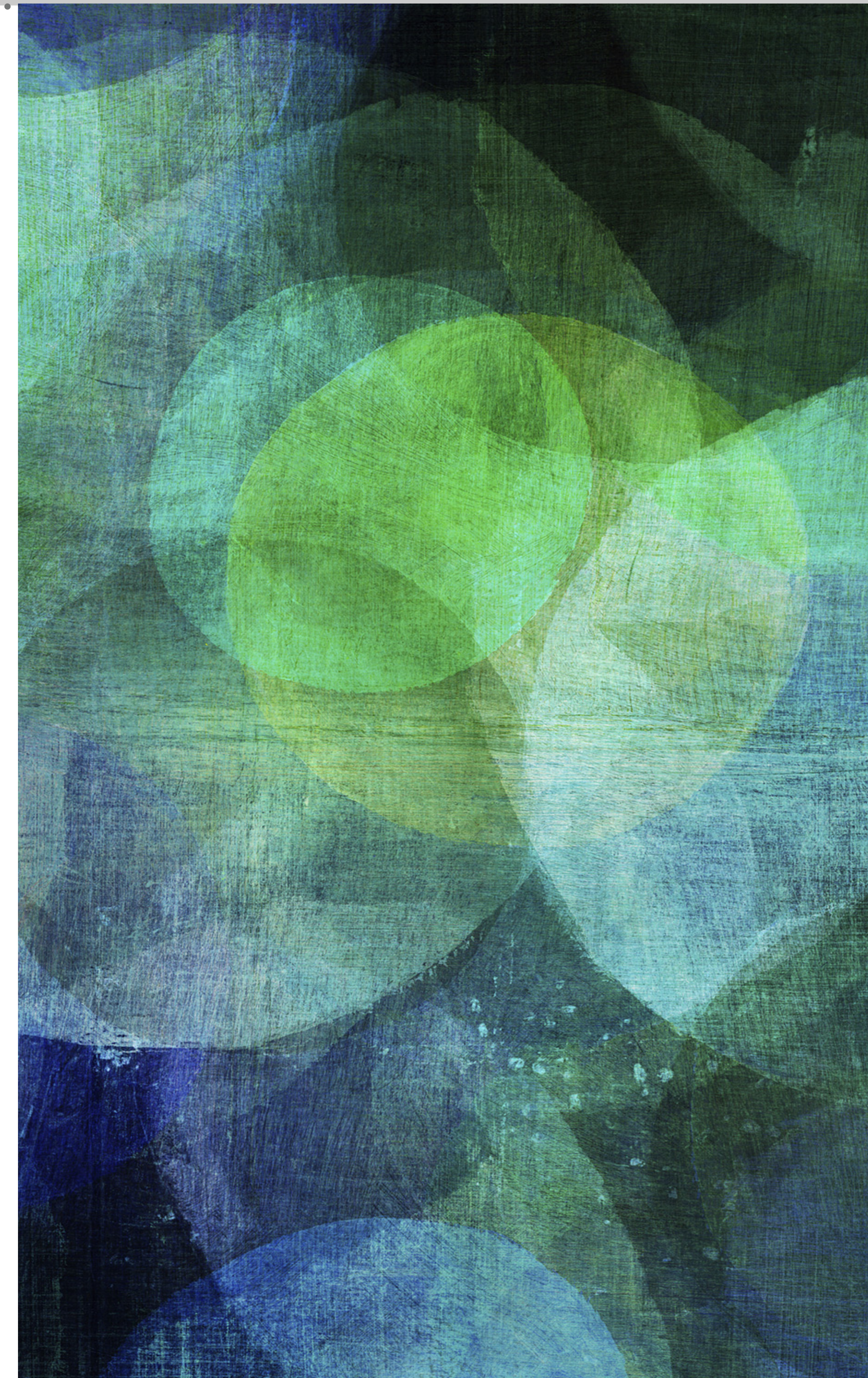
STRATEGIES AND FUTURE TRENDS FOR TRIGGER AND DAQ SYSTEMS IN LHC EXPERIMENTS

F.Pastore (Royal Holloway Un. of London)

THE CONTENTS OF THIS LECTURE

- ➡ Triggering and taking data at the LHC
- ➡ Four experiments, four different use cases and different strategies for T/DAQ
- ➡ Design and future ideas
- ➡ Spotlight upgrade examples



TRIGGERING AND TAKING DATA AT LHC

TDAQ for large discovery experiments

LHC EXPERIMENTS FOR A DISCOVERY MACHINE

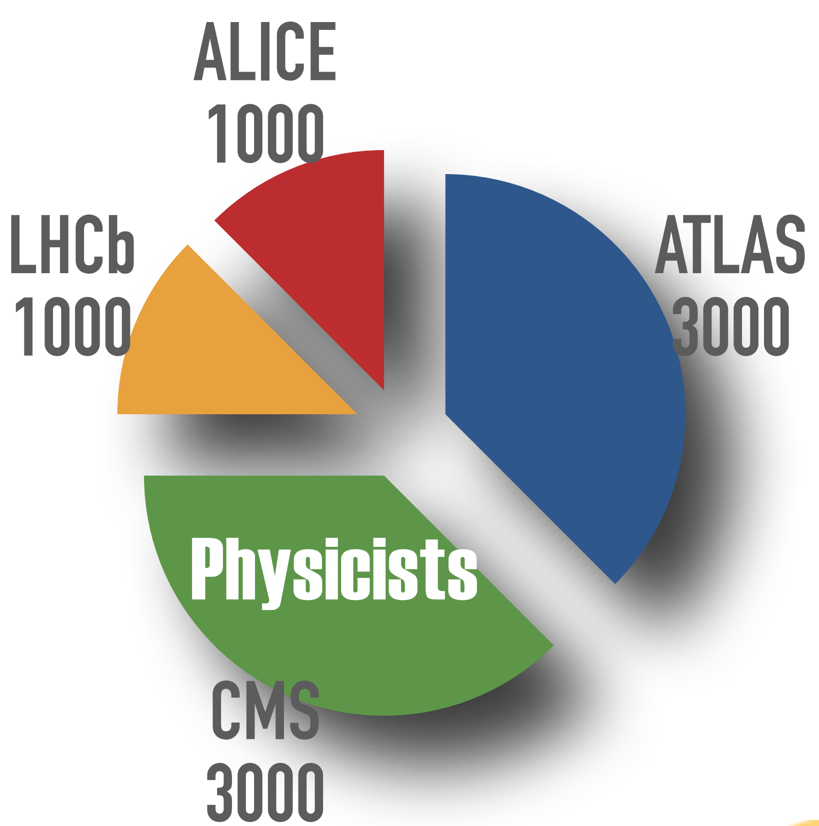
Goal: explore TeV energy scale to find New Physics beyond Standard Model

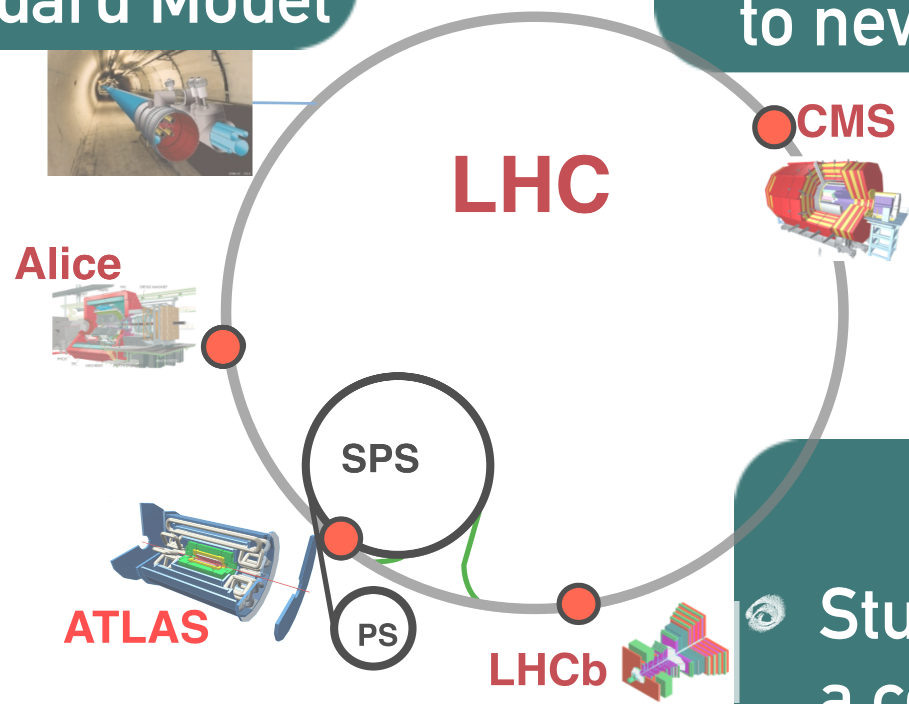
ATLAS & CMS

- 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

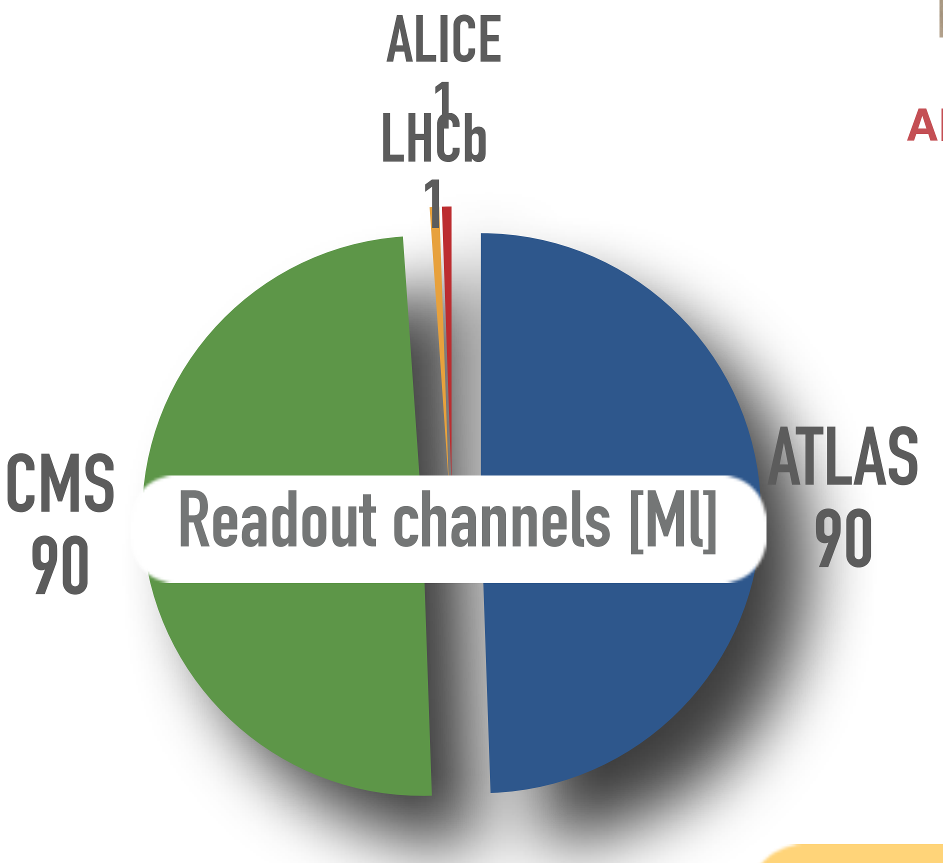
Goal: explore TeV energy scale to find New Physics beyond Standard Model

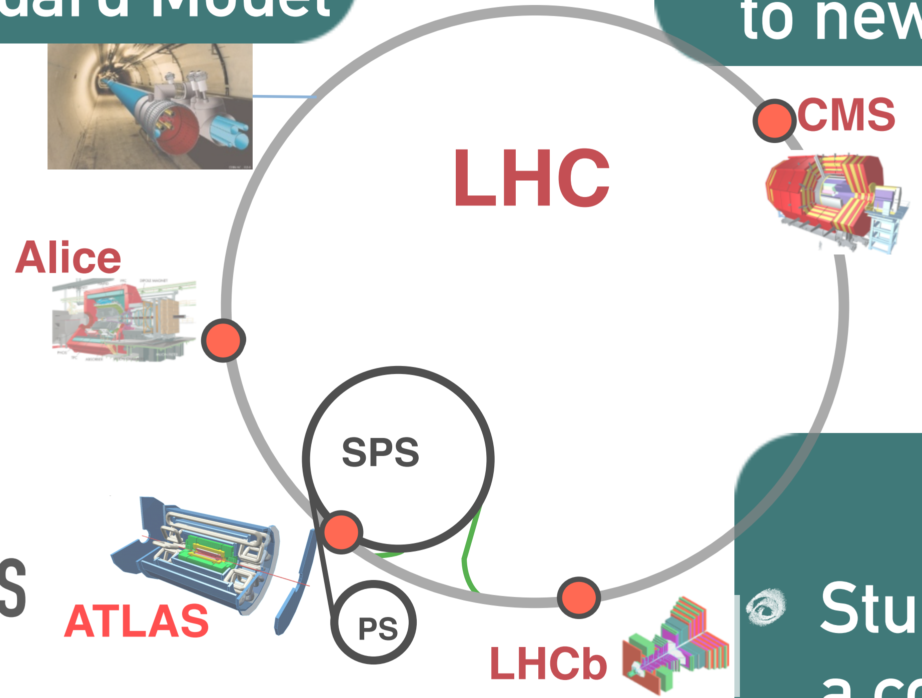
ATLAS & CMS

- 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





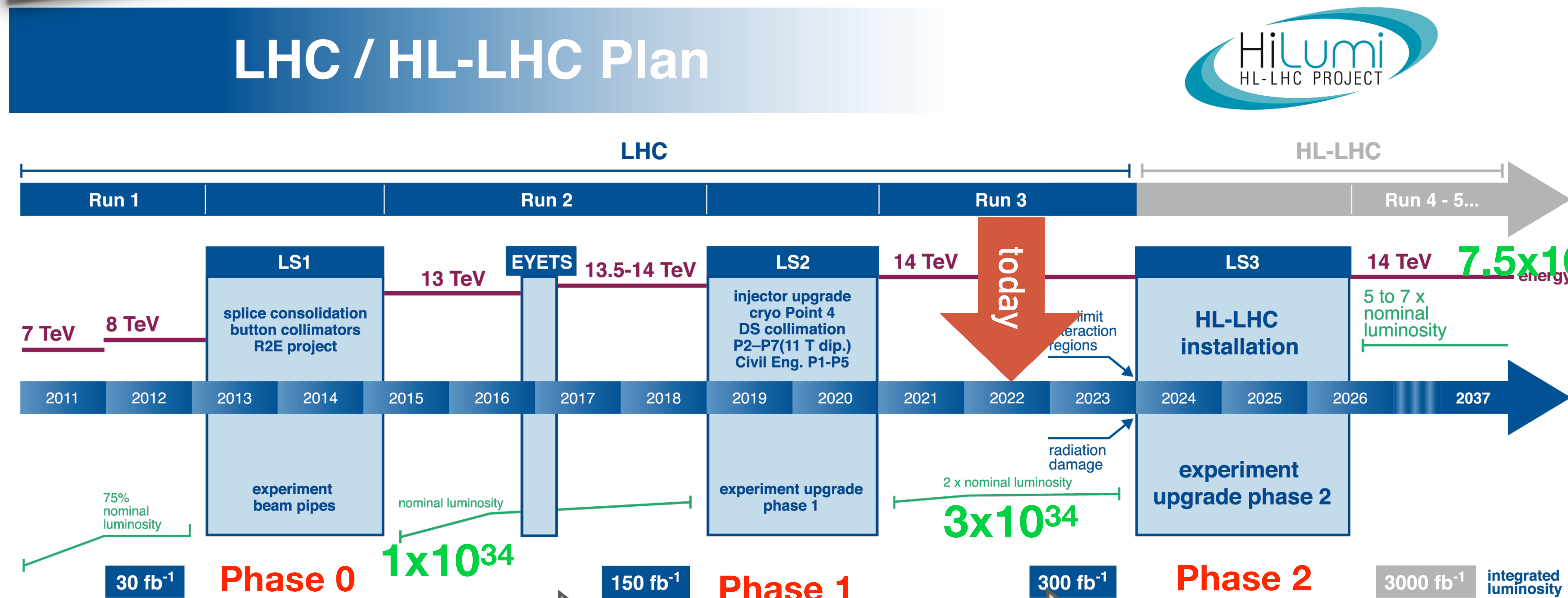
ALICE

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

Proposed: 1992, Approved: 1996, Started: 2009

LHC BECOMING IMPRESSIVELY LUMINOUS

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

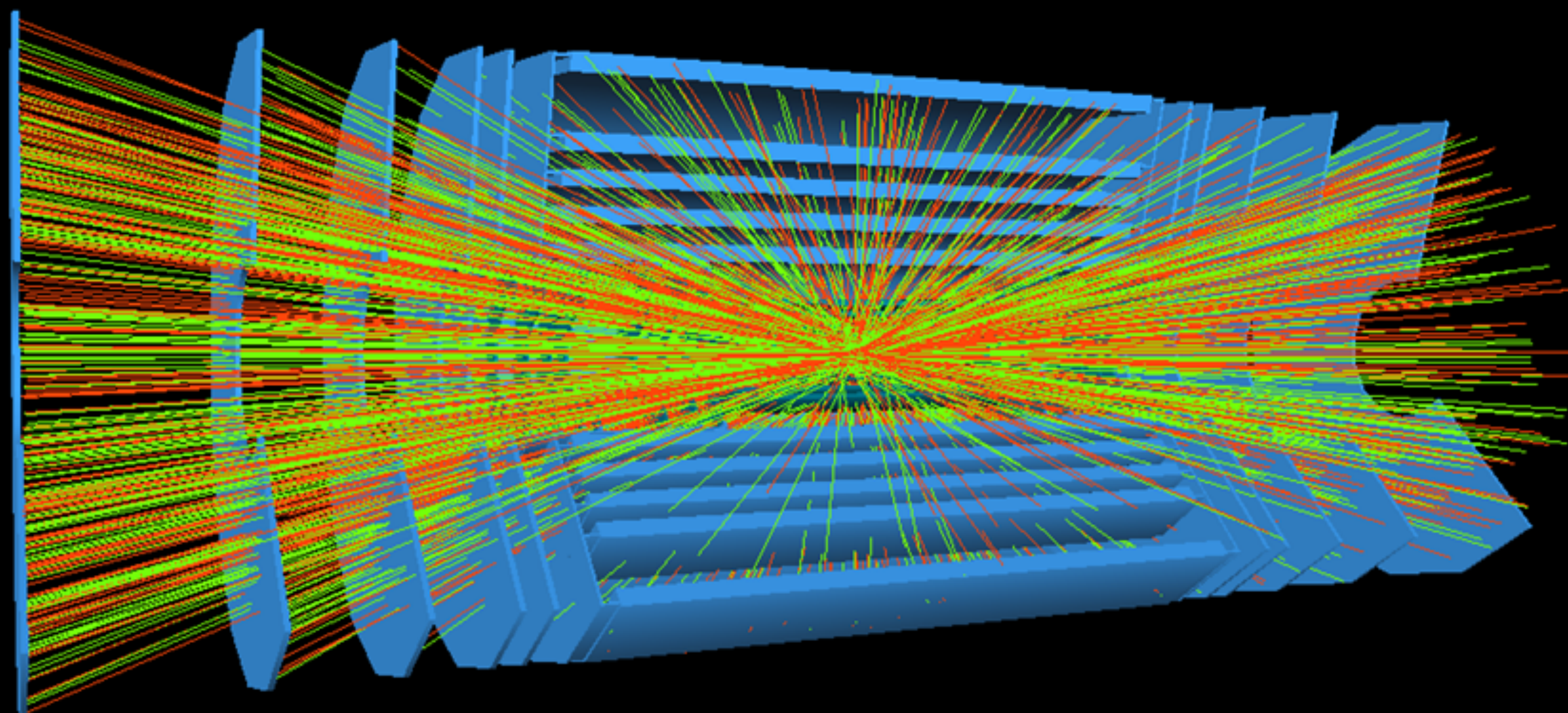


Consolidation for all experiments → Major Upgrade in ALICE and LHCb → Major Upgrade in ATLAS and CMS

➡ Experiments go beyond design specifications ($1x10^{34}$ /cm^2s) and need upgrade as well, 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

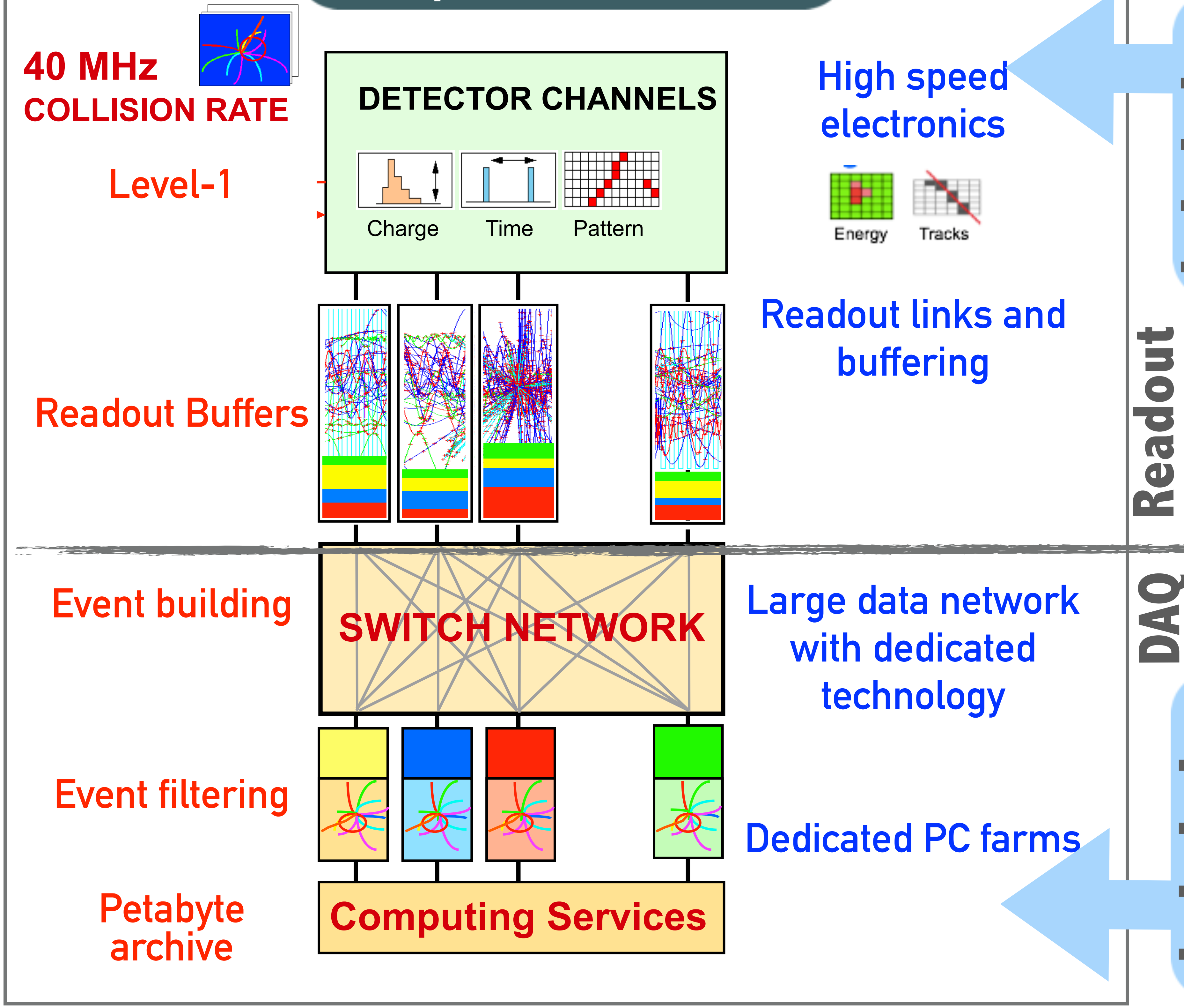


- High Luminosity with collisions close in time and space (1 collision/25ns)
  - fast electronics ➠ fast decisions
  - fine granularity detector ➠ high data volume
- Search for rare physics from hadronic collisions:
  - to store all the possibly relevant data is UNREALISTIC and often UNDESIRABLE
- Three 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



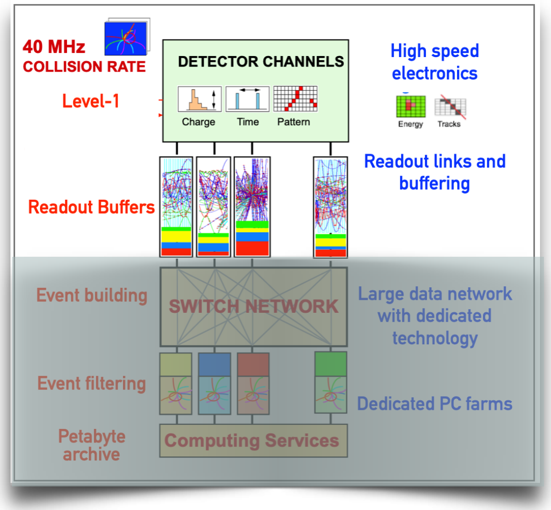
Level-1 triggers

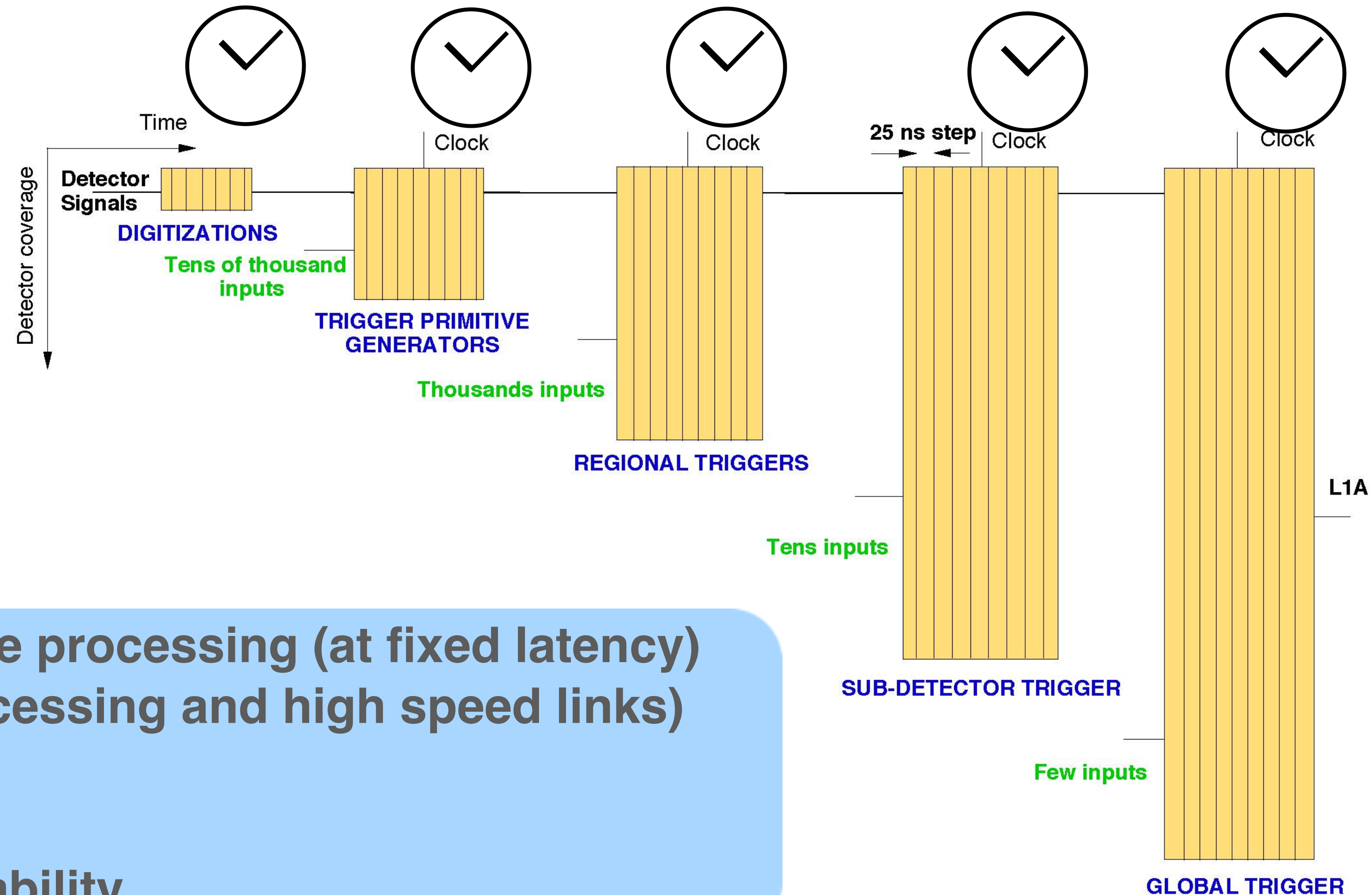
- Set max Readout rate
- Hardware, synchronous
- Readout parallelism
- Latency ~ µsec/event

Higher level triggers

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

LEVEL-1 TRIGGER PRINCIPLES





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

Full synchronisation at 40 MHz (LHC clock)

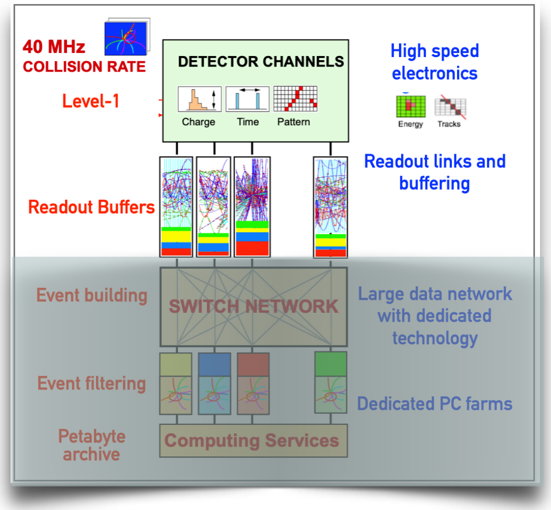
➤ large optical time distribution system

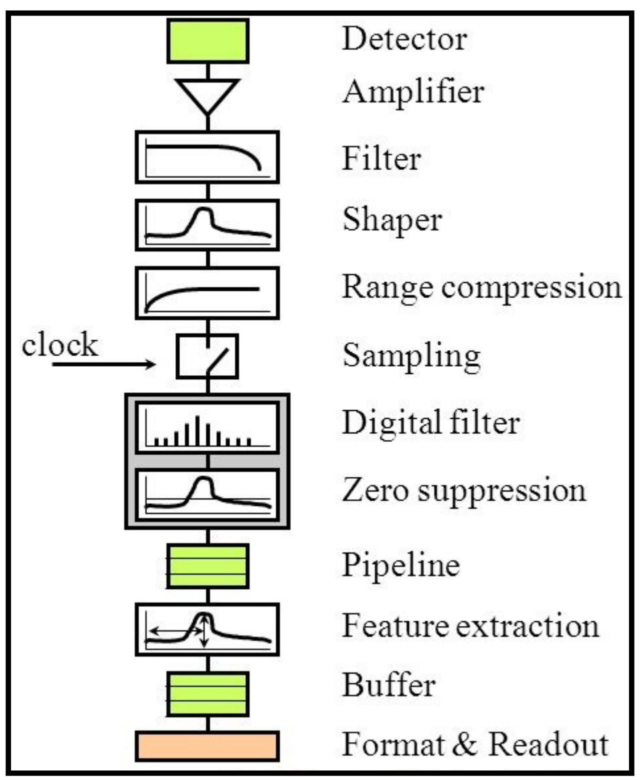
ALICE	No pipeline
ATLAS	2.5 µs
CMS	3 µs
LHCb	4 µs

Fast, robust electronics

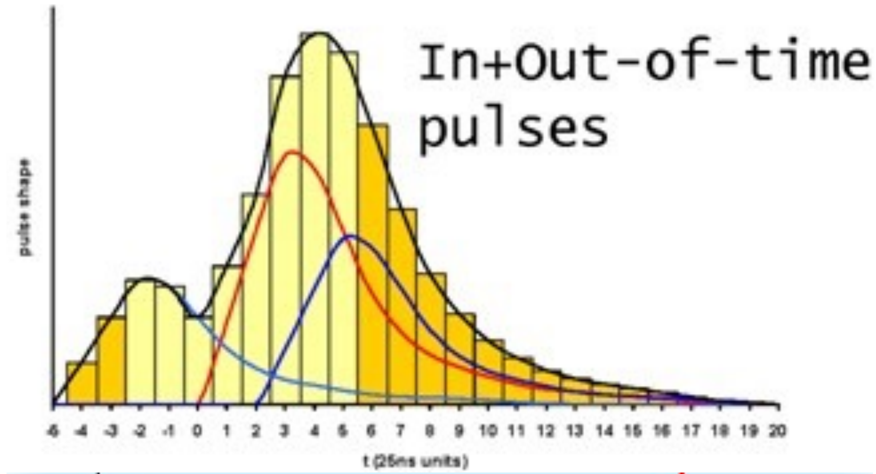
Latency dominated by cable/transmission delay

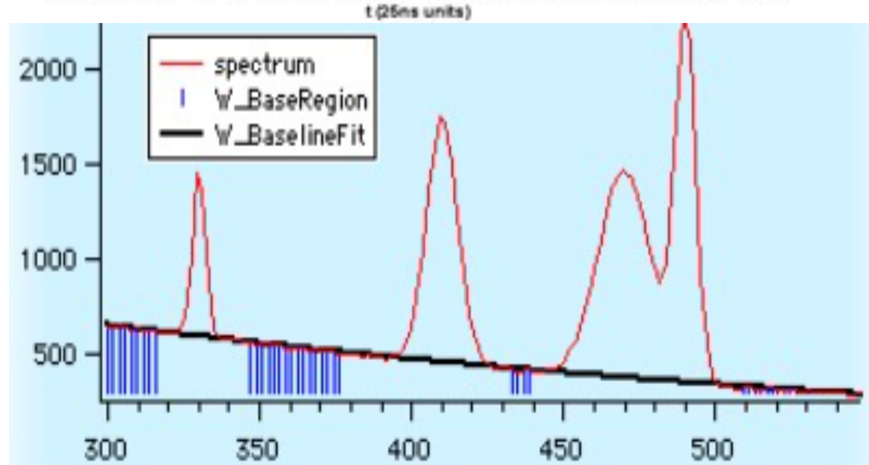
TRIGGER REQUIREMENTS ON FRONT-END ELECTRONICS



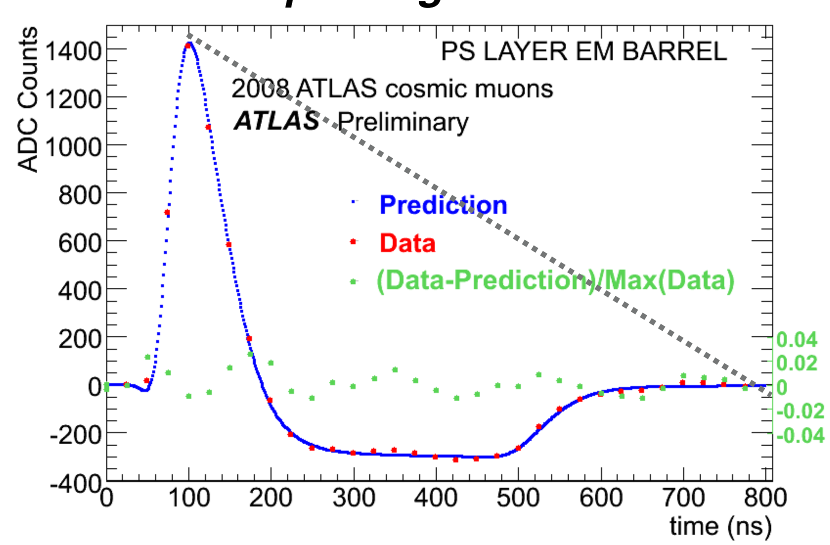


Tight design constraints for trigger & FE





ATLAS Liquid Argon calorimeter

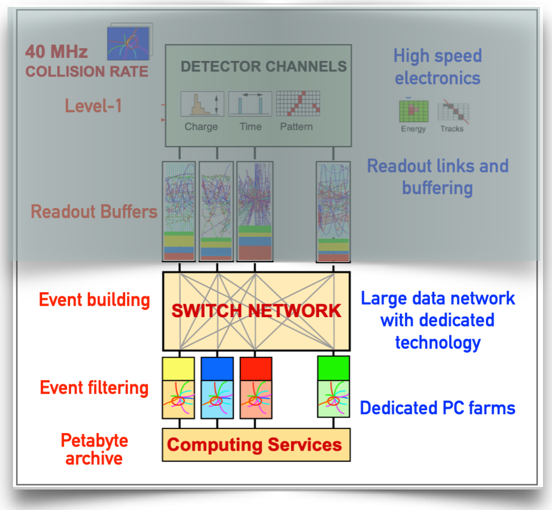


Avoid

- ➡ **Electronic pile-up**
  - ➡ source of dead-time
  - ➡ distortion in pulse
- ➡ **In-time pile-up**
  - ➡ more collisions/BC
  - ➡ Baseline subtraction
- ➡ **Out-of-time pile-up**
  - ➡ BC-identification capability
  - ➡ peak finder algorithms

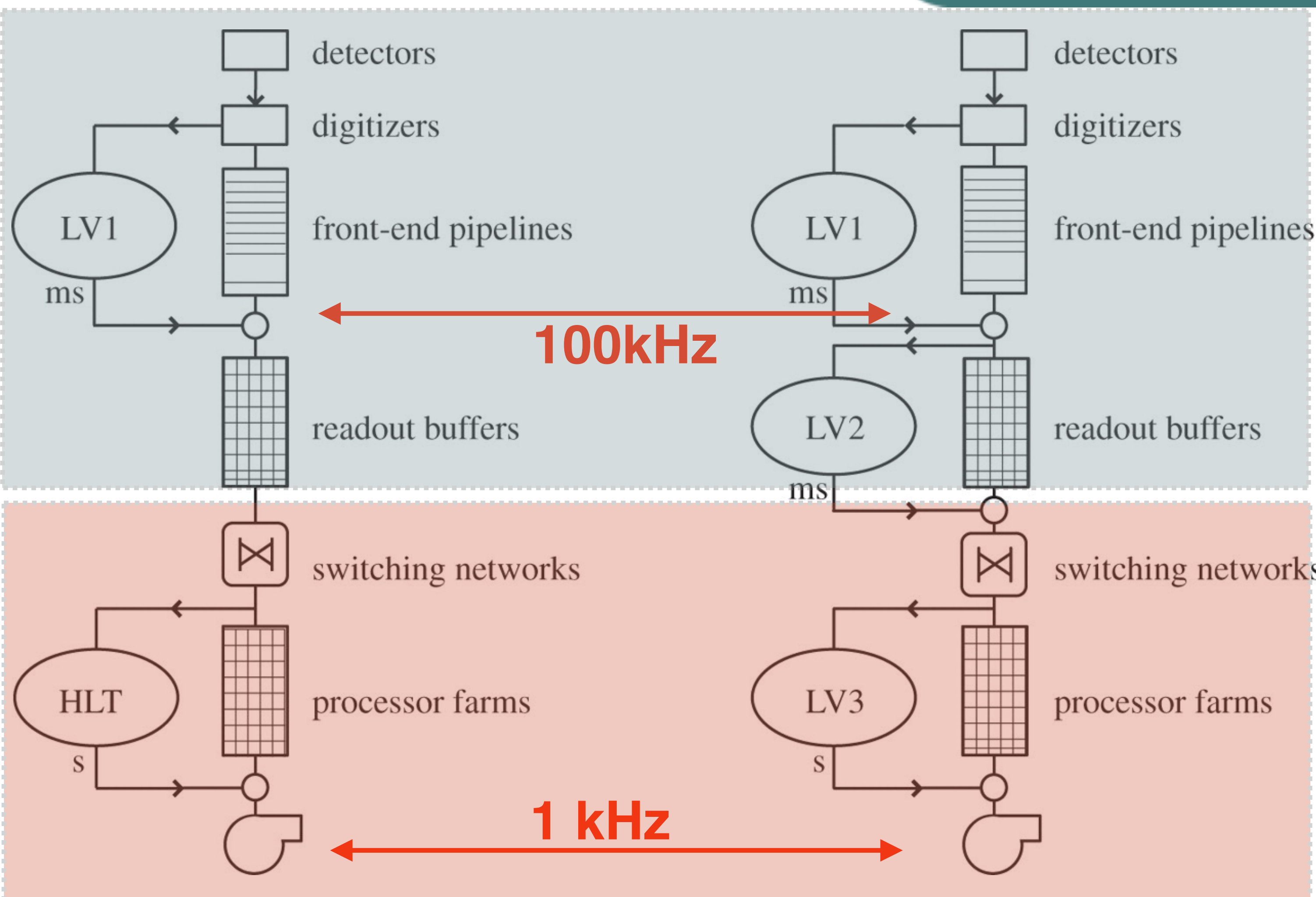
Make it easier with fast, low occupancy and digital detectors

HLT/DAQ REQUIREMENTS



- ➡ 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



DAQ+HLT system

ATLAS/CMS Example

- ➡ **1 MB/event at 100 kHz for O(100ms) HLT latency**
  - ➡ Network: 1 MB*100 kHz = **100 GB/s**
  - ➡ HLT farm: 100 kHz*100 ms = **O(10^4) 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

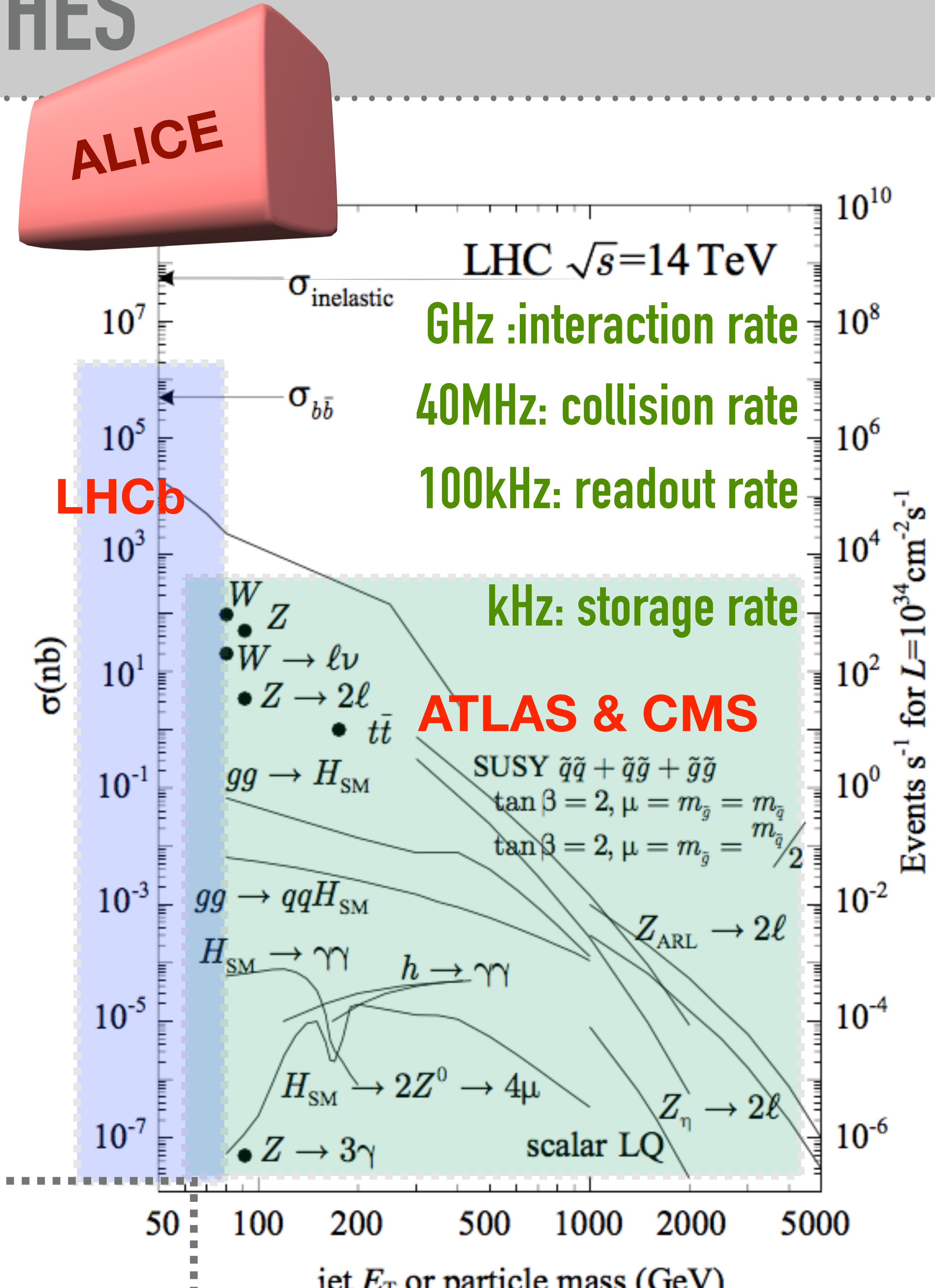
COMPARE 4 EXPERIMENTS

How to maximise physics acceptance

DIFFERENT PHYSICS SEARCHES

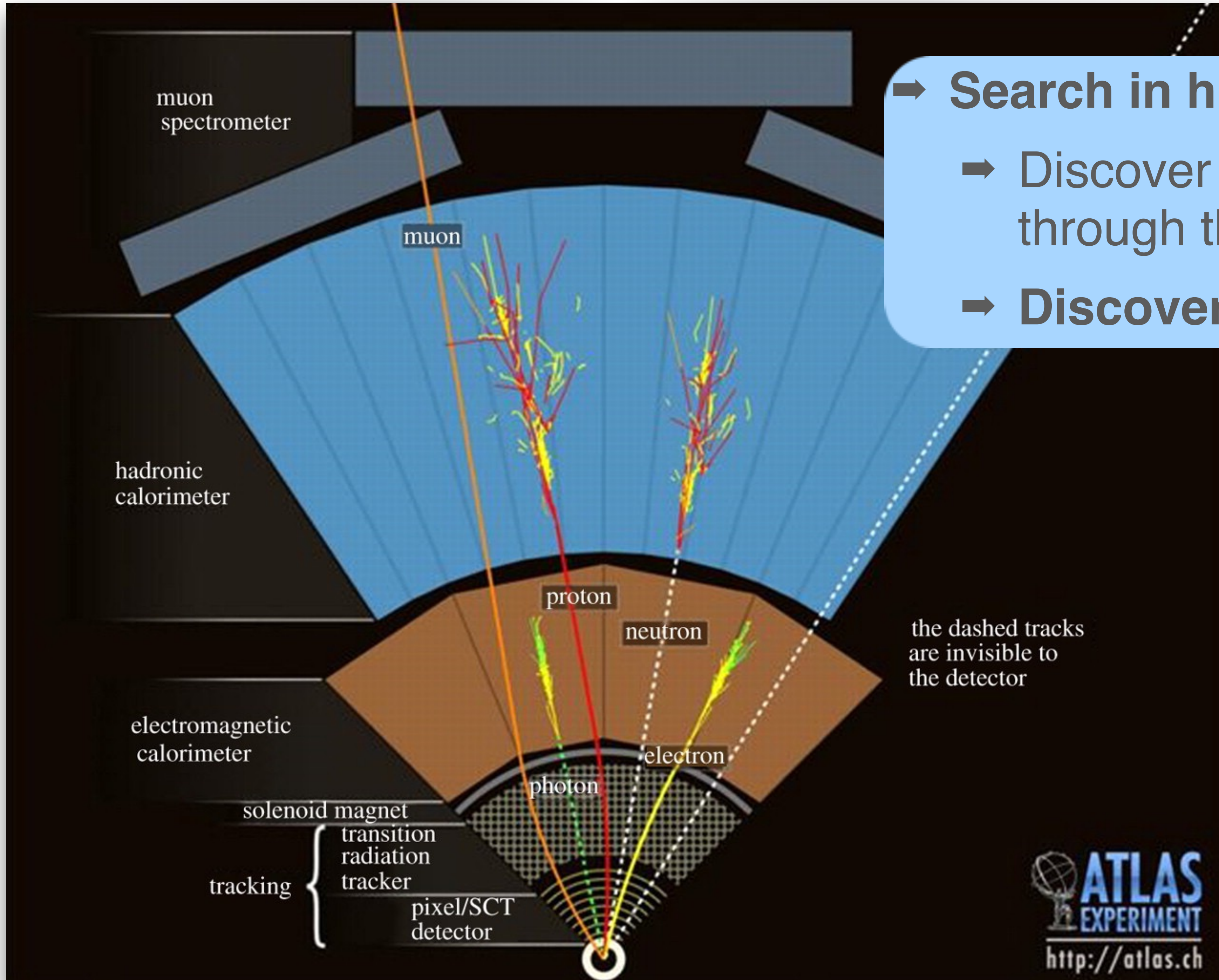
…. and LHC operations

- ✦ **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



- ➡ Expected rates and S/B ratio
- ➡ Signal topology, 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

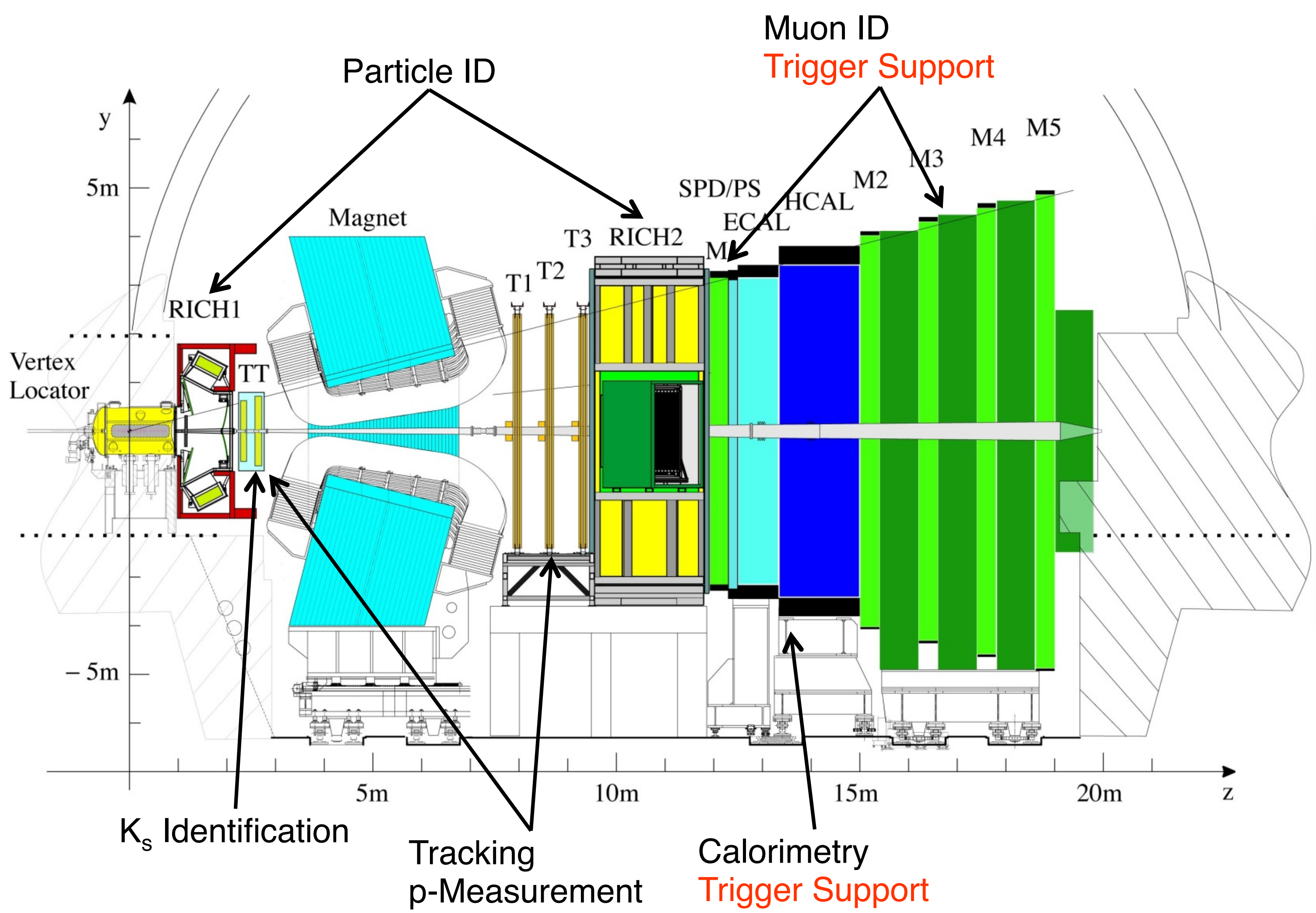
$$\frac{\sigma_{tot}}{\sigma_{H(500\,\mathrm{GeV})}} \approx \frac{100\,mb}{1\,pb} \approx 10^{11}$$

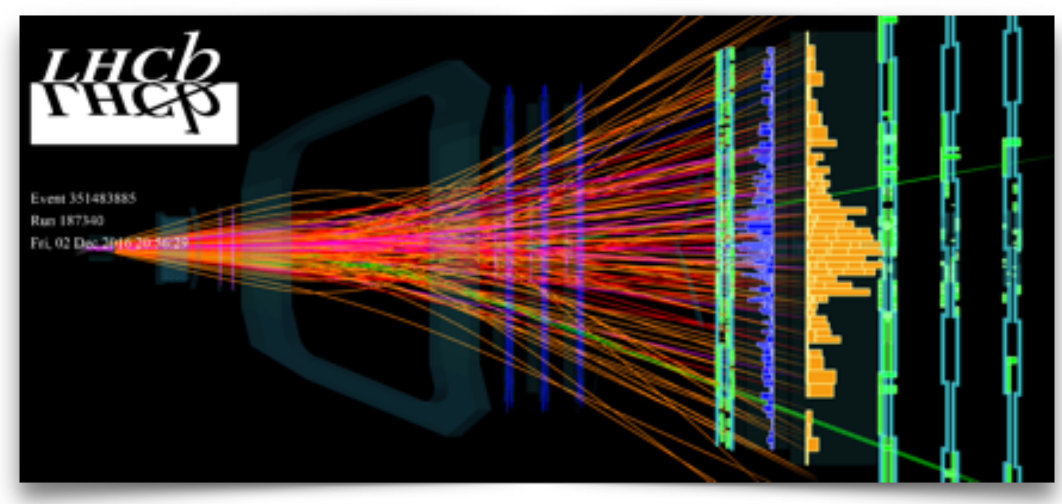
approximately
10^6 rejection

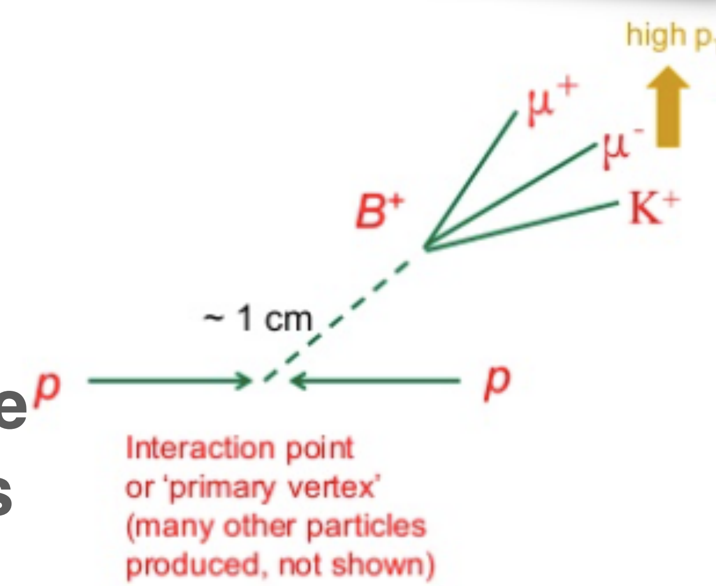
- **Easy selection of signal over background, with high-energy leptons ==> @L1**
  - Against thousands of particles/collisions (typically low momentum jets)
- **Remember: 90M readout channels and full Luminosity ==> 1 MB/event**

LHCB DESIGN PRINCIPLES

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

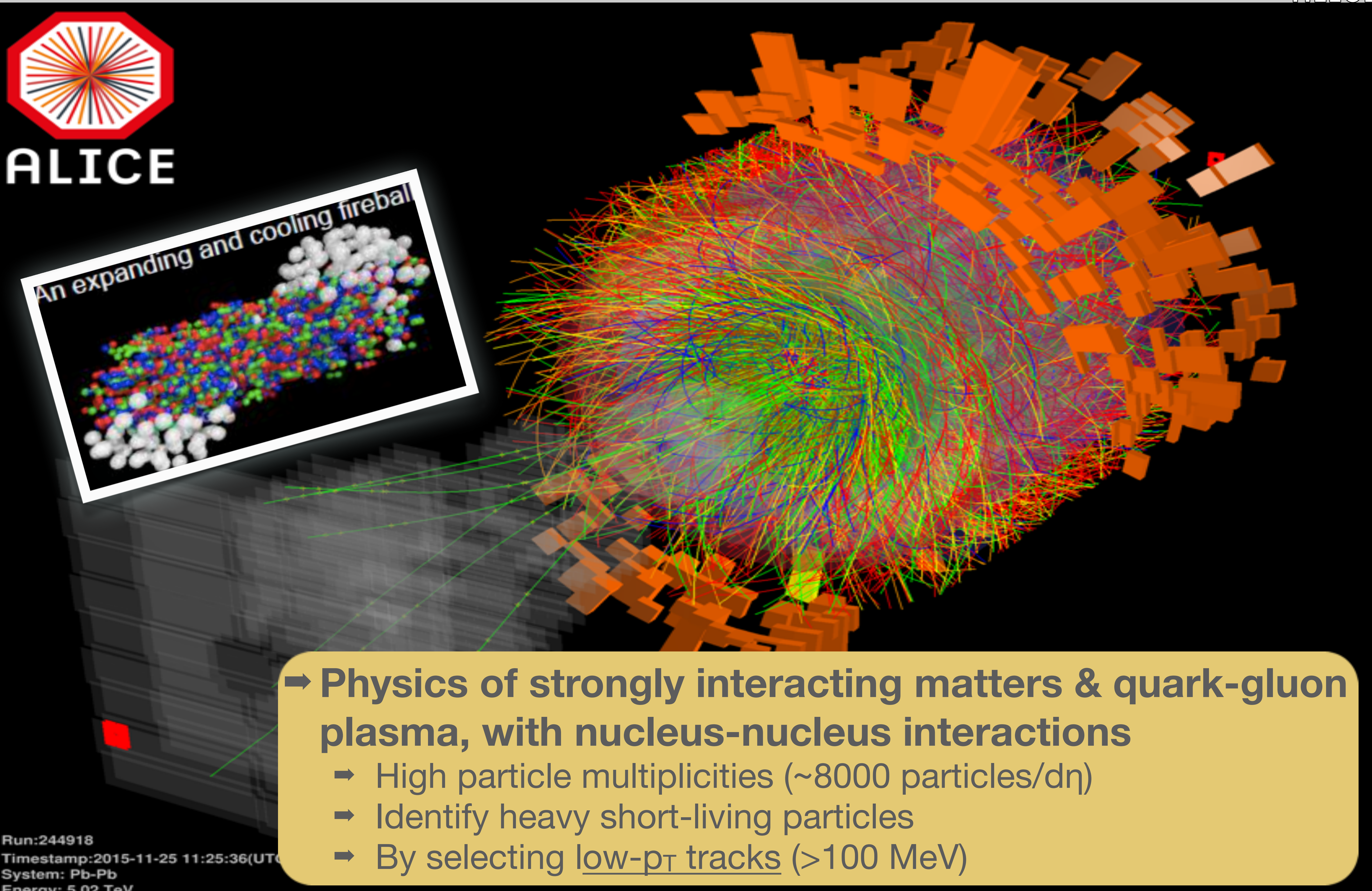






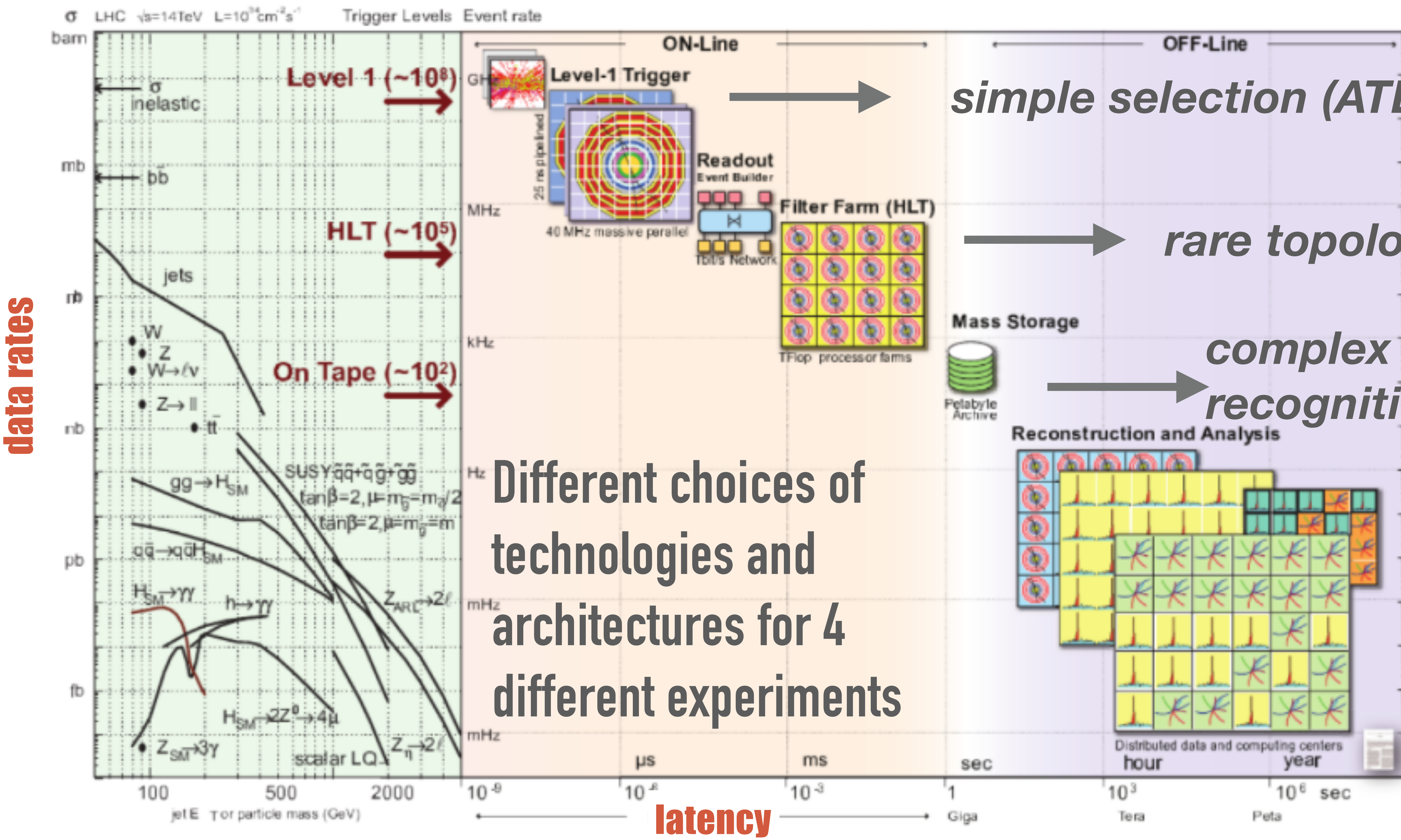
- **Single-arm spectrometer and low L ==> reduced event size**
- **Selection of B mesons ==> in different B-decay topologies**
  - related to high mass and long lifetime of the b-quark

ALICE STRATEGIES



- **Physics of strongly interacting matters & quark-gluon plasma, with nucleus-nucleus interactions**
  - High particle multiplicities (~8000 particles/dη)
  - 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)

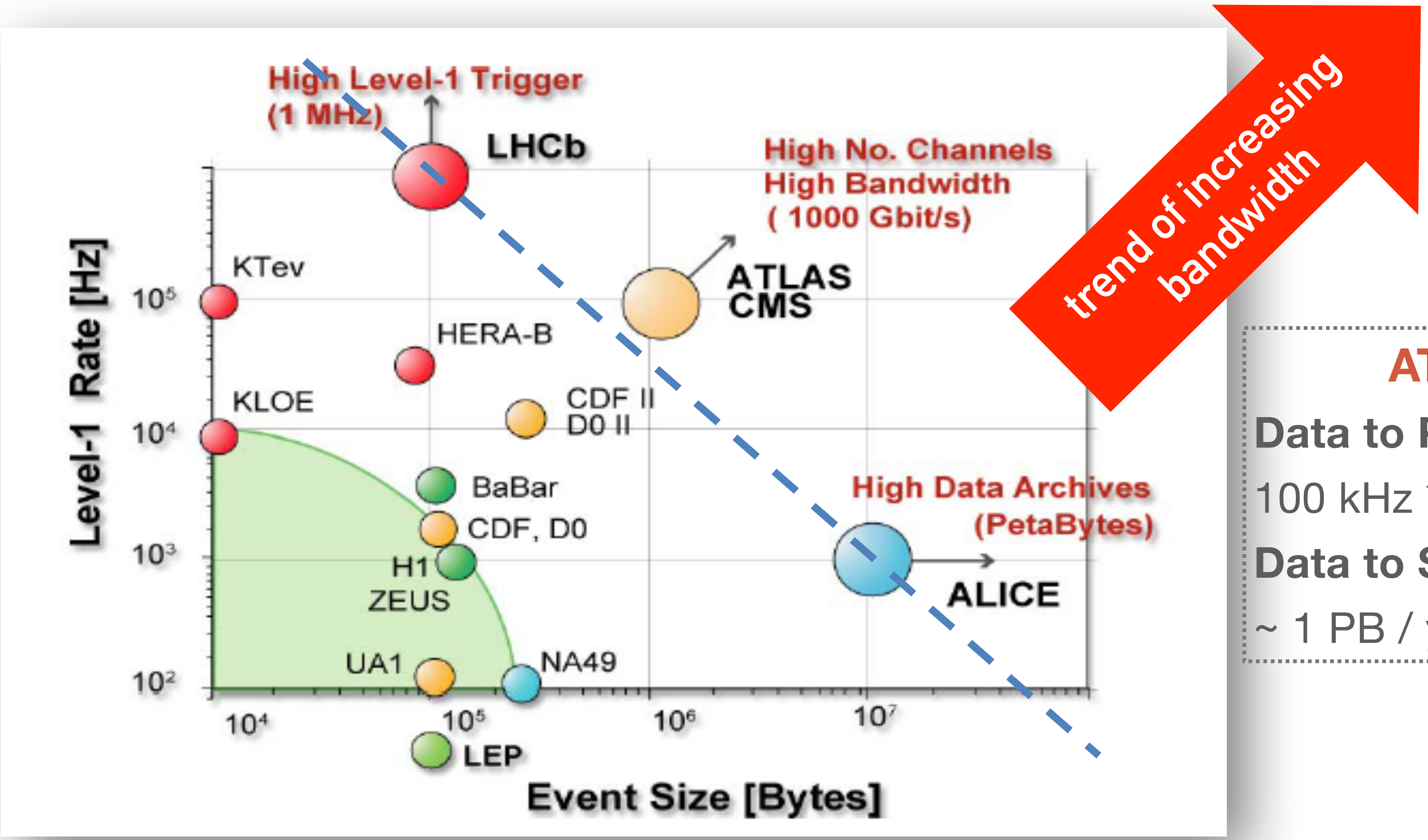
Different choices of technologies and architectures for 4 different experiments

- ➡ **ATLAS/CMS: Trigger power:** reducing the data-flow at the earliest stage
- ➡ **ALICE/LHCb: Large data-flow:** since trigger selectivity is not enough, due to the large irreducible background

READOUT AND DAQ THROUGHPUTS

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

faster L1 electronics



more channels, more complex events

ATLAS/CMS

Data to Process:

100 kHz * 1 MB = 100 GB/s

Data to Store:

~ 1 PB / year /experiment

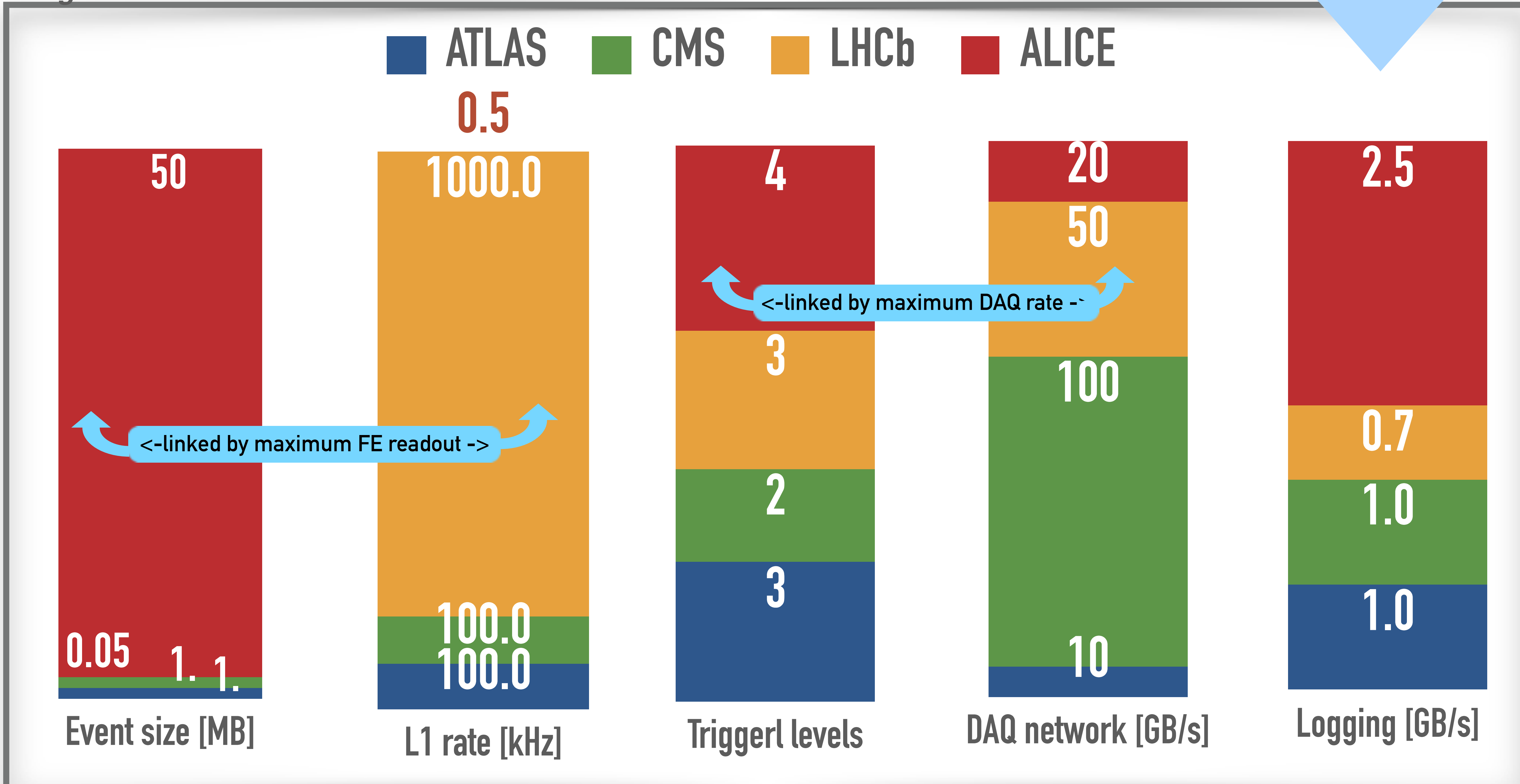
As the data volumes and rates increase, new architectures need to be developed

COMPARING BY NUMBERS

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

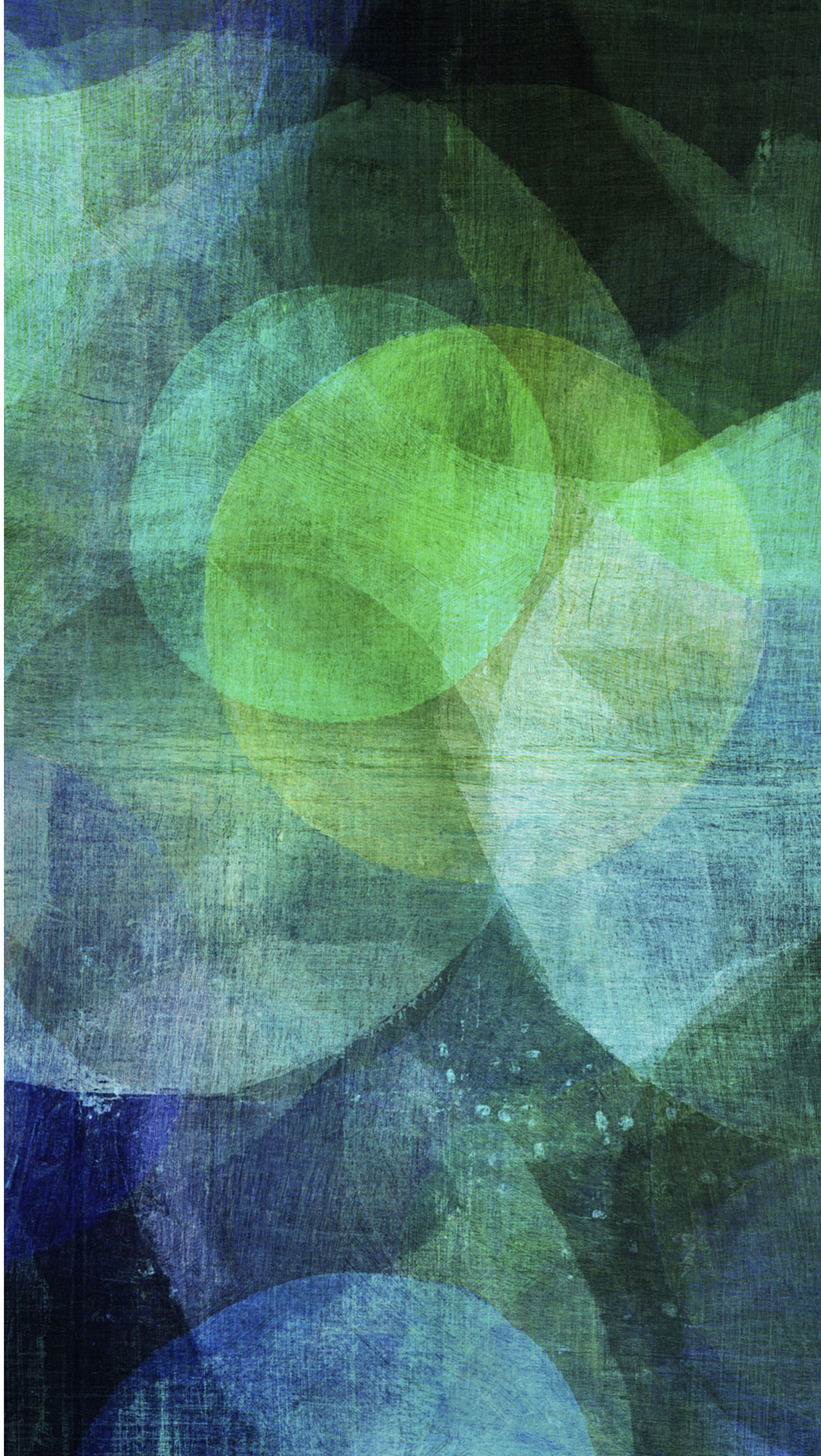
Allowed storage and processing resources

Design values in 2009



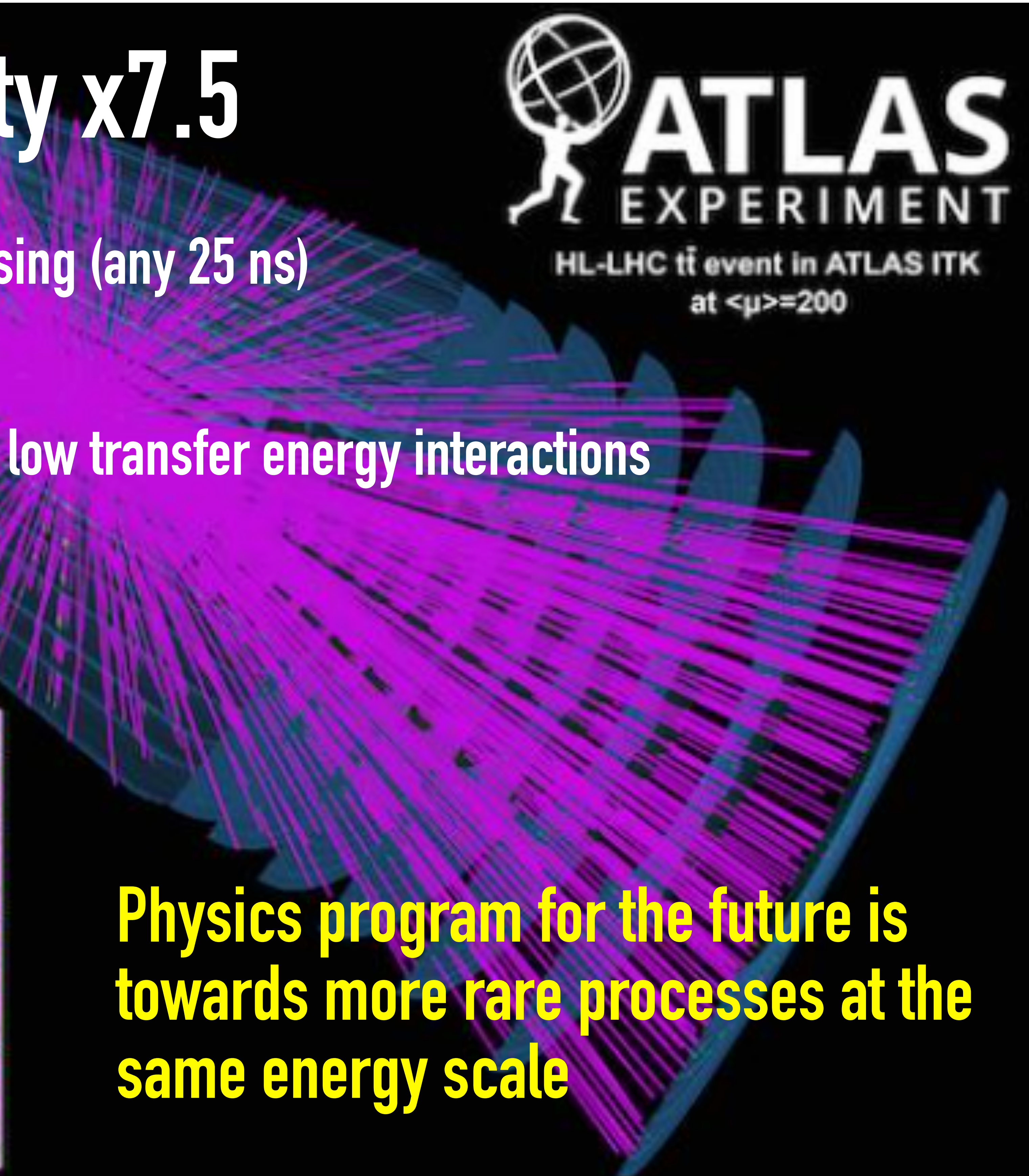
FUTURE TRENDS FOR HIGH-LUMINOSITY

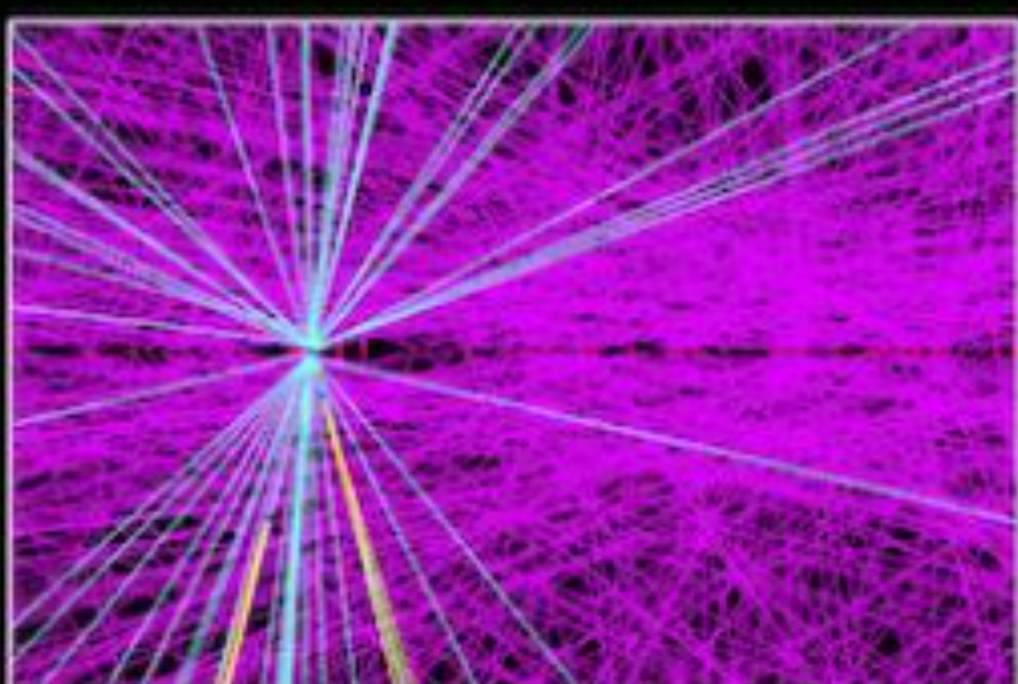
What about … tomorrow?



Design Luminosity x7.5

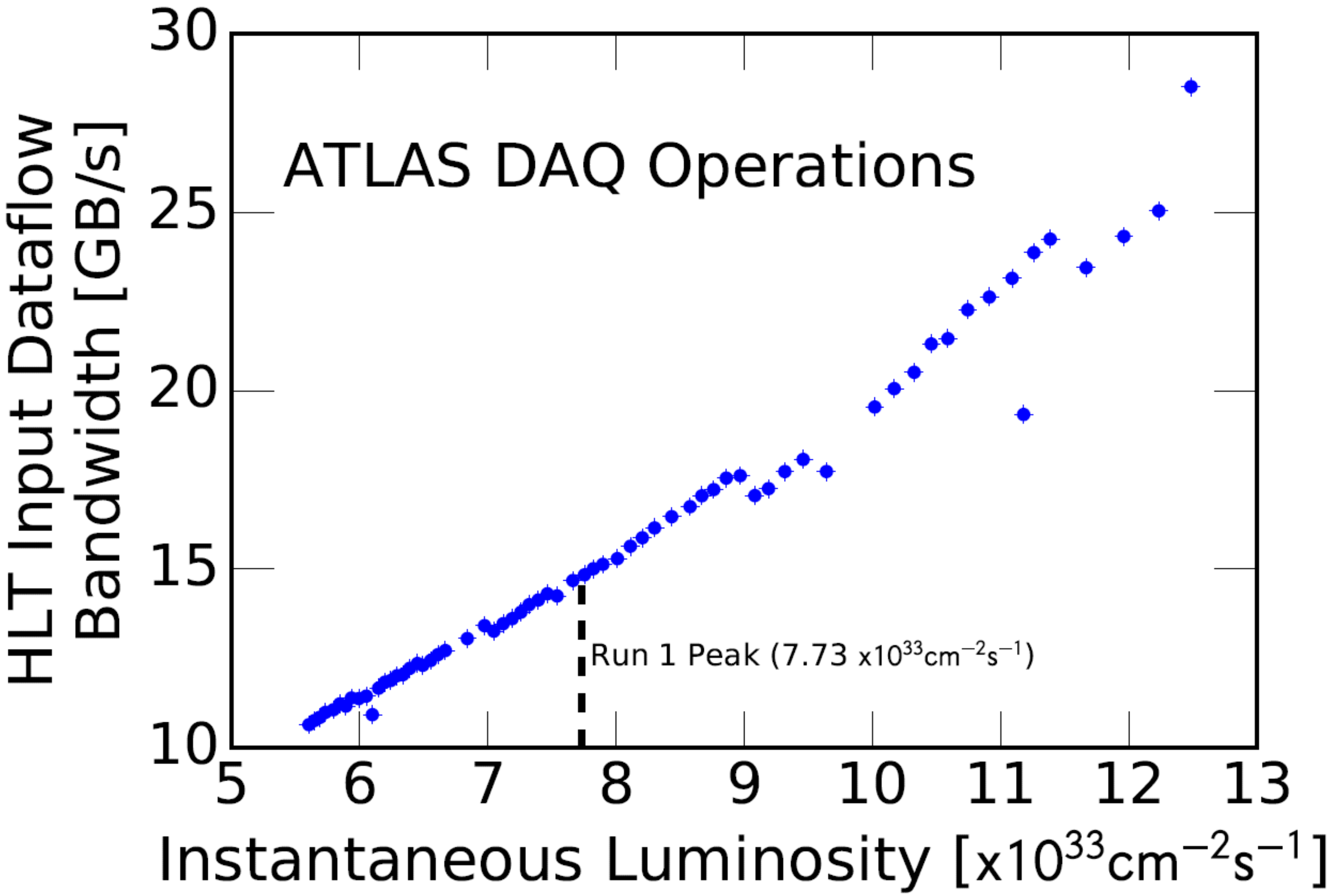
- ➡ 200 collisions per bunch crossing (any 25 ns)
- ➡ ~ 10 000 particles per event
- ➡ Mostly low p_T particles due to low transfer energy interactions



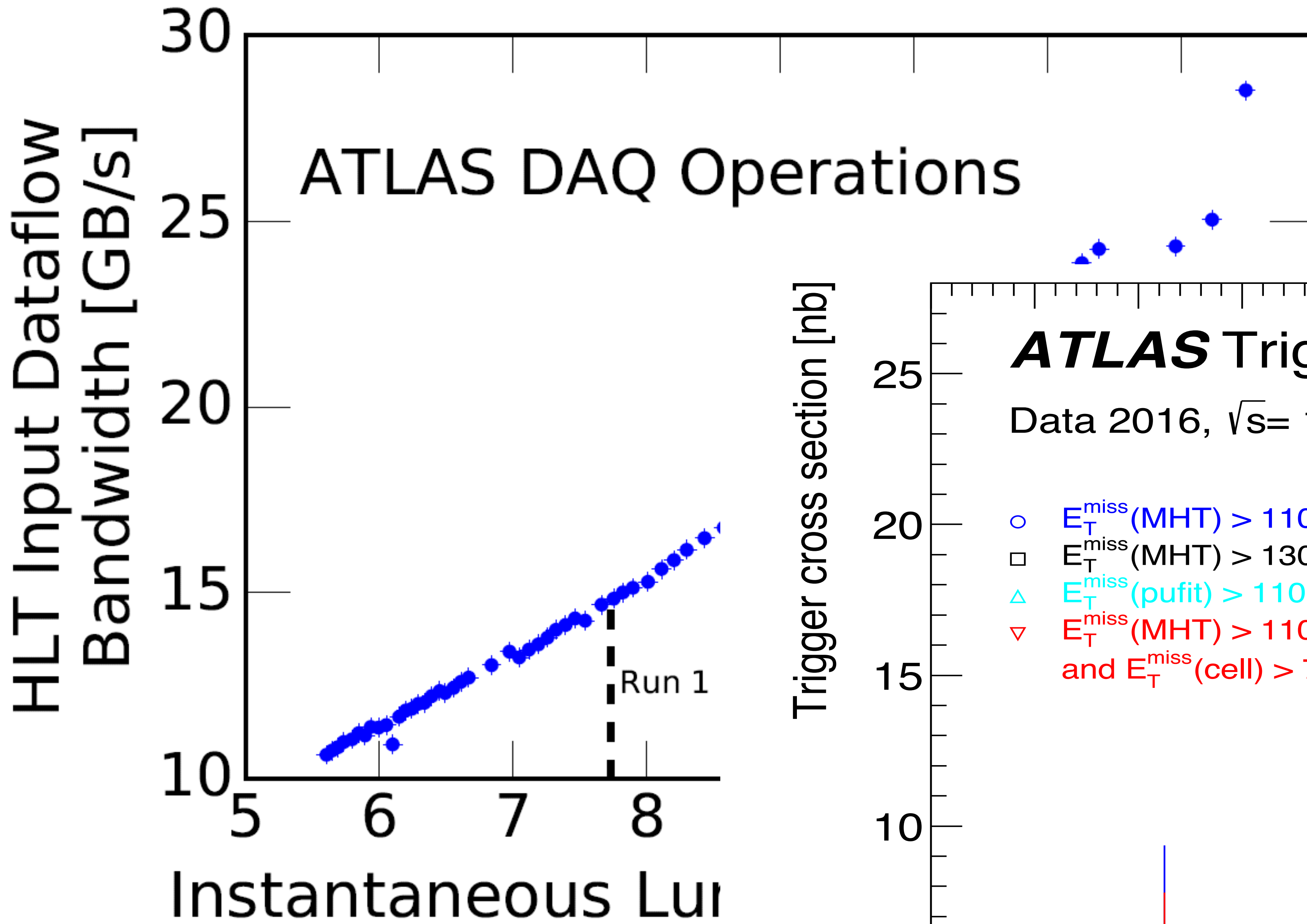


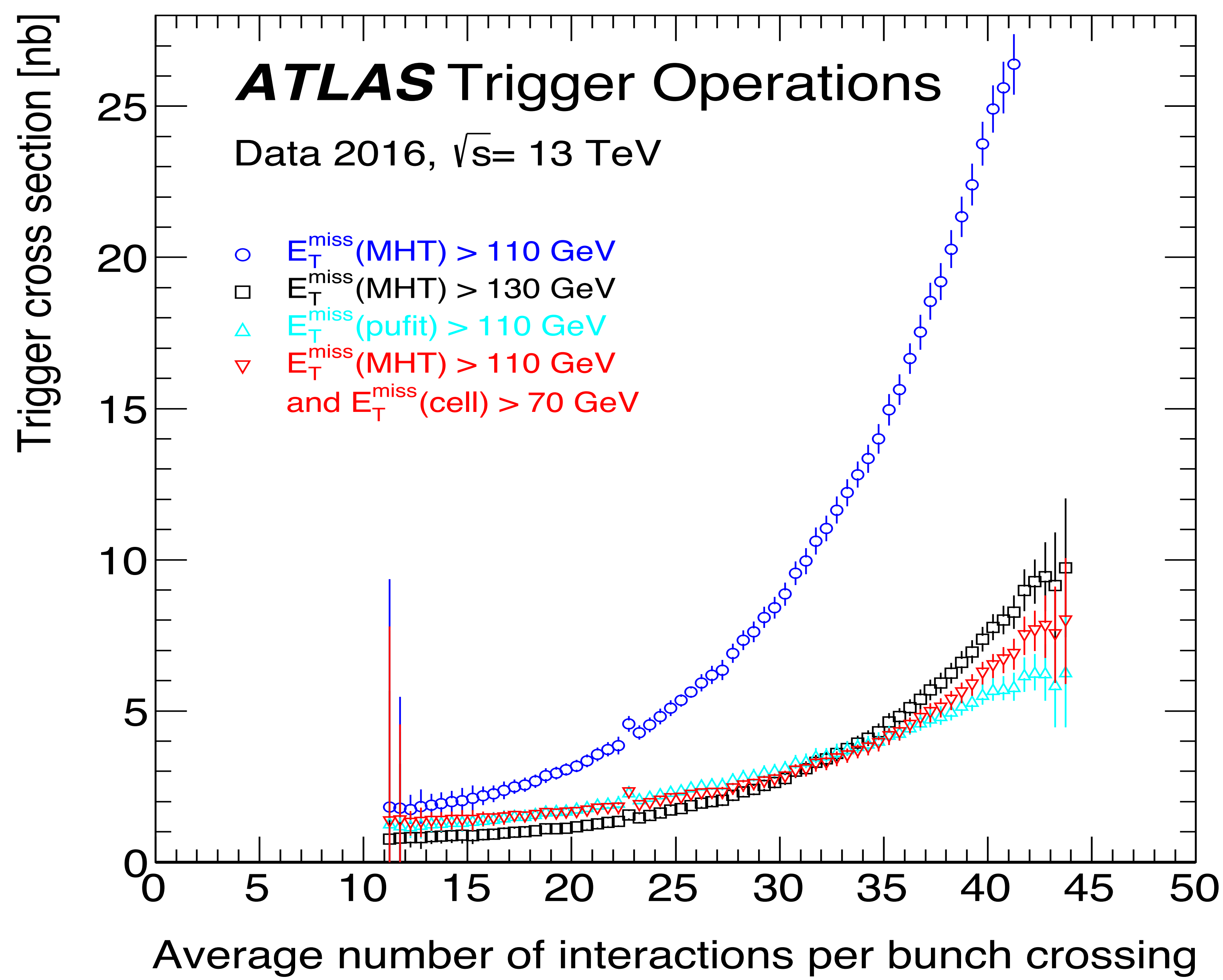
Physics program for the future is towards more rare processes at the same energy scale

WHAT DO YOU EXPECT FOR THE FUTURE?



WHAT DO YOU EXPECT FOR THE FUTURE?





- ➡ **Very large uncertainties to take into account!**

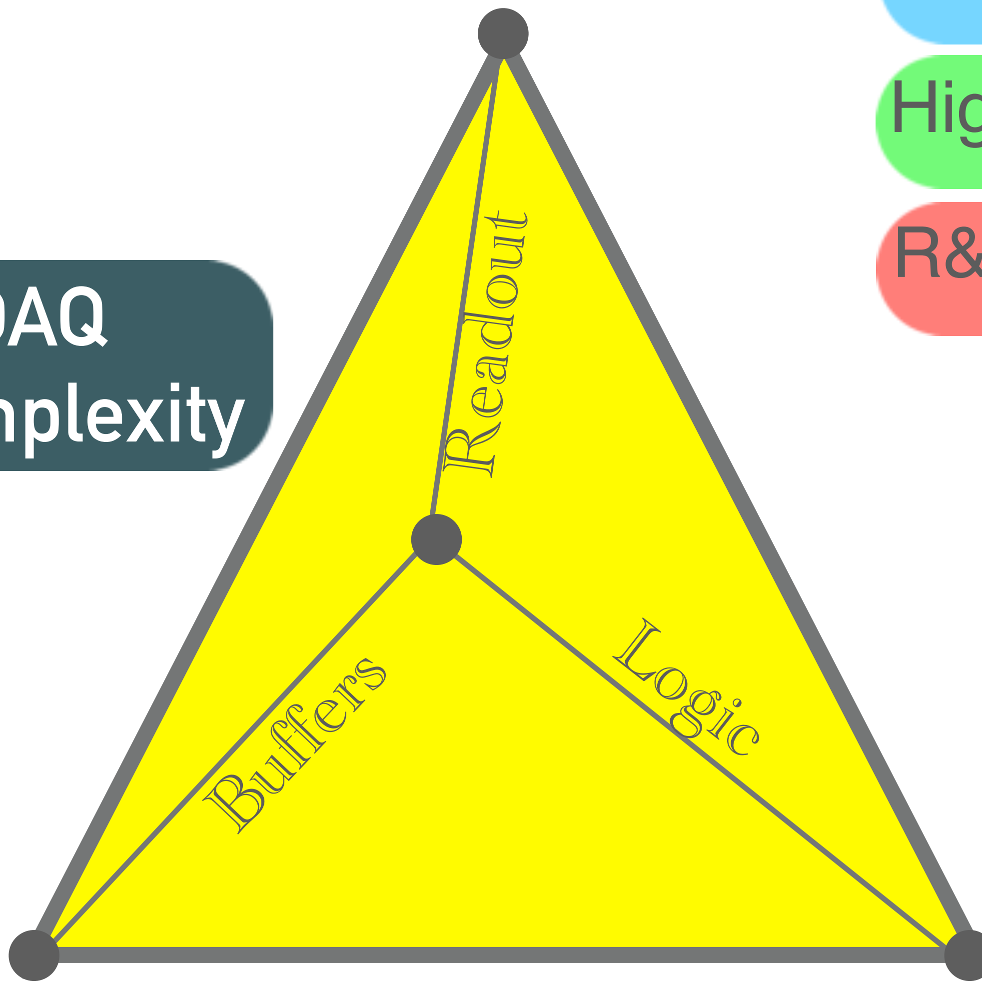
BE SMARTER! INCREASE RESOLUTION FOR BETTER S/B

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

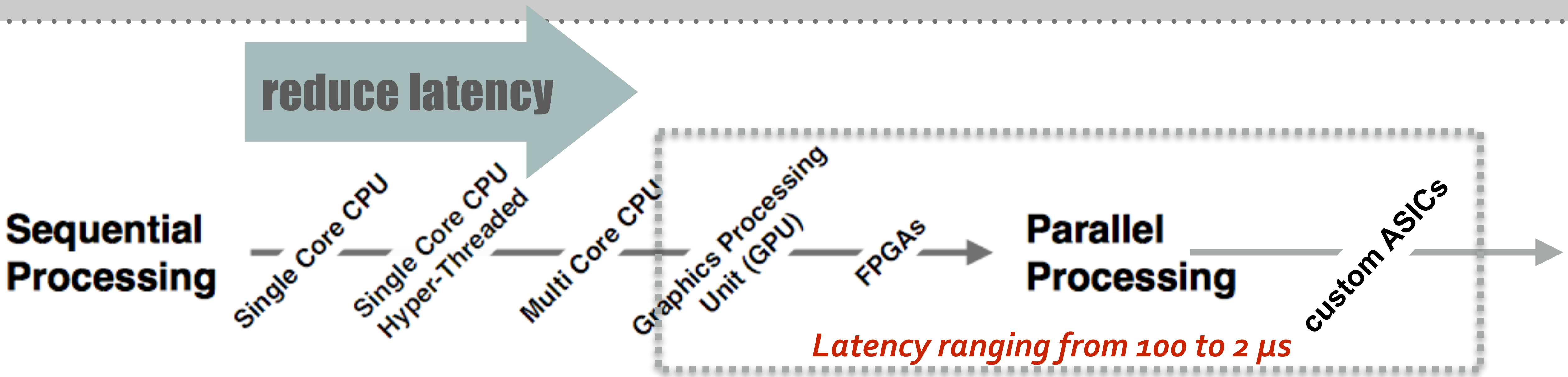
- Refine calibrations, as offline
- large buffers, long latency
- tight: offline=online (LHCb, ALICE)
  soft: decouple trigger/DAQ (ATLAS, CMS)

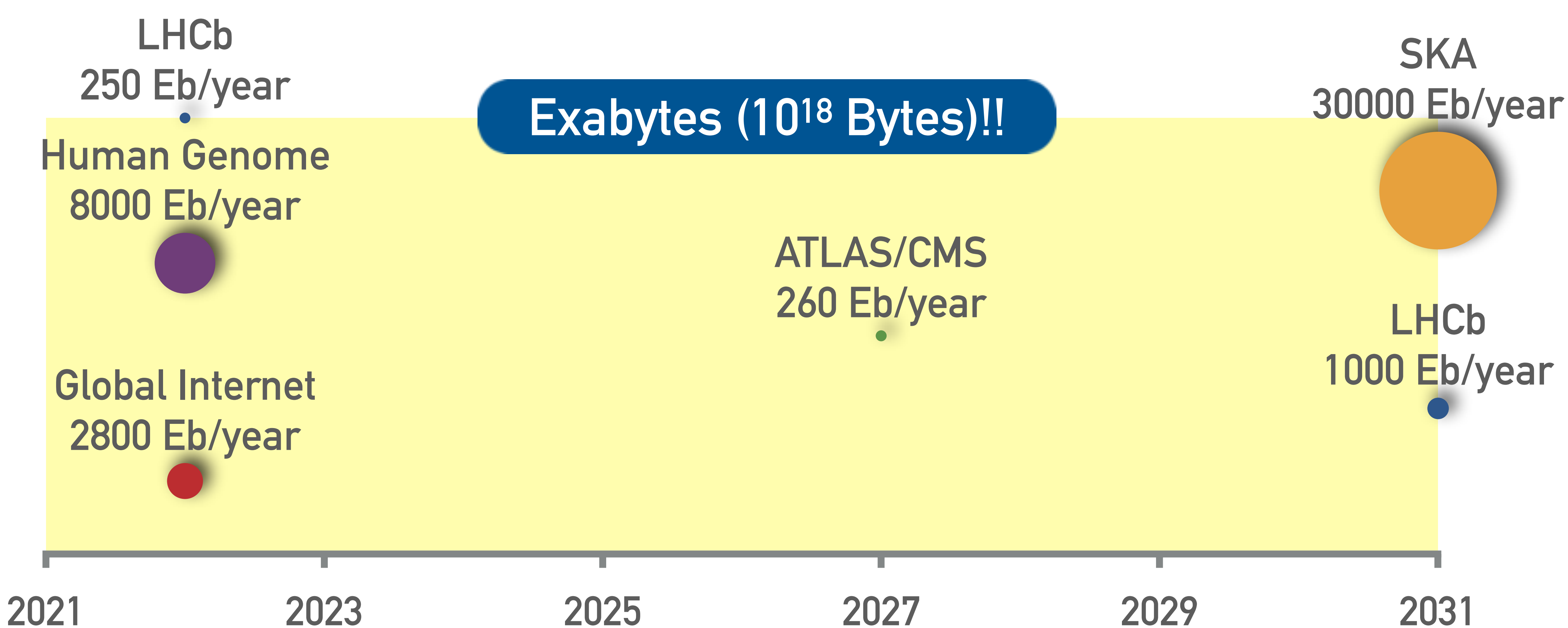
Triggering detectors

- complex ASIC logic
- trigger-driven design
- Hardware Track trigger (CMS)

LHCP-2022

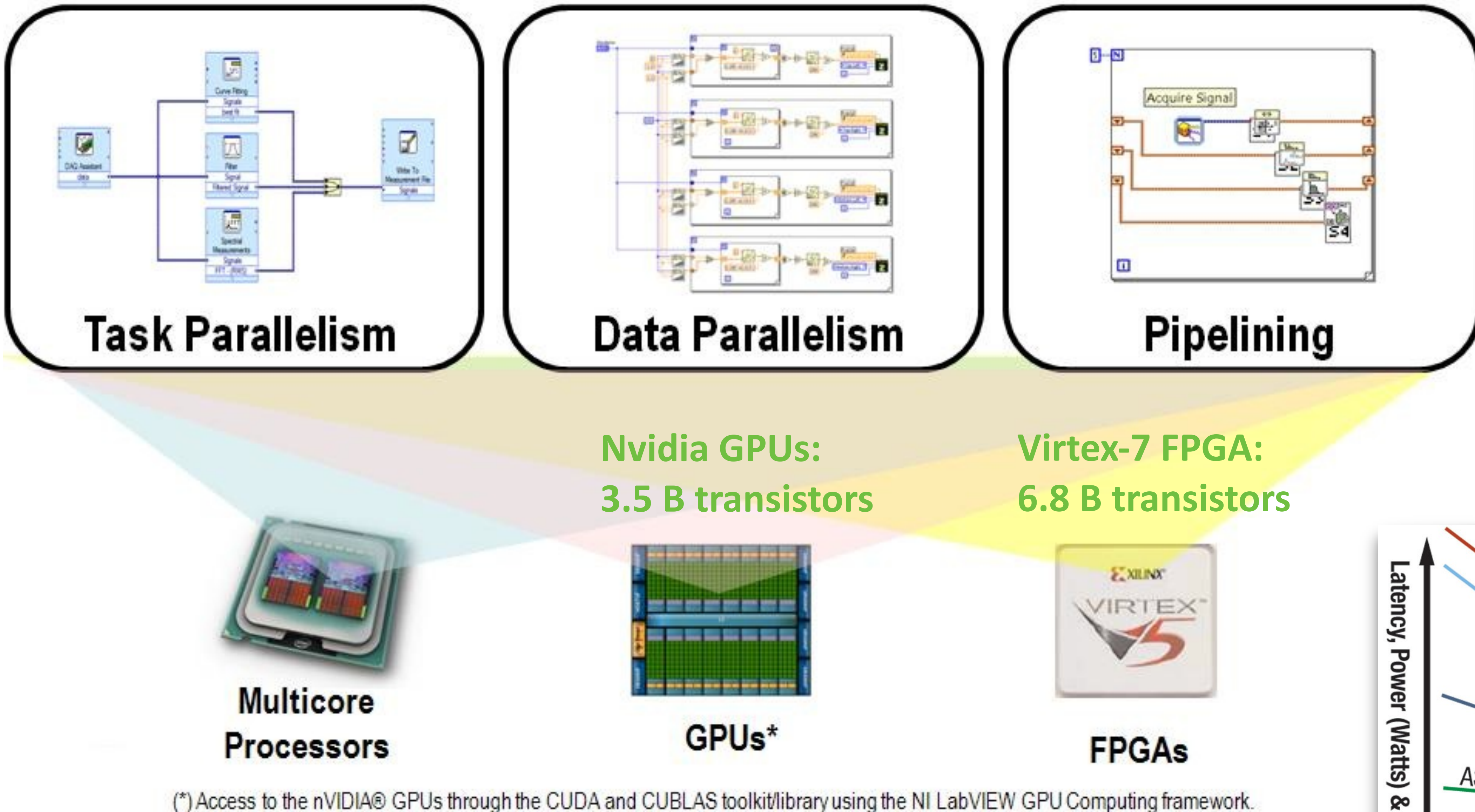
THE REAL-TIME ADVENTURE

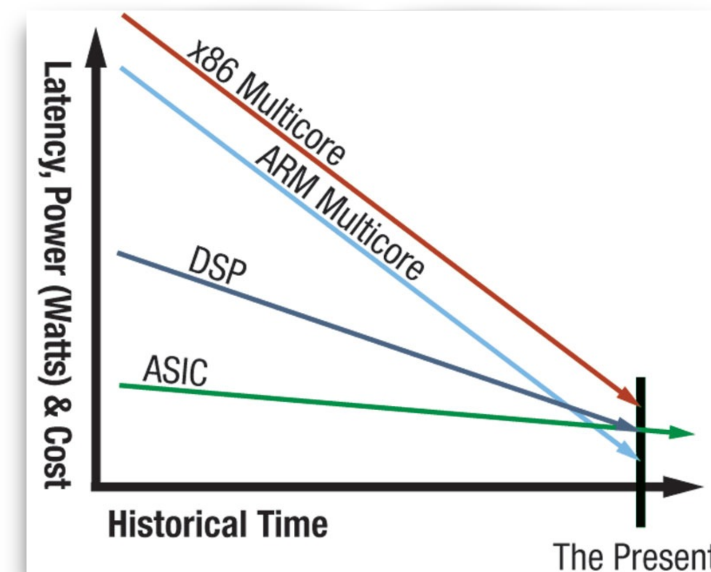




See Openlab workshop

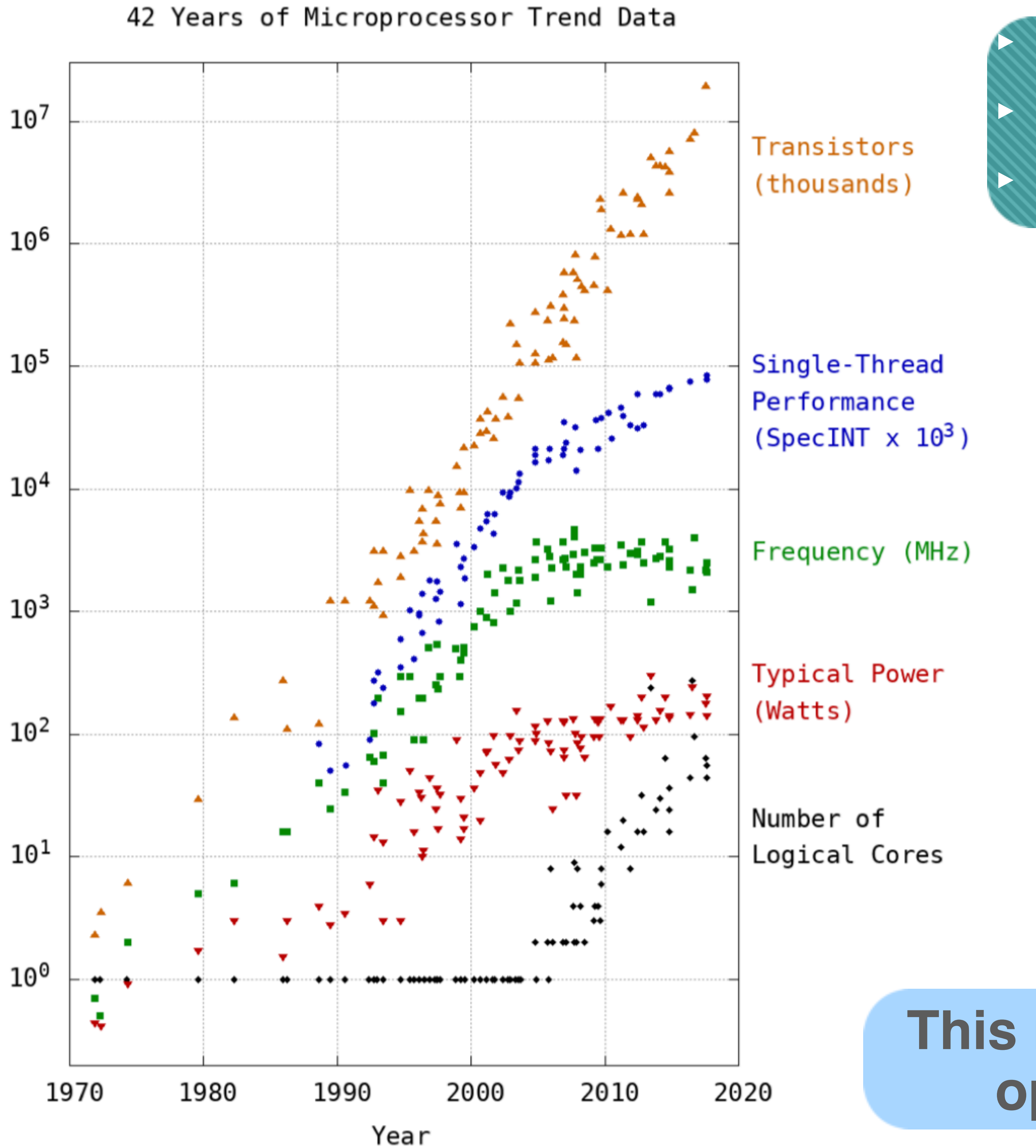
TRENDS: COMBINED TECHNOLOGY





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

TRIGGER SOFTWARE EVOLUTION TO BREAK WALLS



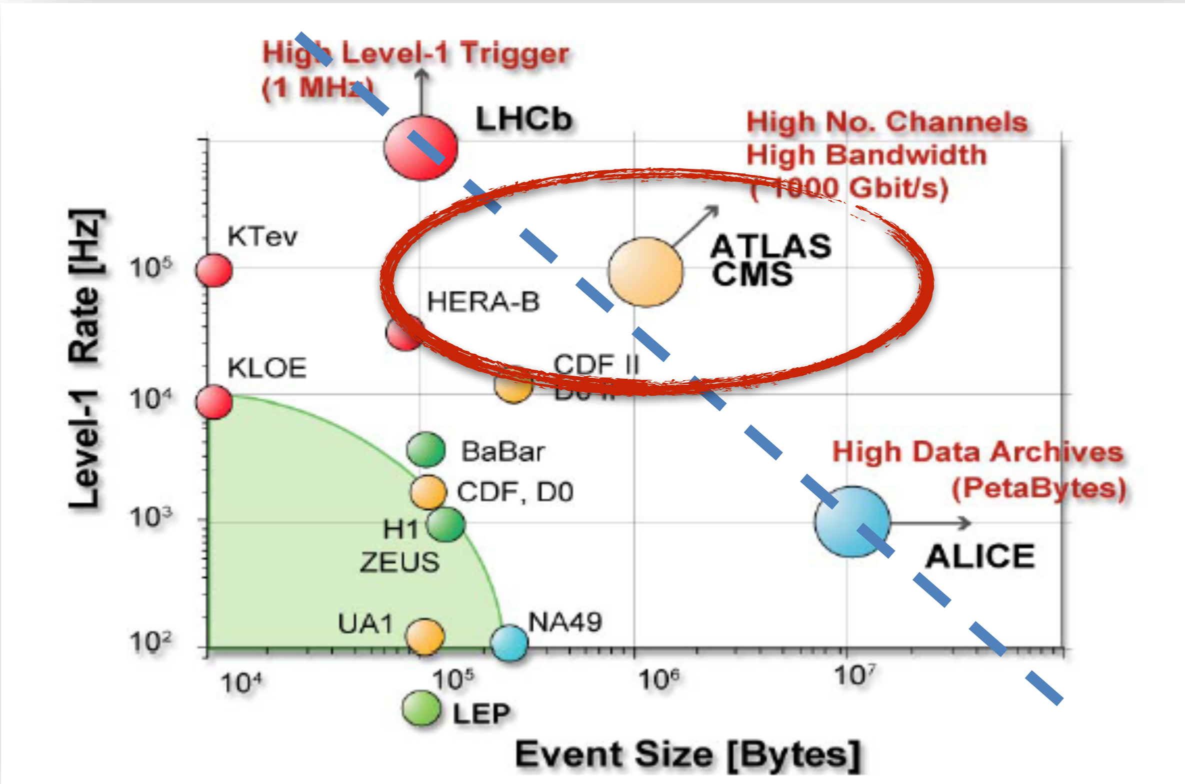
Data Source: https://github.com/karlrupp/microprocessor-trend-data

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

- ➡ **Exploiting CPU h/w, more complicated programming**
  - ➡ vectorisation, low-level memory…
- ➡ **Multithreading processing**
  - ➡ to reduce memory footprint
- ➡ **Use of co-processors, like GPUs:**
  - ➡ High Performance Computing (HPC) often employ GPU architecture to achieve record-breaking results!
  - ➡ data reduction (ALICE/LHCb)
  - ➡ trigger selection (CMS/ATLAS)

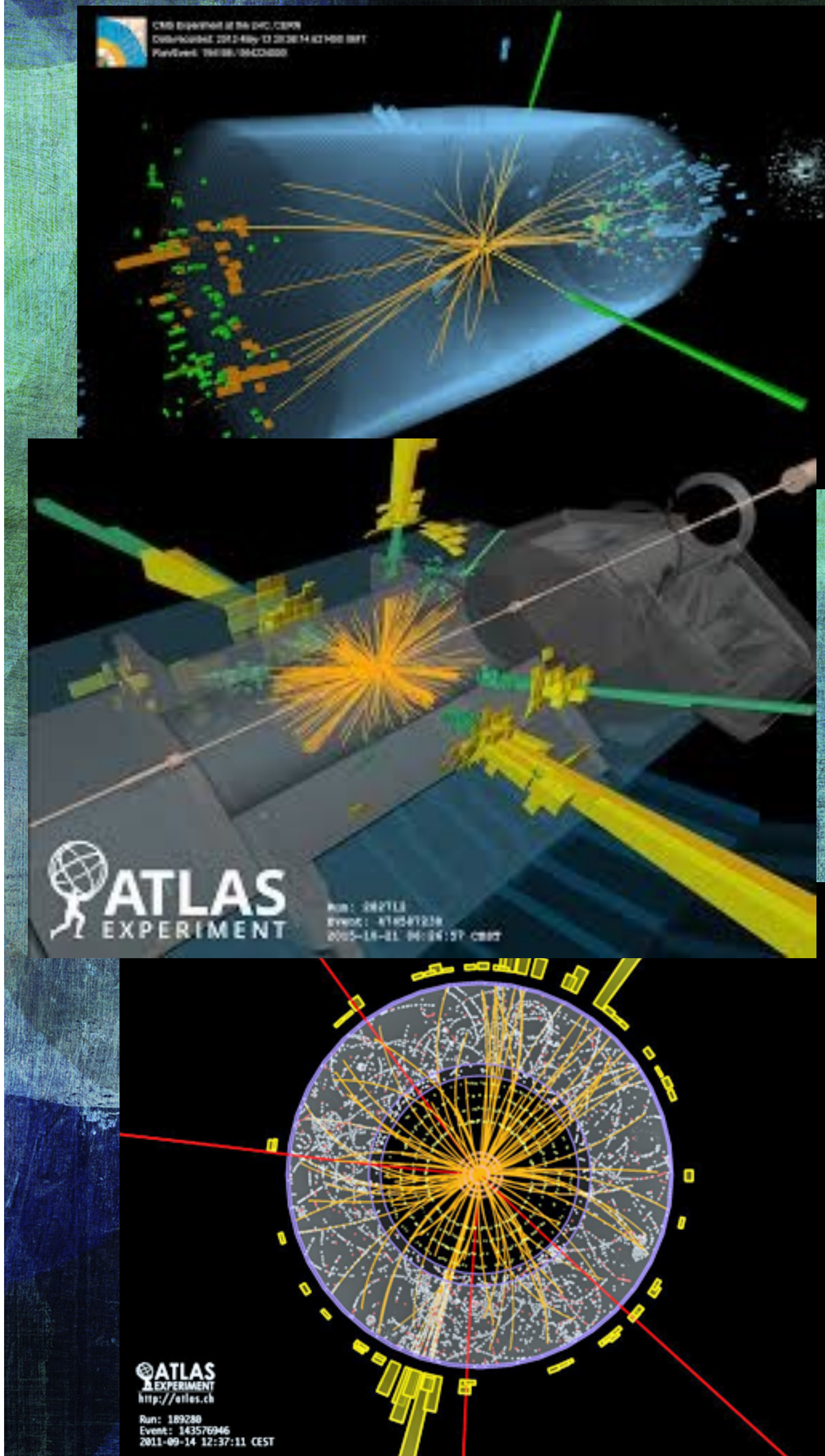
This requires fundamental re-write/ optimization of our software

Read: HPC computing



ATLAS AND CMS

Studying the Standard Model at the high energy frontier

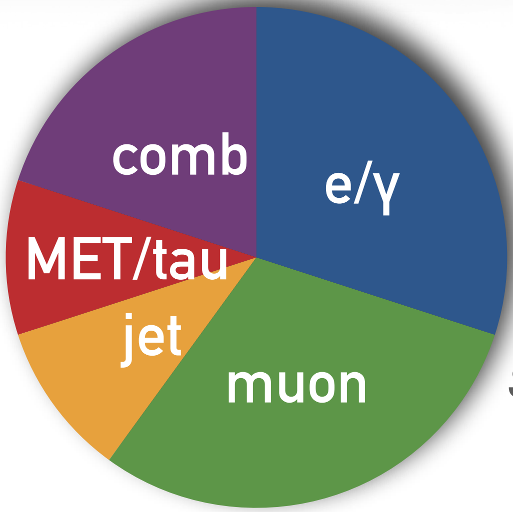


ATLAS & CMS DESIGN PRINCIPLES

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



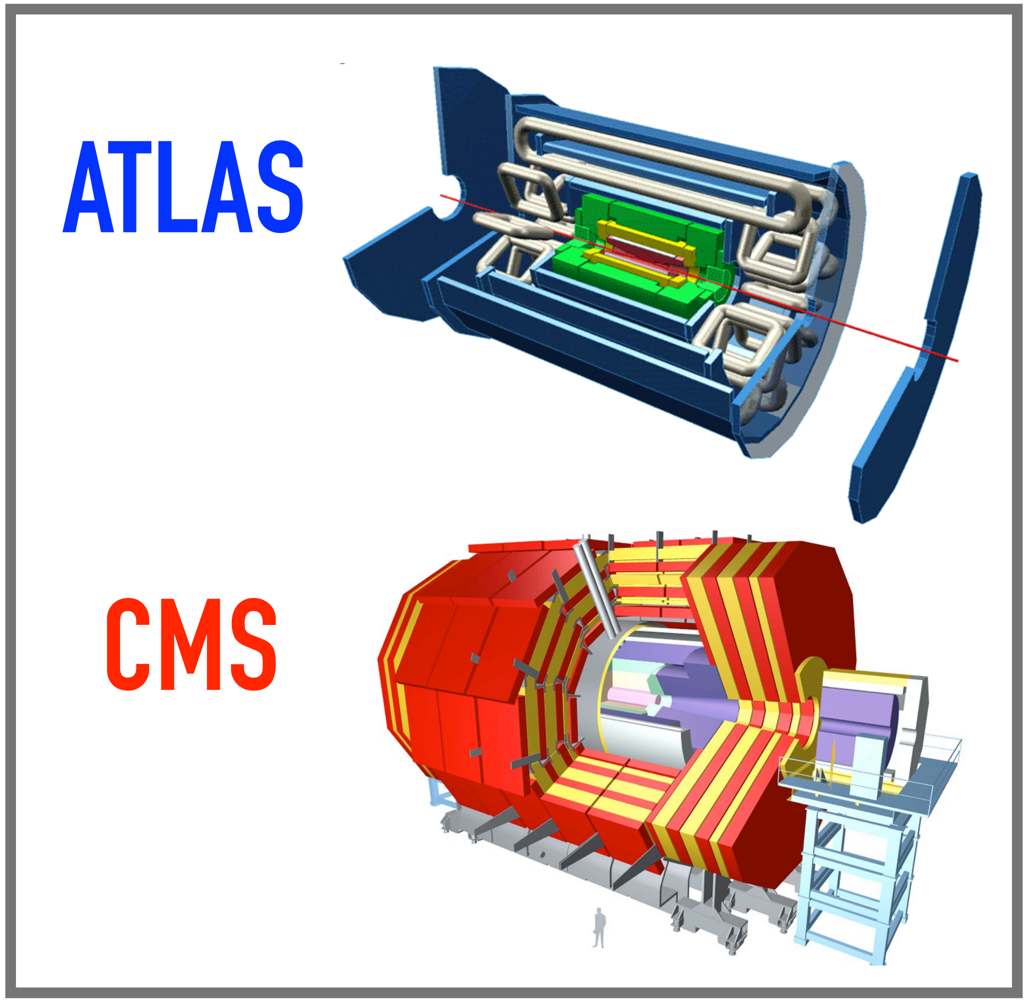
inclusive trigger selections

➡ **Different magnetic field structure**

- ➡ **ATLAS**: 2 T solenoid + Toroids
- ➡ **CMS**: strong 4 T solenoid

➡ **Different DAQ architecture**

- ➡ **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

Cannot do event Building at 100 kHz

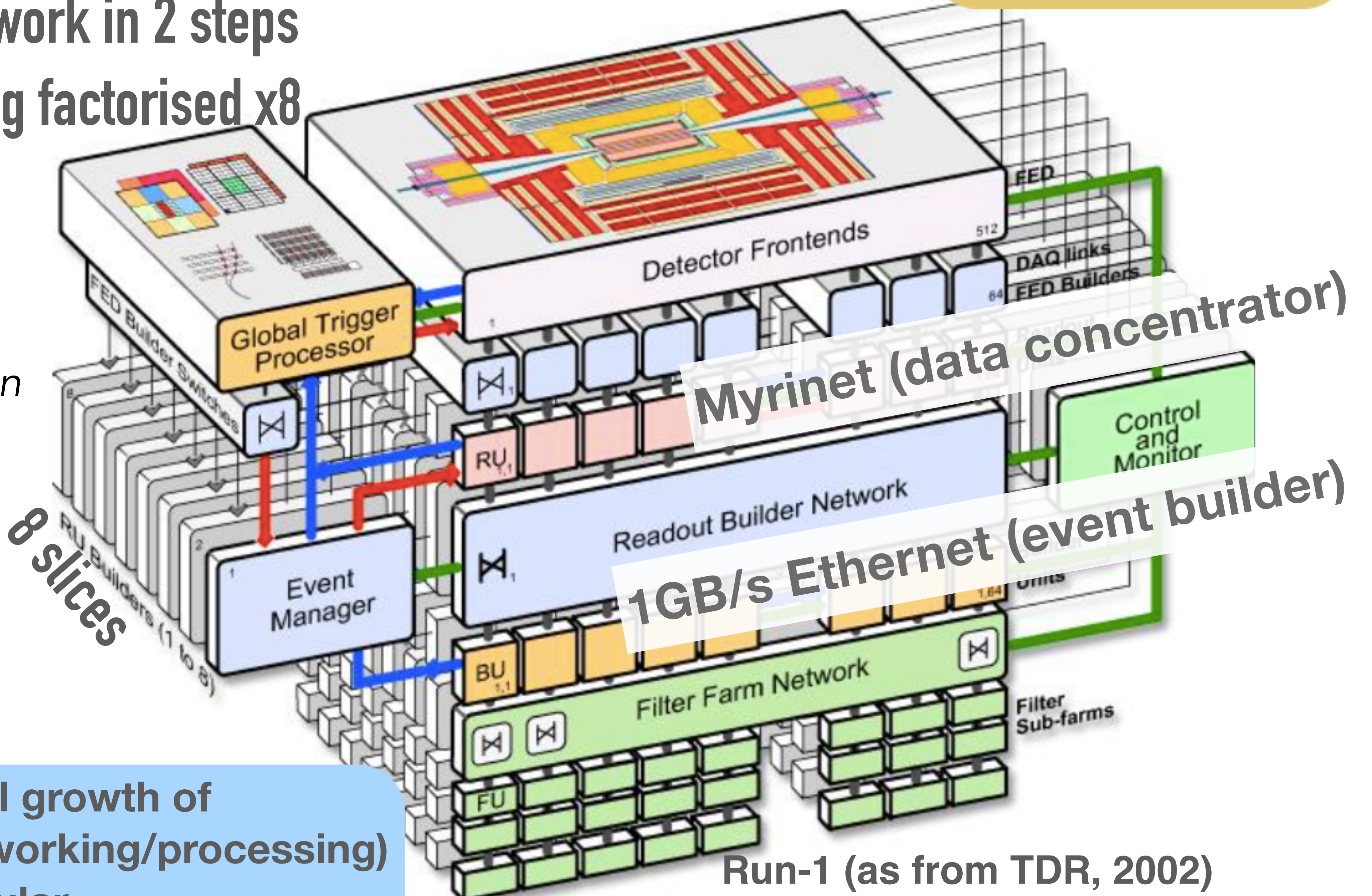
100 GB/s readout network in 2 steps

100 kHz Event Building factorised x8

2 EB networks in blu

Filter network in green

CMS DAQ-1



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

Run-1 (as from TDR, 2002)

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

NETWORK EVOLUTION

Run 1: 100 GB/s network

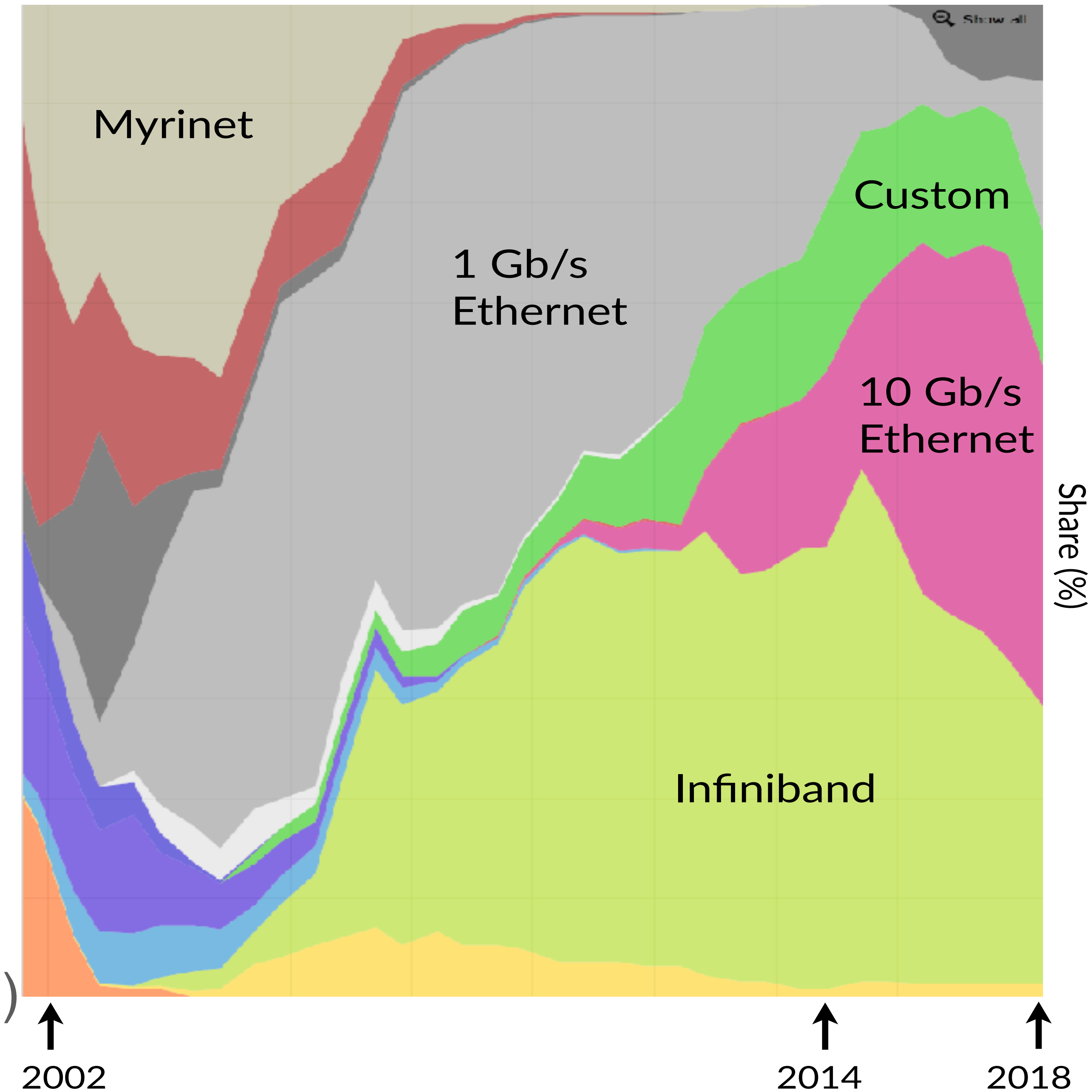
Myrinet widely used when DAQ-1 was designed

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

Run 2: 200 GB/s network

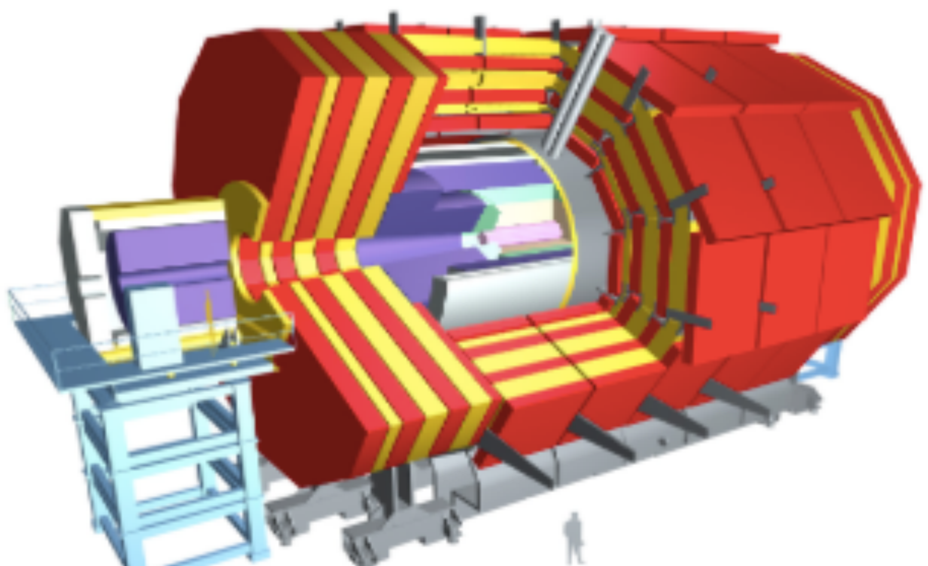
- ➡ Increased event size to 2MB
- ➡ Technology allows single EB network (56 Gbps FDR Infiniband)
- ➡ Myrinet —>10/40 Gbps Ethernet

Top500.org share by interconnect family

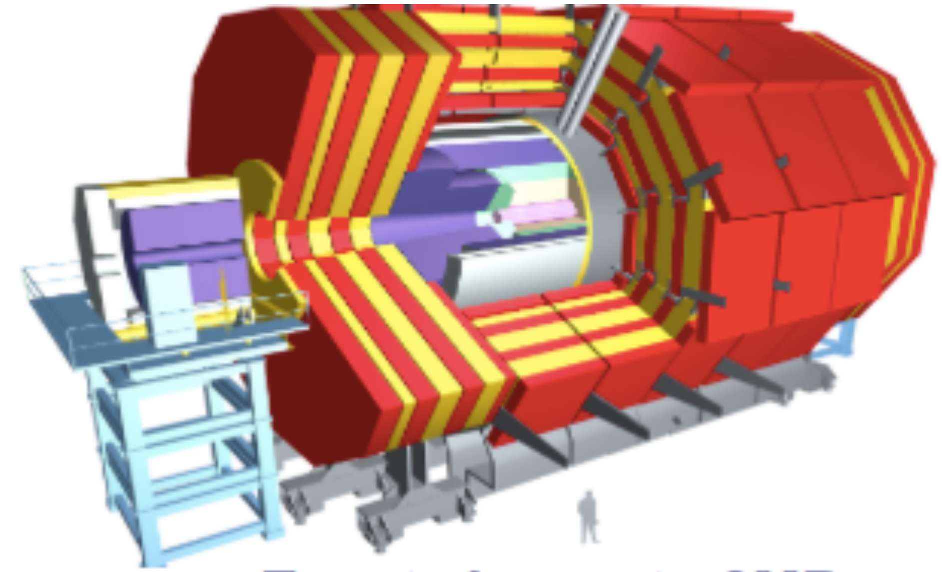


Choose best prize/bitps!

EVOLUTION FROM RUN-1 TO RUN-2

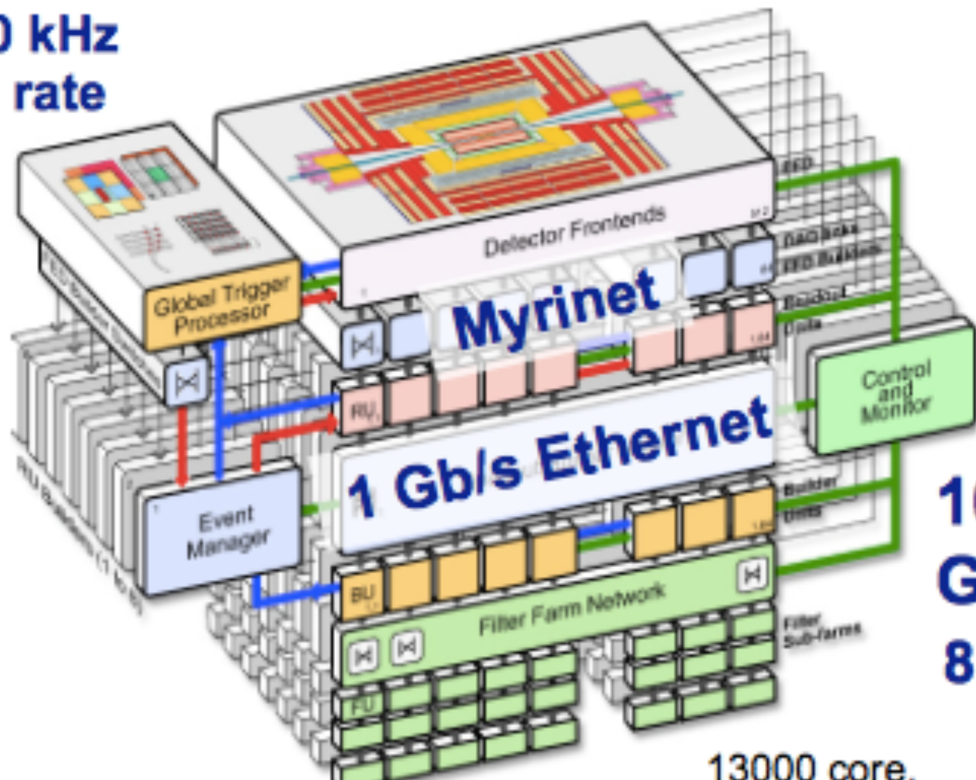


Event size up to 1MB



Event size up to 2MB

100 kHz
L1 rate



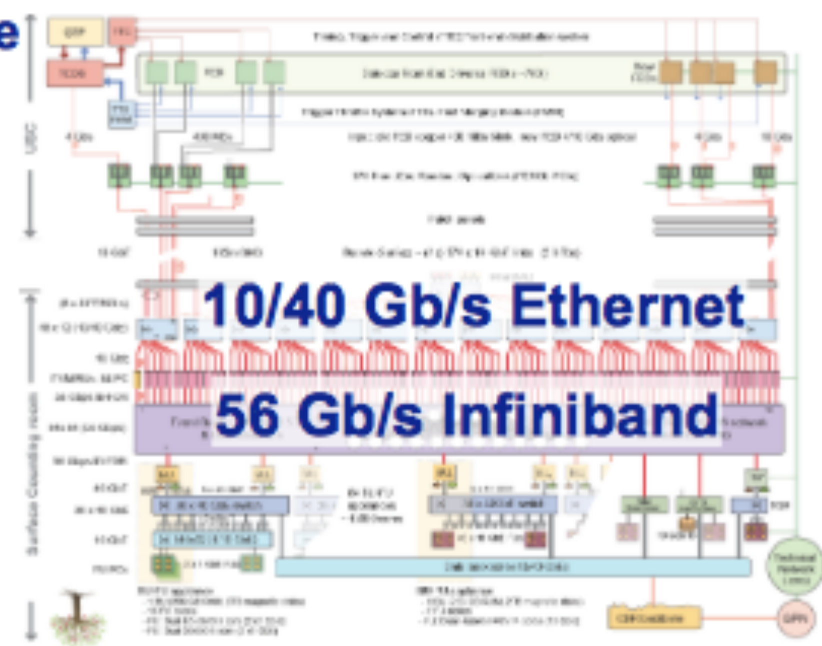
100
GB/s
8 slices

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

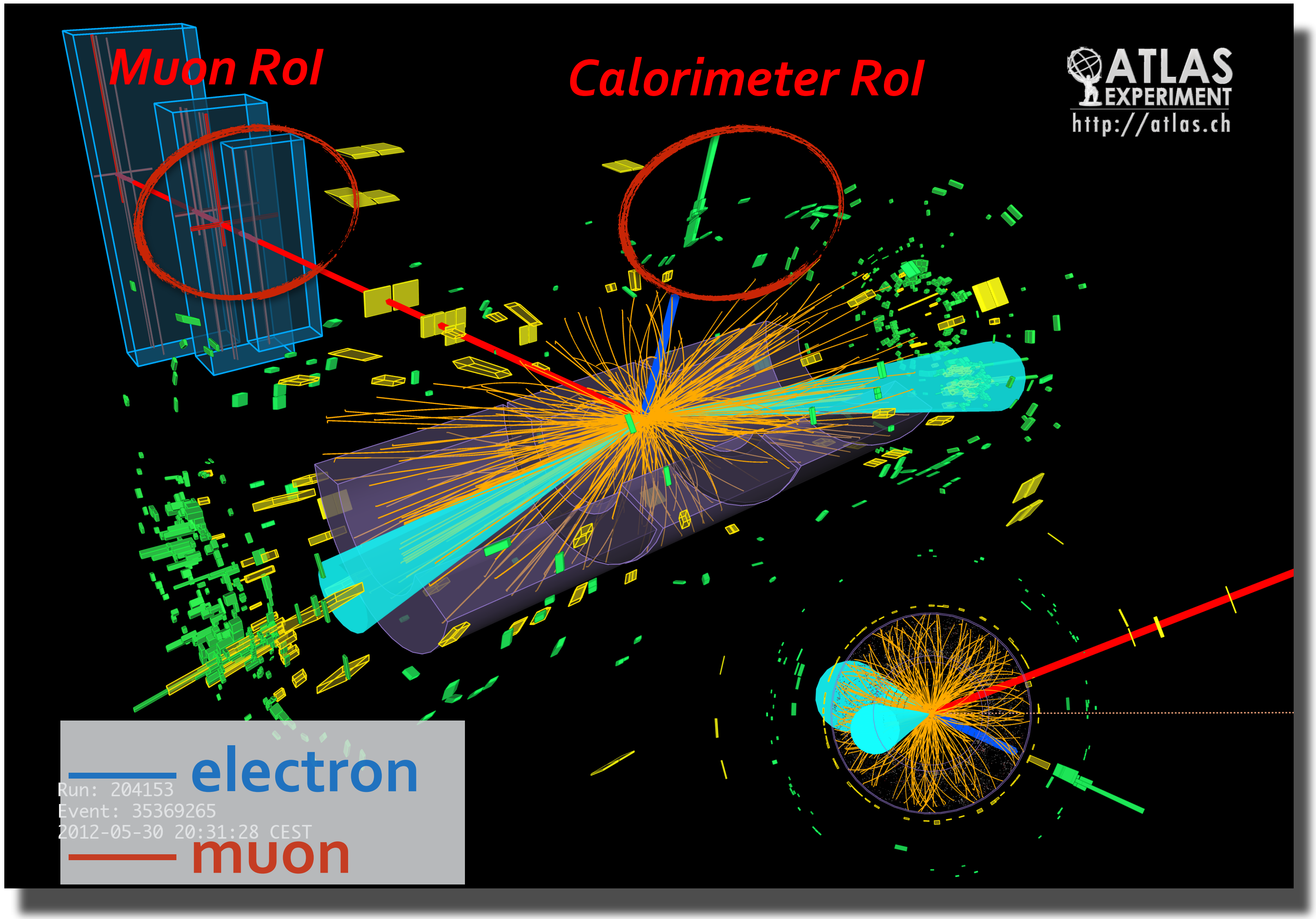
CMS DAQ 2

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)



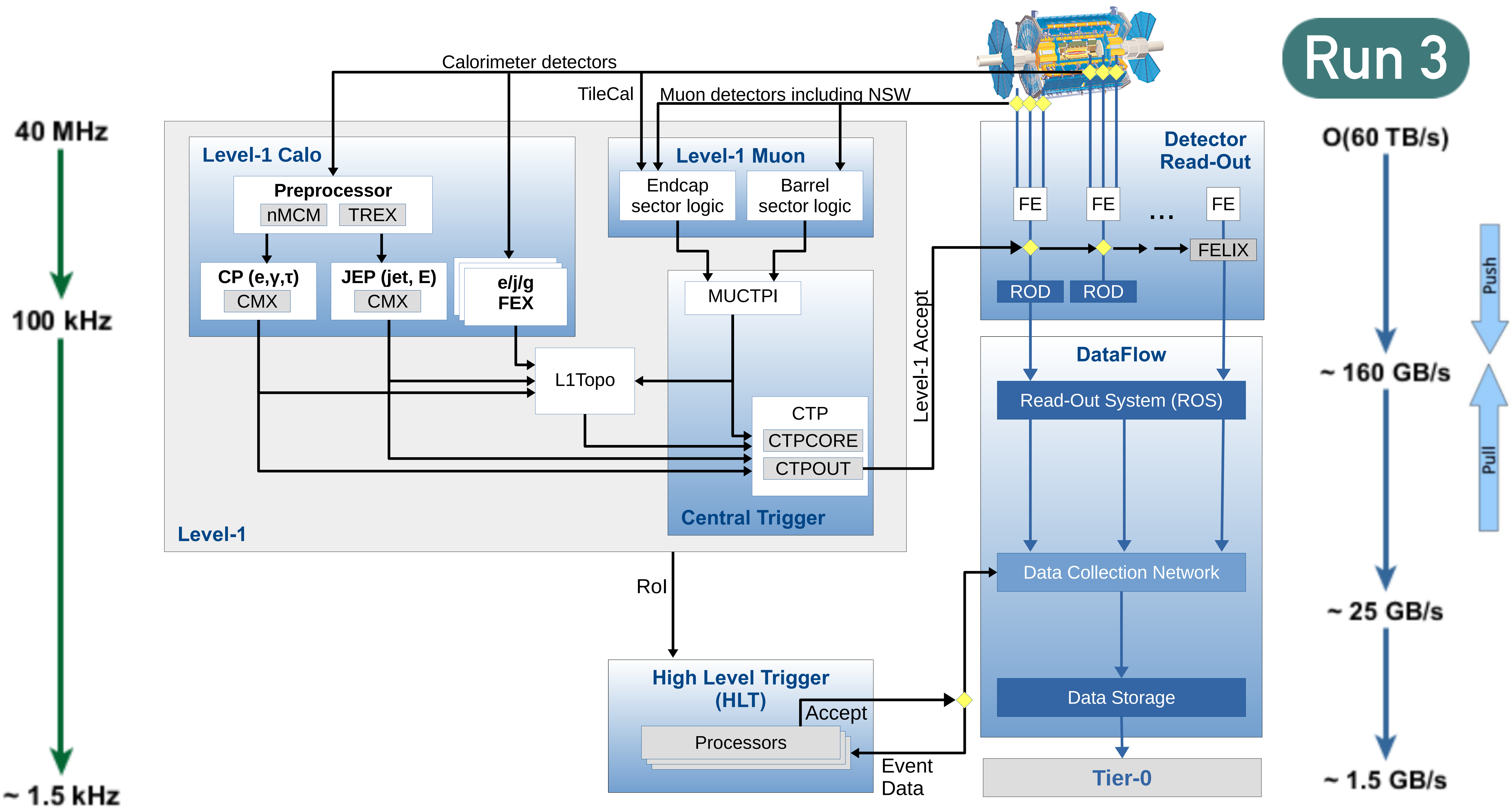
RoI=Region of Interest

- ➡ **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

ATLAS EXPERIMENT

Overall network bandwidth: ~10 GB/s (x10 reduced by regional readout)



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

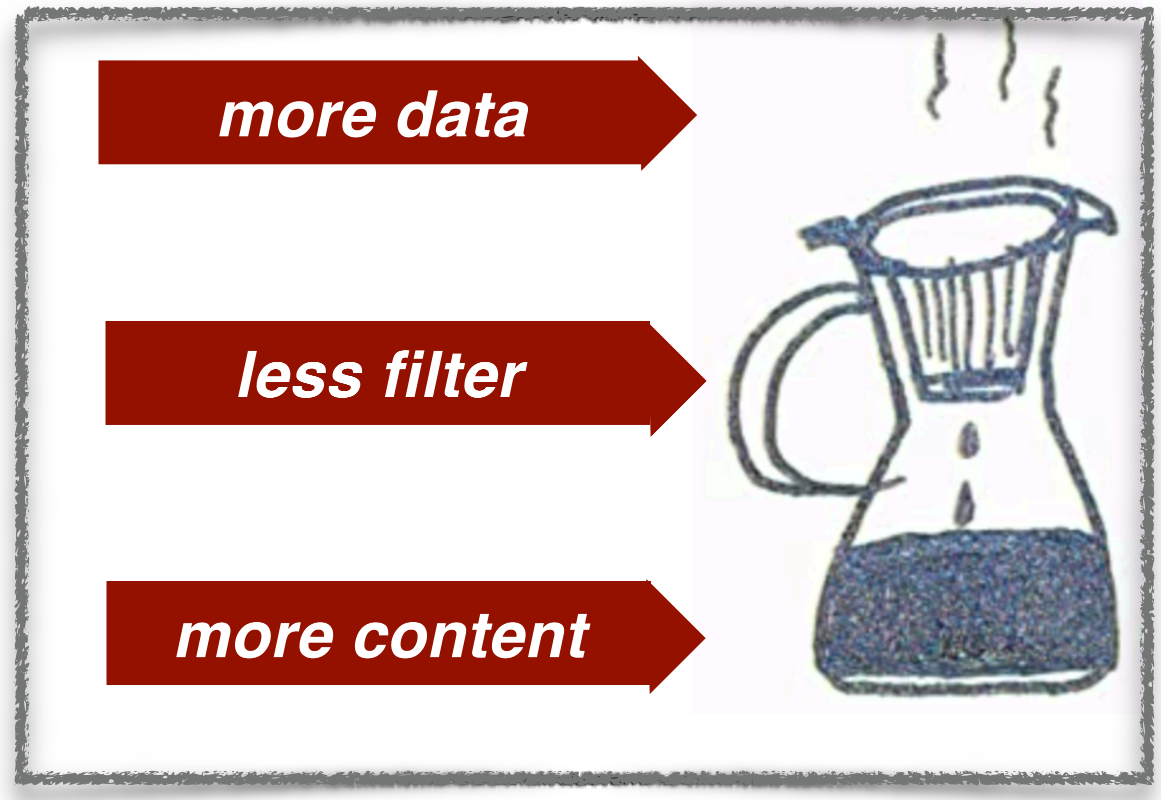
ADDITIONAL COMPLICATION AT HL-LHC

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

x10 higher Luminosity means…

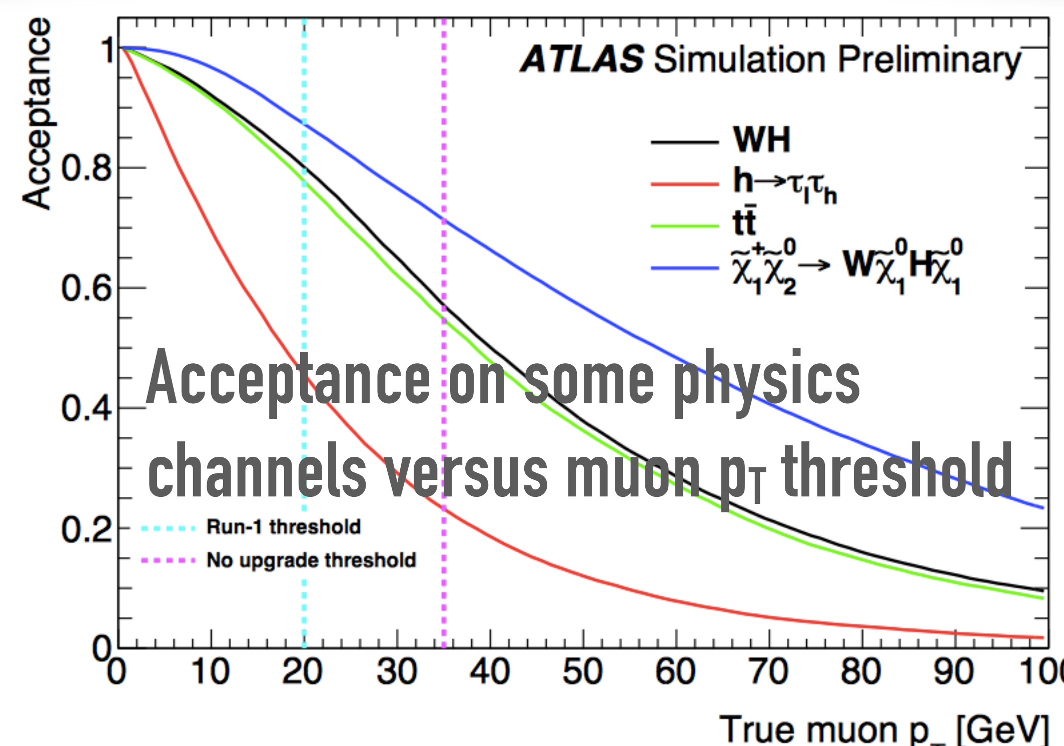
- ➡ **Higher pile-up**
  - ➡ Less rejection power (worse pattern recognition and resolution)
  - ➡ Larger Event size
- ➡ **Larger data rates (new readout/DAQ):**
  - ➡ Readout rate @L1: 0.1 ➠ 1 MHz
  - ➡ DAQ throughput: 1 ➠ 50 Tbps

ATLAS/CMS numbers



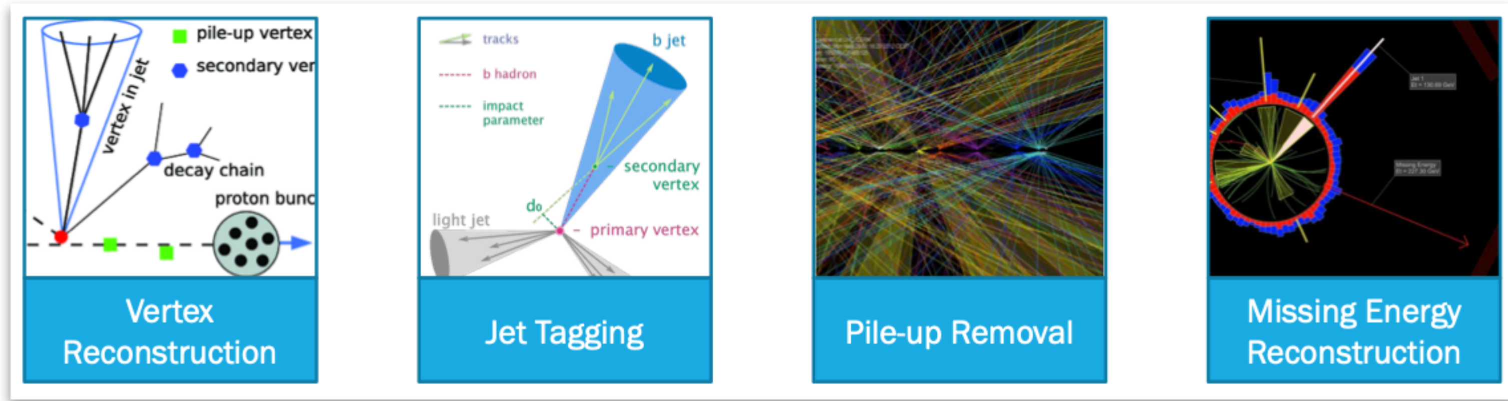
But cannot…

- ➡ **Apply too high thresholds**
  - ➡ Need to maintain physics acceptance
- ➡ **Scale dataflow with Luminosity**
  - ➡ **H/W**: more parallelism ➠ more links ➠ more material and cost
  - ➡ **S/W**: processing time not linear ~ L



TRACK–TRIGGER FOR RUN 4

➡ **Tracking systems provide incredibly high resolution, crucial for controlling rates**

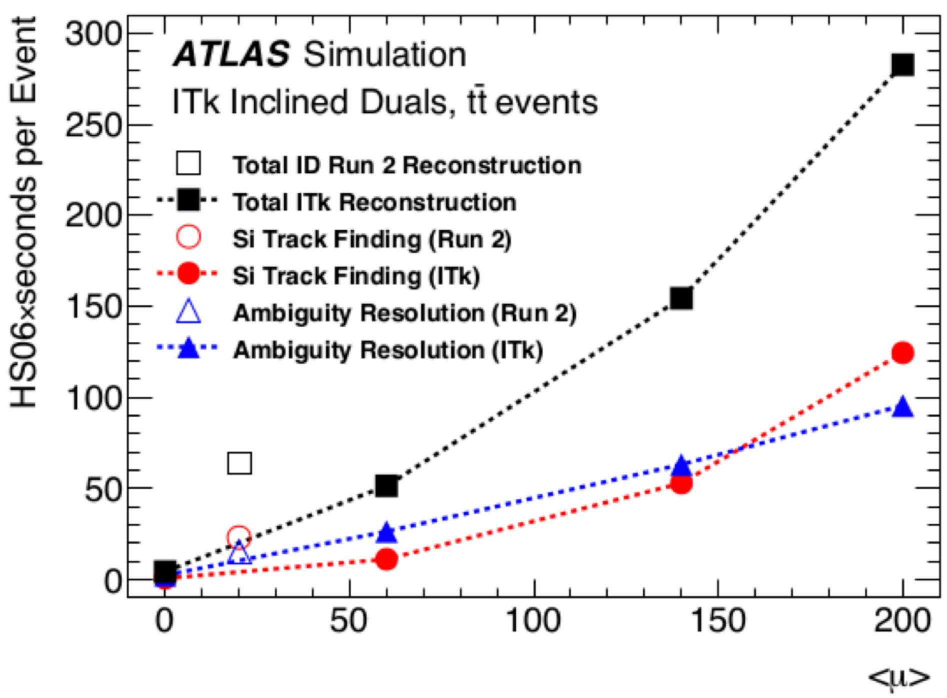


Tracking challenges

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

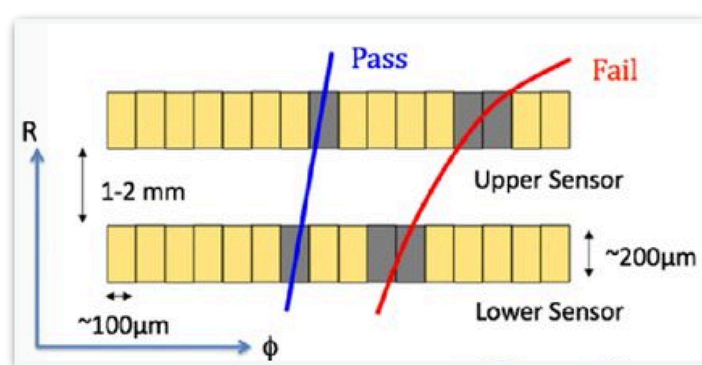
combinatorics scales like L^N

L=luminosity, N=number of layers

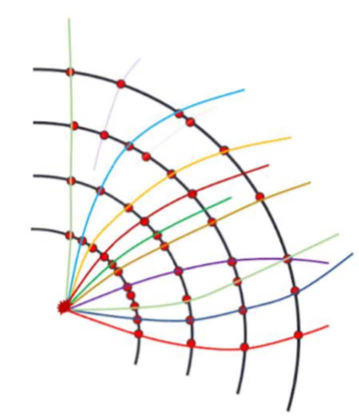


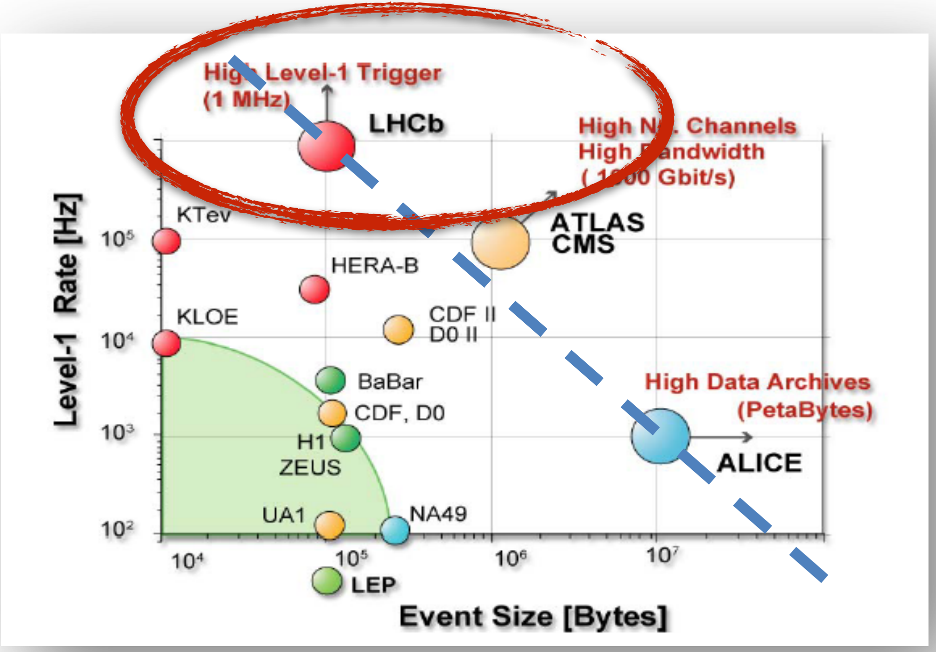
Not feasible @40MHz, nor in few microseconds

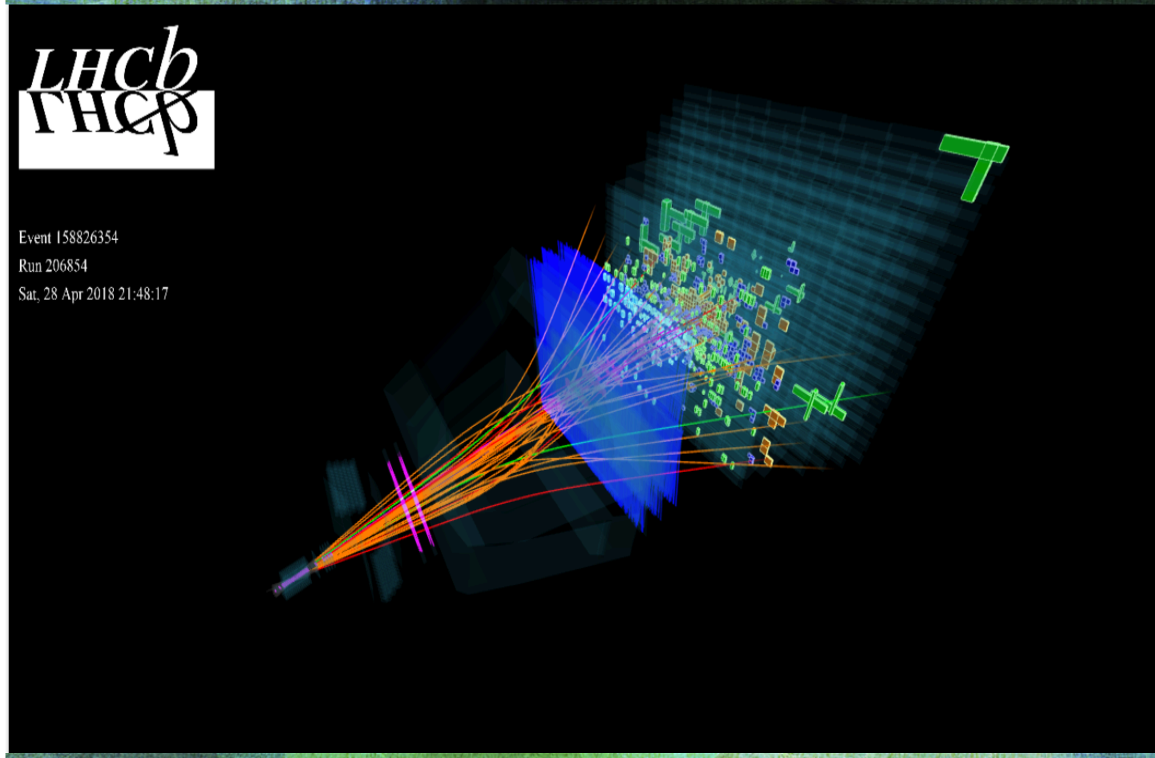
	ATLAS [1]	CMS [2]
data reduction @40MHz	regions from L1 (RoIs)	stubs from hw coincidences
track finding @1MHz	Studying best algorithms to run in FPGAs and/or in GPUs	
track fit @1MHz		
precision tracking @100kHz	optimized offline	optimized offline



stubs in CMS PT modules





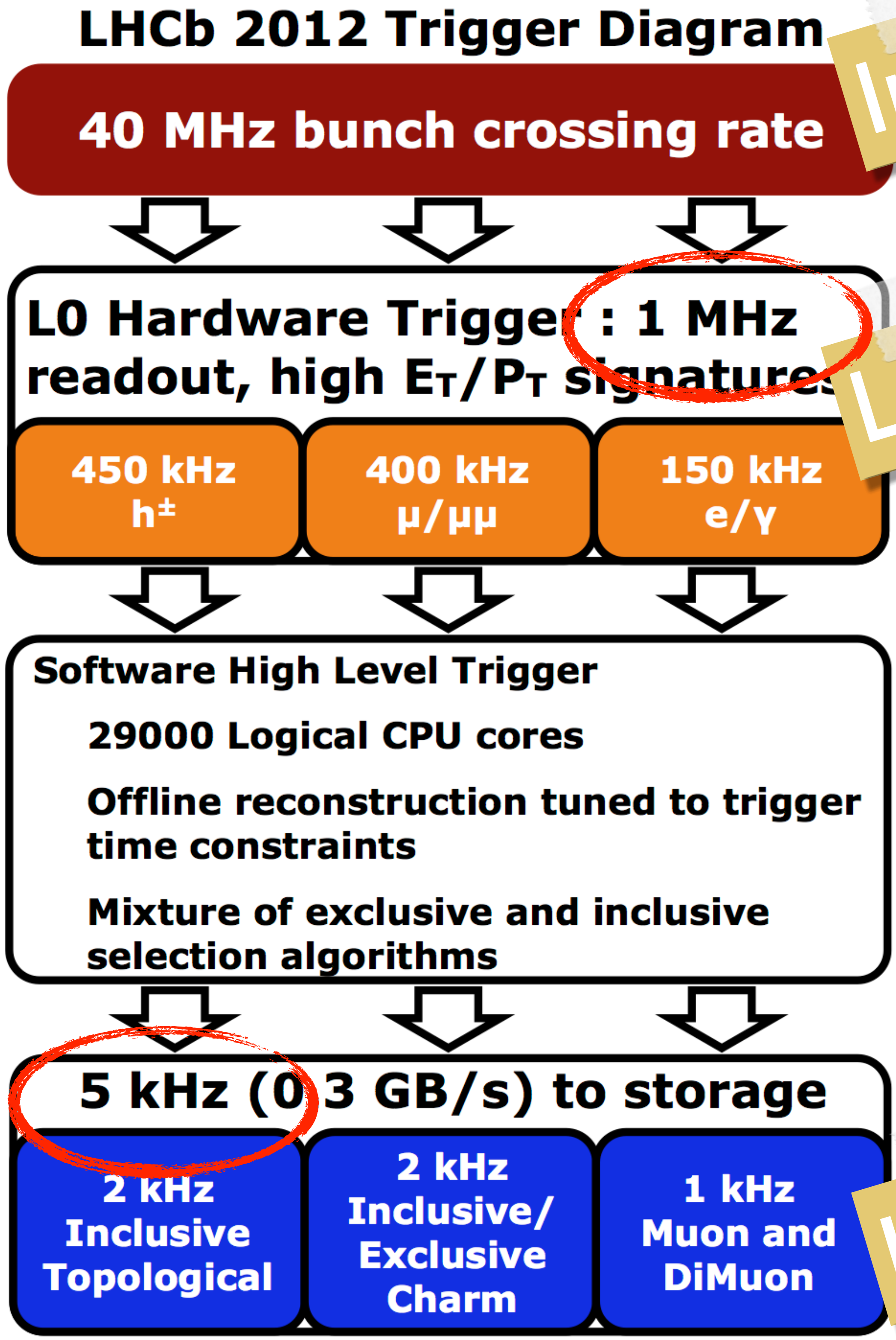


LHCb, THE B-MESON OBSERVATORY

The lightest experiment to study the heavy b-quark

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

LHCB TRIGGER STRATEGY



Input rate

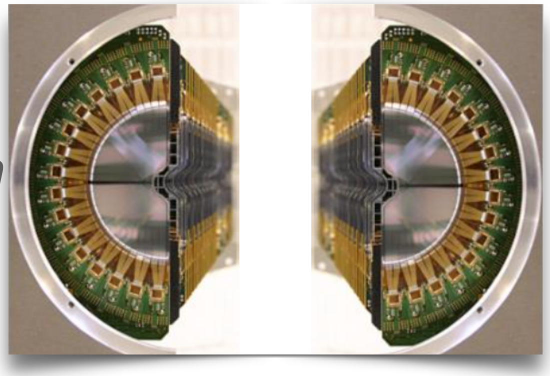
Low input rate and occupancy

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

L0 trigger

- Mainly hadronic triggers
- Reject complex/busy events

simplified vertex reconstruction with the siliconVELO detector

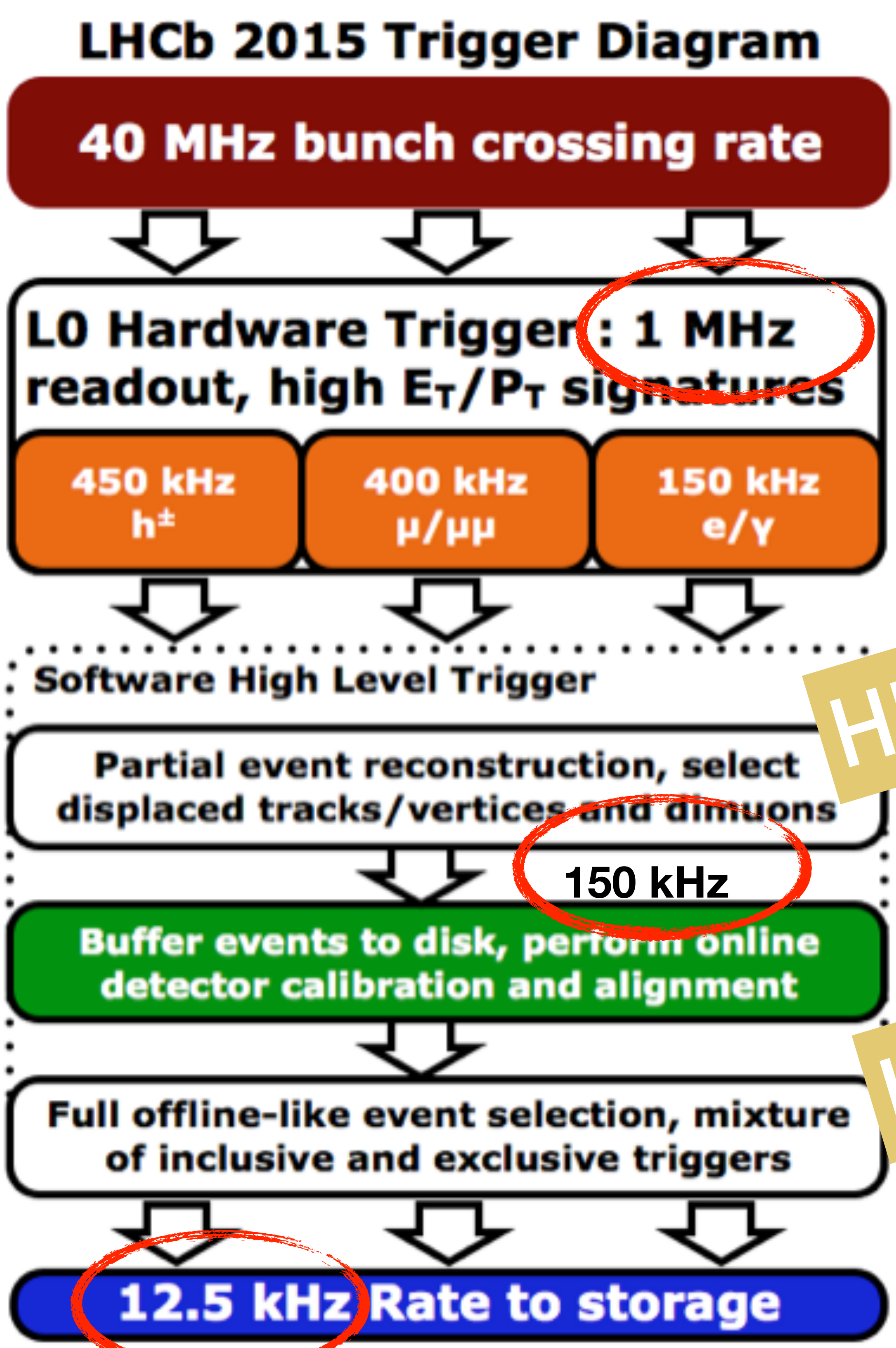


60kB * 1MHz= 60 GB/s readout network

HighLevel

- Multitude of exclusive selections

SCHEMA EVOLUTION



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

Real-time calibration and alignments

==> Split HLT processing into two stages, with a large buffer between (4PB)

Synchronous with DAQ

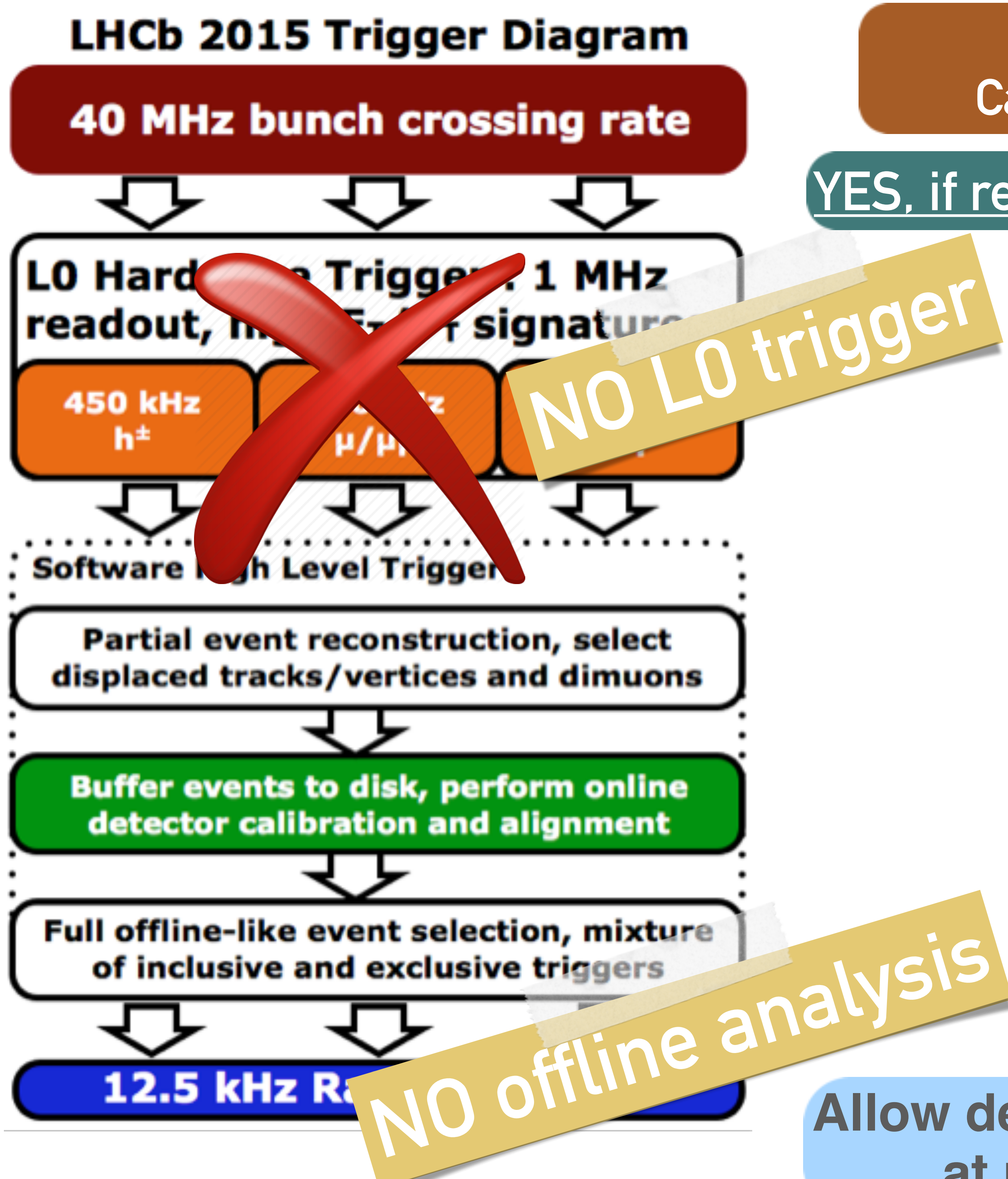
- Use tracks and vertices for impact parameter selection (in 35ms)

Deferred Processing

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

Large benefit from VELO alignments at each fill!

UPGRADES TOWARDS RUN 3

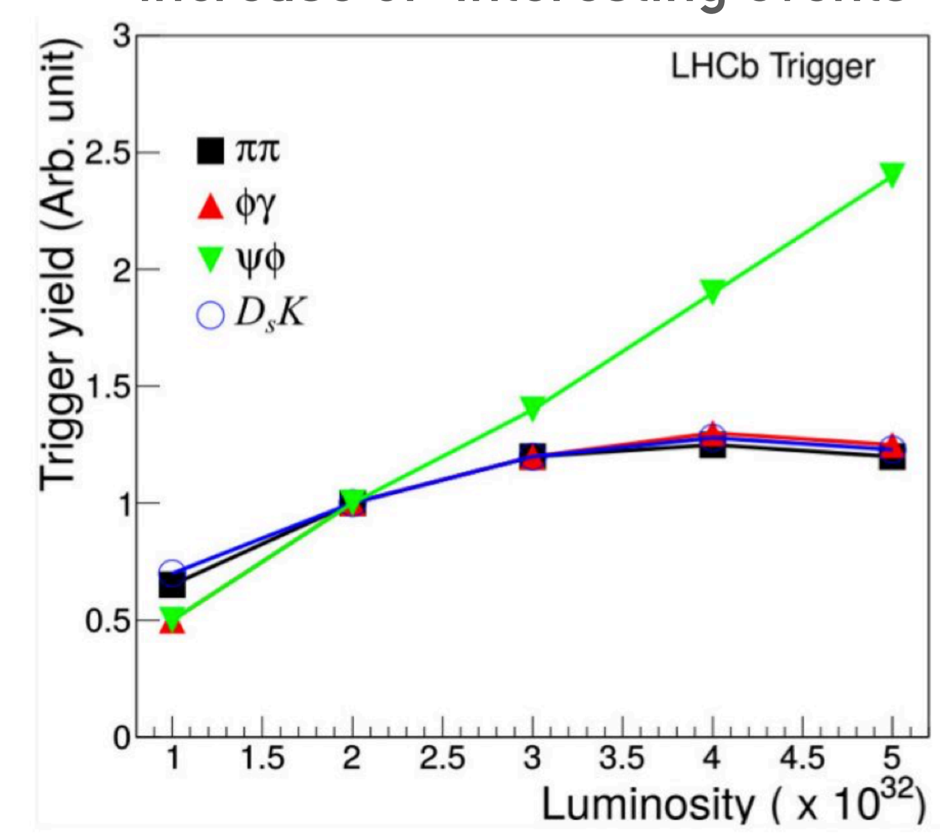


See Phase-I upgrade TDR

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

YES, if remove the limit from L0 -1MHz 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

80 Gbit/s

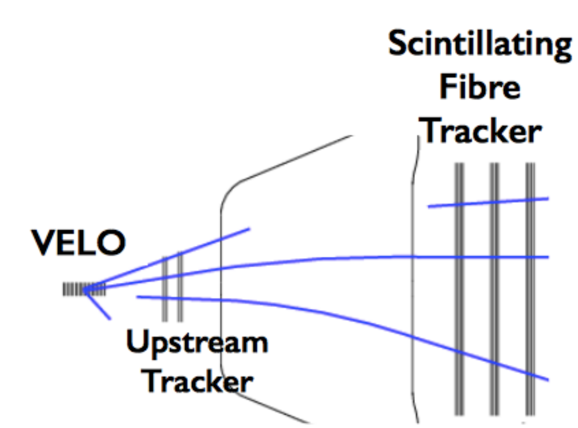
50 kHz

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 ?

- < 6 ms with Run2 HLT (12 cores + 12 hyper threads + 24 GB RAM) ==> ~ 100k cores!
- Modern CPU & co-processors (FPGA/GPU)

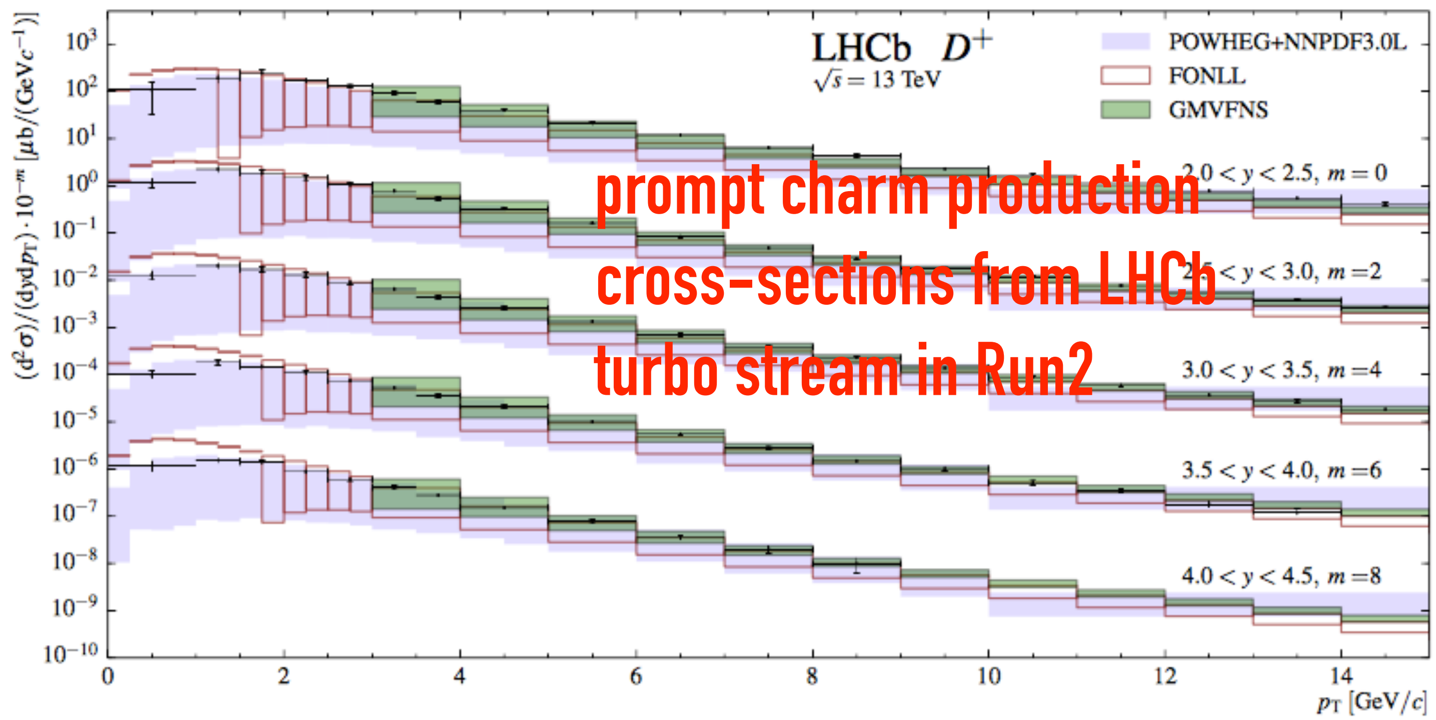


Online Tracking

- 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

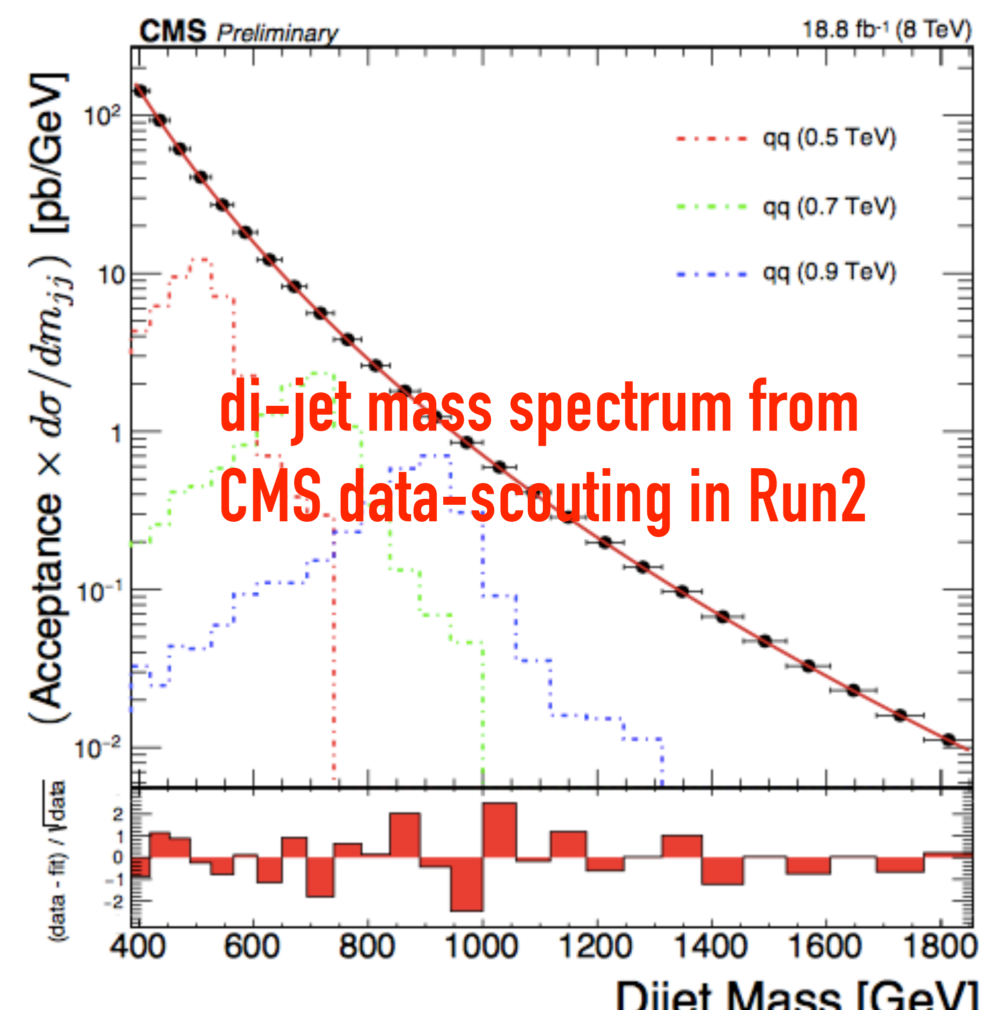
arXiv:2105.04031

A NEW TREND: REAL TIME ANALYSIS



Can we get rid of FrontEnd raw data?

- ➡ **Event size/10 -> x10 rate, for free**
- ➡ **Tested on dedicated data streams in many experiments:**
  - ➡ Full online reconstruction (**LHCb**)
  - ➡ Data scouting (**ATLAS/CMS**)
    - ➡ for some high rate signatures, save only reduced information



➡ Main data stream for LHCb & ALICE upgrade

➡ and be a guidance for all other experiments

HOW TO LIVE WELL WITHOUT A L1 TRIGGER

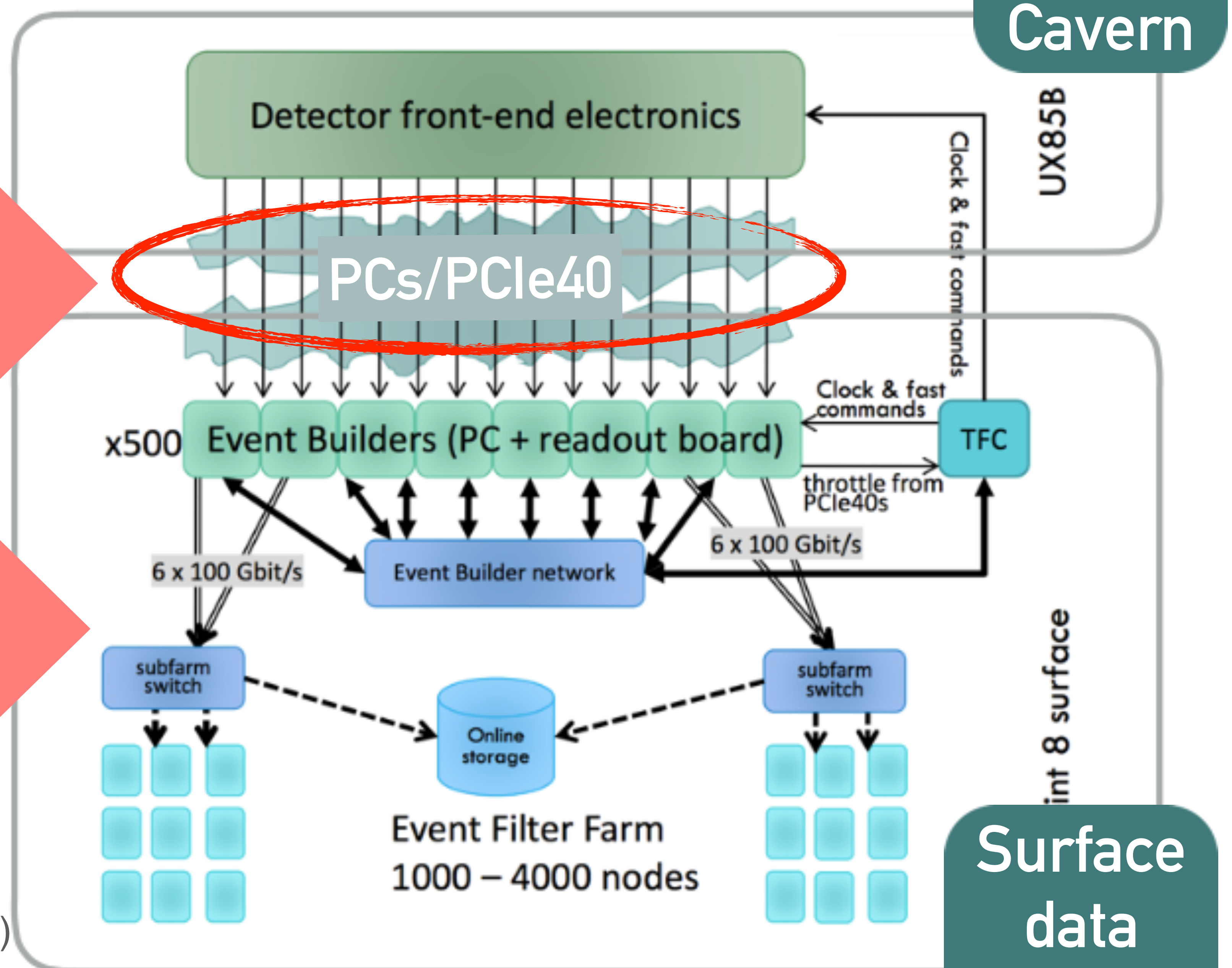
Readout @ 30 MHz
Event size ~ 150kB

PCIe-3: simple protocol, large bandwidth.
PCIe: maximum flexibility in later networking choice.

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

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

- **Decouple Event Building and HLT, in 2 networks**
  - HLT-1 reducing 40Tbit/s => 1 Tbit/s (GPUs)
  - HLT-2 reducing to 80 Gbitps

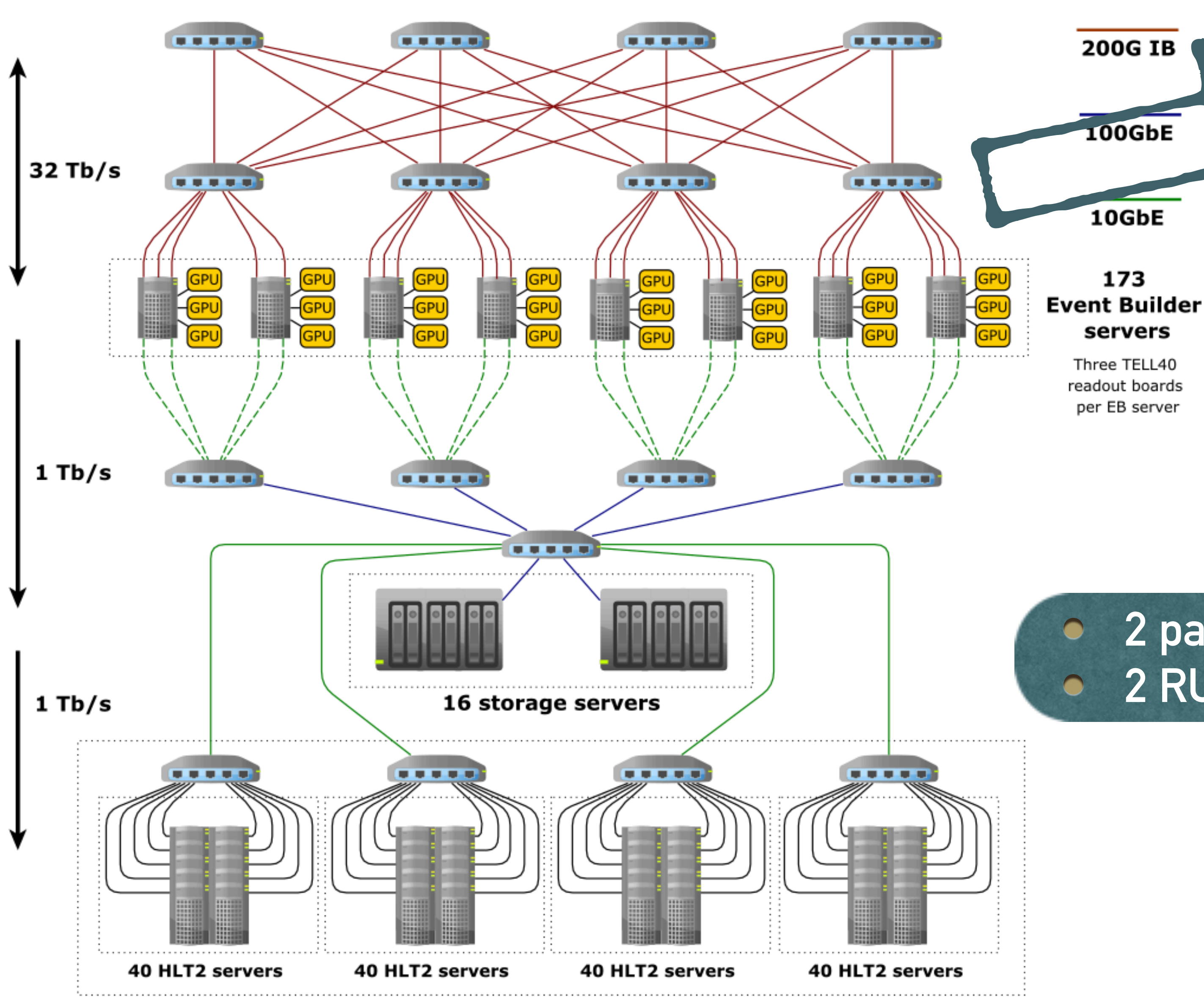


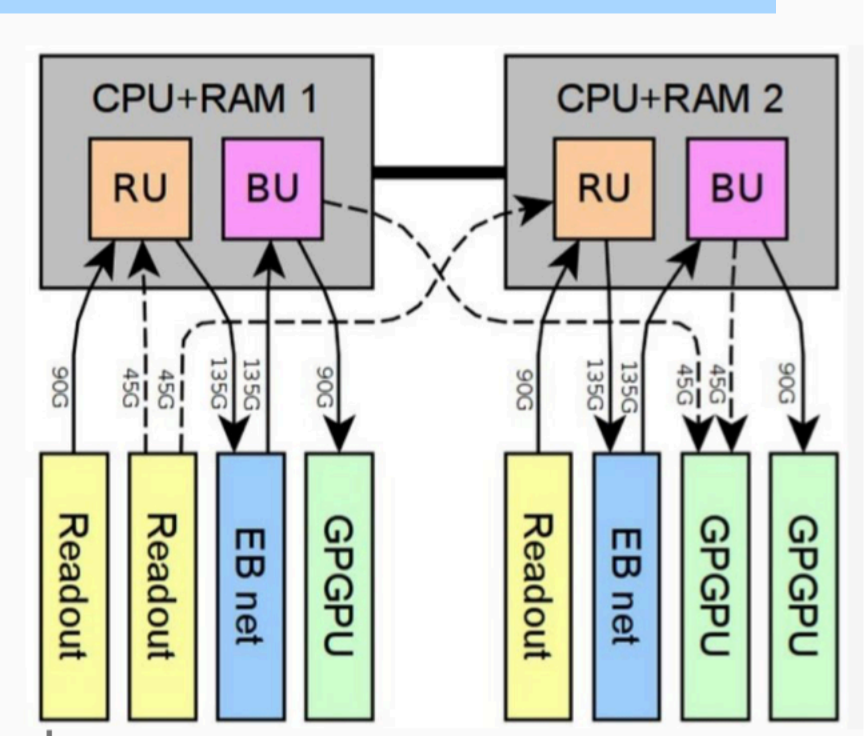
150kB x 30MHz = 40Tbs

Ref for PCIe40

A 2-DIM FOLDED EVENT BUILDING

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





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

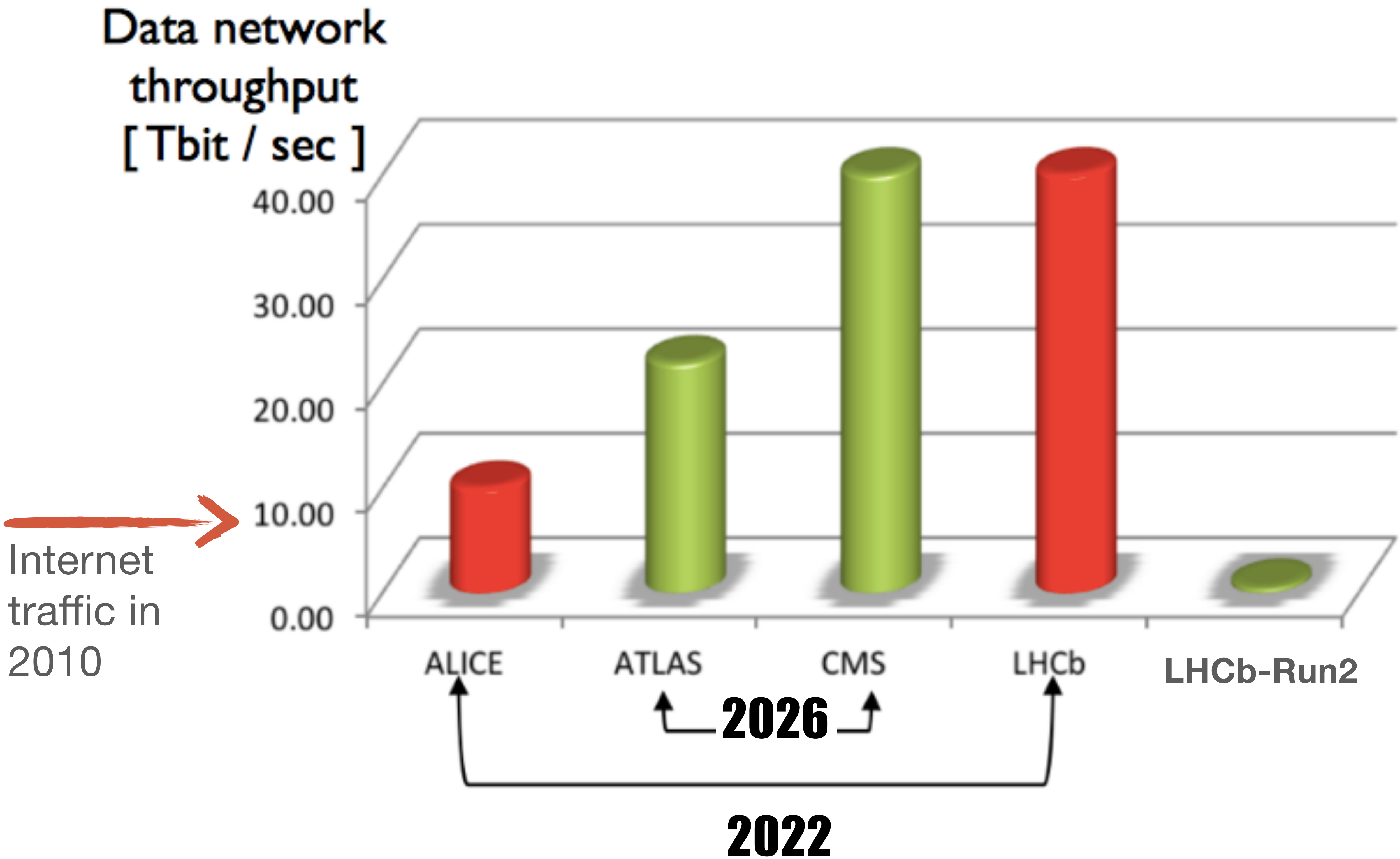
➡ **EB network is oversized: able to manage 64Tb/s (320 network cards x 200Gb/s)**

➡ **Large rejection at HLT1: use O(200) GPU! throughput at ~100kHz**

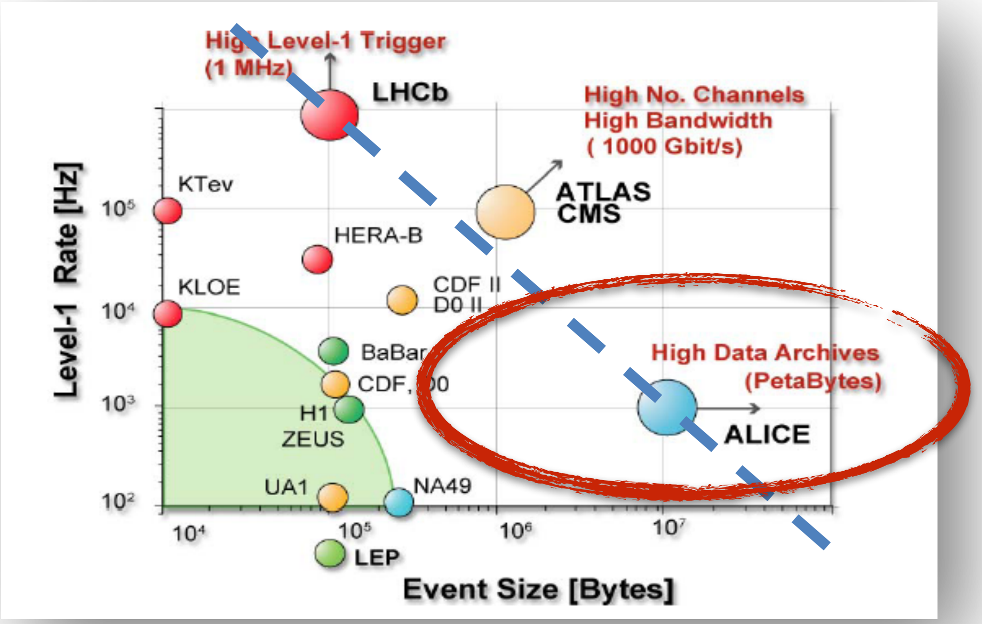
➡ **Storage Buffer HLT1-HLT2 = 40 PB (3000 hard-disks) enough for days**

- ➡ SSD tare faster but have short lifetime wrt high read-write rate, so prefer hard-disks

NETWORK TRAFFIC COMPARISON



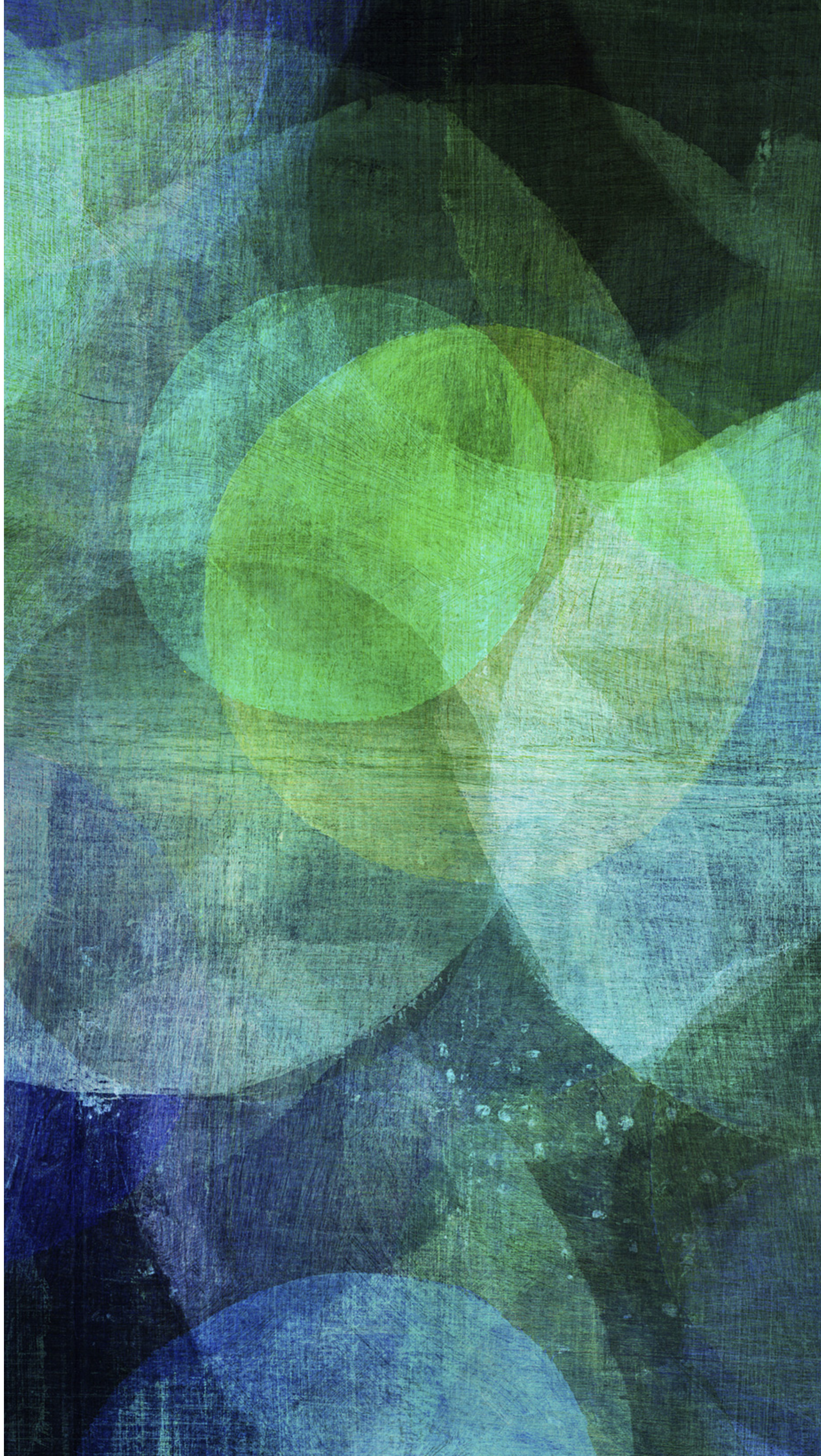
Same data volume as ATLAS/CMS HL-LHC upgrades! But earlier and for less money



ALICE: THE SMALL BIG-BANG

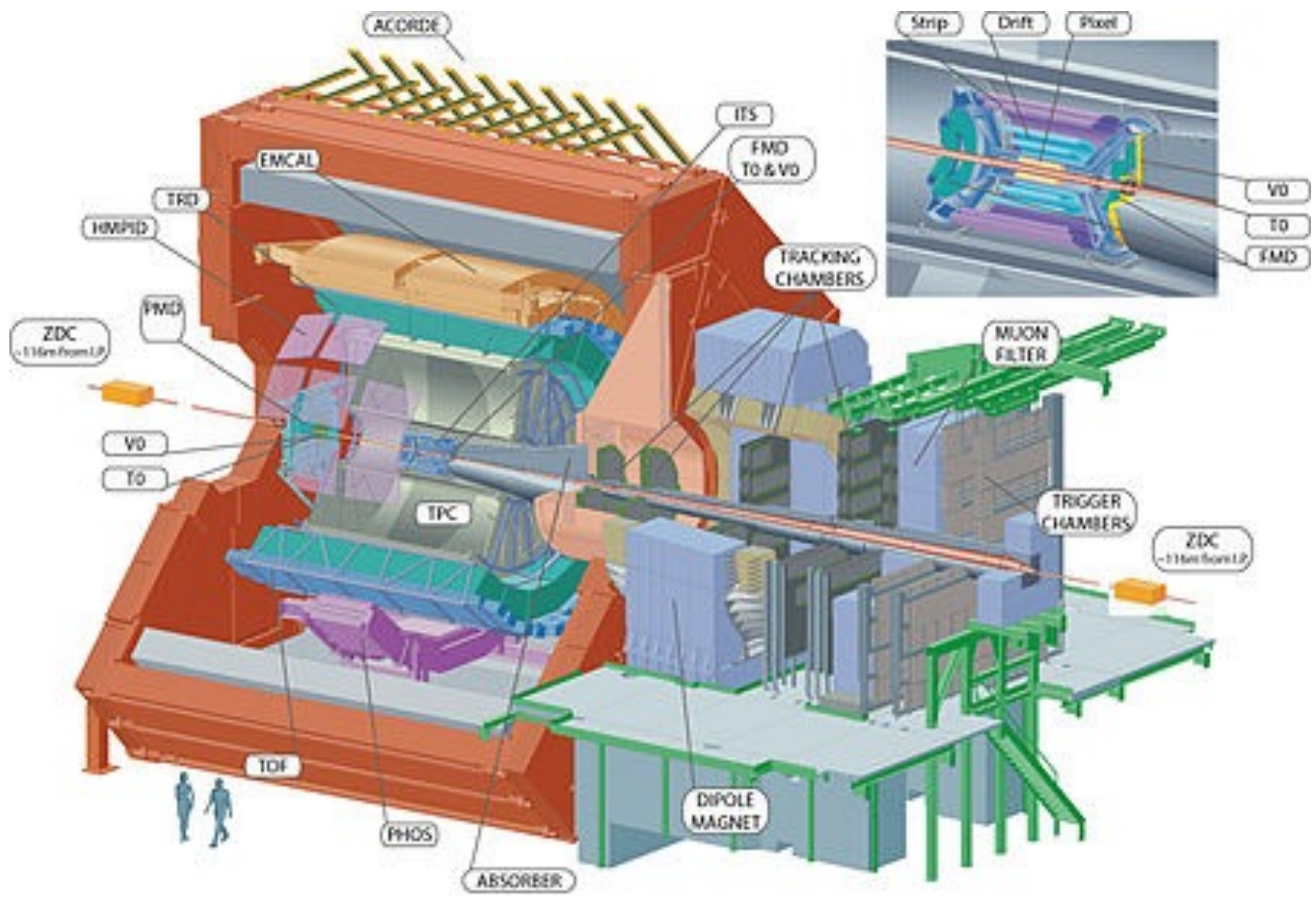
Recording heavy ion collisions

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



DESIGNED FOR HEAVY ION COLLISIONS

- ➡ **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 (> 40MB)**
  - ➡ TPC producing 90% of data
- ➡ **Complex event topology**
  - ➡ low trigger rate: max 3.5 kHz

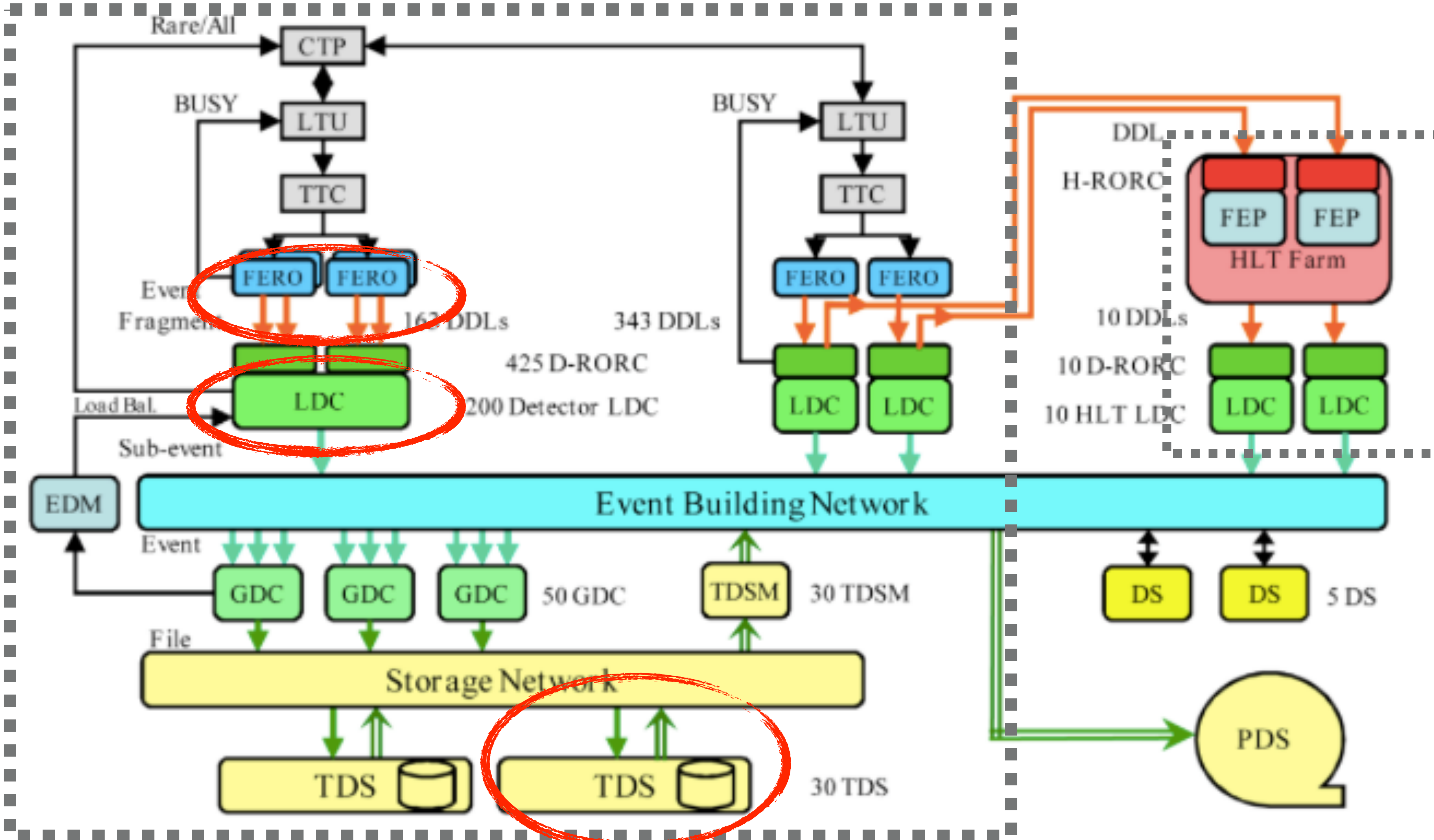


cms = 5.5 TeV per nucleon pair
Pb–Pb collisions at L $=10^{27}$ $cm^{-2}s^{-1}$

➡ Challenges for TDAQ design:

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

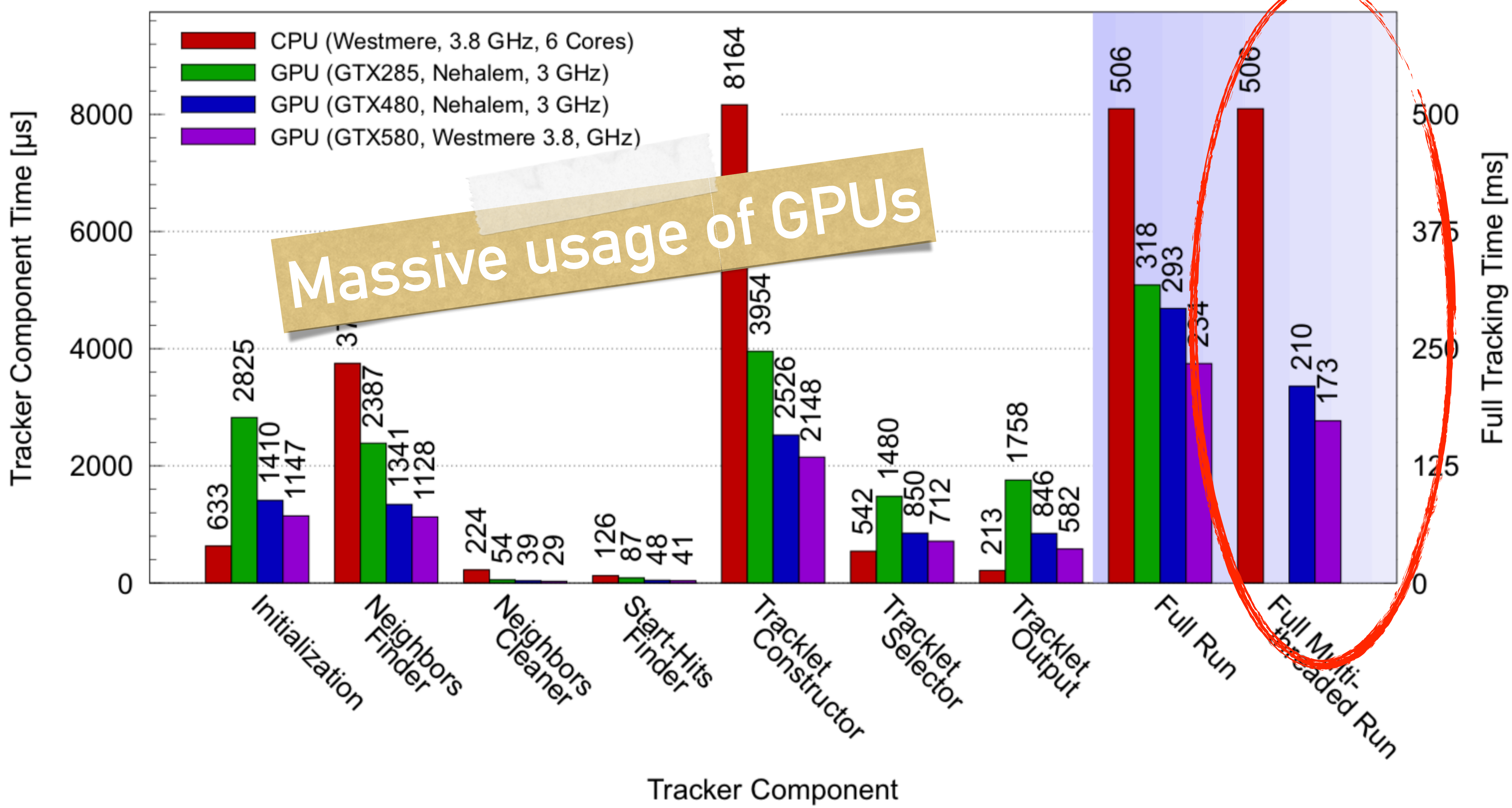
DAQ/HLT ARCHITECTURE IN RUN 1 AND RUN 2



- ➡ **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

RUN 2 EVOLUTION: IN H/W AND S/W

- ➡ x2 readout rate, thanks to data compression in GPUs and FPGAs
- ➡ x2 DAQ throughput, thanks to COTS evolution: 2.5GB/s (2010) ⇒ 6GB/s (2015)



Prove of concepts for next Run!

TOWARDS RUN 3

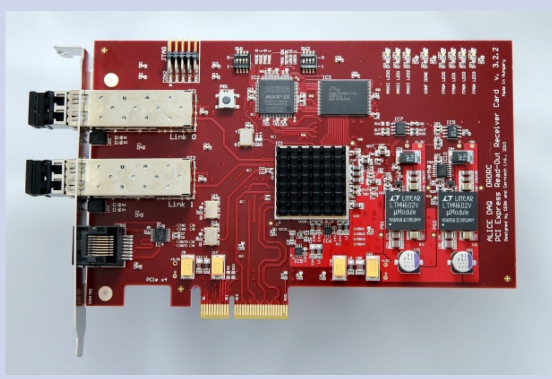
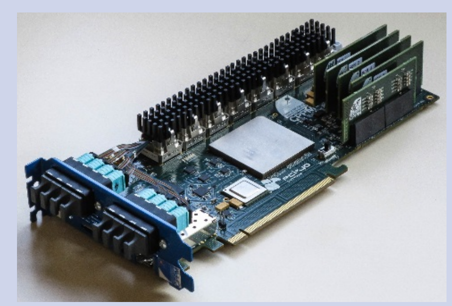
ALICE

➡ **LHC heavy ion programme: <u>extend statistics by x100!</u>**

- ➡ Detector granularity (===> increase event size!)
- ➡ Readout rates (===> faster electronics): ~kHz → 50 kHz
  - ➡ rate very close to TPC readout !!

New TDAQ challenge!

~3TB/s detector readout

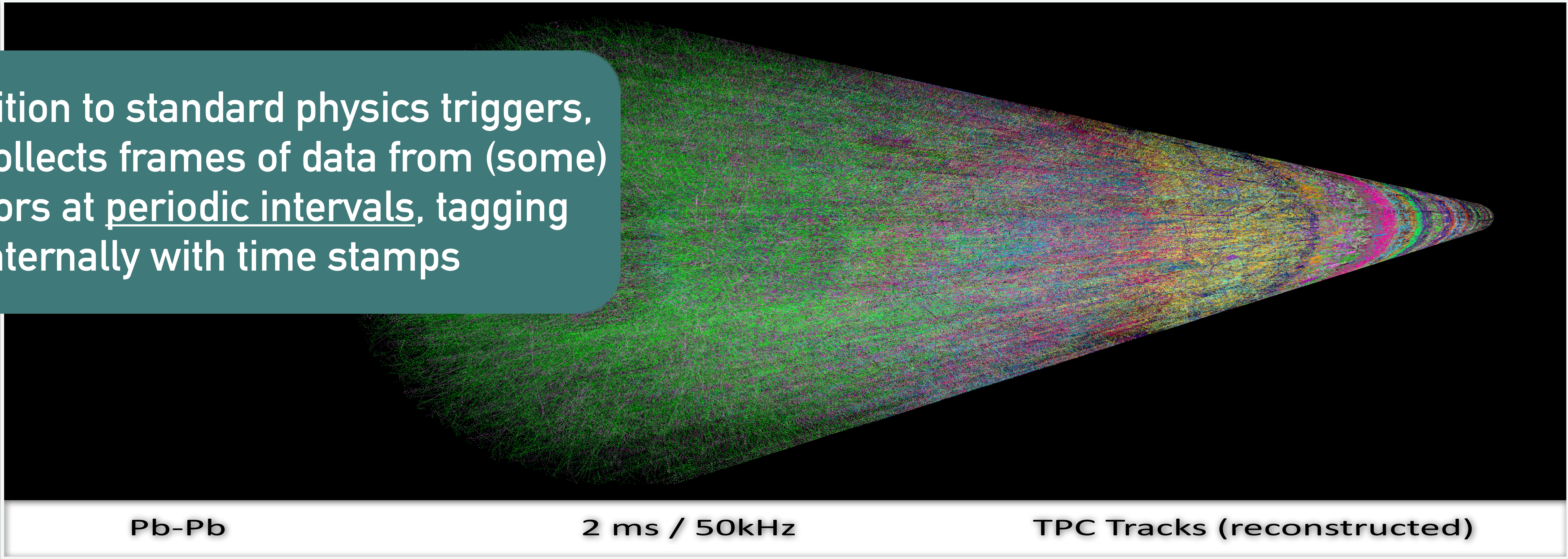
RORC 1	C-RORC	CRU
		
2 ch @ 2 Gb/s PCIe gen.1 x4 (1 GB/s)	12 ch @ up to 6 Gb/s PCIe gen.2 x 8 (4 GB/s)	24 ch @ 5 Gb/s PCIe gen.3 X 16 (16 GB/s)
Custom DDL protocol	Custom DDL protocol (same protocol but faster)	GBT
Protocol handling TPC Cluster Finder	Protocol handling TPC Cluster Finder	Protocol handling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

Run 1 → LS1 → Run 2 → LS 2 → Run 3

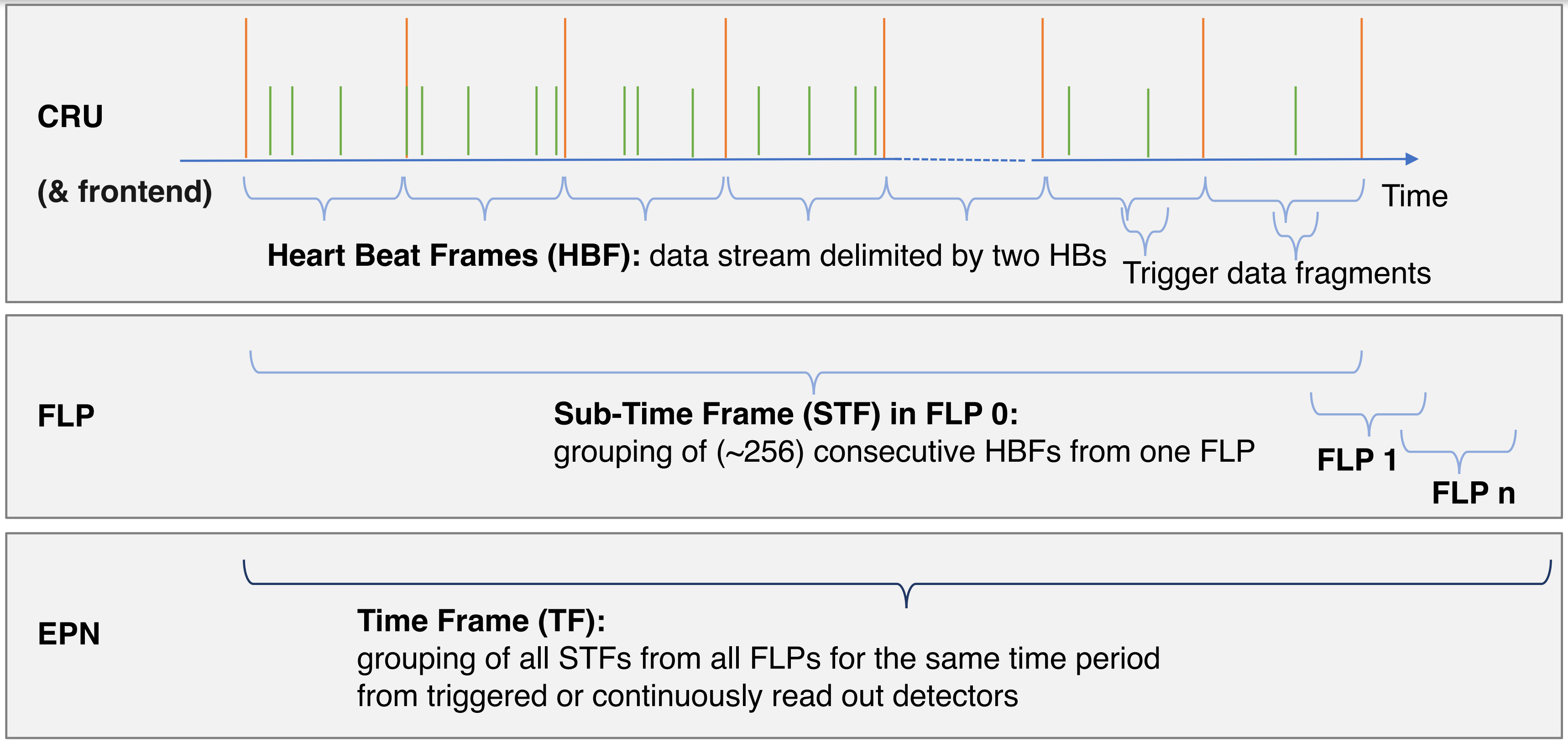
- ➡ **Common Readout Unit**
  - ➡ based on PCIe40 card
- ➡ **Storage bandwidth x O(100)**
  - ➡ Offline reconstruction also challenging due to combinatorics

CONTINUOUS READOUT FOR RUN 3

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



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



RUN 3 DAQ: ONLINE RECONSTRUCTION

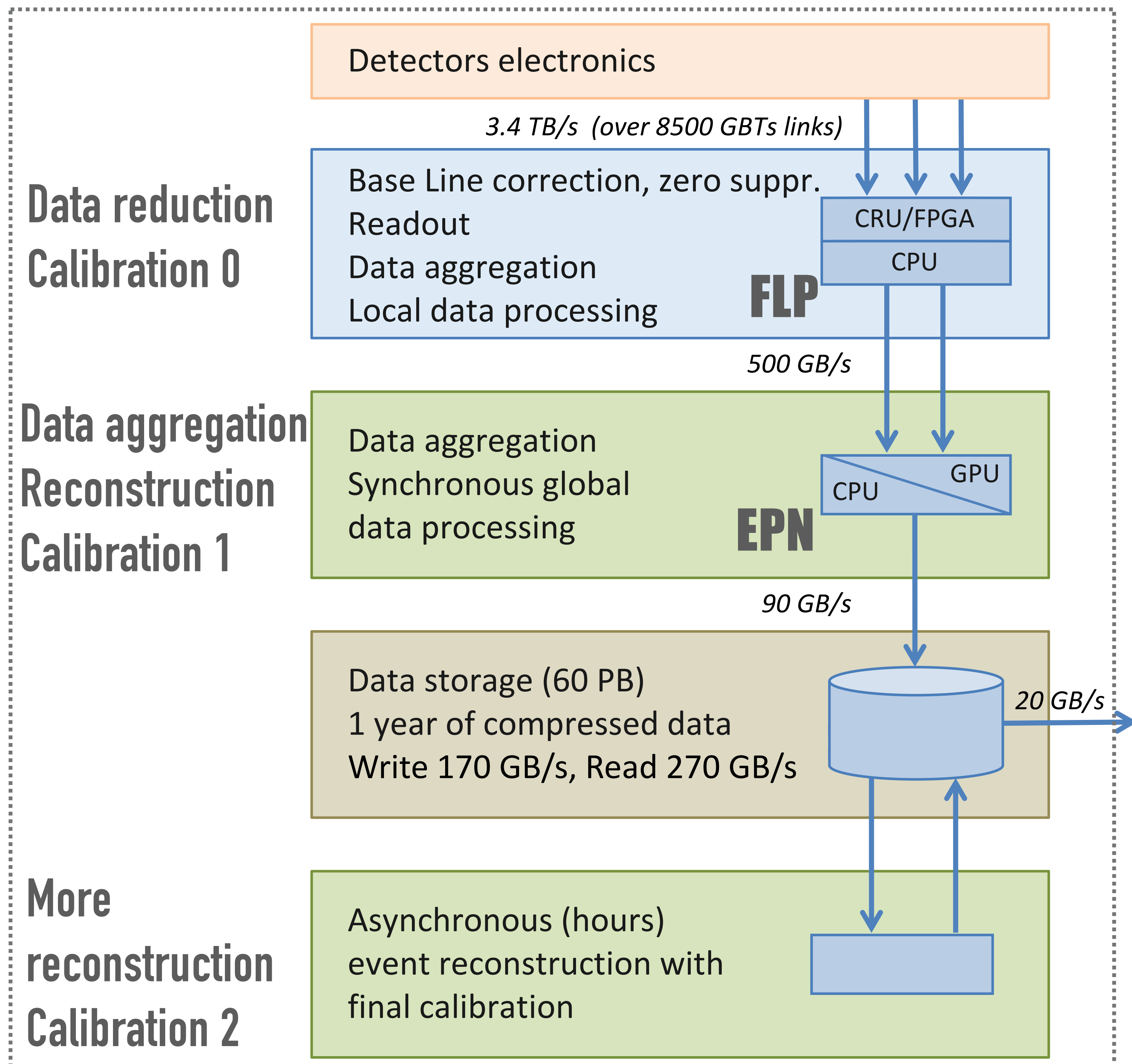
ALICE

Higher rates with smaller data

Reconstruction!

- **Very heterogeneous system**
  - Synchronous: up to EPN (continuous data)
  - Asynchronous with MT and MP
  - 30 s to analyse 20ms-time frame
- **FLP: Data compression in FPGA/CPU**
- **EPN: Track reconstruction in GPUs**
  - 250 EPN servers with 8 GPU-cards
  - Require large-memory GPUs!
  - Moving all reconstruction to GPUs (calorimeter)
- **Common Online and offline software**
  - Same calibrations and resources
- **Store reconstruction, discard raw data**
  - do you 100% trust software?
  - Robust monitoring!

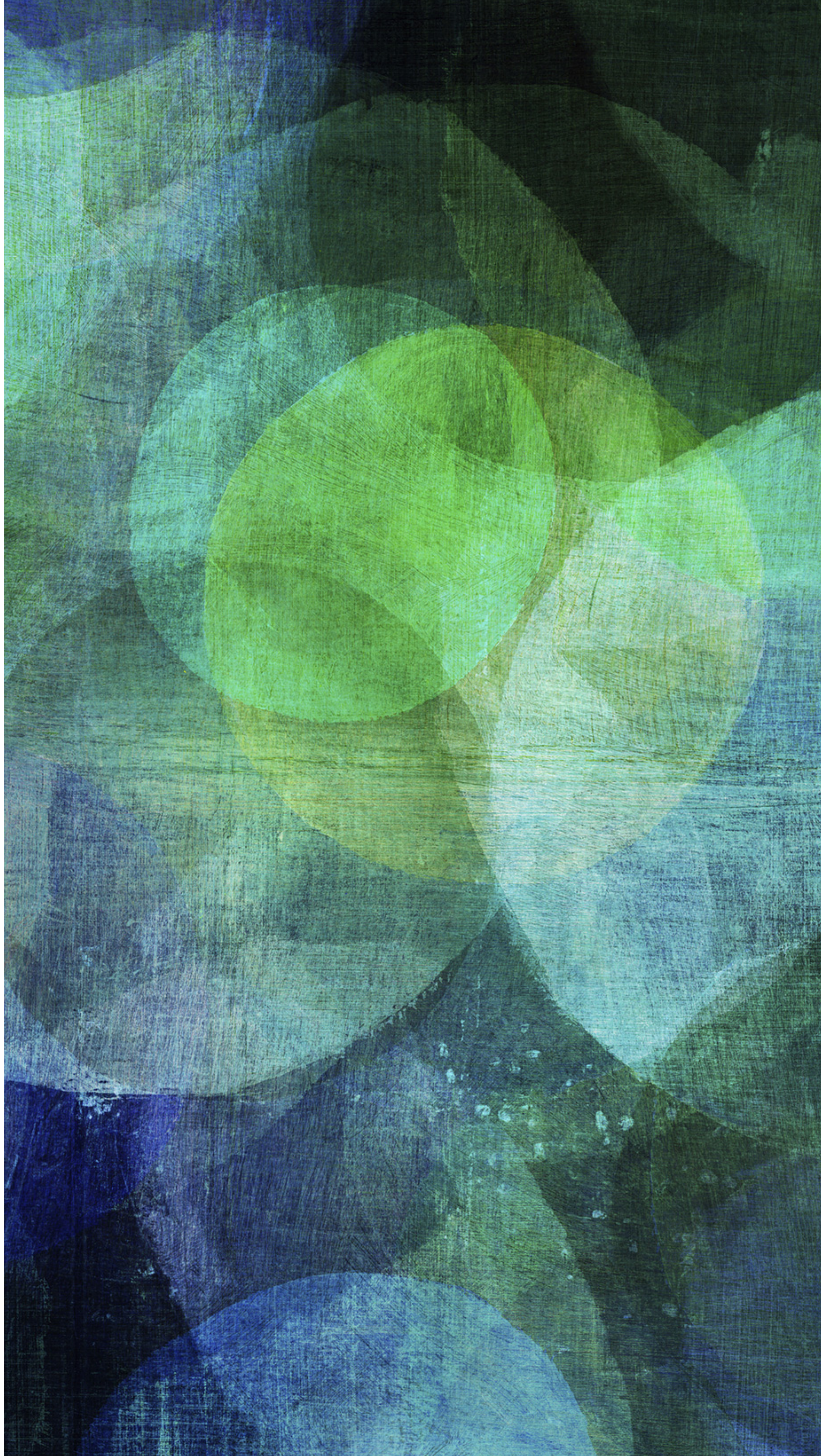
O^2 system



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, <u>towards higher bandwidths and commodity HW</u>**
  - 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…")

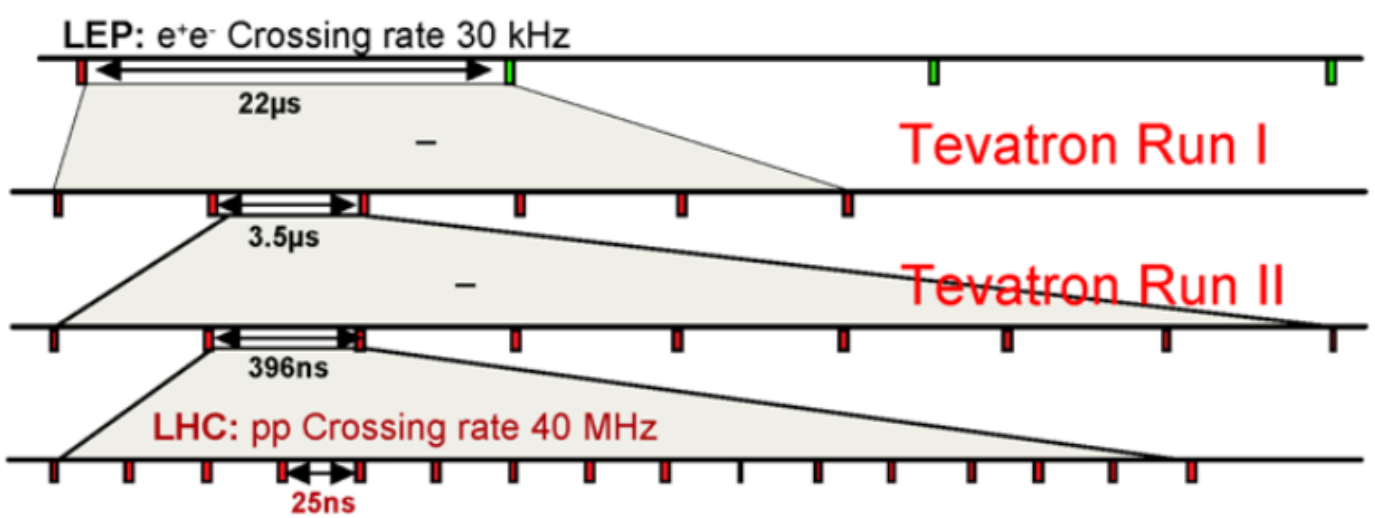
BACK-UP SLIDES



LHC: THE SOURCE

The clock source

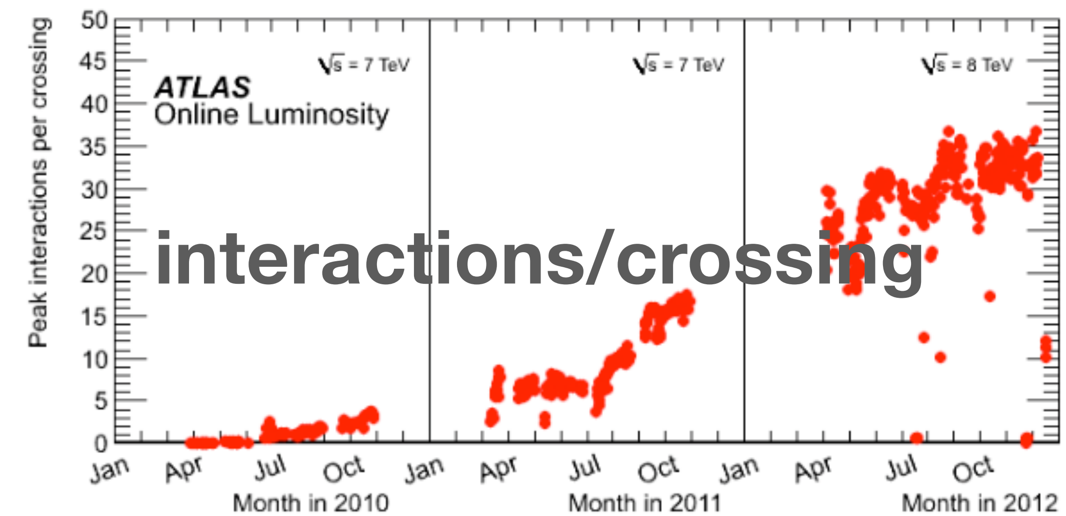
- ~3600 bunches in 27km
- distance bw bunches: 27km/3600 = 7.5m
- distance bw bunches in time: 7.5m/c = 25ns

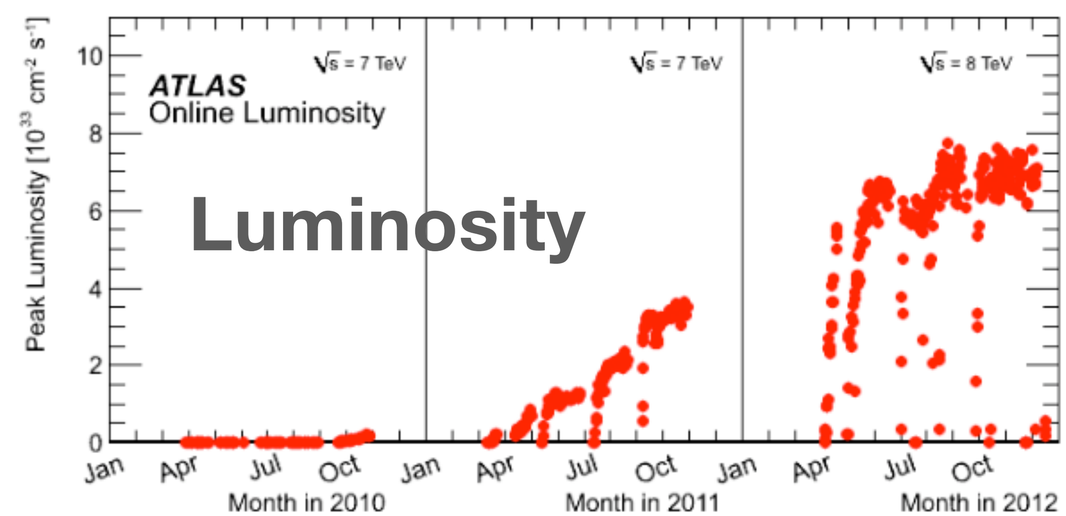


At full Luminosity, every 25ns, ~23 superimposed p-p interaction events

The pile-up source

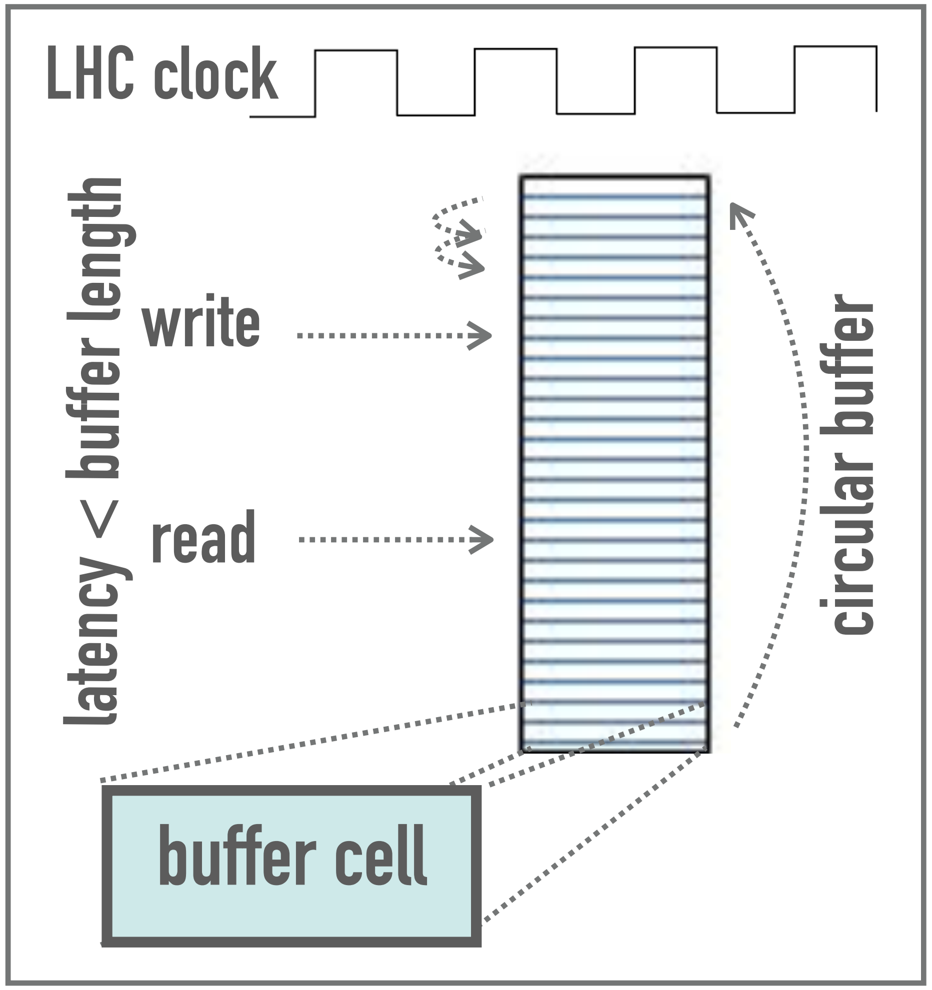
- more collisions/bunch crossing: ~23 at design luminosity





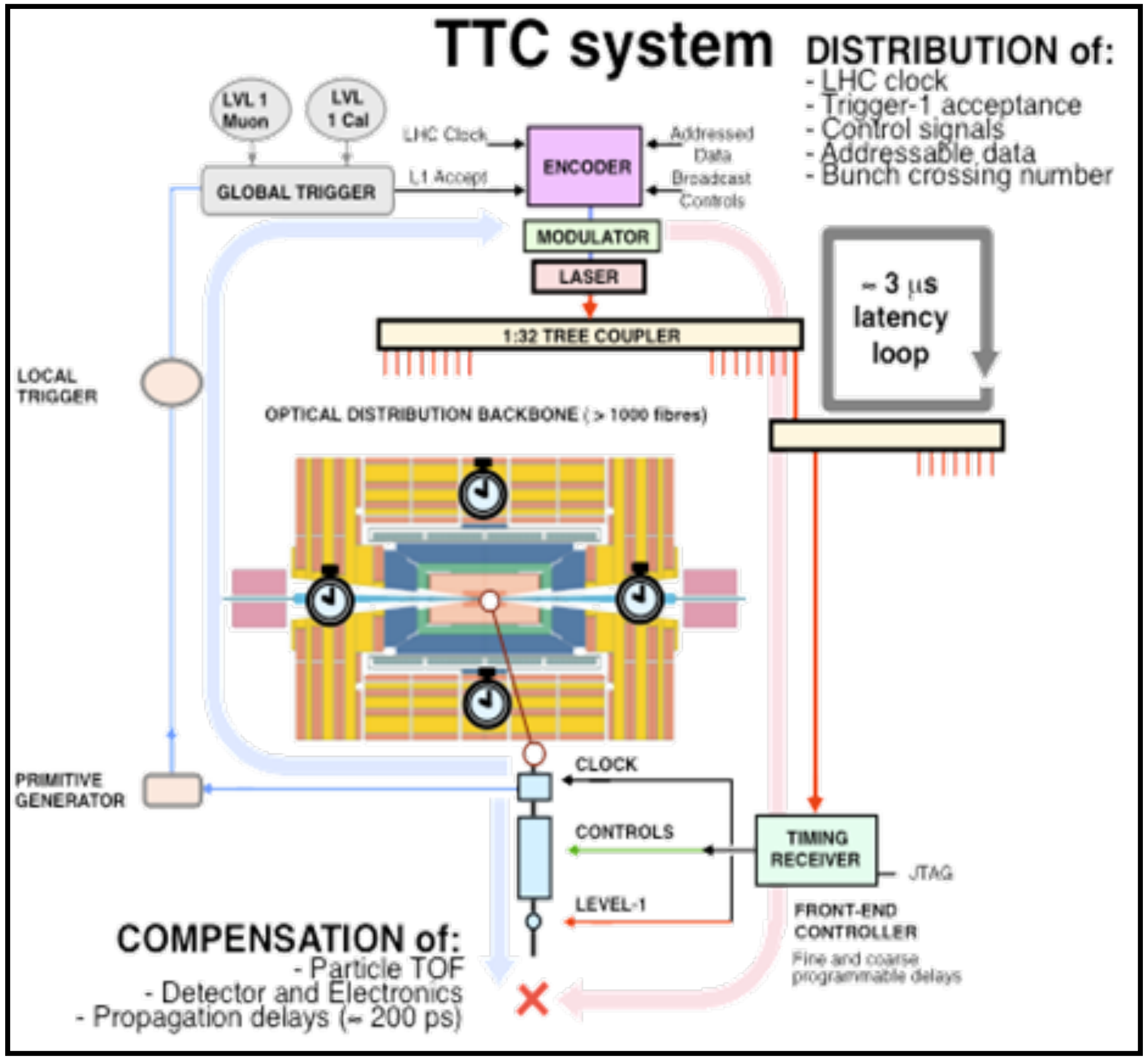
PIPELINED TRIGGERS

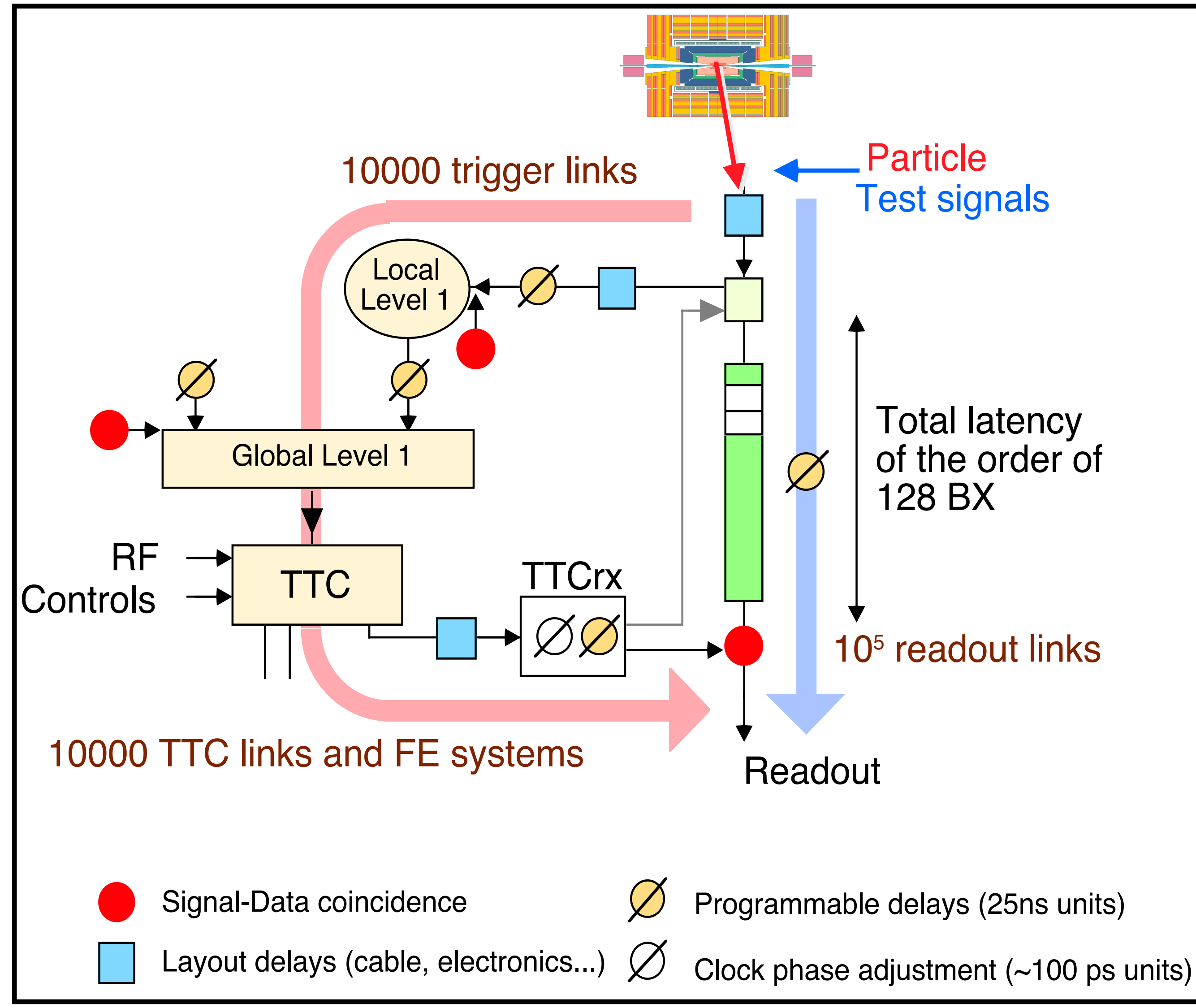
- **Allow trigger decision longer than clock tick (and no deadtime)**
  - Execute trigger selection in defined clocked steps (**fixed latency**)
  - Intermediate storage in stacked buffer cells
  - R/W pointers are moved by clock frequency



- **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



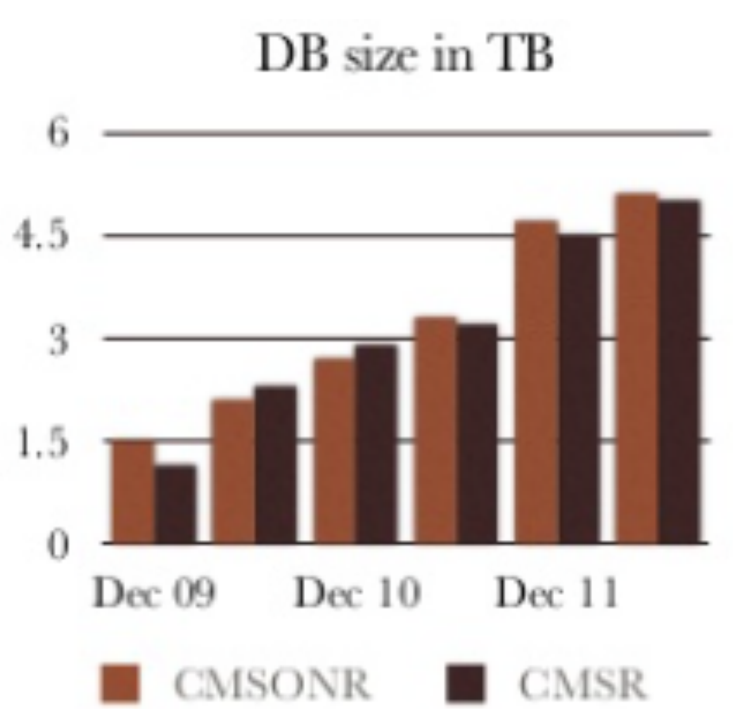


- **Common optical system: TTC**
  - radiation resistance
  - single high power laser
- **Large distribution**
  - experiments with ~10^7 channels

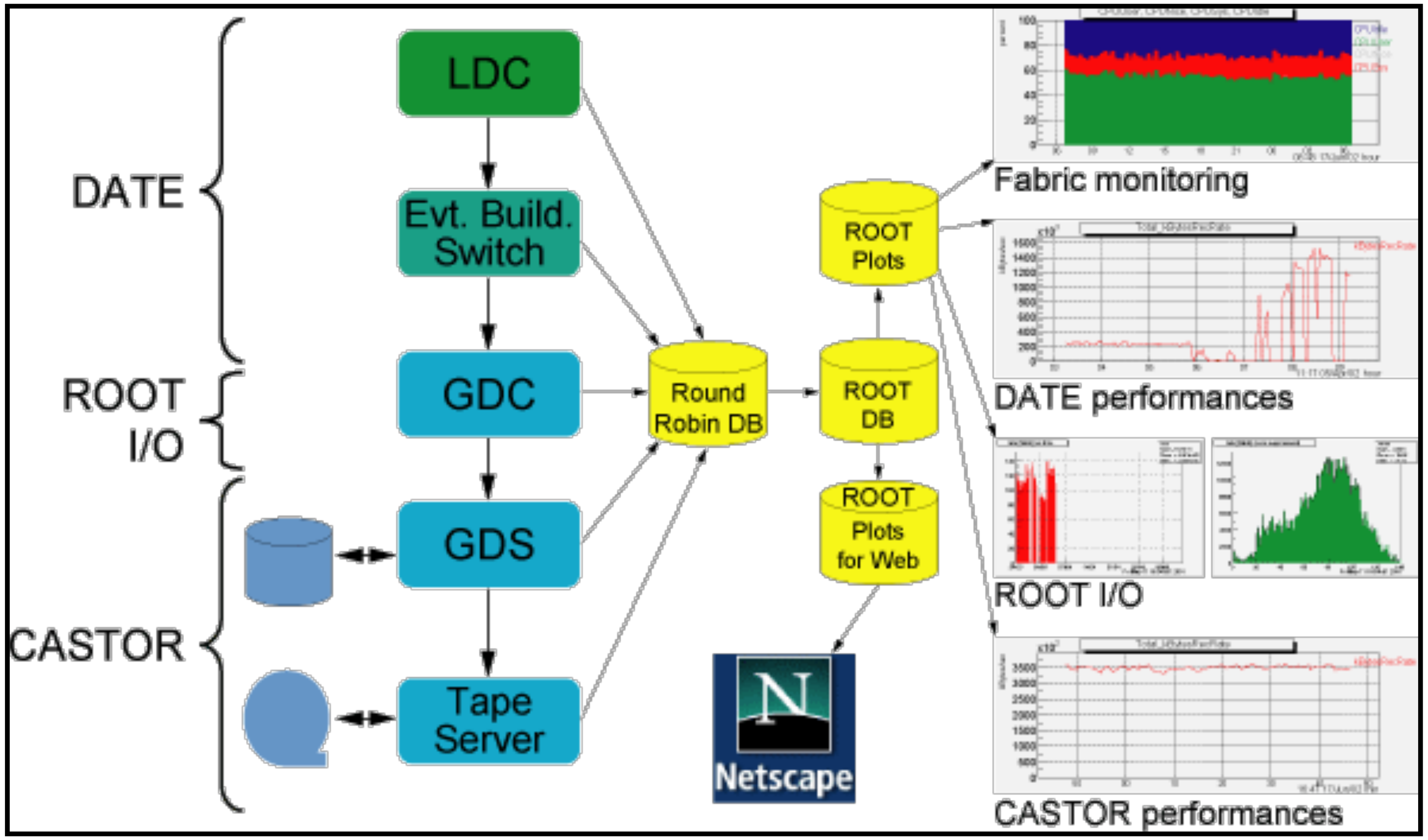
- **Align readout & trigger at (better than) 25ns and correct for**
  - time of flight (25 ns ≈ 7.5m)
  - cable delays (10cm/ns)
  - processing delays (~100 BCs)

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



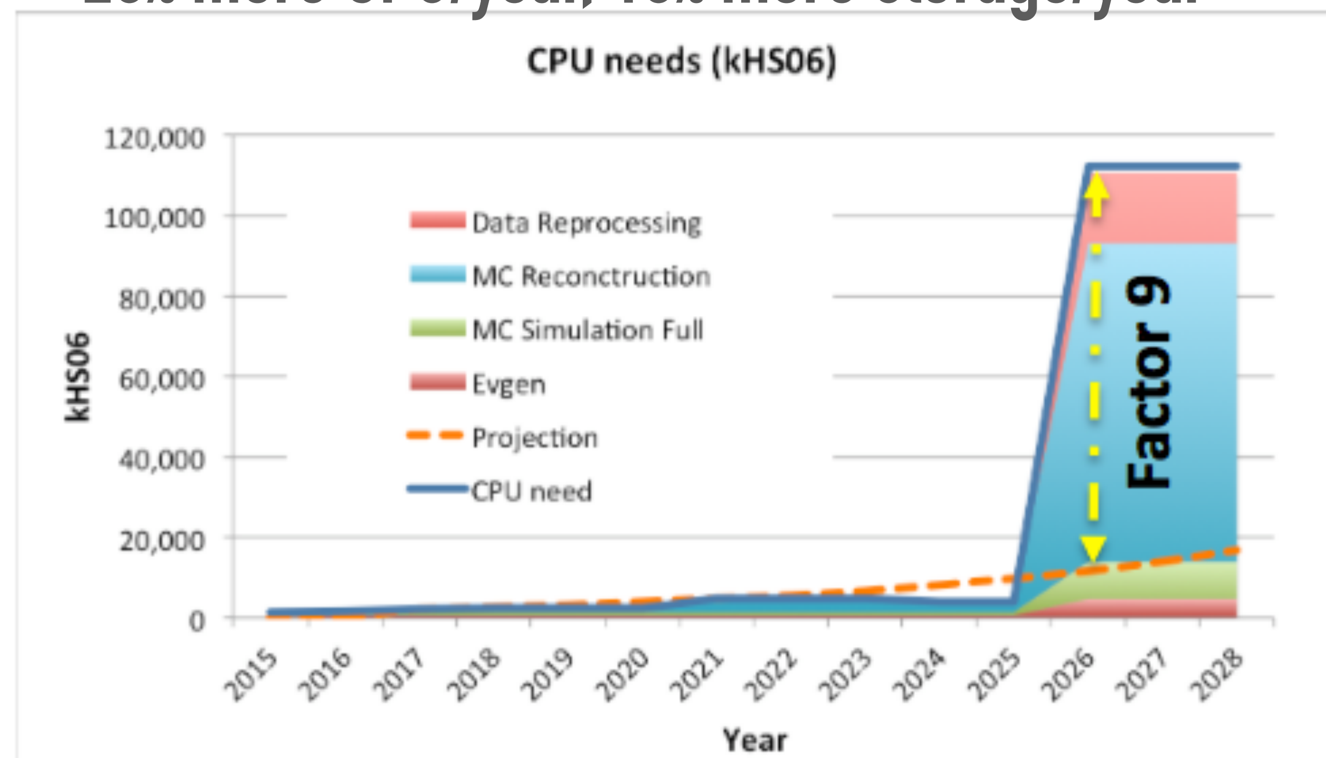
CMS DB grows about 1.5TB/year, condition data only a small fraction

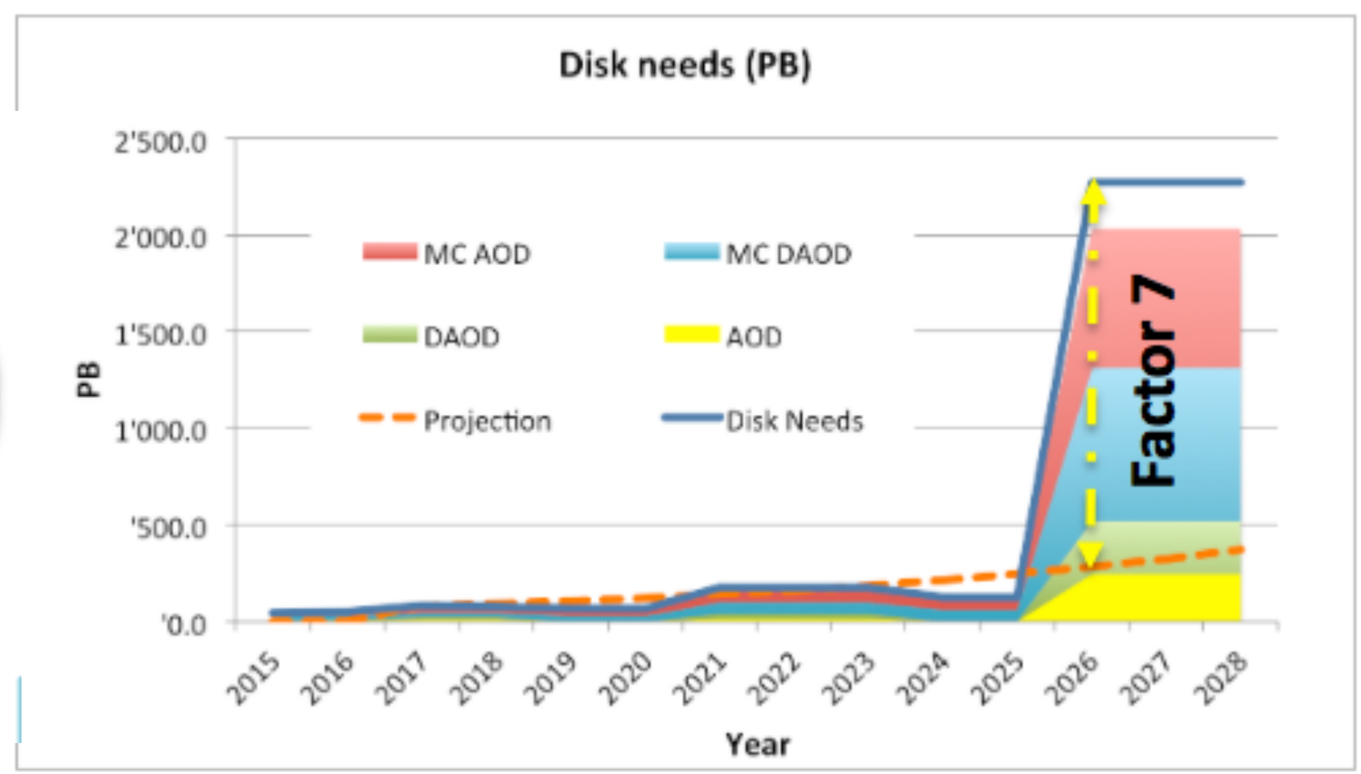


COMPUTING EVOLUTION FOR HL-LHC

- **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

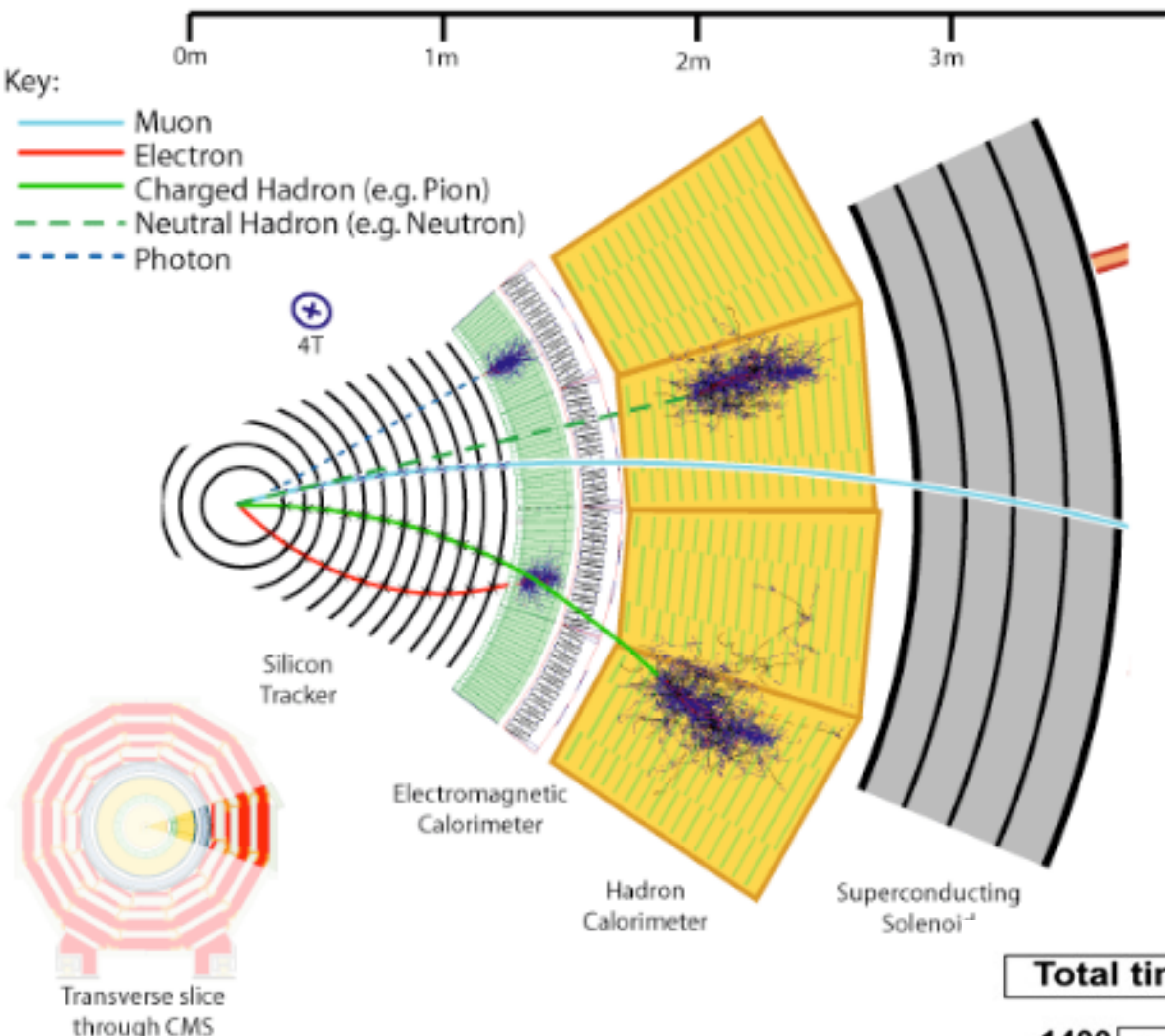
Projection of available resources in HL-LHC: 20% more CPU/year, 15% more storage/year





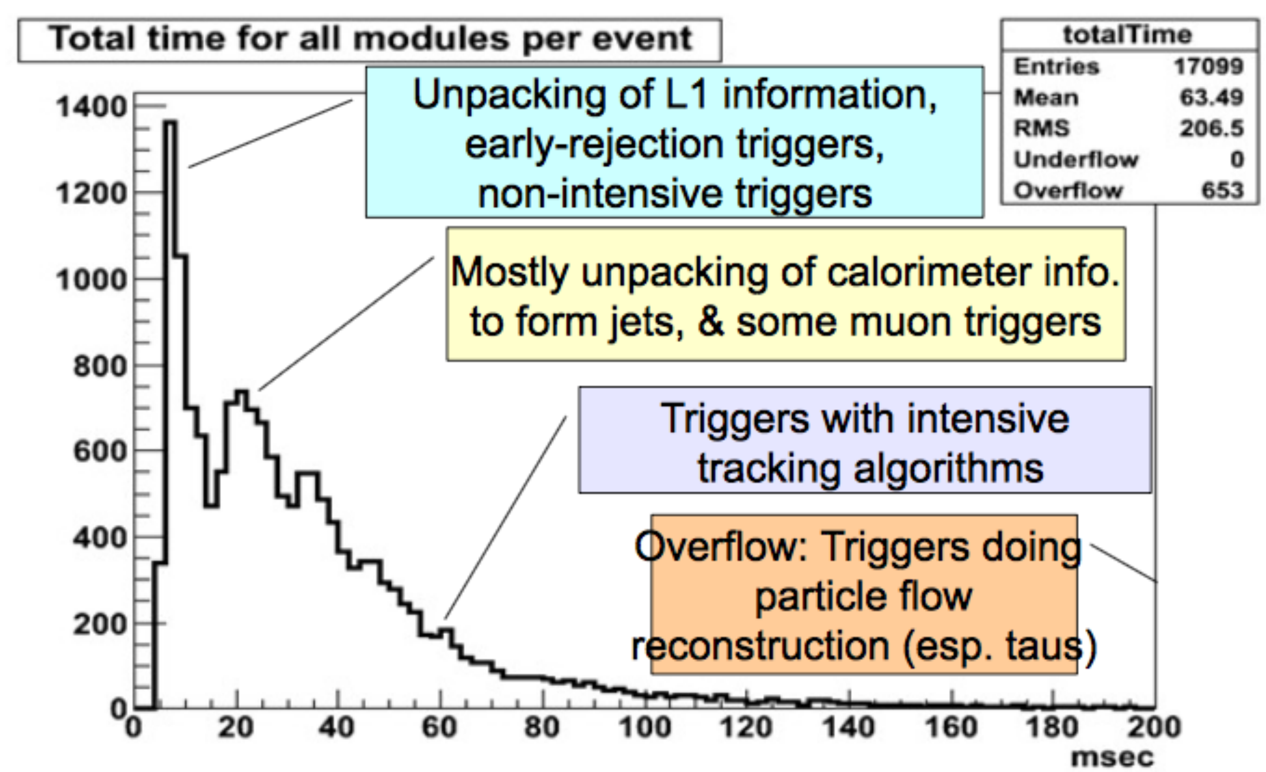
CALORIMETER TRIGGERS

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

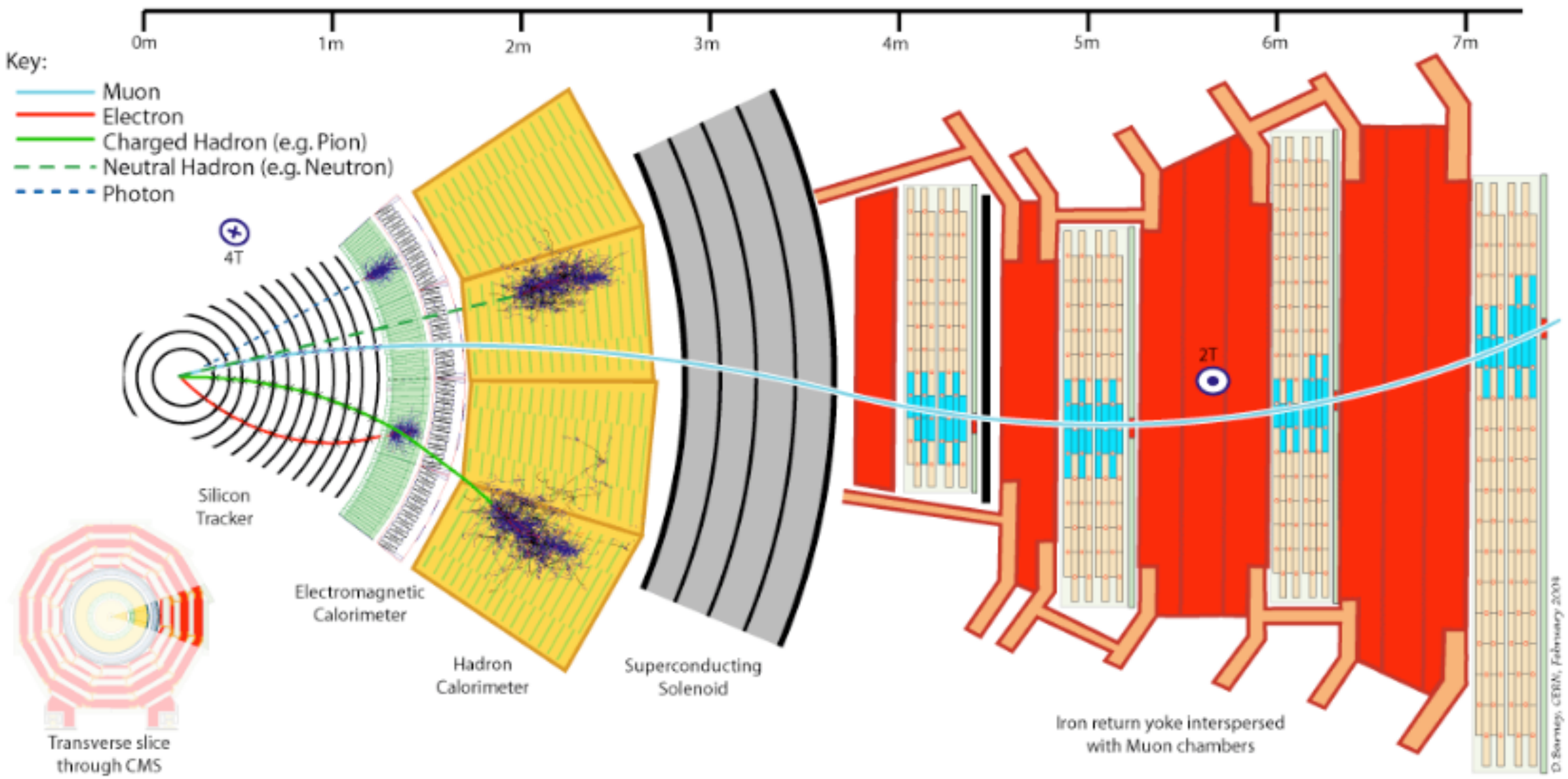


- ➡ **Fast and good resolution (LArg, PbW_4 for e-m)**
- ➡ **First-level processing (40MHz)**
  - ➡ "trigger towers" to reduce data (10-bit range)
  - ➡ sliding-window technique for local maxima
  - ➡ parallel algorithms for cluster shape and energy distribution

- ➡ **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



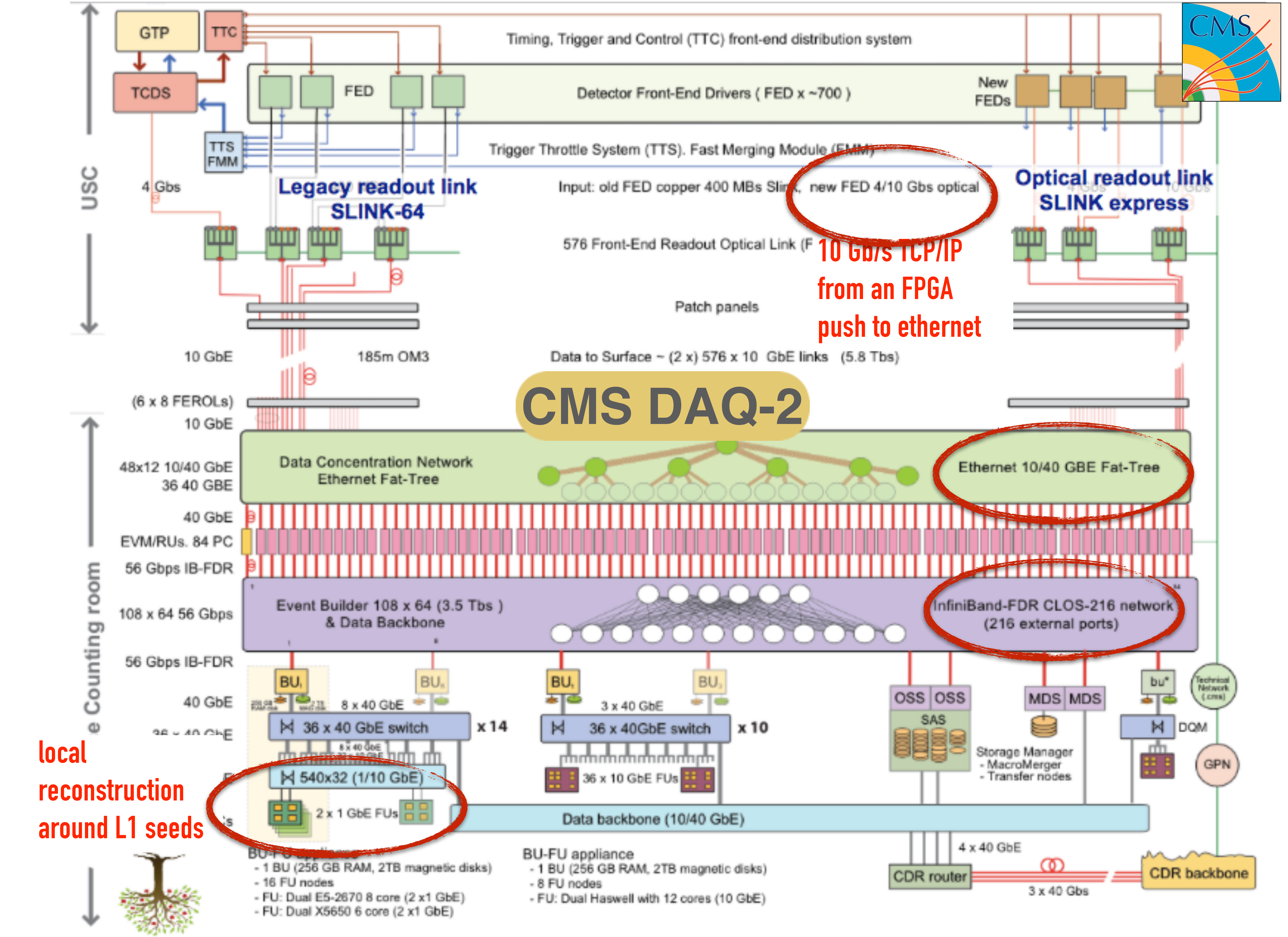
TRIGGERS FOR MUONS



- **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

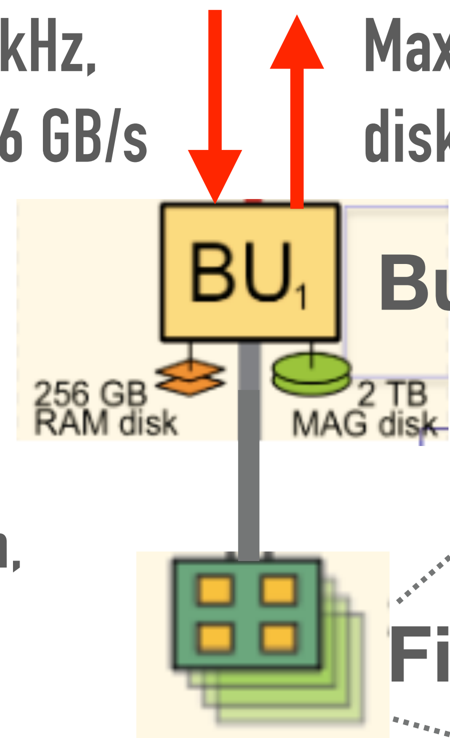

CMS
GTP
TTC
TCDS
TTS FMM
Timing, Trigger and Control (TTC) front-end distribution system
FED
Detector Front-End Drivers (FED x ~700)
New FEDs
Trigger Throttle System (TTS), Fast Merging Module (FMM)
USC
4 Gbs
Legacy readout link
SLINK-64
Input: old FED copper 400 MBs Slink, new FED 4/10 Gbs optical
Optical 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)
CMS DAQ-2
(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
Counting room
40 GbE
BU
8 x 40 GbE
36 x 40 GbE switch
x 14
3 x 40 GbE
36 x 40GbE switch
x 10
540x32 (1/10 GbE)
2 x 1 GbE FUs
36 x 10 GbE FUs
OSS
OSS
SAS
MDS
MDS
Storage Manager
- MacroMerger
- Transfer nodes
bu*
DQM
GPN
local
reconstruction
around L1 seeds
Data backbone (10/40 GbE)
4 x 40 GbE
CDR router
3 x 40 Gbs
CDR backbone
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)

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)

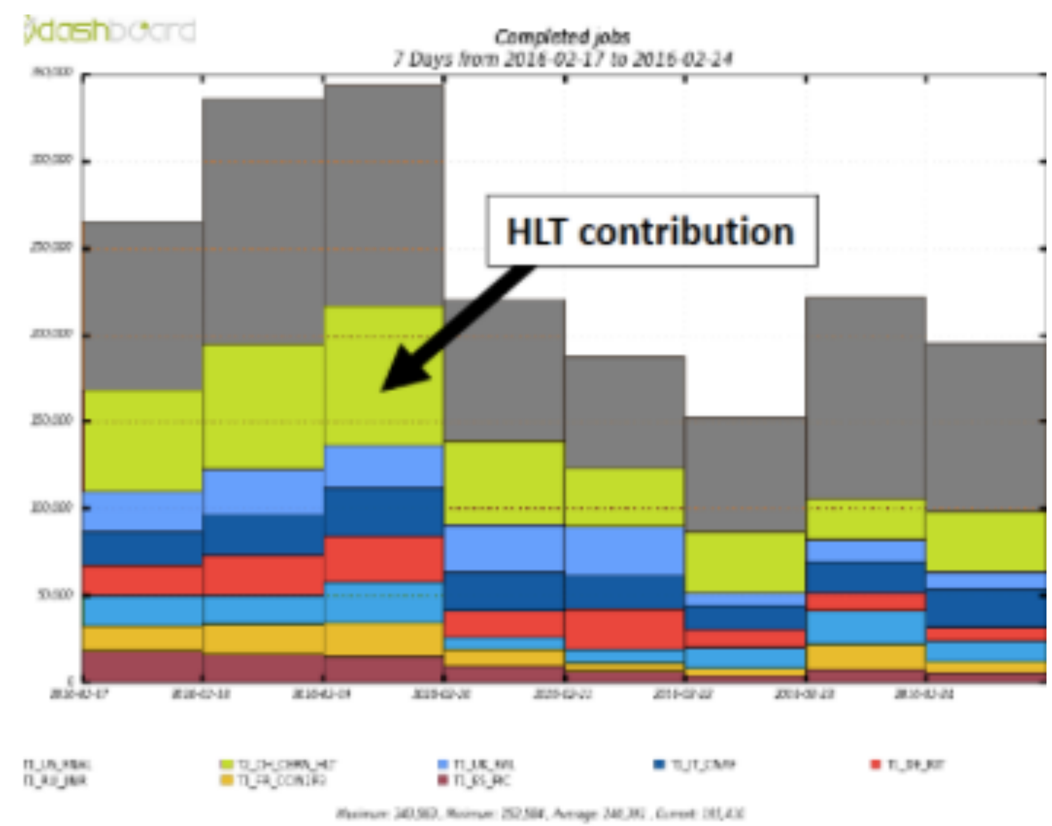
data, status, configuration, latency

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)

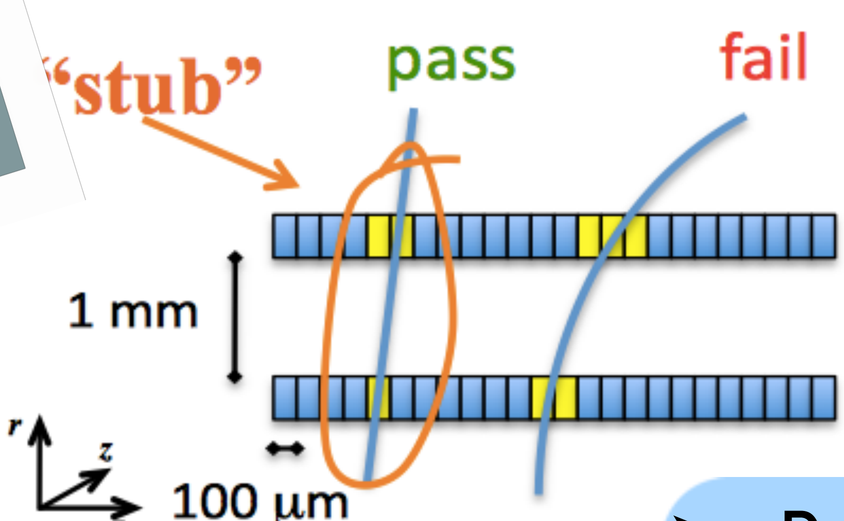
CMS: LOW-P_T TRACK FILTERING

Track filtering (low p_T)

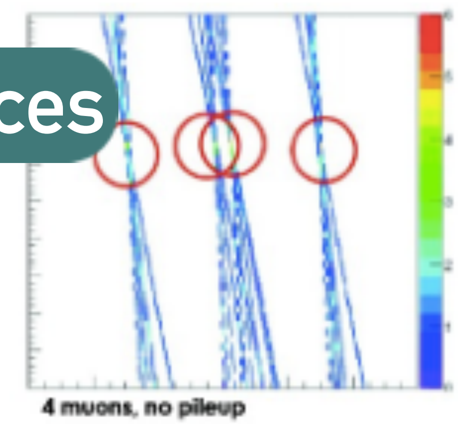
Reduce readout 40 ➟ 1MHz by detector coincidences

- ➡ **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

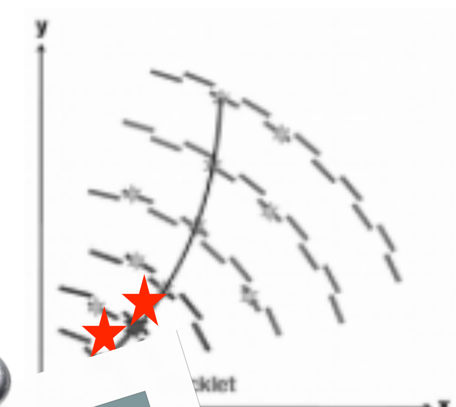
40MHz



Track finding options

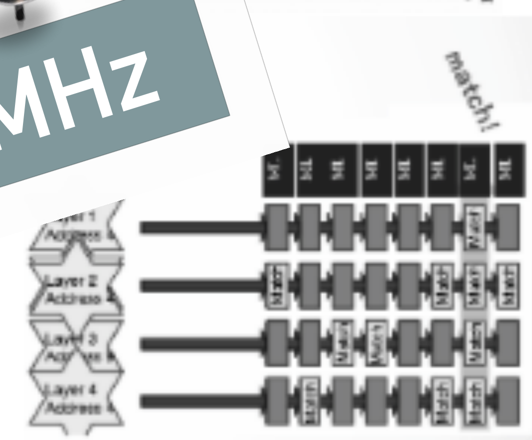


Hough Transform



Tracklets

1MHz

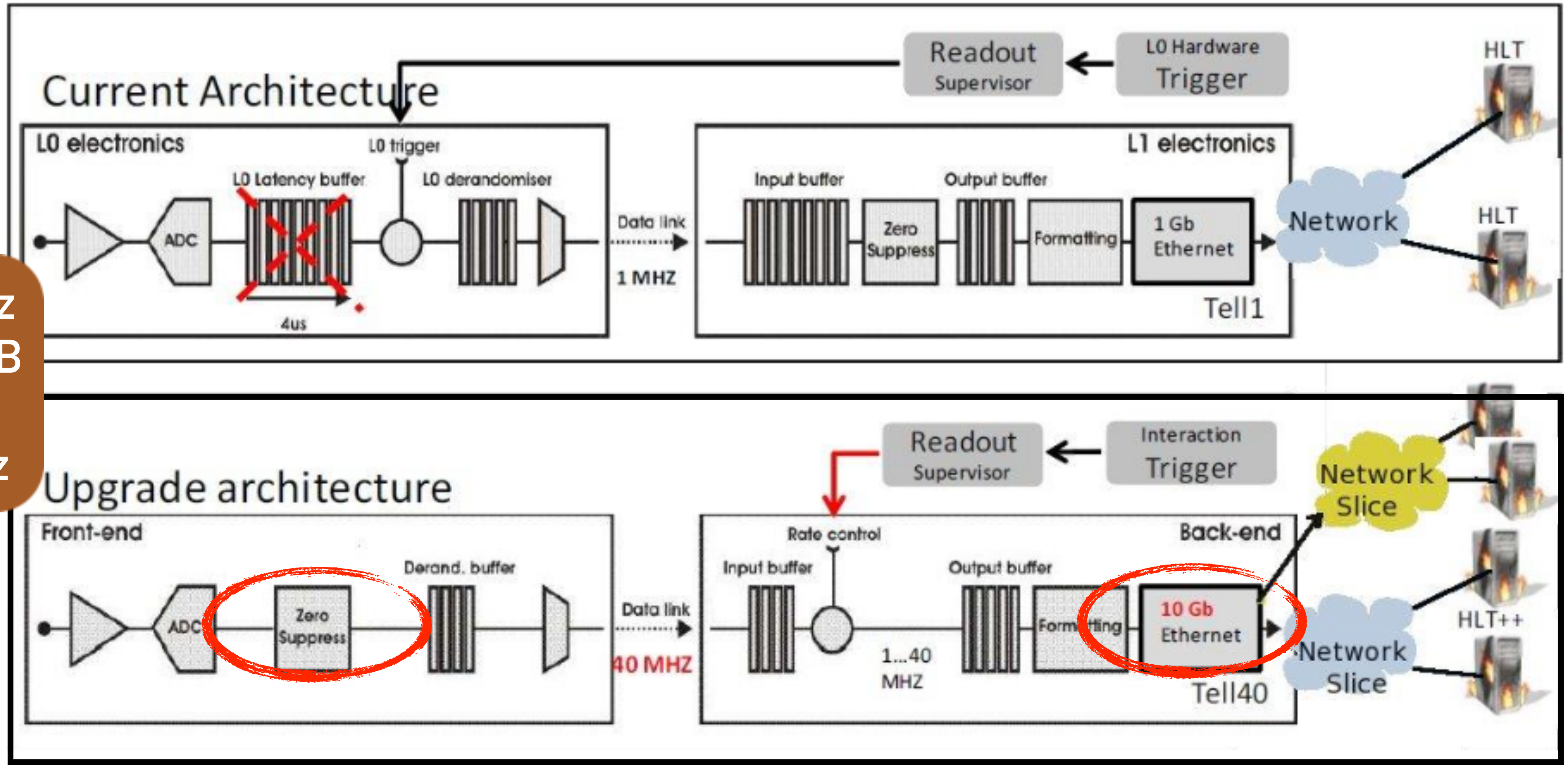


Associative Memories

- ➤ 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

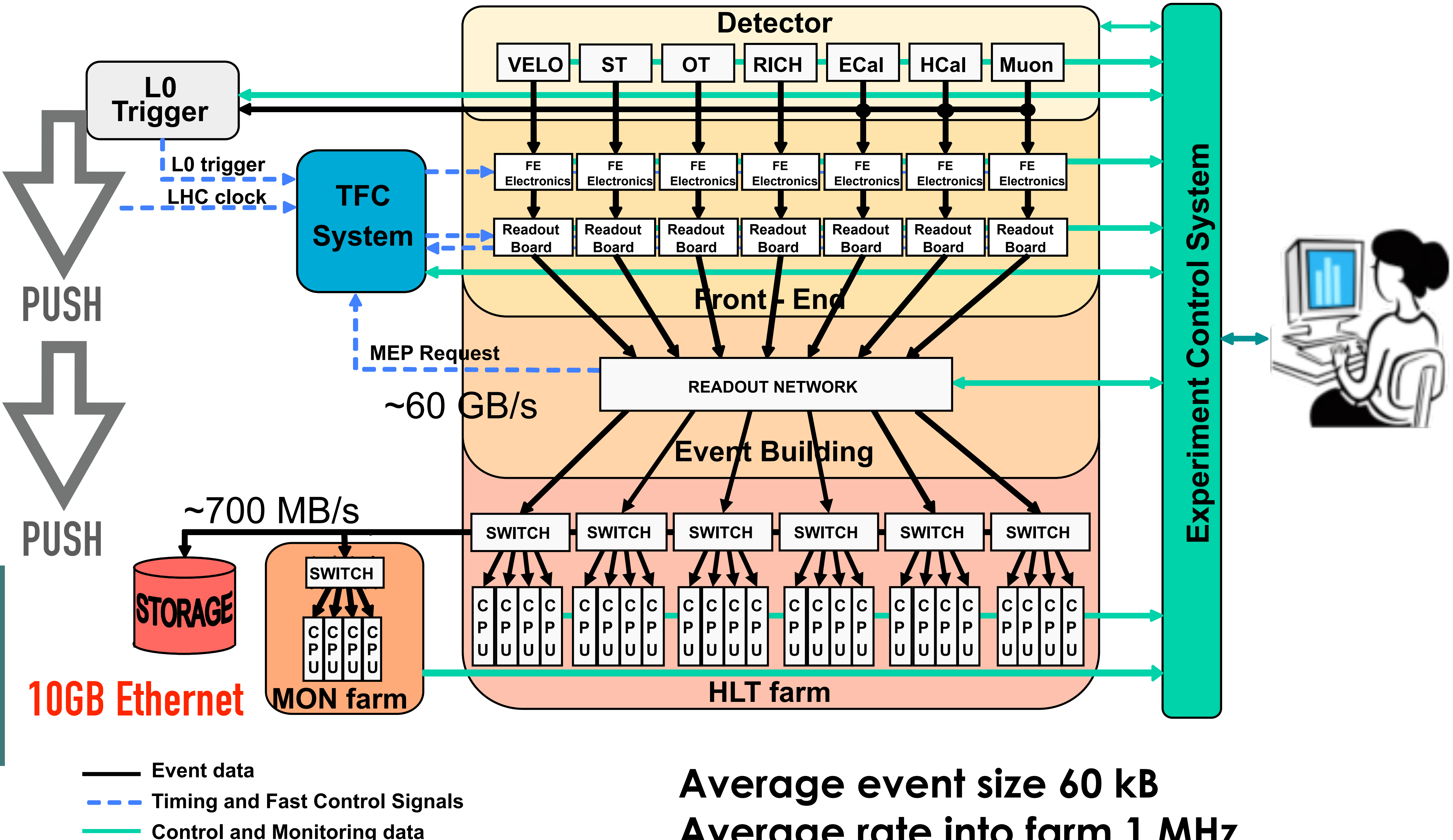


- Need zero-suppressing on front-end electronics
- A single, high performance, custom FPGA-card (PCIe40)
  - 8800 (# VL) * 4.48 Gbit/s (wide mode) => 40 Tbps
- Single board up to 100 Gbits/s (to match DAQ links in 2018)
- Event-builder with 100 Gbit/s technology and data centre-switches

TDAQ ARCHITECTURE IN RUN-2

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

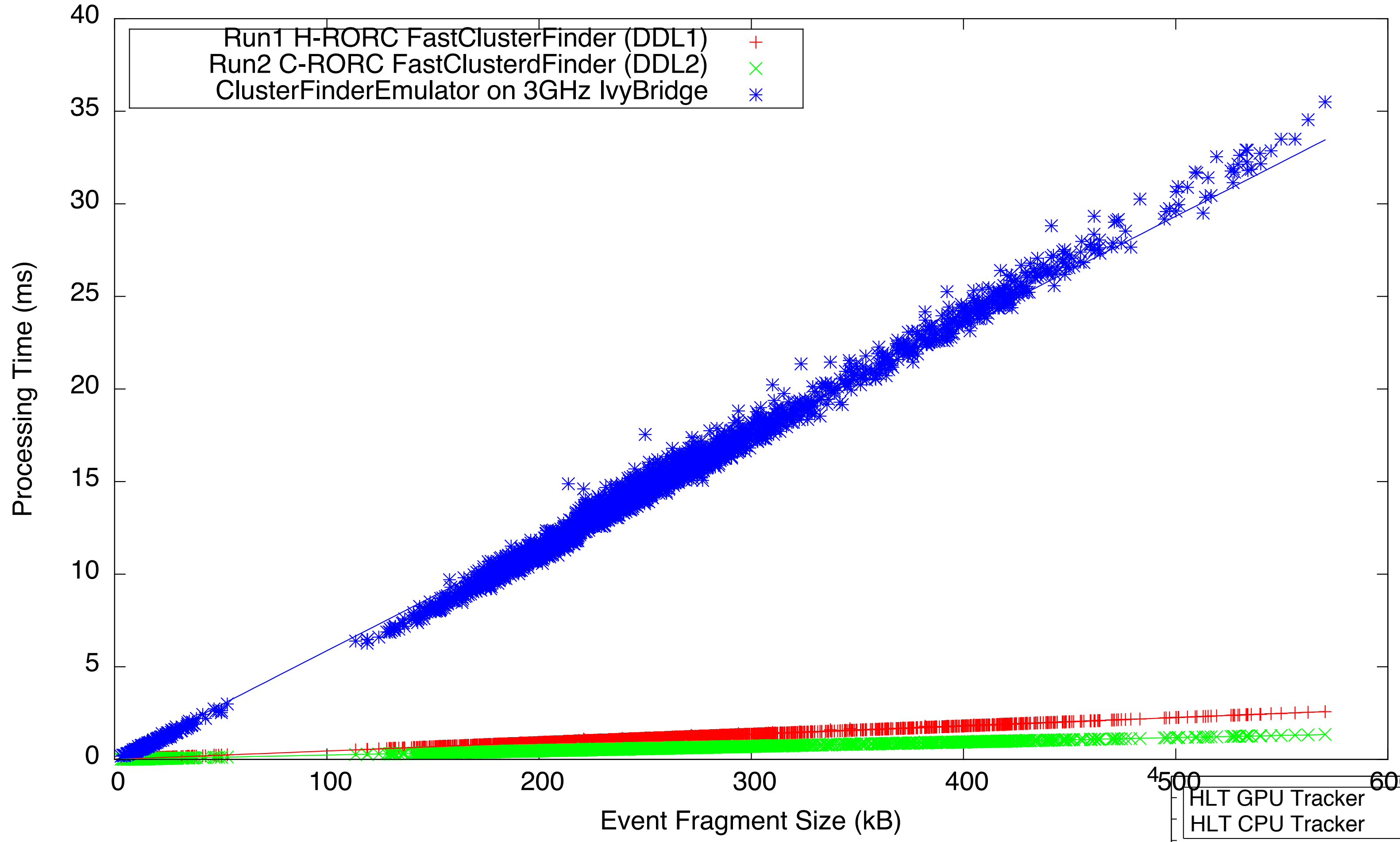
62 sub-farms, total 1780 nodes, with edge-routers (12 Gbps)



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

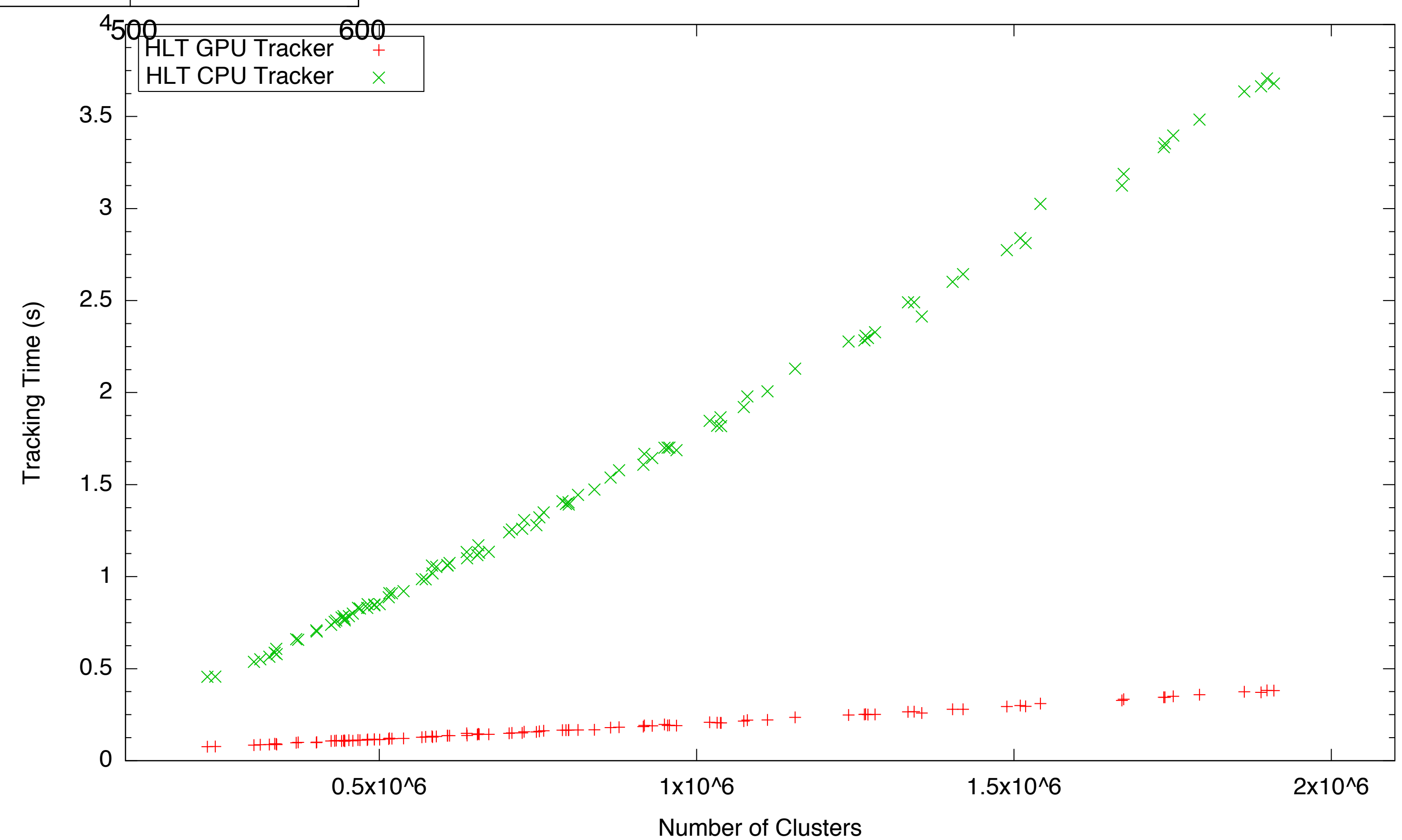
- **Small event, at high rate: ask for optimized transmission**
  - TTC system is used to assign IP addresses to RO boards
  - Ethernet UDP, with 10-15 events packed ⇒ ~ **80 kHz**

HARDWARE ACCELERATION WITH FPGAS AND GPUS

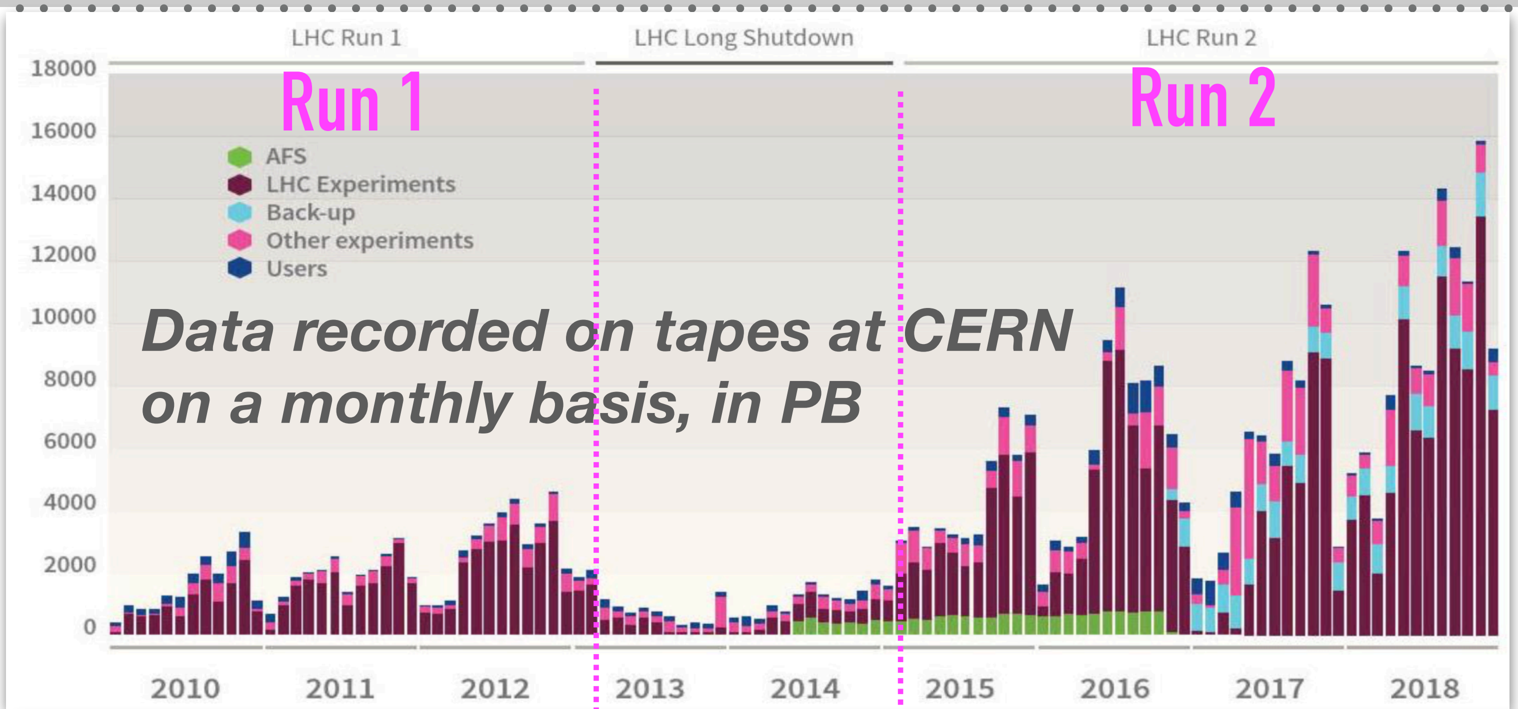


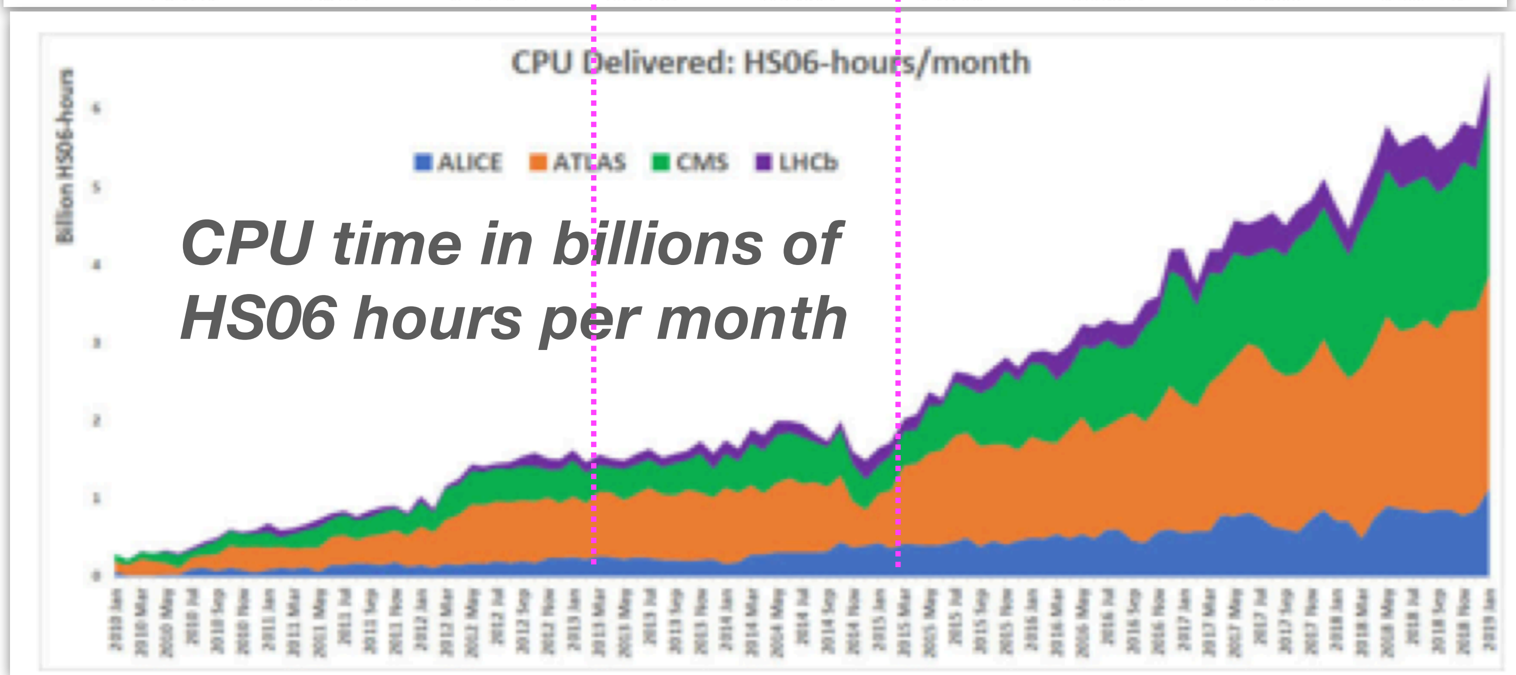
Tracking time of HLT TPC Cellular Automata tracker on Nehalem CPU (6Cores) and NVIDIA Fermi GPU.

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



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