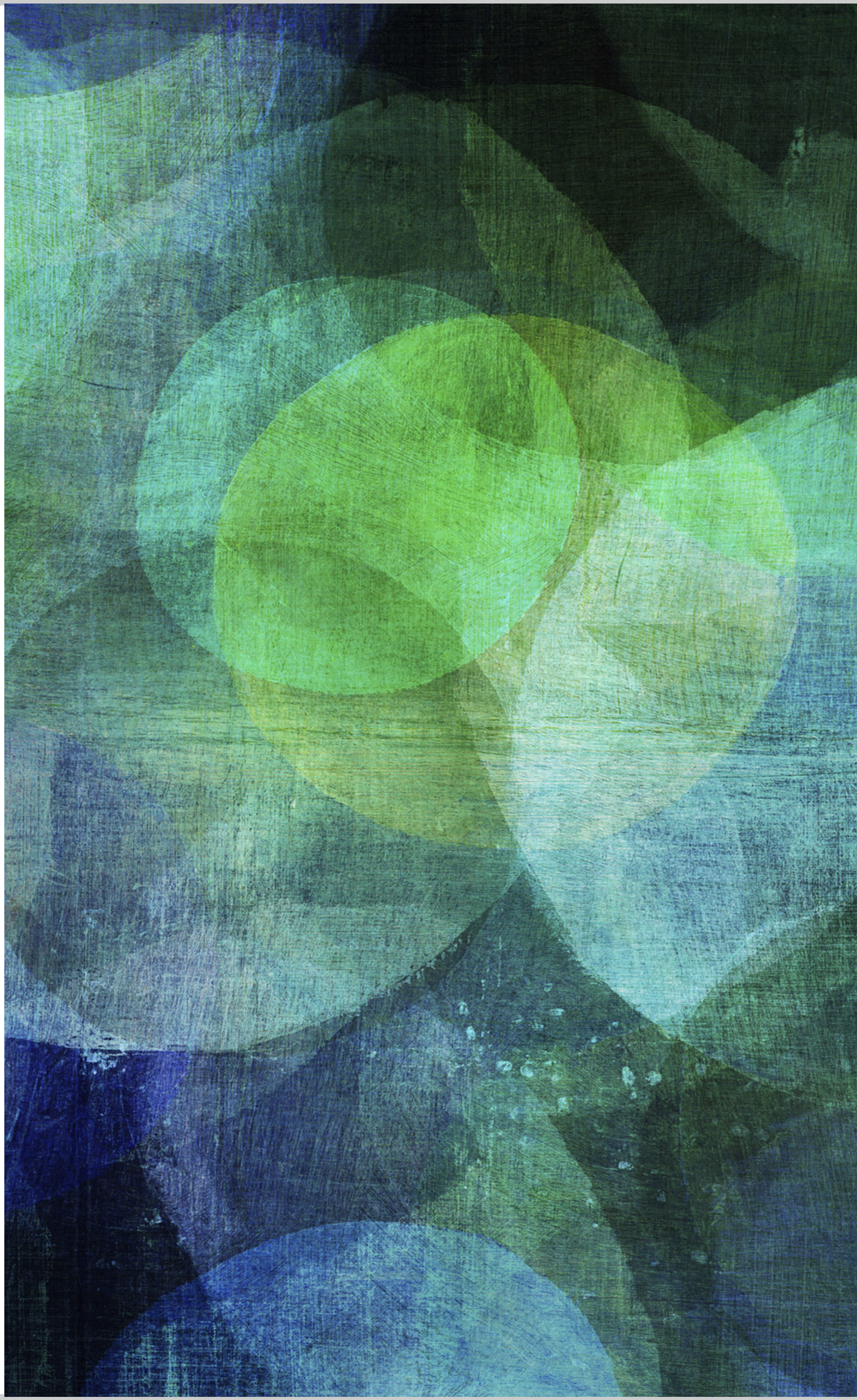
ISOTDAQ 2019 - International School of Trigger and Data AcQuisition

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

F.Pastore (Royal Holloway Un. of London)

THE CONTENTS OF THIS SEMINAR

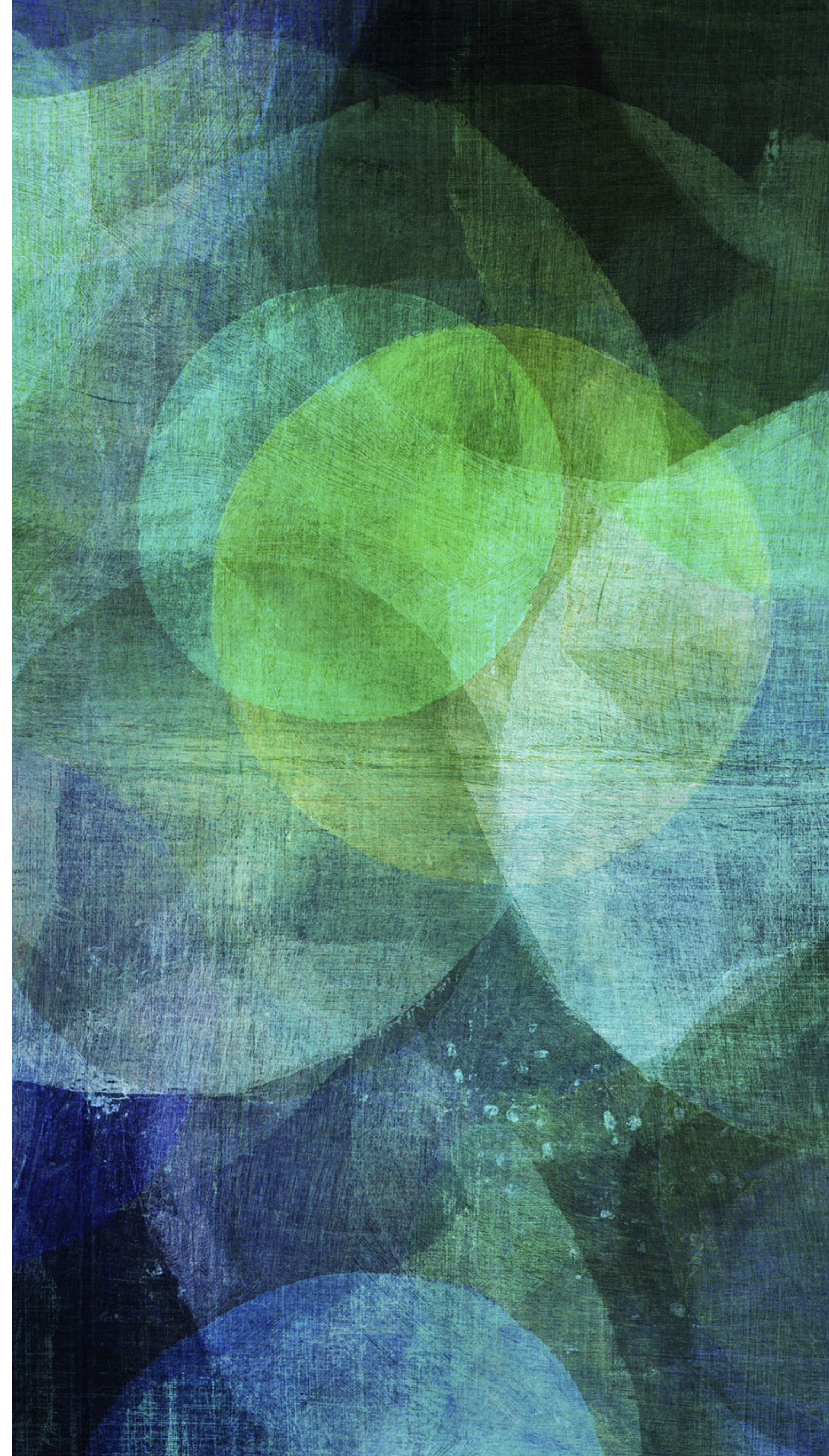
- ➡ LHC environment
- ➡ Trigger and DAQ design for experiments
  - ➡ First-level trigger & electronics
  - ➡ Software triggers and farms
  - ➡ DAQ technology for network and readout
- ➡ High Luminosity LHC: how changing things?
  - ➡ ATLAS, CMS, LHCb, ALICE in different phases
  - ➡ Technology and general trends
- ➡ Spotlight upgrade examples



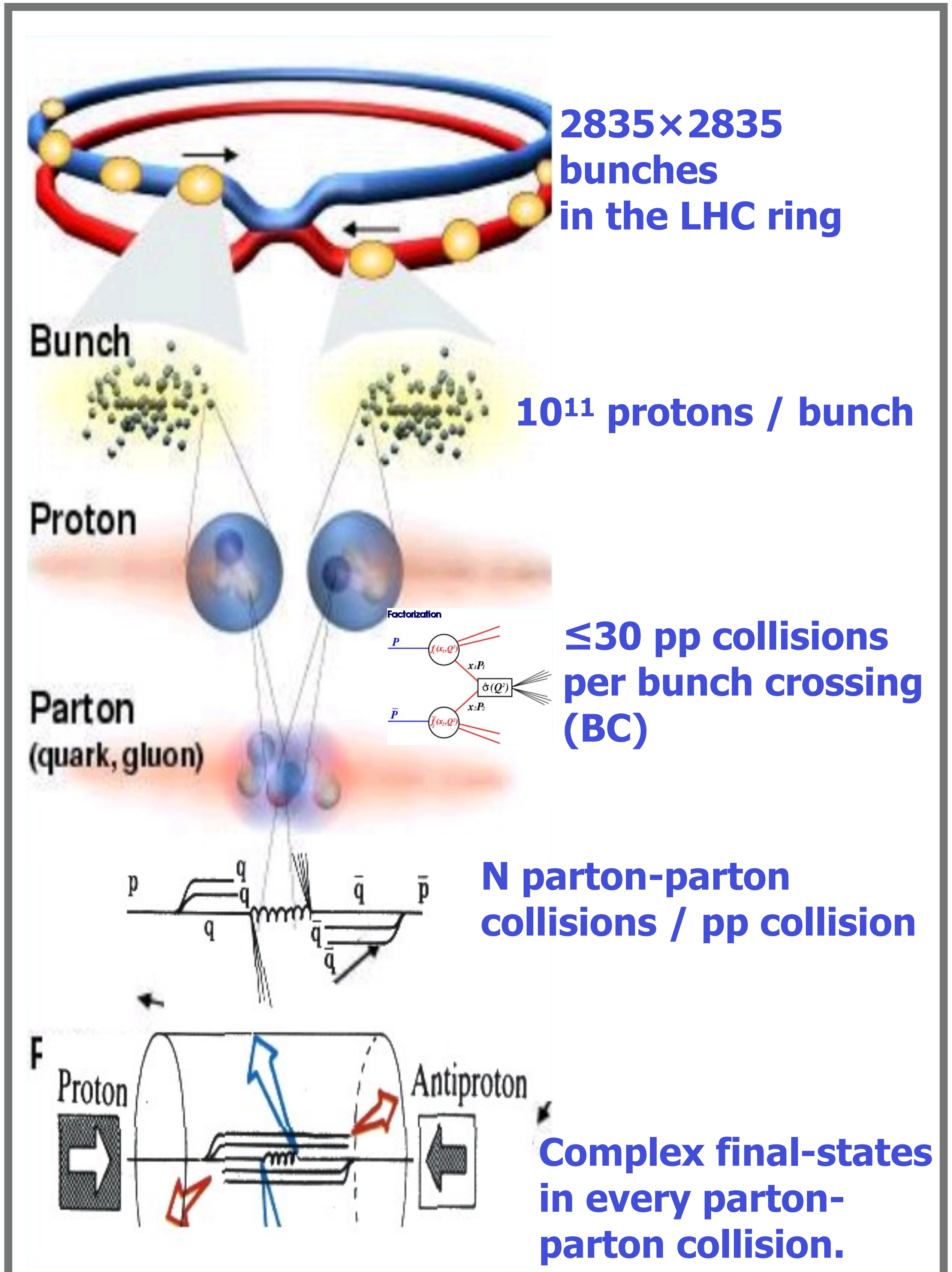
Acknowledgments to a lot of people, also present here

THE LHC PROJECT AND ITS EVOLUTION

What can we do with a boson factory machine?



LHC ENGINE AND ITS PRODUCTS



design parameters

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

$$R = \sigma_{in} \times L$$

- **Why high energy protons?**
  - Discovery potential at high energy
  - But composite particles: abundant not-interesting low momentum transfer interactions (QCD background)
- **Why high luminosity?**
  - Look at very rare processes
  - Close collisions in space and time
    - Large proton bunches (1.5×10^{11})
    - Fixed **frequency:** 40MHz (1/25ns)

Few rare high-E events overwhelmed in abundant low-E background

LHC EXPERIMENTS FOR A DISCOVERY MACHINE

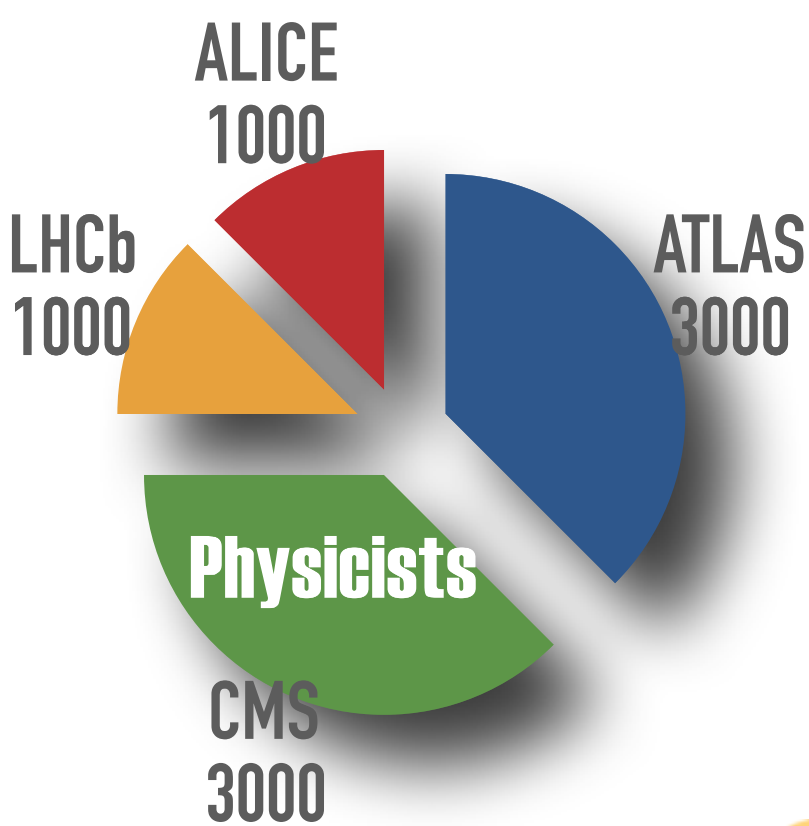
Goal: explore TeV energy scale to find New Physics beyond Standard Model

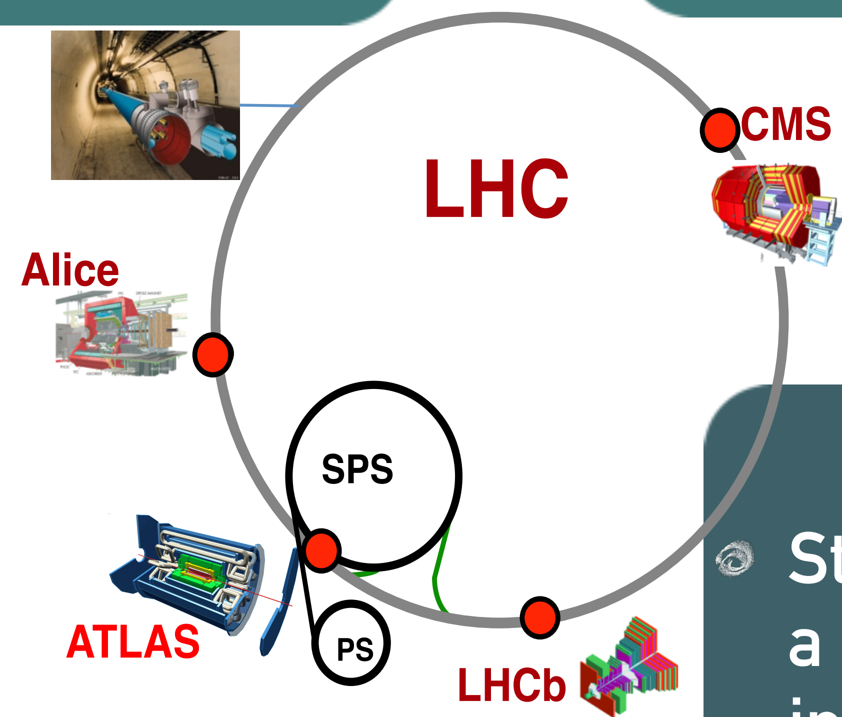
ATLAS & CMS

- 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

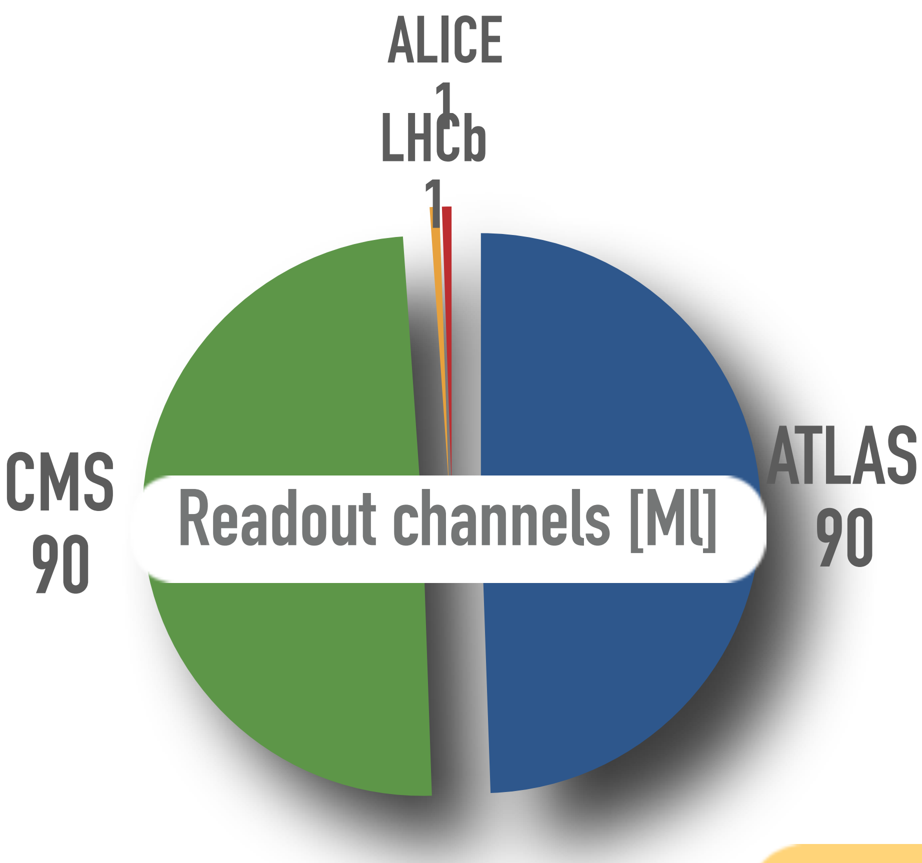
Goal: explore TeV energy scale to find New Physics beyond Standard Model

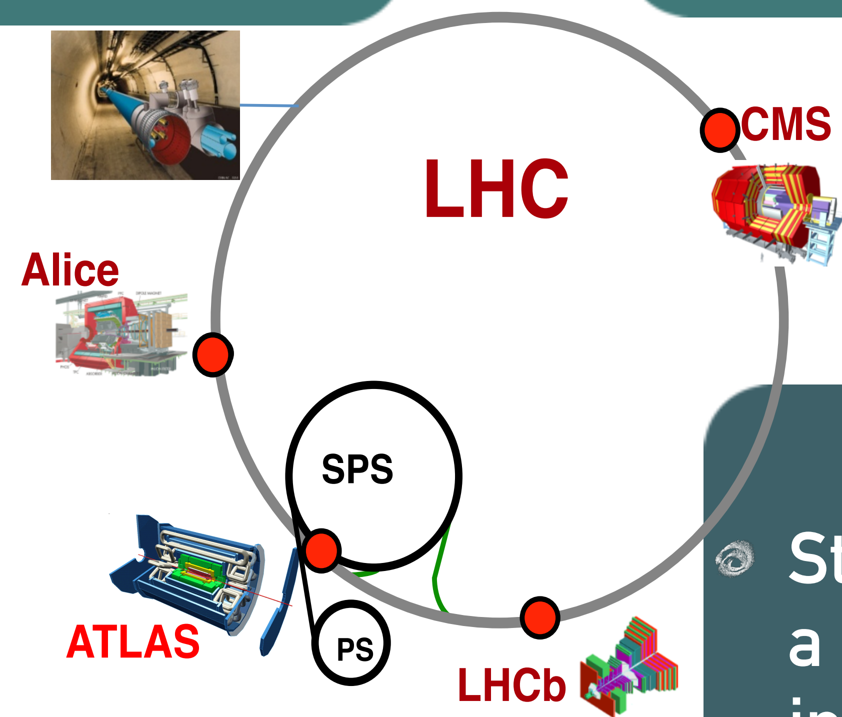
ATLAS & CMS

- 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





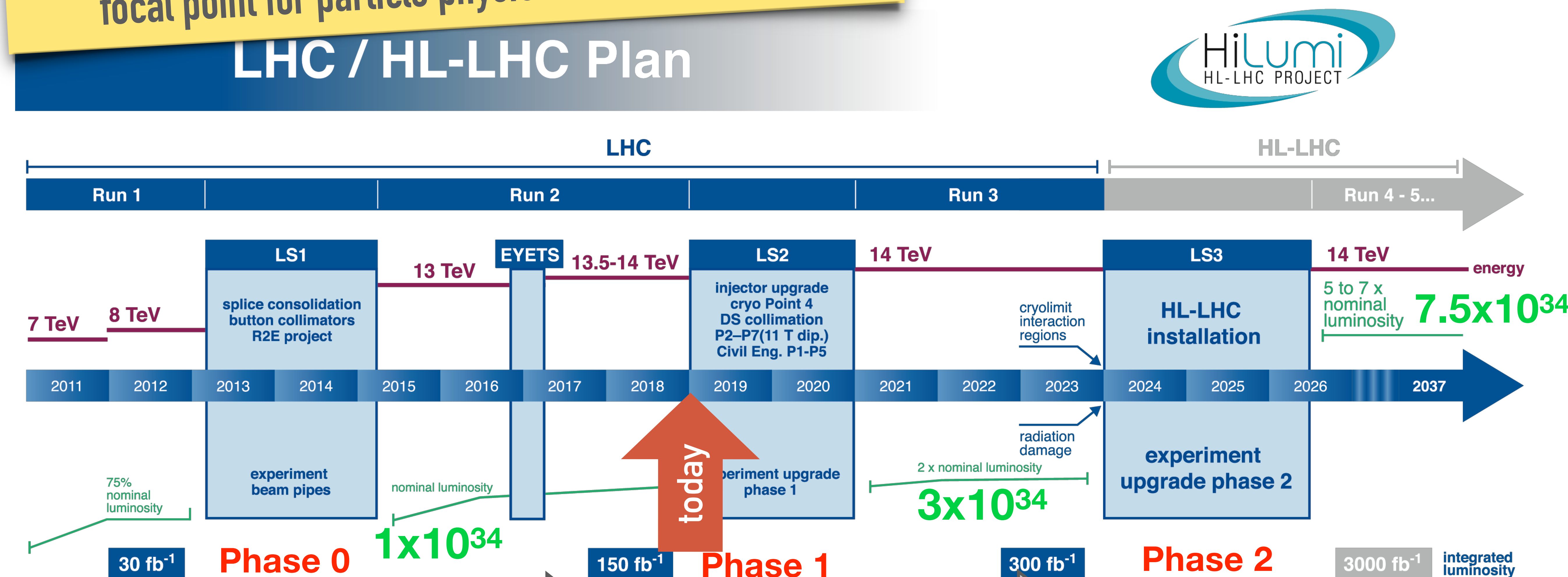
ALICE

- 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”

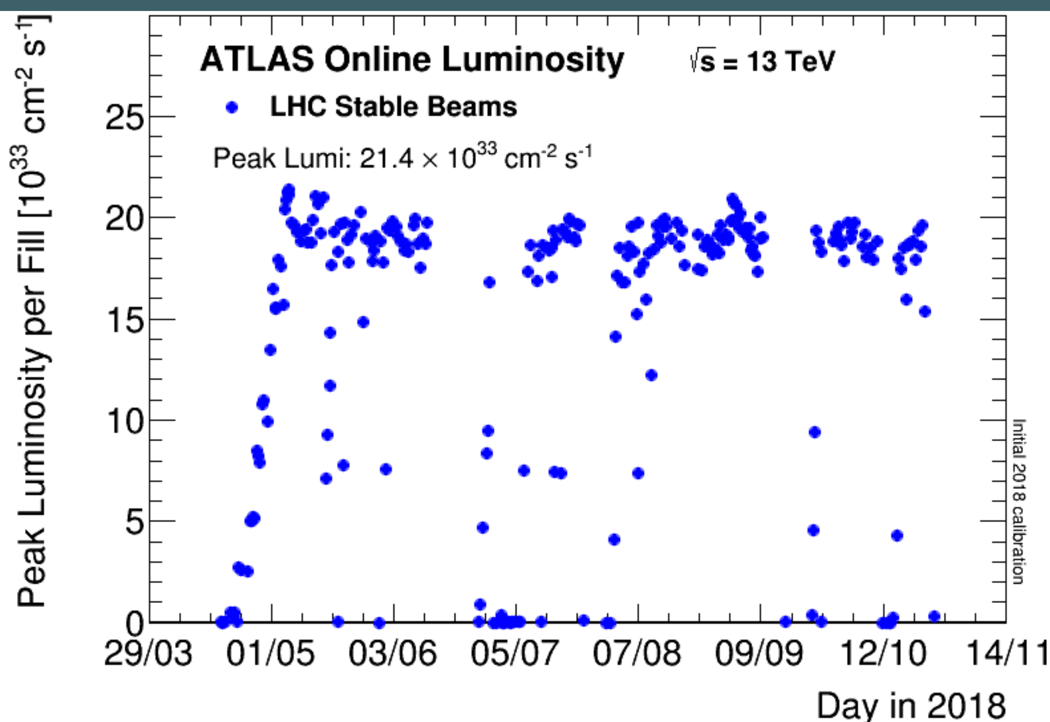


Consolidation for all experiments → Major Upgrade in ALICE and LHCb → Major Upgrade in ATLAS and CMS

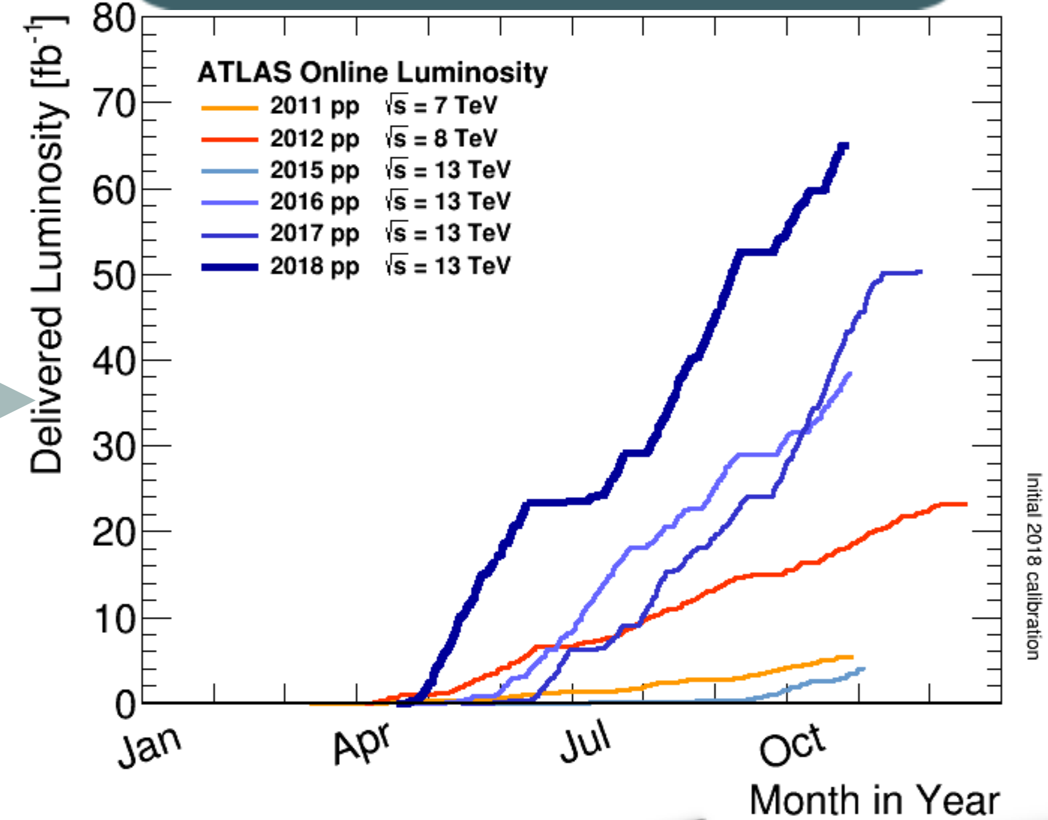
- **Starting from Run 3, requirements will go beyond design specifications**
  - Try to improve or at least maintain performance of present detectors
  - Improve bandwidth and processing capabilities

LHC AS TODAY – END OF RUN2

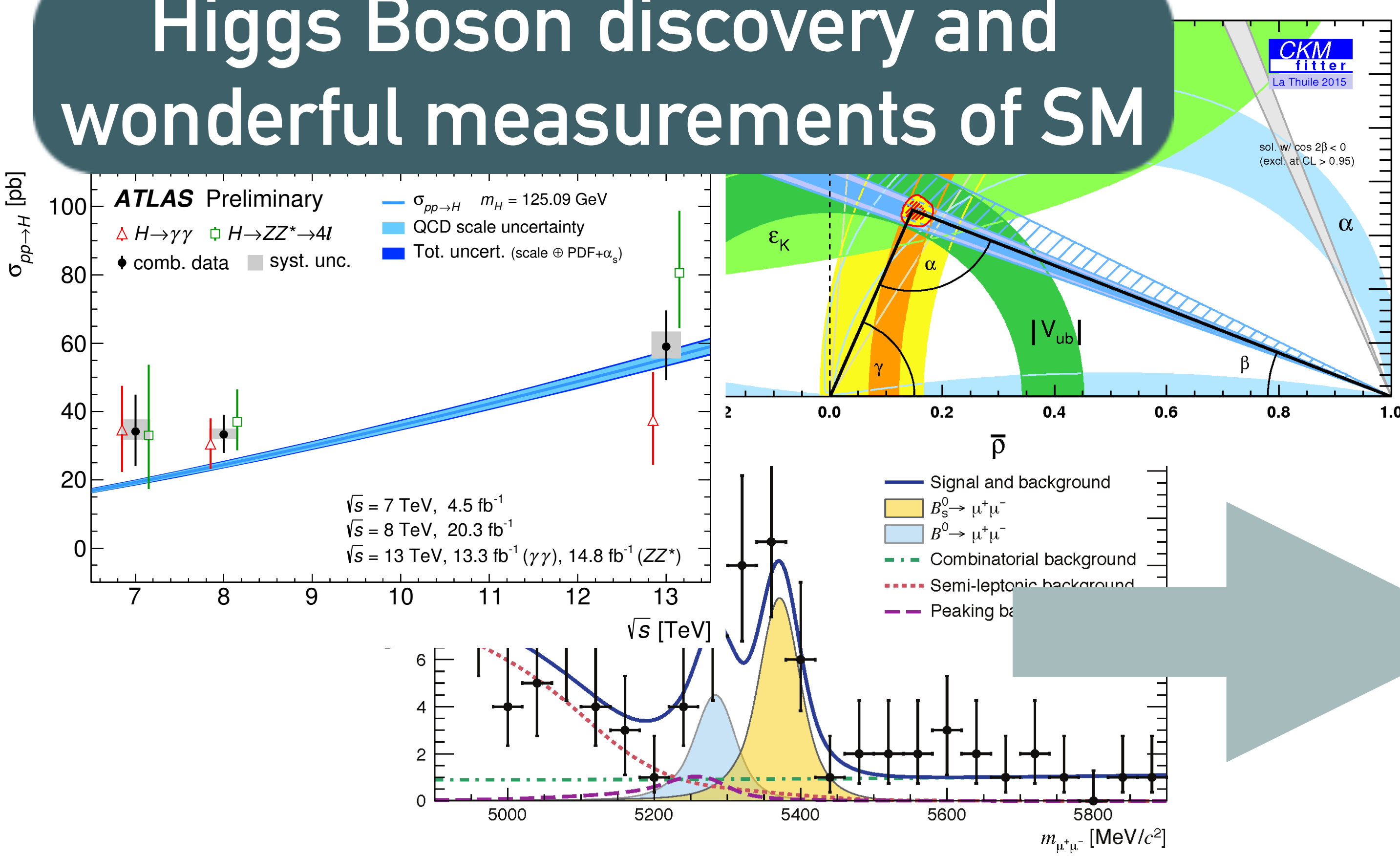
LHC at 2.1×10^{34} /cm²s at $\sqrt{s}$=13 TeV



Collected ~140 fb^{-1}



Higgs Boson discovery and wonderful measurements of SM

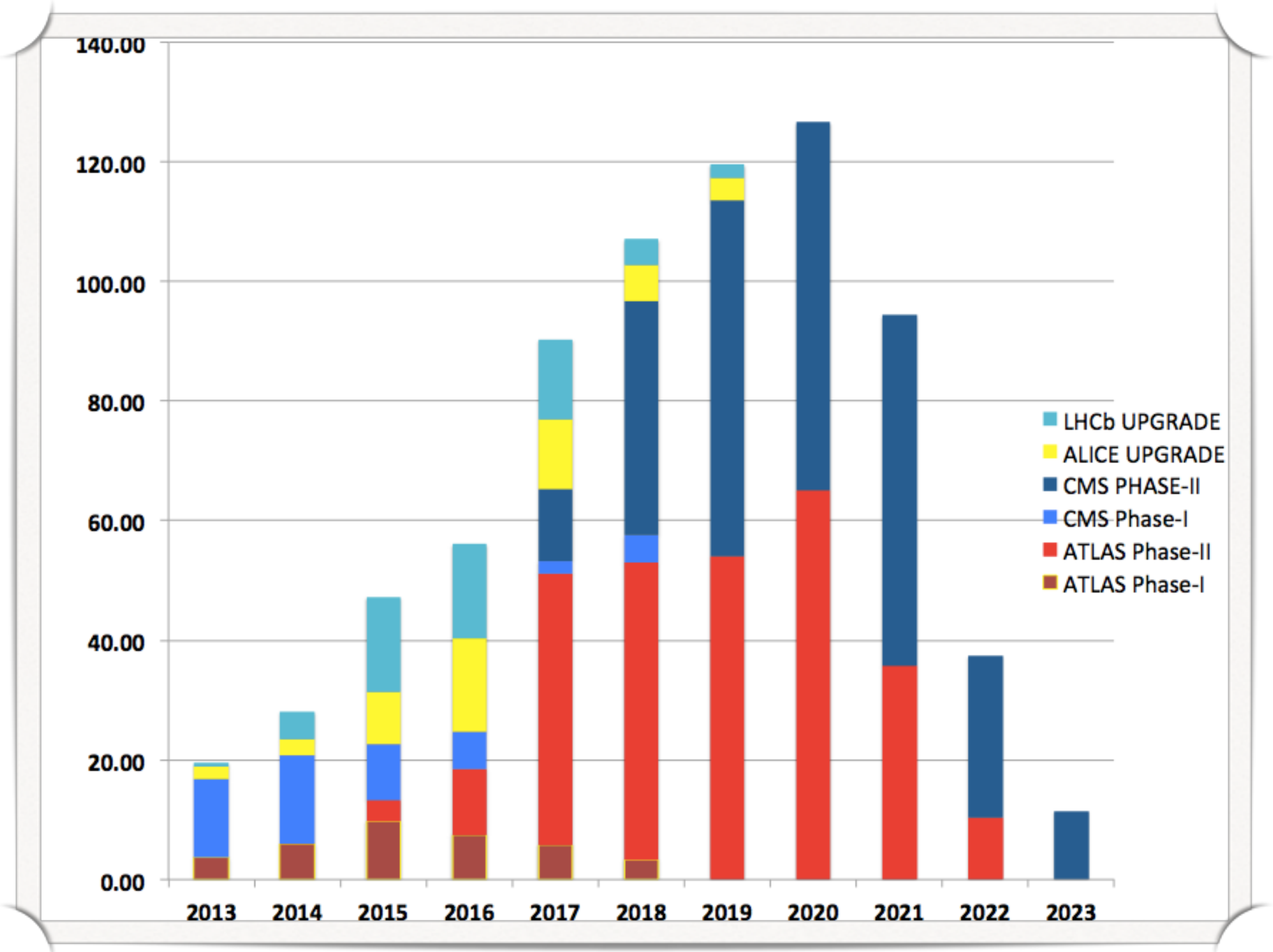


Standard Model is completed!
We have no evidence of New Physics!

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

UPGRADE PHILOSOPHY

Requires right balance between revolutionary approaches and technology evolution, based on physics potential and cost-effectiveness



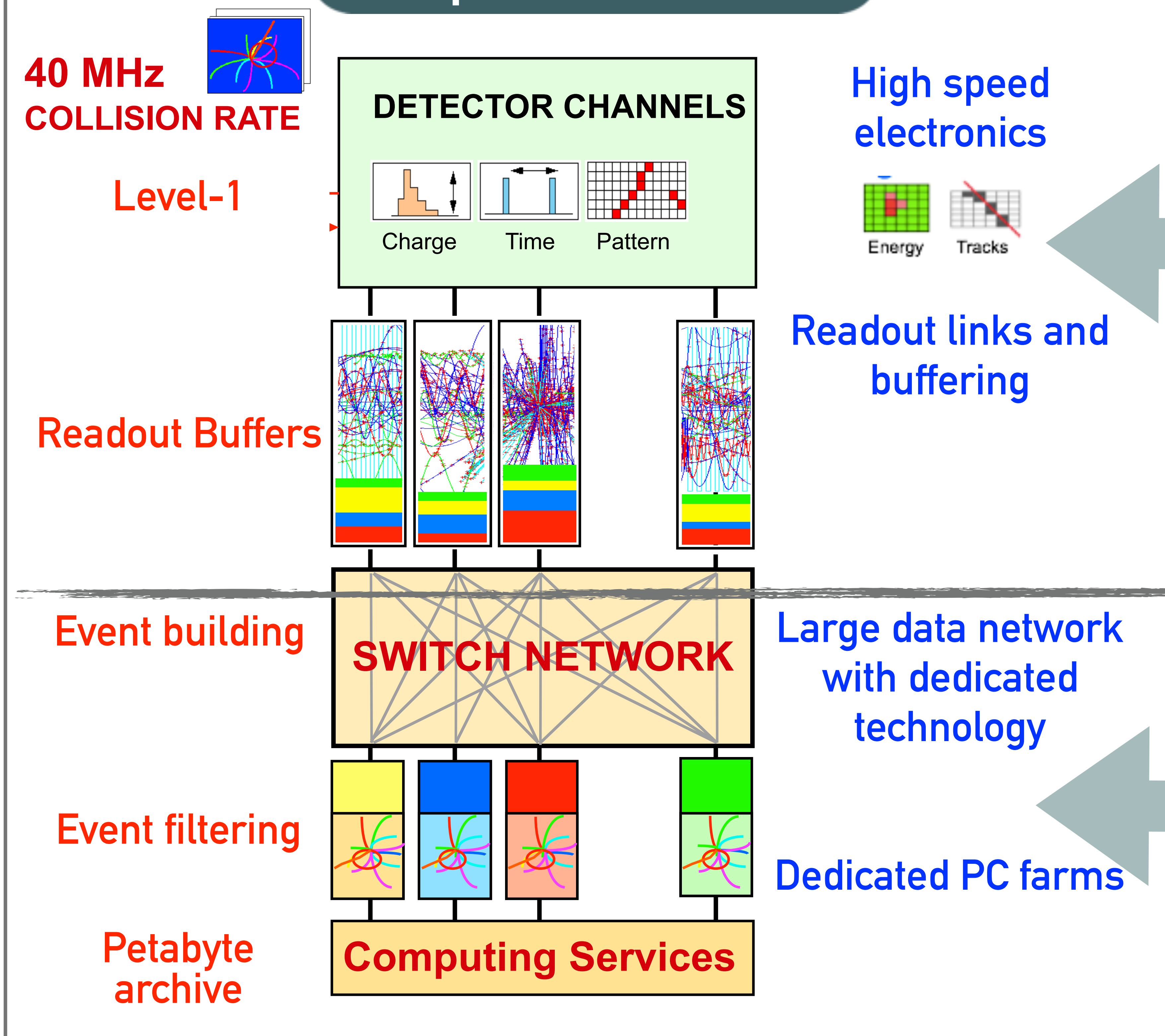
TRIGGERING AND TAKING DATA AT LHC

TDAQ for large discovery experiments

MANY PLAYERS, COMPLEX TDAQ ARCHITECTURES

Buffering and parallelism

Maximum 1-2% deadtime



Level-1 triggers

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

Higher level triggers

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

THE SIZE OF THE TRIGGER AND DAQ SYSTEMS

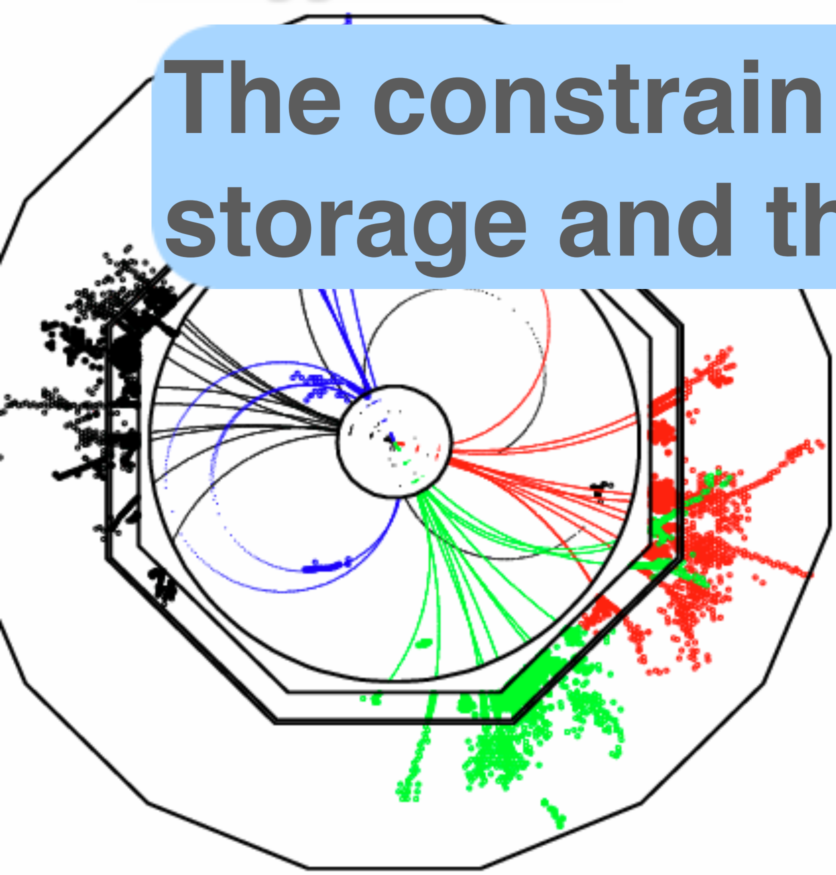
with time constraints → *computing resources*

Found interesting features… — **Identify the interesting process** — ***Trigger***

→ *data is collected…* — **Start data acquisition** — ***DAQ***

→ *data is recorded…* — **Record and Process data** — ***Computing***

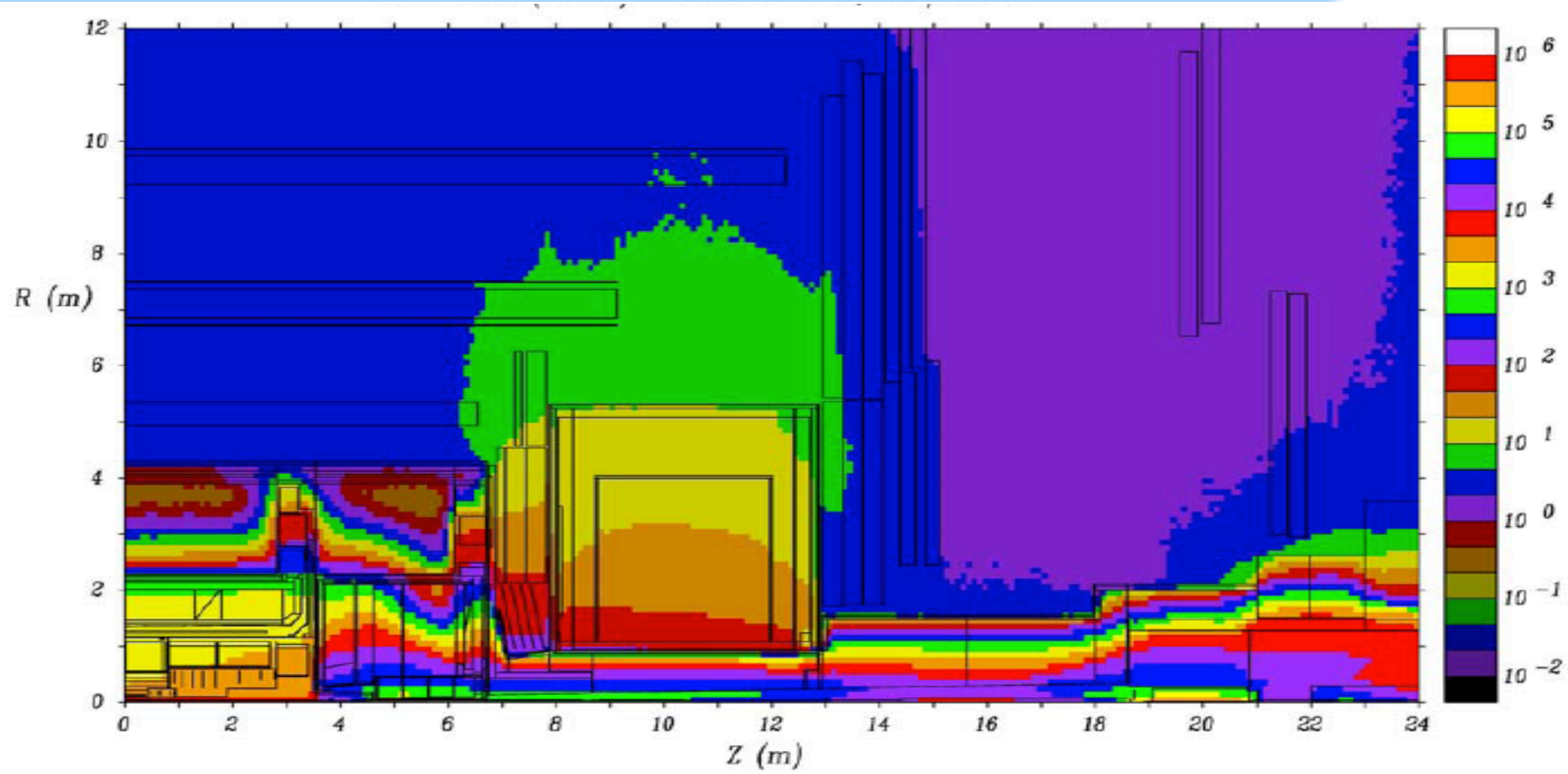
The constrain between trigger and DAQ rate is the storage and the offline computing capabilities



- ➡ **LHC experiments share the CERN budget for computing resources**
- ➡ **The <u>power of the trigger system</u> can be increased when easier selections can be adopted, and consequently reducing the data flow at the earliest stage (ATLAS/CMS)**
- ➡ **If the selectivity of the trigger is not enough, due to the large hadronic background, one bet on large data flow (ALICE/LHCb)**

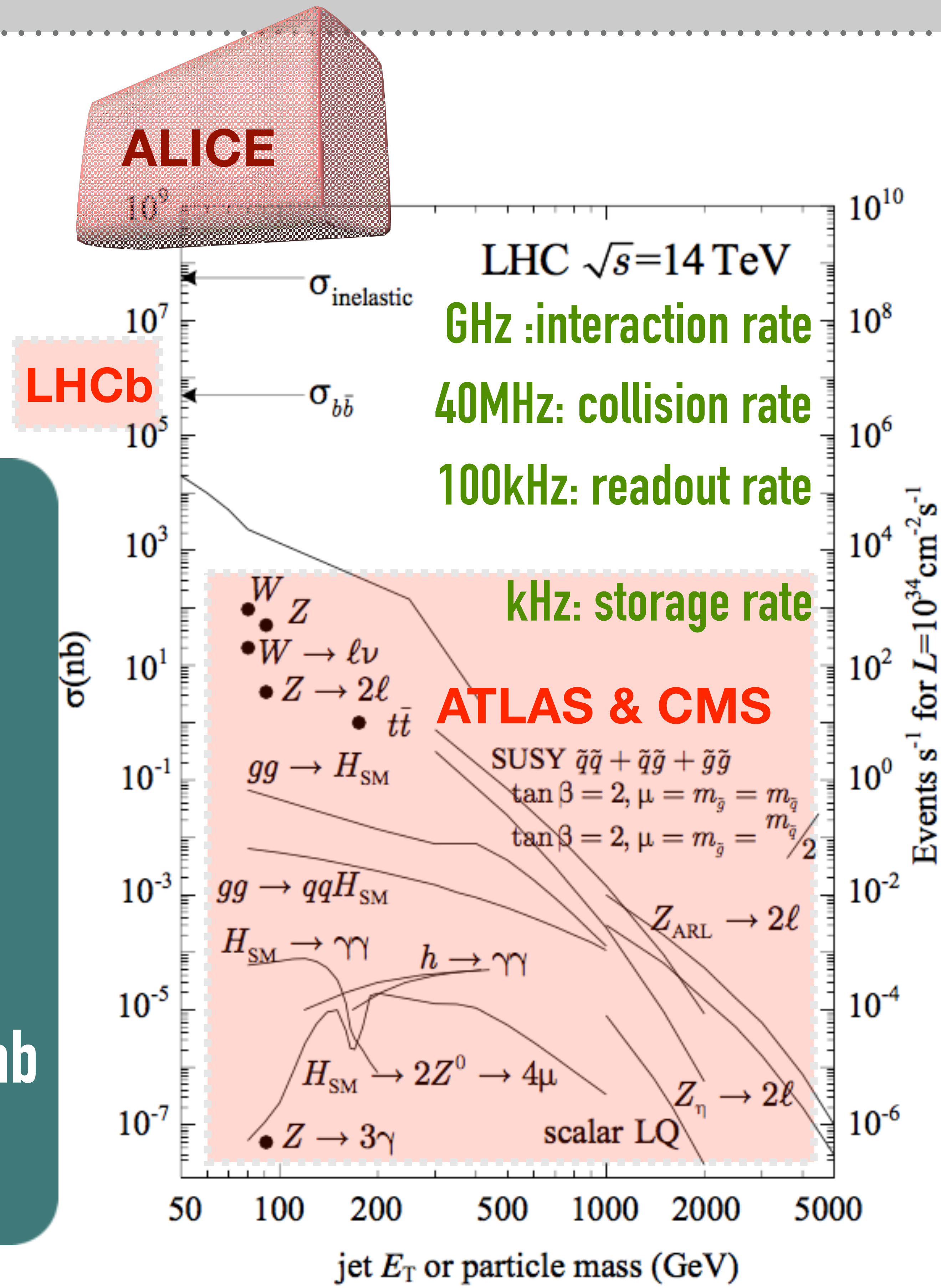
COMMON TDAQ REQUIREMENTS.....

- **Three major TDAQ challenges:**
  - **Search for rare physics**:
    - high rejection or large data collection
  - **Face High Luminosity**:
    - high frequency to resolve individual bunch crossing ➠ **fast electronics**
    - large detectors with fine granularity to avoid pile-up in the same detector element ➠ **high data volume**
  - **Be radiation resistant**



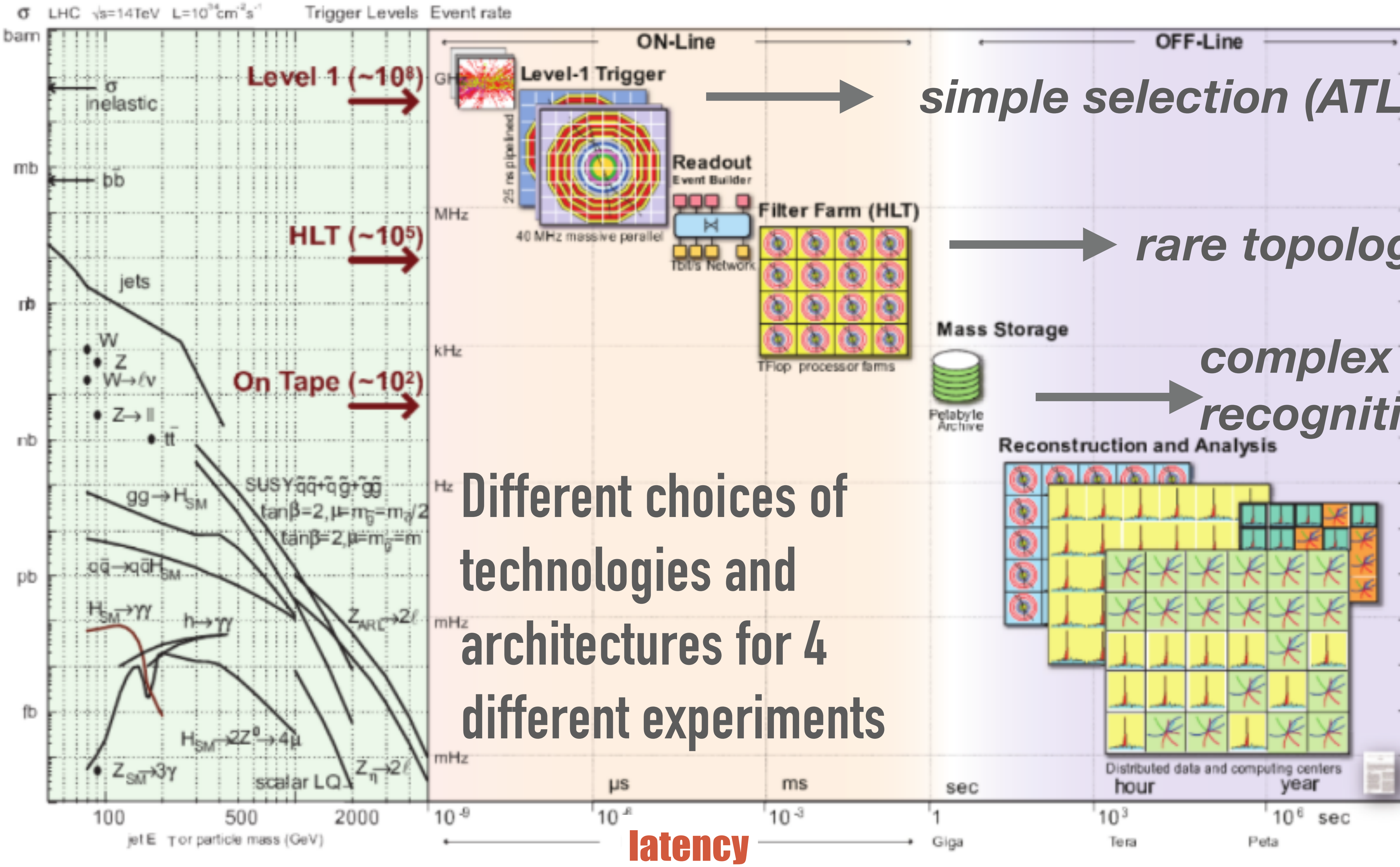
ATLAS cavern while collisions are ongoing

….SHAPED ON DIFFERENT PHYSICS REQUIREMENTS



- ✦ **ATLAS/CMS: p-p collisions @70 mb**
  - ✦ full Luminosity, high rejection
- ✦ **LHCb: p-p collisions**
  - ✦ reduced Luminosity for rare topologies
- ✦ **ALICE: heavy-ion collisions ~2000 mb**
  - ✦ high energy density

PROCESSING LHC DATA IN MULTIPLE STEPS



simple selection (ATLAS, CMS)

rare topology (LHCb)

complex pattern recognition (ALICE)

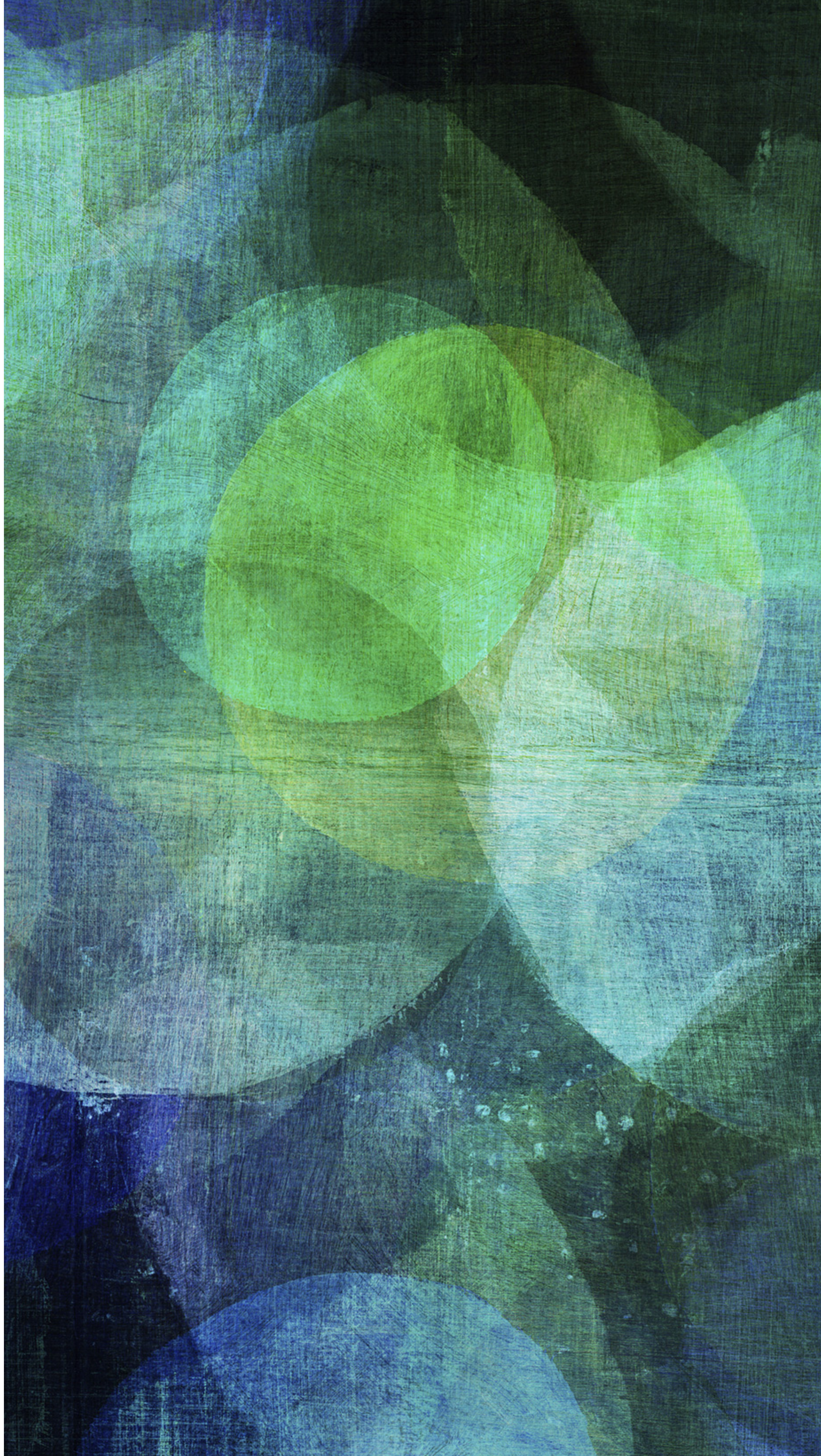
Different choices of technologies and architectures for 4 different experiments

Depending on:

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

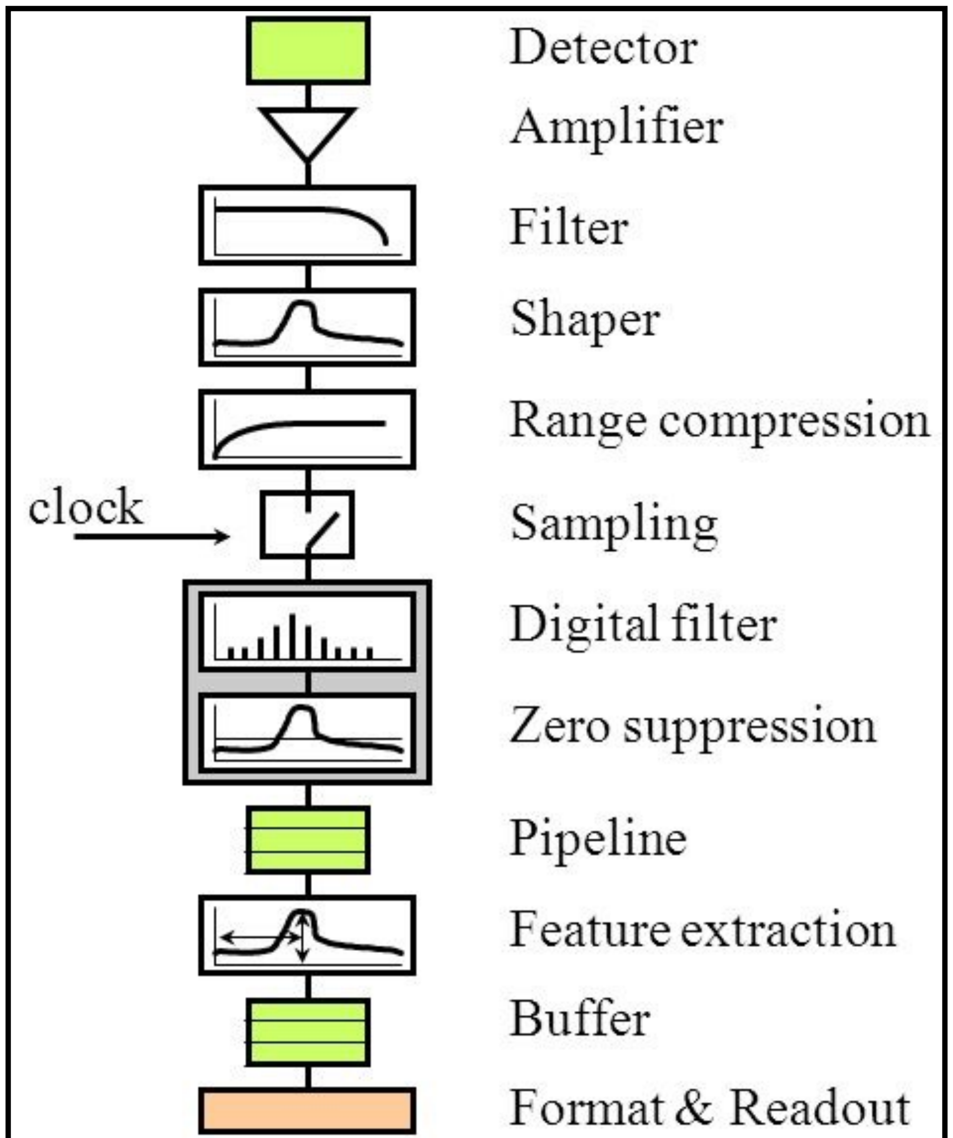
DESIGN PRINCIPLES

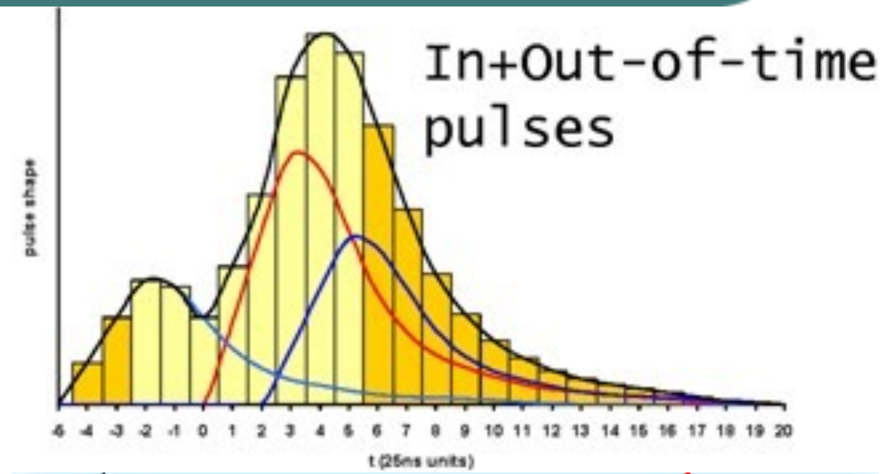
be fast, but robust!

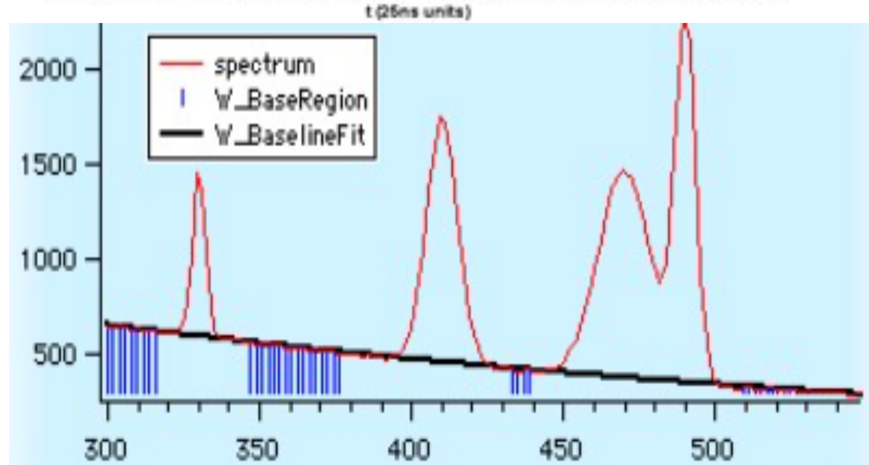


TRIGGER ELECTRONICS AT FRONT-END

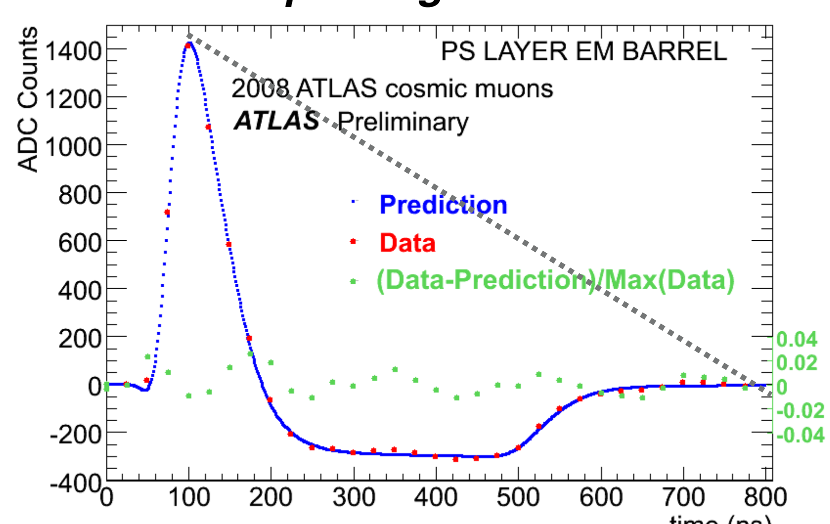
Tight design constraints for trigger/FE







ATLAS Liquid Argon calorimeter

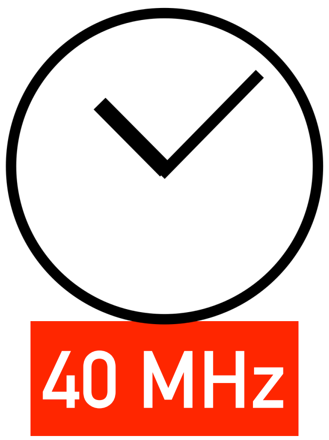


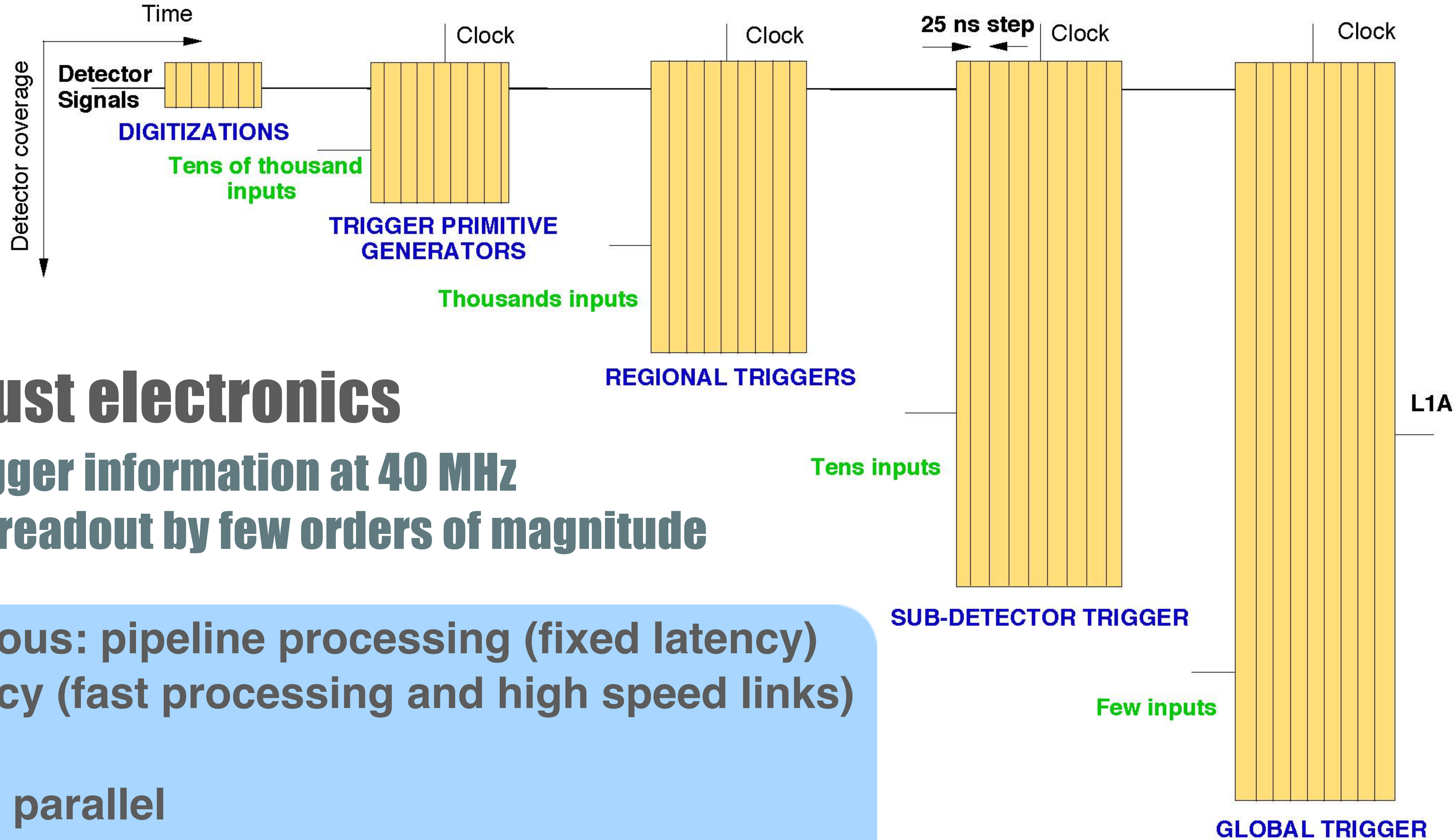
Avoid

- **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 a fast, low occupancy and digital detectors

FIRST LEVEL TRIGGER PRINCIPLES





Fast, robust electronics

Readout trigger information at 40 MHz
Reduce full readout by few orders of magnitude

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

Additional complication: synchronisation

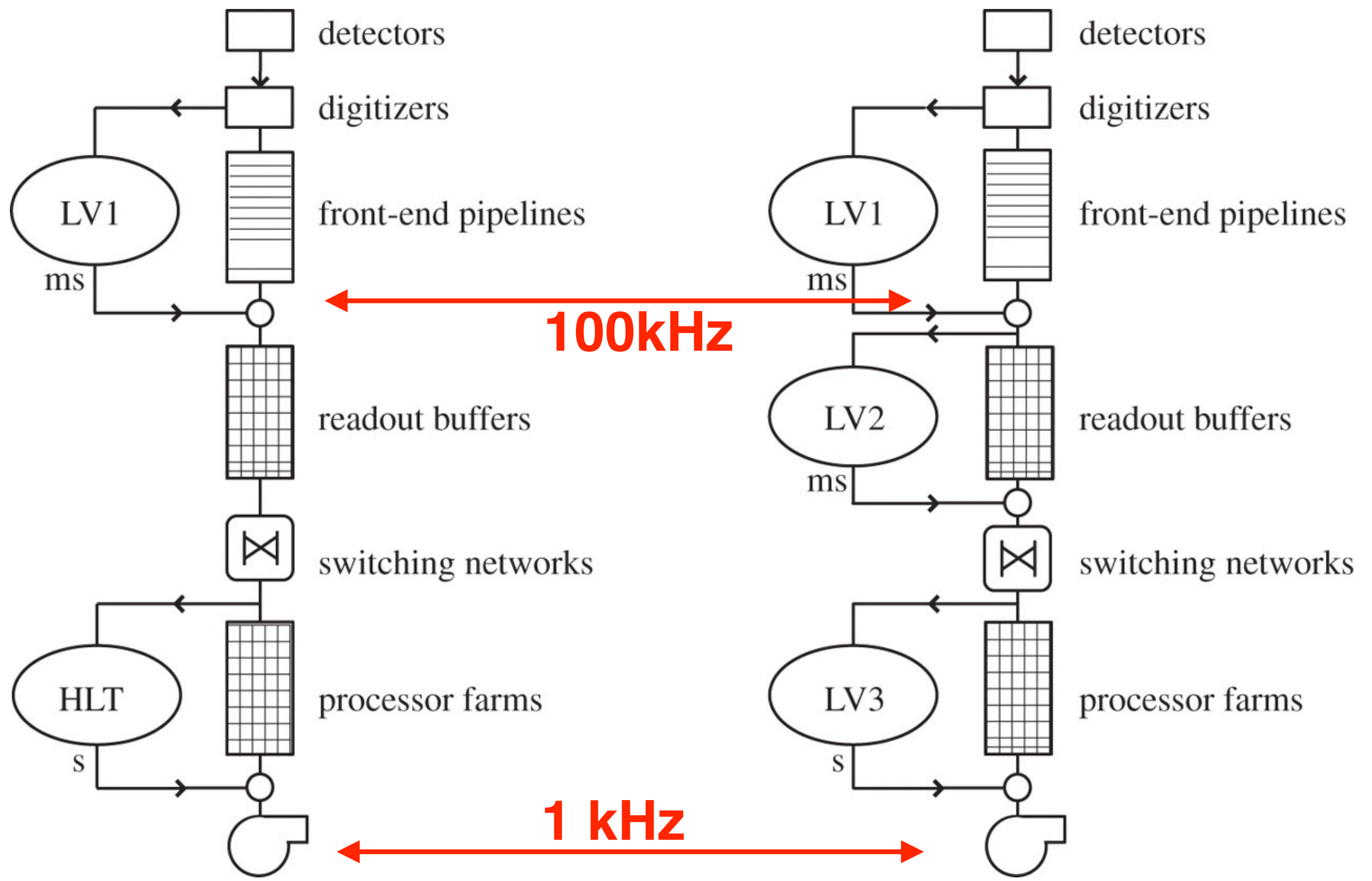
- ➤ BC counted and reset at each LHC turn
- ➤ large optical time distribution system

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

Latency dominated by cable/transmission delay

HLT/DAQ REQUIREMENTS FOR LHC EXPERIMENTS

Storage and processing resources allow order of ~1000 events/s



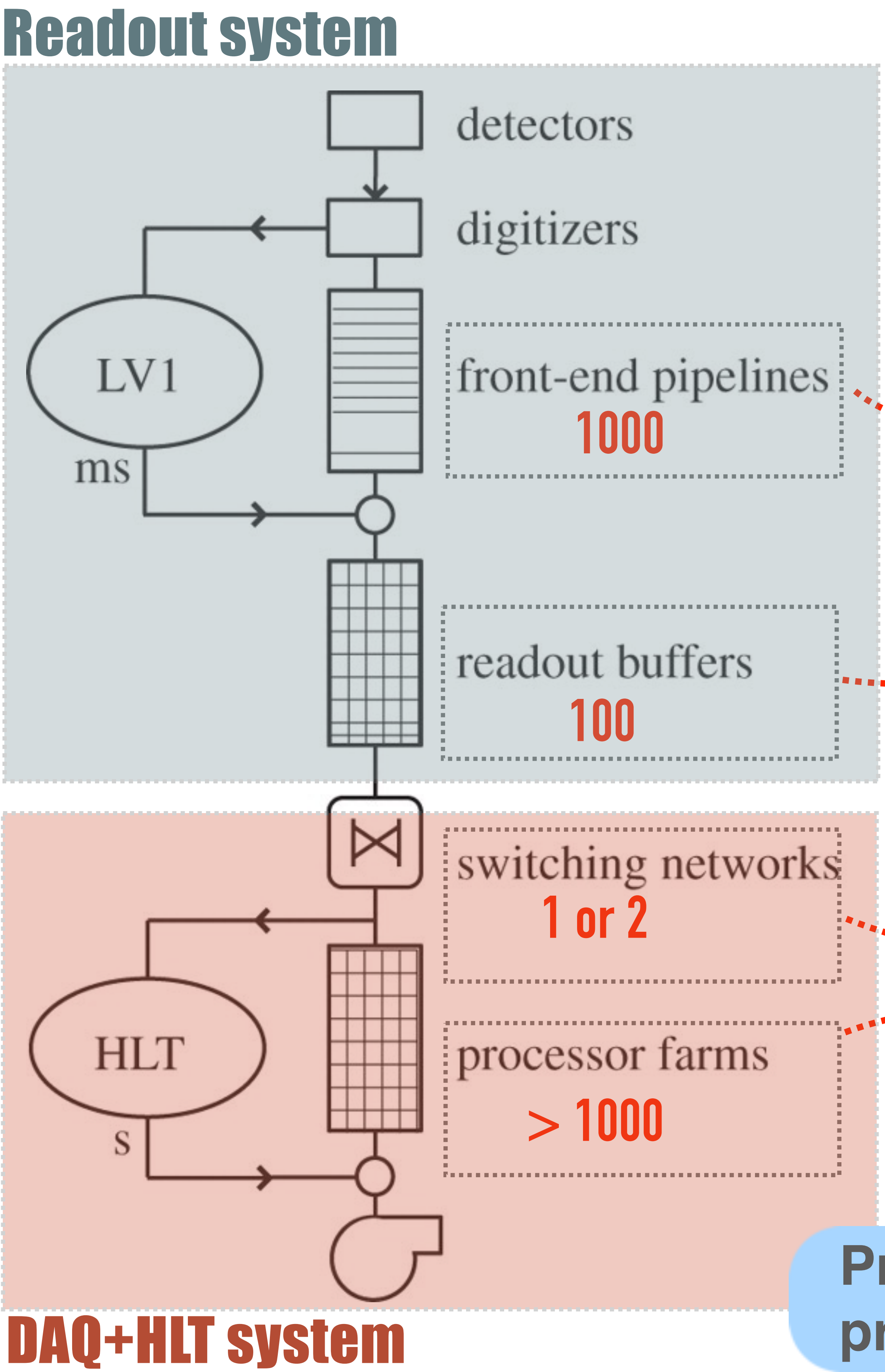
ATLAS/CMS Example

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

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

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

DAQ: HOW MANY COMPONENTS?



shape traffic and free the buffers

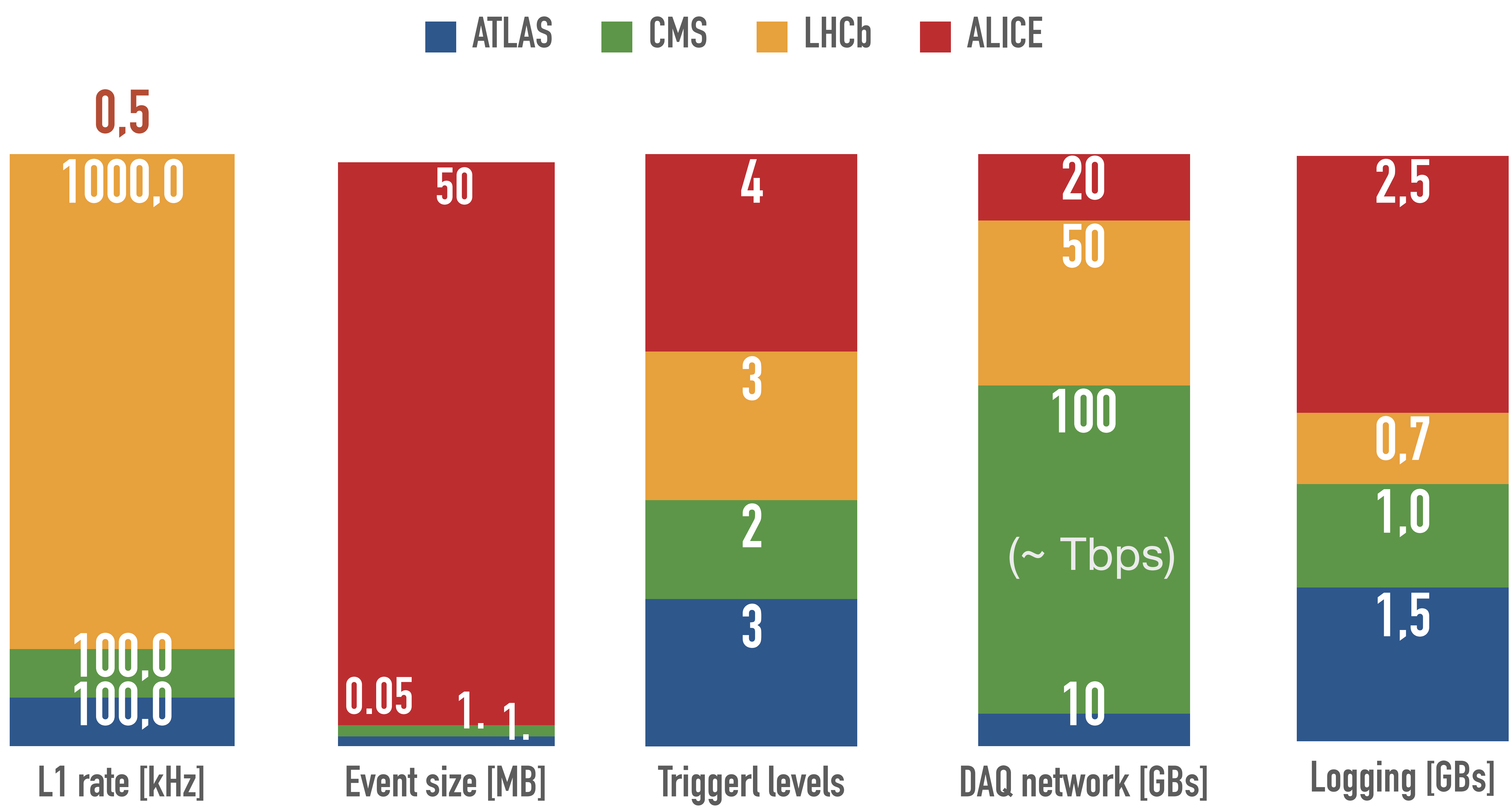
- **Readout links: aggregate data from FE**
  - optical/LVDS 200MB/s, mainly custom
  - can require flow control
- **Readout Units: collect data**
  - commercial or custom NIC
  - interfaced to PC or directly to another network
- **Event building network(s)**
  - scale-free (1Gb Eth towards 10Gb)
  - switching, destination assignment and traffic shaping
- **Processing cores for Event Building (EB) and HLT**

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

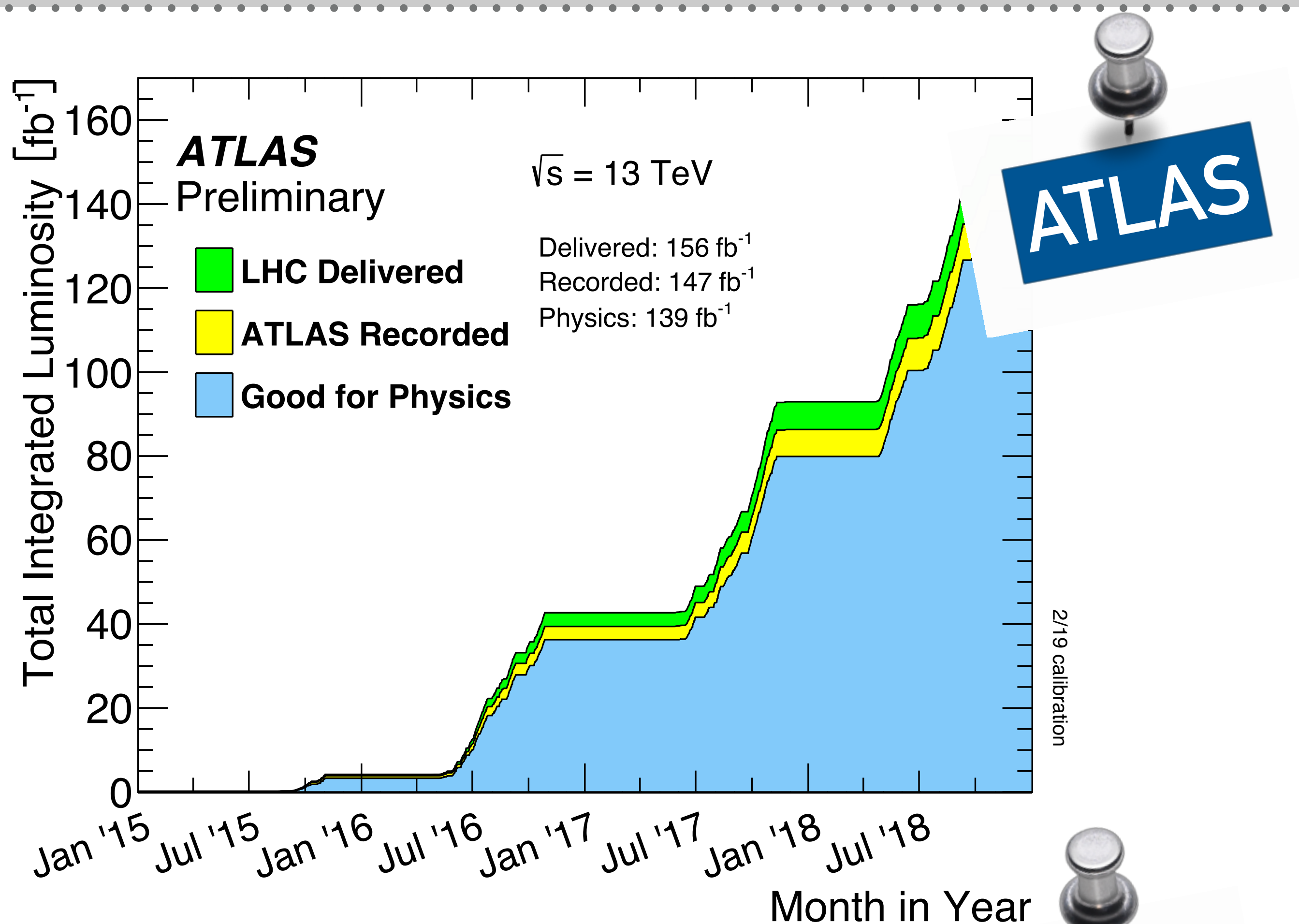
COMPARING LHC EXPERIMENTS DESIGN

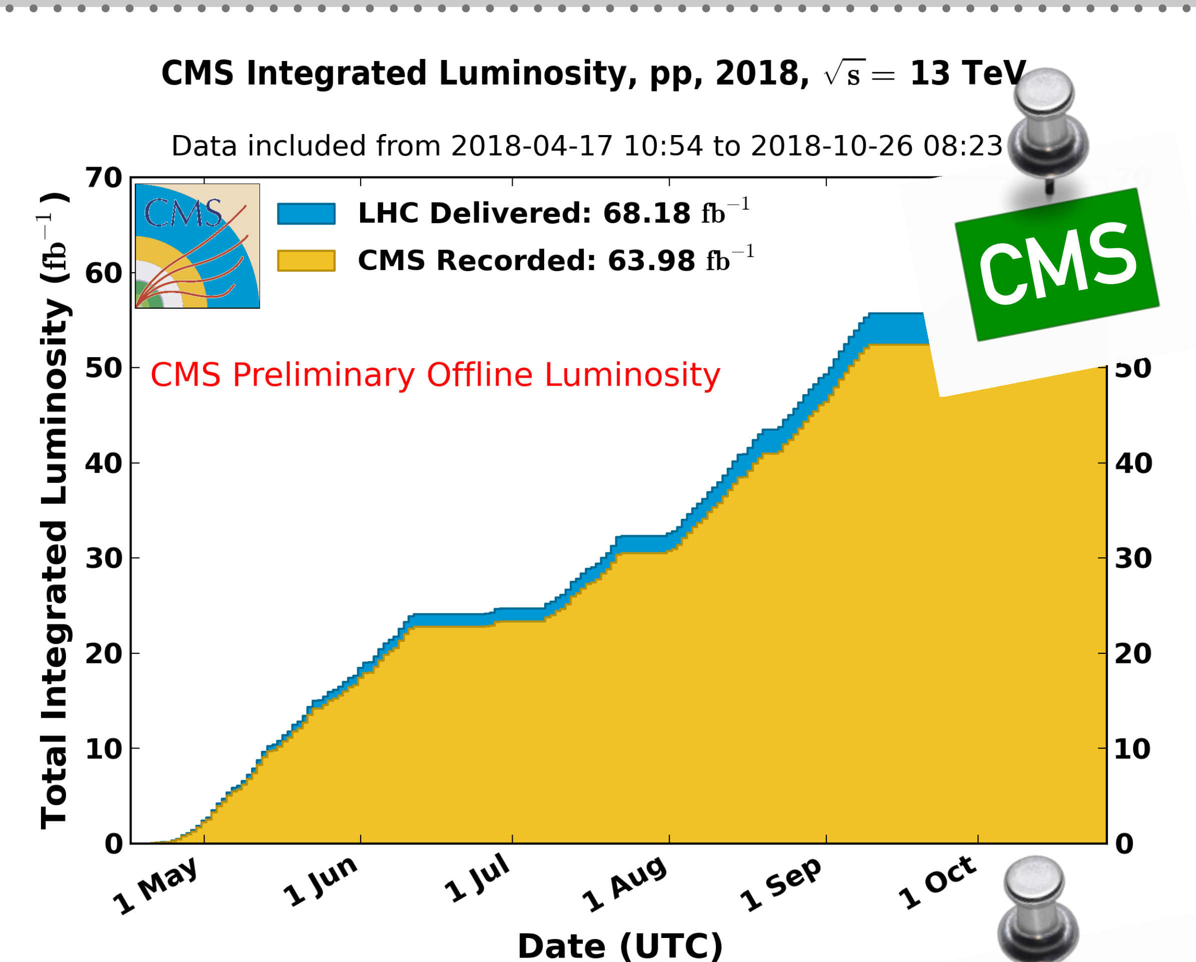
Storage and processing resources allow order of ~1000 events/s

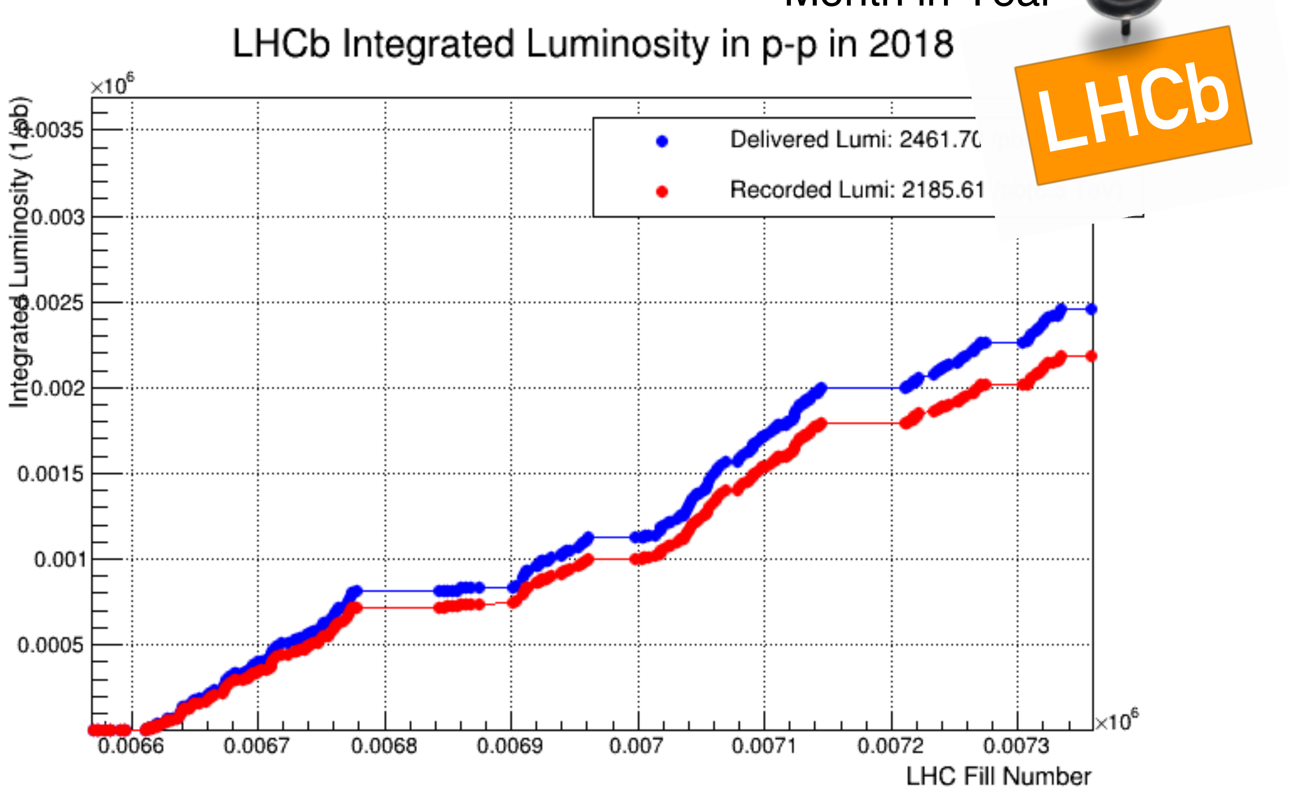
Designed to start in 2009

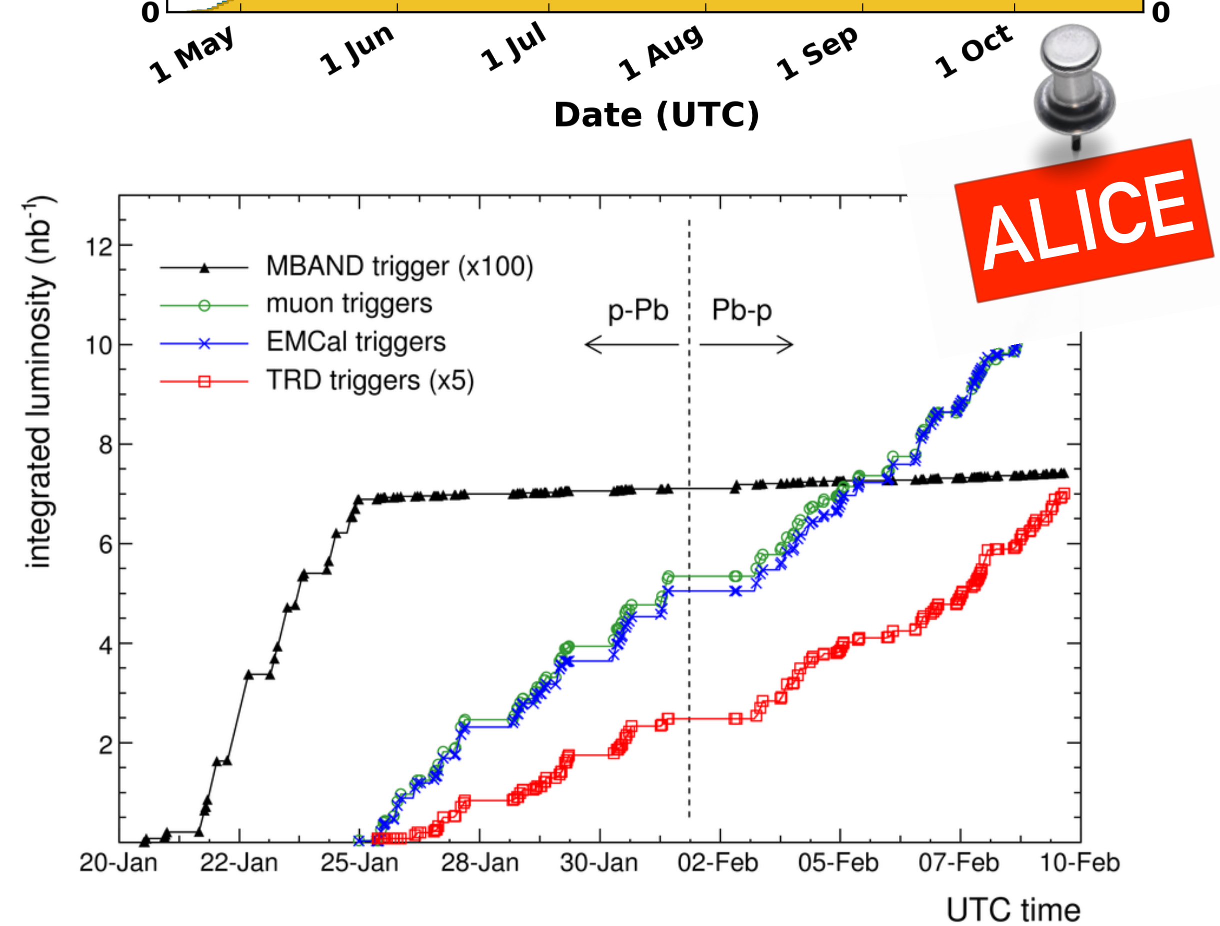


HOW DO THEY BEHAVE NOW? WELL!



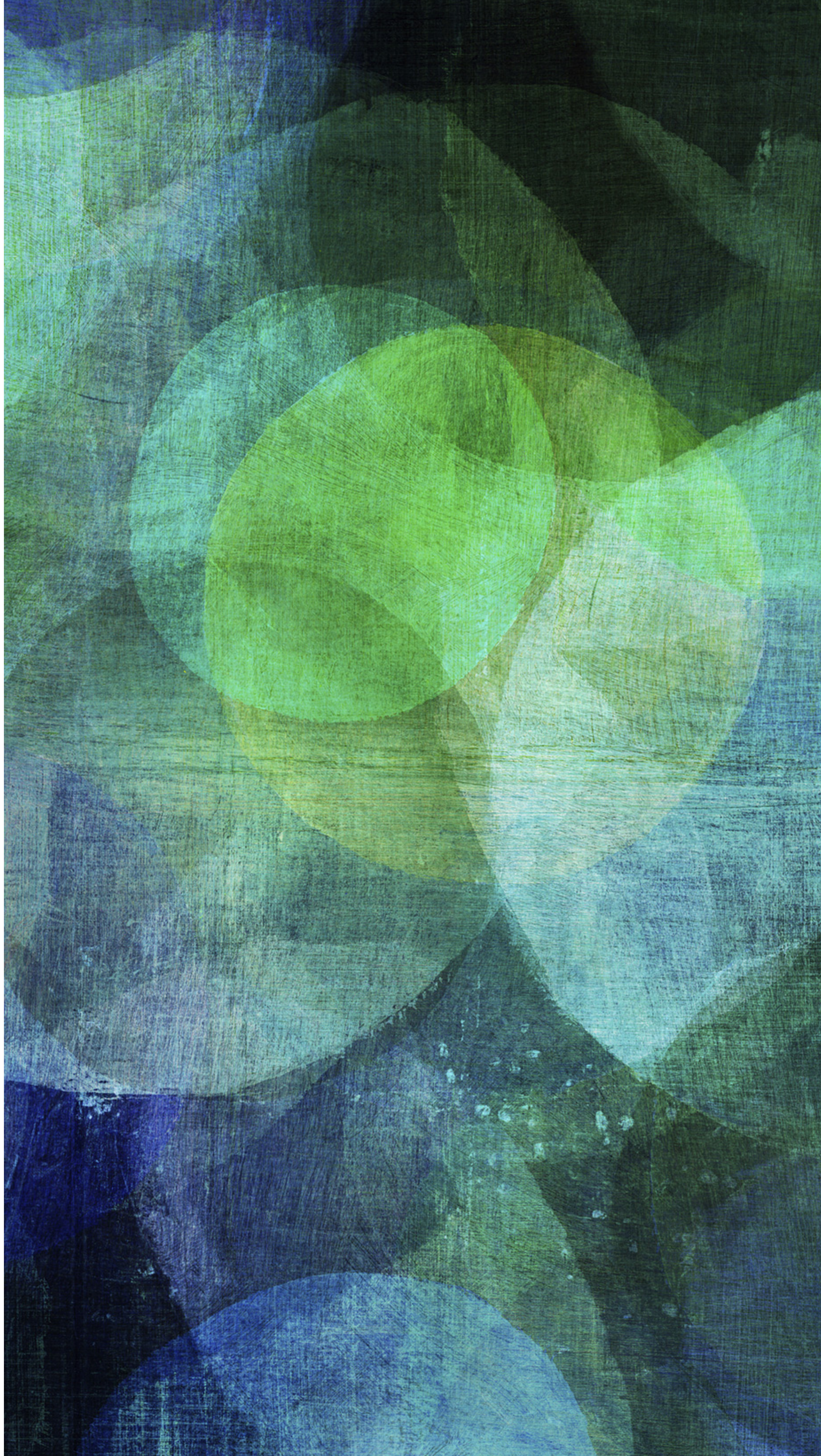






FUTURE TRENDS FOR HIGH-LUMINOSITY

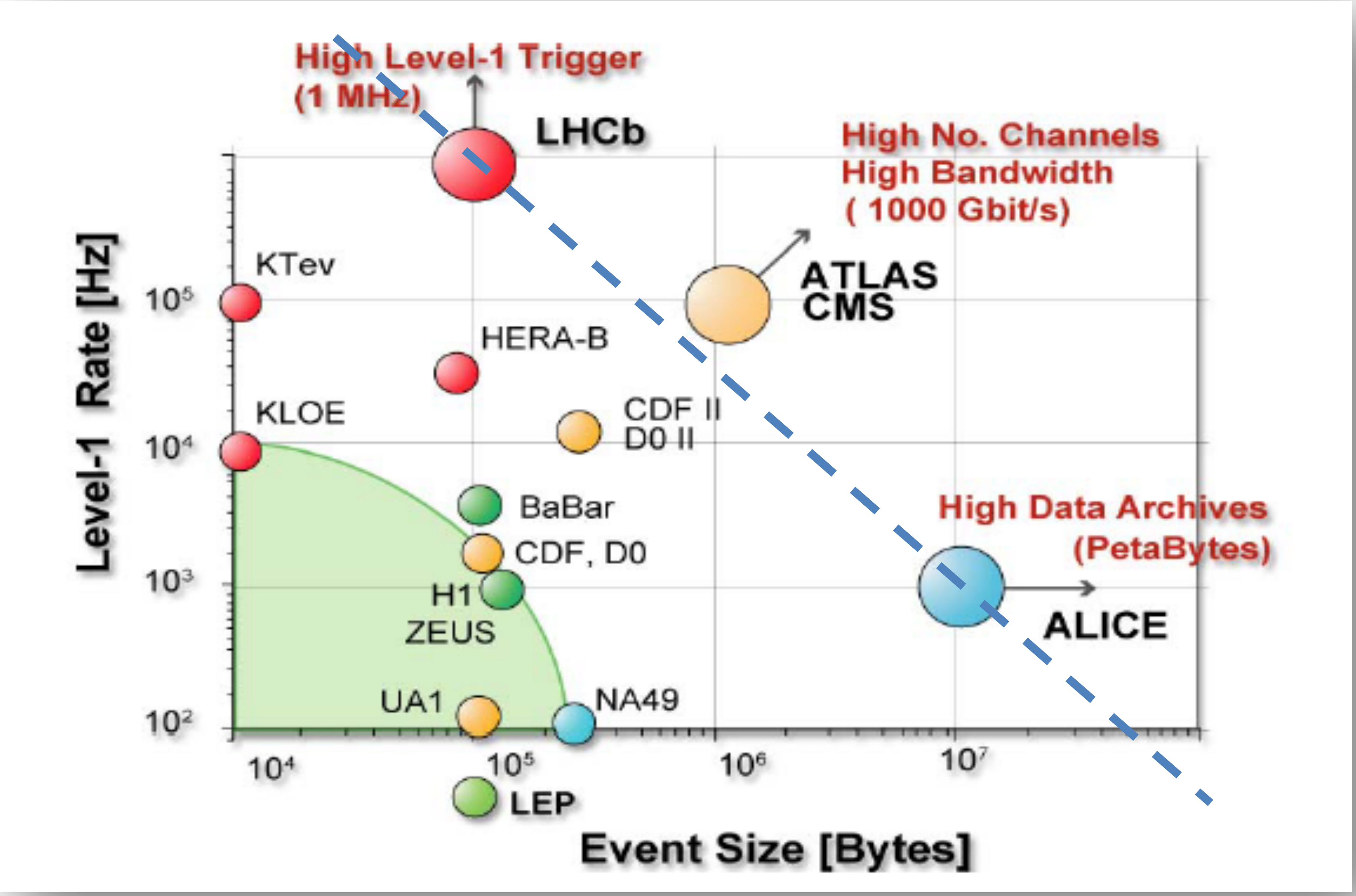
What about … tomorrow?



TRIGGER AND DATA ACQUISITION TRENDS

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

faster L1 electronics



more channels, more complex events

ATLAS/CMS

Data to Process:

100kHz * 1MB = 1Tb/s

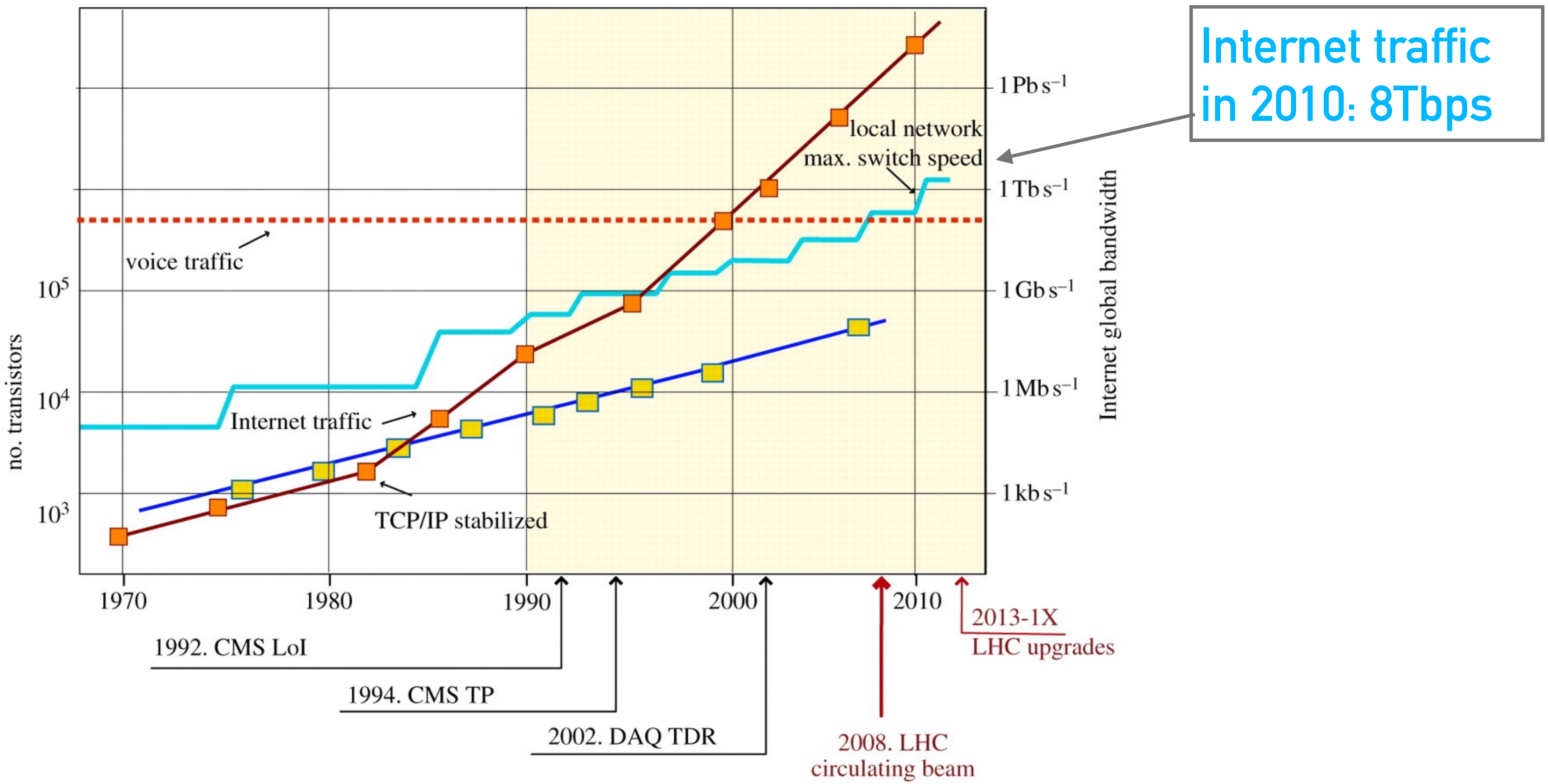
Data to Store:

~ 1 PB / year /experiment

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

WITH SOME EXTRAPOLATIONS FROM THE PAST

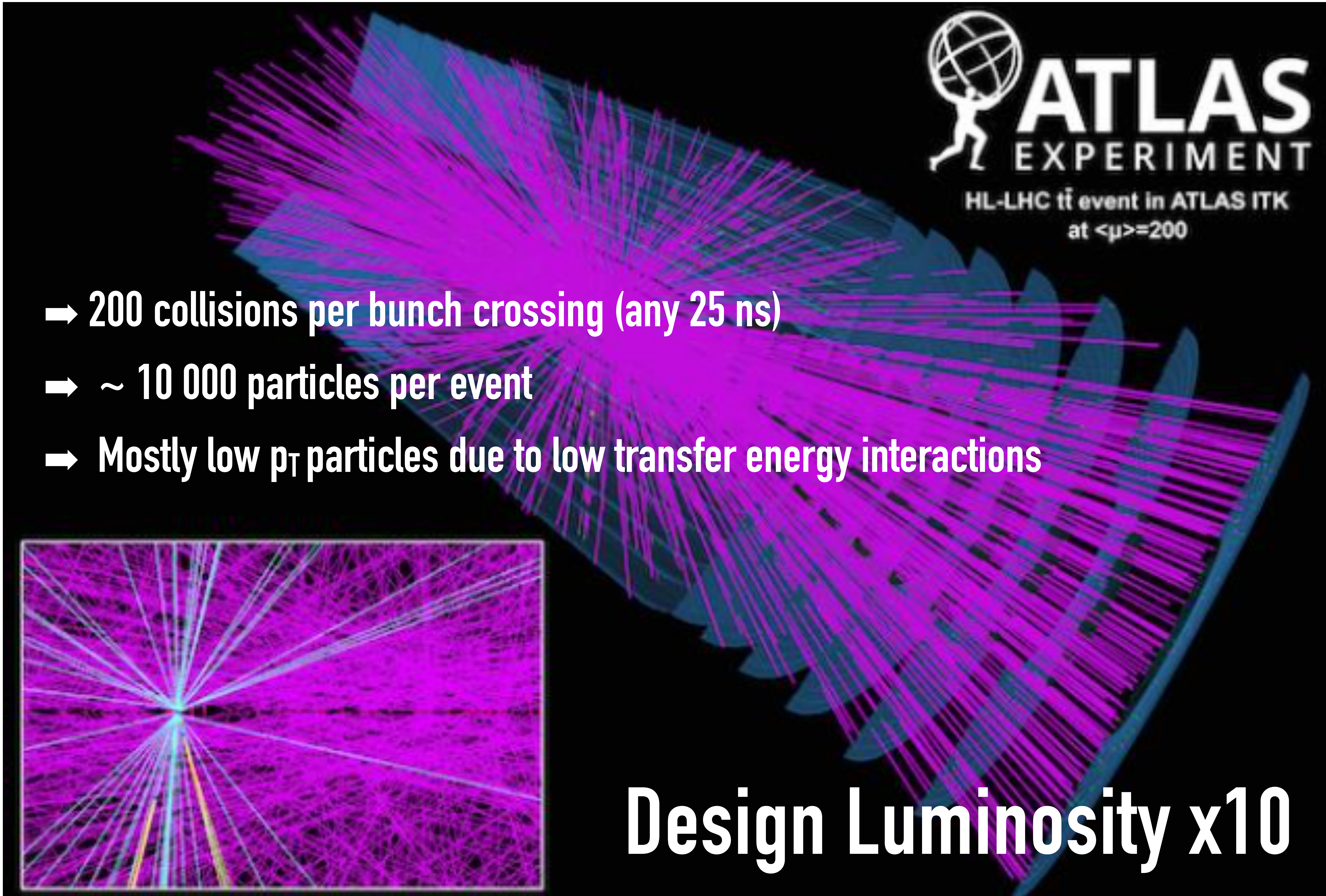
Design in the late 90s, constrained by available technology and budget



➡ Technology (processor speed/memory) grows exponentially

➡ Budget grows linear, and cannot fluctuate too much

ONE EVENT AT HIGH-LUMINOSITY ($L=7.5\times10^{34}$ /CM²/S)

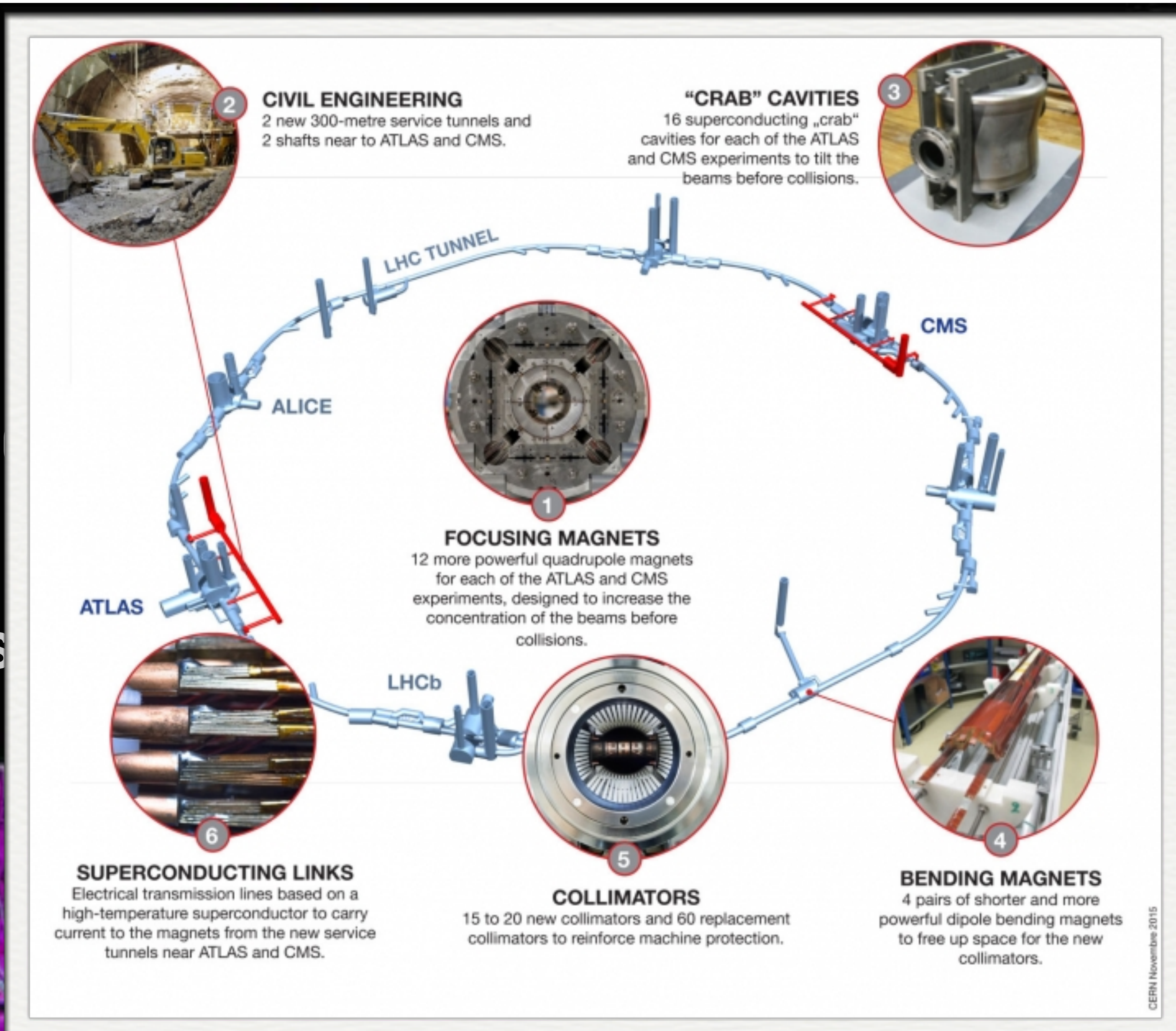


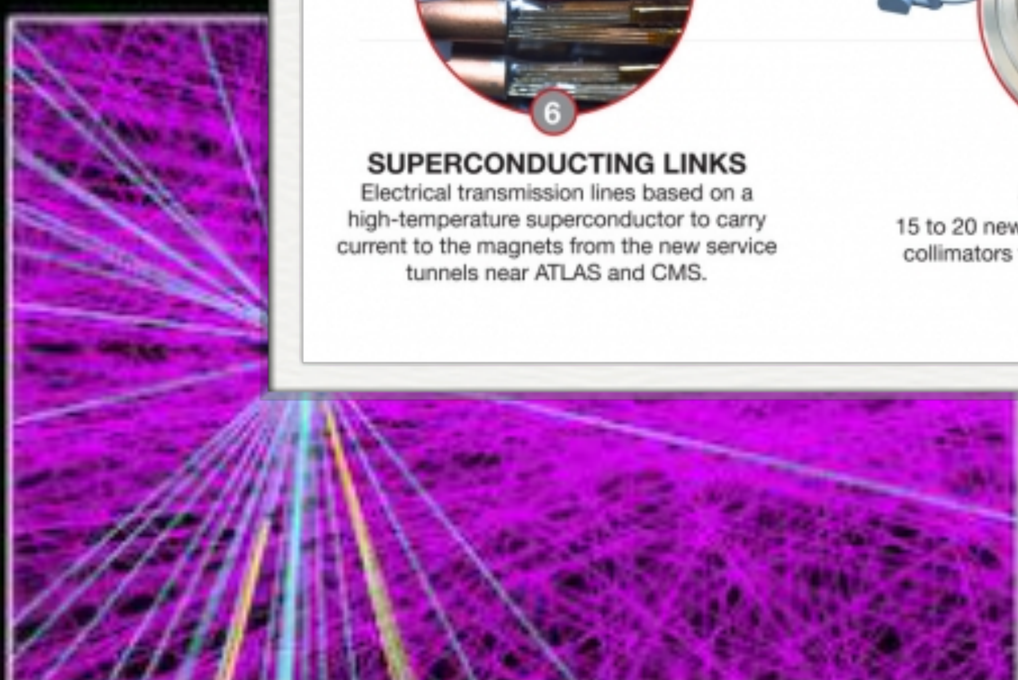
- ➡ 200 collisions per bunch crossing (any 25 ns)
- ➡ ~ 10 000 particles per event
- ➡ Mostly low p_T particles due to low transfer energy interactions

Design Luminosity x10

ONE EVENT AT HIGH-LUMINOSITY (L=7.5X10^{34} /CM^{2}/S)

- 200
- ~ 1
- Mos ... ons





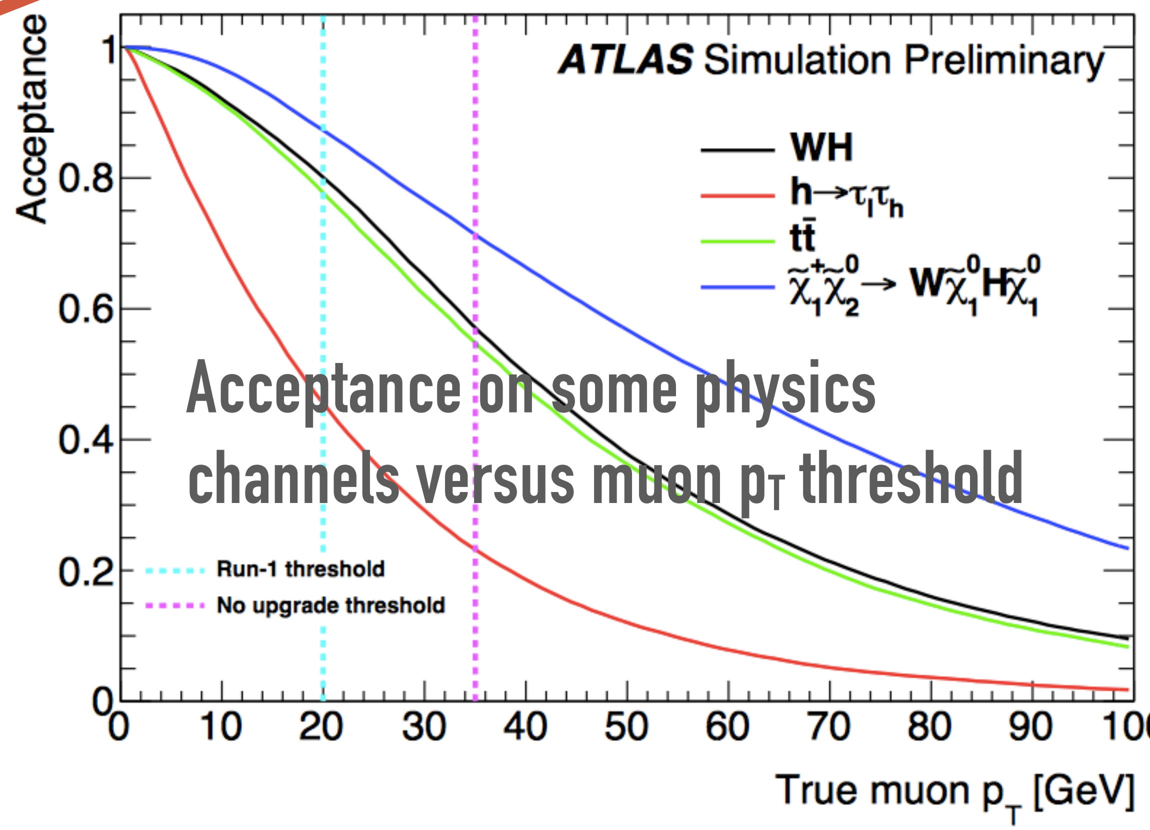
Design Luminosity x10

TDAQ CHALLENGE AT HL-LHC

New readout/DAQ architecture

Higher x10 Luminosity means…

- ➡ **Higher pile-up (40 ➠ 200)**
  - ➡ Less rejection (worse pattern recognition and resolution)
  - ➡ Larger Event size (x5)
- ➡ **Higher rates**
  - ➡ Readout rate @L1: 0.1 ➠ 1 MHz
  - ➡ DAQ throughput: 1 ➠ 50 Tbps

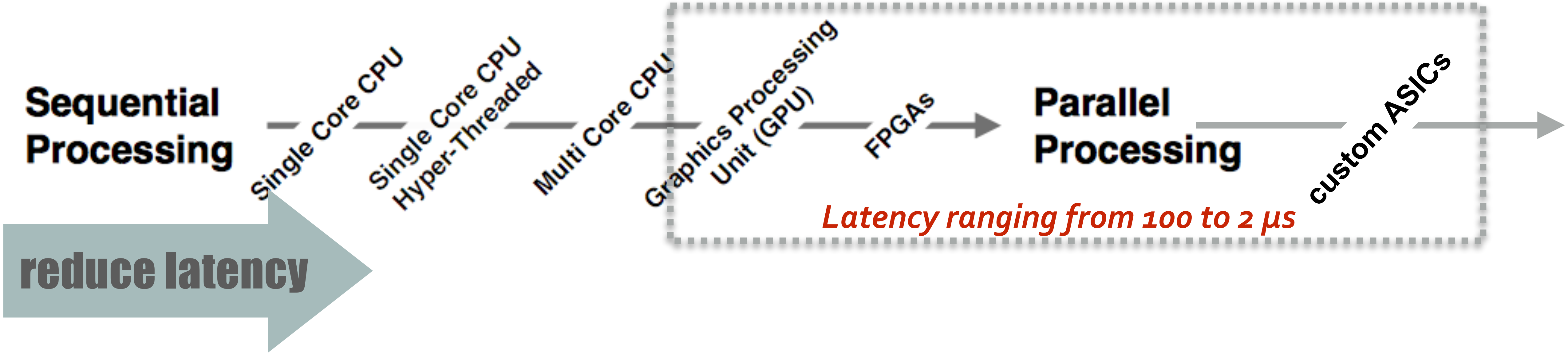


But cannot…

- ➡ **Apply too high thresholds**
  - ➡ Need to maintain physics acceptance
- ➡ **Scale dataflow with Luminosity**
  - ➡ **H/W**: short latency ➠ more parallelism ➠ more links ➠ more material and cost
  - ➡ **S/W**: processing time not scaling linearly with L, event complexity is dominant

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

TRIGGER PROCESSORS: TECHNOLOGY TRENDS



Nowadays

- ➡ **Push digital IC on a single chip (SoC)**
  - ➡ Higher complexity ⇒ higher chip density ⇒ smaller size (transistors and memory): **32 nm ⇒ ⇒ ⇒ 10 nm**

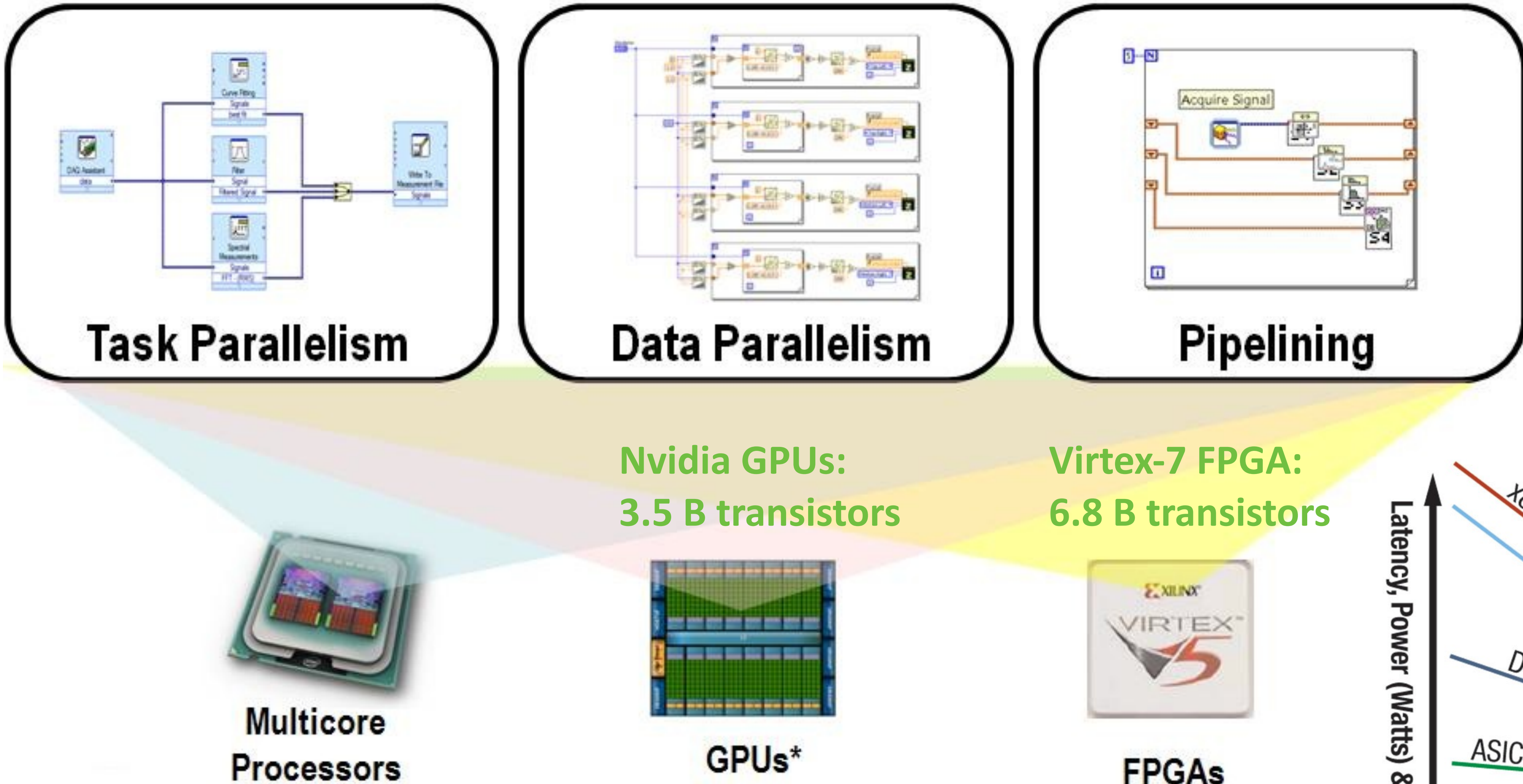
Tomorrow

- ➡ **Limited by the Power Wall for**
  - ➡ High frequency clocks (20MHz to 20 GHz and beyond)
  - ➡ Low noise
- ➡ Analog interference on digital electronics (noise, cross-talk, reflections)
- ➡ **Current technology could not be simply scaled**
  - ➡ Significant improvements/breakthroughs: aggressive R&D

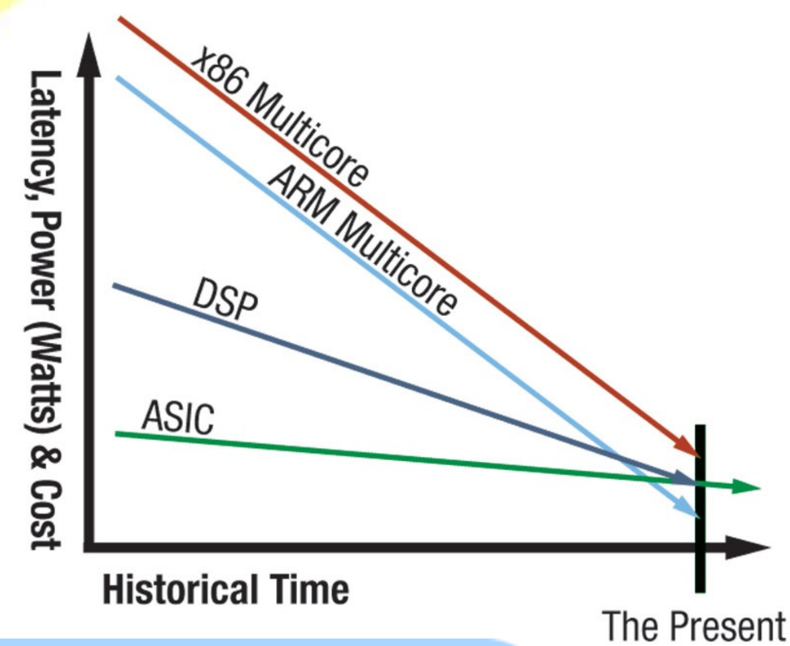
The golden time for "easy" digital electronics is over

High-Speed Digital Design: A Handbook of Black Magic

TRENDS: COMBINED TECHNOLOGY

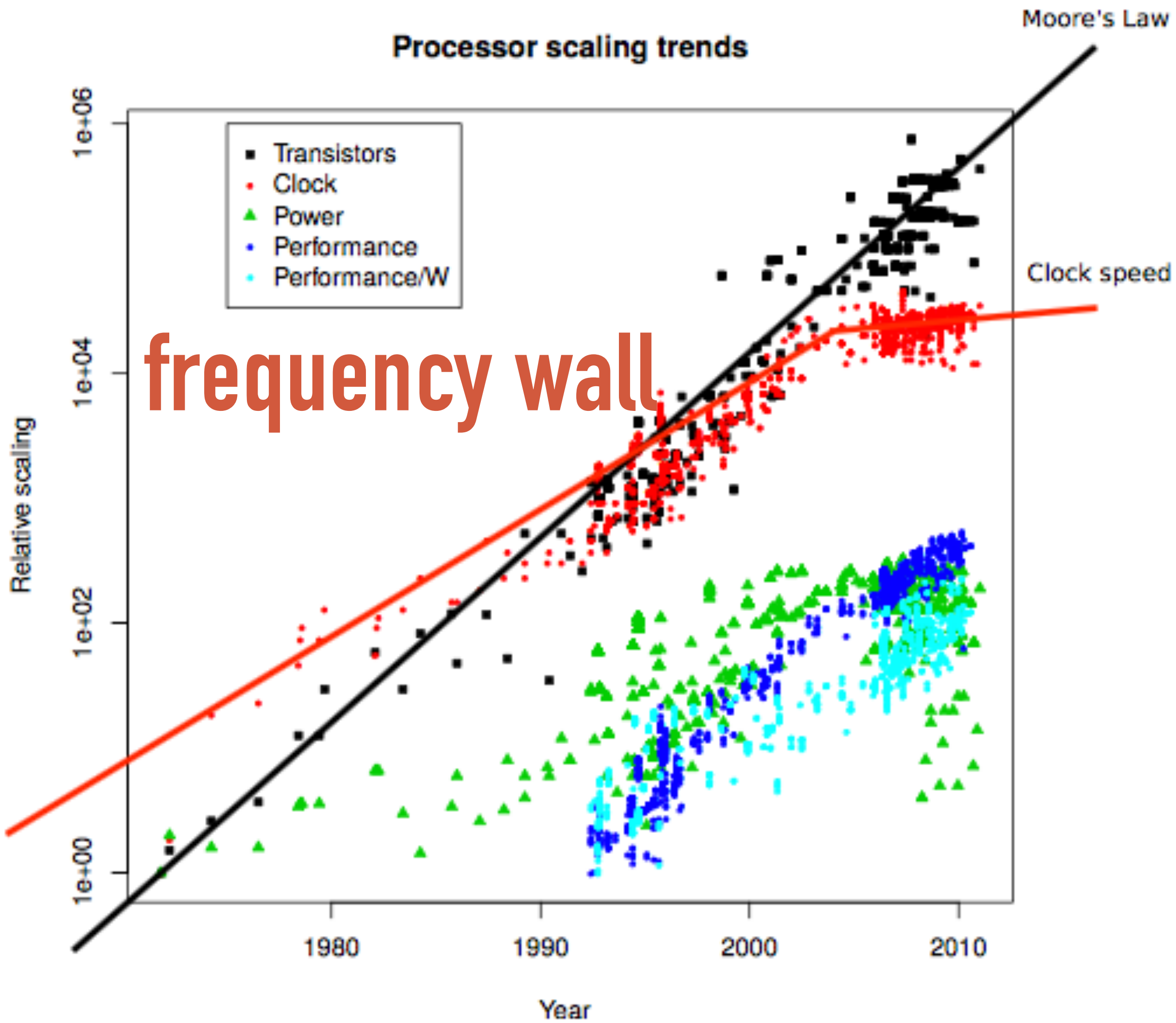


(*) Access to the nVIDIA® GPUs through the CUDA and CUBLAS toolkit/library using the NI LabVIEW GPU Computing framework.



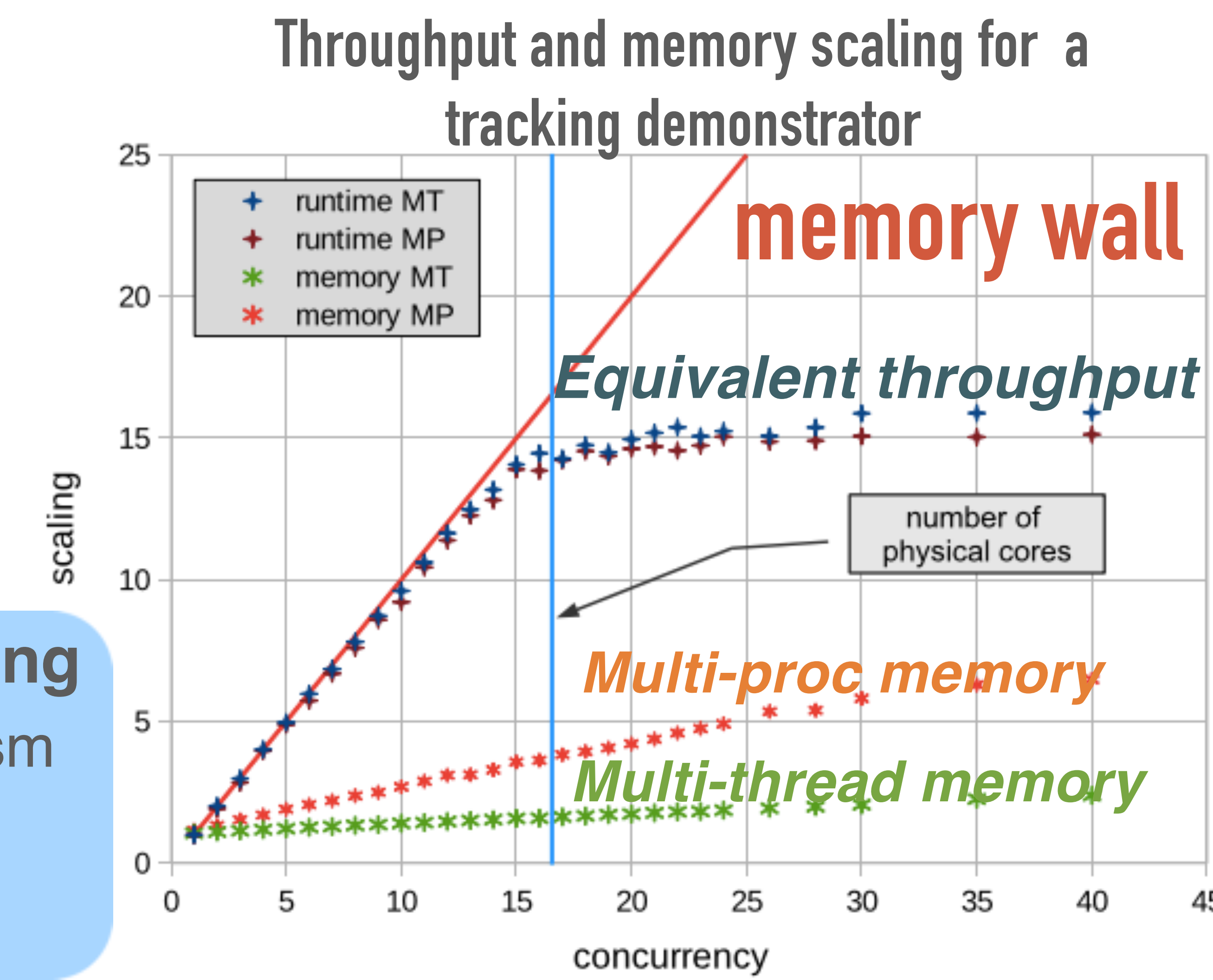
The right choice can be combining the best of both worlds by analysing which strengths of FPGA, GPU and CPU best fit the different demands of the application.

TRIGGER SOFTWARE EVOLUTION TO BREAK WALLS



Higher pile-up means more needs

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



- ➡ **Move towards multithreaded processing**
  - ➡ Multiple events in flight, sub-event parallelism
  - ➡ Exploiting CPU h/w, but more complicated (vectors, memory sharing…)

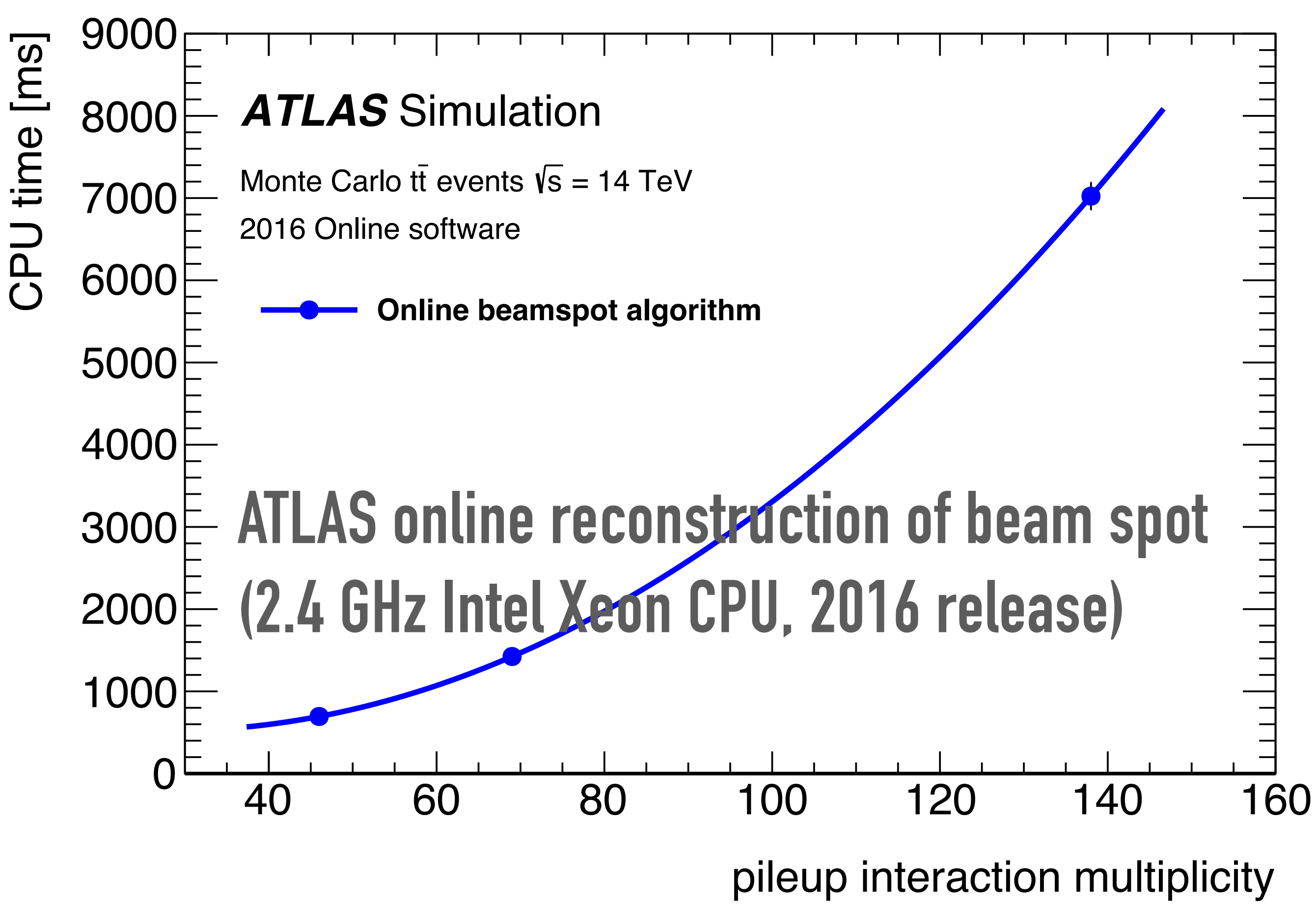
Evolution in programming paradigms, tools and libraries

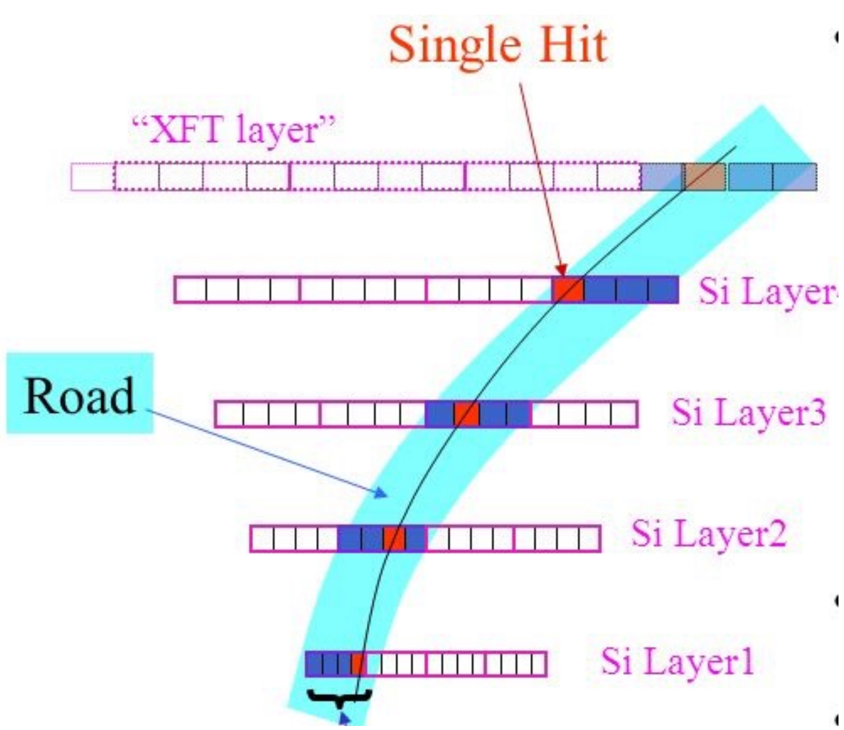
WITH SOFTWARE/HARDWARE HYBRID SOLUTIONS

- ➡ **Mainly driven by big software developments**
  - ➡ Hardware/software interplay (compilators)
  - ➡ Algorithms and parallelisation
- ➡ **Tracking dominates CPU time**
  - ➡ Hardware pattern recognition
  - ➡ Software: seeded precision tracking
    - ➡ Use of accelerators, e.g. GPU

Tracking challenges

- ➤ Readout ~800M channels (in few microseconds)
- ➤ Solve enormous combinatorics due to high occupancy (10^4 hits/BC)





combinatorics scales like L^N

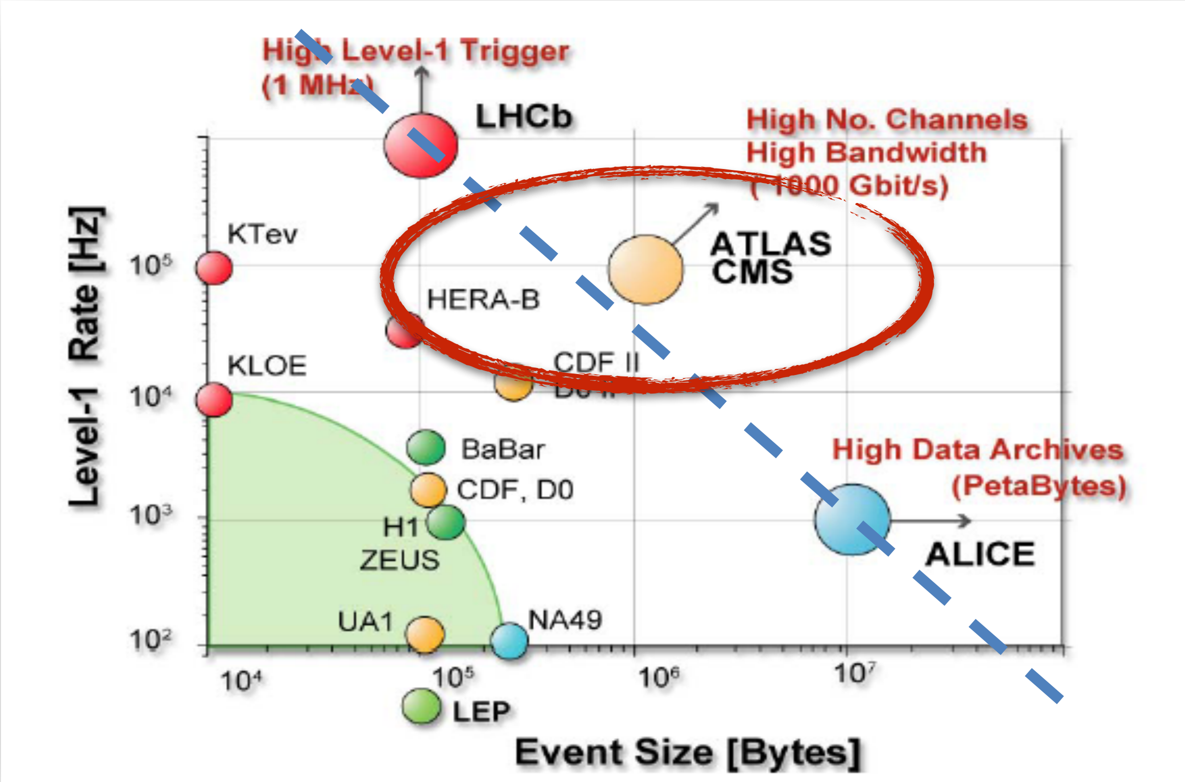
L=luminosity, N=number of layers

TRIGGER GOAL: INCREASE RESOLUTION FOR BETTER S/B

As early as possible (40MHz?)*

Approach	Solution	Implementations
High detector granularity	➡ **High speed electronics/links**	➡ New detectors FrontEnd
Closer to offline ➡ share algorithms ➡ **BUT** calibrations are slow	➡ **online-offline merging** ➡ more parallelism	➡ **tight**: offline is online (**LHCb**, **ALICE**) ➡ **soft**: decouple trigger & DAQ (**ATLAS, CMS**)
Vertex silicon trackers ➡ **BUT** 800M channels ➡ **AND** large combinatorics	➡ **Hardware track trigger**	➡ regional readout (**ATLAS**) ➡ detector coincidence (**CMS**)

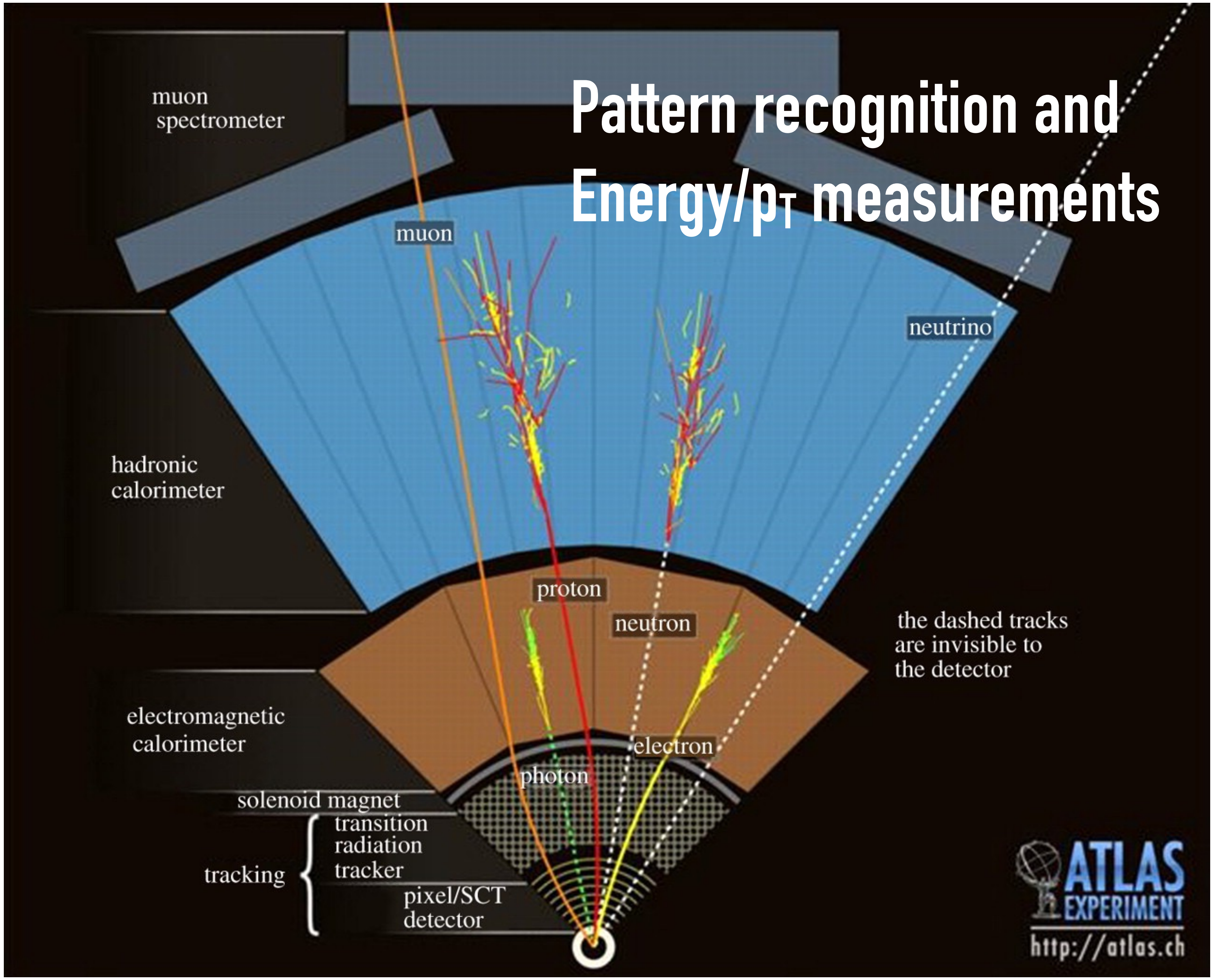
To slow down the scaling of the data flow



ATLAS AND CMS

Studying the Standard Model at the high energy frontier

ATLAS/CMS TRIGGER STRATEGY



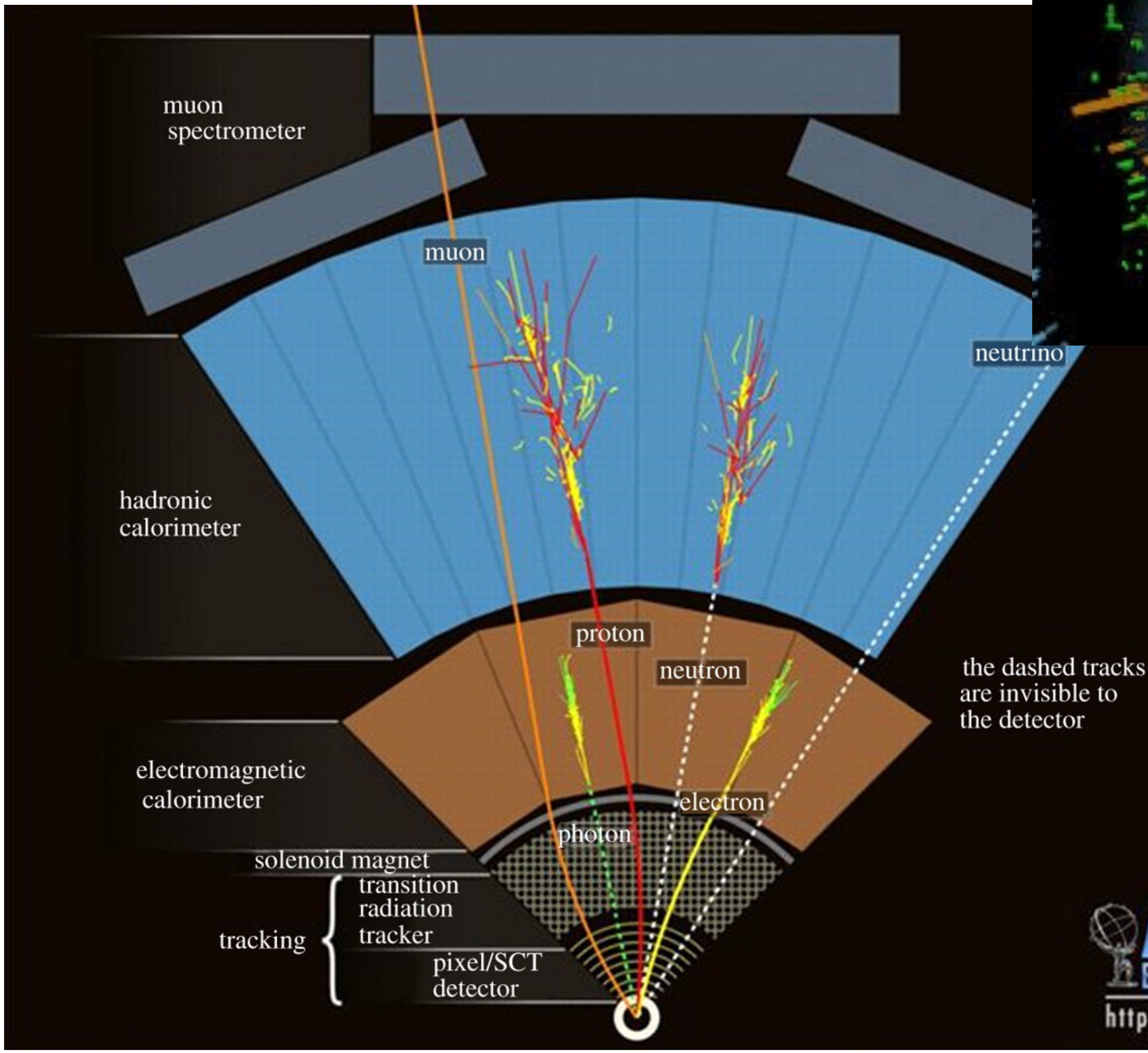
$$\frac{\sigma_{tot}}{\sigma_{H(500\,\mathrm{GeV})}} \approx \frac{100\,mb}{1\,pb} \approx 10^{11}$$

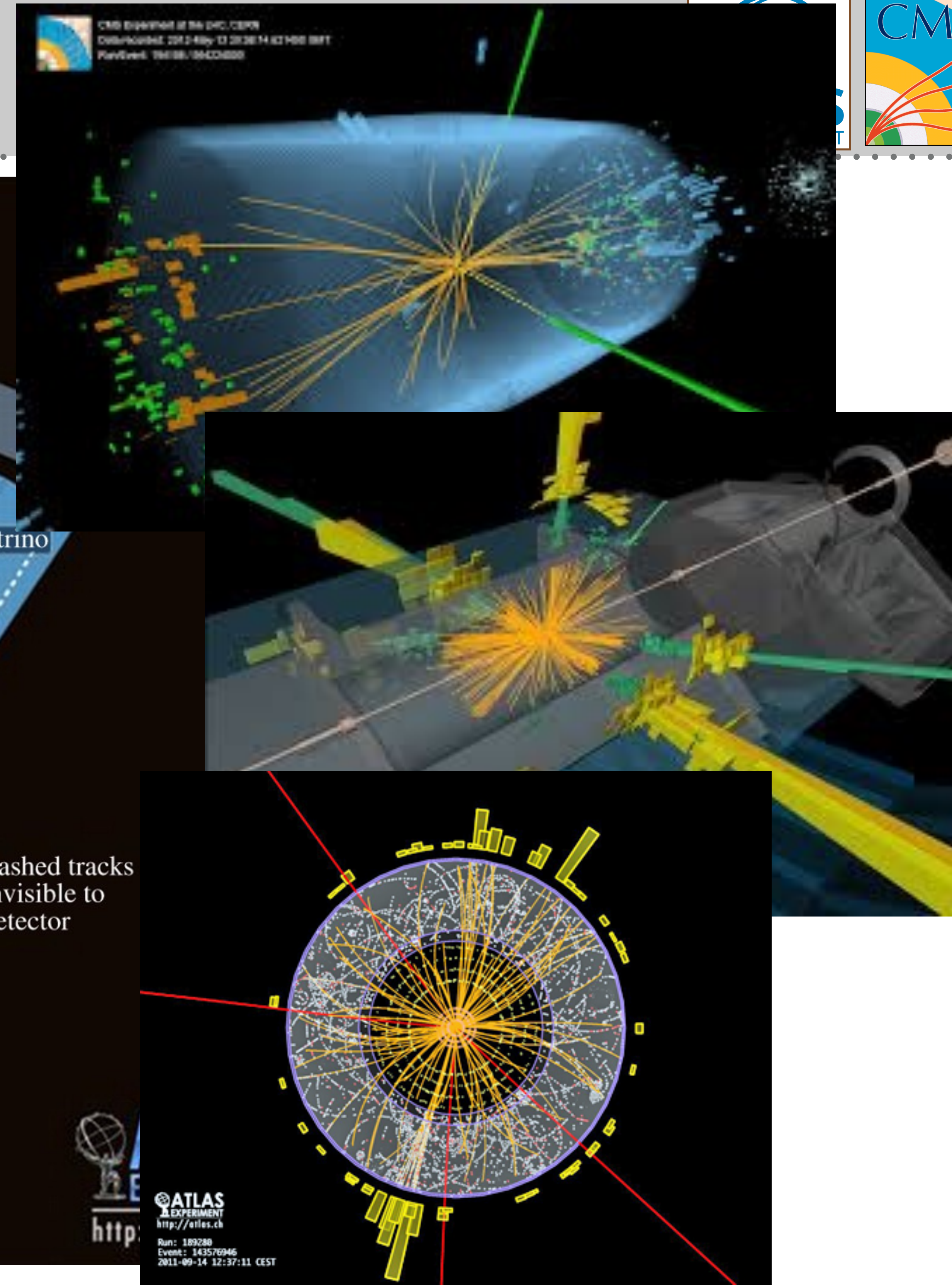
approximately 10^6 rejection

- ➡ Higher the energy, higher the mass of particles to discover
- ➡ Easy selection of signal over background
  - ➡ High p_T particles

- ➡ **Expected thousands of particles/collisions**
- ➡ **Typically hadrons with $p_T \sim 1$ GeV (low momentum jets)**

WHAT DO THEY SEE?





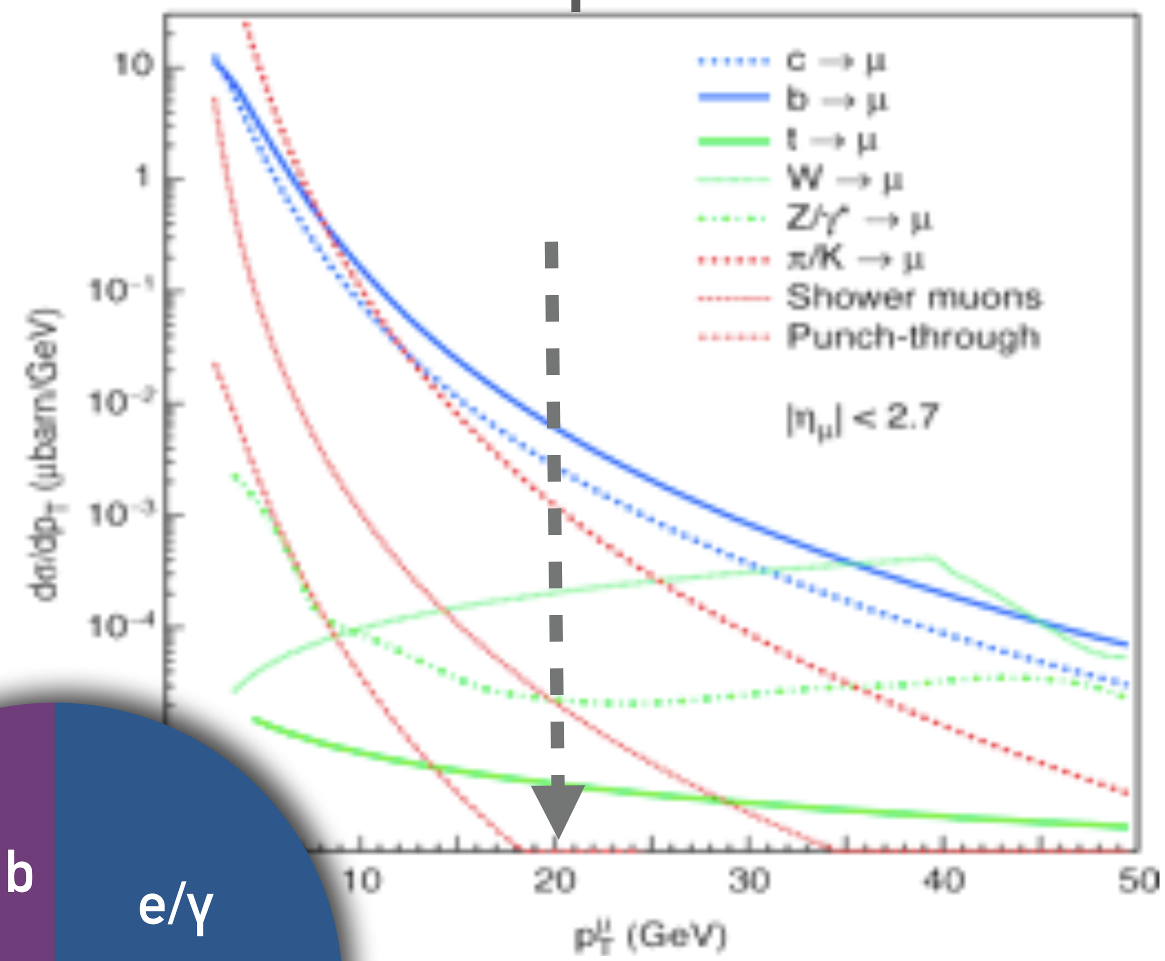
- ➡ **Pattern recognition much easier in calorimeter and muon system**
- ➡ Cannot reconstruct all tracks at 40MHz, neither at 100 kHz

Lepton identification far more easy in hadron colliders

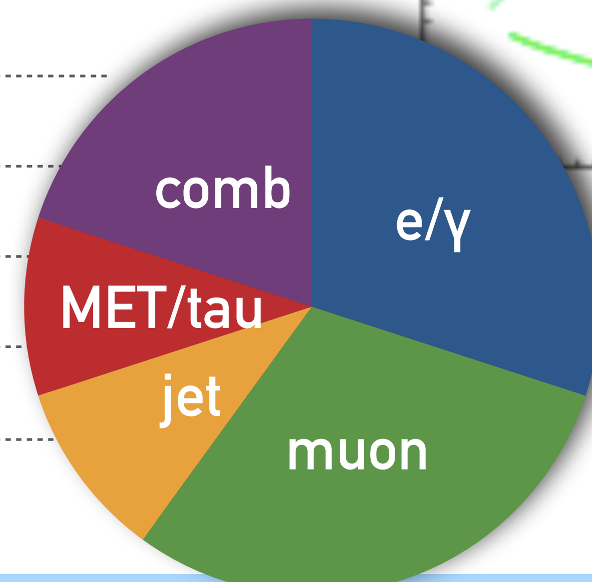
TRIGGER STRATEGIES

- Mainly lepton signs
- Wide physics program (more than 1000 selections in the menu)
- Target: same thresholds in HL-LHC

Inclusive muon spectrum at 14 TeV



	L1 p_T threshold	rate @10^{34}
e/γ	30 GeV	10-20 kHz
2 e/γ	20 GeV	5 kHz
muon	20 GeV	10 kHz
2 mu	6 GeV	1 kHz
jet	300 GeV	200 Hz
jet+ETmis	100 GeV, E_{Tmiss}>100GeV	500 Hz
4-jet	100 GeV	200 Hz



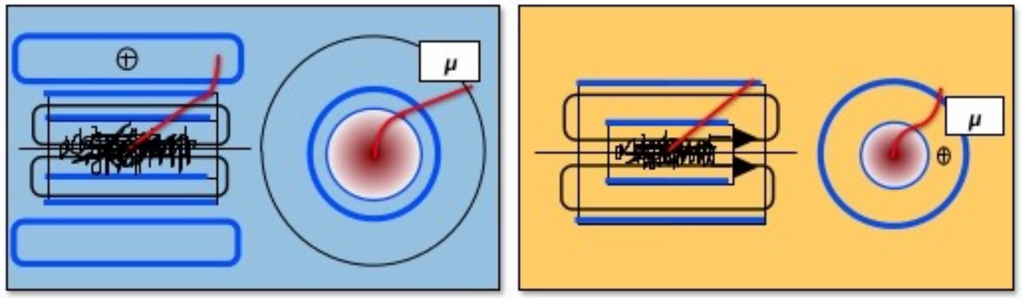
➡ **Inclusive trigger**, with sufficiently low thresholds to be sensitive to decay products of new particles and to Z and W decays

➡ Need to understand several sources of **background** and low energy spectrums

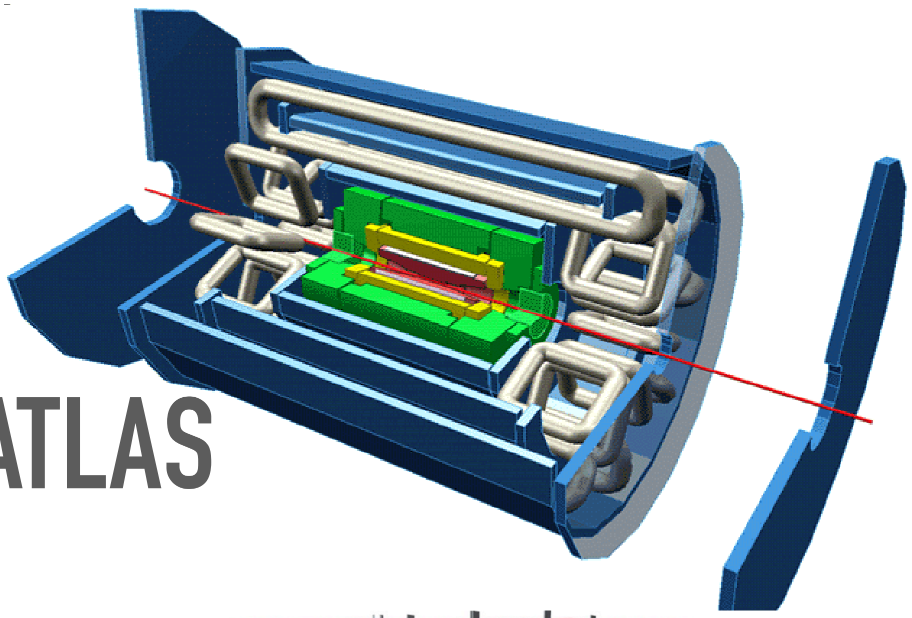
ATLAS & CMS DESIGN PRINCIPLES

Same physics plans, different competitive approaches for detectors and DAQ

- **Different magnetic field structure**
  - **ATLAS**: 2 T solenoid + Toroids
  - **CMS**: 4 T solenoid

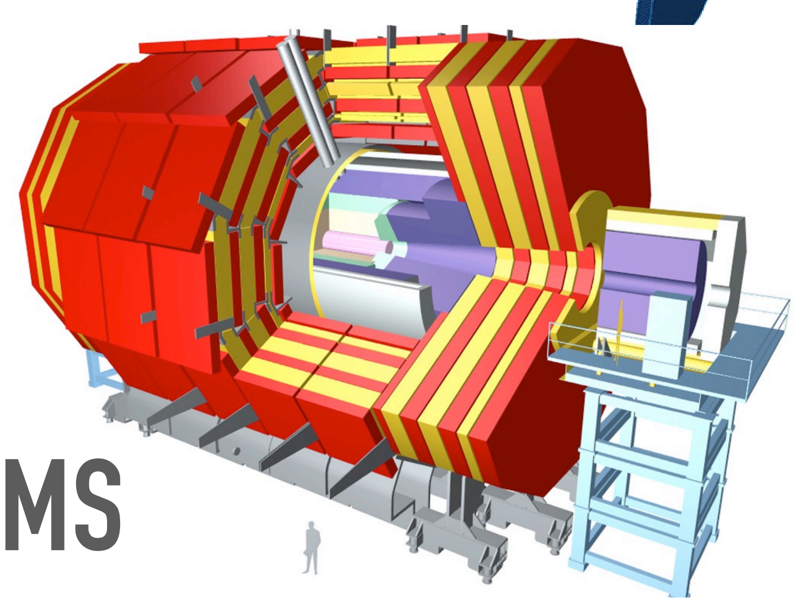


- **Different muon system**
  - **ATLAS**: air-core toroid, minimising MS, standalone muon reconstruction, fast dedicated trigger detectors (RPC/TGC, 10 ns)
  - **CMS**: high bending power and instrumented return yoke, 2 independent trigger systems (DT/CSC + RPC)



ATLAS

- **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

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

CMS: 2-STAGE EVENT BUILDING

Cannot do EB at 100kHz

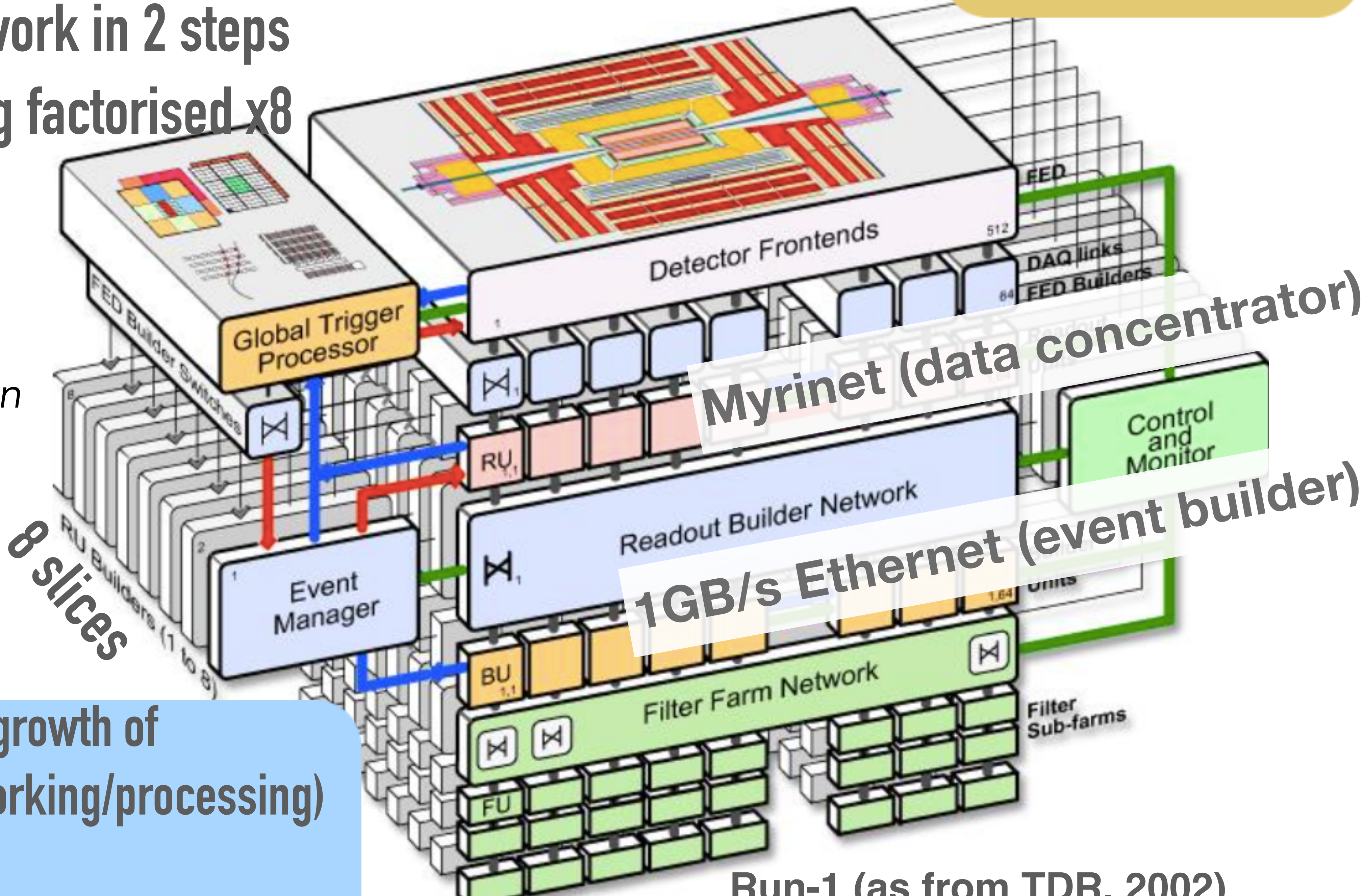
100GB/s readout network in 2 steps

100kHz Event Building factorised x8

CMS DAQ-1

2 EB networks in blu

Filter network in green



- Bet on exponential growth of technologies (networking/processing)
- Scalable, 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

Run1: 100 GB/s network

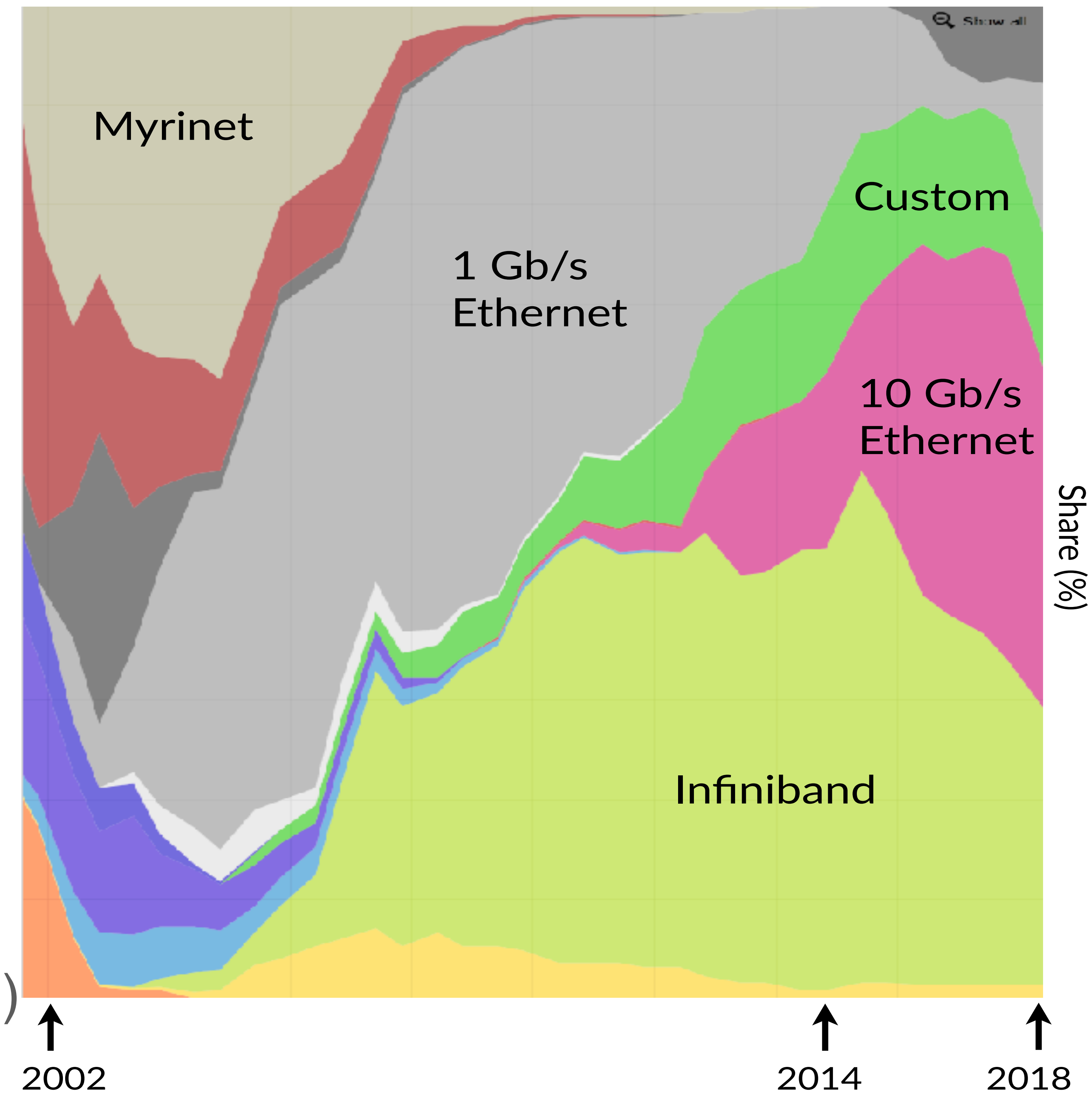
Myrinet widely used when DAQ-1 was designed

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

Run2: 200 GB/s network

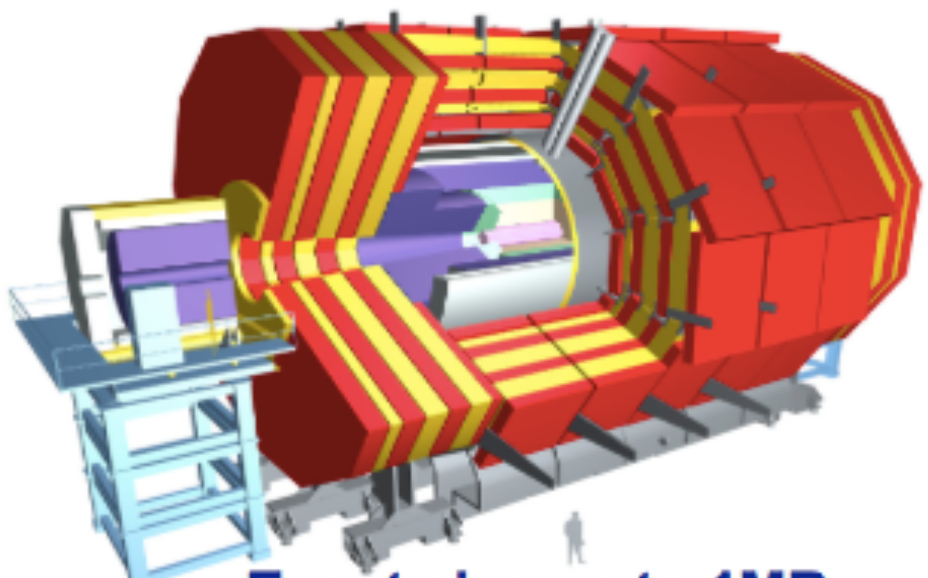
- ➡ Increased event size to 2MB
- ➡ Technology allows single EB network (56 Gbps FDR Infiniband)
- ➡ Myrinet —>10/40 Gbps Ethernet

Top500.org share by interconnect family

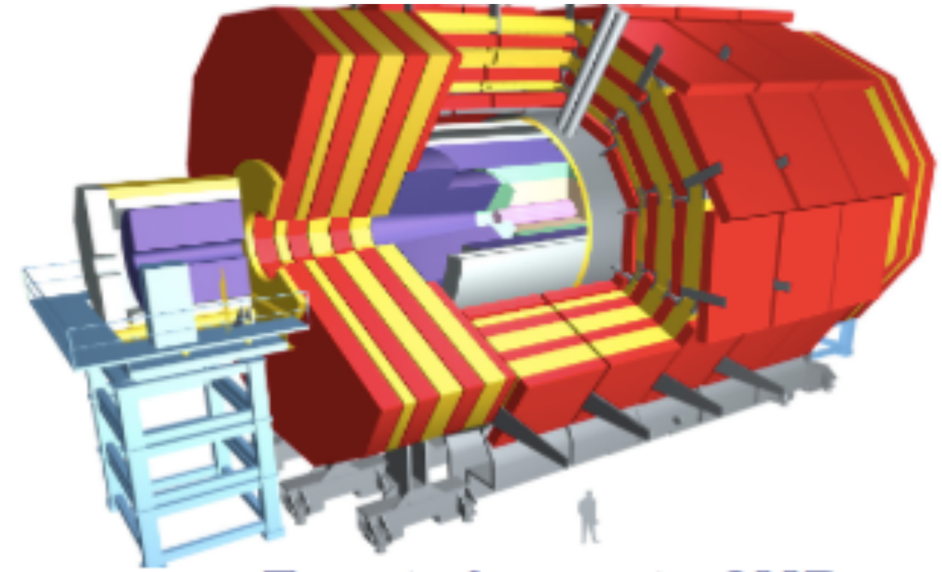


Choose best prize/bitps

EVOLUTION FROM RUN-1 TO RUN-2

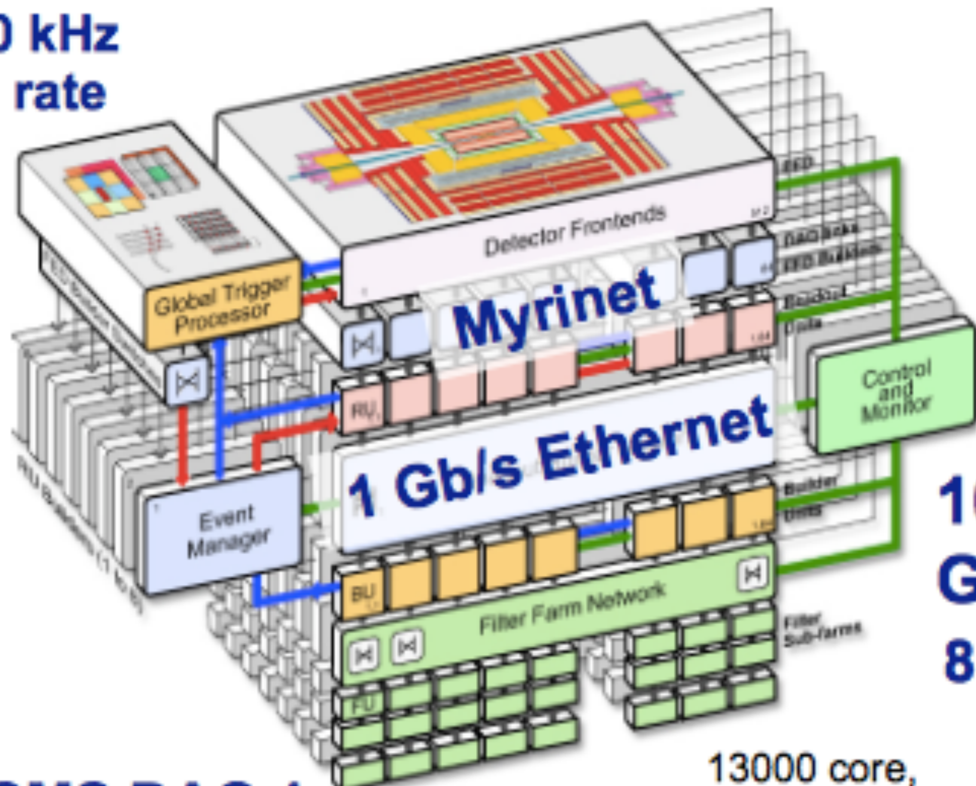


Event size up to 1MB



Event size up to 2MB

100 kHz
L1 rate



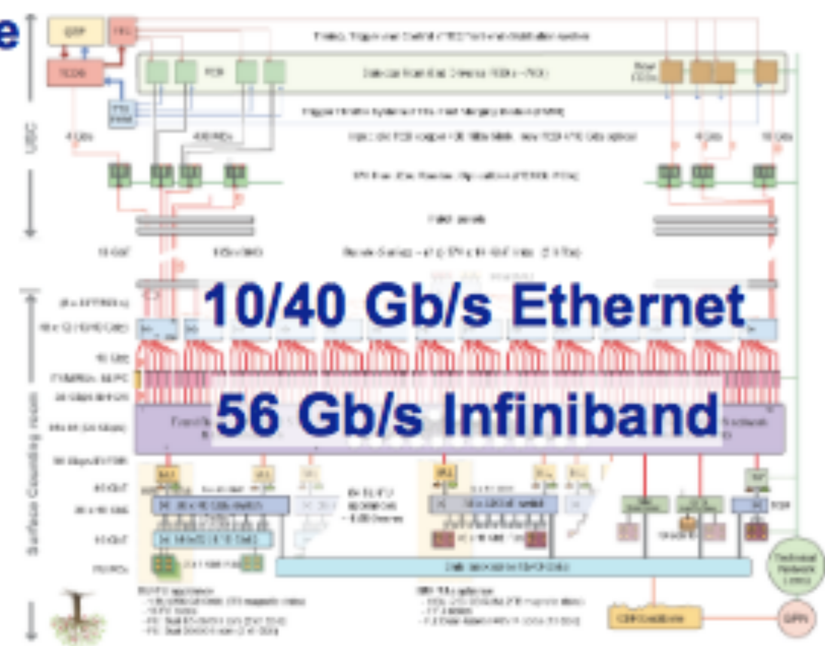
100
GB/s
8 slices

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

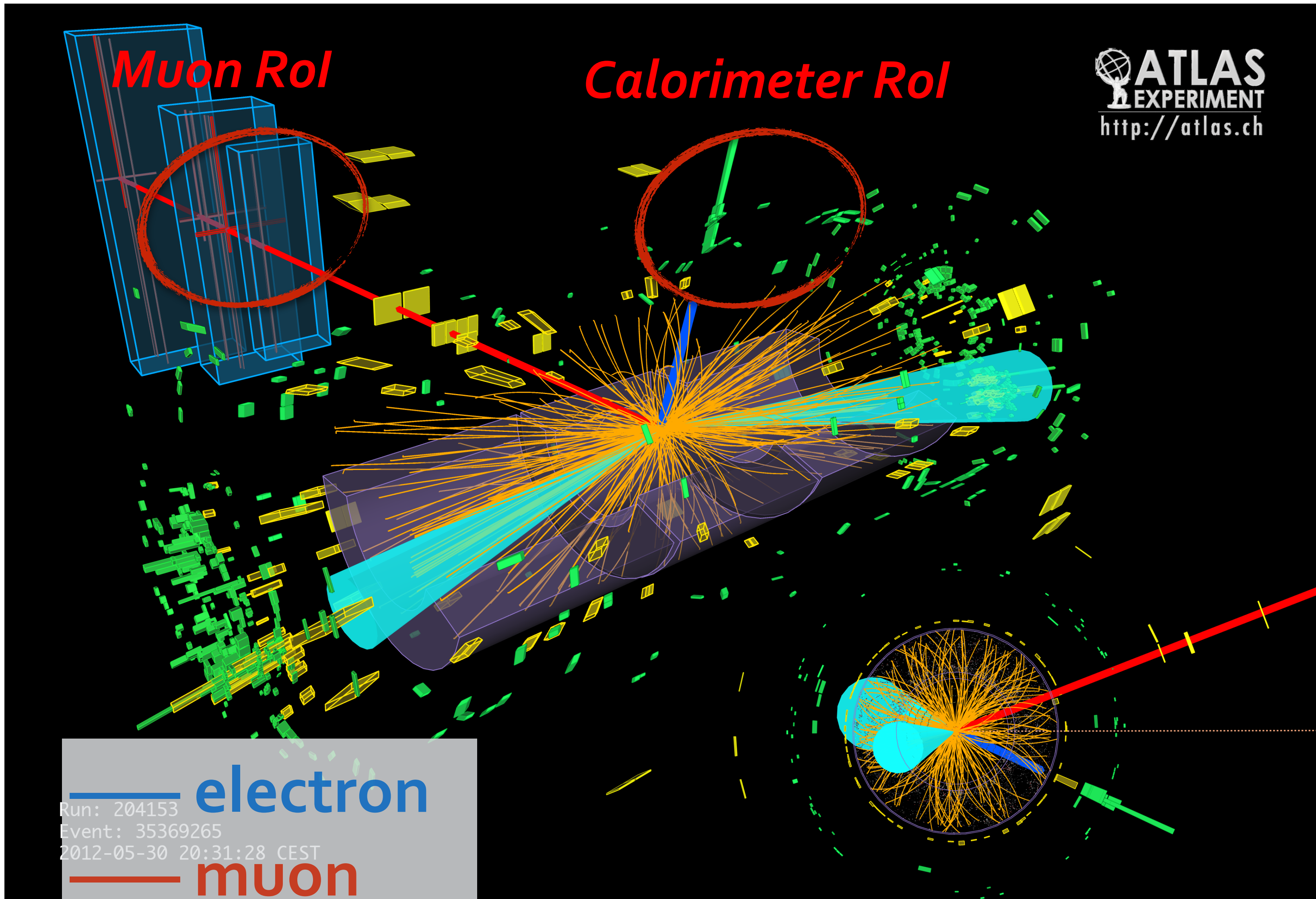
CMS DAQ 2

16000+ core,
900 host
filter farm

~ 3-6 GB/s to storage

ATLAS: REGION OF INTEREST (ROI) DATAFLOW

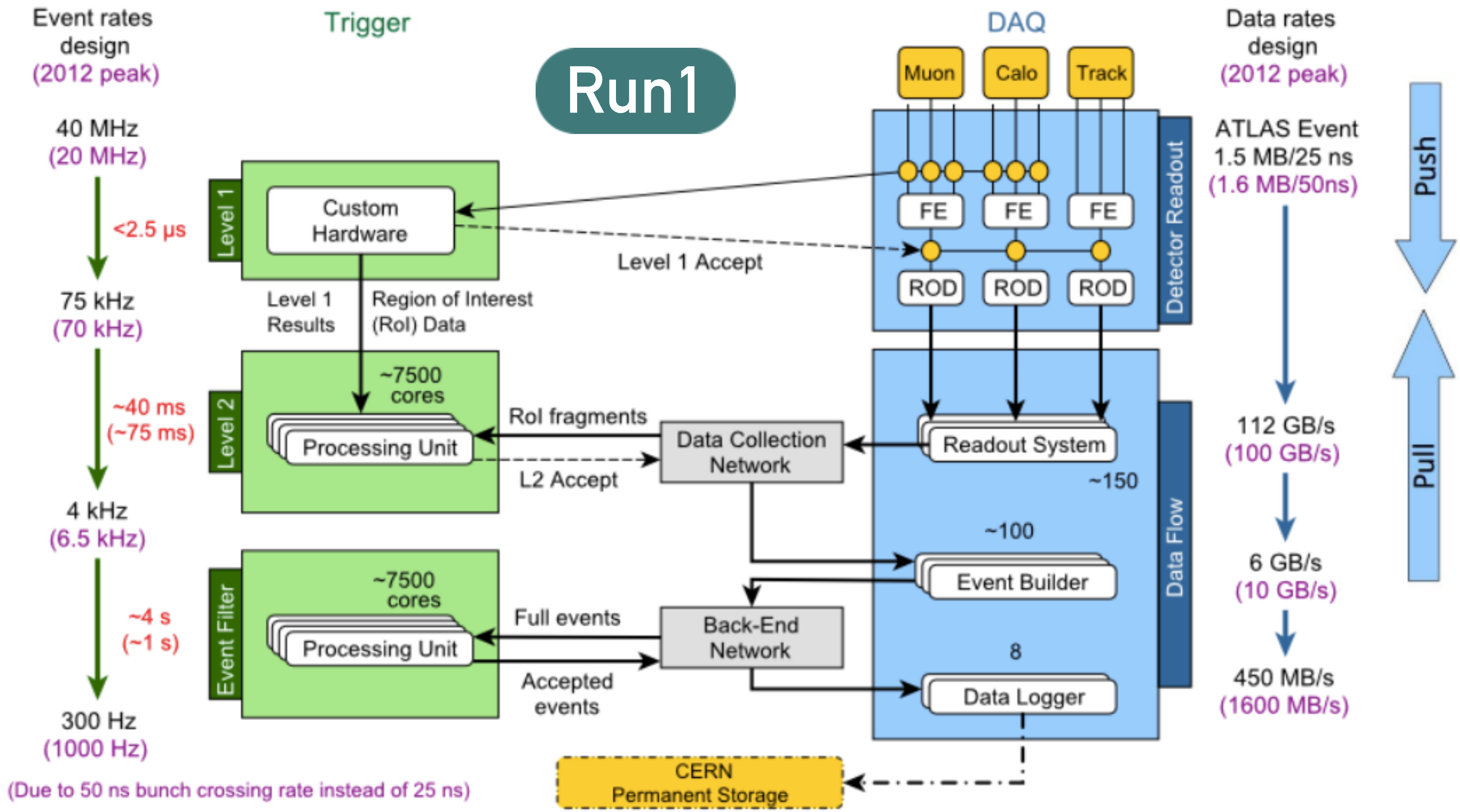
HLT selections based on regional readout and reconstruction, seeded by L1 trigger objects



- ➡ 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

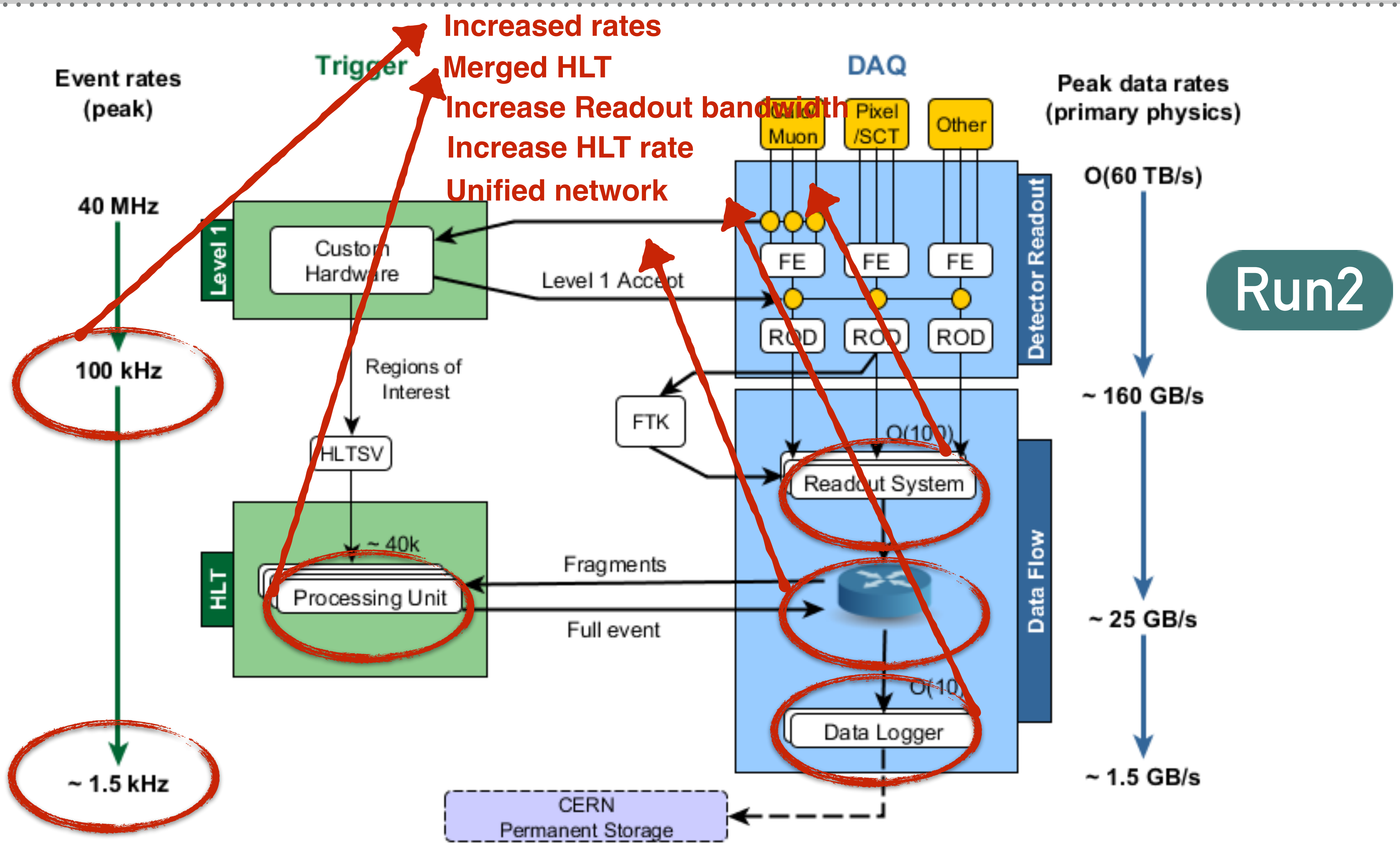
ATLAS: SEEDED RECONSTRUCTION HLT

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



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

NEW TDAQ ARCHITECTURE FOR RUN-2



- **New architecture with 2 levels only allows more flexibility**
  - New: network architecture, Readout System (PCIe boards), trigger detectors

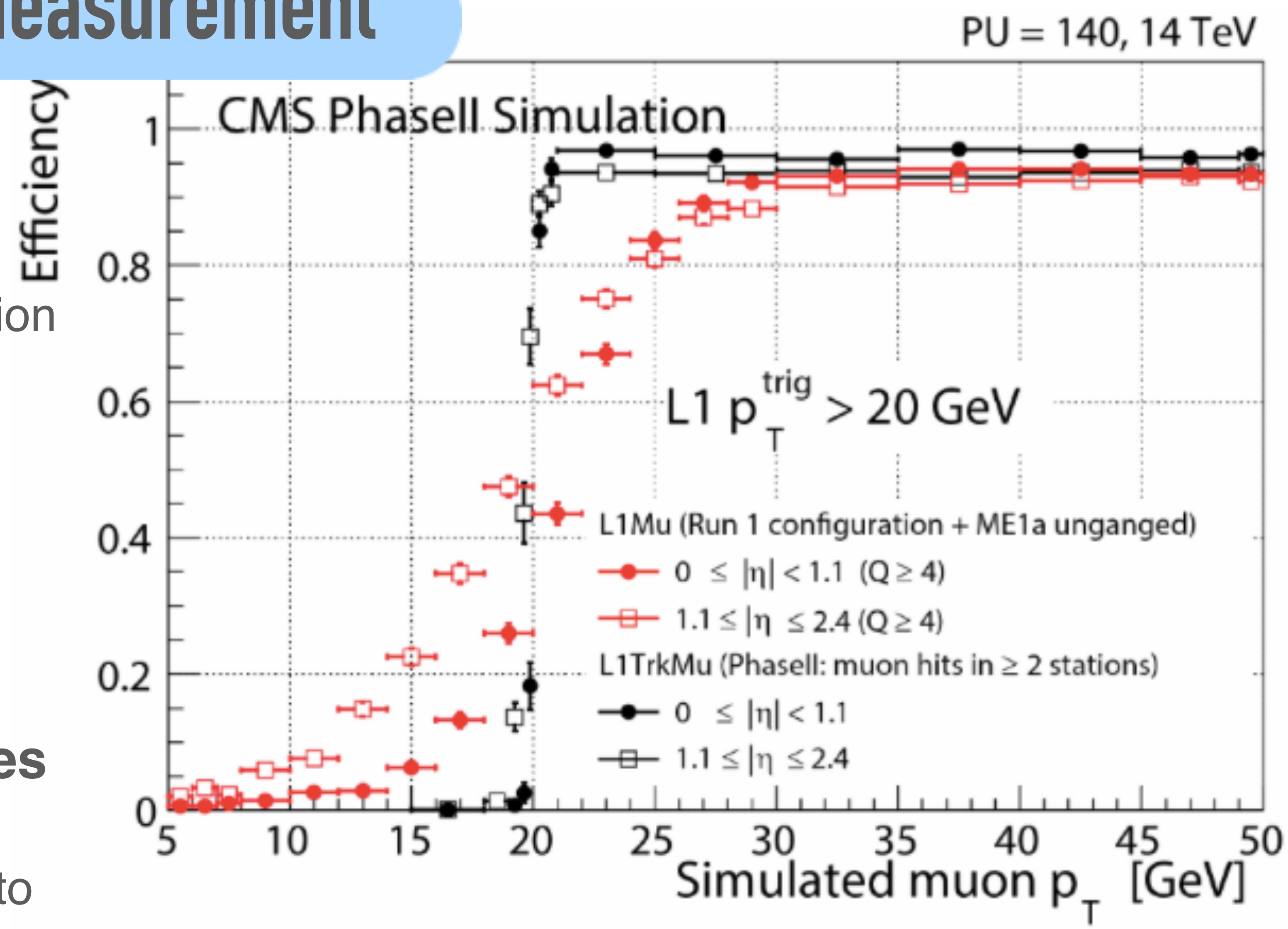
Increasing resolution on p_T measurement

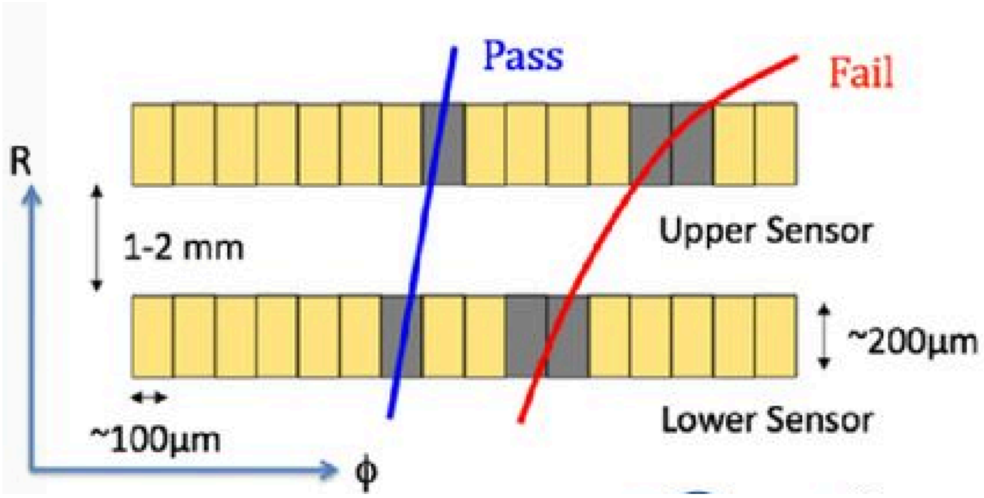
- **Main goals:**
  - Rejecting hadrons/jets mimicking leptons
  - Selecting particles from same interaction

- **Global tracking not feasible at 40MHz so reduce to 1MHz with:**
  - regional readout (**ATLAS**)
  - detector coincidence (**CMS**)

- **Event at 1MHz the strategy includes two steps:**
  - track filtering: a first pattern matching to reduce combinatorics
  - track fitting: linearised algorithms on dedicated processors

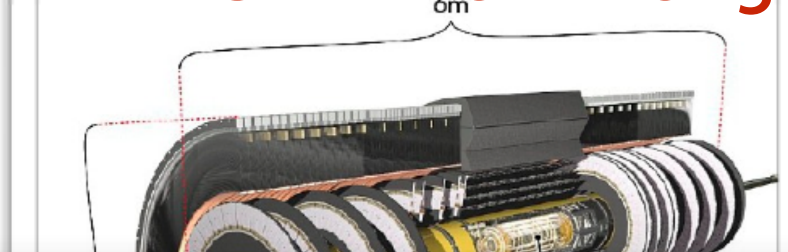
- **Algorithms can run on fast modern electronics (FPGAs/ASICs)**

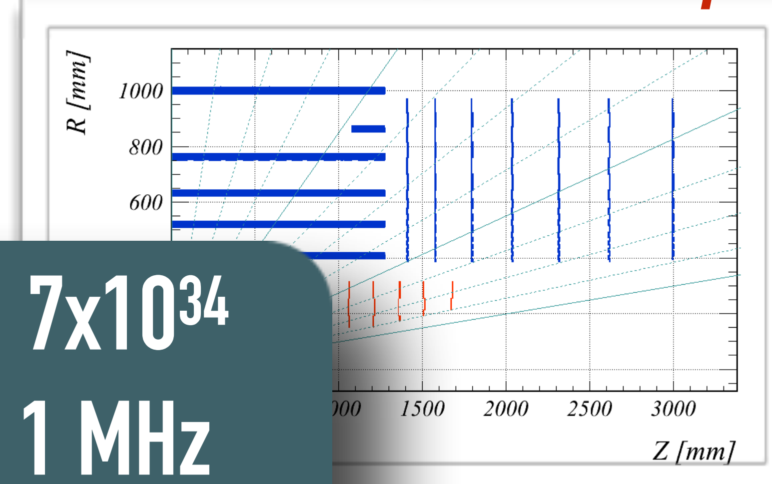




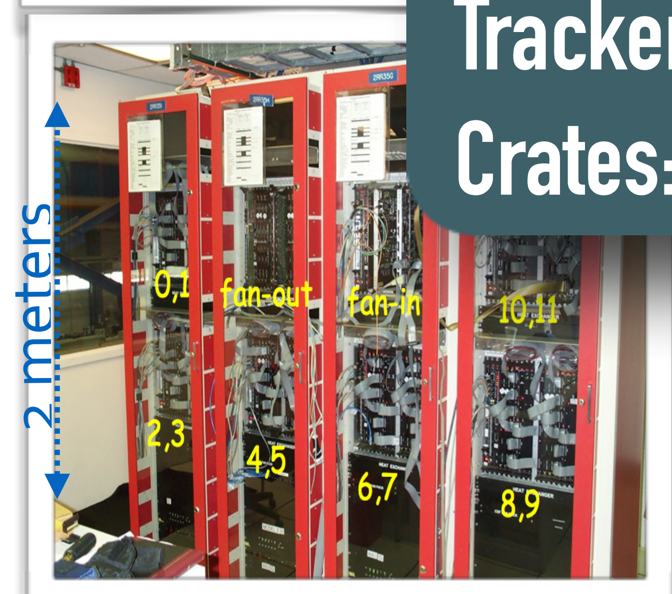
TREND OF TRACKING TRIGGER SYSTEMS

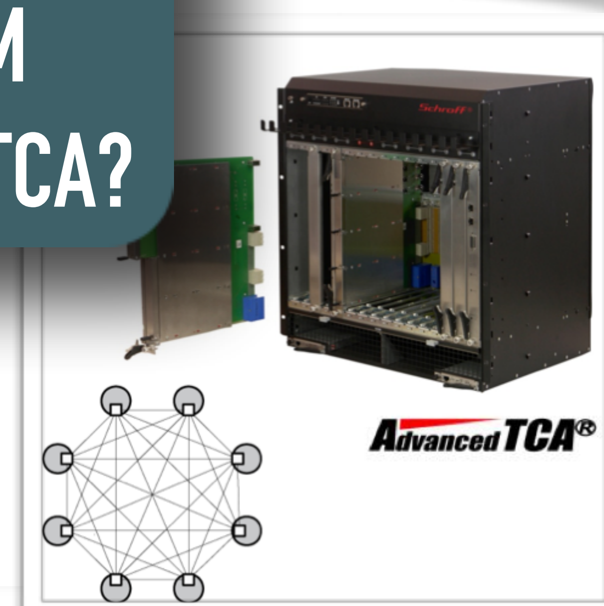
CDF- SVX II	ATLAS FTK Run2-Run 3	ATLAS HW-TT Run 4

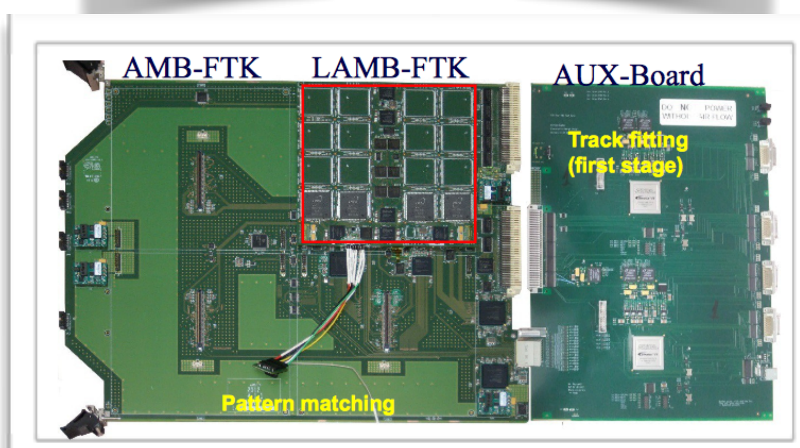


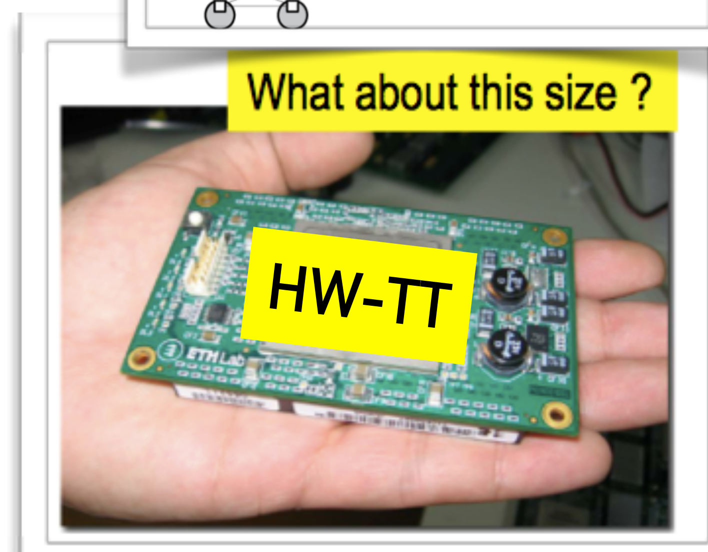


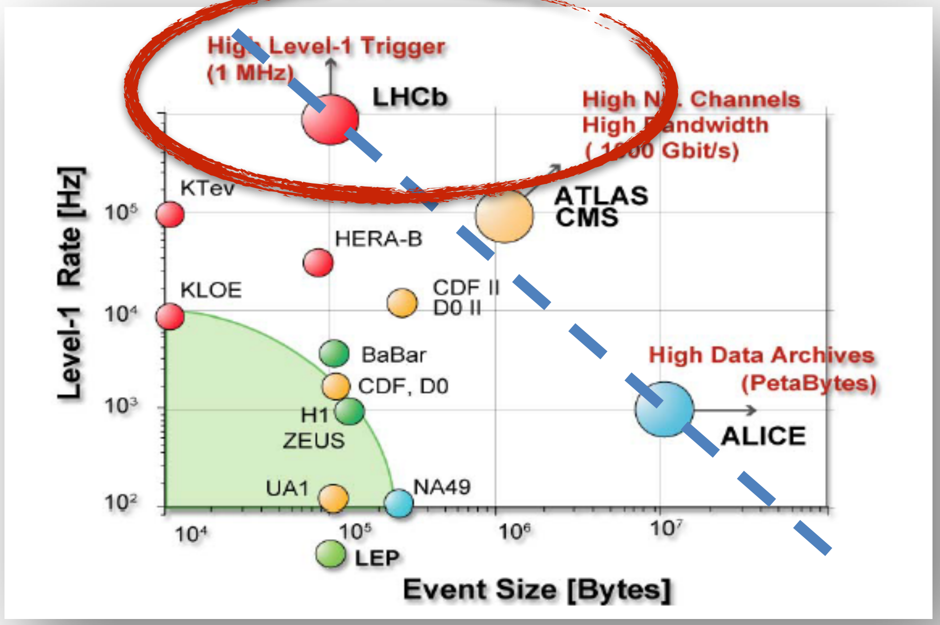
Luminosity:	$3x10^{32}$ →	$3x10^{34}$ →	$7x10^{34}$
L1 rate:	30 kHz →	100 kHz →	1 MHz
Tracker channels:	0.2M →	100M →	800M
Crates:	10 VME →	13 ATCA →	50ATCA?







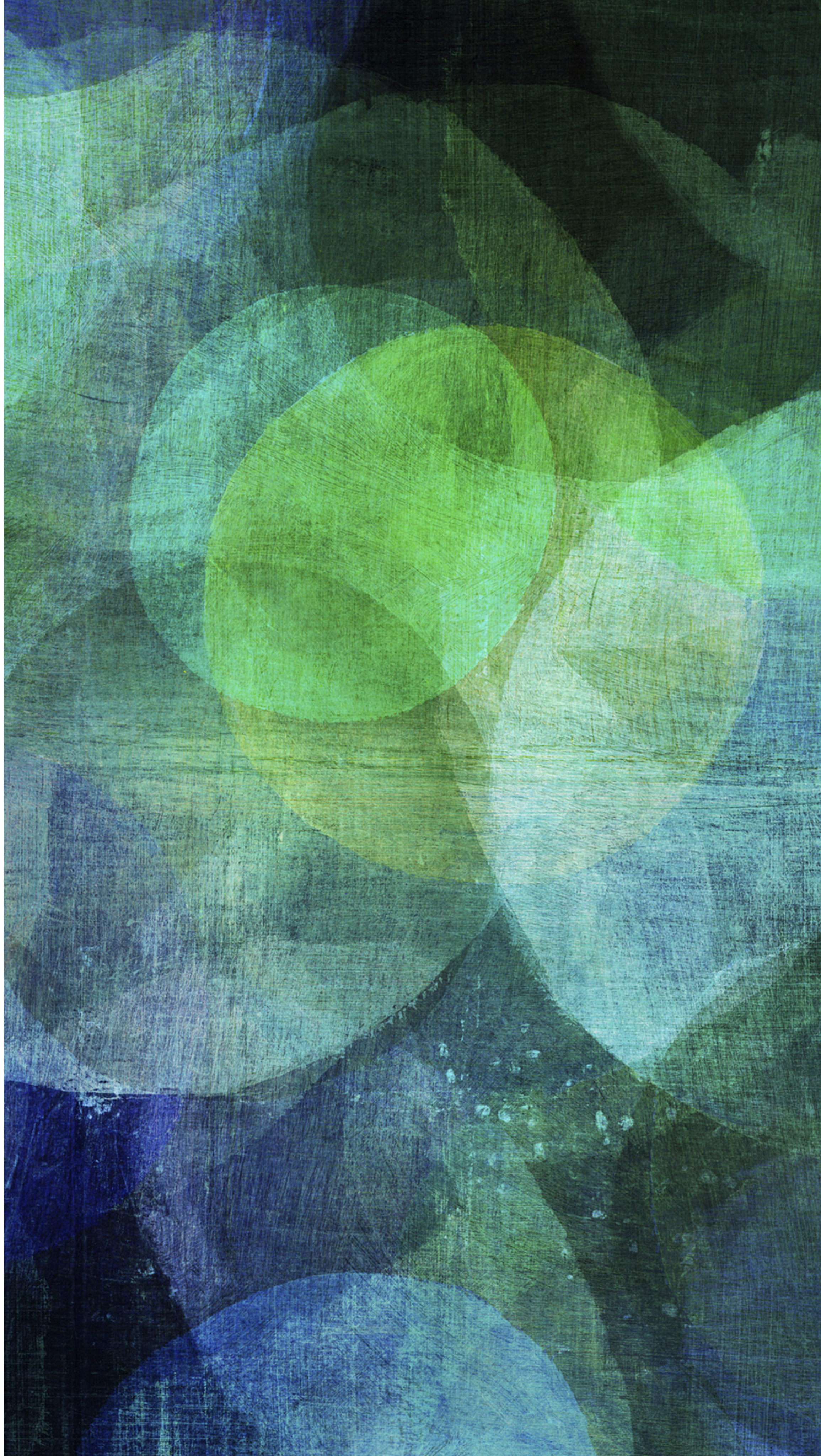




LHCb, THE B-MESON OBSERVATORY

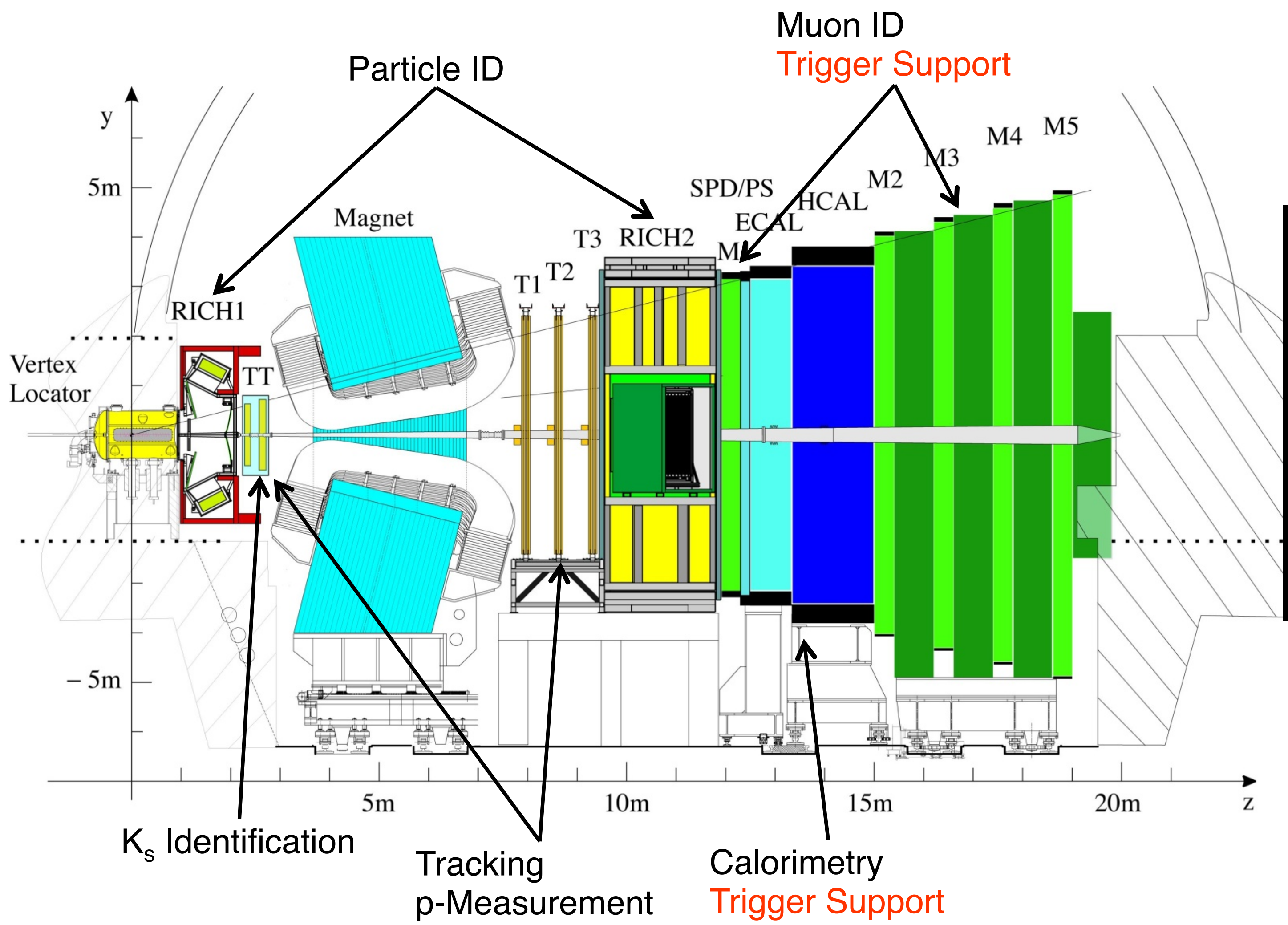
The lightest experiment to study the heavy b-quark

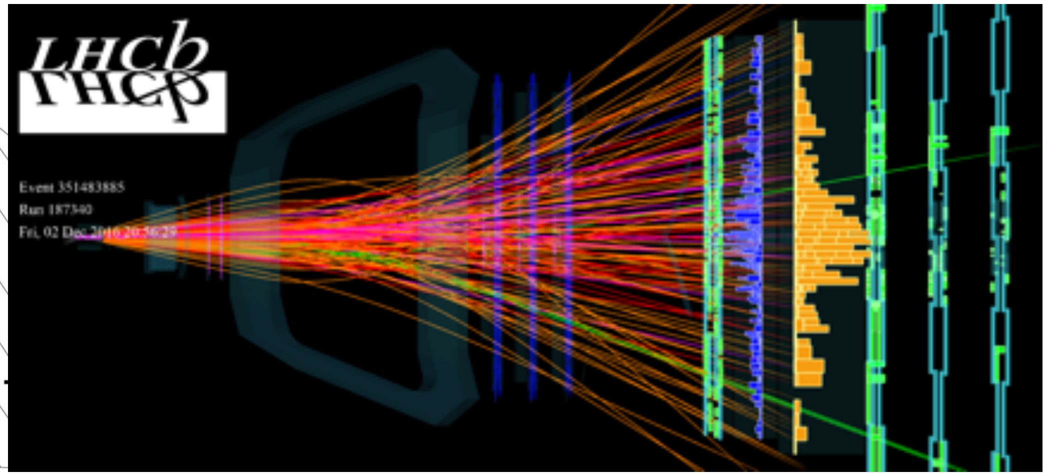
http://lhcb-public.web.cern.ch/lhcb-public/

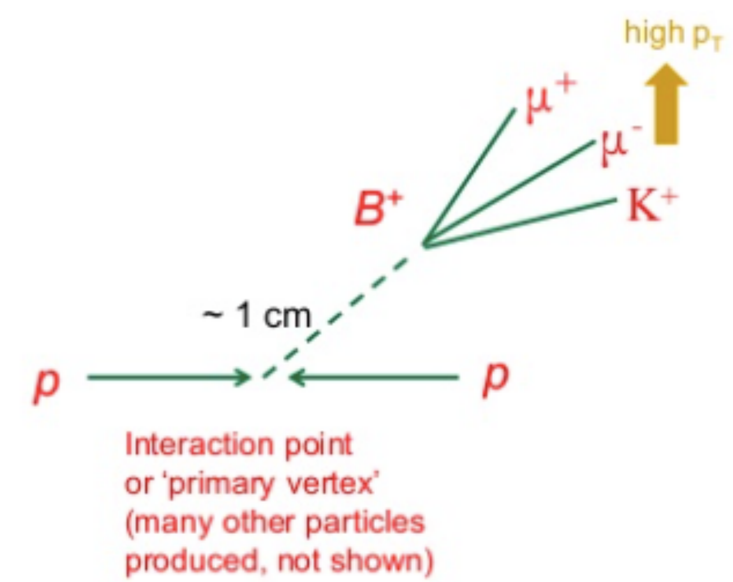


LHCB DESIGN PRINCIPLES

- ➡ **Precision measurements of CPV and rare decays in the B system**
  - ➡ Large σ_{BB}~500 μb, but still σ_{BB}/σ_{Tot} ~ 5x10^{-3}
  - ➡ Interesting B decays quite rare: BR ~ 10^{-5}

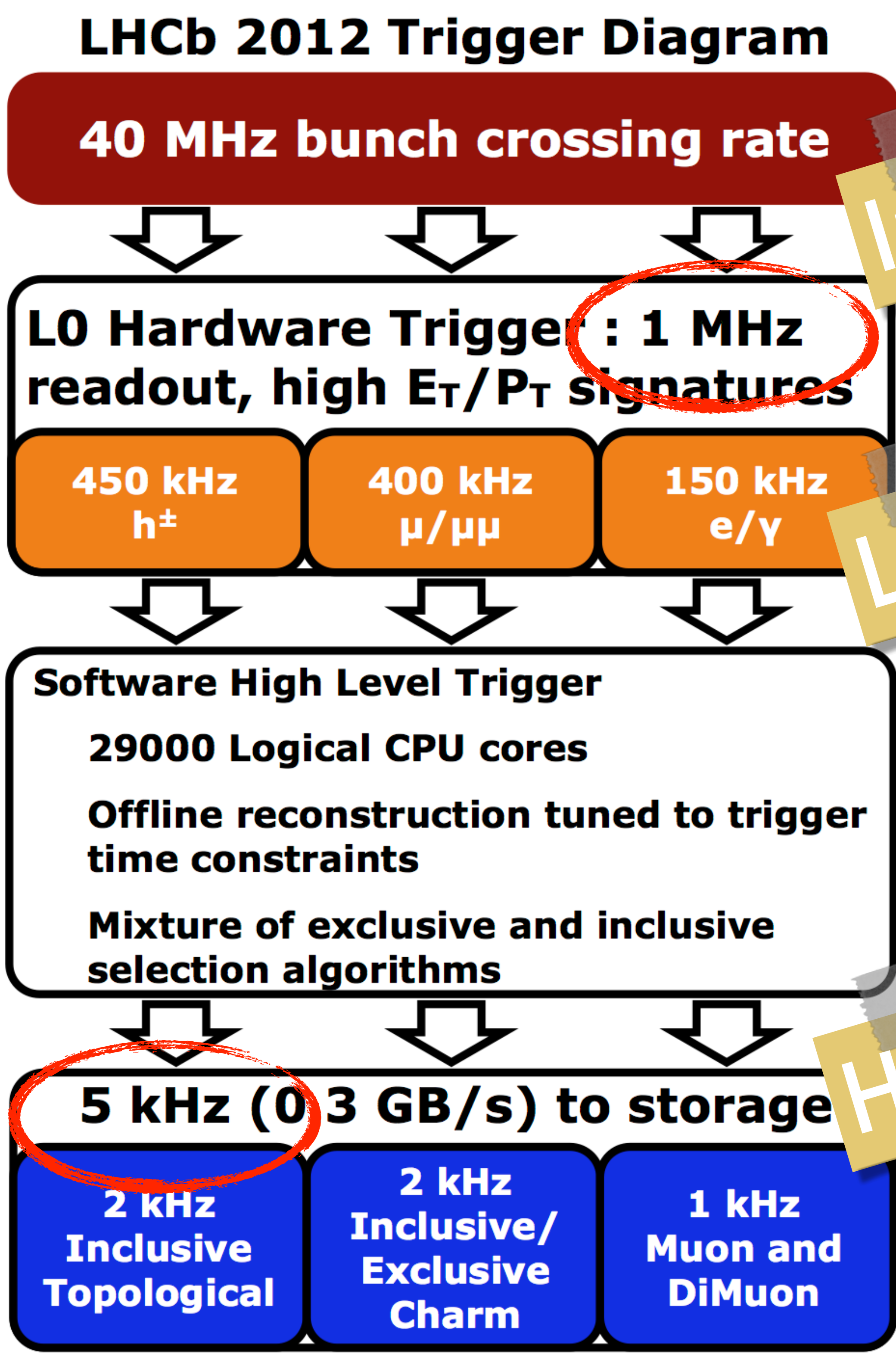






- ➡ Single forward arm spectrometer (reduced acceptance)
- ➡ Selection of B mesons using **p_T and impact parameter**, related to high mass and long lifetime of the b-quark

LHCB TRIGGER STRATEGY



Input rate

4μs latency

L0 trigger

HighLevel

Low input rate and occupancy

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

- Enhances B content with high E_T particles and reject complex events
- Mainly hadronic triggers

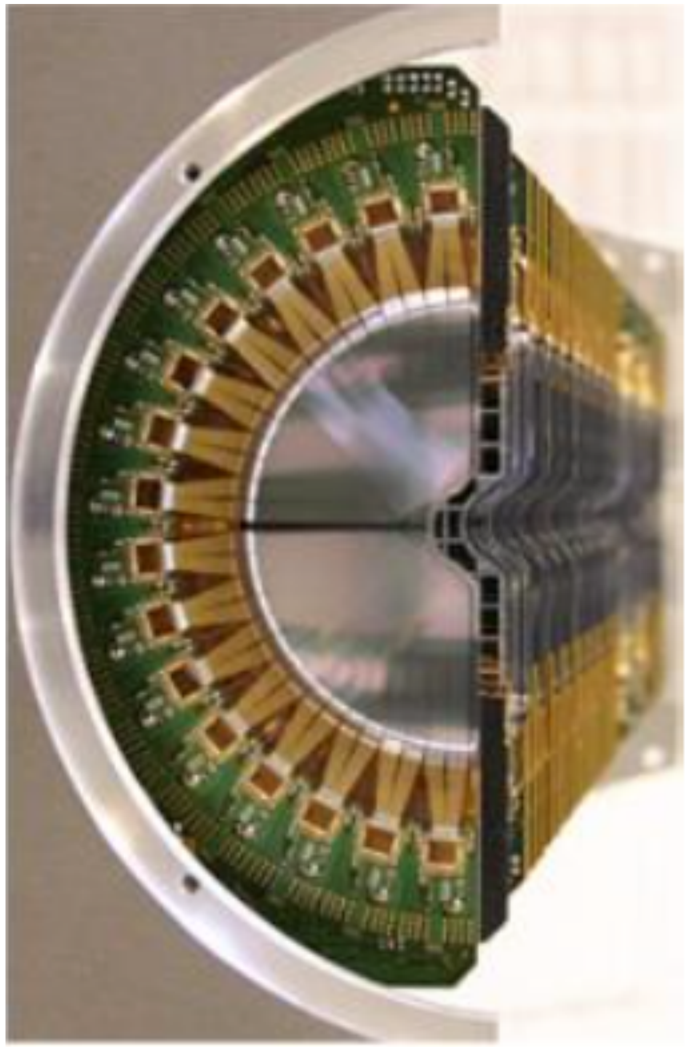
60kB * 1MHz= 60 GB/s readout network

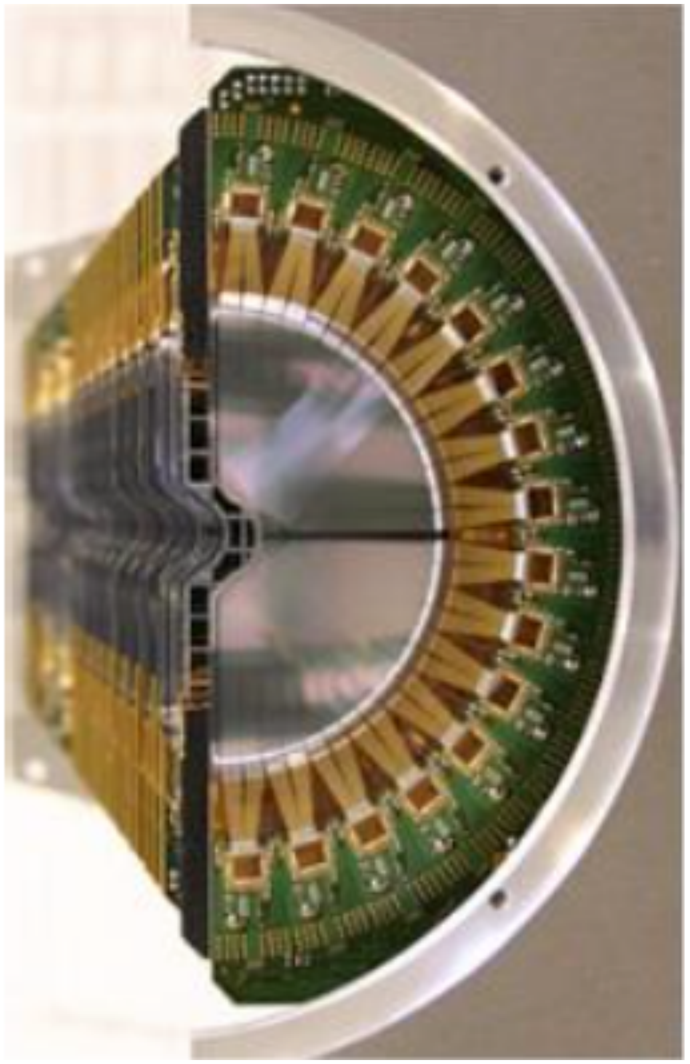
- Inclusive selections (for calibration, alignments and systematics)
- Multitude of exclusive selections

Run1: collected 3 fb^{-1} (~300×10^9 b-antib pairs)

LEVEL-0 PILE-UP SYSTEM

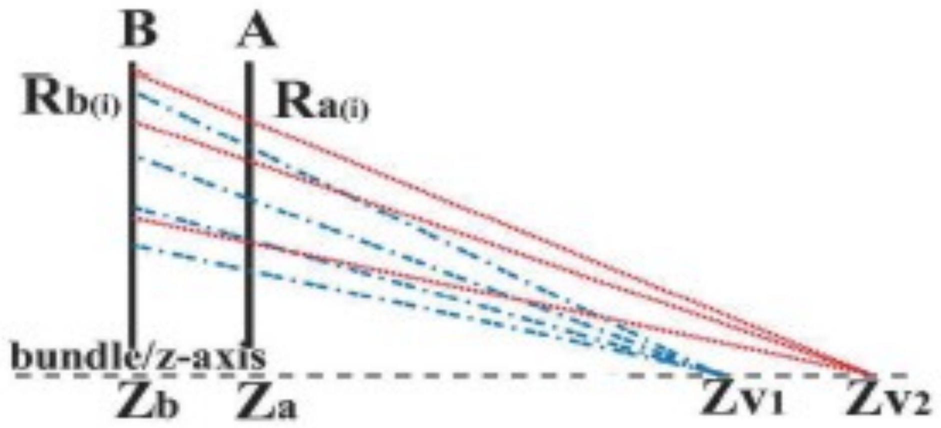
VELO silicon detector 8mm from the beam,
for secondary vertex reconstruction





Suppress events with multiple primary interactions:

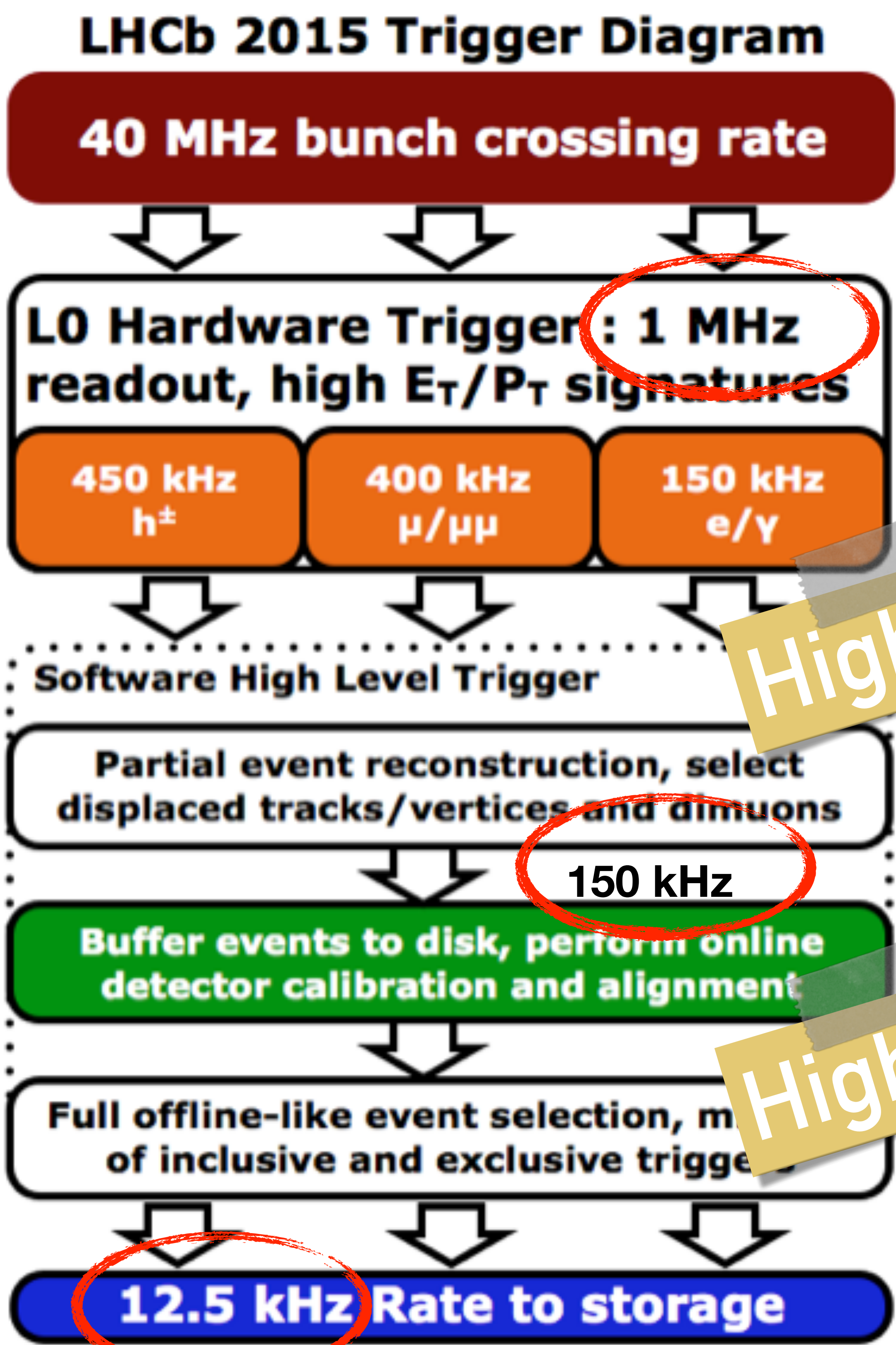
- easiest reconstruction
- reducing: event size, bandwidth and processing



- Two dedicated layers to perform **simplified vertex reconstruction**
- Moves by 29 mm at every fill. Re-alignment required

Typical performance:
60% efficiency identifying double interactions with 95% purity

SCHEMA EVOLUTION FOR RUN2



Defer processing when there are no beam

—> Optimise CPU usage (70% idle)

With large buffer between two stages (4PB) can perform real-time calibration and alignments

Large benefit from VELO alignments at each fill!

Synchronous with DAQ

- ✦ Tracks and vertices for impact parameter (in 35ms)

Decouple HLT2 from DAQ

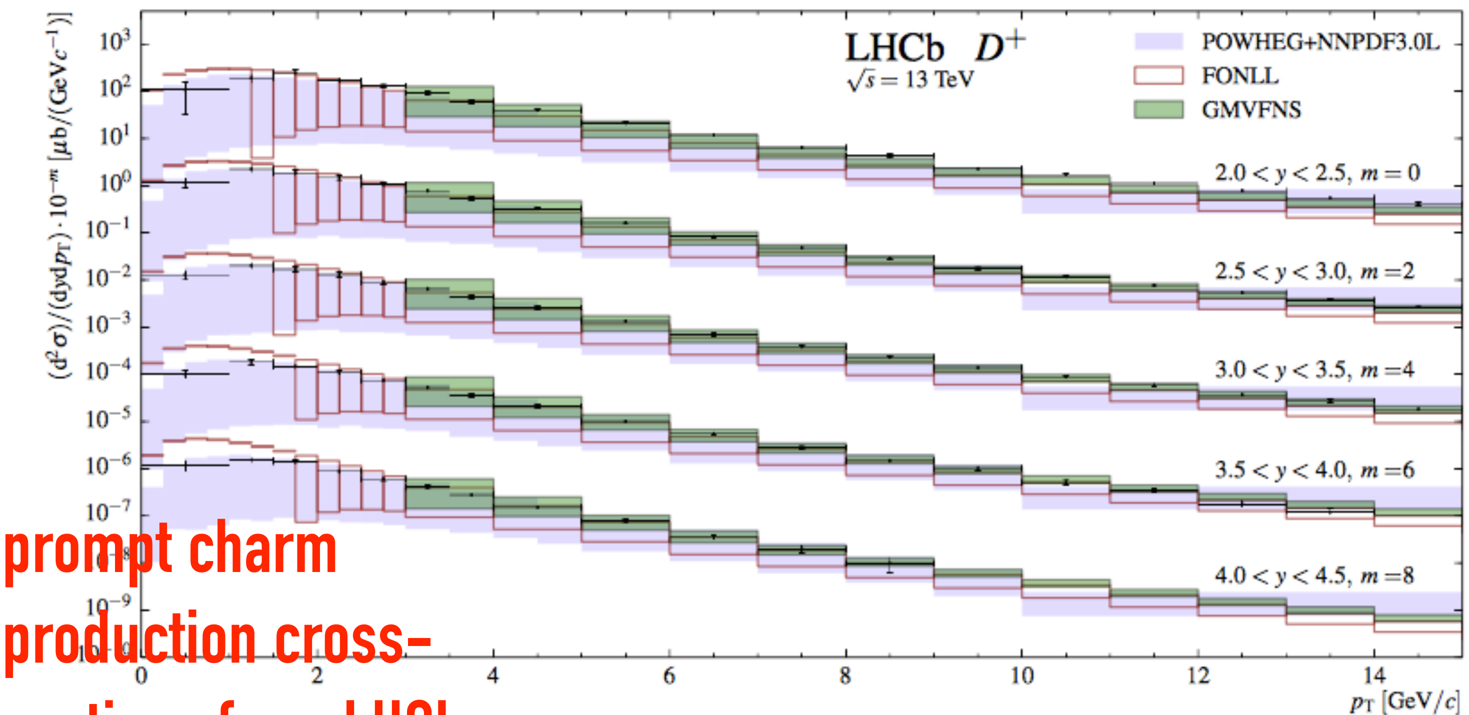
Deferred Processing

- ✦ Reconstruct with offline-like calibrations (in 350ms), becoming real time physics analysis
  - ➤ Machine learning (BDT) to separate charm/ beauty decays

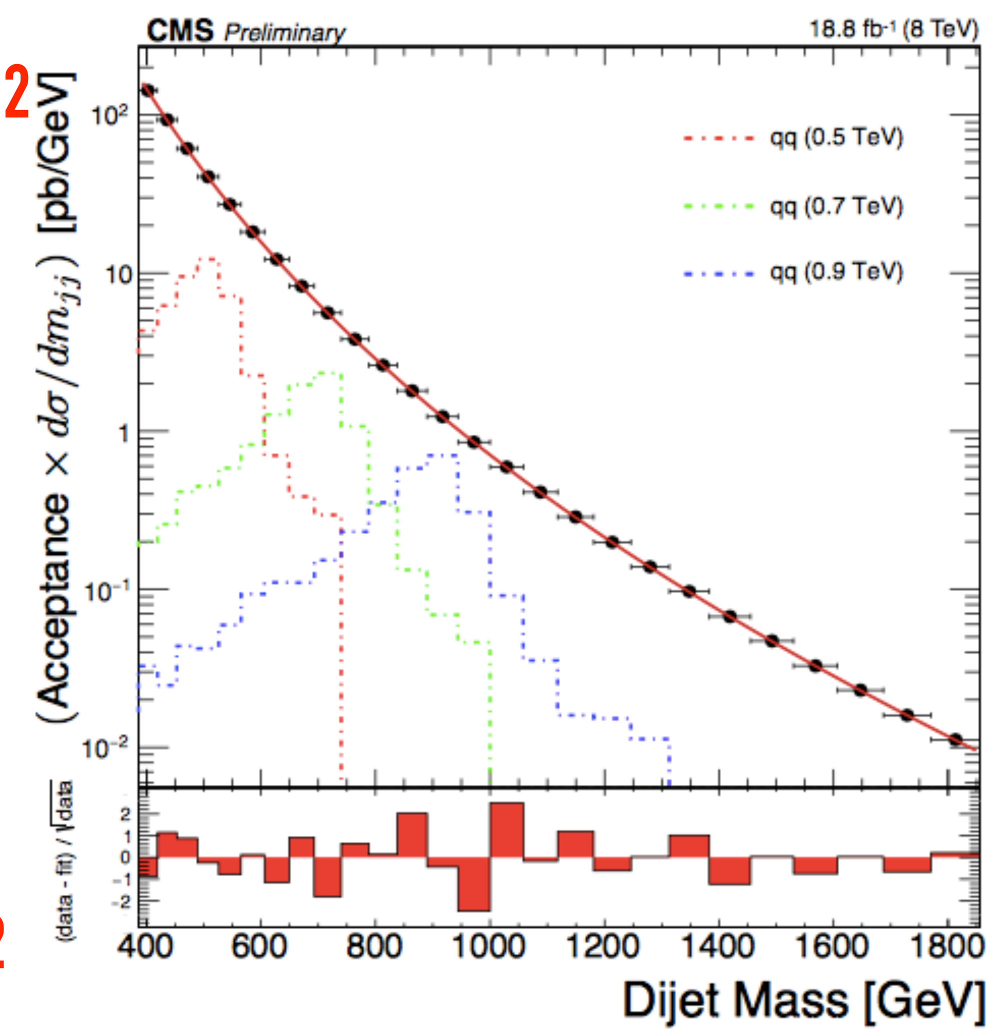
A NEW TREND: REAL TIME ANALYSIS

Can we get rid of FrontEnd raw data?

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



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

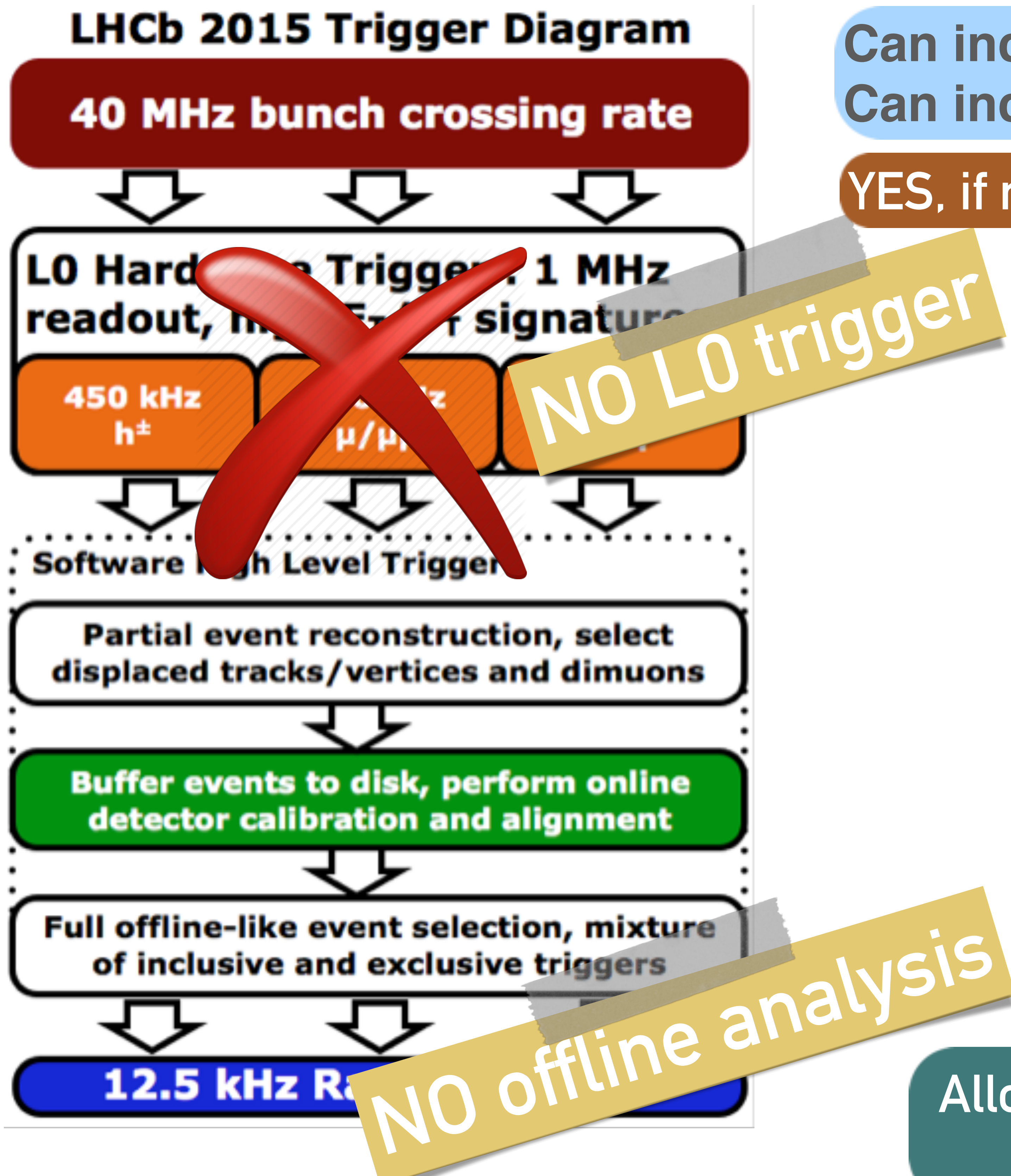


di-jet mass spectrum from CMS data-scouting in Run2

➡ Main data stream for LHCb&ALICE upgrade

➡ and be a guidance for all other experiments

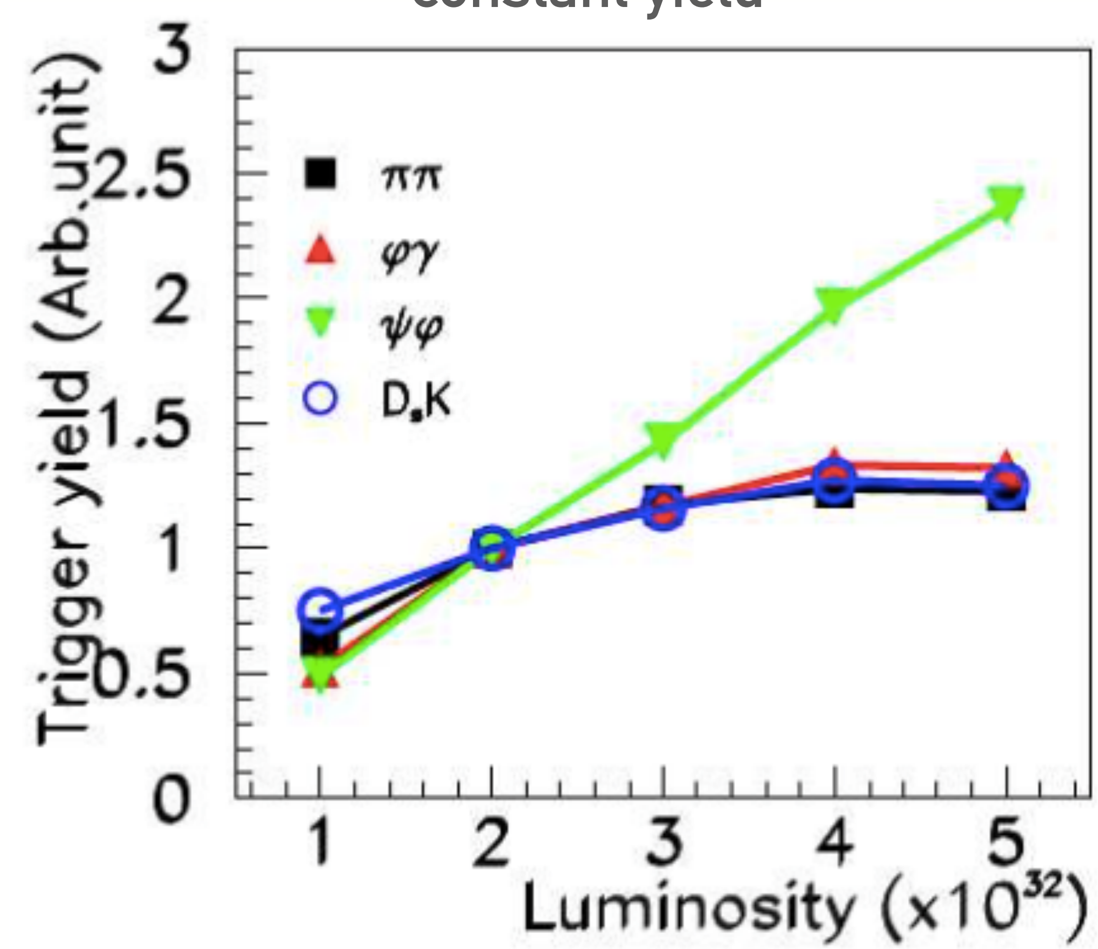
UPGRADES TOWARDS 2020 (RUN3)



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

YES, if remove the limit from L0 1MHz readout!

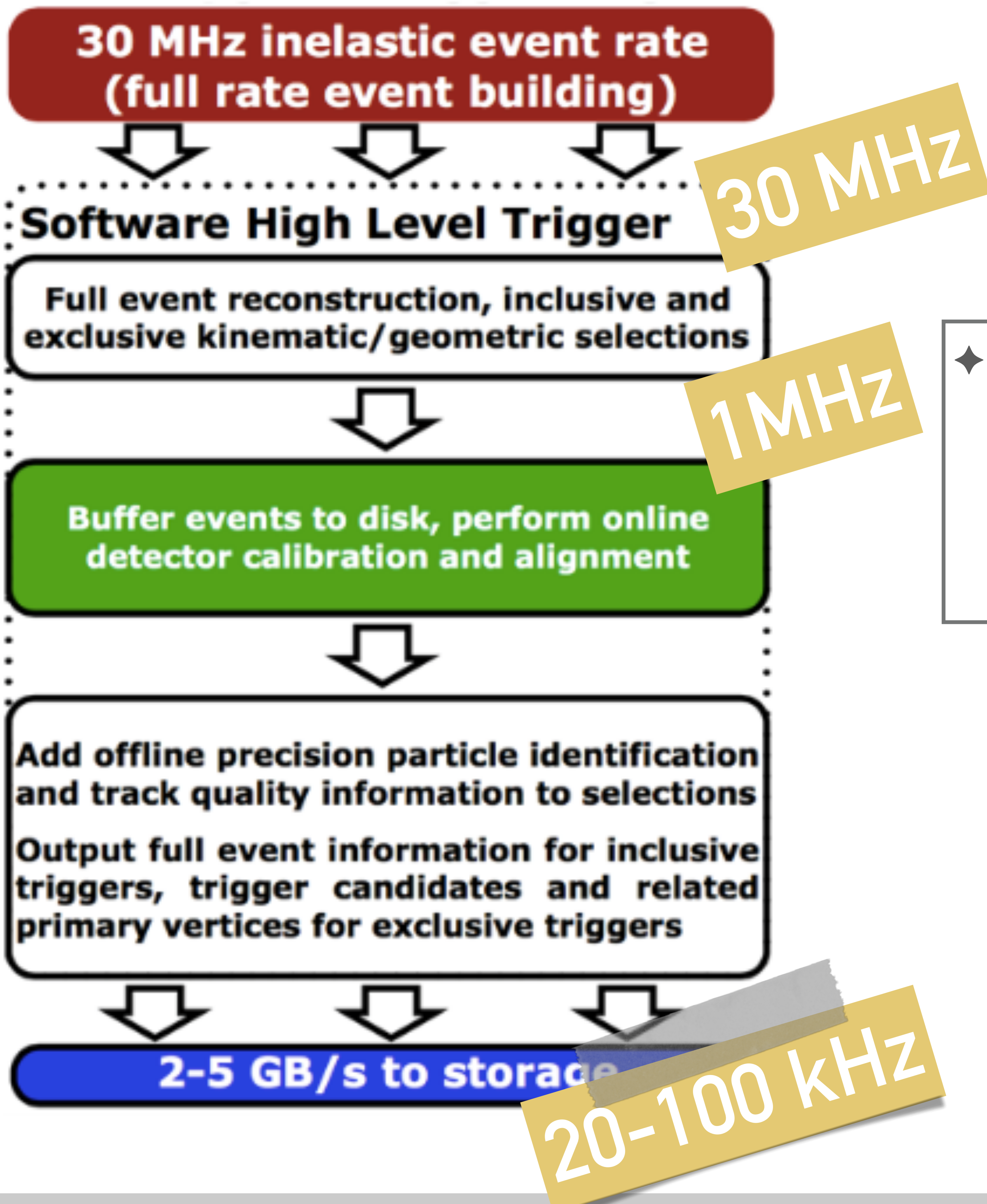
Any further increase in Luminosity for almost constant yield



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

See Phase-I upgrade TDR

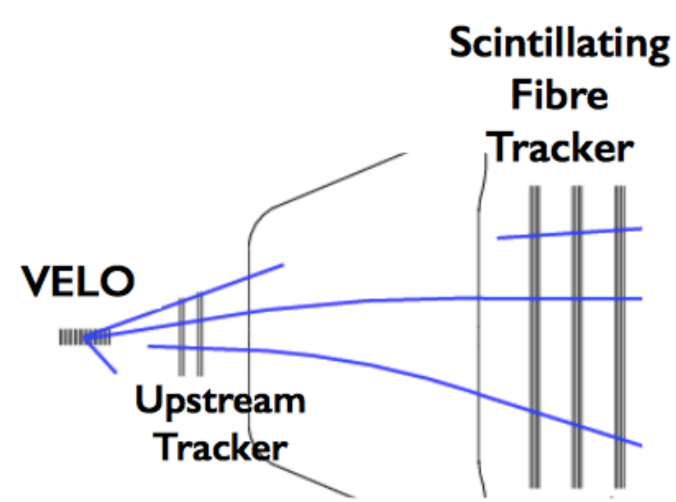
CONTINUOUS READOUT?

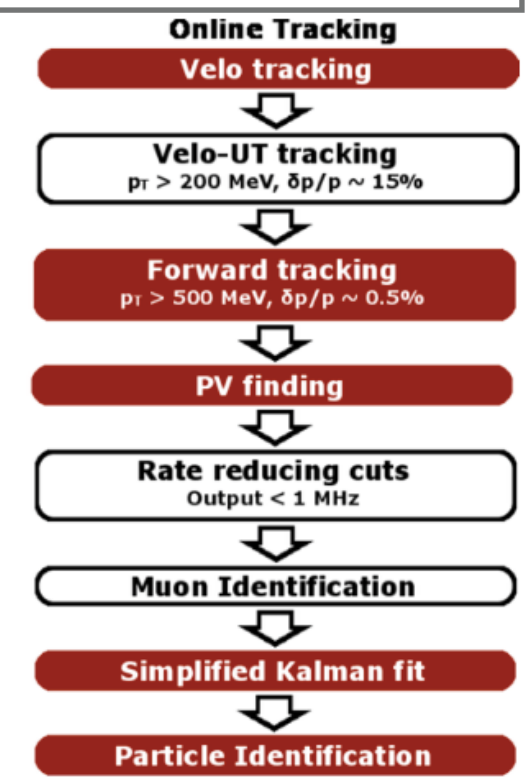


Key strategy: reduce data size and suppress pileup

- FE readout, Event Building and HLT at 30 MHz by design

- Tracking at ~30 MHz ?
  - < 6 ms with current HLT (12 cores + 12 hyper threads + 24 GB RAM) ==> ~ 100k cores!
  - Need to exploit modern CPU architectures & co-processor technologies (FPGA/GPU)



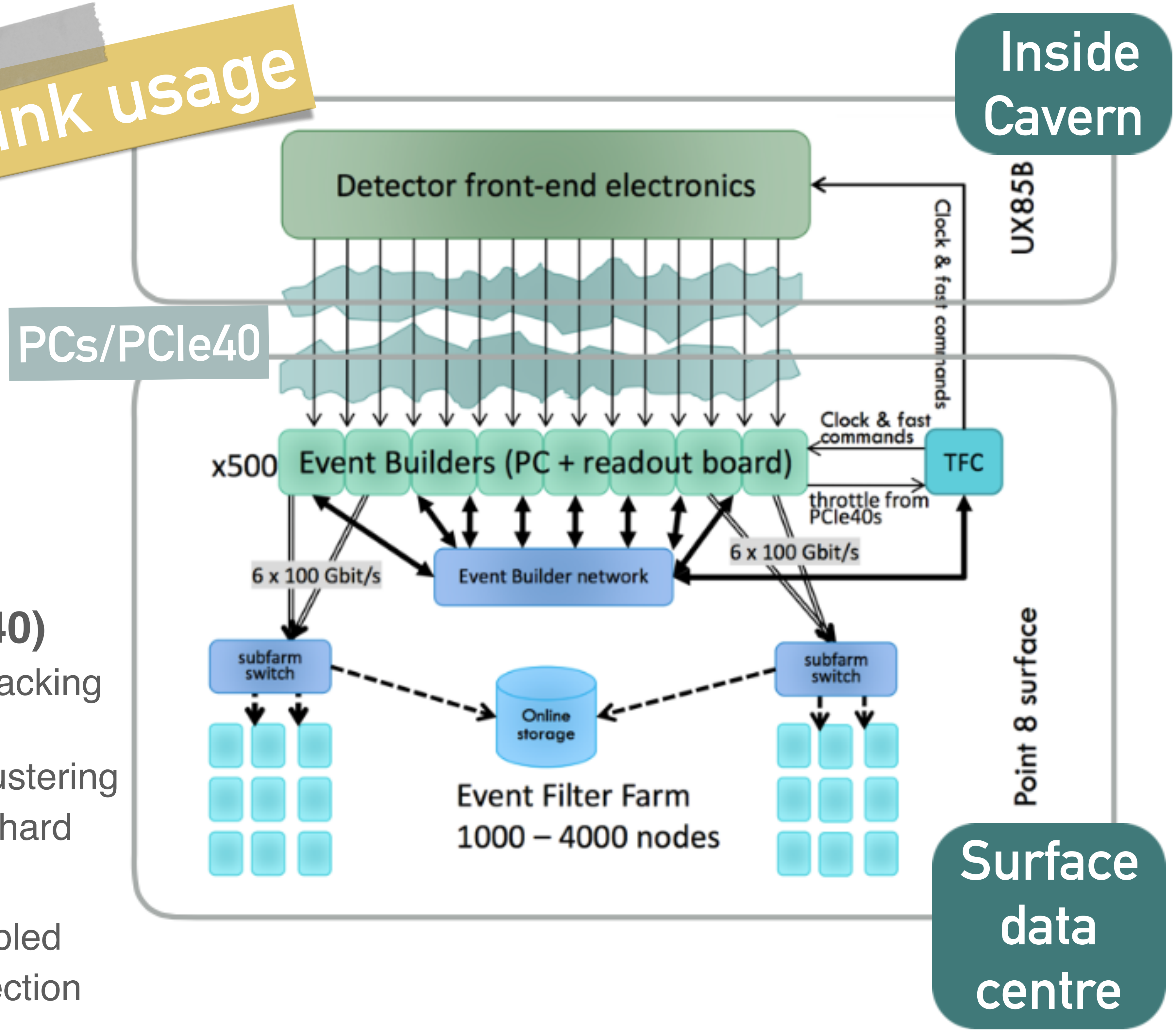


HOW TO LIVE WELL WITHOUT A L1 TRIGGER

Massive link usage

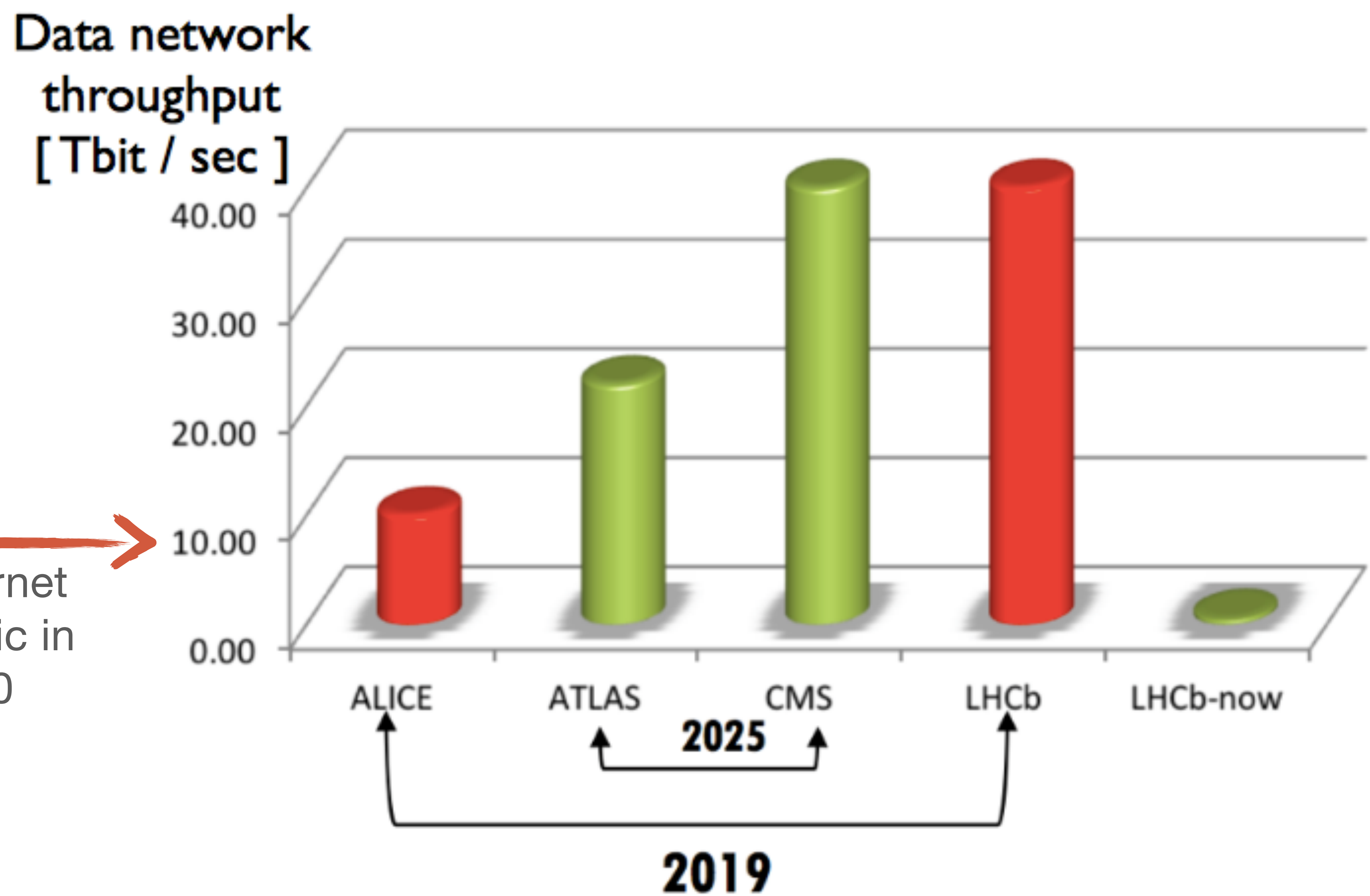
Readout @ 30 MHz
Event size ~ 150kB

DAQ network < 40 Tbit/s
Record: 100 kHz

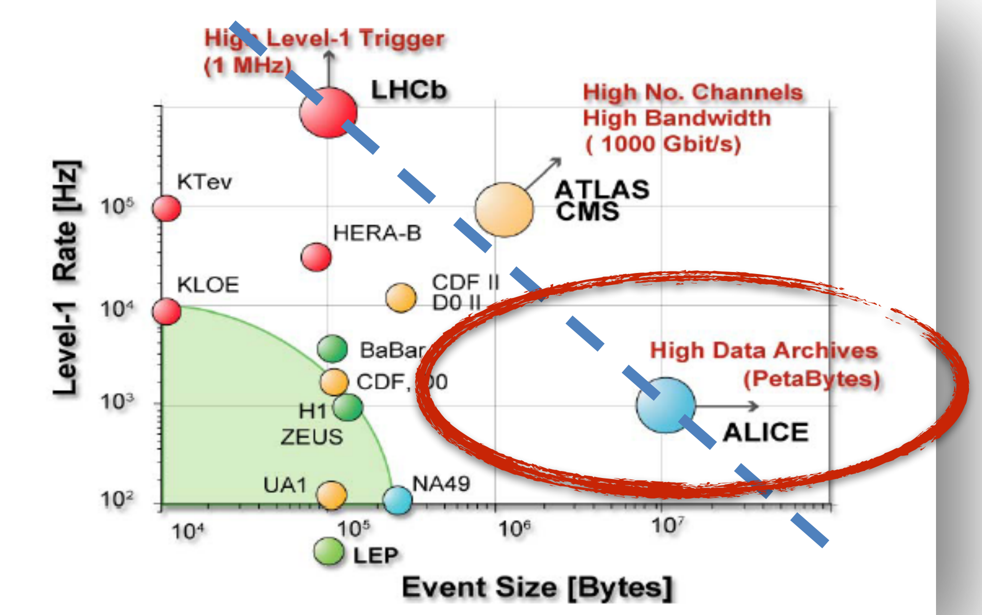


- ➡ **Data reduction in custom readout FPGA-card (PCIe40)**
  - ➡ Each sub detectors with its packing algorithm
  - ➡ i.e.: zero-suppression and clustering
  - ➡ **~10,000 GBT** (4.8 Gb/s) rad-hard
- ➡ **DataFlow:**
  - ➡ EB and HLT networks decoupled
  - ➡ EB with dedicated Data Collection network, in the same card
  - ➡ scalable up to 400 x 100Gbps links

NETWORK TRAFFIC COMPARISON



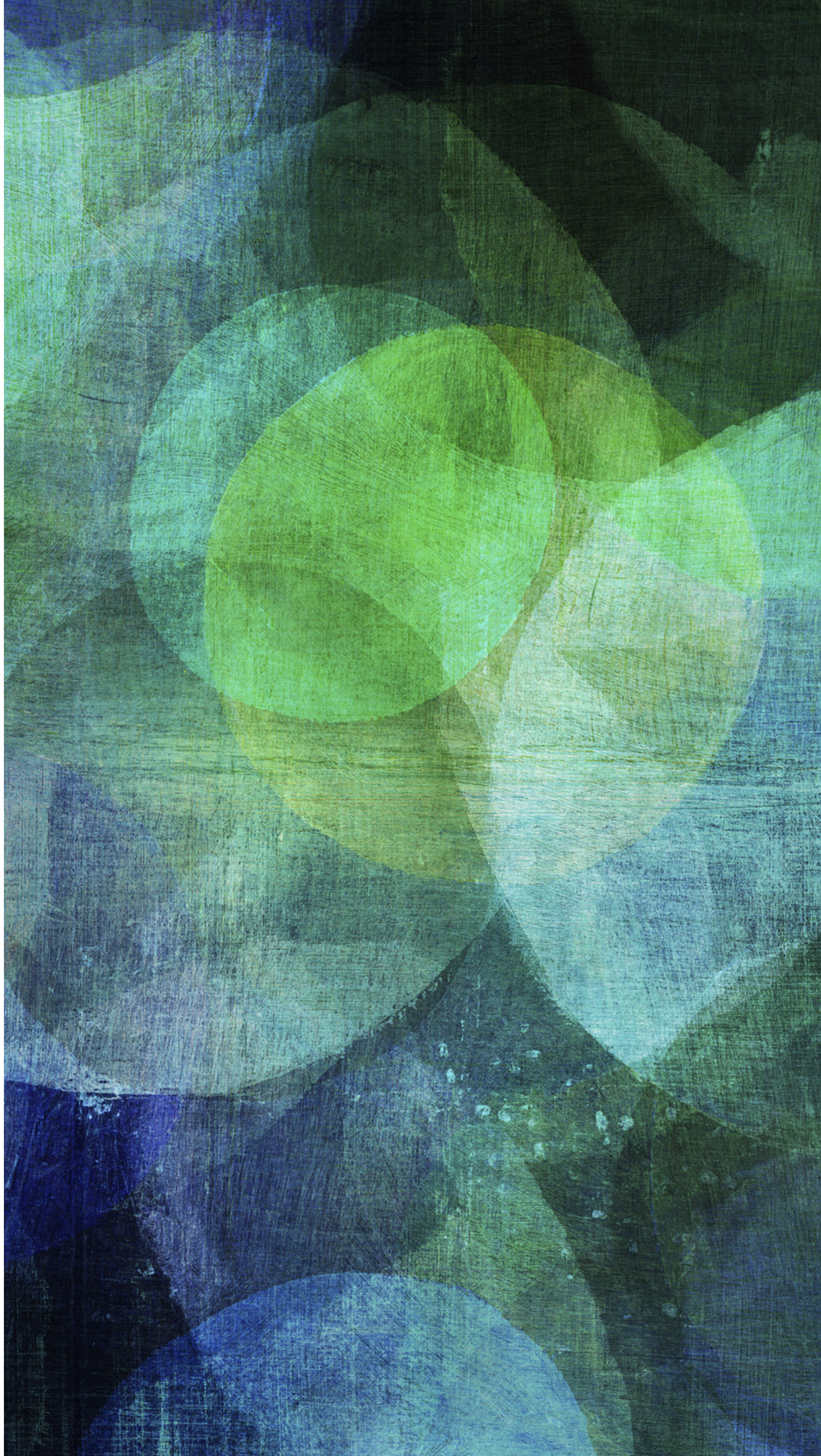
Internet traffic in 2010



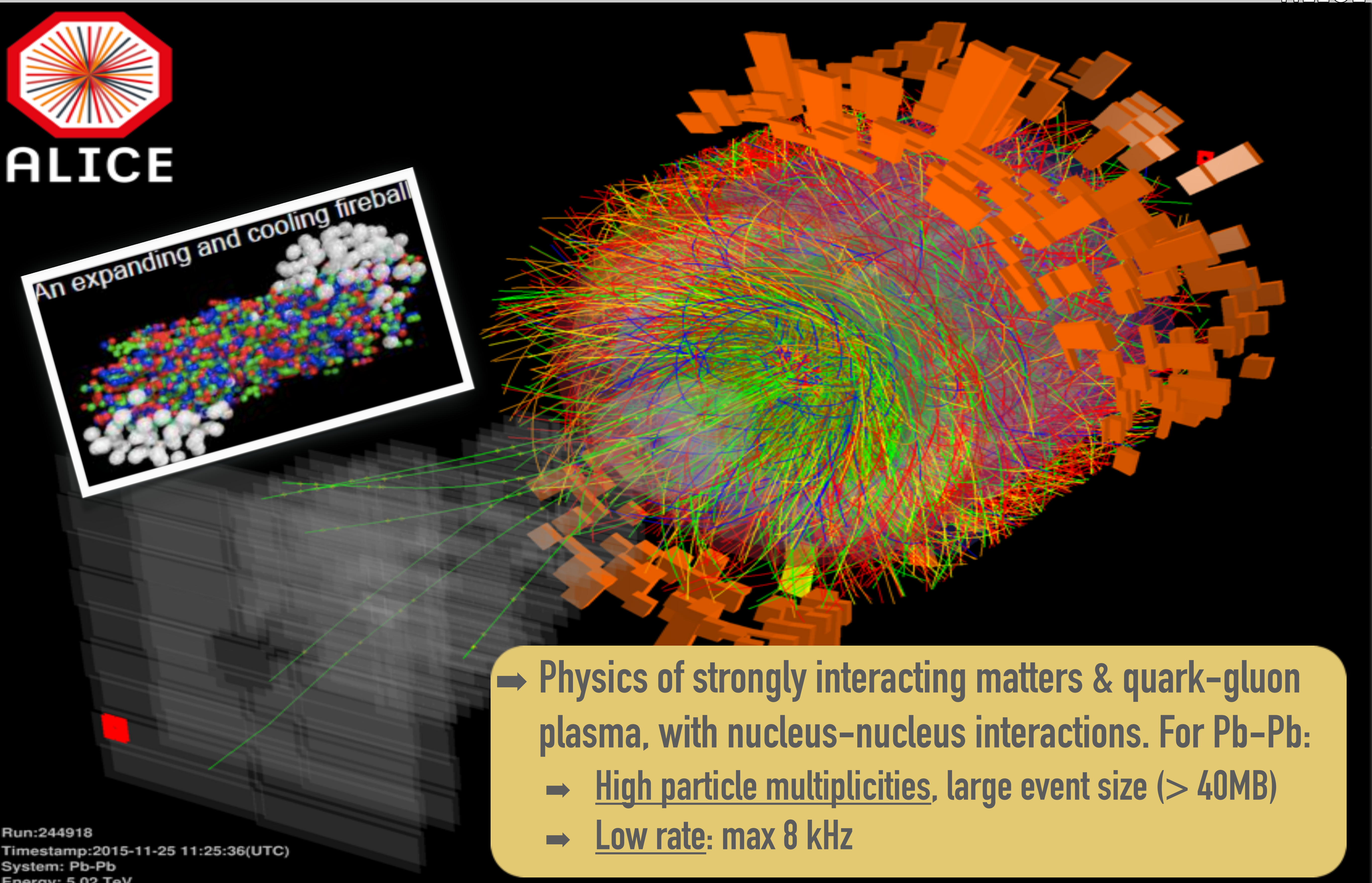
ALICE: THE SMALL BIG-BANG

Recording heavy ion collisions

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



TYPICALLY…

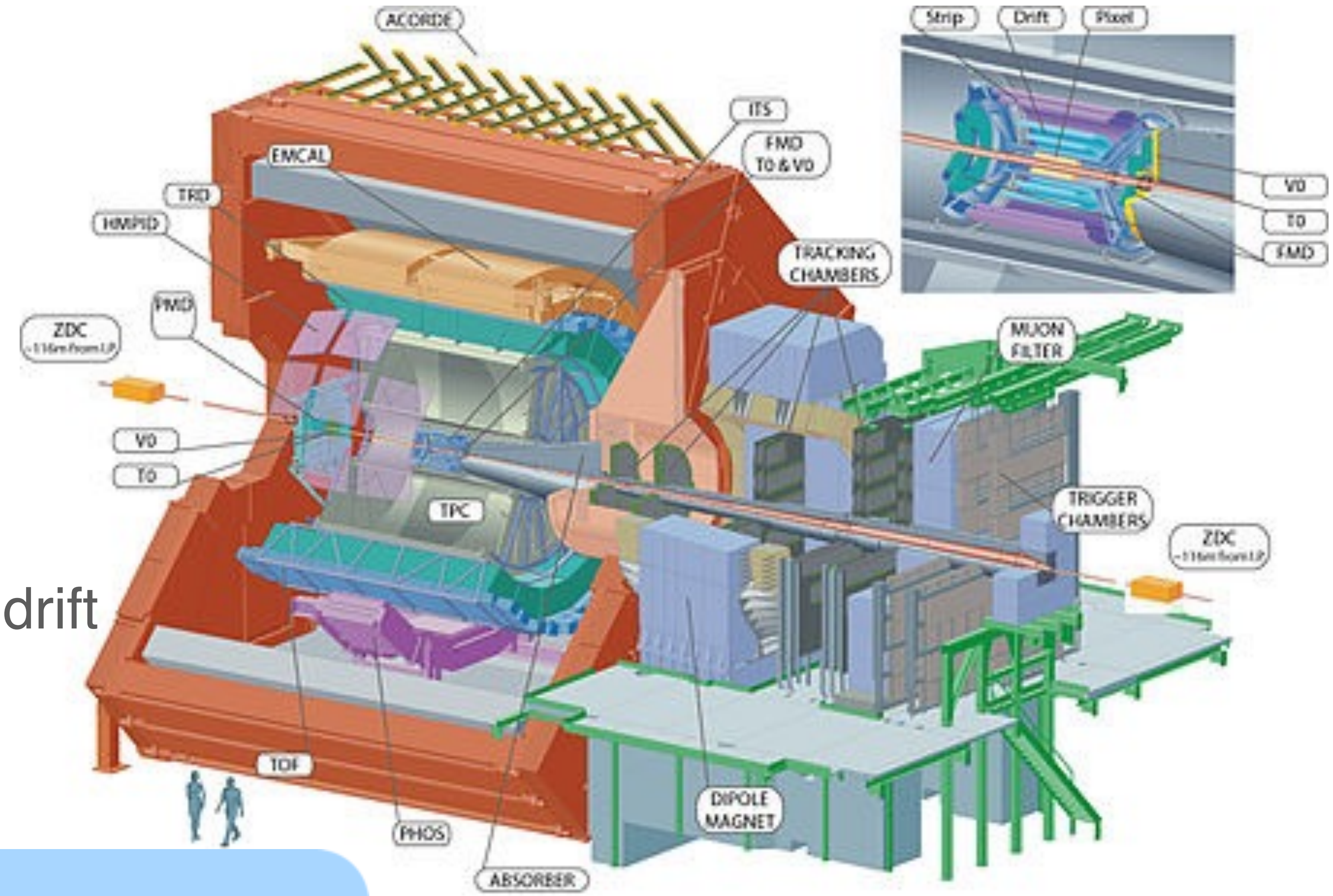


- ➡ Physics of strongly interacting matters & quark-gluon plasma, with nucleus-nucleus interactions. For Pb-Pb:
  - ➡ <u>High particle multiplicities</u>, large event size (> 40MB)
  - ➡ <u>Low rate</u>: max 8 kHz

DESIGNED FOR HEAVY ION COLLISIONS

➡ **Strategy**: identify short-living particles (hyperons) **through low-p_T tracks** (>100MeV)

- ➡ **19 different detectors**
  - ➡ (~8000 particles/dη)
- ➡ **slow but high-granularity detectors**: TPC and silicon drift
  - ➡ with low rate readout rate

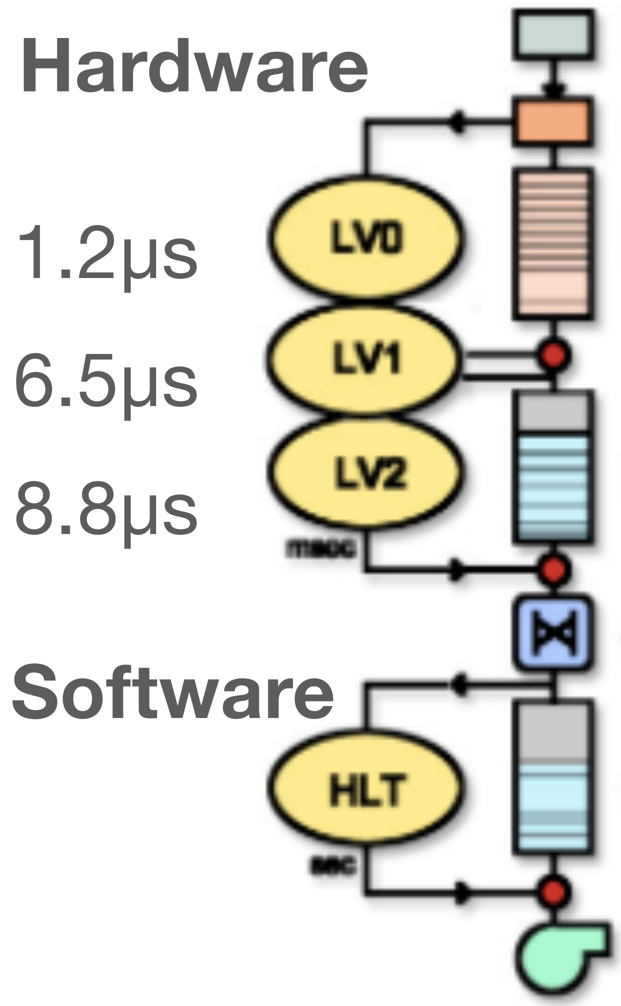


➡ Challenges for DAQ design:

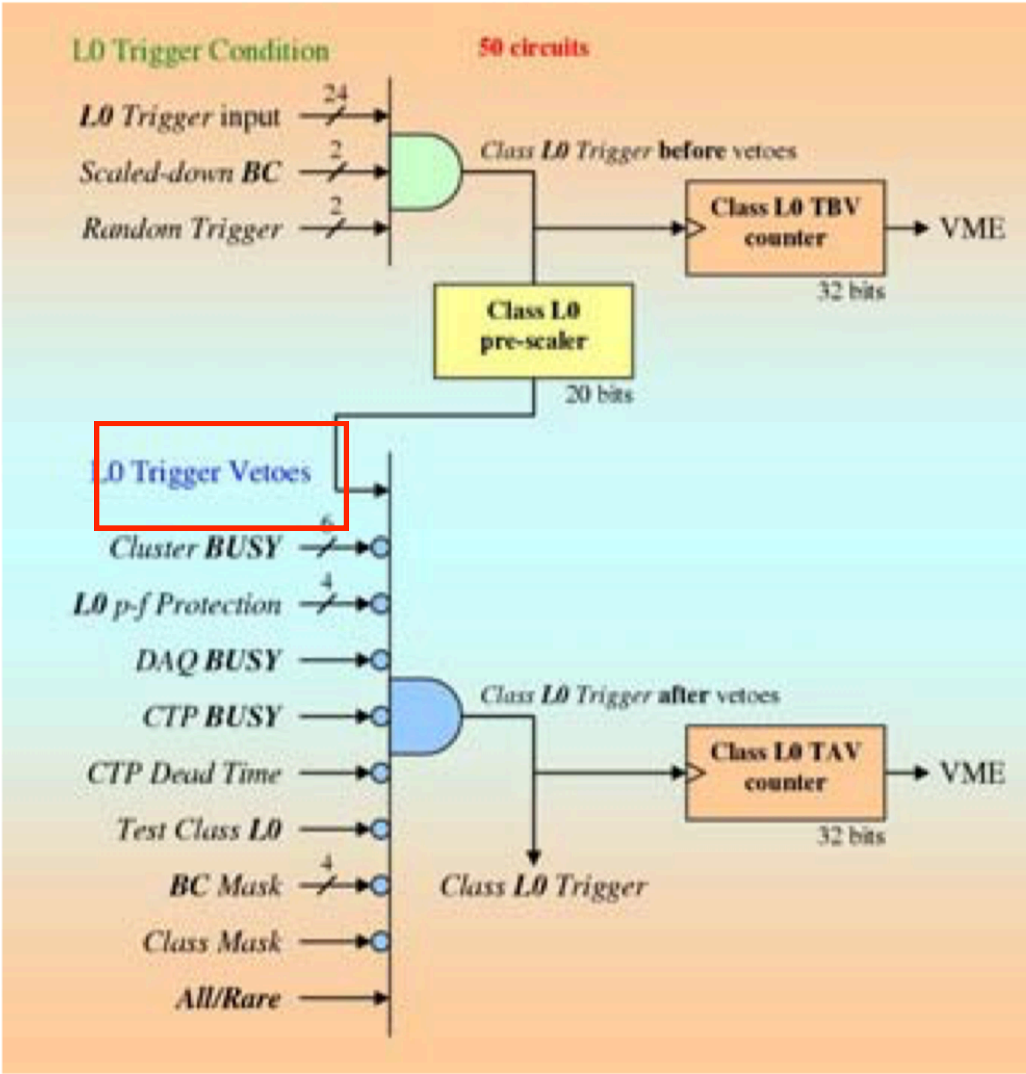
- ➡ detector readout: up to ~50 GB/s
  - ➡ TPC producing 90% of data
- ➡ storage: for Pb–Pb 1.2TB/s (pp: 100 MB/s)

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

A 4-LEVEL TRIGGER FOR HIGH OCCUPANCY EVENTS



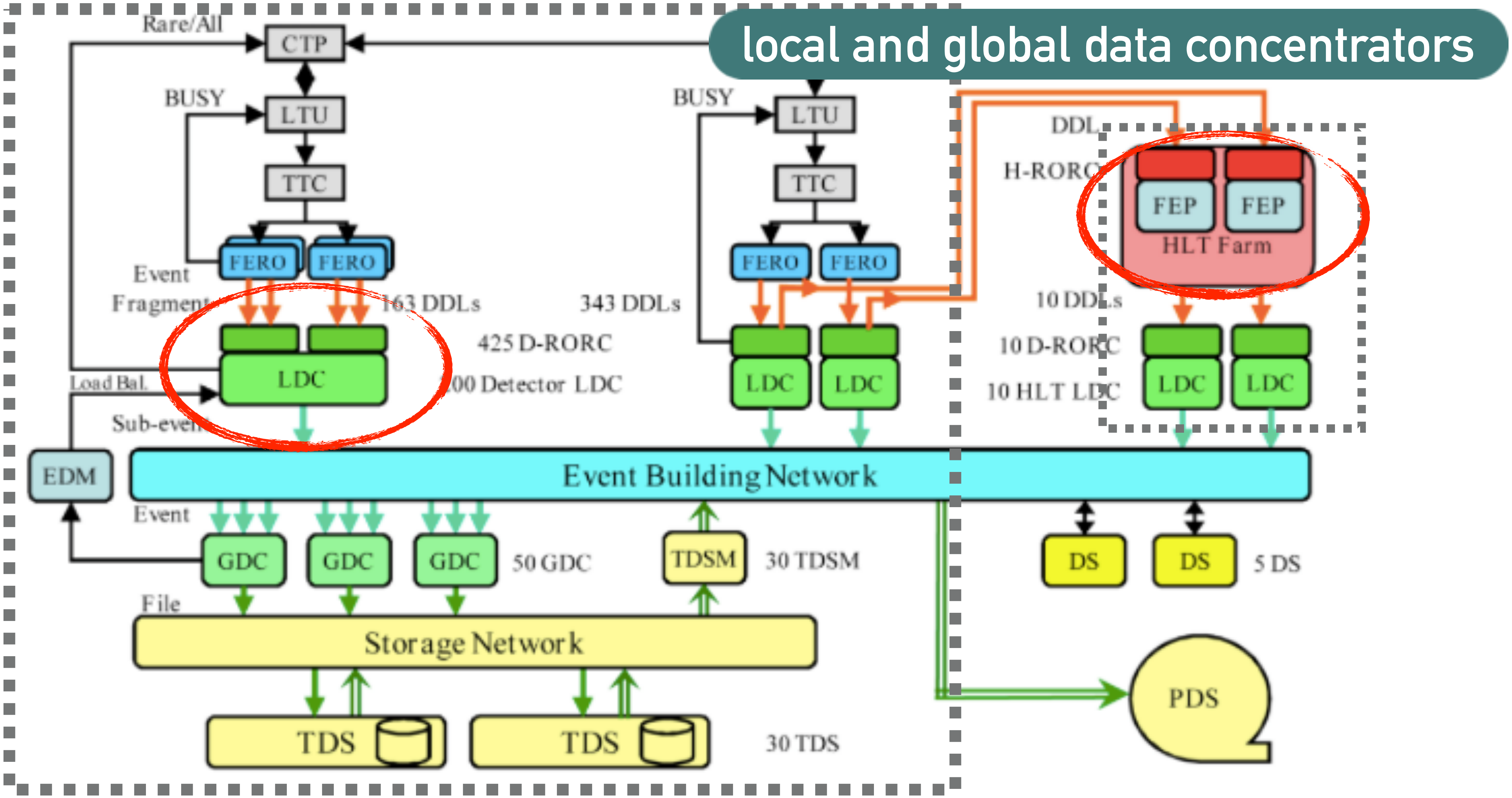
- **Detectors with different latencies for readout/ signal**
  - TPC ~ 100μs, but some need early probe < 1.2 μs
- **Trigger strategy for high occupancy events**
  - Search for topologies
  - Each detector into global decision, without geometrical match



total 60 inputs: 24 L0; 24 L1; 12 L2

- **Special trigger features to avoid deadtime (using Veto-logic)**
  - Dynamic readout (read what is needed)
  - Past-future protection (avoid pile-up for TPC)
  - Rare trigger handling (when DAQ buffers ~full, restrict the global trigger conditions)
- **Multitude of signals: large configuration system and safe error handling**

DAQ/HLT ARCHITECTURE IN RUN1 AND RUN2

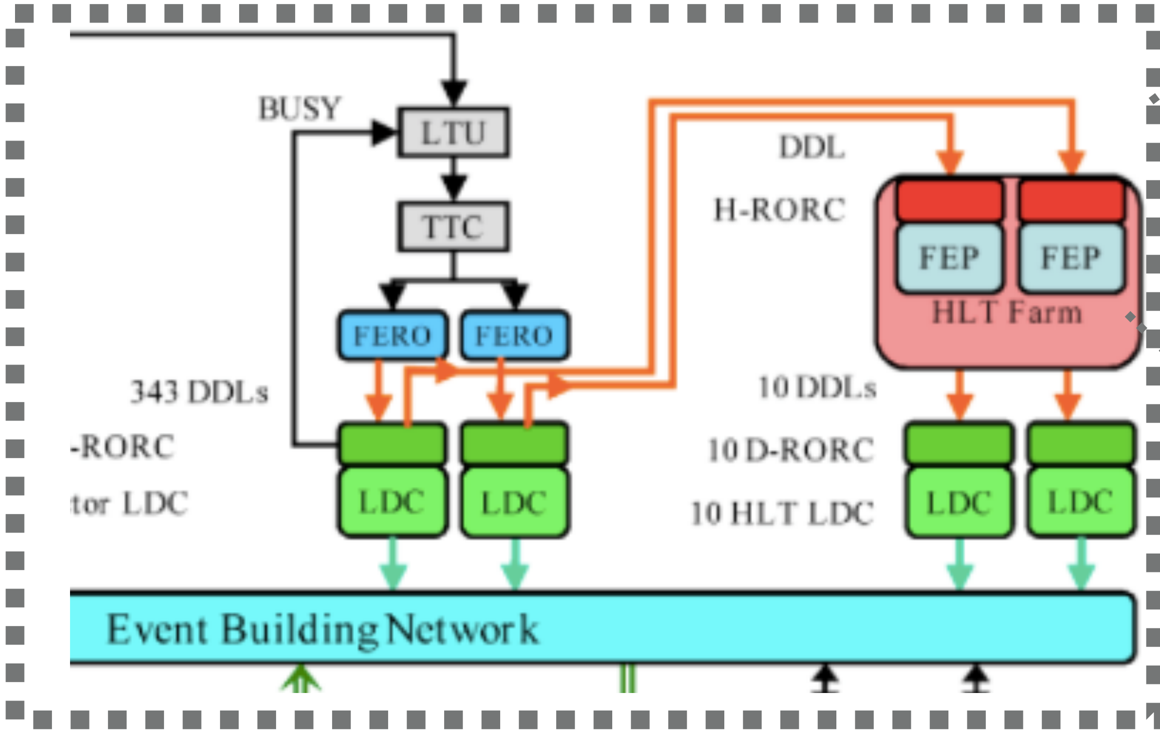


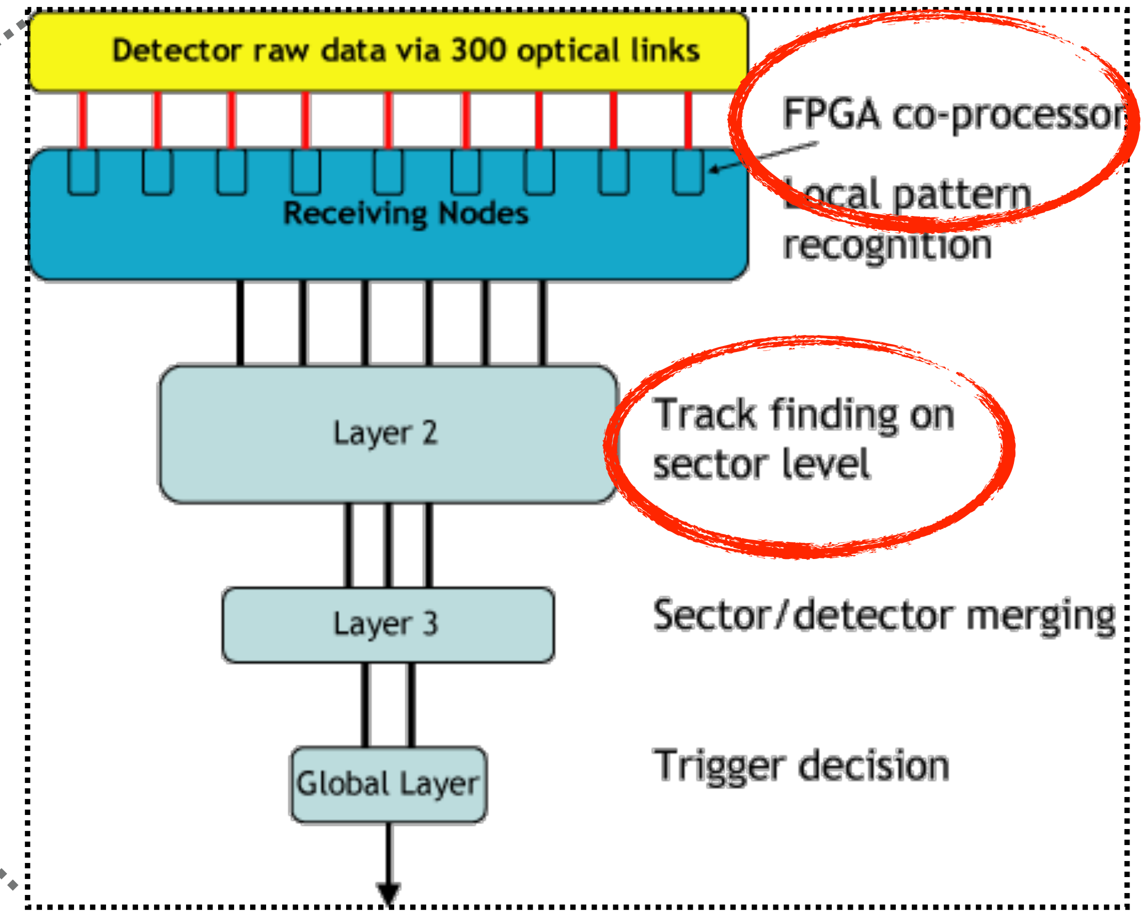
- **Total traffic from detector FE: ~20 GB/s**
  - **400 DDL** point-to-point optical links to **RORC** (6Gbps) directly into PC memory at 200 MB/s (DMA)
- **HLT and DAQ decoupled (EB not waiting for HLT decision)**
  - HLT as any other sub-detector in DAQ

SOFTWARE TRIGGER ARCHITECTURE

HLT reduces ~20 GB/s ⇒ ~4 GB/s

Challenge: large data, decision in few 100ms



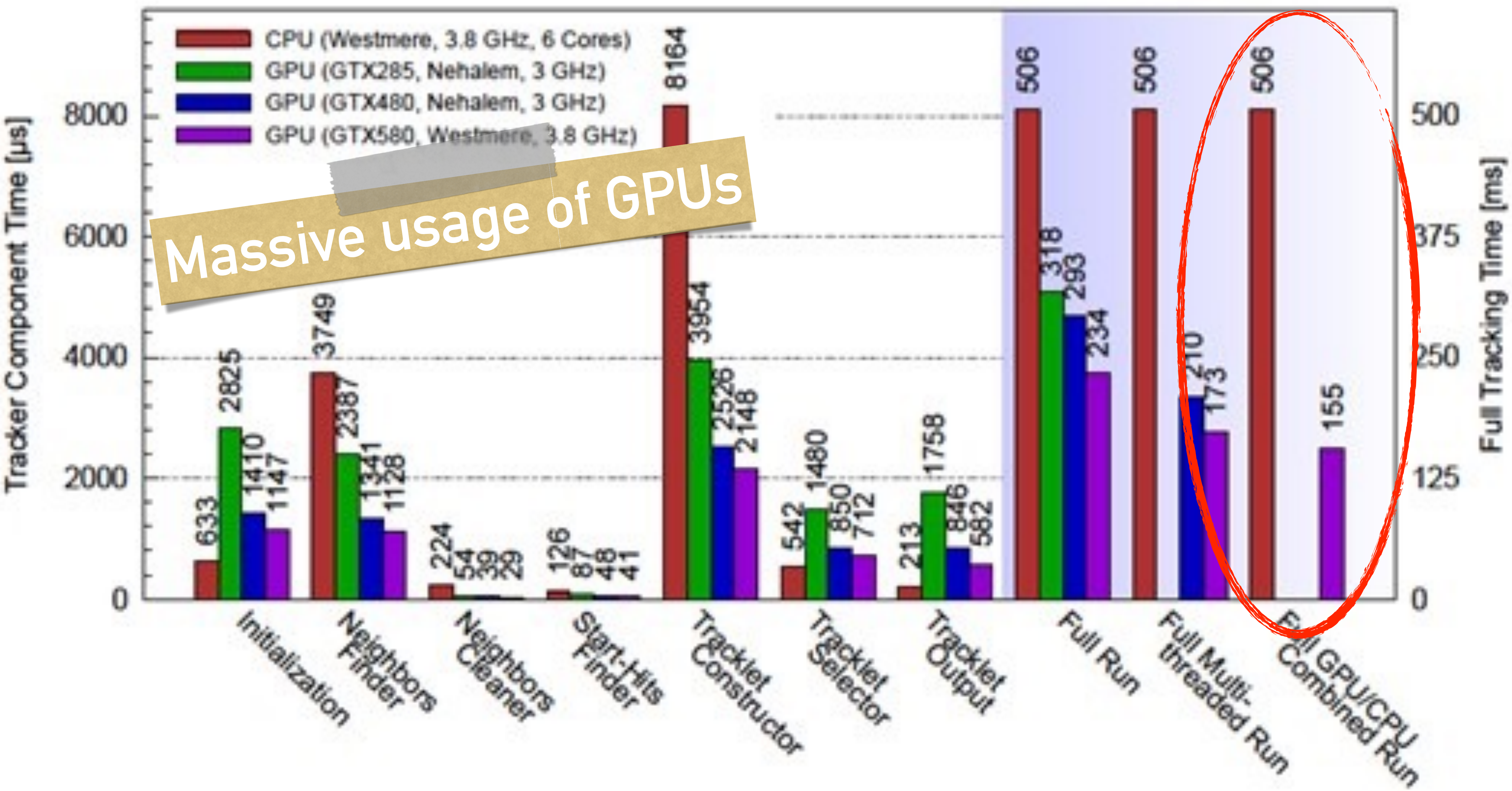


	Data Format	Data Reduction Factor	Event Size (MByte)
	Raw Data	1	700
FEE	Zero Suppression	35	20
HLT	Clustering & Compression	5-7	~3
	Remove clusters not associated to relevant tracks	2	1.5
	Data format optimization	2-3	<1

- ➡ **Local reconstruction & compression**
  - ➡ **FPGA for advanced TPC data compression and cluster-finding** (factor x4 reduction)
  - ➡ **GPU for tracking**: cellular automaton/ Kalman filter algorithms

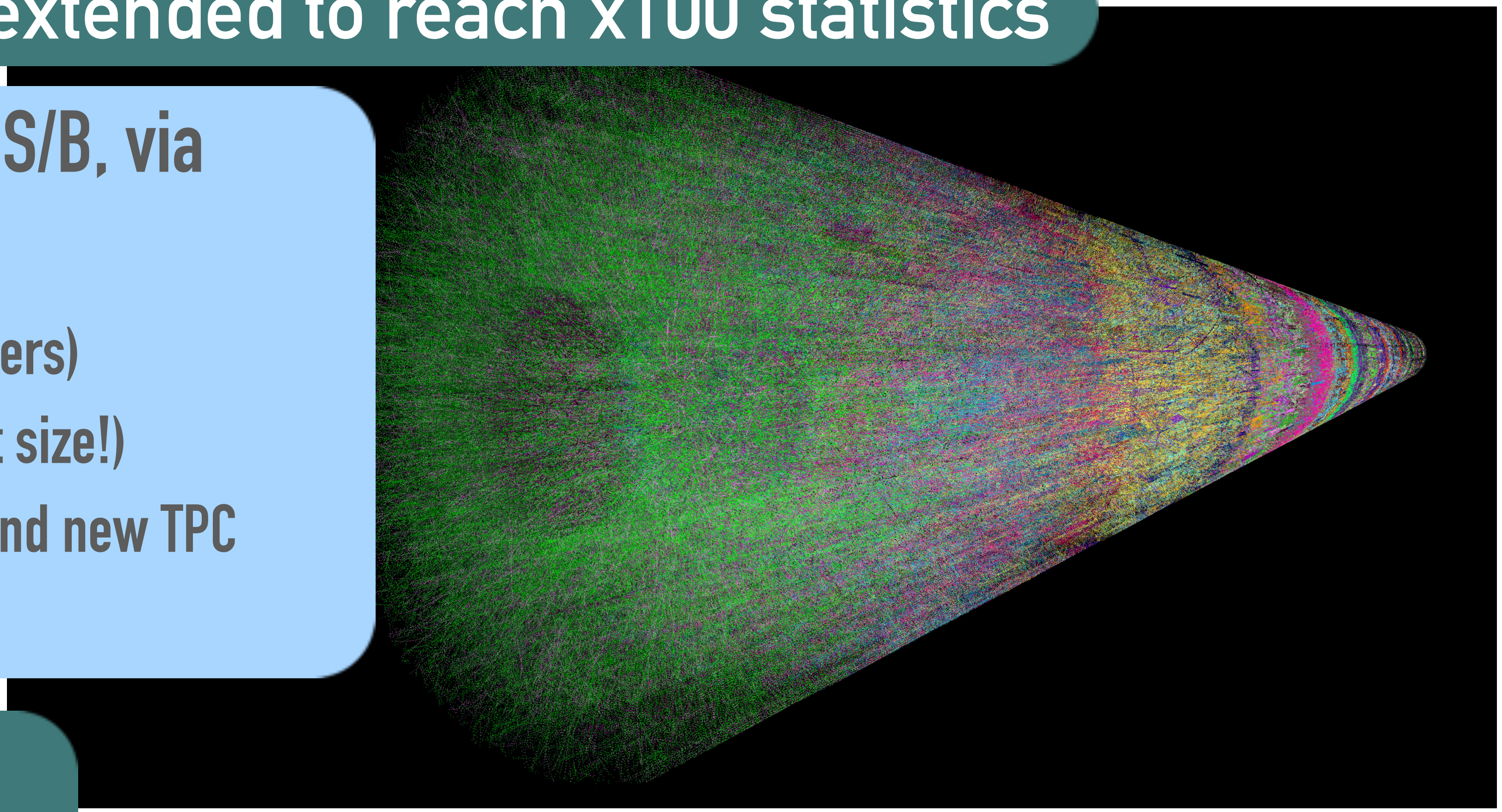
RUN2 TRIGGER

- ➡ Readout rate x2 for TPC and TRD (thanks to compression)
- ➡ Increase DAQ throughput (thanks to COTS): 2.5GB/s (2010) ⇒ 6GB/s (2015)



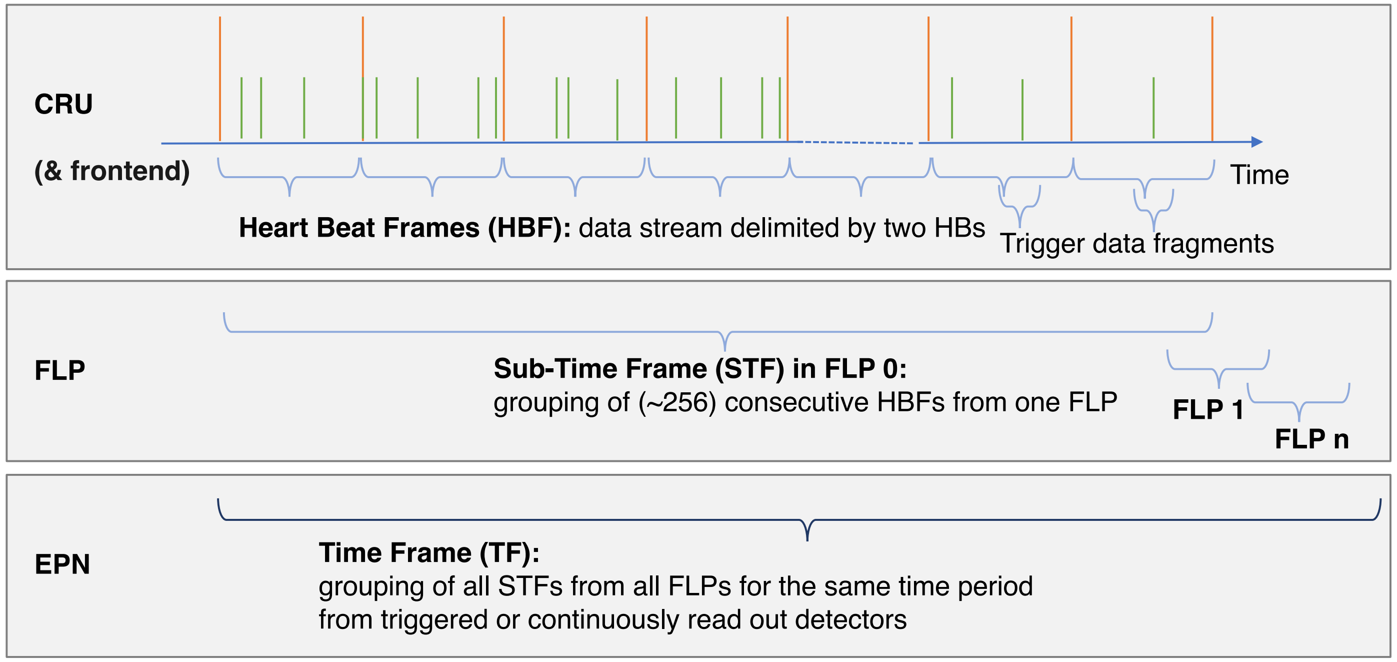
LHC heavy ion programme extended to reach x100 statistics

- ➡ Access rare physics for with low S/B, via complex probes at low p_T
  - ➡ Increase vertex/tracking (-> new trackers)
  - ➡ Increase detector granularity (-> event size!)
  - ➡ Higher readout rates: new electronics and new TPC readout with GEM (up to 50kHz)



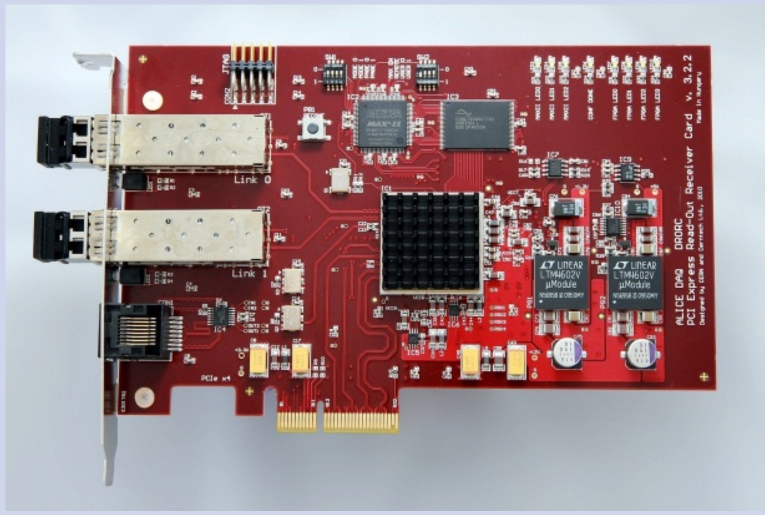
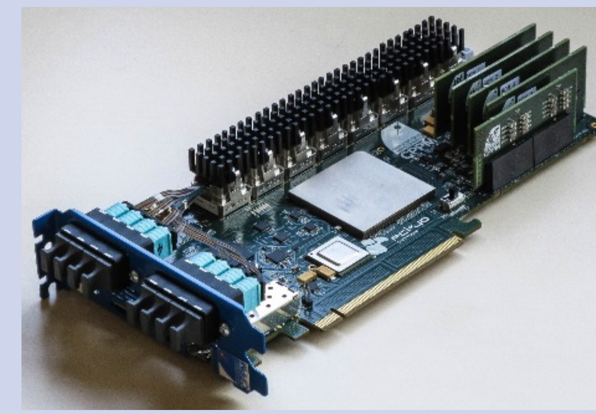
To maintain acceptance, overcome classical trigger concepts

- ➡ **Trigger-less continuous read-out**
  - ➡ Triggering techniques very inefficient if not impossible in most cases
- ➡ **Heart Beat (HB): issued in continuous & triggered modes to all detectors**
  - ➡ 1 per orbit, 89.4 μs: ~10 kHz
  - ➡ based on Time-framing: 1 every ~20 ms: **~50 Hz** (1 TF = ~256 HBF)



ALICE READOUT EVOLUTION

~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

RUN3 DAQ STRATEGY: ONLINE RECONSTRUCTION

ALICE

~3TB/s detector readout
Storage bandwidth x O(100)
Offline reconstruction also challenging

Higher rates with smaller data

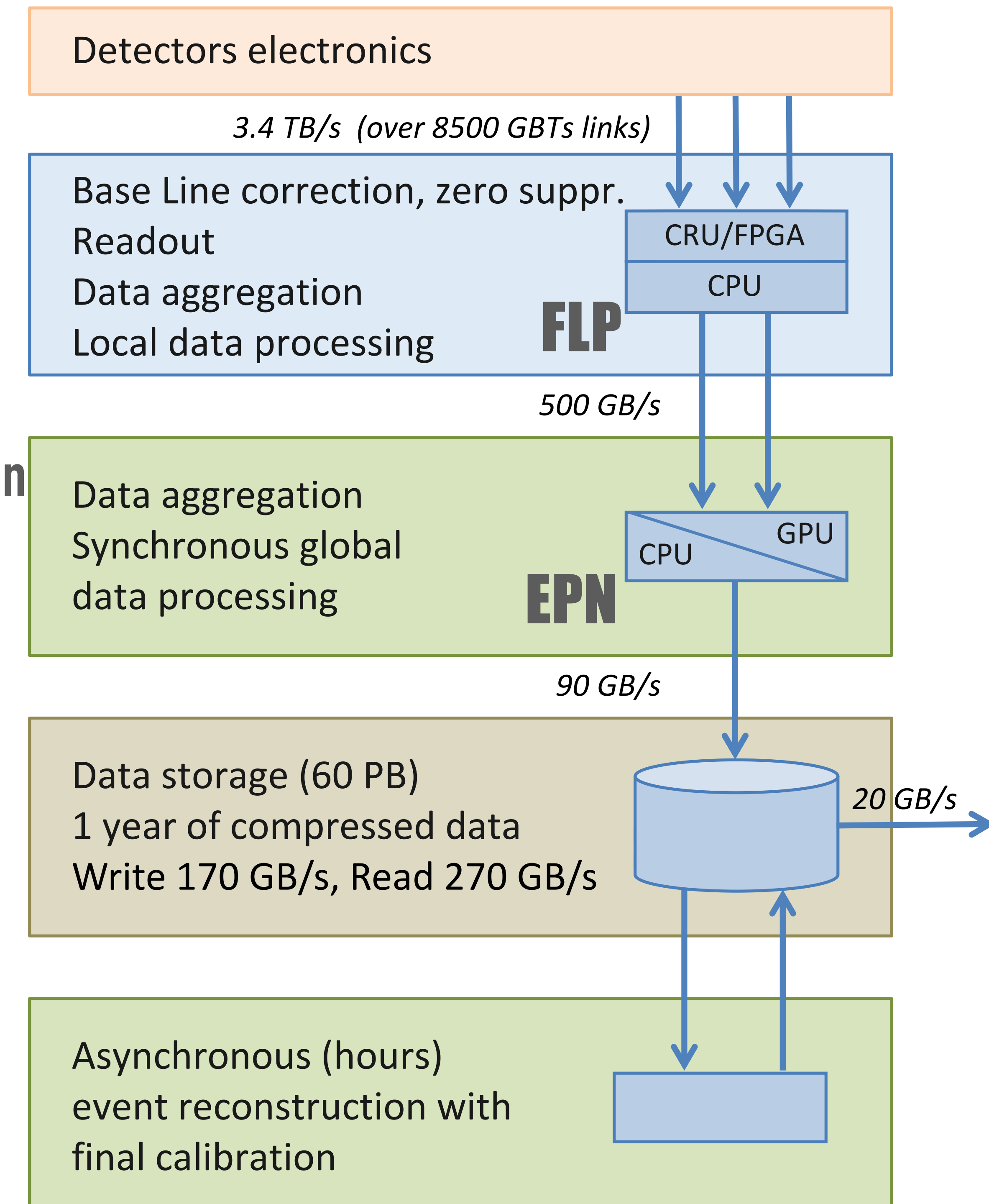
- **Very heterogeneous system**
- **Data compression in FPGA/CPU**
  - 270 First level processors (FLP)
- **More data aggregation forming tracks in GPUs**
  - 1500 Event Processing Nodes (EPN)
- **Store only reconstruction results, discard raw data**
  - 100% trust software?
- **Much tighter coupling between online and offline reconstruction software (O²) sharing:**
  - calibration constants
  - resources

O² system

Data reduction
Calibration 0

Data aggregation
Reconstruction
Calibration 1

More reconstruction
Calibration 2



SUMMARY OF SUMMARIES

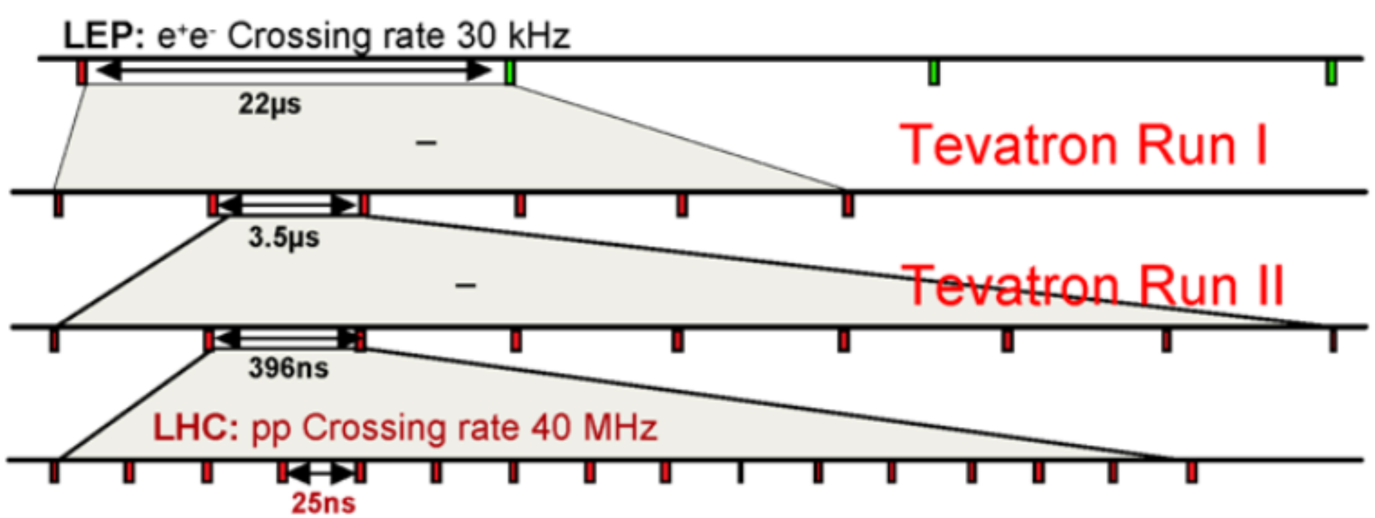
- ➡ **Among the largest and most complex TDAQ systems have to cope with current and future LHC Luminosity**
- ➡ **Scalability not obvious, may need some breakthrough in technology**
  - ➡ Moore's law still valid for processors but needs more effort to be exploited
  - ➡ Hopefully tick-tock model can be extended for the future
- ➡ **All LHC experiments break the limits of their design and are upgrading (between 2019-2024)**
  - ➡ **ATLAS/CMS** drives high rate readout and Event Building, still based on robust trigger selections
  - ➡ **LHCb** pioneer online-offline merging with large data throughputs
  - ➡ **ALICE** drives the GPU evolution and data compression
- ➡ **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

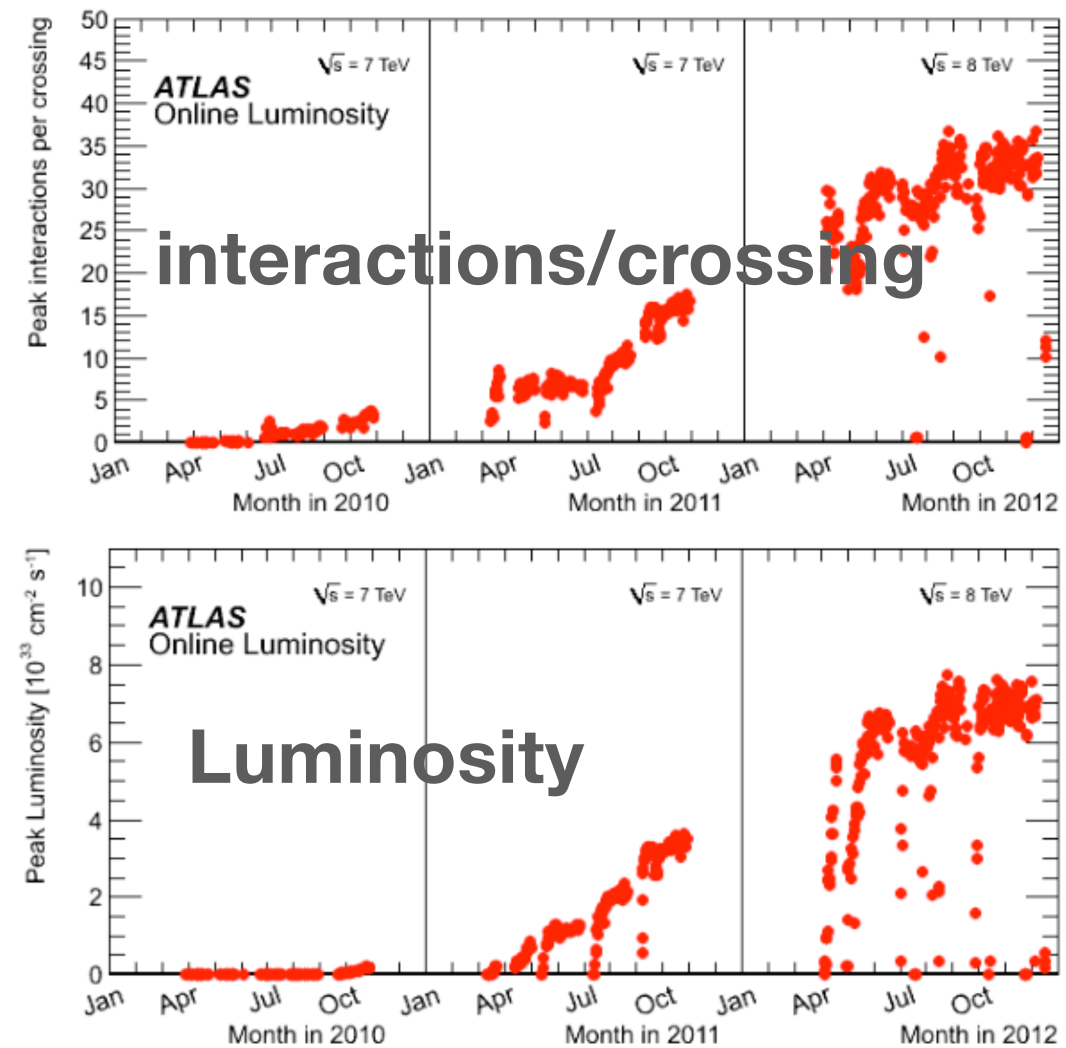
- ~3600 bunches in 27km
- distance bw bunches: 27km/3600 = 7.5m
- distance bw bunches in time: 7.5m/c = 25ns



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

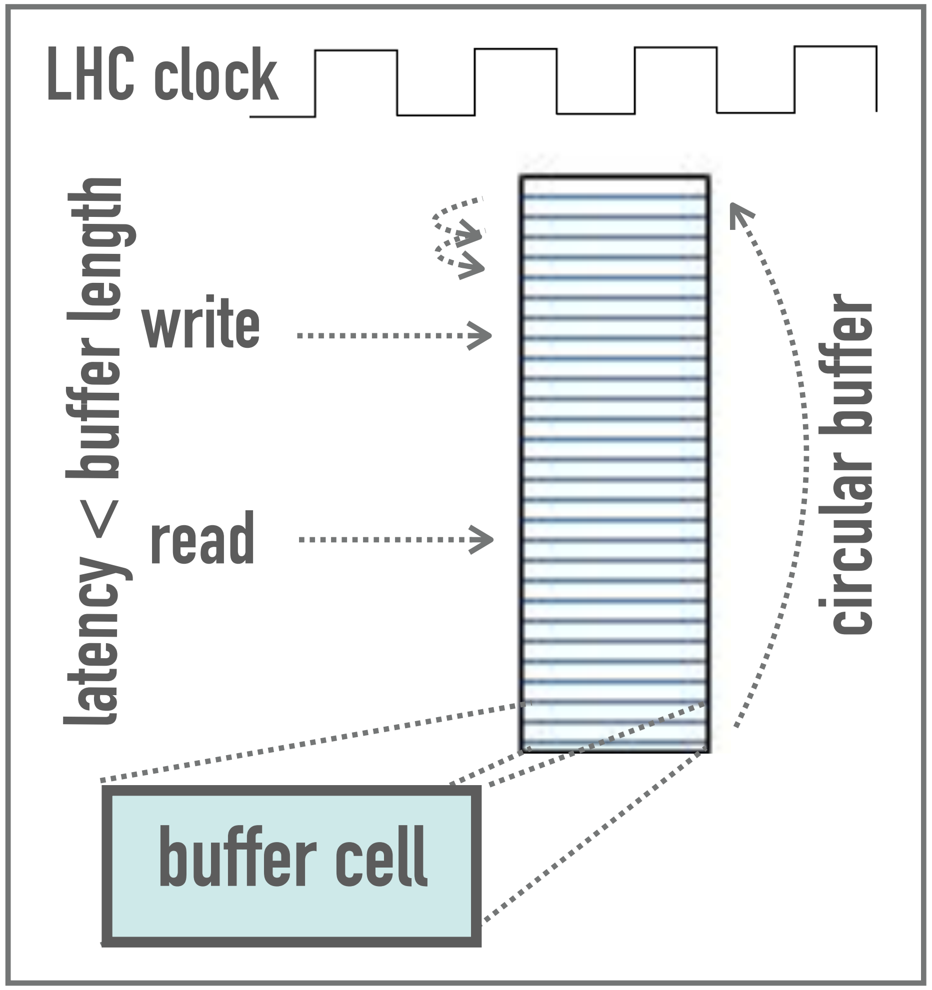
The pile-up source

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



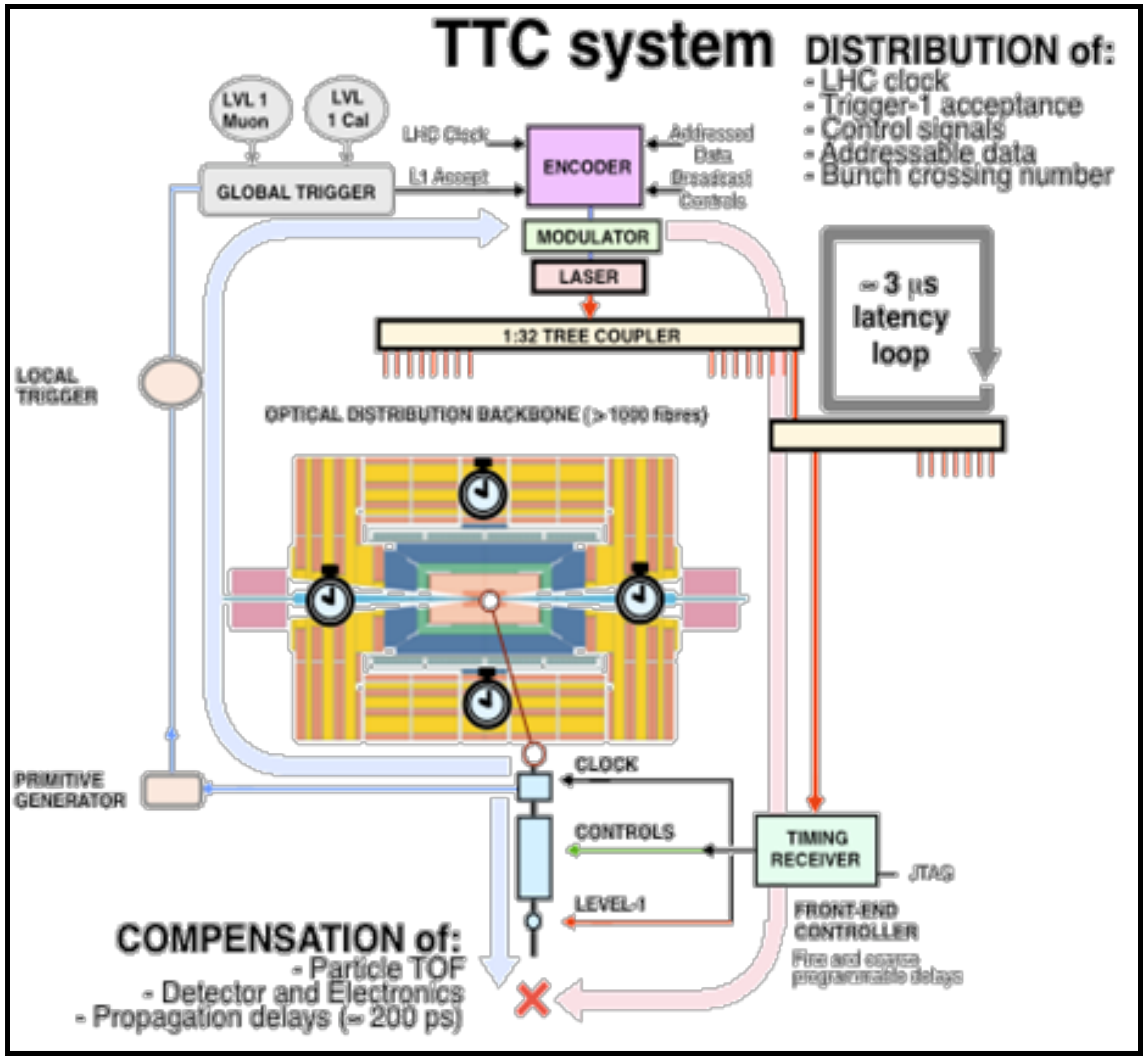
PIPELINED TRIGGERS

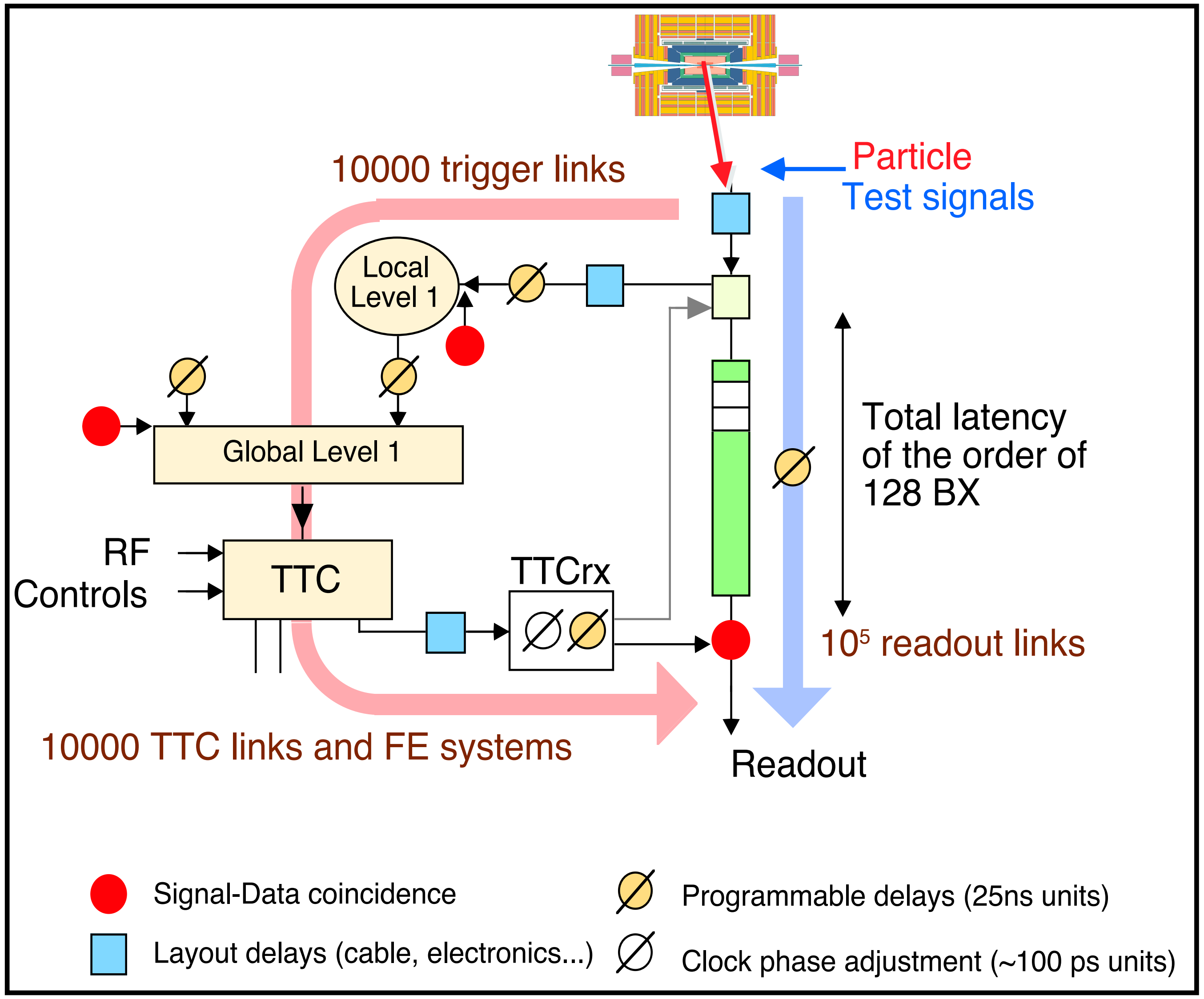
- **Allow trigger decision longer than clock tick (and no deadtime)**
  - Execute trigger selection in defined clocked steps (**fixed latency**)
  - Intermediate storage in stacked buffer cells
  - R/W pointers are moved by clock frequency



- **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



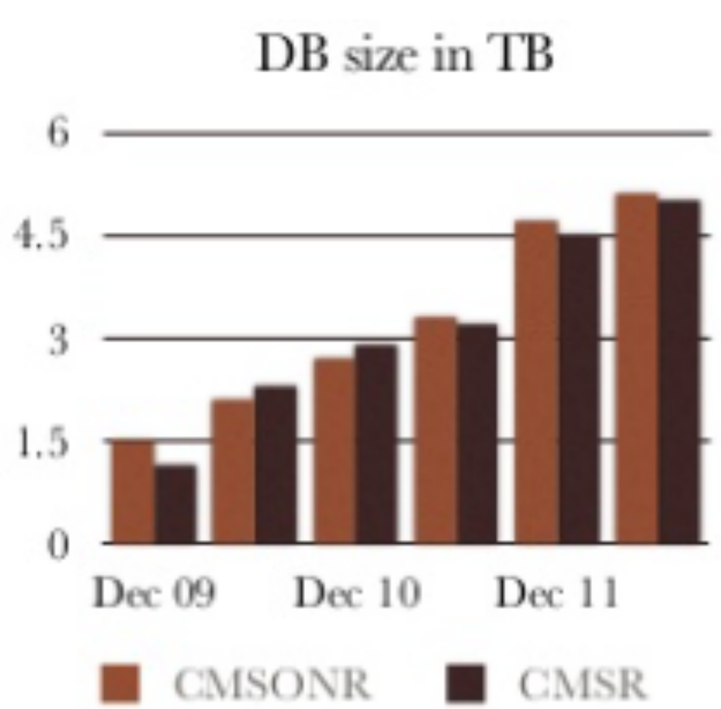


- ➡ **Common optical system: TTC**
  - ➡ radiation resistance
  - ➡ single high power laser
- ➡ **Large distribution**
  - ➡ experiments with ~10^7 channels

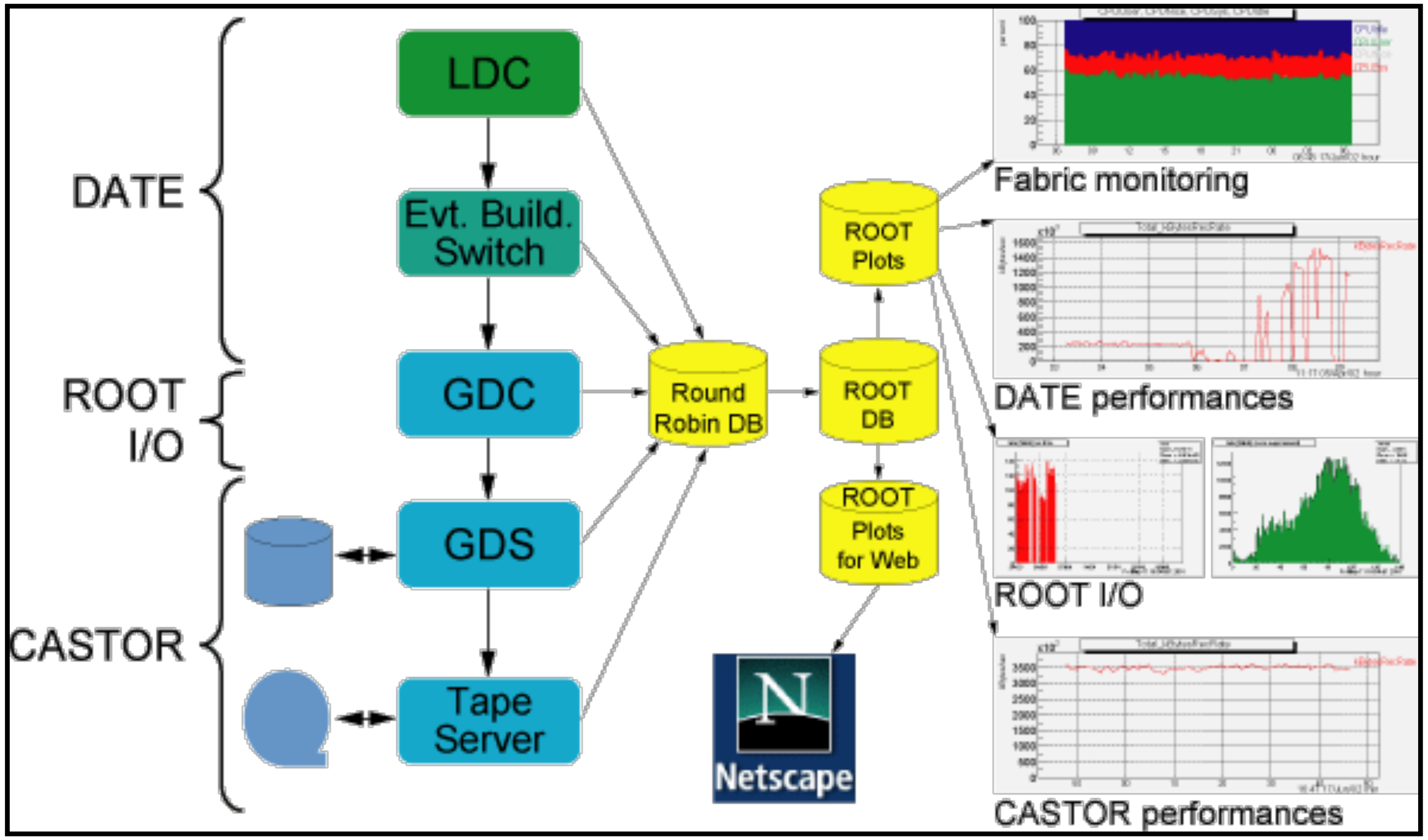
- ➡ **Align readout & trigger at (better than) 25ns and correct for**
  - ➡ time of flight (25 ns ≈ 7.5m)
  - ➡ cable delays (10cm/ns)
  - ➡ processing delays (~100 BCs)

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

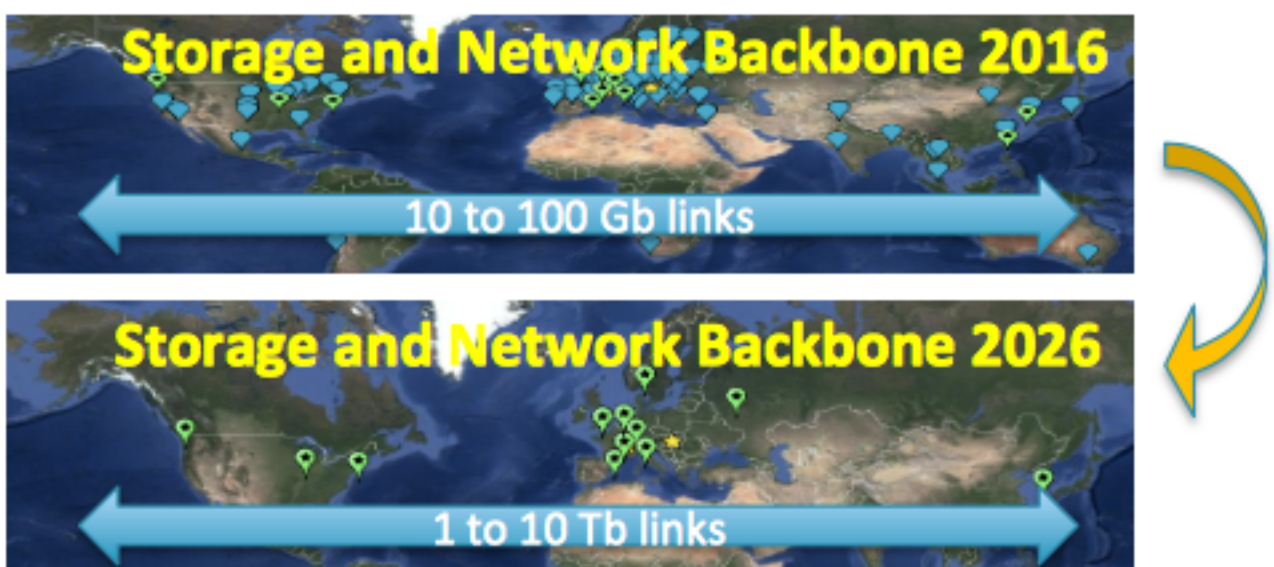


CMS DB grows about 1.5TB/year, condition data only a small fraction

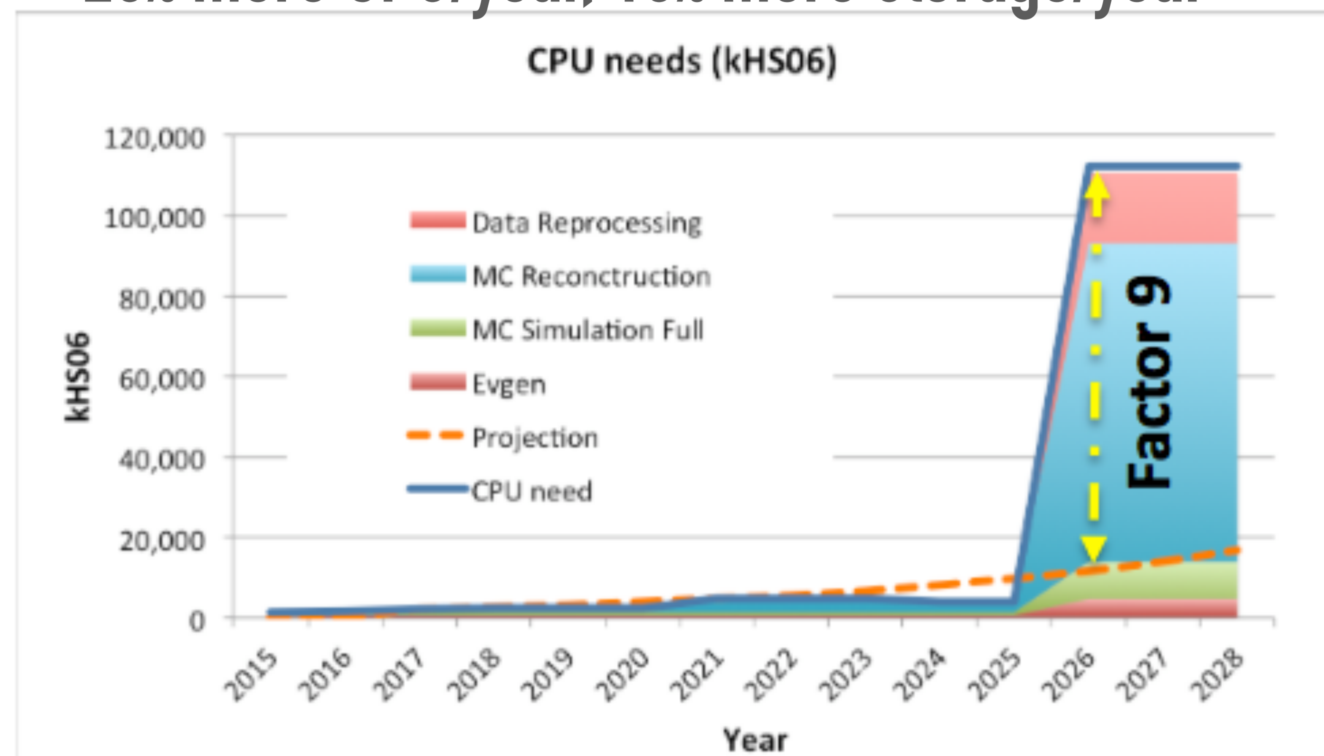


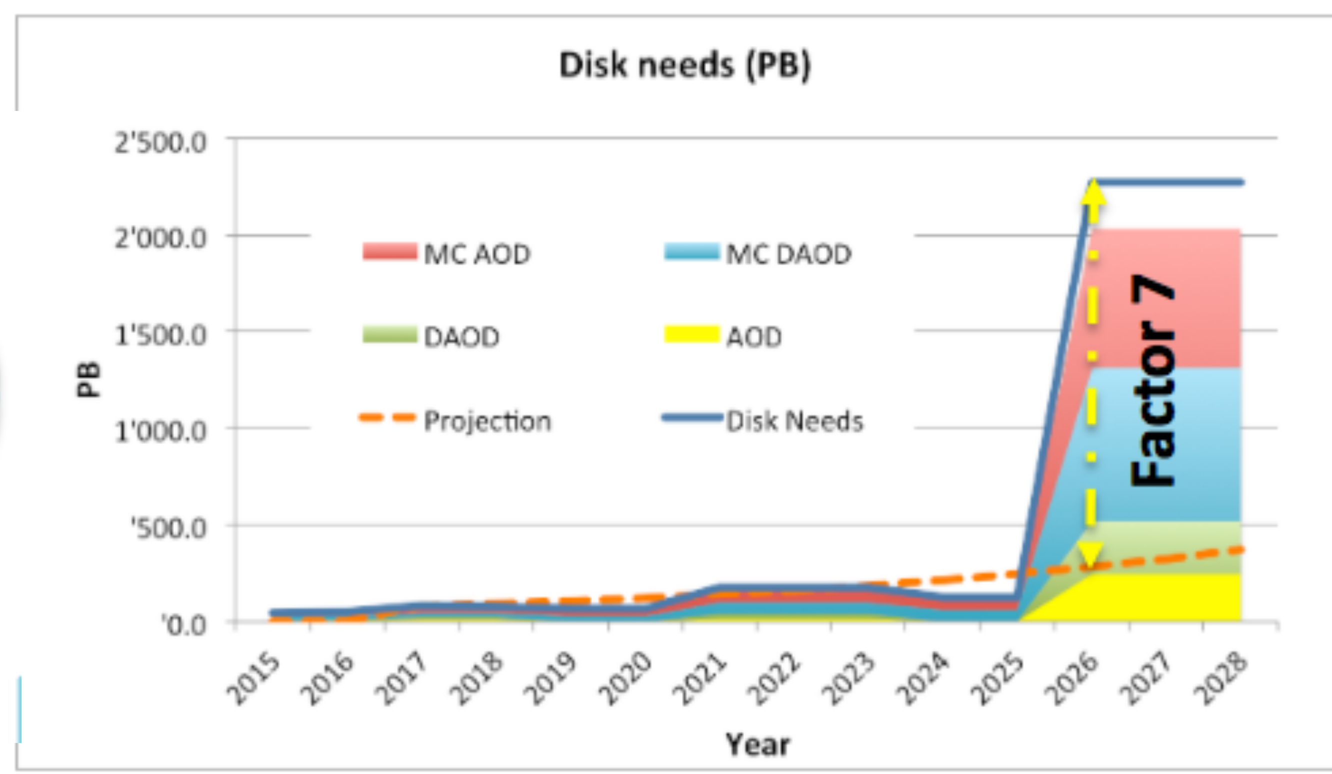
COMPUTING EVOLUTION FOR HL-LHC

- 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



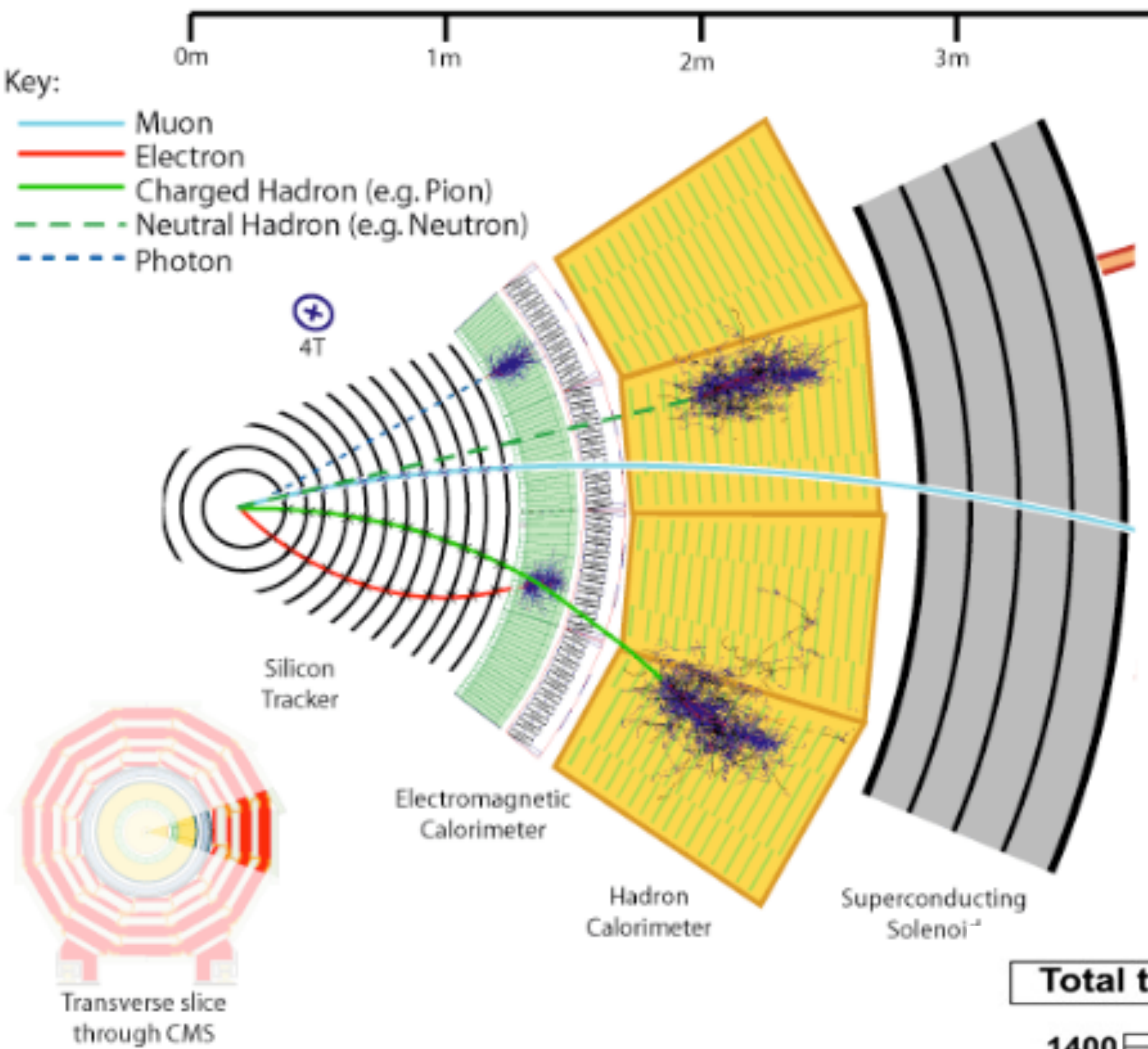
Projection of available resources in HL-LHC: 20% more CPU/year, 15% more storage/year





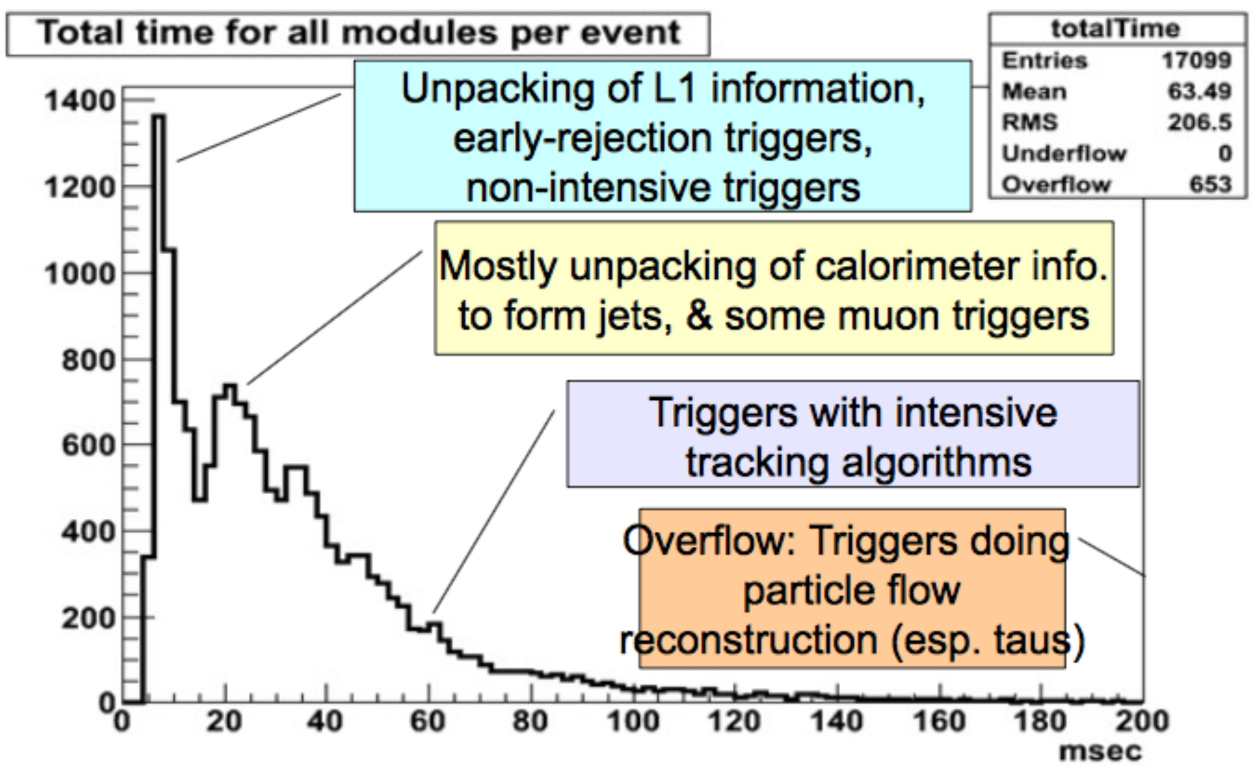
CALORIMETER TRIGGERS

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

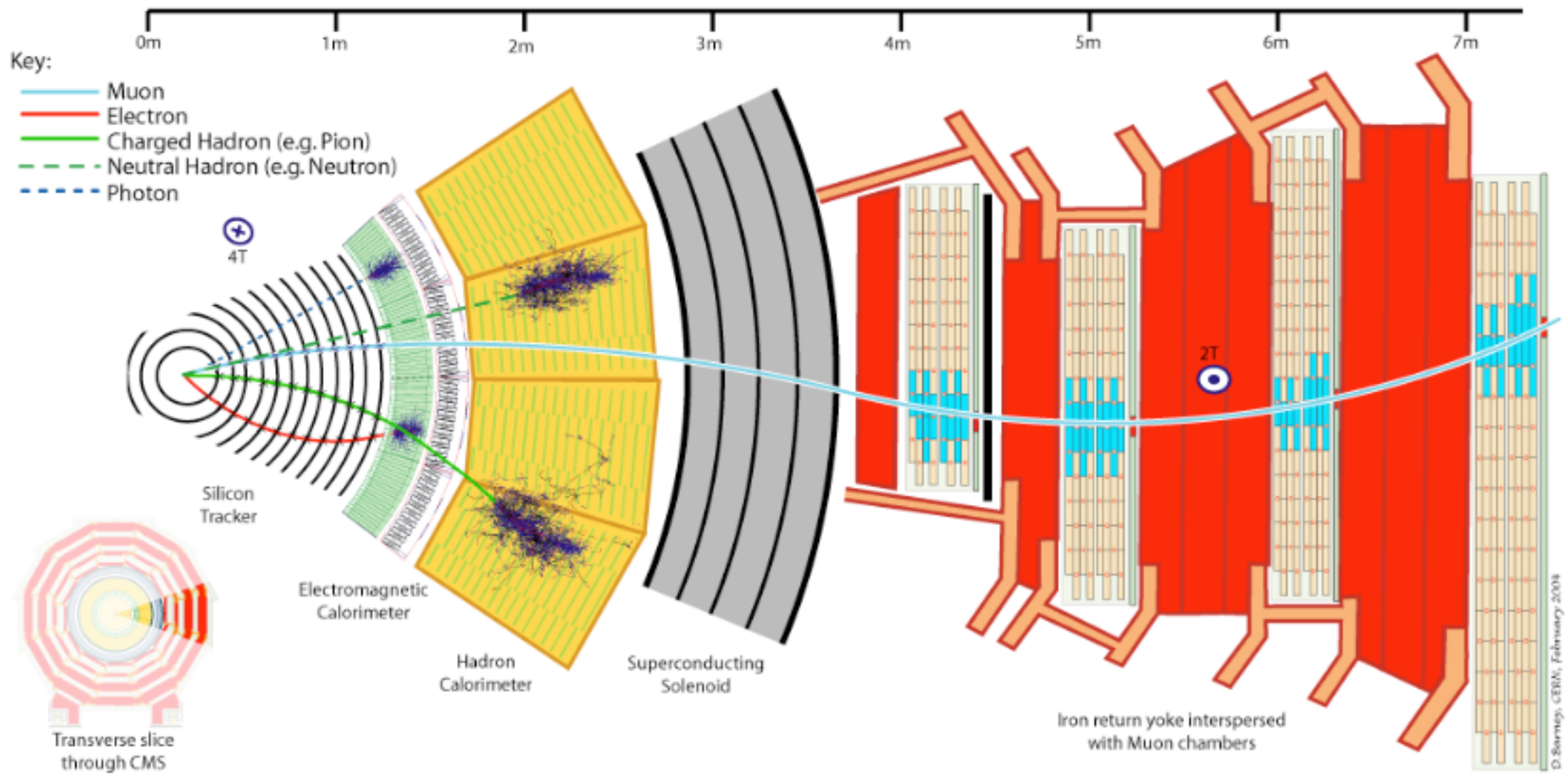


- ➡ **Fast and good resolution (LArg, PbW_4 for e-m)**
- ➡ **First-level processing (40MHz)**
  - ➡ "trigger towers" to reduce data (10-bit range)
  - ➡ sliding-window technique for local maxima
  - ➡ parallel algorithms for cluster shape and energy distribution

- ➡ **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:**

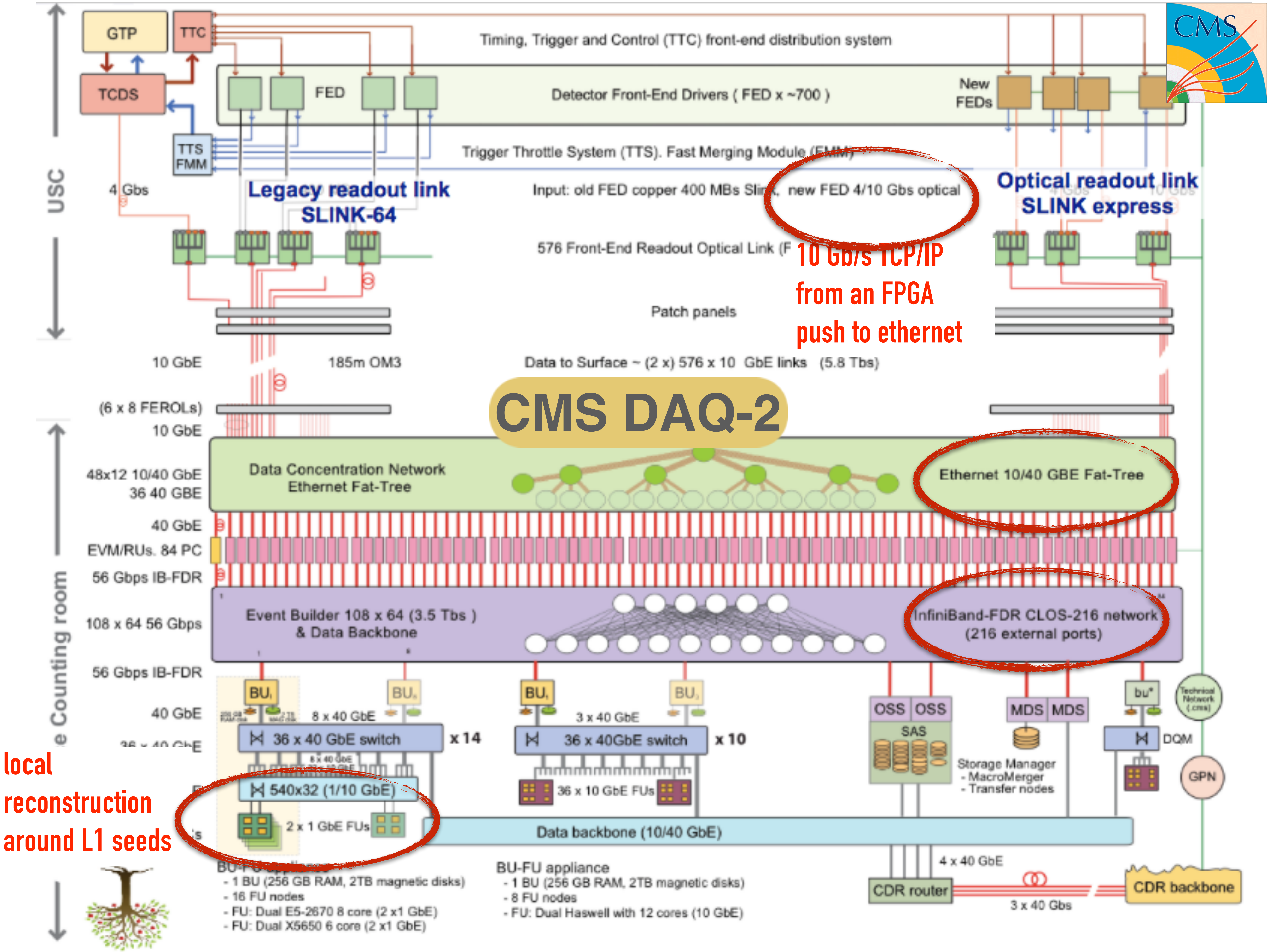
- ➡ low occupancy for fast pattern recognition
- ➡ optimal time-resolution for BC-identification

➡ **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

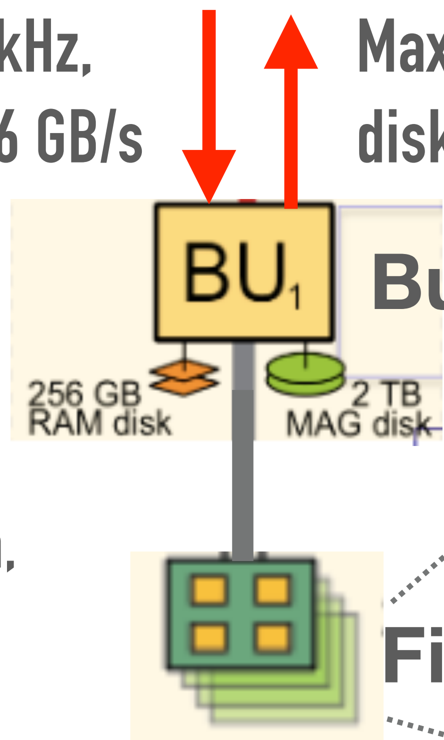

CMS DAQ-2
GTP
TTC
TCDS
TTS FMM
Timing, Trigger and Control (TTC) front-end distribution system
FED
Detector Front-End Drivers (FED x ~700)
New FEDs
Trigger Throttle System (TTS). Fast Merging Module (FMM)
USC
4 Gbs
Legacy readout link
SLINK-64
Input: old FED copper 400 MBs Slink, new FED 4/10 Gbs optical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

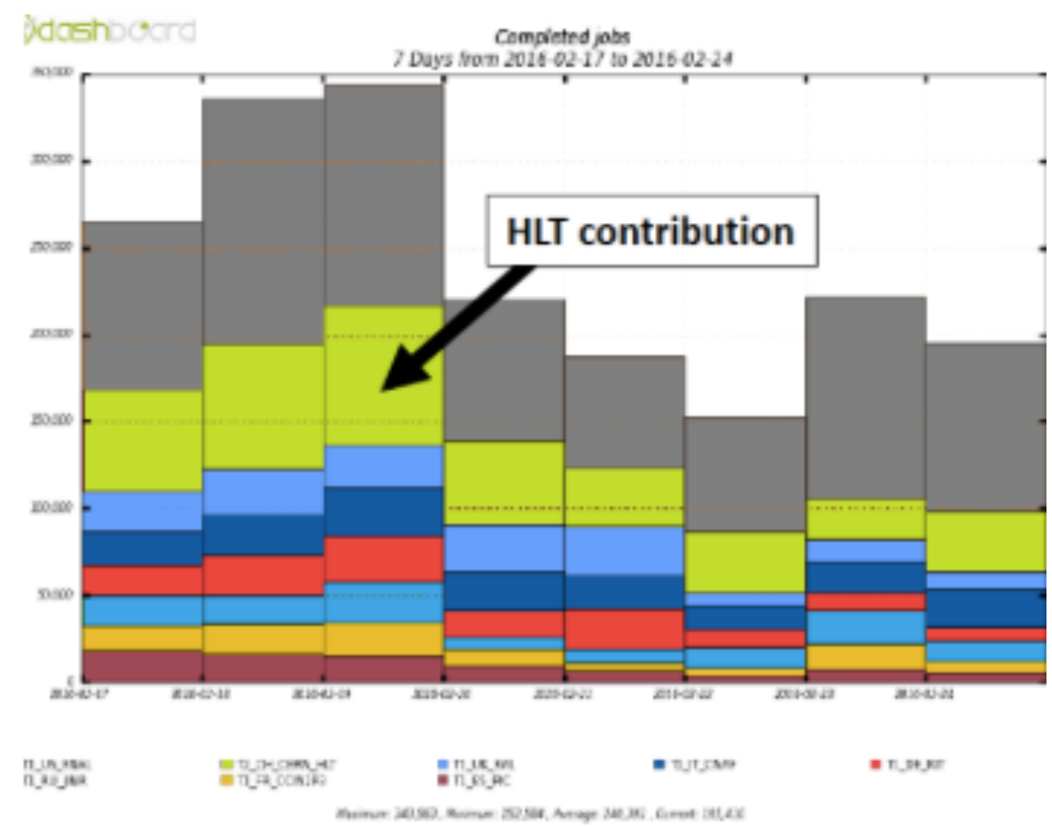
data,
status,
configuration,
latency

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

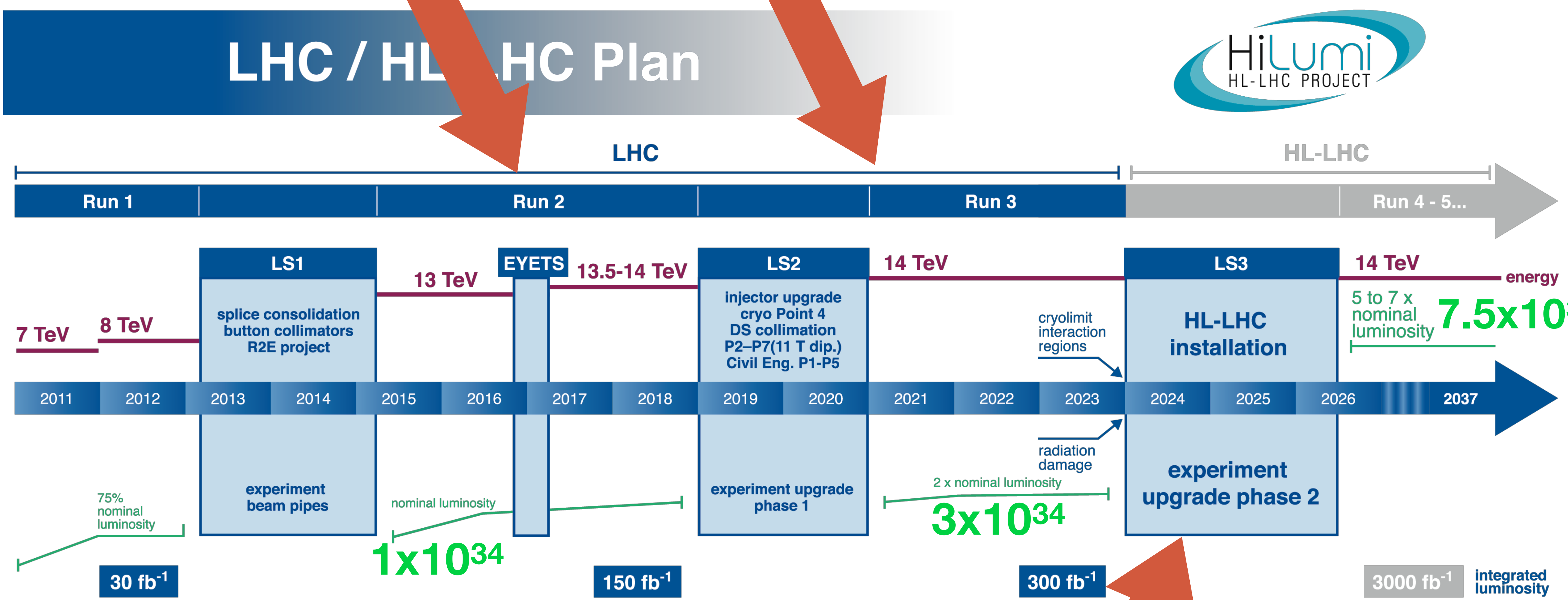


File-based communication

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

PHASES OF ATLAS/CMS TDAQ EVOLUTION

- **Run 2: optimising existing system for increasing luminosity**
- **Run 3: Add more flexibility, without major architectural changes**



- **Run 4: Major upgrade to ensure appropriate rejection**
  - Expected L1 over the limit allowed by detector FE (1MHz readout, 10x today)
  - A new tracker will be available…

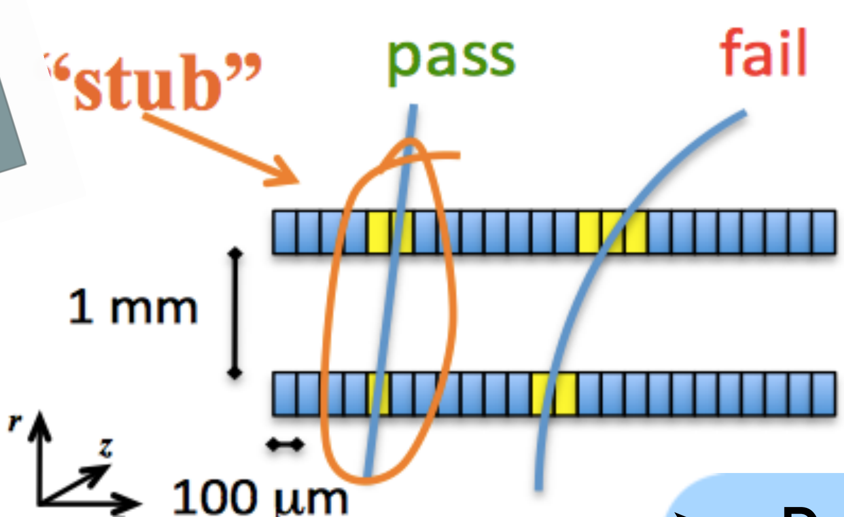
CMS: LOW-P_T TRACK FILTERING

Track filtering (low p_T)

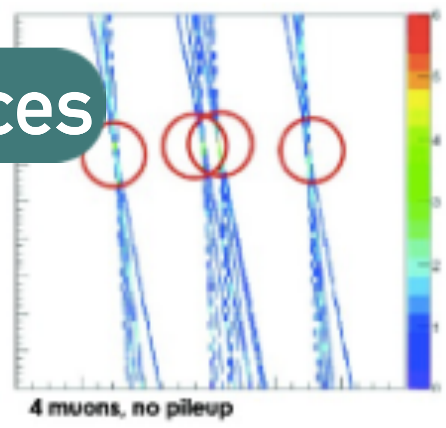
Reduce readout 40 ➟1MHz by detector coincidences

- ➡ **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

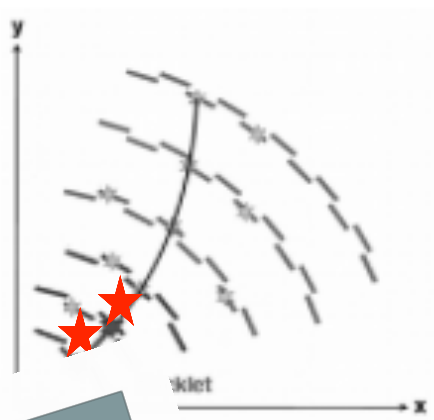
40MHz



Track finding options

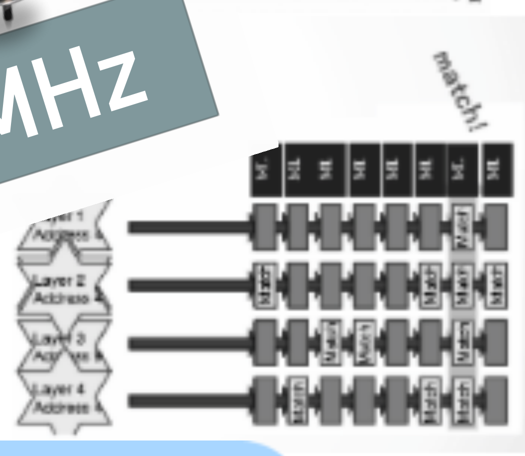


Hough Transform



Tracklets

1MHz



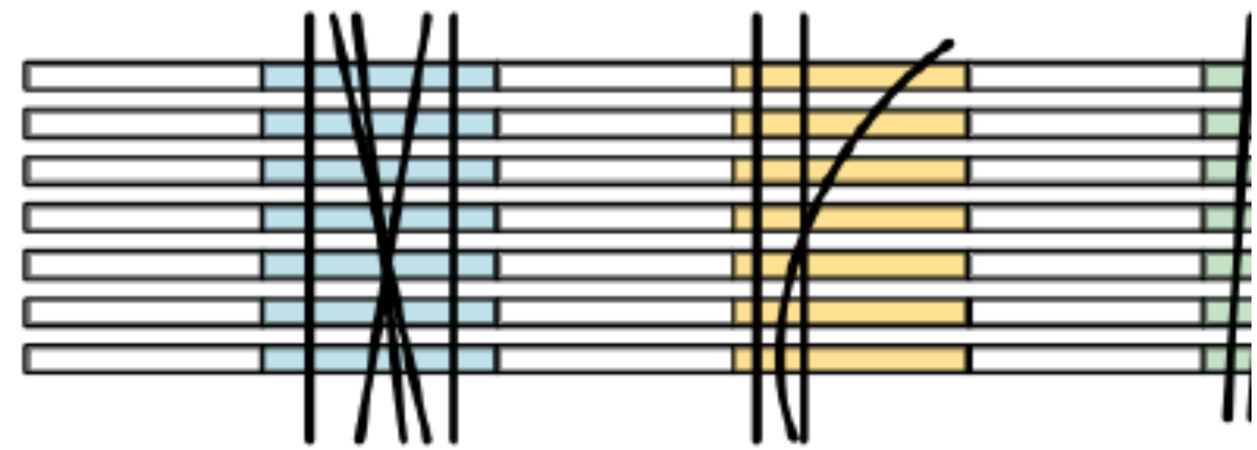
Associative Memories

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

ATLAS: EVOLUTION OF FAST TRACK TRIGGER

- **Based on current FTK system**
  - **Track-filtering:** pattern-recognition with AM
  - **Track-fitting:** linearised algorithms in FPGAs

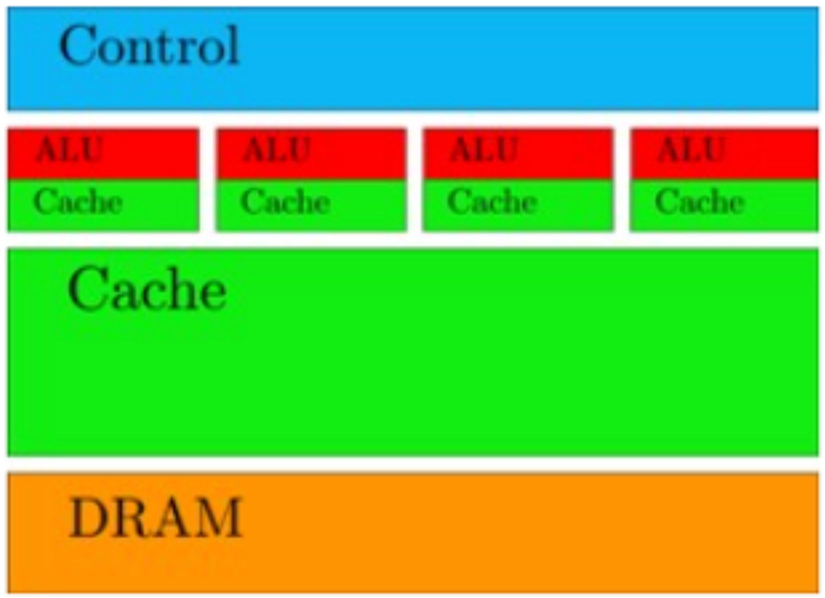
Associative Memories



AM2020:
28nm technology
250 MHz clock

- **Can either select before HLT or help HLT decision (single or double-level architecture)**
  - Depending on rates (and luminosity)
  - May need a short latency (30 μs) system if L0 rate grows up to 4MHz

- **Fast Readout speed on the silicon detectors (in 30 us latency)**
- **Massively parallel, O(500) boards, with 1-4 MHz input rate**
  - New generation chips (AM2020), 0.5 Million patterns each (total ~Billion)


Control
ALU
Cache
ALU
Cache
ALU
Cache
ALU
Cache
Cache
DRAM

CPU


DRAM

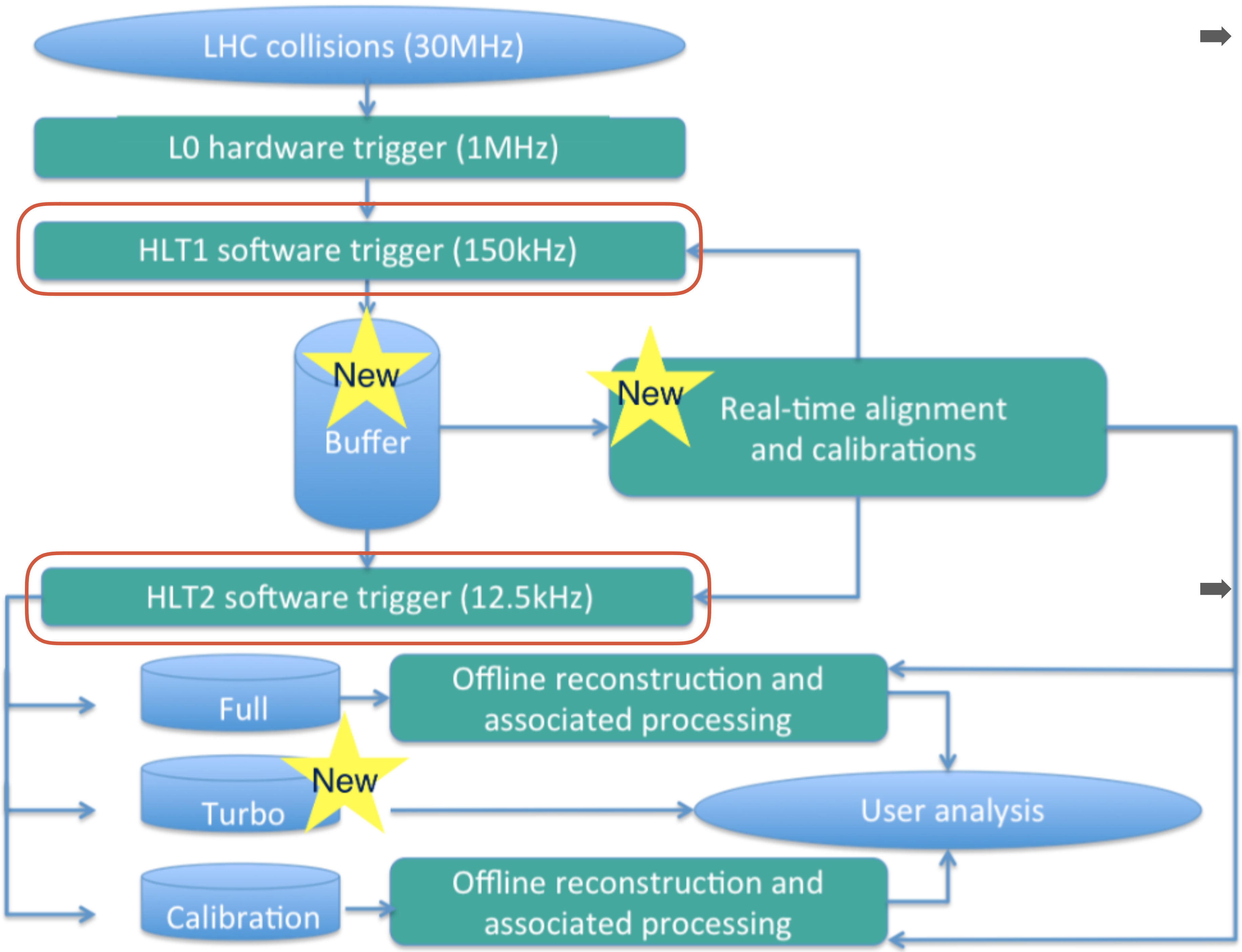
GPU


intel
Xeon® 7500


TESLA

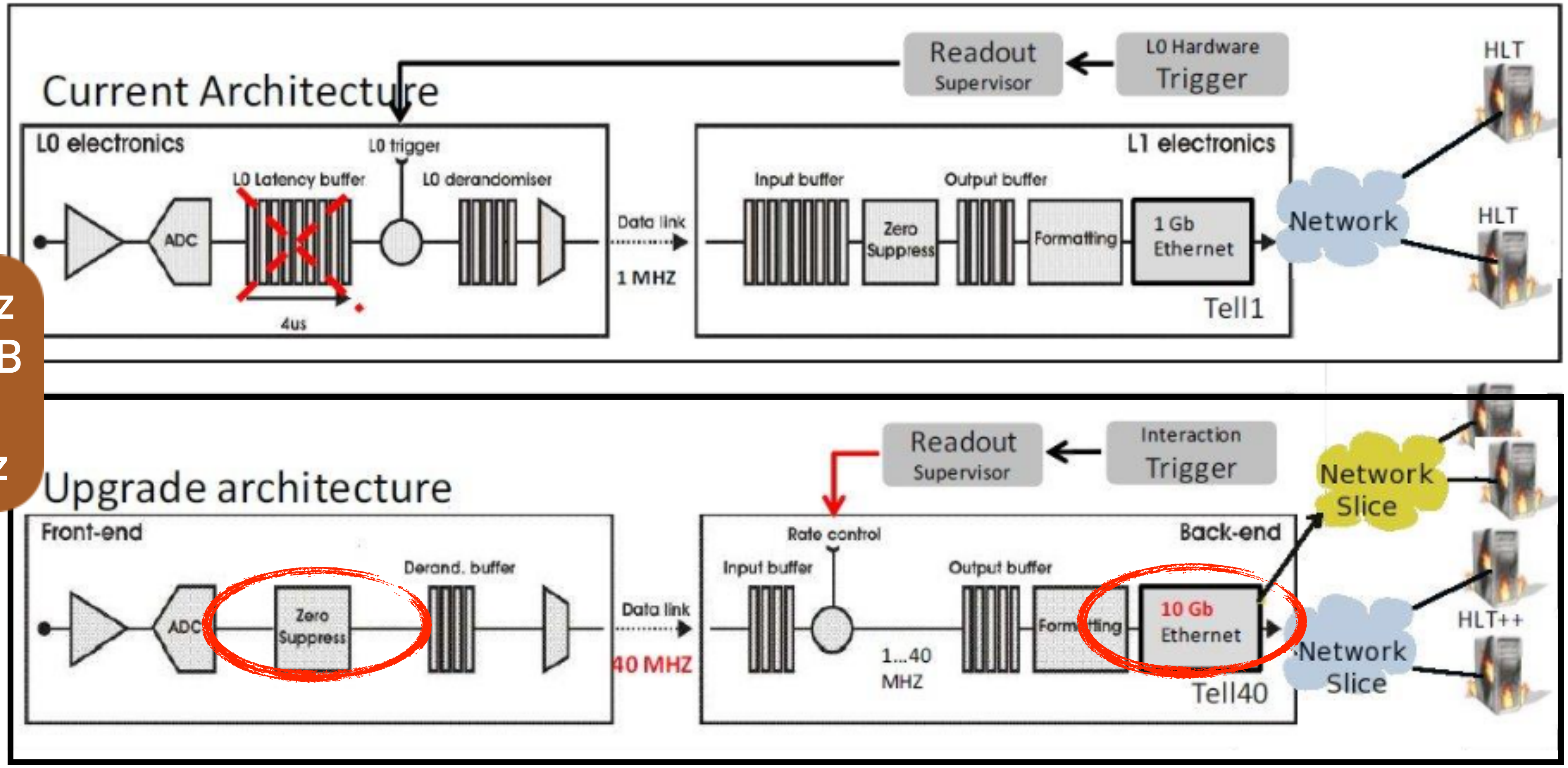
REAL-TIME ANALYSIS IN RUN-2

Large benefit from VELO alignments at each fill!



- **Calibrations/ Alignments**
  - align ~1700 detector components
  - calculate ~2000 calibration constants
  - within a few minutes
- **"turbo stream"**
  - Offline quality obtained in HLT-2
  - More than 200 selections
  - Run2 results at EPS-HEP conference just a week after end of data-taking

HOW TO LIVE WELL WITHOUT A L1 TRIGGER



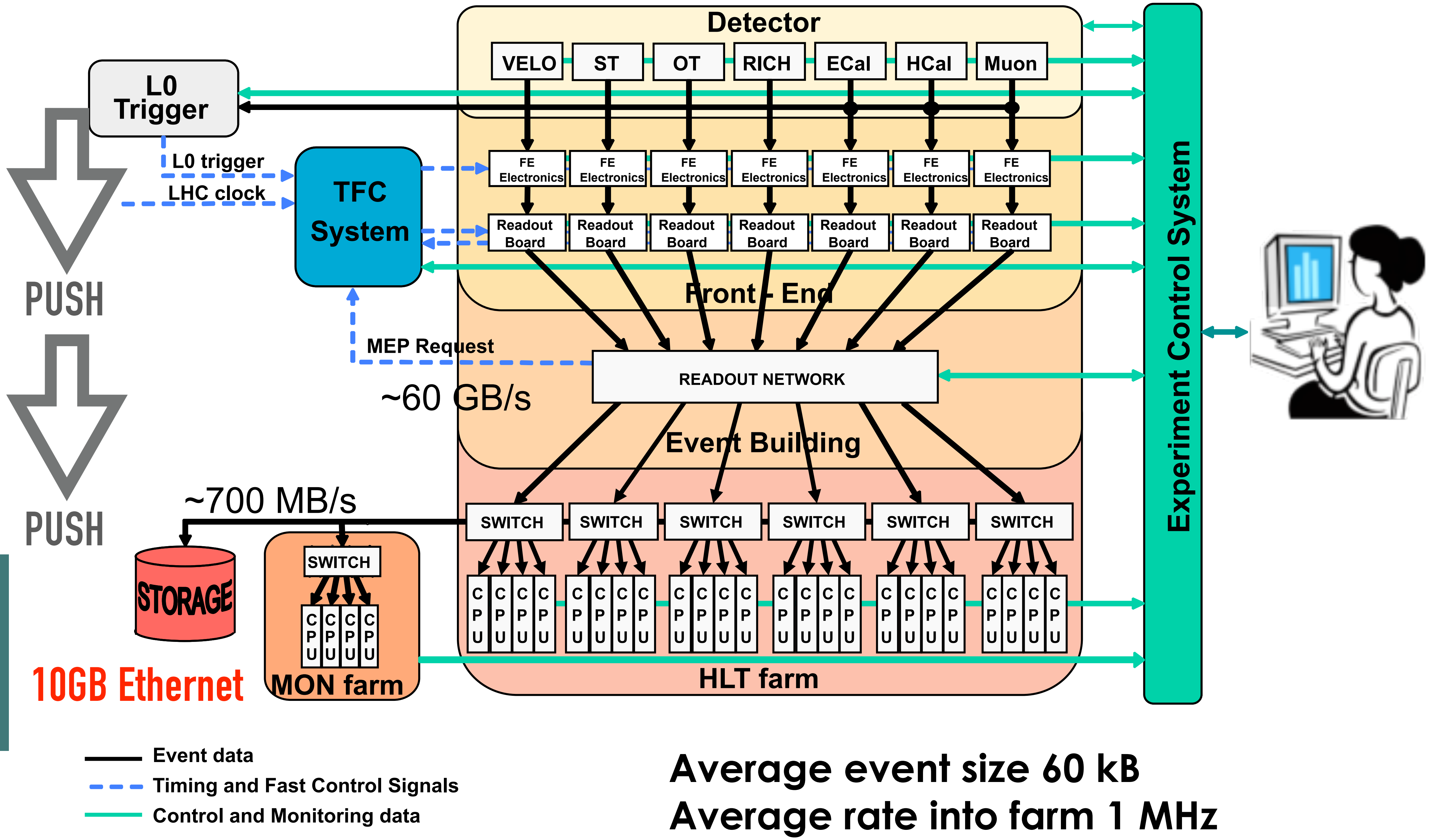
Readout: 40 MHz
Event size: 100kB
DAQ: 40 Tbit/s
Record: 100 kHz

- ➡ **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)



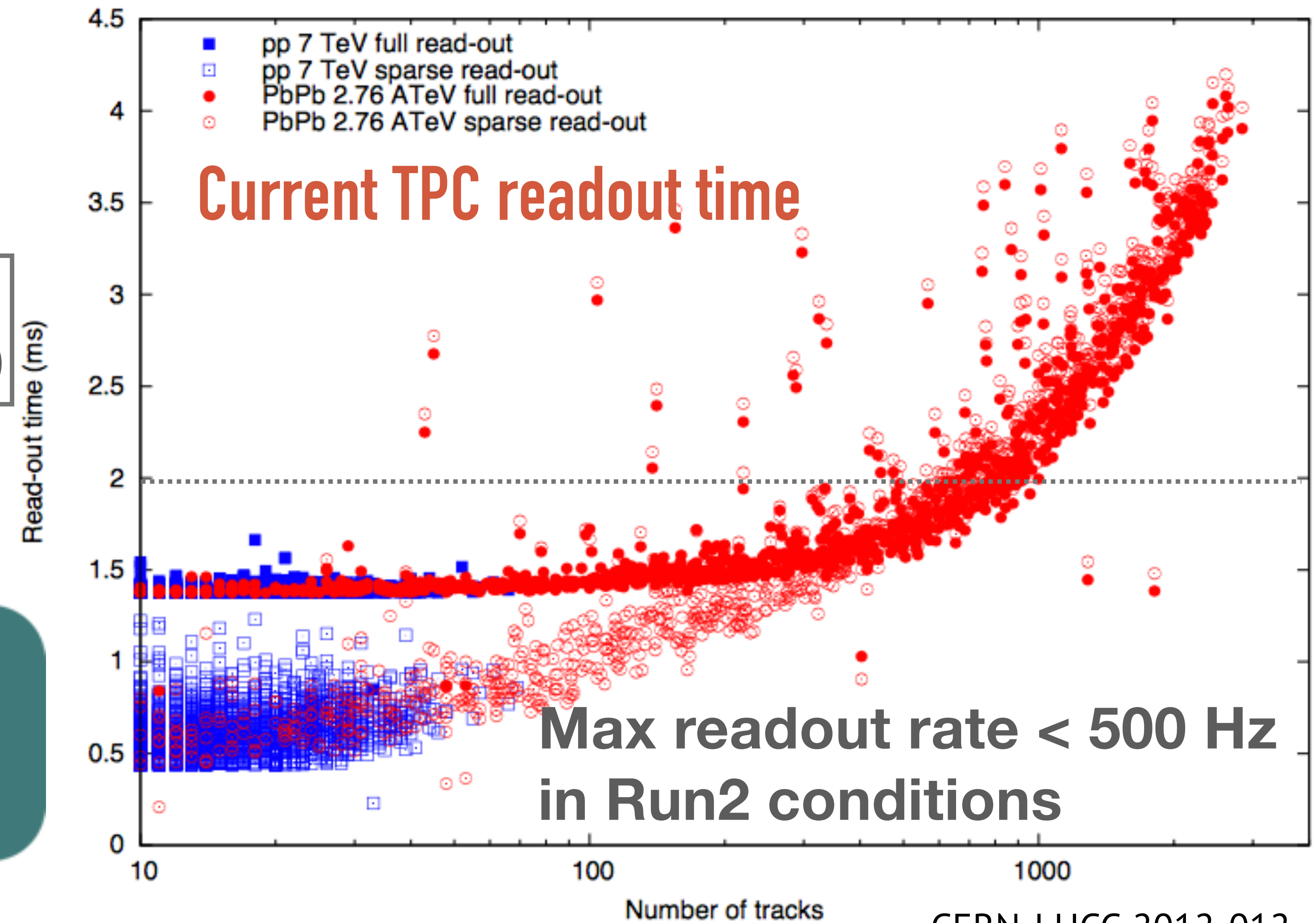
Average event size 60 kB
Average rate into farm 1 MHz
Average rate to tape ~12 kHz

- **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**

- ➡ LHC heavy ion programme extended to reach x100 statistics
- ➡ Access rare physics for dynamics of condensed QCD, via complex probes at low p_T
  - ➡ Increase vertex and tracking capabilities at low momentum (new trackers)
  - ➡ Increase detector granularity (event size!)
  - ➡ Higher readout rates: new electronics, TPC readout with GEM (no gate)

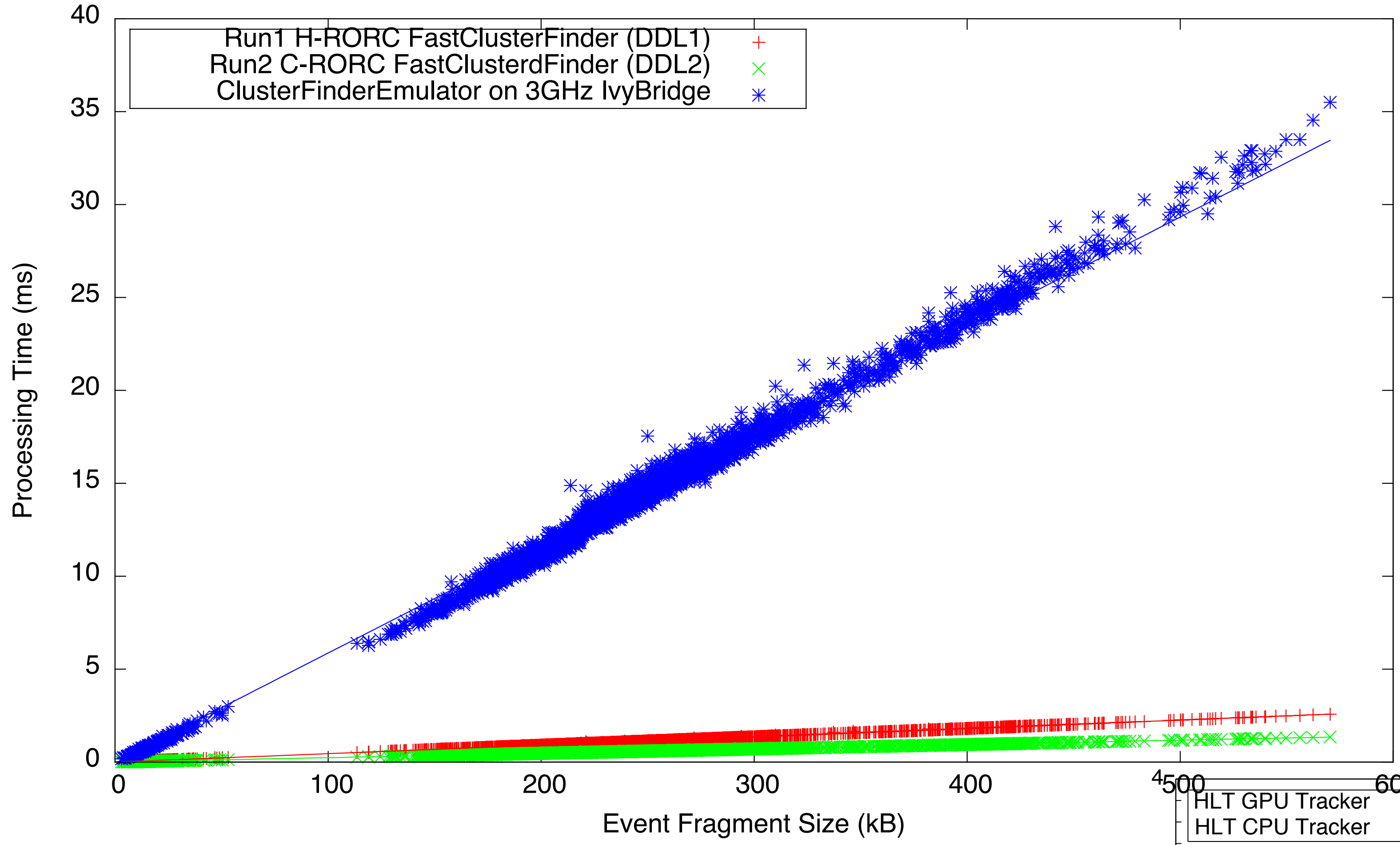
- ➡ **Requirements for DAQ**
  - ➡ Pb-Pb rate: **~kHz → 50 kHz** (23 MB/event)
  - ➡ → ~TB/s detector readout
  - ➡ → Storage bandwidth x O(100)
  - ➡ Offline reconstruction also challenging

To maintain acceptance, overcome classical trigger concept



CERN-LHCC-2012-012

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.