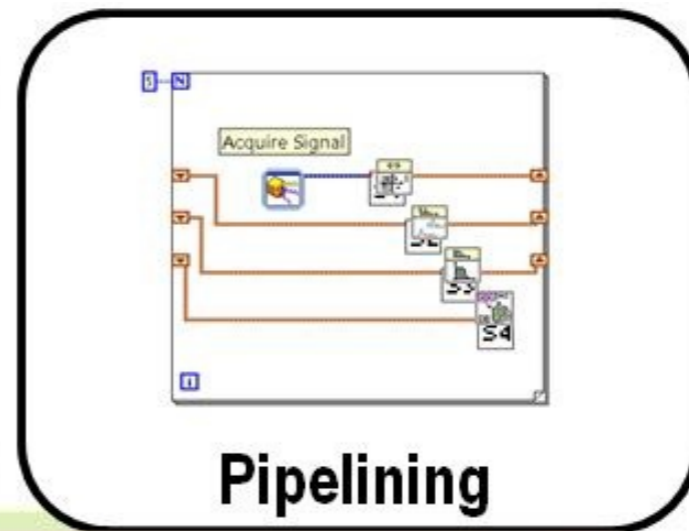
# TRENDS: COMBINED TECHNOLOGY



**Task Parallelism**



**Data Parallelism**

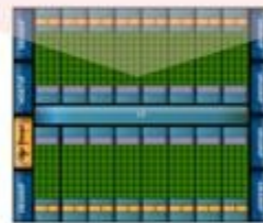


**Pipelining**



**Multicore Processors**

**Nvidia GPUs:  
3.5 B transistors**



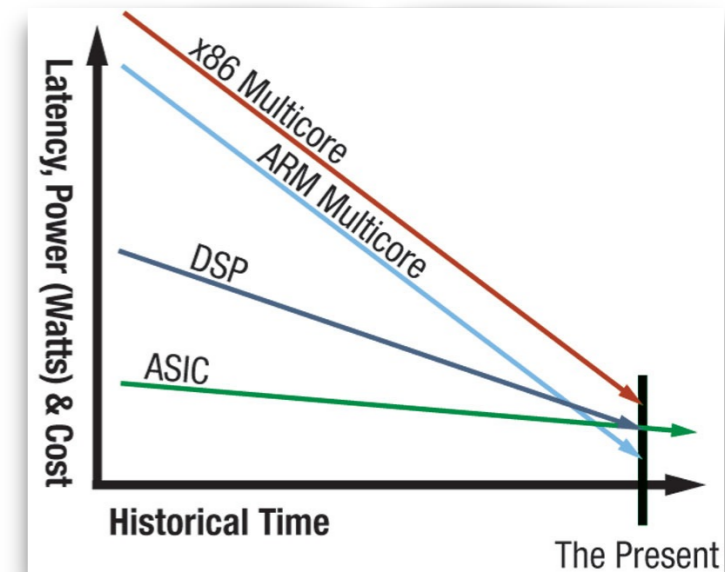
**GPUs\***

**Virtex-7 FPGA:  
6.8 B transistors**



**FPGAs**

(\*) Access to the nVIDIA® GPUs through the CUDA and CUBLAS toolkit/library using the NI LabVIEW GPU Computing framework.



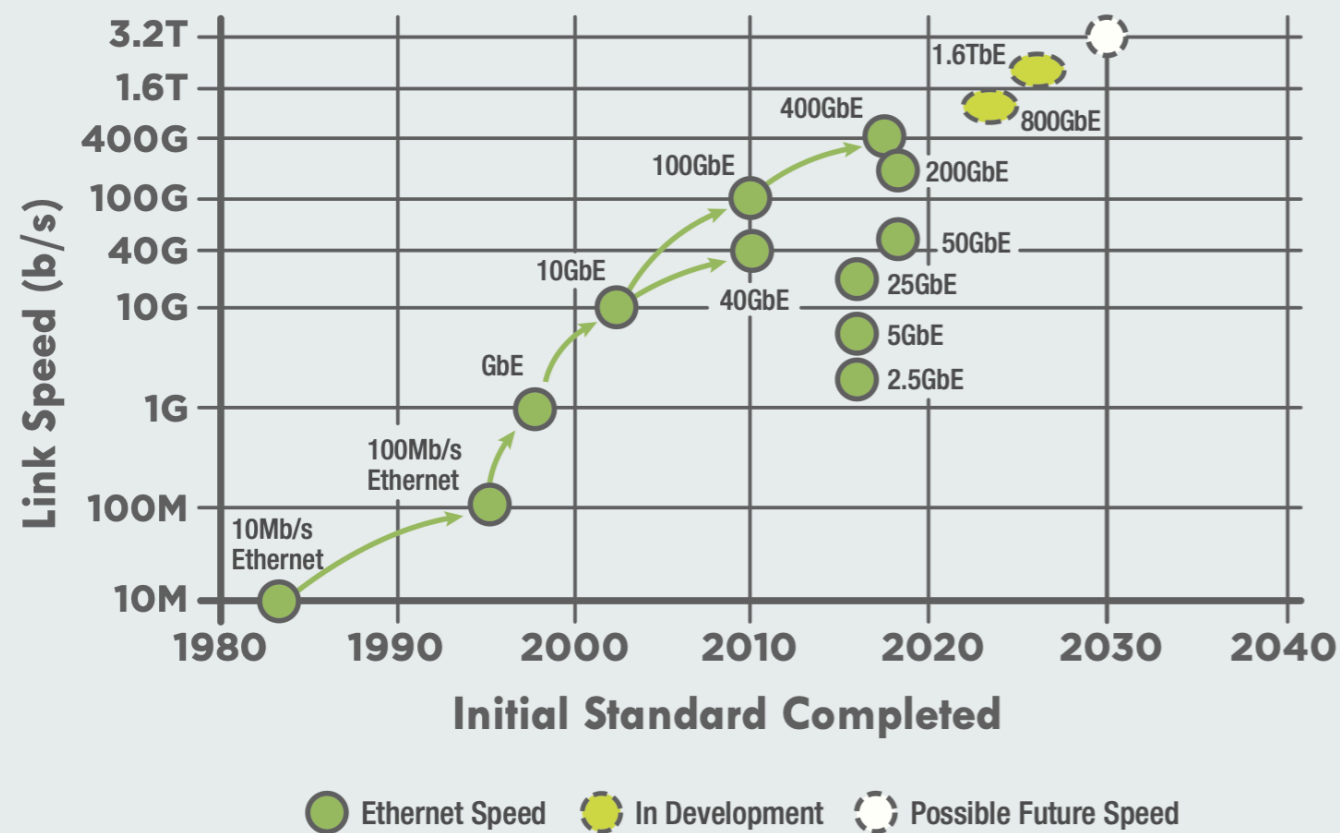
**The right choice can be combining the best of both worlds by analysing which strengths of FPGA, GPU and CPU best fit the different demands of the application**



# GENERAL TDAQ TRENDS

## Use COTS for network .... and processing

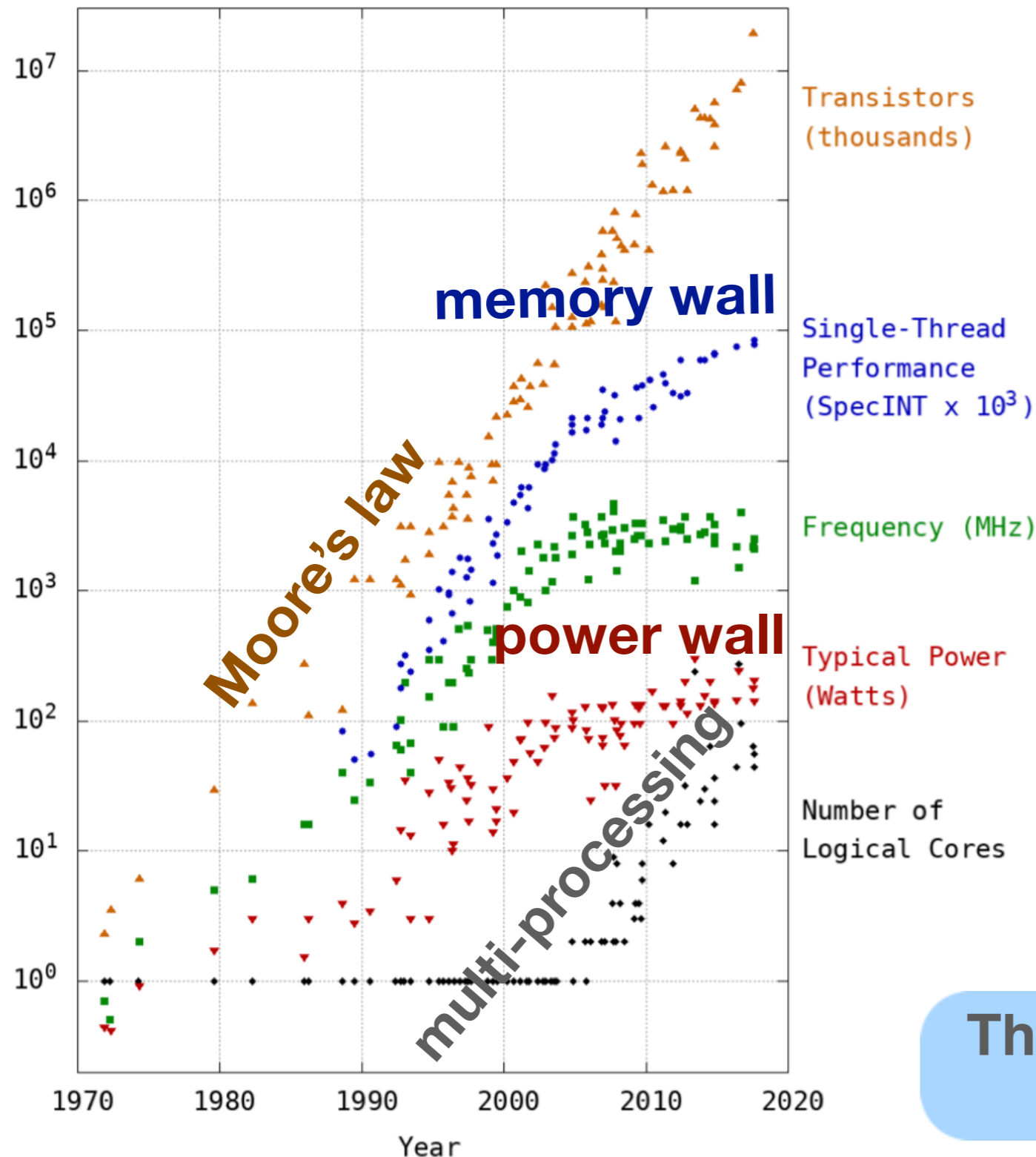
### ETHERNET SPEEDS



- ➔ Deal with dataflow instead of latency
  - ➔ **decouple** DAQ from High Level Triggers
  - ➔ decouple dataflow from storage, with temporary buffers
- ➔ Use **networks** as soon as possible
  - ➔ toward commercial bidirectional point-to-multipoint architecture
- ➔ Increase data **aggregation** at the Event Building
  - ➔ reducing request rates on DAQ software
  - ➔ per time-frame, per orbit instead of per-event

# EVOLUTION OF PROCESSING POWER TO BREAK WALLS

42 Years of Microprocessor Trend Data



Data Source: <https://github.com/karlrupp/microprocessor-trend-data>

- ▶ CPU frequencies are plateauing
- ▶ Local memory/core is decreasing
- ▶ Number of cores is increasing

- ➔ Exploiting CPU h/w, with more complicated programming
  - ➔ Vectorisation, low-level memory...
- ➔ Multithreading processing
  - ➔ To reduce memory footprint
- ➔ Use of co-processors:
  - ➔ High Performance Computing (HPC) often employ GPU architecture to achieve record-breaking results!
- ➔ Examples in LHC experiments:
  - ➔ data reduction ([ALICE & LHCb](#))
  - ➔ trigger selection ([CMS & ATLAS](#))

This requires fundamental re-write/  
optimization of our software

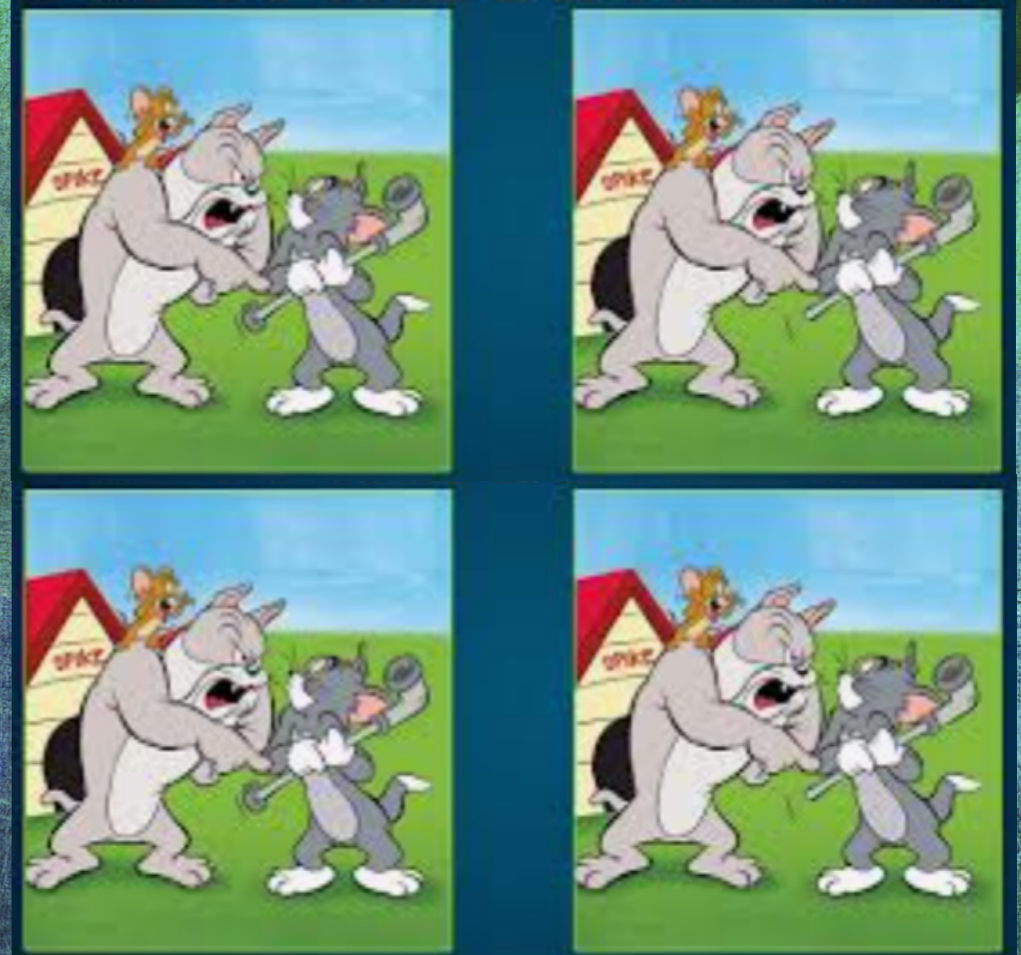
*Read: HPC computing*



# QUATTRO ESPERIMENTI A CONFRONTO

.....  
*How to maximise physics  
acceptance*

spot the differences



# LHC EXPERIMENTS FOR A DISCOVERY MACHINE

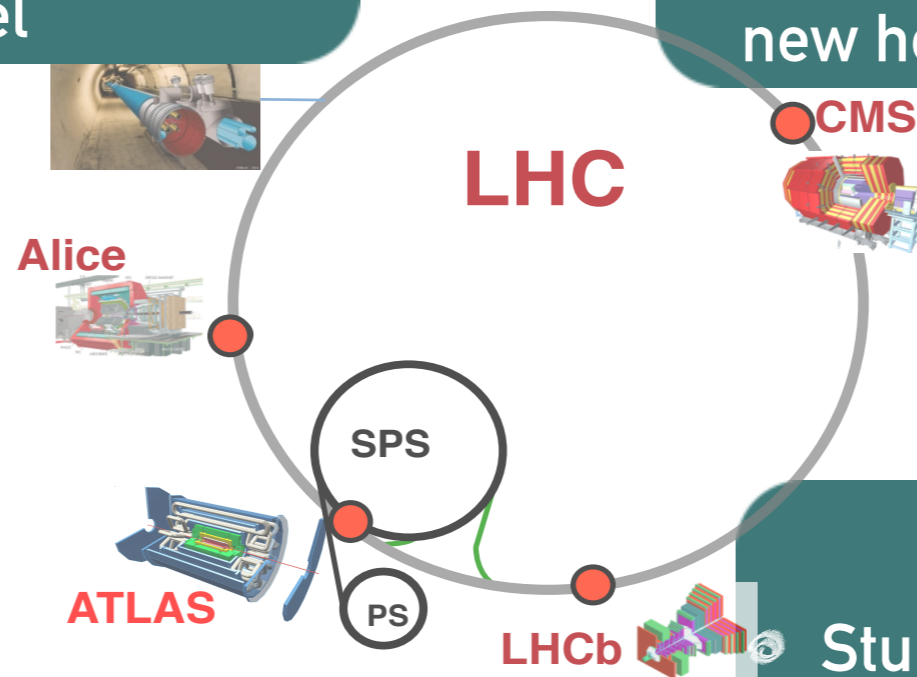
Goal: explore TeV energy scale to find New Physics beyond Standard Model

## ATLAS & CMS

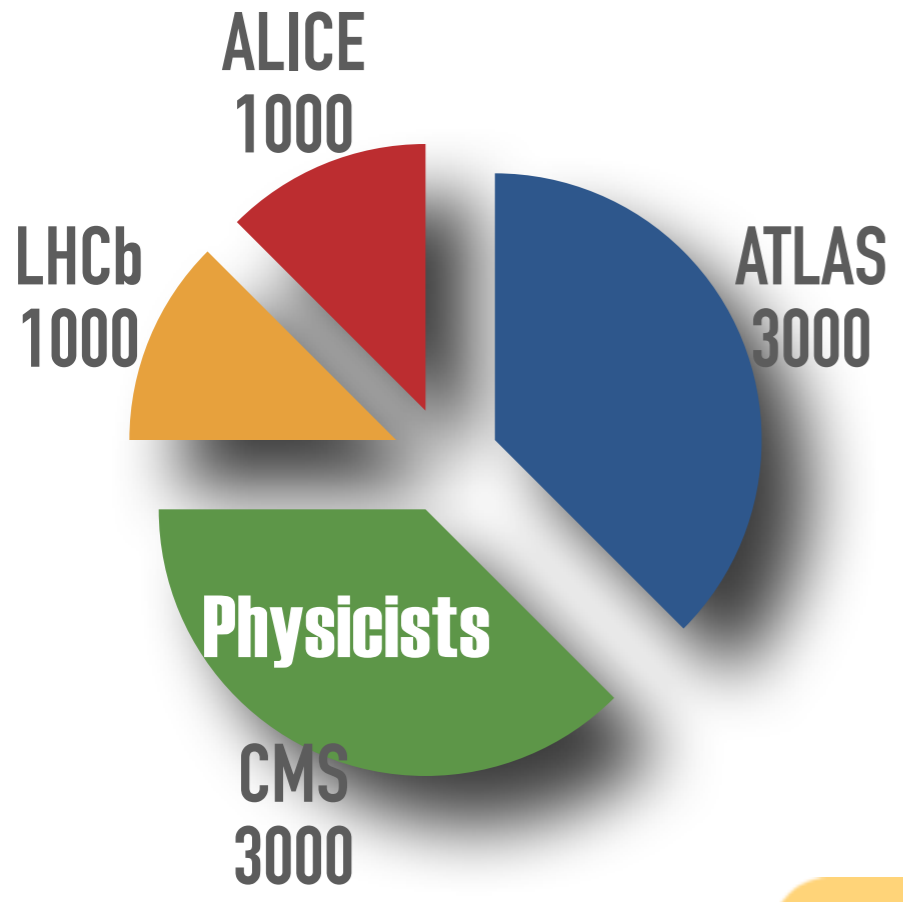
- Completing the Standard Model and probing the Higgs sector
- Extending the reach for new physics beyond the Standard Model

## LHCb

- Study CP violation and rare decays in b- and c-quark sector
- Search for deviations of SM due to new heavy particles



- ## ALICE
- Studying quark-gluon plasma, a complex system of strongly interacting matter produced by heavy ion collisions



Proposed: 1992, Approved: 1996, Started: 2009



# LHC EXPERIMENTS FOR A DISCOVERY MACHINE

Goal: explore TeV energy scale to find New Physics beyond Standard Model

## ATLAS & CMS

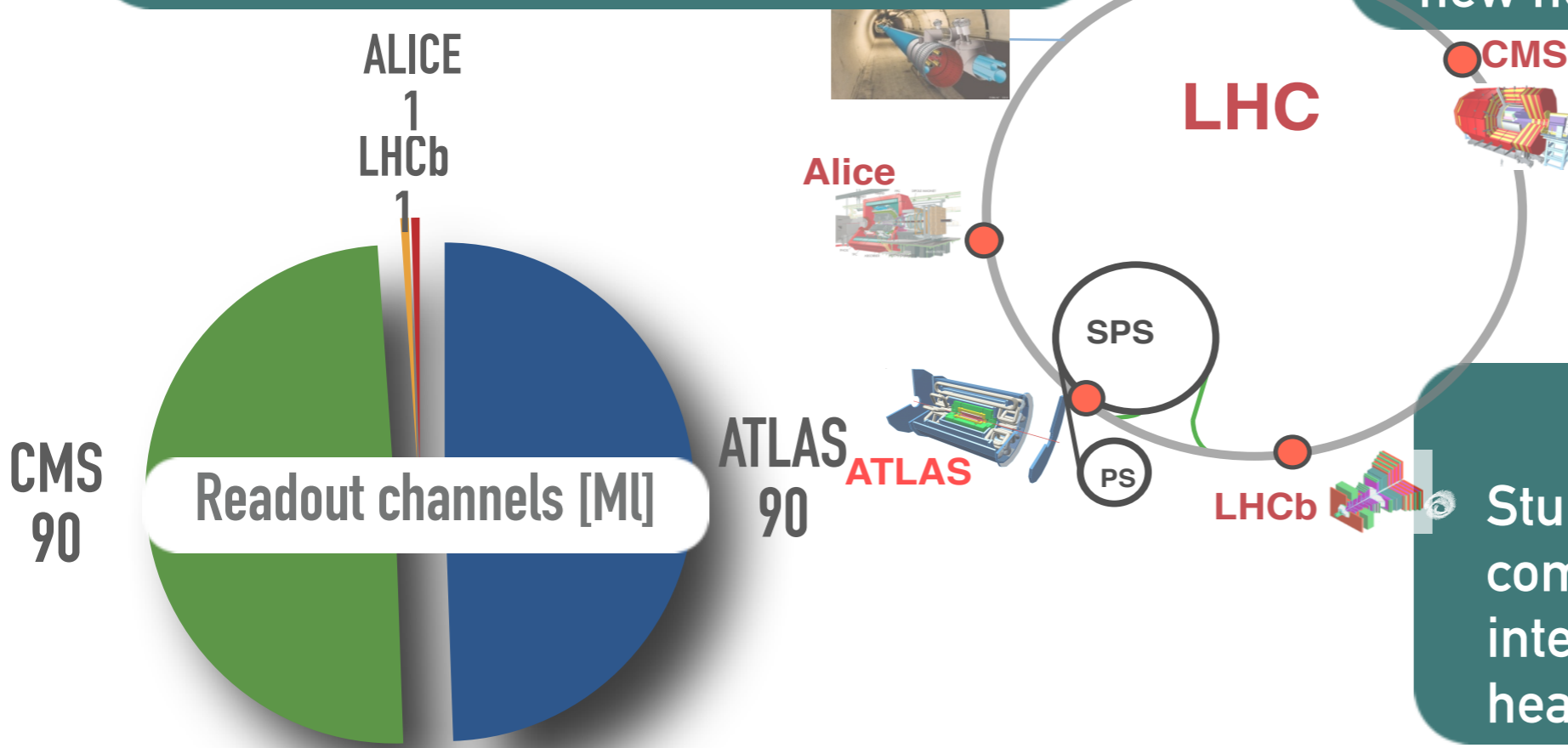
- Completing the Standard Model and probing the Higgs sector
- Extending the reach for new physics beyond the Standard Model

## LHCb

- Study CP violation and rare decays in b- and c-quark sector
- Search for deviations of SM due to new heavy particles

## ALICE

- Studying quark-gluon plasma, a complex system of strongly interacting matter produced by heavy ion collisions



Proposed: 1992, Approved: 1996, Started: 2009

# DIFFERENT PHYSICS SEARCHES

.... and LHC operations

✦ **ATLAS/CMS: p-p collisions at full Luminosity**

✦ search in high energy scale

✦ **LHCb: p-p collisions at reduced Luminosity**

✦ search complex topologies of b-quark decays

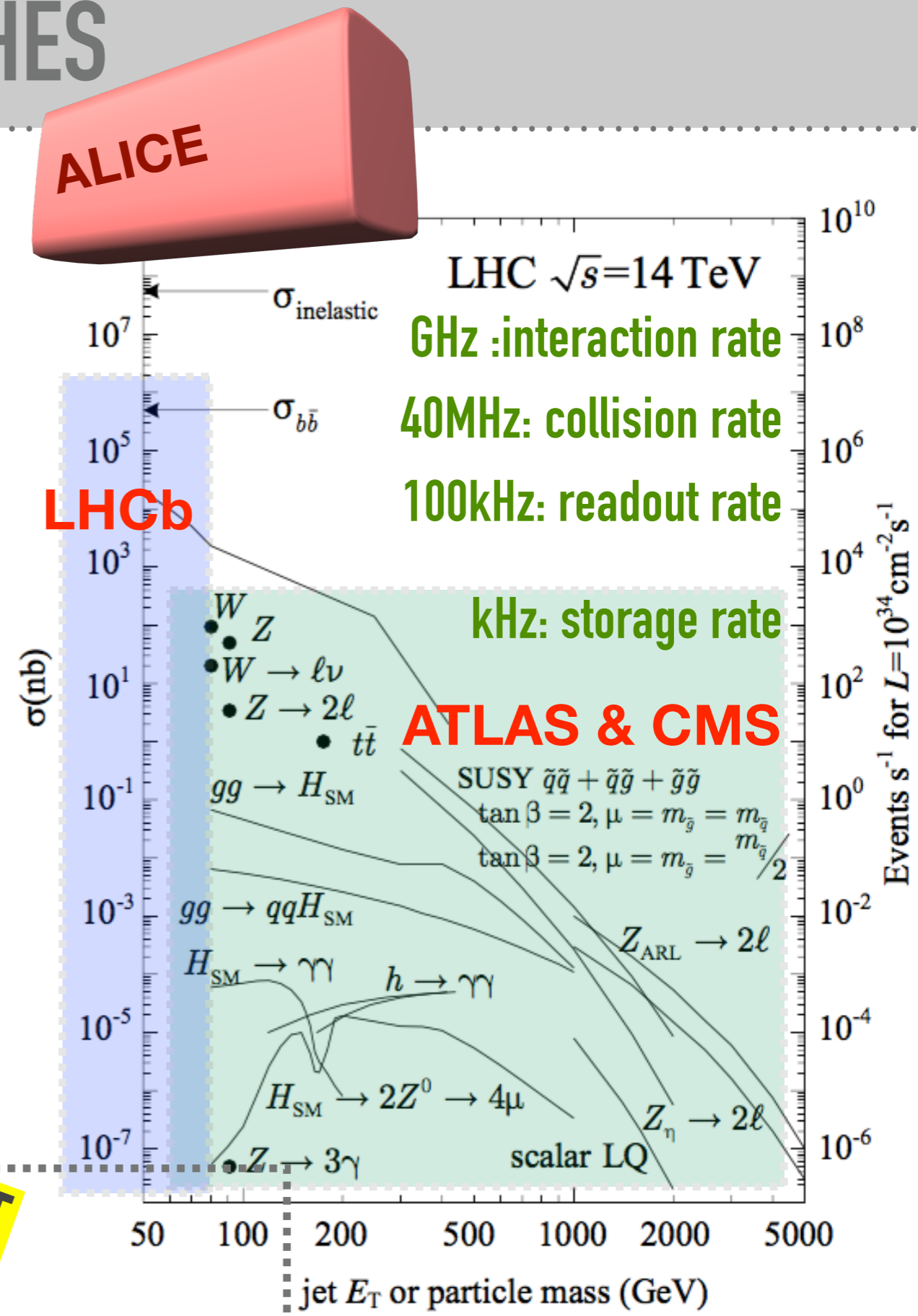
✦ **ALICE: heavy-ion collisions ~2000 mb**

✦ search in high energy density



- ➔ Expected rates and S/B ratio
- ➔ Signal topology and complexity
- ➔ Size of event (number of channels, particle multiplicity)

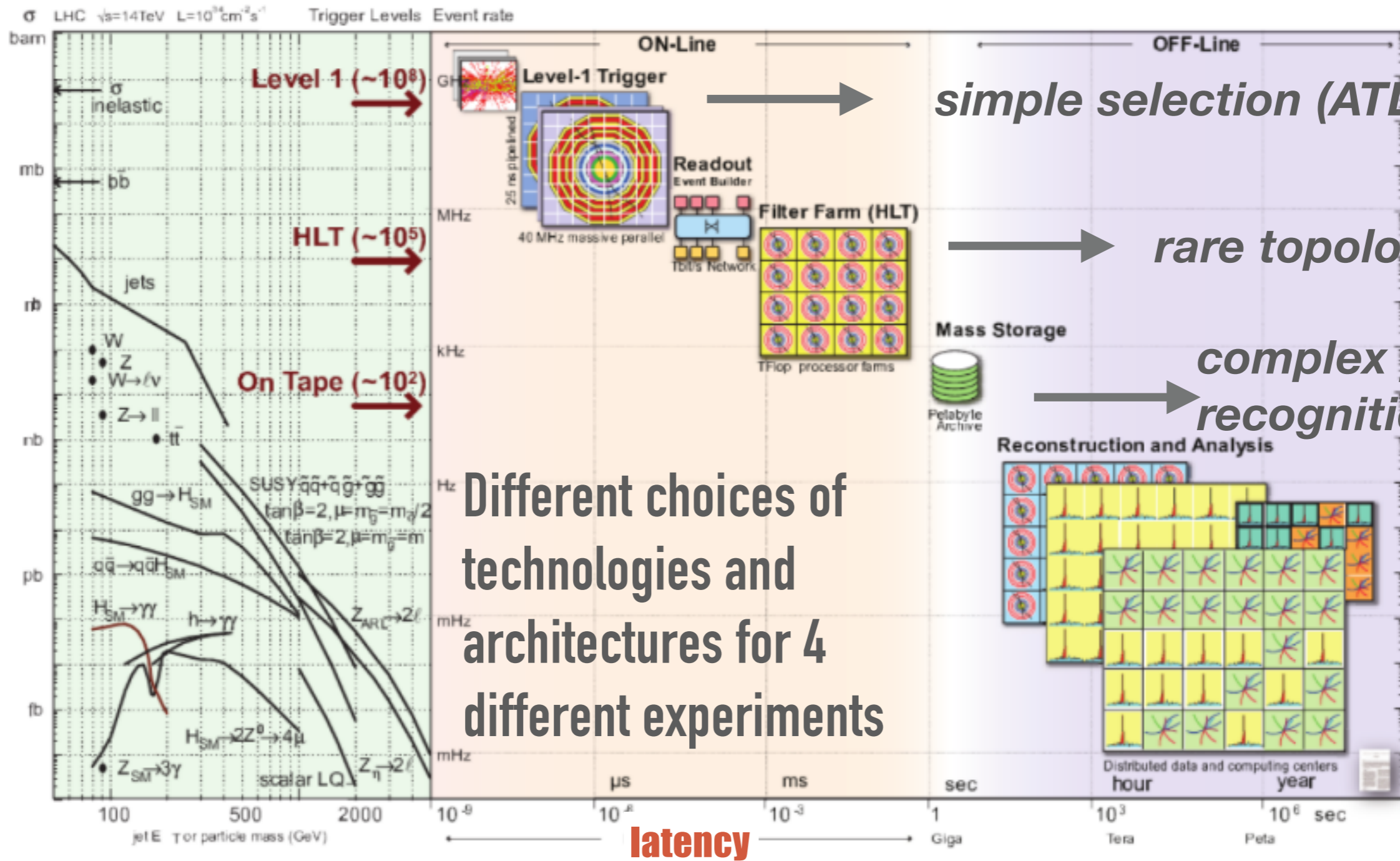
**DIFFERENT**





# ENHANCED TRIGGER SELECTIONS

data rates



Different choices of technologies and architectures for 4 different experiments

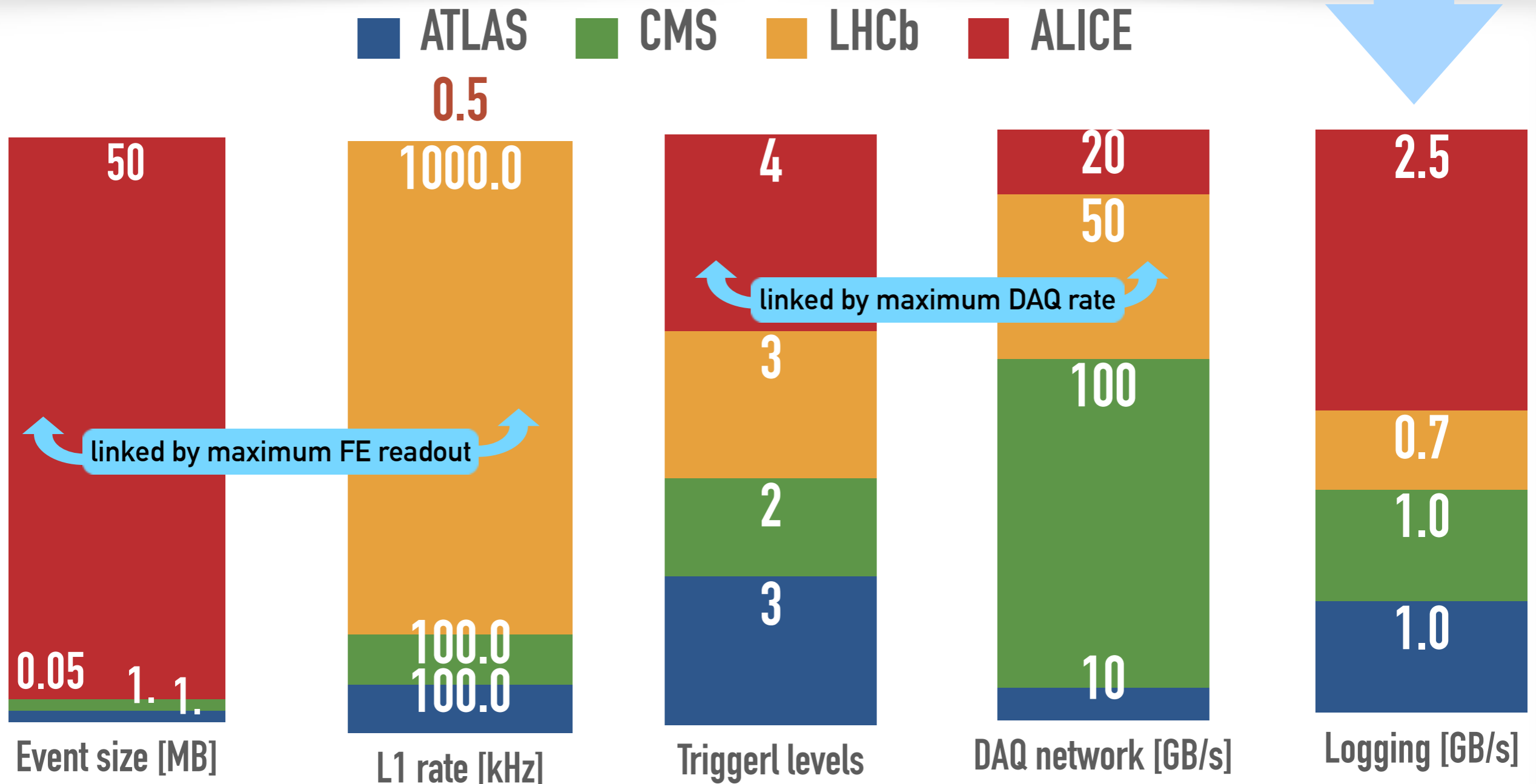
- ➔ **ATLAS/CMS: Trigger power:** reducing the data-flow at the earliest stage
- ➔ **ALICE/LHCb: Large data-flow:** low trigger selectivity due to large irreducible background

# COMPARING BY NUMBERS

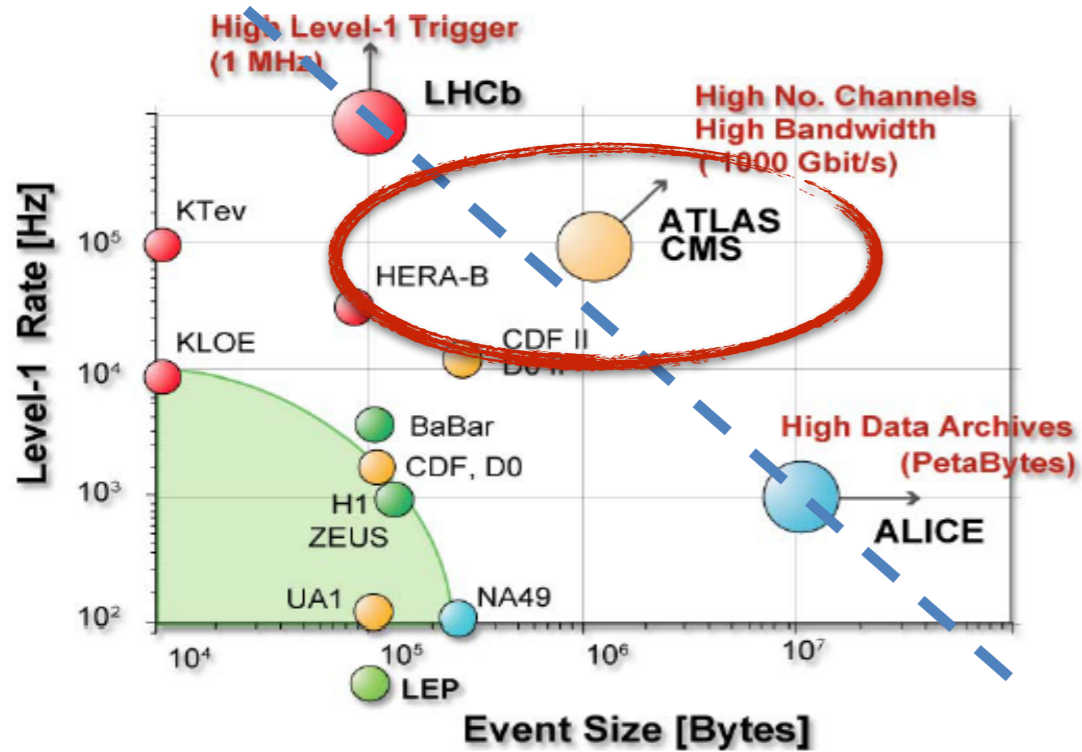
LHC experiments share the same CERN budget for computing resources, which is the constrain between trigger and DAQ power

Allowed storage and processing resources

Design values in 2009

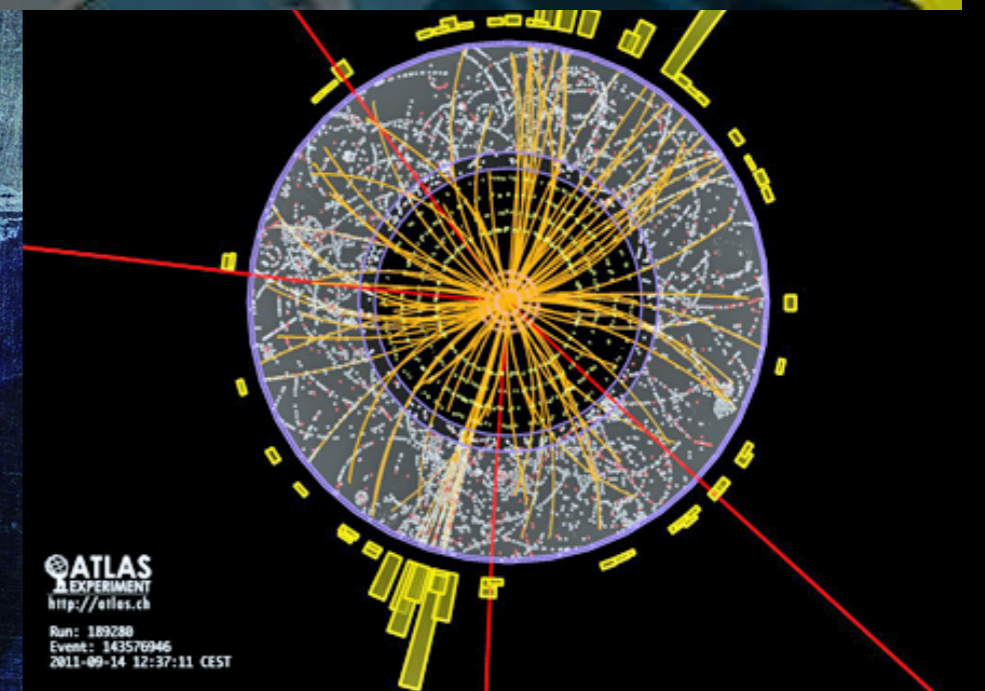
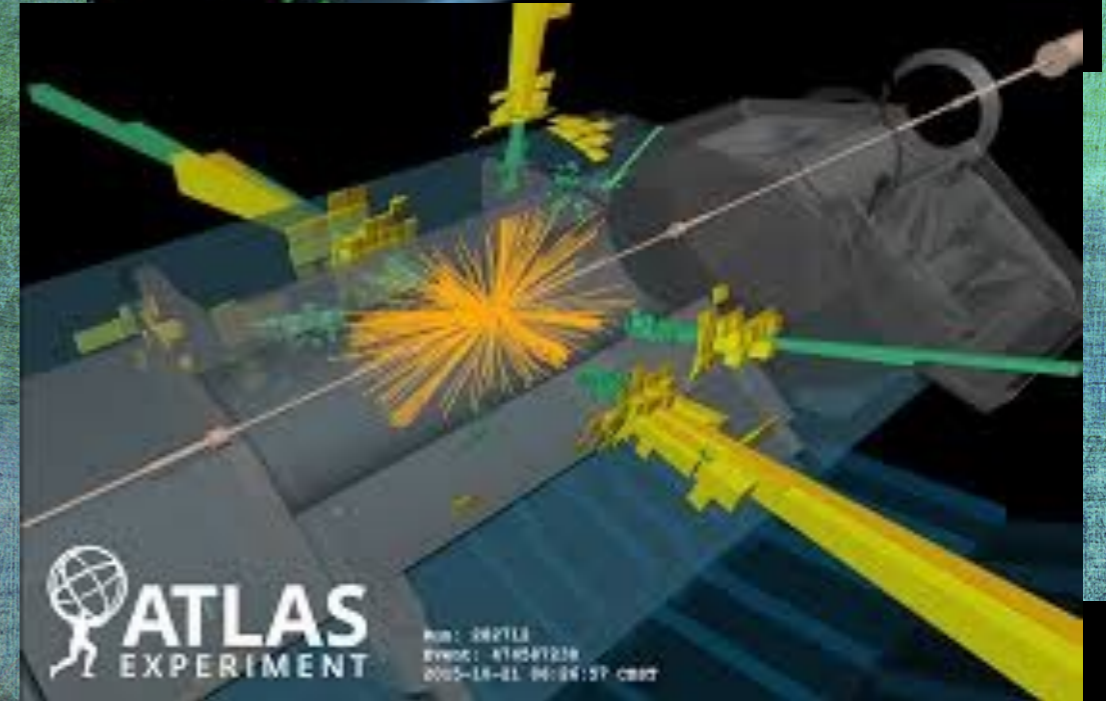
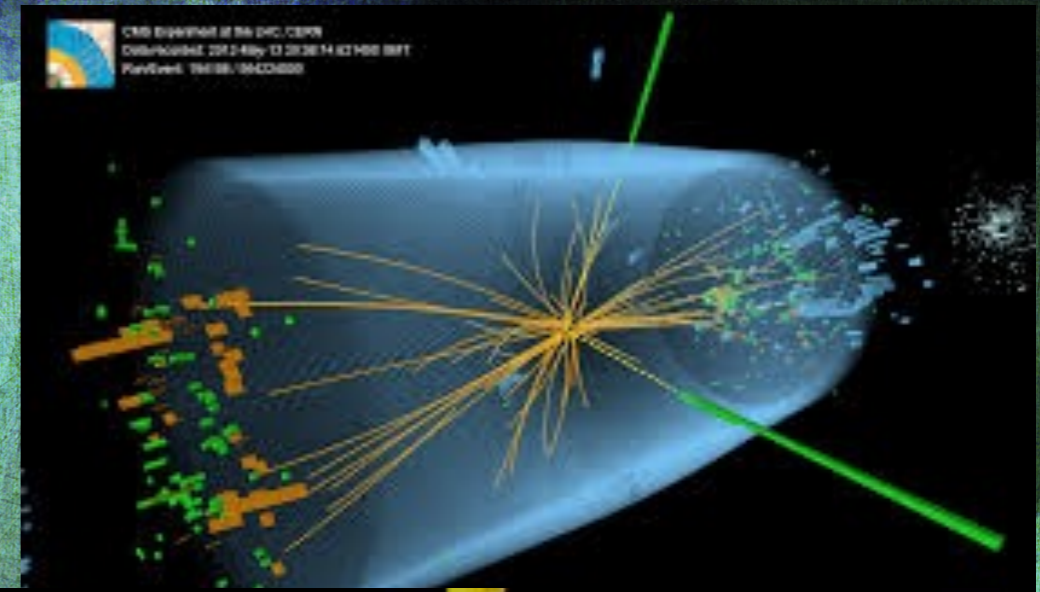






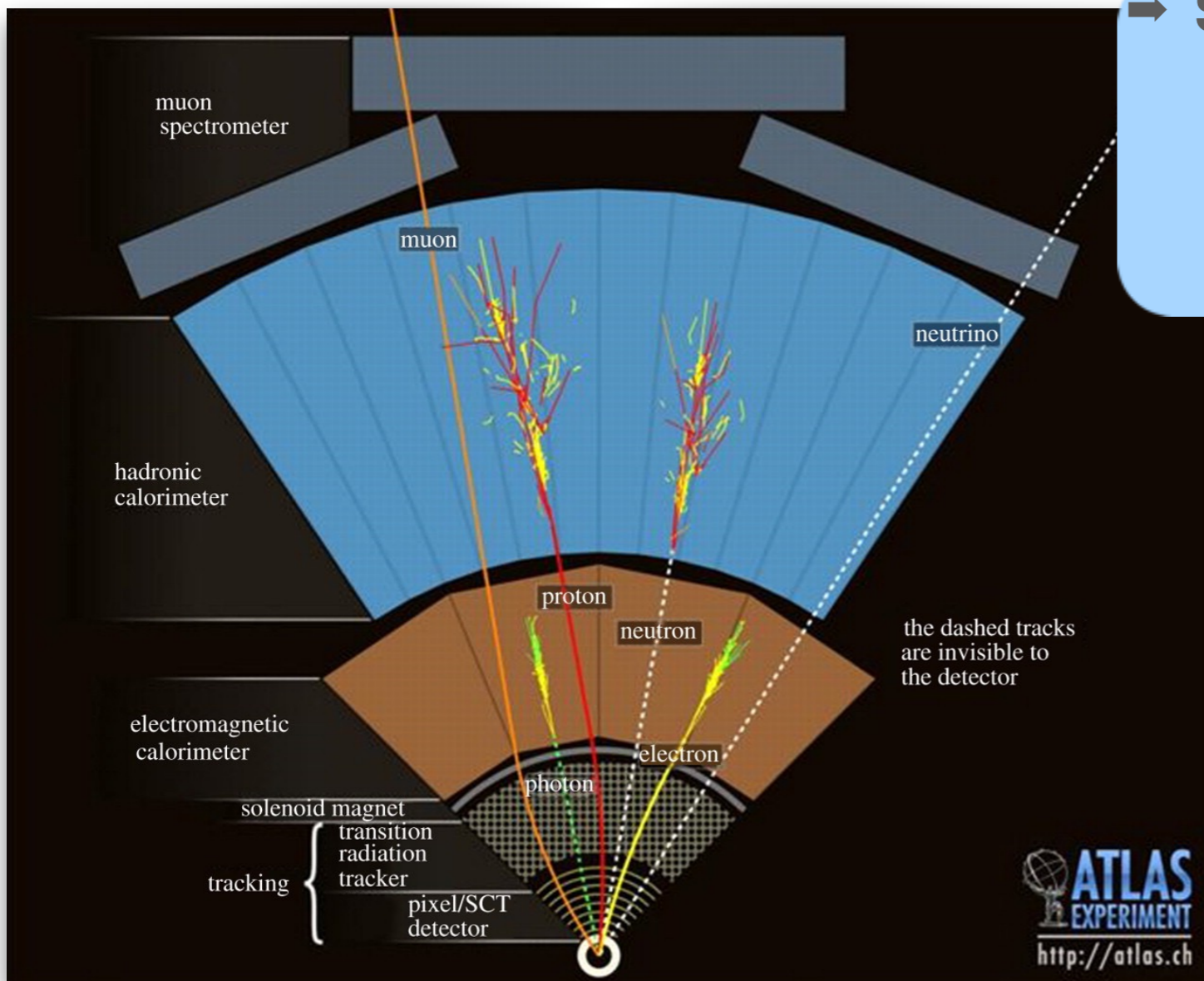
# ATLAS AND CMS

*Studying the Standard Model at the high energy frontier*





- Search in high-energy scale
  - Discover large mass particles through their high-energy products
  - **Discovery** = inclusive selections



$$\frac{\text{everything}}{\text{Higgs}} = \frac{\sigma_{tot}}{\sigma_{H(500\text{GeV})}} \approx \frac{100\text{ mb}}{1\text{ pb}} \approx 10^{11}$$

**approximately  
10<sup>6</sup> rejection**

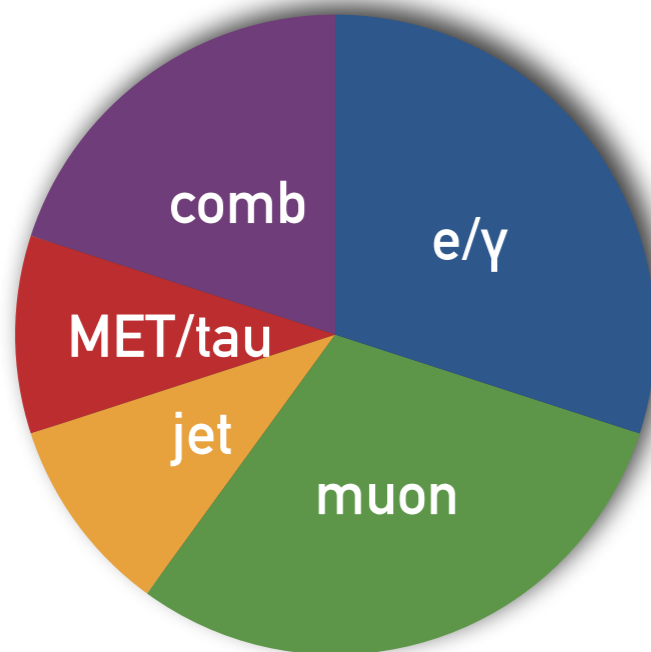
- Easy selection of high-energy leptons over background ==> @L1
  - Against thousands of particles/collisions (typically low momentum jets)
- Remember: 90M readout channels and full Luminosity ==> 1 MB/event



Same physics plans, different competitive approaches for detectors and DAQ

→ Same trigger strategy and data rates

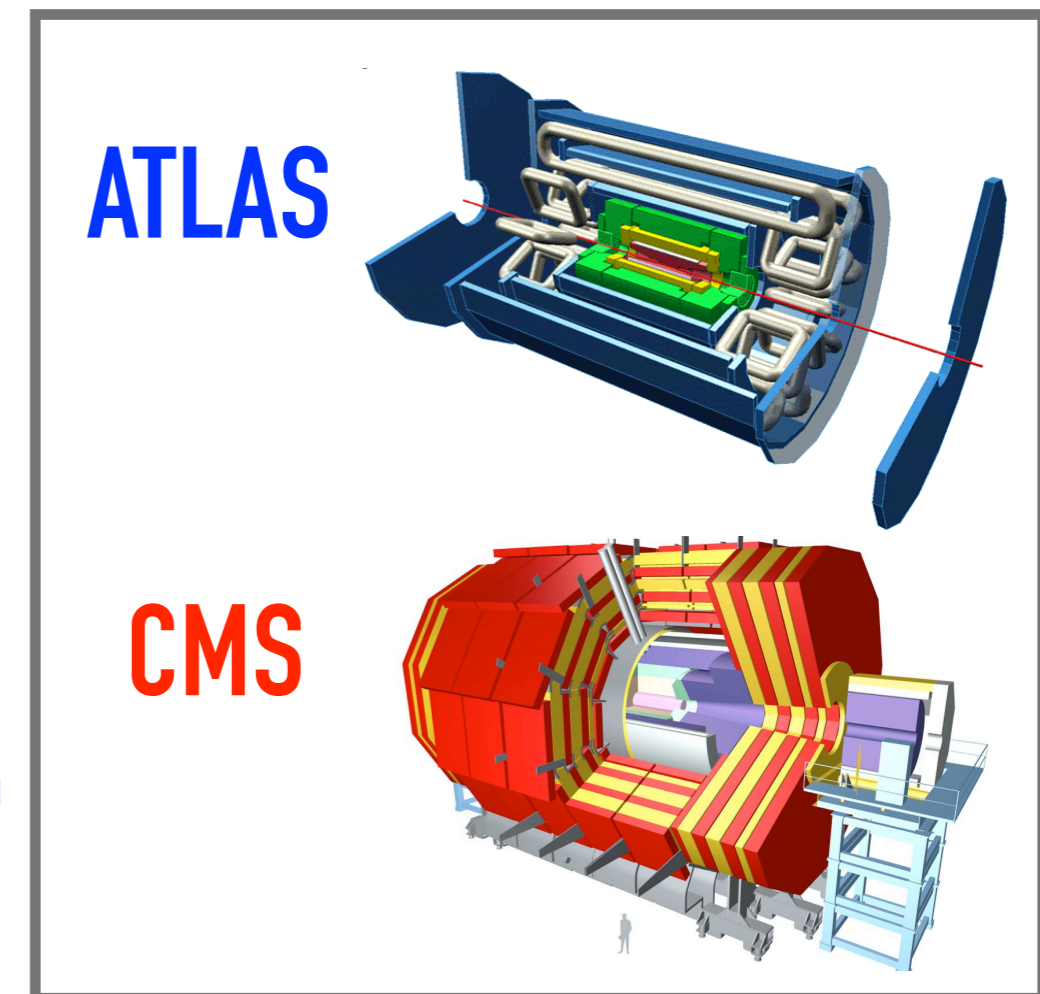
1 MB \* 100 kHz = 100 GB/s readout network



*inclusive trigger selections*

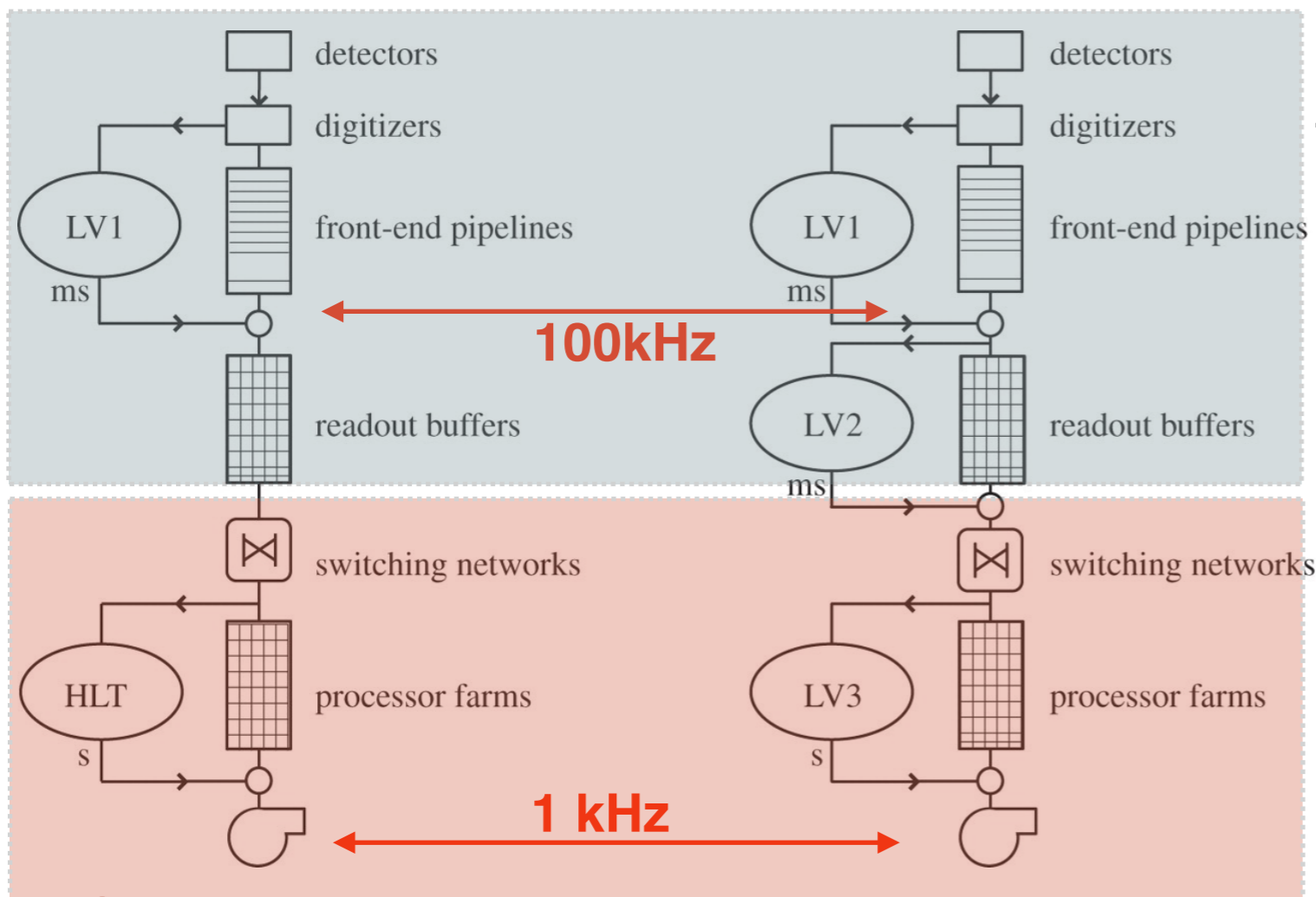
→ Different DAQ architectures

- **ATLAS**: minimise data flow bandwidth with multiple levels and regional readout
- **CMS**: large bandwidth, invest on commercial technologies for processing and communication



## Final storage and processing resources (at Tier0) allow order of few GB/s output

*Evolved from 1GB/s to current almost 5GB/s*



### DAQ+HLT system

### Network and Farm size

- ➔ **1MB/event at 100kHz for O(100ms) HLT latency**
  - ➔ Network:  $1\text{MB} \cdot 100\text{kHz} = 100\text{GB/s}$
  - ➔ HLT farm:  $100\text{kHz} \cdot 100\text{ms} = \mathbf{O(10^4) \text{ CPU cores}}$
- ➔ Can add intermediate steps (level-2) to reduce resources, at cost of complexity (at ms scale)

*See S.Cittolin, DOI: 10.1098/rsta.2011.0464*



# CMS: 2-STAGE EVENT BUILDING IN RUN 1



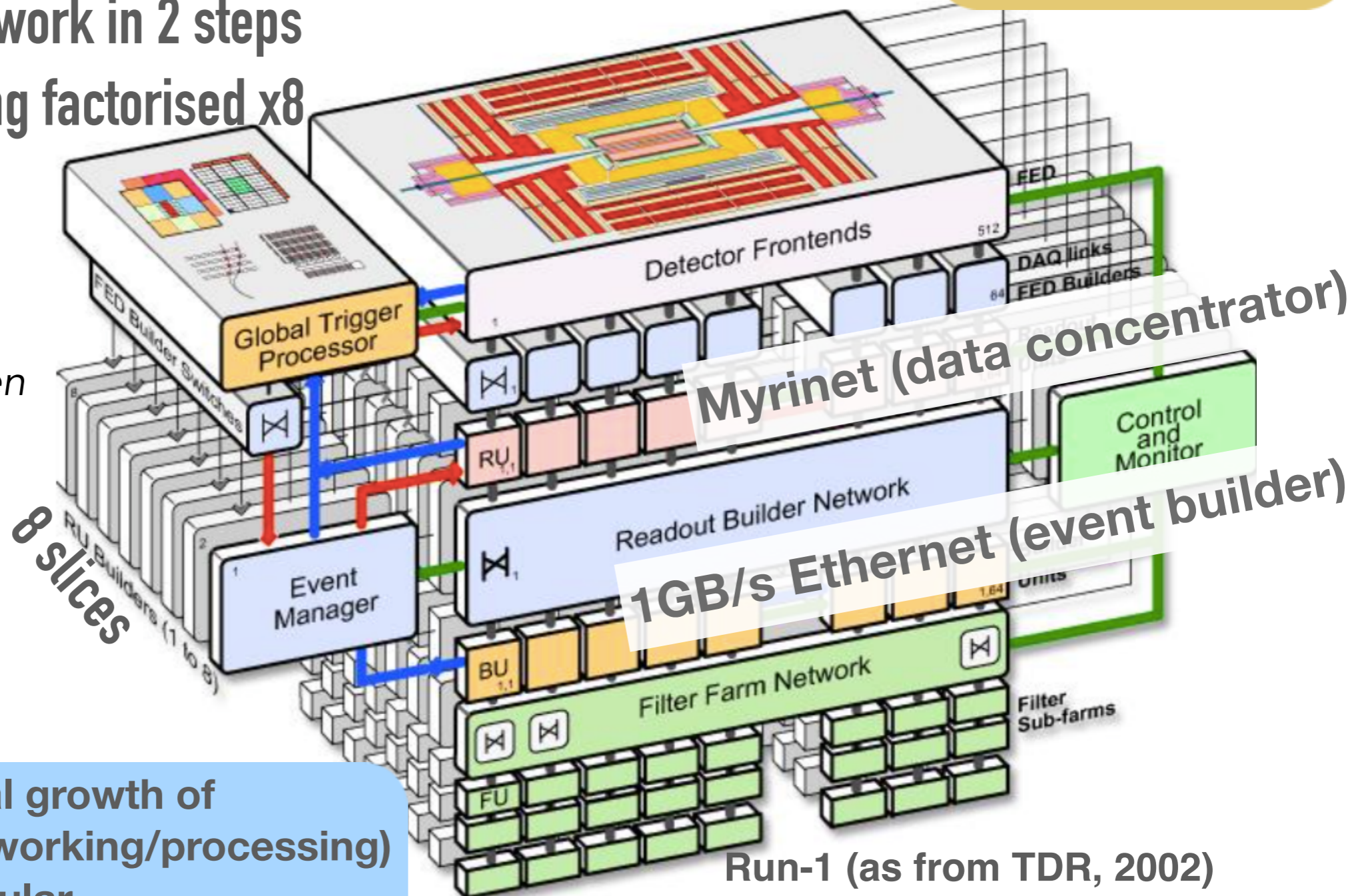
Cannot do Event Building at 100 kHz

CMS DAQ-1

100 GB/s readout network in 2 steps

100 kHz Event Building factorised x8

2 EB networks in blue  
Filter network in green



- ➔ Bet on exponential growth of technologies (networking/processing)
- ➔ Scalable and modular
  - ➔ Independent development of two network technologies

Run-1 (as from TDR, 2002)

- ➔ Myrinet + 1GB Ethernet
- ➔ 1-stage building: 1200 cores (2C)
- ➔ HLT: ~13,000 cores
- ➔ 18 TB memory @ 100kHz: ~90ms/event

# NETWORK EVOLUTION

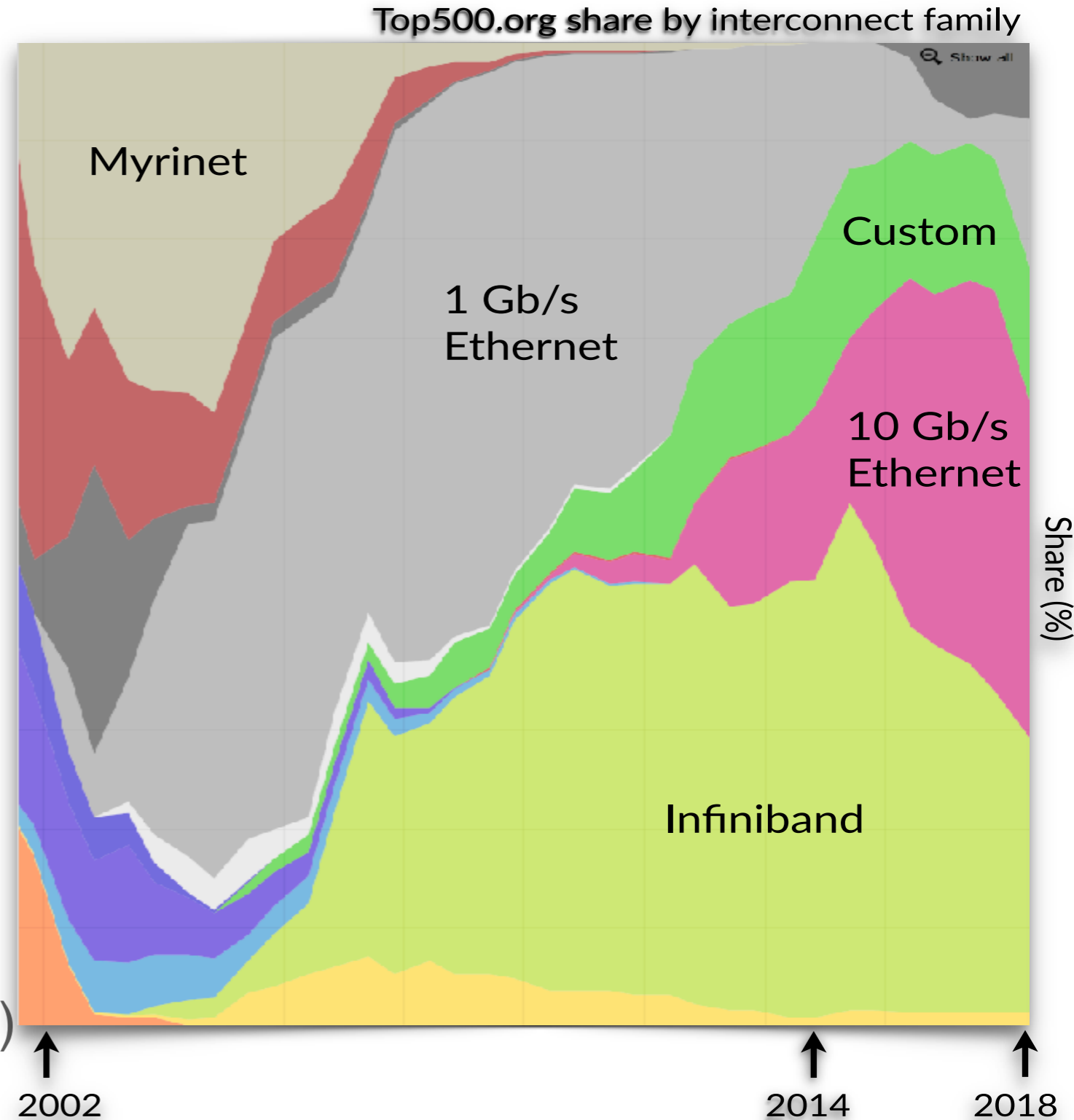
## Run 1: 100 GB/s network

**Myrinet widely used when DAQ-1 was designed**

- ➔ high throughput, low overhead
- ➔ direct access to OS
- ➔ flow control included
- ➔ new generation supporting 10GBE

## Run 2: 200 GB/s network

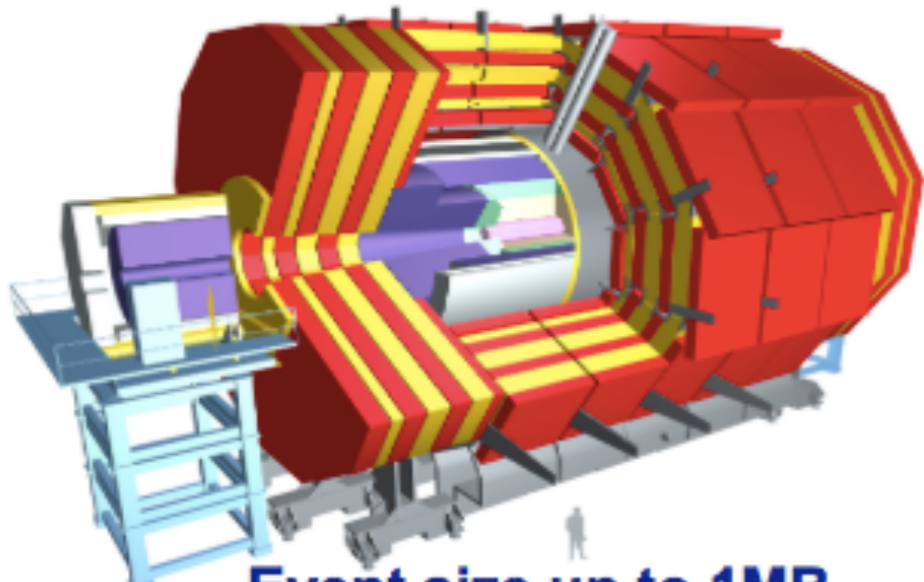
- ➔ Increased event size to 2MB
- ➔ Technology allows single EB network (56 Gbps FDR Infiniband)
- ➔ Myrinet → >10/40 Gbps Ethernet



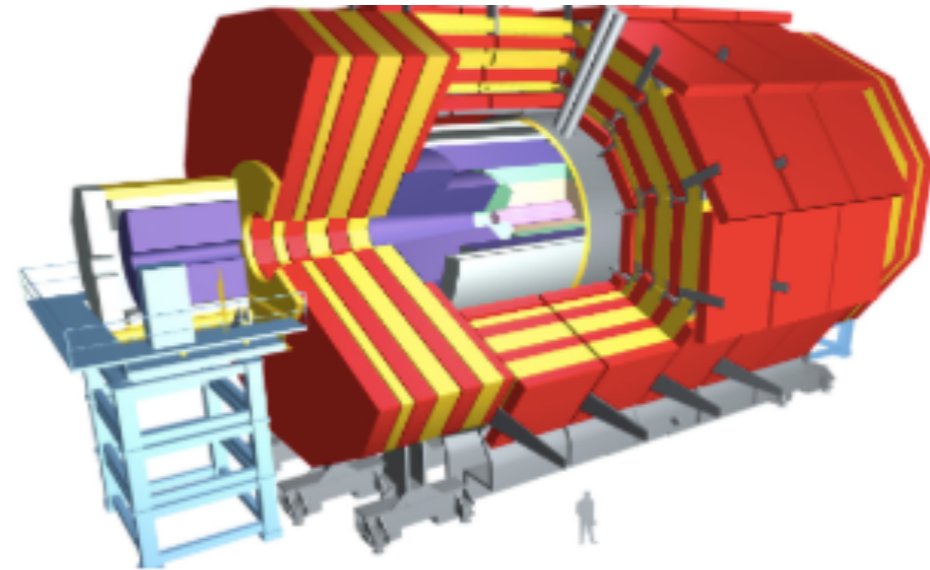
**Choose best prize/bitps!**



# EVOLUTION FROM RUN-1 TO RUN-2

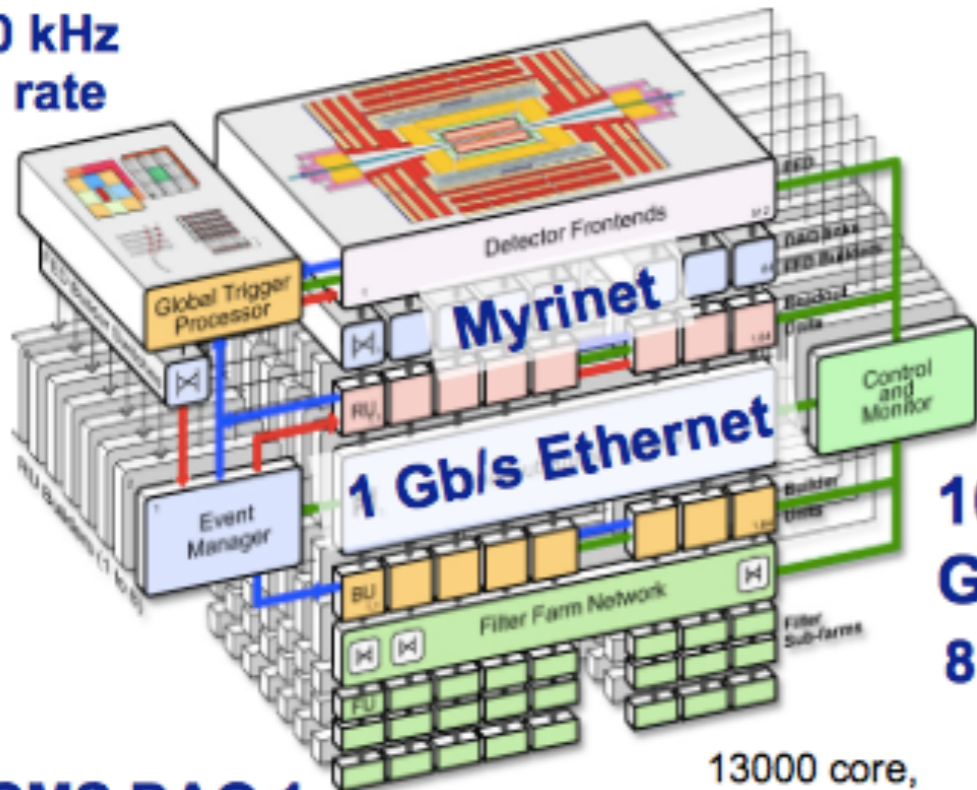


Event size up to 1MB



Event size up to 2MB

100 kHz  
L1 rate

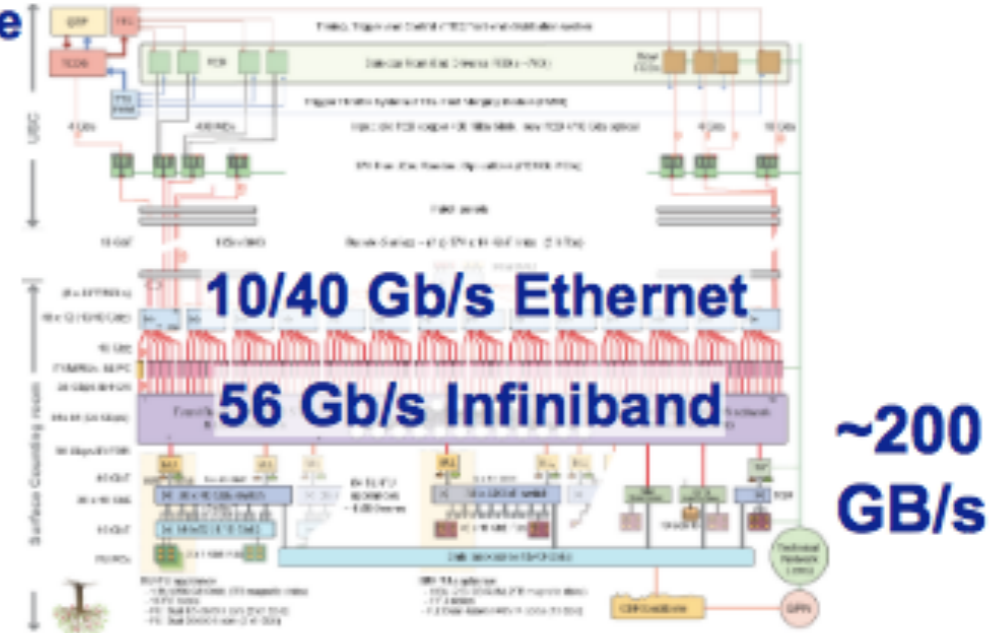


CMS DAQ 1

13000 core,  
1260 host  
filter farm

max. 1.2 GB/s to storage

100 kHz  
L1 rate



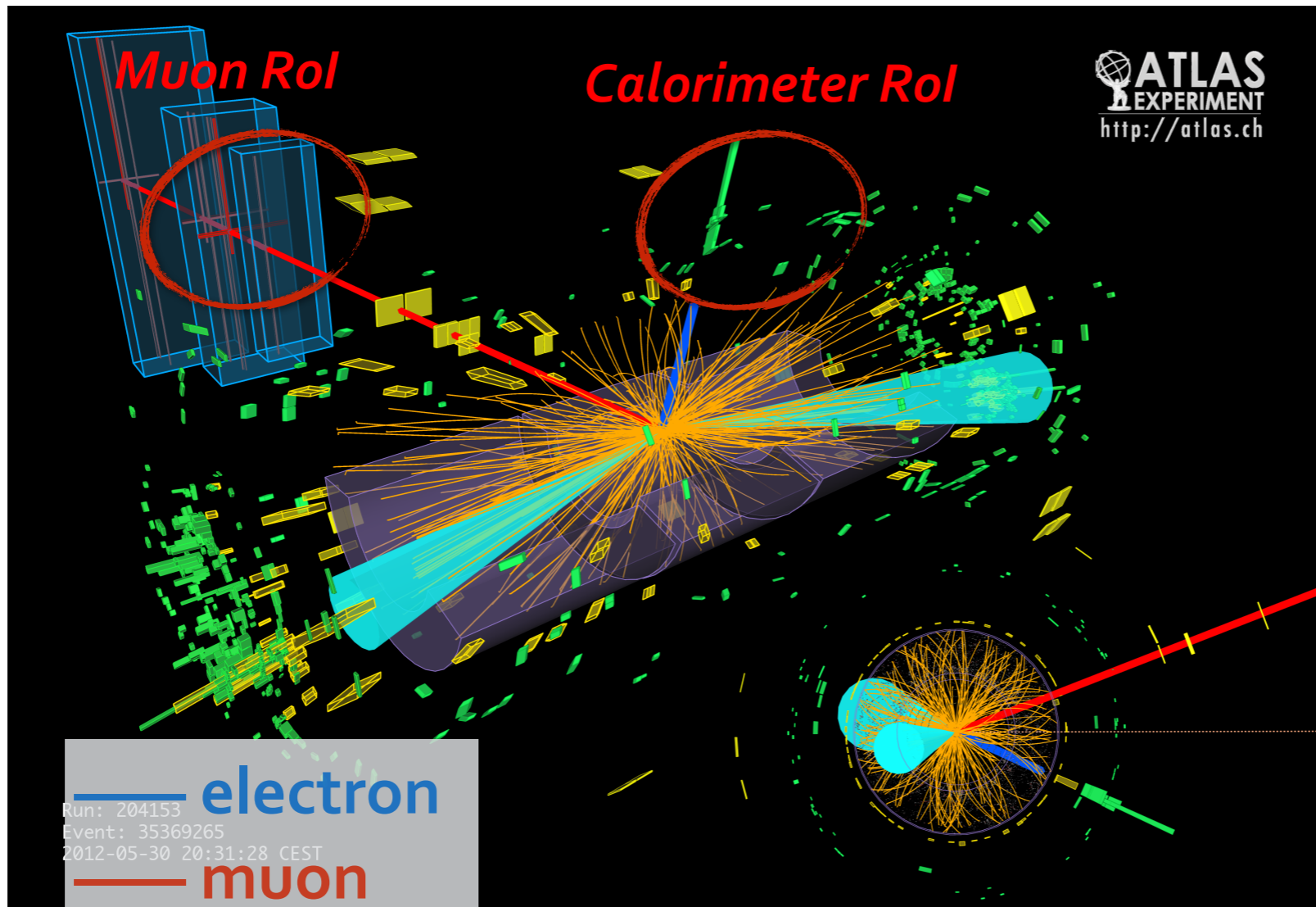
CMS DAQ 2

1 slice

16000+ core,  
900 host  
filter farm

~ 3-6 GB/s to storage

HLT selections based on regional readout and reconstruction,  
seeded by L1 trigger objects (RoI)



RoI=Region of Interest

- Total amount of RoI data is minimal: a few % of the Level-1 throughput
  - one order of magnitude smaller readout network ...
  - ... at the cost of a higher control traffic and reduced scalability

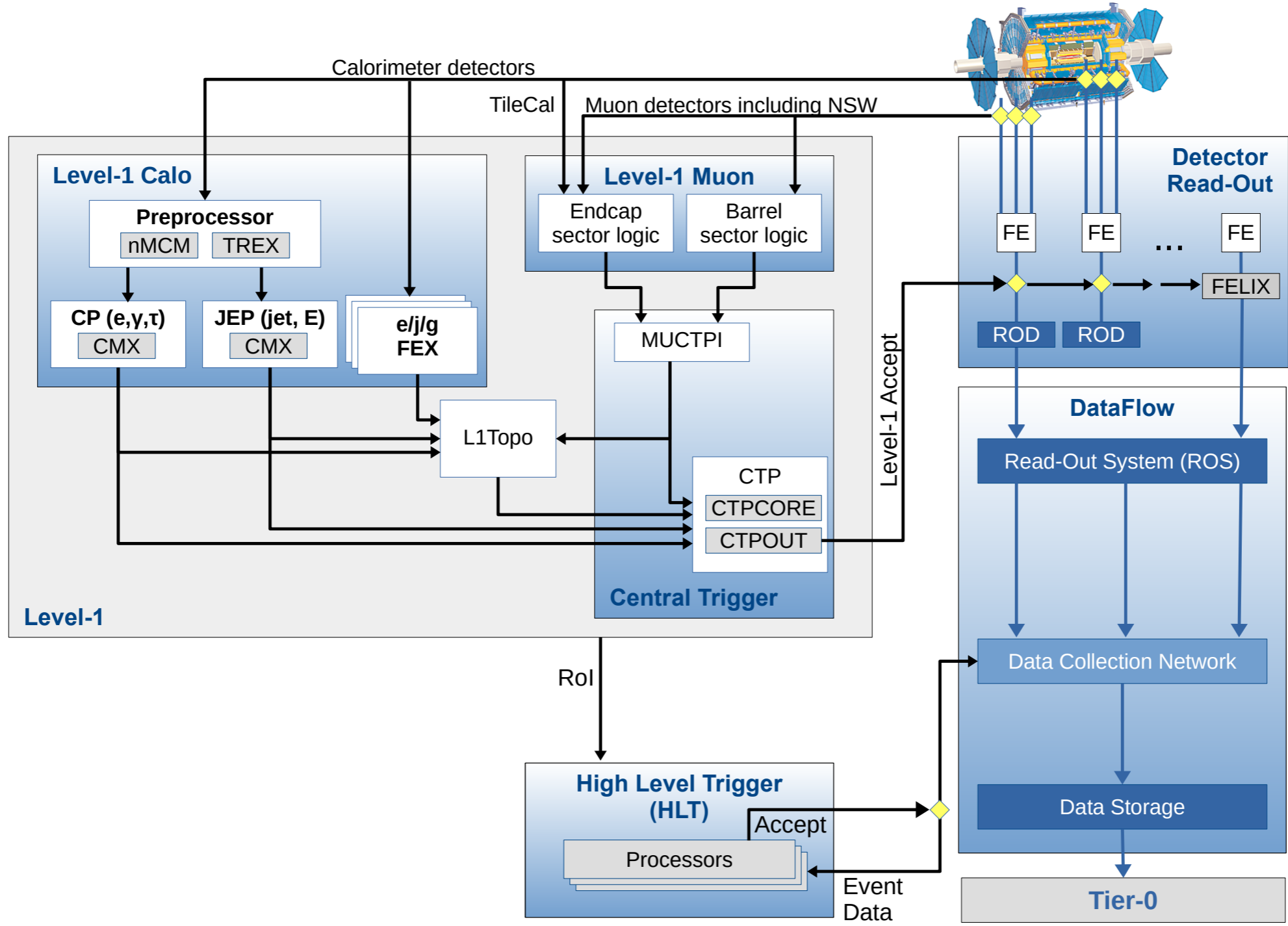


# ATLAS REGIONAL TDAQ ARCHITECTURE

Overall network bandwidth:  $\sim 10$  GB/s (x10 reduced by regional readout)

Run 3

40 MHz  
↓  
100 kHz  
↓  
~ 1.5 kHz



$O(60$  TB/s)  
↓  
 $\sim 160$  GB/s  
↓  
 $\sim 25$  GB/s  
↓  
 $\sim 1.5$  GB/s

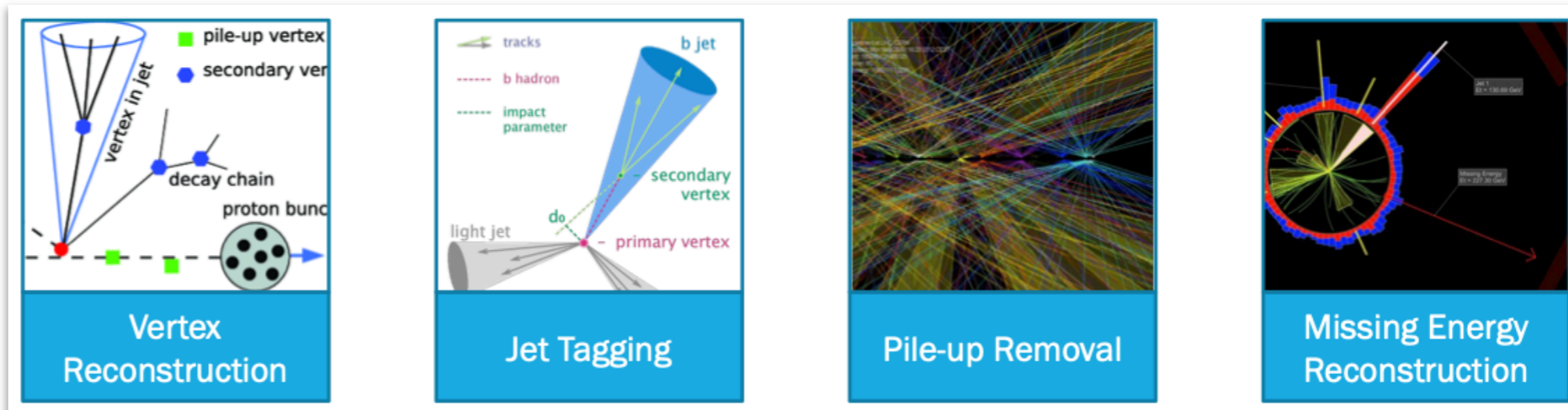
Push  
↑  
Pull

complex data router to forward different parts of the detector data, based on the trigger type

# TRACK-TRIGGER IS KEY FOR HL-LHC (RUN 4)



Silicon tracking systems provide incredibly high resolution, crucial for controlling rates



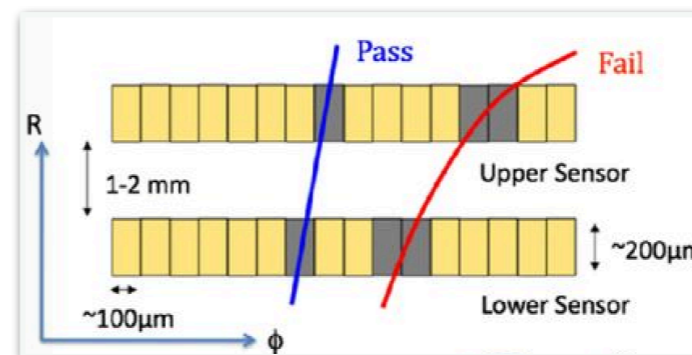
## Tracking challenges

- Readout ~800M channels, ~50 Tbps
- Combinatorics ( $10^4$  hits/BC)

combinatorics scales like  $L^N$   
 L=luminosity, N=number of layers

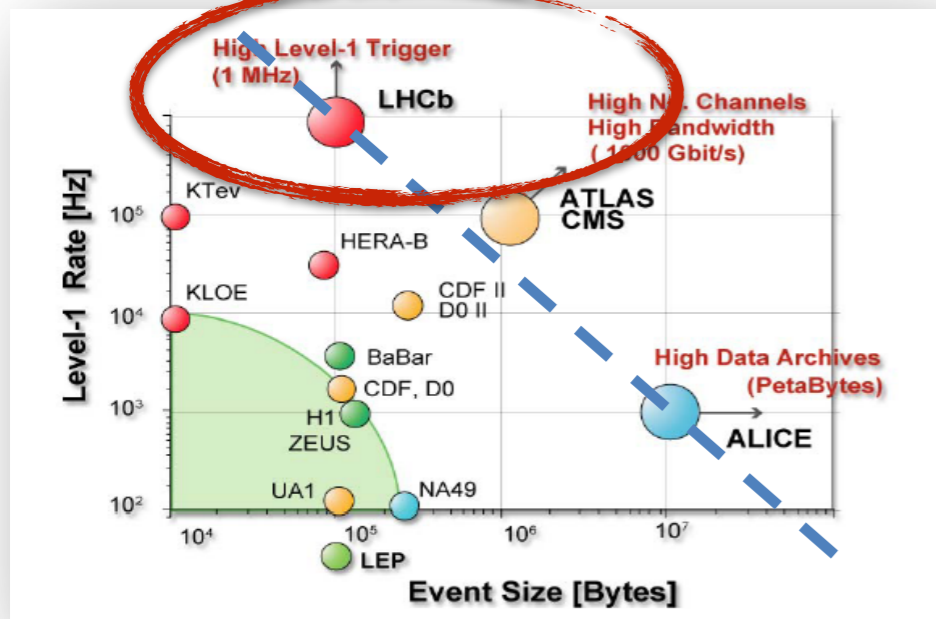
Tracking reconstruction not feasible @40MHz, nor in few microseconds

	ATLAS [1]	CMS [2]
<i>data reduction @40MHz</i>	regions from L1 (Rols)	stubs from hw coincidences
<i>track finding @1MHz</i>	Studying best algorithms to run in FPGAs and/or in GPUs	
<i>track fit @1MHz</i>		
<i>precision tracking @100kHz</i>	optimized offline	optimized offline



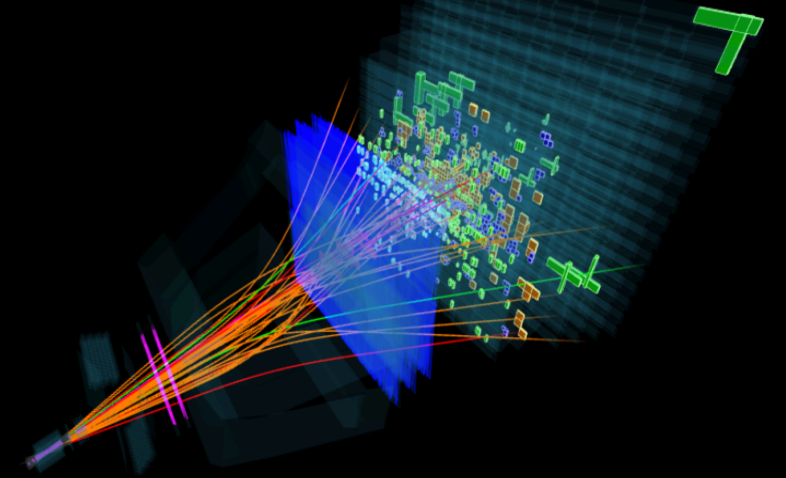
stubs in CMS PT modules





LHCb  
HERA-B

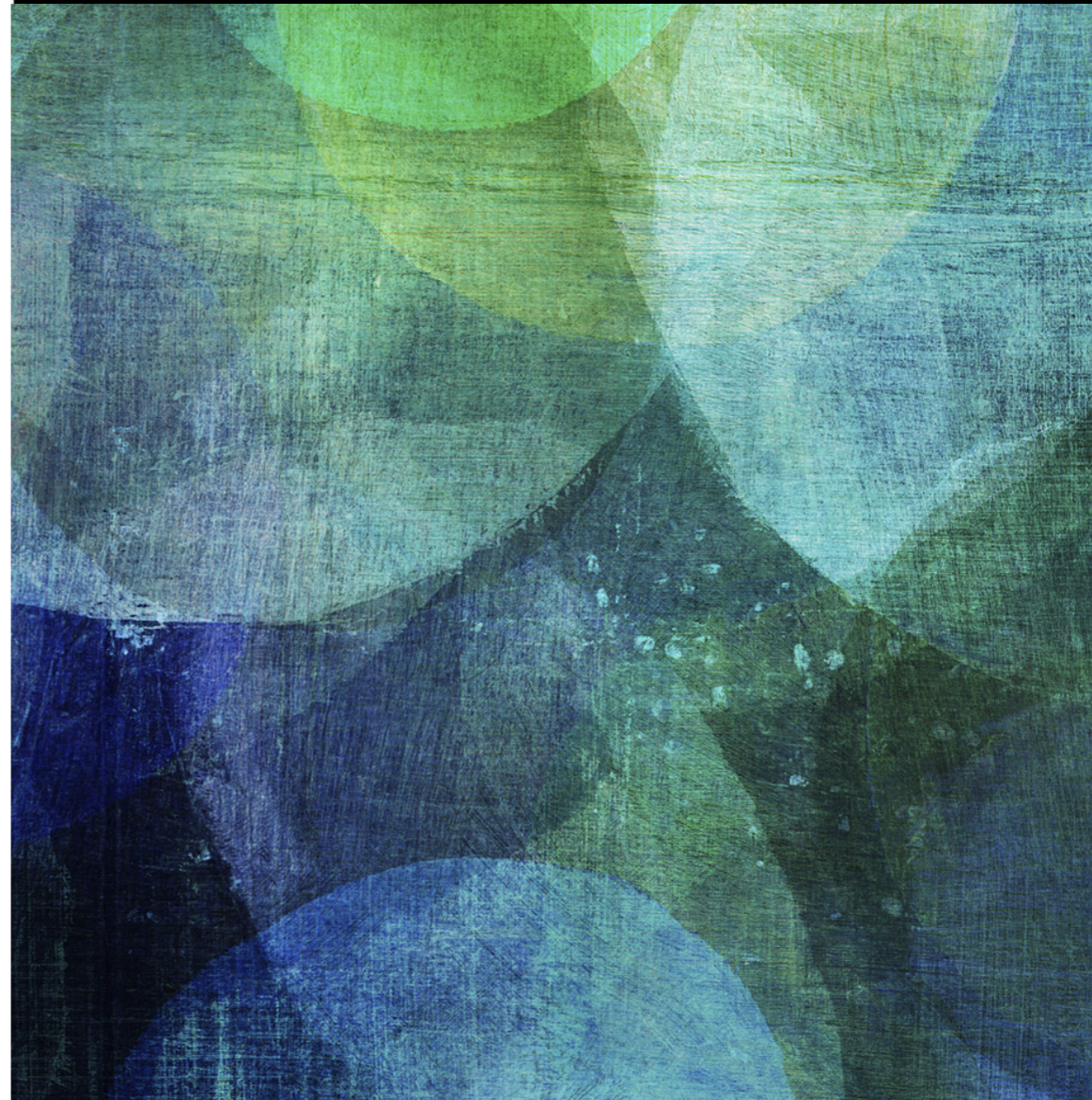
Event 158826354  
Run 206854  
Sat, 28 Apr 2018 21:48:17



# LHCb, THE B-MESON OBSERVATORY

*The lightest experiment to study the heavy b-quark*

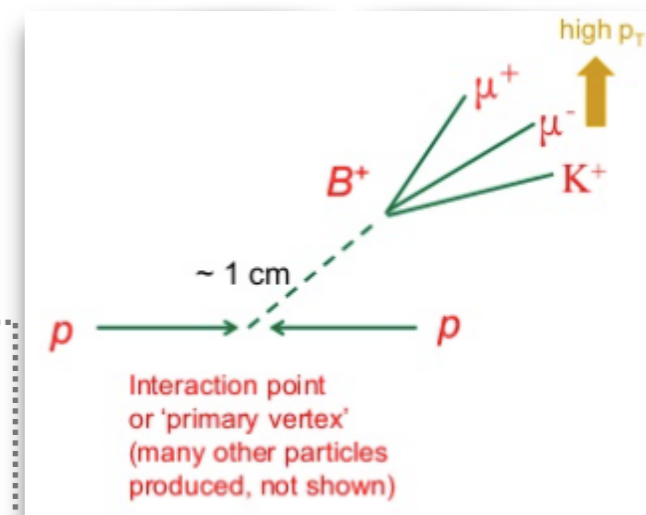
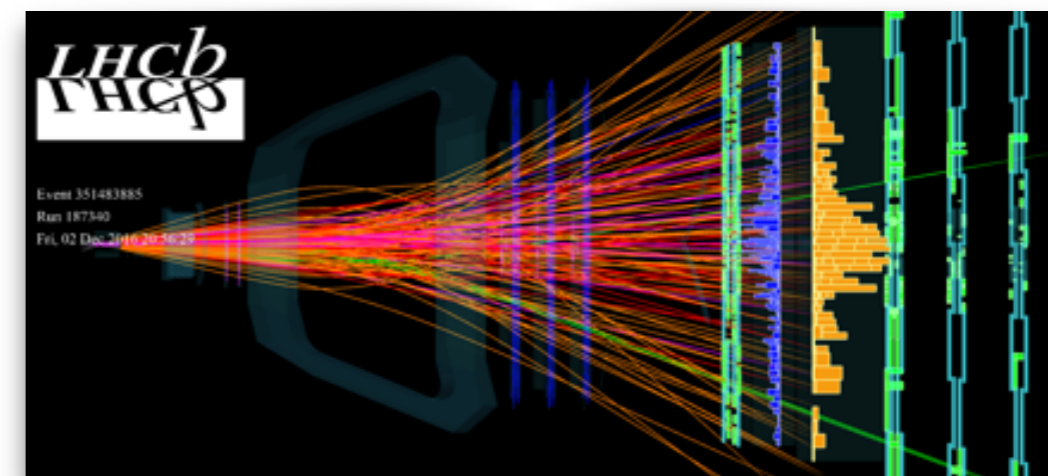
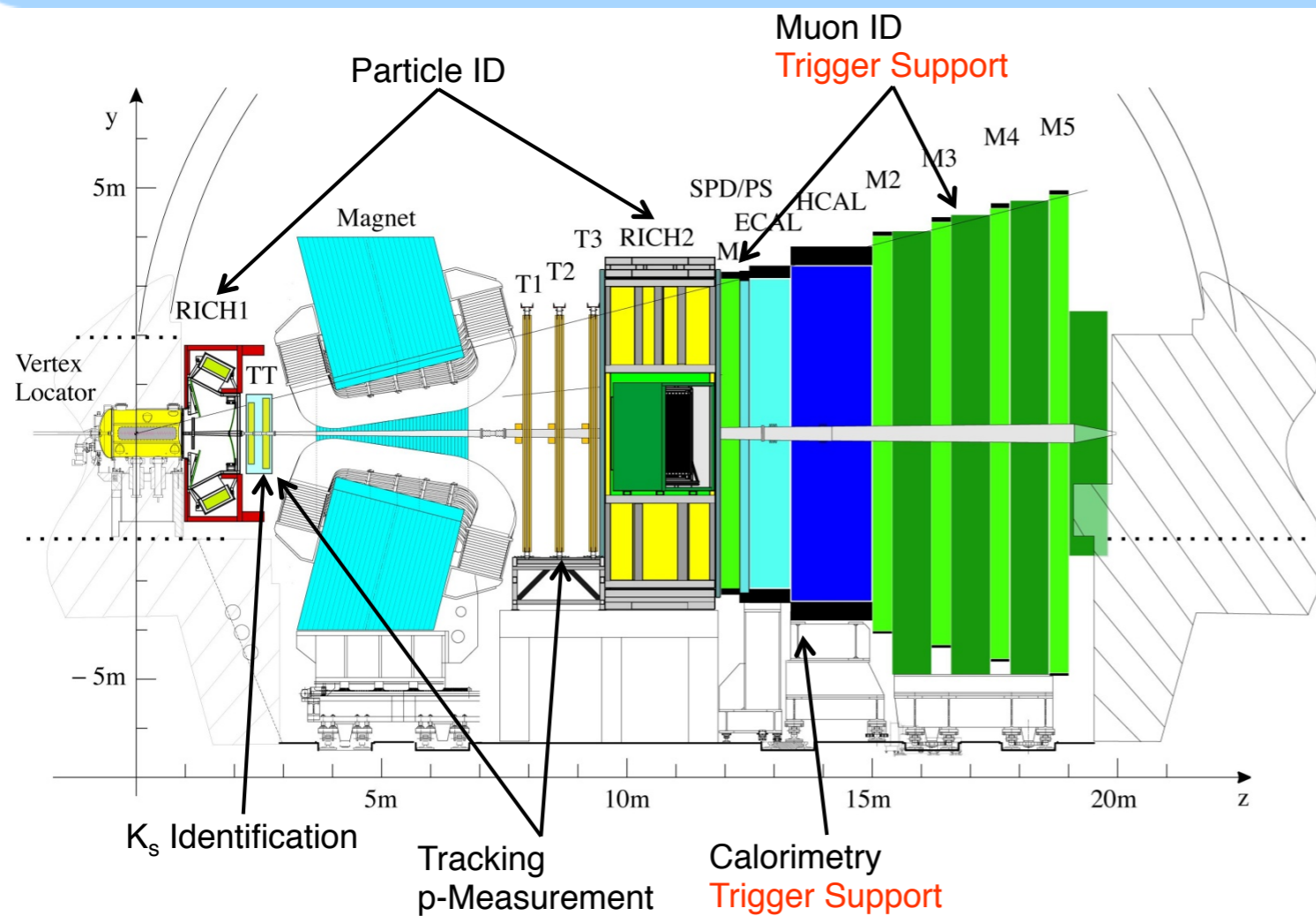
<http://lhcb-public.web.cern.ch/lhcb-public/>





## → Precision measurements and rare decays in the B system

- Large production ( $\sigma_{BB} \sim 500 \mu\text{b}$ ), but still  $\sigma_{BB}/\sigma_{\text{Tot}} \sim 5 \times 10^{-3}$
- Interesting B decays are quite rare ( $\text{BR} \sim 10^{-5}$ )



- Single-arm spectrometer and low L ⇒ **reduced event size**
- Selection of B mesons ⇒ **search for B-decay topologies**
- related to high mass and long lifetime of the b-quark

## LHCb 2012 Trigger Diagram

**40 MHz bunch crossing rate**

Input rate

### Low input rate and occupancy

- ◆ Limited acceptance: 10 MHz
- ◆ Limited Luminosity =  $2 \times 10^{32} \text{cm}^{-2}\text{s}^{-1}$

**L0 Hardware Trigger : 1 MHz readout, high  $E_T/P_T$  signature**

L0 trigger

- ◆ Select Bs in hadronic triggers
- ◆ Reject complex/busy events

450 kHz  
 $h^\pm$

400 kHz  
 $\mu/\mu\mu$

150 kHz  
 $e/\gamma$

### Software High Level Trigger

29000 Logical CPU cores

Offline reconstruction tuned to trigger time constraints

Mixture of exclusive and inclusive selection algorithms

60kB \* 1MHz = 60 GB/s readout network

**5 kHz (0.3 GB/s) to storage**

High Level

- ◆ Multitude of **exclusive selections**

2 kHz  
Inclusive  
Topological

2 kHz  
Inclusive/  
Exclusive  
Charm

1 kHz  
Muon and  
DiMuon



# SCHEMA EVOLUTION

## LHCb 2015 Trigger Diagram

40 MHz bunch crossing rate

L0 Hardware Trigger: 1 MHz readout, high  $E_T/P_T$  signatures

450 kHz  $h^\pm$

400 kHz  $\mu/\mu\mu$

150 kHz  $e/\gamma$

### Software High Level Trigger

Partial event reconstruction, select displaced tracks/vertices and dimuons

150 kHz

Buffer events to disk, perform online detector calibration and alignment

Full offline-like event selection, mixture of inclusive and exclusive triggers

12.5 kHz Rate to storage

HLT-1

HLT-2

Can increase efficiency on B-hadrons?  
YES, use more precision!!

Real-time calibration and alignments

Synchronous with DAQ

Use tracks for selections on B-decay vertices (in 35ms)

Split with a large buffer (4PB)!

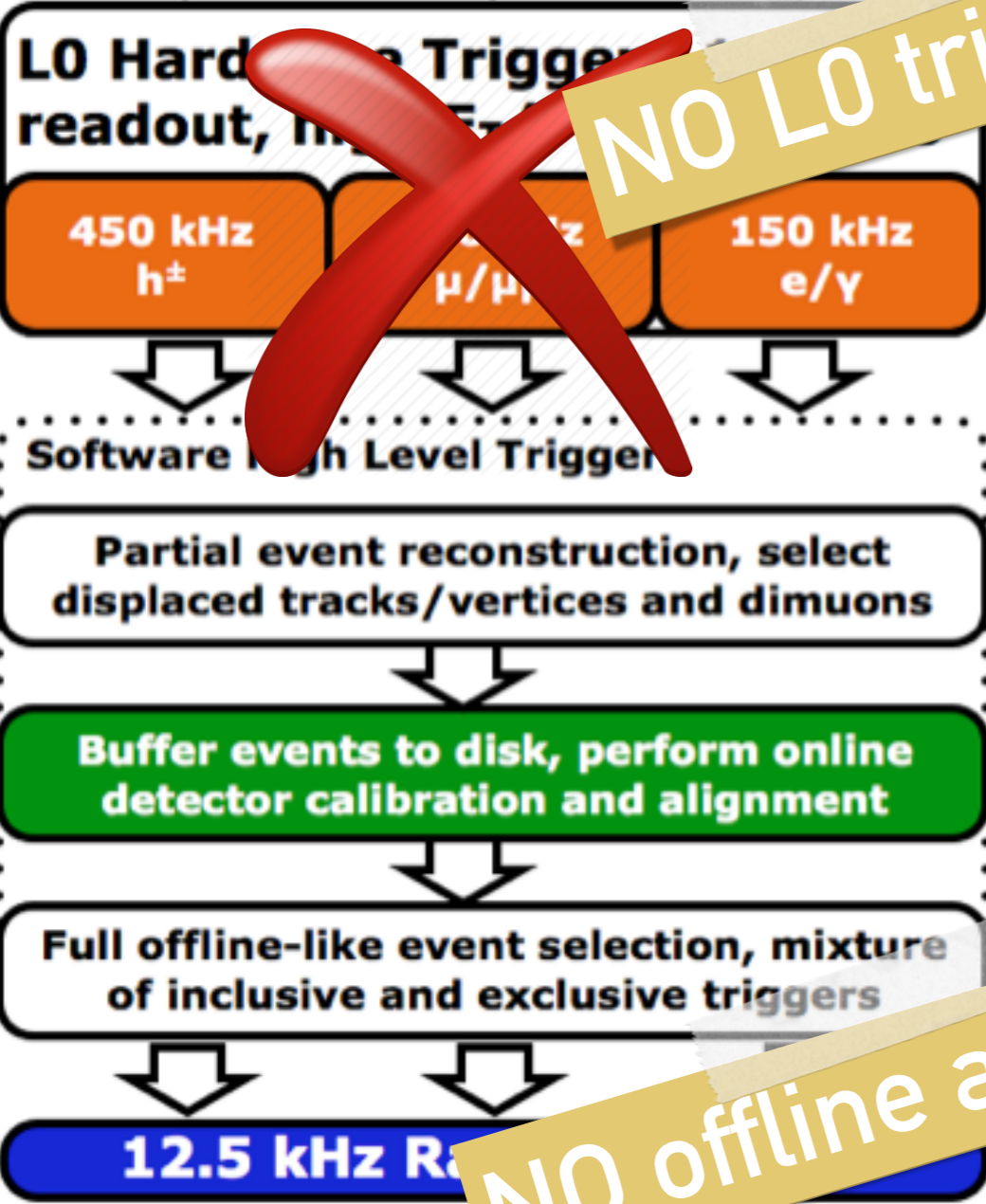
Deferred Processing

Reconstruct with offline-like calibrations (in 350ms), becoming real-time physics analysis

# UPGRADES FOR RUN 3

## LHCb 2015 Trigger Diagram

40 MHz bunch crossing rate



**NO L0 trigger**



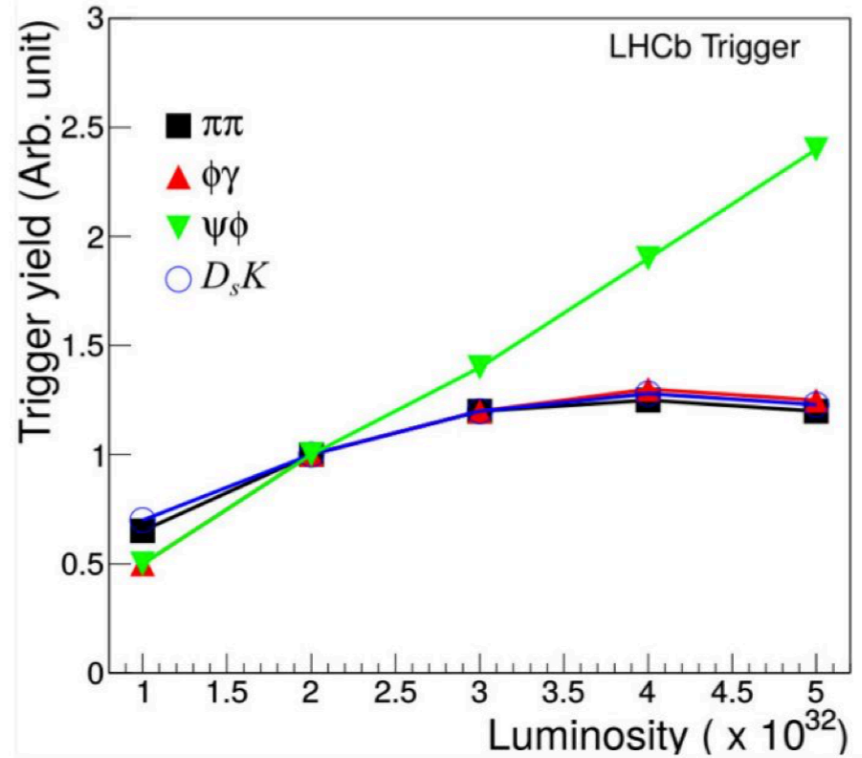
**NO offline analysis**

Can increase luminosity x10 ?  
Can increase b-hadron efficiency x2?



**YES, remove limit from L0 -1 MHz readout!**

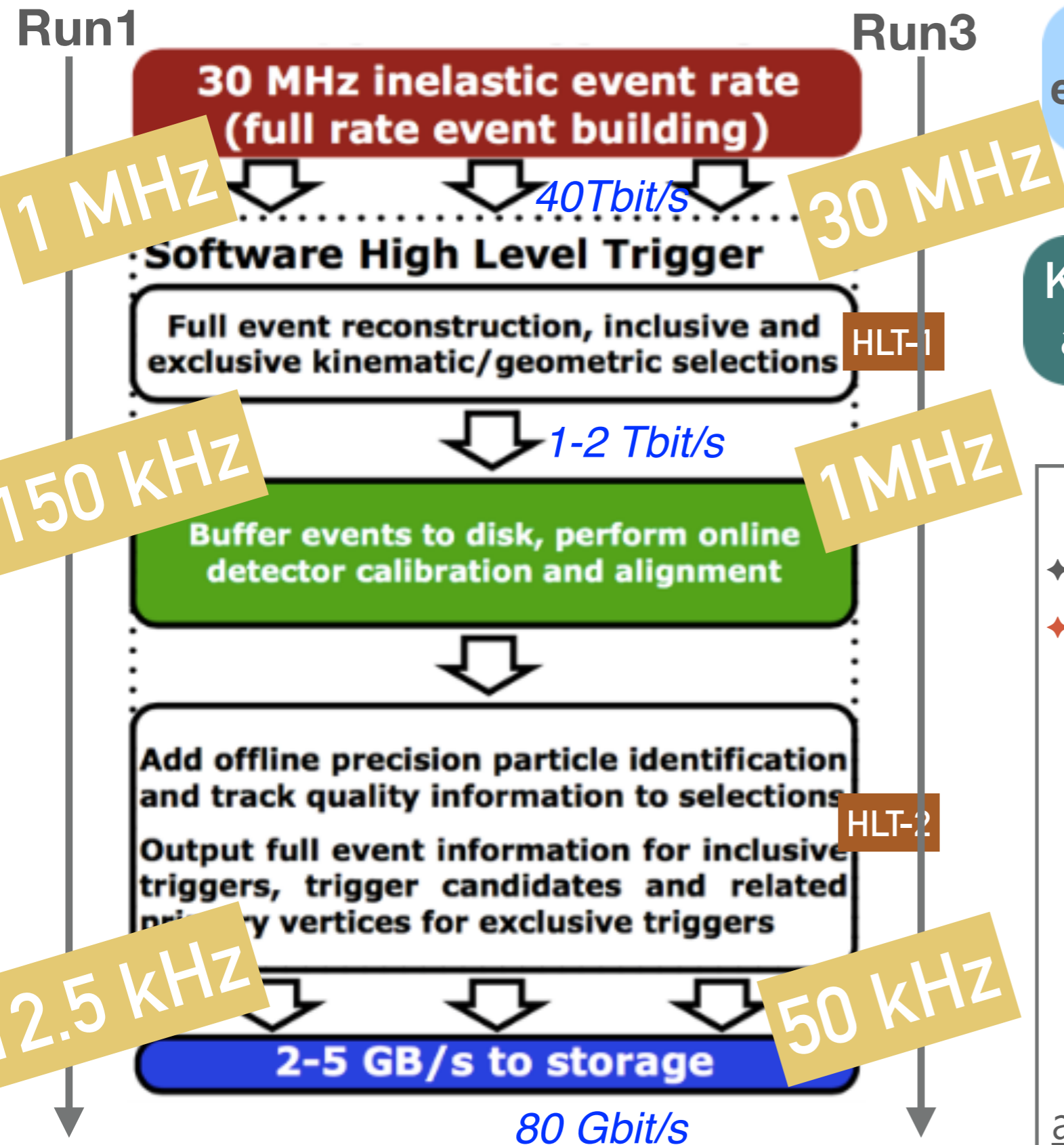
Increase in luminosity does not lead to increase of "interesting events"



Allow detector readout and reconstruction at unprecedented rate: **30MHz !!**



# TRIGGER-LESS?

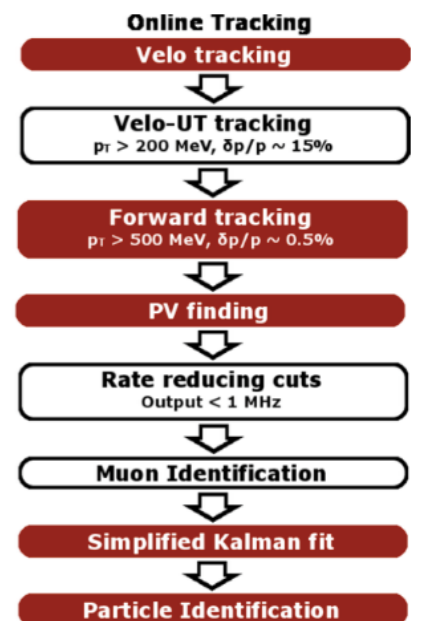
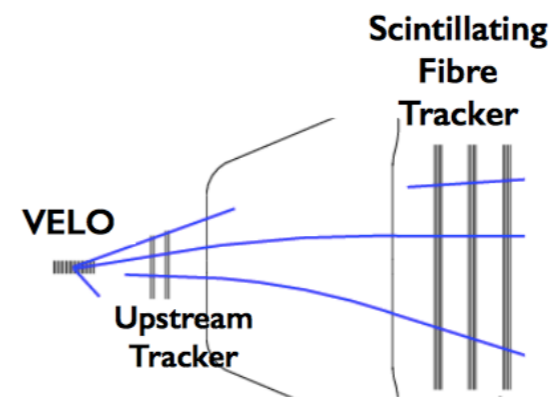


From Run1 to Run3, TDAQ system evolved to handle more readout rate

Key strategy: reduce data size at FE and suppress pileup with tracking

## Tracking at ~30 MHz?

- Run2: ~ 100k cores < 6 ms
- Run3: modern CPU & co-processors (FPGA/GPU)



[arXiv:2105.04031](https://arxiv.org/abs/2105.04031)

# LHCB IN RUN3: NETWORK IS DATAFLOW

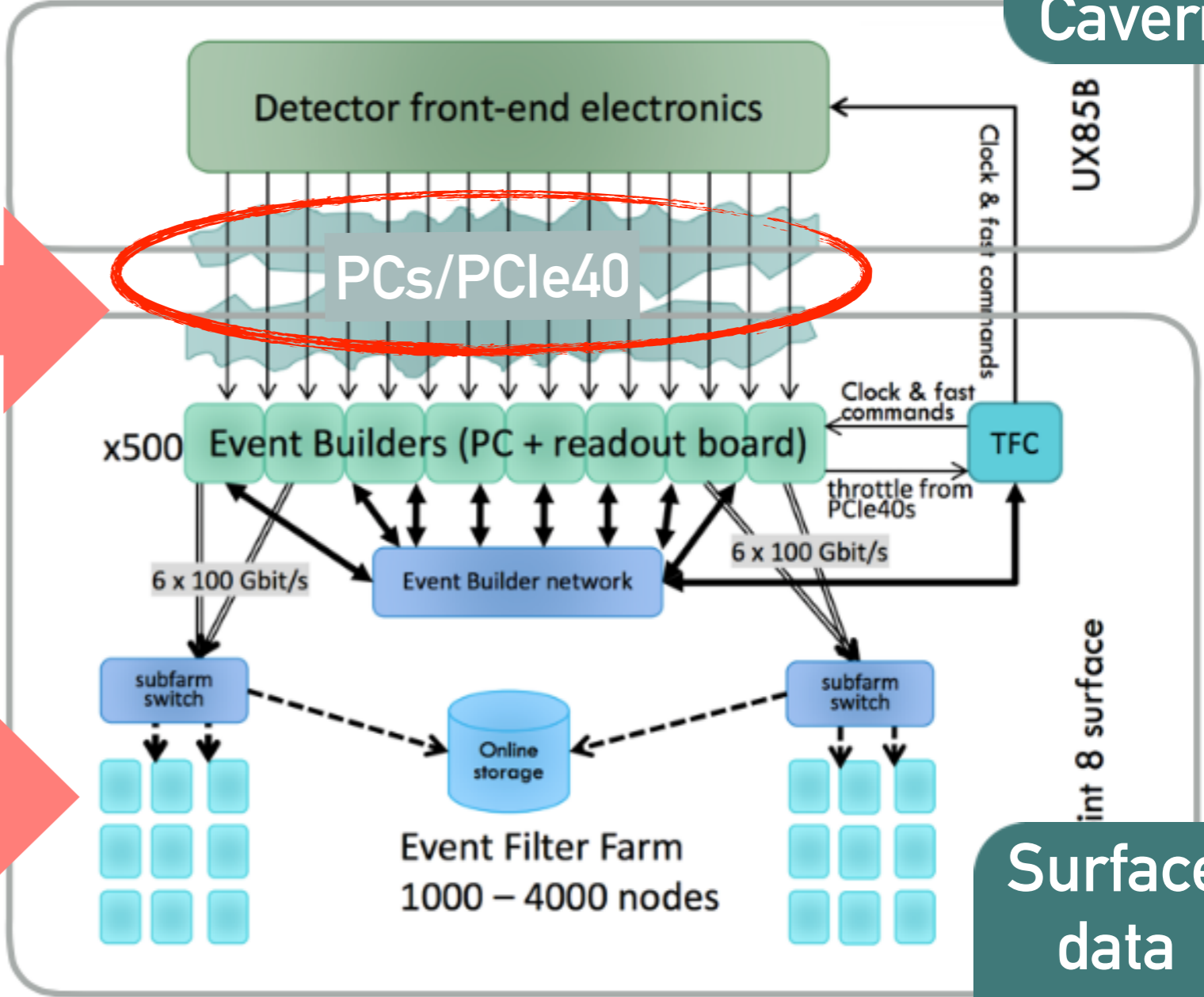
Readout @ 30 MHz  
Event size ~ 150kB

$$150\text{kB} \times 30\text{MHz} = 40\text{Tbs}$$

Inside Cavern

- Data reduction with custom FPGA-card (**PCle40**), also used in ALICE
  - Data-packing for sub-detectors (zero-suppression, clustering)
- Data pushed to the Event Building with **massive link usage**:
  - ~10,000 GBT (4.8 Gb/s, rad-hard)

DAQ network < 40 Tbit/s  
Record rate: <100 kHz



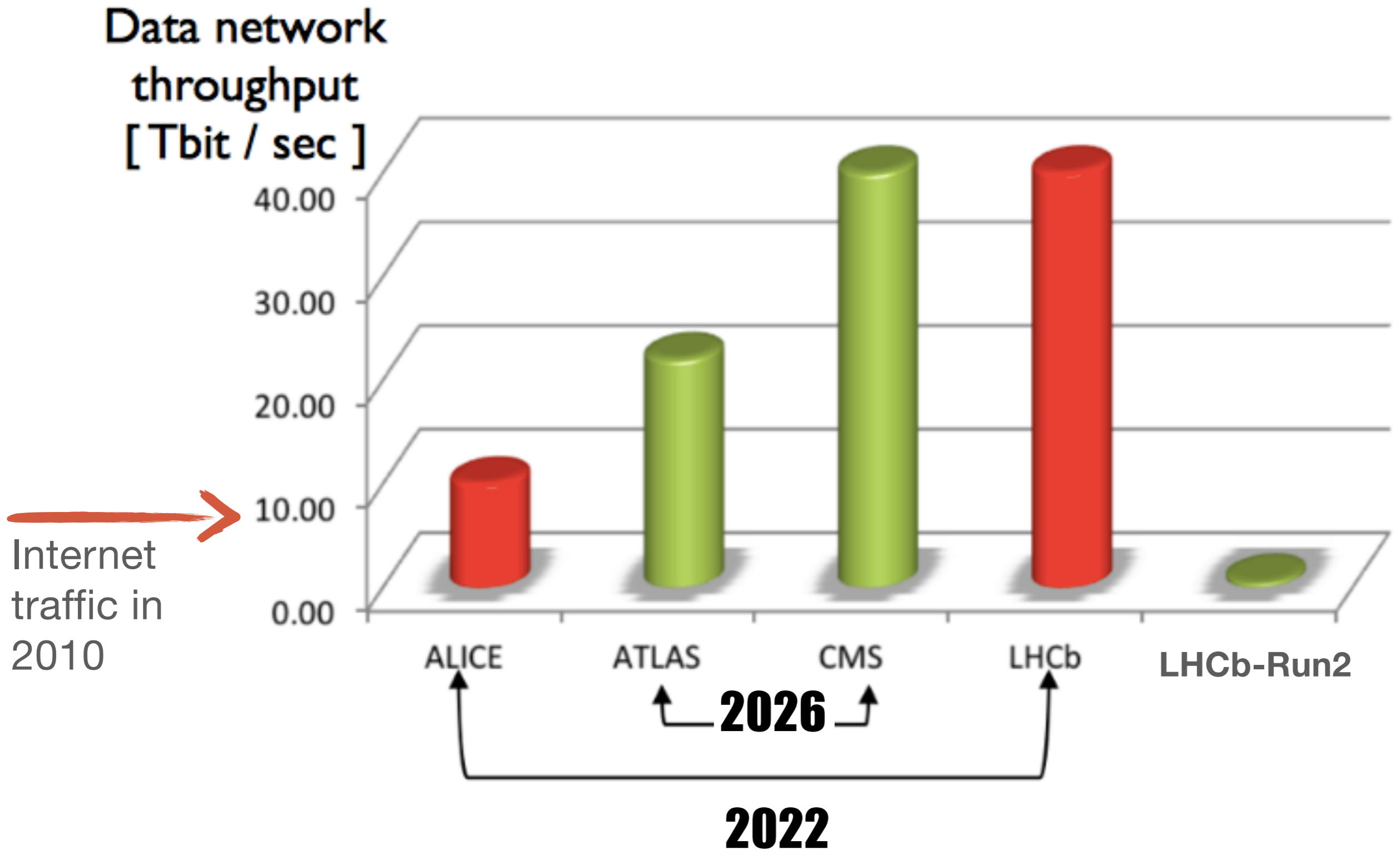
Surface data centre

PCle-gen3: simple protocol, large bandwidth  
PCle: maximum flexibility in later networking choice

*Ref for PCle40*

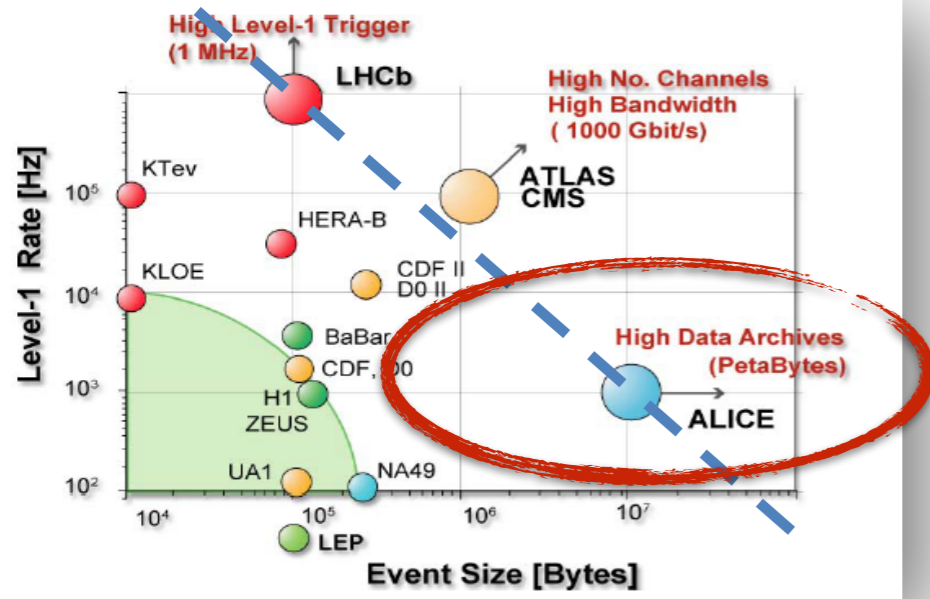


# NETWORK TRAFFIC COMPARISON



Same data volume as ATLAS/CMS HL-LHC upgrades! But earlier and for less money

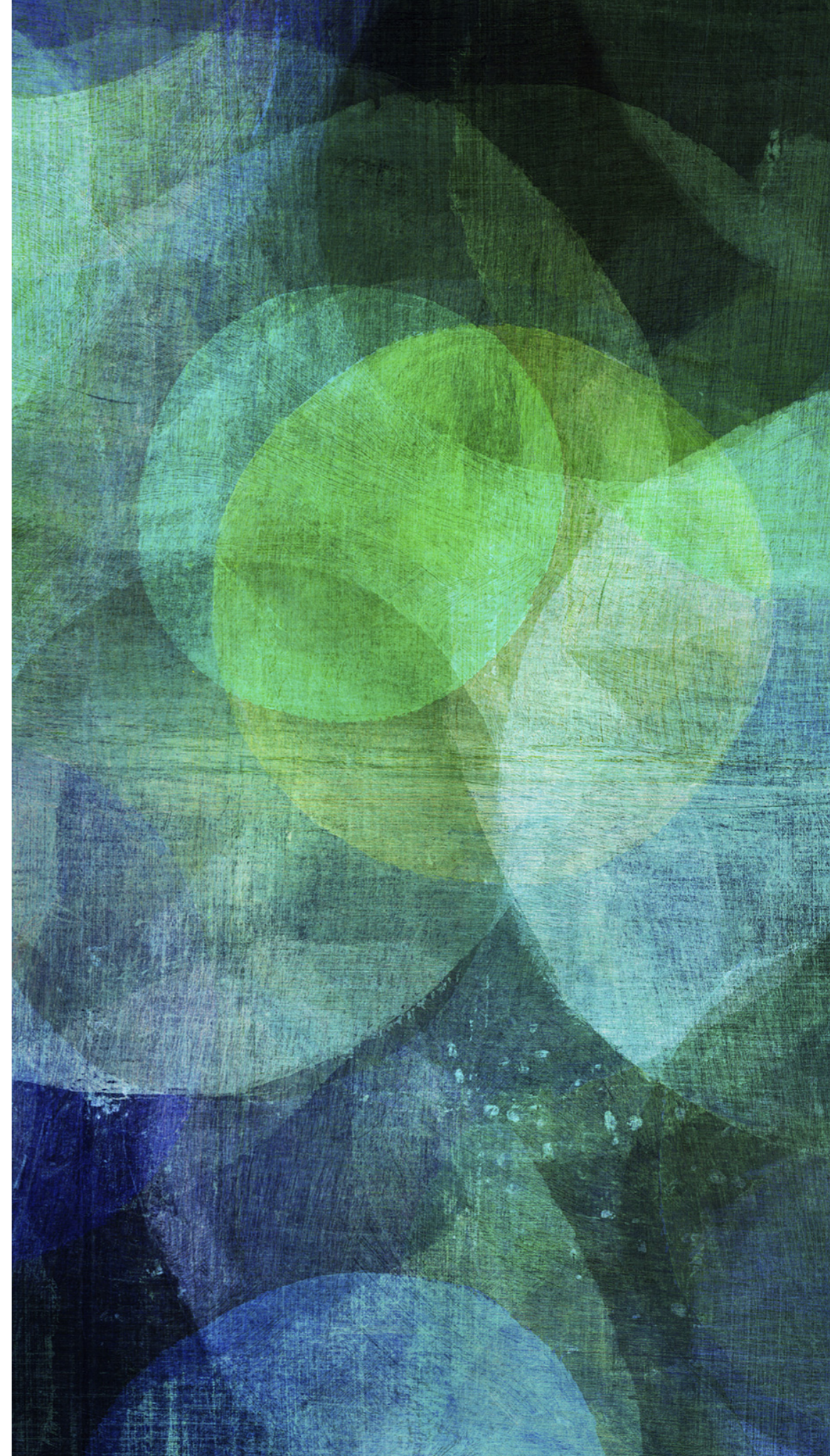




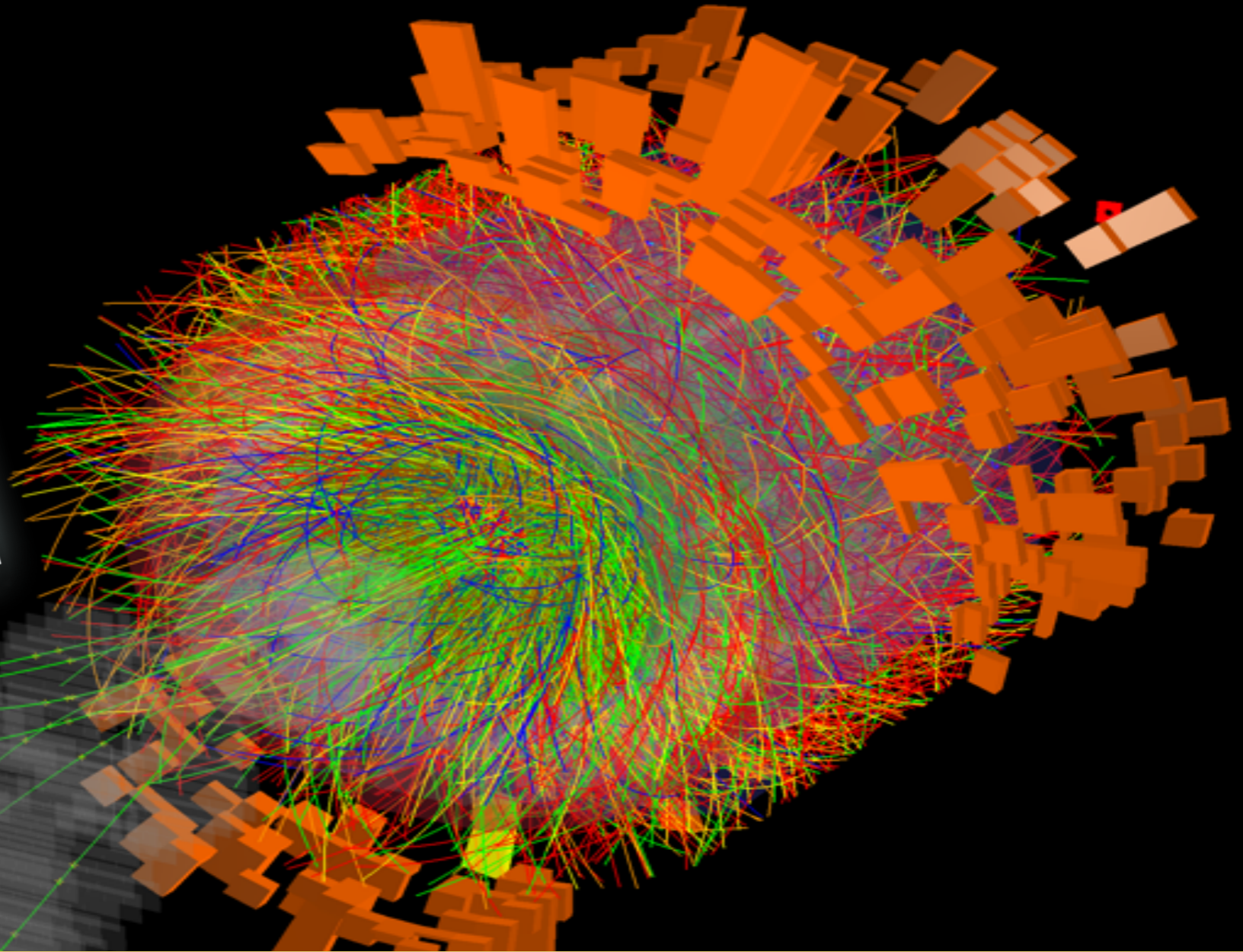
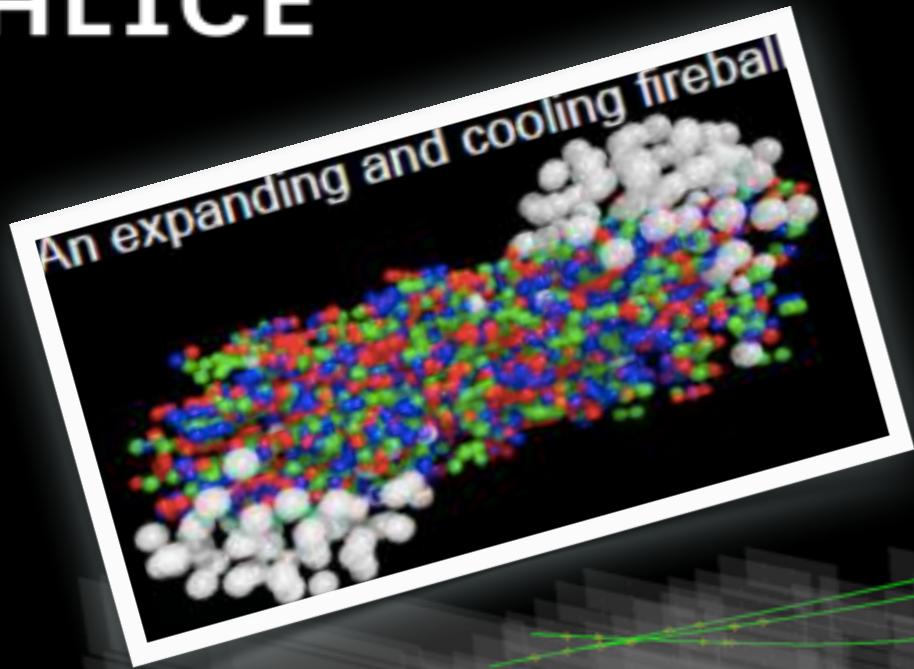
# ALICE: THE SMALL BIG-BANG

*Recording heavy ion collisions*

<http://alice-daq.web.cern.ch>







- **Physics of strongly interacting matters & quark-gluon plasma, with nucleus-nucleus interactions**
  - High particle multiplicities ( $\sim 8000$  particles/d $\eta$ )
  - Identify heavy short-living particles
  - By selecting low- $p_T$  tracks ( $> 100$  MeV)

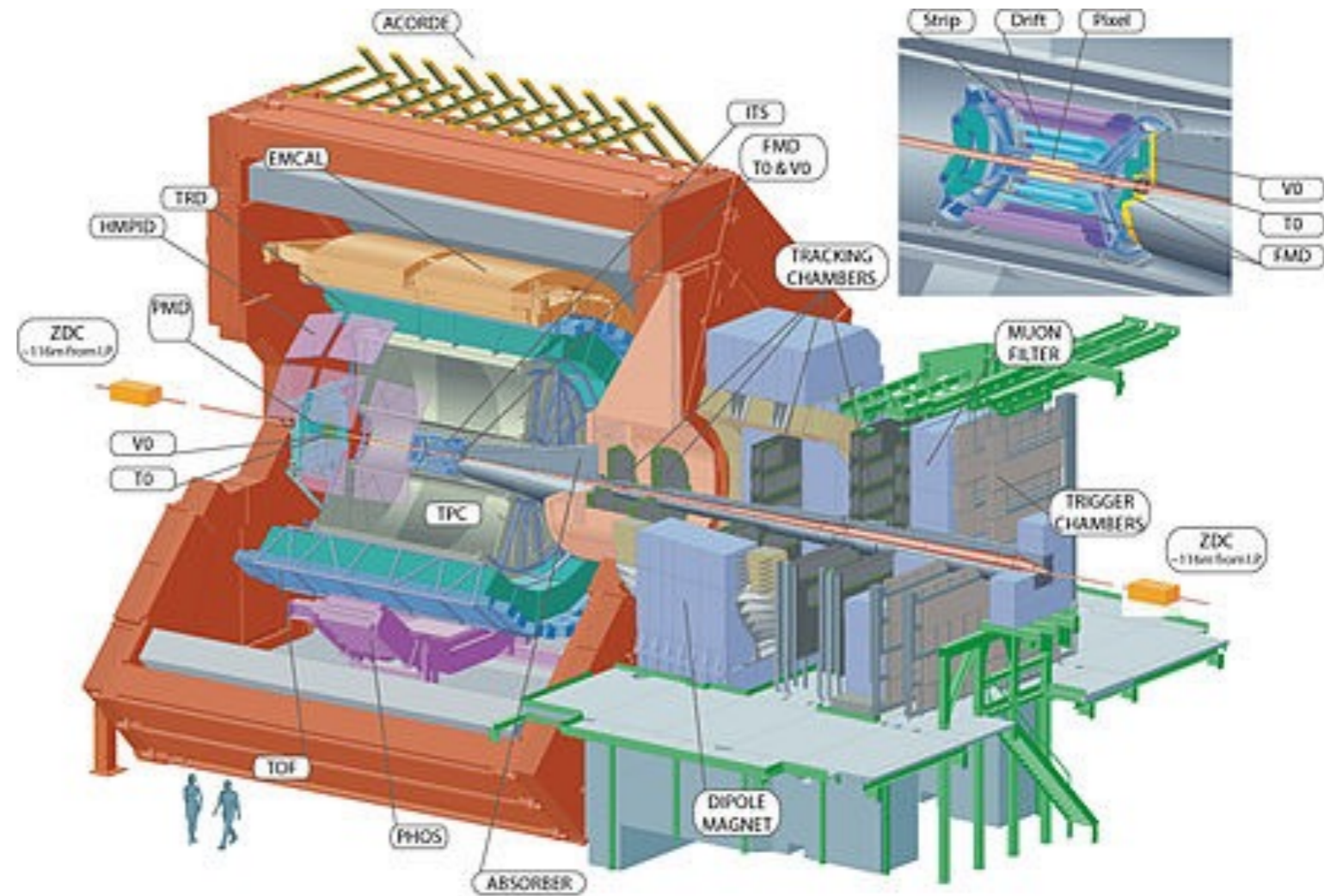


# DESIGNED FOR HEAVY ION COLLISIONS



ALICE

- ➔ 19 different detectors
- ➔ With high-granularity and timing information
  - ➔ in particular the Time Projection Chamber (TPC) has very high occupancy, and slow response
- ➔ Large event size ( $> 40\text{MB}$ )
  - ➔ TPC producing 90% of data
- ➔ Complex event topology
  - ➔ low trigger rate: max 3.5 kHz



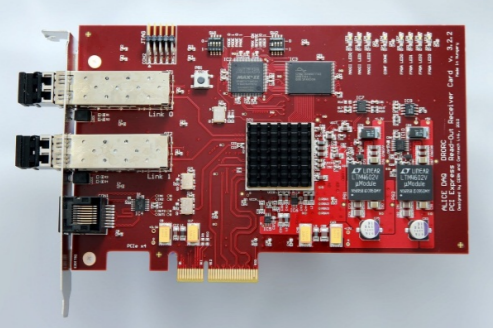

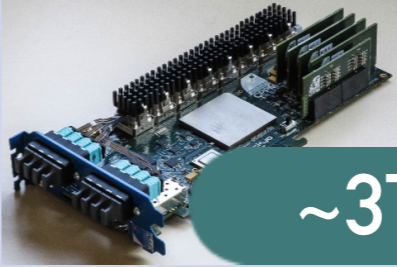
cms = 5.5 TeV per nucleon pair  
Pb–Pb collisions at  $L = 10^{27} \text{ cm}^{-2}\text{s}^{-1}$

- ➔ **Challenges for TDAQ design:**
  - ➔ detector readout: up to  $\sim 50 \text{ GB/s}$
  - ➔ storage:  $1.2 \text{ TB/s}$  (Pb–Pb)



- ➔ **LHC heavy ion programme: extend statistics by x100!**
  - ➔ Increase detector granularity (==> increase event size!)
  - ➔ Increase storage bandwidth x O(100)
    - ➔ Offline reconstruction also challenging due to combinatorics
  - ➔ Increase readout rates ~kHz → 50 kHz (==> need new and faster electronics)
    - ➔ Rate very close to TPC readout !!

## New TDAQ challenges!

RORC 1	C-RORC	CRU
		
2 ch @ 2 Gb/s PCIe gen.1 x4 (1 GB/s)	12 ch @ up to 6 Gb/s PCIe gen.2 x 8 (4 GB/s)	24 ch @ 5 Gb/s PCIe gen.3 X 16 (16 GB/s)
Custom DDL protocol	Custom DDL protocol (same protocol but faster)	GBT
Protocol handling TPC Cluster Finder	Protocol handling TPC Cluster Finder	Protocol handling TPC Cluster Finder Common-Mode correction Zero suppression

~3TB/s detector readout

New Common Readout Unit (CRU), based on PCIe40 card

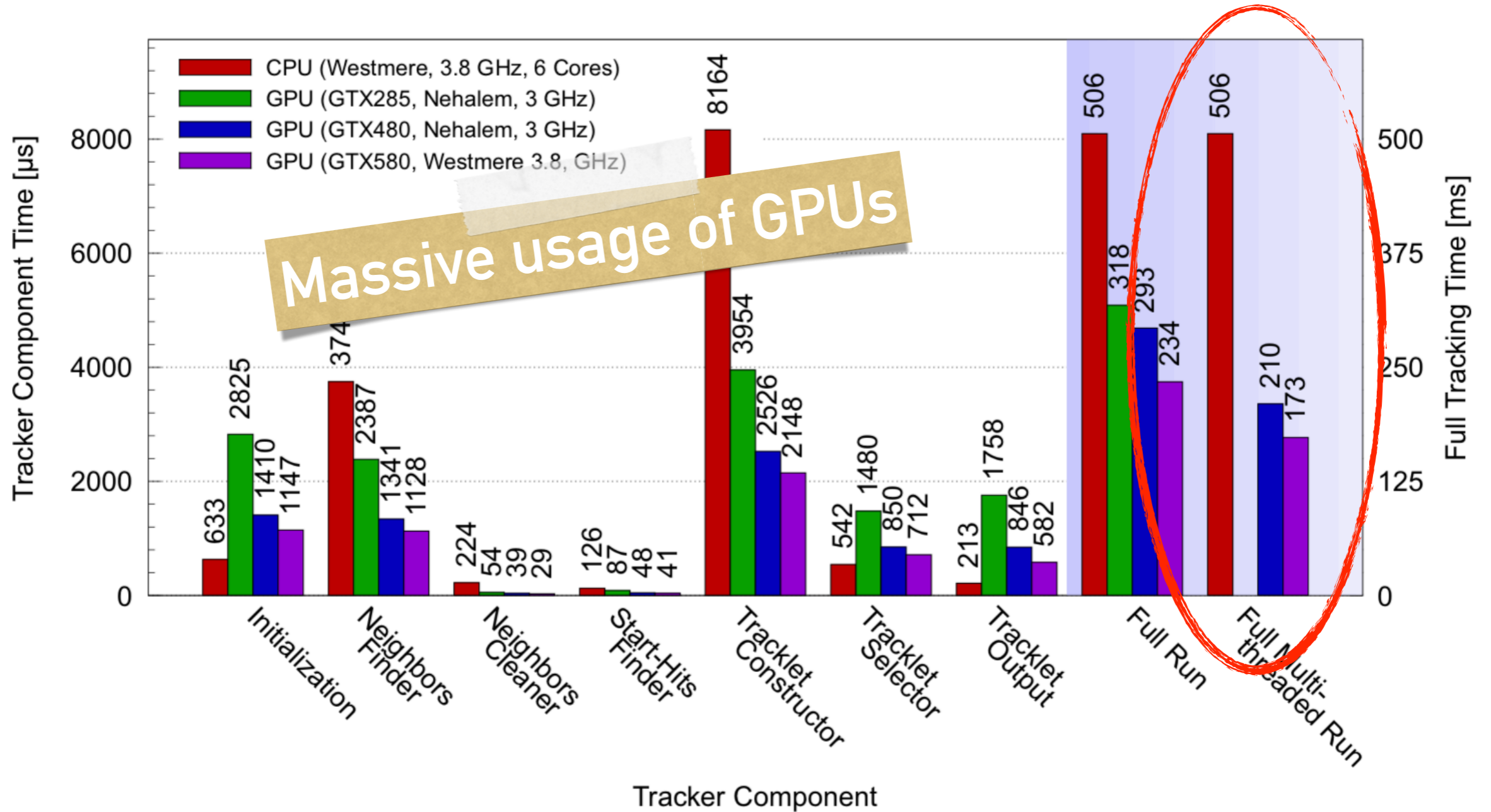


# INCREASING THROUGHPUTS WITH COTS



ALICE

- ➔ Data compression in GPUs and FPGAs ==> x2 readout rate
- ➔ Network evolution: 2.5GB/s (2010) => 6GB/s (2015) ==> x2 DAQ throughput



Tracking processing based on GPUs since Run1!



# RUN 3 DAQ: ONLINE RECONSTRUCTION



Higher rates with smaller data?

Store reconstruction,  
discard raw data

## Very heterogeneous system

- Synchronous, with continuous data
  - Data compression in FPGA/CPU
  - 30s to analyse 20ms-time frame

- Asynchronous, reconstruction in GPUs
  - 250 EPN servers with 8 GPU-cards
  - Require large-memory GPUs!

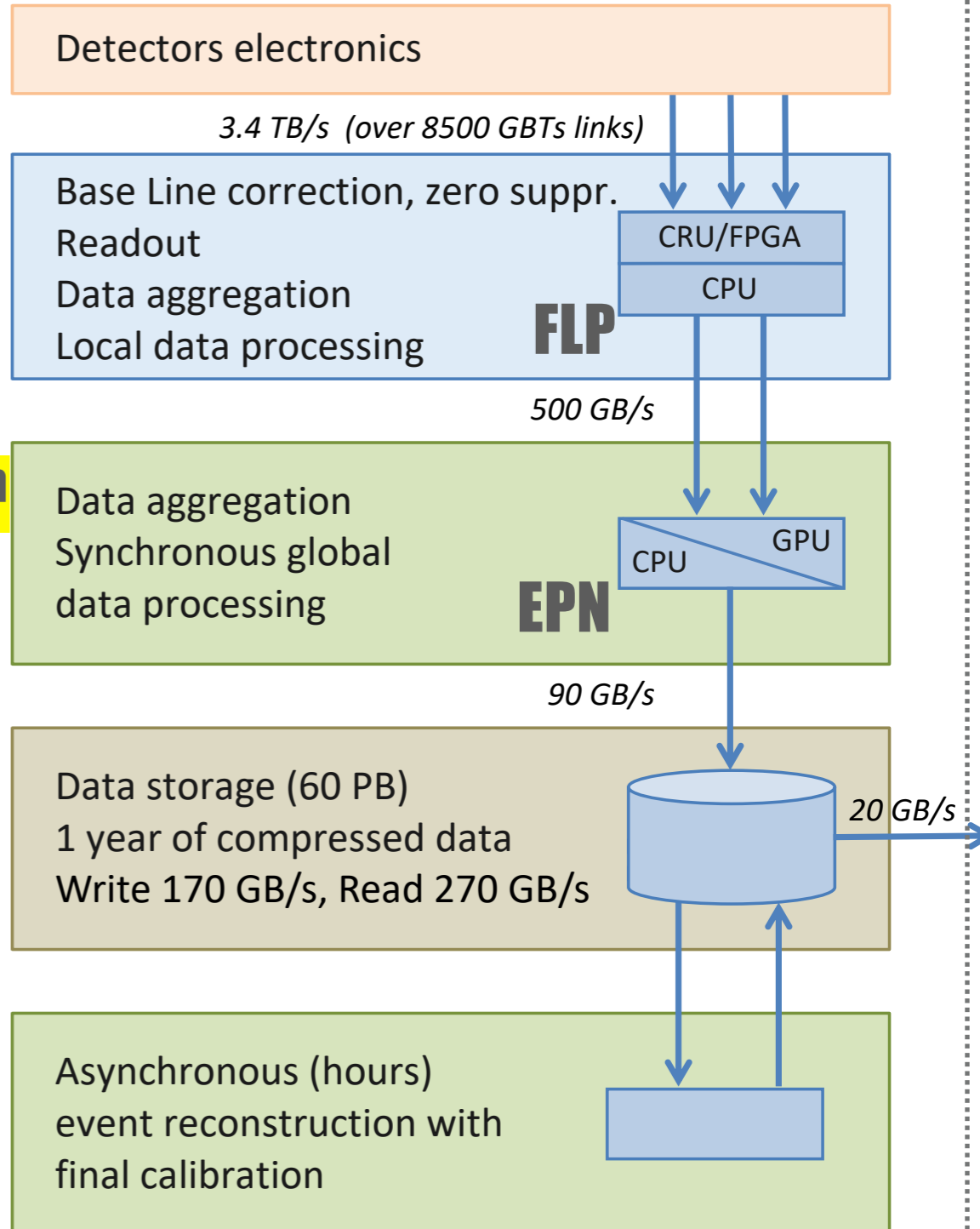
**O<sup>2</sup> system**

- Common online/offline software
  - Same calibrations and resources

**Data reduction**  
**Calibration 0**

**Data aggregation**  
**Reconstruction**  
**Calibration 1**

**More reconstruction**  
**Calibration 2**



# SUMMARY OF THE SUMMARIES

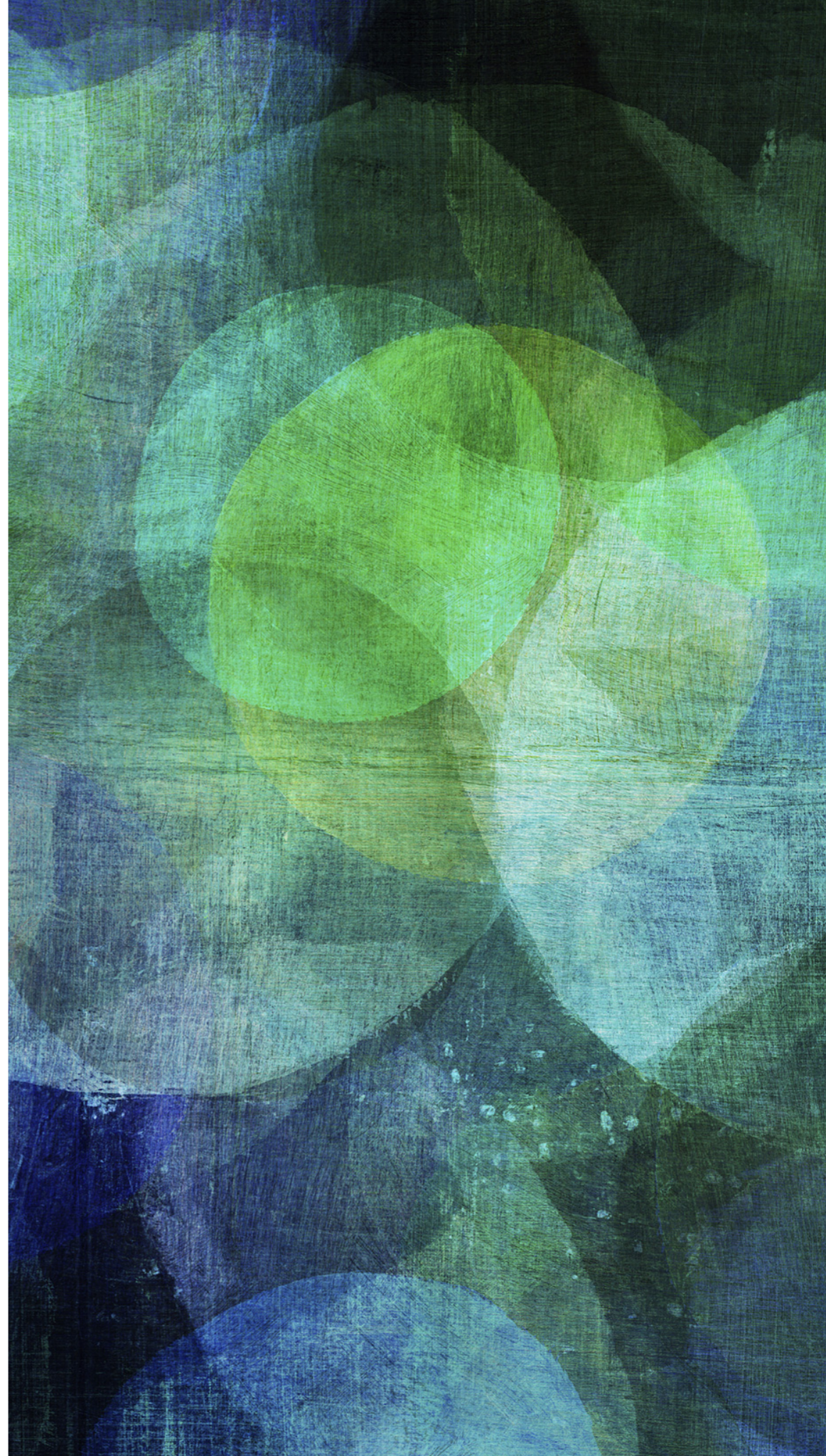
---

- **LHC experiments are among the largest and most complex TDAQ systems in HEP, to cope with a very difficult environment (always top LHC Luminosity)**
- **Continuous upgrade following the LHC luminosity, with different approaches**
  - **ATLAS/CMS** high-rate readout and Event Building, based on robust trigger selections
  - **LHCb** pioneer online-offline merging with large data throughputs
  - **ALICE** drives the GPU evolution and data compression
- **With a general trend, towards higher bandwidths and commodity HW**
  - Scalability not obvious. Challenge remains for front-end and back-end technologies and efficient (cost, time, power) computing farms
  - Moore's law still valid for processors but needs more effort to be exploited
- **Each experiment trying to gain advantage from others' developments**
  - joined efforts already started for hardware/software
  - sometimes stealing ideas (“... but we can do better than that...”)



# BACK-UP SLIDES

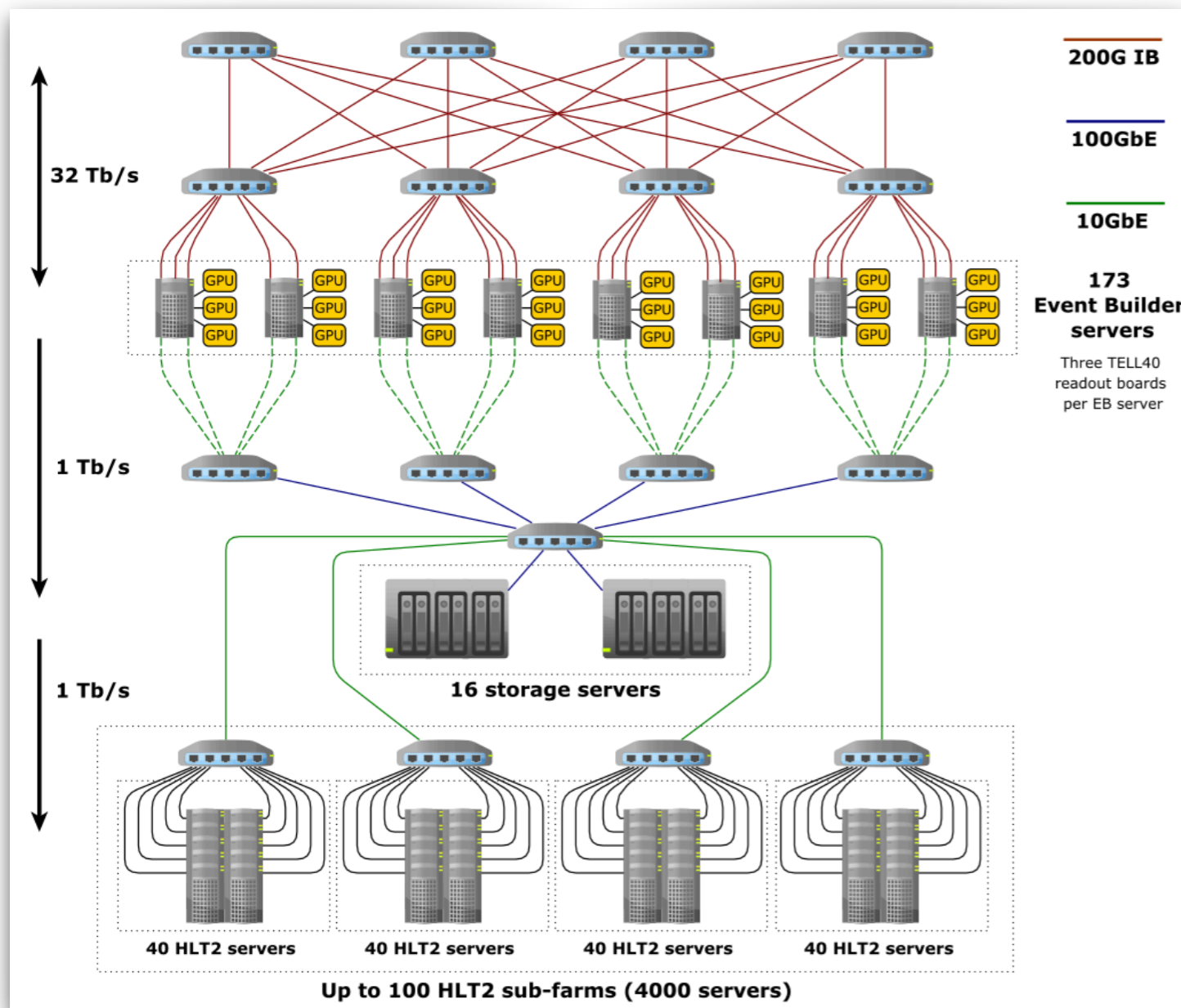
---





# A 2-DIM FOLDED EVENT BUILDING

Large farm of equal nodes with 8 PCIe40 boards, specialised by firmware

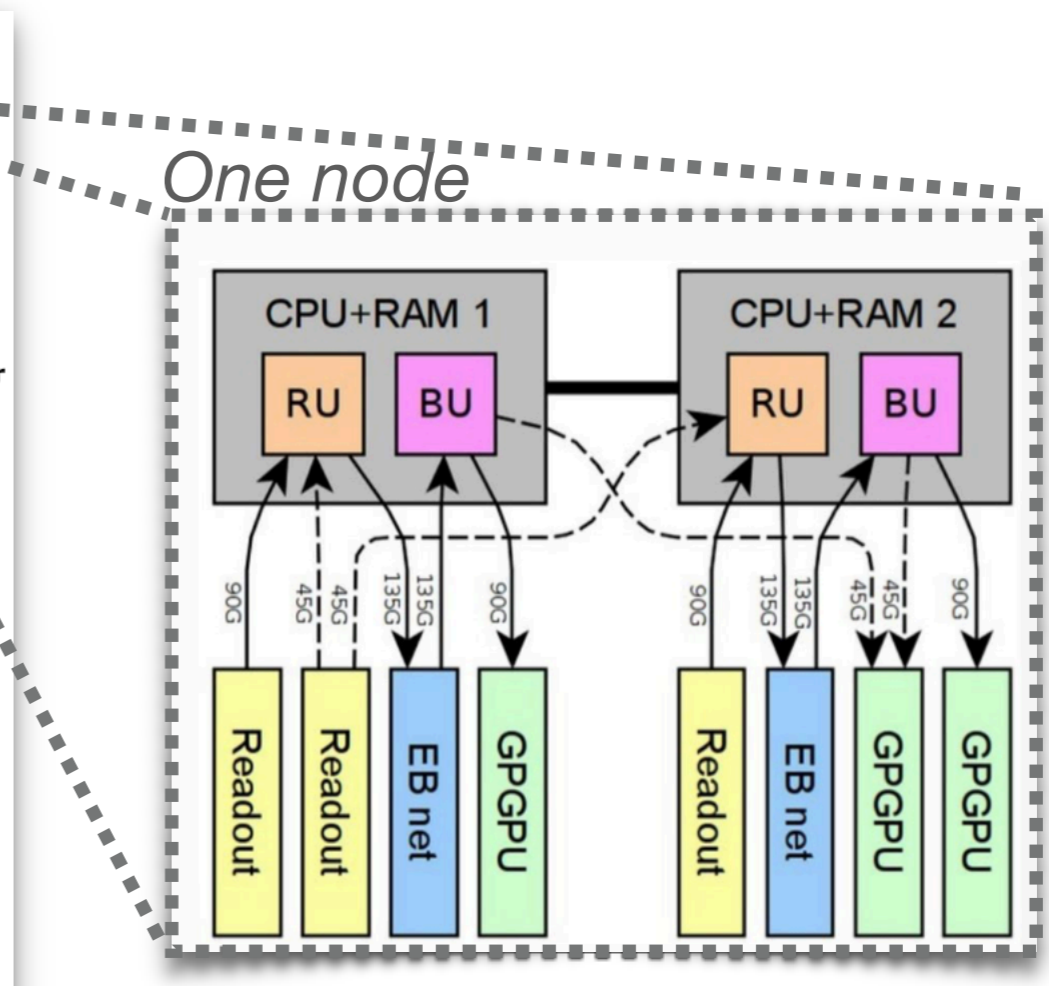
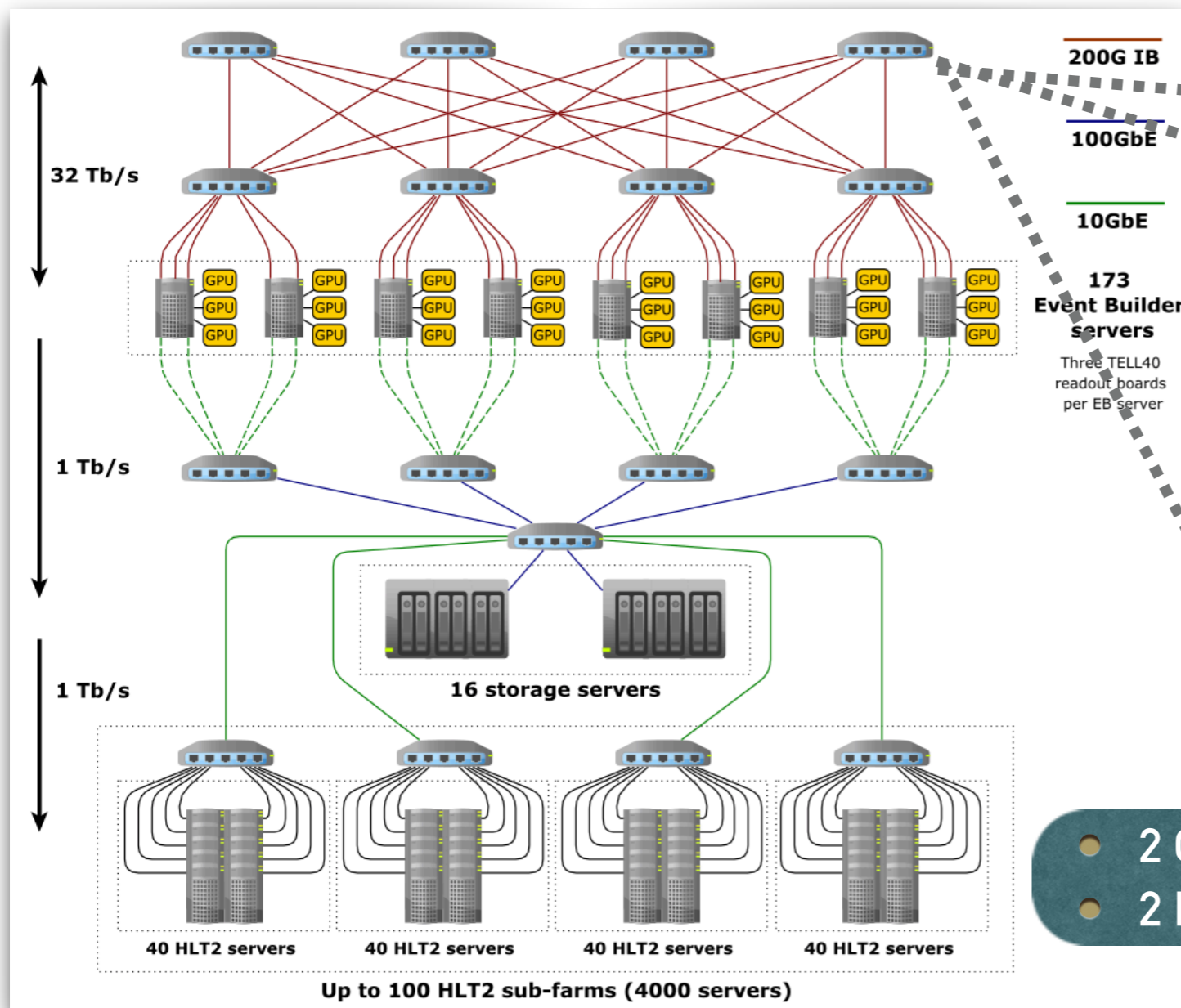


- ➔ EB network is oversized: able to manage 64Tb/s (320 network cards x 200Gb/s)
- ➔ Large rejection at HLT1: use O(200) GPU! throughput at ~100kHz
- ➔ Storage Buffer HLT1-HLT2 = 40 PB (3000 hard-disks) enough for days
  - ➔ SSD faster but have short lifetime wrt high read-write rate, so prefer hard-disks



# A 2-DIM FOLDED EVENT BUILDING

Large farm of equal nodes with 8 PCIe40 boards, specialised by firmware



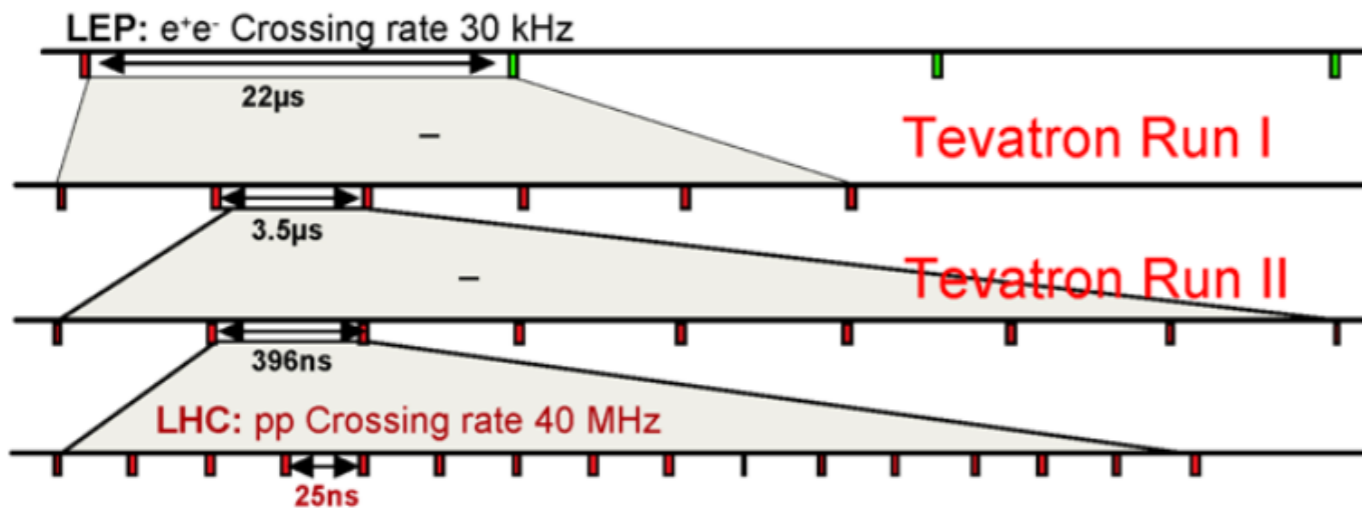
- 2 CPUs with large RAM (up to 512 GB!)
- 2 RU, 2 BU, 2 infiniband NIC (200 Gb/s), 1-3 GPUs

- ➔ EB network is oversized: able to manage 64Tb/s (320 network cards x 200Gb/s)
- ➔ Large rejection at HLT1: use O(200) GPU! throughput at ~100kHz
- ➔ Storage Buffer HLT1-HLT2 = 40 PB (3000 hard-disks) enough for days
  - ➔ SSD faster but have short lifetime wrt high read-write rate, so prefer hard-disks

# LHC: THE SOURCE

## The clock source

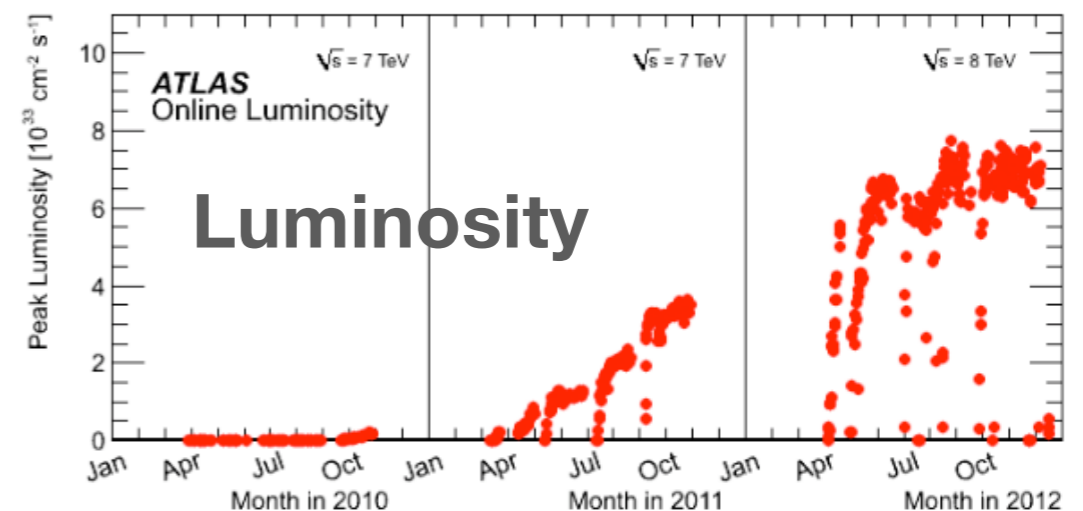
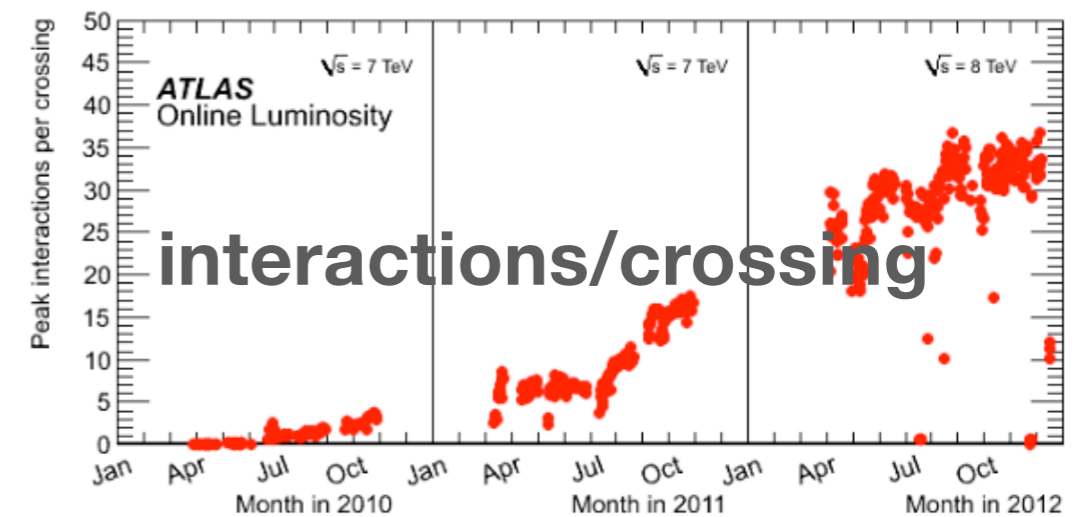
- ~3600 bunches in 27km
- distance bw bunches:  $27\text{km}/3600 = 7.5\text{m}$
- distance bw bunches in time:  $7.5\text{m}/c = 25\text{ns}$



At full Luminosity, every 25ns,  
~23 superimposed p-p  
interaction events

## The pile-up source

- more collisions/bunch crossing:  
~23 at design luminosity





# PIPELINED TRIGGERS

- ➔ **Allow trigger decision longer than clock tick (and no deadtime)**
  - ➔ Execute trigger selection in defined clocked steps (**fixed latency**)
  - ➔ Intermediate storage in stacked buffer cells
  - ➔ R/W pointers are moved by clock frequency

- ➔ **Tight design constraints for trigger/FE**

- ➔ **Analog/digital pipelines**

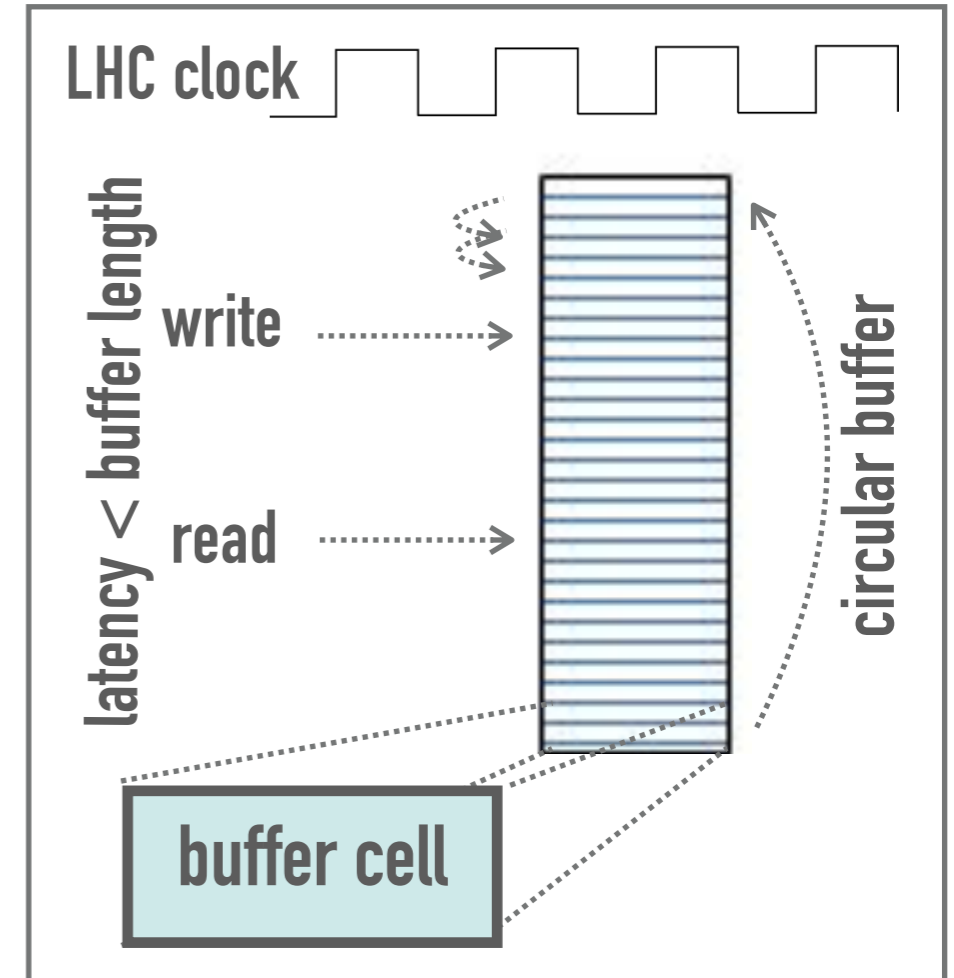
- ➔ Analog: built from switching capacitors
- ➔ Digital: registers/FIFO/...

- ➔ **Full digitisation before/after L1A**

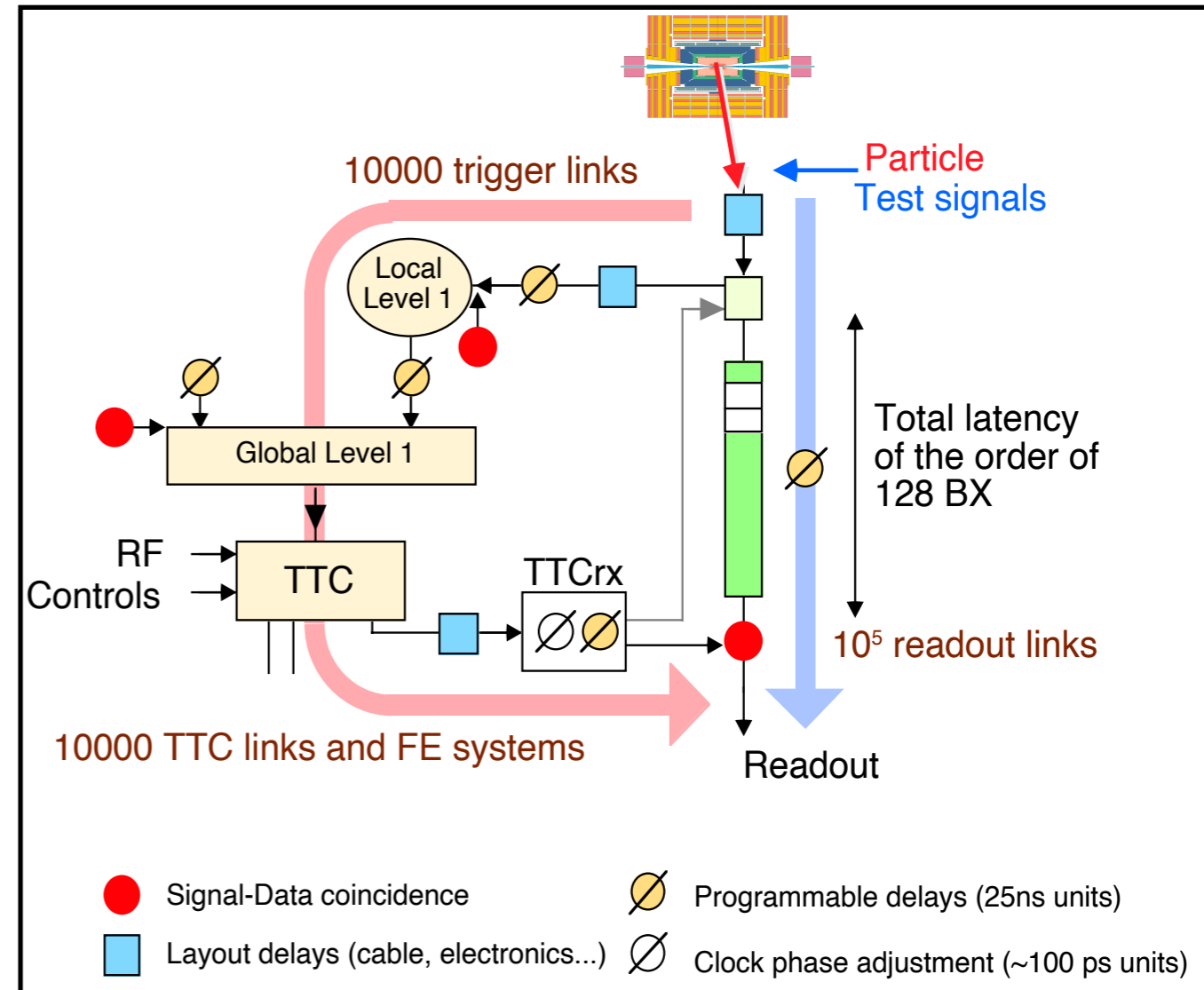
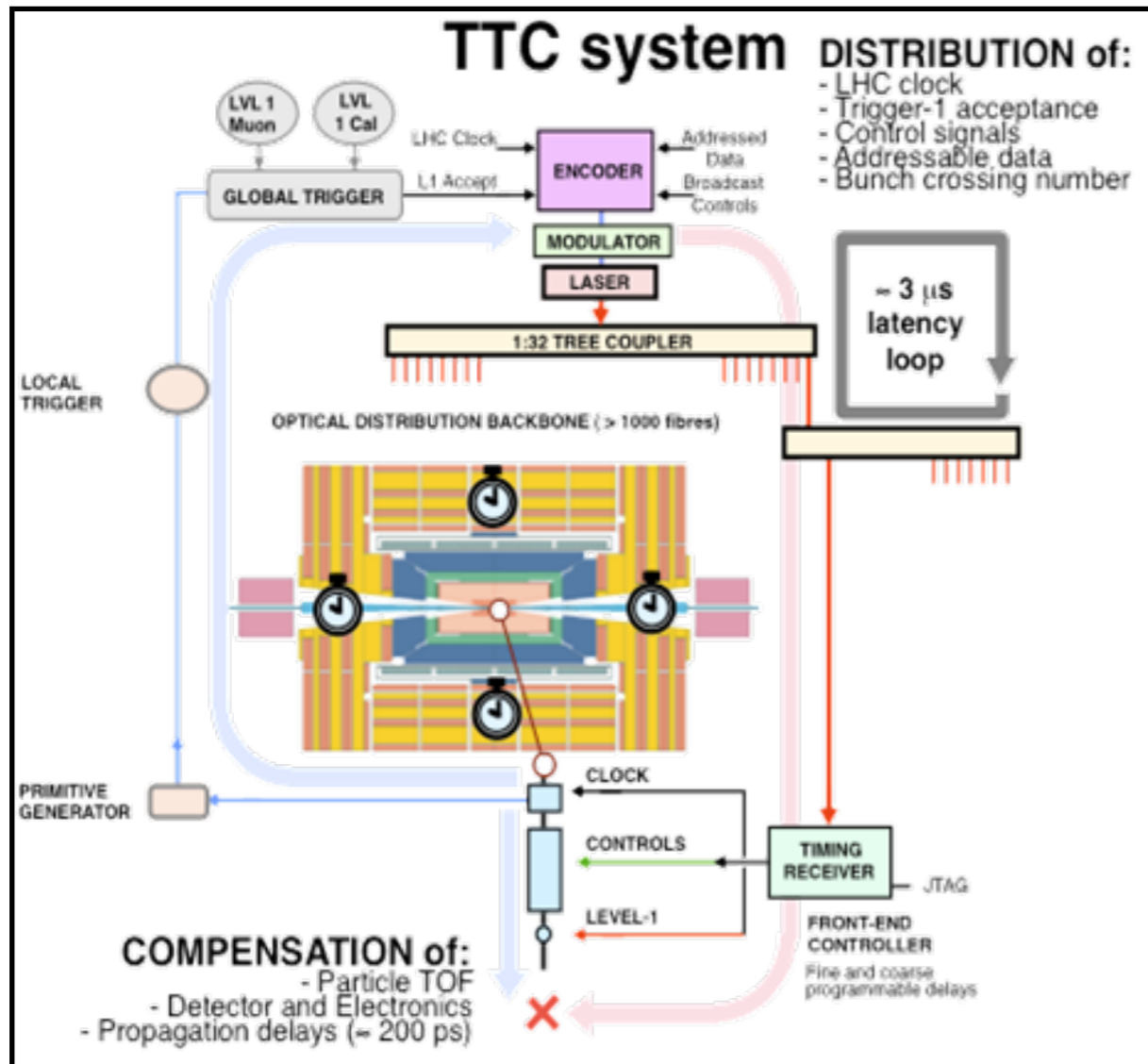
- ➔ Fast DC converters (power consumption!)

- ➔ **Additional complication: synchronisation**

- ➔ BC counted and reset at each LHC turn
- ➔ large optical time distribution system



# LOCAL TIMING AND ADJUSTMENTS

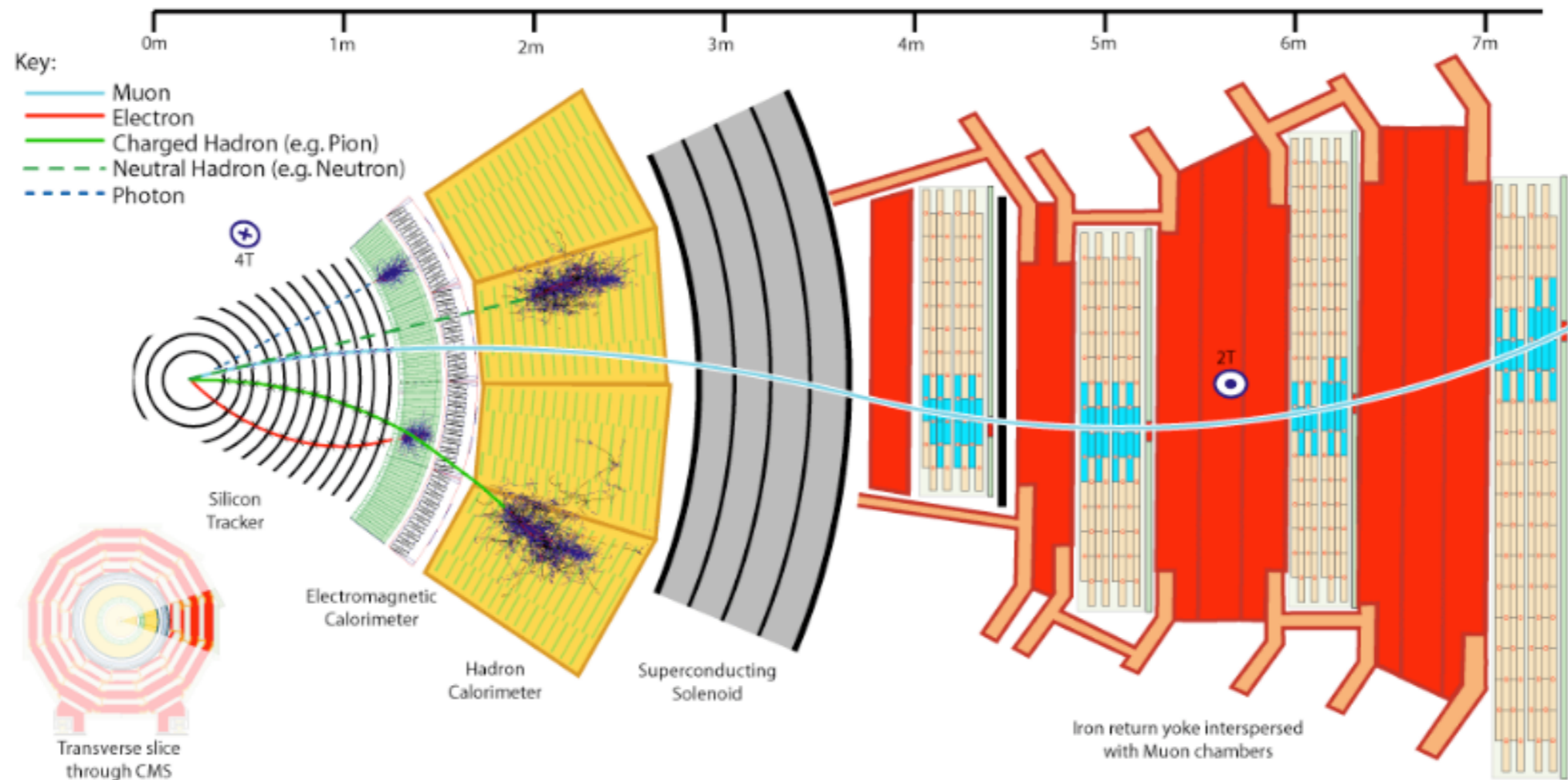


- ➔ **Common optical system: TTC**
  - ➔ radiation resistance
  - ➔ single high power laser
- ➔ **Large distribution**
  - ➔ experiments with  $\sim 10^7$  channels

- ➔ **Align readout & trigger at (better than) 25ns and correct for**
  - ➔ time of flight ( $25 \text{ ns} \approx 7.5\text{m}$ )
  - ➔ cable delays ( $10\text{cm/ns}$ )
  - ➔ processing delays ( $\sim 100$  BCs)



# TRIGGERS FOR MUONS



## ➔ Dedicated detectors:

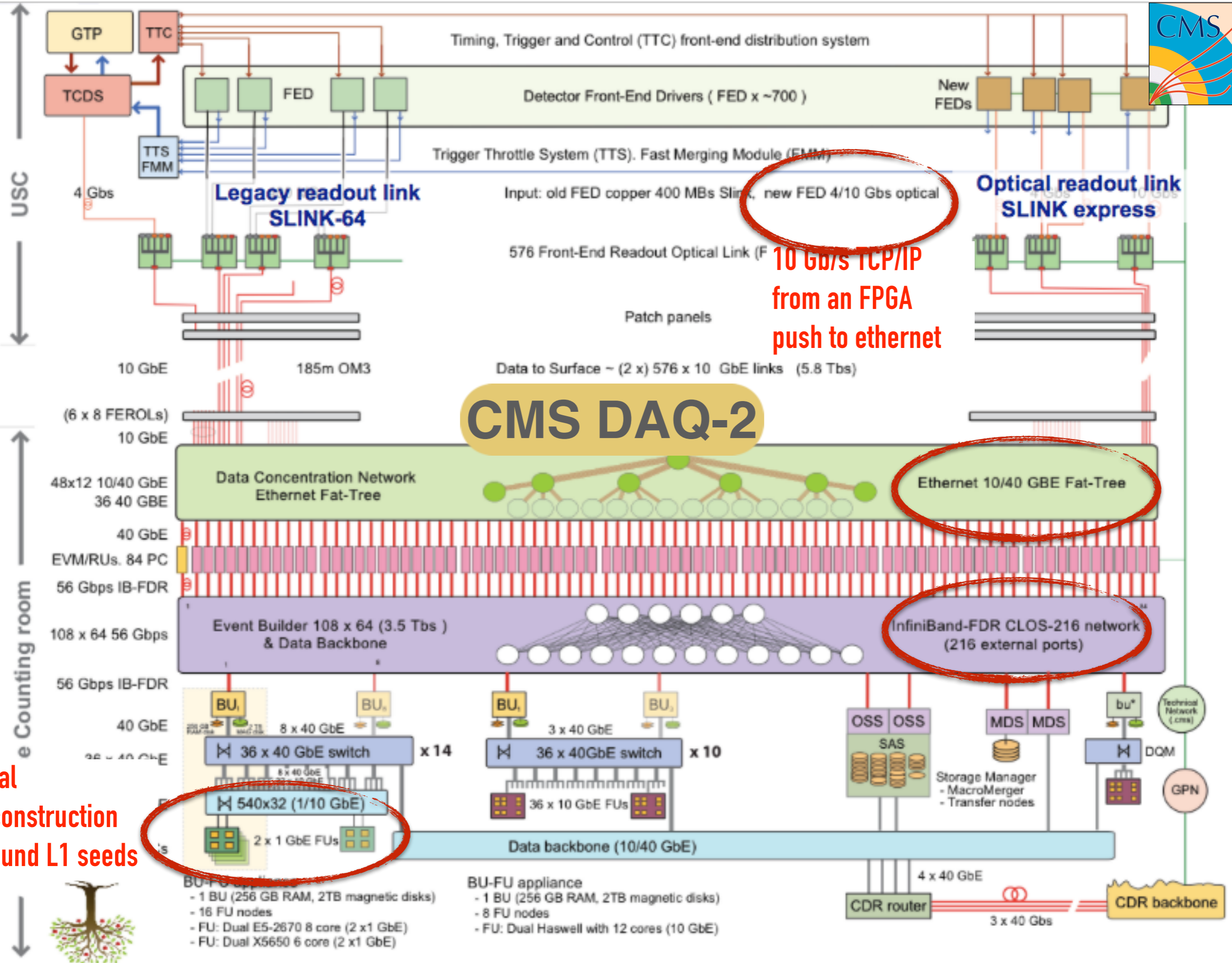
- ➔ low occupancy for fast pattern recognition
- ➔ optimal time-resolution for BC-identification

## ➔ L1 processing (40 MHz)

- ➔ pattern matching with patterns stored in buffers
- ➔ simplified fit of track segments

## ➔ High level processing (100 kHz)

- ➔ full detector resolutions
- ➔ match segments with tracks in the ID
- ➔ isolation



10 Gb/s TCP/IP from an FPGA push to ethernet

# CMS DAQ-2

local reconstruction around L1 seeds



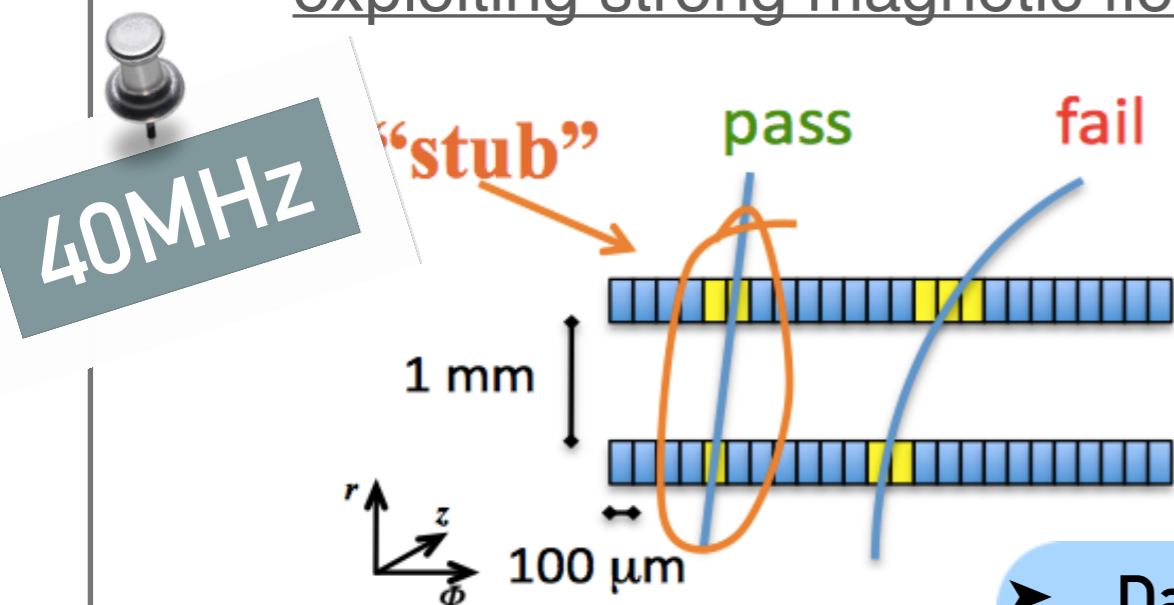




## Track filtering (low $p_T$ )

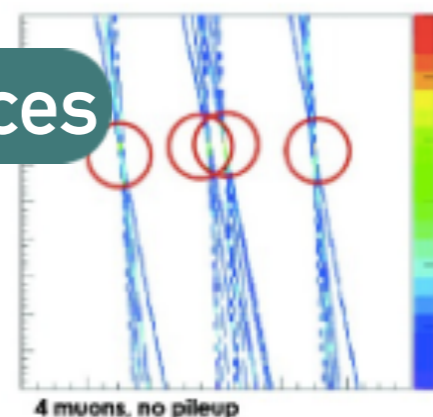
Reduce readout 40  $\rightarrow$  1 MHz by detector coincidences

- ➔ **Special outer tracker modules**
  - ➔ two layers of silicon at few mm
  - ➔ using cluster width and stacked trackers
- ➔ **Design tracker to have coherent  $p_T$  threshold in the full volume**
  - ➔ exploiting strong magnetic field of CMS

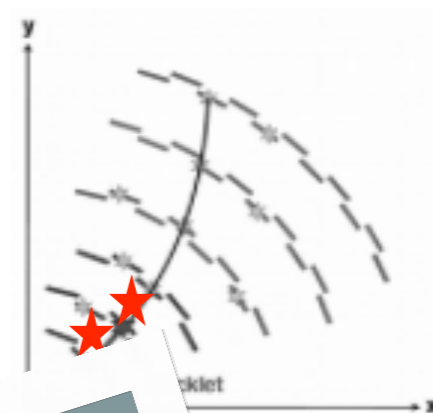


40MHz

## Track finding options

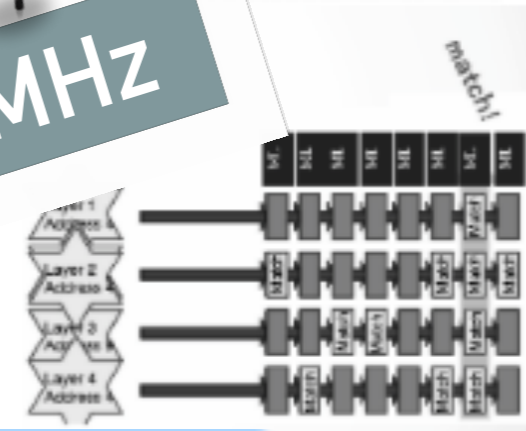


Hough Transform



Tracklets

1MHz

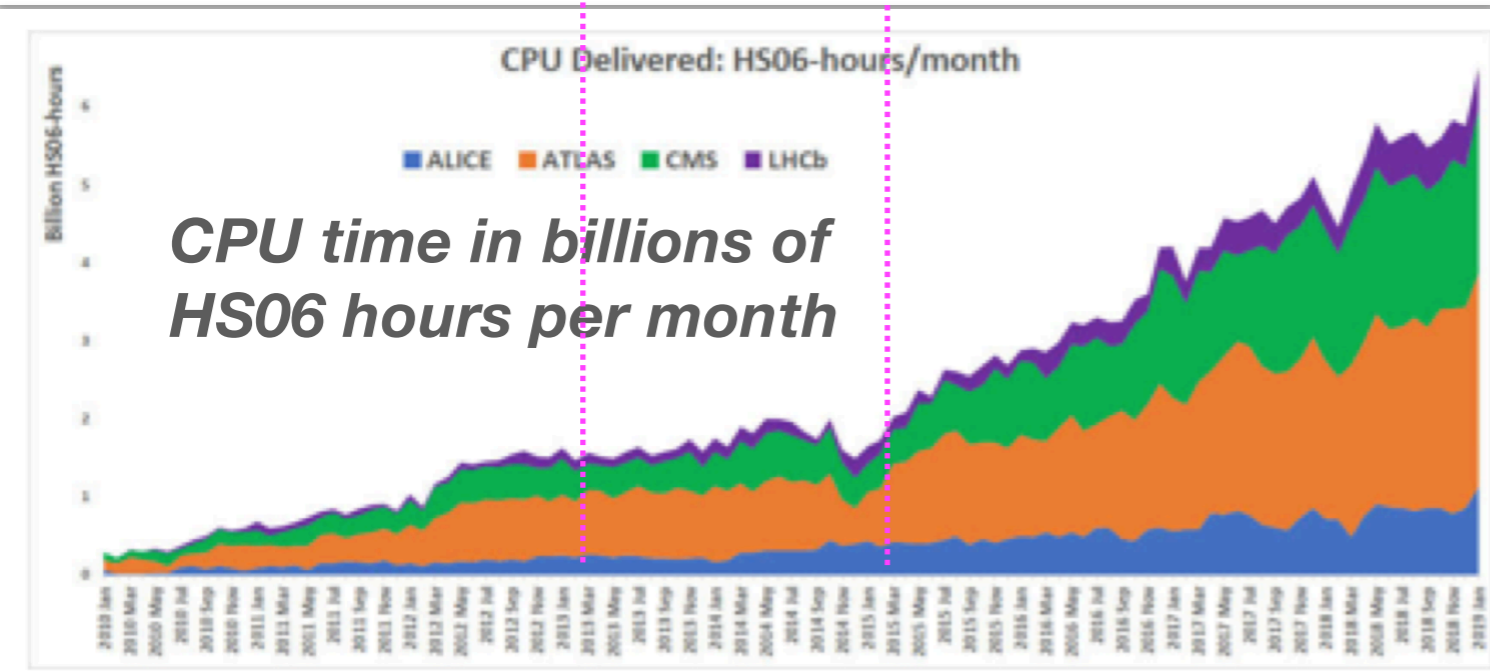
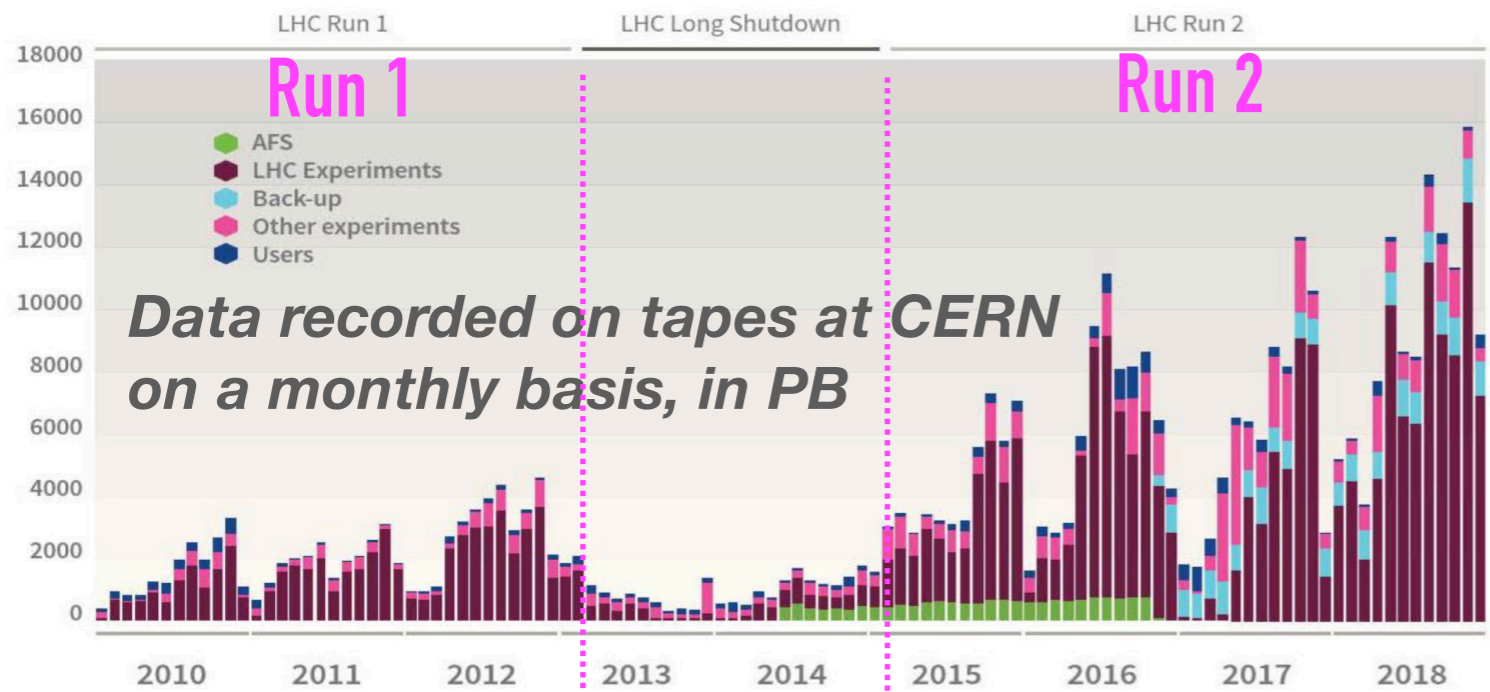


Associative Memories

- Data rates > 50-100 Tbps
- Latency: 4+1  $\mu$ s
- Three R&D efforts: FPGA/ASIC



# LHC COMPUTING TOWARDS NEW PARADIGMS



## Run1 + Run2

- **Data storage**
  - 339 PB on tapes, 173 PB on disks
- **Global CPU time delivered by Worldwide LHC Computing Grid (WLCG)**
  - about 900,000 cores

## Run 3

- **Evolution of current technologies and current (flat) funding is ok**

## Run 4

- **Linear increase of digitisation time**
- **Factorial increase of reconstruction time**
- **Larger events, lots of more memory**



see [Ref]

→ **Need factor 2-3 more storage and computing resources for HL-LHC**

→ new developments and R&D projects for data management and processing, SW multithreading, new computing models and data compression