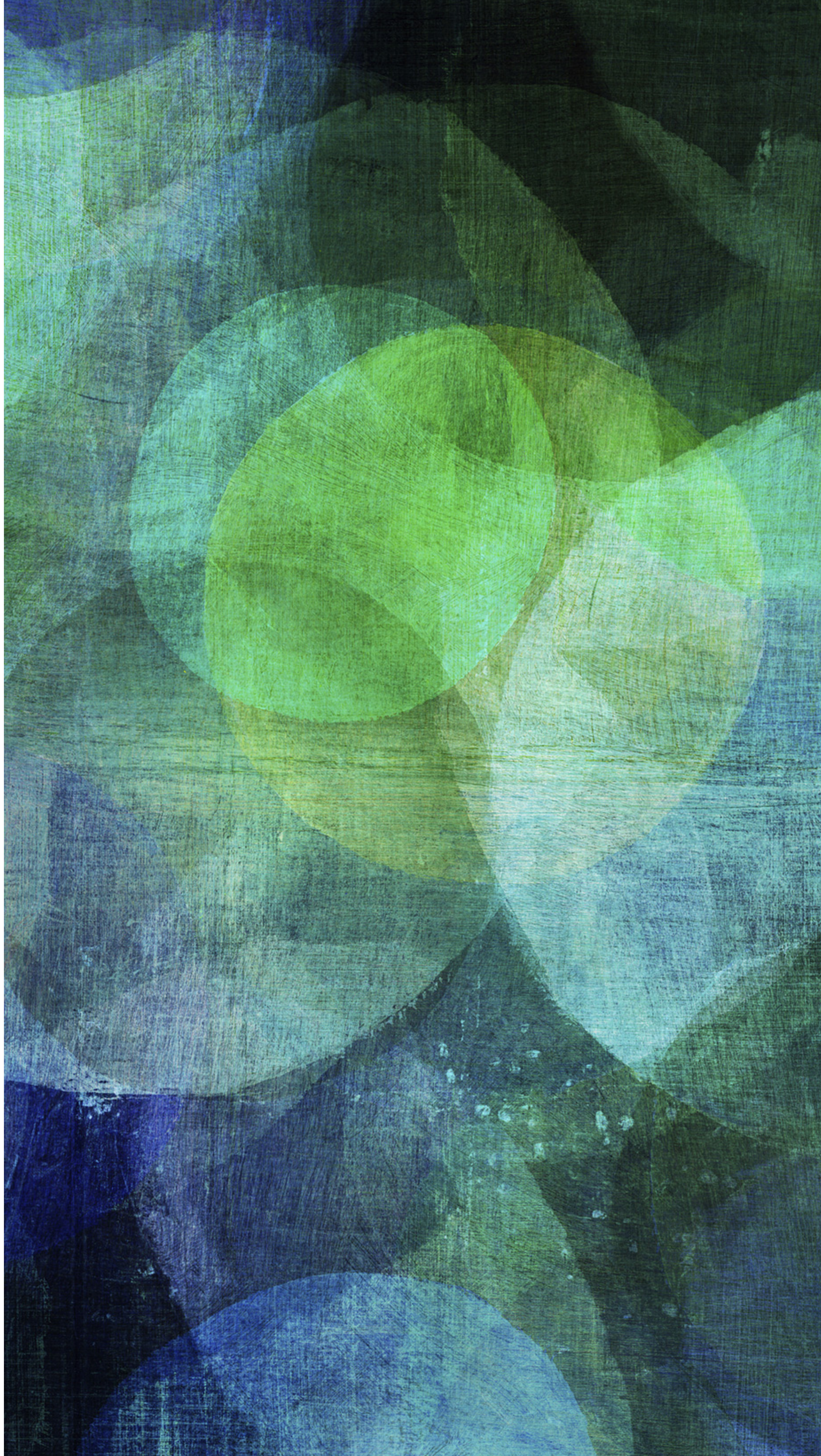
STRATEGIES AND FUTURE TRENDS FOR TRIGGER AND DAQ SYSTEMS IN LHC EXPERIMENTS

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

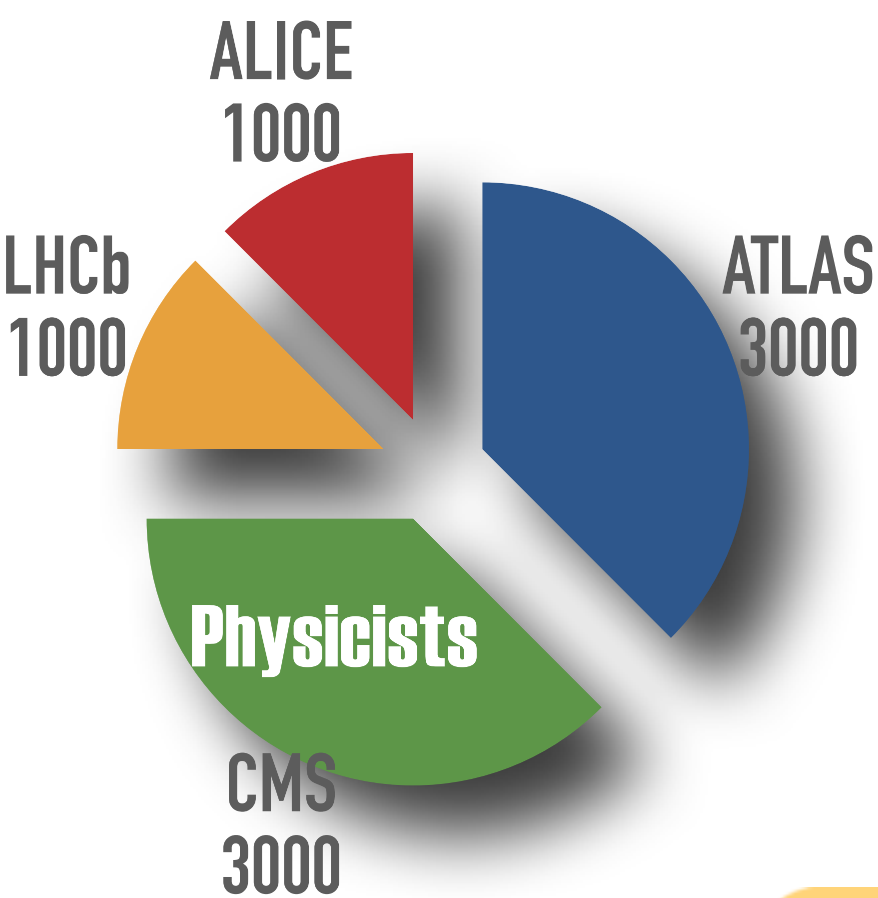
TRIGGERING AND TAKING DATA AT LHC

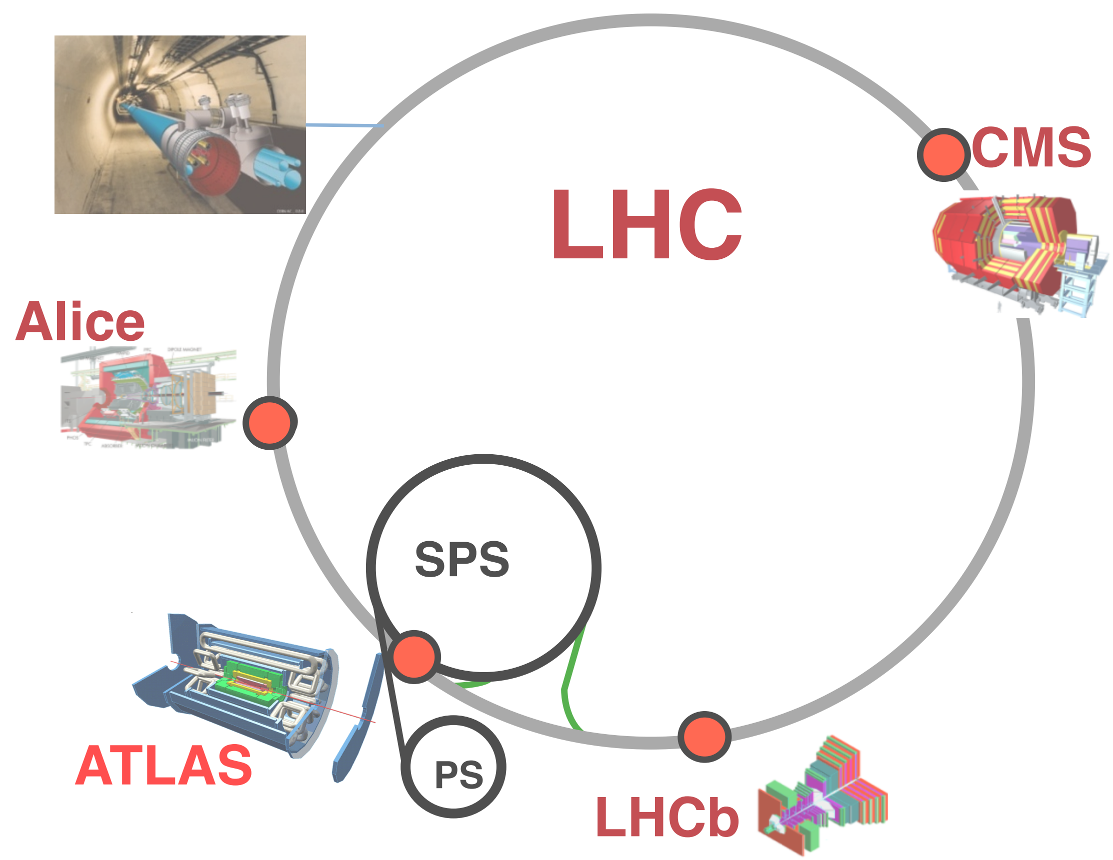
TDAQ for large discovery experiments



LHC EXPERIMENTS FOR A DISCOVERY MACHINE

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

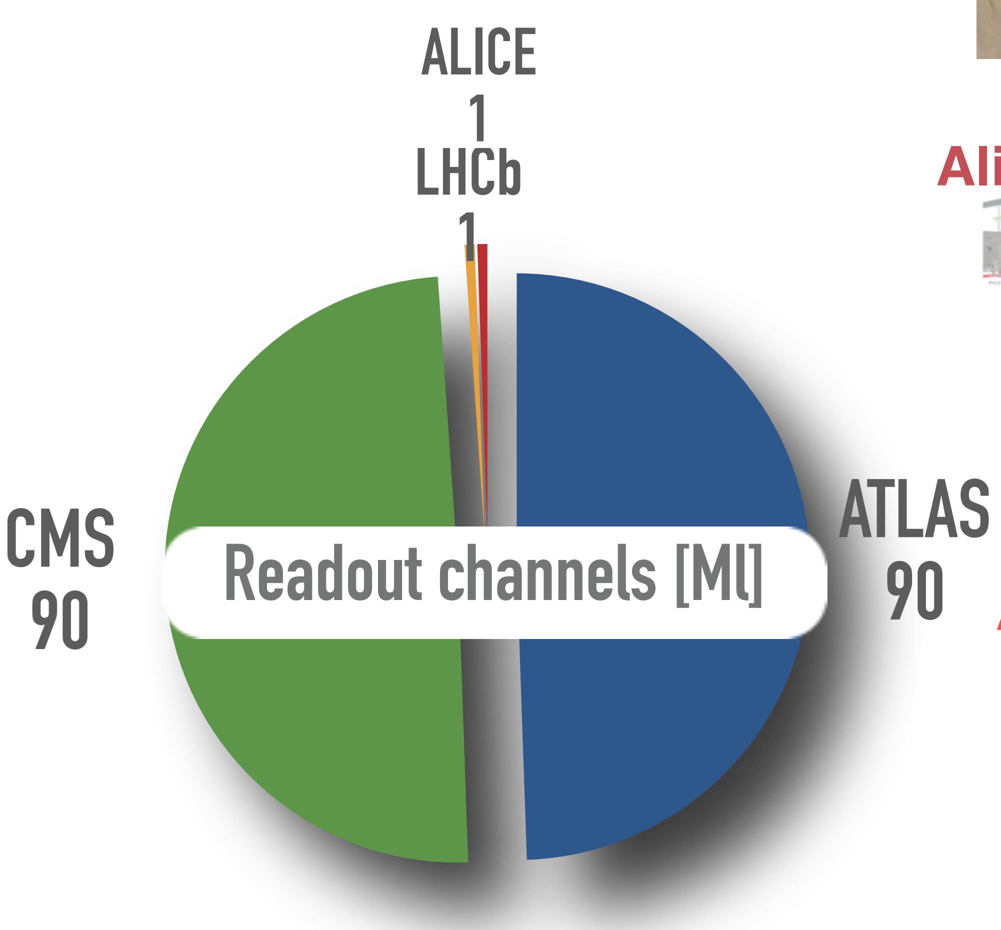


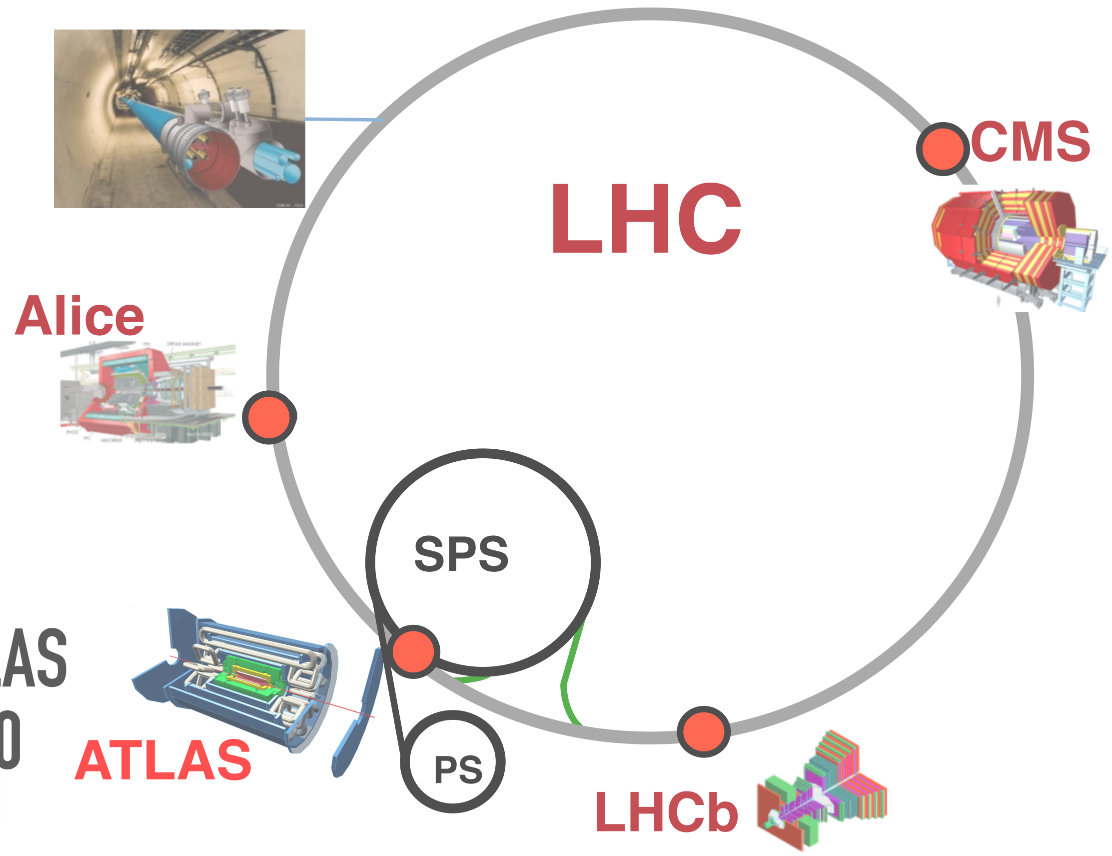


Proposed: 1992, Approved: 1996, Started: 2009

LHC EXPERIMENTS FOR A DISCOVERY MACHINE

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





Proposed: 1992, Approved: 1996, Started: 2009

LHC EXPERIMENTS FOR A DISCOVERY MACHINE

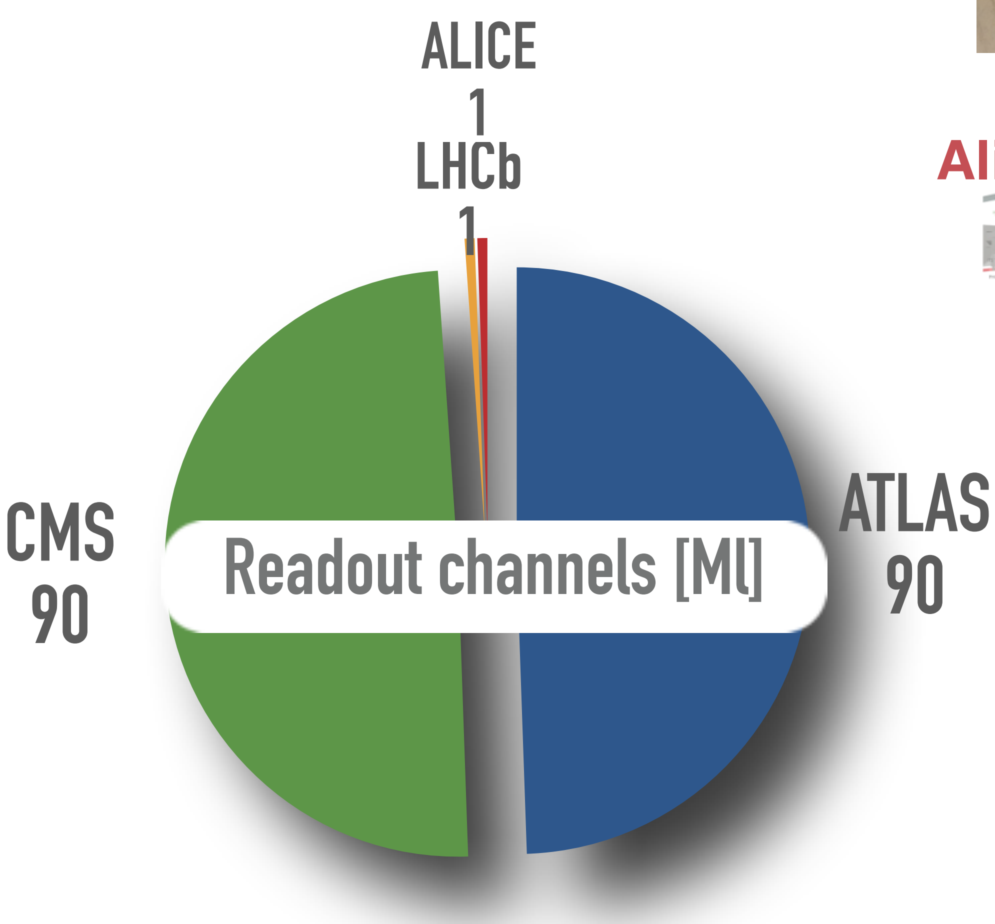
Goal: explore TeV energy scale to find New Physics beyond Standard Model

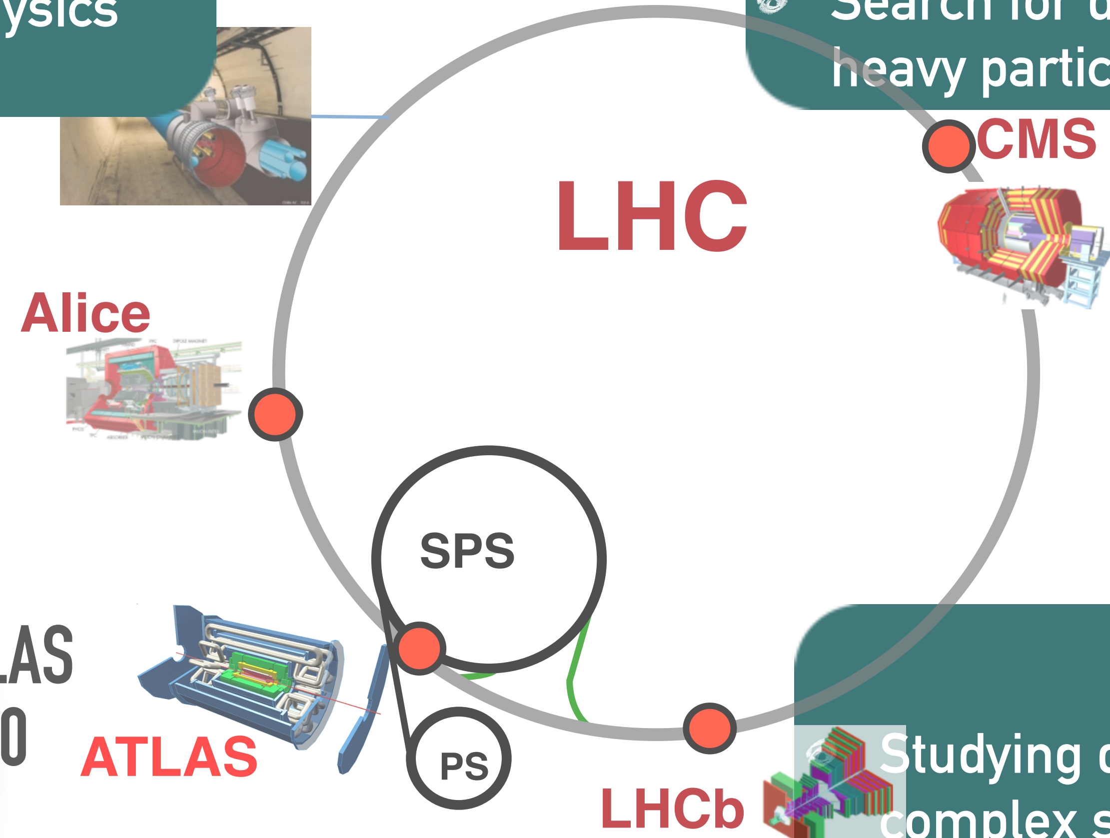
ATLAS & CMS

- 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





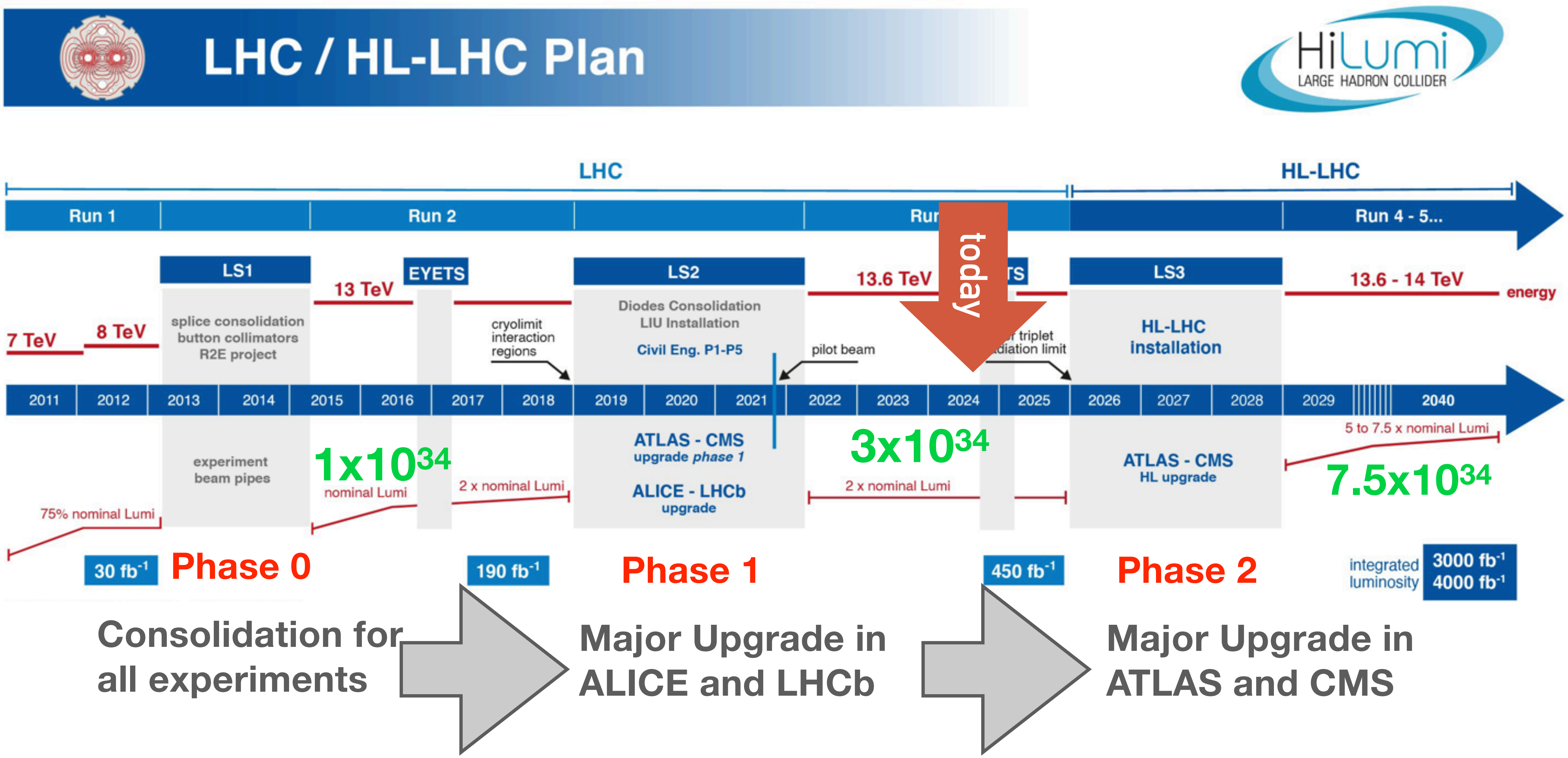
ALICE

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

Proposed: 1992, Approved: 1996, Started: 2009

LHC BECOMING IMPRESSIVELY LUMINOUS

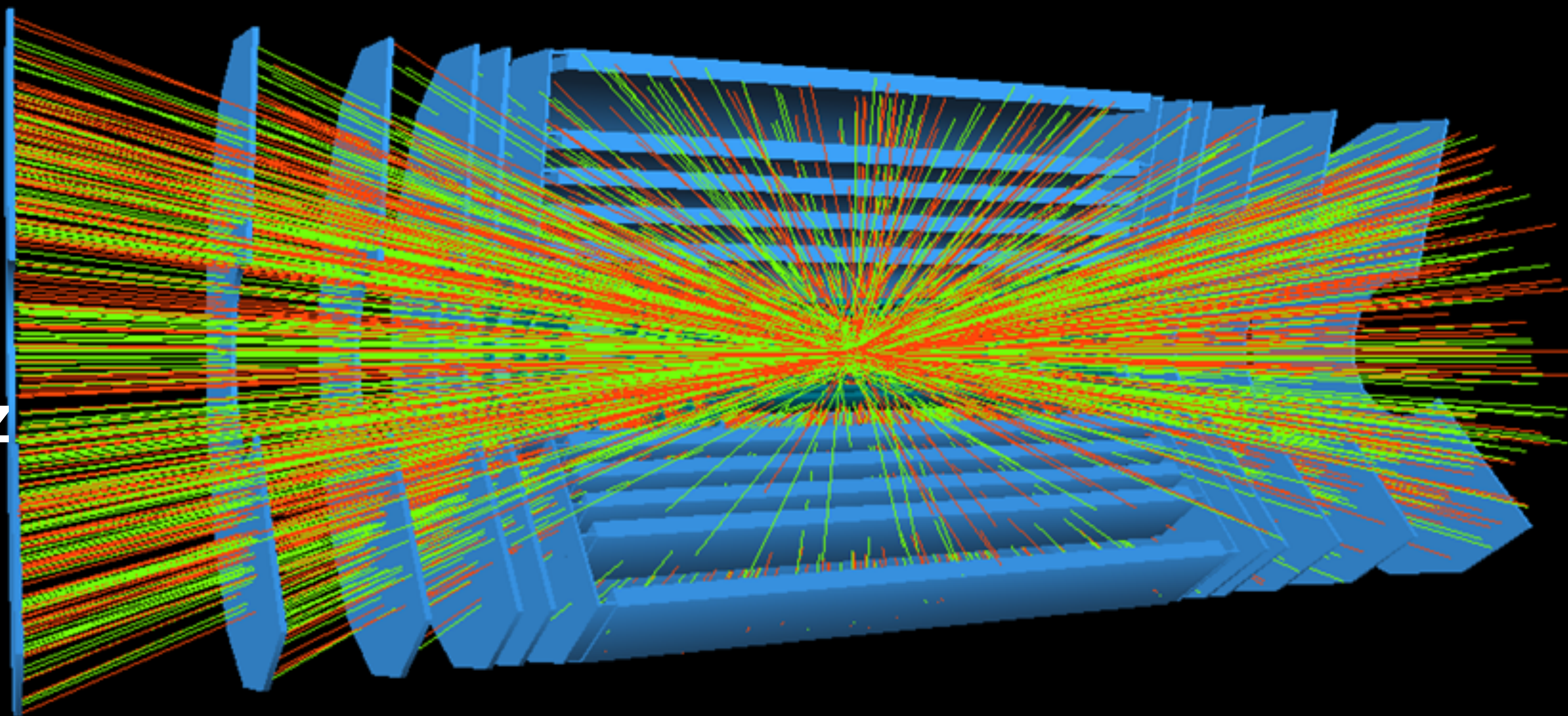
European Council (2014): "CERN is the strong European focal point for particle physics in next 20 years"



➡ Experiments upgraded as the luminosity increases, to improve or at least maintain the design performance

LHC DATA DELUGE

p-p collisions
E_{cms} = 13-14 TeV
L = 10^{34} /cm^2 s
BC clock = 40 MHz



- High Luminosity with collisions close in time and space (1 collision/25ns)
  - fast electronics ➠ fast decisions
  - fine granularity detector ➠ high data volume
- 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!

MANY PLAYERS, COMPLEX TDAQ ARCHITECTURES

Buffering and parallelism

Maximum 1-2% deadtime

40 MHz
COLLISION RATE

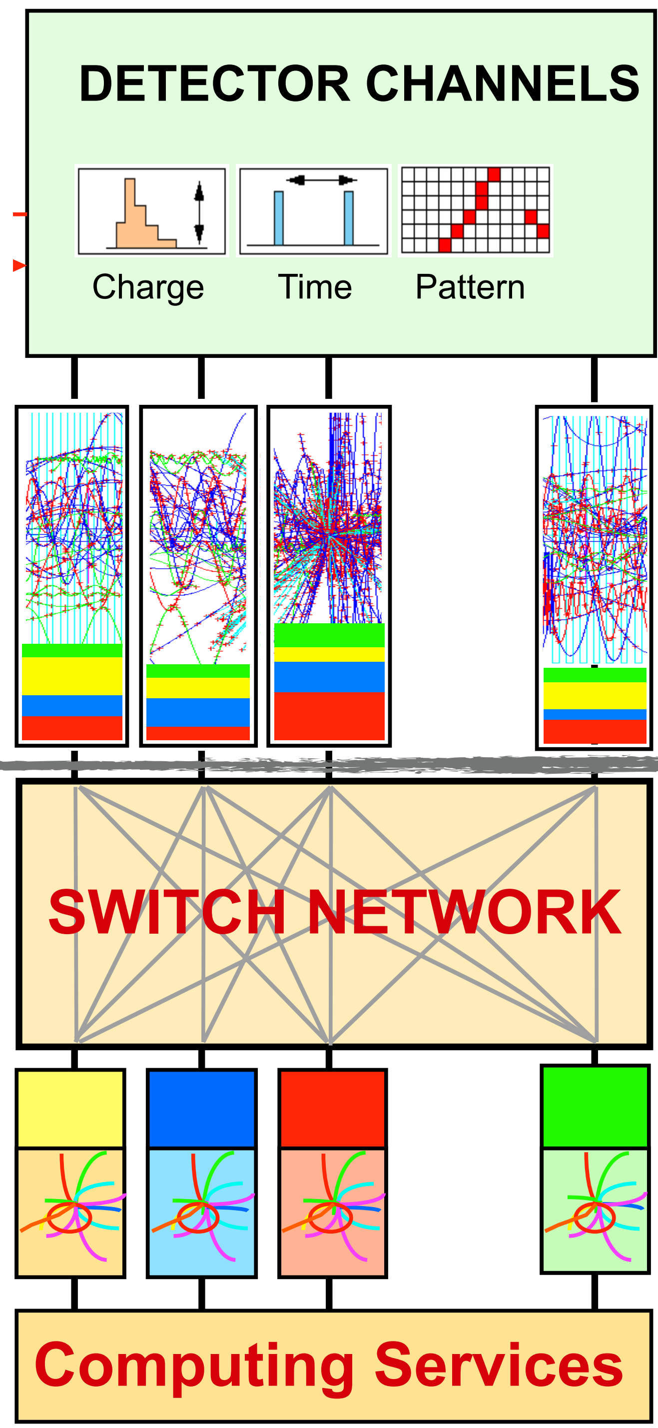
Level-1

Readout Buffers

Event building

Event filtering

Petabyte archive



High speed electronics

Readout links and buffering

Large data network with dedicated technology

Dedicated PC farms

Level-1 triggers

- Set max Readout rate
- Hardware, synchronous
- Readout parallelism
- Latency ~ µsec/event

Readout

L1

DAQ

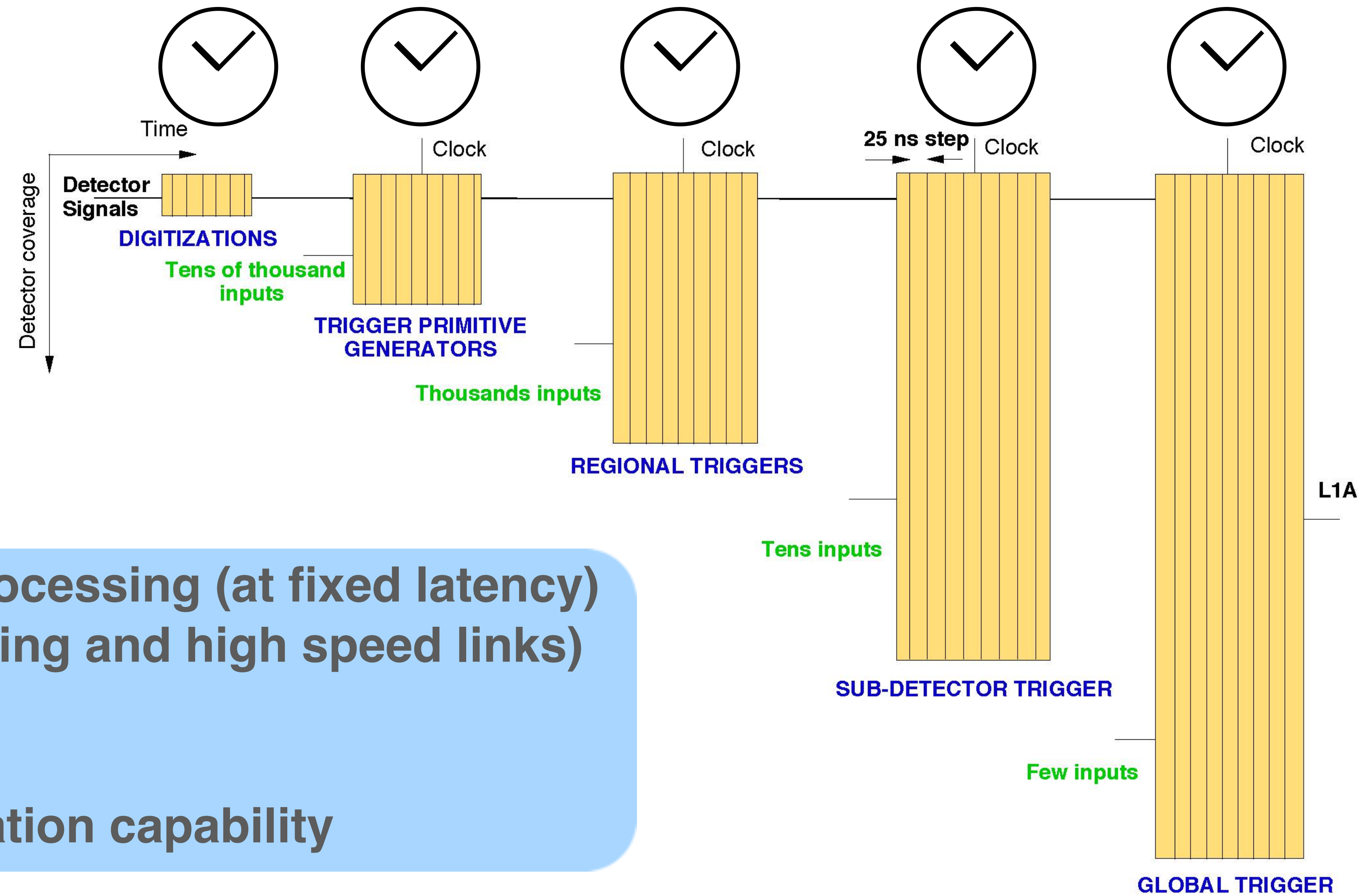
HLT

Higher level triggers

- Set max storage rate
- Software, asynchronous
- Event parallelism
- Latency < 1 sec/event

LEVEL-1 TRIGGER PRINCIPLES

L1



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

Full synchronisation at 40 MHz (LHC clock)

➤ large optical time distribution system

Fast, robust electronics

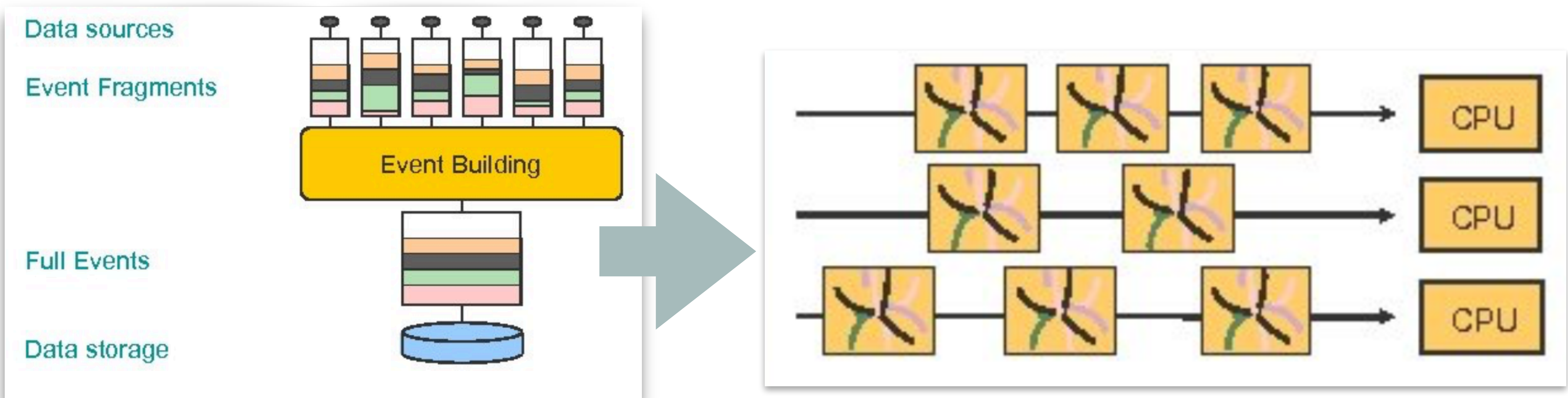
Latency dominated by cable/transmission delay

ALICE	No pipeline
ATLAS	2.5 µs
CMS	3 µs
LHCb	4 µs

HLT/DAQ REQUIREMENTS

HLT

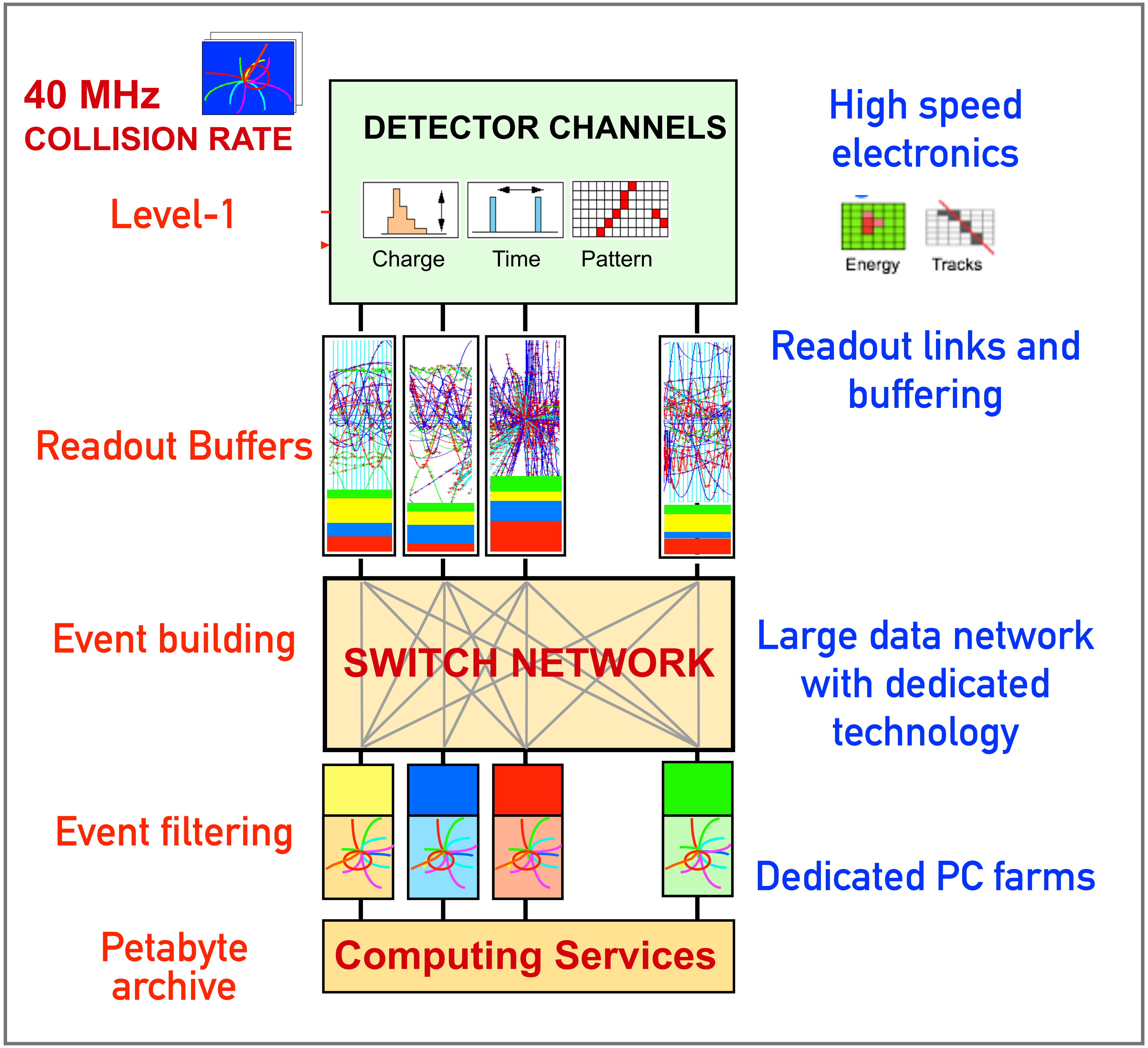
- ➡ **Robustness and redundancy**
- ➡ **Scalability to adapt to Luminosity, detectors,…**
- ➡ **Flexibility (10-years experiments)**
- ➡ **Based on commercial products**
- ➡ **Limited cost**



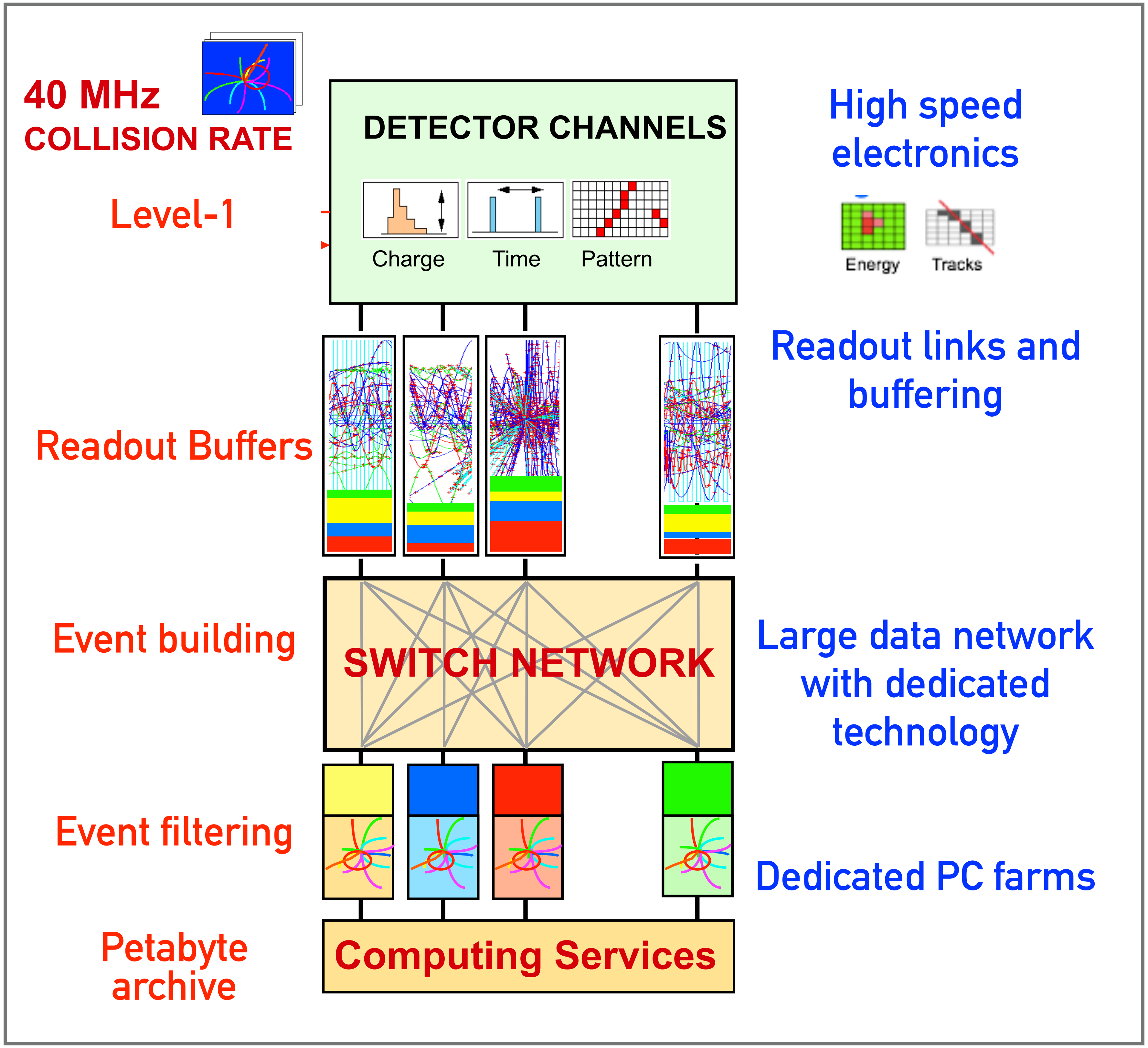
- ➤ Event Building and Event Filter farms on networks
  - ➤ farm processing: one event per processor (larger latency, but scalable)
  - ➤ additional networks regulates the CPU assignment

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

EXAMPLE: SOME NUMBERS FOR ATLAS AND CMS

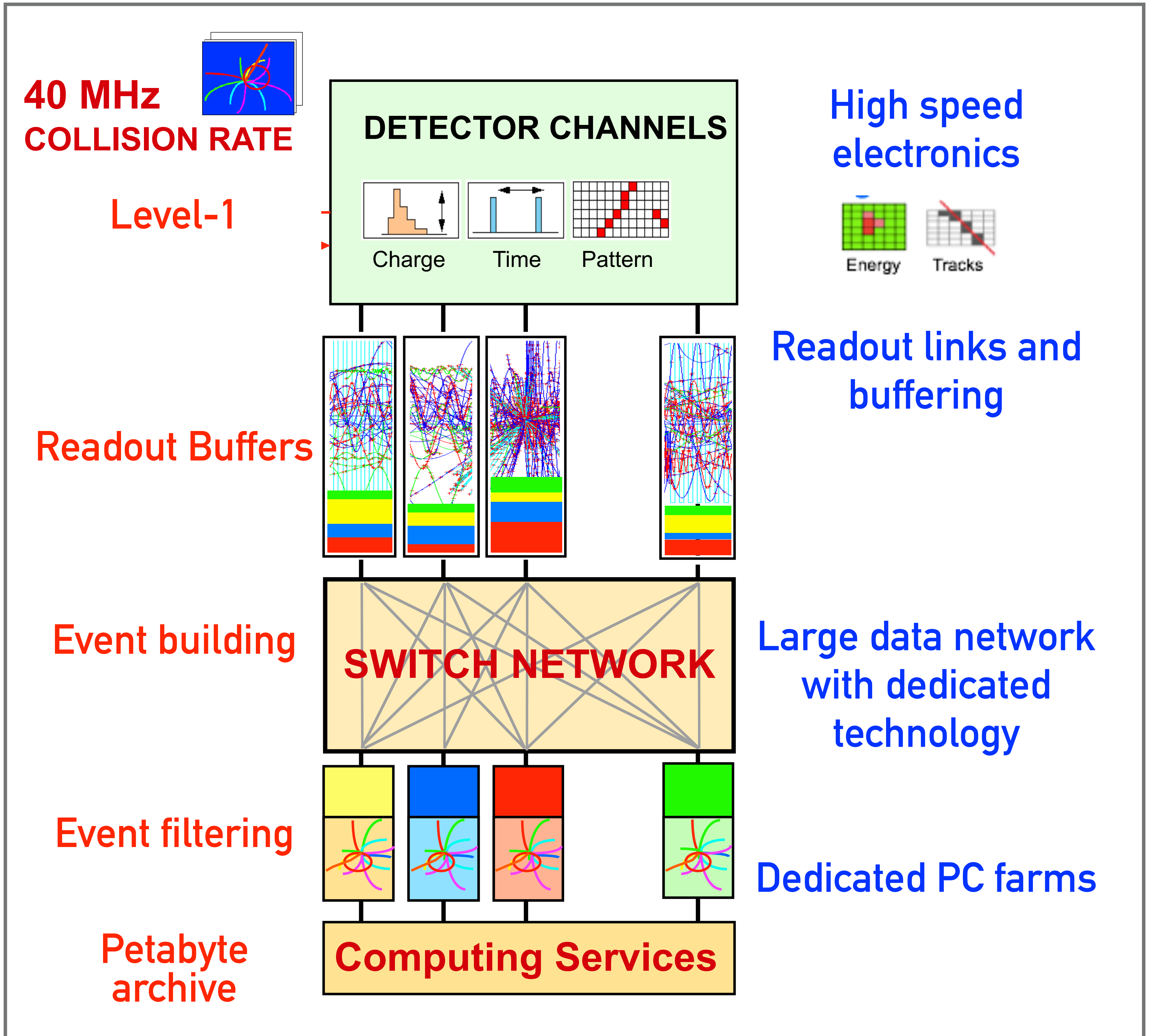


EXAMPLE: SOME NUMBERS FOR ATLAS AND CMS



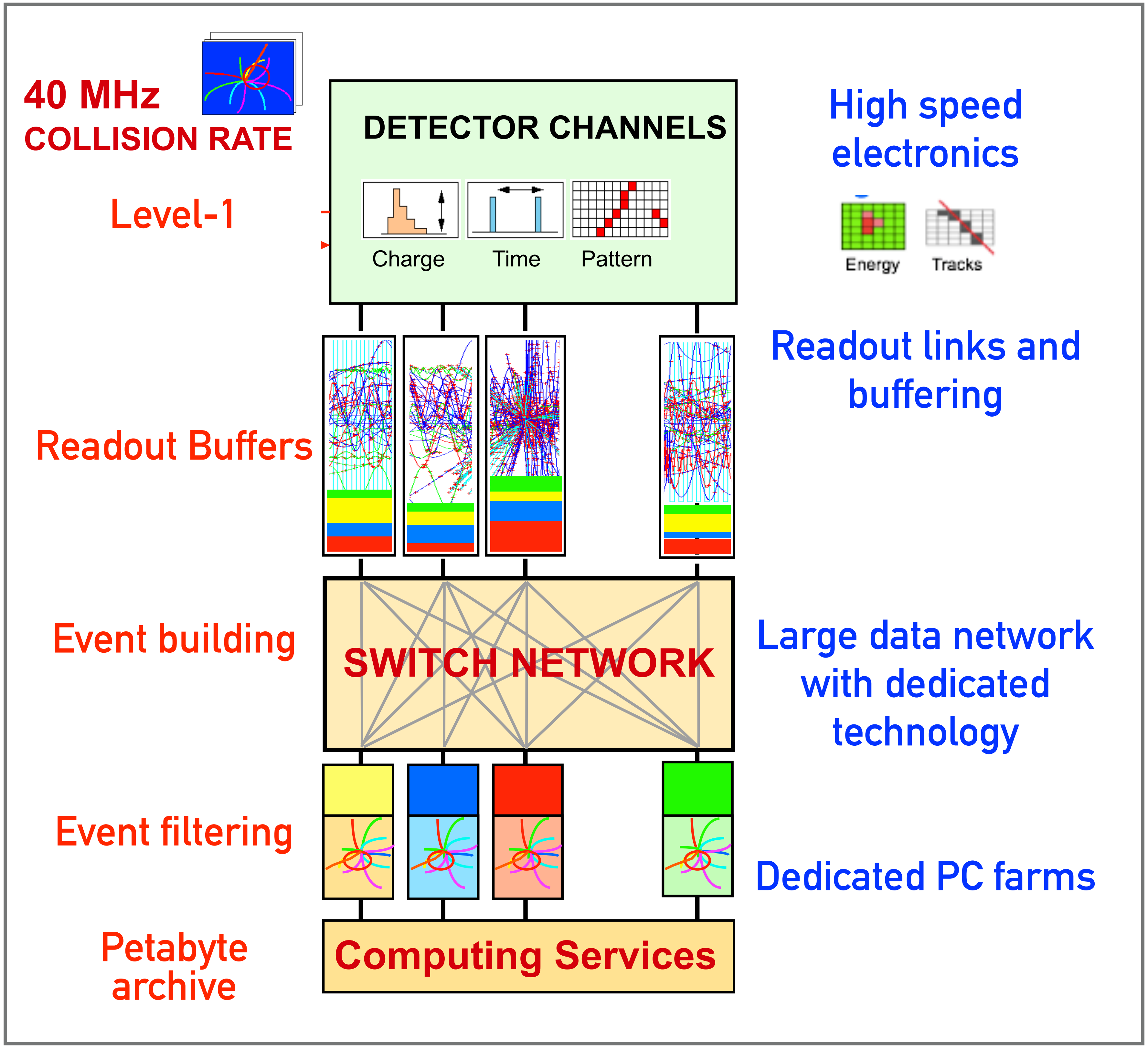
➡ Event size = 1 MB

EXAMPLE: SOME NUMBERS FOR ATLAS AND CMS



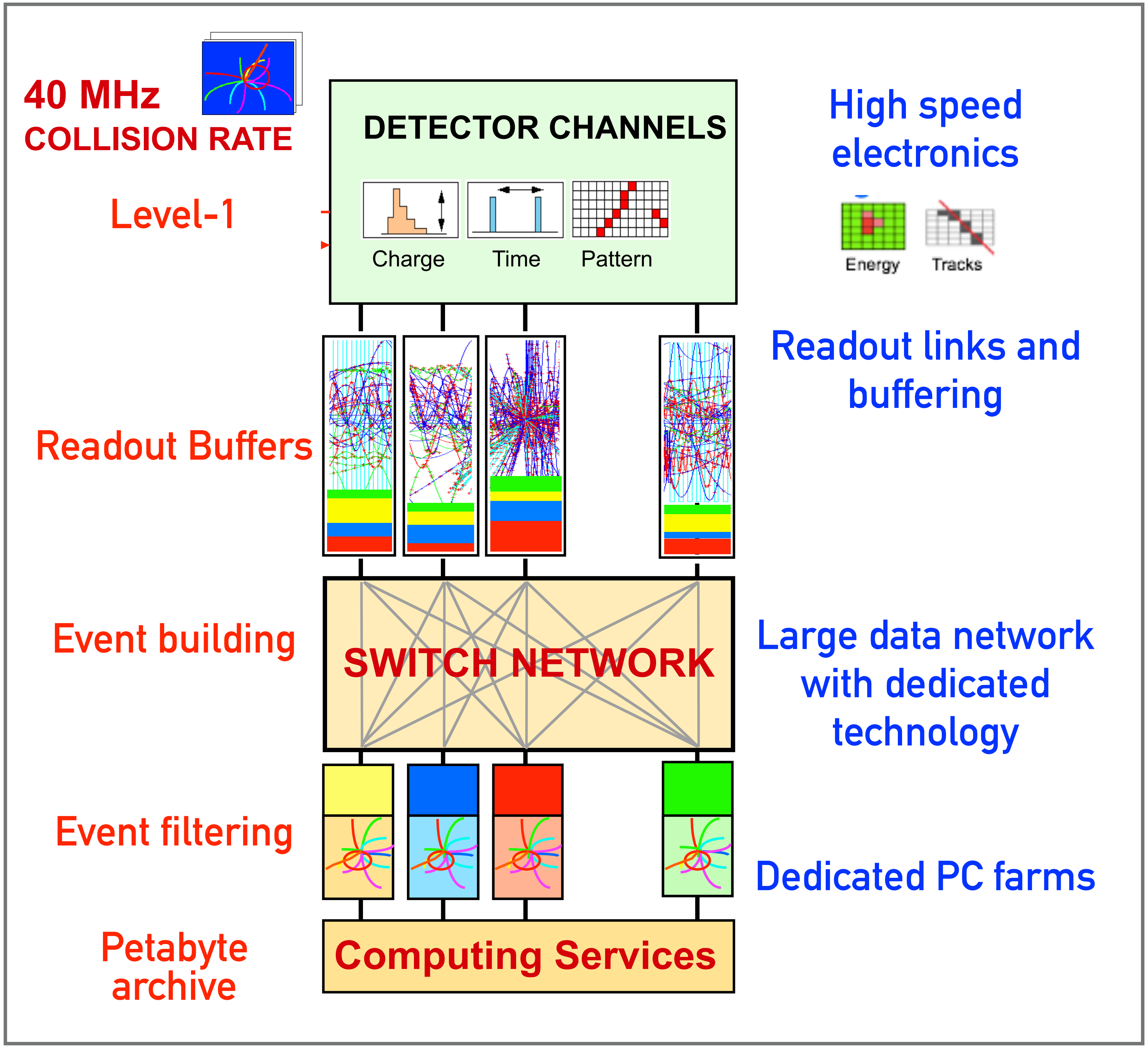
- ➡ Event size = 1 MB
- ➡ L1 rate = 100 kHz

EXAMPLE: SOME NUMBERS FOR ATLAS AND CMS



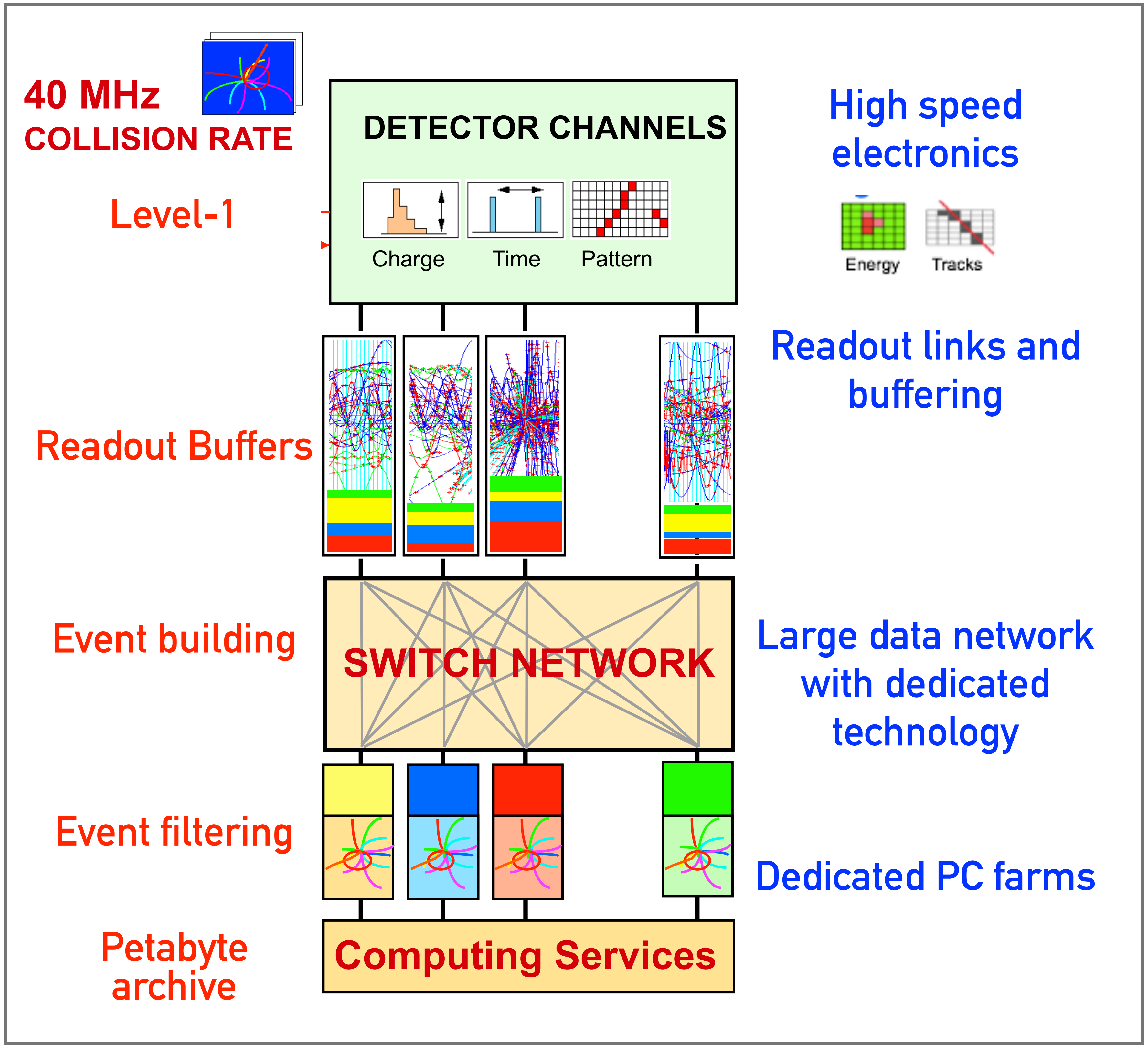
- ➡ Event size = 1 MB
- ➡ L1 rate = 100 kHz
- ➡ DAQ network size = ?

EXAMPLE: SOME NUMBERS FOR ATLAS AND CMS



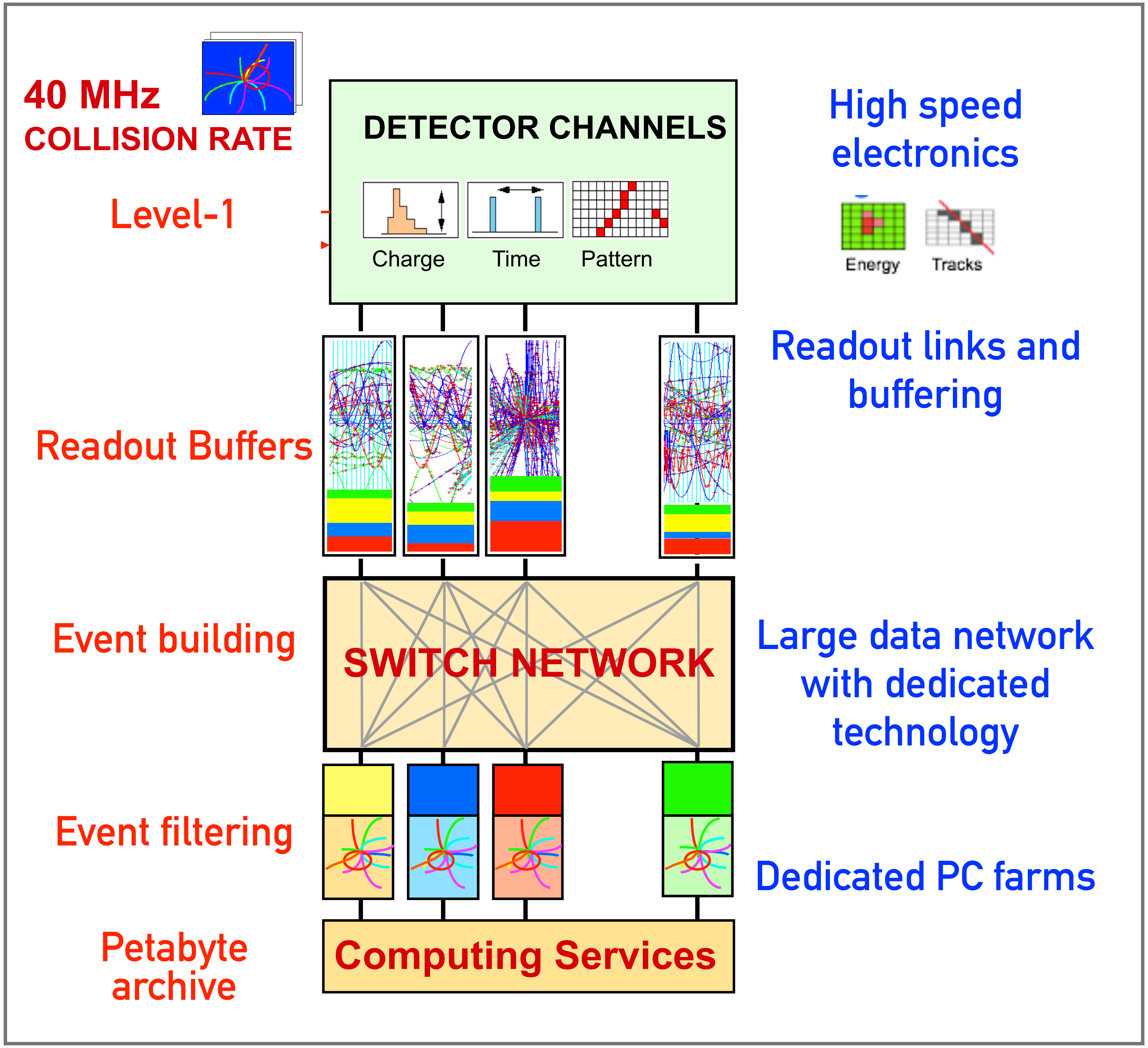
- ➡ Event size = 1 MB
- ➡ L1 rate = 100 kHz
- ➡ DAQ network size = ?
  - ➡ 1 MB x 100 kHz =

EXAMPLE: SOME NUMBERS FOR ATLAS AND CMS



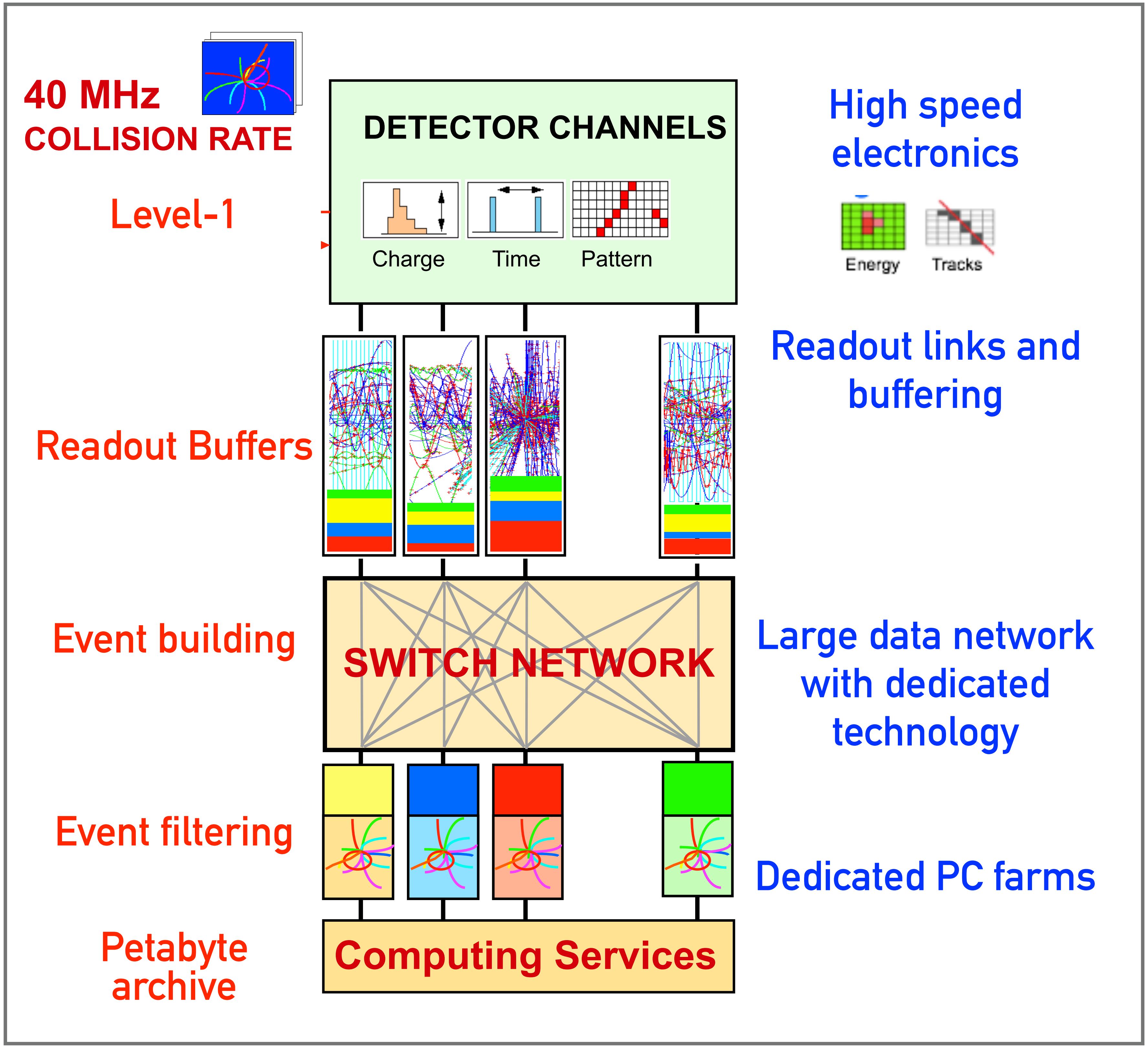
- Event size = 1 MB
- L1 rate = 100 kHz
- DAQ network size = ?
  - 1 MB x 100 kHz =
  - 100 GB/s (~Tbps)

EXAMPLE: SOME NUMBERS FOR ATLAS AND CMS



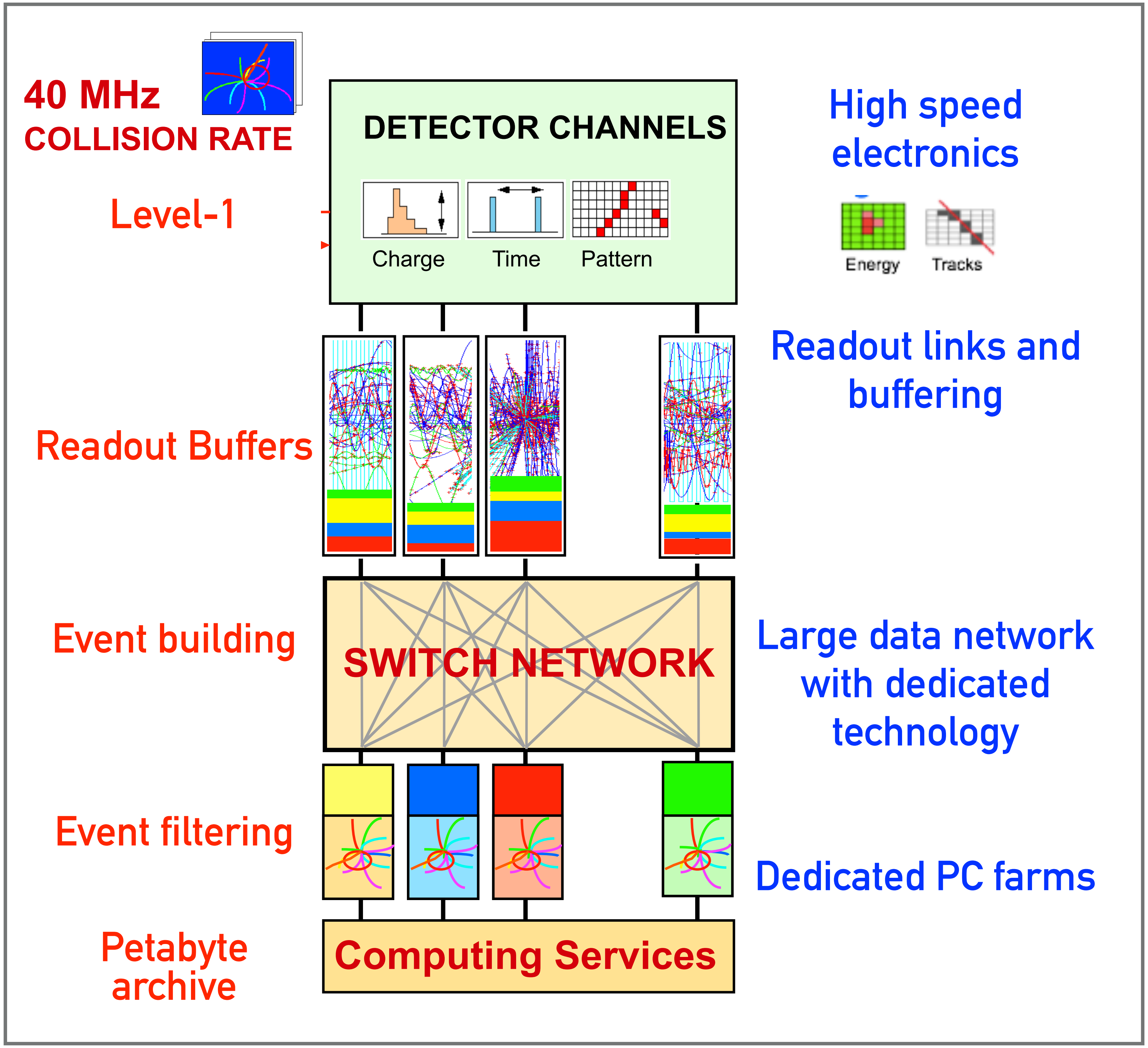
- Event size = 1 MB
- L1 rate = 100 kHz
- DAQ network size = ?
  - 1 MB x 100 kHz =
  - 100 GB/s (~Tbps)
- L1 latency = 2.5 μs

EXAMPLE: SOME NUMBERS FOR ATLAS AND CMS



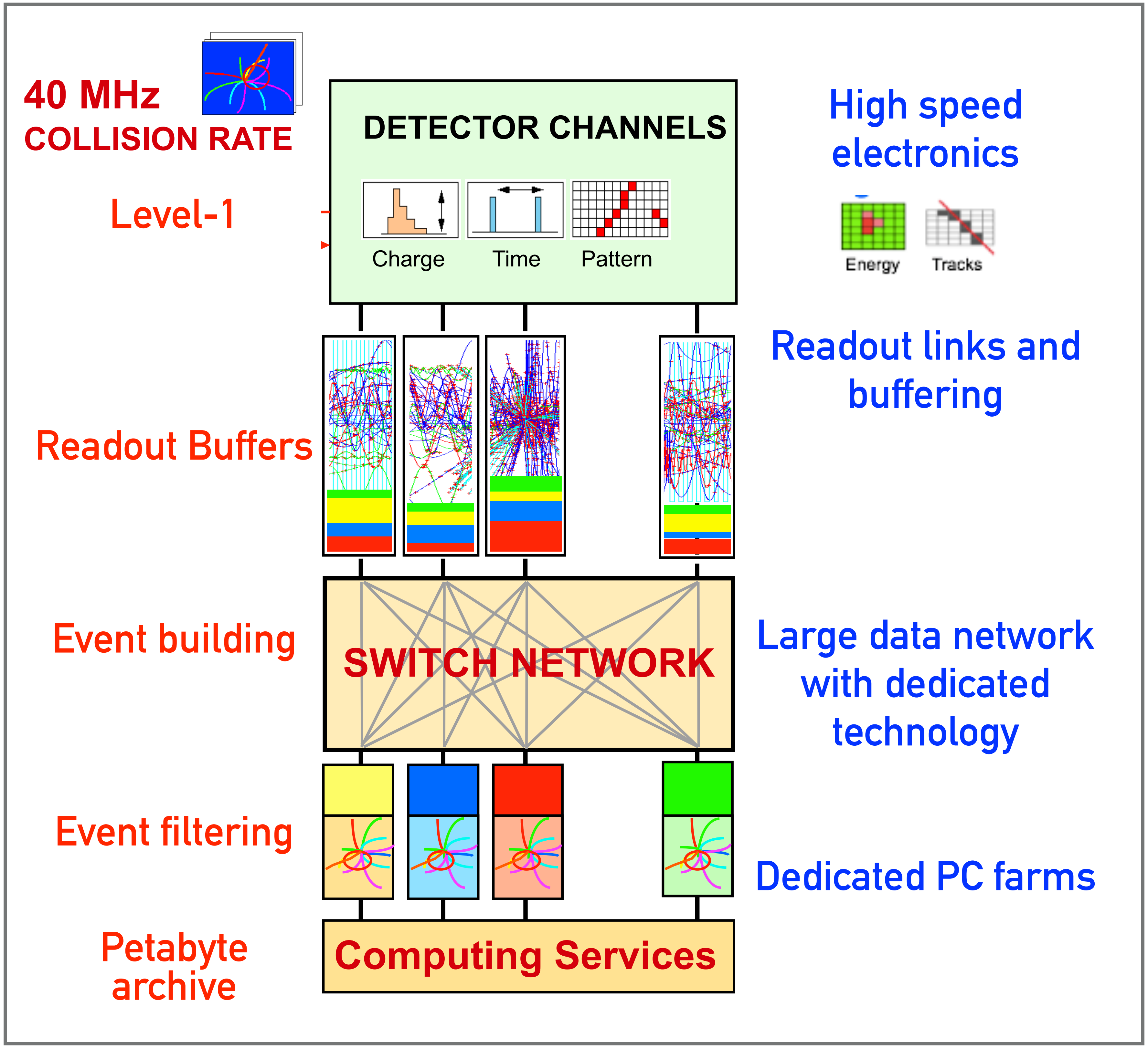
- Event size = 1 MB
- L1 rate = 100 kHz
- DAQ network size = ?
  - 1 MB x 100 kHz =
  - 100 GB/s (~Tbps)
- L1 latency = 2.5 μs
- L1 buffer size = ?

EXAMPLE: SOME NUMBERS FOR ATLAS AND CMS



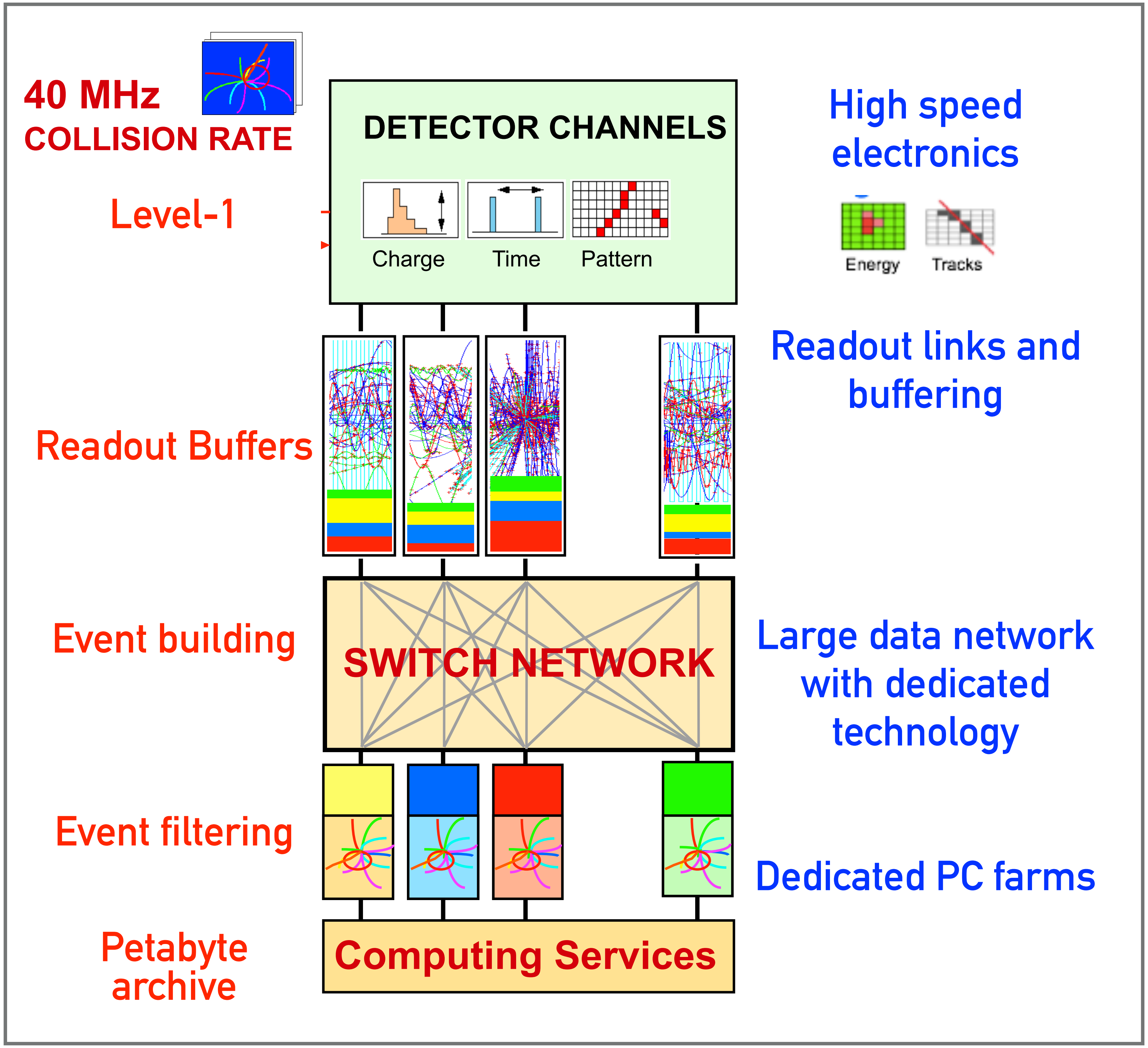
- Event size = 1 MB
- L1 rate = 100 kHz
- DAQ network size = ?
  - 1 MB x 100 kHz =
  - 100 GB/s (~Tbps)
- L1 latency = 2.5 μs
- L1 buffer size = ?
  - 2.5 μs / 25 ns = 100

EXAMPLE: SOME NUMBERS FOR ATLAS AND CMS



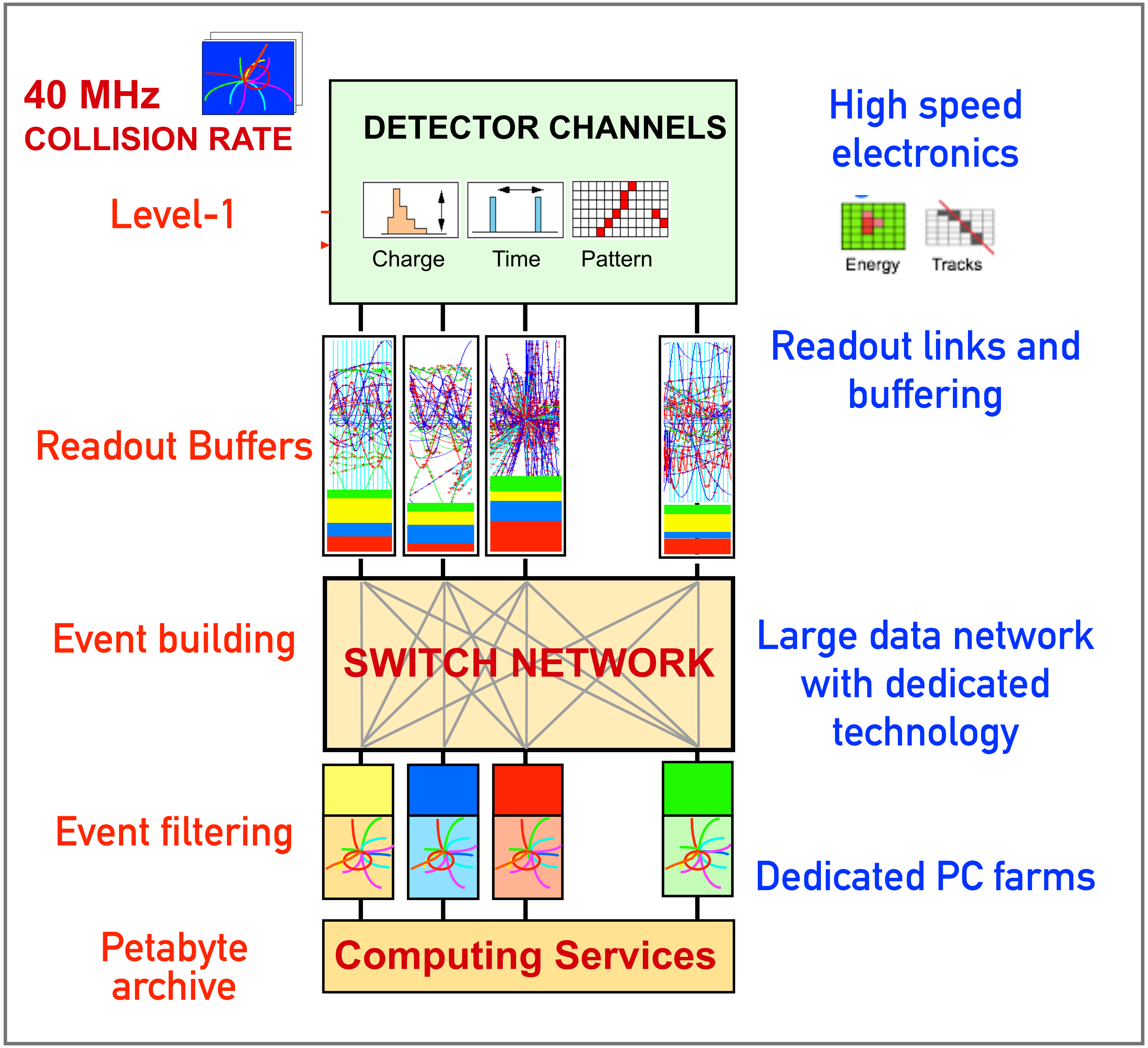
- Event size = 1 MB
- L1 rate = 100 kHz
- DAQ network size = ?
  - 1 MB x 100 kHz =
  - 100 GB/s (~Tbps)
- L1 latency = 2.5 μs
- L1 buffer size = ?
  - 2.5 μs / 25 ns = 100
  - 100 events

EXAMPLE: SOME NUMBERS FOR ATLAS AND CMS



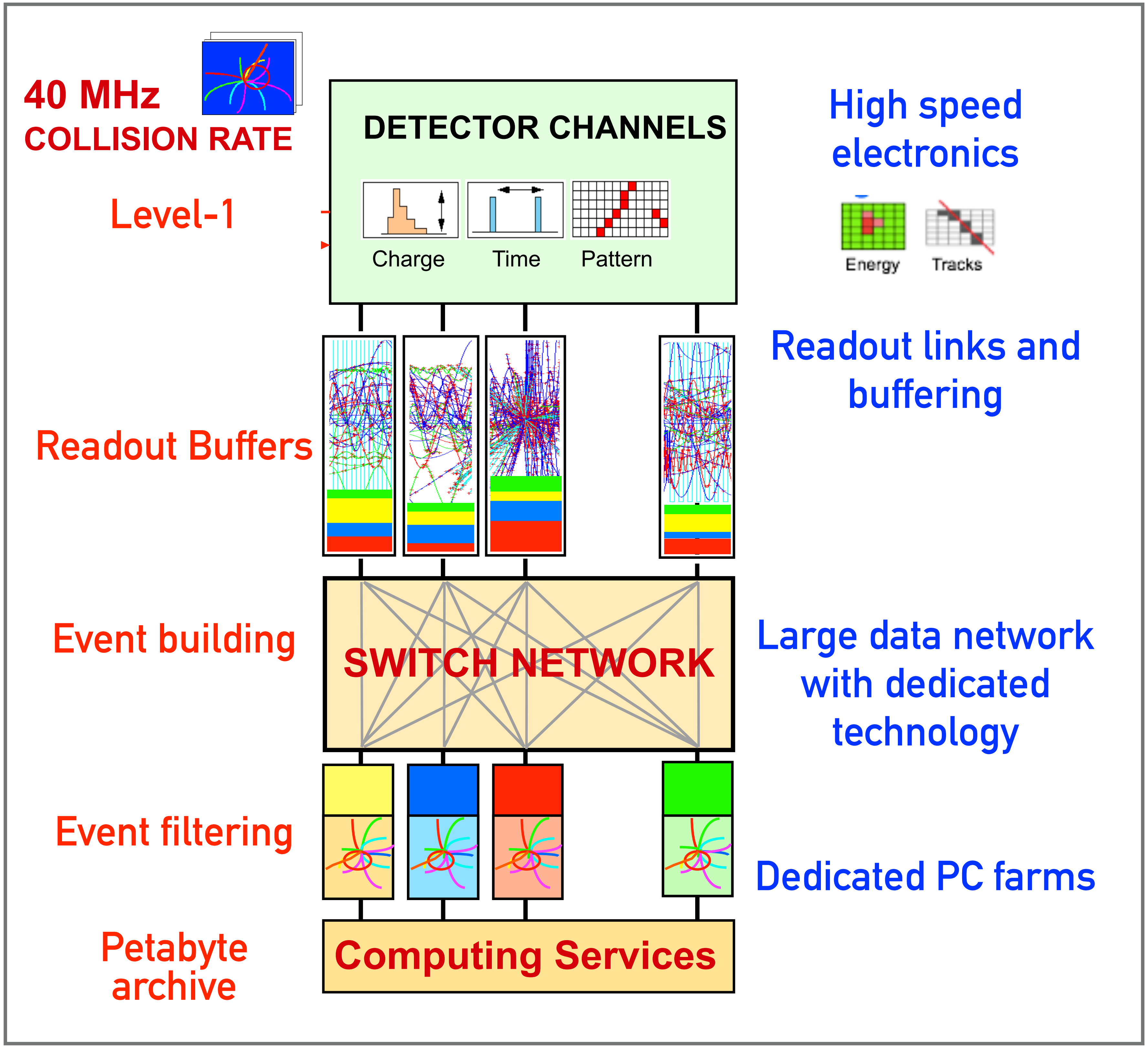
- Event size = 1 MB
- L1 rate = 100 kHz
- DAQ network size = ?
  - 1 MB x 100 kHz =
  - 100 GB/s (~Tbps)
- L1 latency = 2.5 μs
- L1 buffer size = ?
  - 2.5 μs / 25 ns = 100
  - 100 events
- HLT latency ~ 100 ms

EXAMPLE: SOME NUMBERS FOR ATLAS AND CMS



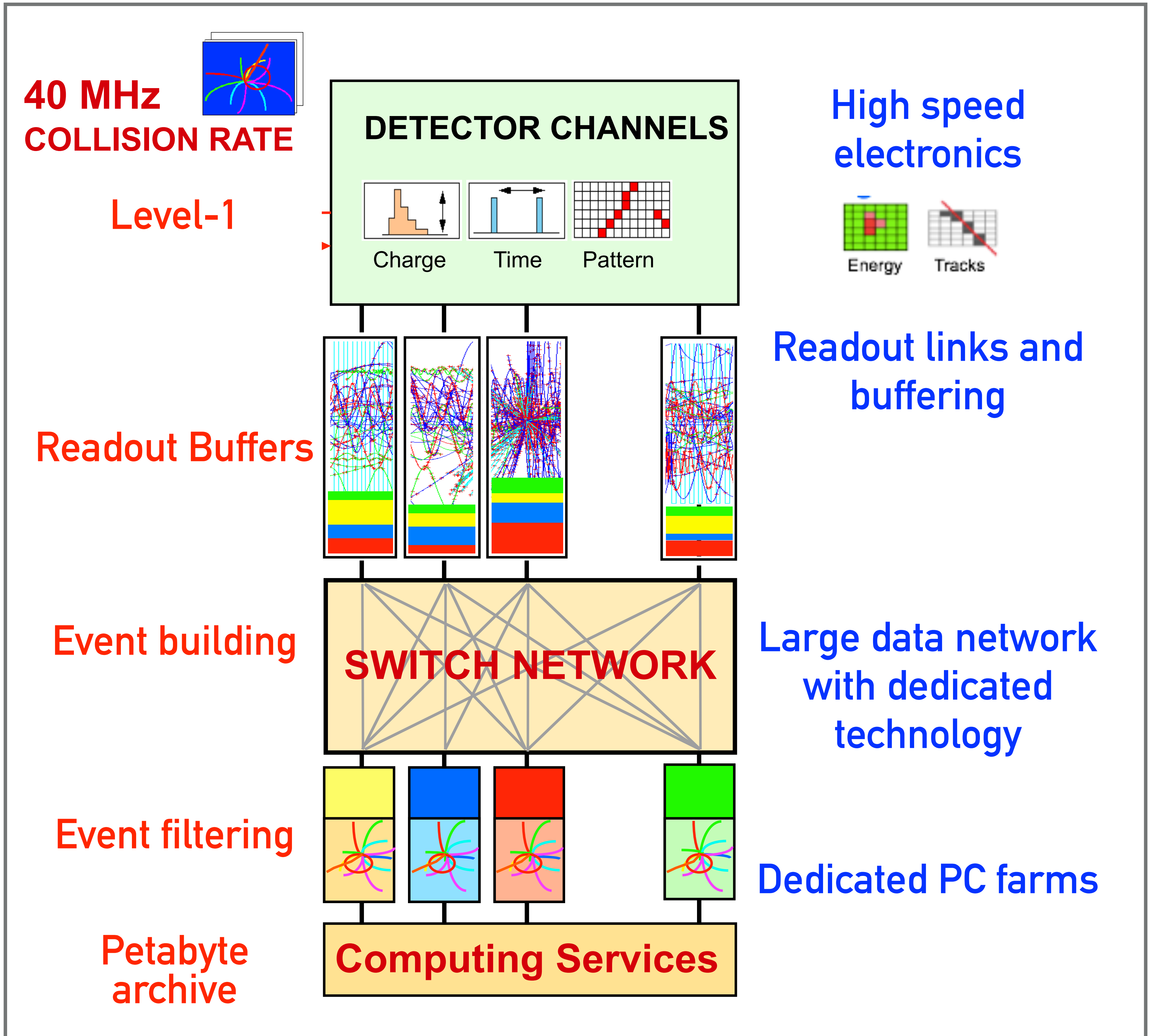
- Event size = 1 MB
- L1 rate = 100 kHz
- DAQ network size = ?
  - 1 MB x 100 kHz =
  - 100 GB/s (~Tbps)
- L1 latency = 2.5 μs
- L1 buffer size = ?
  - 2.5 μs / 25 ns = 100
  - 100 events
- HLT latency ~ 100 ms
- HLT farm size = ?

EXAMPLE: SOME NUMBERS FOR ATLAS AND CMS



- ➡ **Event size = 1 MB**
- ➡ **L1 rate = 100 kHz**
- ➡ **DAQ network size = ?**
  - ➡ 1 MB x 100 kHz =
  - ➡ 100 GB/s (~Tbps)
- ➡ **L1 latency = 2.5 μs**
- ➡ **L1 buffer size = ?**
  - ➡ 2.5 μs / 25 ns = 100
  - ➡ 100 events
- ➡ **HLT latency ~ 100 ms**
- ➡ **HLT farm size = ?**
  - ➡ 100 kHz x 100 ms =

EXAMPLE: SOME NUMBERS FOR ATLAS AND CMS



- Event size = 1 MB
- L1 rate = 100 kHz
- DAQ network size = ?
  - 1 MB x 100 kHz =
  - 100 GB/s (~Tbps)
- L1 latency = 2.5 μs
- L1 buffer size = ?
  - 2.5 μs / 25 ns = 100
  - 100 events
- HLT latency ~ 100 ms
- HLT farm size = ?
  - 100 kHz x 100 ms =
  - O(10^4) CPU cores

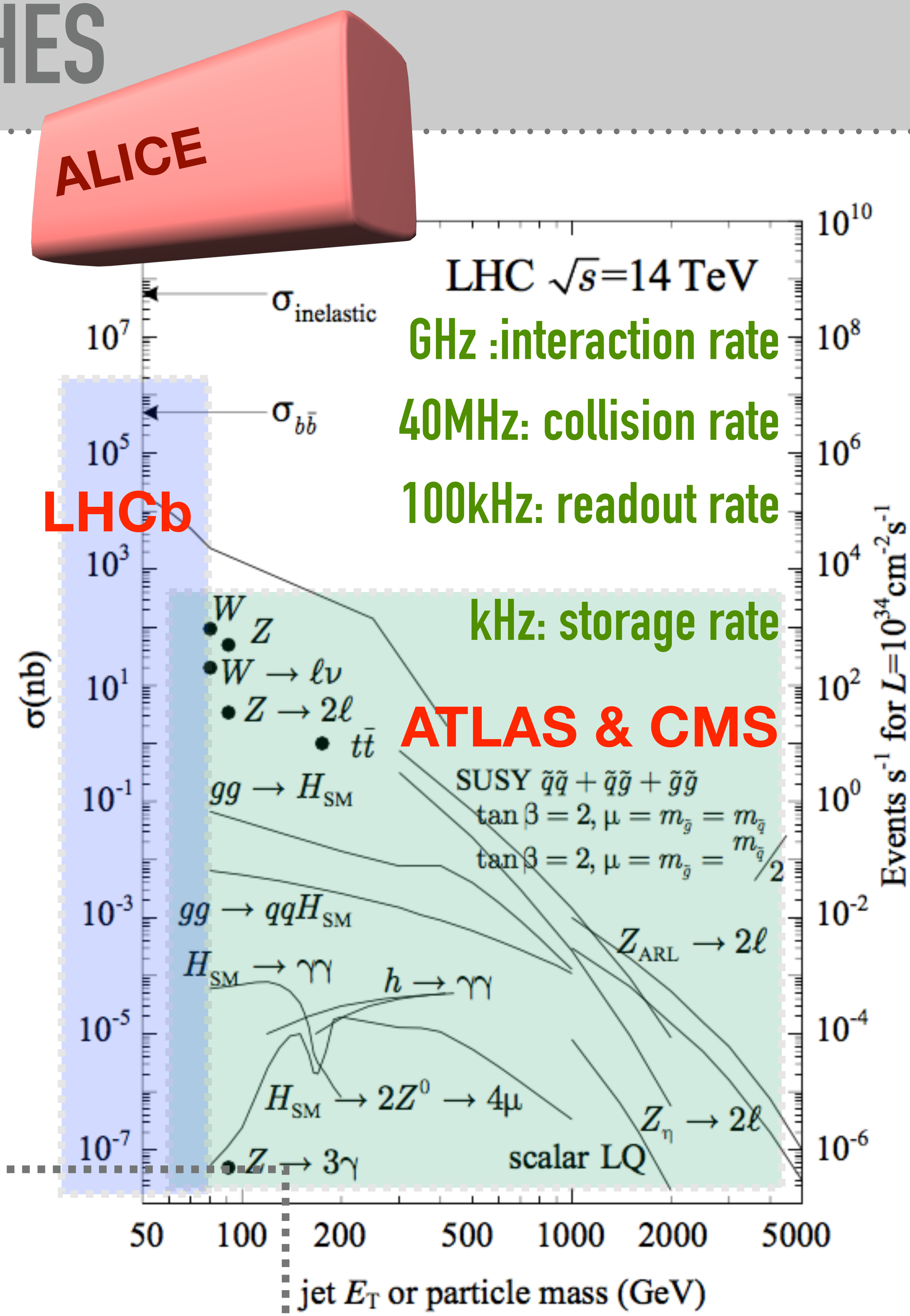
COMPARE 4 EXPERIMENTS

How to maximise physics acceptance

DIFFERENT PHYSICS SEARCHES

…. and LHC operations

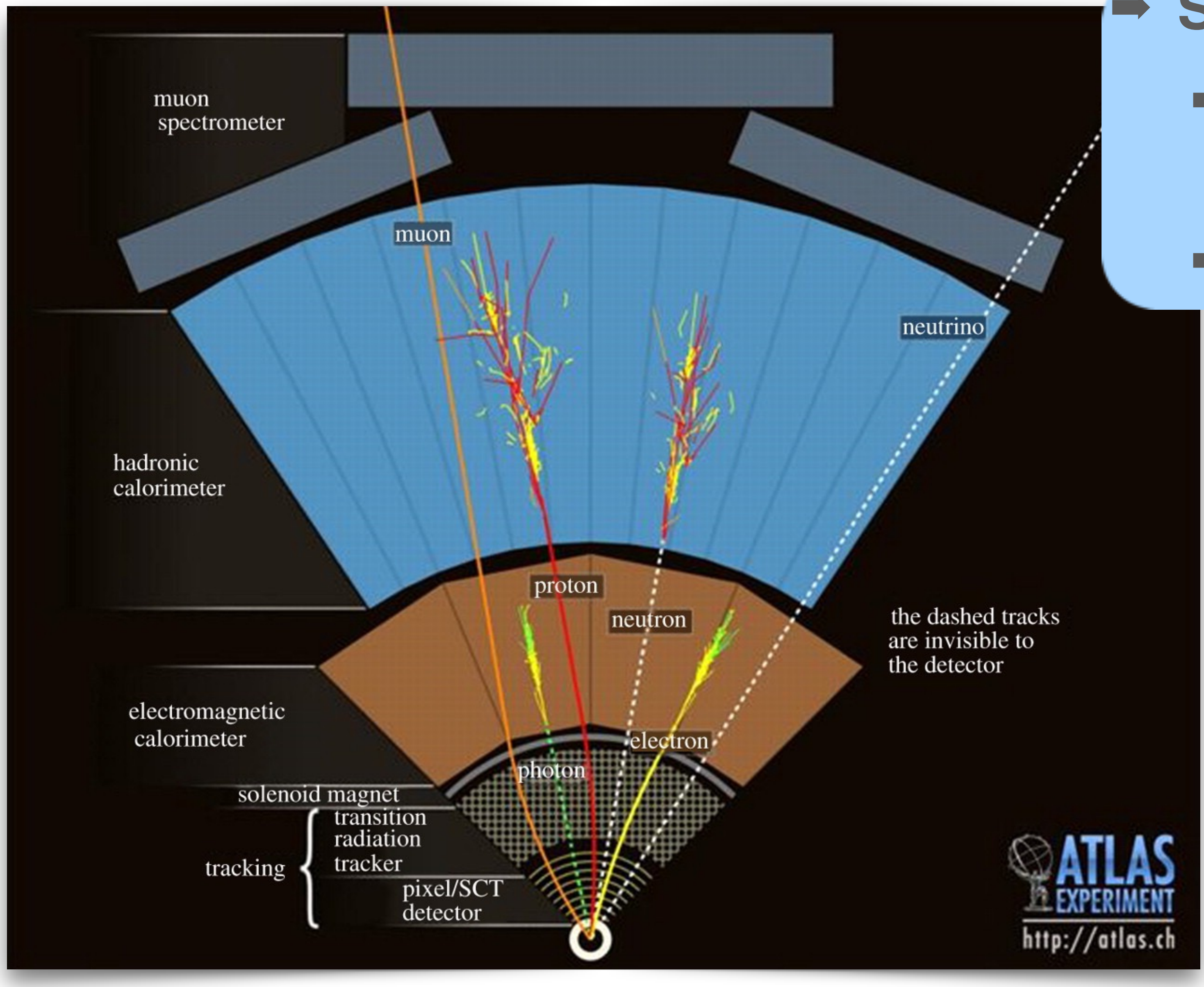
- **ATLAS/CMS: p-p collisions at full Luminosity**
  - search in high energy scale
- **LHCb: p-p collisions at reduced Luminosity**
  - search complex topologies of b-quark decays
- **ALICE: heavy-ion collisions ~2000 mb**
  - search in high energy density



Differ in

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

ATLAS/CMS TRIGGER STRATEGY



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

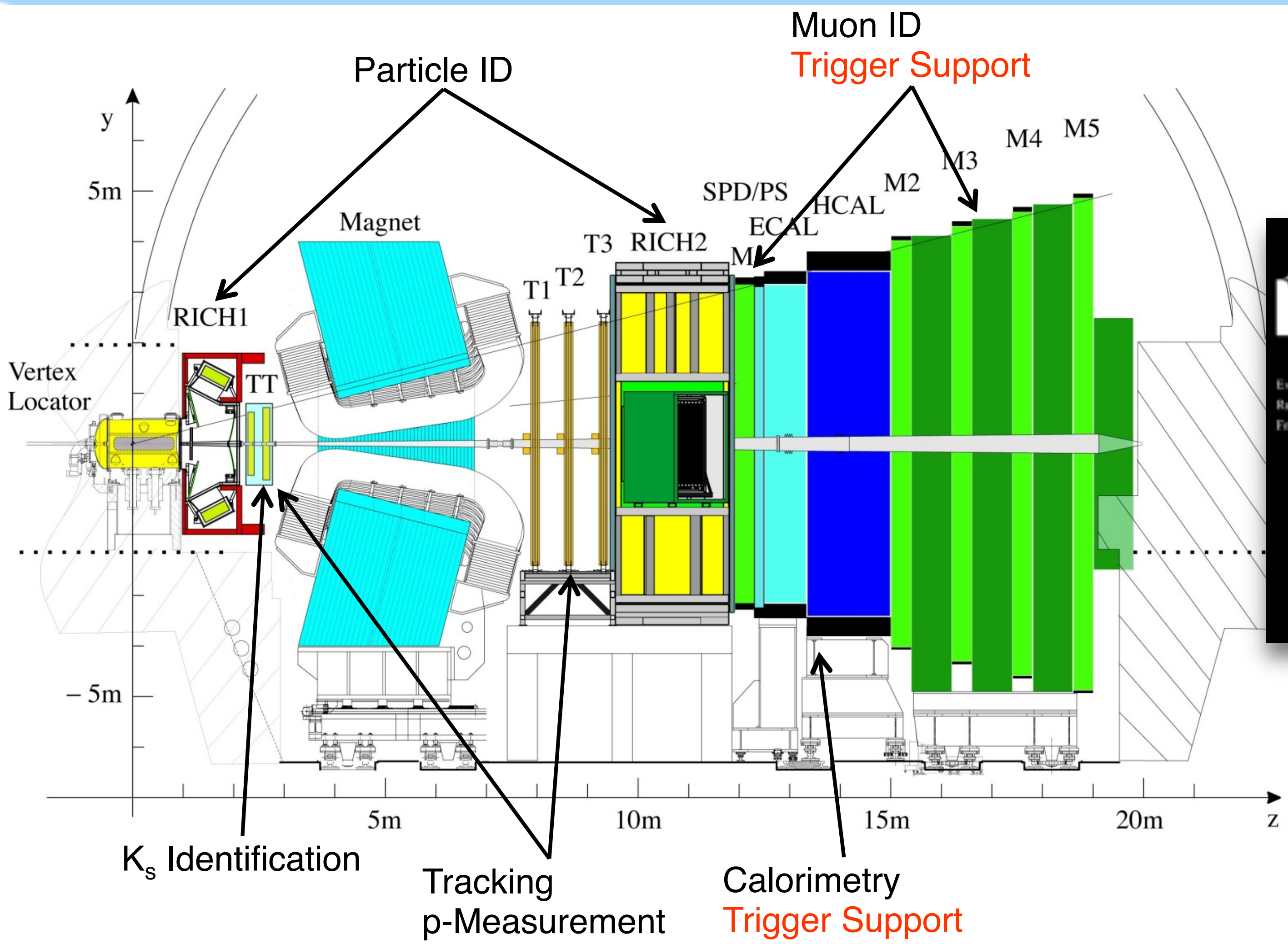
$$\frac{everything}{Higgs} = \frac{\sigma_{tot}}{\sigma_{H(500\,\mathrm{GeV})}} \approx \frac{100\,mb}{1\,pb} \approx 10^{11}$$

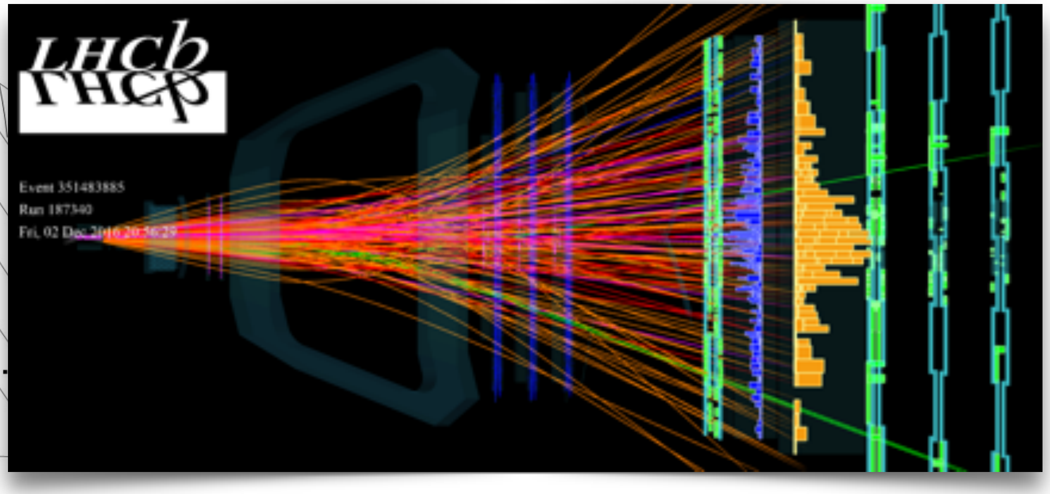
approximately 10^6 rejection is needed

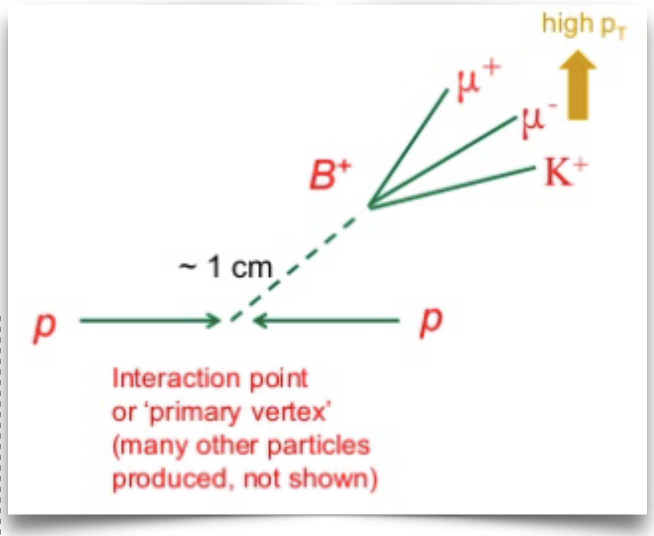
- **Easy selection of high-energy <u>leptons (e/μ)</u> ==> powerful L1**
  - Against thousands of particles/collisions (typically low momentum jets)
- **Remember: 90 M readout channels and full Luminosity ==> 1 MB/event**

LHCB DESIGN PRINCIPLES

➡ **Precision measurements and rare decays in the B system**
- ➡ Large production (σ_{BB} ~ 500 μb), but still σ_{BB}/σ_{Tot} ~ 5 x 10^{-3}
- ➡ Interesting B decays are quite <u>rare</u> (BR ~ 10^{-5})

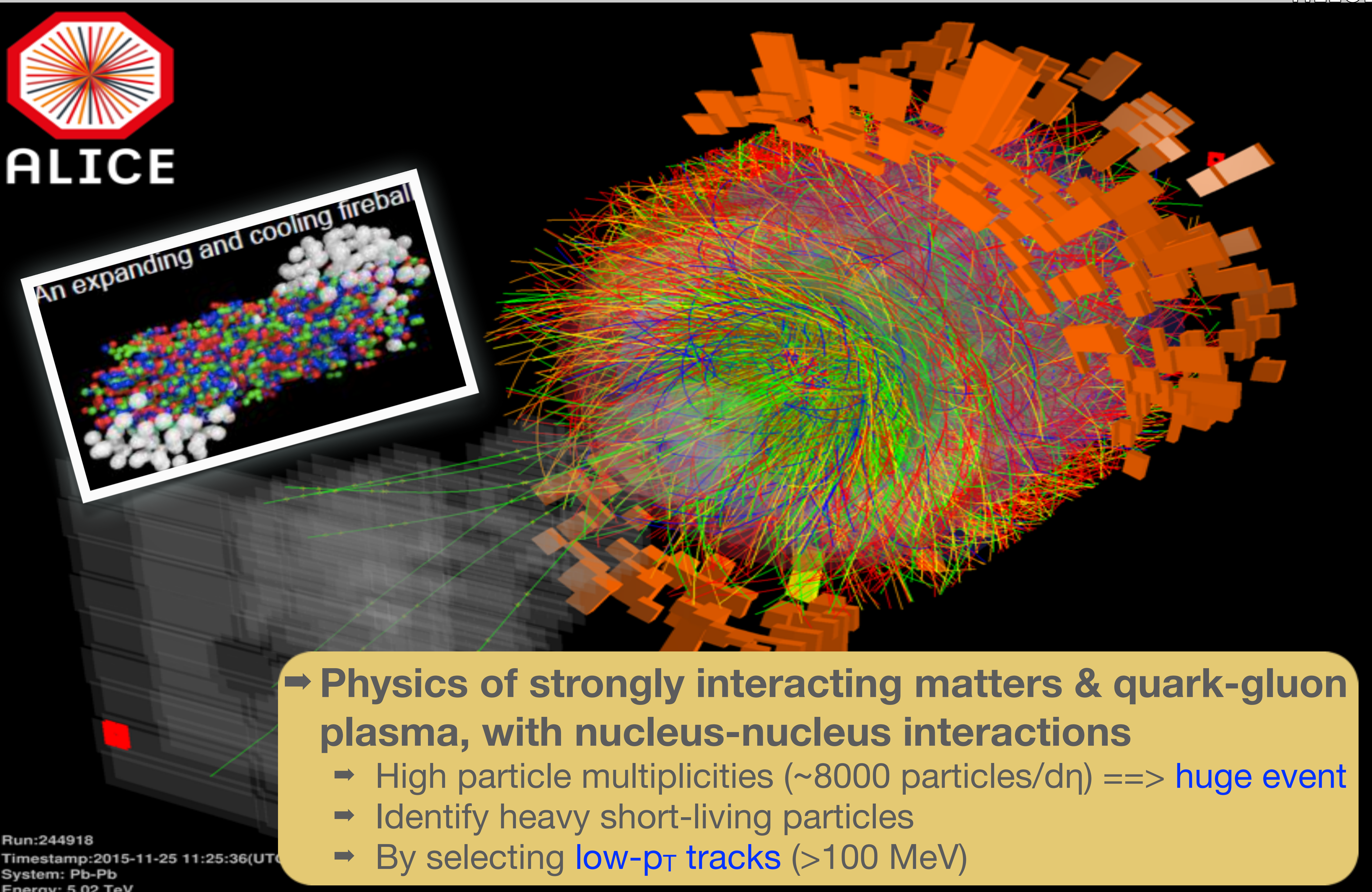






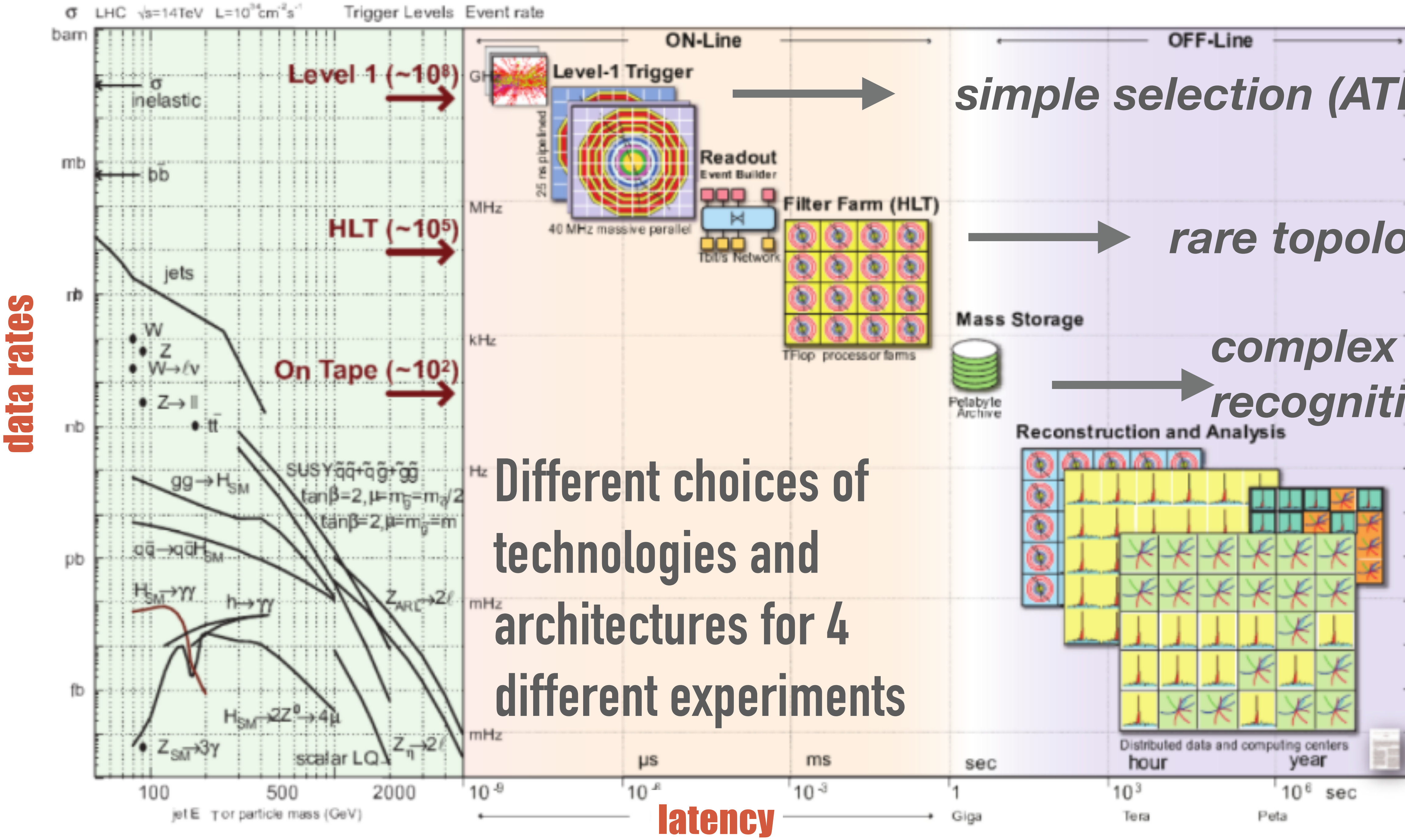
- ➡ **Single-arm spectrometer and low L ==> reduced event size**
- ➡ **Selection of B mesons ==> search for B-decay topologies**
  - ➡ related to high mass and long lifetime of the b-quark

ALICE STRATEGIES



- ➡ **Physics of strongly interacting matters & quark-gluon plasma, with nucleus-nucleus interactions**
  - ➡ High particle multiplicities (~8000 particles/dη) ==> huge event
  - ➡ Identify heavy short-living particles
  - ➡ By selecting low-p_T tracks (>100 MeV)

ENHANCED TRIGGER SELECTIONS



simple selection (ATLAS, CMS)

rare topology (LHCb)

complex pattern recognition (ALICE)

Different choices of technologies and architectures for 4 different experiments

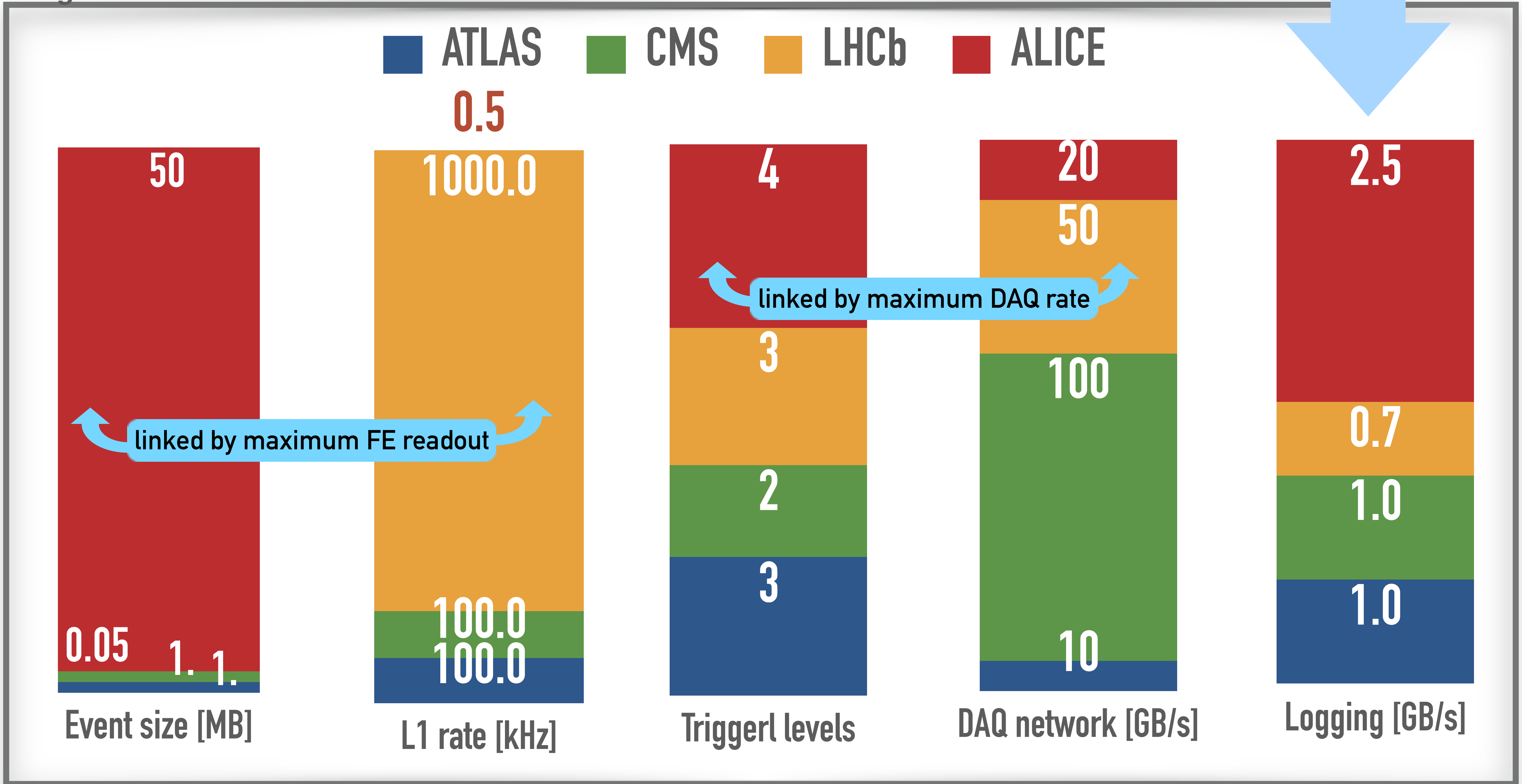
- ➡ **ATLAS/CMS: Trigger power:** reduce the data-flow at the earliest stage
- ➡ **ALICE/LHCb: Large data-flow:** low trigger selectivity due to large irreducible background

COMPARING BY NUMBERS

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

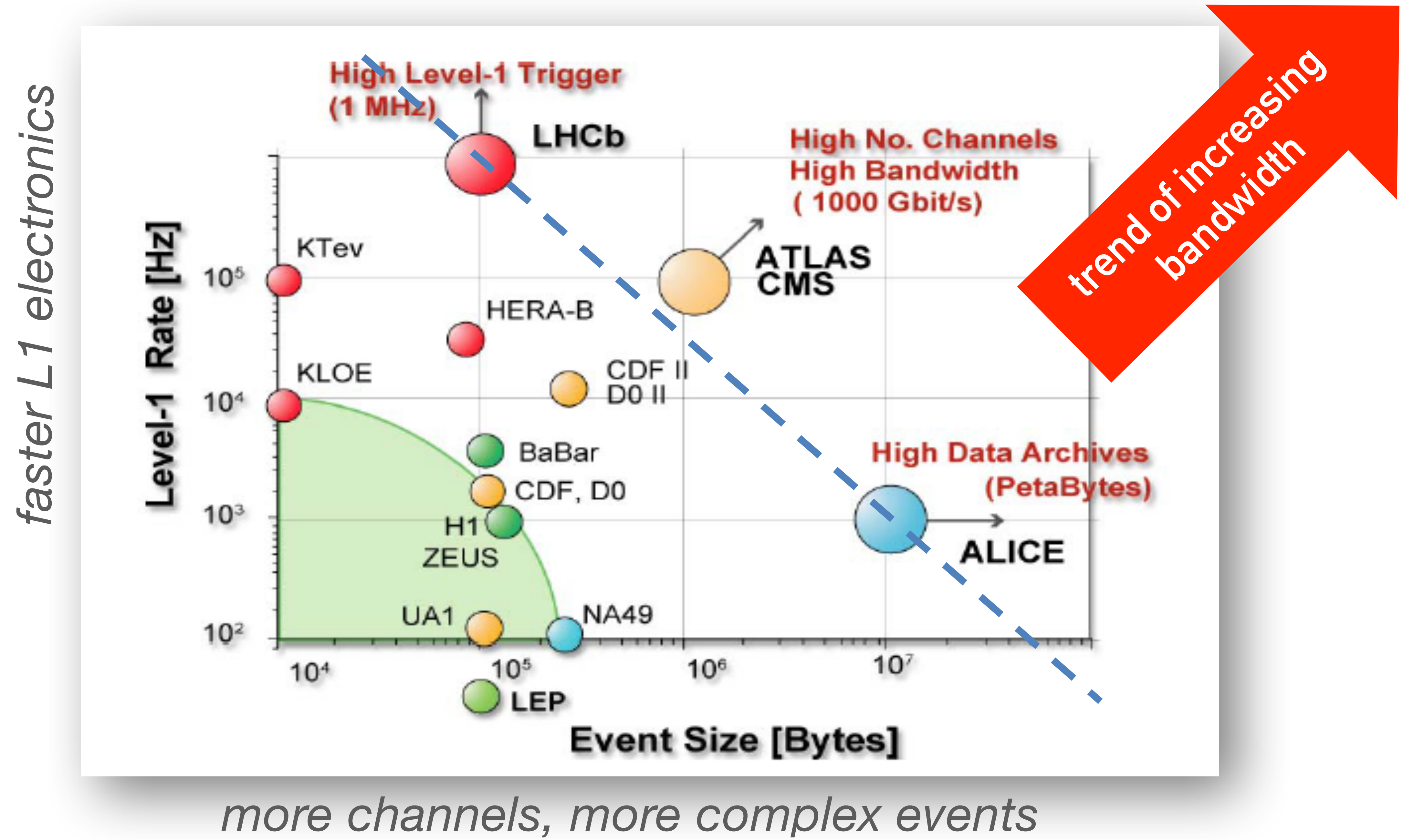
Allowed storage and processing resources

Design values in 2009



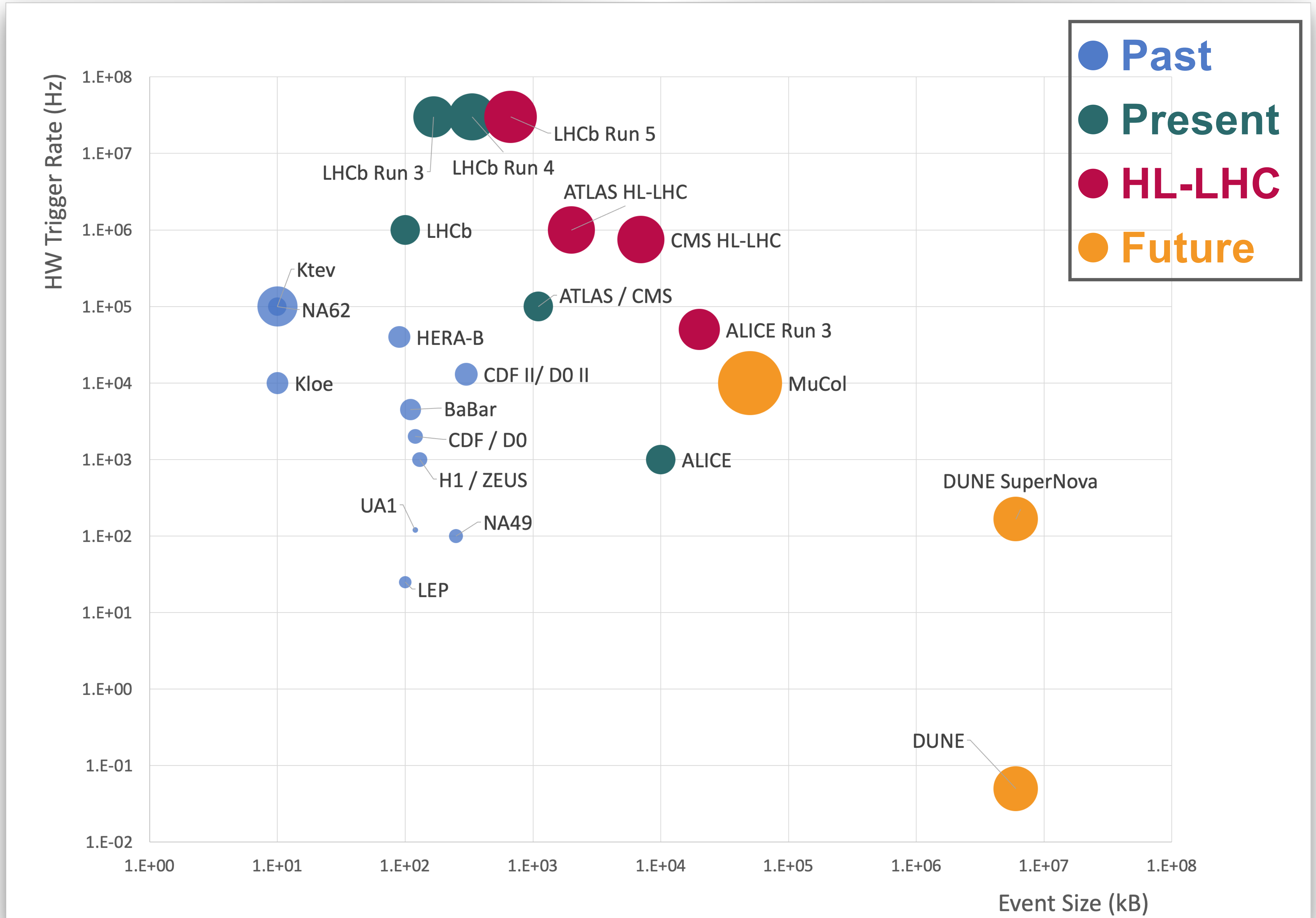
READOUT AND DAQ THROUGHPUTS

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



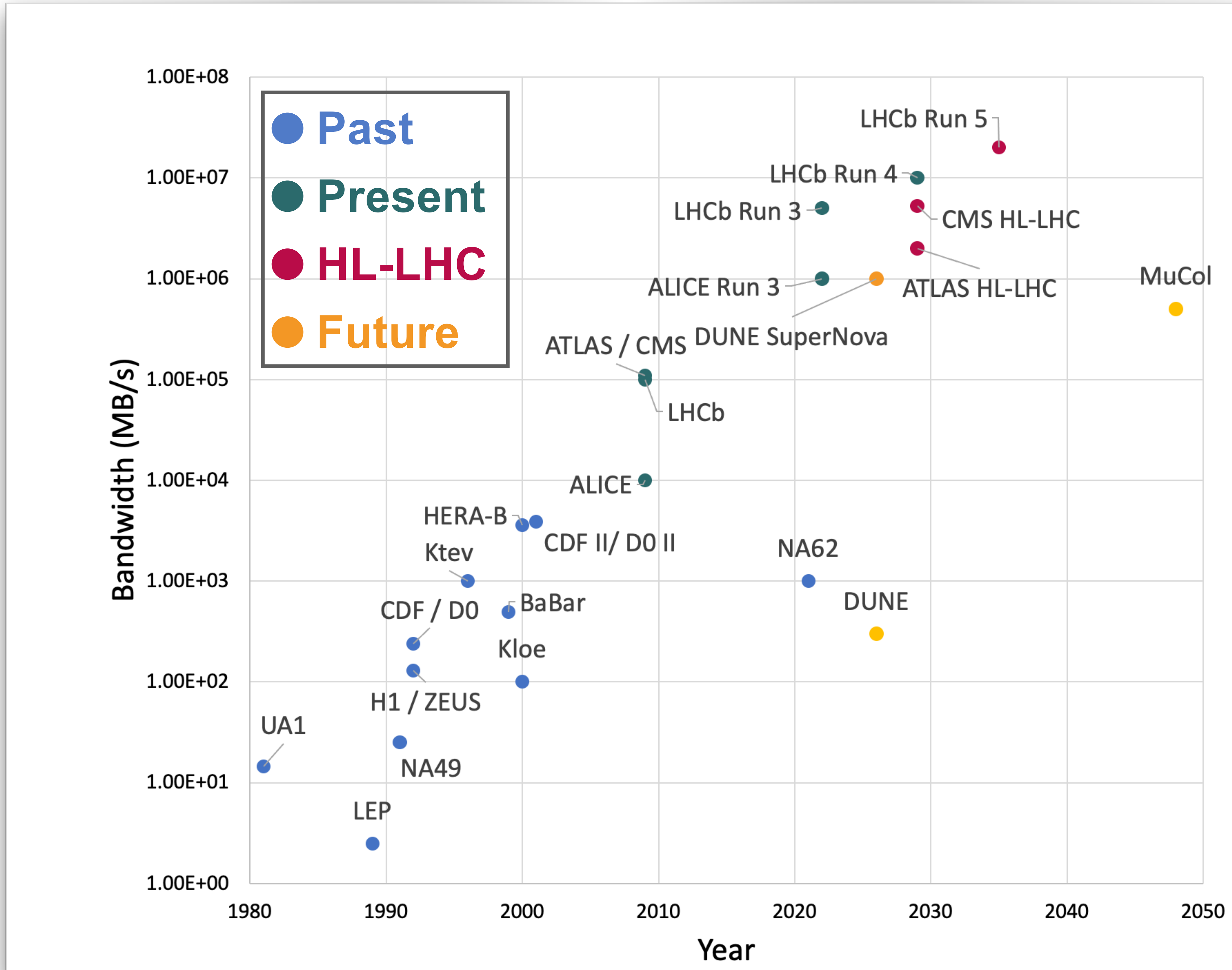
As the data volumes and rates increase, new architectures need to be developed

UPDATED FIGURE!



Courtesy of A.Cerri

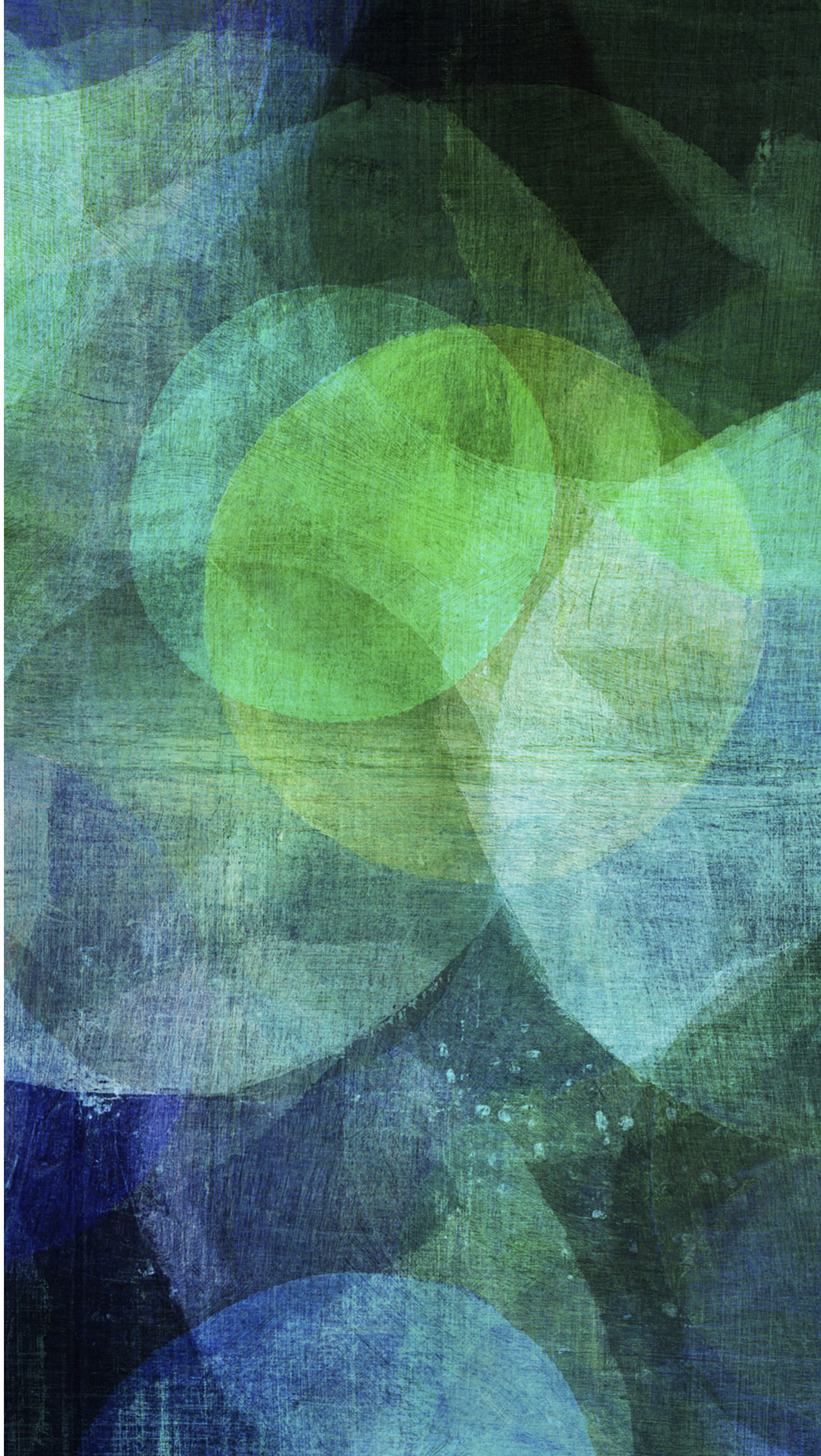
LOOKING FOR MORE DATA IN THE FUTURE



Courtesy of A.Cerri

FUTURE TRENDS FOR HIGH-LUMINOSITY

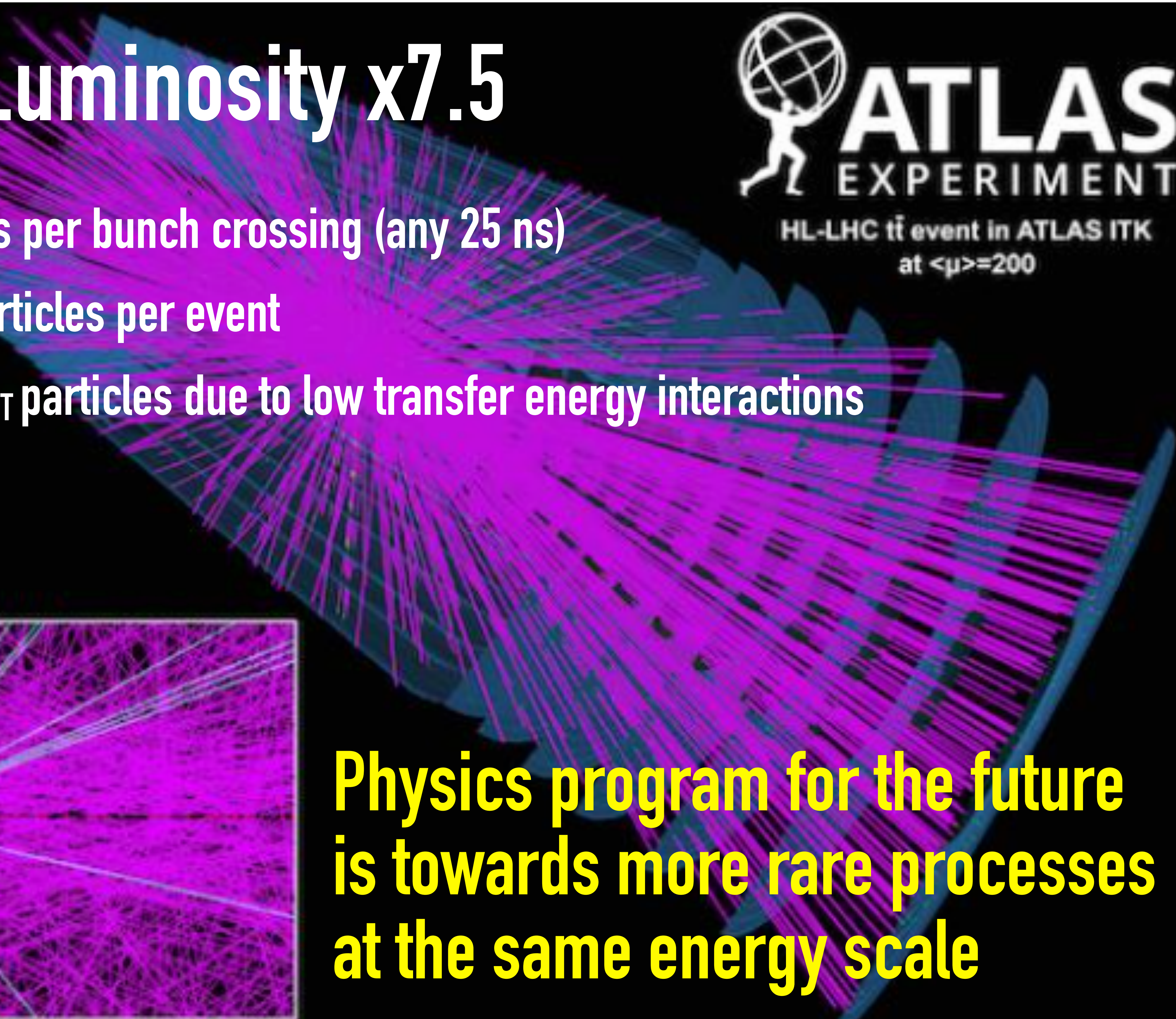
What about … tomorrow?



ONE EVENT AT HIGH-LUMINOSITY LHC (HL-LHC)

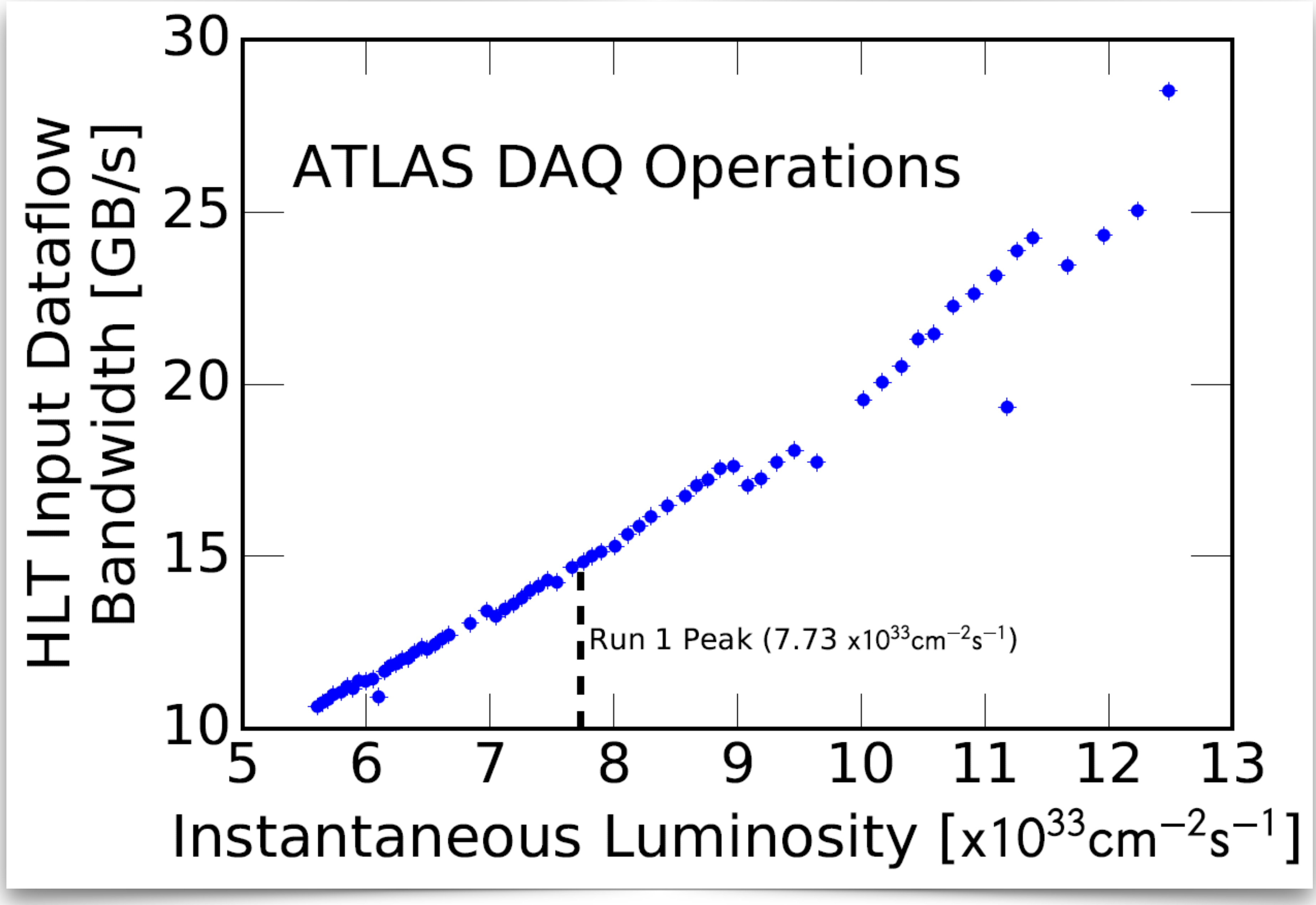
Design Luminosity x7.5

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

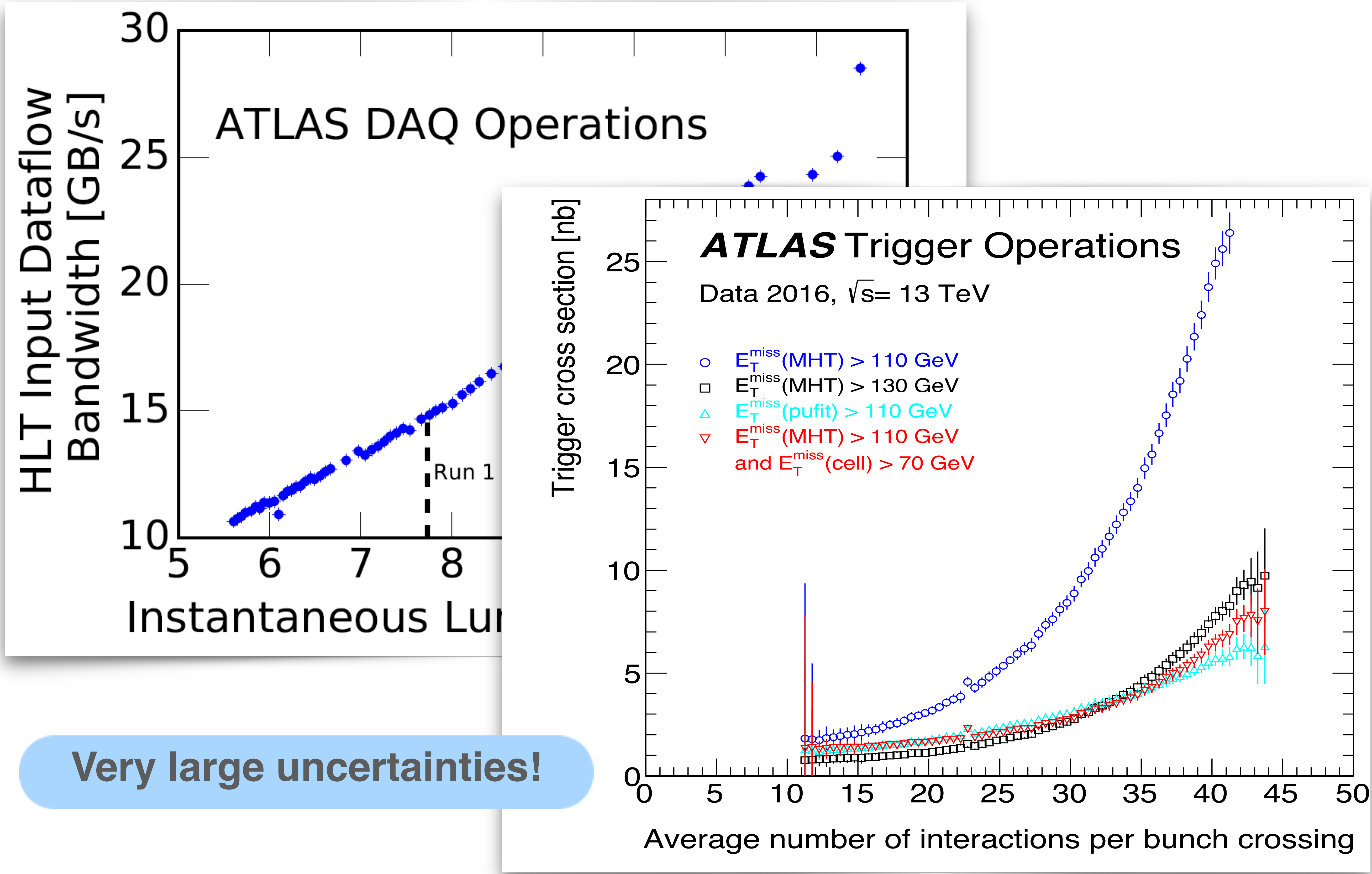


Physics program for the future is towards more rare processes at the same energy scale

WHAT DO YOU EXPECT FOR THE FUTURE?



WHAT DO YOU EXPECT FOR THE FUTURE?

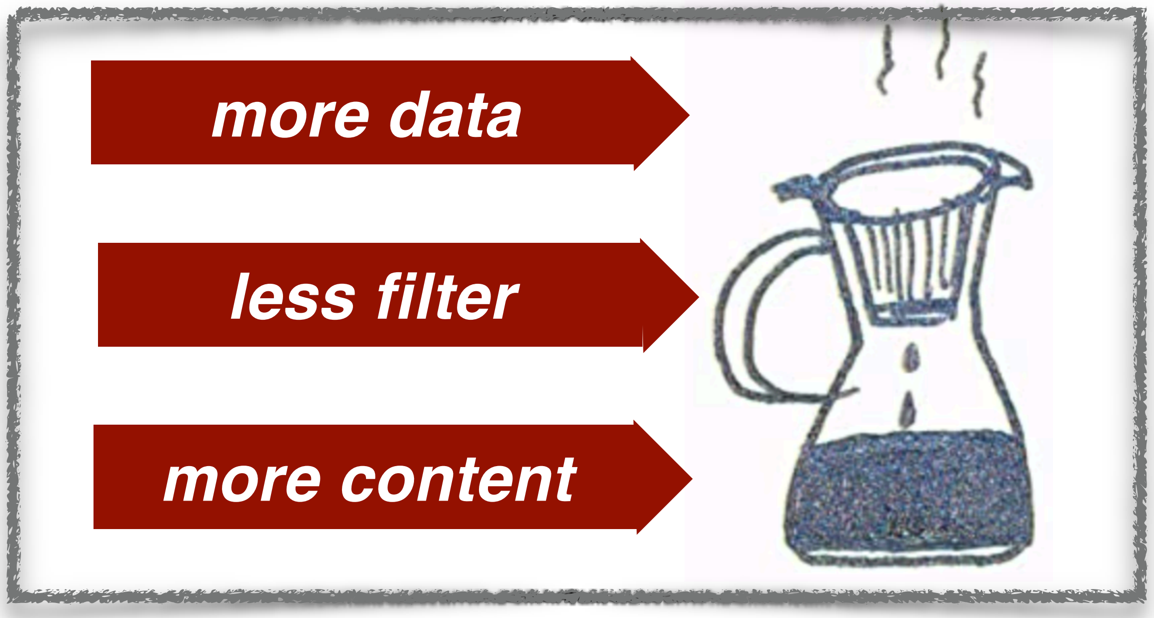


Very large uncertainties!

ADDITIONAL COMPLICATION AT HL-LHC

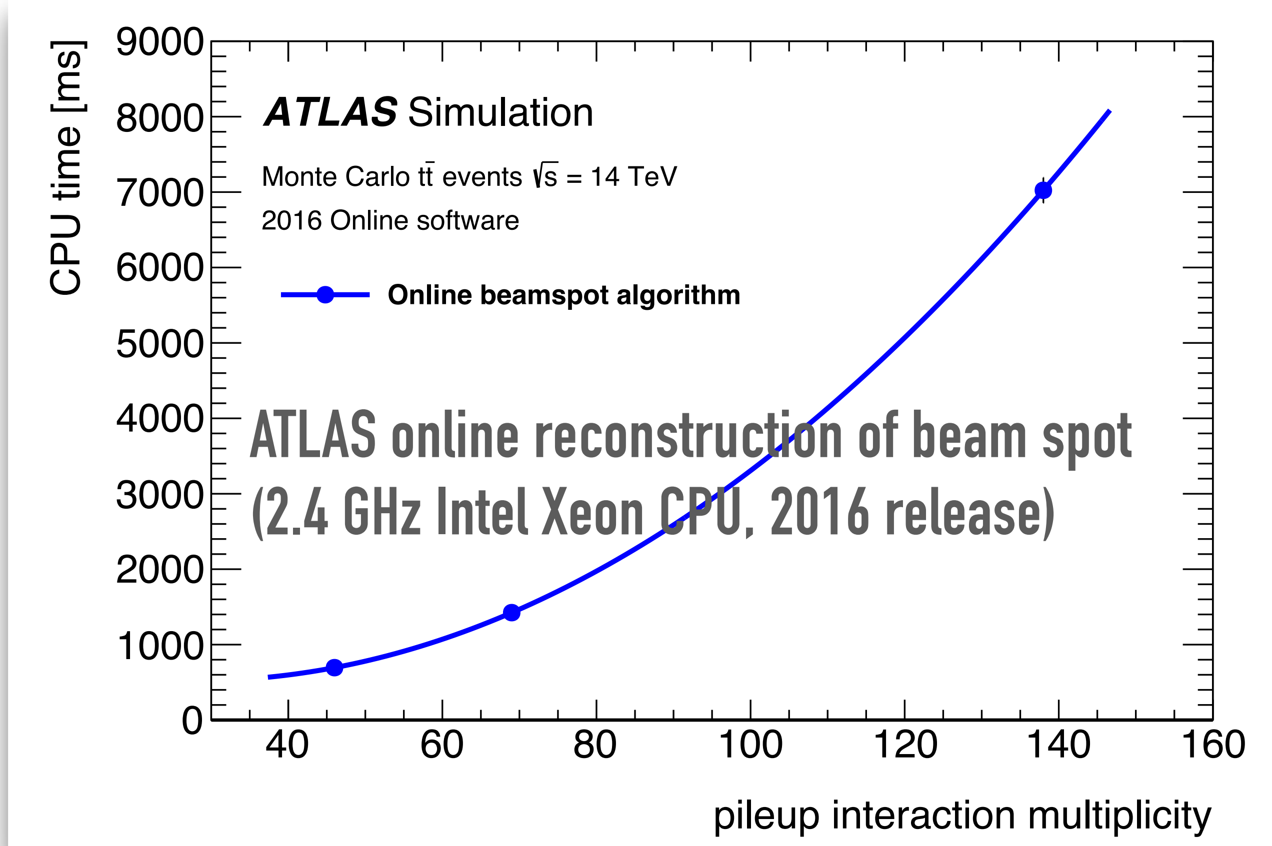
x10 higher Luminosity means…

- ➡ **More interactions per BC (pile-up)**
  - ➡ Less rejection power
    - ➡ worse pattern recognition and resolution
  - ➡ Larger event size (x5)
- ➡ **Larger data rates: i.e. ATLAS/CMS**
  - ➡ Readout rate @L1: 0.1 ➠ 1 MHz
  - ➡ DAQ network: 1 ➠ 50 Tbps



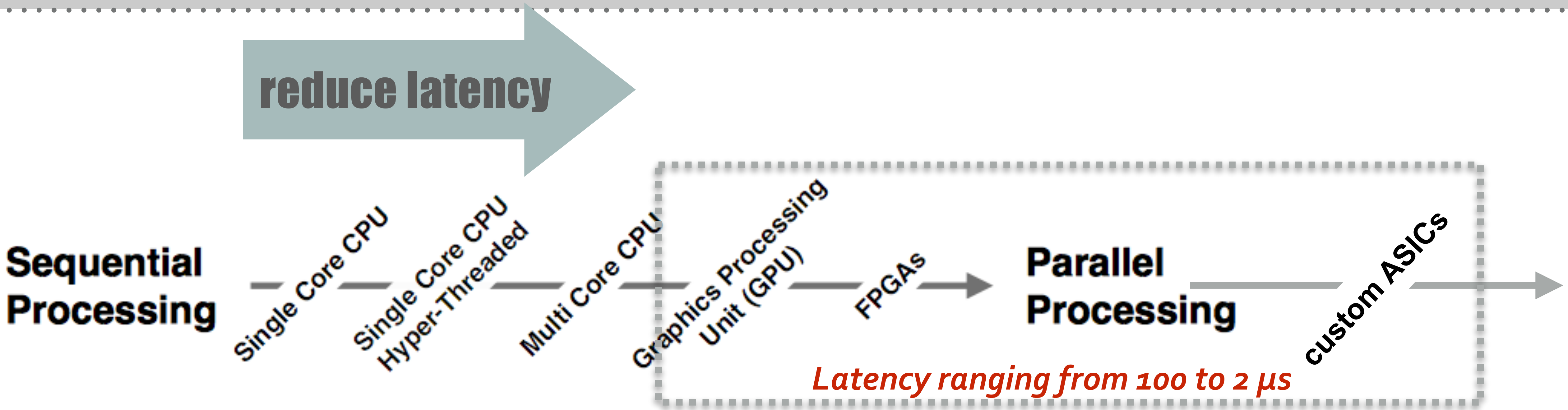
But cannot…

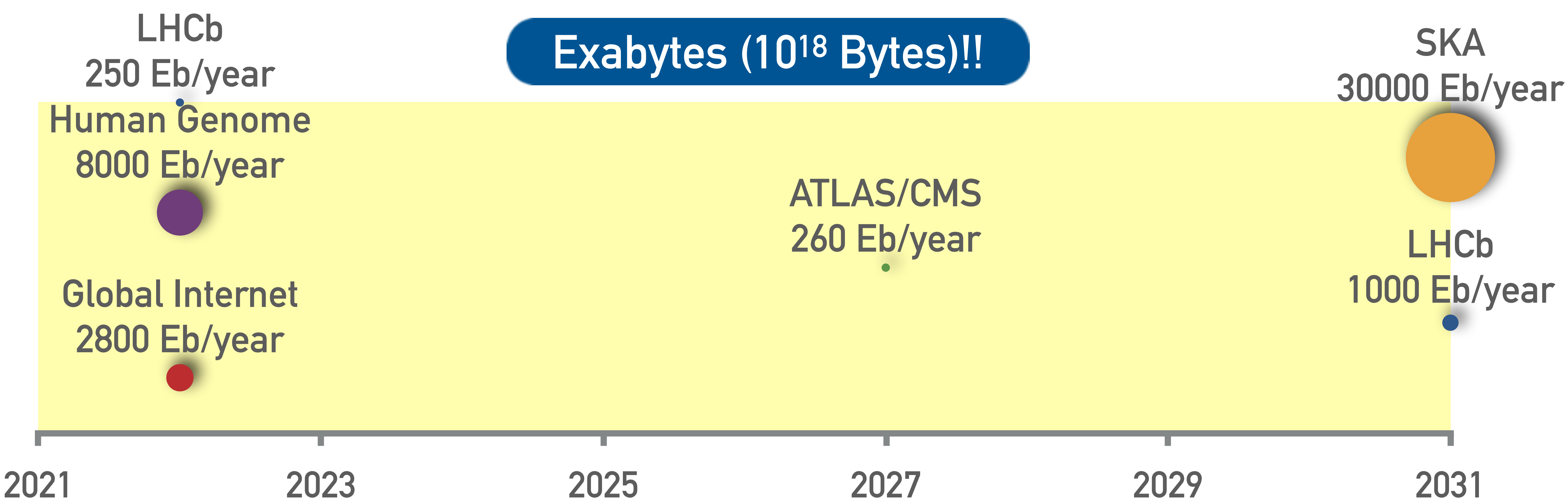
- ➡ **Increase trigger thresholds**
  - ➡ Need to maintain efficiency
- ➡ **Scale dataflow with Luminosity**
  - ➡ **H/W**: more parallelism ➠ more links ➠ more material and cost
  - ➡ **S/W**: processing time not linear ~ L



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

THE REAL-TIME ADVENTURE

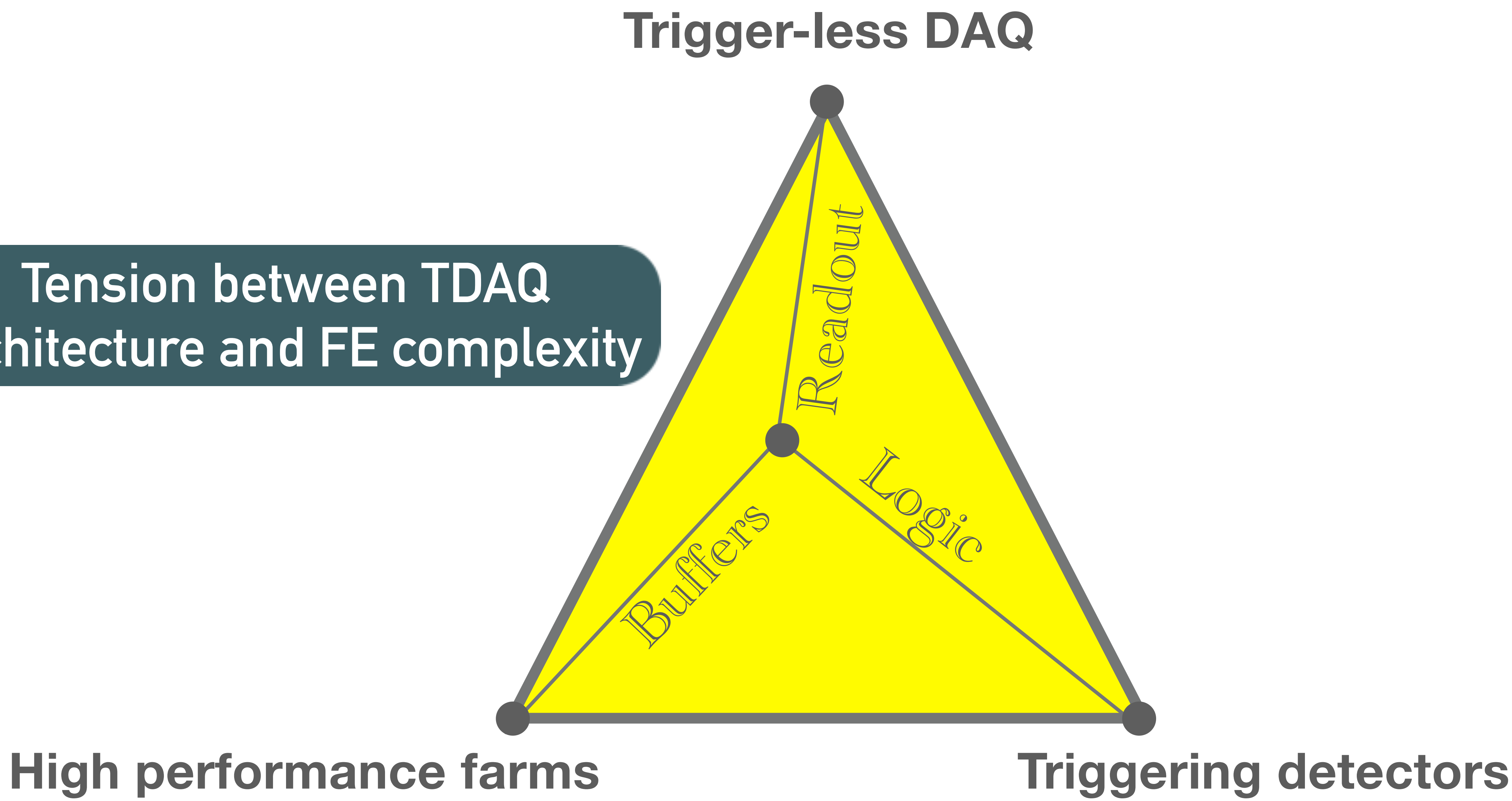




See Openlab workshop 2022

BE SMARTER! INCREASE RESOLUTION FOR BETTER S/B

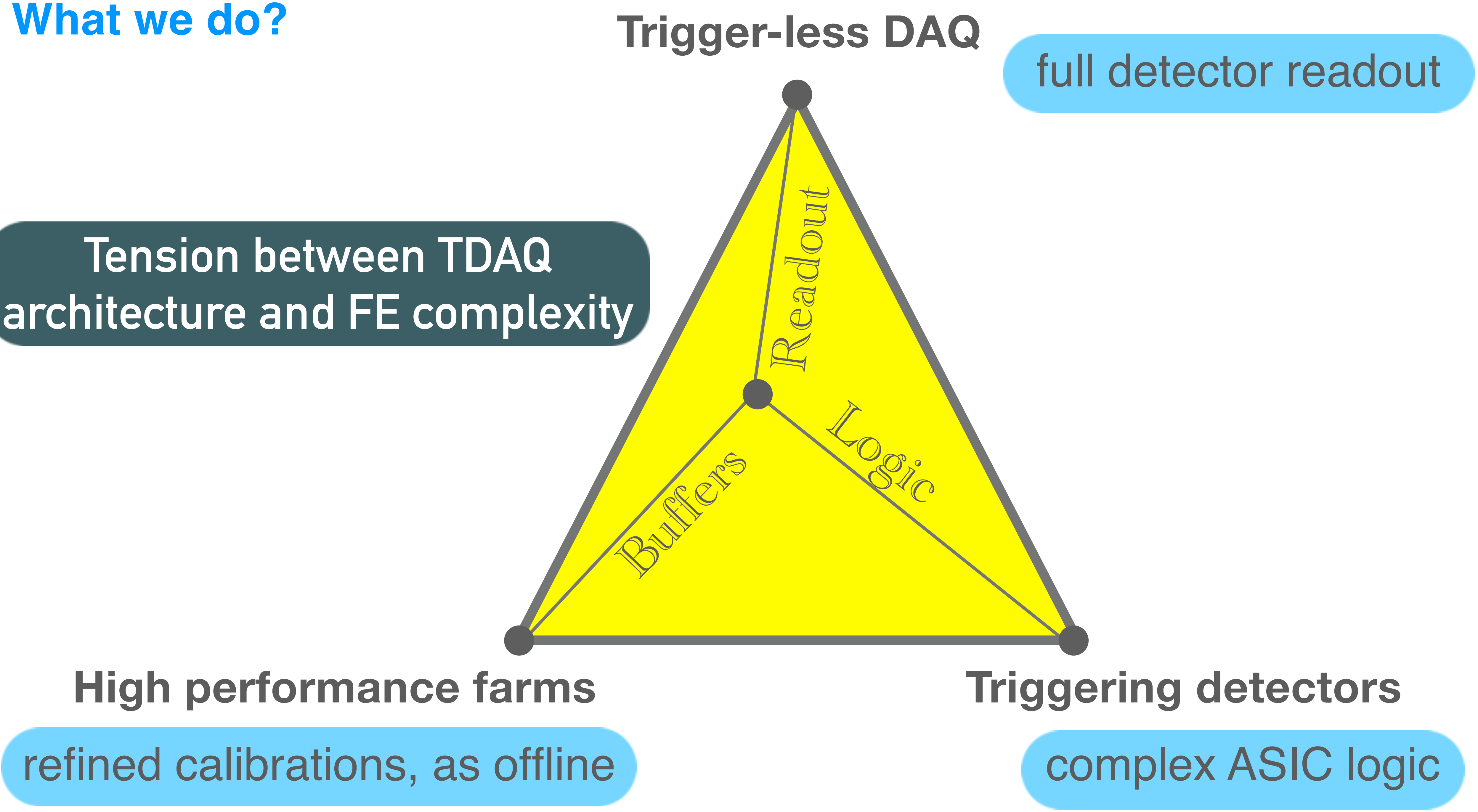
Tension between TDAQ architecture and FE complexity



BE SMARTER! INCREASE RESOLUTION FOR BETTER S/B

What we do?

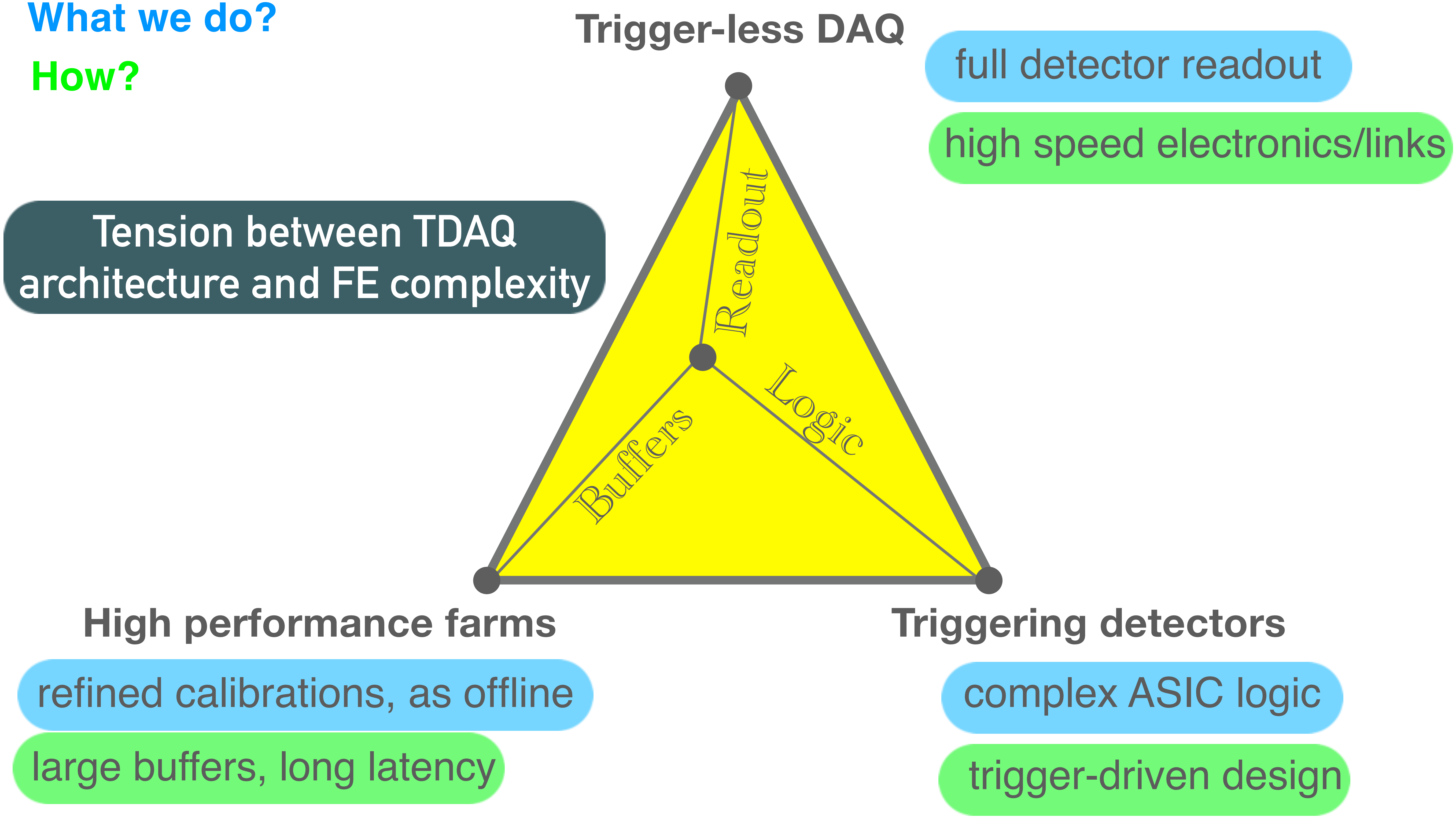
Tension between TDAQ architecture and FE complexity



BE SMARTER! INCREASE RESOLUTION FOR BETTER S/B

What we do?

How?



BE SMARTER! INCREASE RESOLUTION FOR BETTER S/B

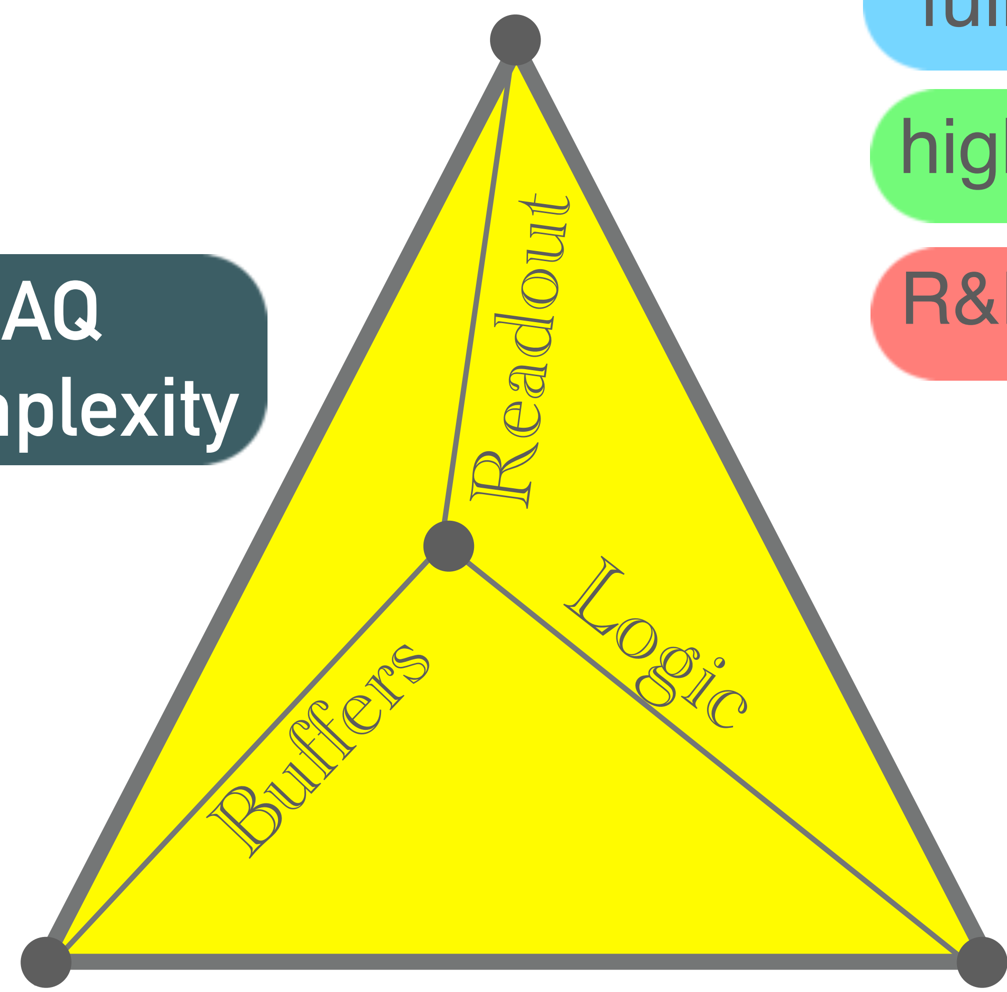
What we do?

How?

Example

Tension between TDAQ architecture and FE complexity

Trigger-less DAQ



full detector readout

high speed electronics/links

R&D on detectors Front-End

High performance farms

refined calibrations, as offline

large buffers, long latency

tight: offline=online (LHCb, ALICE)
soft: decouple trigger/DAQ (ATLAS, CMS)

Triggering detectors

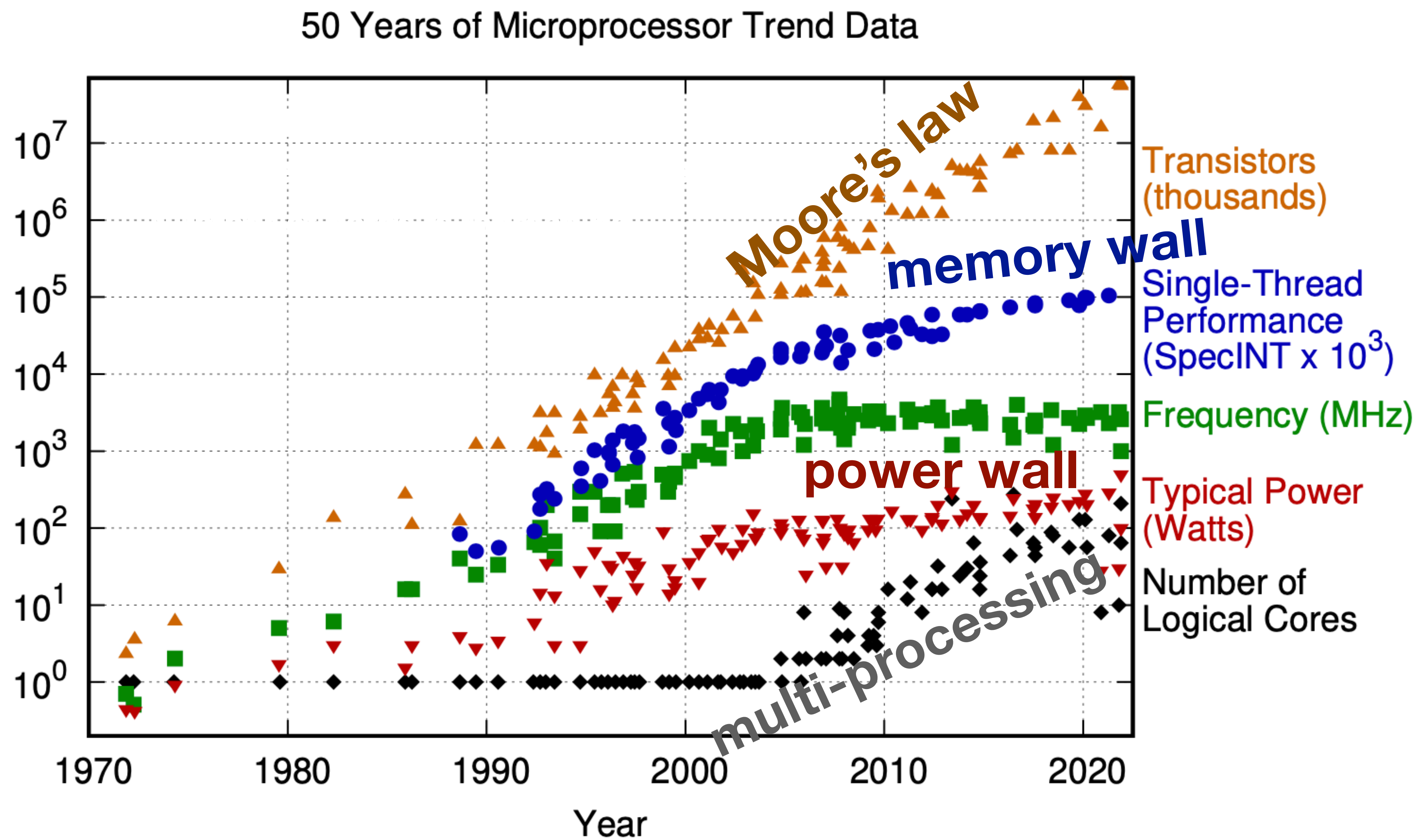
complex ASIC logic

trigger-driven design

hardware track trigger (CMS)

LHCP-2022

EVOLUTION OF PROCESSING POWER TO BREAK THE WALLS



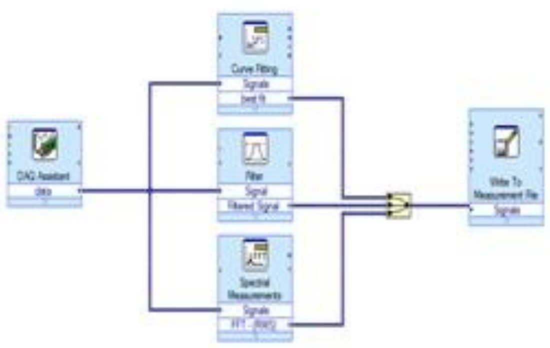
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2021 by K. Rupp
See Reference

- ➡ **Multi-threading processing on CPU**
- ➡ **Use of co-processors (GPU/FPGA)**
- ➡ **Exploit CPU h/w at low level**
  - ➡ Vectorisation, low-level memory…

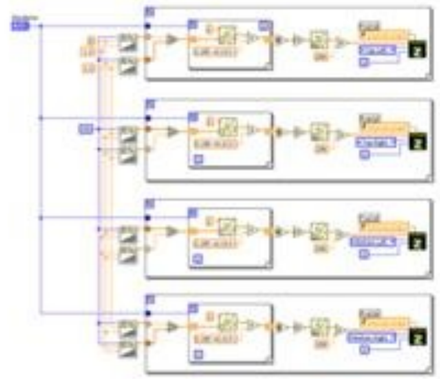
This requires fundamental re-write/ optimization of the software

See news from HPC computing (2022)

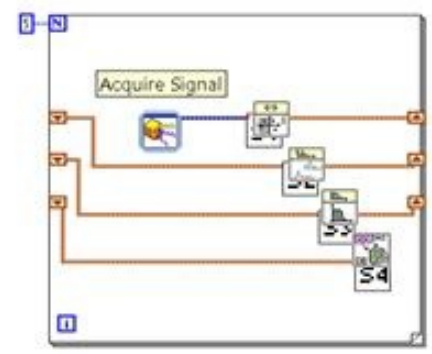
TRENDS: COMBINED TECHNOLOGY



Task Parallelism



Data Parallelism

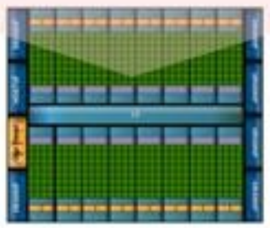


Pipelining

Nvidia GPUs:
3.5 B transistors

Virtex-7 FPGA:
6.8 B transistors

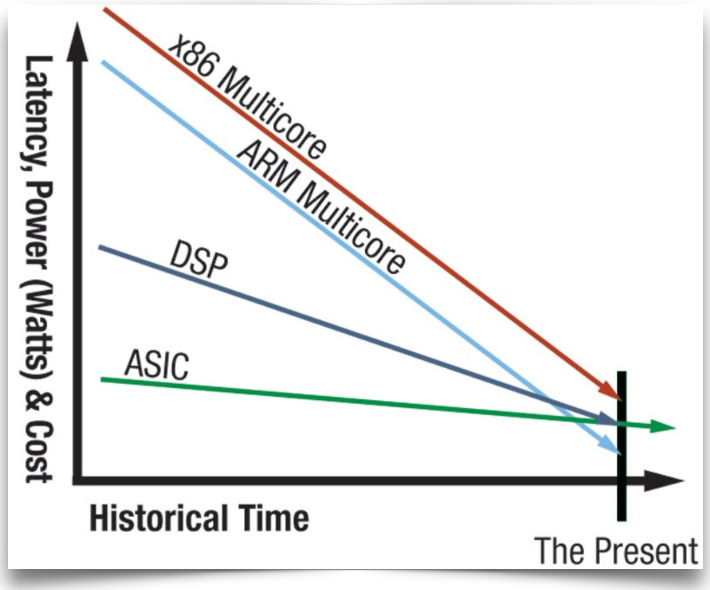
Multicore Processors



GPUs*

FPGAs

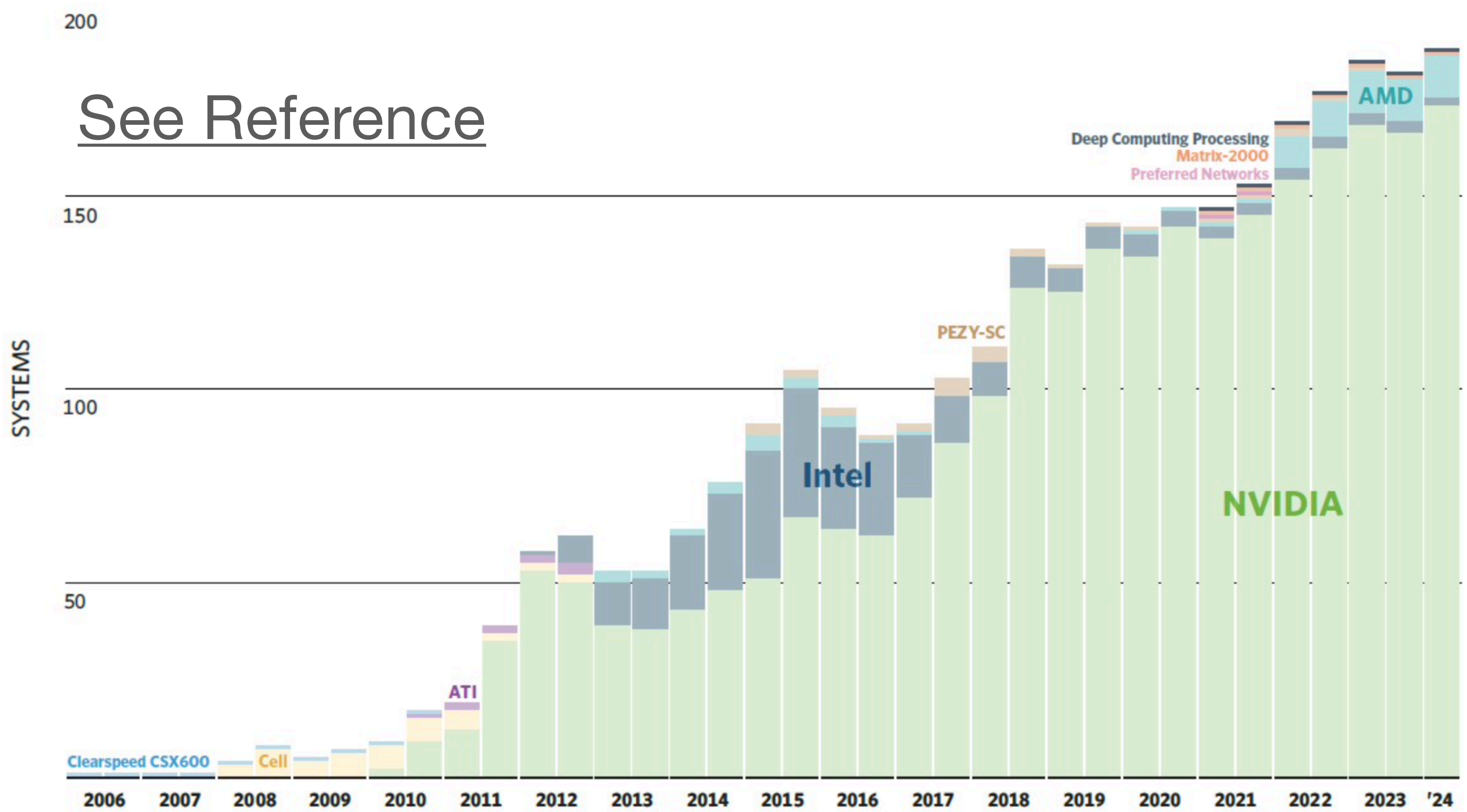
(*) Access to the nVIDIA® GPUs through the CUDA and CUBLAS toolkit/library using the NI LabVIEW GPU Computing framework.



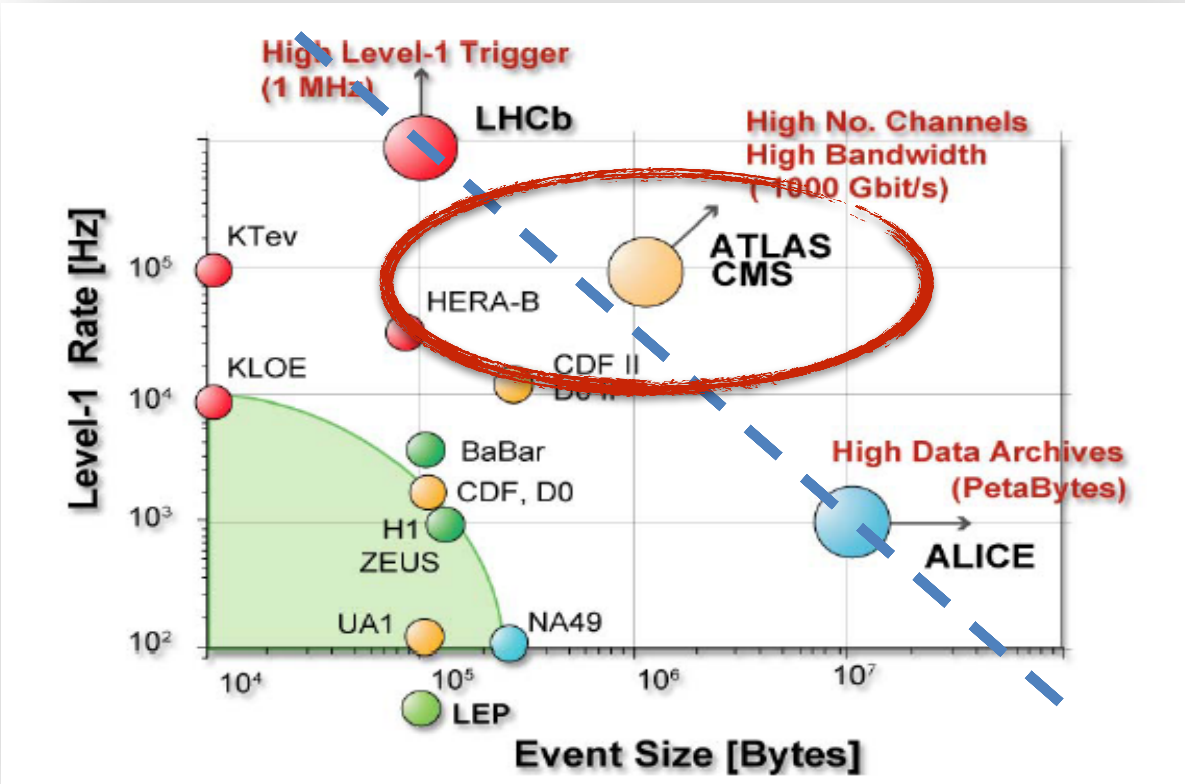
The right choice can be combining the best of both worlds by analysing which strengths of FPGA, GPU and CPU best fit the different demands of the application

TOWARDS THE EXA-SCALE COMPUTING

- **Scientific computing is the third paradigm, complementing theory and experiment**
  - Global scientific facilities (e.g., LIGO, LHC, Vera Rubin Observatory, Square Kilometer Array)
- **Future trends in HPC focusing on:**
  - Rise of massive scale commercial clouds (Google Kubernetes, server-less computing,….)
  - Evolution of semiconductor technology (chip size and packaging, see Amazon Graviton 3)

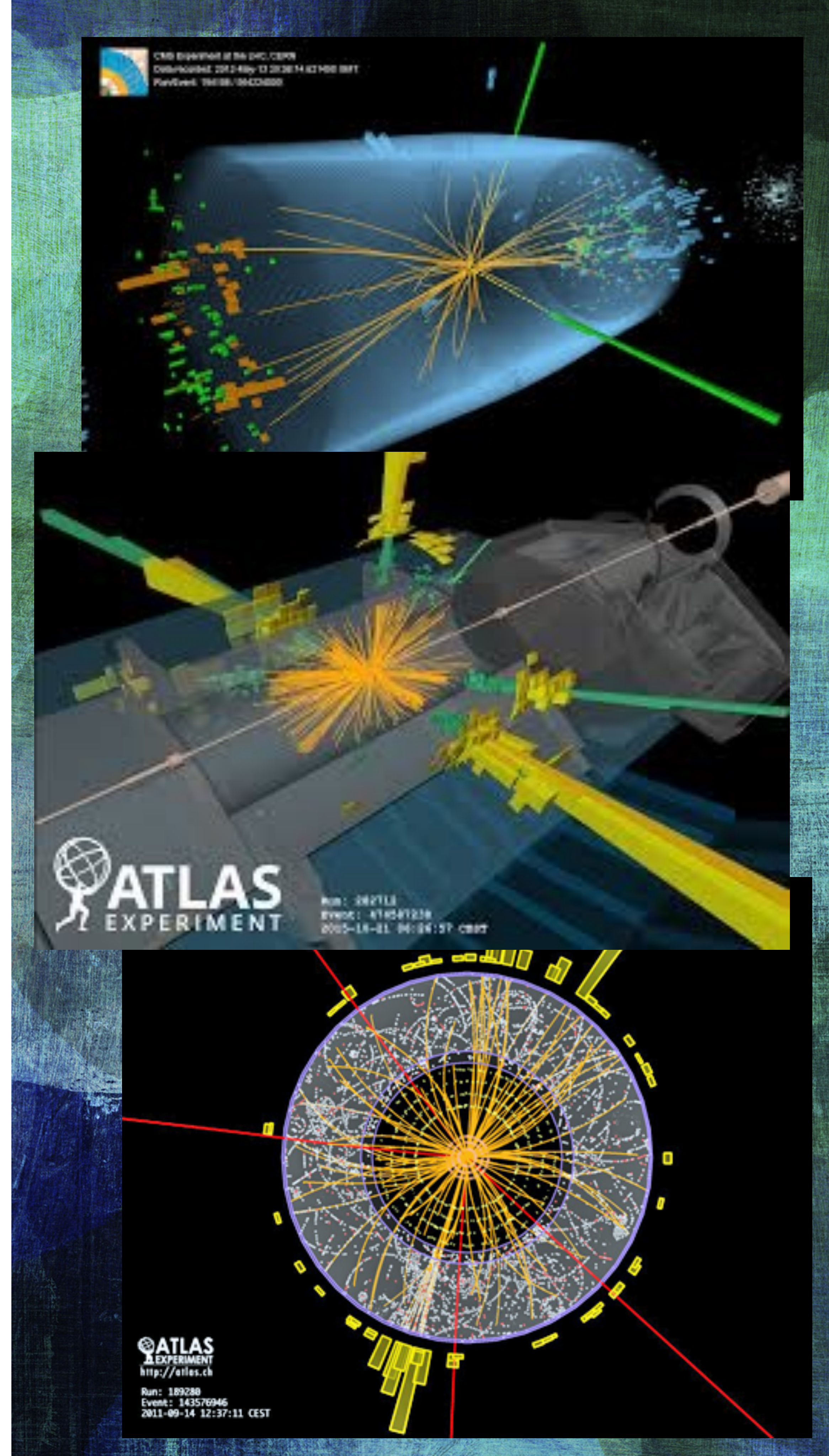


TOP500 today largely examples of a commodity monoculture: nodes with server-class microprocessors + GPUs



ATLAS AND CMS

Studying the Standard Model at the high energy frontier



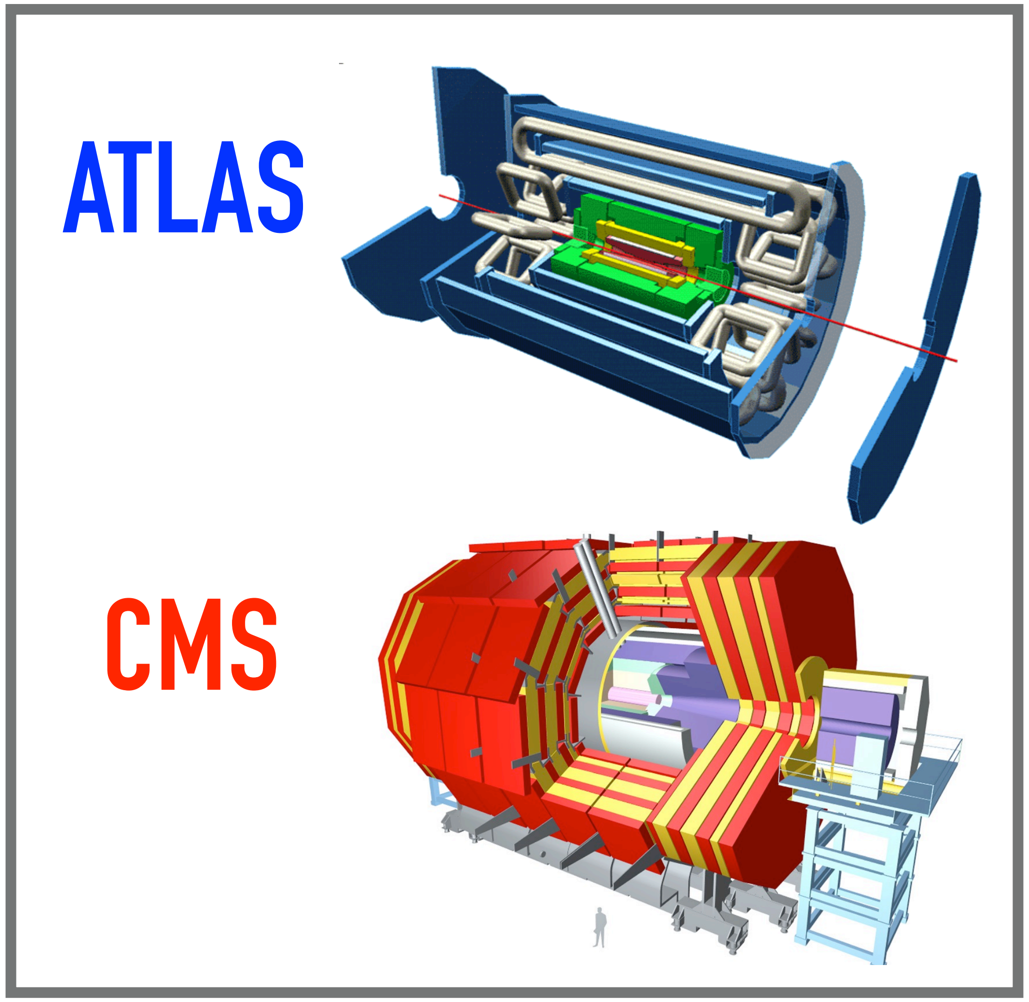
ATLAS & CMS DESIGN PRINCIPLES

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

- ➡ **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



CMS: 2-STAGE EVENT BUILDING IN RUN 1

CMS DAQ-1

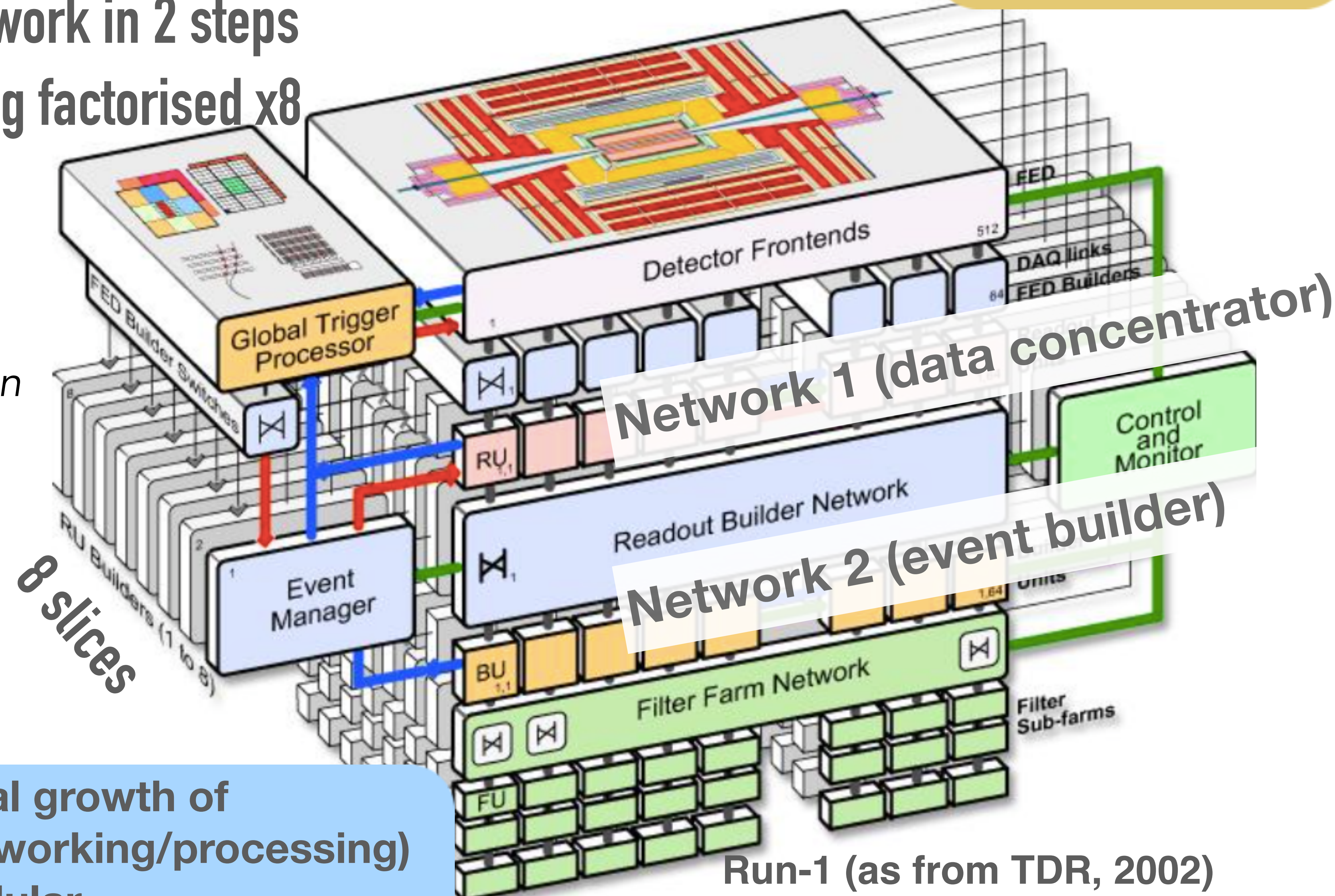
Cannot do Event Building at 100 kHz

100 GB/s readout network in 2 steps

100 kHz Event Building factorised x8

2 EB networks in blu

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

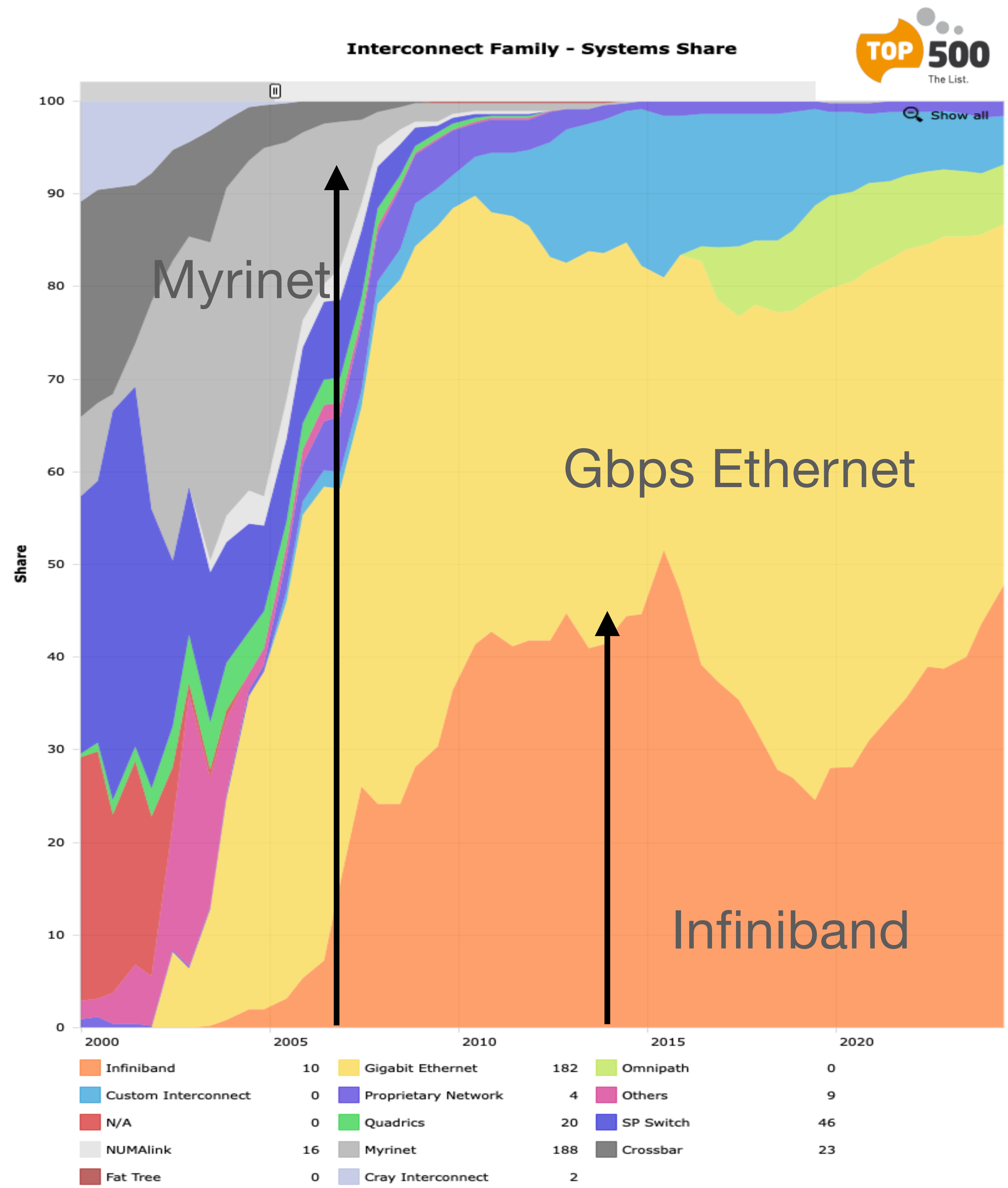
Myrinet widely used when DAQ-1 was designed

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

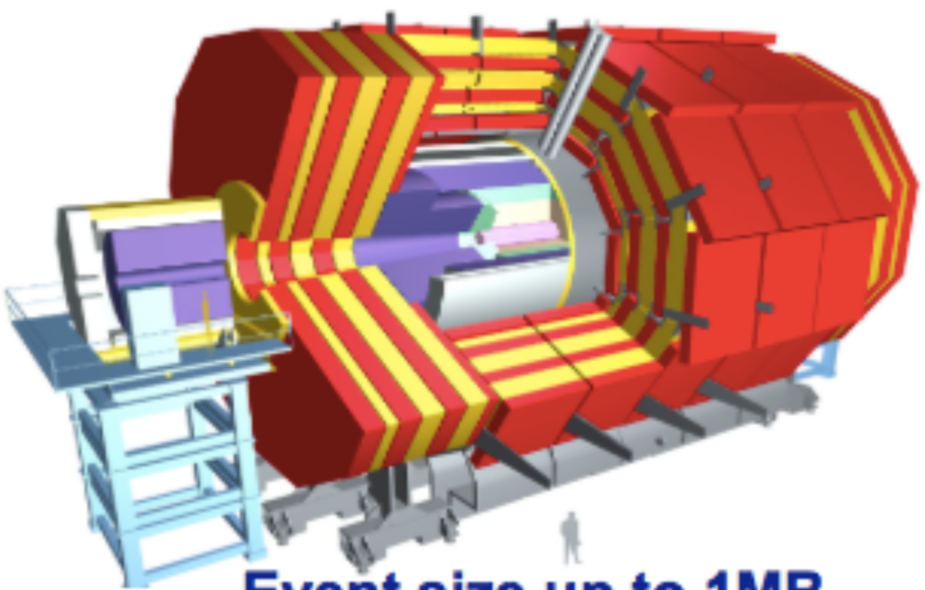
Run 2: 200 GB/s network

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

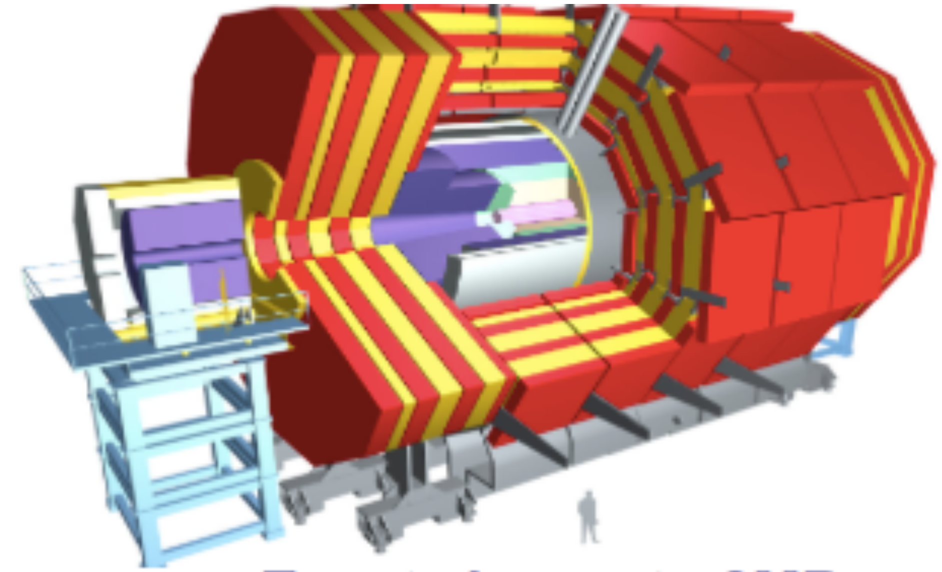
Choose best prize/bitps!



EVOLUTION FROM RUN-1 TO RUN-2

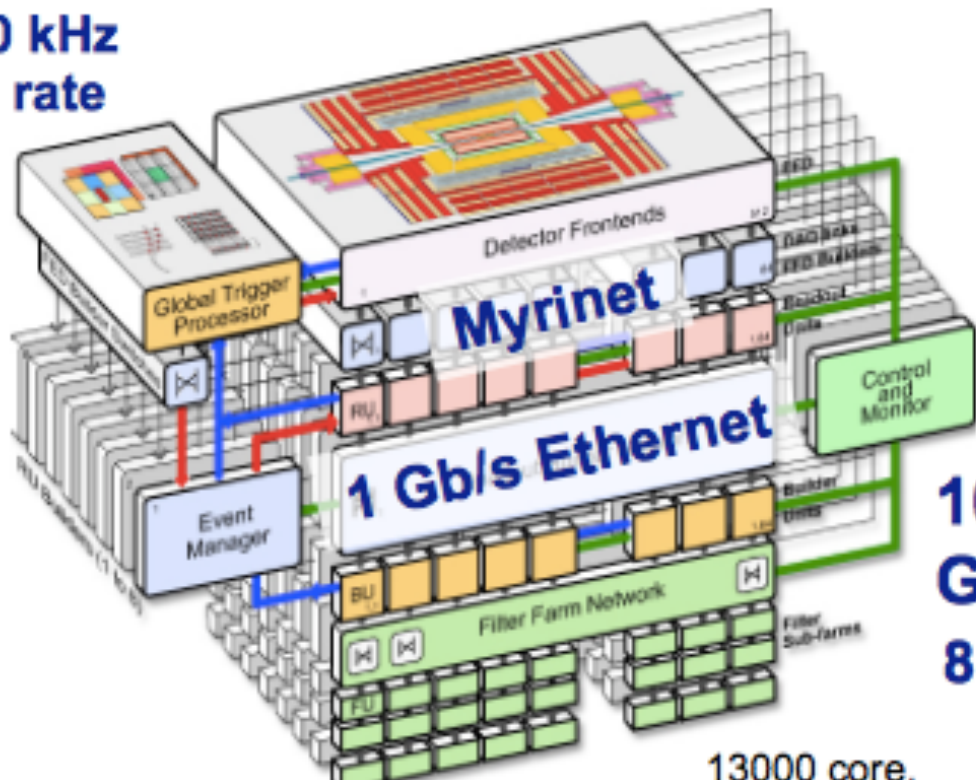


Event size up to 1MB



Event size up to 2MB

100 kHz
L1 rate



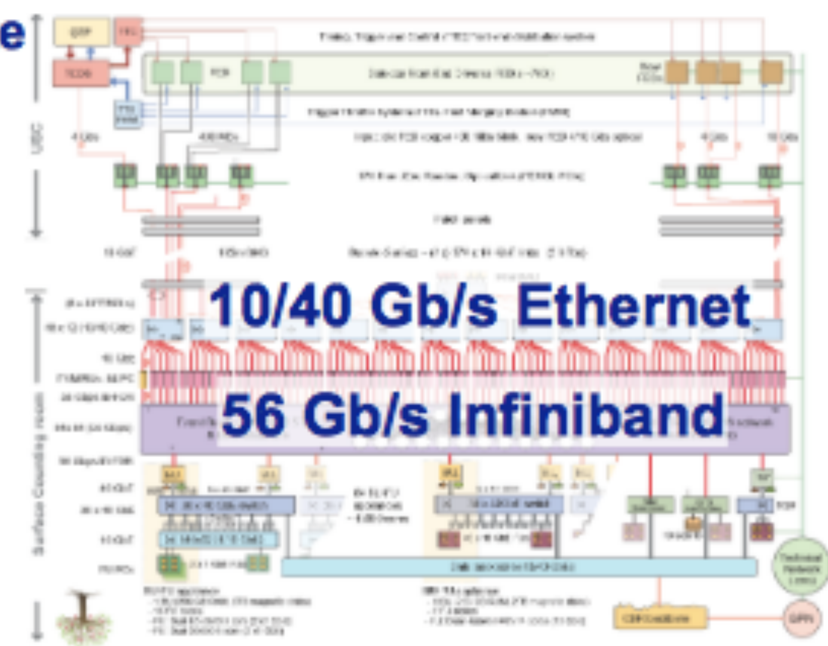
100
GB/s
8 slices

CMS DAQ 1

13000 core,
1260 host
filter farm

max. 1.2 GB/s to storage

100 kHz
L1 rate



~200
GB/s

1 slice

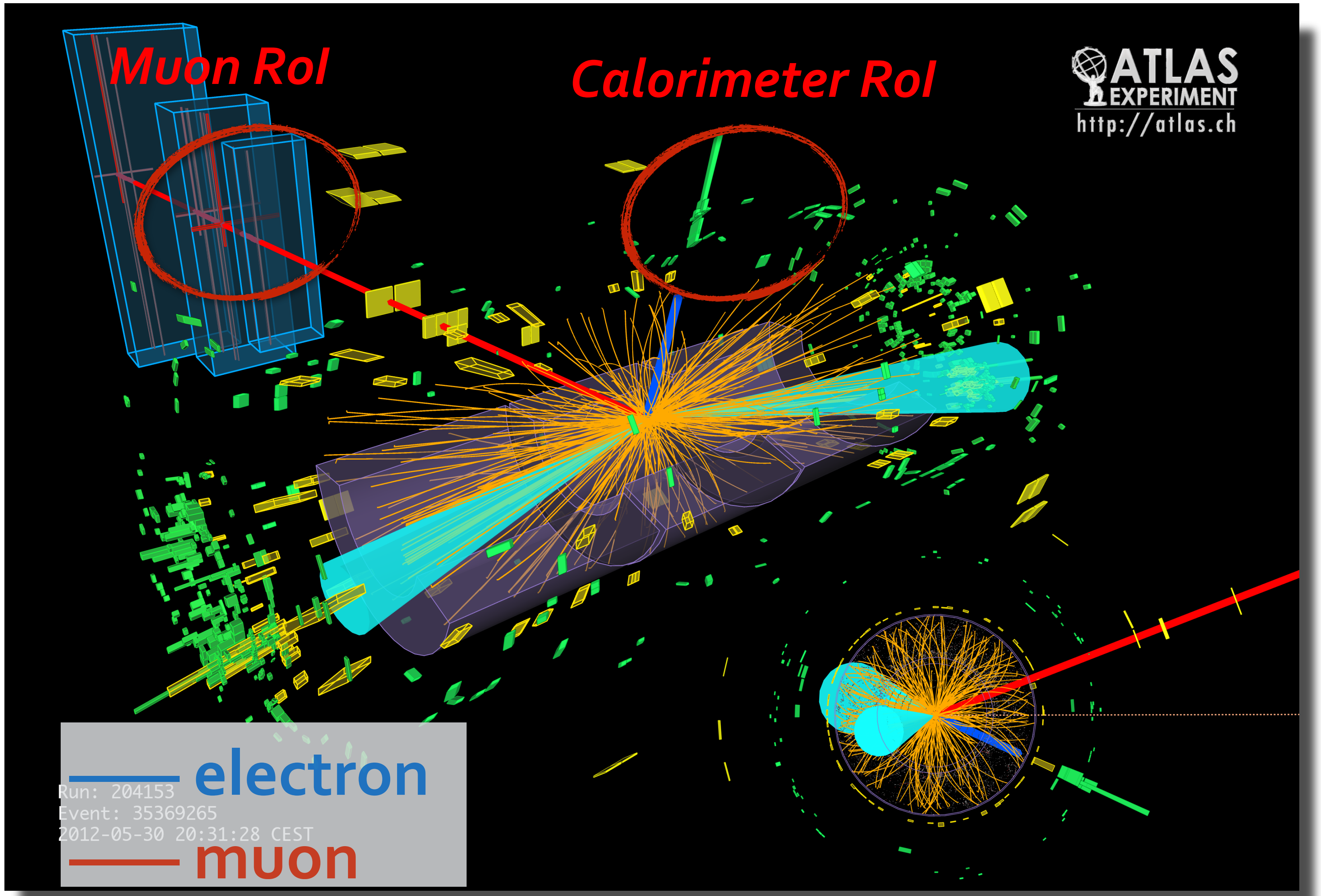
CMS DAQ 2

16000+ core,
900 host
filter farm

~ 3-6 GB/s to storage

ATLAS: REGION OF INTEREST (ROI) DATAFLOW

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



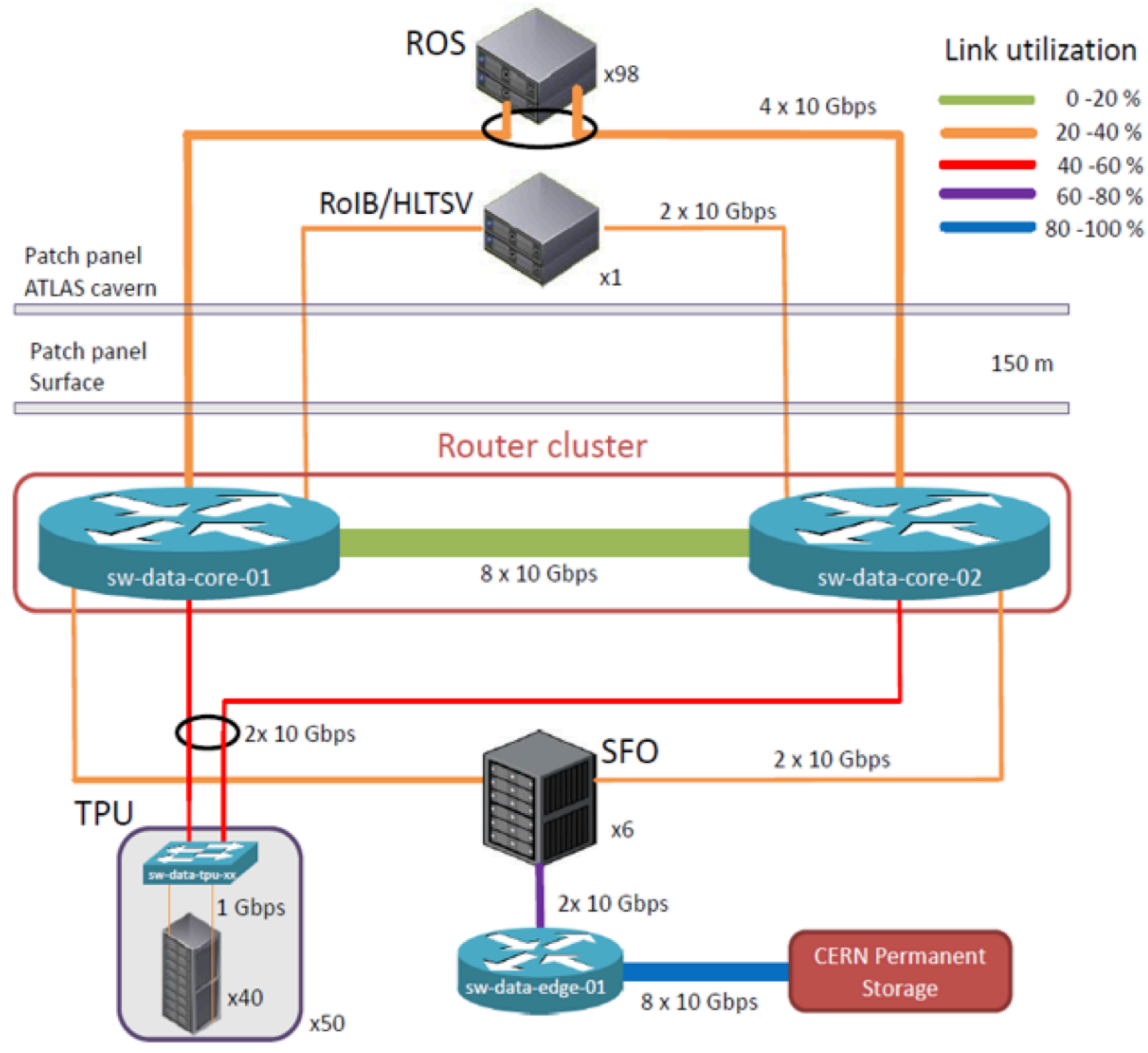
RoI=Region of Interest

- **Regional readout data is a few % of the total data @Level-1**
  - one order of magnitude smaller readout network …
  - … at the cost of a higher control traffic and reduced scalability

ATLAS REGIONAL TDAQ ARCHITECTURE

Run 2-Run 3

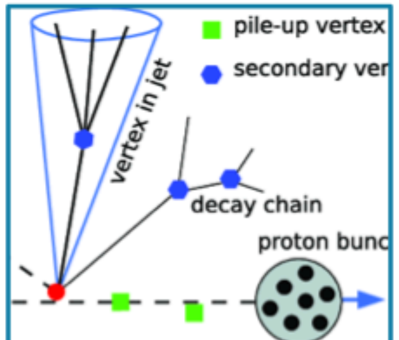
Overall network bandwidth: ~10 GB/s (instead of 100 GB/s due to regional readout)



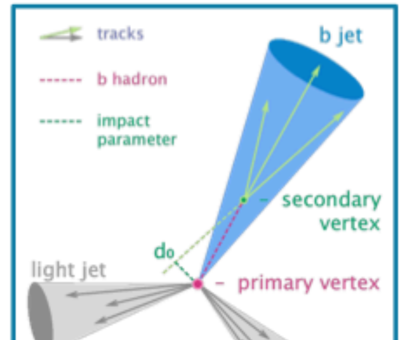
complex data router to forward different parts of the detector data, based on the trigger type

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

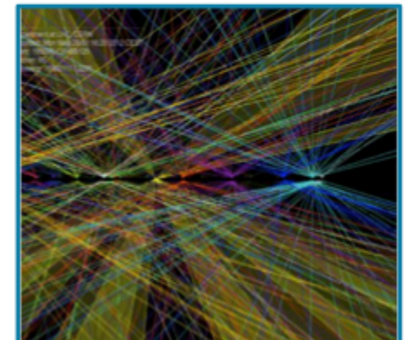
Silicon tracking systems provide incredibly high resolution, crucial for controlling rates

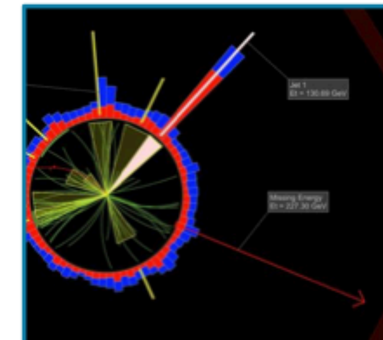


Vertex Reconstruction



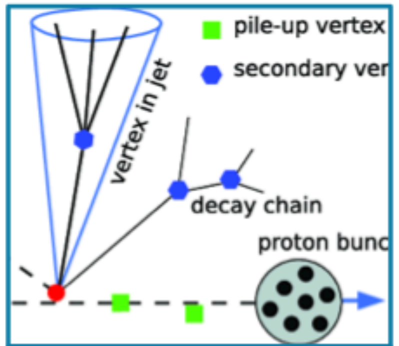
Jet Tagging


Pile-up Removal

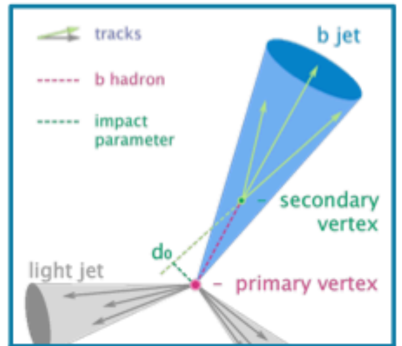

Missing Energy Reconstruction

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

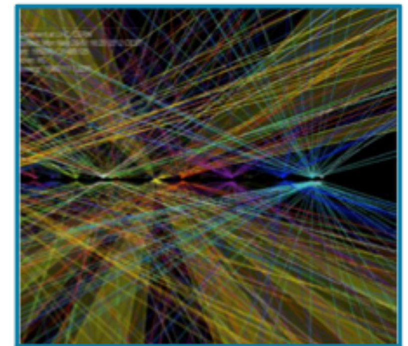
Silicon tracking systems provide incredibly high resolution, crucial for controlling rates



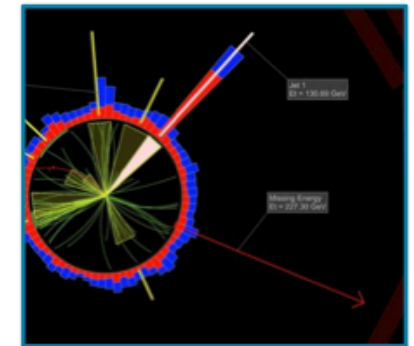
Vertex Reconstruction



Jet Tagging



Pile-up Removal



Missing Energy Reconstruction

Tracking challenges

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

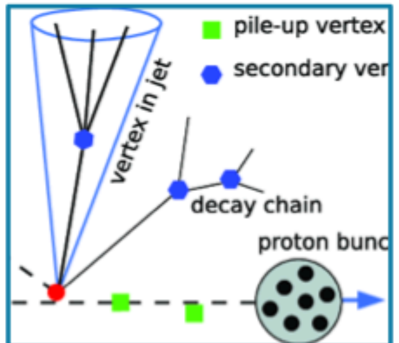
combinatorics scales like L^N

L=luminosity, N=number of layers

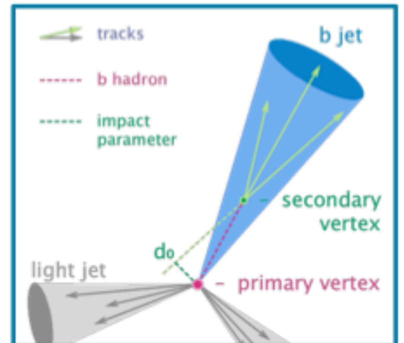
Tracking reconstruction not feasible @40MHz, nor in few microseconds

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

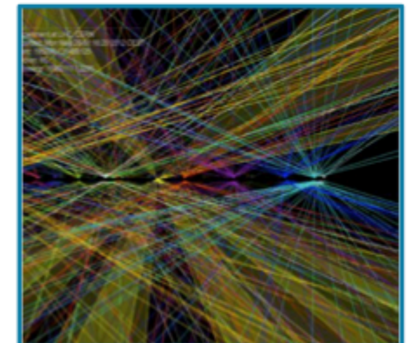
Silicon tracking systems provide incredibly high resolution, crucial for controlling rates



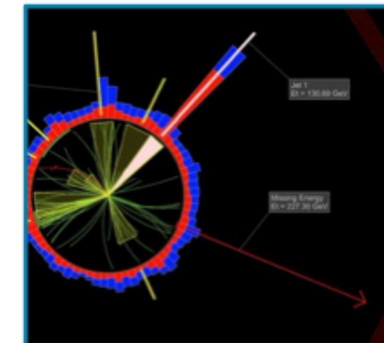
Vertex Reconstruction



Jet Tagging



Pile-up Removal



Missing Energy Reconstruction

Tracking challenges

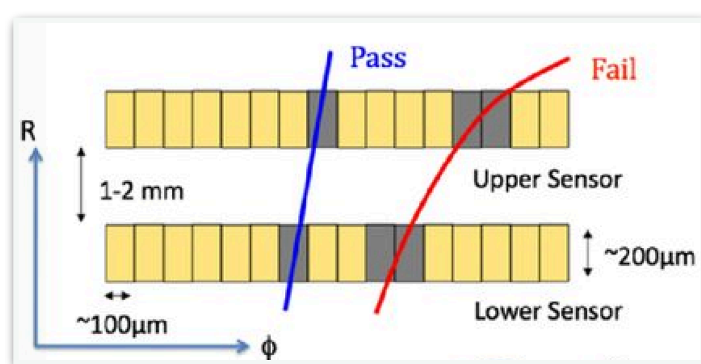
- Readout ~800M channels, ~50 Tbps
- Combinatorics (10^4 hits/BC)

combinatorics scales like L^N

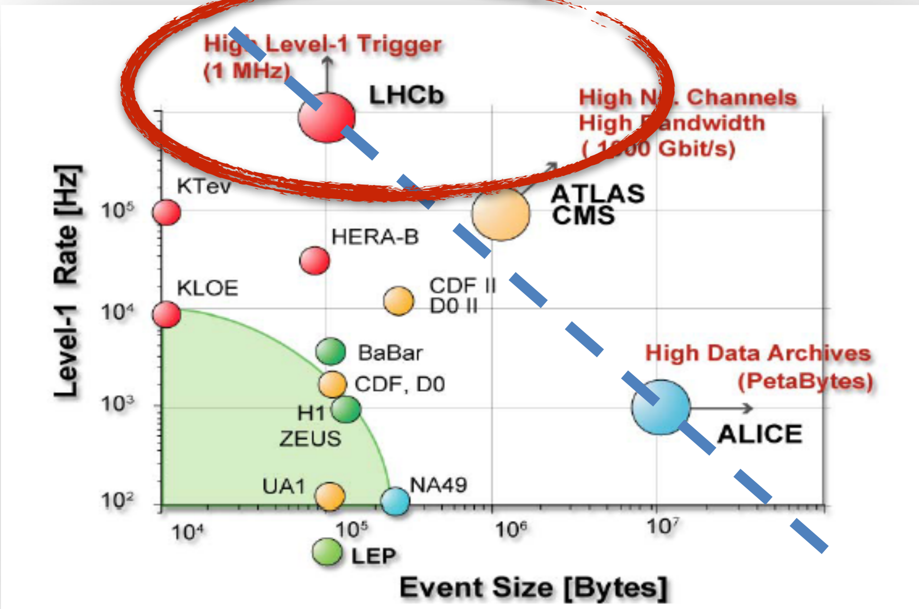
L=luminosity, N=number of layers

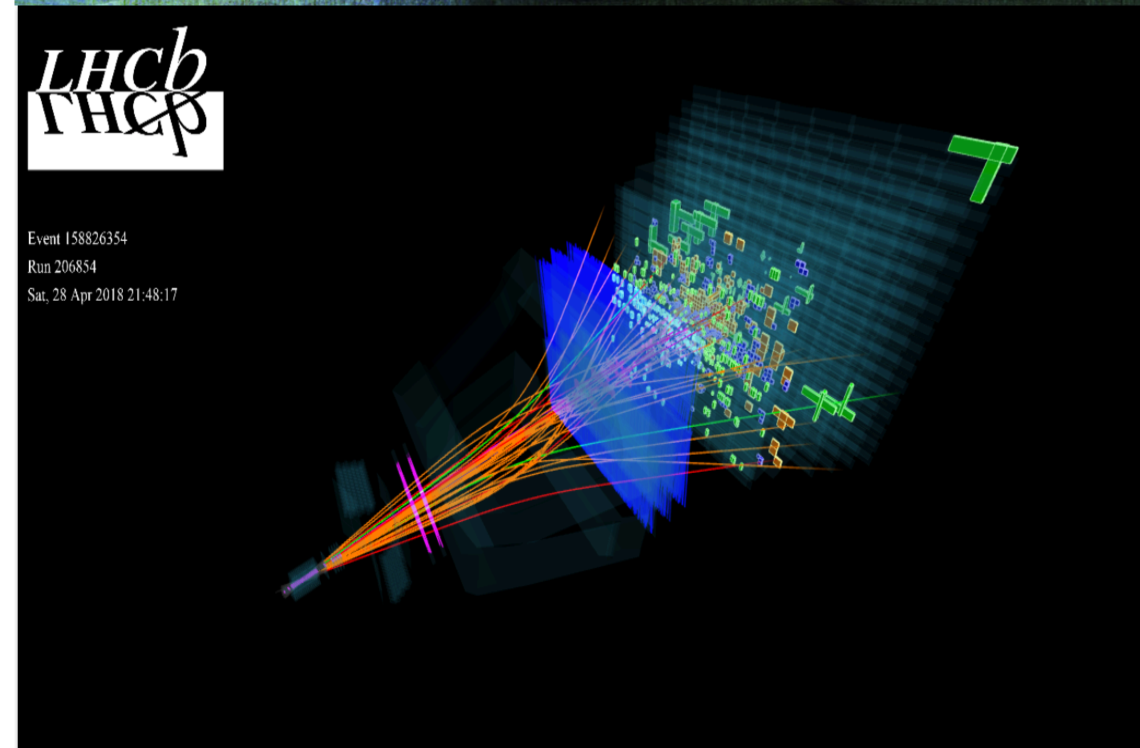
Tracking reconstruction not feasible @40MHz, nor in few microseconds

	ATLAS [1]	**CMS [2]**
data reduction @40 MHz	regions @L1 (RoIs)	h/w coincidences (stubs) @L1
fast tracking @1 MHz	algorithms on FPGAs and/or GPUs	
precision tracking @100 kHz	optimized, as offline	



stubs in CMS PT modules



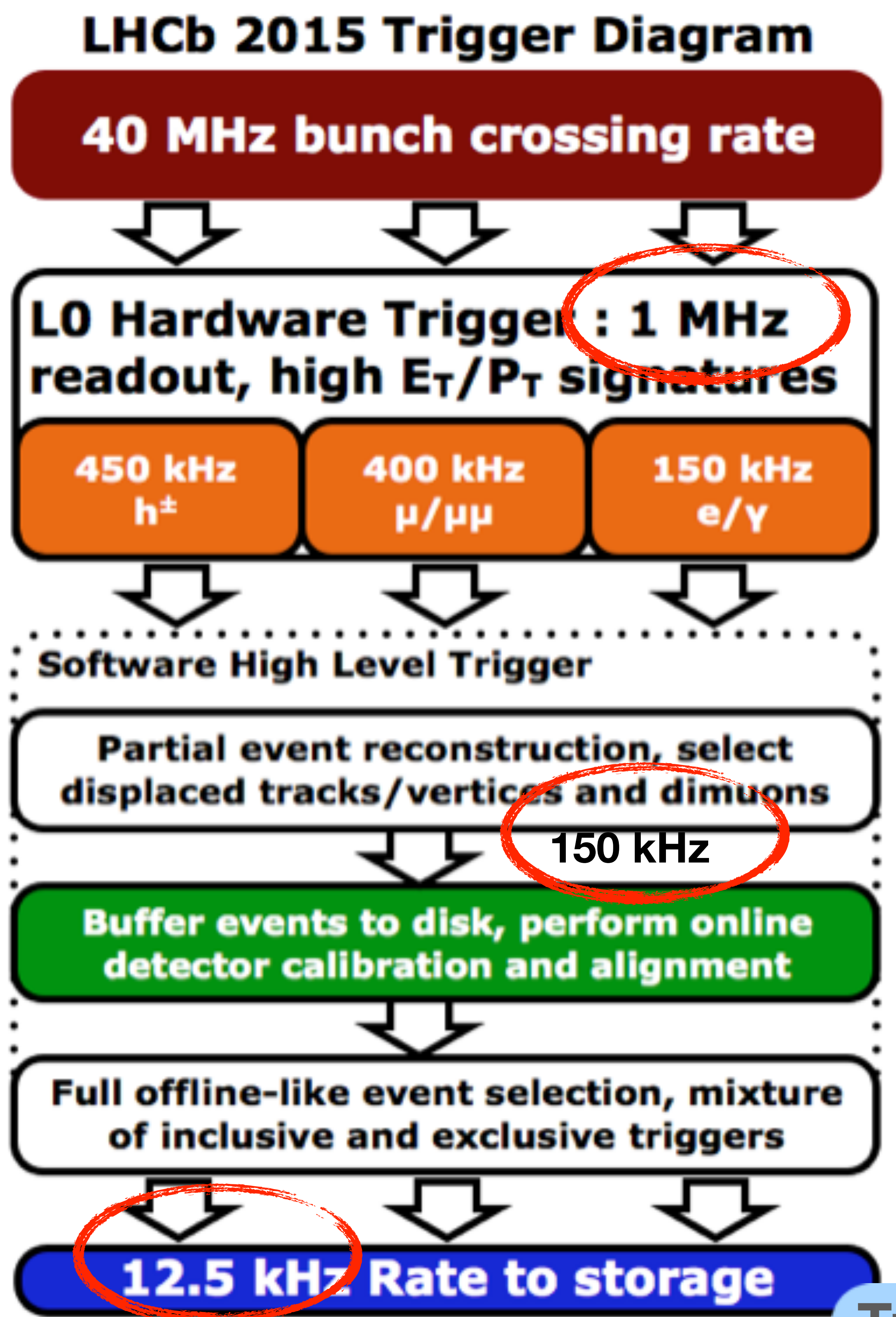


LHCb, THE B-MESON OBSERVATORY

The lightest experiment to study the heavy b-quark

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

LHCB TRIGGER STRATEGY



small event size @ 10 MHz

- Limited Luminosity ==> 2 x 10^{32} $cm^{-2}s^{-1}$

L0 hardware trigger @ 1MHz

- Select B hadrons
- Reject complex/busy events

60 kB x 1 MHz = 60 GB/s readout network

Software HL trigger in two stages

Multitude of exclusive B-selections

Split in 2 stages with large buffer (4PB)!
(3000 hard-disks, enough for days)

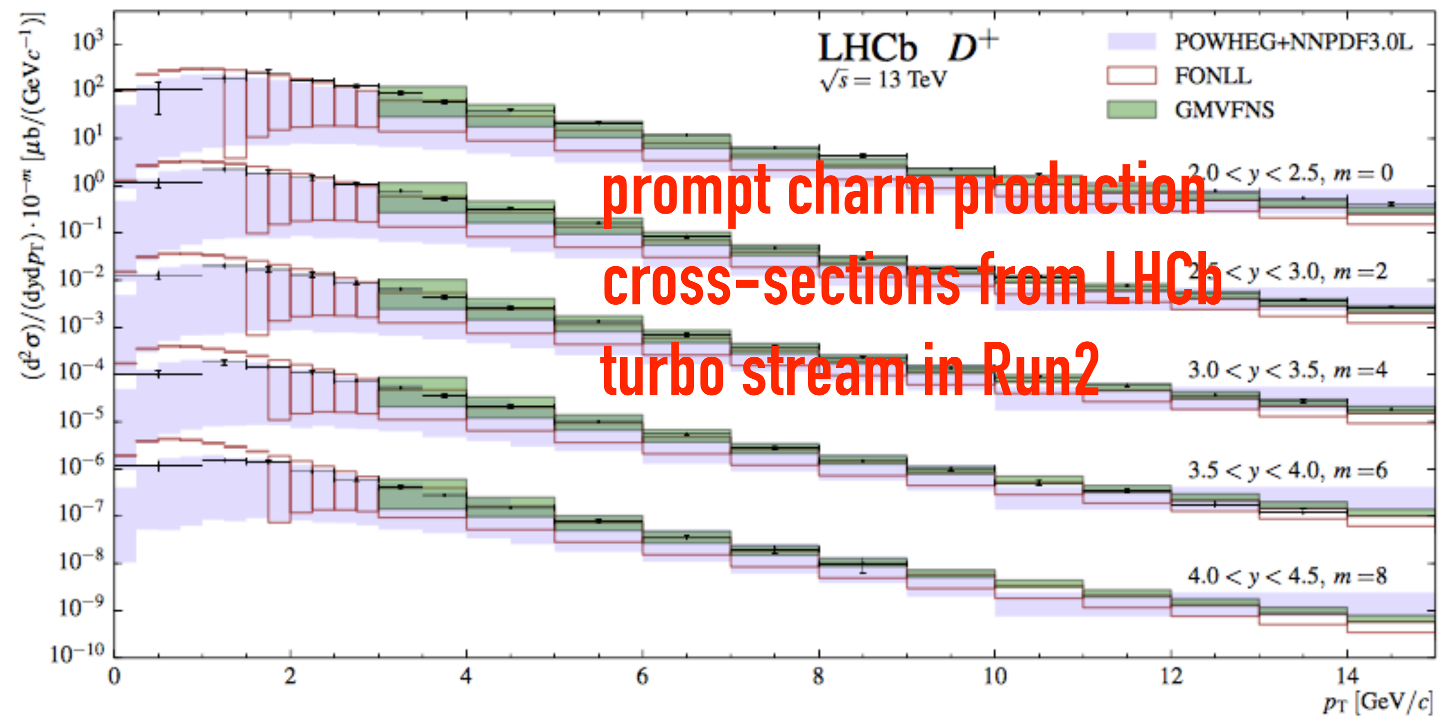
HLT-1: Synchronous with DAQ - GPUs @100 kHz
Fast tracks for B-decay vertices (in 35 ms)

HLT-2: Deferred Processing
Reconstruct with real-time calibrations and alignments (in 350 ms)

Trigger becomes a real-time physics analysis

A NEW TREND: REAL TIME ANALYSIS

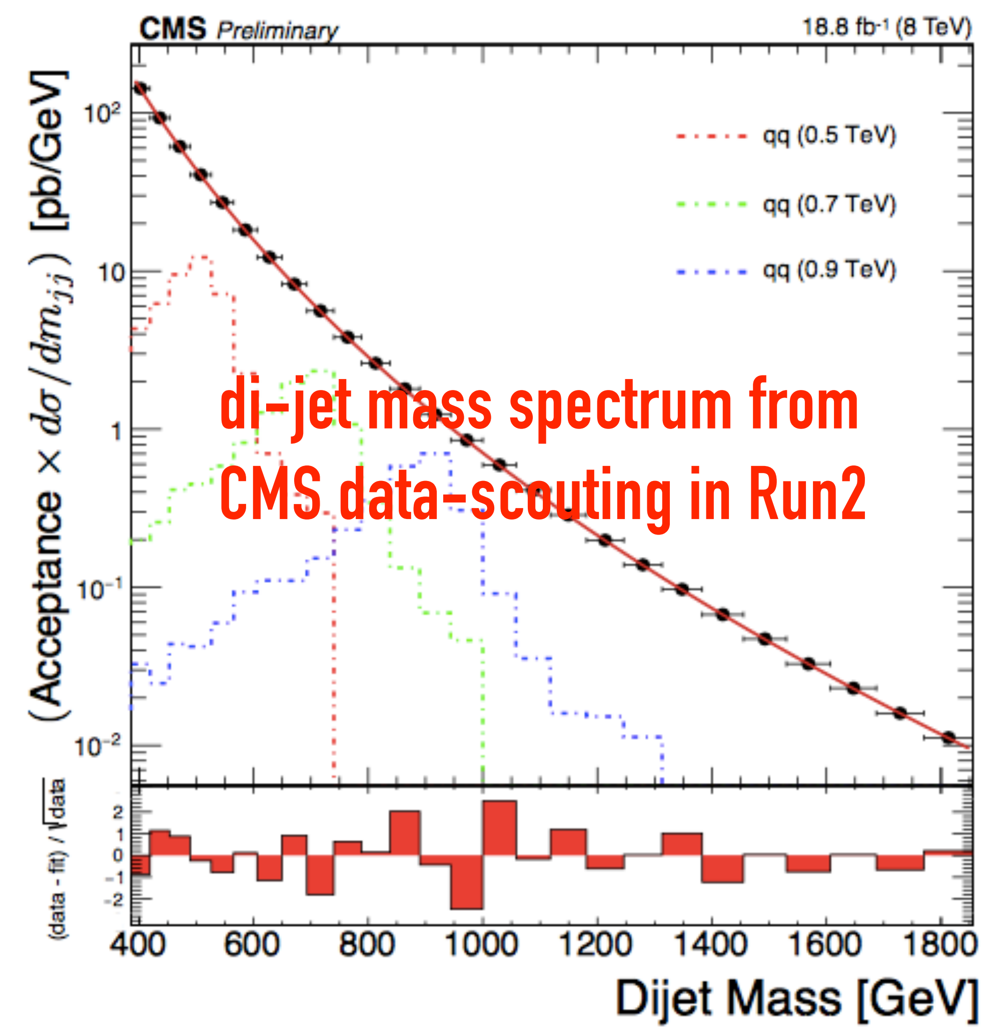
Can we get rid of FrontEnd raw data?



prompt charm production cross-sections from LHCb turbo stream in Run2

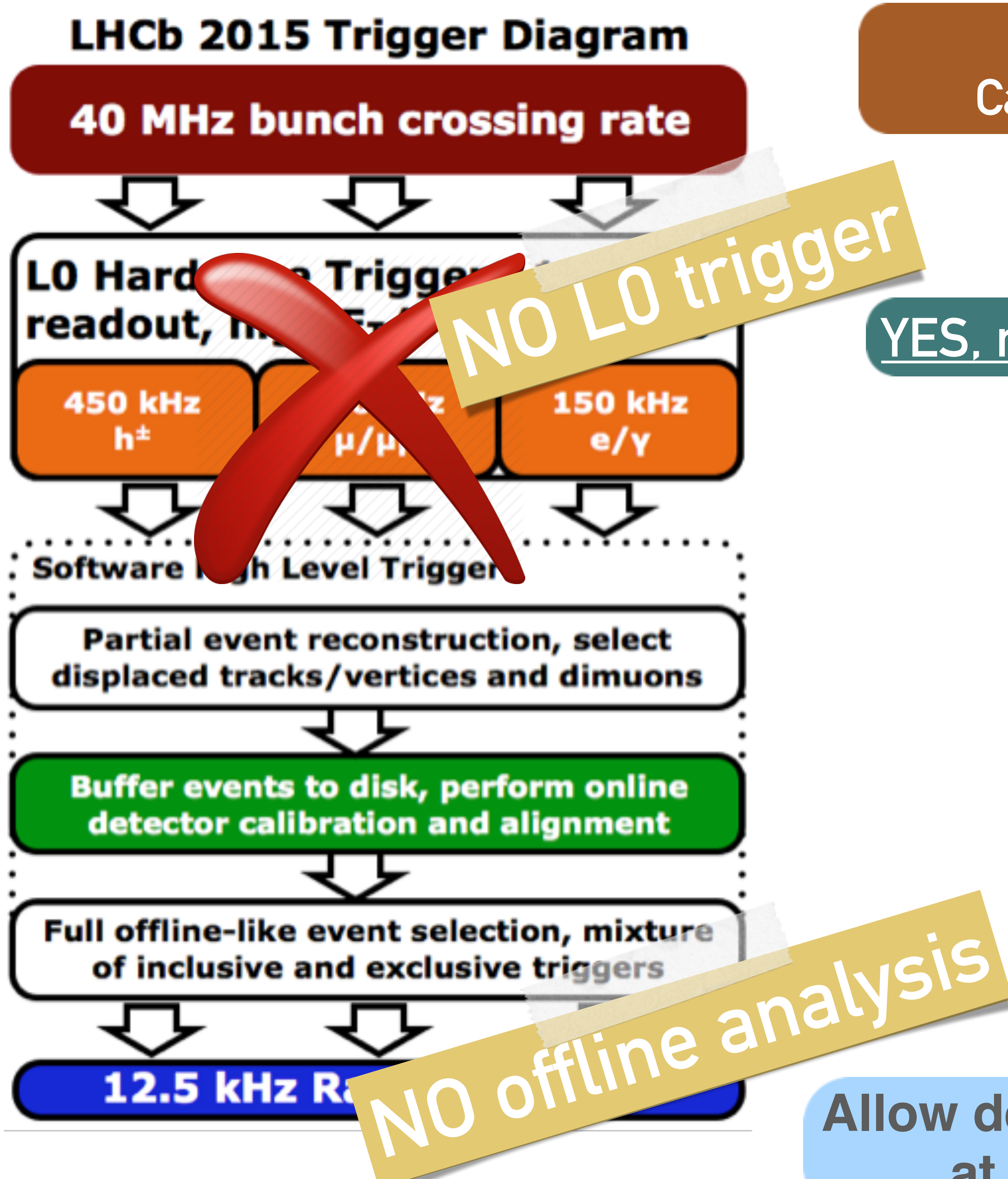
- ➡ **Event size/10 -> rate x 10 for free**
- ➡ **Adopted by all experiments:**
  - ➡ Full online reconstruction (**LHCb/ALICE**)
  - ➡ On dedicated data streams (**ATLAS/CMS**)
    - ➡ for some high rate signatures, save only reduced information

turbo-stream
data-scouting
Trigger-Level-Analysis



di-jet mass spectrum from CMS data-scouting in Run2

UPGRADES FOR RUN 3

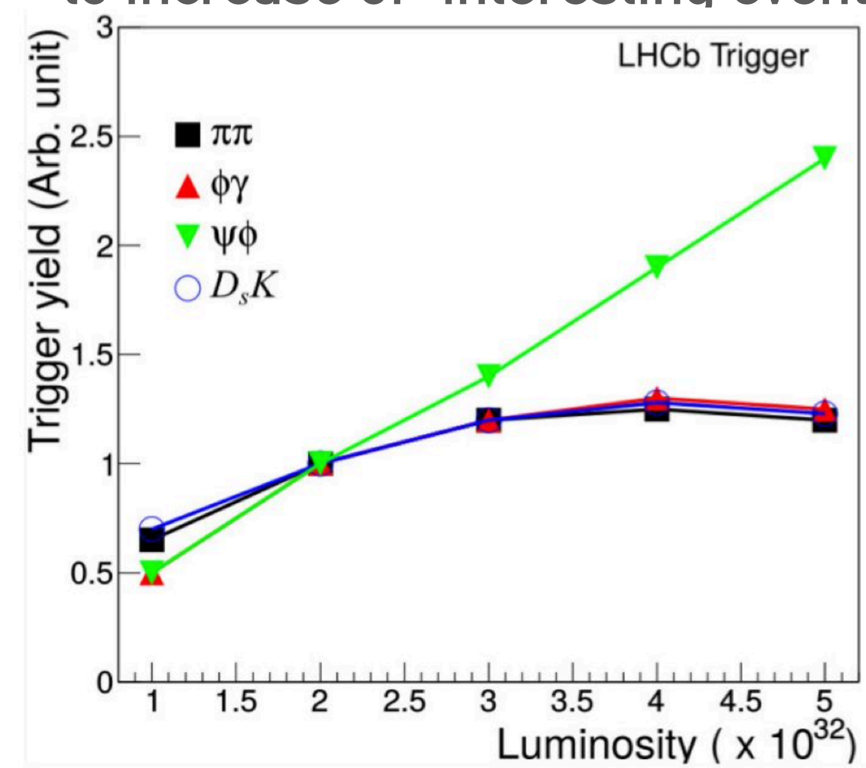


See Phase-I upgrade TDR

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?

30 MHz inelastic event rate (full rate event building)

40Tbit/s

30 MHz

Software High Level Trigger

Full event reconstruction, inclusive and exclusive kinematic/geometric selections — HLT-1

1-2 Tbit/s

1MHz

Buffer events to disk, perform online detector calibration and alignment

Add offline precision particle identification and track quality information to selections

Output full event information for inclusive triggers, trigger candidates and related primary vertices for exclusive triggers — HLT-2

2-5 GB/s to storage

50 kHz

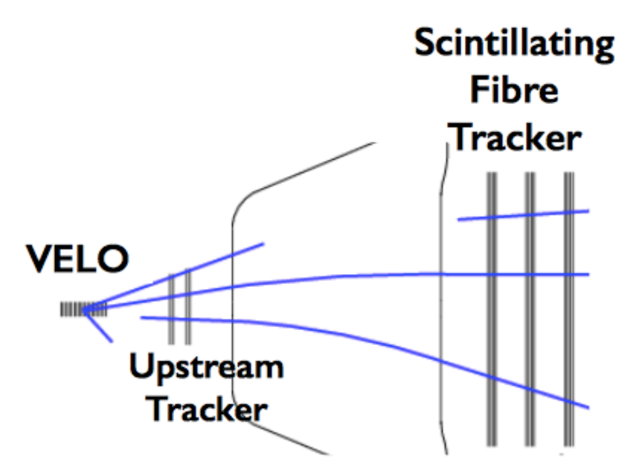
80 Gbit/s

FE readout & Event Building at 30 MHz (~40 Tbit/s)

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)



Online Tracking

- Velo tracking
- Velo-UT tracking (p_T > 200 MeV, $\delta p/p \sim 15\%$)
- Forward tracking (p_T > 500 MeV, $\delta p/p \sim 0.5\%$)
- PV finding
- Rate reducing cuts (Output < 1 MHz)
- Muon Identification
- Simplified Kalman fit
- Particle Identification

arXiv:2105.04031

HOW TO LIVE WELL WITHOUT A L1 TRIGGER

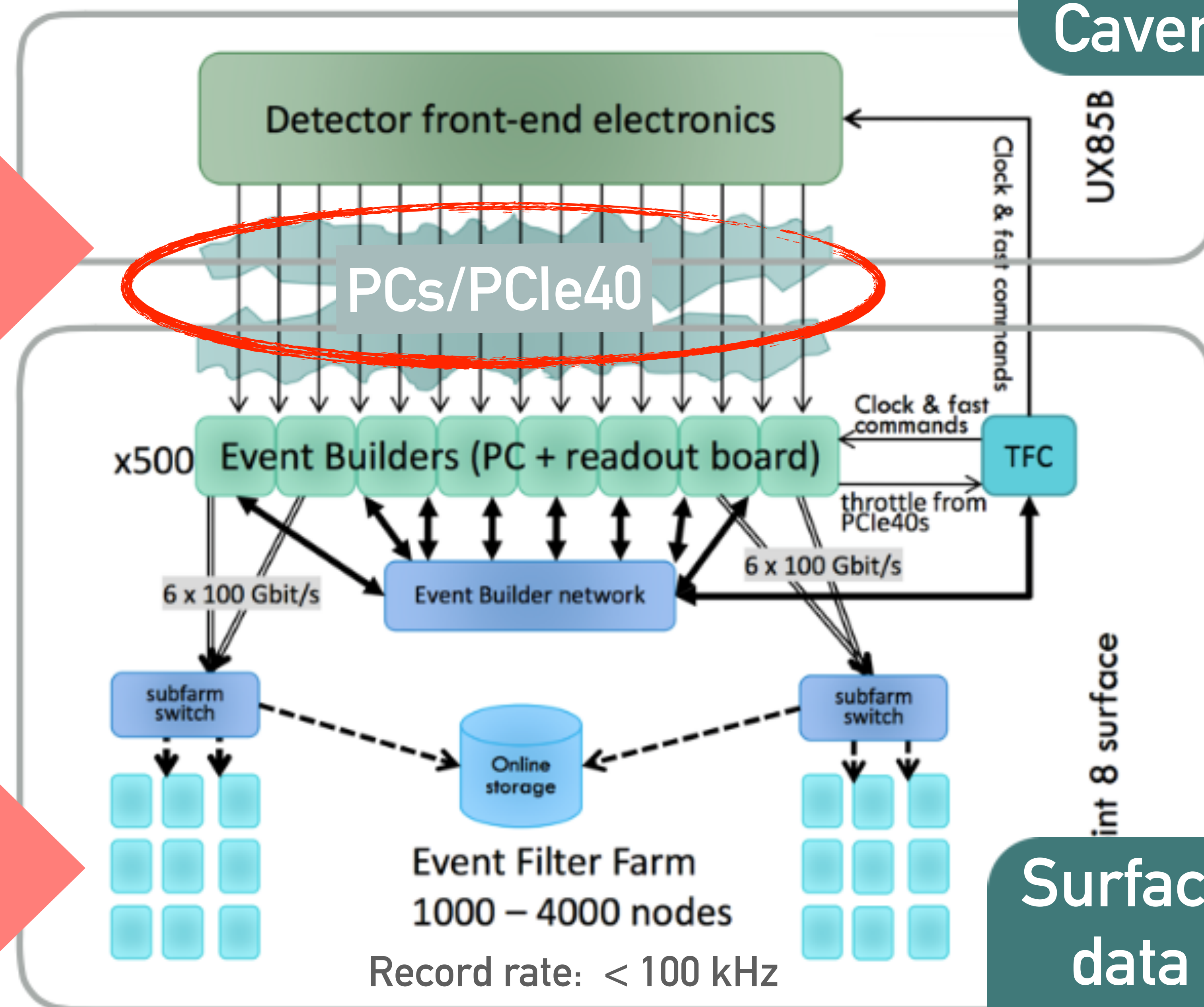
150 kB x 30 MHz = 40 Tbps

Readout @ 30 MHz
Event size ~ 150kB

- **Data reduction:**
  - Custom FPGA-card (**PCIe40**)
  - Data-packing for sub-detectors (zero-suppression, clustering)
- **Massive link usage:**
  - ~10,000 GBT (4.8 Gb/s, rad-hard)

DAQ network < 40 Tbps
Record rate: < 100 kHz

Inside Cavern

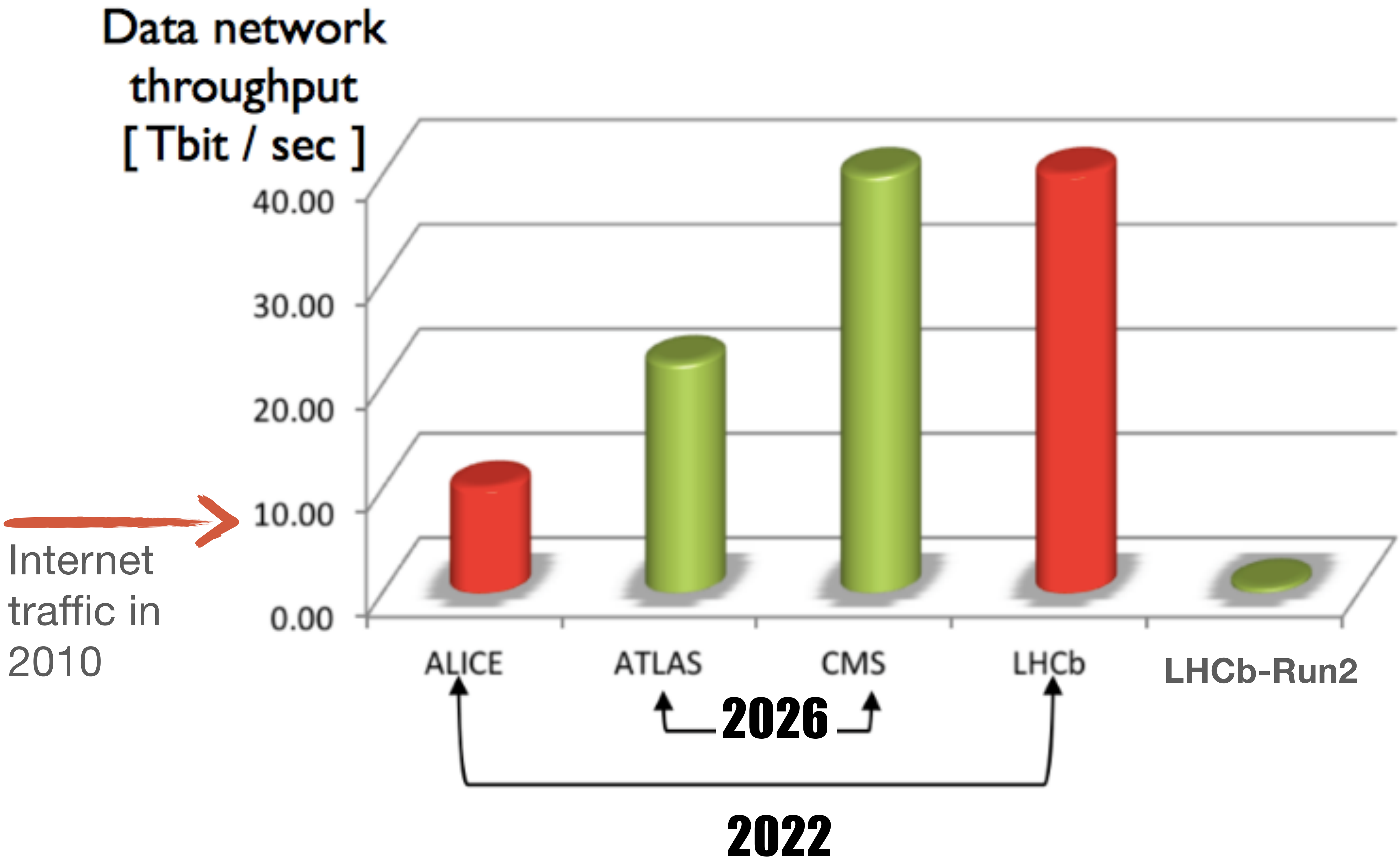


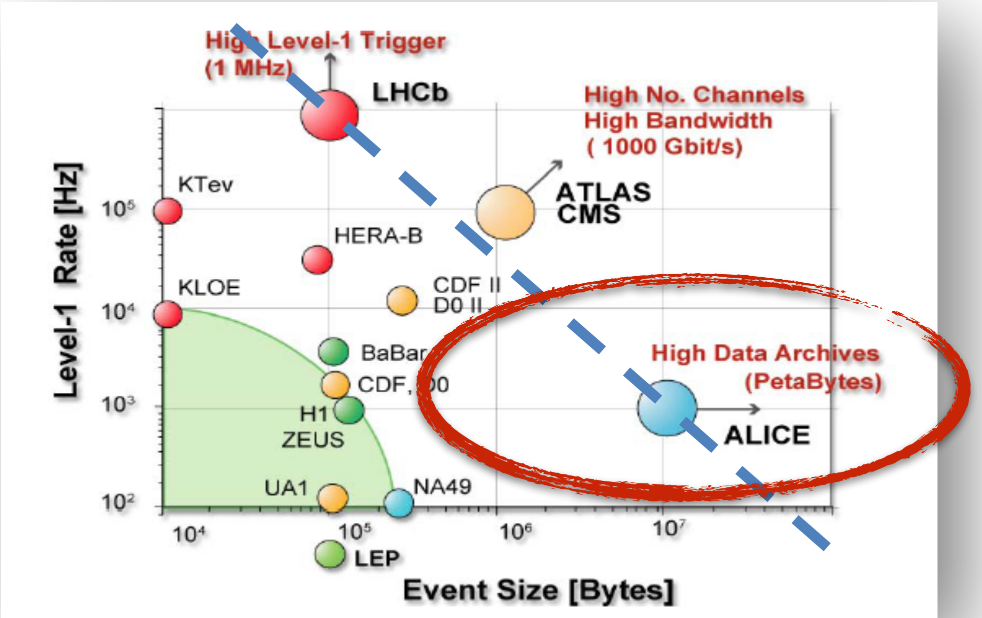
Surface data centre

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

Ref for PCIe40

NETWORK TRAFFIC COMPARISON

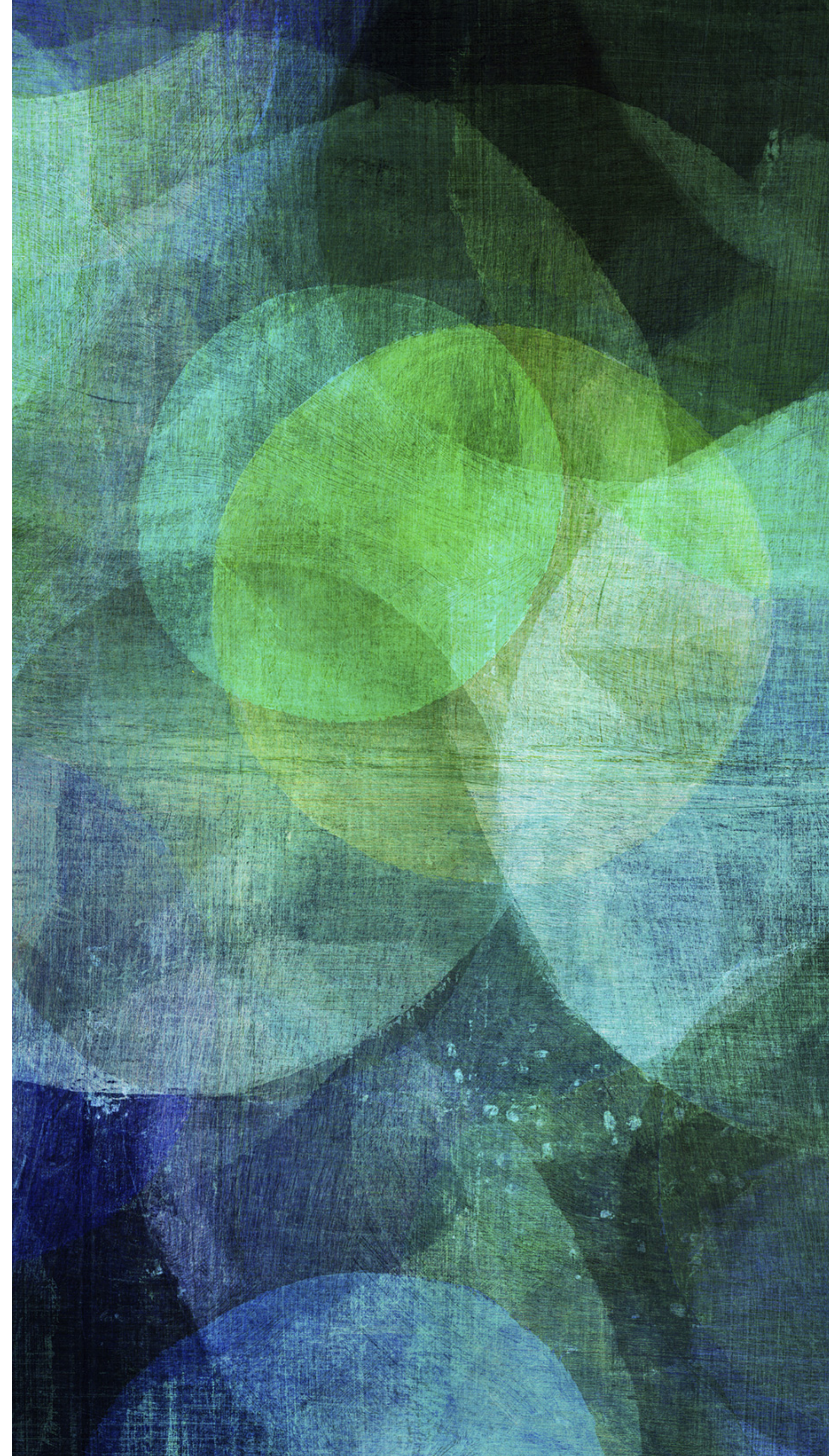




ALICE: THE SMALL BIG-BANG

Recording heavy ion collisions

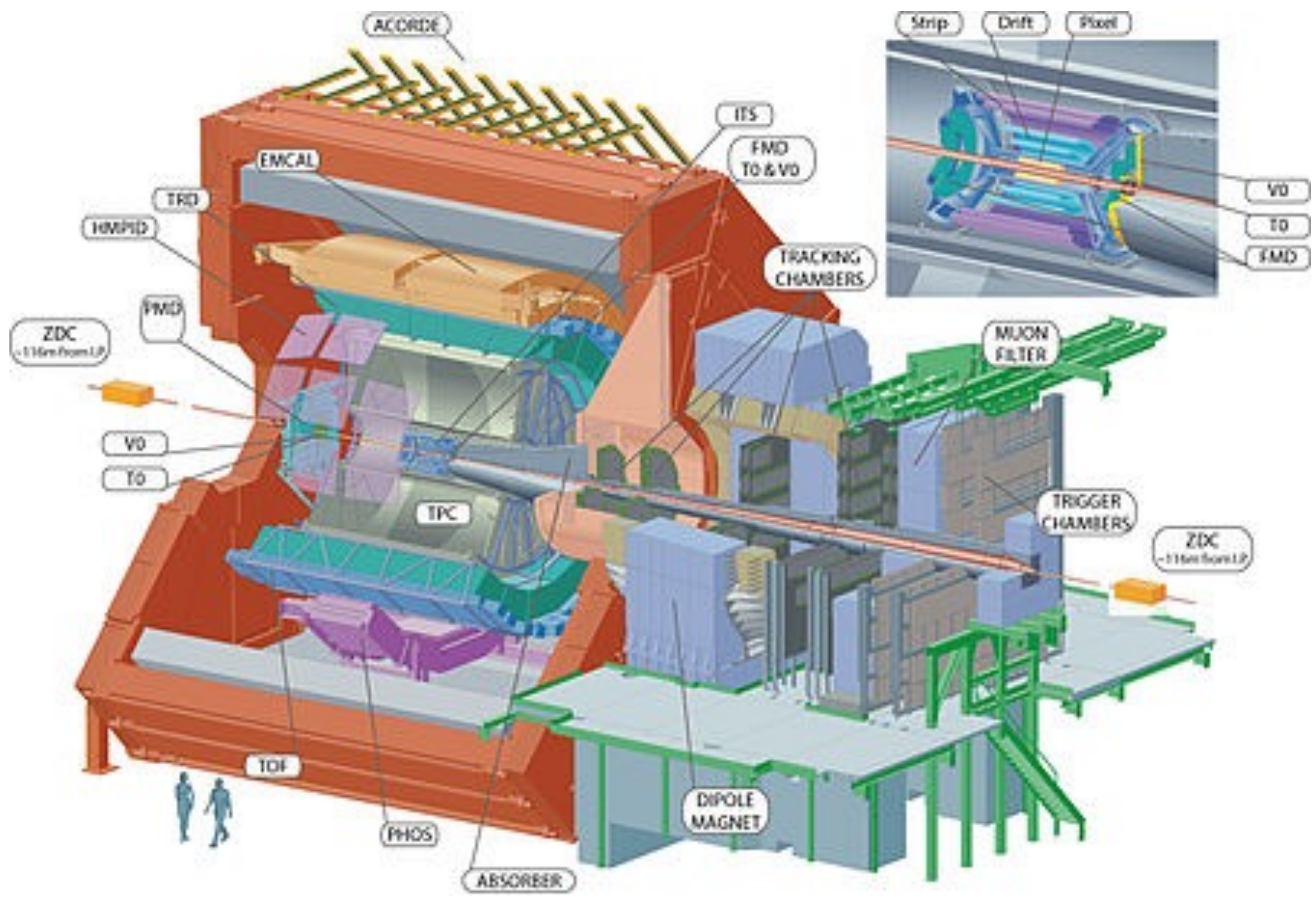
http://alice-daq.web.cern.ch



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 MB)**
  - TPC producing 90% of data
- **Complex event topology**
  - low trigger rate: max 3.5 kHz

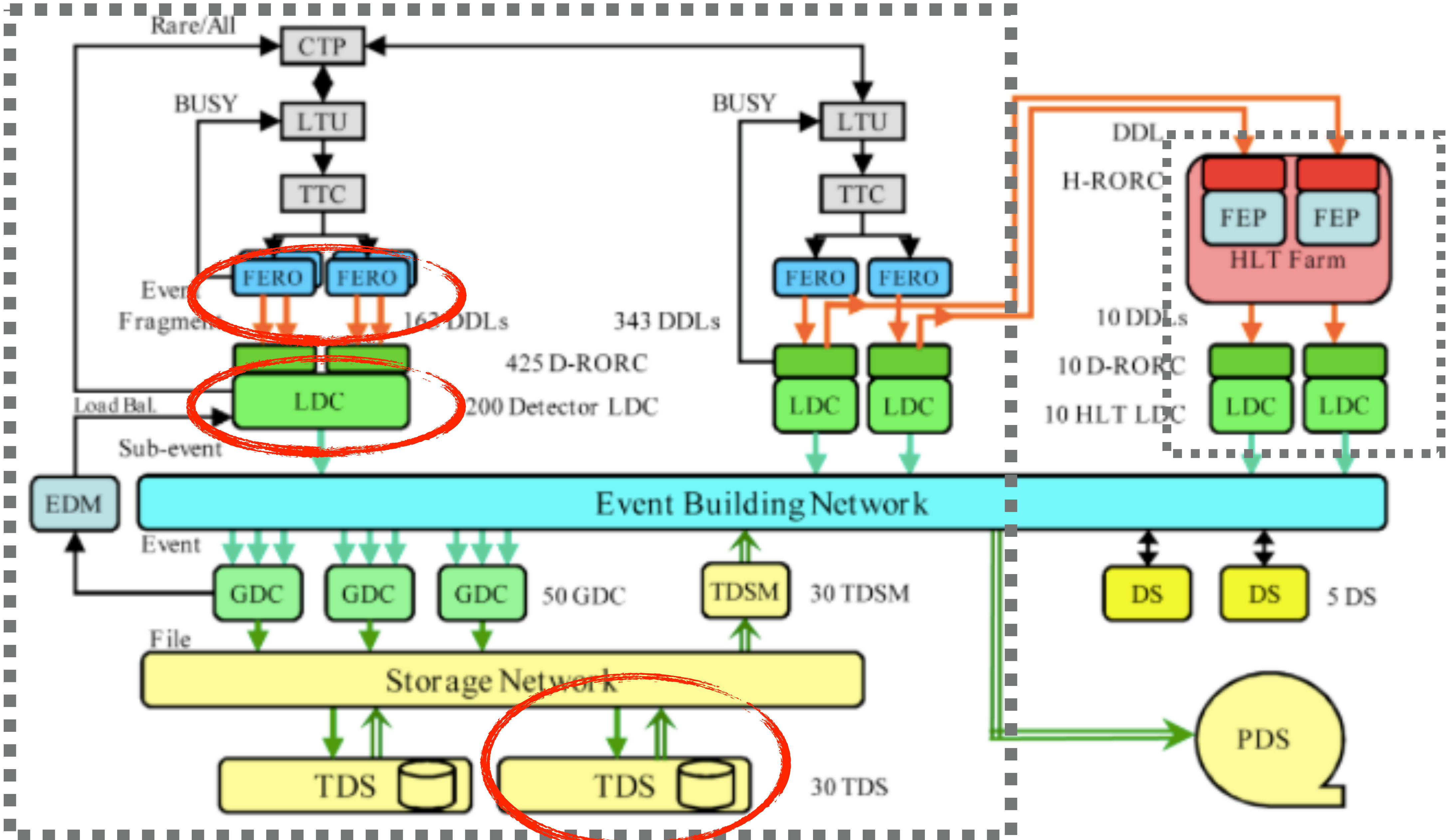


cms = 5.5 TeV per nucleon pair
Pb–Pb collisions at L $=10^{27}$ cm^{-2}s^{-1}

➡Challenges for TDAQ design:

- detector readout: up to ~50 GB/s
- storage: 1.2 TB/s (Pb-Pb)

READOUT DATA CONCENTRATORS



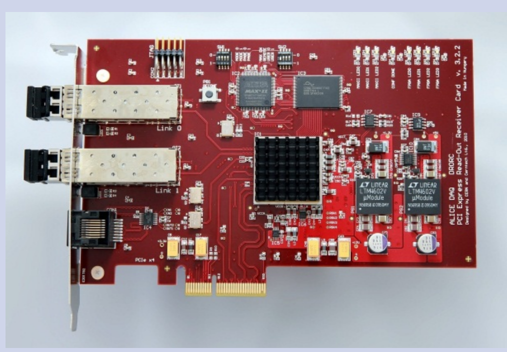
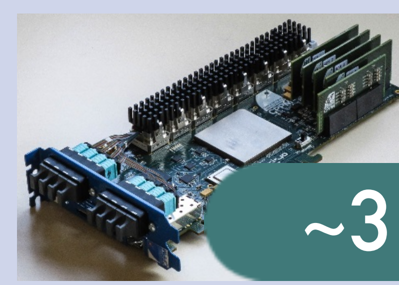
- ➡ **Dataflow with local (LDC) and global (GDC) data concentrators**
  - ➡ Detector readout (~20 GB/s) with point-to-point optical links (DDL, max 6Gb/s)
  - ➡ Rate to the LDCs can go above 13 GB/s
- ➡ **Transient Data Storage (TDS)**
  - ➡ Before the Permanent Data Storage (PDS) and publish via the Grid

PREPARING FOR RUN 3

ALICE

- ➡ **LHC heavy ion programme extended the 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

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

~3 TB/s detector readout

Run 1 → LS1 → Run 2 → LS 2 → Run 3

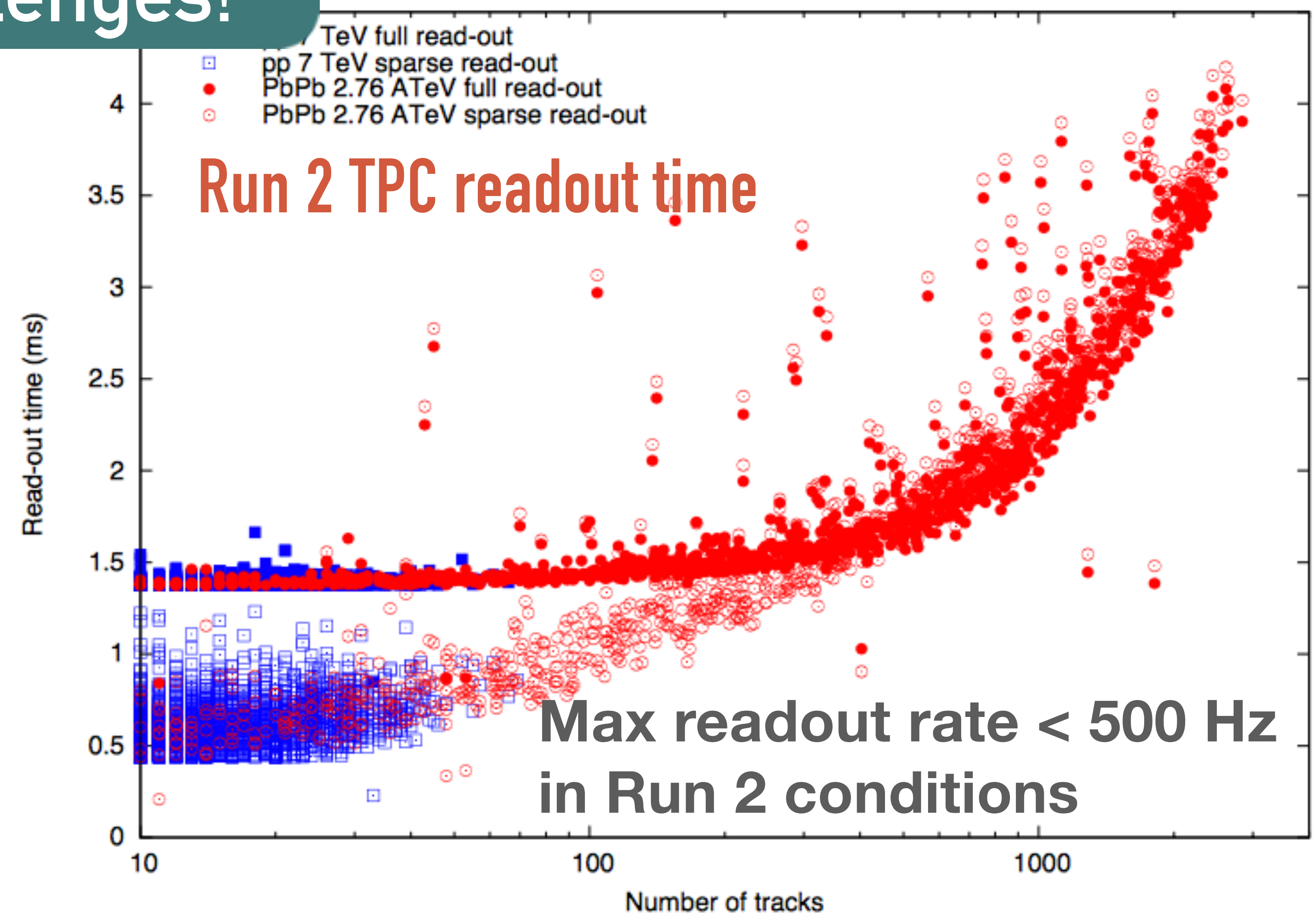
New Common Readout Unit (CRU), based on PCIe40 card

ALICE

- ➡ **LHC heavy ion programme extended the 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

New TDAQ challenges!

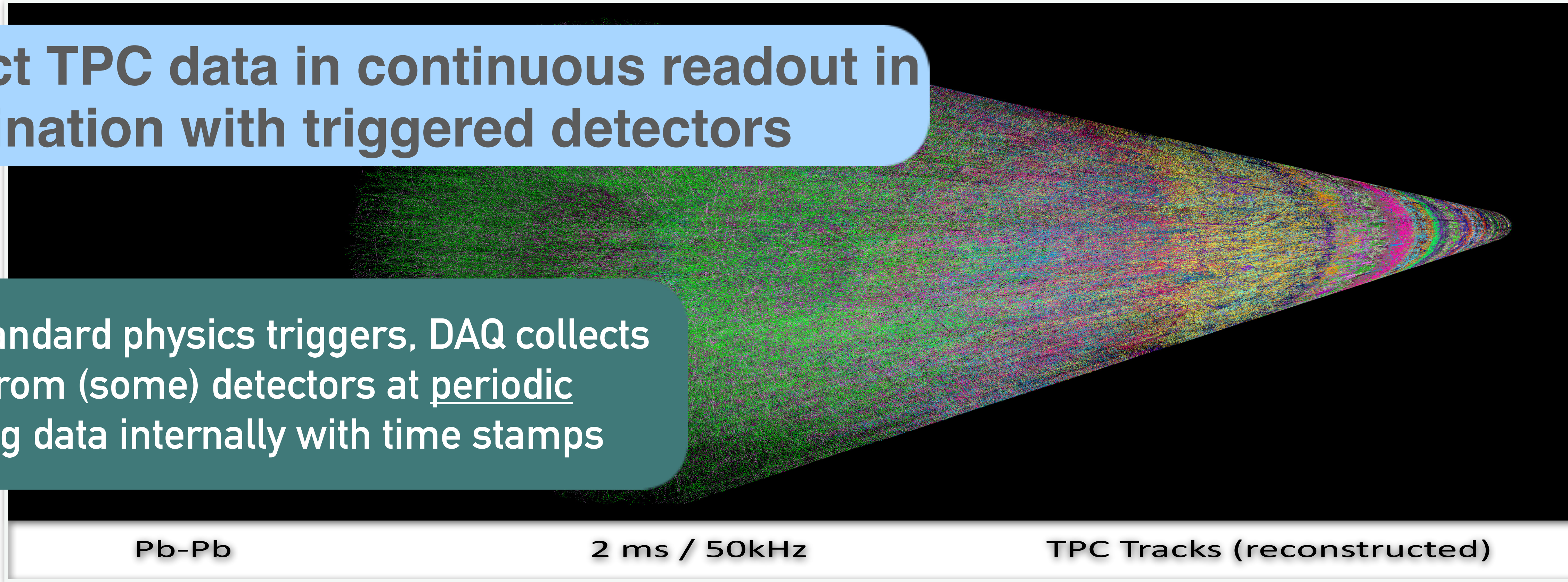
- ➡ **Need new and faster electronics**
  - ➡ in particular TPC readout with GEM, no gate
- ➡ **The readout rate is very close to TPC readout !!**



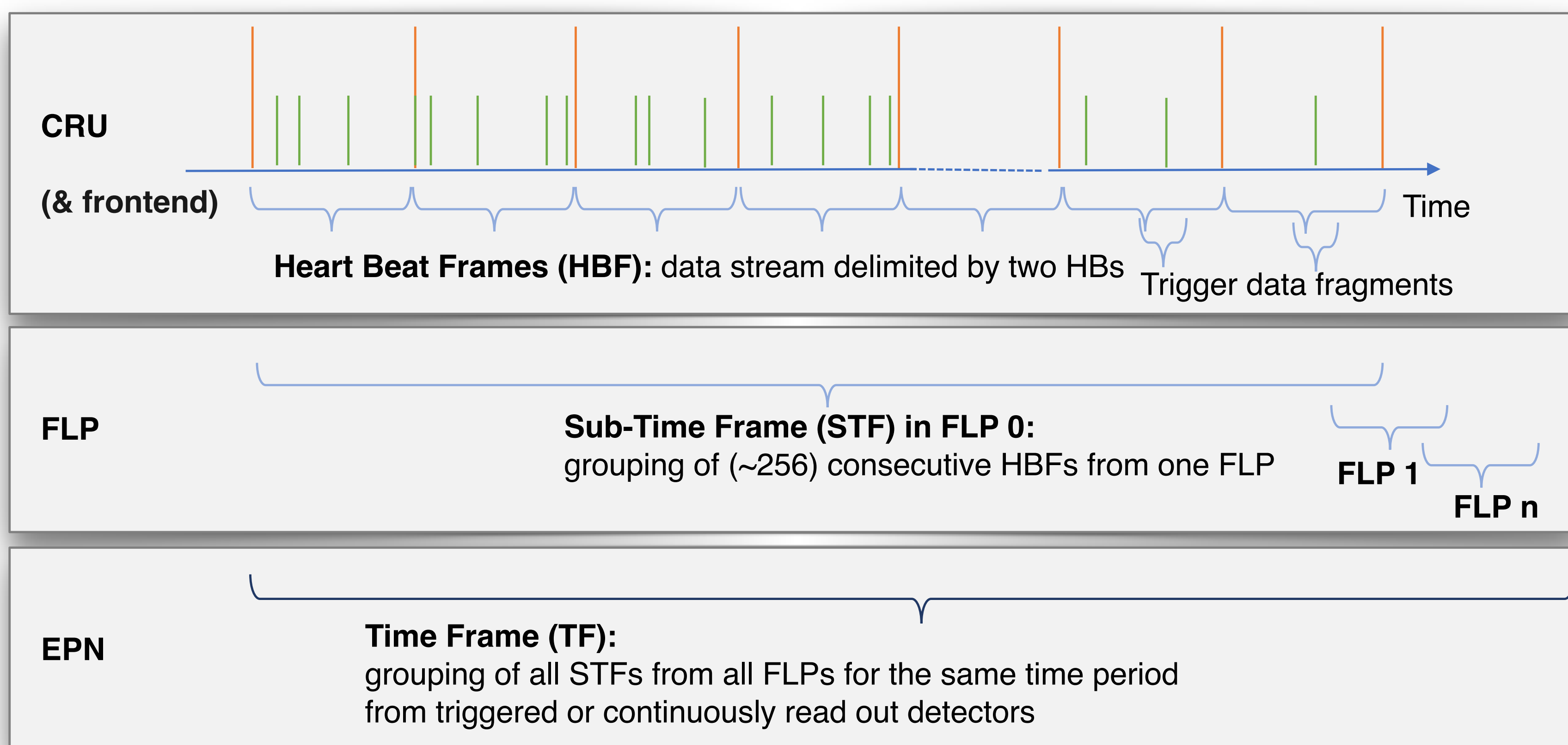
CONTINUOUS READOUT FOR RUN 3

Reconstruct TPC data in continuous readout in combination with triggered detectors

In addition to standard physics triggers, DAQ collects frames of data from (some) detectors at <u>periodic intervals</u>, tagging data internally with time stamps



- **Heart Beat (HB) issued in continuous & triggered modes**
  - group data into time intervals to allow synchronisation between different detectors
  - 1 per LHC orbit, 89.4 μs: <u>~10 kHz</u>
- **Grouped in Time-Frames:**
  - 1 every ~20 ms: **<u>~50 Hz</u>** (1 TF = ~256 HBF)



RUN 3 DAQ: ONLINE RECONSTRUCTION

ALICE

Higher rates with smaller data? → **Store reconstruction, discard raw data**

Very heterogeneous system

- **Synchronous, continuous readout**
  - Data compression in **FPGA**/**CPU**
  - gain x2 readout rate

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

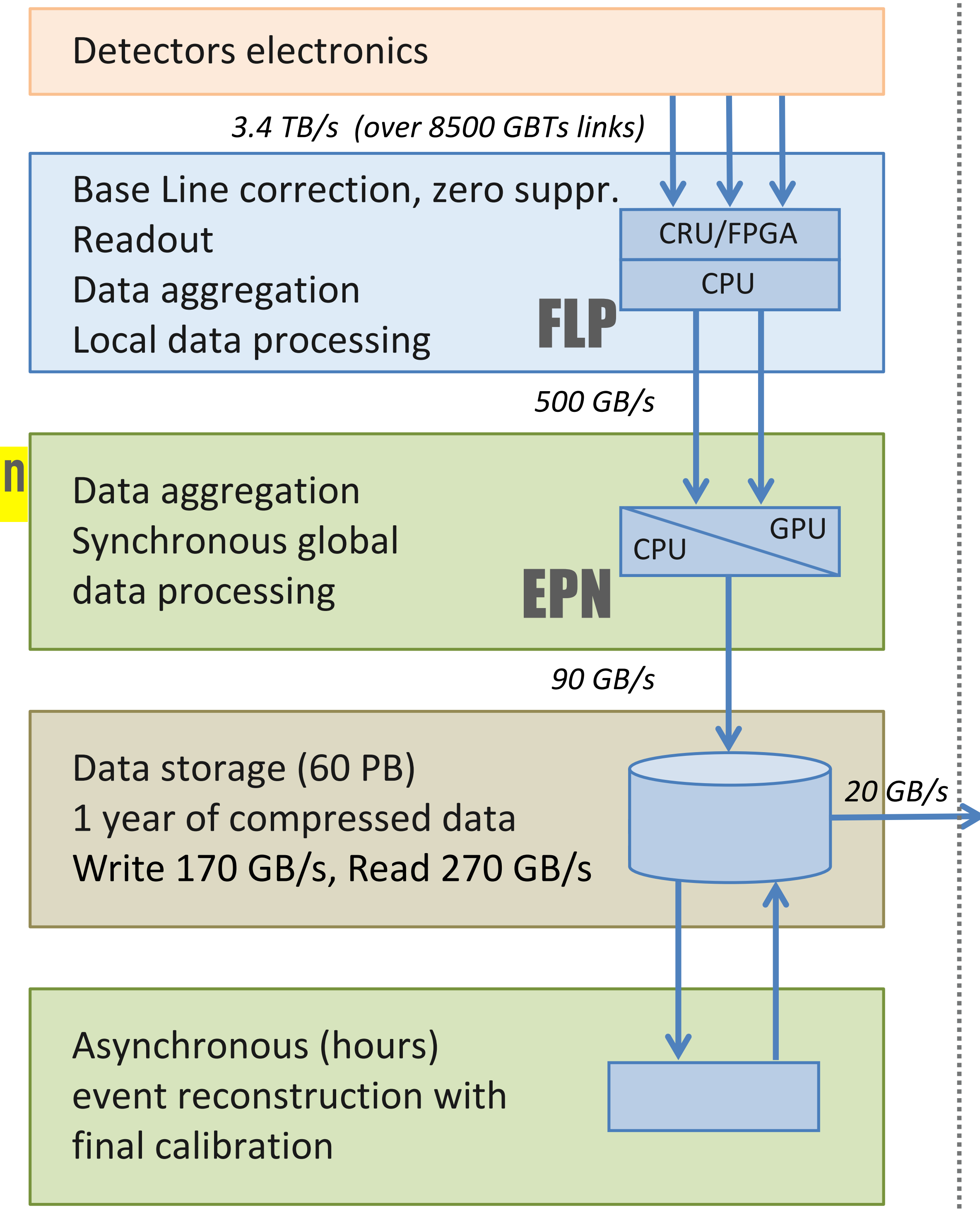
O^2 system

- **Common online/offline software**
  - Same calibrations and resources

Data reduction
Calibration 0

Data aggregation
Reconstruction
Calibration 1

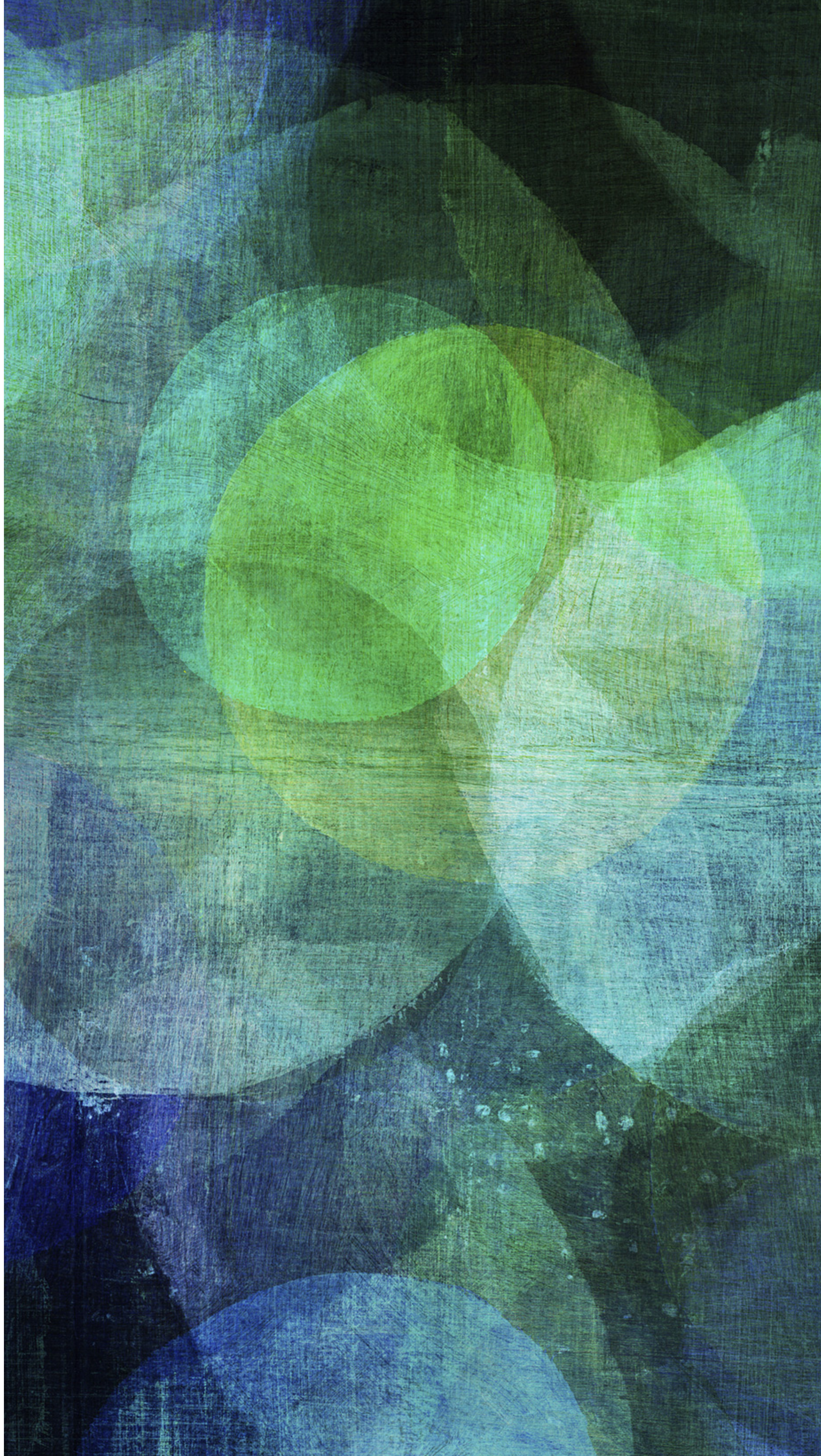
More reconstruction
Calibration 2



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 comodity 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…")

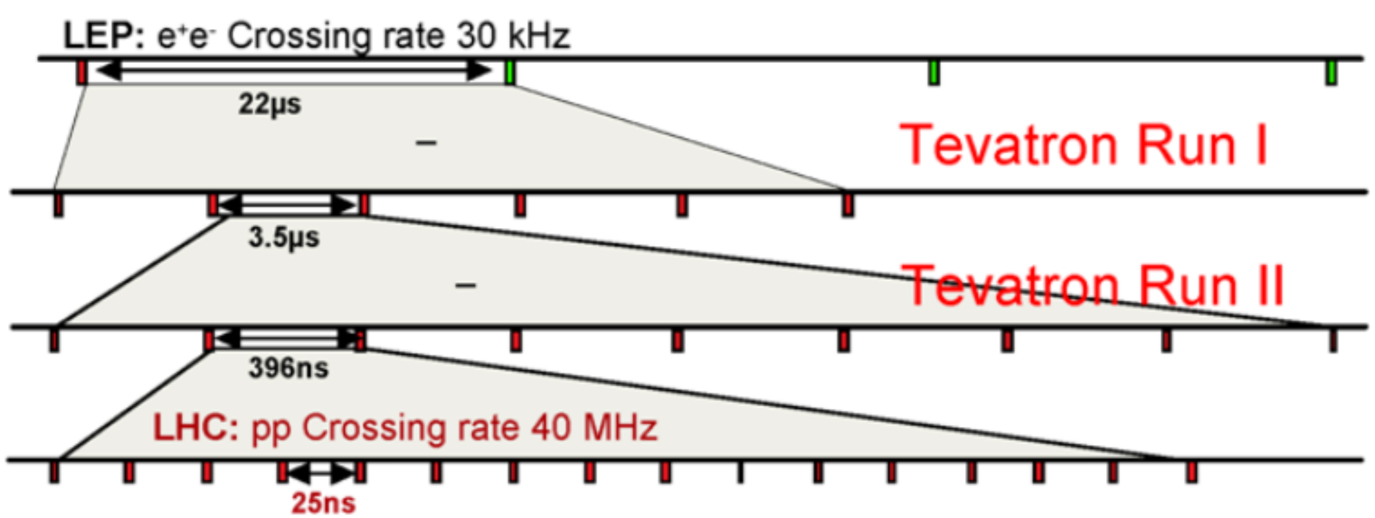
BACK-UP SLIDES



LHC: THE SOURCE

The clock source

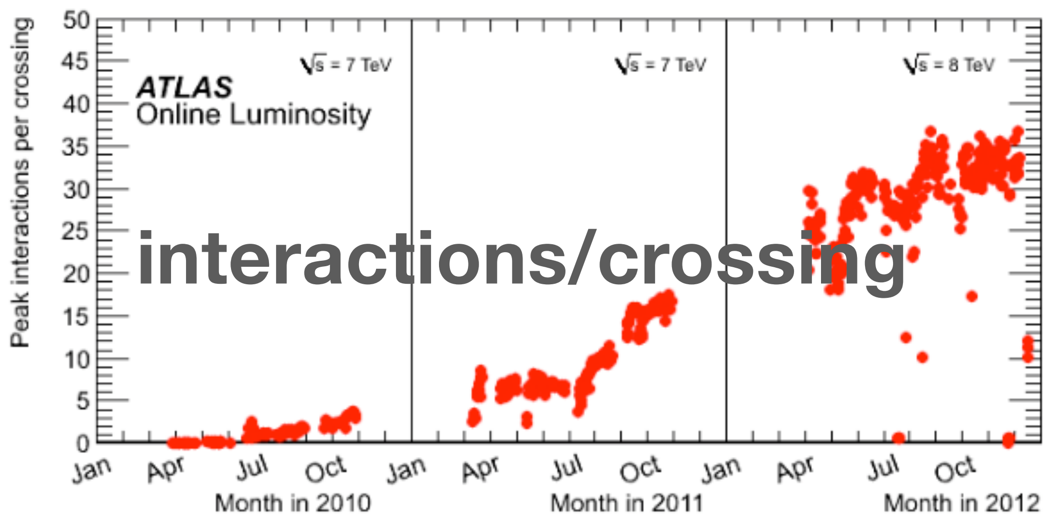
- ~3600 bunches in 27km
- distance bw bunches: 27km/3600 = 7.5m
- distance bw bunches in time: 7.5m/c = 25ns

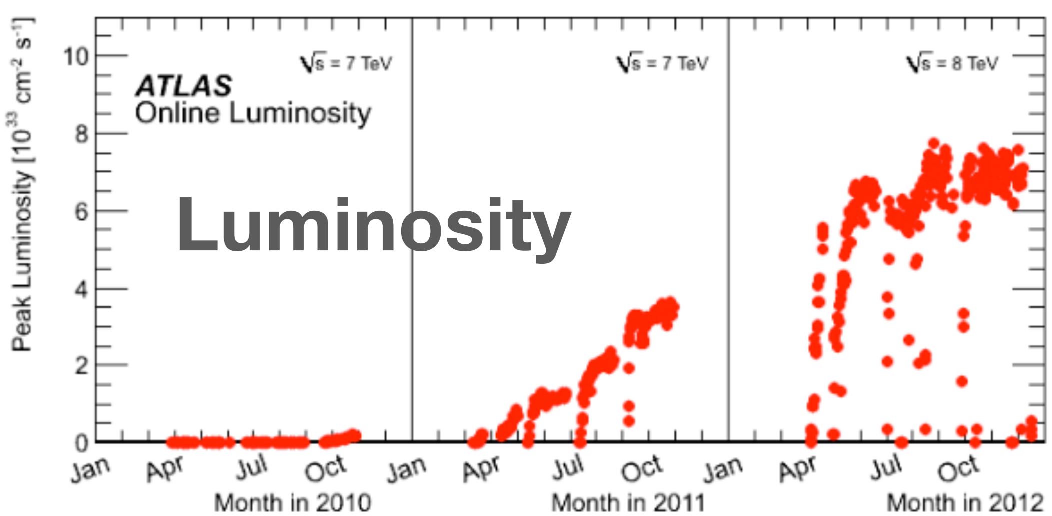


At full Luminosity, every 25ns, ~23 superimposed p-p interaction events

The pile-up source

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

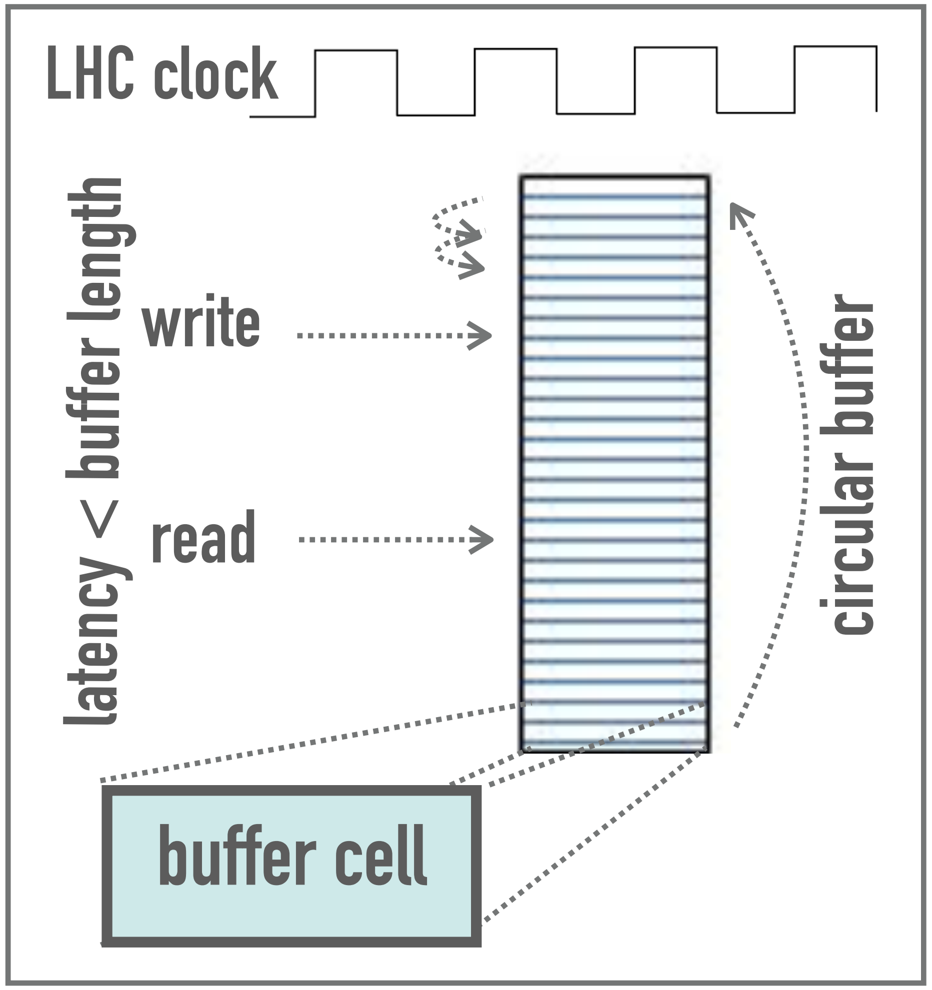




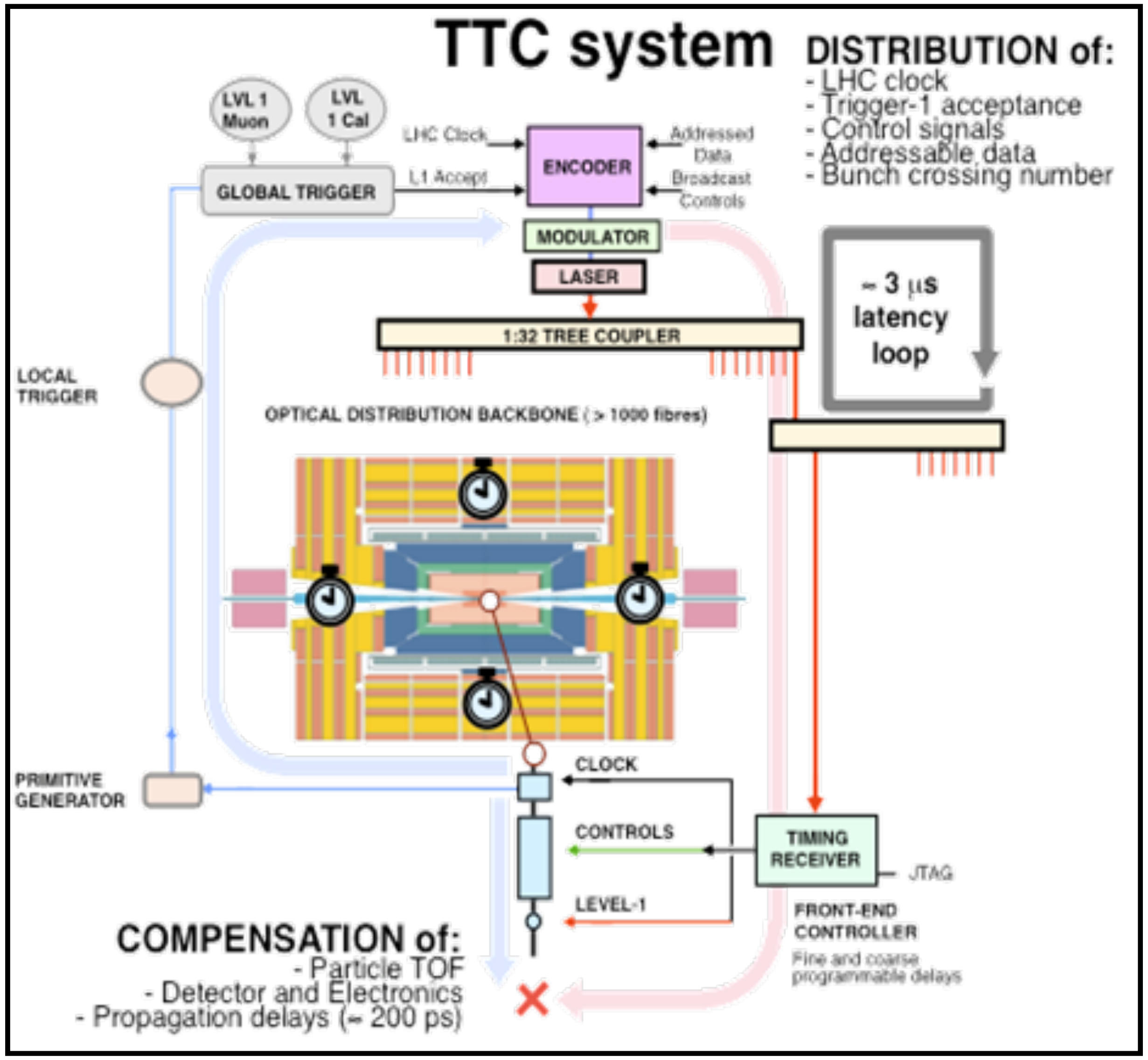
PIPELINED TRIGGERS

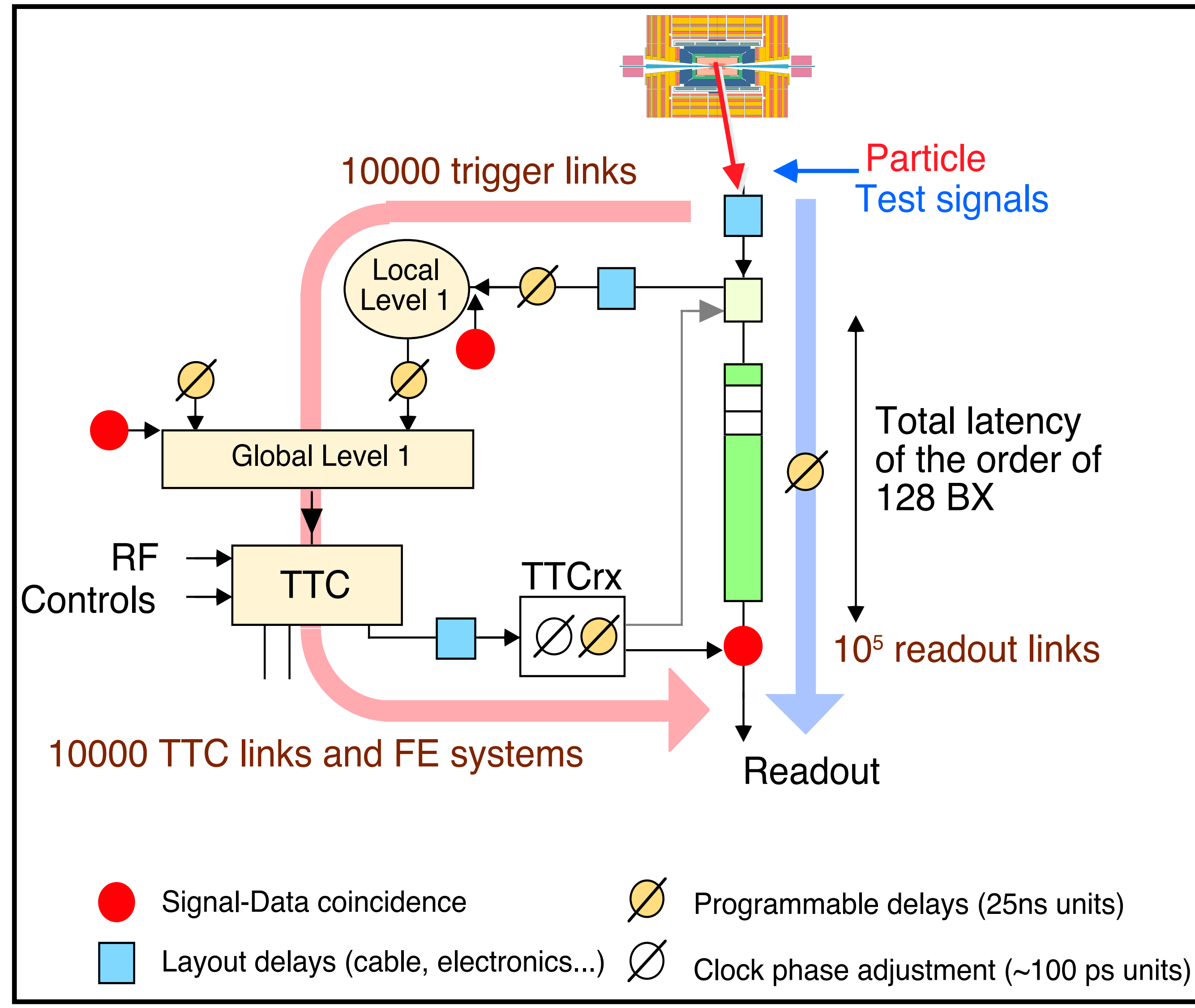
- **Allow trigger decision longer than clock tick (and no deadtime)**
  - Execute trigger selection in defined clocked steps (**fixed latency**)
  - Intermediate storage in stacked buffer cells
  - R/W pointers are moved by clock frequency

- **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



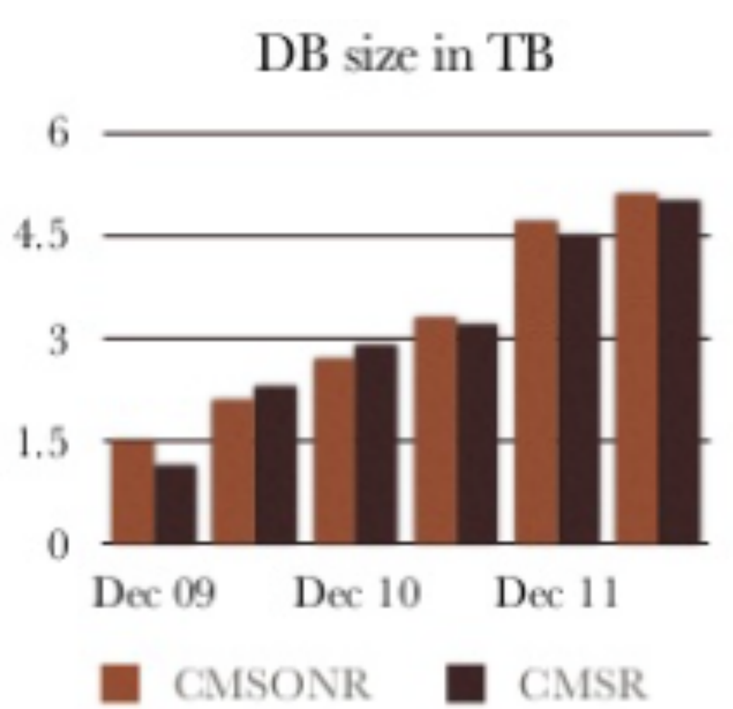


- ➡ **Common optical system: TTC**
  - ➡ radiation resistance
  - ➡ single high power laser
- ➡ **Large distribution**
  - ➡ experiments with ~10^7 channels

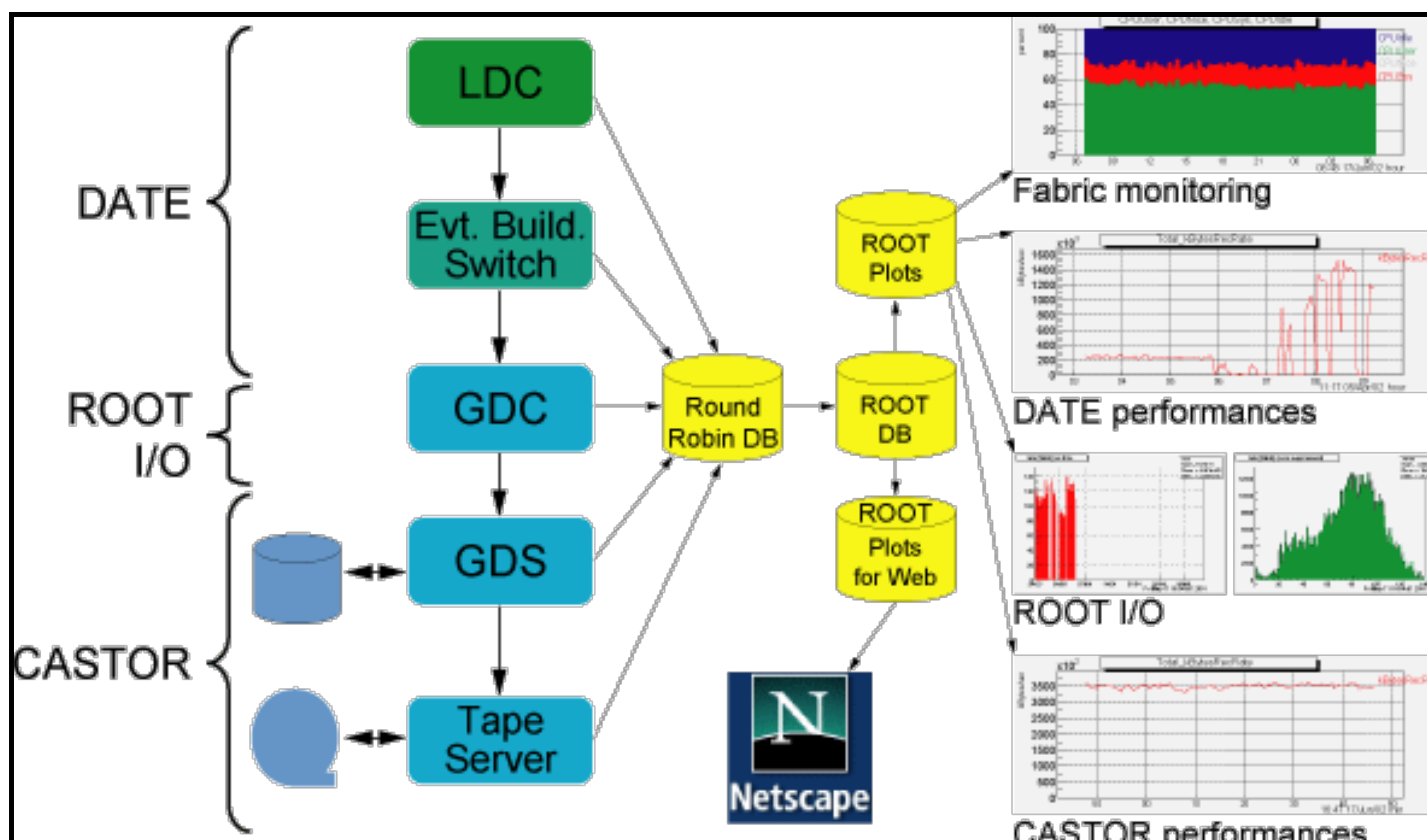
- ➡ **Align readout & trigger at (better than) 25ns and correct for**
  - ➡ time of flight (25 ns ≈ 7.5m)
  - ➡ cable delays (10cm/ns)
  - ➡ processing delays (~100 BCs)

LAST, BUT NOT LEAST

- **Multiple Databases: configuration, condition, both online and offline**
  - Use (Frontier) caches to minimise access to Oracle servers
- **Monitoring and system administration**
  - thousands of nodes and network connections
  - advanced tools of monitoring and management
  - support software updates and rolling replacement of hardware

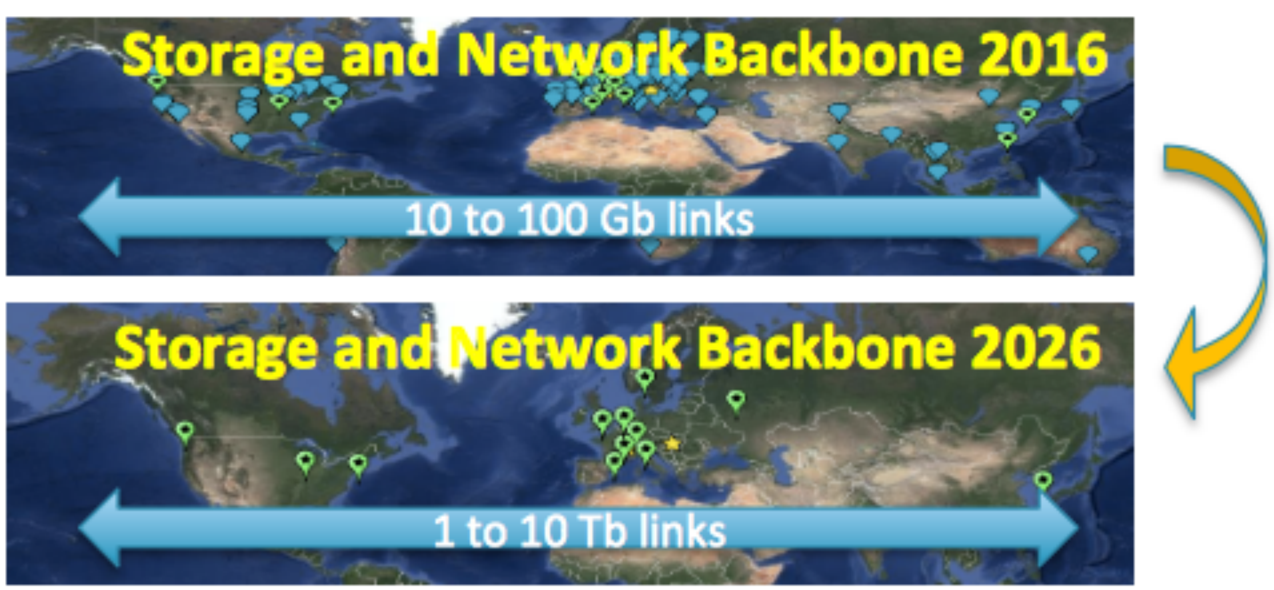


CMS DB grows about 1.5TB/year, condition data only a small fraction

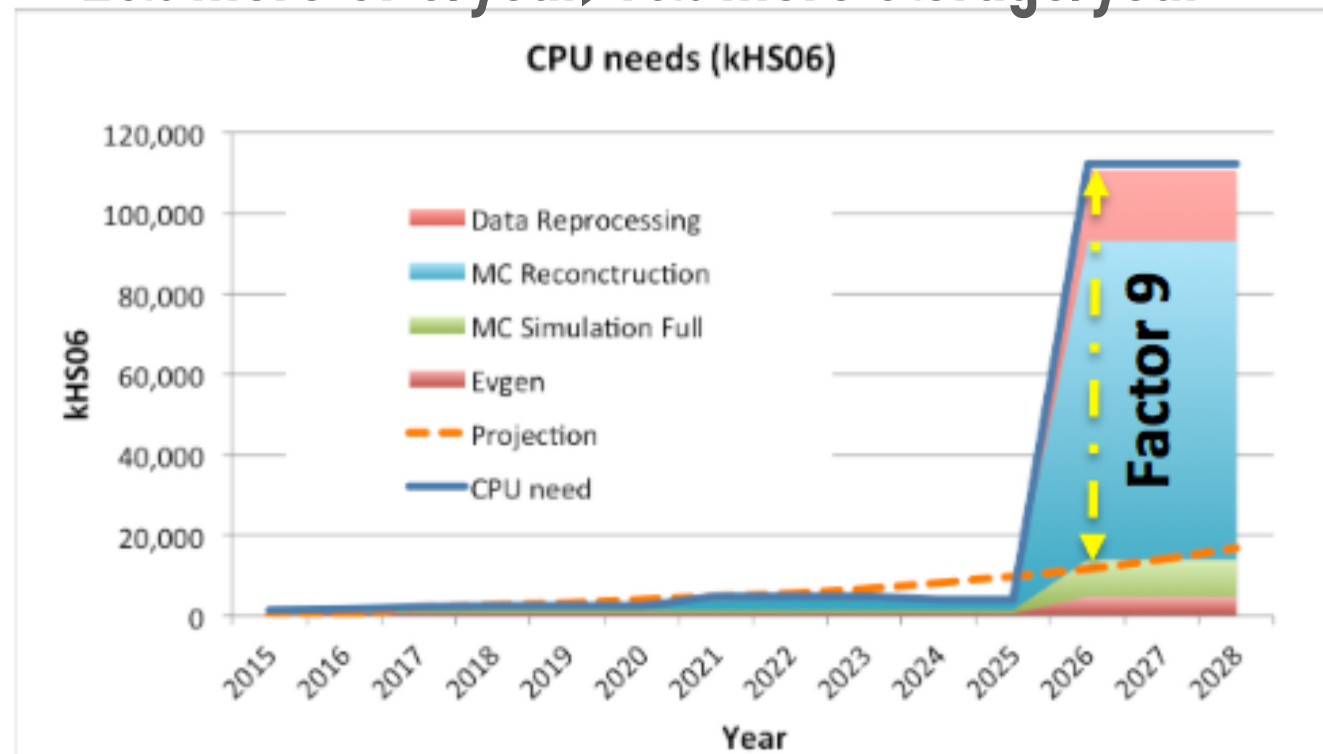


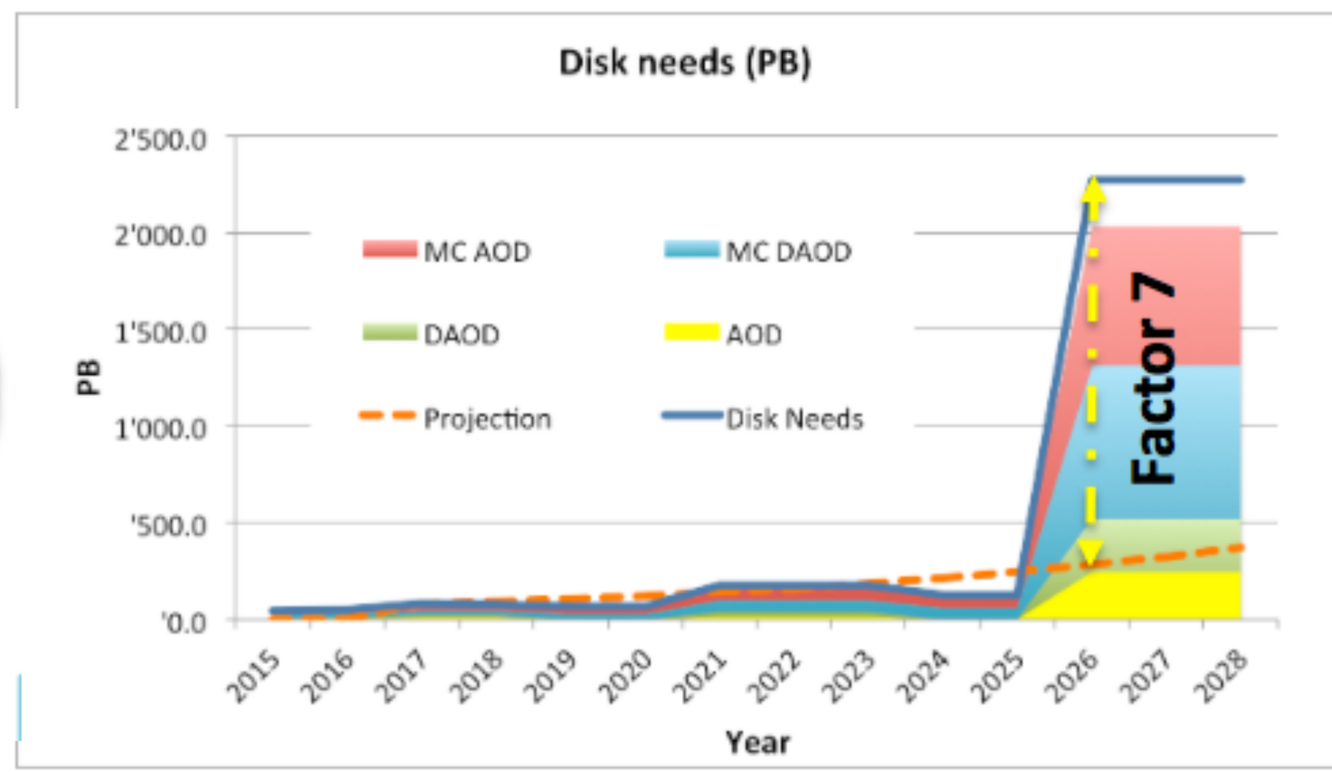
COMPUTING EVOLUTION FOR HL-LHC

- ➡ **Re-thinking of distributed data management, distributed storage and data access.**
- ➡ **A network driven data model allows to reduce the amount of storage, particularly for disk**
  - ➡ Tape today costs 4 times less than disk
- ➡ **Computing infrastructure in HL-LHC**
  - ➡ Network-centric infrastructure
  - ➡ Storage and computing loosely coupled
  - ➡ Storage on fewer data centers in WLCG
  - ➡ Heterogeneous computing facilities (Grid/Cloud/HPC/ ...) everywhere



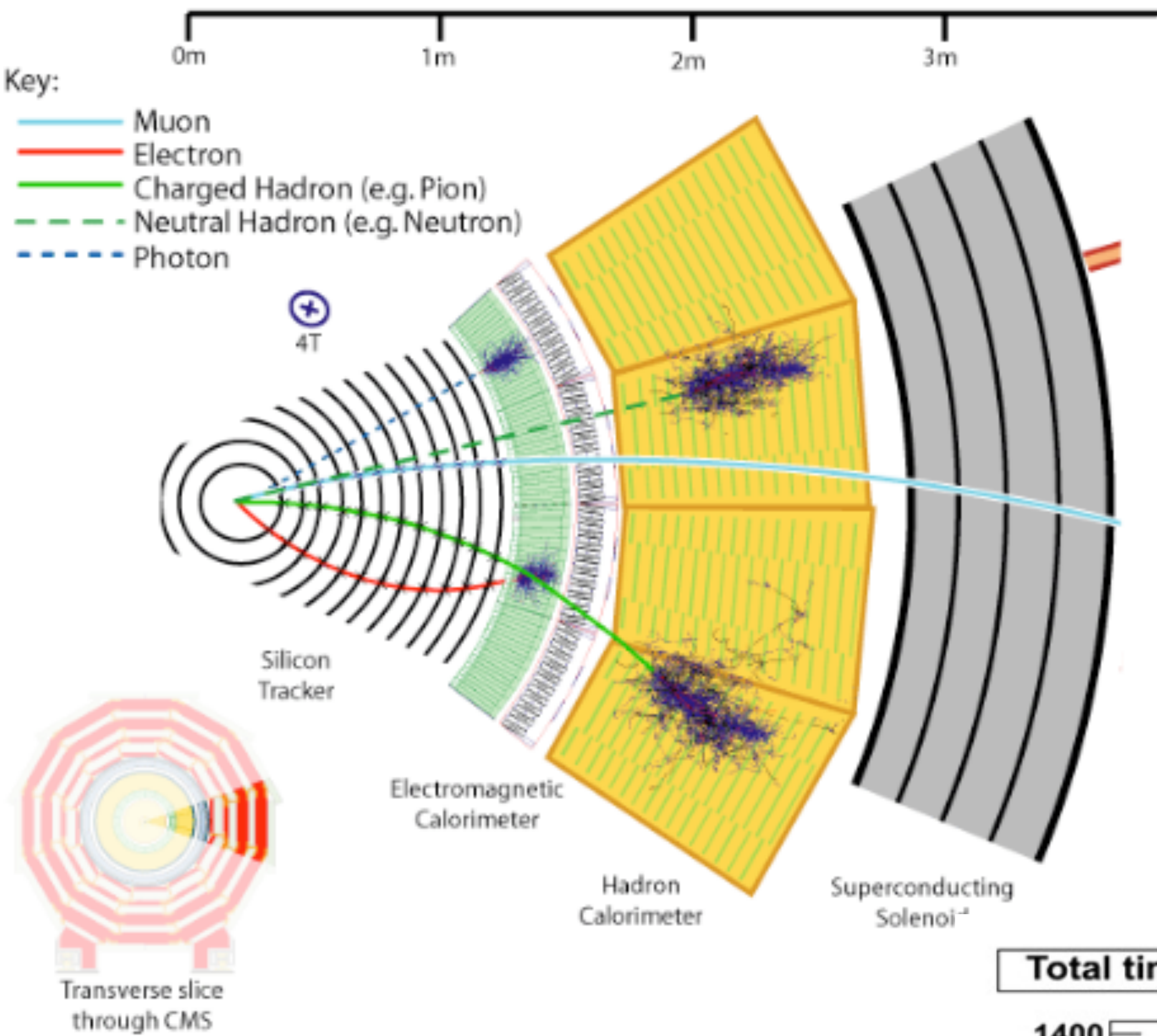
Projection of available resources in HL-LHC: 20% more CPU/year, 15% more storage/year





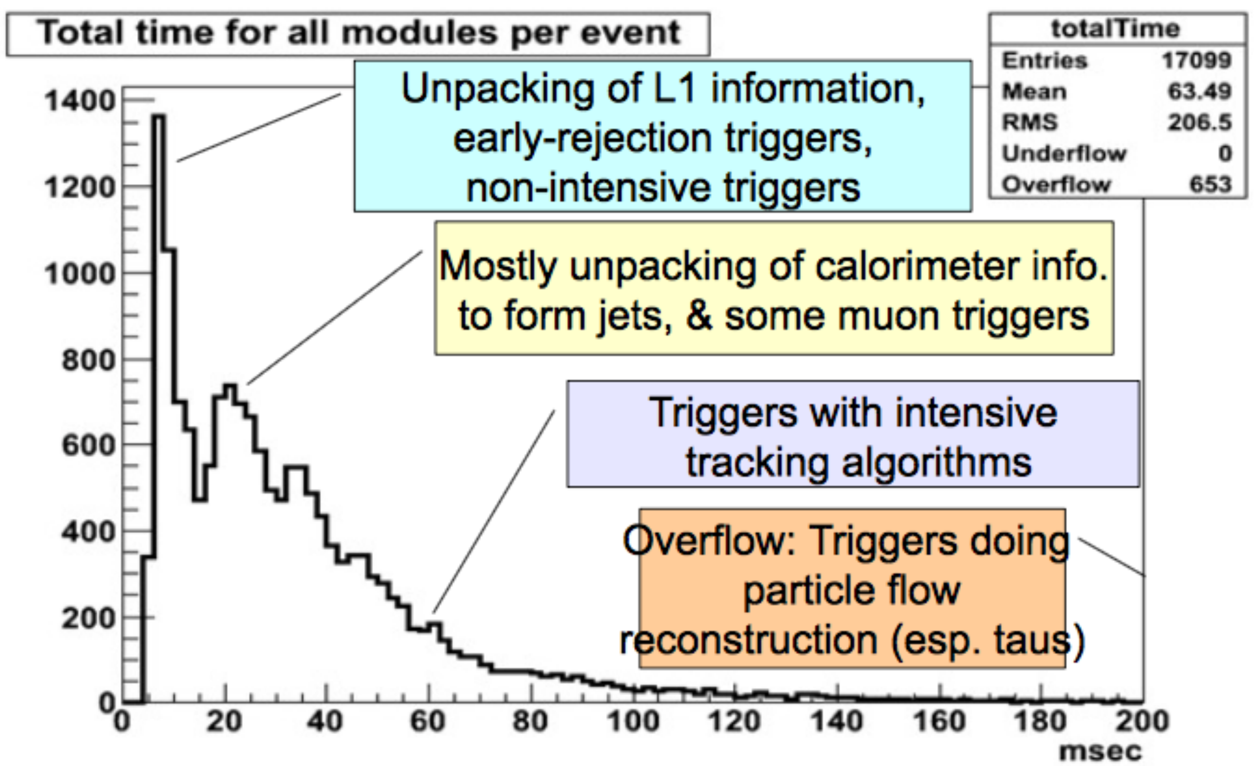
CALORIMETER TRIGGERS

electrons,
photons, taus,
jets,
total energy,
missing energy
Isolation

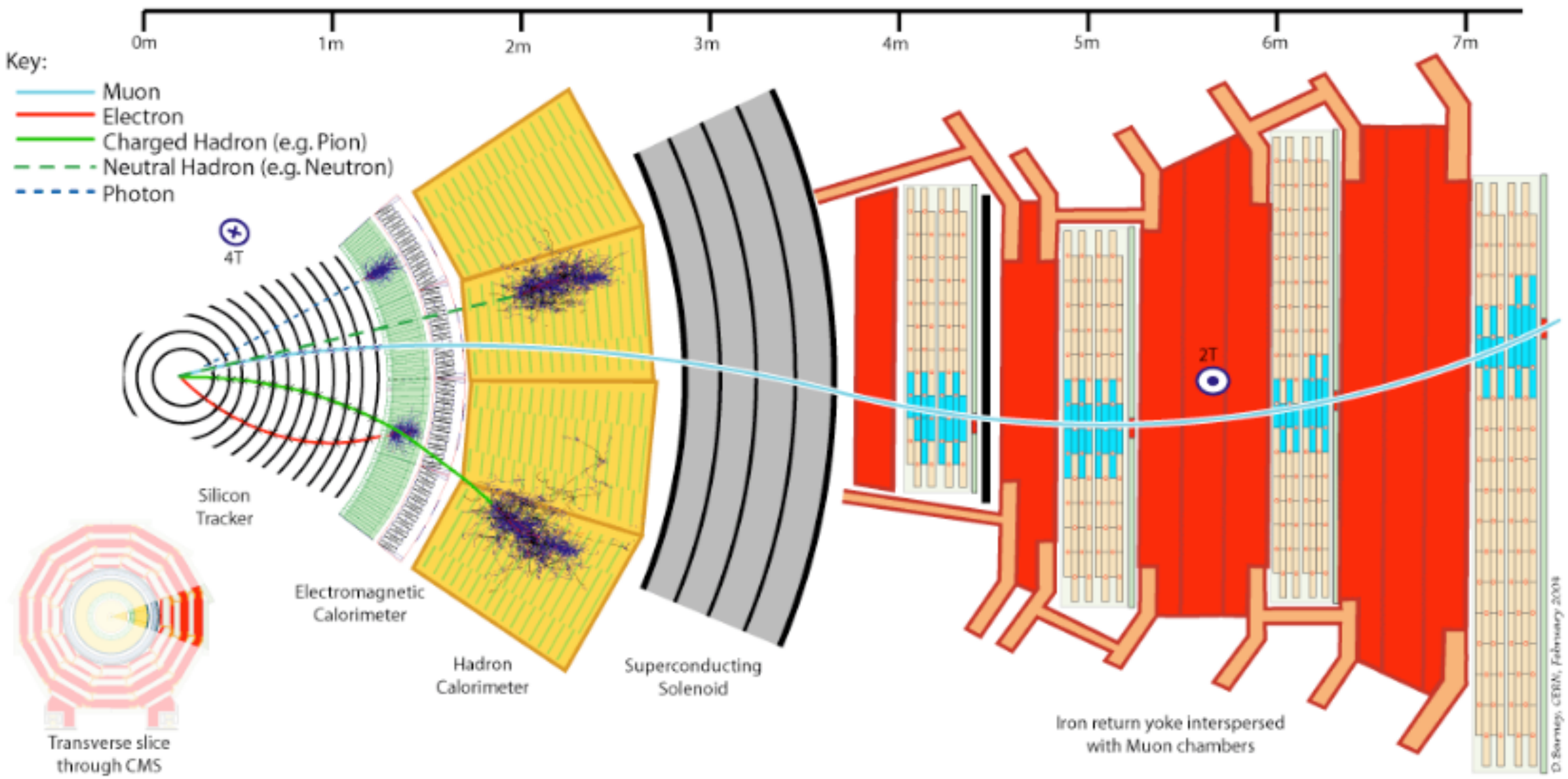


- ➡ **Fast and good resolution (LArg, PbW_4 for e-m)**
- ➡ **First-level processing (40MHz)**
  - ➡ "trigger towers" to reduce data (10-bit range)
  - ➡ sliding-window technique for local maxima
  - ➡ parallel algorithms for cluster shape and energy distribution

- ➡ **High-level processing (100 kHz)**
  - ➡ regional tracking in the inner detectors
  - ➡ bremsstrahlung recovery
  - ➡ measure activity in cones (with tracks/ clusters) to isolate e/jets
  - ➡ jet algorithms



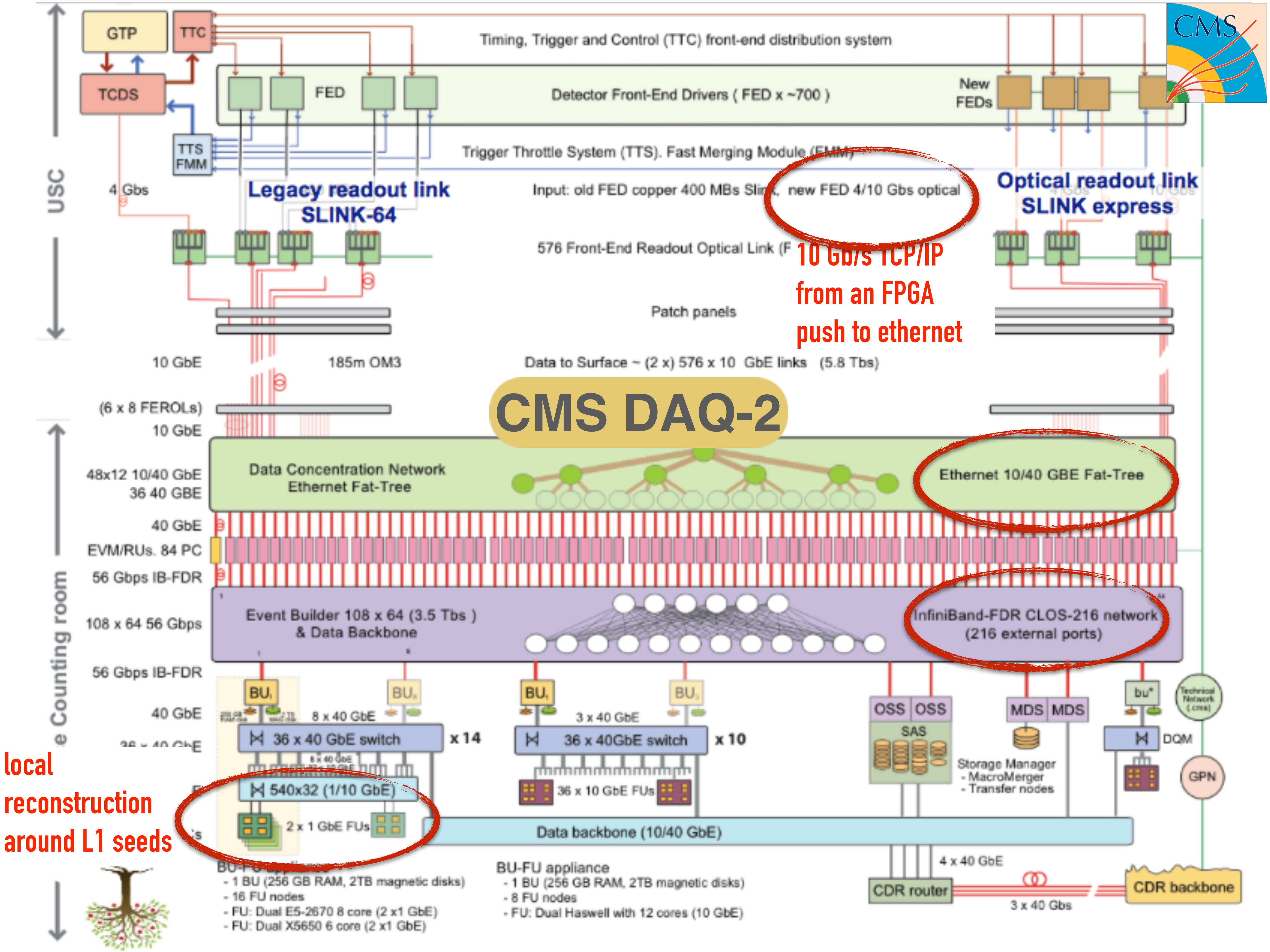
TRIGGERS FOR MUONS



- **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

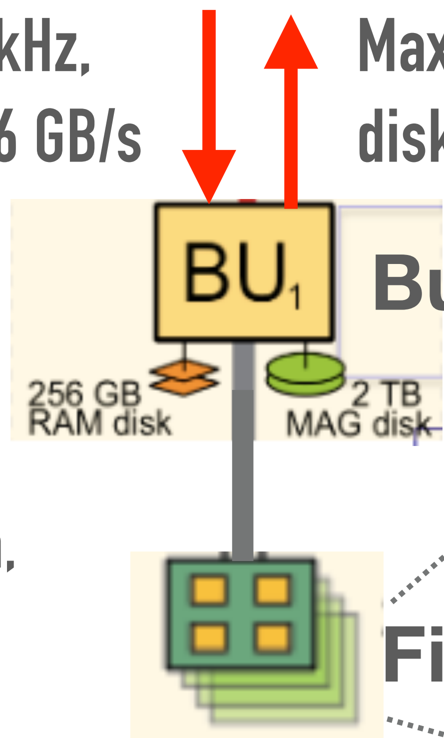

CMS DAQ-2
GTP
TTC
TCDS
Timing, Trigger and Control (TTC) front-end distribution system
FED
Detector Front-End Drivers (FED x ~700)
New FEDs
TTS FMM
Trigger Throttle System (TTS). Fast Merging Module (FMM)
USC
4 Gbs
Legacy readout link SLINK-64
Input: old FED copper 400 MBs Slink, new FED 4/10 Gbs optical
Optical readout link SLINK express
576 Front-End Readout Optical Link (F
10 Gb/s TCP/IP from an FPGA push to ethernet
Patch panels
10 GbE
185m OM3
Data to Surface ~ (2 x) 576 x 10 GbE links (5.8 Tbs)
(6 x 8 FEROLs)
10 GbE
48x12 10/40 GbE
36 40 GBE
Data Concentration Network Ethernet Fat-Tree
Ethernet 10/40 GBE Fat-Tree
40 GbE
EVM/RUs. 84 PC
56 Gbps IB-FDR
108 x 64 56 Gbps
Event Builder 108 x 64 (3.5 Tbs) & Data Backbone
InfiniBand-FDR CLOS-216 network (216 external ports)
56 Gbps IB-FDR
40 GbE
Counting room
BU
8 x 40 GbE
36 x 40 GbE switch
x 14
540x32 (1/10 GbE)
2 x 1 GbE FUs
3 x 40 GbE
36 x 40GbE switch
x 10
36 x 10 GbE FUs
OSS
OSS
SAS
MDS
MDS
Storage Manager
- MacroMerger
- Transfer nodes
bu*
DQM
GPN
Data backbone (10/40 GbE)
local reconstruction around L1 seeds
BU-FU appliance
- 1 BU (256 GB RAM, 2TB magnetic disks)
- 16 FU nodes
- FU: Dual E5-2670 8 core (2 x1 GbE)
- FU: Dual X5650 6 core (2 x1 GbE)
BU-FU appliance
- 1 BU (256 GB RAM, 2TB magnetic disks)
- 8 FU nodes
- FU: Dual Haswell with 12 cores (10 GbE)
4 x 40 GbE
CDR router
3 x 40 Gbs
CDR backbone

EVOLUTION OF THE FILTER FARM

Full readout, but regional reconstruction in HLT seeded by L1 trigger objects

Max 2kHz, 2.2–2.6 GB/s

Max 150 MB/s (into 4x disk RAID0 array)



Building Unit (BU)

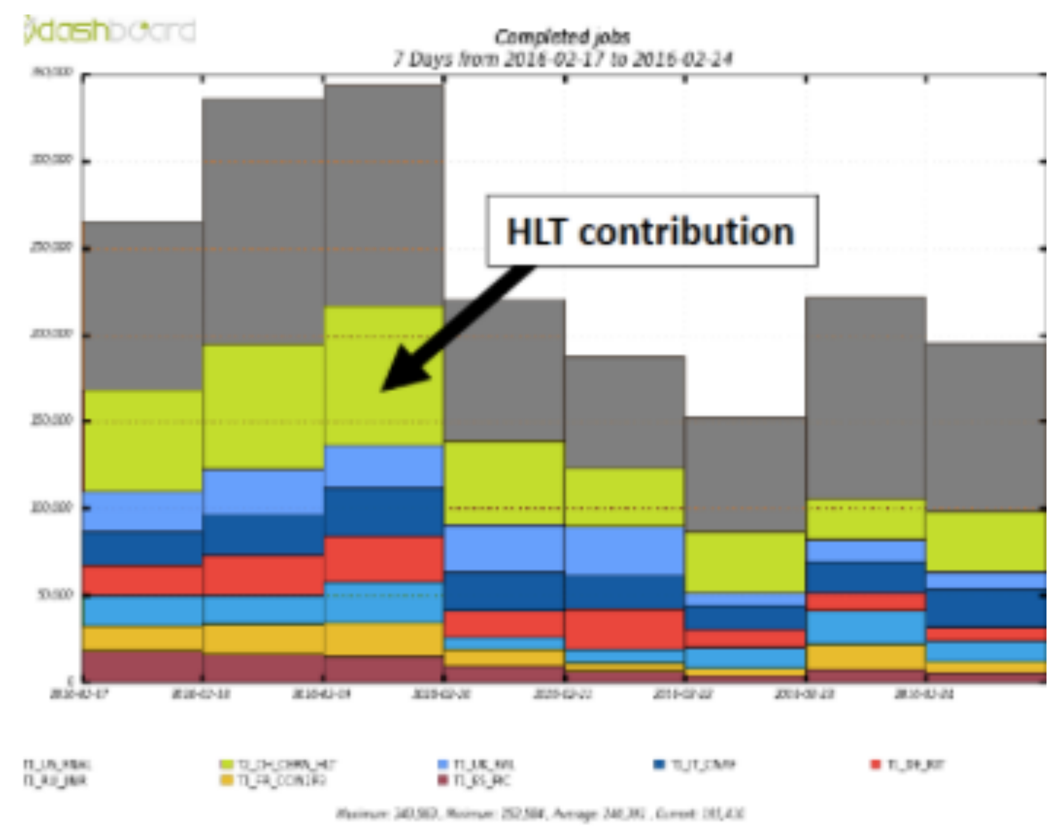
data, status, configuration, latency

Filter Unit (FU)

Every data file accompanied by a metadata in JSON files

Integrated Cloud capability (New!)

- ➡ Added ability to run WLCG grid jobs in FUs during stops/interfill



File-based communication

- ➡ HLT and DAQ completely decoupled
- ➡ Network filesystem used as transport (and resource arbitration) protocol (LUSTRE FS)

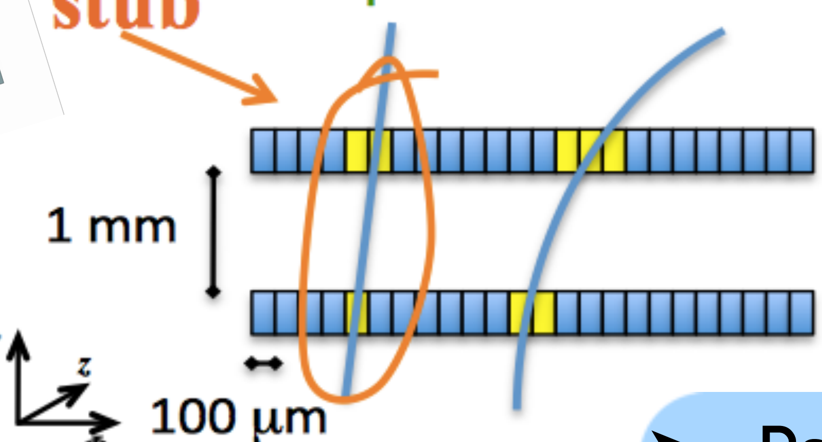
CMS: LOW-p_T TRACK FILTERING

Track filtering (low p_T)

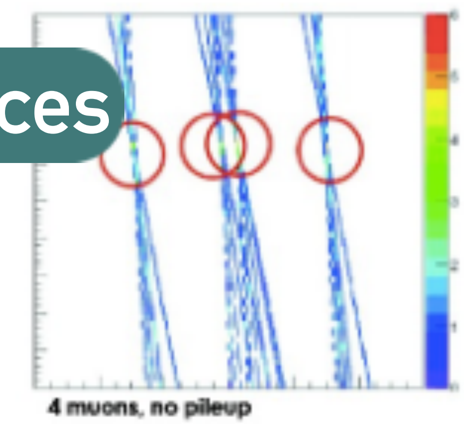
Reduce readout 40 ➟1MHz by detector coincidences

- ➡ **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

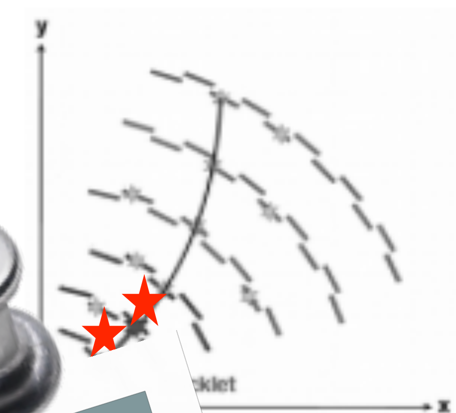
40MHz



Track finding options

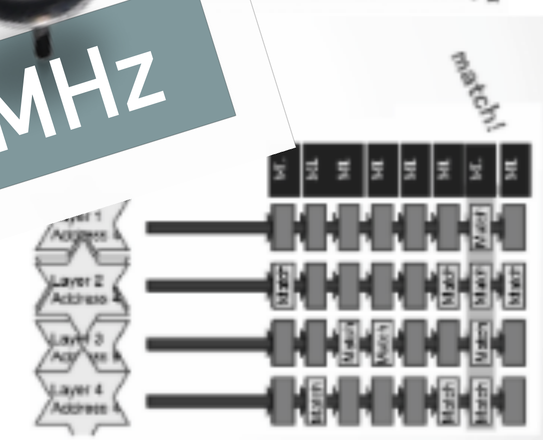


Hough Transform



Tracklets

1MHz

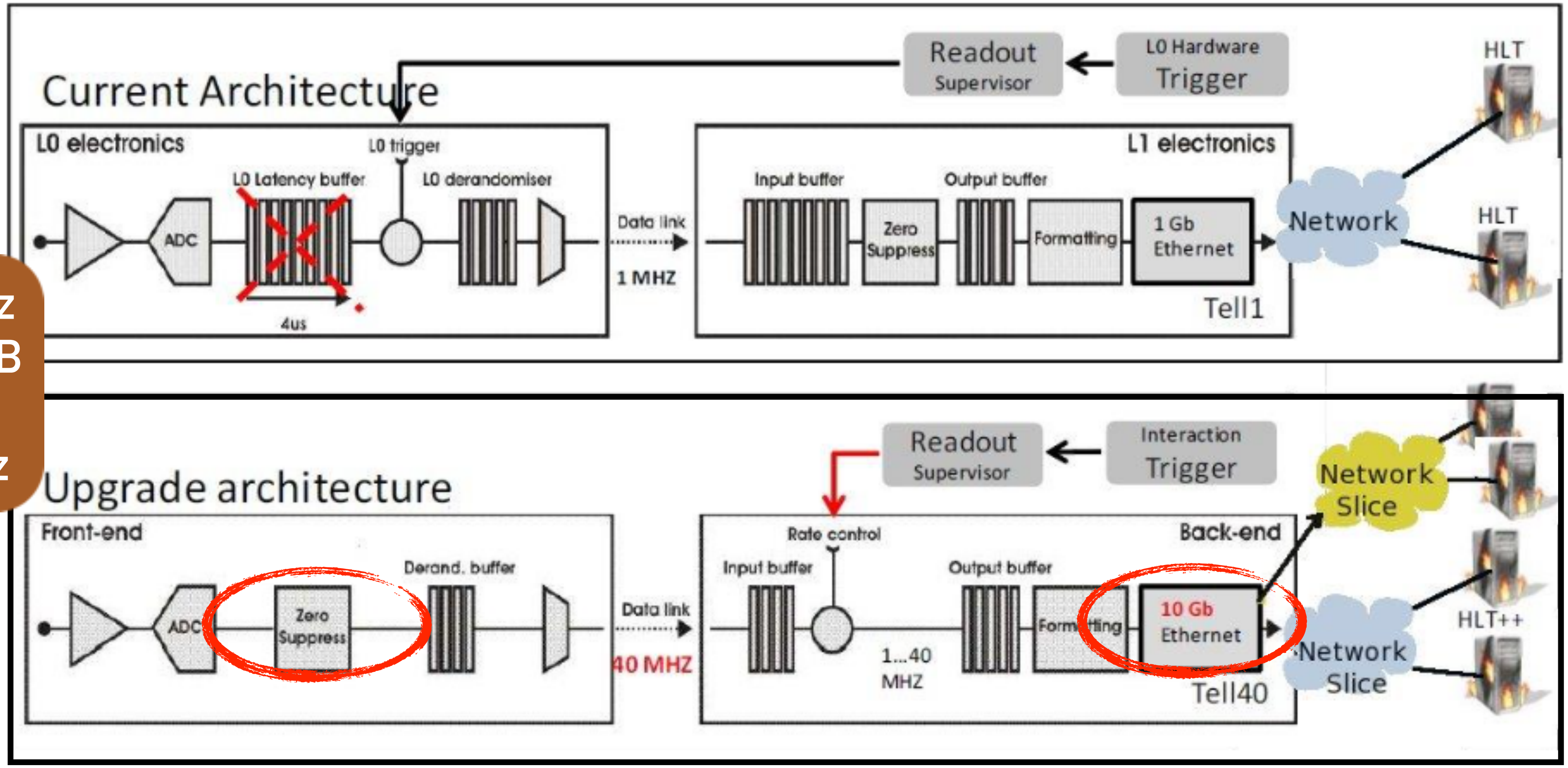


Associative Memories

- ➤ Data rates > 50-100 Tbps
- ➤ Latency: 4+1 µs
- ➤ Three R&D efforts: FPGA/ASIC

HOW TO LIVE WELL WITHOUT A L1 TRIGGER

Readout: 40 MHz
Event size: 100kB
DAQ: 40 Tbit/s
Record: 100 kHz

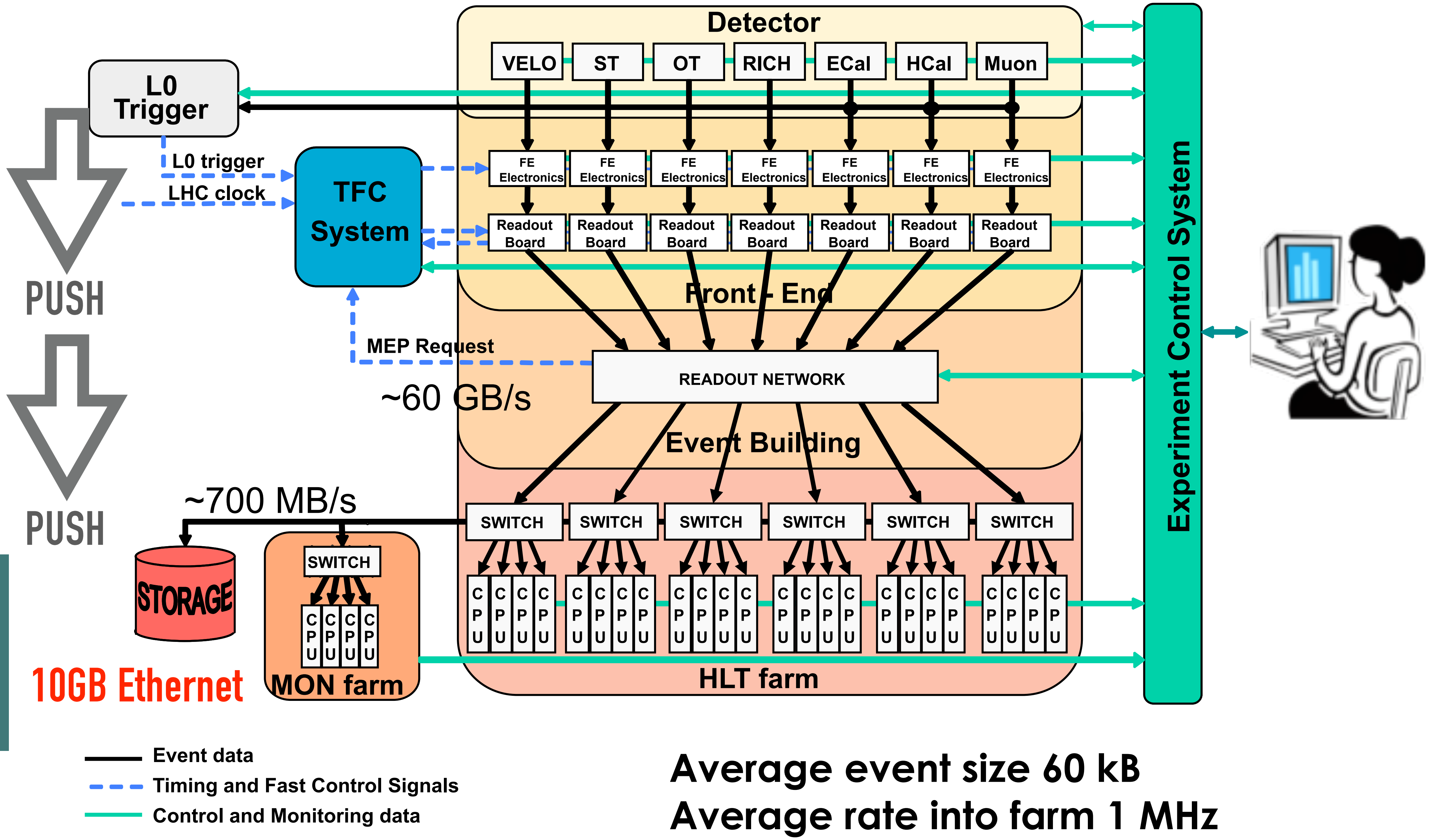


- ➡ **Need zero-suppressing on front-end electronics**
- ➡ **A single, high performance, custom FPGA-card (PCIe40)**
  - ➡ 8800 (# VL) * 4.48 Gbit/s (wide mode) => 40 Tbps
- ➡ **Single board up to 100 Gbits/s (to match DAQ links in 2018)**
- ➡ **Event-builder with 100 Gbit/s technology and data centre-switches**

TDAQ ARCHITECTURE IN RUN-2

Deep buffering in the readout network (overloaded x300 at L0A)

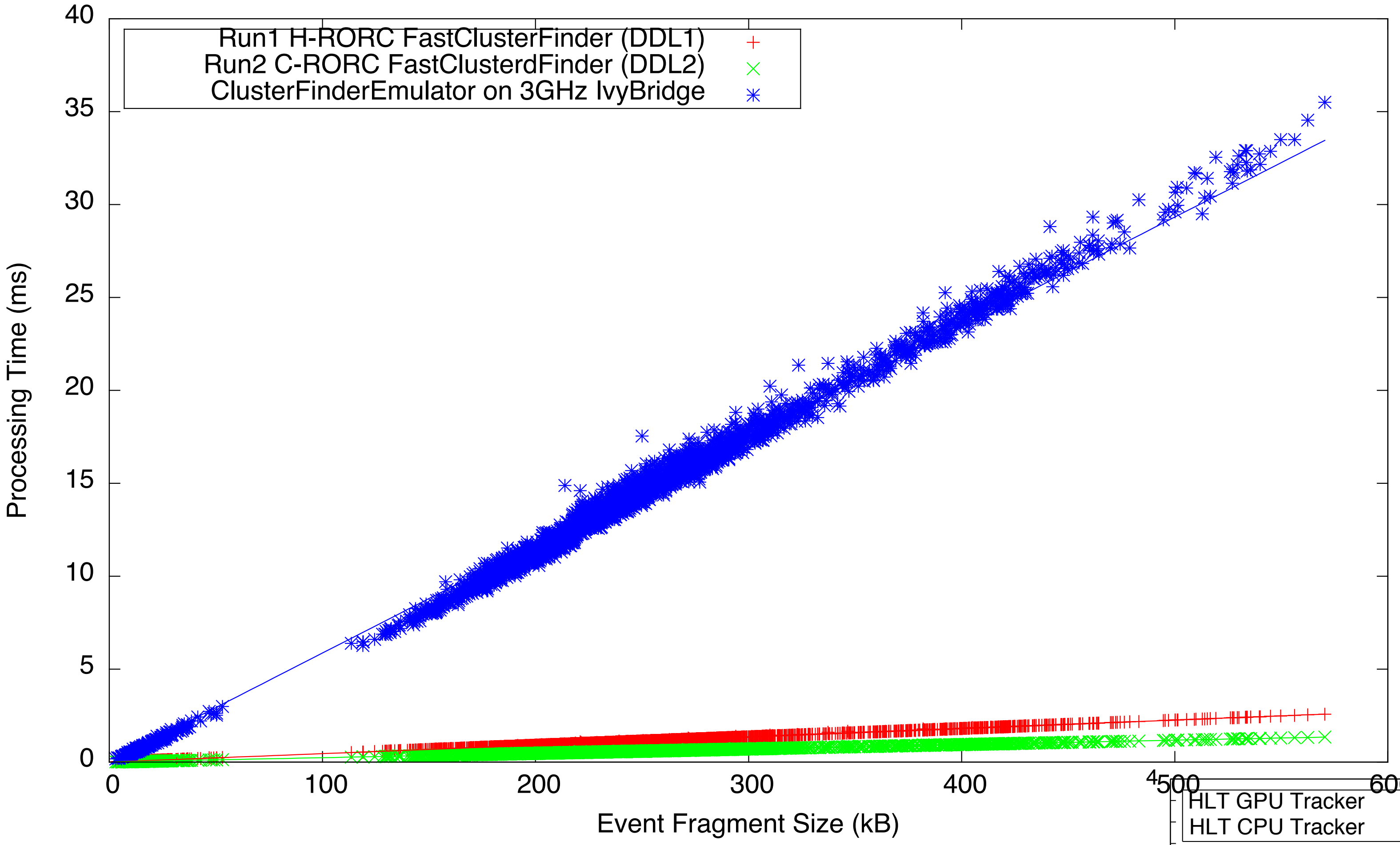
62 sub-farms, total 1780 nodes, with edge-routers (12 Gbps)



Average event size 60 kB
Average rate into farm 1 MHz
Average rate to tape ~12 kHz

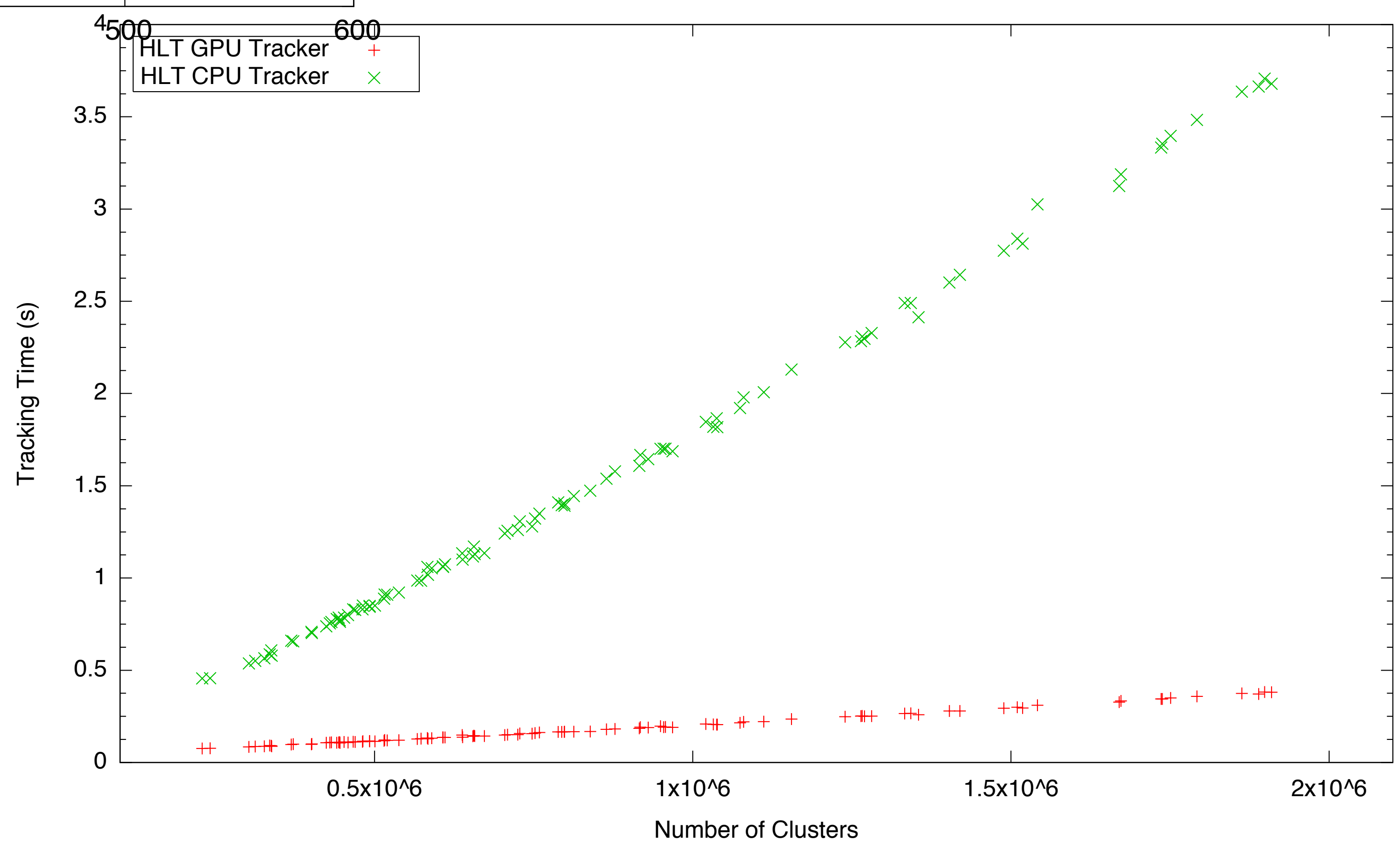
- **Small event, at high rate: ask for optimized transmission**
  - TTC system is used to assign IP addresses to RO boards
  - Ethernet UDP, with 10-15 events packed ⇒ ~ **80 kHz**

HARDWARE ACCELERATION WITH FPGAS AND GPUS

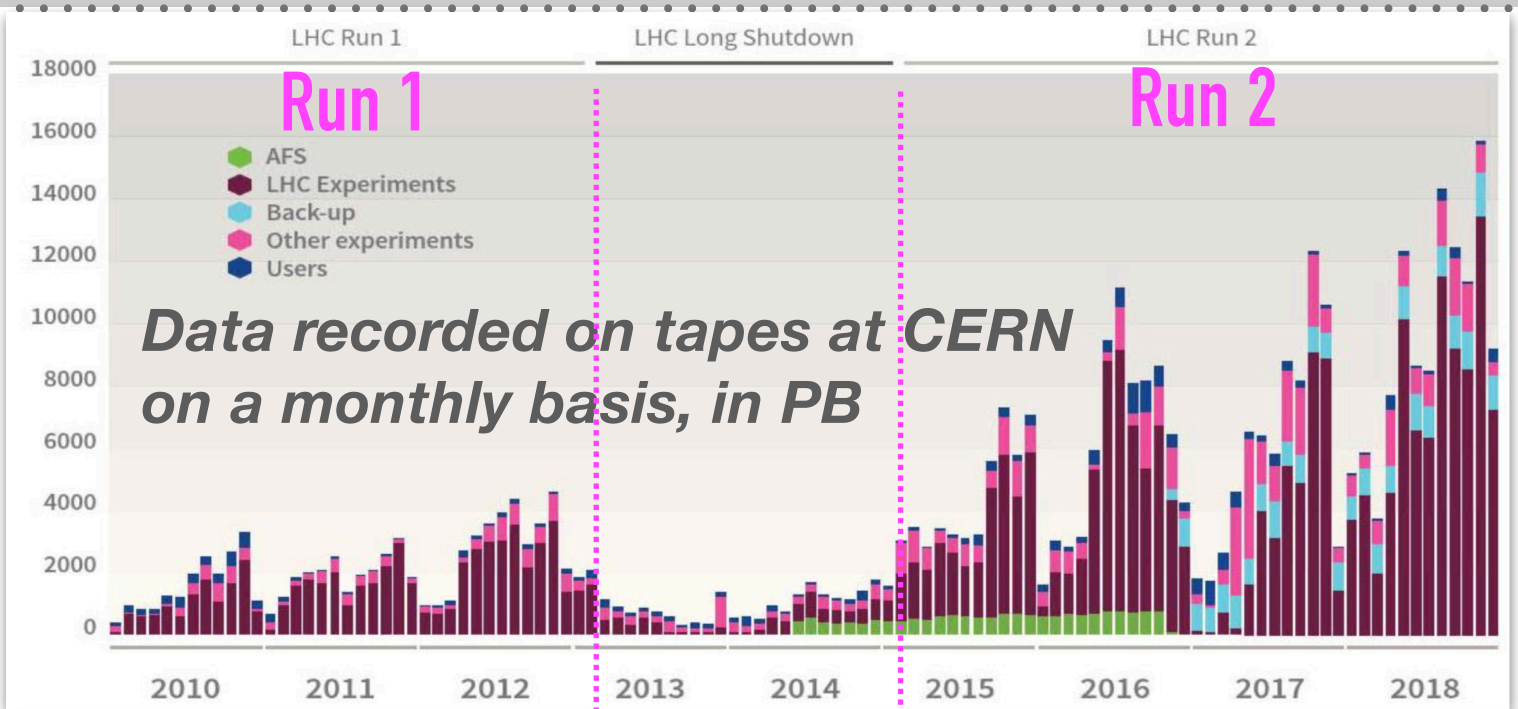


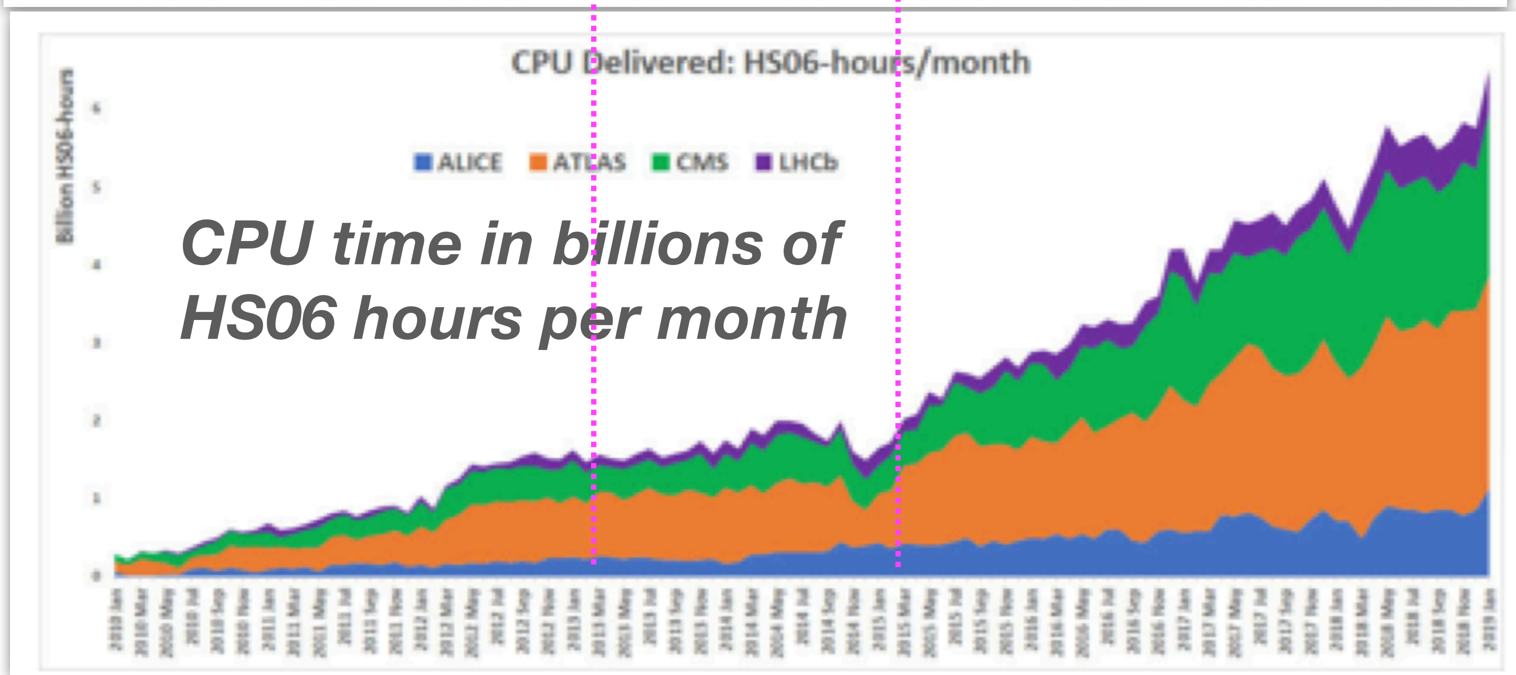
Performance of the FPGA-based FastClusterFinder algorithm for DDL1 (Run1) and DDL2 (Run2) compared to the software implementation on a recent server PC.

Tracking time of HLT TPC Cellular Automata tracker on Nehalem CPU (6Cores) and NVIDIA Fermi GPU.



LHC COMPUTING TOWARDS NEW PARADIGMS





see [Ref]

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**

➡ **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