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

FROM SMALL TO LARGE T/DAQ SYSTEMS

OUTLINE

- ➡ Examples of small experiments with their limits
- ➡ Overview of LHC experiments and their upgrade
- ➡ Future TDAQ systems (Dune/Proto-Dune)

OUTLINE

- ➡ **Examples of small experiments with their limits**
- ➡ Overview of LHC experiments and their upgrade
- ➡ Future TDAQ systems (Dune/Proto-Dune)

SOME NOMENCLATURE

SOME NOMENCLATURE

- **Data Size**
  - Summing up data from all Front-End channels
    - 100 M channels of silicon detectors give few MB/event
  - depends on detector granularity (number of channels), on detector technology (single bit versus drift-time or TPC) and pile-up level

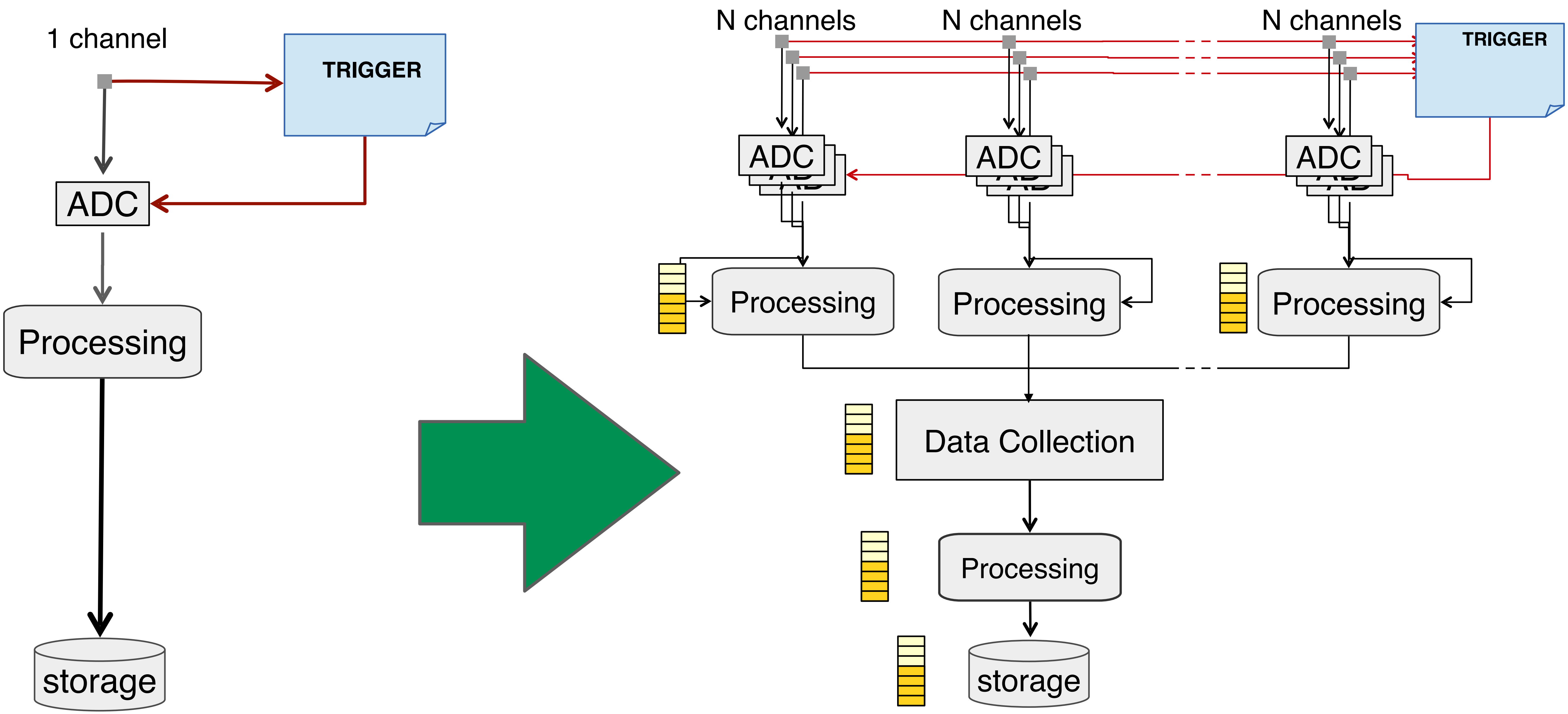
SOME NOMENCLATURE

- **Data Size**
  - Summing up data from all Front-End channels
    - 100 M channels of silicon detectors give few MB/event
  - depends on detector granularity (number of channels), on detector technology (single bit versus drift-time or TPC) and pile-up level
- **Data Rate**
  - Front-End readout rate
  - LHC clock gives about 40 M evt/s at 13 TeV
  - Pierre Auger Observatory: about 1 evt/100years/km at EeV

SOME NOMENCLATURE

- **Data Size**
  - Summing up data from all Front-End channels
    - 100 M channels of silicon detectors give few MB/event
  - depends on detector granularity (number of channels), on detector technology (single bit versus drift-time or TPC) and pile-up level
- **Data Rate**
  - Front-End readout rate
  - LHC clock gives about 40 M evt/s at 13 TeV
  - Pierre Auger Observatory: about 1 evt/100years/km at EeV
- **DAQ bandwidth**
  - 40MHz x 1 MB = 40TB/s
    - too much data!
  - select and record only the most important events

RECAP ON T/DAQ SYSTEMS



- ➡ **Two independent paths for trigger and DAQ**
- ➡ **Segmented Readout and trigger to allow parallel processing**
- ➡ **Included buffers at each stage to control dead-time**
- ➡ **How to scale these systems?**

ONE SMALL EXPERIMENT: NA59 @SPS

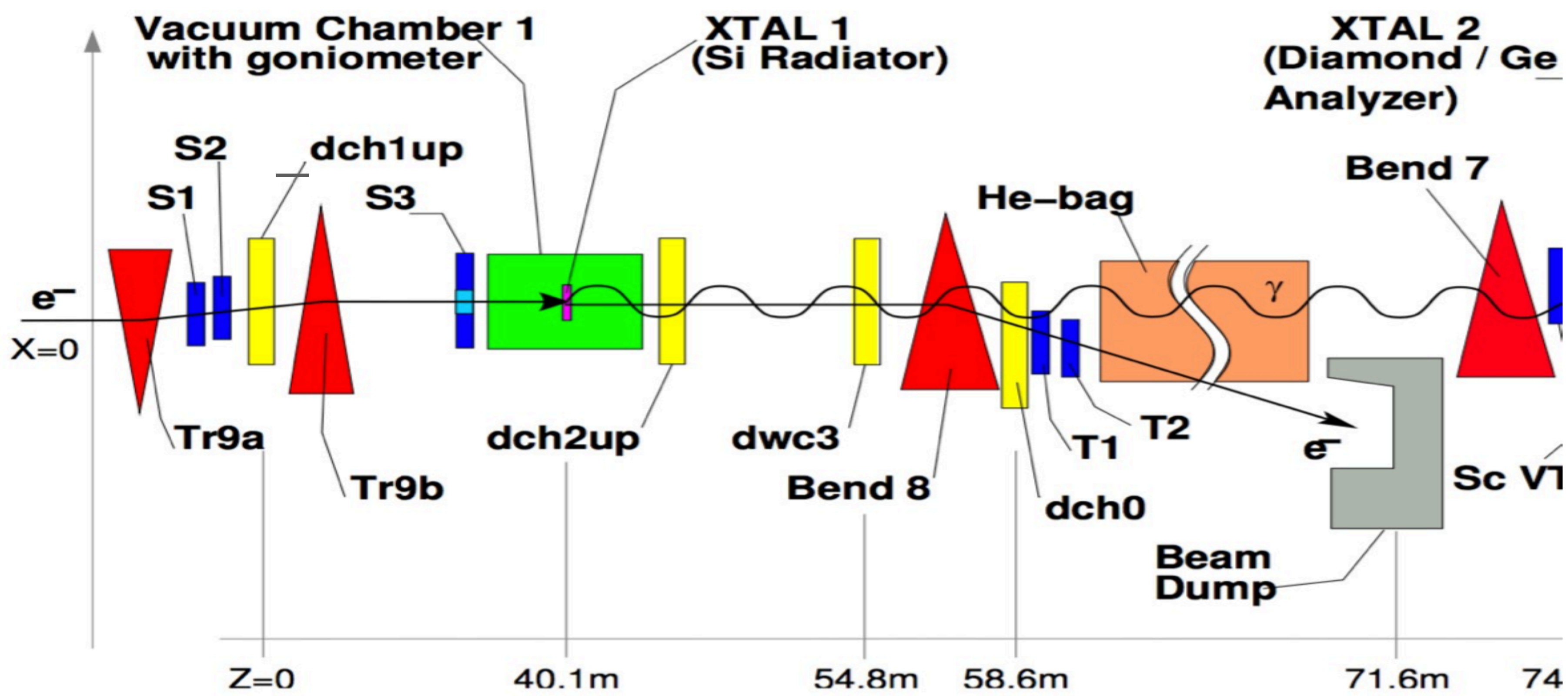
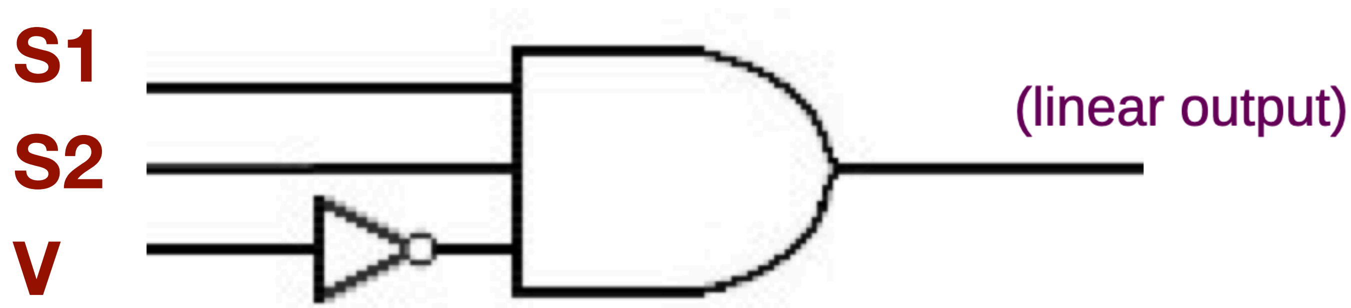


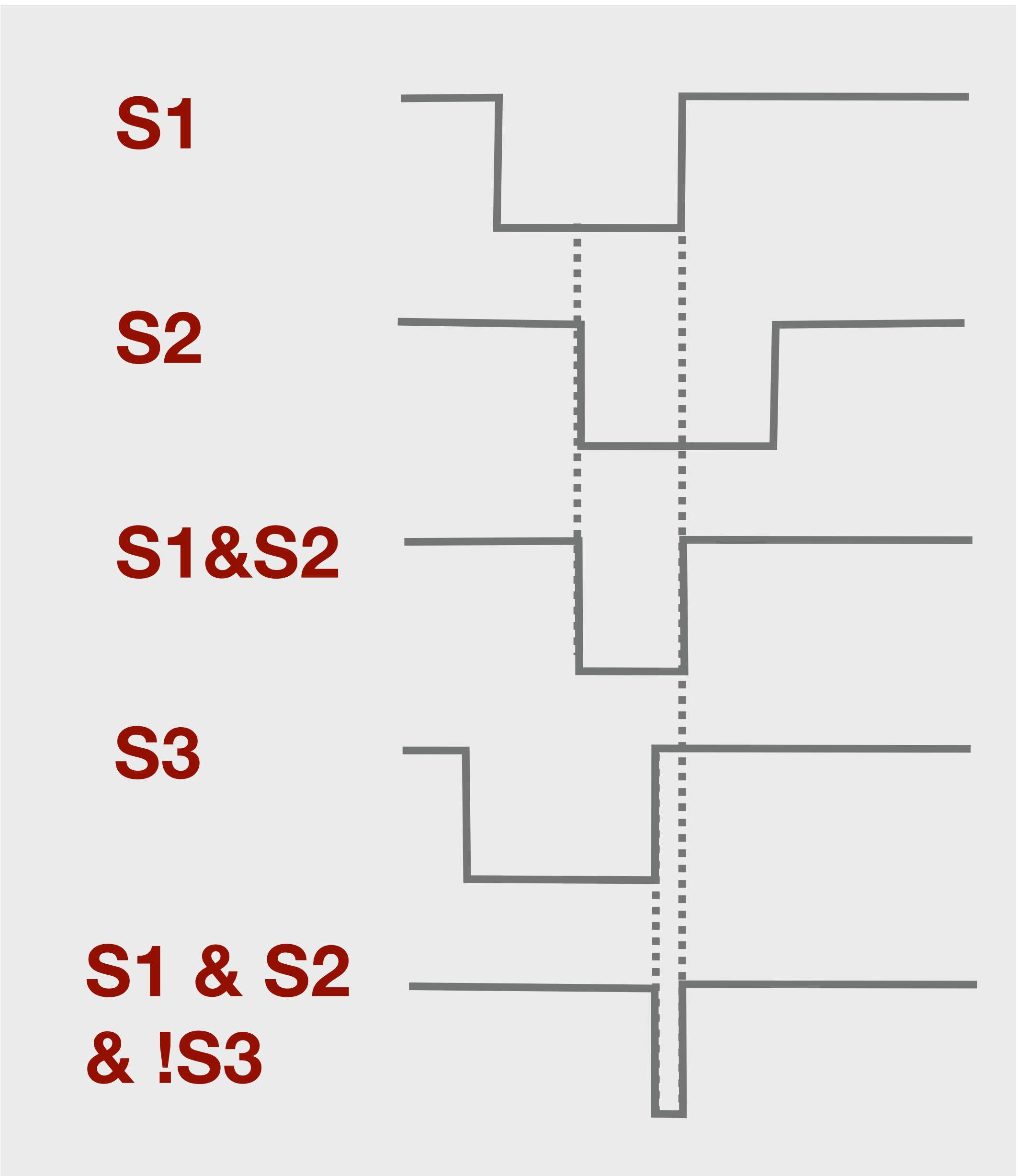
Fig. 1. Setup of the Na59 Experiment

- Trigger event with an electron at the correct incident angle wrt crystal
- Three scintillators S1, S2 and S3 ensure the arrival of the beam within the acceptance of the crystal
  - Input **N1 = S1 & S2 & !S3** ---> an electron is coming and it is not away from the central axis
  - use S3 as veto (anti-coincidence)
- After the magnet, two scintillators to tag the electron out of the beam
  - **N2 = N1 x (T1 || T2)** ---> the electron radiated a photon and was diverted by the magnet

TRIGGER TIMING

- Simple coincidence and veto logic can be broken if signals are not formed correctly



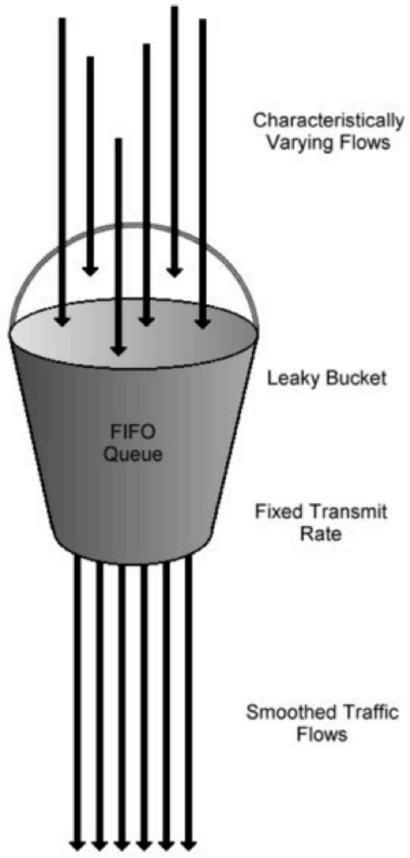


- **Signals are random/independent**
- **Can fluctuate in duration and jitter**
  - Need preliminary **timing alignment** between signals
    - e.g adding delays to faster signals
  - Need **forming output signals** with known width
    - fix width of output signal at each step

WHAT TO SCALE

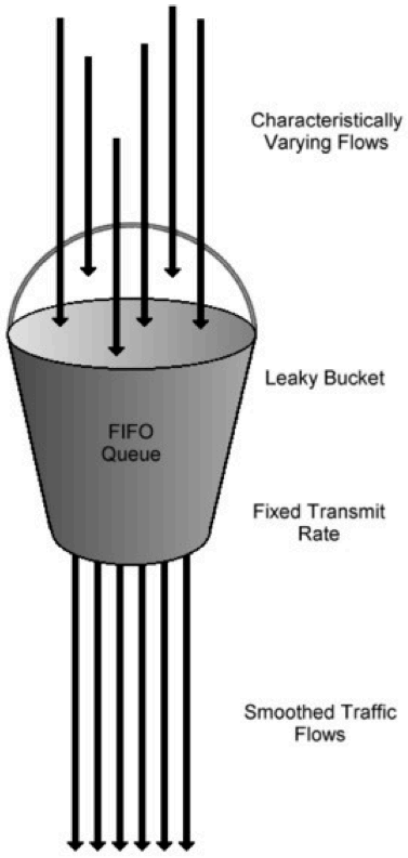
- ➡ **Step 1: Increasing rate**
- ➡ **Step 2: Increasing sensors**
- ➡ **Step 3: Multiple front-ends**
- ➡ **Step 4: Multi-level trigger**
- ➡ **Step 5: Data-flow control**

1 – INCREASING RATE



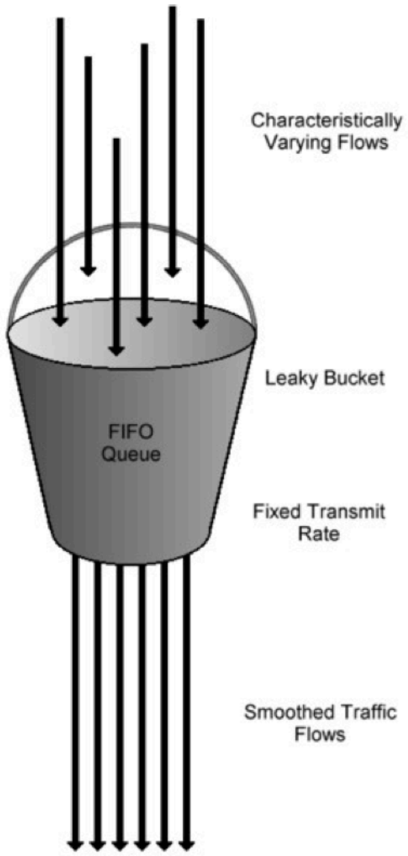
1 – INCREASING RATE

- **If two signals arrive very close in time**
  - detector signals overlap (ask you detector expert, are you sure the detector is good at that rate? is your FE fast enough?)
  - can have dead-time if not added any … FIFO!



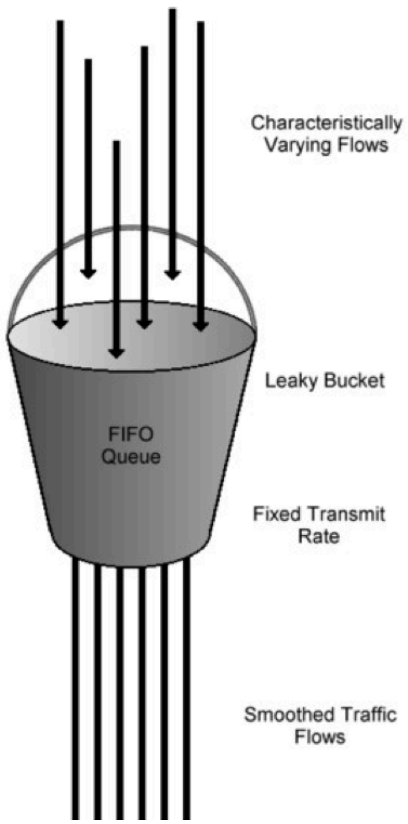
1 – INCREASING RATE

- **If two signals arrive very close in time**
  - detector signals overlap (ask you detector expert, are you sure the detector is good at that rate? is your FE fast enough?)
  - can have dead-time if not added any … FIFO!
- **Is derandomization enough?**
  - if FE readout windows overlap
    - add artificial dead-time to protect the FrontEnd (**simple deadtime**)
  - if FE buffers overflow in case of trigger bursts
    - add artificial dead-time (**complex deadtime**)



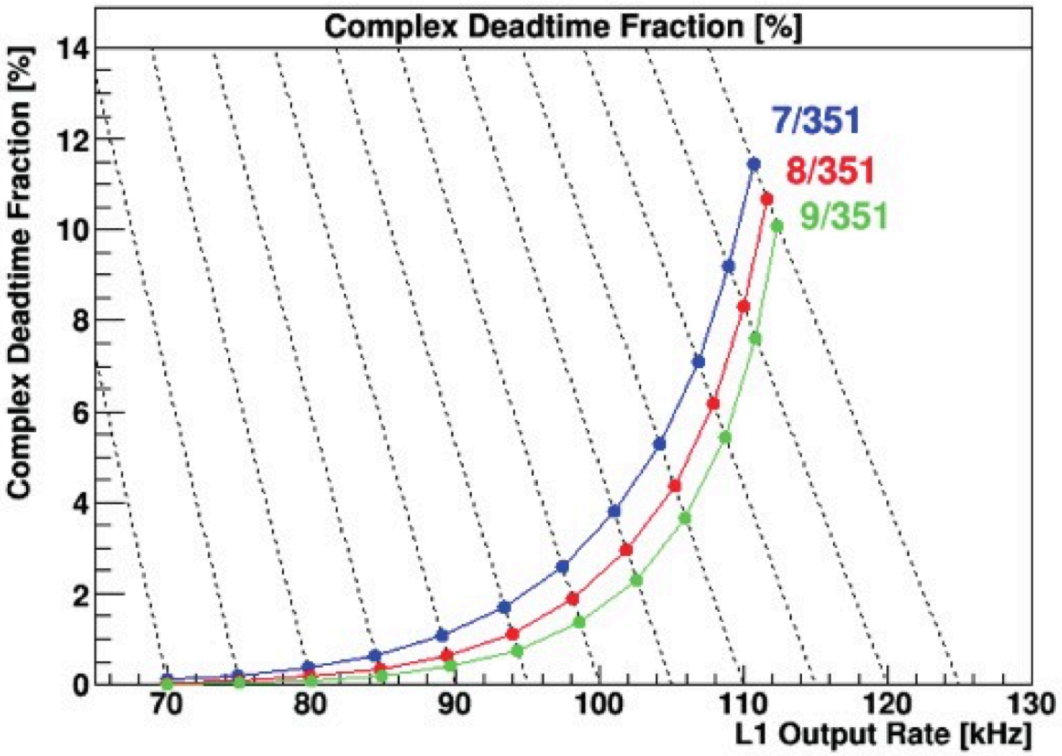
1 – INCREASING RATE

- **If two signals arrive very close in time**
  - detector signals overlap (ask you detector expert, are you sure the detector is good at that rate? is your FE fast enough?)
  - can have dead-time if not added any … FIFO!
- **Is derandomization enough?**
  - if FE readout windows overlap
    - add artificial dead-time to protect the FrontEnd (**simple deadtime**)
  - if FE buffers overflow in case of trigger bursts
    - add artificial dead-time (**complex deadtime**)

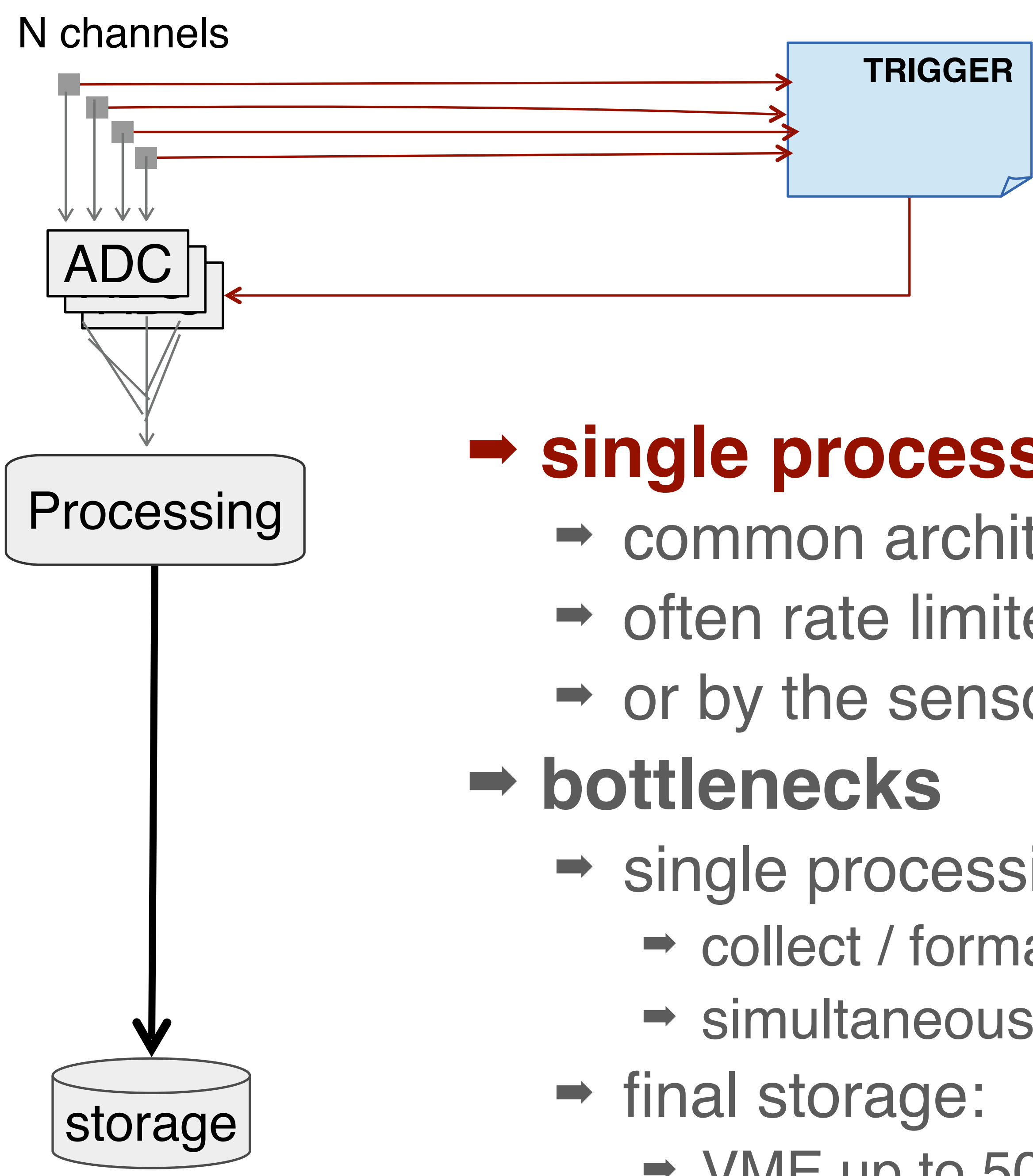


Leaky bucket (LAr readout)

- **Example in ATLAS @Run2: 90 kHz < 2%**
  - Simple deadtime: 4 LHC BC [100 ns] after any L1 trigger
  - Complex deadtime: leaky-bucket algorithms x4 detectors
    - two params: bucket size (in number of events), /readout time (in BC units)
    - i.e. 9 / 351 for LAr readout



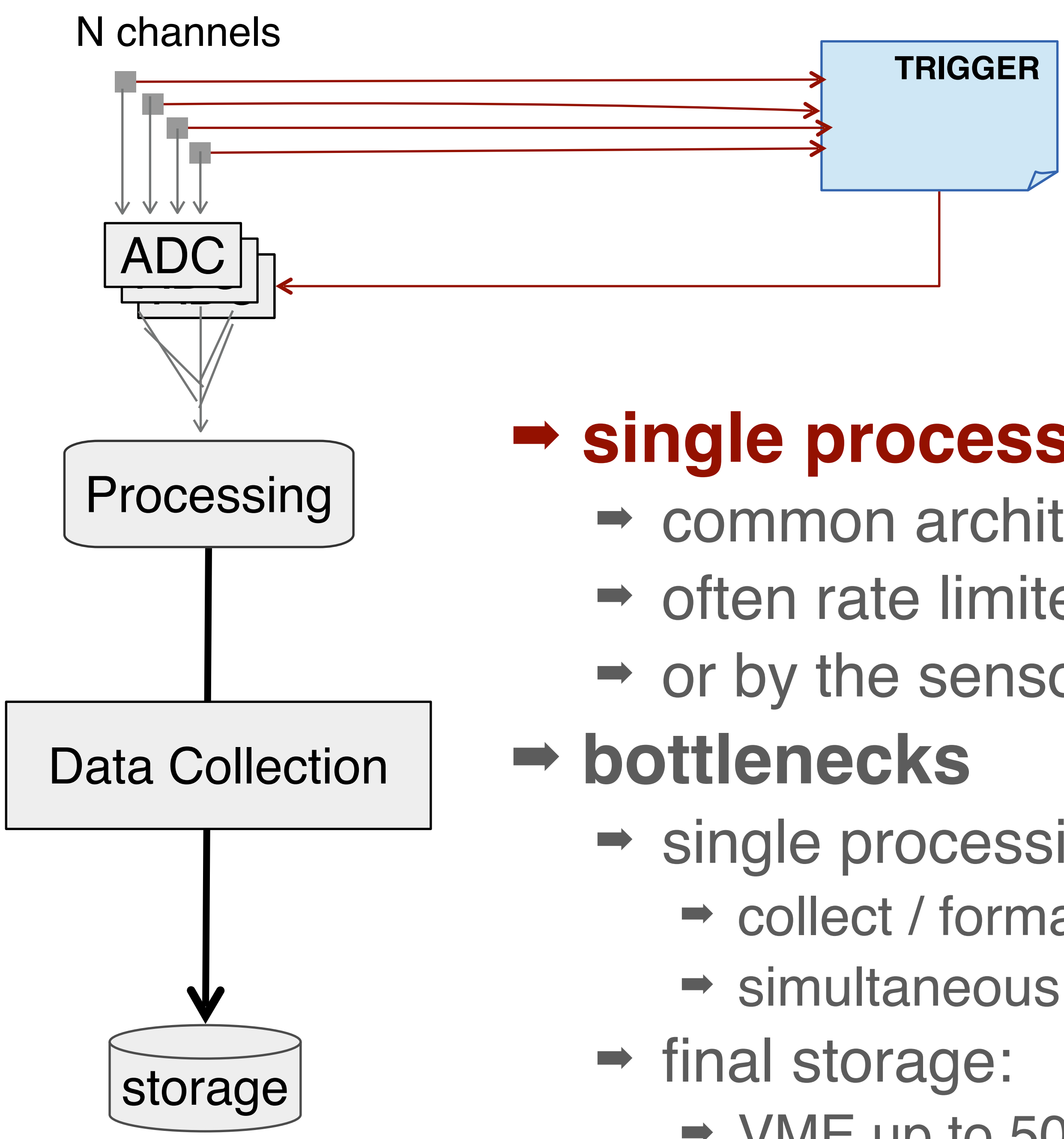
2 – INCREASING NUMBER OF CHANNELS



- ➡ more sensors ==> more granularity
- ➡ multiple digitisers ==> parallelism

- ➡ **single processing system**
  - ➡ common architecture in **test-beams and small experiments**
  - ➡ often rate limited by (interesting) physics itself, not TDAQ
  - ➡ or by the sensors
- ➡ **bottlenecks**
  - ➡ single processing unit
    - ➡ collect / format / compress data can be heavy
    - ➡ simultaneously writing storage
  - ➡ final storage:
    - ➡ VME up to 50MB/s → 1TB in 6h
    - ➡ too many disks in one week!
- ➡ **decouple storage from processing unit (PU)**
  - ➡ dedicated "Data Collection" unit to format, compress and store

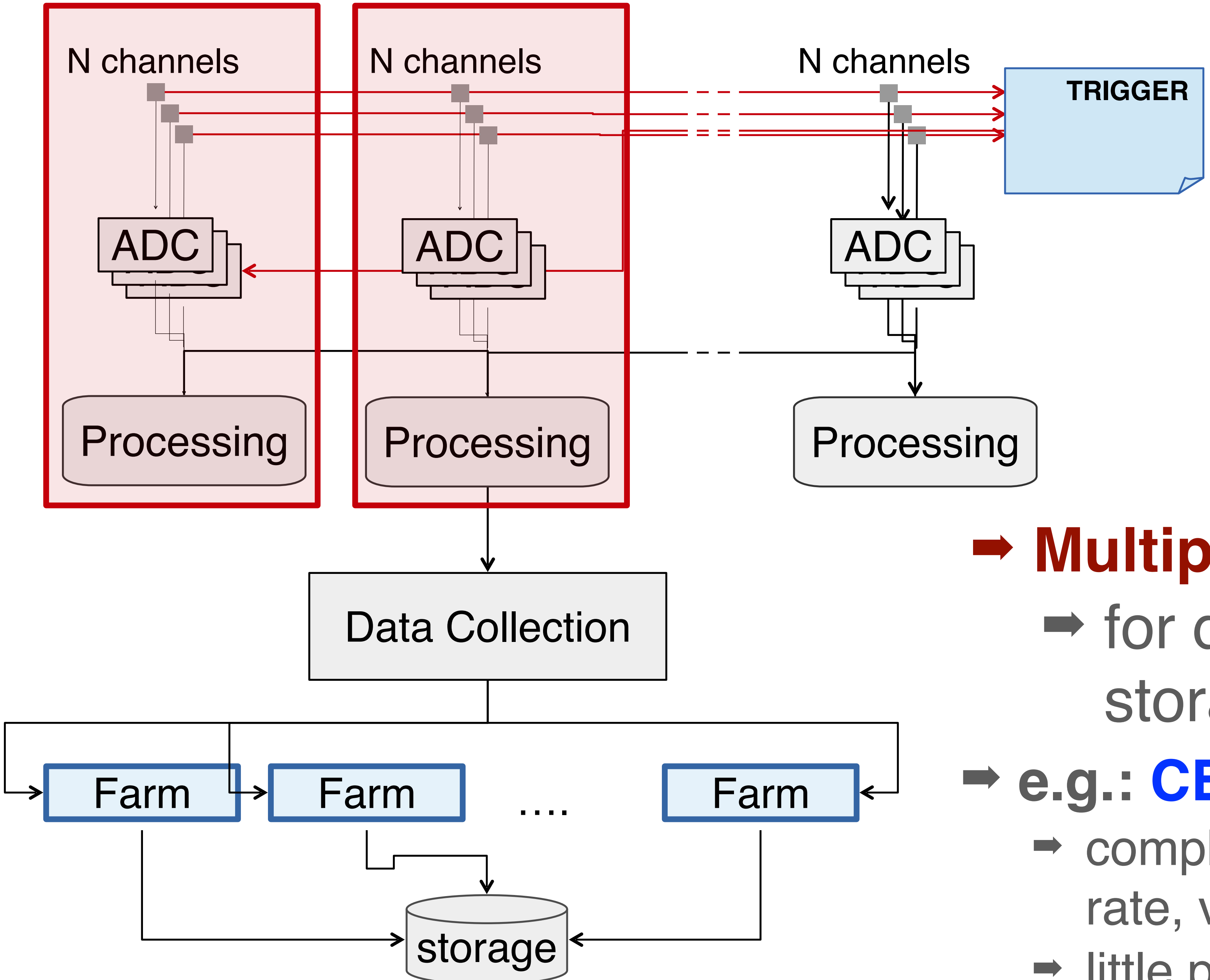
2 – INCREASING NUMBER OF CHANNELS



- ➡ more sensors ==> more granularity
- ➡ multiple digitisers ==> parallelism

- ➡ **single processing system**
  - ➡ common architecture in **test-beams and small experiments**
  - ➡ often rate limited by (interesting) physics itself, not TDAQ
  - ➡ or by the sensors
- ➡ **bottlenecks**
  - ➡ single processing unit
    - ➡ collect / format / compress data can be heavy
    - ➡ simultaneously writing storage
  - ➡ final storage:
    - ➡ VME up to 50MB/s → 1TB in 6h
    - ➡ too many disks in one week!
- ➡ **decouple storage from processing unit (PU)**
  - ➡ dedicated “Data Collection” unit to format, compress and store

3 – INCREASING FROND-END ELEMENTS



- ➡ **LEP**
- ➡ 10^5 channels
- ➡ 22µs crossing rate – no event overlap
- ➡ single interaction

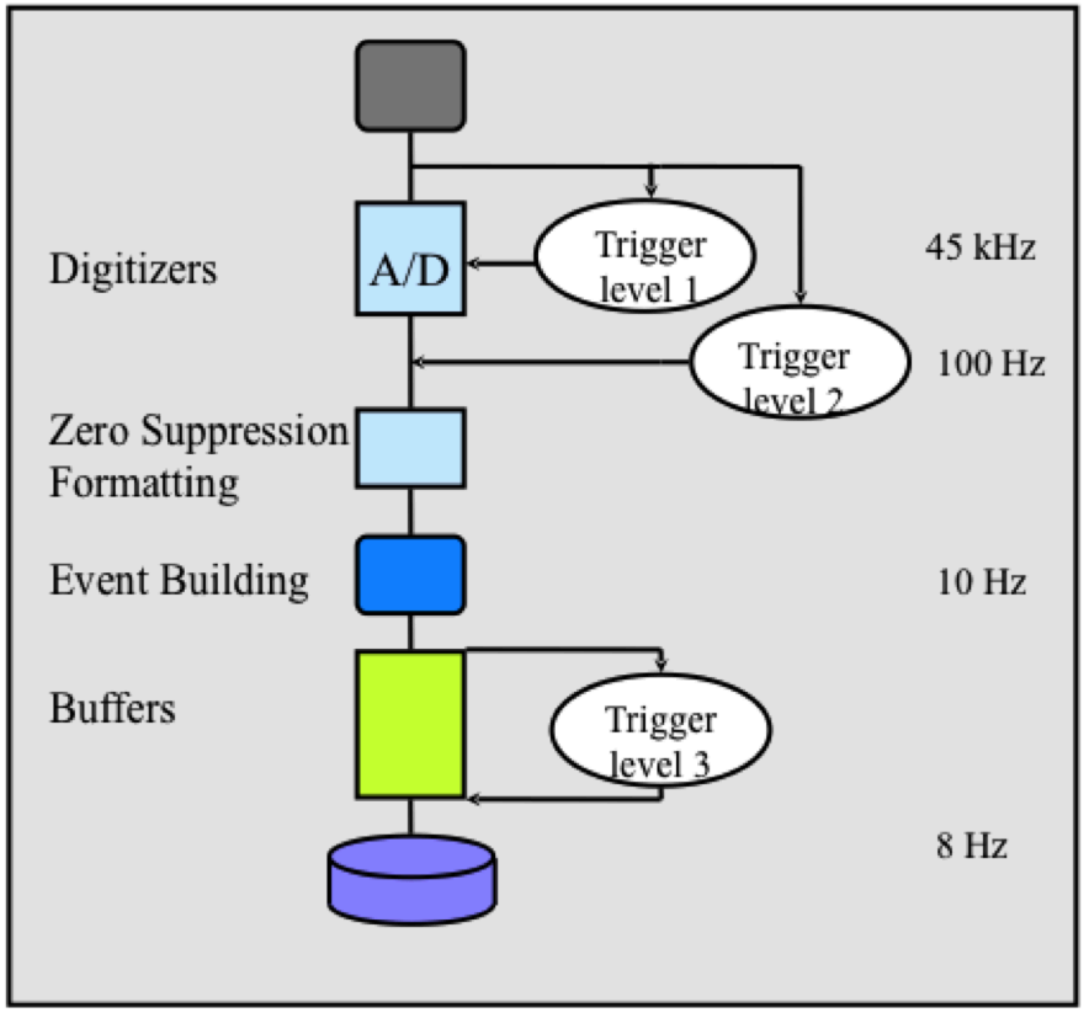
➡ **Multiple processing units**
- ➡ for data processing and storage

➡ **e.g.: CERN LEP experiments**
- ➡ complex detectors, moderate trigger rate, very little background
- ➡ little pileup, limited channel occupancy

➡ **simpler, slow gas-based main trackers**

4 - MULTI-LEVEL TRIGGER

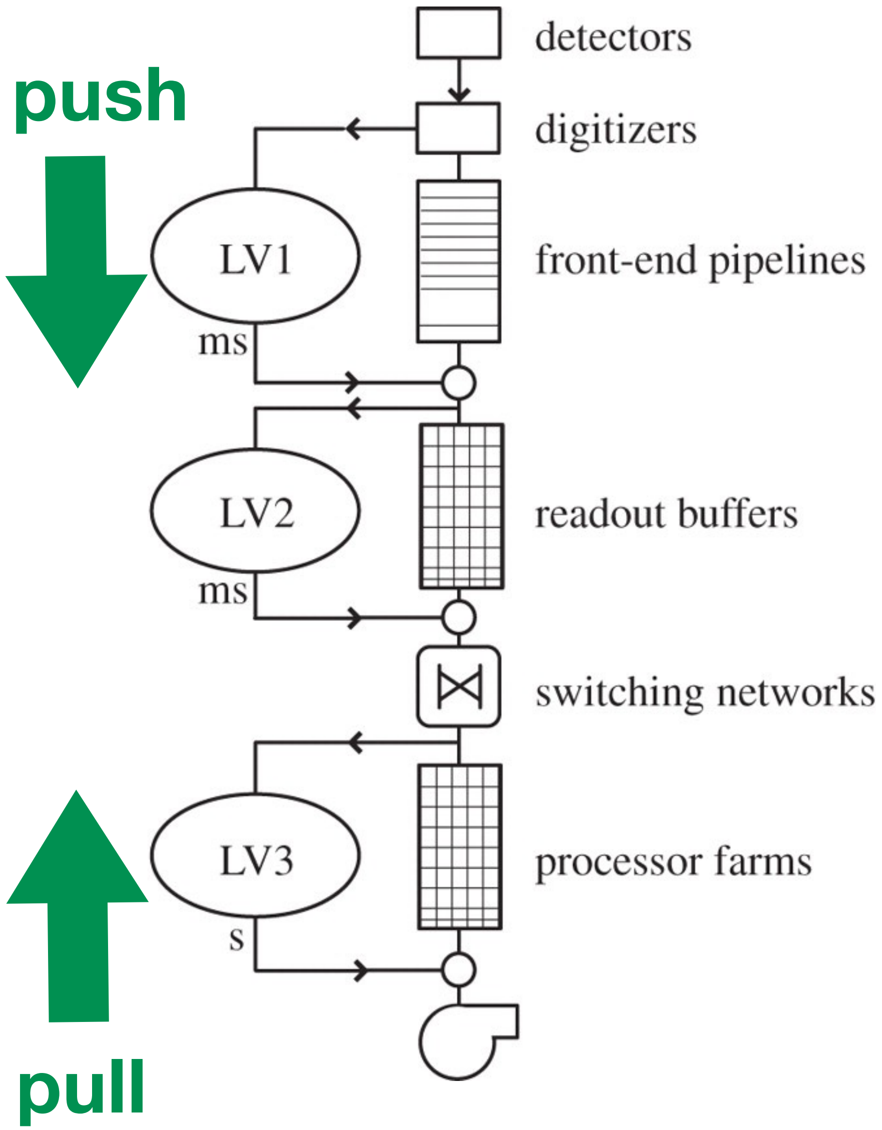
- More channels + more rate + more data to process online ==> longer latency
  - single level trigger not enough
- Add **High level triggers** with longer latency
  - more complex filters
  - more data (for example silicon detectors)



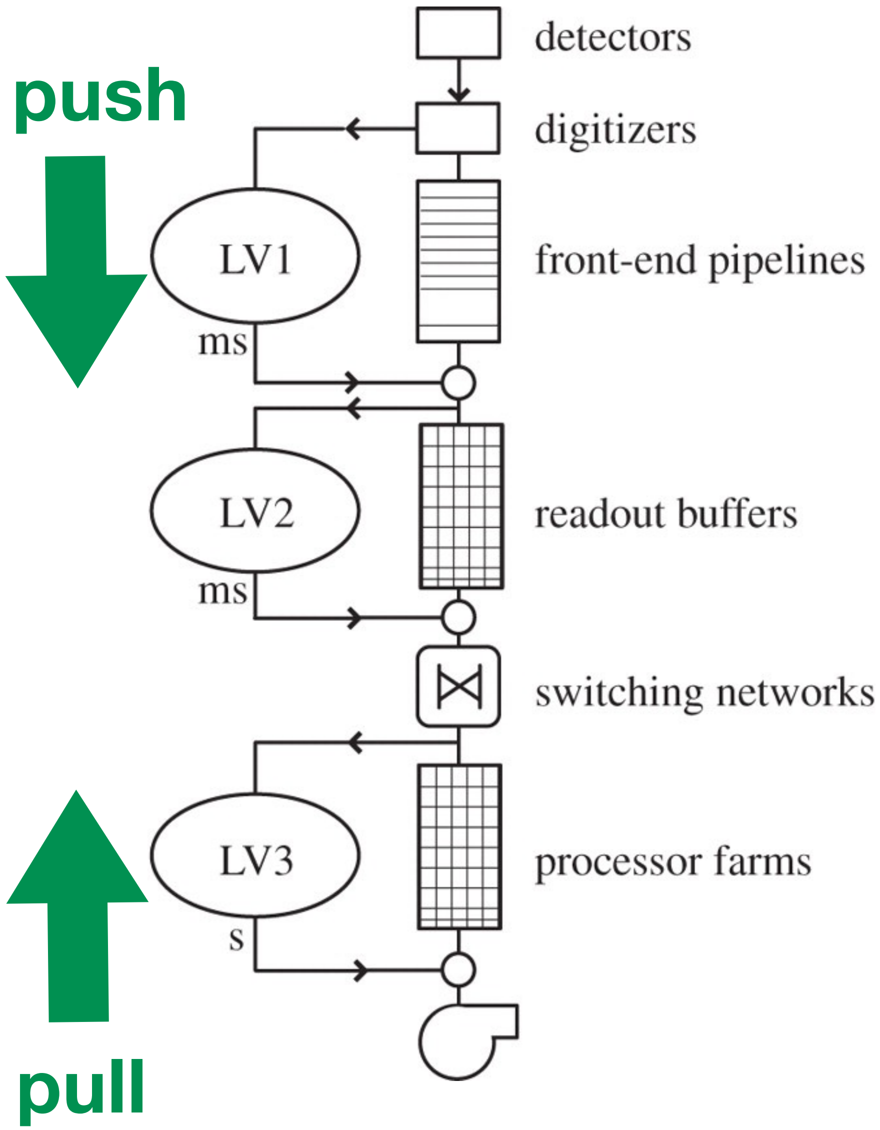
- **Recall on trigger system architectures**
- **Real time system**
  - must respond within some **fixed latency**
  - → Latency = Max Latency
  - → over fluctuations bad, will create deadtime
- **Non-real-time system**
  - responds as soon as it's available
  - → Latency = **Mean Latency**
  - → over fluctuations fine, shouldn't create deadtime

- **LEP**
- 10^5 channels
- 22µs crossing rate – no event overlap
- single interaction
- L1 ~10^3 Hz
- L2 ~10^2 Hz
- L3 ~10 Hz
- 100kB/ev → 1MB/s

5 – DATAFLOW CONTROL

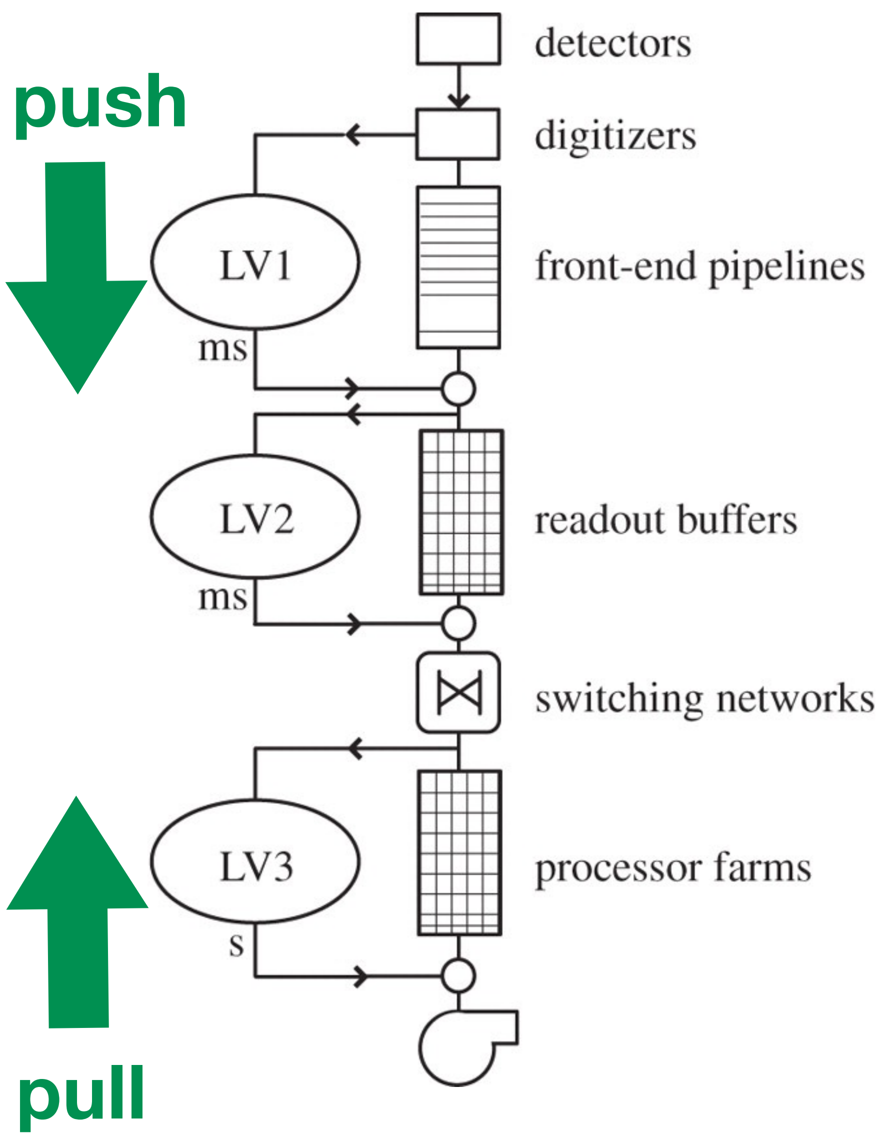


5 – DATAFLOW CONTROL



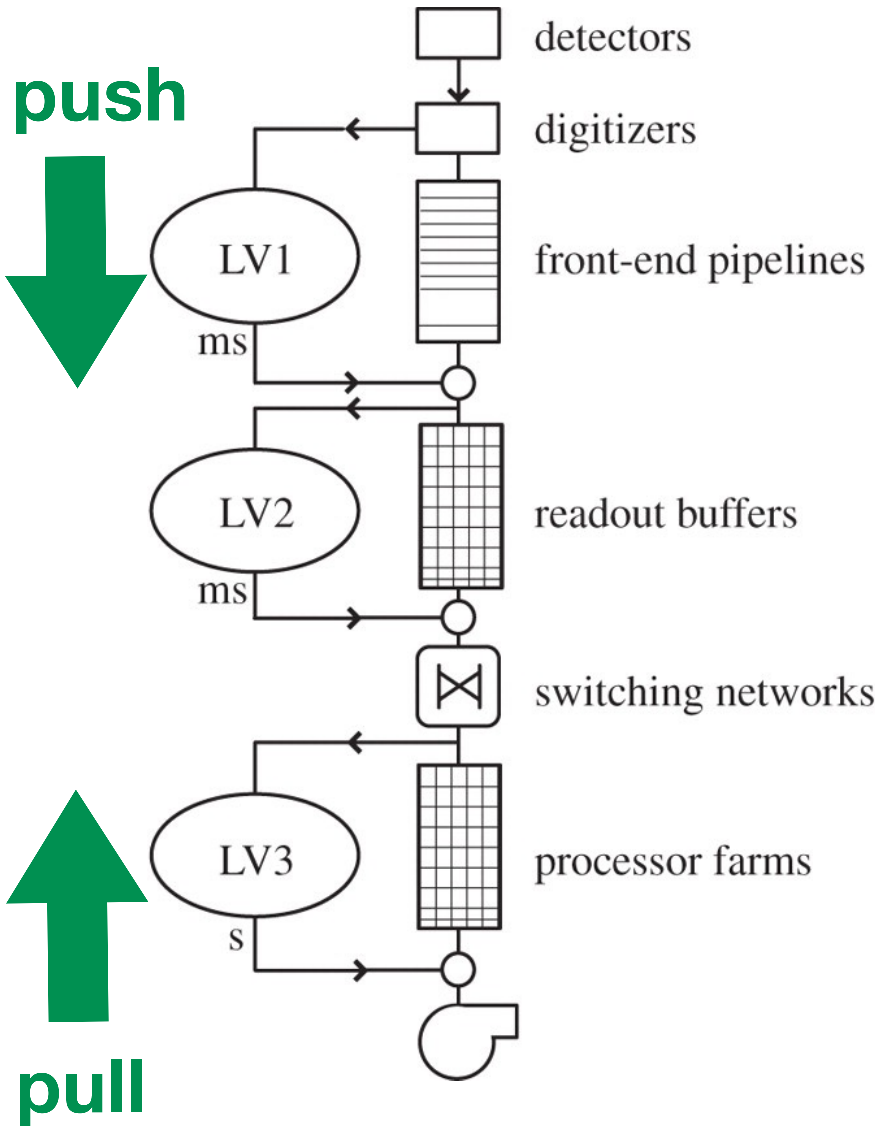
- ➡ **Buffers are not the "final solution"**
  - ➡ Can overflow, with bursts and unusual event sizes
  - ➡ In these cases
    - ➡ discard data locally or
    - ➡ exert "**back-pressure**", i. e. ask previous level(s) to block dataflow

5 – DATAFLOW CONTROL



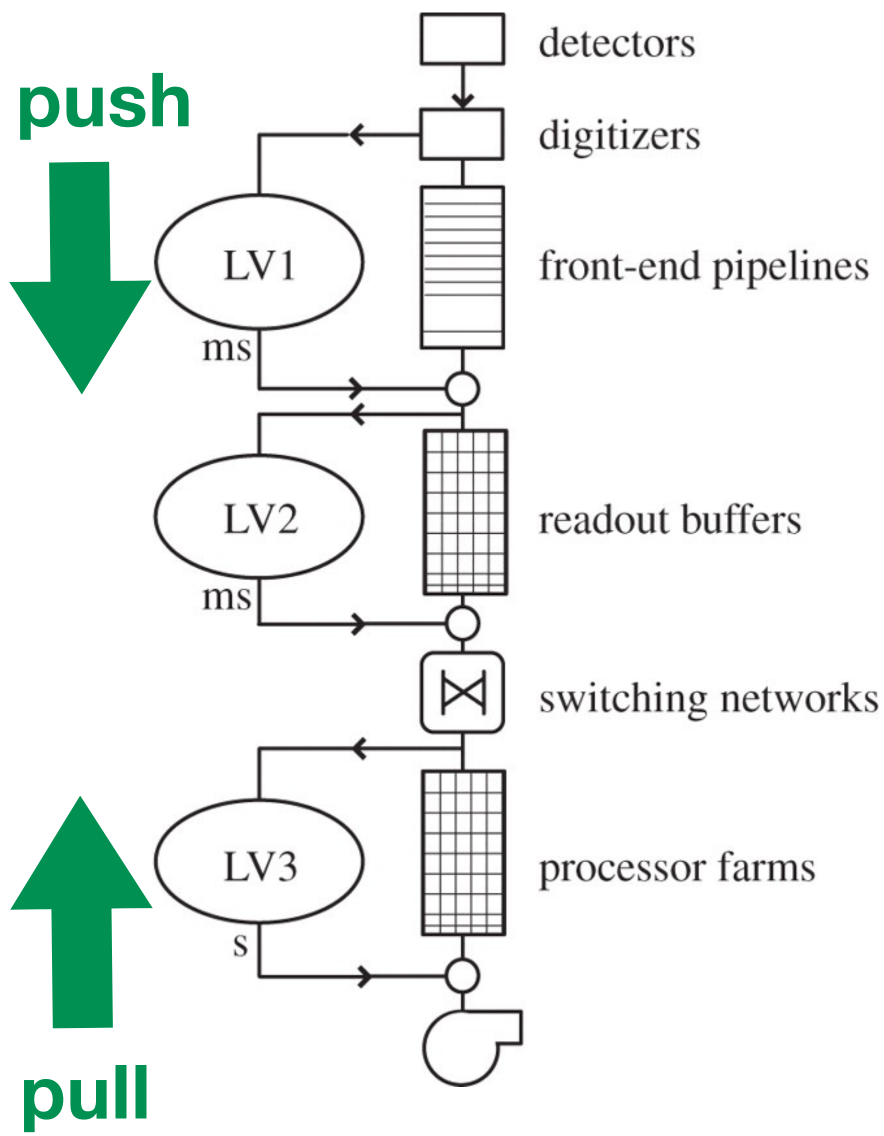
- **Buffers are not the "final solution"**
  - Can overflow, with bursts and unusual event sizes
  - In these cases
    - discard data locally or
    - exert "**back-pressure**", i. e. ask previous level(s) to block dataflow
- **Throughput optimization means avoiding dead-time due to back-pressure**
  - using knowledge of the input buffer state

5 – DATAFLOW CONTROL



- **Buffers are not the "final solution"**
  - Can overflow, with bursts and unusual event sizes
  - In these cases
    - discard data locally or
    - exert "**back-pressure**", i. e. ask previous level(s) to block dataflow
- **Throughput optimization means avoiding dead-time due to back-pressure**
  - using knowledge of the input buffer state
- **Who controls the flow?**

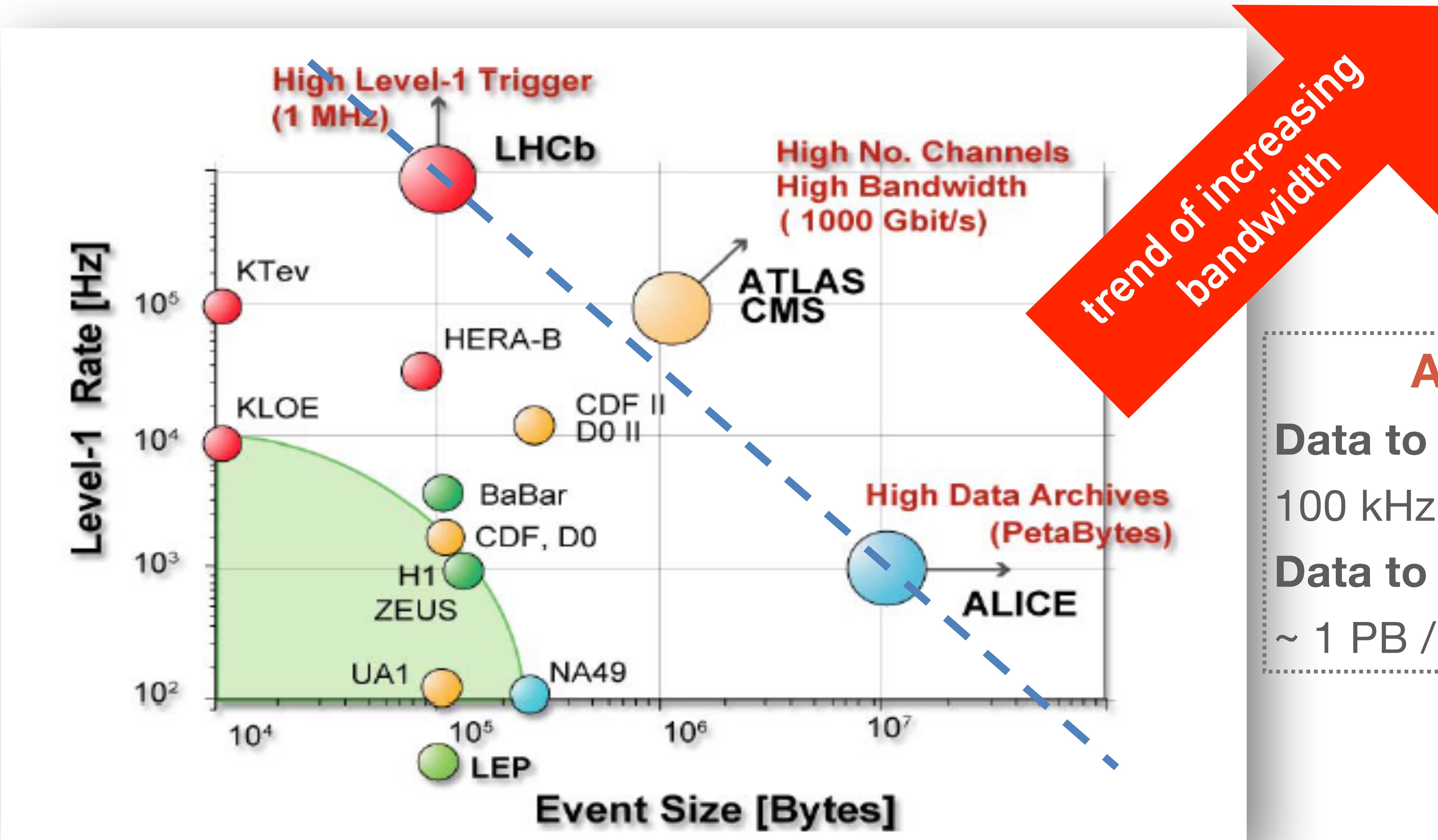
5 – DATAFLOW CONTROL



- **Buffers are not the "final solution"**
  - Can overflow, with bursts and unusual event sizes
  - In these cases
    - discard data locally or
    - exert "**back-pressure**", i. e. ask previous level(s) to block dataflow
- **Throughput optimization means avoiding dead-time due to back-pressure**
  - using knowledge of the input buffer state
- **Who controls the flow?**
- **FE (push) or EB (pull)**
  - **Push**: Events are sent as soon as data are available to the sender (for example round-robin algorithm) ==> Busy or Throttle
  - **Pull** : events are required by a given destination processes (may need an event manager) ==> back-pressure
  - **Push-Pull** ==> busy and back-pressure

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

faster L1 electronics



more channels, more complex events

ATLAS/CMS

Data to Process:

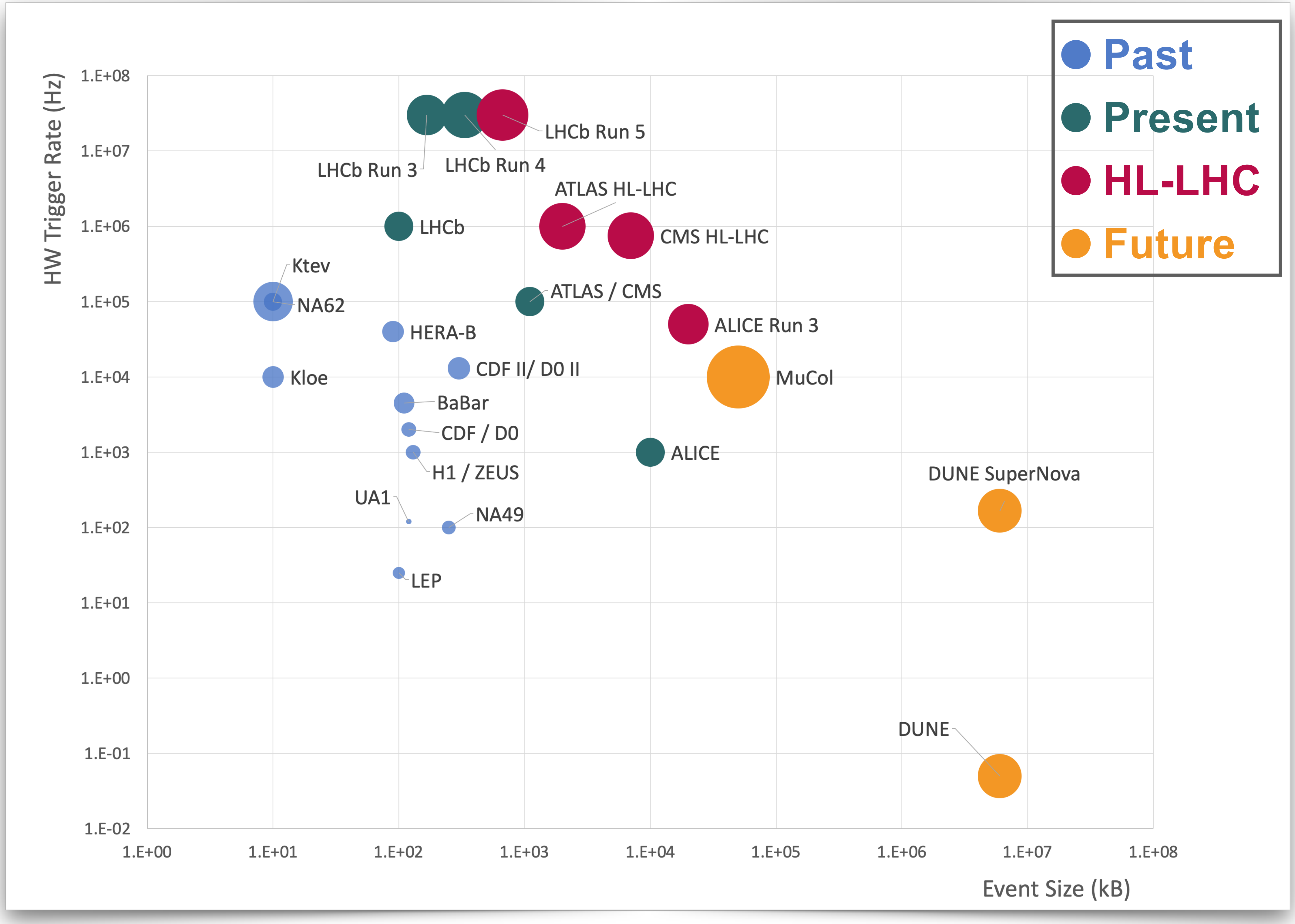
100 kHz * 1 MB = 100 GB/s

Data to Store:

~ 1 PB / year /experiment

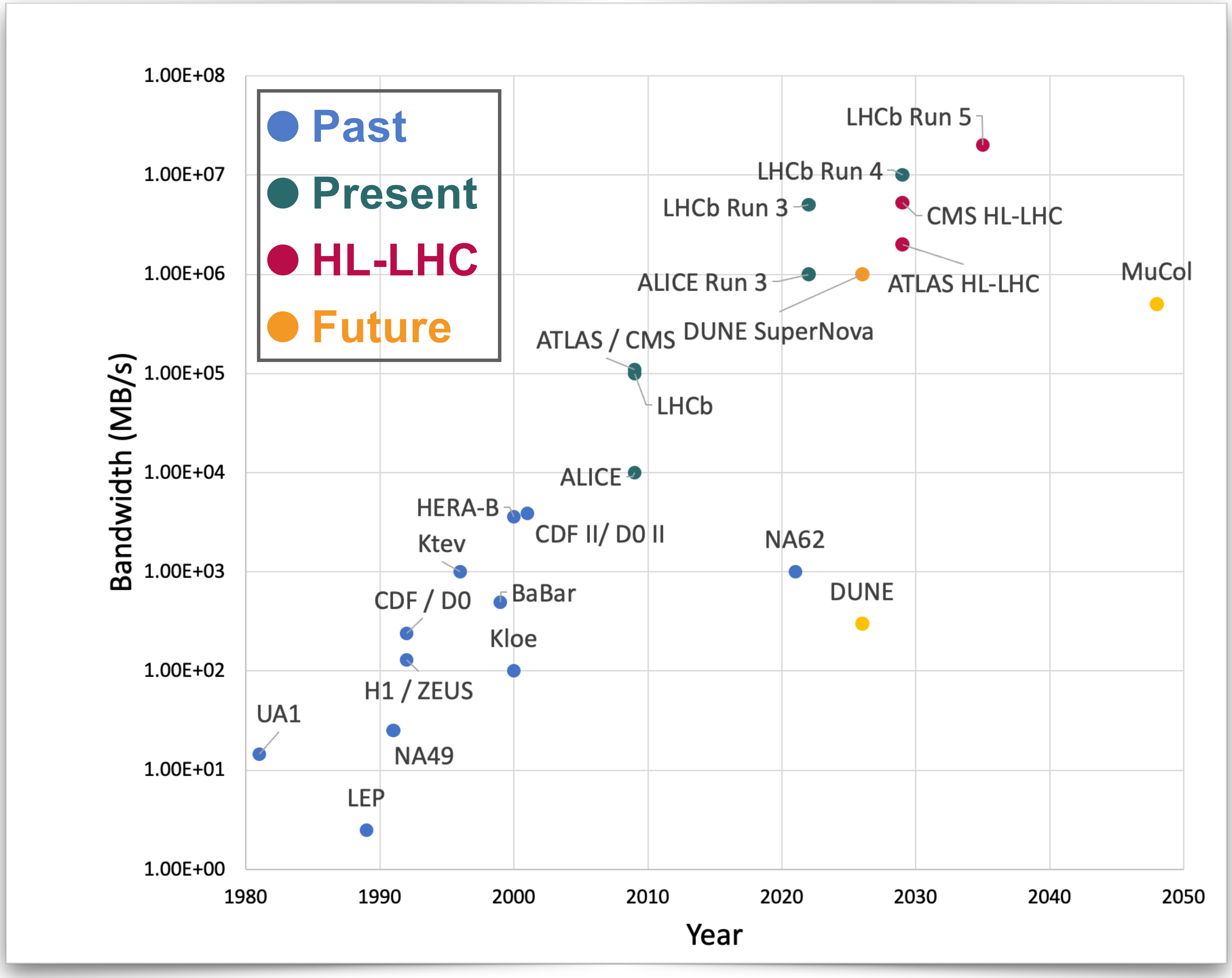
As the data volumes and rates increase, new architectures need to be developed

UPDATED FIGURE!



Courtesy of A.Cerri

LOOKING FOR MORE DATA IN THE FUTURE

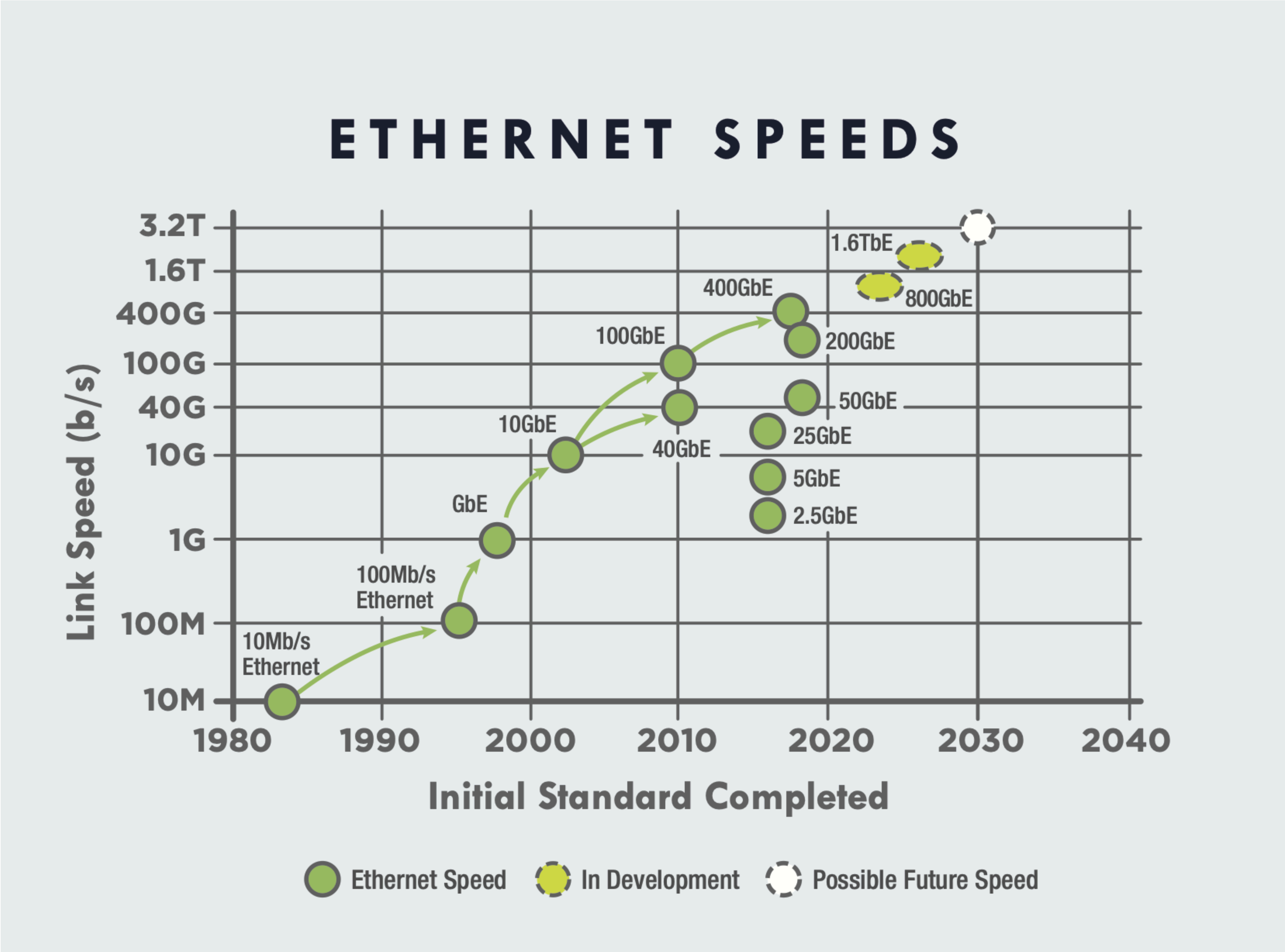


Courtesy of A.Cerri

GENERAL T/DAQ TRENDS

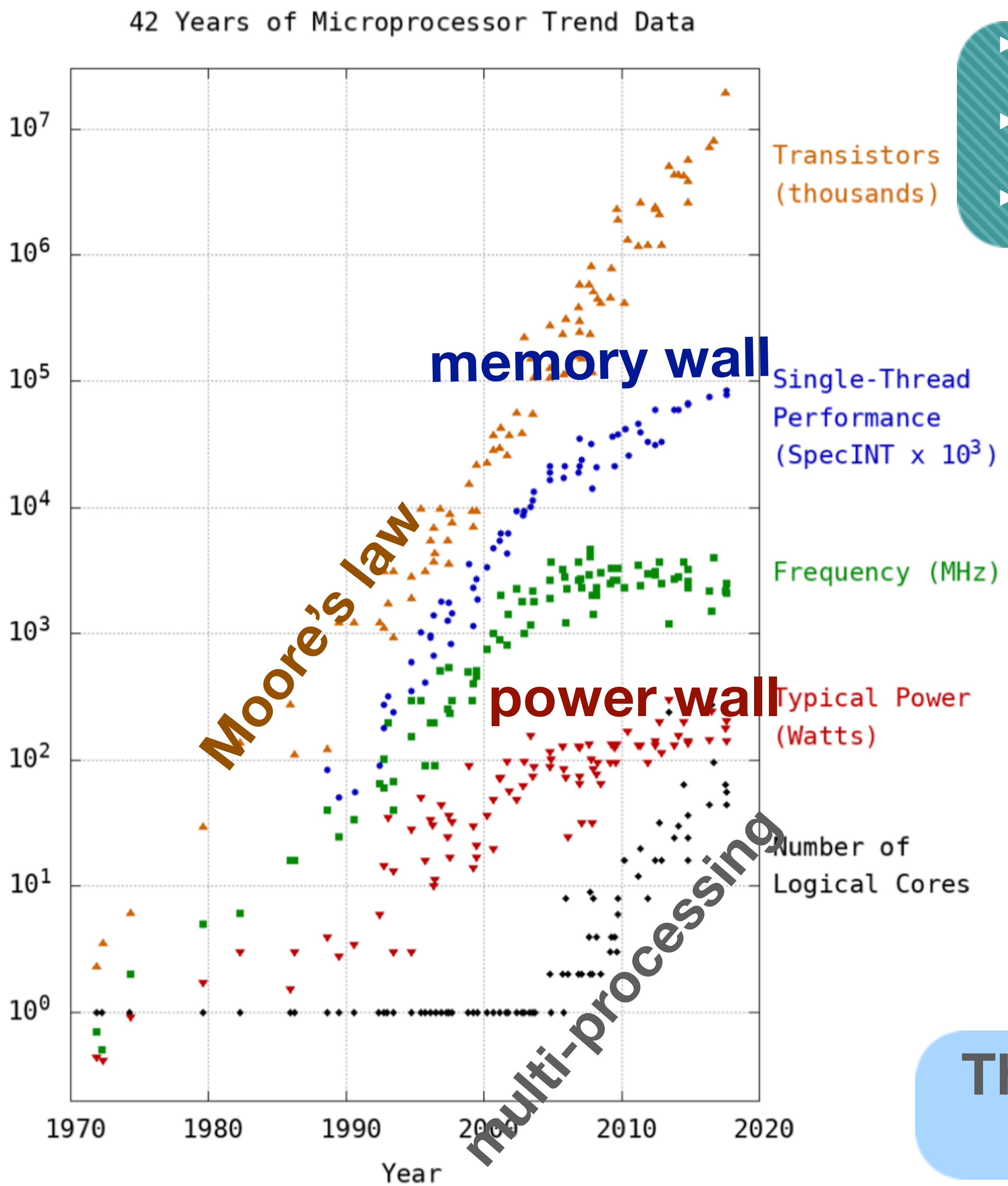
- **Increasing readout channels, and front-end cards, distributed in multi-level three structure**
- **Integrate synchronous low-latency in Front-End**
  - limitations do not disappear, but decouple (factorise)
- **Deal with dataflow instead of latency**
  - **decouple** DAQ from High Level Triggers
  - decouple dataflow from storage, with temporary buffers
  - Use COTS network and processing
- **Use networks as soon as possible**
  - toward commercial bidirectional point-to-multipoint architecture
- **Use "network" design already at small scale**
  - easily get high performance with commercial components
- **Increase data aggregation at the Event Building**
  - reducing request rates on DAQ software
  - per time-frame, per orbit instead of per-event

CLEAR WHY?



NEXT ETHERNET ERA
ethernet alliance

EVOLUTION OF PROCESSING POWER 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, with more complicated programming**
  - ➡ Vectorisation, low-level memory…

➡ **Multithreading processing**
  - ➡ To reduce memory footprint

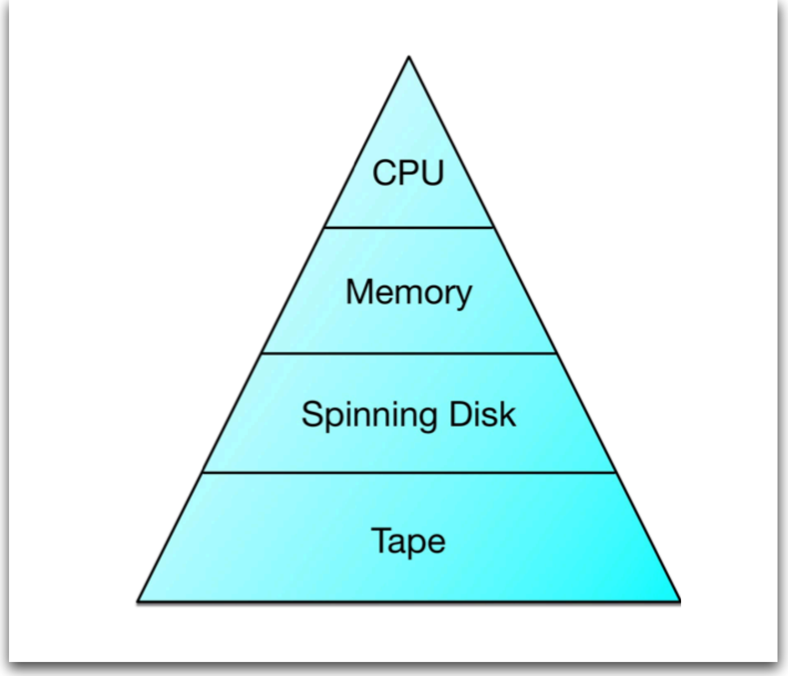
➡ **Use of co-processors:**
  - ➡ High Performance Computing (HPC) often employ GPU architecture to achieve record-breaking results!

This requires fundamental re-write/ optimization of our software

Read: HPC computing

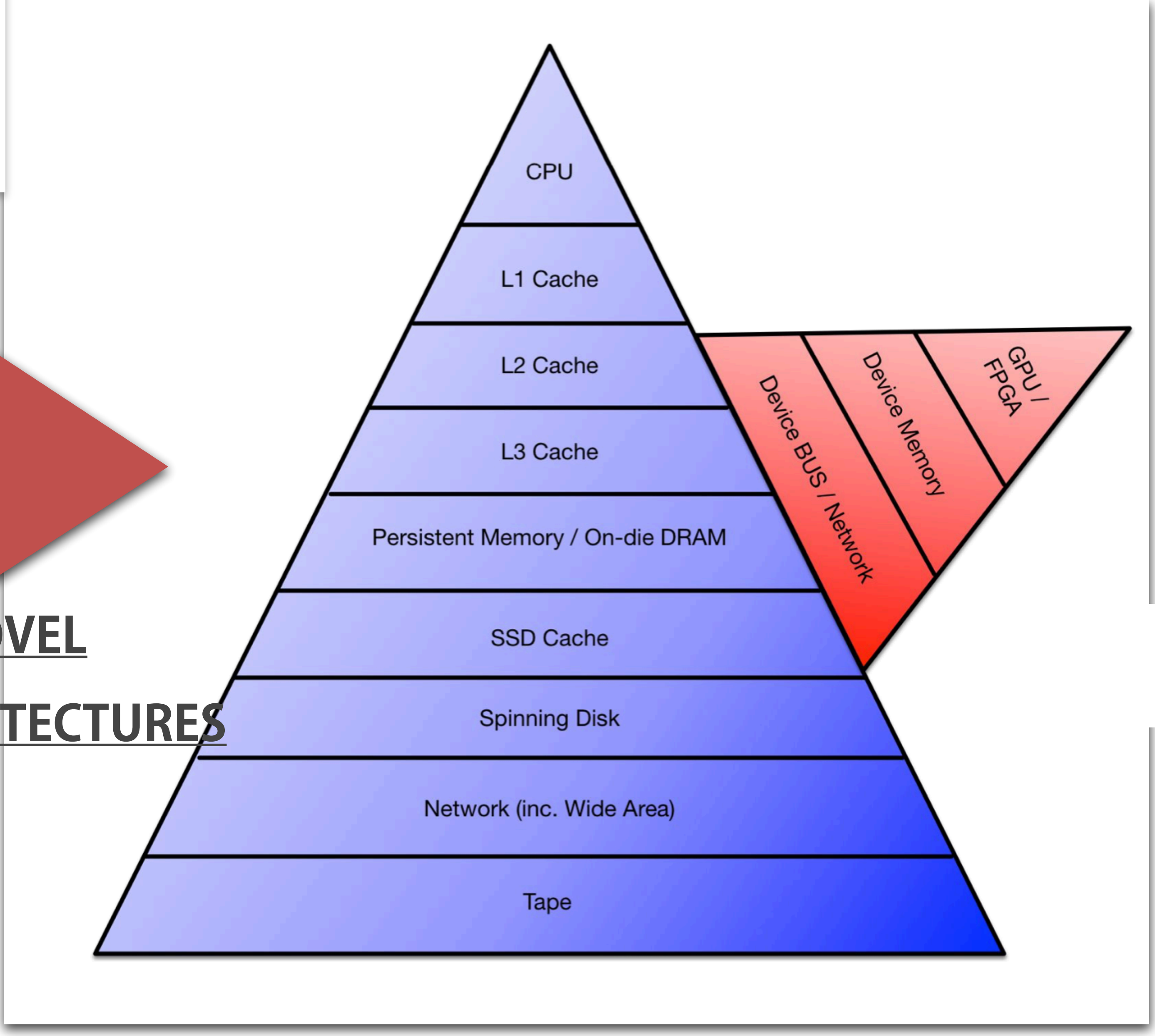
(TRIGGER) SOFTWARE EVOLUTION TO BREAK WALLS

"We're approaching the limits of computer power – we need new programmers now"
John Naughton, Guardian



EXPLOSION OF NOVEL COMPUTER ARCHITECTURES

See LP-2022 slides from Graeme Stewart



- ➡ **Exploiting CPU hardware in new architectures**
  - ➡ more complicated programming (vectorisation, memory sharing…)
- ➡ **Exploit more efficiently instruction level parallelism (ILP)**

EXASCALE COMPUTING

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

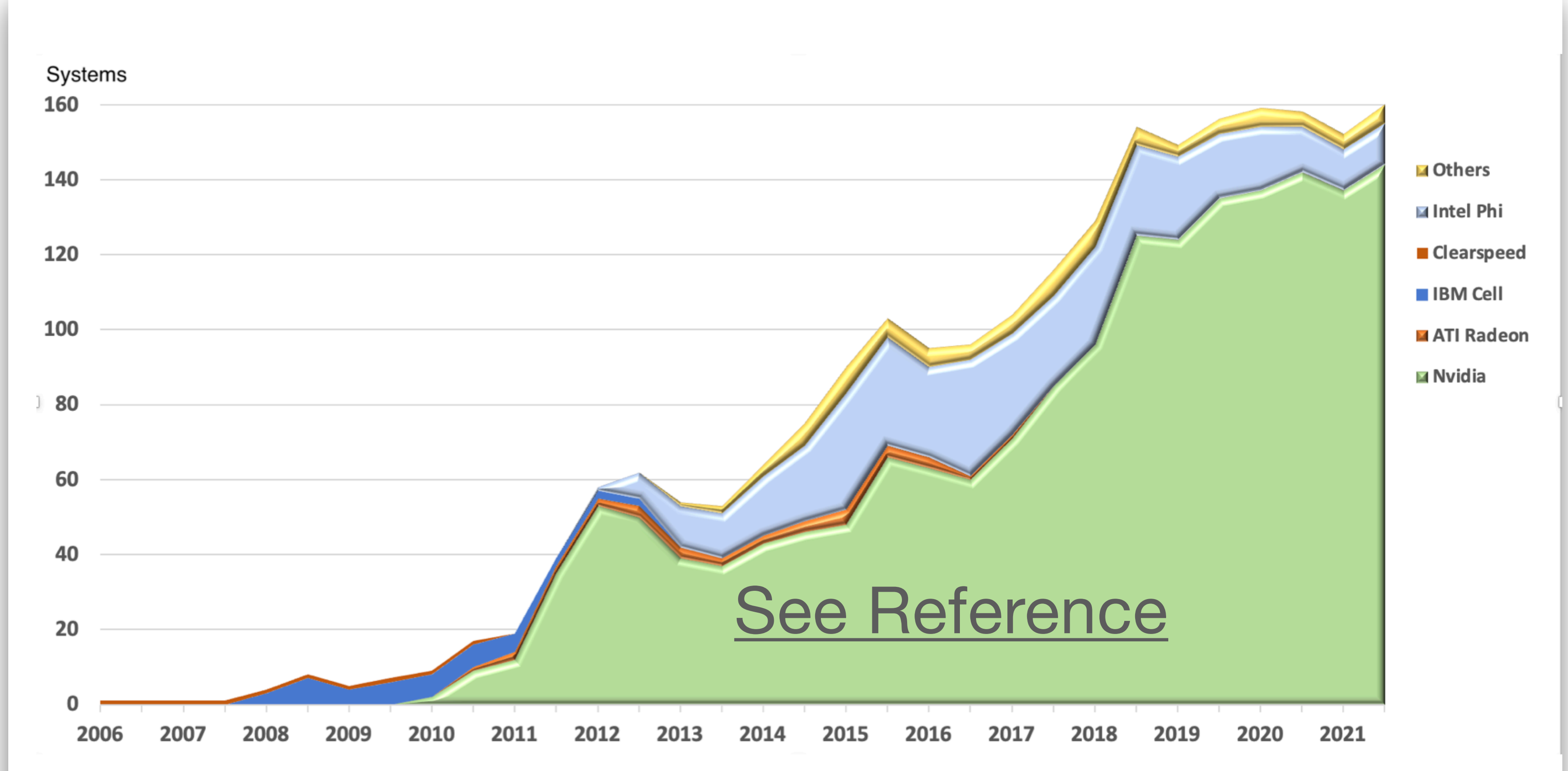
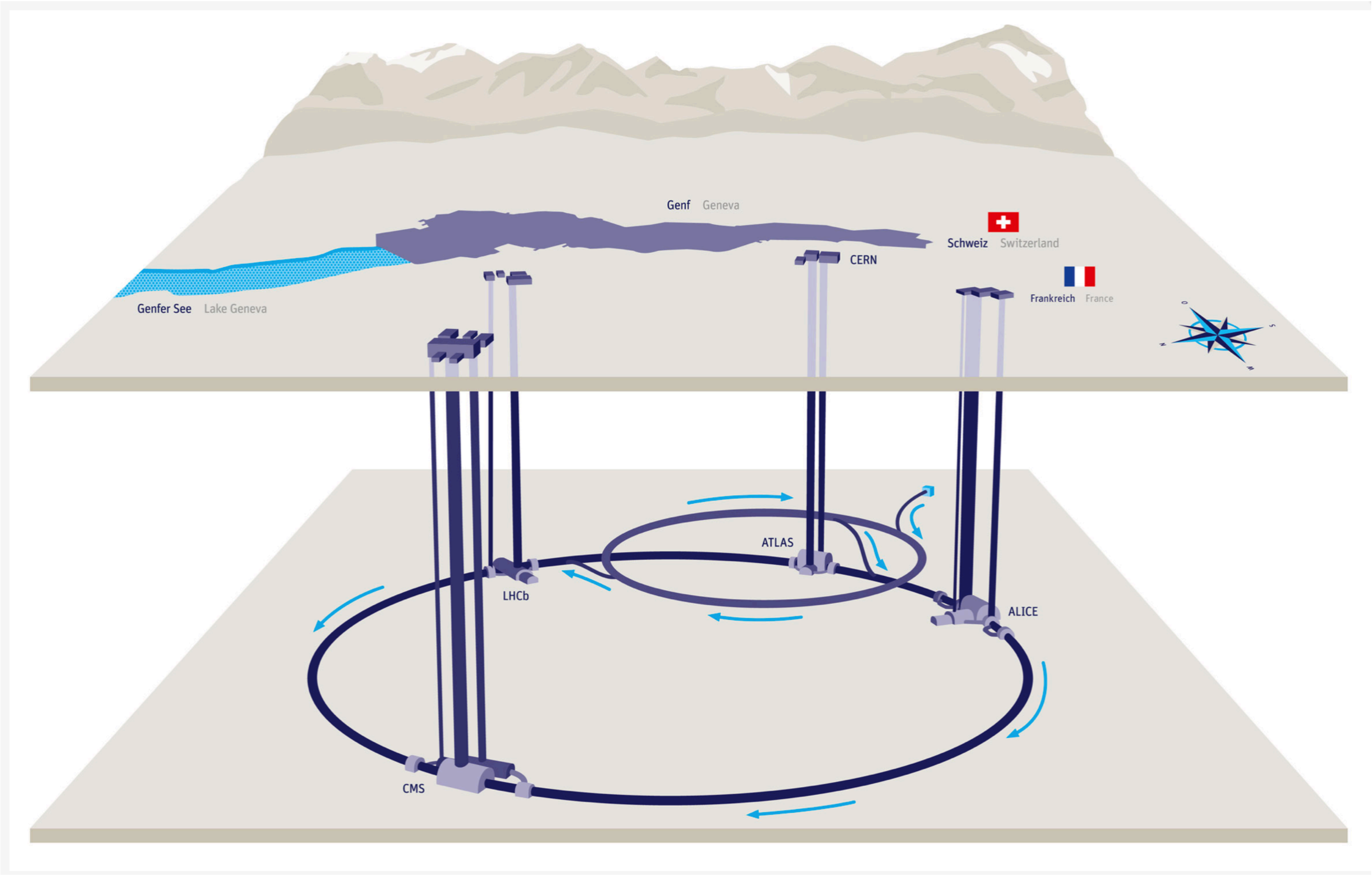


Figure 5: Systems Using GPU Accelerators on the TOP500

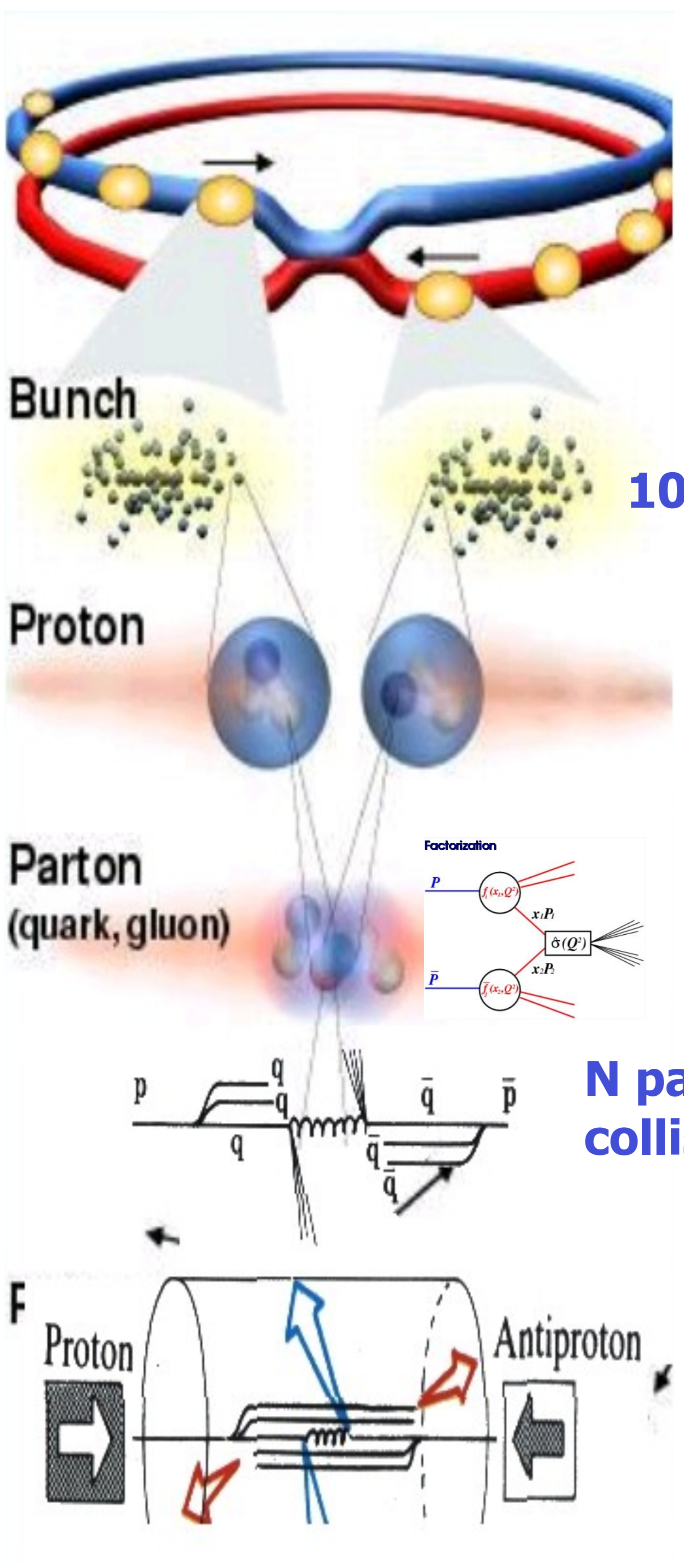
TOP500 today largely examples of a commodity monoculture: nodes with server-class microprocessors + GPUs

OUTLINE

- ➡ Examples of small experiments with their limits
- ➡ **Overview of LHC experiments and their upgrade**
- ➡ Future TDAQ systems (Dune/Proto-Dune)



LHC ENGINE AND ITS CHALLENGES



2835×2835 bunches in the LHC ring

10^{11} protons / bunch

≤30 pp collisions per bunch crossing (BC)

N parton-parton collisions / pp collision

Complex final-states in every parton-parton collision.

E_{cms} = 14 TeV
L = 10^{34} /cm² s
BC clock = 40 MHz

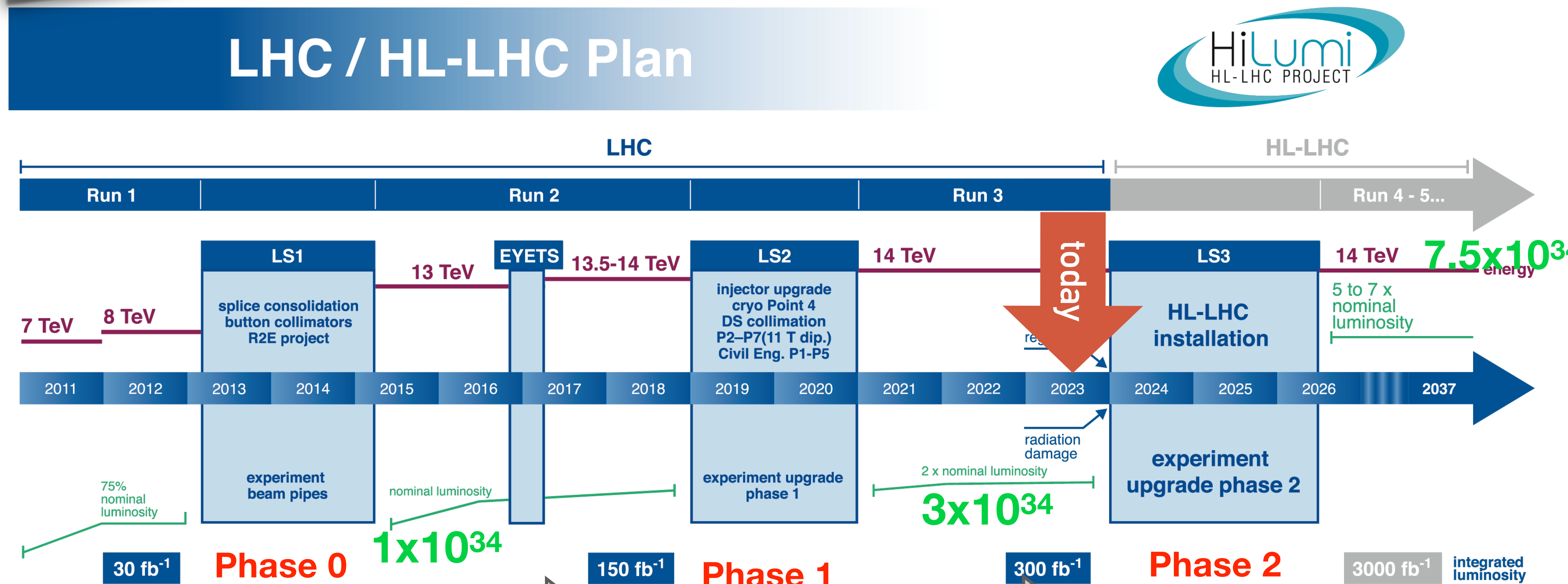
Search for rare events overwhelmed in abundant low-energy particles

Three major challenges for T/DAQ

- **Face High Luminosity**:
  - **fast electronics,** to resolve in time
  - fine granularity detector, to resolve in space ➠ **high data volume**
- **Search for rare physics**:
  - high rejection or large data collection
- **Be radiation resistant:**
  - very costly for electronics ==> survive up to 100 Mrad= 1 MGy

LHC 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 the initial design specifications ($1x10^{34}/cm^2s$) and need upgrade to improve, or at least maintain, the design performance

LHC EXPERIMENTS FOR A DISCOVERY MACHINE

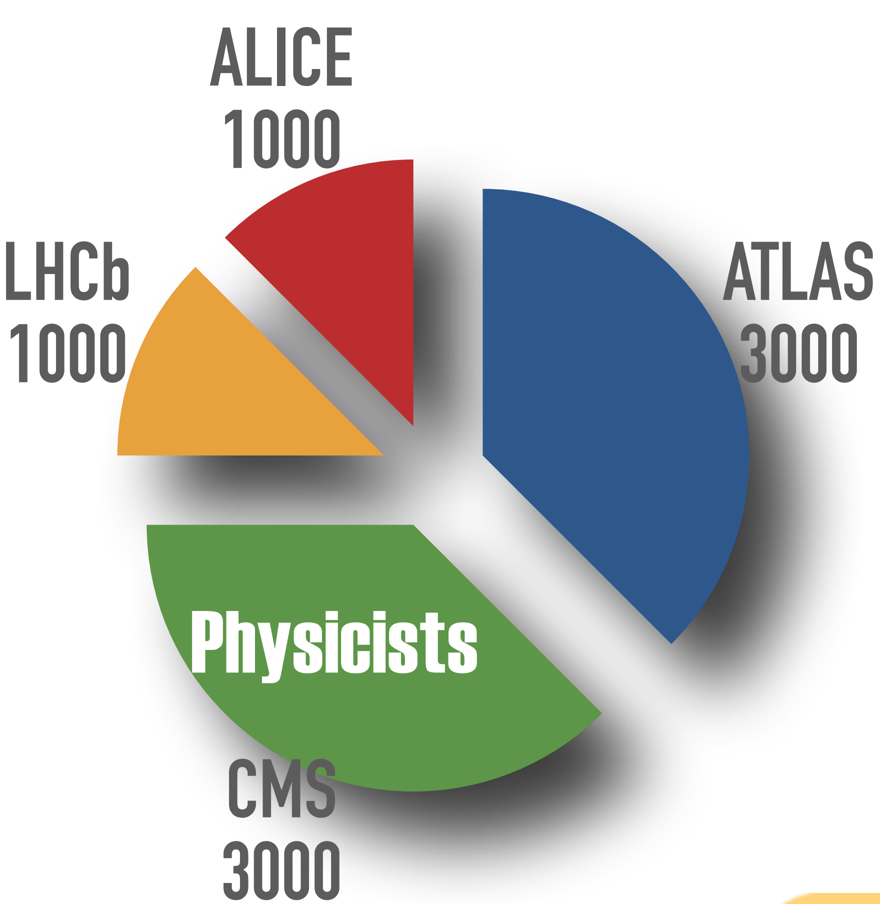
Goal: explore TeV energy scale to find New Physics beyond Standard Model

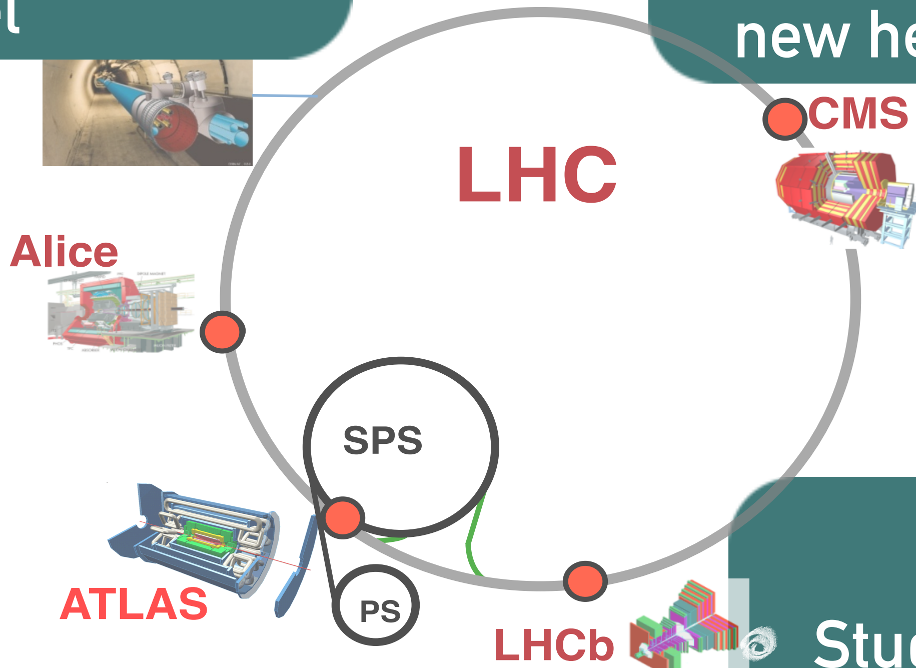
ATLAS & CMS

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

LHCb

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





ALICE

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

Proposed: 1992, Approved: 1996, Started: 2009

LHC EXPERIMENTS FOR A DISCOVERY MACHINE

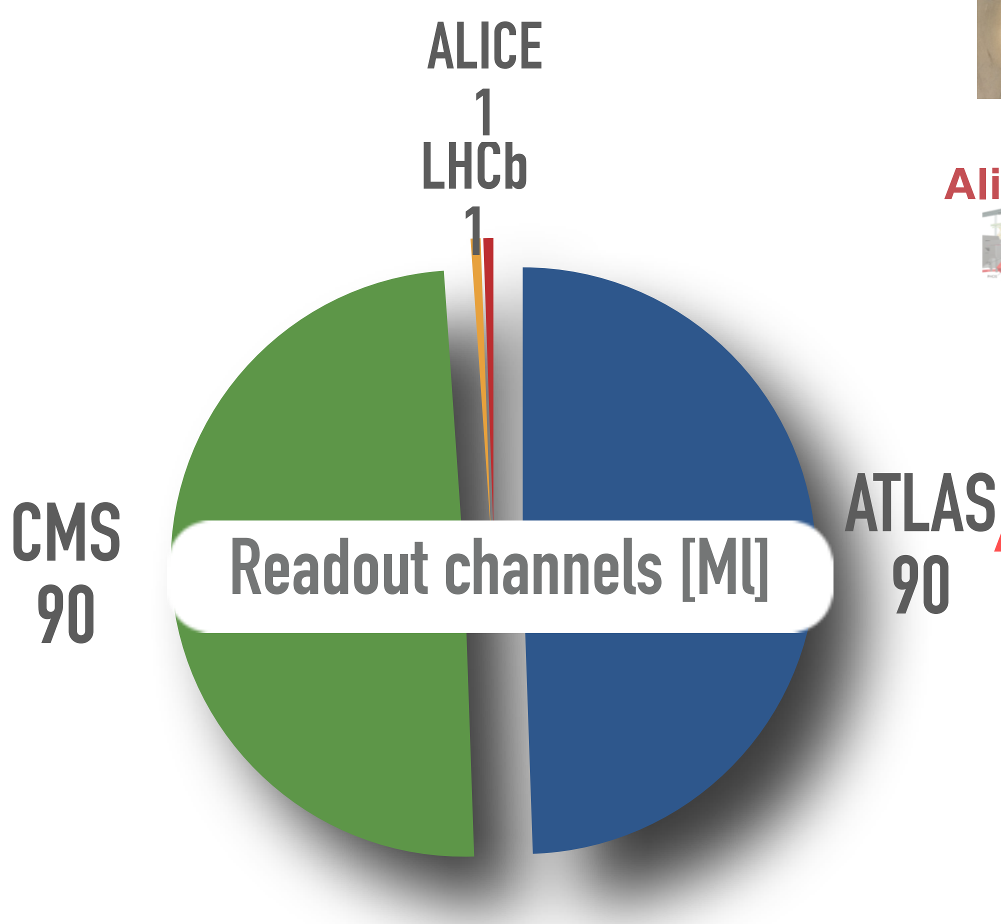
Goal: explore TeV energy scale to find New Physics beyond Standard Model

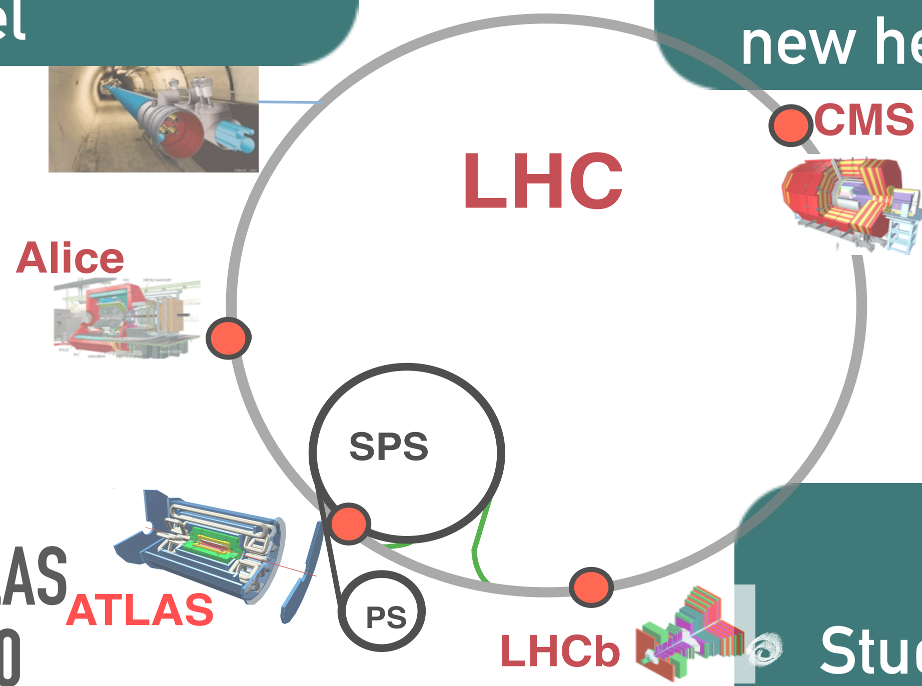
ATLAS & CMS

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

LHCb

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





ALICE

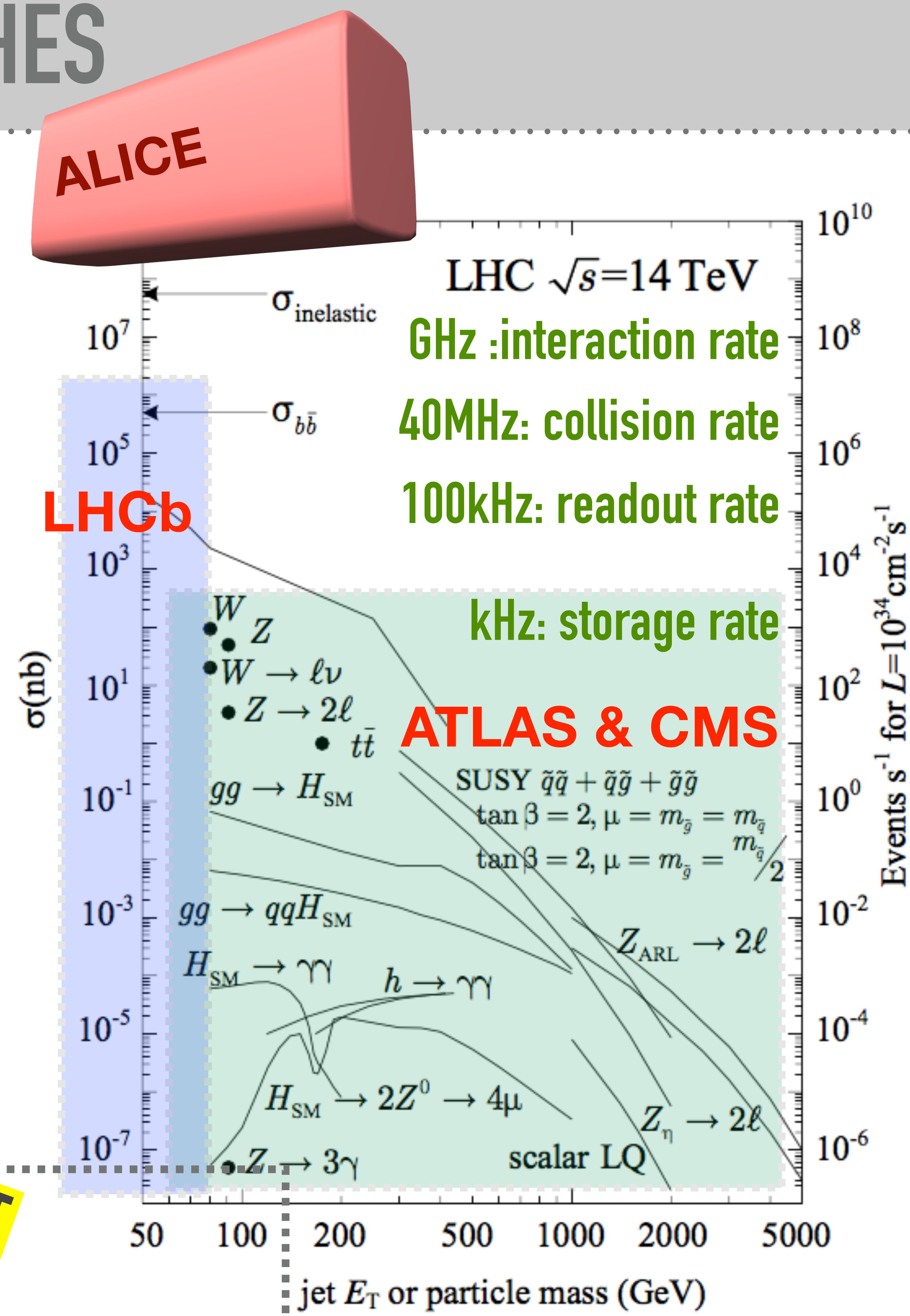
- Studying quark-gluon plasma, a complex system of strongly interacting matter produced by heavy ion collisions

Proposed: 1992, Approved: 1996, Started: 2009

DIFFERENT PHYSICS SEARCHES

…. and LHC operations

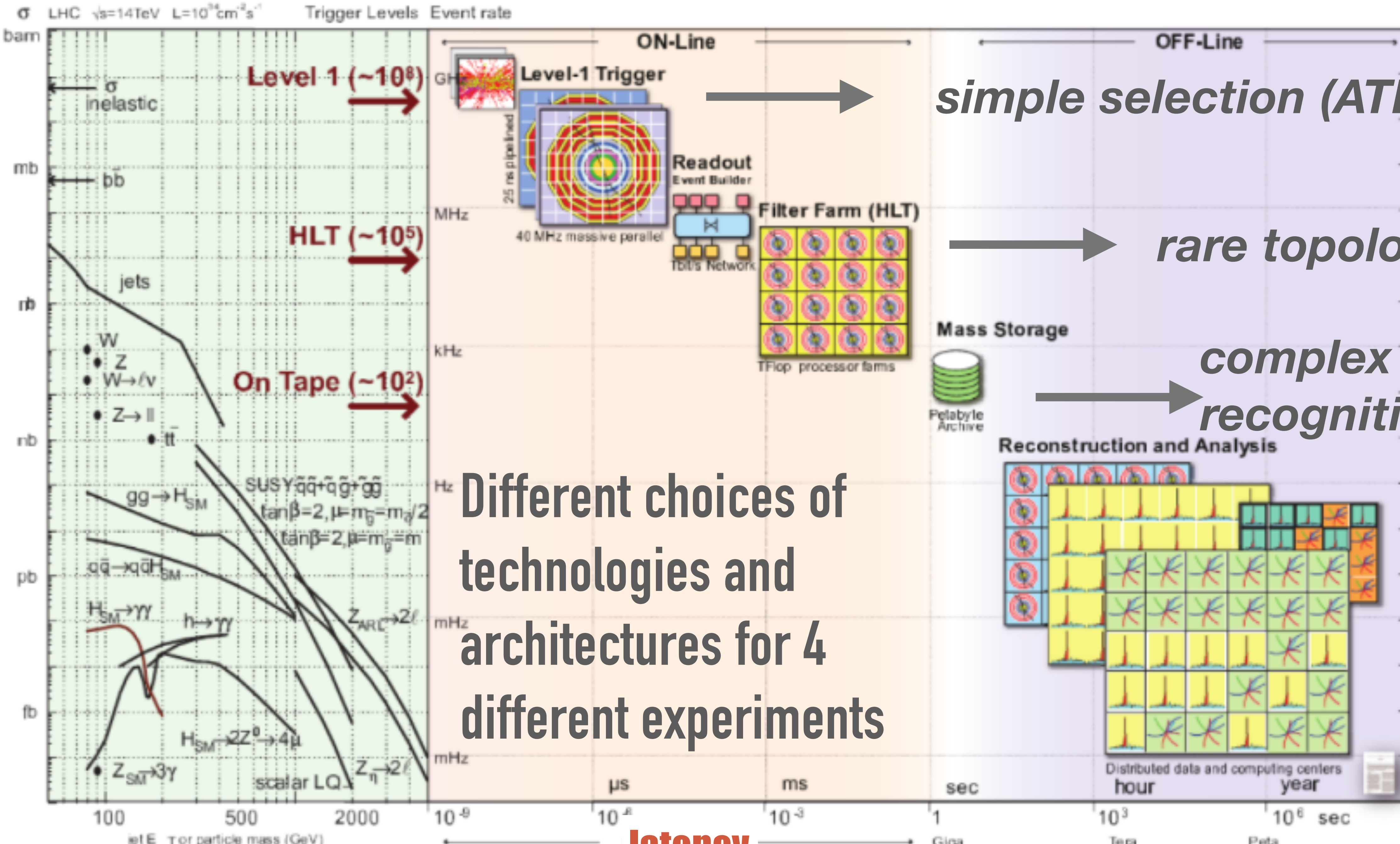
- **ATLAS/CMS: p-p collisions at full Luminosity**
  - search in high energy scale
- **LHCb: p-p collisions at reduced Luminosity**
  - search complex topologies of b-quark decays
- **ALICE: heavy-ion collisions ~2000 mb**
  - search in high energy density



DIFFERENT

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

ENHANCED TRIGGER SELECTIONS



data rates

latency

simple selection (ATLAS, CMS)

rare topology (LHCb)

complex pattern recognition (ALICE)

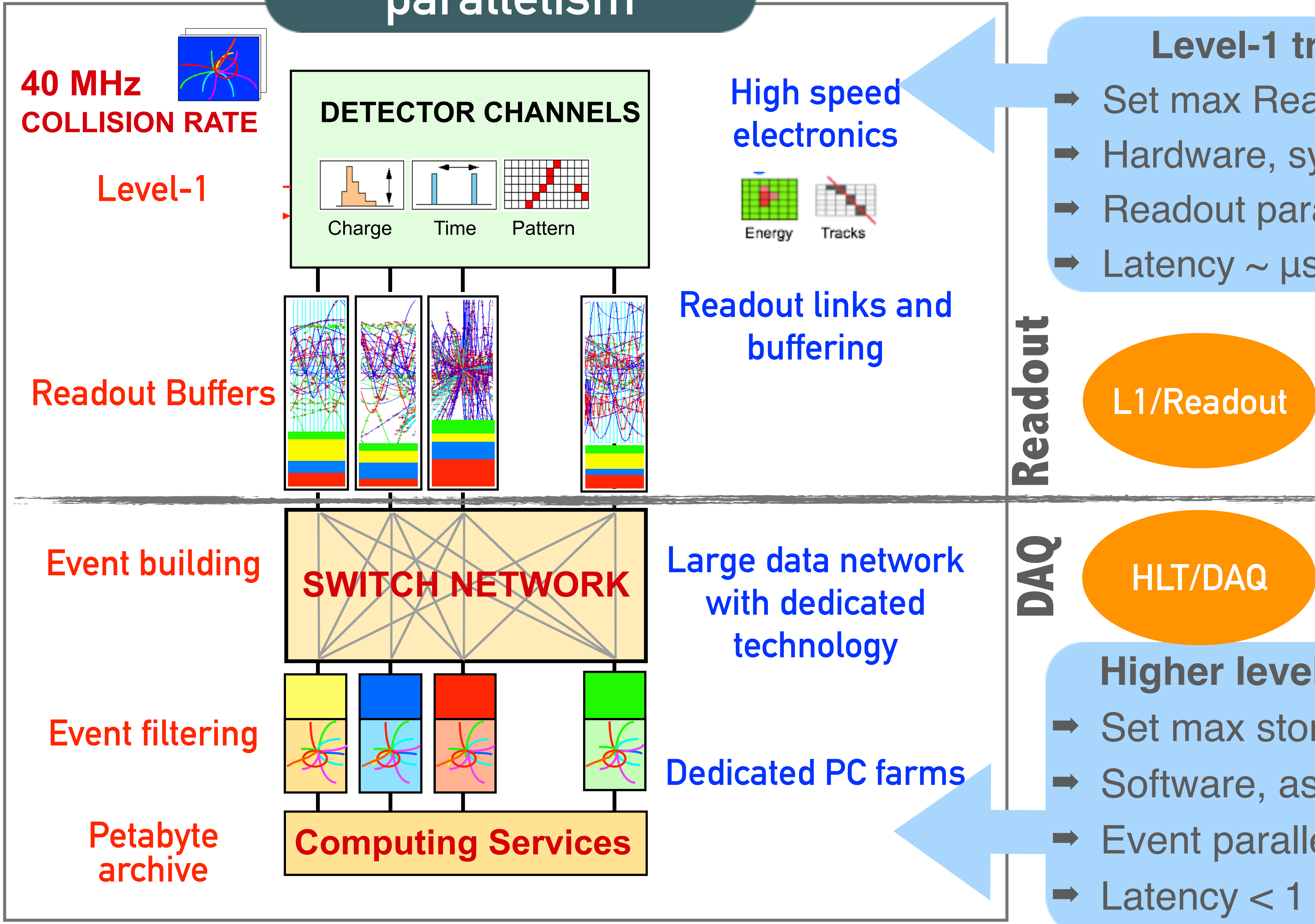
Different choices of technologies and architectures for 4 different experiments

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

MANY PLAYERS, COMPLEX TDAQ ARCHITECTURES

Buffering and parallelism

Maximum 1-2% deadtime



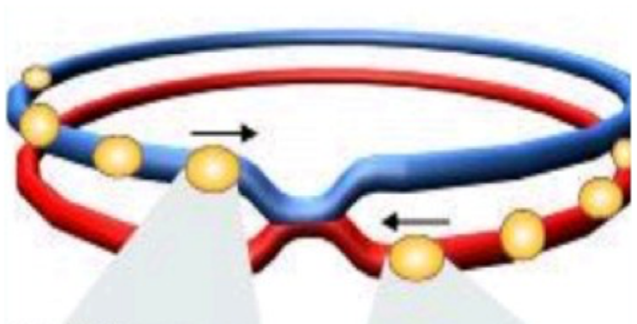
Level-1 triggers

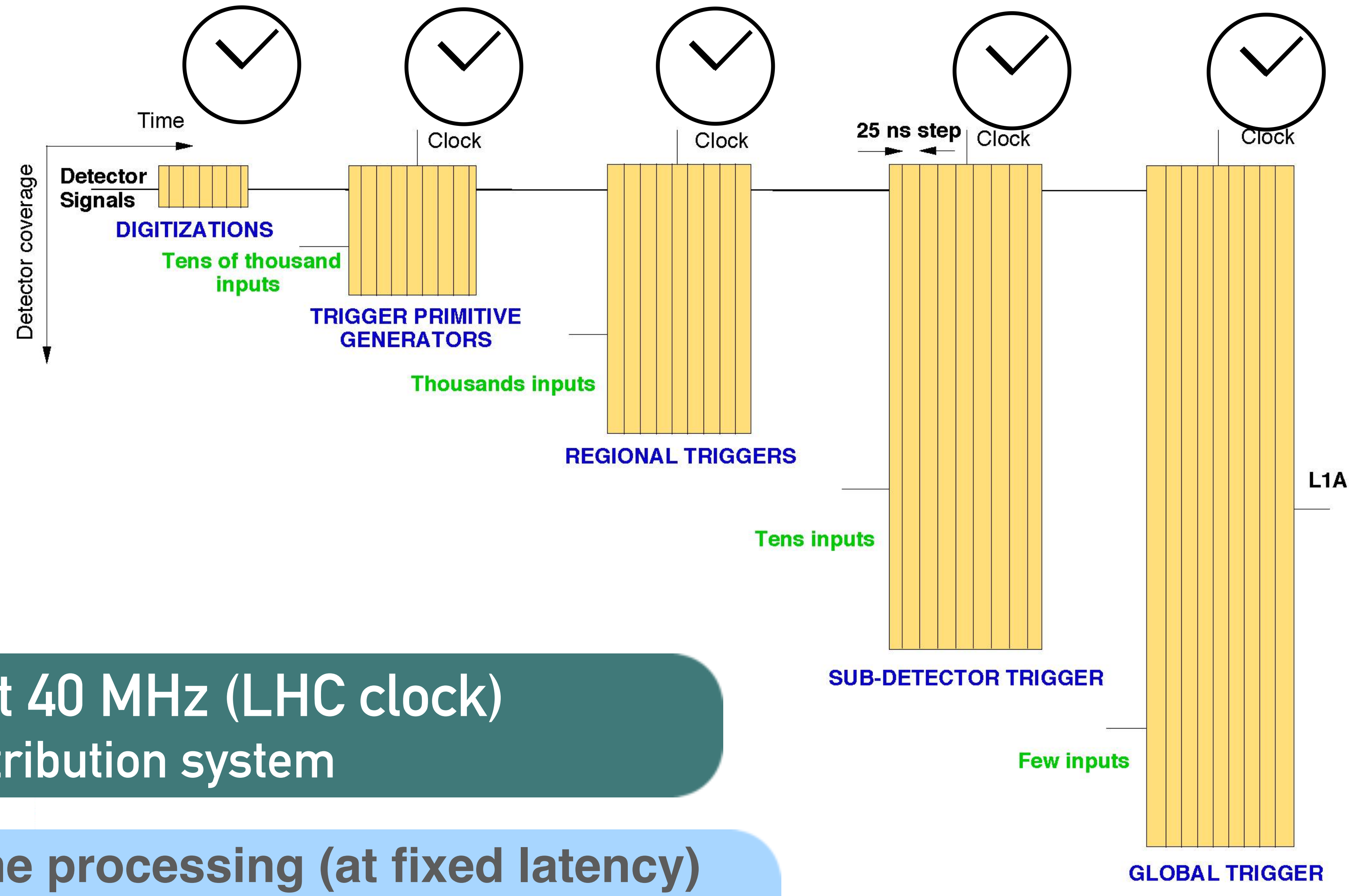
- ➡ Set max Readout rate
- ➡ Hardware, synchronous
- ➡ Readout parallelism
- ➡ Latency ~ μsec/event

Higher level triggers

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

LEVEL-1 TRIGGER REQUIREMENTS





Full synchronisation at 40 MHz (LHC clock)

- ➤ large optical time distribution system

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

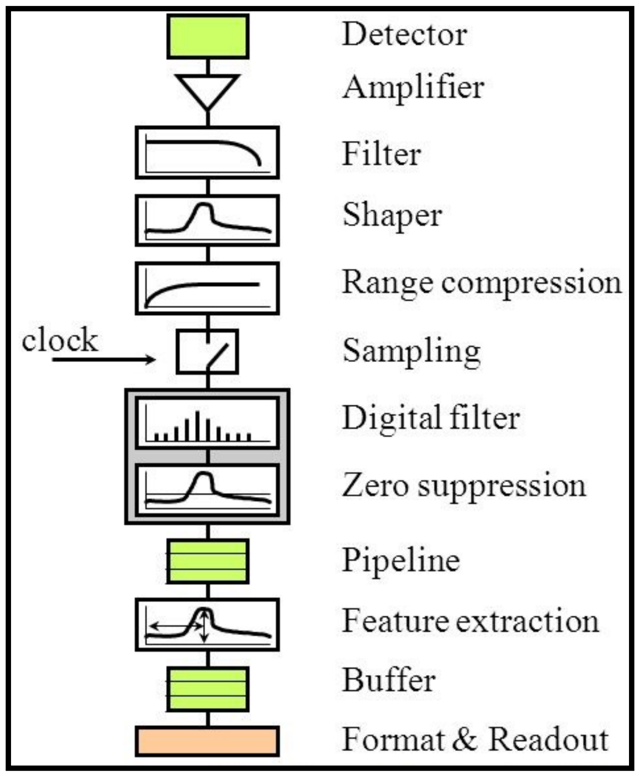
ALICE	No pipeline
ATLAS	2.5 µs
CMS	3 µs
LHCb	4 µs

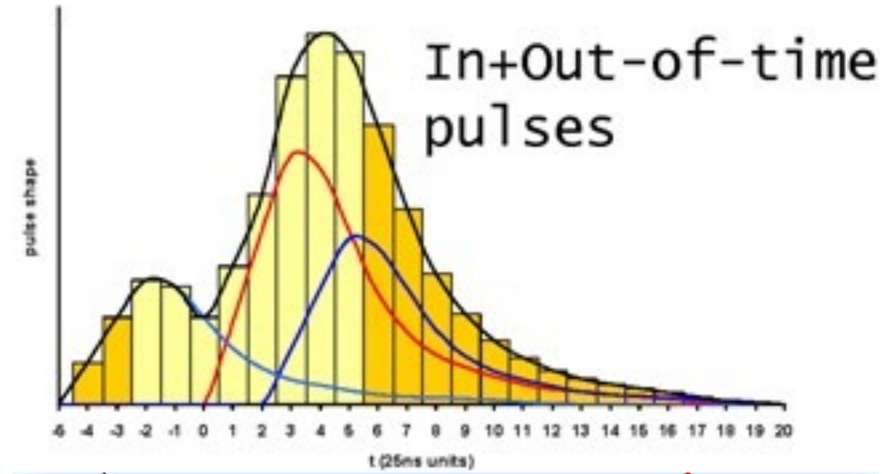
Fast, robust electronics

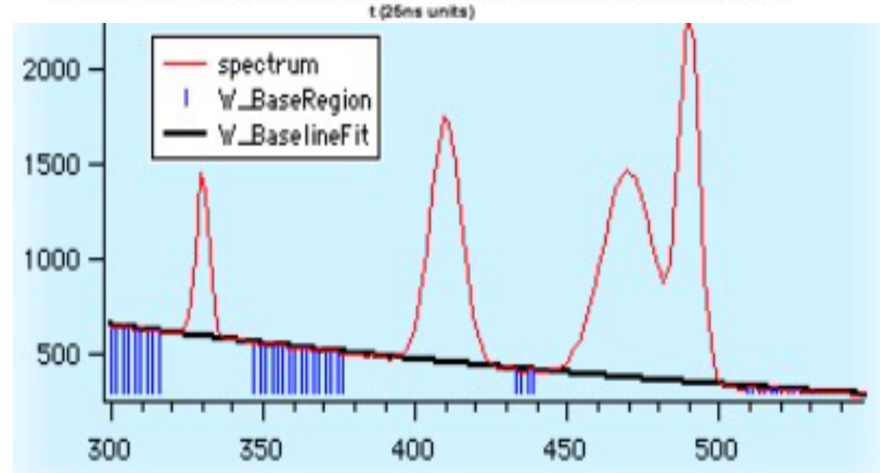
Latency dominated by cable/transmission delay

TRIGGER REQUIREMENTS ON FRONT-END ELECTRONICS

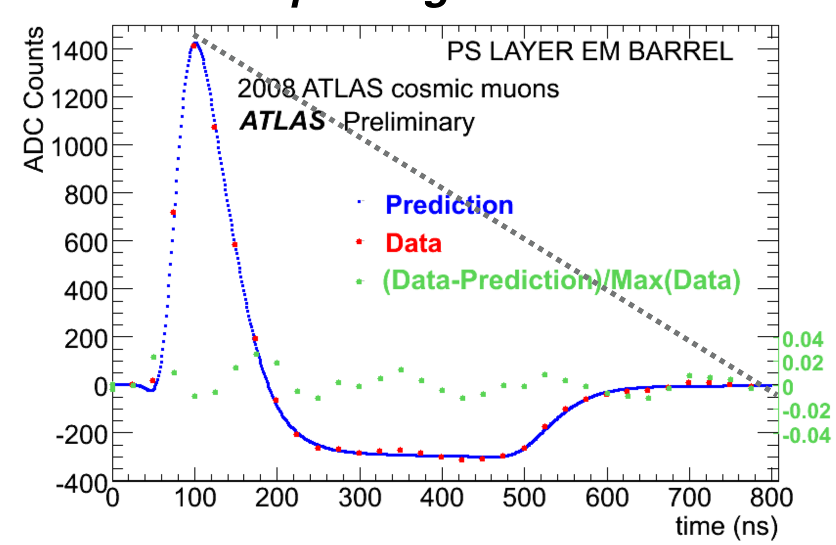
Tight design constraints for trigger & FE







ATLAS Liquid Argon calorimeter



Avoid

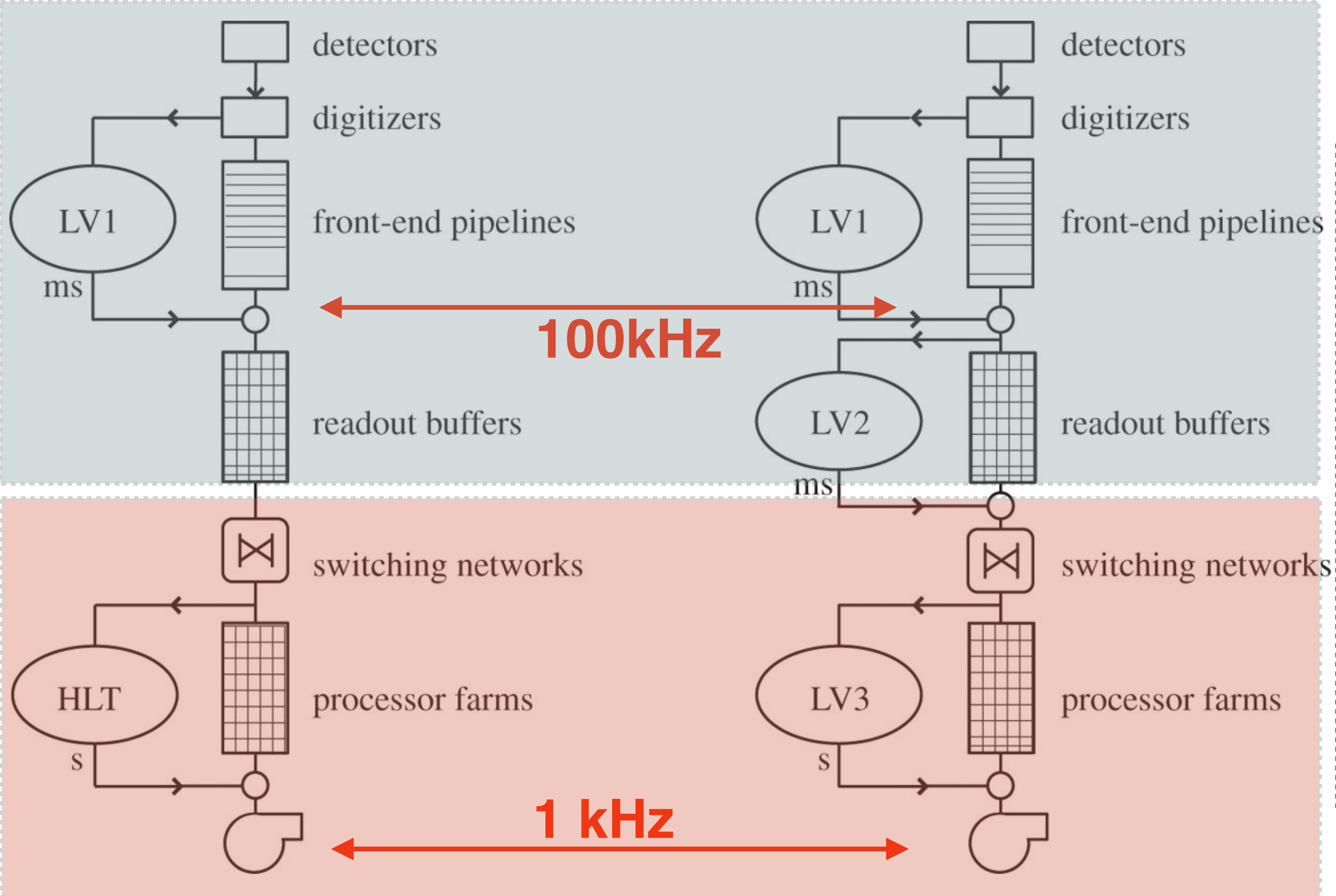
- ➡ **Electronic pile-up**
  - ➡ source of dead-time
  - ➡ distortion in pulse
- ➡ **In-time pile-up**
  - ➡ more collisions/BC
  - ➡ Baseline subtraction
- ➡ **Out-of-time pile-up**
  - ➡ BC-identification capability
  - ➡ peak finder algorithms

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)

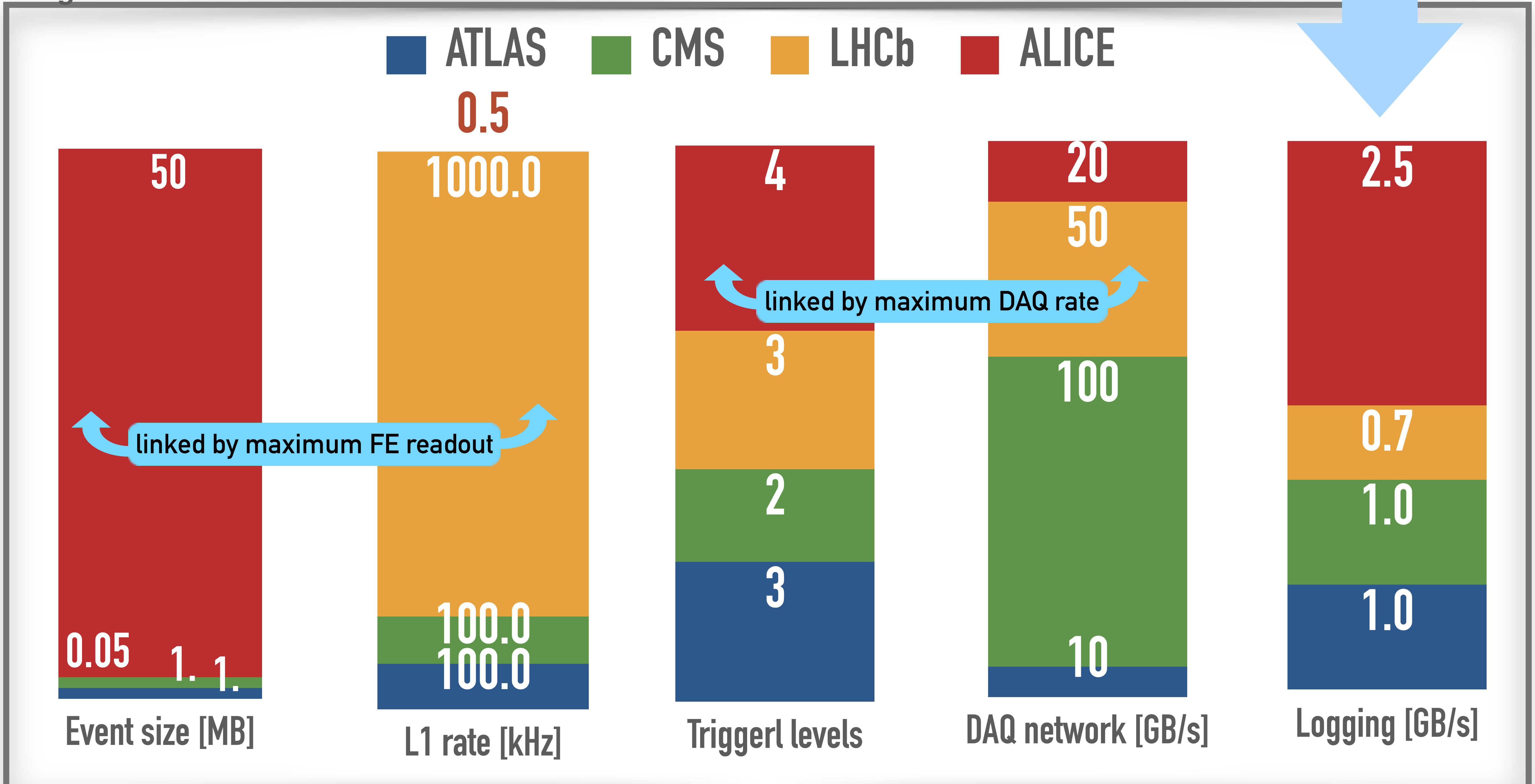
See S.Cittolin, DOI: 10.1098/rsta.2011.0464

COMPARING BY NUMBERS

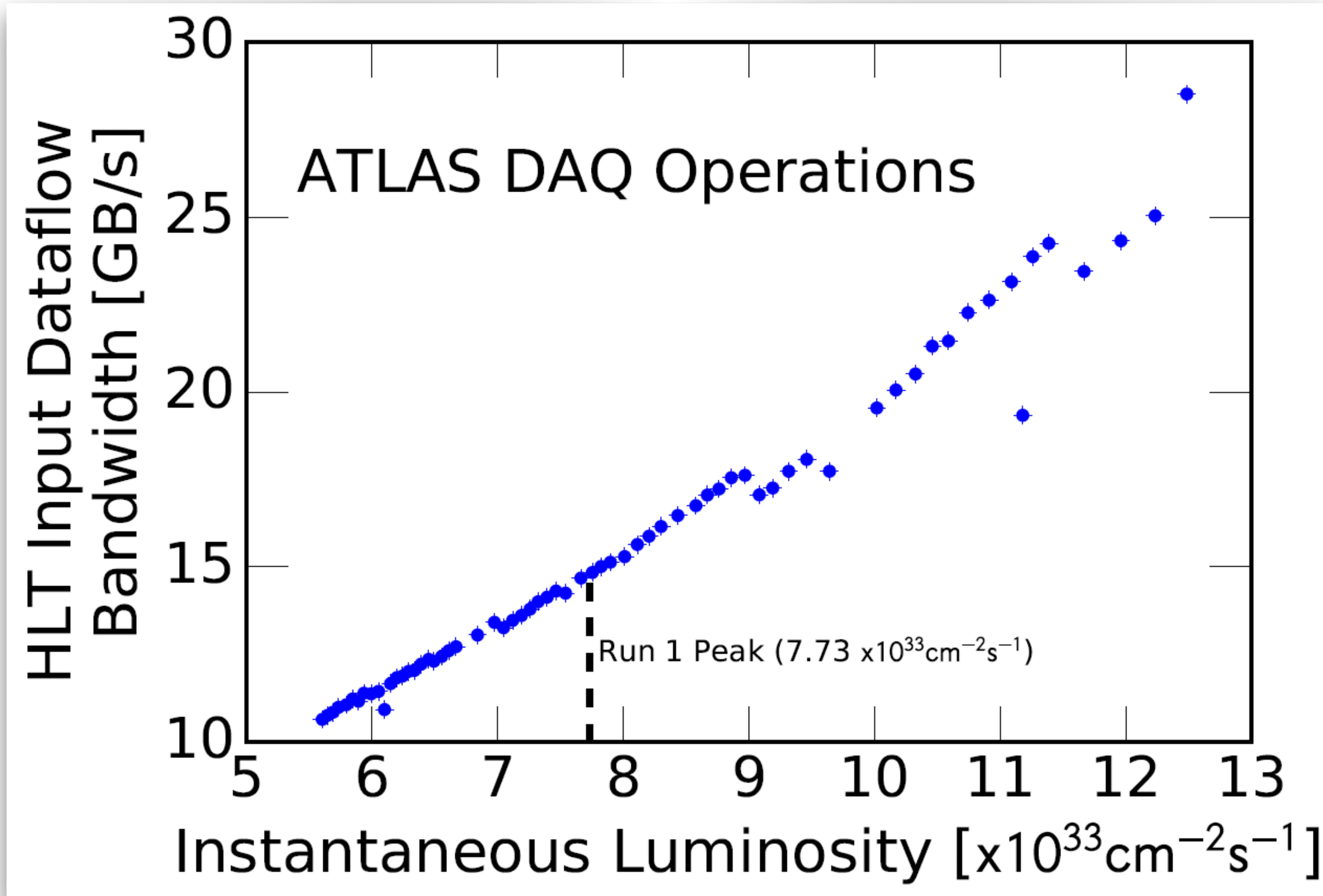
LHC experiments share the same CERN budget for computing resources, which is the constrain between trigger and DAQ power

Allowed storage and processing resources

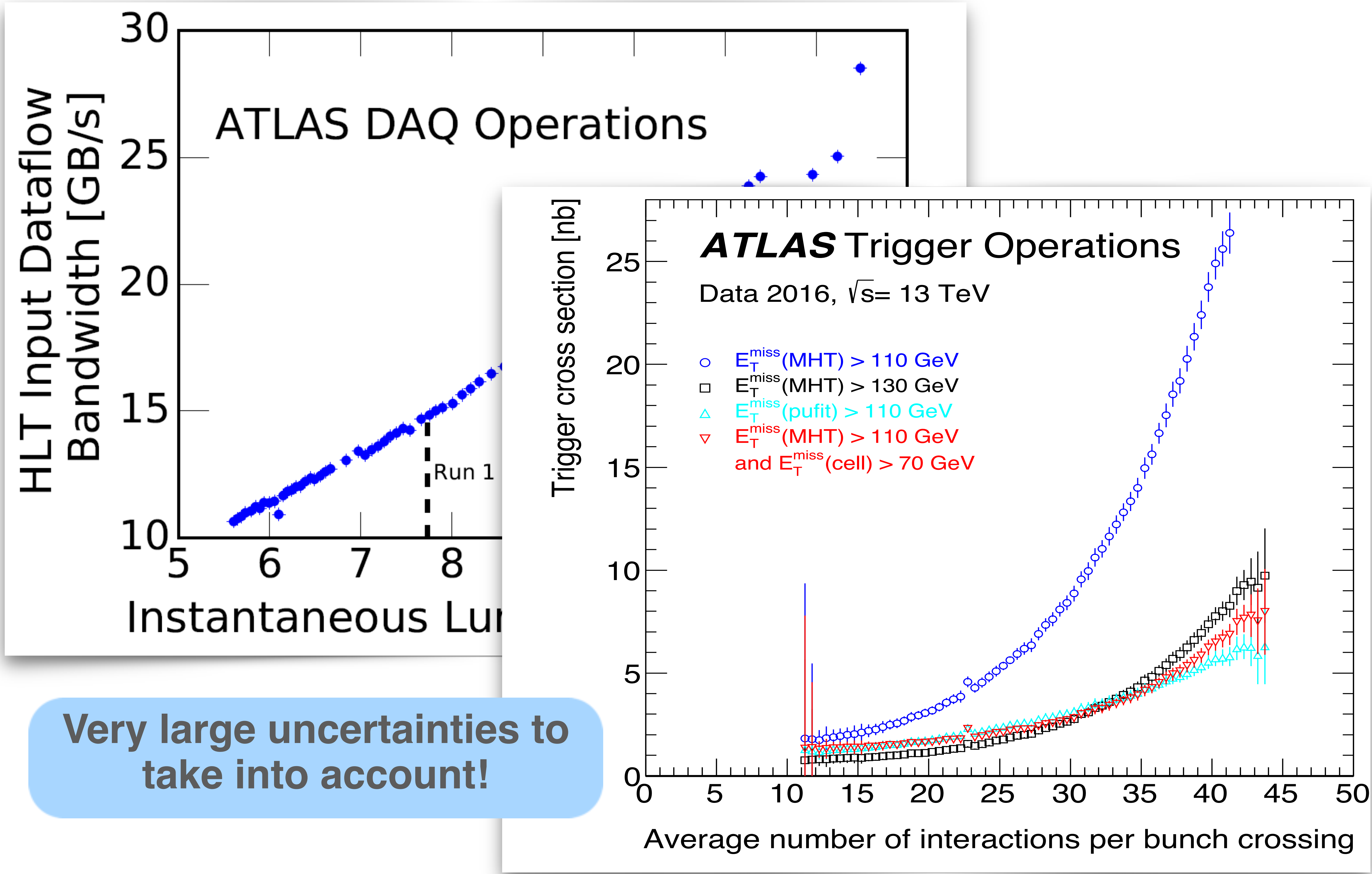
Design values in 2009



WHAT DO YOU EXPECT FOR THE FUTURE?



WHAT DO YOU EXPECT FOR THE FUTURE?



Very large uncertainties to take into account!

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

Design Luminosity x7.5

HL-LHC $t\bar{t}$ event in ATLAS ITK at <μ>=200

- ➡ 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

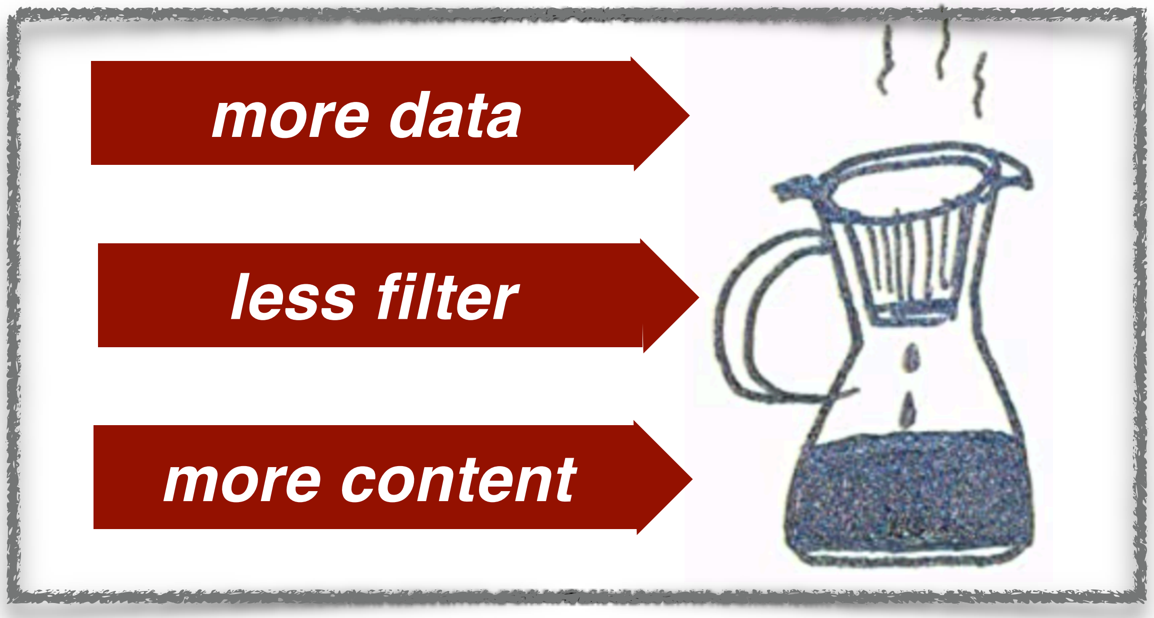
ADDITIONAL COMPLICATION AT HL-LHC

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

x10 higher Luminosity means…

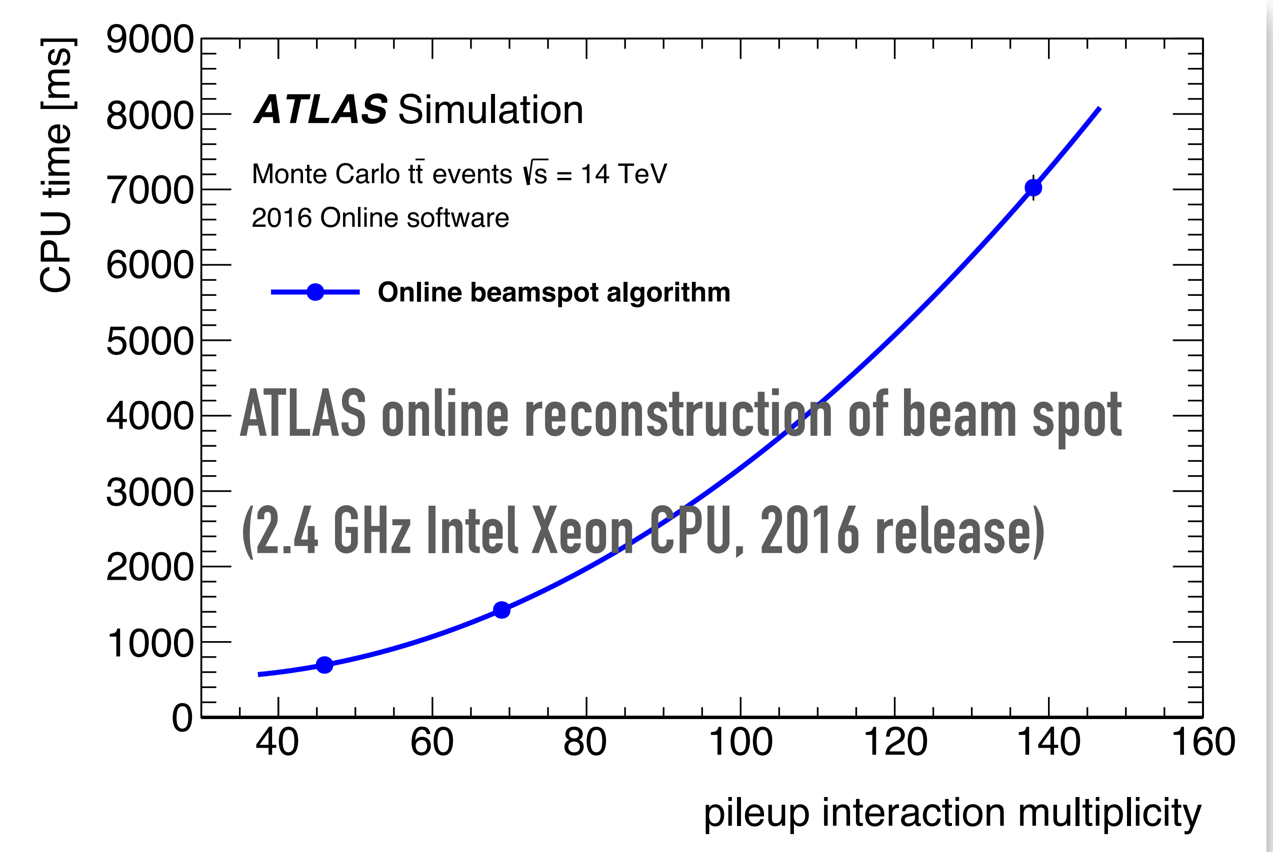
- ➡ **More interactions per BC (pile-up)**
  - ➡ Less rejection power (worse pattern recognition and resolution)
  - ➡ Larger event size
- ➡ **Larger data rates:**
  - ➡ FE readout rate @L1: 0.1 ➠ 1 MHz
  - ➡ DAQ throughput: 1 ➠ 50 Tbps

ATLAS/CMS numbers

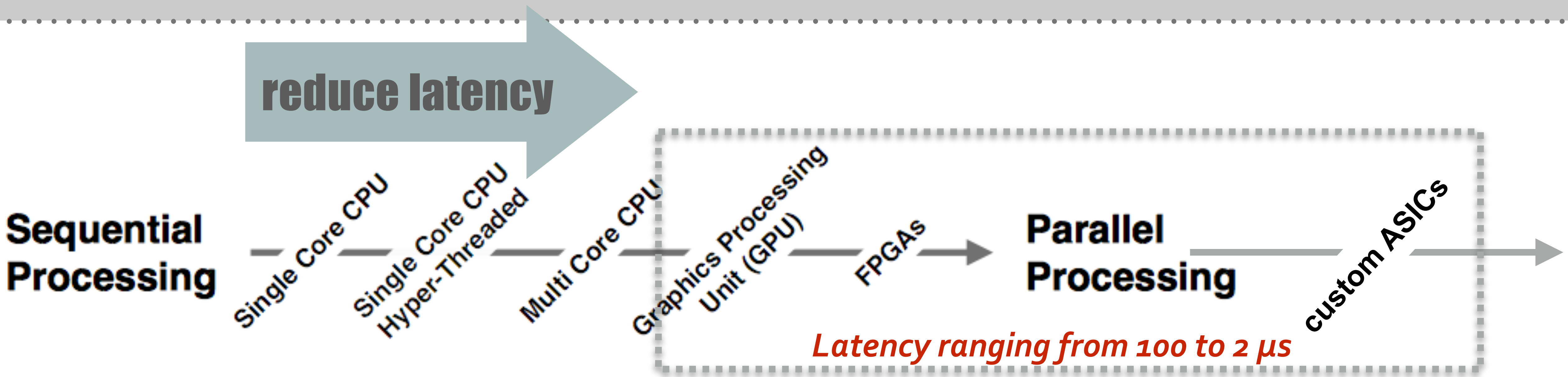


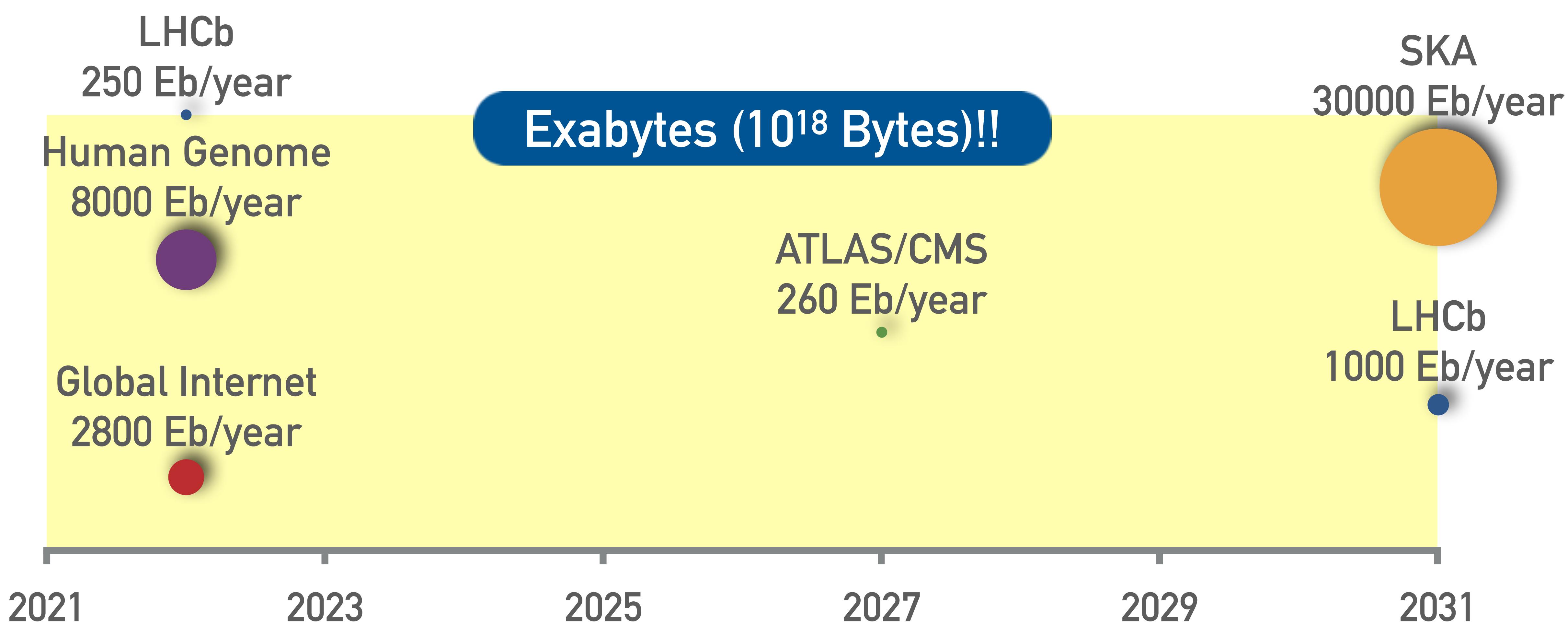
But cannot…

- ➡ **Increase trigger 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



THE REAL-TIME ADVENTURE

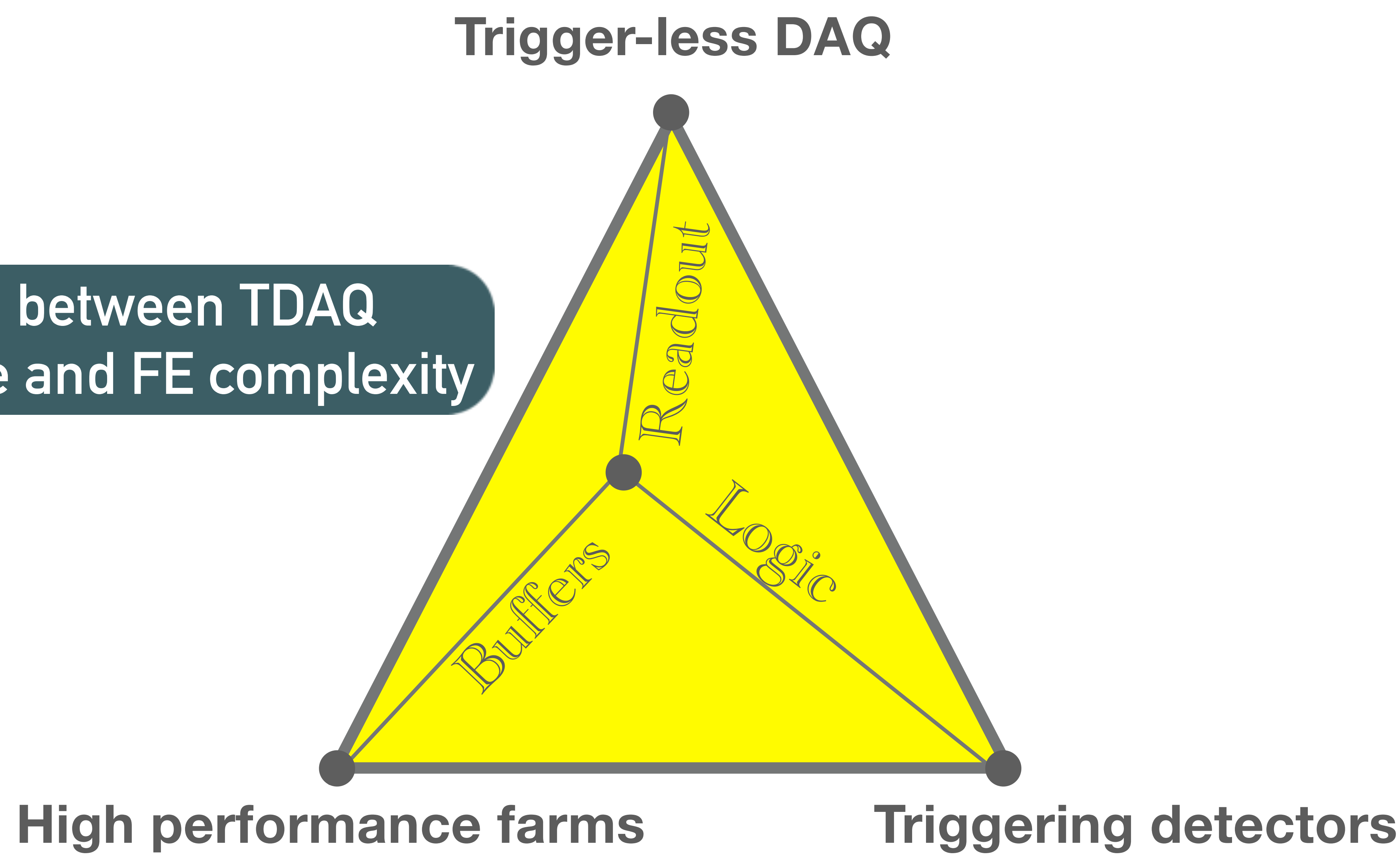




See Openlab workshop

BE SMARTER! INCREASE RESOLUTION FOR BETTER S/B

Tension between TDAQ architecture and FE complexity



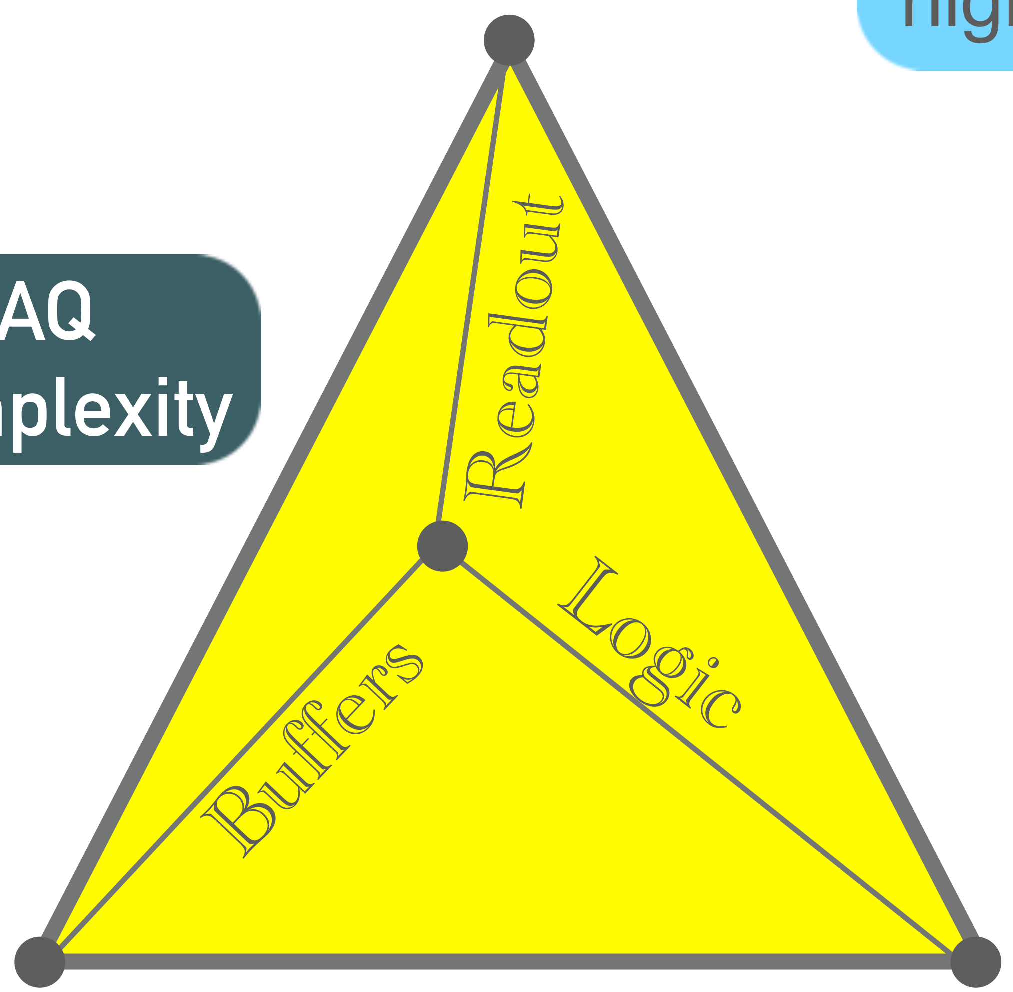
BE SMARTER! INCREASE RESOLUTION FOR BETTER S/B

What we do?

Trigger-less DAQ

high detector granularity

Tension between TDAQ architecture and FE complexity



High performance farms

Triggering detectors

refine calibrations, as offline

complex ASIC logic

BE SMARTER! INCREASE RESOLUTION FOR BETTER S/B

What we do?

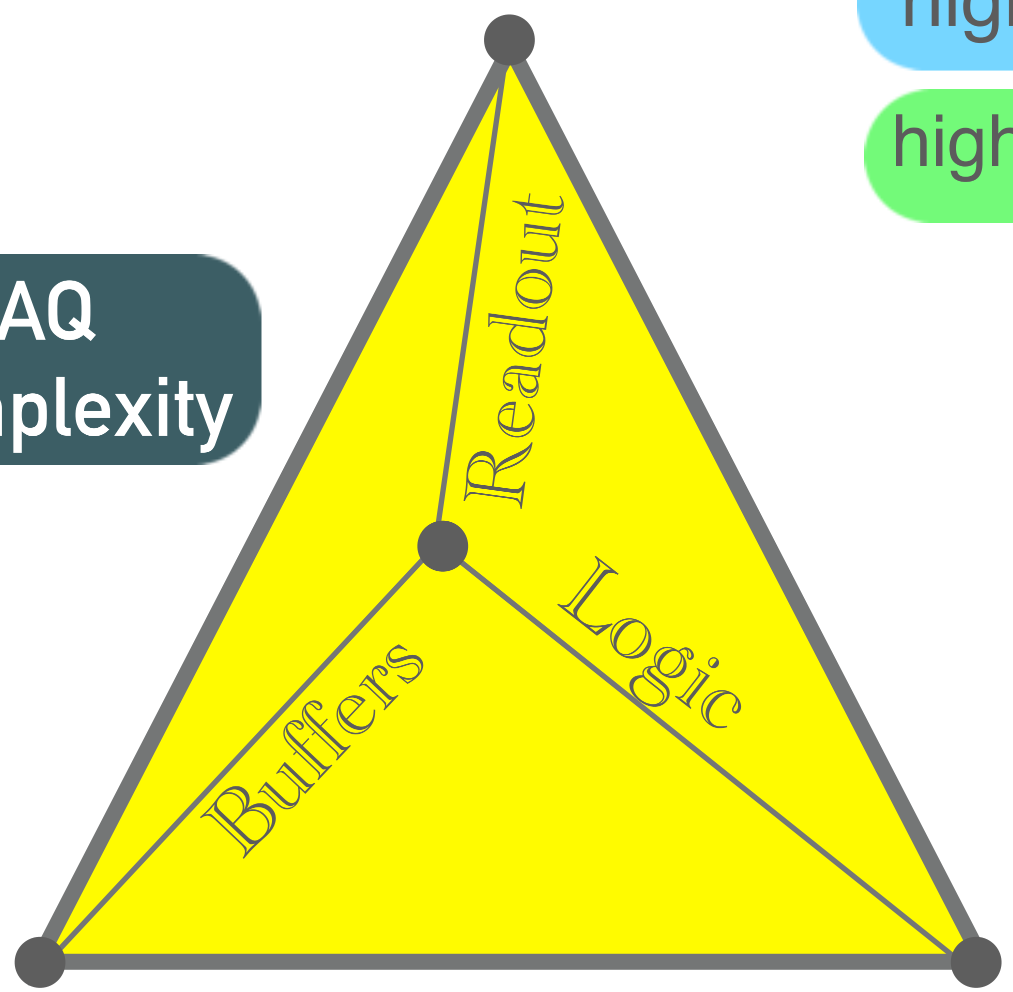
How?

Tension between TDAQ architecture and FE complexity

Trigger-less DAQ

high detector granularity

high speed electronics/links



High performance farms

refine calibrations, as offline

large buffers, long latency

Triggering detectors

complex ASIC logic

trigger-driven design

LHCP-2022

BE SMARTER! INCREASE RESOLUTION FOR BETTER S/B

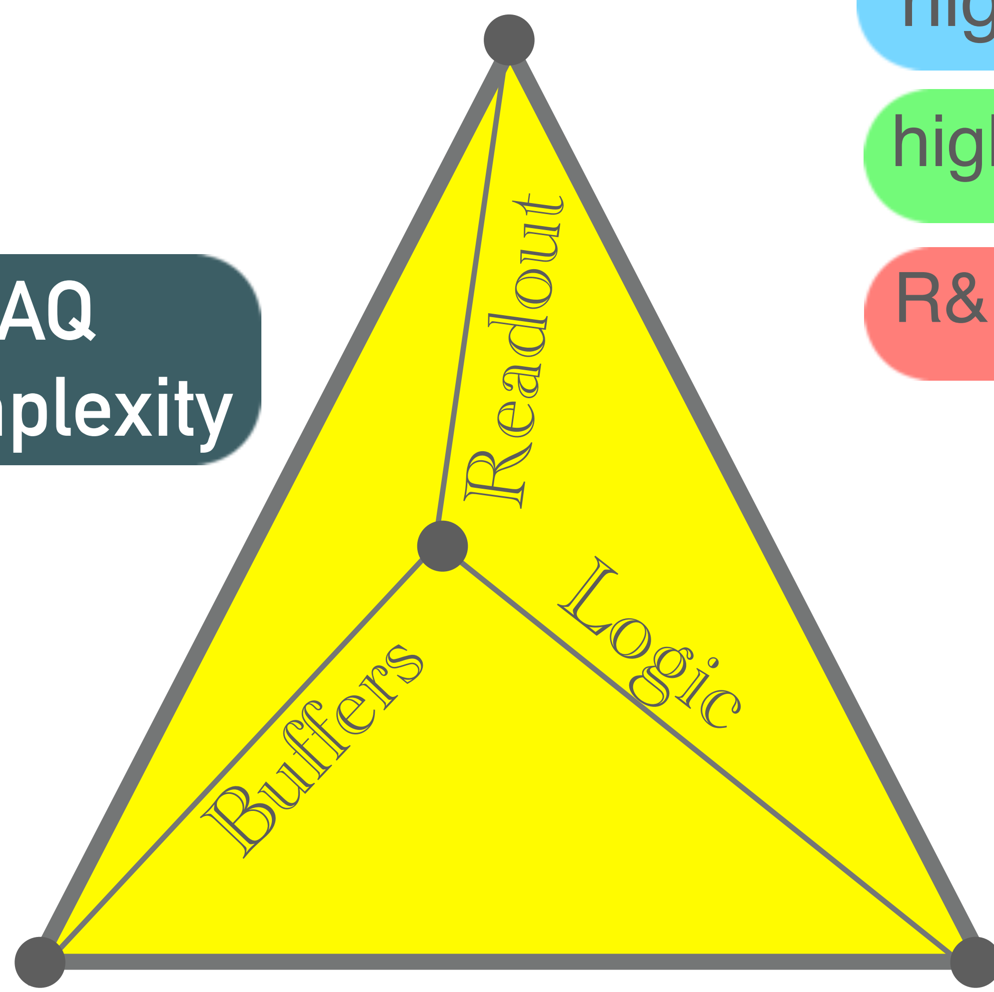
What we do?

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

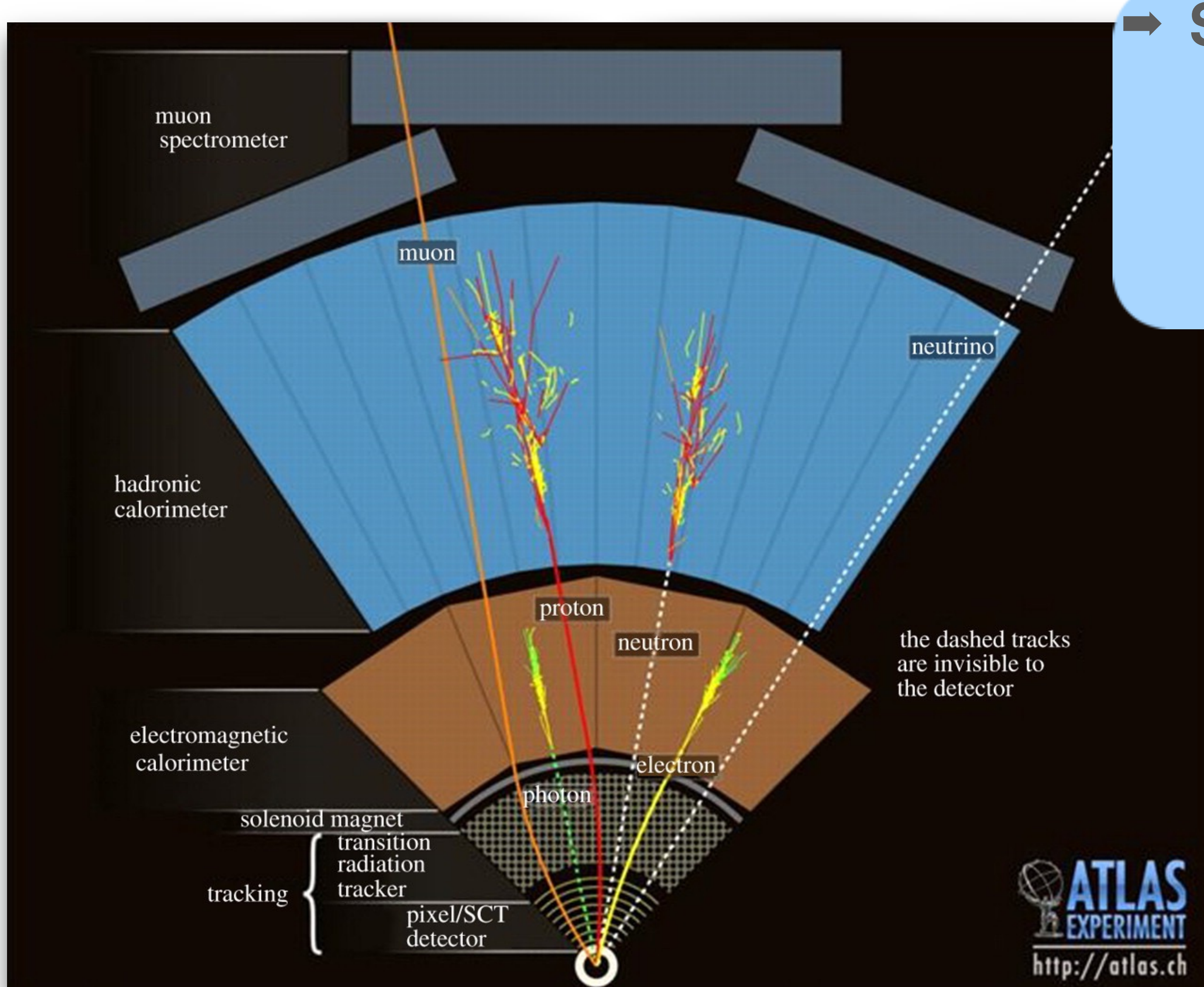
trigger-driven design

hardware track trigger (CMS)

COMPARE 4 EXPERIMENTS

How to maximise physics acceptance

ATLAS & CMS TRIGGER STRATEGY



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

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

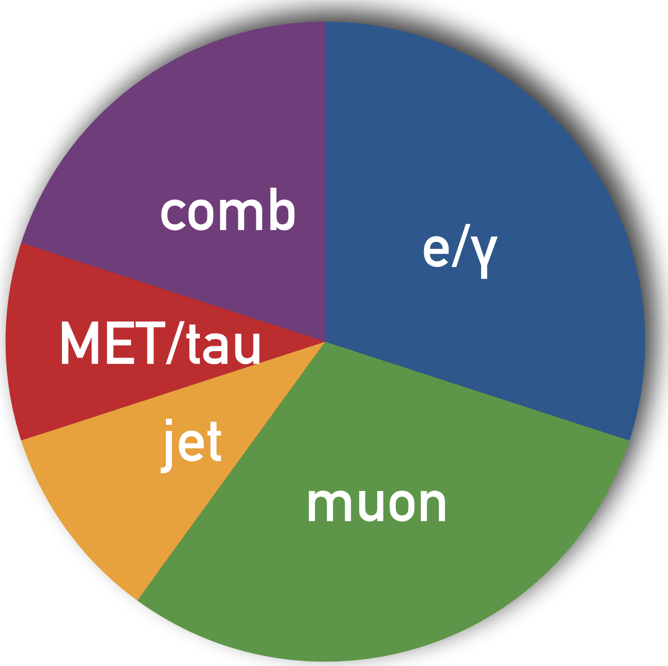
approximately 10^6 rejection

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

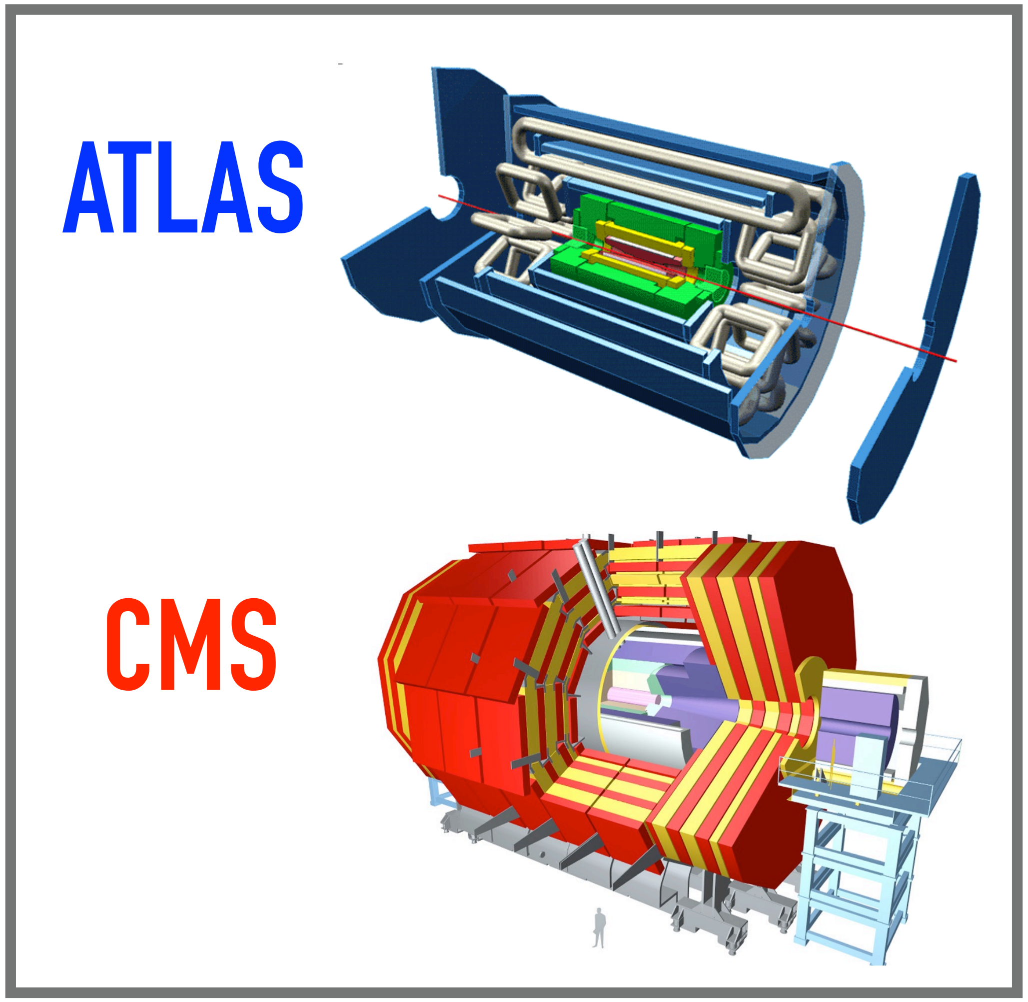
Same physics plans, different competitive approaches for detectors and DAQ

➡ **Same trigger strategy and data rates**

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



inclusive trigger selections



➡ **Different DAQ architectures**

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

EXAMPLE: NETWORK EVOLUTION IN CMS

Run 1: 100 GB/s network

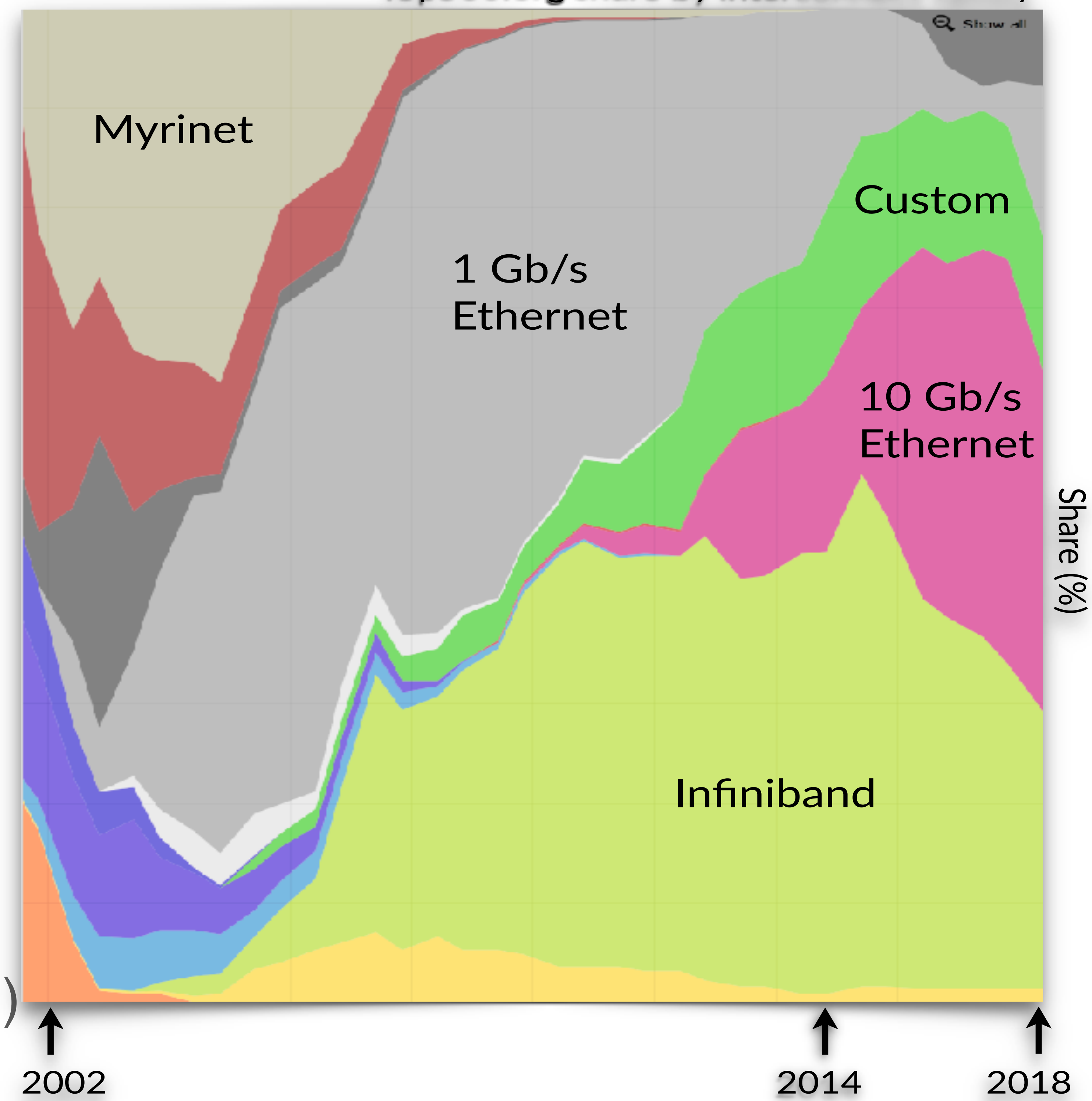
Myrinet widely used when DAQ-1 was designed

- ➡ 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

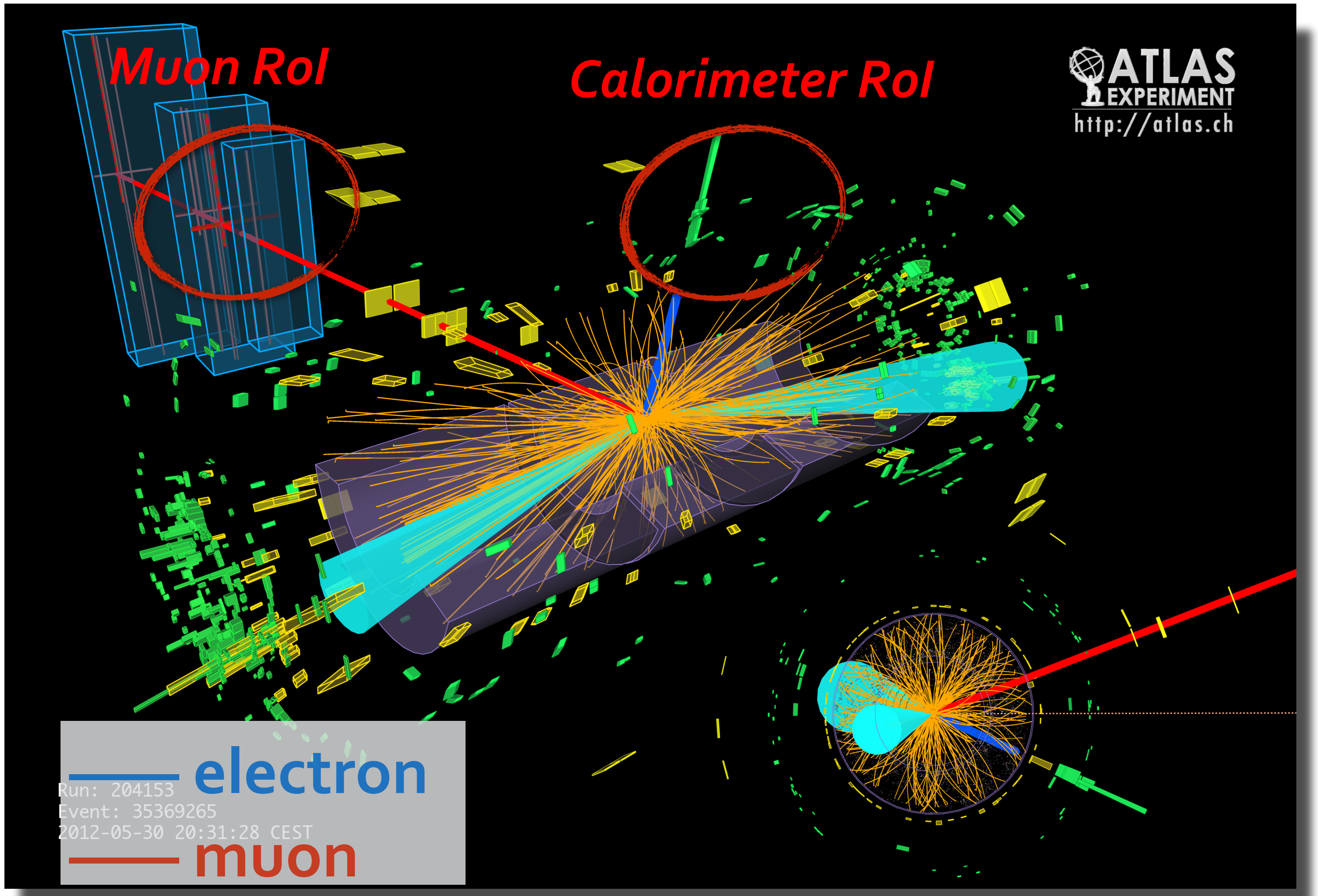
Top500.org share by interconnect family



Choose best prize/bitps!

ATLAS: REGION OF INTEREST (ROI) DATAFLOW

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

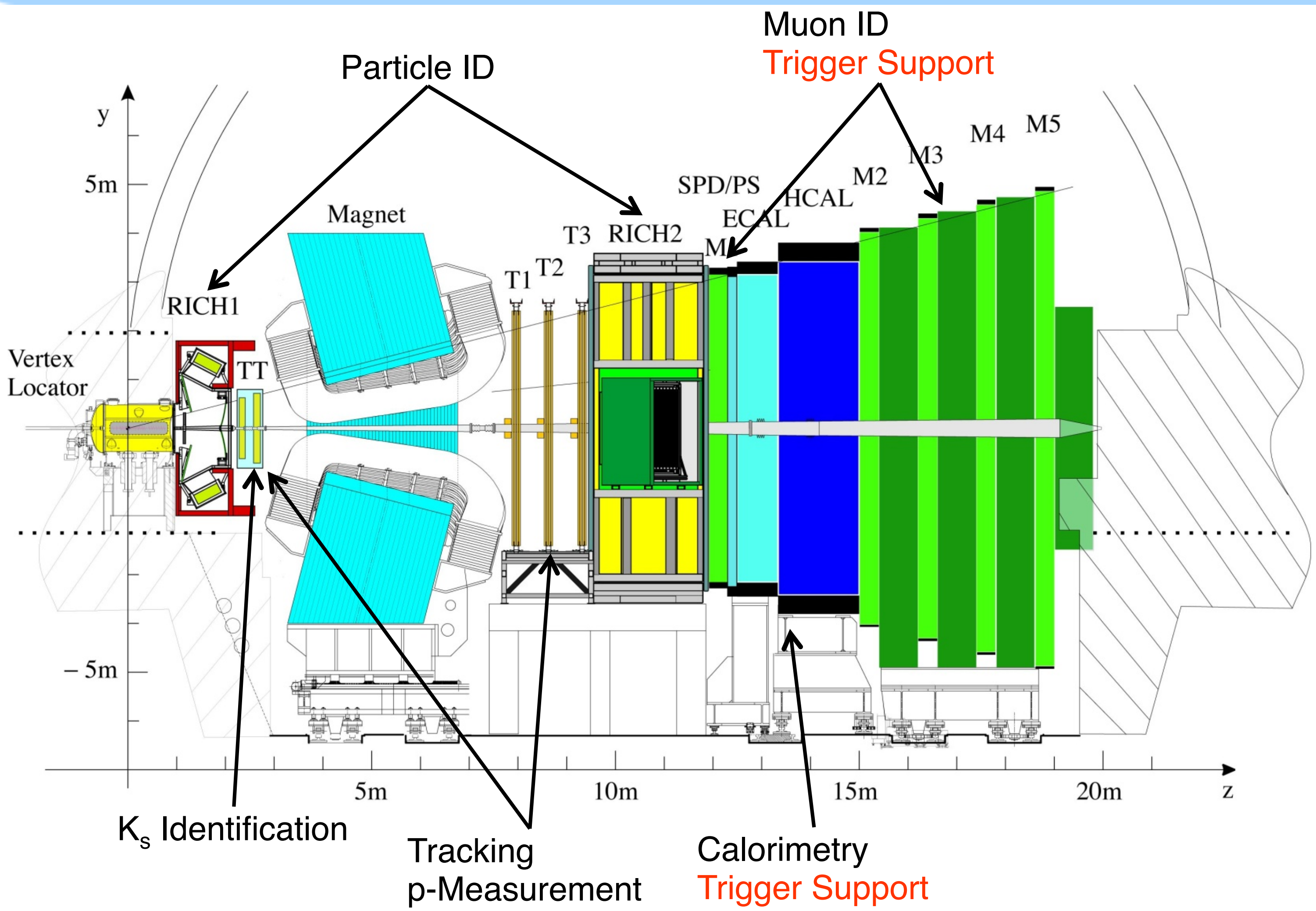


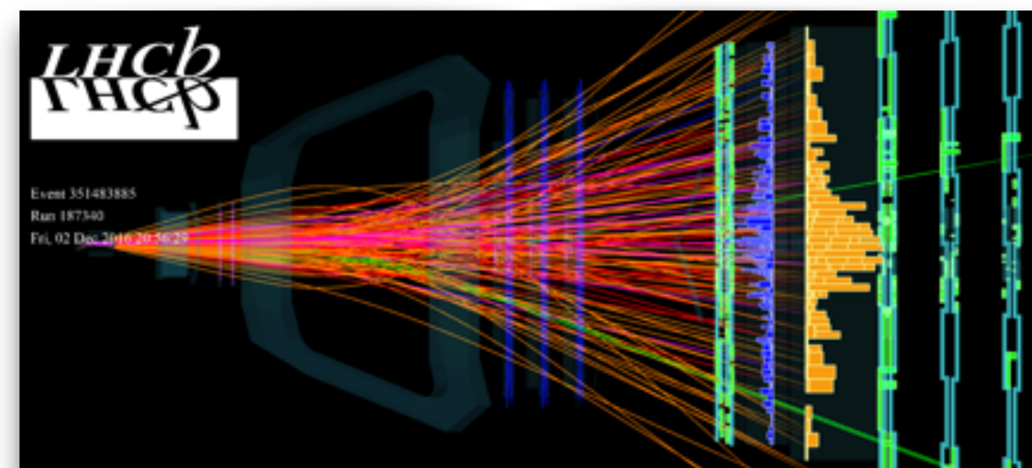
RoI=Region of Interest

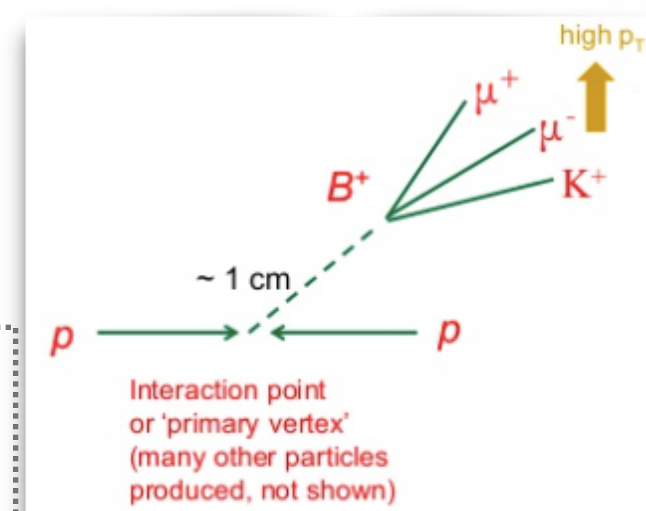
- ➡ **Total amount of RoI data is minimal: a few % of the Level-1 throughput**
  - ➡ one order of magnitude smaller readout network …
  - ➡ … at the cost of a higher control traffic and reduced scalability

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 <u>rare</u> (BR ~ 10^{-5})

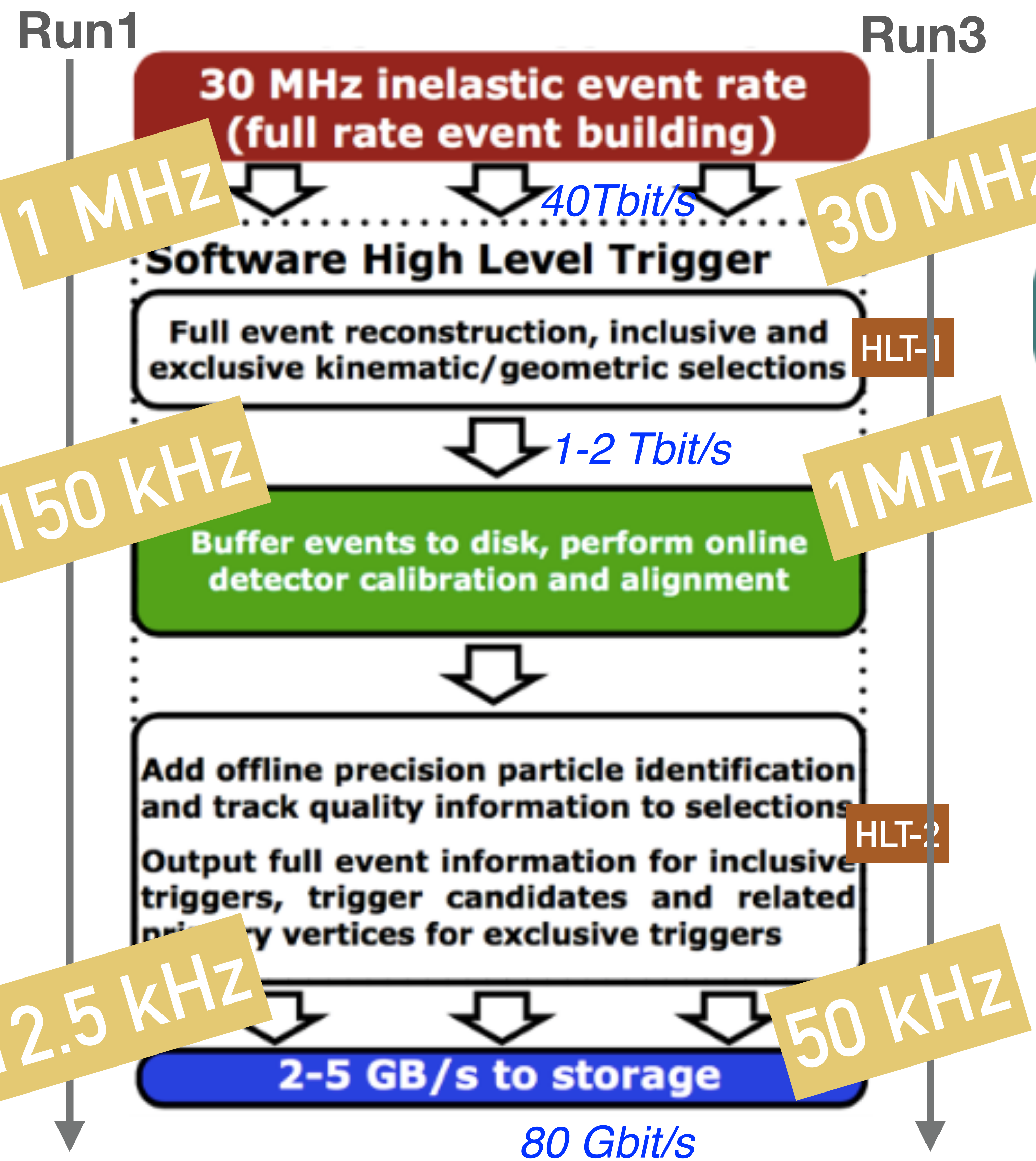






- ➡ **Single-arm spectrometer and low L ==> reduced event size**
- ➡ **Selection of B mesons ==> search for B-decay topologies**
  - ➡ related to high mass and long lifetime of the b-quark

TRIGGER-LESS?

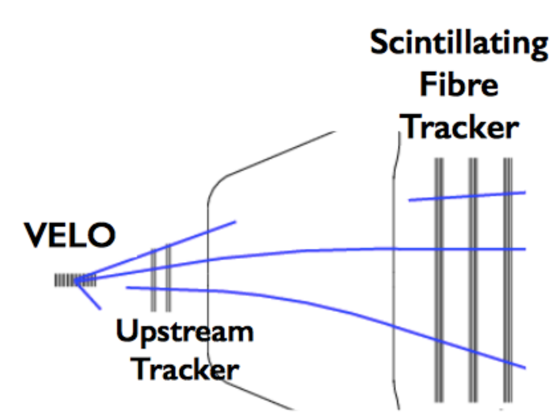


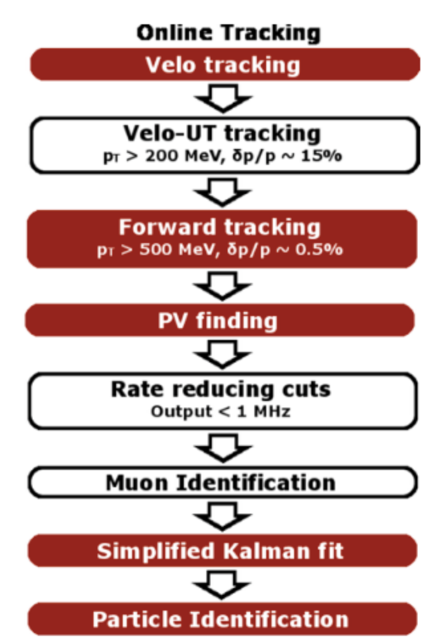
From Run1 to Run3, TDAQ system evolved to handle more readout rate

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

Tracking at ~30 MHz ?

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





arXiv:2105.04031

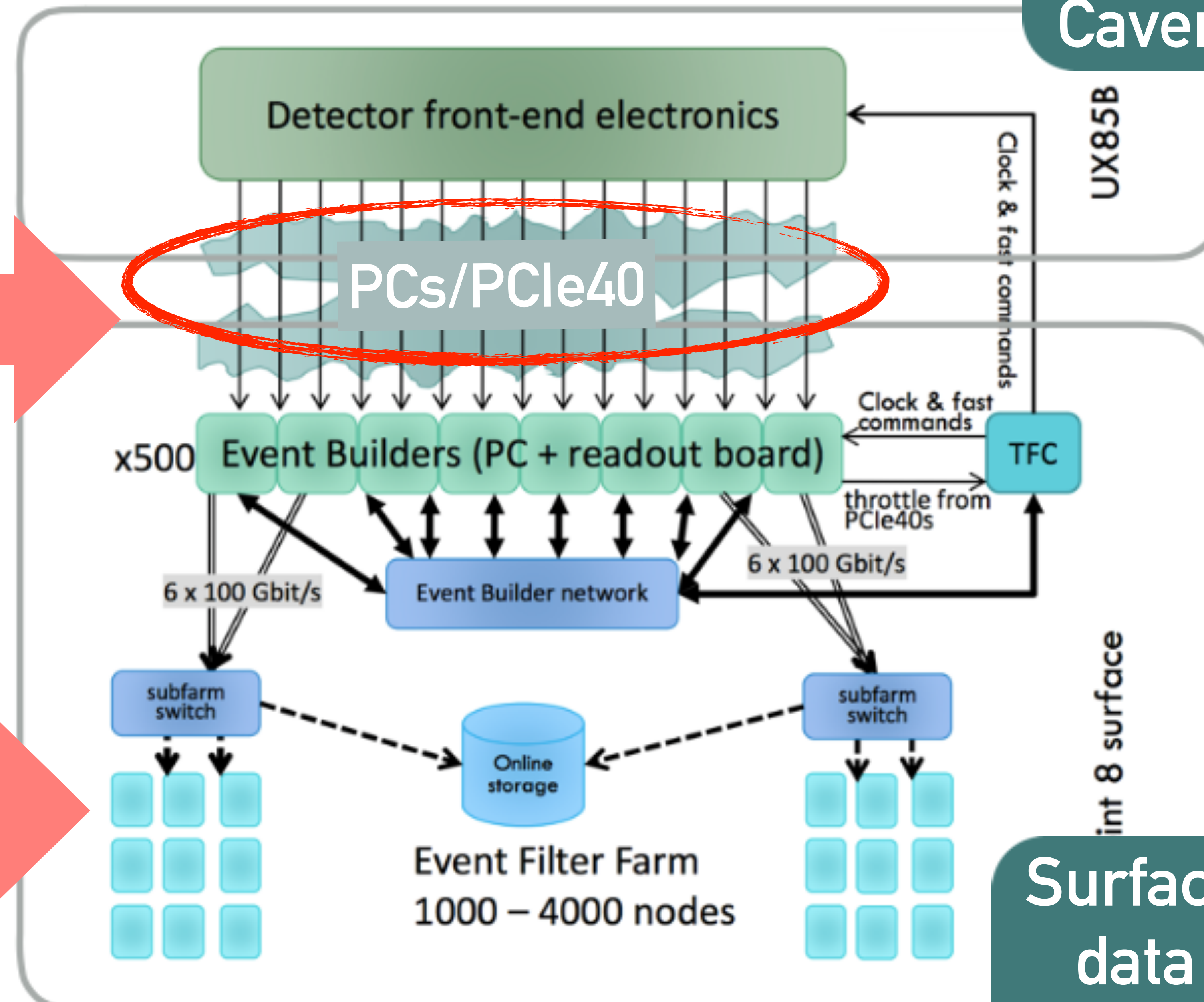
LHCB IN RUN3: NETWORK IS DATAFLOW

Readout @ 30 MHz
Event size ~ 150kB

150kB x 30MHz = 40Tbs

Inside Cavern

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



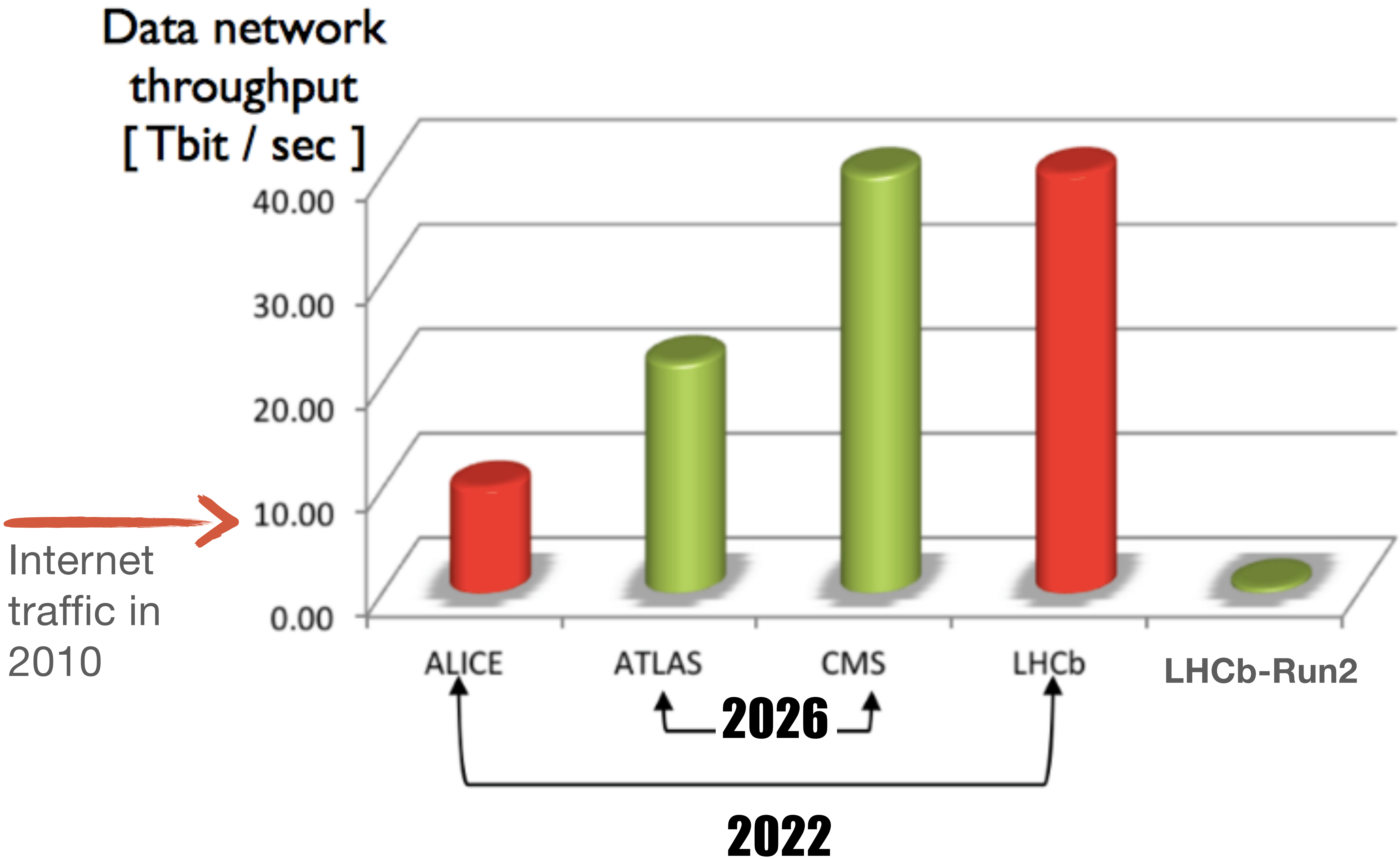
Surface data centre

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

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

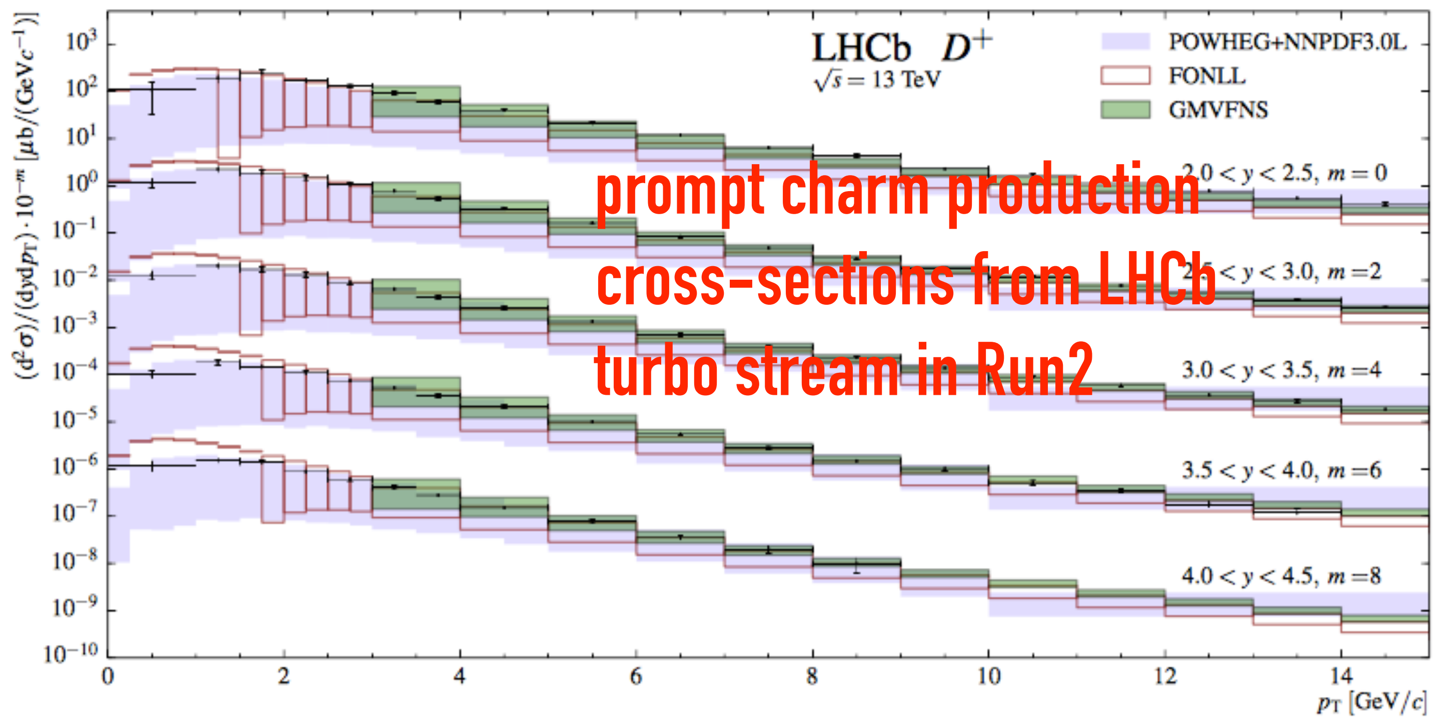
Ref for PCIe40

NETWORK TRAFFIC COMPARISON



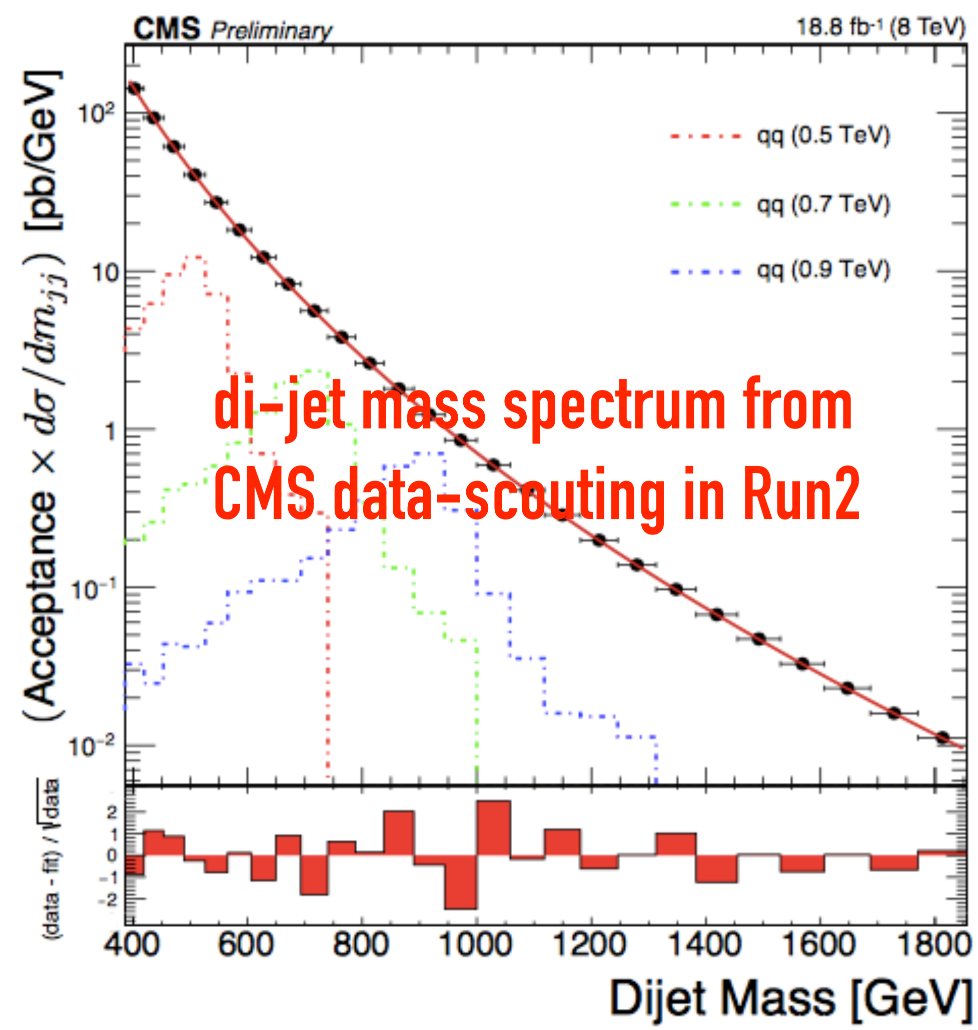
Same data volume as ATLAS/CMS HL-LHC upgrades! But earlier and for less money

A NEW TREND: REAL TIME ANALYSIS



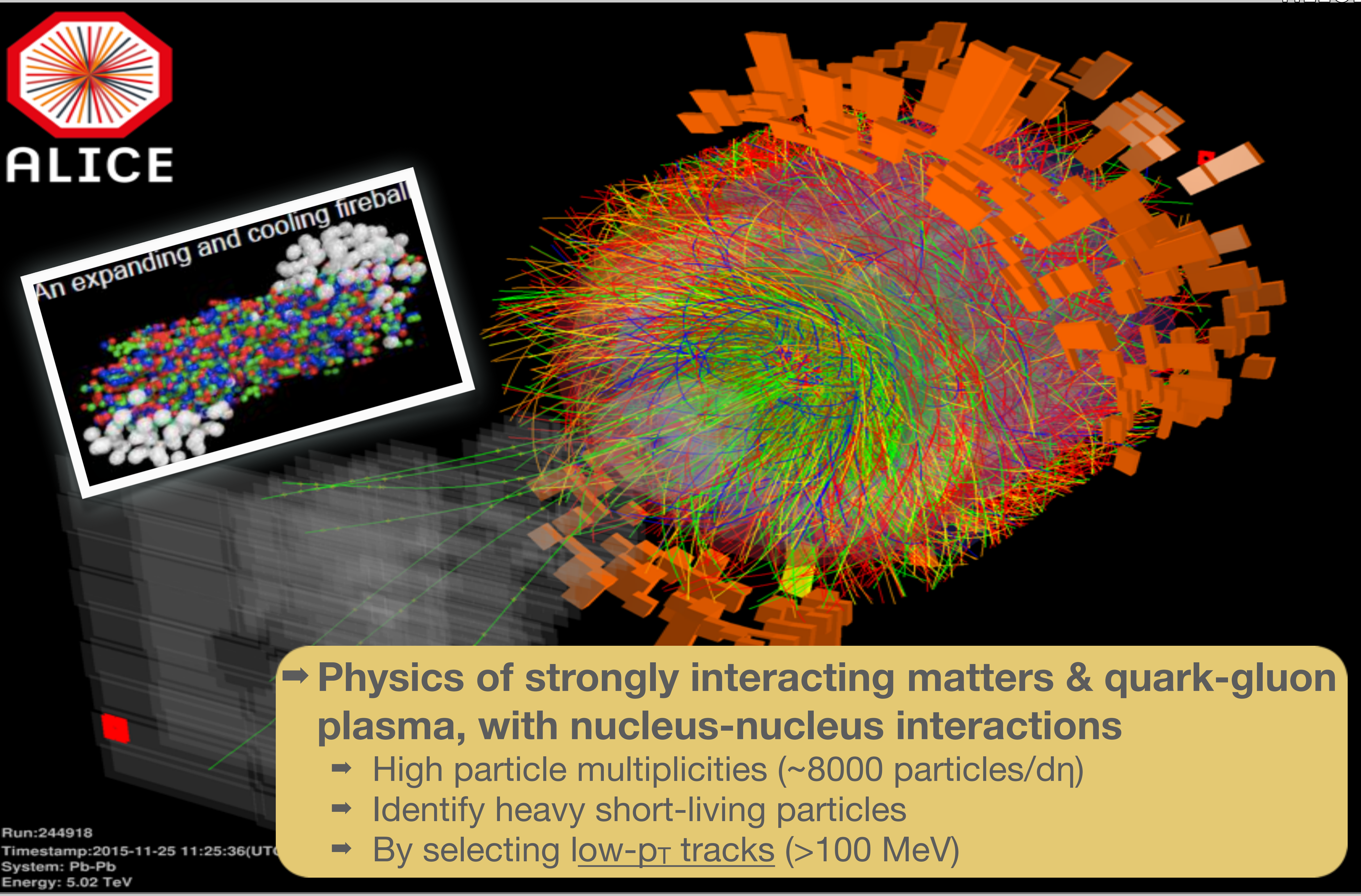
Can we get rid of FrontEnd raw data?

- ➡ **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

ALICE STRATEGIES

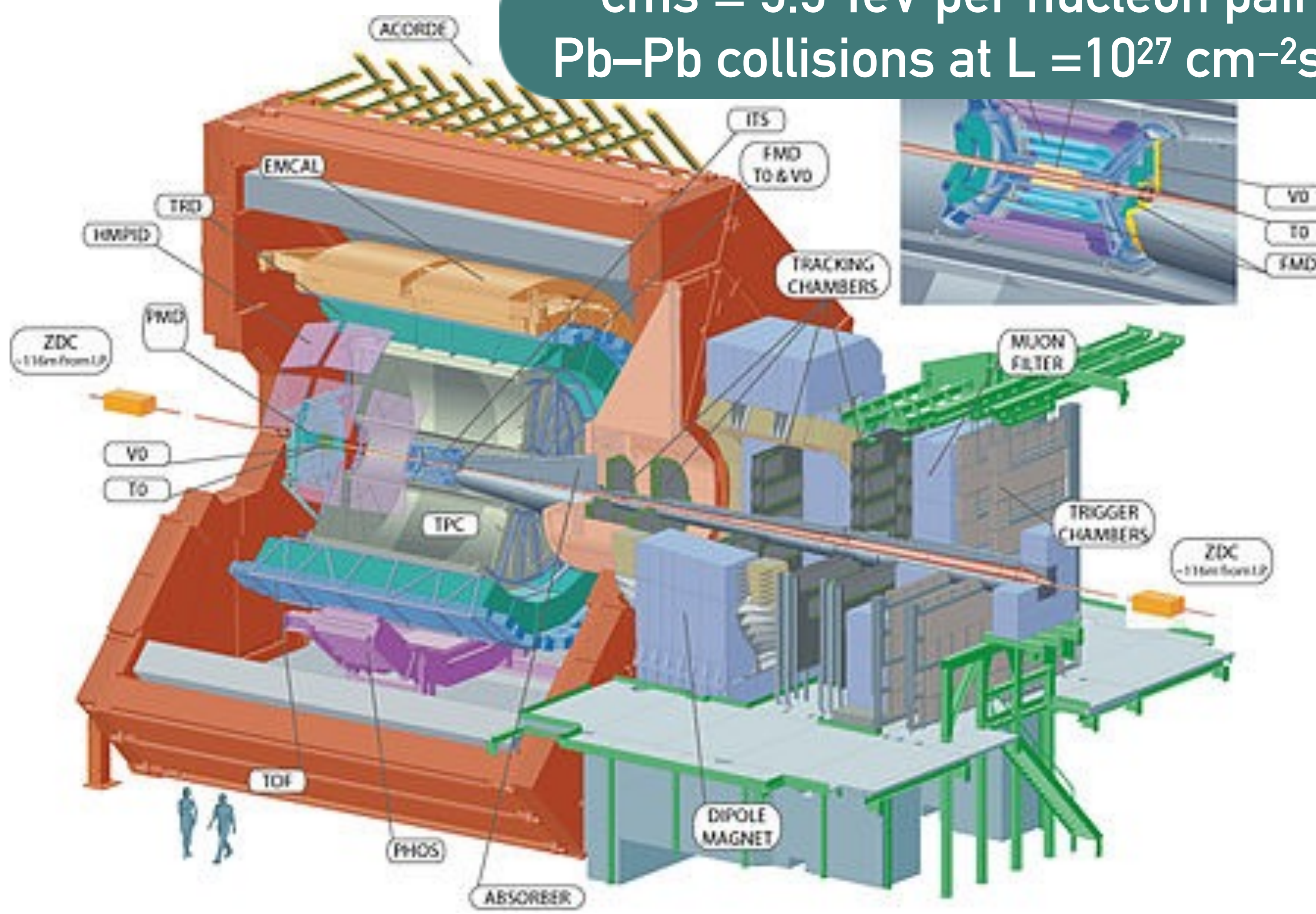


➡ **Physics of strongly interacting matters & quark-gluon plasma, with nucleus-nucleus interactions**
- ➡ High particle multiplicities (~8000 particles/dη)
- ➡ Identify heavy short-living particles
- ➡ By selecting <u>low-p_T tracks</u> (>100 MeV)

DESIGNED FOR HEAVY ION COLLISIONS

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

- ➡ **19 different detectors**
- ➡ **With high-granularity and timing information**
  - ➡ Time Projection Chamber (**TPC**): very high occupancy, and slow response
- ➡ **Large event size (> 40MB)**
  - ➡ TPC producing 90% of data
- ➡ **Complex event topology**
  - ➡ low trigger rate: ~ kHz



➡ **Challenges for TDAQ and evolution:**

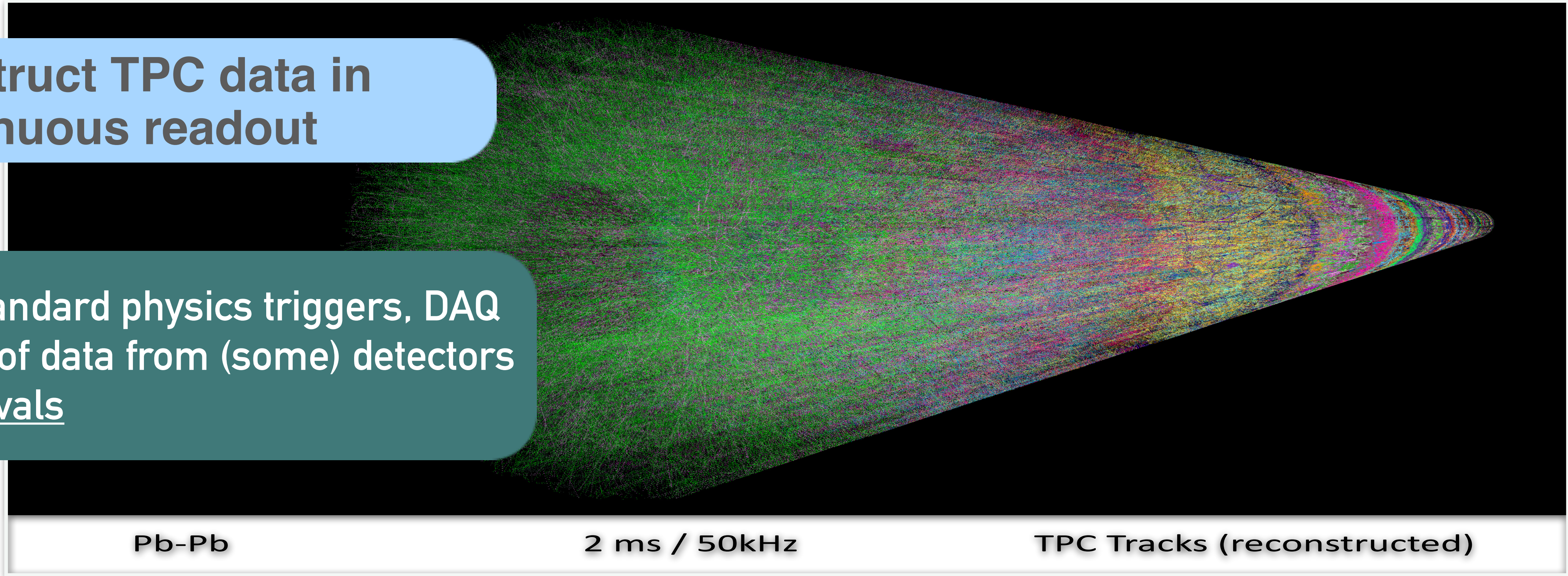
- ➡ detector readout: up to ~50 GB/s ==> x100 for Run3
- ➡ storage: 1.2 TB/s (Pb-Pb) ==> x100 for Run3

How can we increate the readout rate, when it's close to TPC readout?

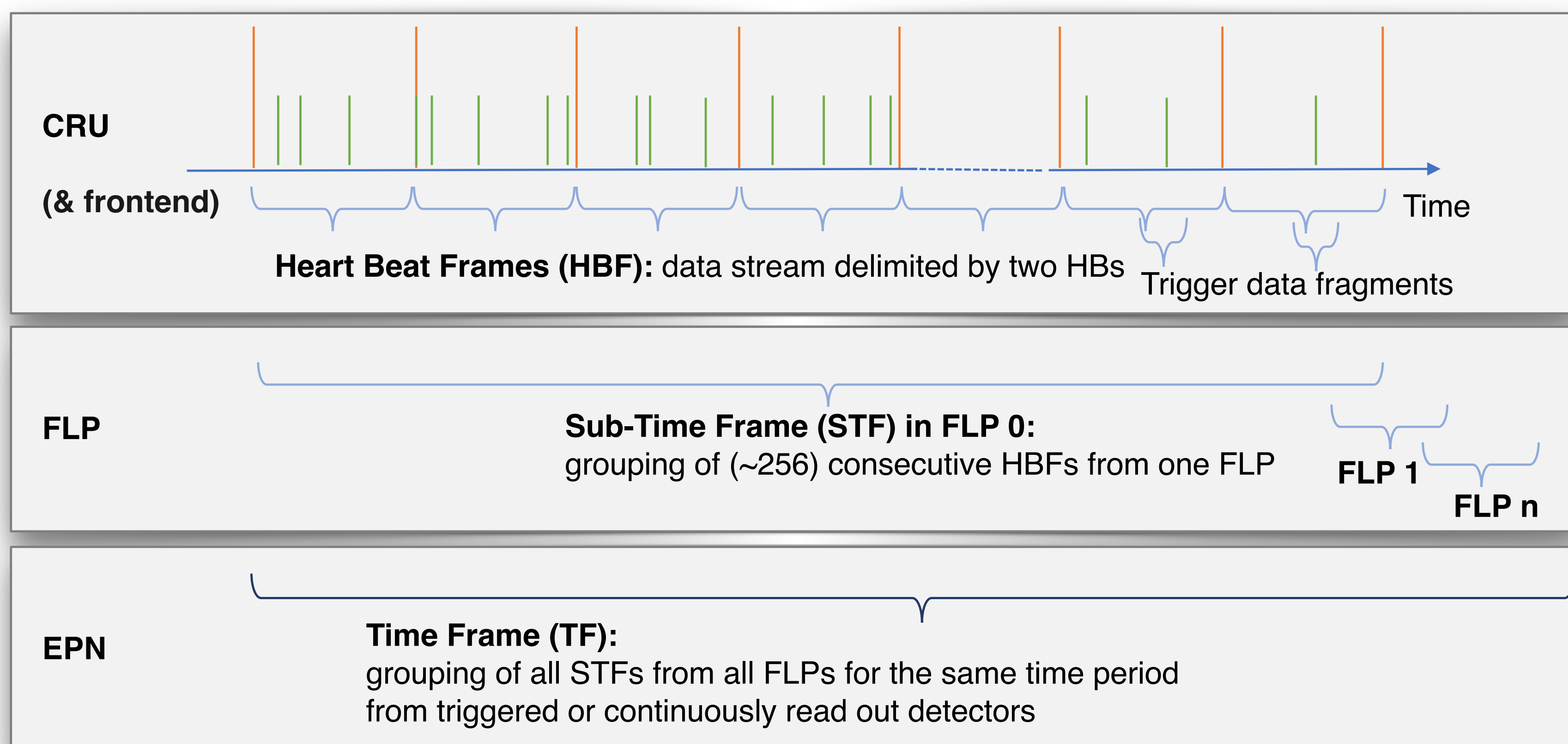
CONTINUOUS READOUT FOR RUN 3

Reconstruct TPC data in continuous readout

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

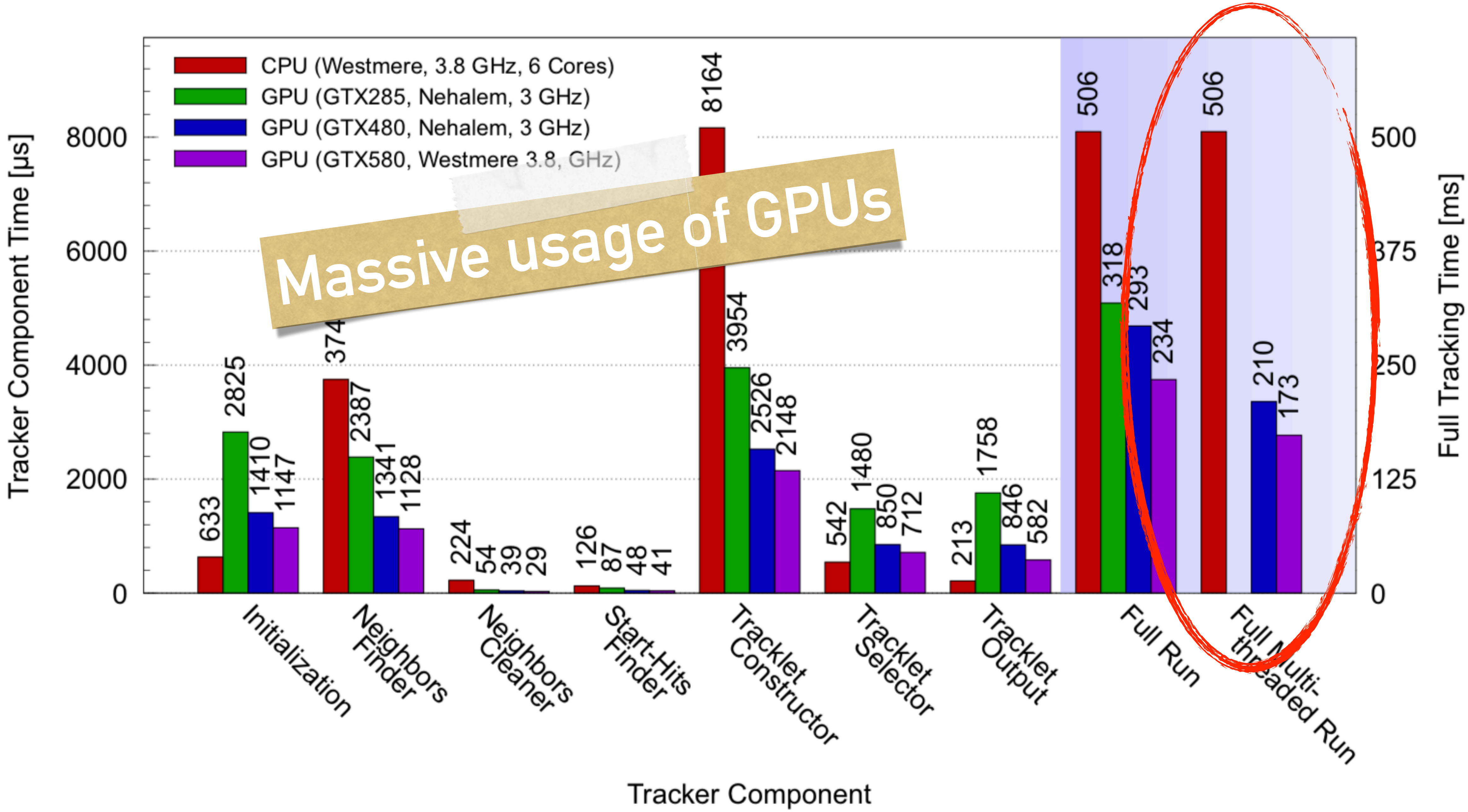


- **Heart Beat (HB) issued in continuous & triggered modes**
  - subdivision of data into time intervals to allow synchronisation between different detectors
  - 1 per LHC orbit, 89.4 μs: <u>~10 kHz</u>
- **Grouped in Time-Frames:**
  - 1 every ~20 ms: <u>**~50 Hz**</u> (1 TF = ~256 HBF)



INCREASING THROUGHPUTS WITH COTS

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



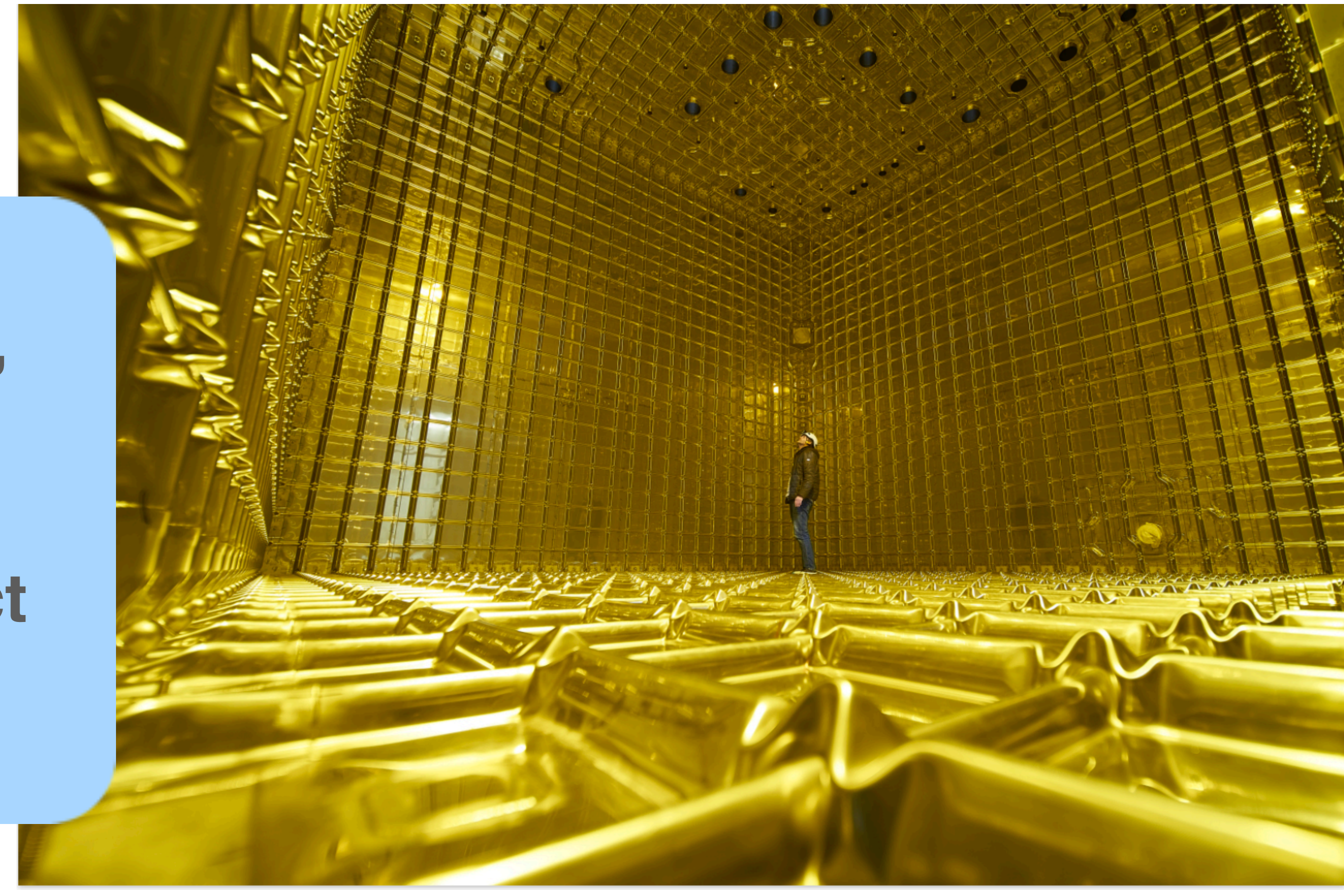
Tracking processing based on GPUs since Run1!

OUTLINE

- ➡ Examples of small experiments with their limits
- ➡ Overview of LHC experiments and their upgrade
- ➡ **Future TDAQ systems (Dune/Proto-Dune)**

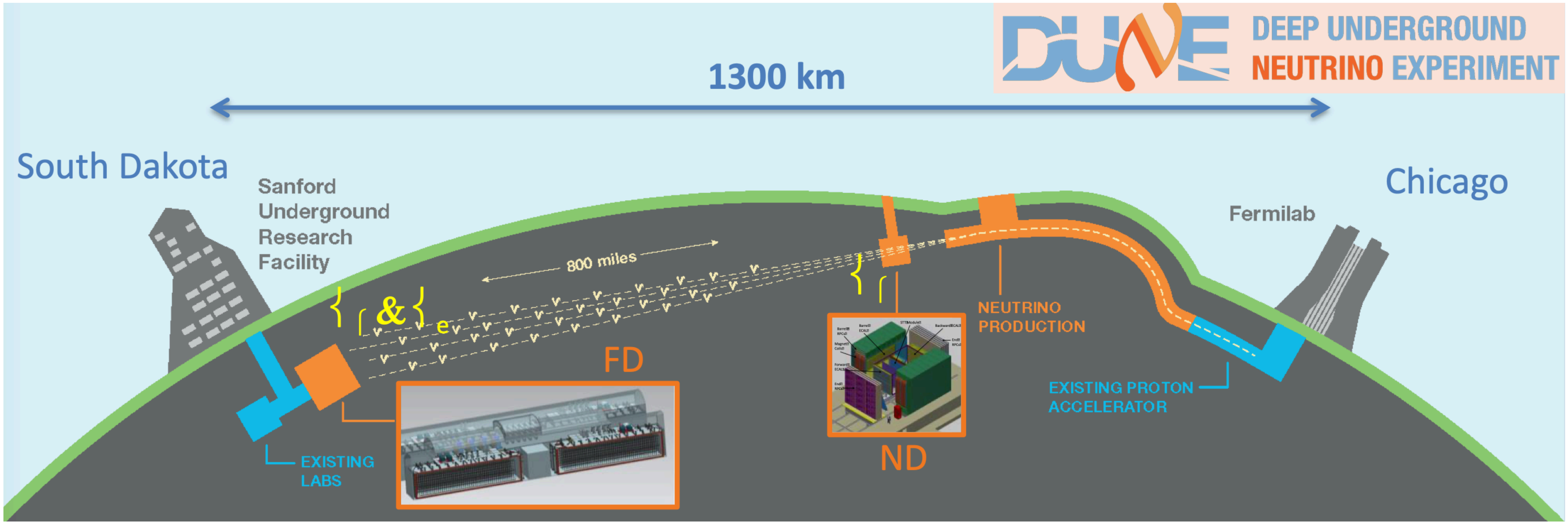
- **The next generation project for neutrino physics**
  - the experiment does not exist (ready for 2030)
  - the TDAQ of the experiment does not exist
- **Consider here design inputs:**
  - have a broad understanding of what the experiment wants to achieve
  - understand the detection principles and front-end electronics
  - understand the constraints in which the TDAQ will live

- **http://dunescience.org**
- **DUNE Collaboration : 1317 members, 208 institutions, 33 Countries**
- **Strong International partnership to build a mega neutrino science project based in US**
- **see recent CERN colloquium**



A view of the ProtoDUNE cryostat at CERN (Image: CERN)

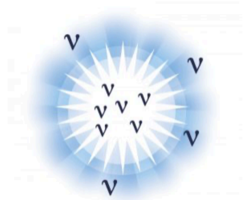
DUNE FACILITY AND DETECTORS



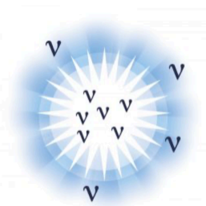
- **Two detectors on a muon-neutrino beam @Long-Baseline Neutrino Facility**
  - One near the source of the beam, at Fermilab (**ND**), to characterise the beam & systematics
  - One, much larger, 1300 km downstream, 1.48 km underground (**FD**)
    - Massive Liquid Argon Time Projection Chambers (70-kton, slow) + photon detectors (fast)
    - *the best particle imaging capability*
- **No quick access and no large host lab in the area !**

- **Prototypes at CERN Neutrino Platform (proto-DUNE)**
  - 2 prototypes, 1/20th the size of planned DUNE
    - *the largest liquid-argon neutrino detector in the world!*
  - Collected 4M events in 2018- 2020 from both cosmic rays and a beam

DUNE TRIGGERS AND READOUT



DUNE TRIGGERS AND READOUT

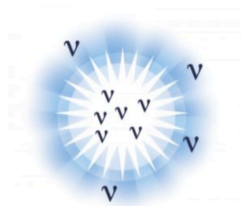


- **Extended physics cases:**
  - Origin of matter: measure **neutrino oscillations** on large distances and unfold CPV from matter effects
    - trigger: neutrino beam -> external trigger possible
  - Unification of forces: search for **proton decay**
    - trigger: very local, rare signature
  - Black hole formation: observe neutrinos from **supernova collapse**
    - Very distributed, rare signature

DUNE TRIGGERS AND READOUT

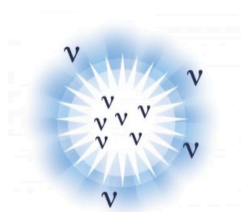
- **Extended physics cases:**
  - Origin of matter: measure **neutrino oscillations** on large distances and unfold CPV from matter effects
    - trigger: neutrino beam -> external trigger possible
  - Unification of forces: search for **proton decay**
    - trigger: very local, rare signature
  - Black hole formation: observe neutrinos from **supernova collapse**
    - Very distributed, rare signature
- **TDAQ active at “all” times, mixing readout strategies**
  - local readout for photon detectors, sampling @ 150 MHz
  - continuous readout for TPC, sampling @ 2 MHz
  - post-readout system combines data fragments into time windows of interesting detector regions
    - data reordering appears to be the biggest CPU consumer

DUNE TRIGGERS AND READOUT



- **Extended physics cases:**
  - Origin of matter: measure **neutrino oscillations** on large distances and unfold CPV from matter effects
    - trigger: neutrino beam -> external trigger possible
  - Unification of forces: search for **proton decay**
    - trigger: very local, rare signature
  - Black hole formation: observe neutrinos from **supernova collapse**
    - Very distributed, rare signature
- **TDAQ active at "all" times, mixing readout strategies**
  - local readout for photon detectors, sampling @ 150 MHz
  - continuous readout for TPC, sampling @ 2 MHz
  - post-readout system combines data fragments into time windows of interesting detector regions
    - data reordering appears to be the biggest CPU consumer
- **Adding all up, TDAQ has to sustain readout of ~5 TB/s**
  - TPC: 384 k channels (12 bit ADC) @ 2 MHz = 9.2 Tb/s (dominates)

DUNE TRIGGERS AND READOUT



- **Extended physics cases:**
  - Origin of matter: measure **neutrino oscillations** on large distances and unfold CPV from matter effects
    - trigger: neutrino beam -> external trigger possible
  - Unification of forces: search for **proton decay**
    - trigger: very local, rare signature
  - Black hole formation: observe neutrinos from **supernova collapse**
    - Very distributed, rare signature
- **TDAQ active at "all" times, mixing readout strategies**
  - local readout for photon detectors, sampling @ 150 MHz
  - continuous readout for TPC, sampling @ 2 MHz
  - post-readout system combines data fragments into time windows of interesting detector regions
    - data reordering appears to be the biggest CPU consumer
- **Adding all up, TDAQ has to sustain readout of ~5 TB/s**
  - TPC: 384 k channels (12 bit ADC) @ 2 MHz = 9.2 Tb/s (dominates)
- **Sounds very much like HL-LHC...**

DIFFERENCE WITH COLLIDER EXPERIMENTS

DIFFERENCE WITH COLLIDER EXPERIMENTS

- **Differently from LHC, time frames varies a lot**
  - from few ms to ~100s for the supernova core collapse
  - Data corresponding to a trigger can have size ranging << 1 GB to ~100 TB!

DIFFERENCE WITH COLLIDER EXPERIMENTS

- **Differently from LHC, time frames varies a lot**
  - from few ms to ~100s for the supernova core collapse
  - Data corresponding to a trigger can have size ranging << 1 GB to ~100 TB!
- **The rate of events varies widely from few Hz to <<1/month**

DIFFERENCE WITH COLLIDER EXPERIMENTS

- **Differently from LHC, time frames varies a lot**
  - from few ms to ~100s for the supernova core collapse
  - Data corresponding to a trigger can have size ranging << 1 GB to ~100 TB!
- **The rate of events varies widely from few Hz to <<1/month**
- **The trigger selection need to accumulate data from detectors over several seconds**
  - readout needs very large buffers to accommodate the long decision latency
  - fast storage of 3.5 TB with a sequential write performance @ 25 GBps

DIFFERENCE WITH COLLIDER EXPERIMENTS

- **Differently from LHC, time frames varies a lot**
  - from few ms to ~100s for the supernova core collapse
  - Data corresponding to a trigger can have size ranging << 1 GB to ~100 TB!
- **The rate of events varies widely from few Hz to <<1/month**
- **The trigger selection need to accumulate data from detectors over several seconds**
  - readout needs very large buffers to accommodate the long decision latency
  - fast storage of 3.5 TB with a sequential write performance @ 25 GBps
- **Complexity and size are similar, but uptime is much larger (100% instead of ~30%)**

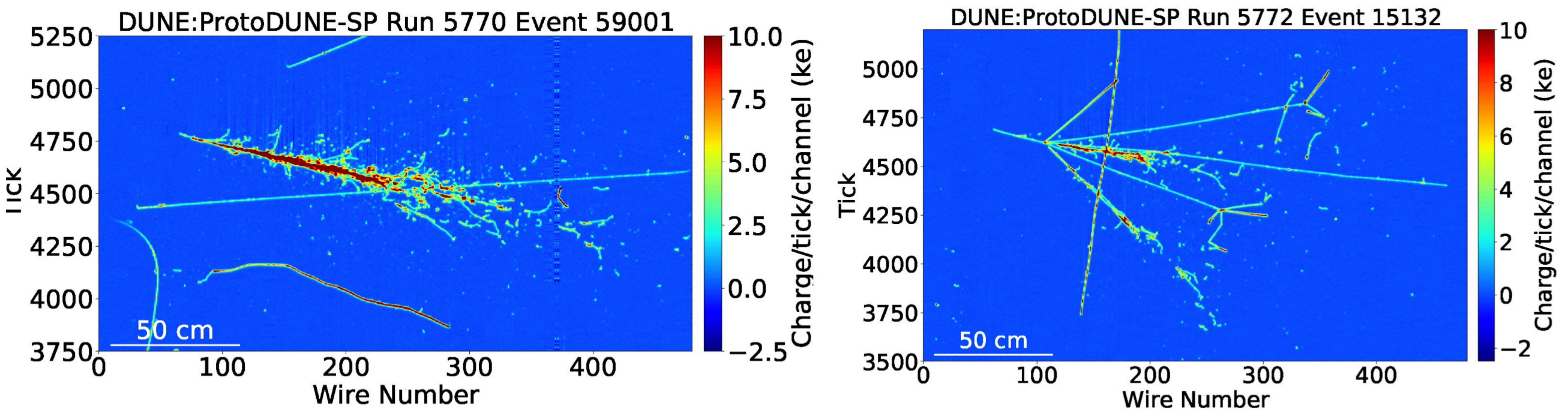
DIFFERENCE WITH COLLIDER EXPERIMENTS

- **Differently from LHC, time frames varies a lot**
  - from few ms to ~100s for the supernova core collapse
  - Data corresponding to a trigger can have size ranging << 1 GB to ~100 TB!
- **The rate of events varies widely from few Hz to <<1/month**
- **The trigger selection need to accumulate data from detectors over several seconds**
  - readout needs very large buffers to accommodate the long decision latency
  - fast storage of 3.5 TB with a sequential write performance @ 25 GBps
- **Complexity and size are similar, but uptime is much larger (100% instead of ~30%)**
- **Limited accessibility makes things more complex**

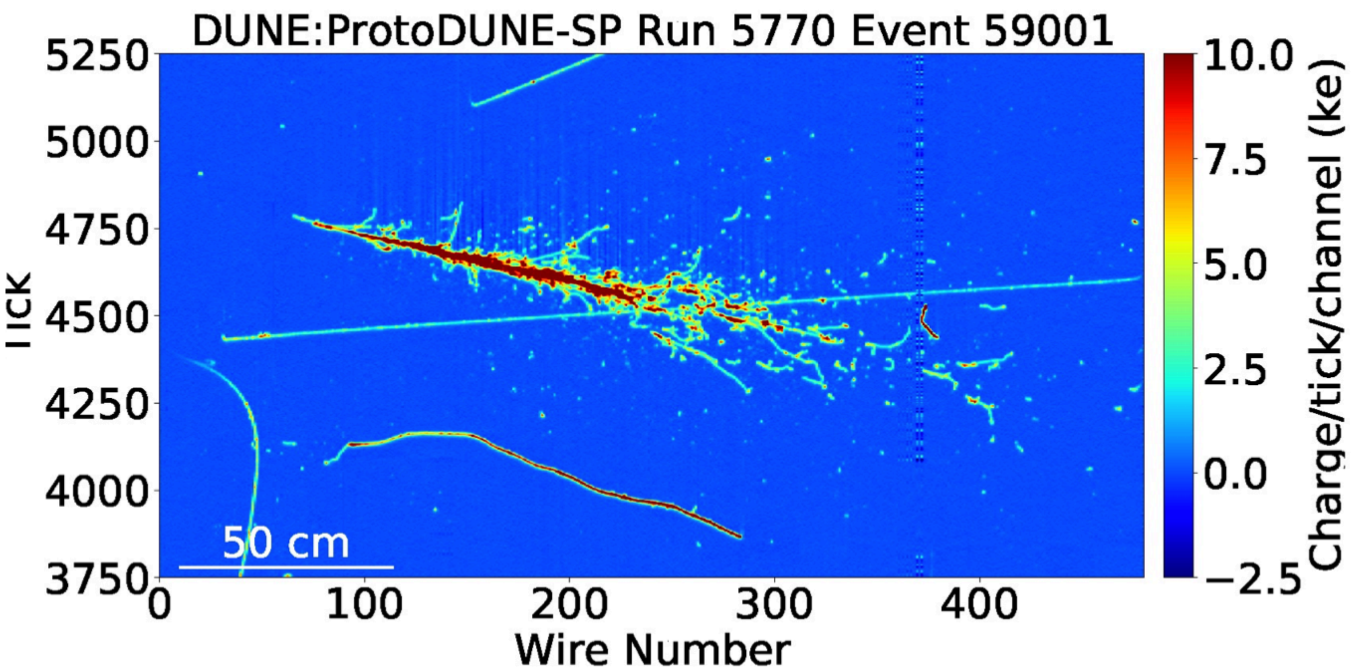
DIFFERENCE WITH COLLIDER EXPERIMENTS

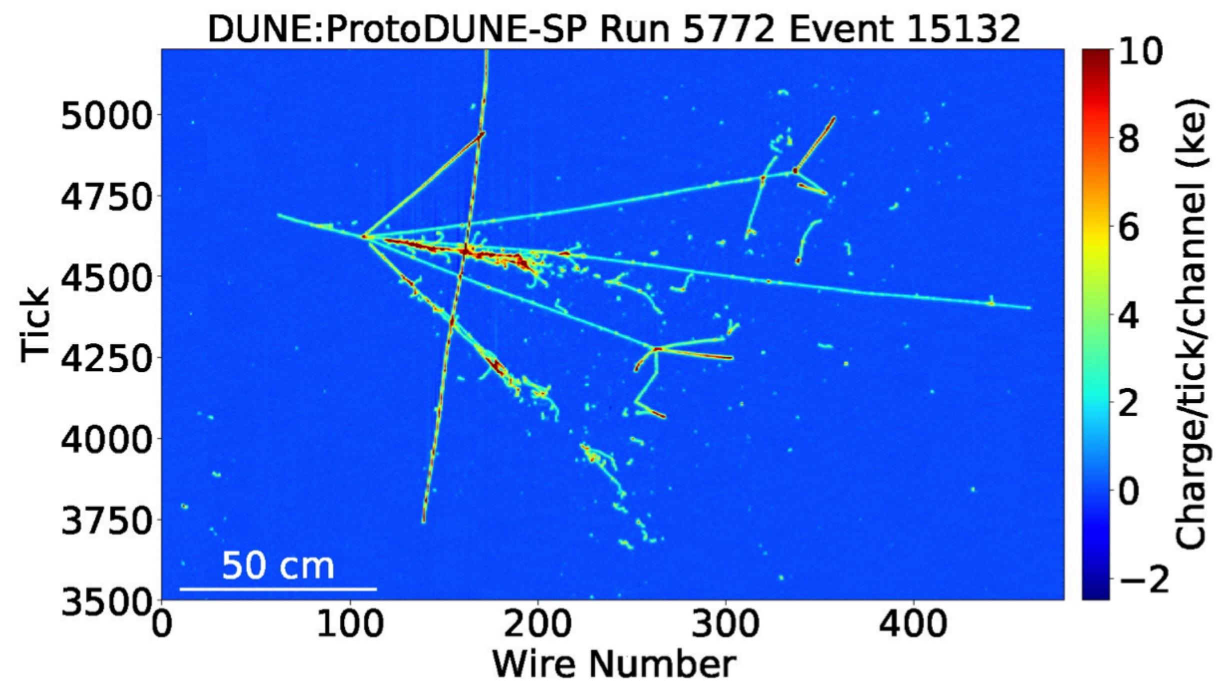
- **Differently from LHC, time frames varies a lot**
  - from few ms to ~100s for the supernova core collapse
  - Data corresponding to a trigger can have size ranging << 1 GB to ~100 TB!
- **The rate of events varies widely from few Hz to <<1/month**
- **The trigger selection need to accumulate data from detectors over several seconds**
  - readout needs very large buffers to accommodate the long decision latency
  - fast storage of 3.5 TB with a sequential write performance @ 25 GBps
- **Complexity and size are similar, but uptime is much larger (100% instead of ~30%)**
- **Limited accessibility makes things more complex**
- **The control and monitoring system will have a predominant role for the success of the DUNE TDAQ**
  - Automated anomaly detection and recovery
  - Remote monitoring and control

DUNE SOFTWARE TRIGGER



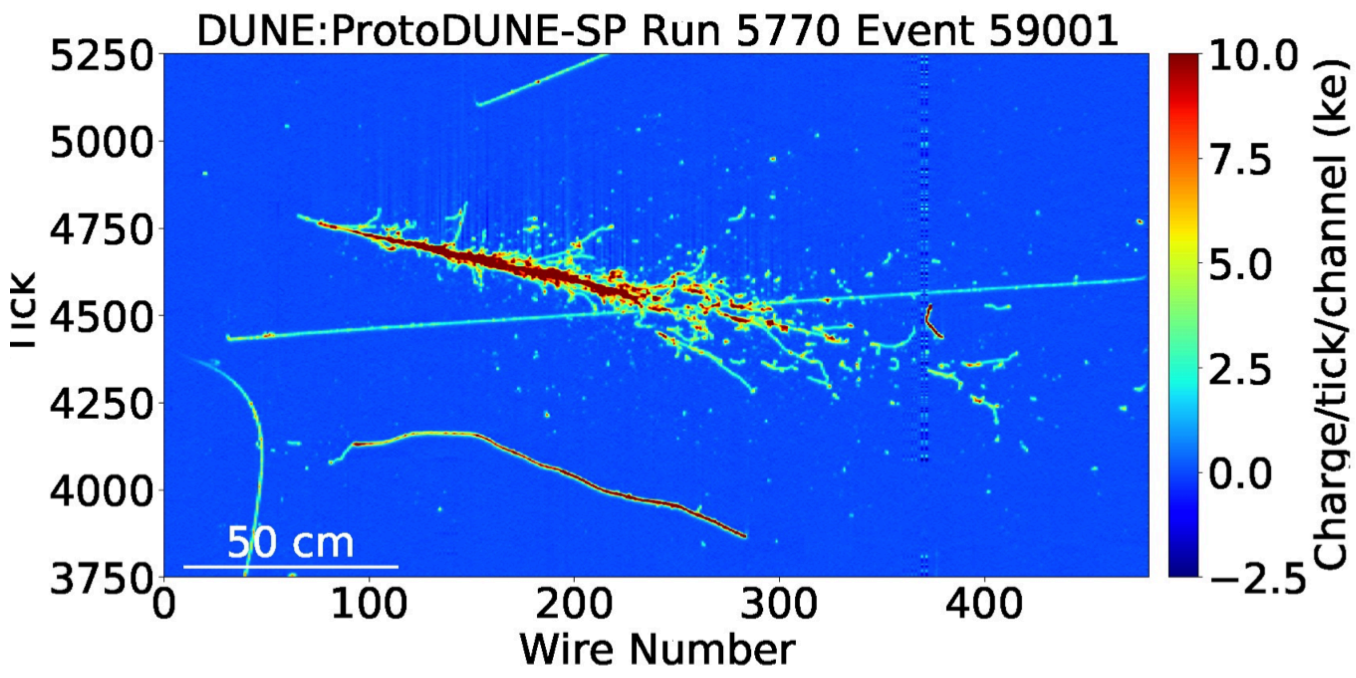
DUNE SOFTWARE TRIGGER

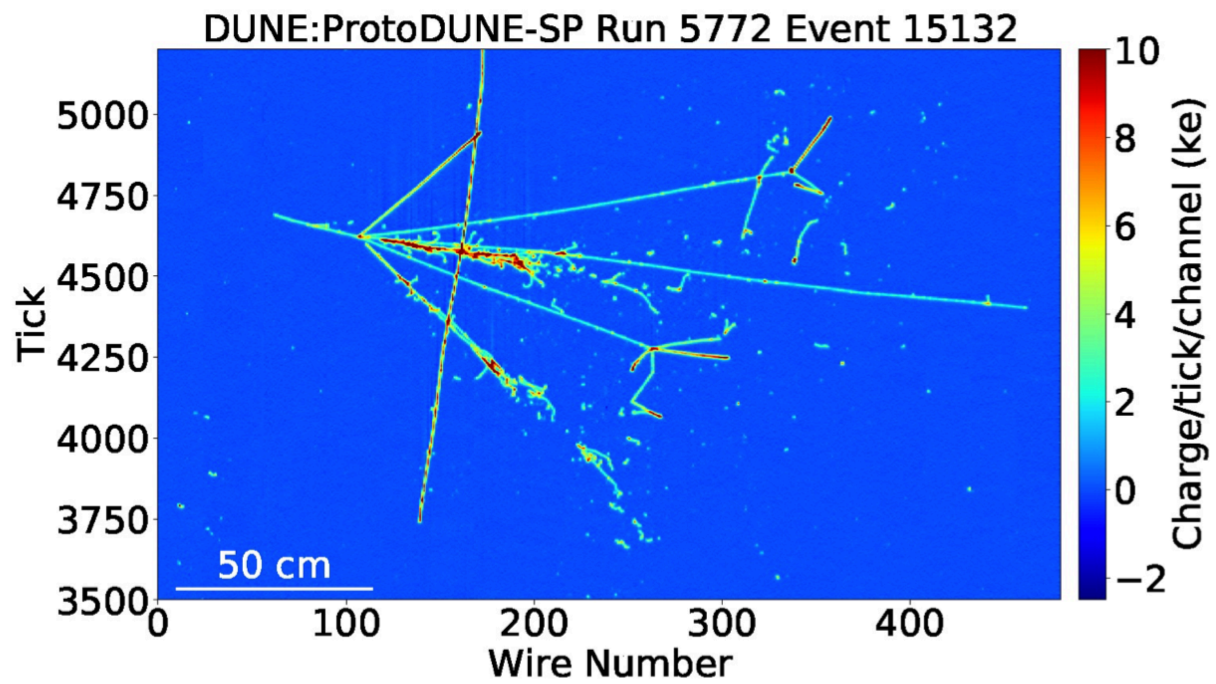




➡ **Why?**

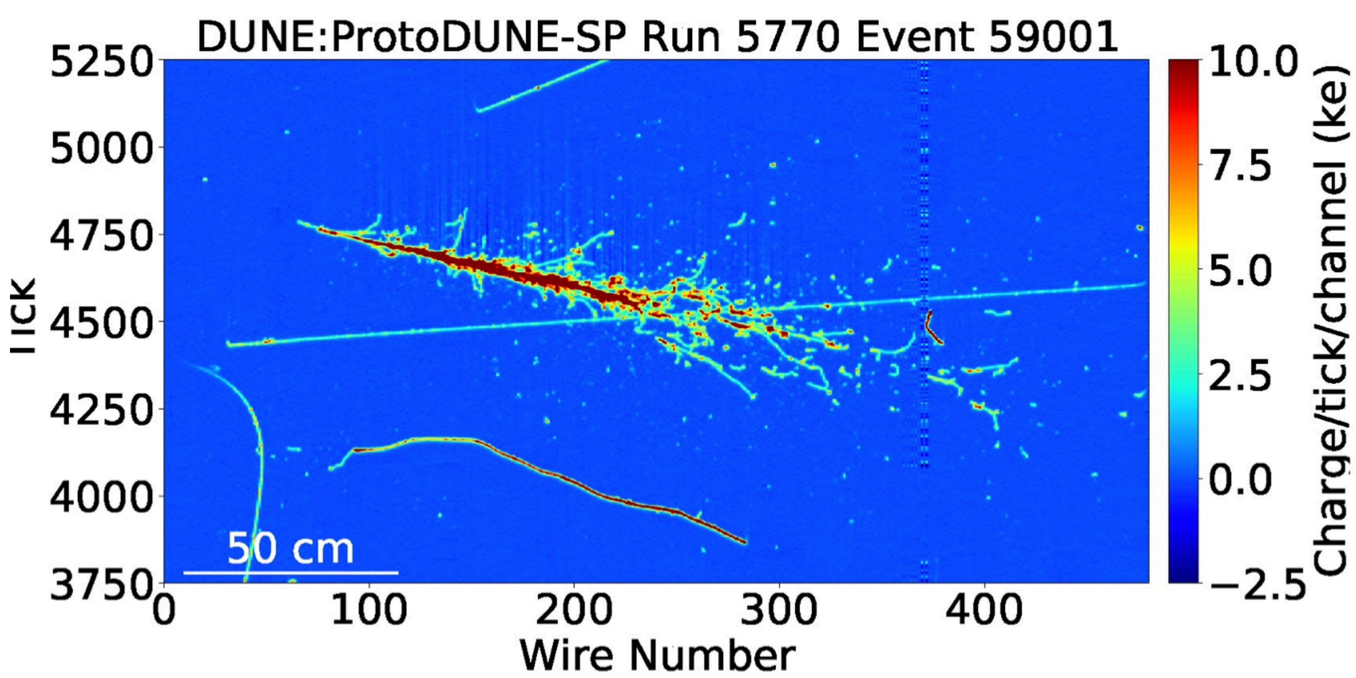
DUNE SOFTWARE TRIGGER

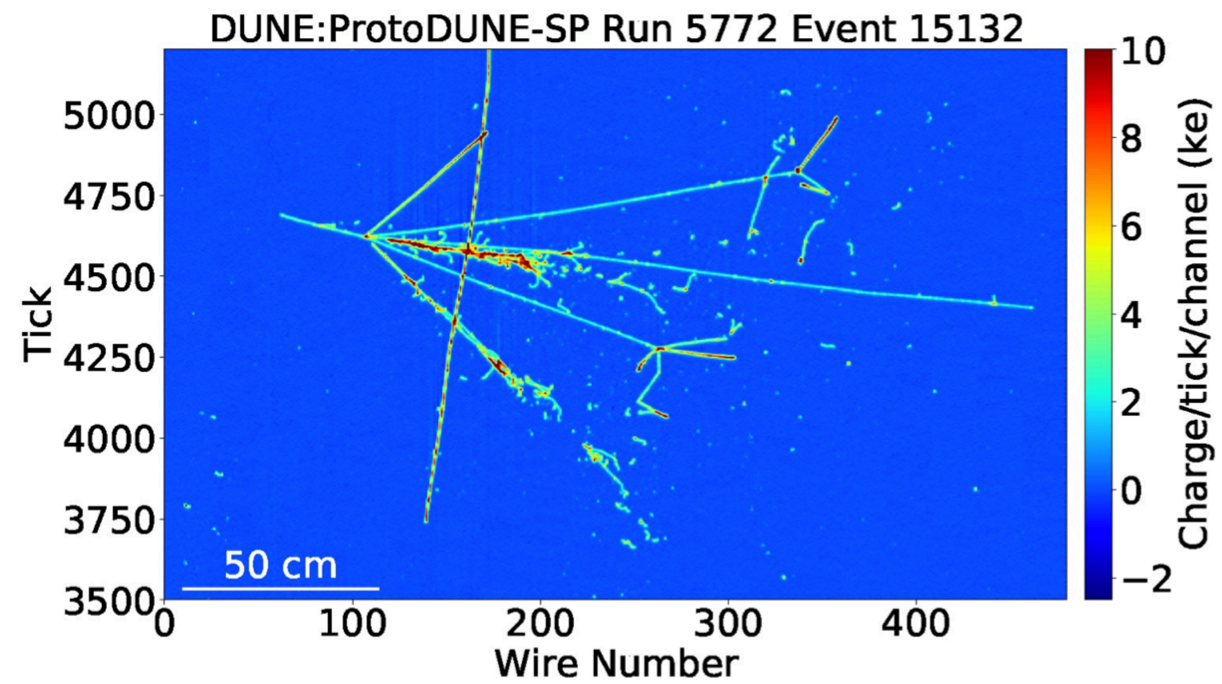




- ➡ **Why?**
- ➡ **TPC Information is very rich**
  - ➡ triggering algorithms are more sophisticated than what a hardware trigger could do

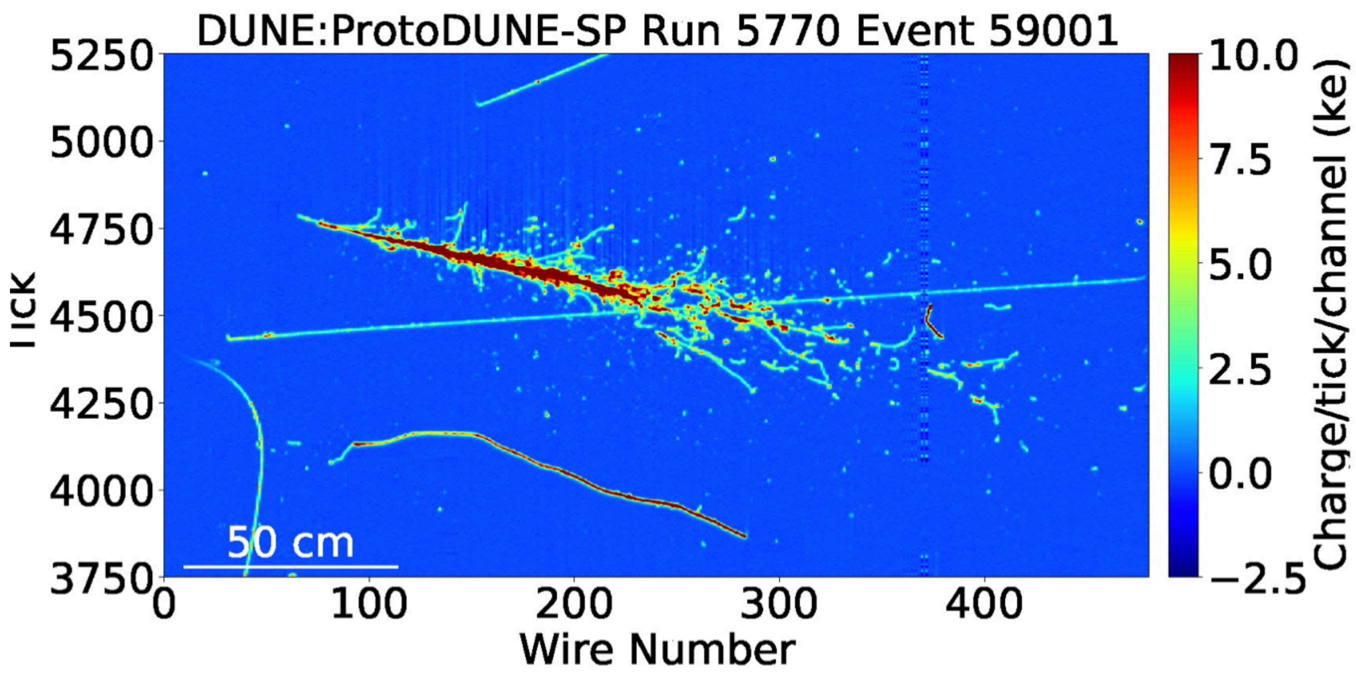
DUNE SOFTWARE TRIGGER

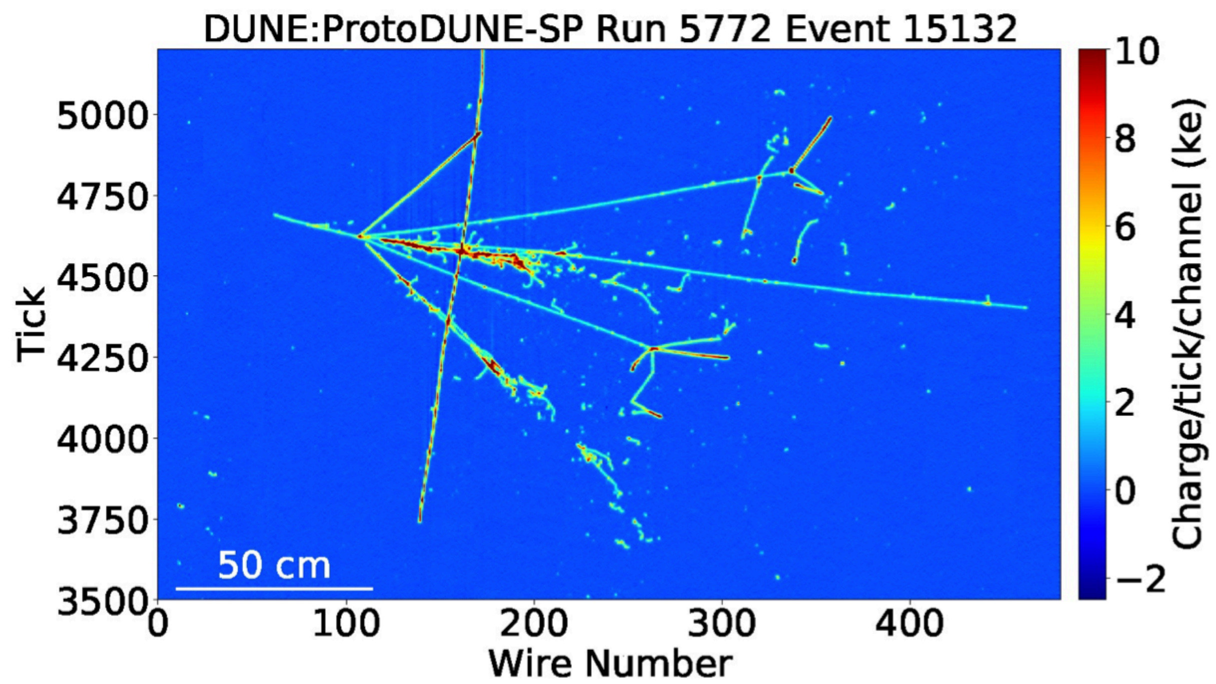




- **Why?**
- **TPC Information is very rich**
  - triggering algorithms are more sophisticated than what a hardware trigger could do
- **TPC is also very slow and u/g rates are very low...**
  - Plenty of time to make decisions, large buffers add more time
  - Not naturally “friendly” to a hardware approach

DUNE SOFTWARE TRIGGER





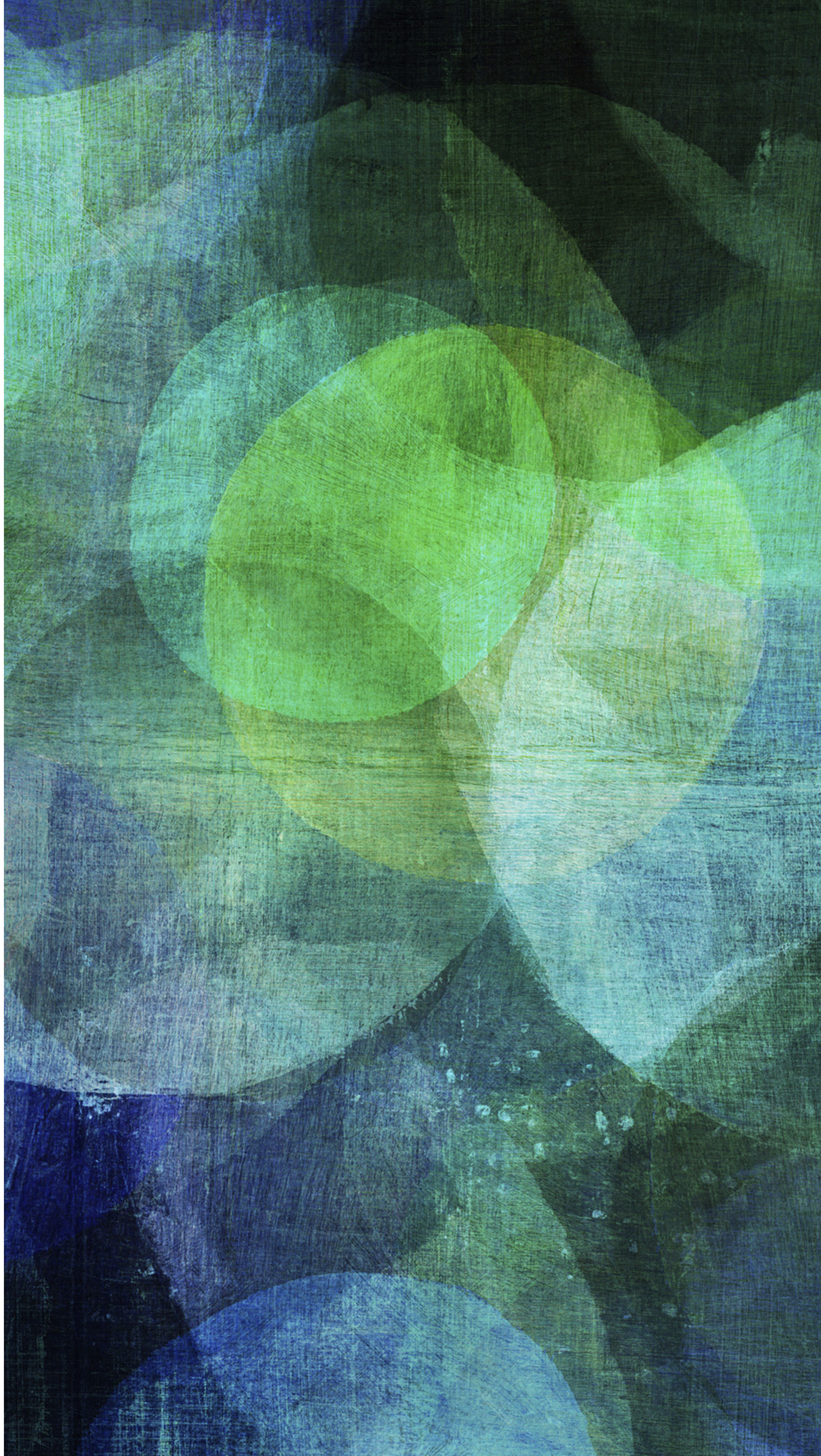
- ➡ **Why?**
- ➡ **TPC Information is very rich**
  - ➡ triggering algorithms are more sophisticated than what a hardware trigger could do
- ➡ **TPC is also very slow and u/g rates are very low…**
  - ➡ Plenty of time to make decisions, large buffers add more time
  - ➡ Not naturally “friendly” to a hardware approach
- ➡ **Want out-of-beam triggering for broad program**
  - ➡ And beam information may be slow to arrive anyway

IT'S ALL ABOUT PHYSICS

- **The knowledge of hardware and software technologies is becoming critical in our community**
  - thanks to this school we try to keep a high level
- **The physics goals depends on technology and innovation**
  - Particle physicists must monitor technological trends and make innovation (especially true in TDAQ field)
- **Not always easy to make extrapolations for the future**

- **[Snowmass 2022 report]**
  - *"Modern computing architectures and emerging technologies are changing the way we do particle physics"*
  - *"Machine learning was essentially not a part of the 2013 Snowmass report"*
- **[ATLAS TDR, 2003]**
  - *"Thanks to the Moore law, in 2007 our event selection farm will be based on 8 GHz CPUs"*
- **[Ken Olsen, Founder of DEC, 1977]**
  - *"There is no reason anyone would want a computer at home."*

BACK-UP SLIDES



CMS: 2-STAGE EVENT BUILDING IN RUN 1

Cannot do Event Building at 100 kHz

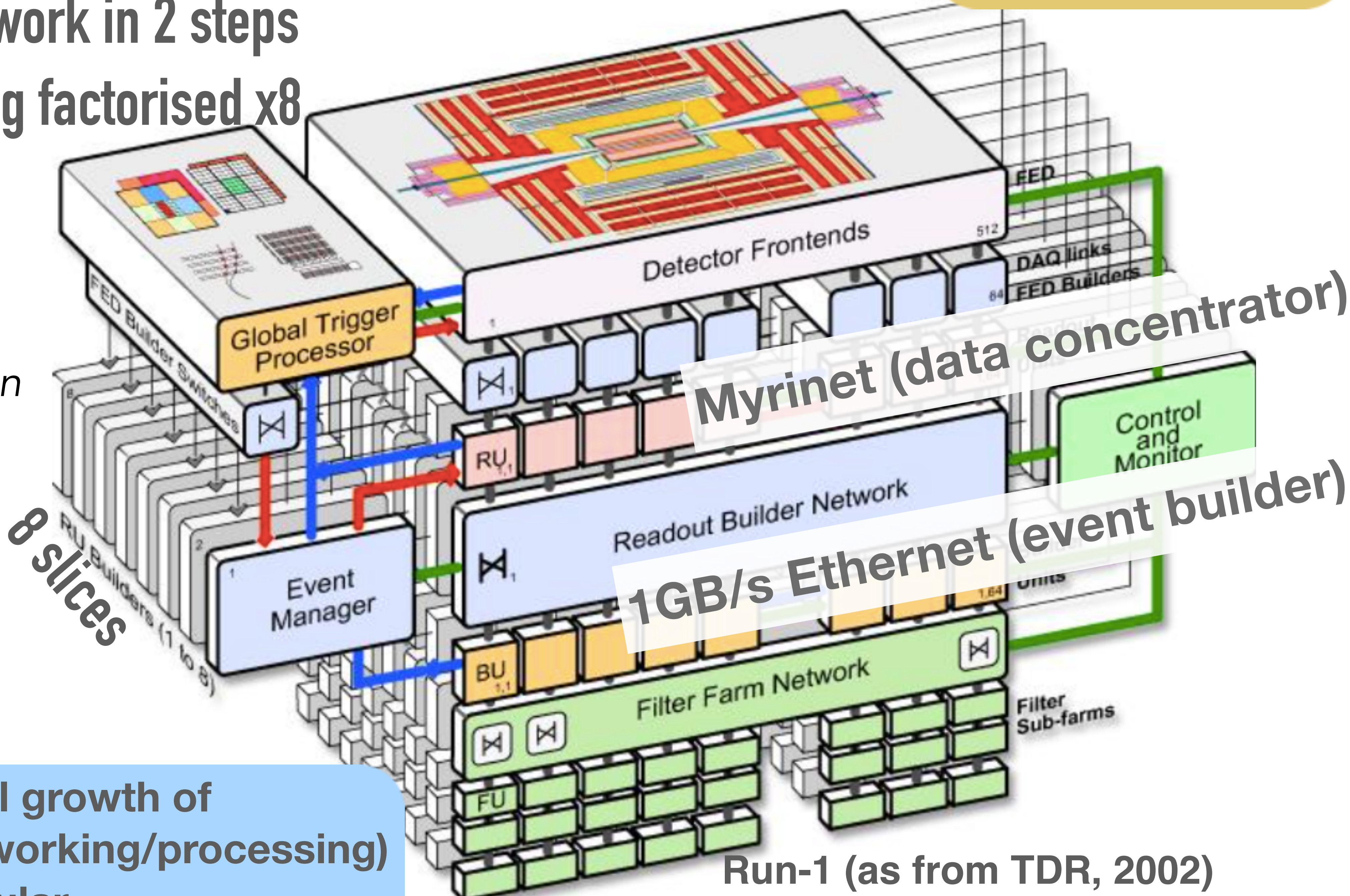
CMS DAQ-1

100 GB/s readout network in 2 steps

100 kHz Event Building factorised x8

2 EB networks in blu

Filter network in green

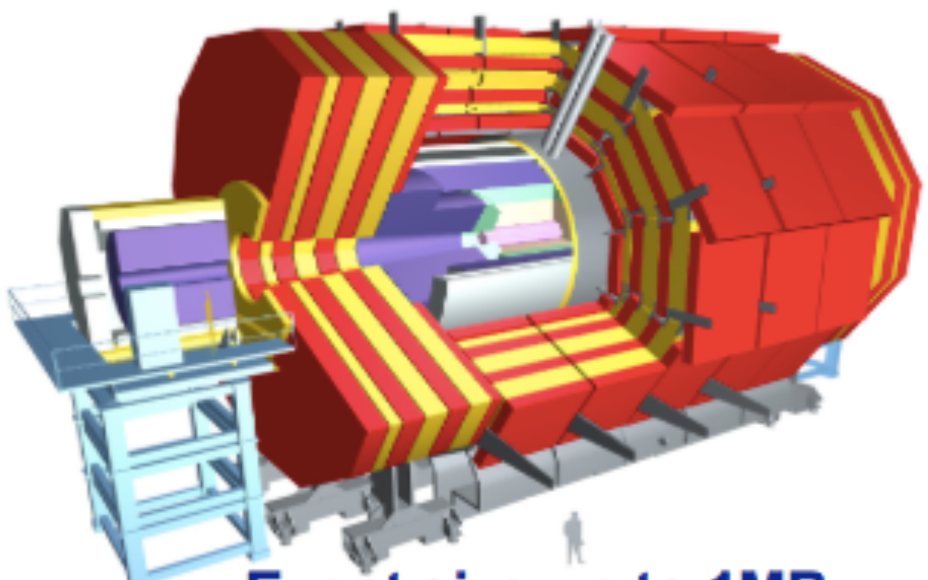


- **Bet on exponential growth of technologies (networking/processing)**
- **Scalable and modular**
  - Independent development of two network technologies

Run-1 (as from TDR, 2002)

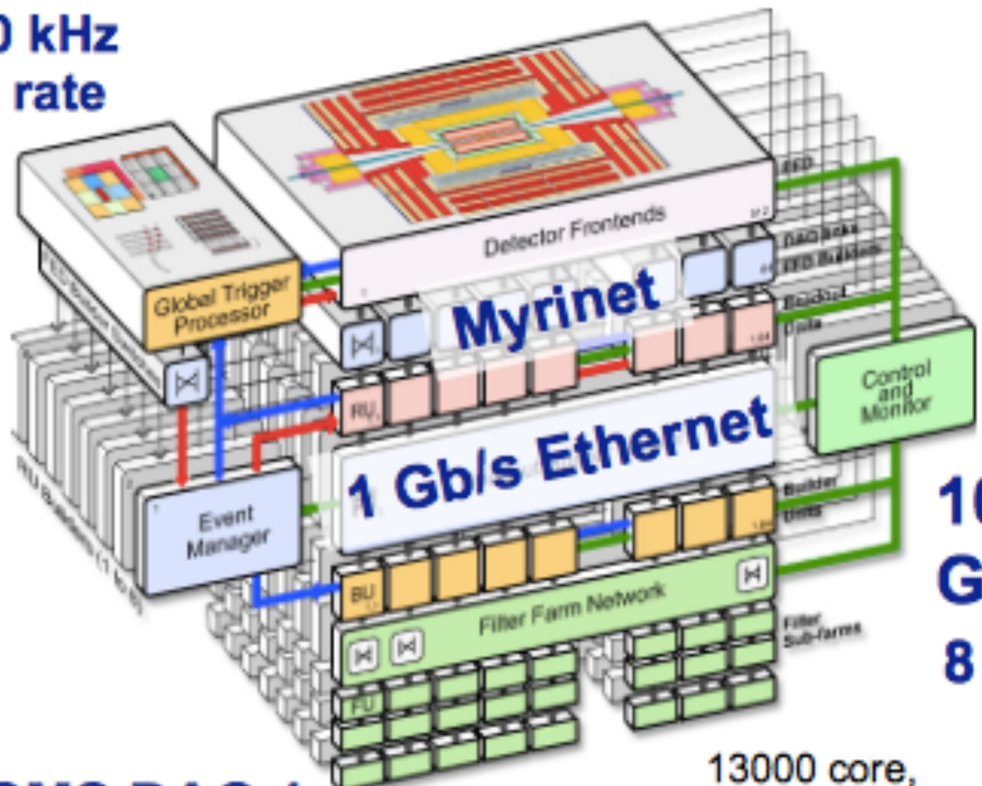
- Myrinet + 1GBEthernet
- 1-stage building: 1200 cores (2C)
- HLT: ~13,000 cores
- 18 TB memory @100kHz: ~90ms/event

EVOLUTION FROM RUN-1 TO RUN-2



Event size up to 1MB

100 kHz
L1 rate

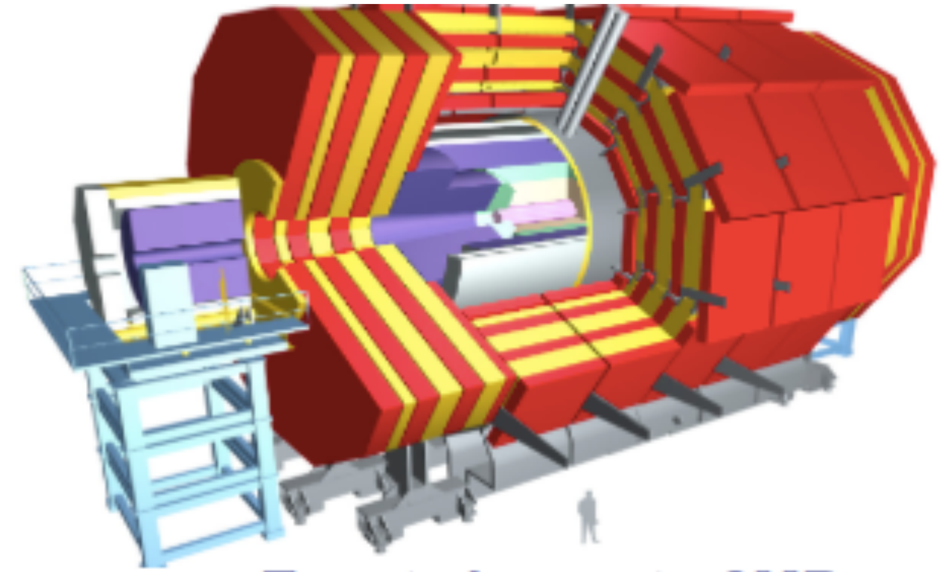


100 GB/s
8 slices

CMS DAQ 1

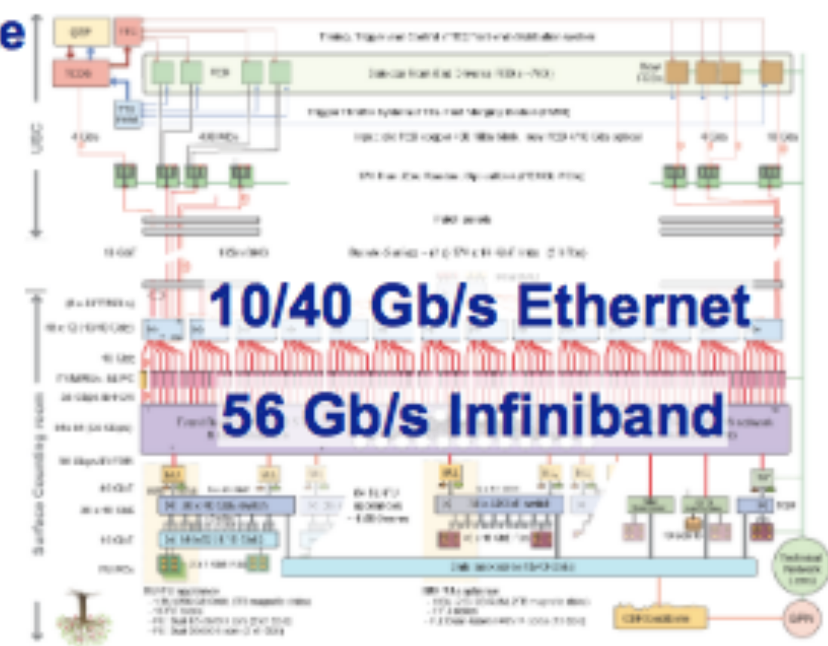
13000 core,
1260 host
filter farm

max. 1.2 GB/s to storage



Event size up to 2MB

100 kHz
L1 rate



~200 GB/s

1 slice

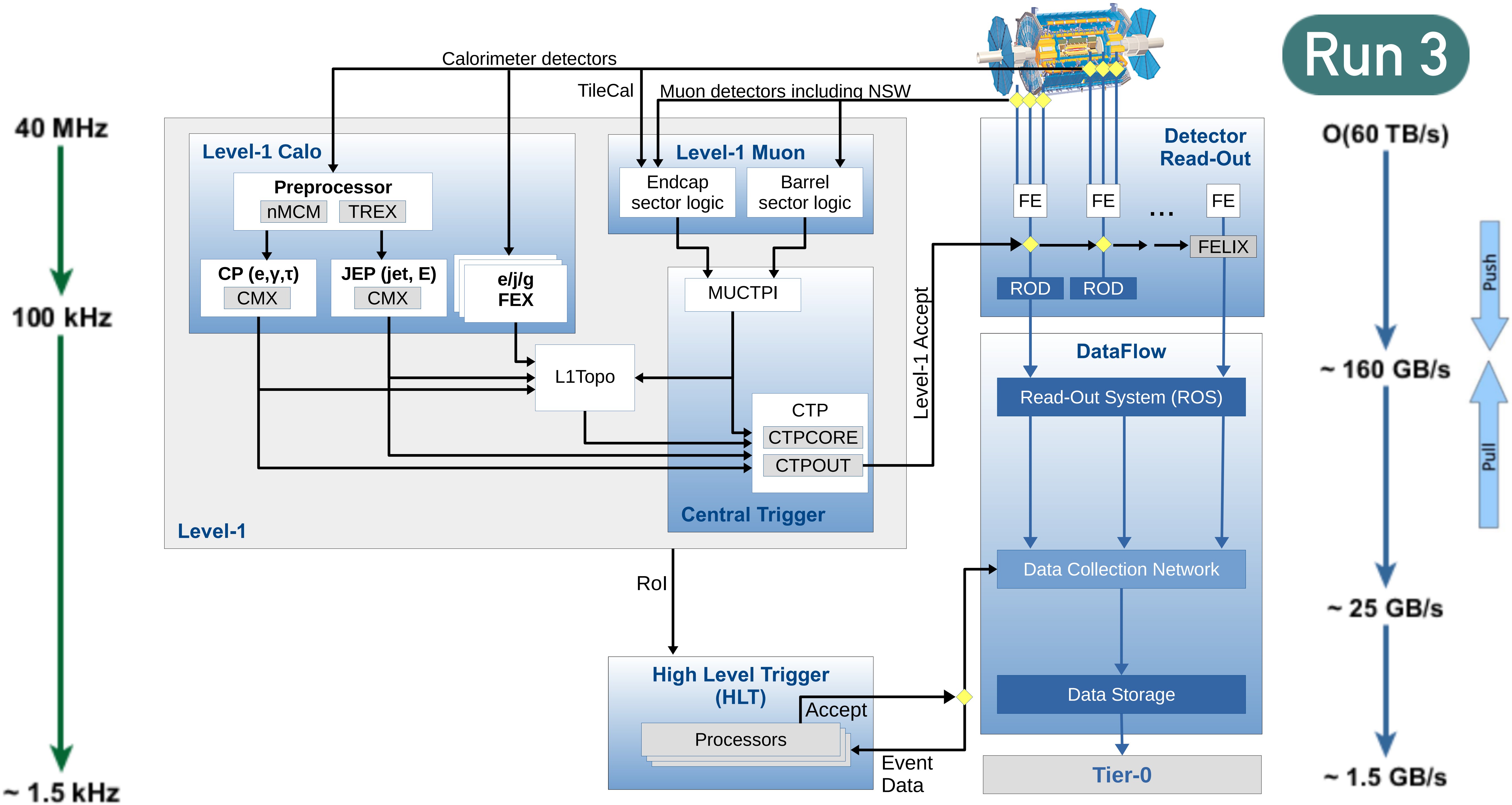
CMS DAQ 2

16000+ core,
900 host
filter farm

~ 3-6 GB/s to storage

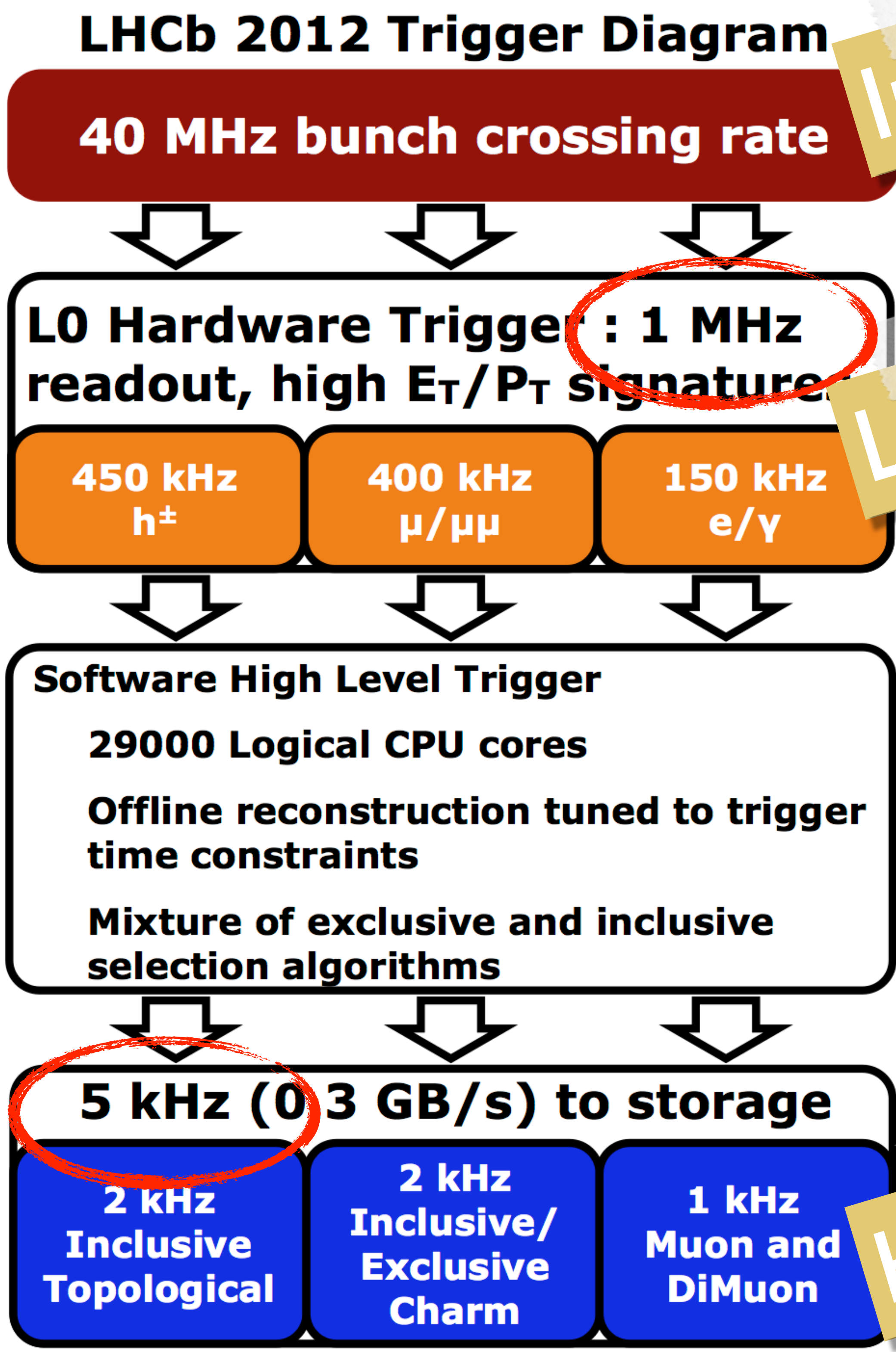
ATLAS REGIONAL TDAQ ARCHITECTURE

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

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

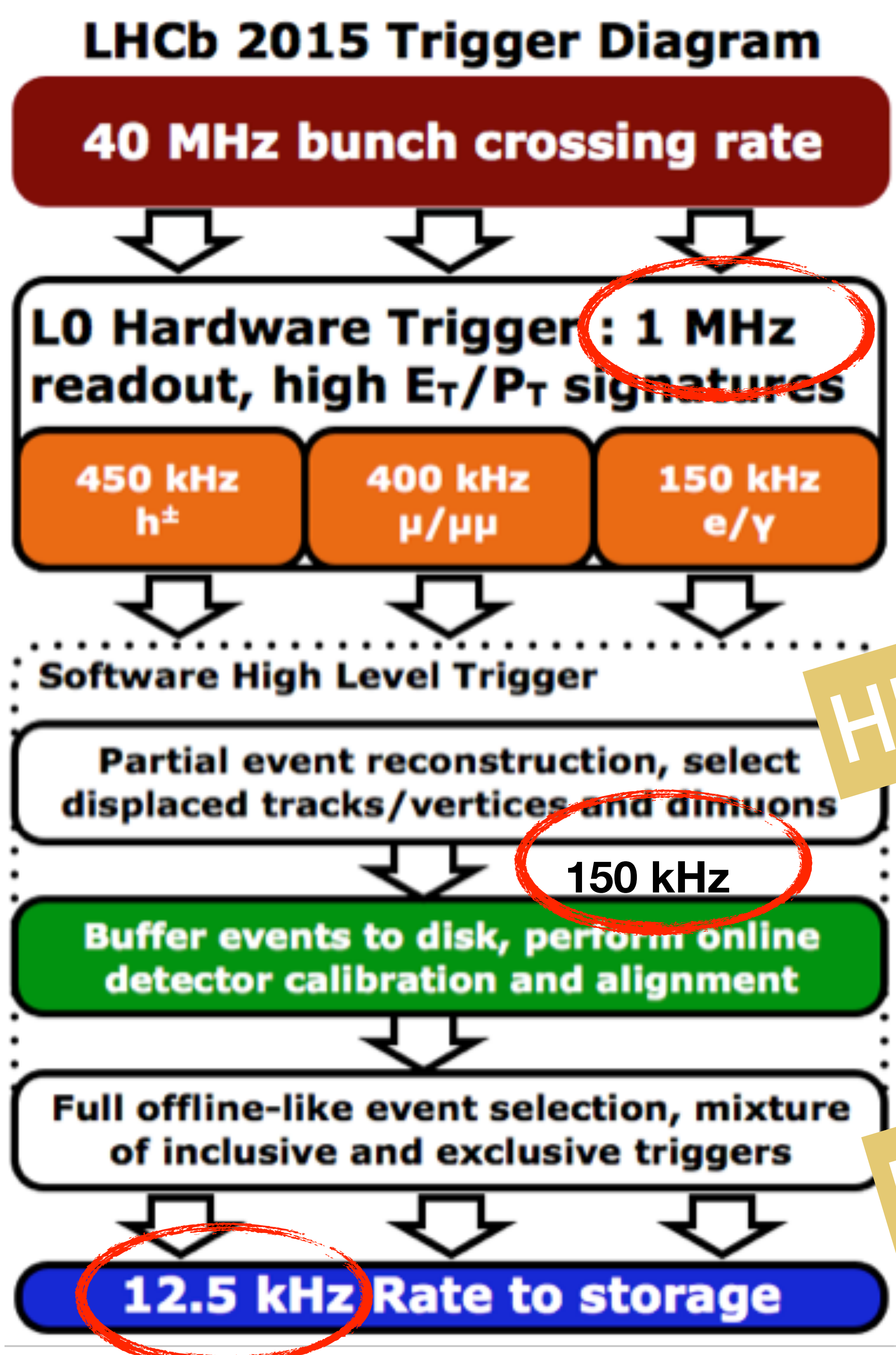
- Select Bs in hadronic triggers
- Reject complex/busy events

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

High Level

- Multitude of exclusive selections

SCHEMA EVOLUTION



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

Real-time calibration and alignments

HLT-1

Synchronous with DAQ

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

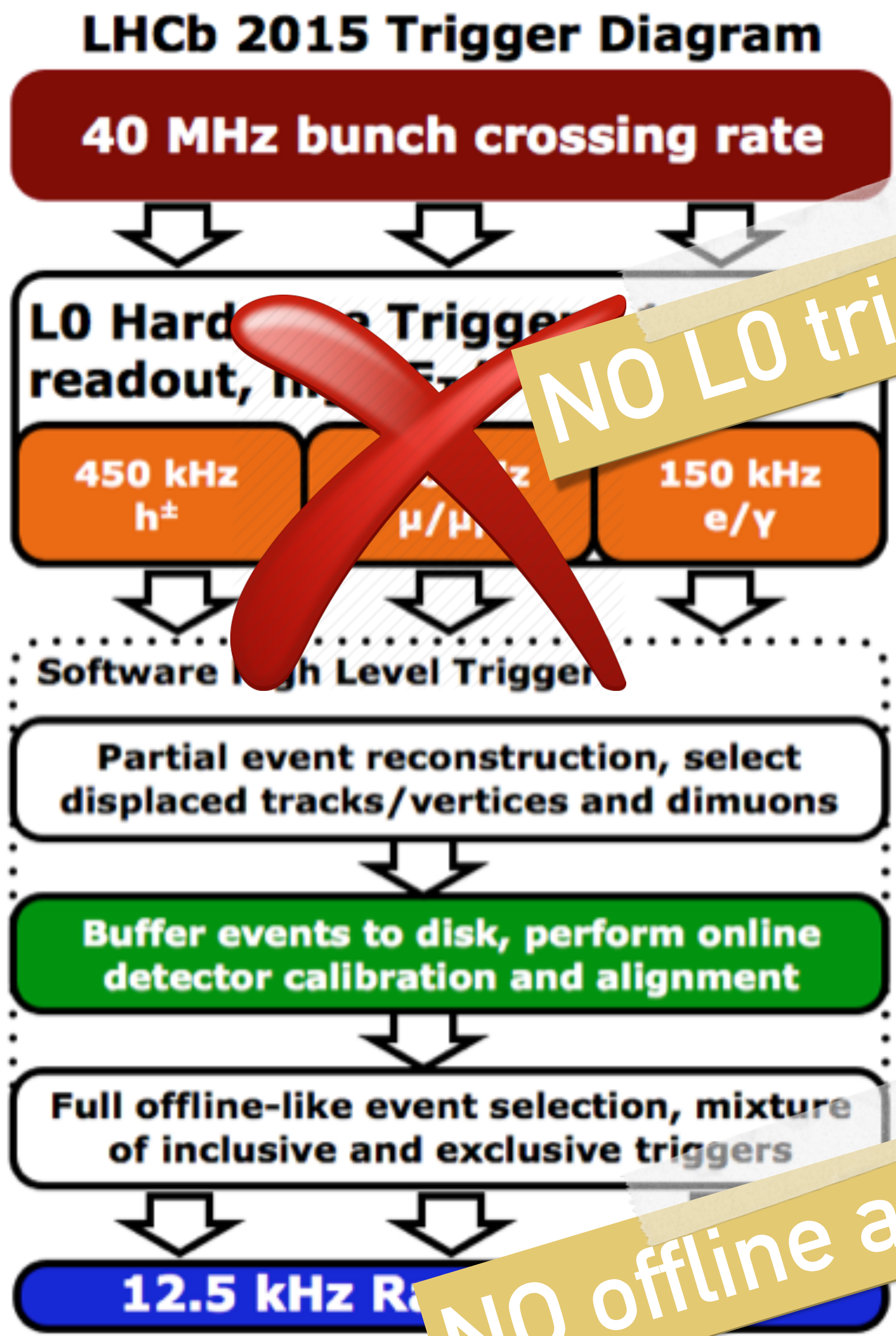
Split with a large buffer (4PB)!

HLT-2

Deferred Processing

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

UPGRADES FOR RUN 3



NO L0 trigger

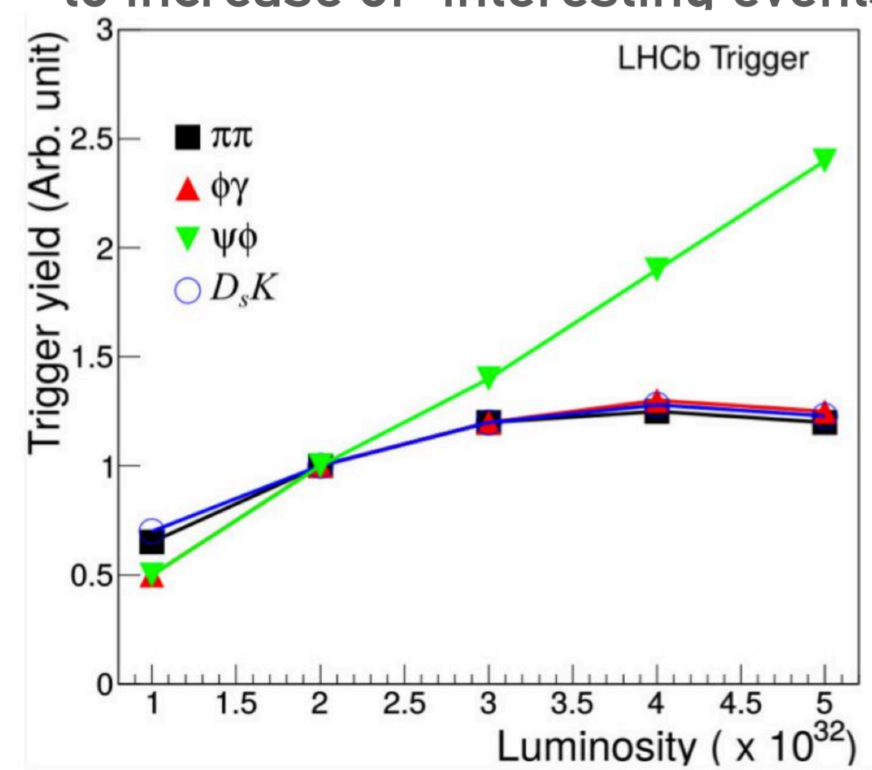
NO offline analysis

See Phase-I upgrade TDR

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

YES, remove limit from L0 -1MHz readout!

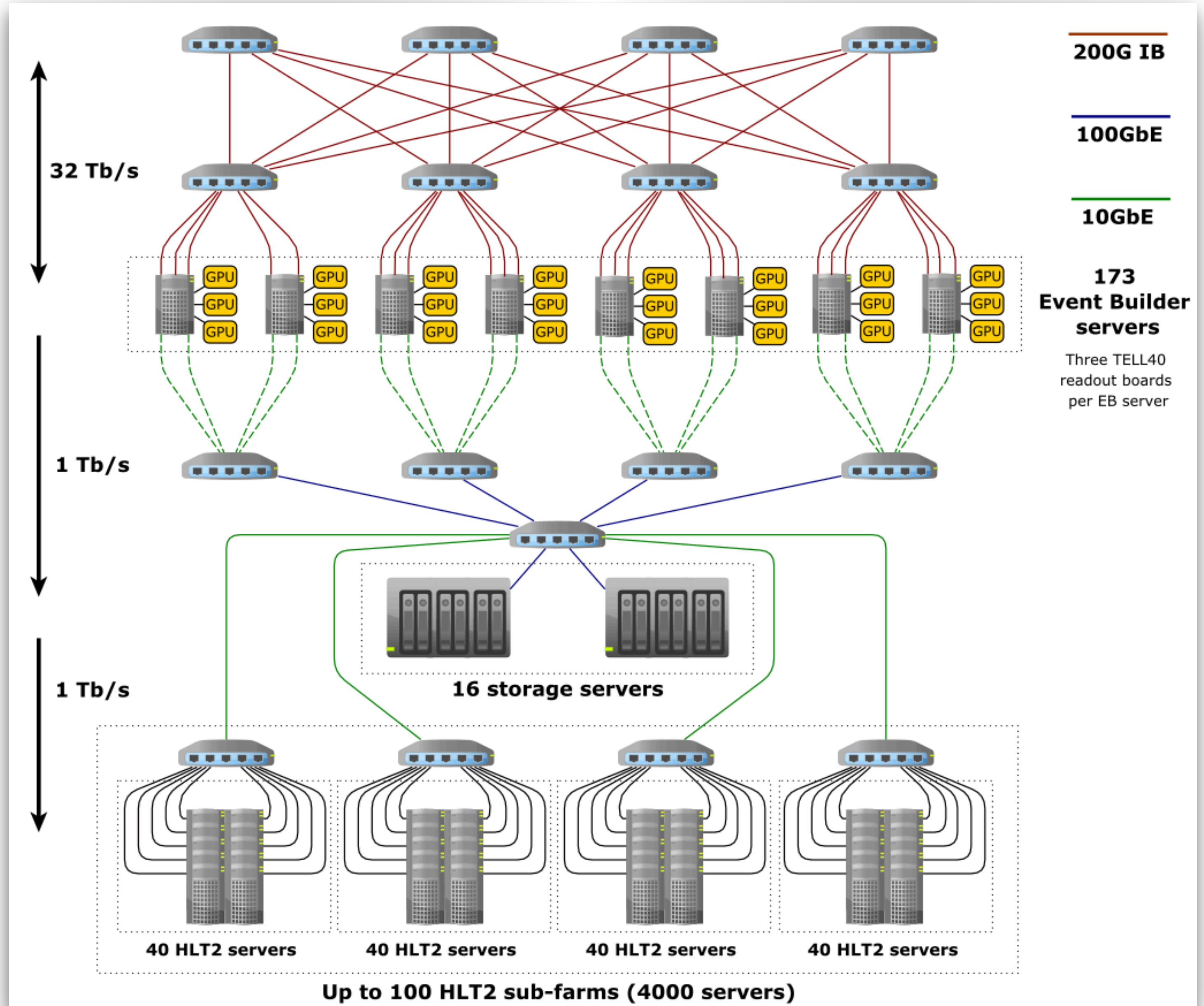
Increase in luminosity does not lead to increase of "interesting events"



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

A 2-DIM FOLDED EVENT BUILDING

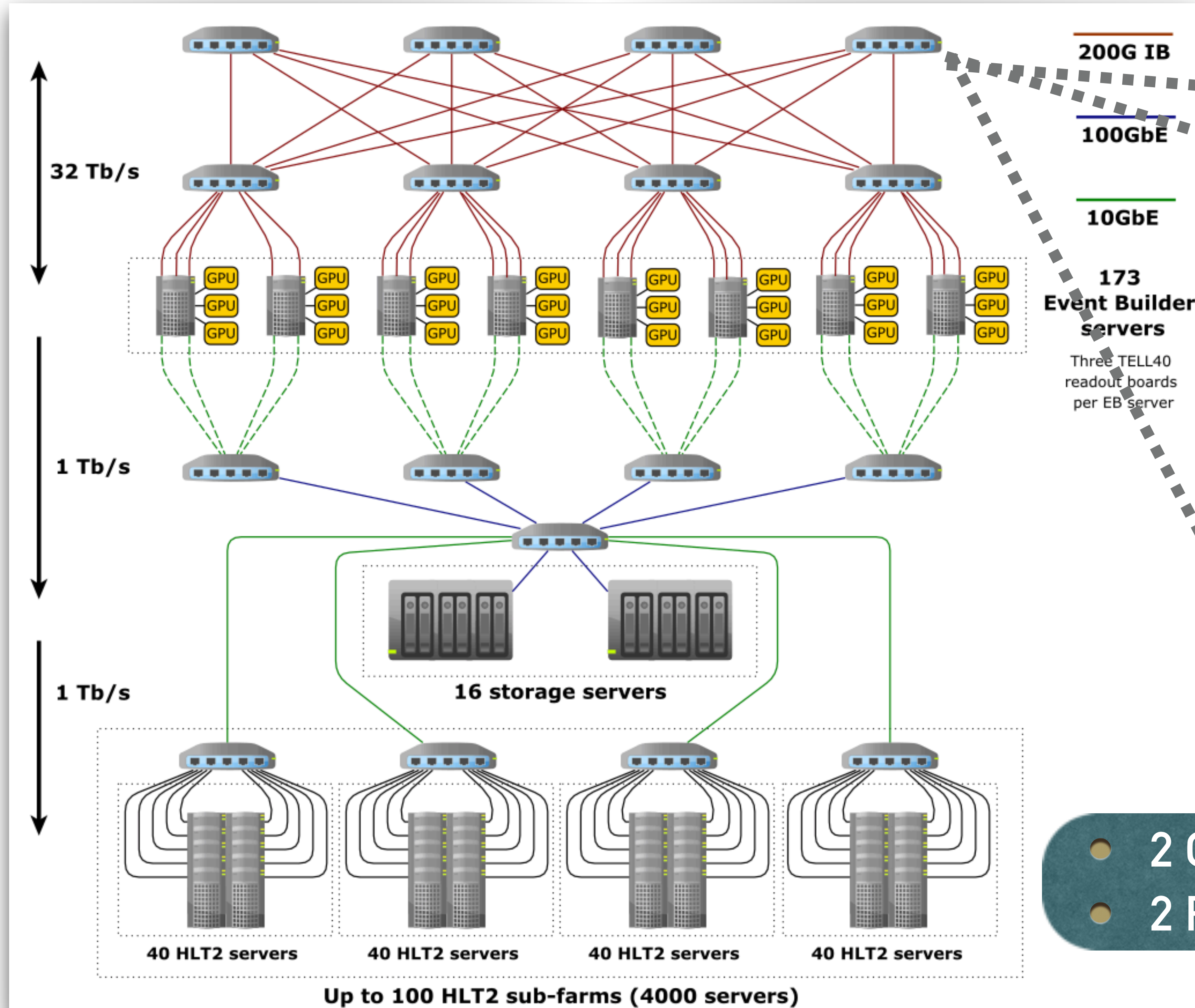
Large farm of equal nodes with 8 PCIe40 boards, specialised by firmware

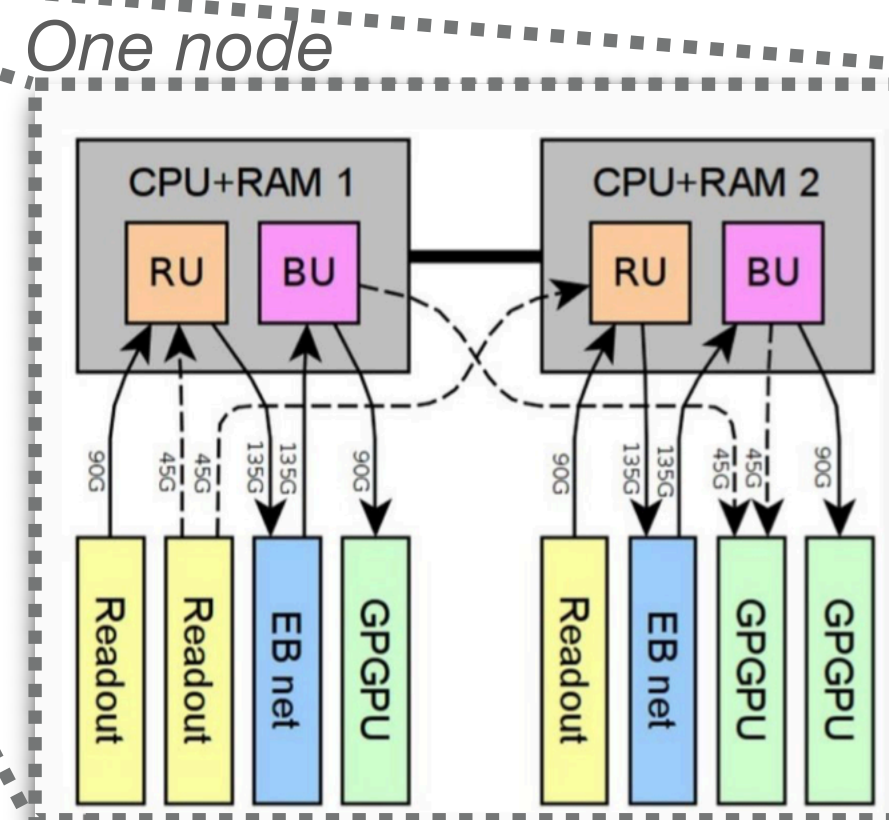


- ➡ **EB network is oversized: able to manage 64Tb/s (320 network cards x 200Gb/s)**
- ➡ **Large rejection at HLT1: use O(200) GPU! throughput at ~100kHz**
- ➡ **Storage Buffer HLT1-HLT2 = 40 PB (3000 hard-disks) enough for days**
  - ➡ SSD faster but have short lifetime wrt high read-write rate, so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

A 2-DIM FOLDED EVENT BUILDING

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





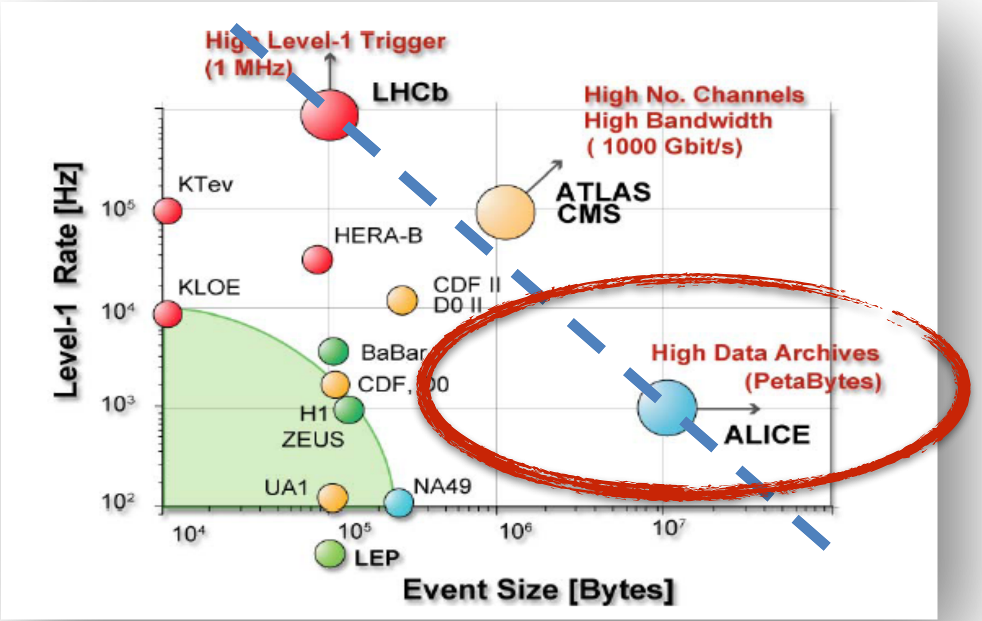
- 2 CPUs with large RAM (up to 512 GB!)
- 2 RU, 2 BU, 2 infiniband NIC (200 Gb/s), 1-3 GPUs

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

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

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

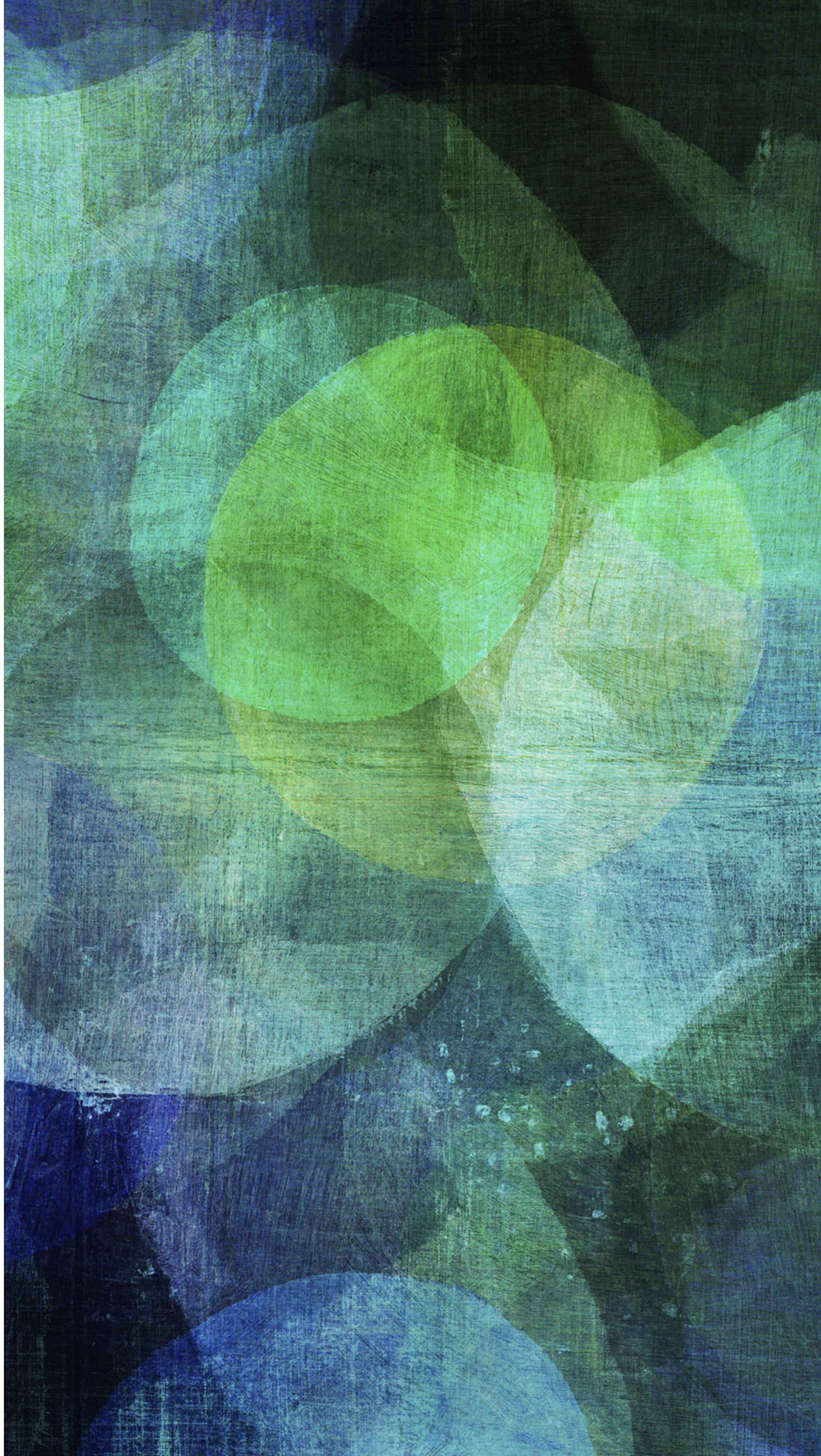
  ➡ SSD faster but have short lifetime wrt high read-write rate, so prefer hard-disks



ALICE: THE SMALL BIG-BANG

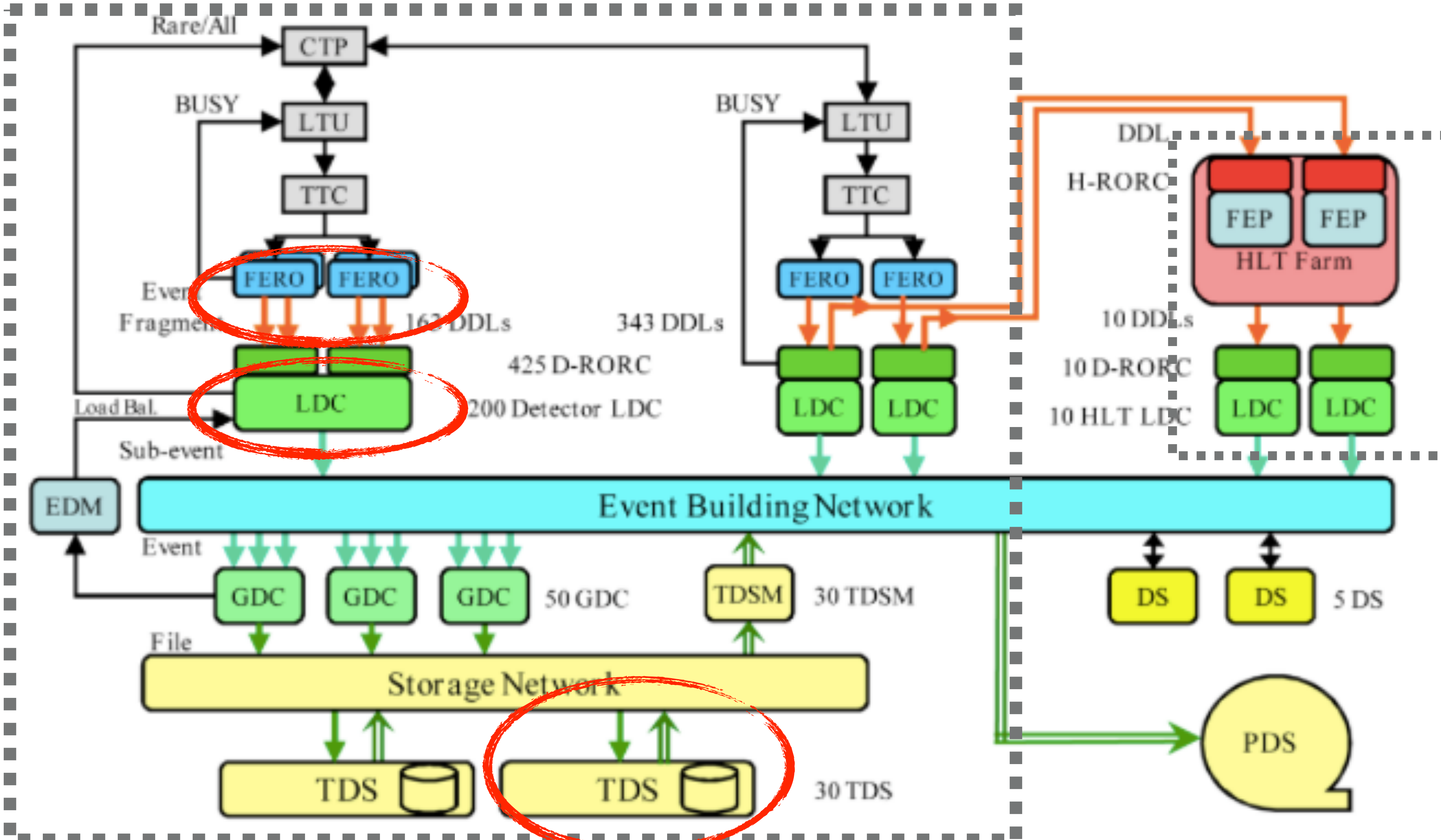
Recording heavy ion collisions

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



READOUT DATA CONCENTRATORS

ALICE



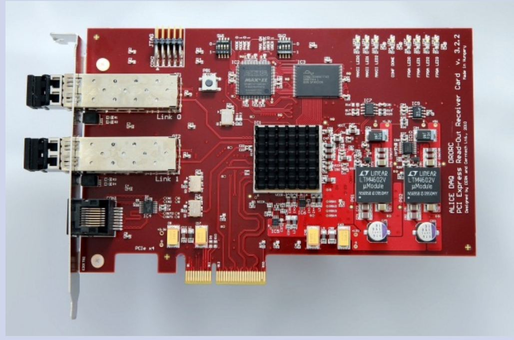
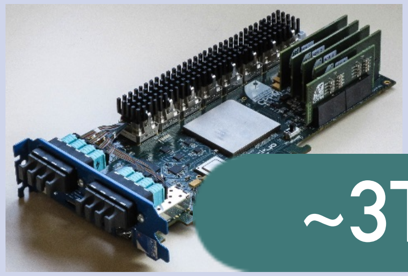
- ➡ **Dataflow with local (LDC) and global (GDC) data concentrators**
  - ➡ Detector readout (~20 GB/s) with point-to-point optical links (DDL, max 6Gb/s)
  - ➡ Rate to the LDCs can go above 13 GB/s
- ➡ **Transient Data Storage (TDS)**
  - ➡ Before the Permanent Data Storage (PDS) and publish via the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

UPGRADING TO RUN 3

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

- ➡ Increase detector granularity (===> increase event size!)
- ➡ Increase storage bandwidth x O(100)
  - ➡ Offline reconstruction also challenging due to combinatorics
- ➡ Increase readout rates ~kHz → 50 kHz (===> need new and faster electronics)
  - ➡ Rate very close to TPC readout !!

New TDAQ challenges!

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

Run 1 → LS1 → Run 2 → LS 2 → Run 3

~3TB/s detector readout

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

RUN 3 DAQ: ONLINE RECONSTRUCTION

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

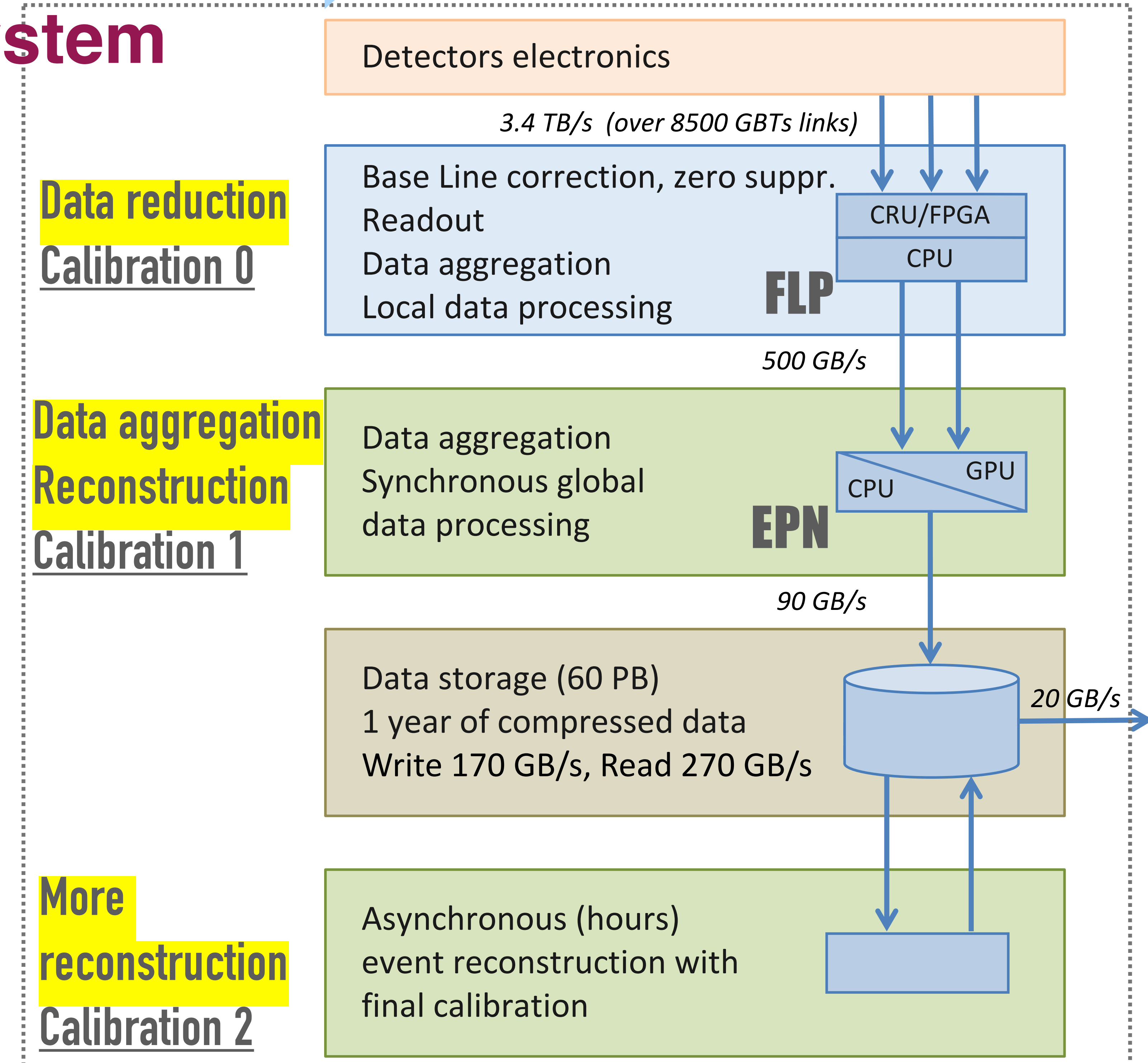
Very heterogeneous system

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

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

O² system

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



SUMMARY OF THE SUMMARIES

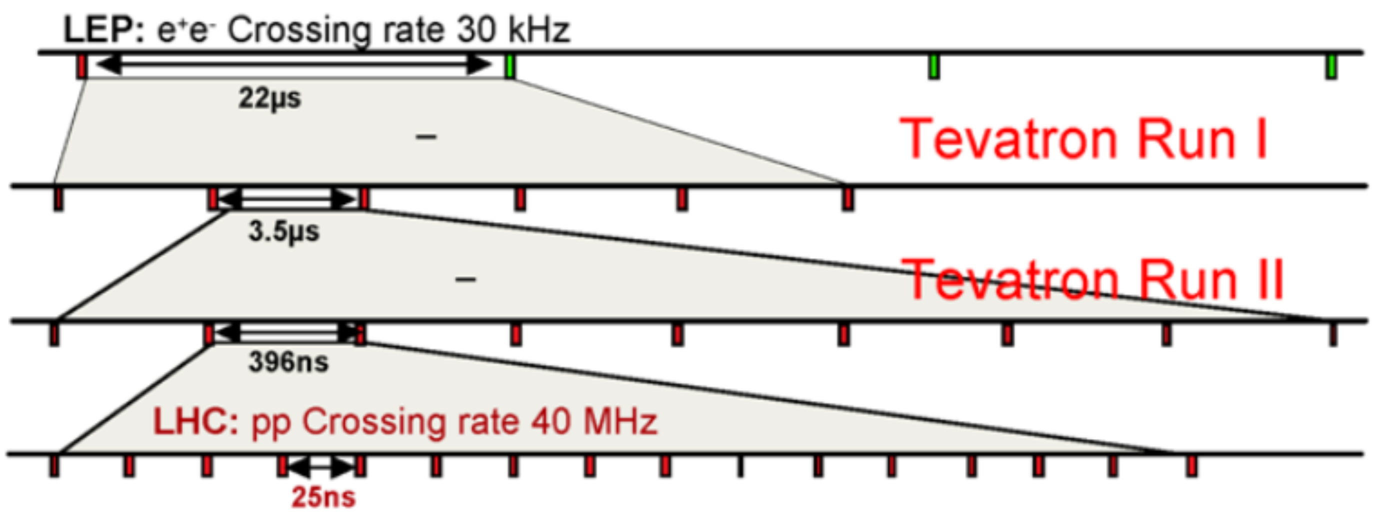
- **LHC experiments are among the largest and most complex TDAQ systems in HEP, to cope with a very difficult environment (always top LHC Luminosity)**
- **Continuous upgrade following the LHC luminosity, with different approaches**
  - **ATLAS/CMS** high-rate readout and Event Building, based on robust 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 comodity 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

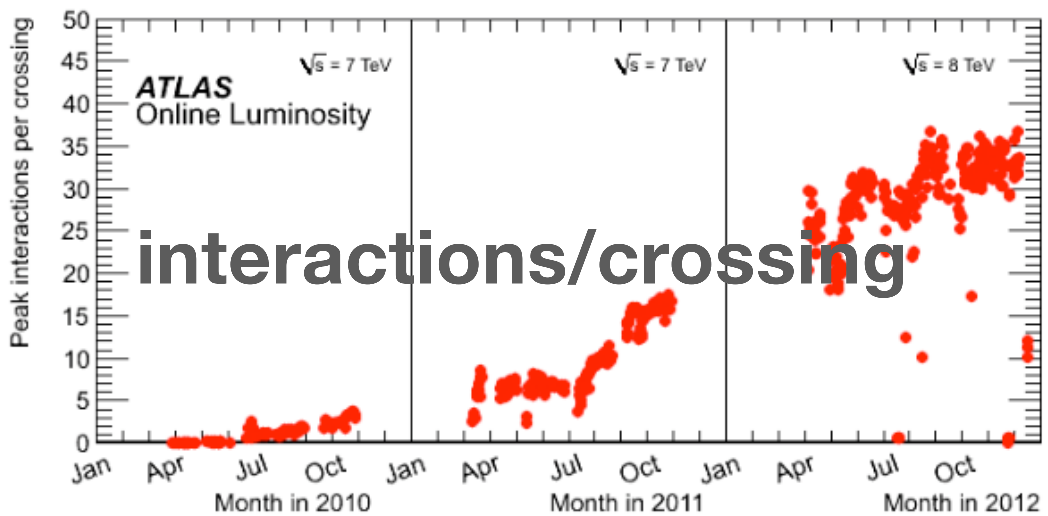
- ~3600 bunches in 27km
- distance bw bunches: 27km/3600 = 7.5m
- distance bw bunches in time: 7.5m/c = 25ns

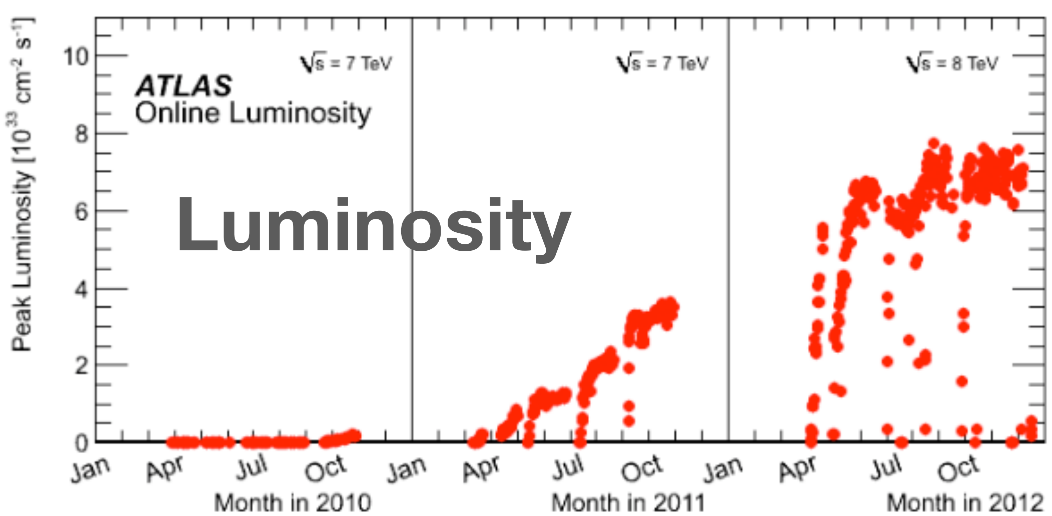


At full Luminosity, every 25ns, ~23 superimposed p-p interaction events

The pile-up source

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

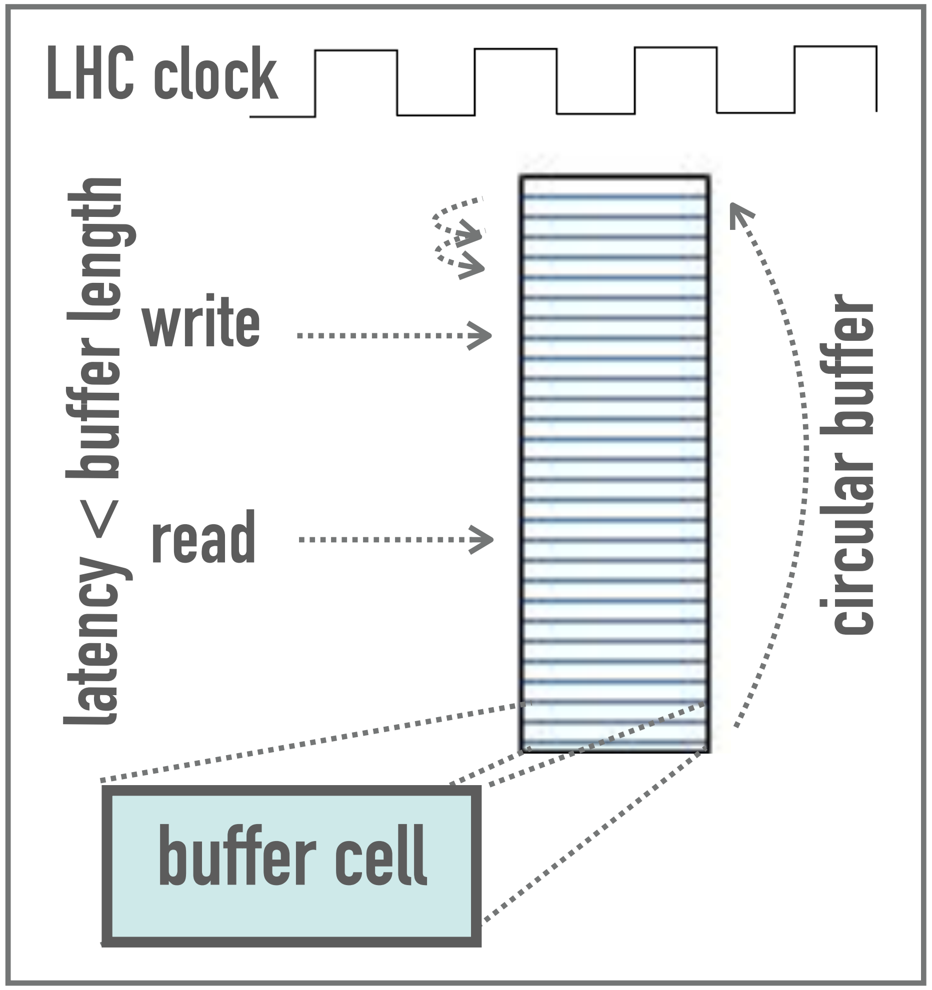




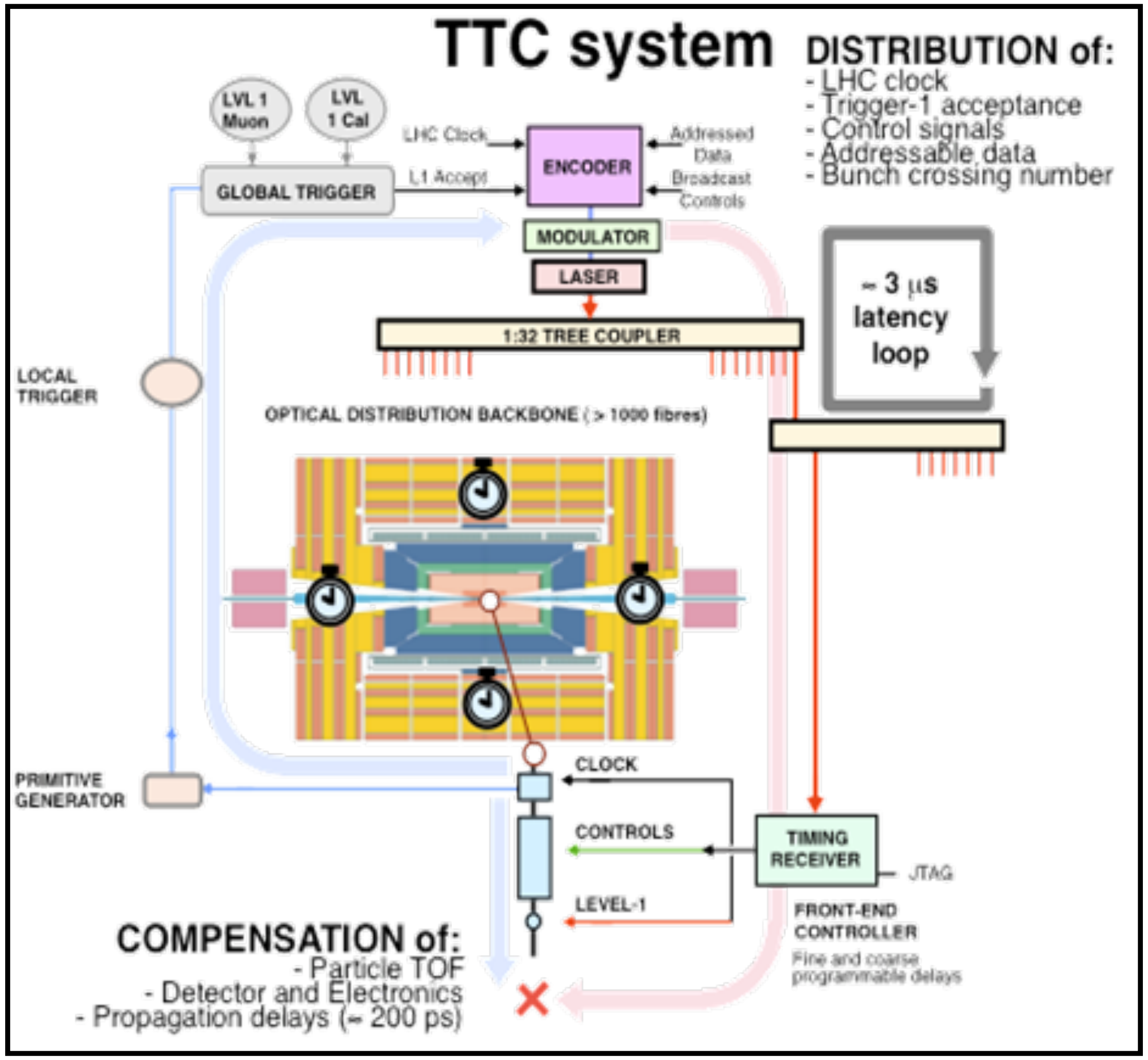
PIPELINED TRIGGERS

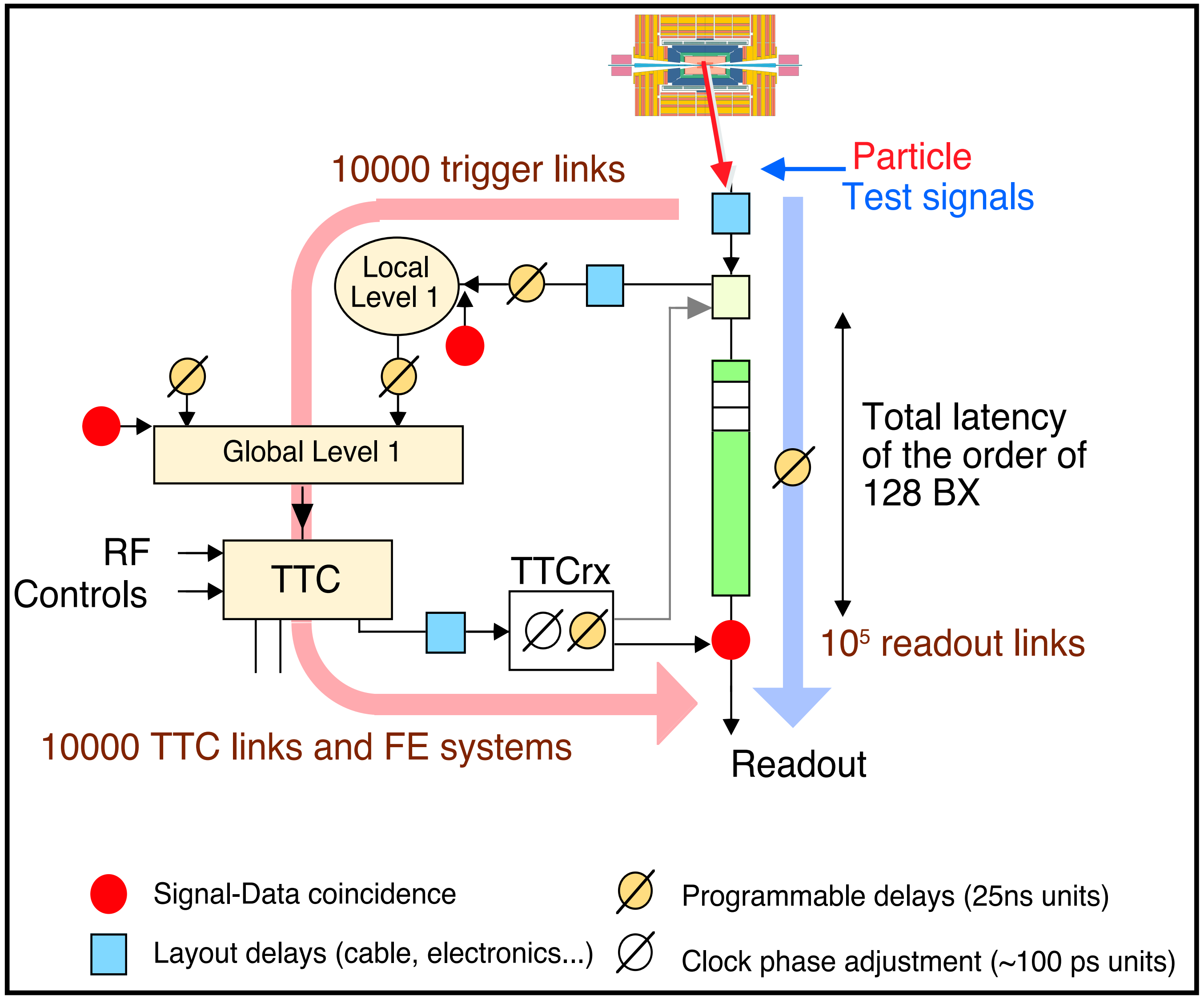
- ➡ **Allow trigger decision longer than clock tick (and no deadtime)**
  - ➡ Execute trigger selection in defined clocked steps (**fixed latency**)
  - ➡ Intermediate storage in stacked buffer cells
  - ➡ R/W pointers are moved by clock frequency

- ➡ **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



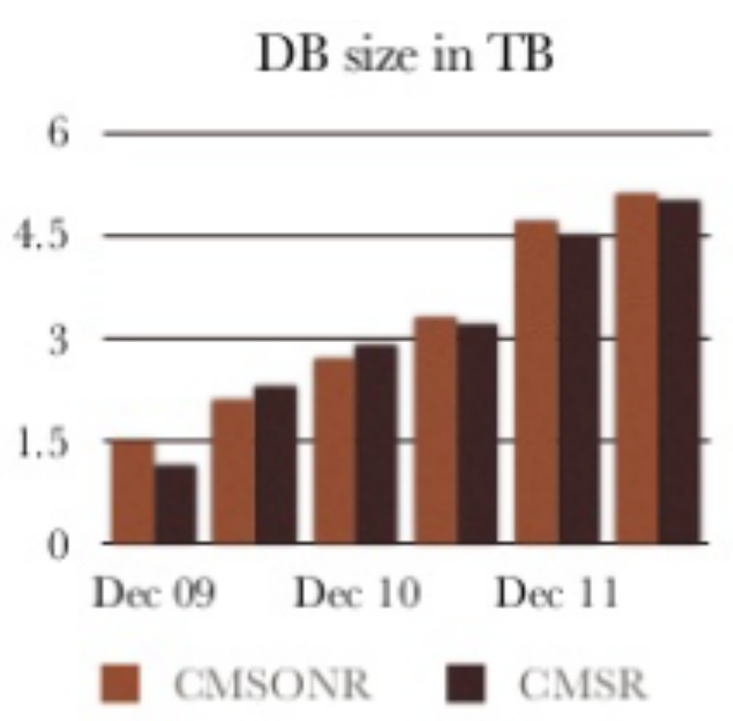


- **Common optical system: TTC**
  - radiation resistance
  - single high power laser
- **Large distribution**
  - experiments with ~10^7 channels

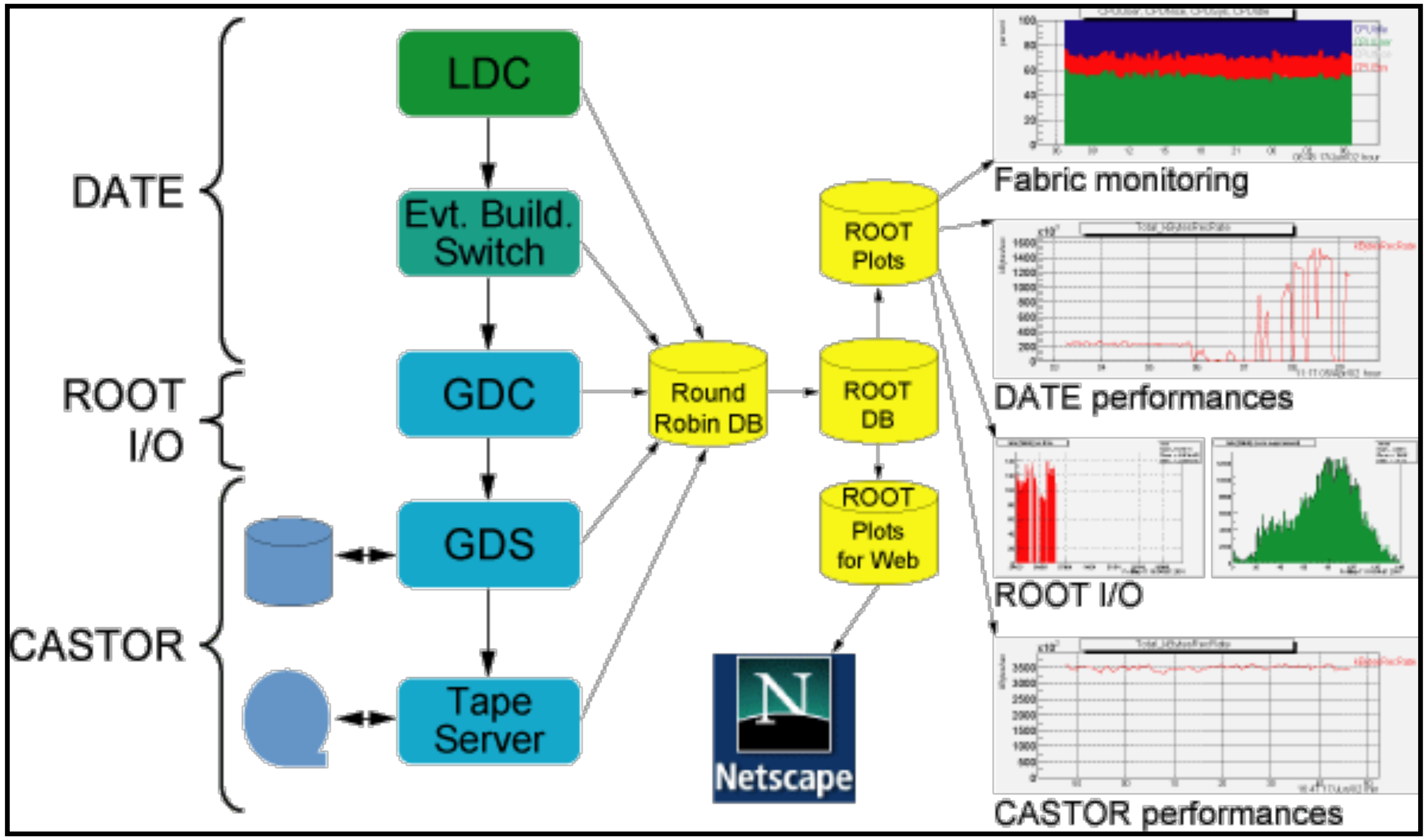
- **Align readout & trigger at (better than) 25ns and correct for**
  - time of flight (25 ns ≈ 7.5m)
  - cable delays (10cm/ns)
  - processing delays (~100 BCs)

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

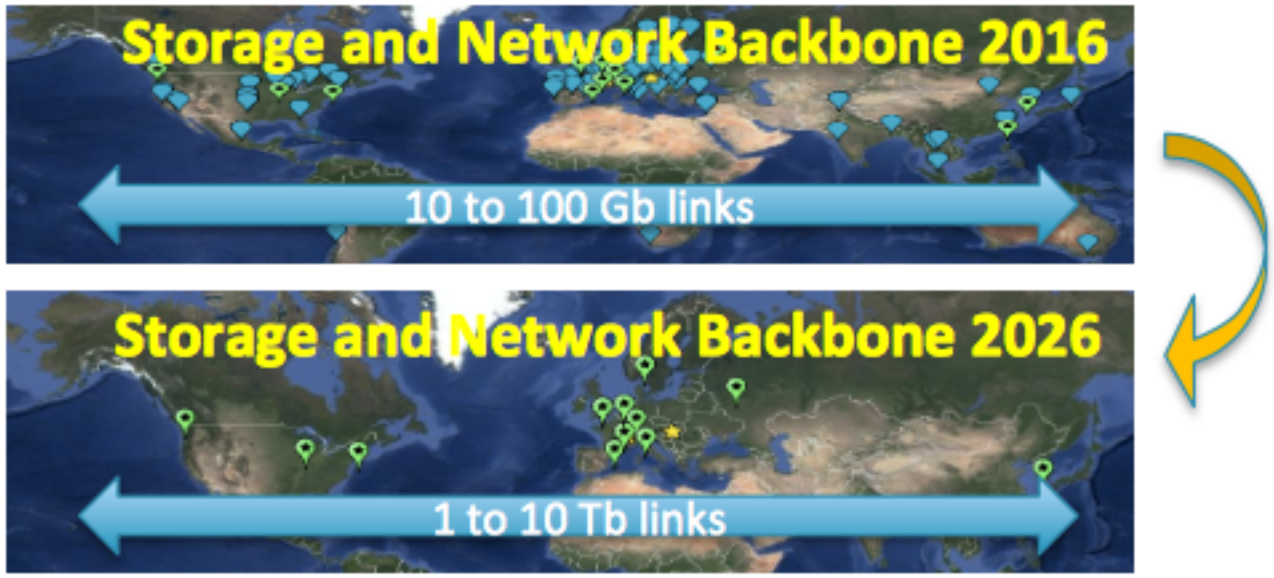


CMS DB grows about 1.5TB/year, condition data only a small fraction

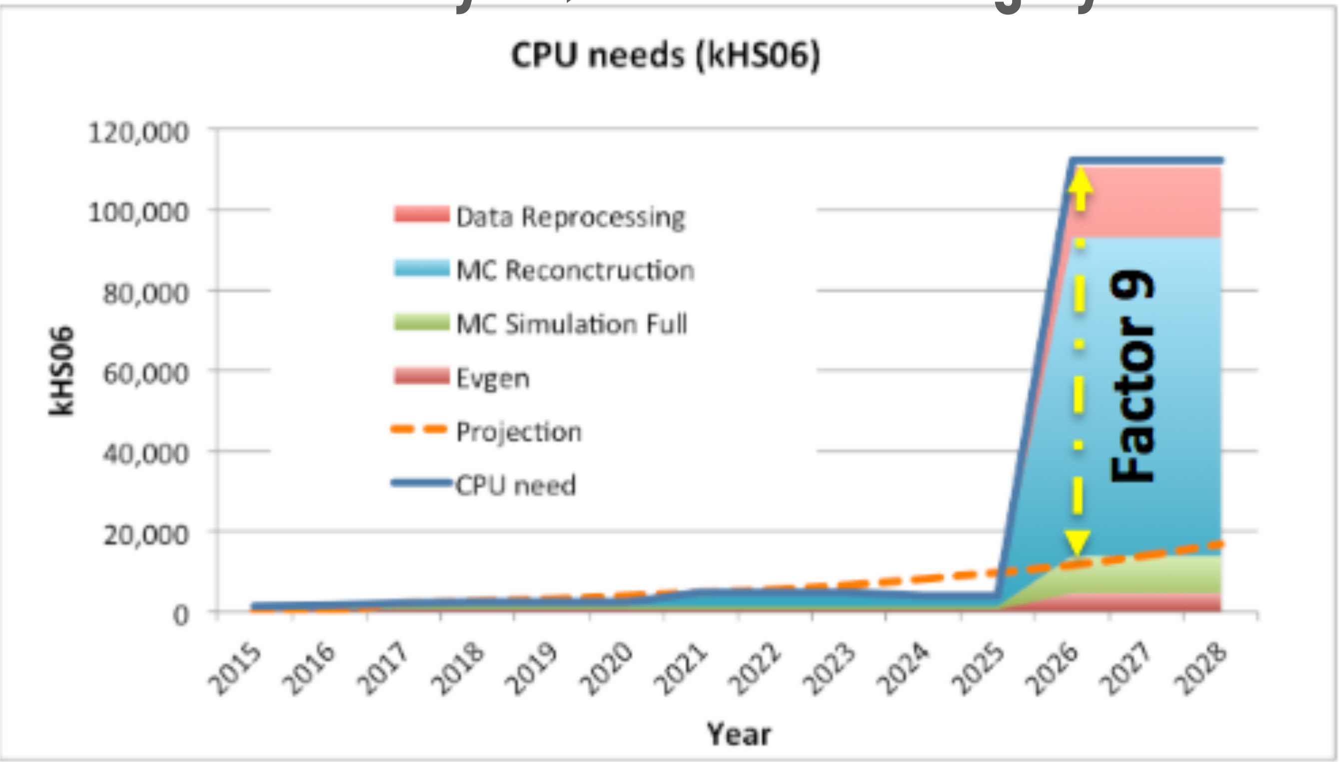


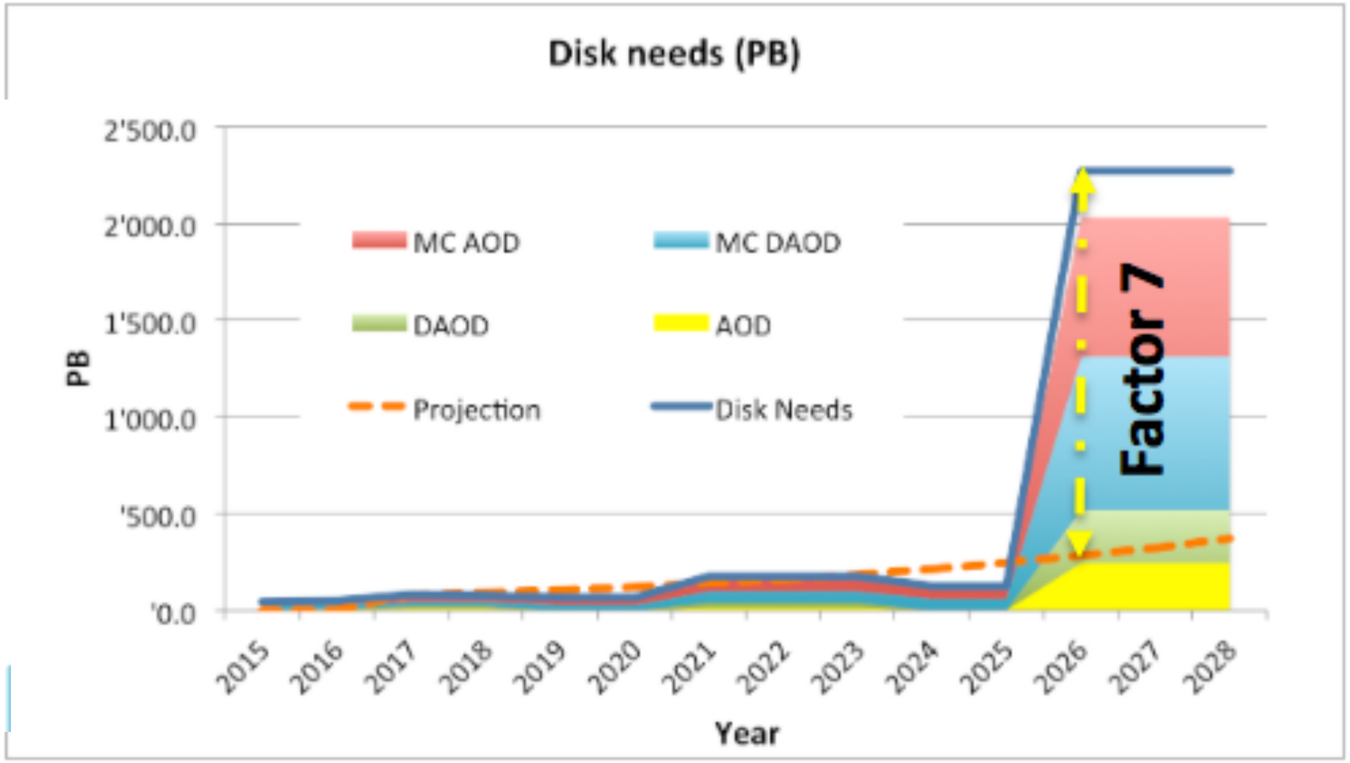
COMPUTING EVOLUTION FOR HL-LHC

- **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



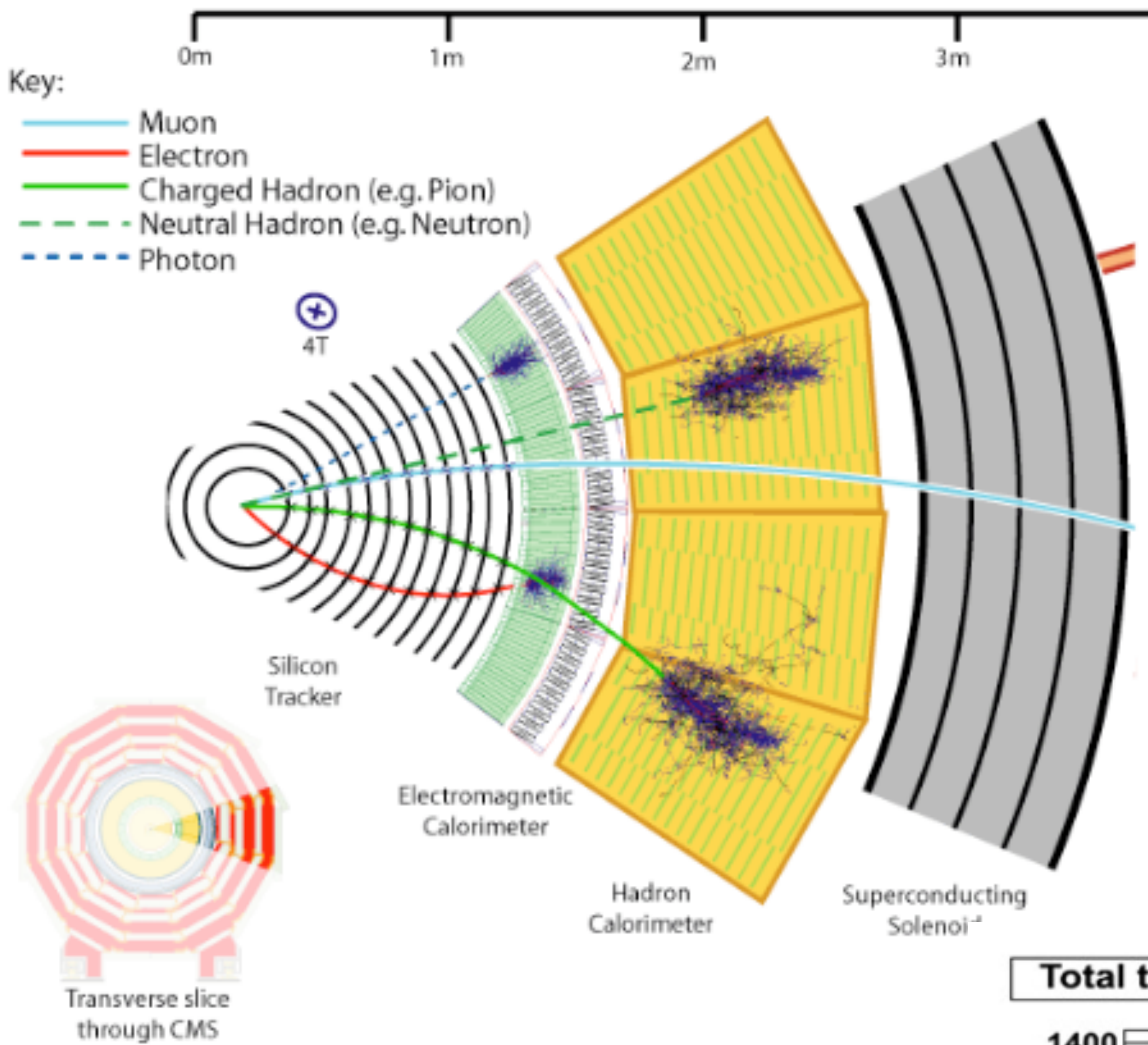
Projection of available resources in HL-LHC:
20% more CPU/year, 15% more storage/year





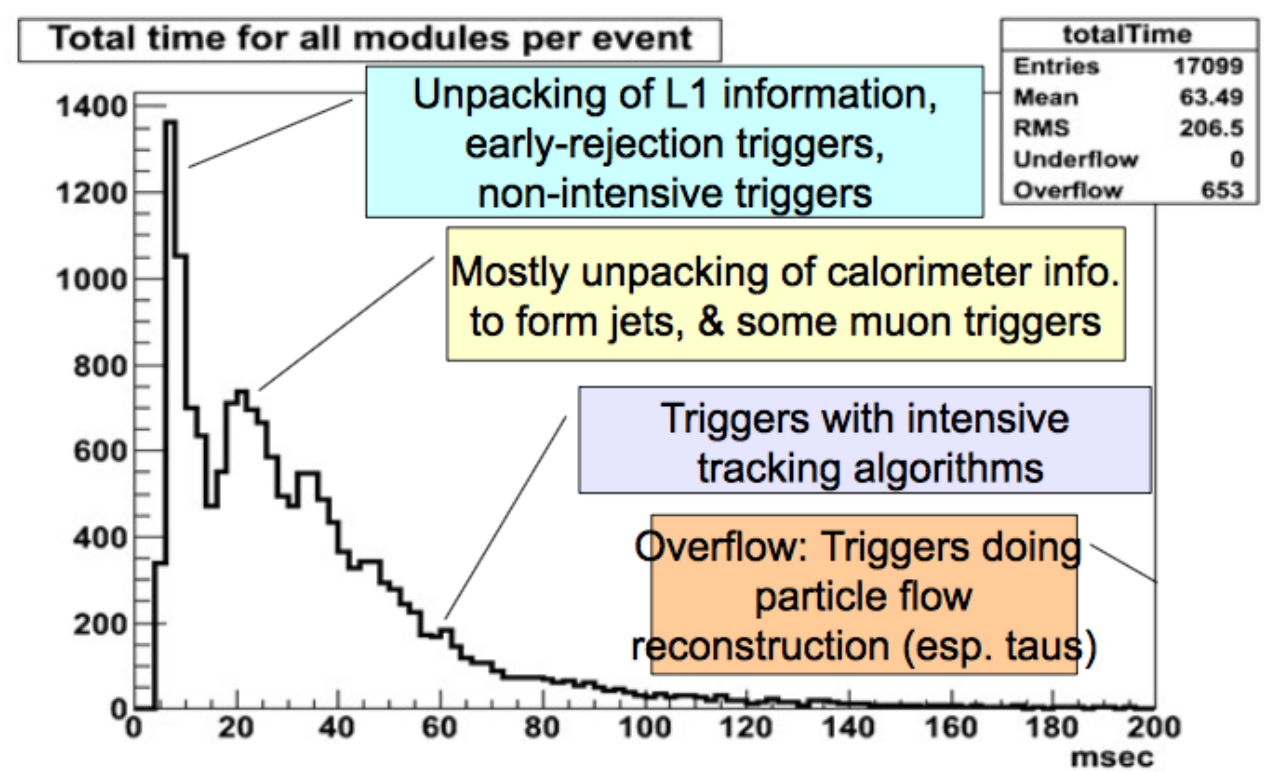
CALORIMETER TRIGGERS

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

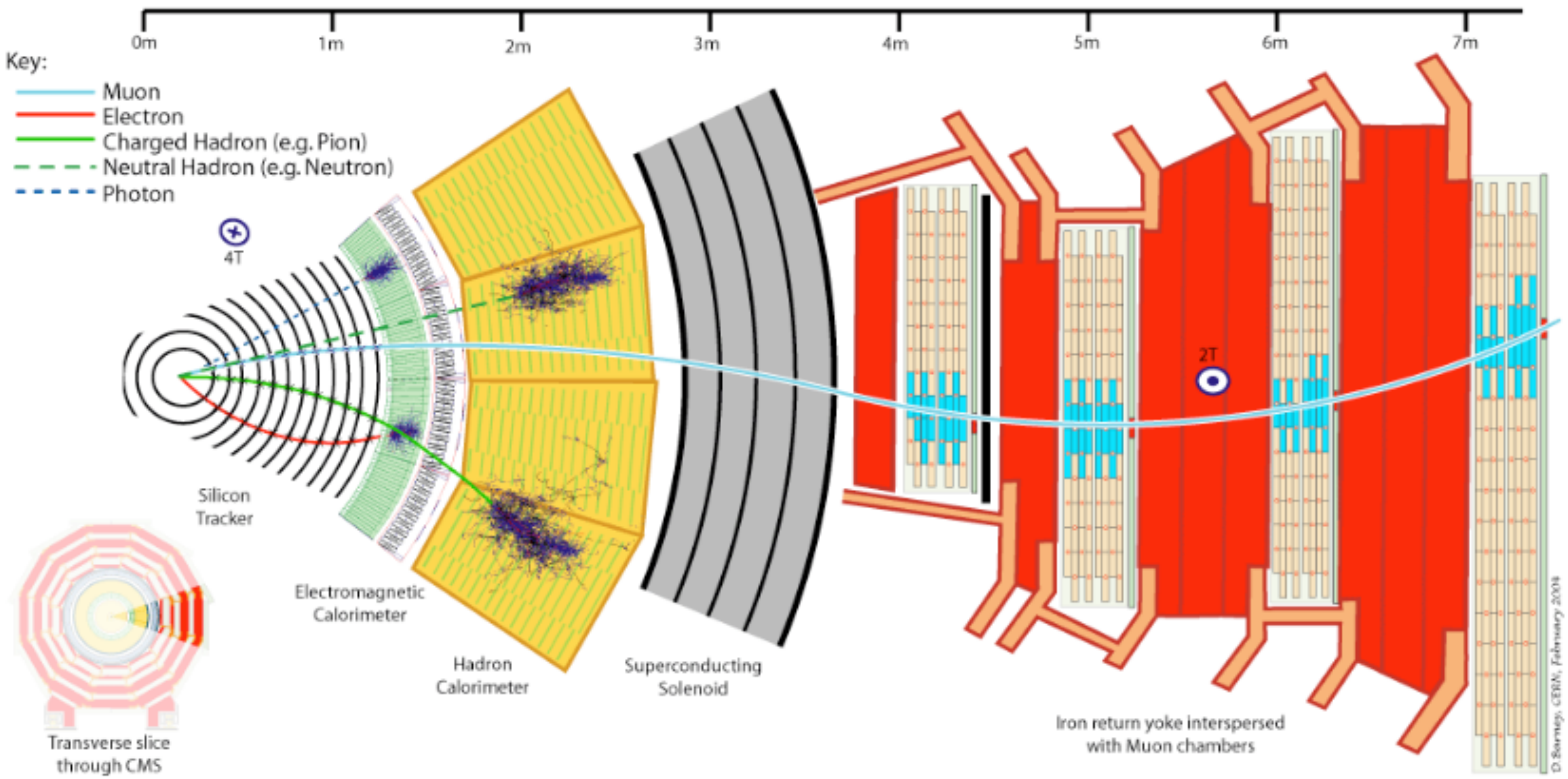


- ➡ **Fast and good resolution (LArg, PbW_4 for e-m)**
- ➡ **First-level processing (40MHz)**
  - ➡ "trigger towers" to reduce data (10-bit range)
  - ➡ sliding-window technique for local maxima
  - ➡ parallel algorithms for cluster shape and energy distribution

- ➡ **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



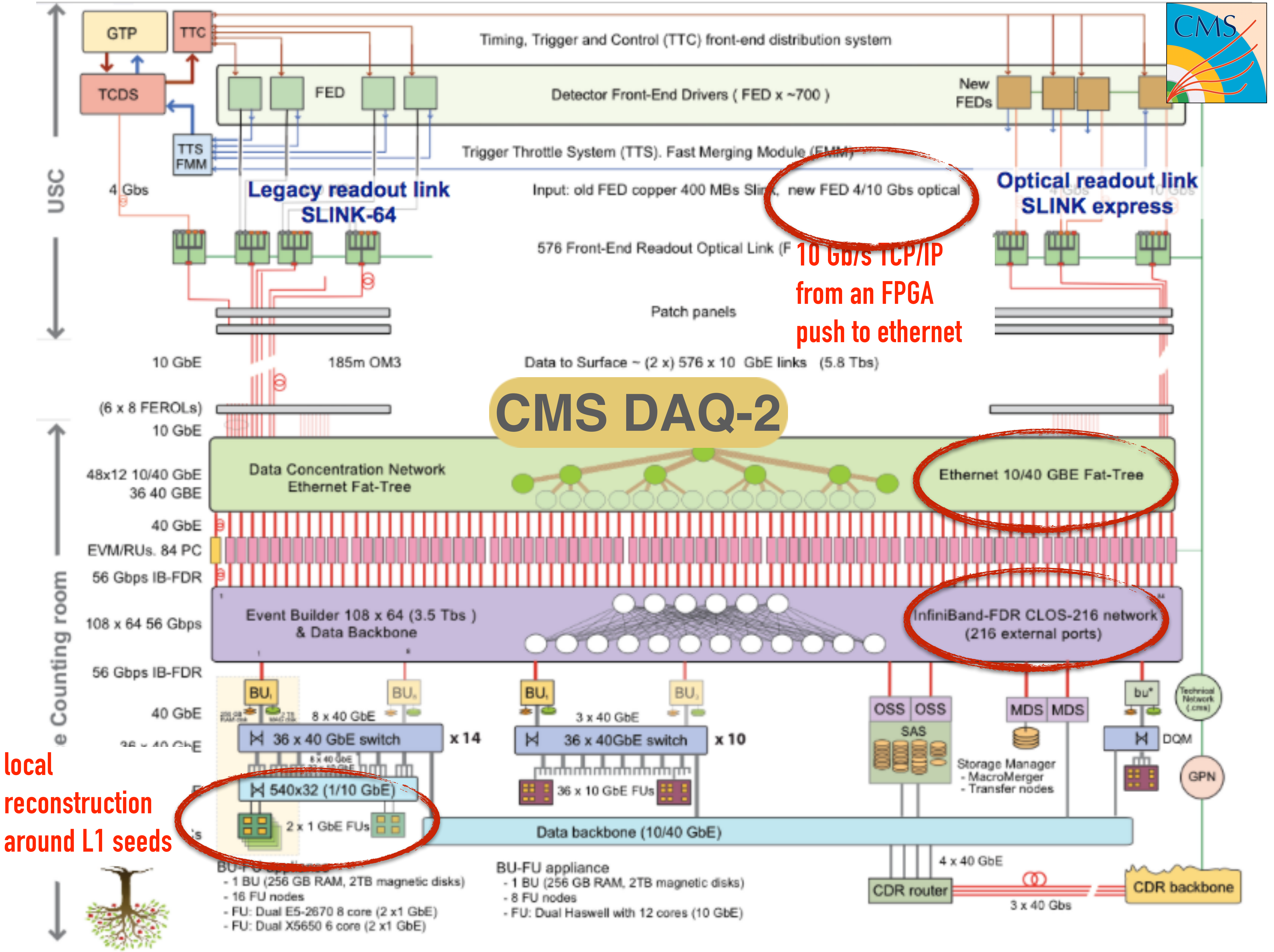
TRIGGERS FOR MUONS



- **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



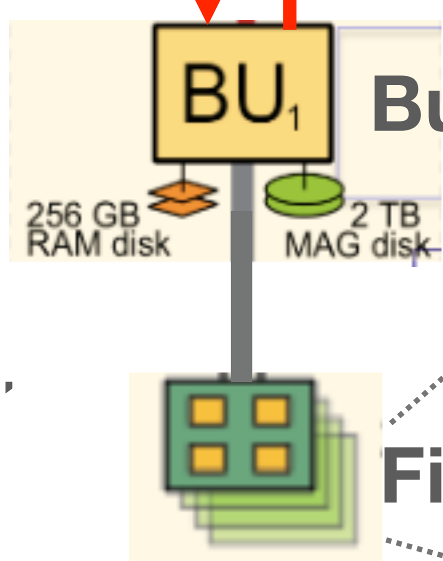
GTP
TTC
TCDS
Timing, Trigger and Control (TTC) front-end distribution system
FED
Detector Front-End Drivers (FED x ~700)
New FEDs
TTS FMM
Trigger Throttle System (TTS). Fast Merging Module (FMM)
USC
4 Gbs
Legacy readout link SLINK-64
Input: old FED copper 400 MBs Slink, new FED 4/10 Gbs optical
Optical readout link SLINK express
576 Front-End Readout Optical Link (F
10 Gb/s TCP/IP from an FPGA push to ethernet
Patch panels
10 GbE
185m OM3
Data to Surface ~ (2 x) 576 x 10 GbE links (5.8 Tbs)
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
40 GbE
Counting room
8 x 40 GbE
36 x 40 GbE switch
x 14
3 x 40 GbE
36 x 40GbE switch
x 10
OSS
OSS
SAS
MDS
MDS
DQM
GPN
540x32 (1/10 GbE)
2 x 1 GbE FUs
36 x 10 GbE FUs
Storage Manager
- MacroMerger
- Transfer nodes
Data backbone (10/40 GbE)
local reconstruction around L1 seeds
BU-FU appliance
- 1 BU (256 GB RAM, 2TB magnetic disks)
- 16 FU nodes
- FU: Dual E5-2670 8 core (2 x1 GbE)
- FU: Dual X5650 6 core (2 x1 GbE)
BU-FU appliance
- 1 BU (256 GB RAM, 2TB magnetic disks)
- 8 FU nodes
- FU: Dual Haswell with 12 cores (10 GbE)
4 x 40 GbE
CDR router
3 x 40 Gbs
CDR backbone

EVOLUTION OF THE FILTER FARM

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

Max 2kHz, 2.2–2.6 GB/s

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



Building Unit (BU)

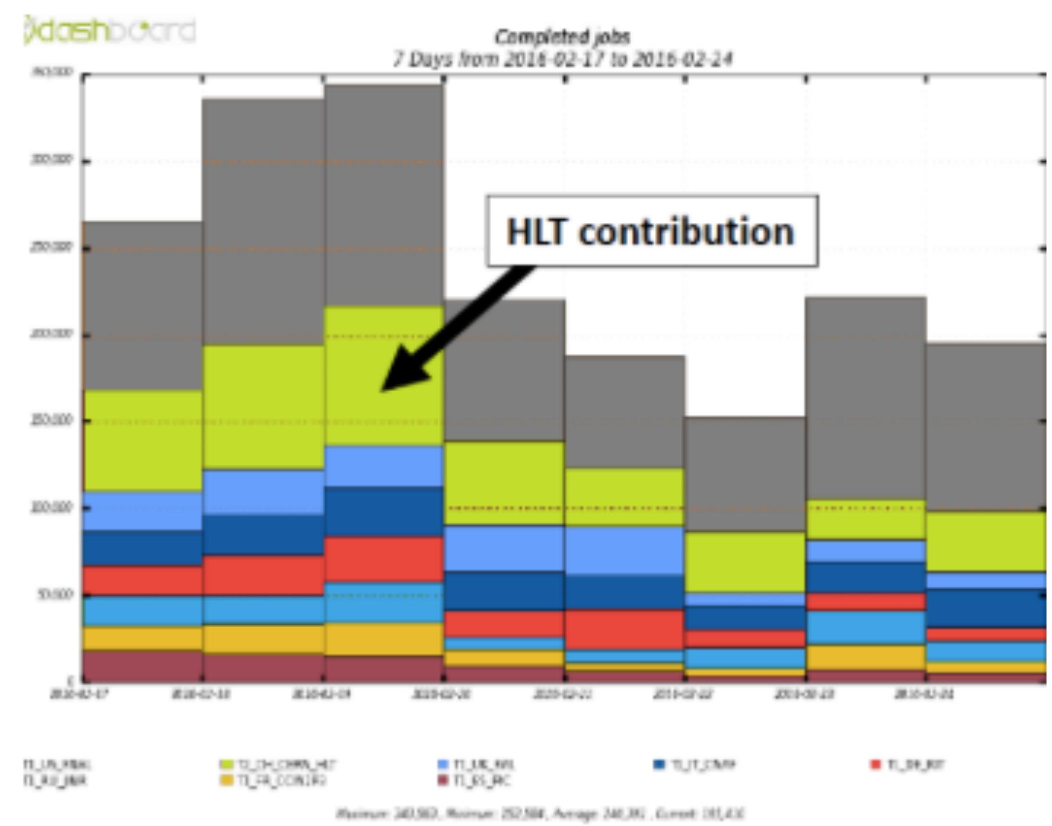
data, status, configuration, latency

Filter Unit (FU)

Every data file accompanied by a metadata in JSON files

Integrated Cloud capability (New!)

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



File-based communication

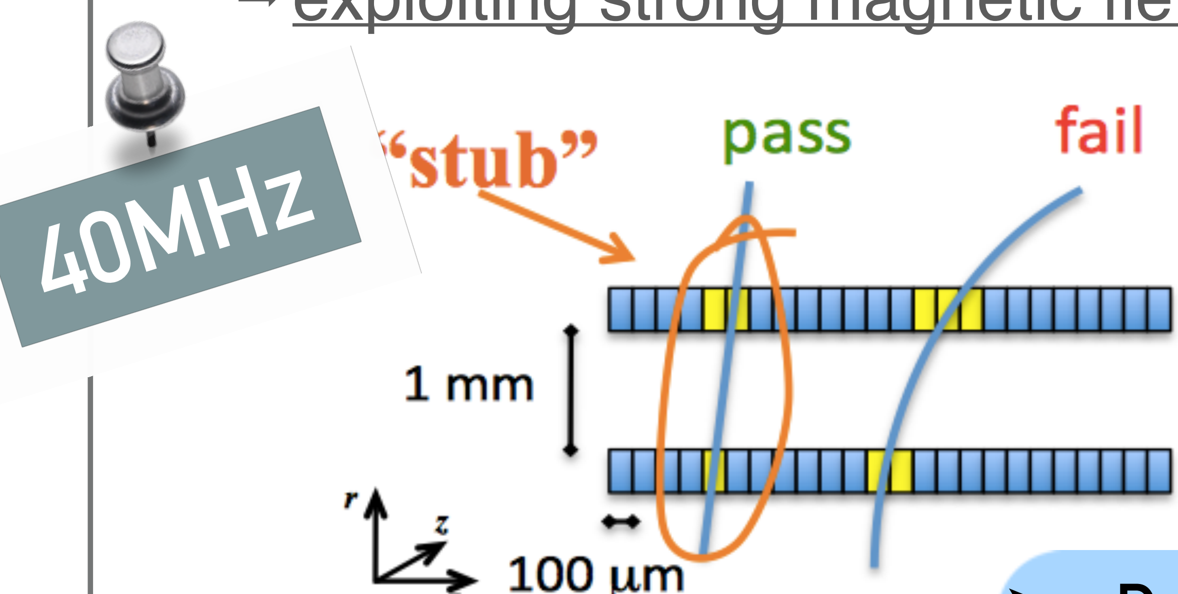
- ➡ HLT and DAQ completely decoupled
- ➡ Network filesystem used as transport (and resource arbitration) protocol (LUSTRE FS)

CMS: LOW-P_T TRACK FILTERING

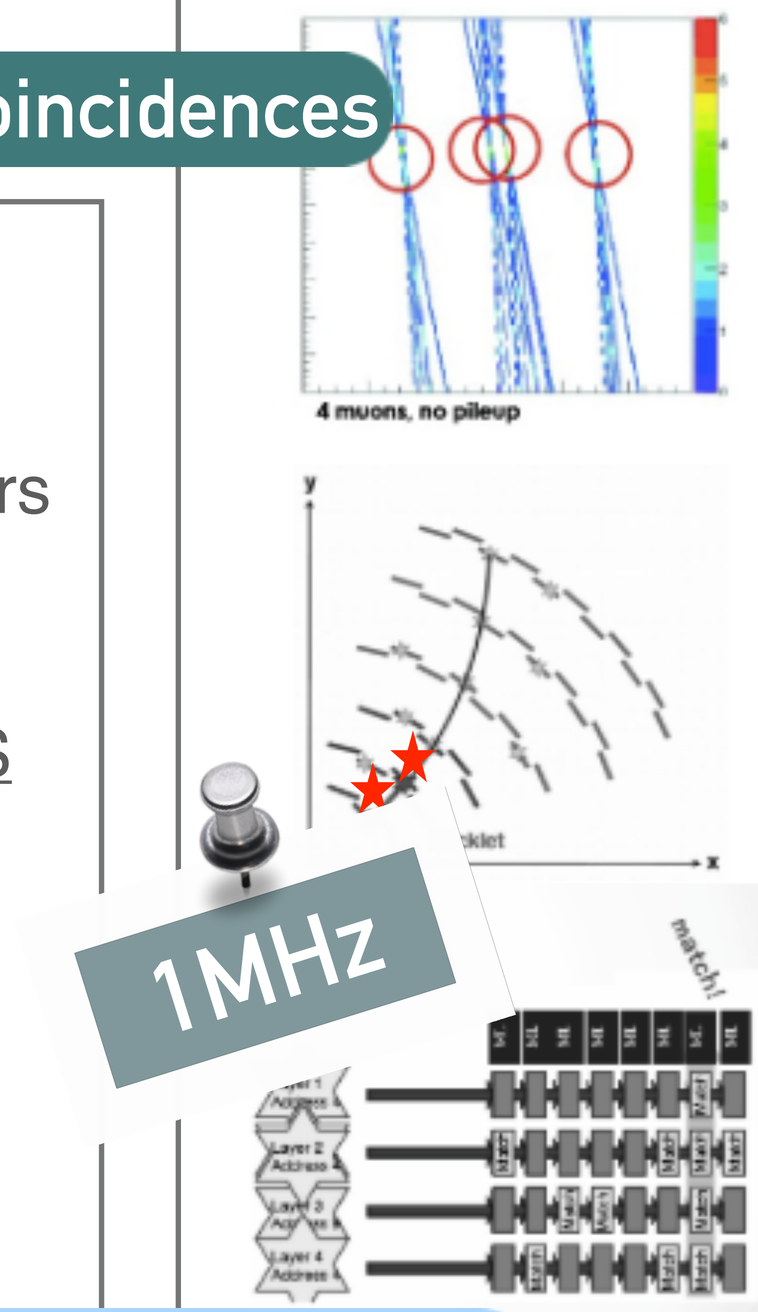
Track filtering (low p_T)

Reduce readout 40 ➟1MHz by detector coincidences

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



Track finding options



Hough Transform

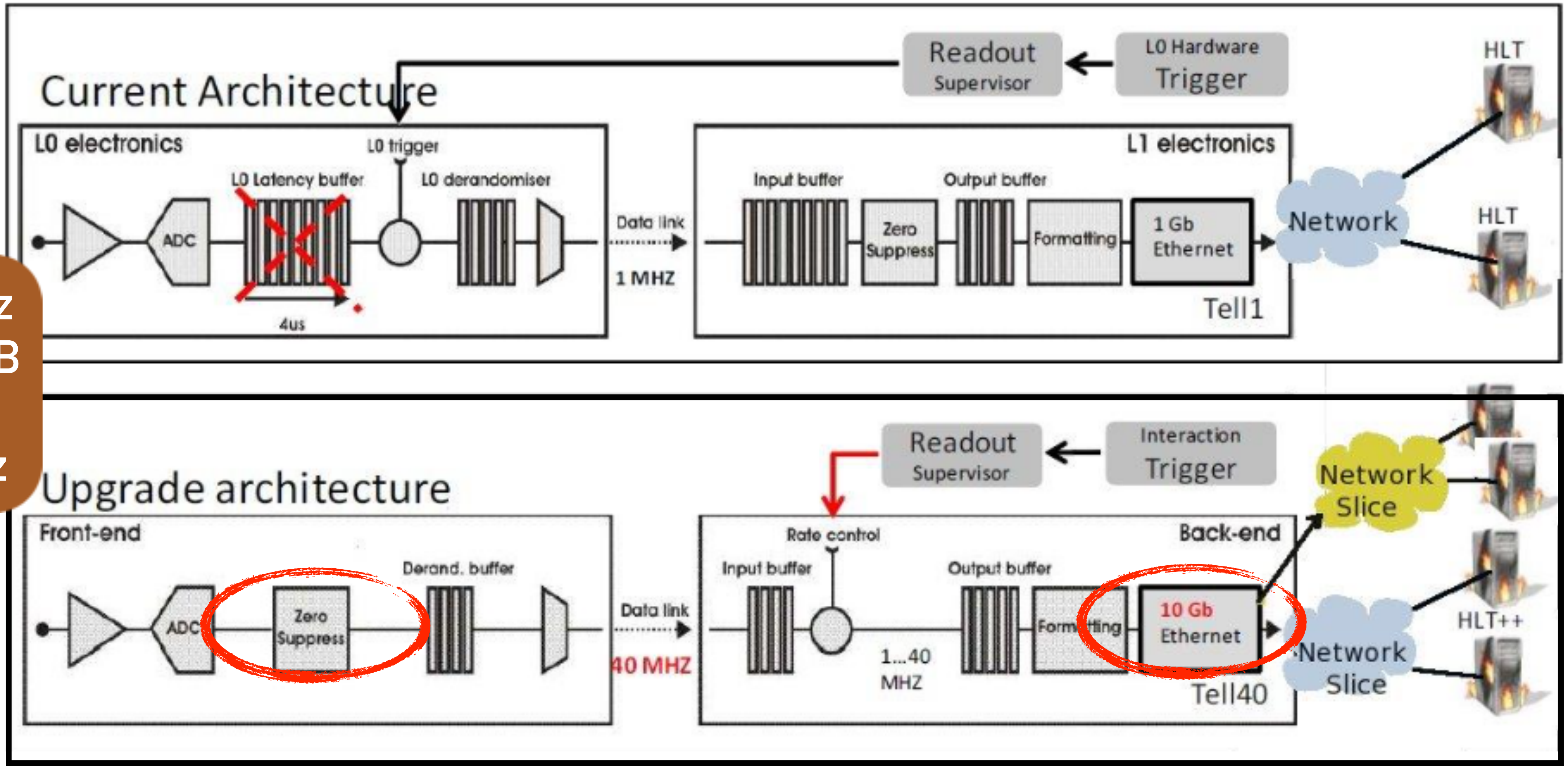
Tracklets

Associative Memories

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

HOW TO LIVE WELL WITHOUT A L1 TRIGGER

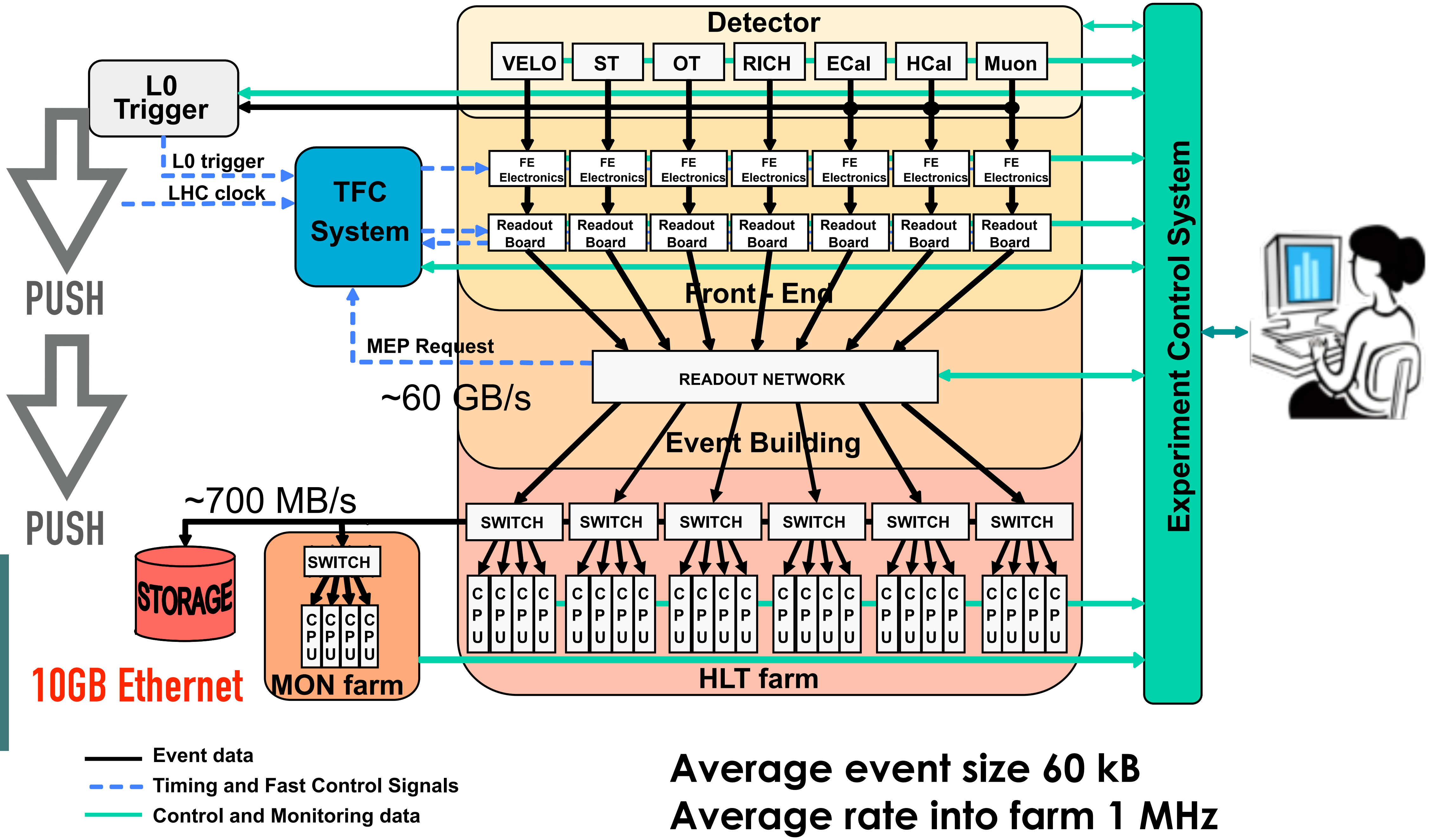
Readout: 40 MHz
Event size: 100kB
DAQ: 40 Tbit/s
Record: 100 kHz



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

TDAQ ARCHITECTURE IN RUN-2

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

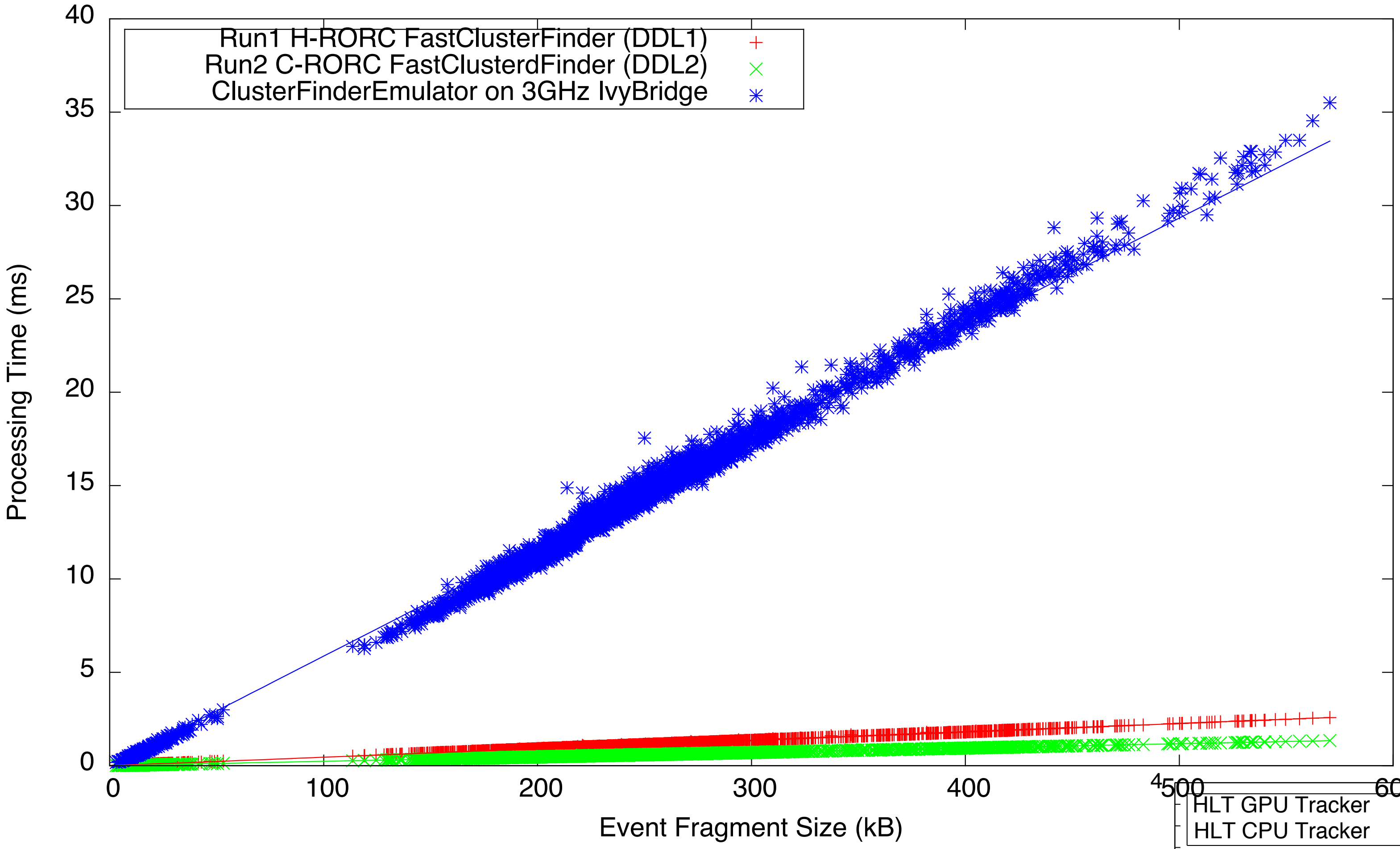


62 sub-farms, total 1780 nodes, with edge-routers (12 Gbps)

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

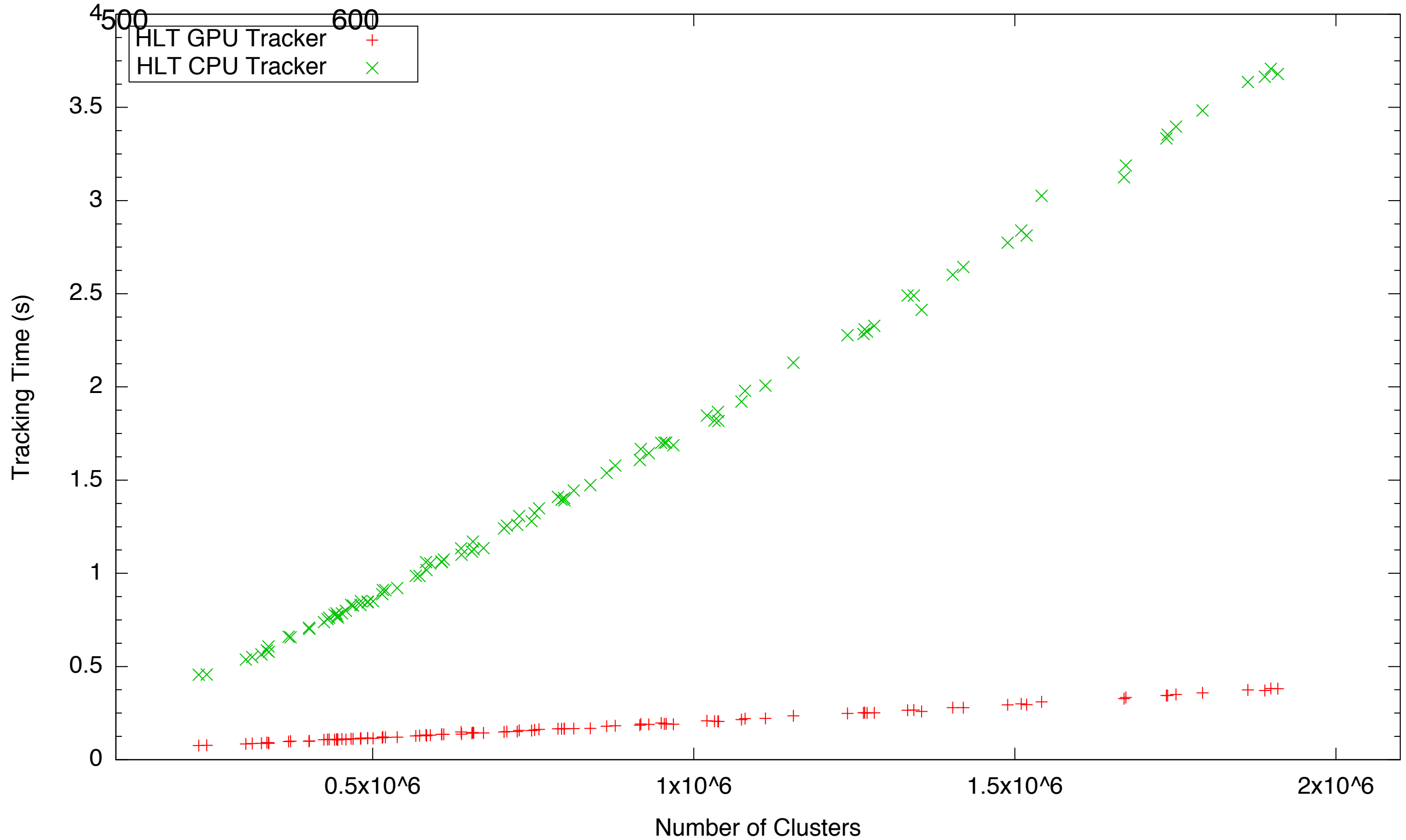
- **Small event, at high rate: ask for optimized transmission**
  - TTC system is used to assign IP addresses to RO boards
  - Ethernet UDP, with 10-15 events packed ⇒ ~ **80 kHz**

HARDWARE ACCELERATION WITH FPGAS AND GPUS

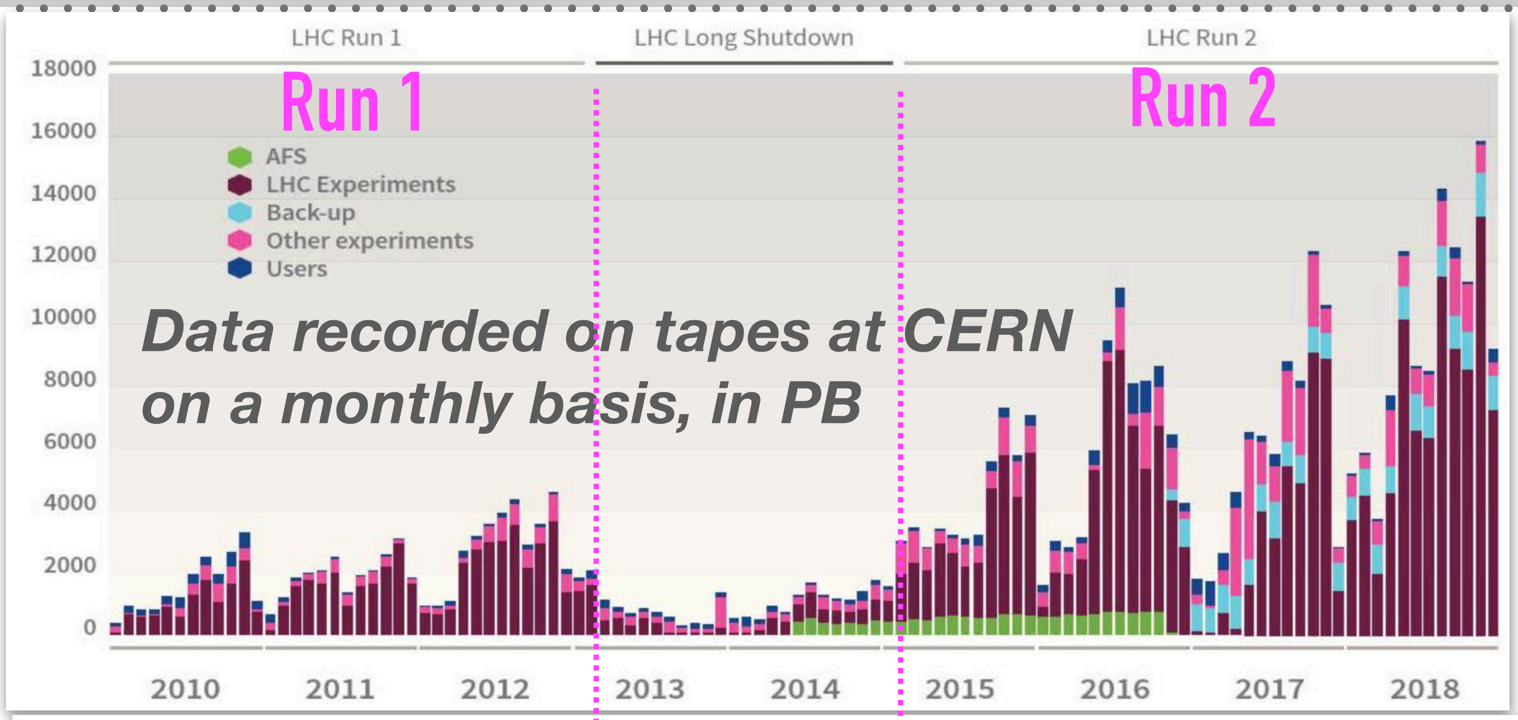


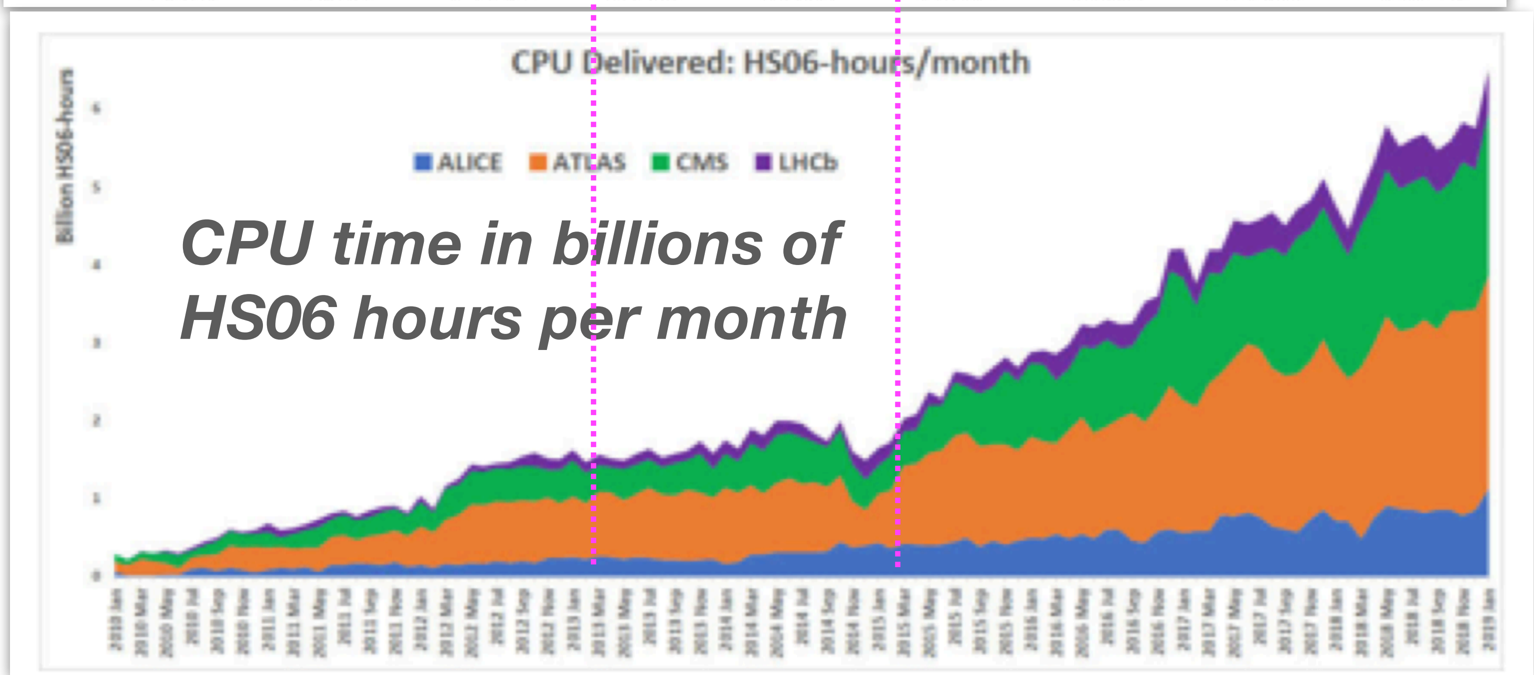
Performance of the FPGA-based FastClusterFinder algorithm for DDL1 (Run1) and DDL2 (Run2) compared to the software implementation on a recent server PC.

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



LHC COMPUTING TOWARDS NEW PARADIGMS





see [Ref]

Run1 + Run2

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

Run 3

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

Run 4

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

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

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