



Data-Flow: current and future data-transmission applications in the data-acquisition systems at the LHC

INFIERI 2015, Hamburg

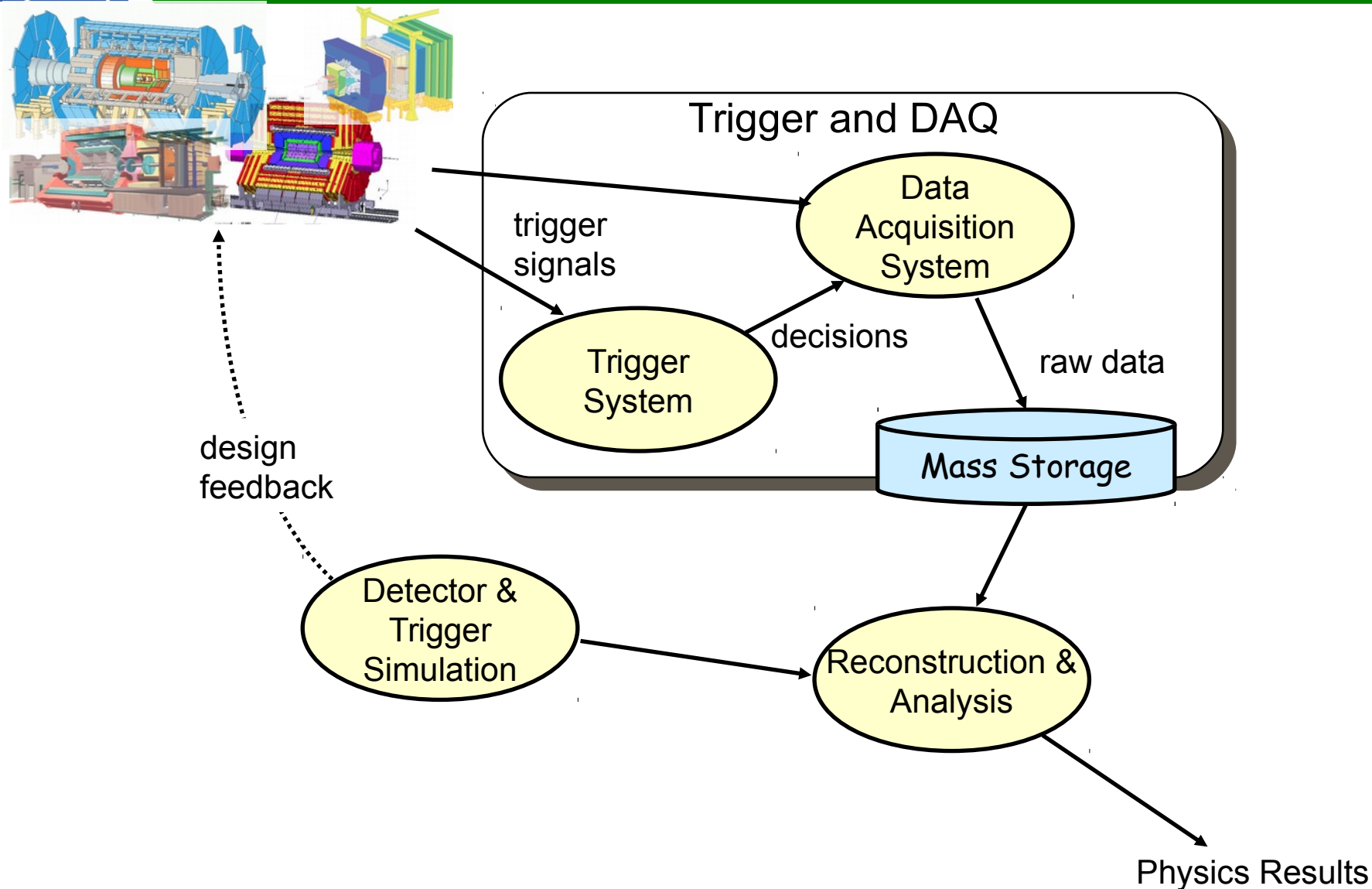
**W.Vandelli CERN/PH-ADT
Wainer.Vandelli@cern.ch**

- ➔ Introduction
- ➔ Basis Data-Acquisition (DAQ) principles
 - efficiency and dead-time
- ➔ Scaling it up
 - architecture and data-flow
- ➔ DAQ at the LHC
 - challenges and designs
- ➔ Event Building and Networking
 - DAQ-specific workloads and technology limits
- ➔ Coping with LHC upgrade programme

Introduction

- ➔ Data-flow is a sub-system of large data-acquisition systems
 - connects and safely transport data between the other subsystems
- ➔ Will need to understand DAQ before we can discuss the data movements
- ➔ Data acquisition is **not an exact science**. It is an alchemy of electronics, computer science, networking and physics
 - funding and manpower matter as well
 - there is not ONE solution
 - each experiment has it own peculiarities and legacies
 - experience, risk perception, technology expectations, ...

General Overview



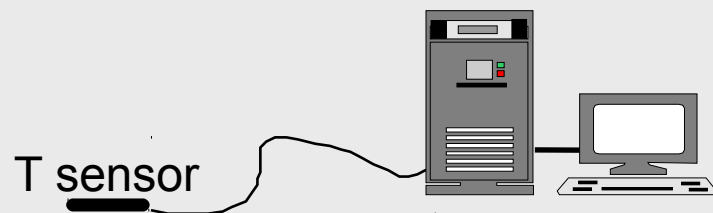
➔ Overall the main role of T & DAQ is to process the signals generated in a detector, storing the interesting information on a permanent storage



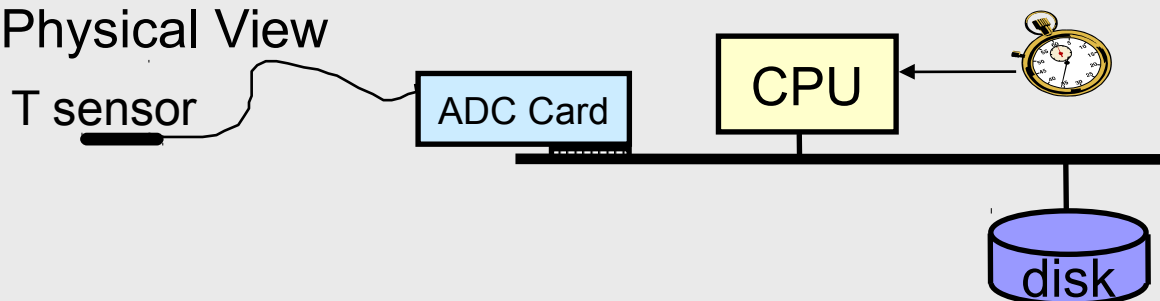
DAQ & Trigger

Basic DAQ: periodic trigger

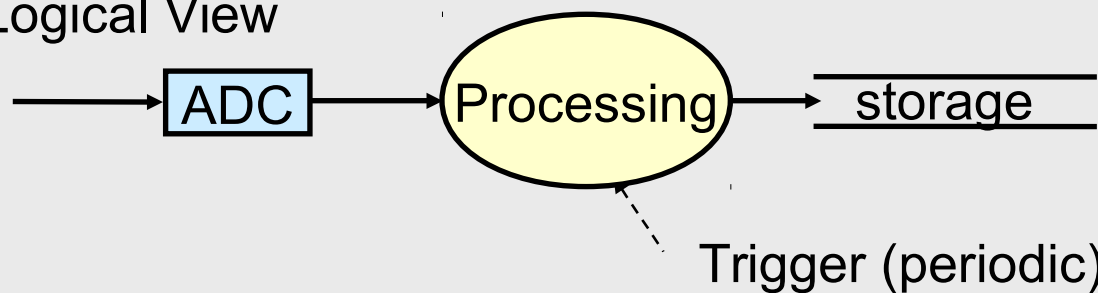
External View



Physical View



Logical View



→ Measure temperature at a fixed frequency

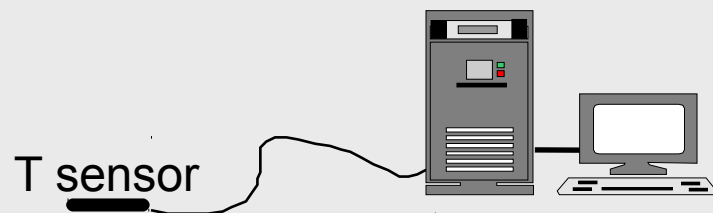
→ ADC performs analog to digital conversion

– our front-end electronics

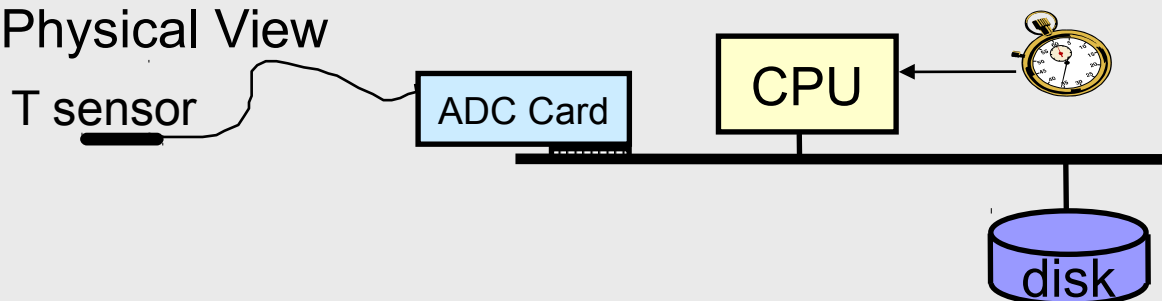
→ CPU does readout and processing

Basic DAQ: periodic trigger

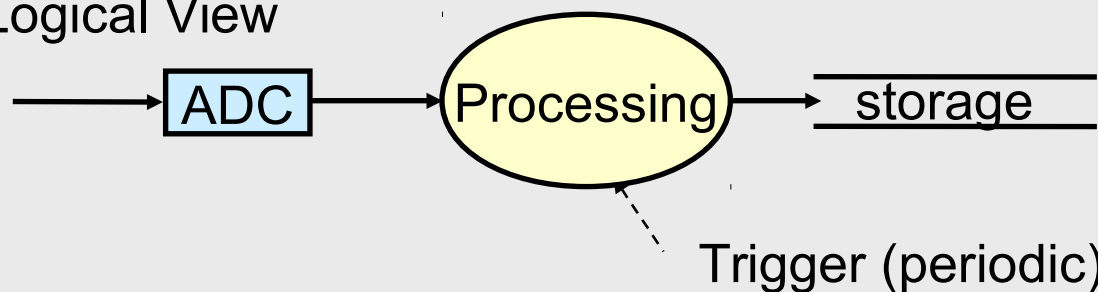
External View



Physical View



Logical View



→ Measure temperature at a fixed frequency

→ The system is clearly limited by the time to process an "event"

→ Example $\tau=1\text{ ms}$ to

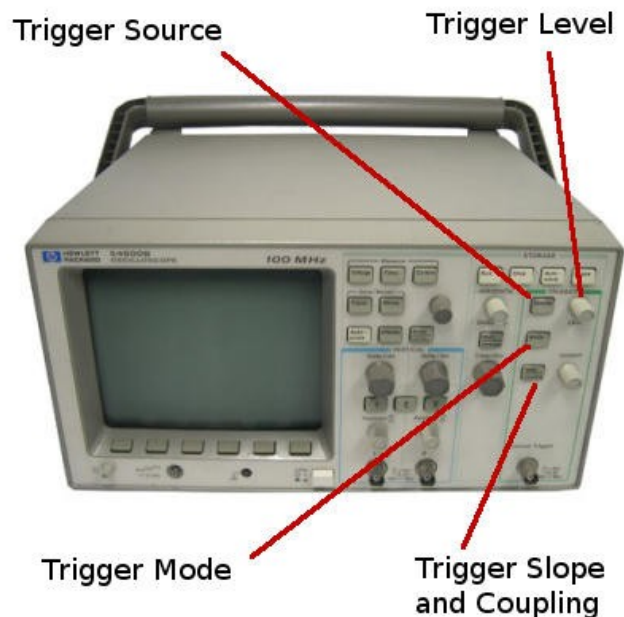
- ADC conversion
- +CPU processing
- +Storage

→ Sustain $\sim 1/1\text{ ms}=1\text{ kHz}$ ***periodic trigger*** rate

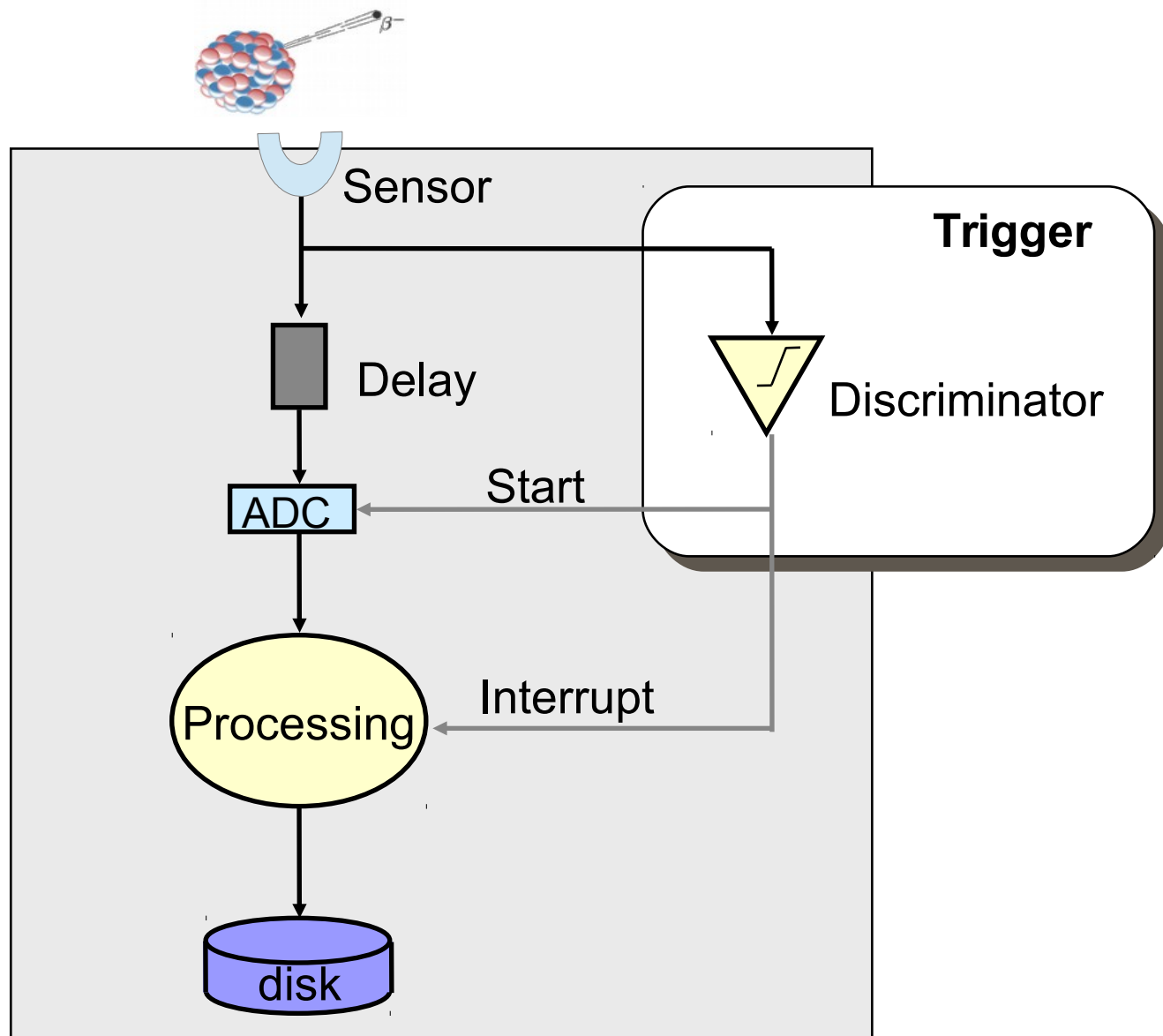
What is a “trigger”?

- ➔ A “trigger” is a system that rapidly decides, based on “simple” criteria, if an interesting event took place, initiating the data-acquisition process
- ➔ Simple, rapid, selective are the trigger keywords
- ➔ Relative parameters that depend on the operating conditions
 - in a multi-level trigger system the last level is normally way slower and more complex than the first one

The oscilloscope trigger does exactly this.
Informs the instrument to initiate the internal signal acquisition and visualization



Basic DAQ: physics trigger



→ Measure β decay properties

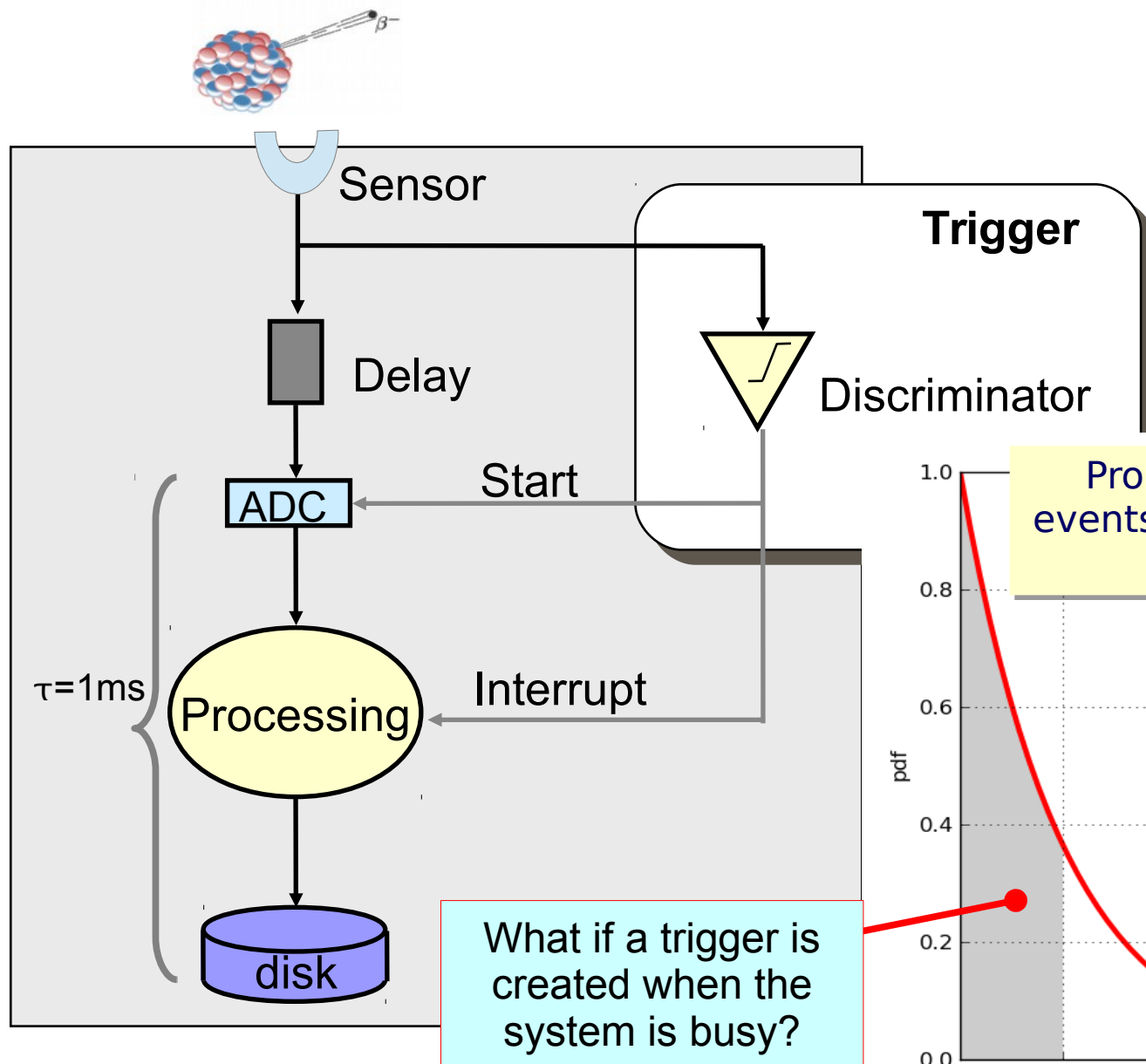
→ Events are asynchronous and unpredictable

- need a **physics** trigger

→ Delay compensates for the **trigger latency**

- time needed to reach a decision

Basic DAQ: real trigger

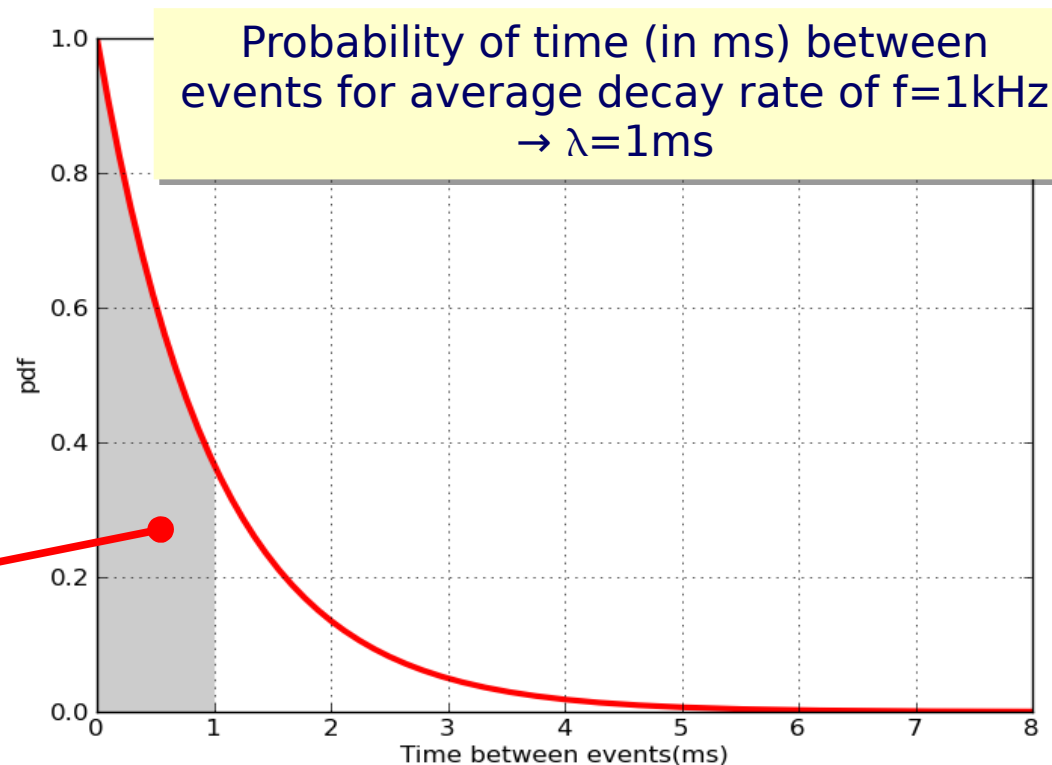


→ Measure β decay properties

- need a **physics** trigger

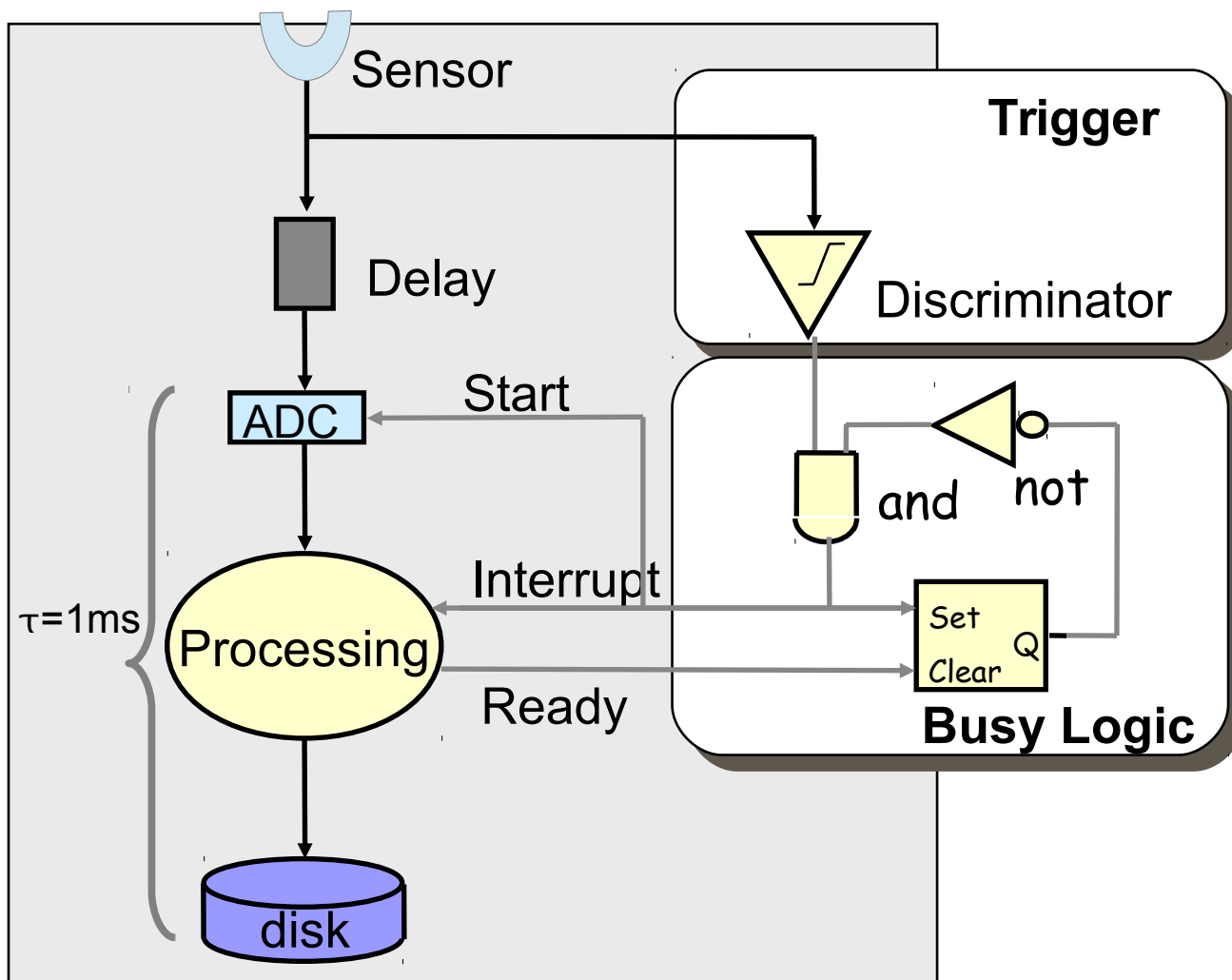
→ Stochastic process

- fluctuations



Basic DAQ: real trigger & busy logic


 $f=1\text{kHz}$
 $1/f=\lambda=1\text{ms}$



→ Busy logic avoids triggers while processing

→ Which (average) DAQ rate can we achieve now?

- reminder: $\tau=1\text{ms}$ was sufficient to run at 1kHz with a clock trigger

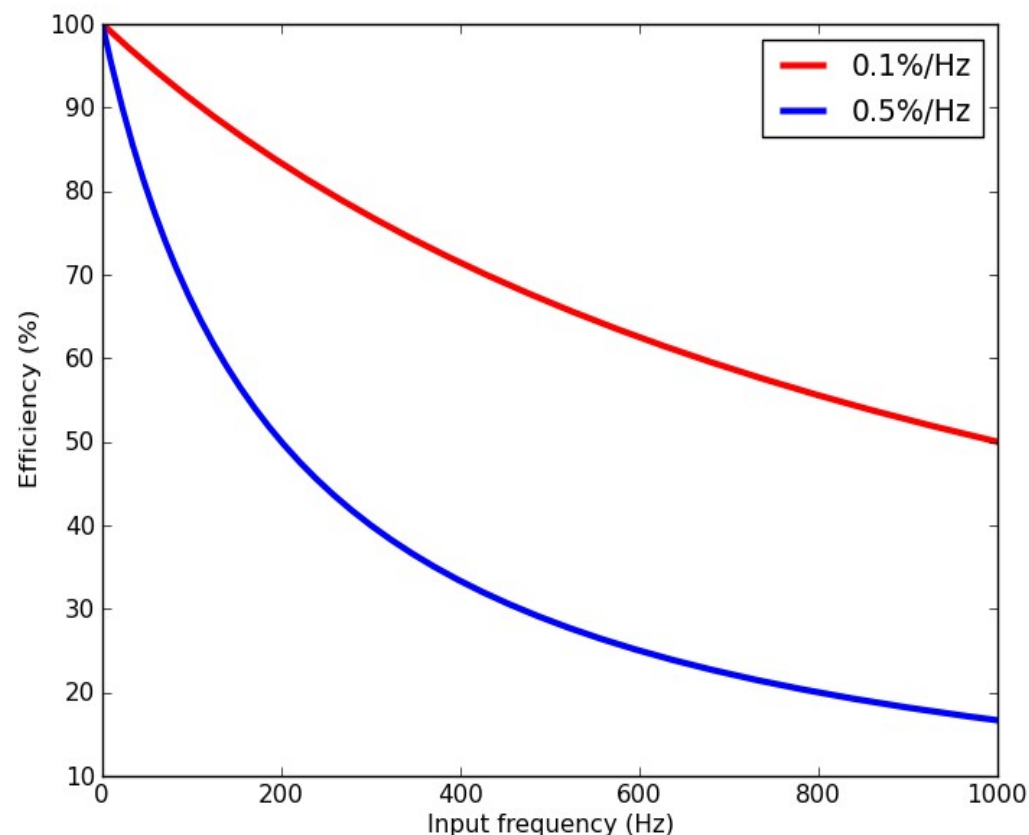
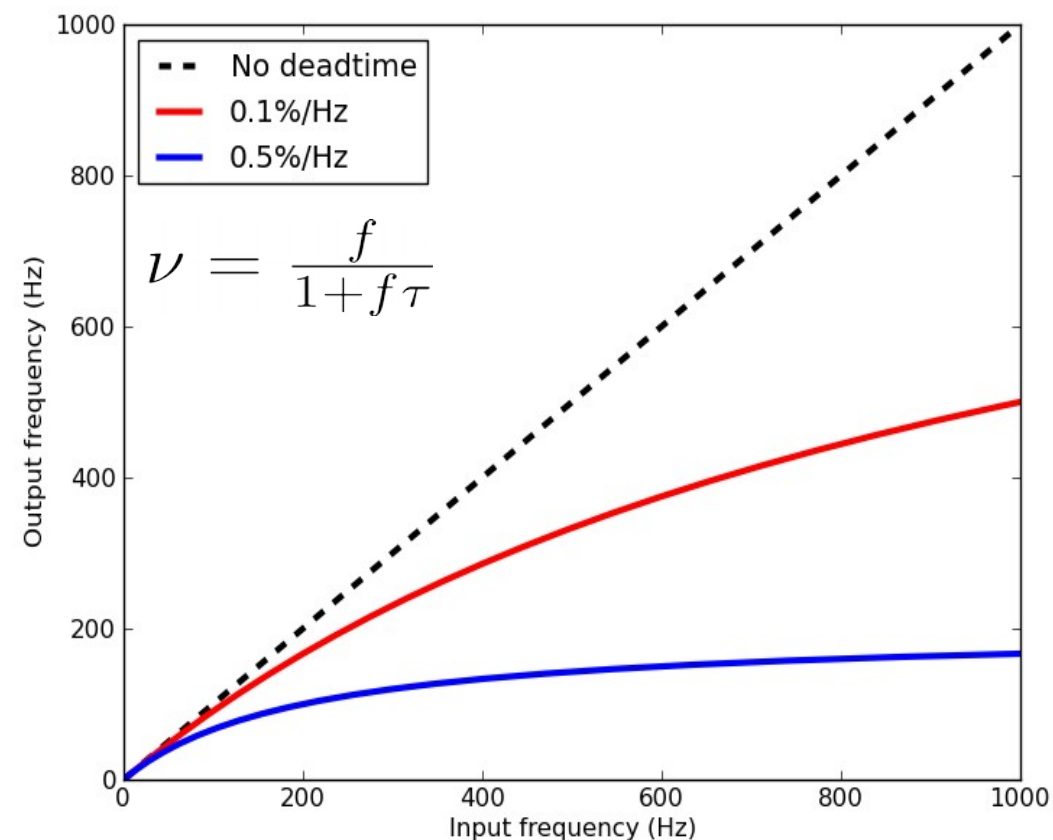
Define ν as average DAQ frequency

$\nu\tau \rightarrow$ DAQ system is busy - $(1 - \nu\tau) \rightarrow$ DAQ system is free

$$f(1 - \nu\tau) = \nu \rightarrow \nu = \frac{f}{1 + f\tau} < f$$

$$\epsilon = \frac{N_{saved}}{N_{tot}} = \frac{1}{1 + f\tau} < 100\%$$

- ➔ Define DAQ deadtime (d) as the time the system requires to process an event, without being able to handle other triggers. In our example $d=0.1\%/Hz$
- ➔ Due to the fluctuations introduced by the stochastic process the efficiency will always be less 100%
 - in our specific example, $d=0.1\%/Hz$, $f=1kHz \rightarrow \nu=500Hz$, $\epsilon=50\%$



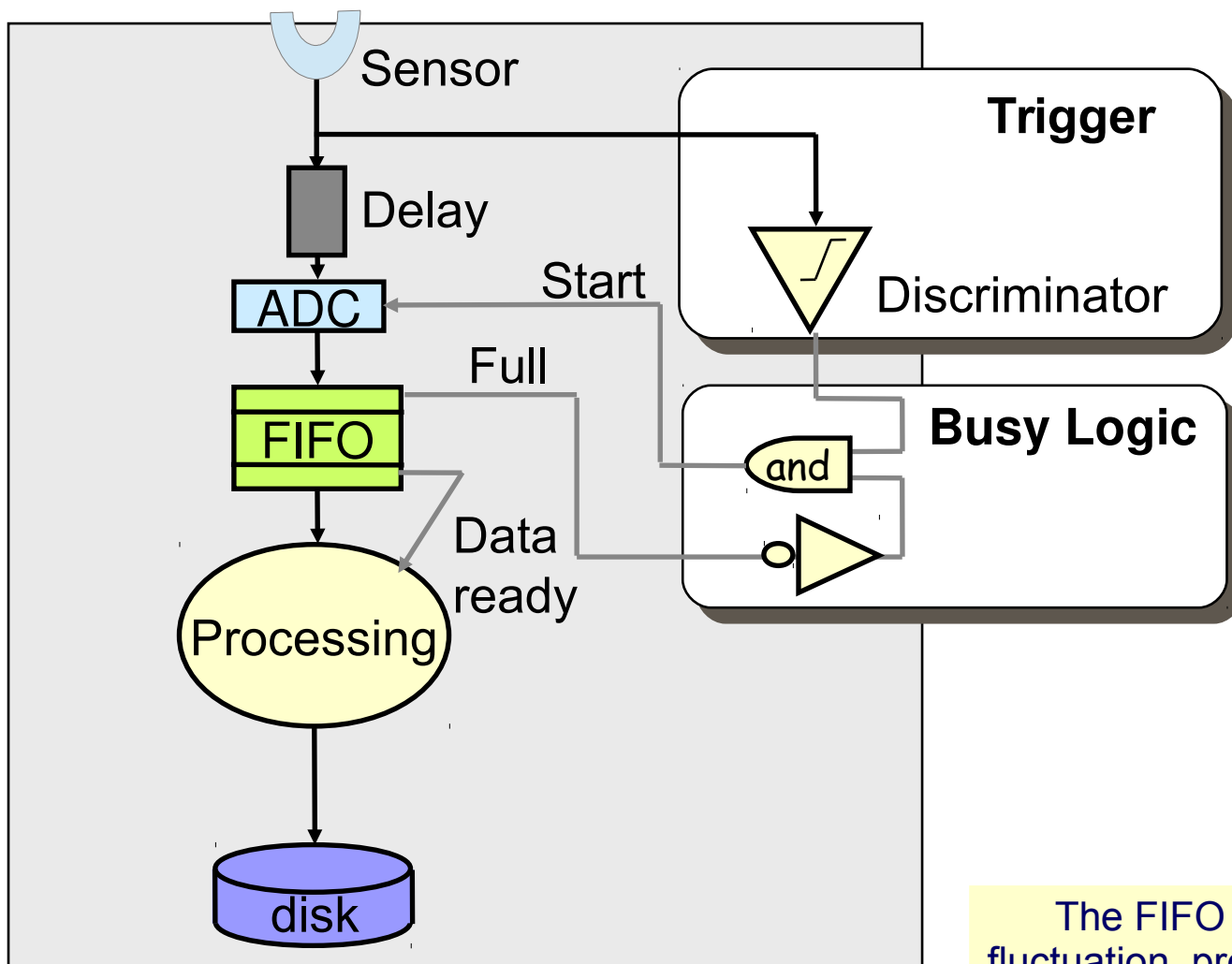
➔ If we want to obtain $\nu \sim f$ ($\epsilon \sim 100\%$) $\rightarrow f\tau \ll 1 \rightarrow \tau \ll \lambda$

- $f = 1\text{kHz}$, $\epsilon = 99\% \rightarrow \tau < 0.1\text{ms} \rightarrow 1/\tau > 10\text{kHz}$

➔ In order to cope with the input signal fluctuations, we have to over-design our DAQ system by a factor 10. This is very inconvenient! Can we mitigate this effect?

Basic DAQ: De-randomization

$f=1\text{kHz}$
 $1/f=\lambda=1\text{ms}$



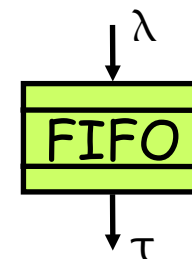
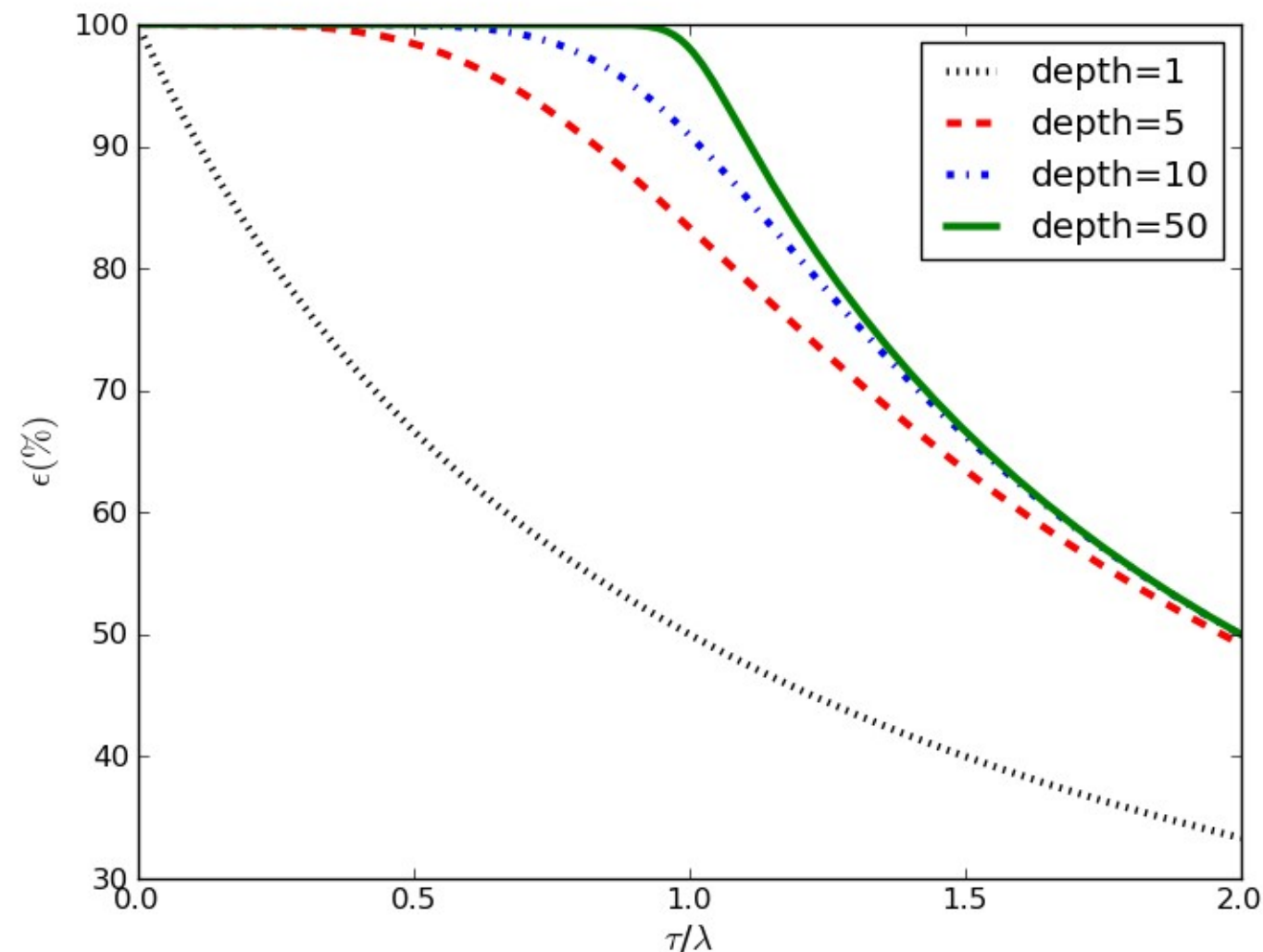
→ First-In First-Out

- buffer area organized as a queue
- depth: number of cells
- implemented in HW and SW



→ FIFO introduces an additional latency on the data path

The FIFO absorbs and smooths the input fluctuation, providing a ~steady (De-randomized) output rate

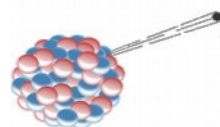


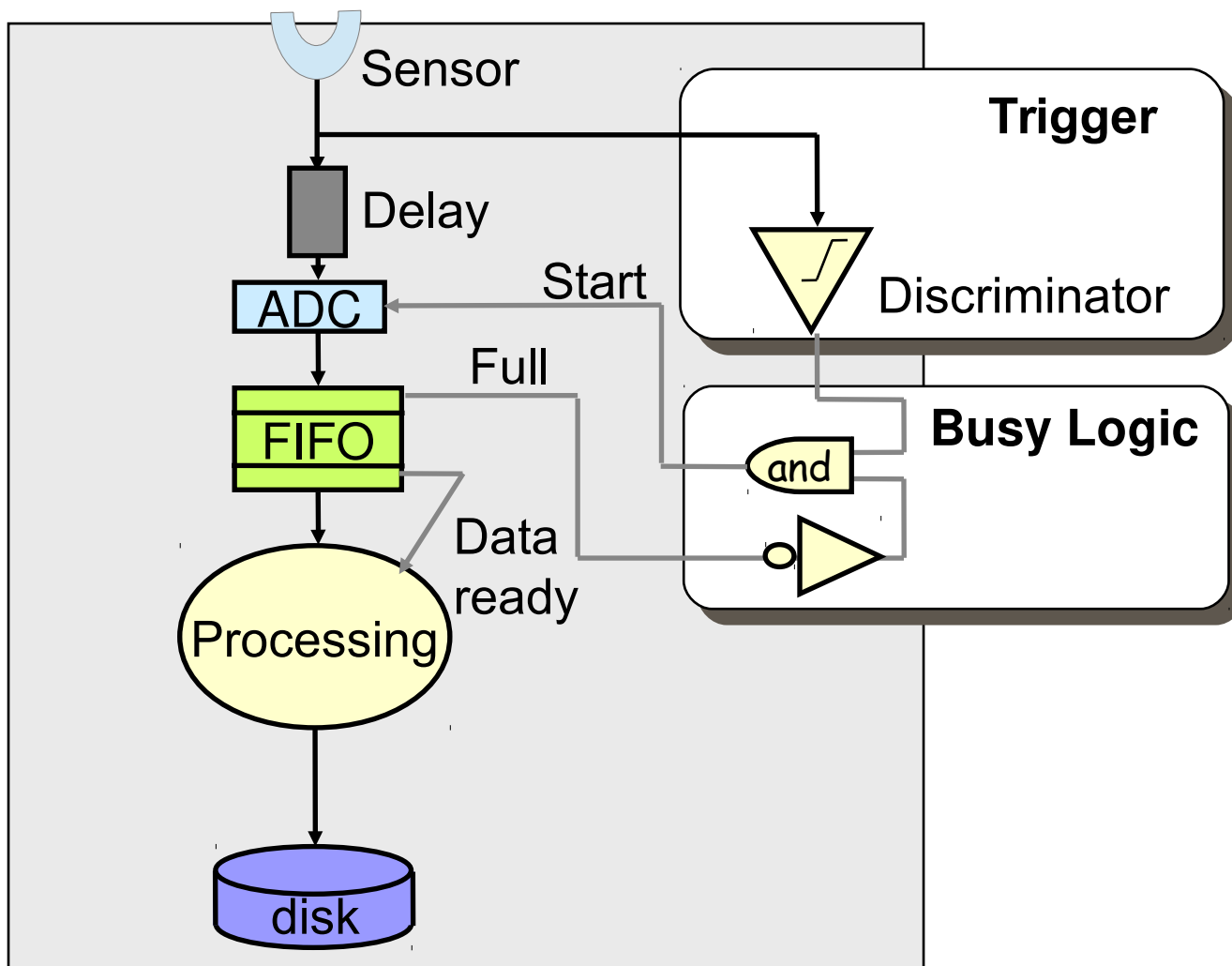
→ We can now attain a FIFO efficiency
~100% with $\tau \sim \lambda$

- moderate buffer size

Analytic calculation possible for very simple systems only. Otherwise simulations must be used.

De-randomization: summary


 $f=1\text{kHz}$
 $1/f=\lambda=1\text{ms}$



→ Almost 100% efficiency and minimal deadtime are achieved if

- ADC is able to operate at rate $\gg f$
- data processing and storing operates at $\sim f$

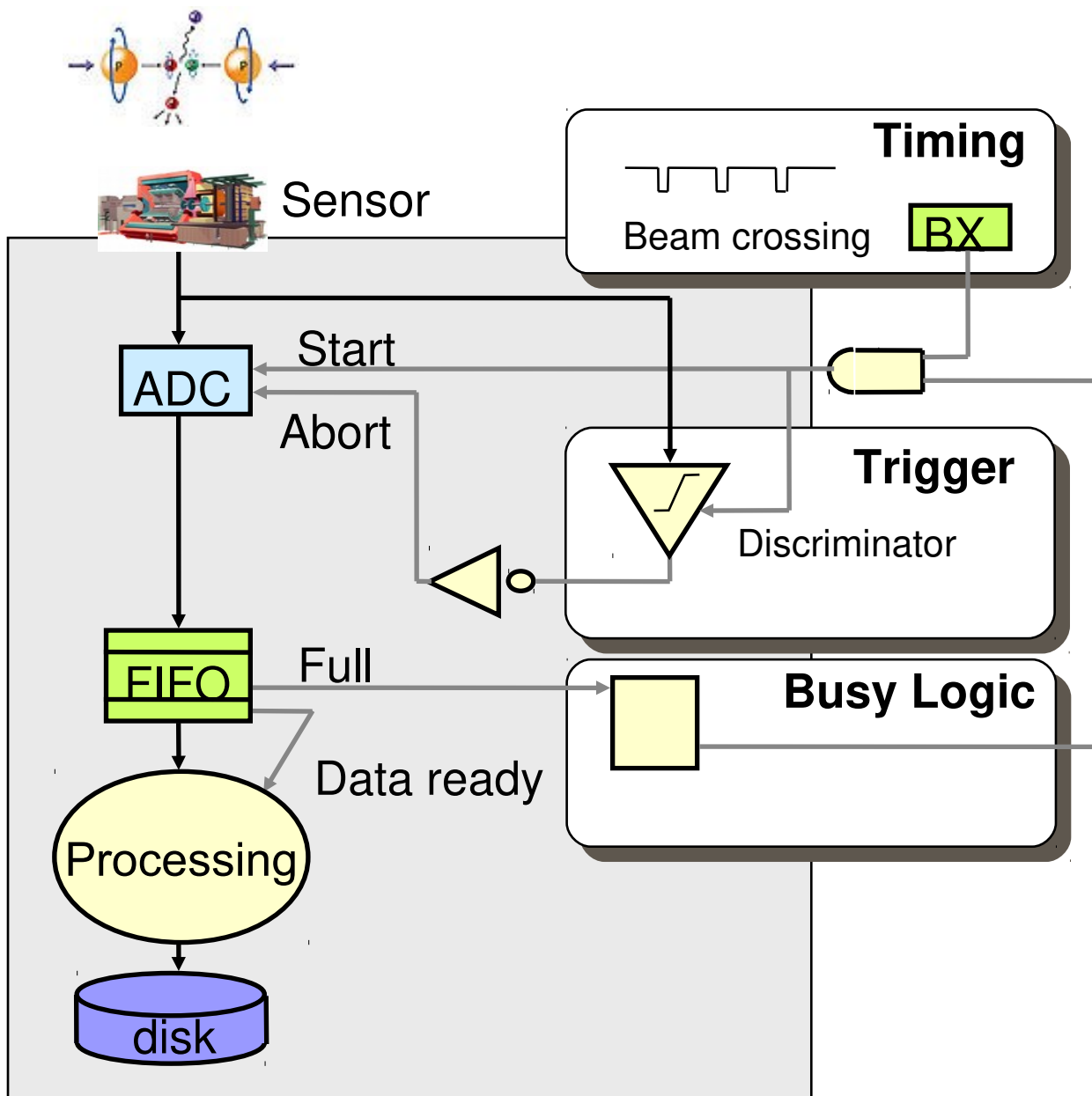
→ The FIFO decouples the low latency front-end from the data processing

- minimize the amount of "unnecessary" fast components

→ Could the delay be replaced with a "FIFO"?

- analog pipelines → Heavily used in LHC DAQs

Basic DAQ: collider mode



- Particle collisions are synchronous
- Trigger rejects uninteresting events
- Even if collisions are synchronous, the triggers (i.e. good events) are **unpredictable**
- De-randomization is still needed

Multi-level trigger systems

→ Sometime impossible to take a proper decision in a single place

- too long decision time
- too far
- too many inputs

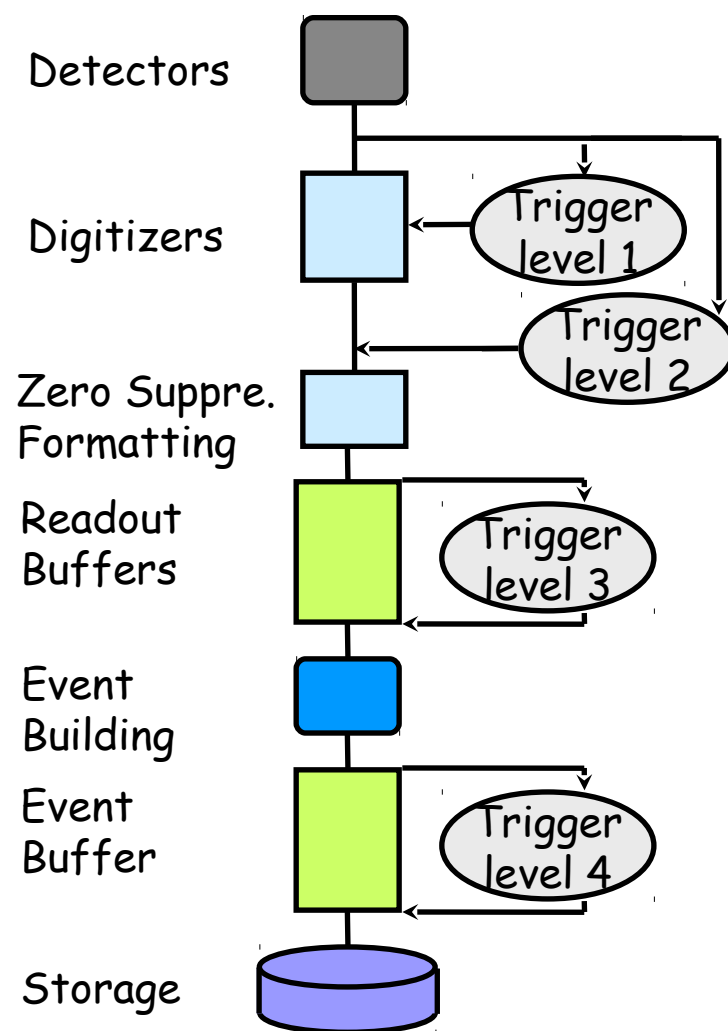
→ Distribute the decision burden in a hierarchical structure

- usually $\tau_{N+1} \gg \tau_N$, $f_{N+1} \ll f_N$

→ At the DAQ level, proper buffering must be provided for every trigger level

- absorb latency
- de-randomize

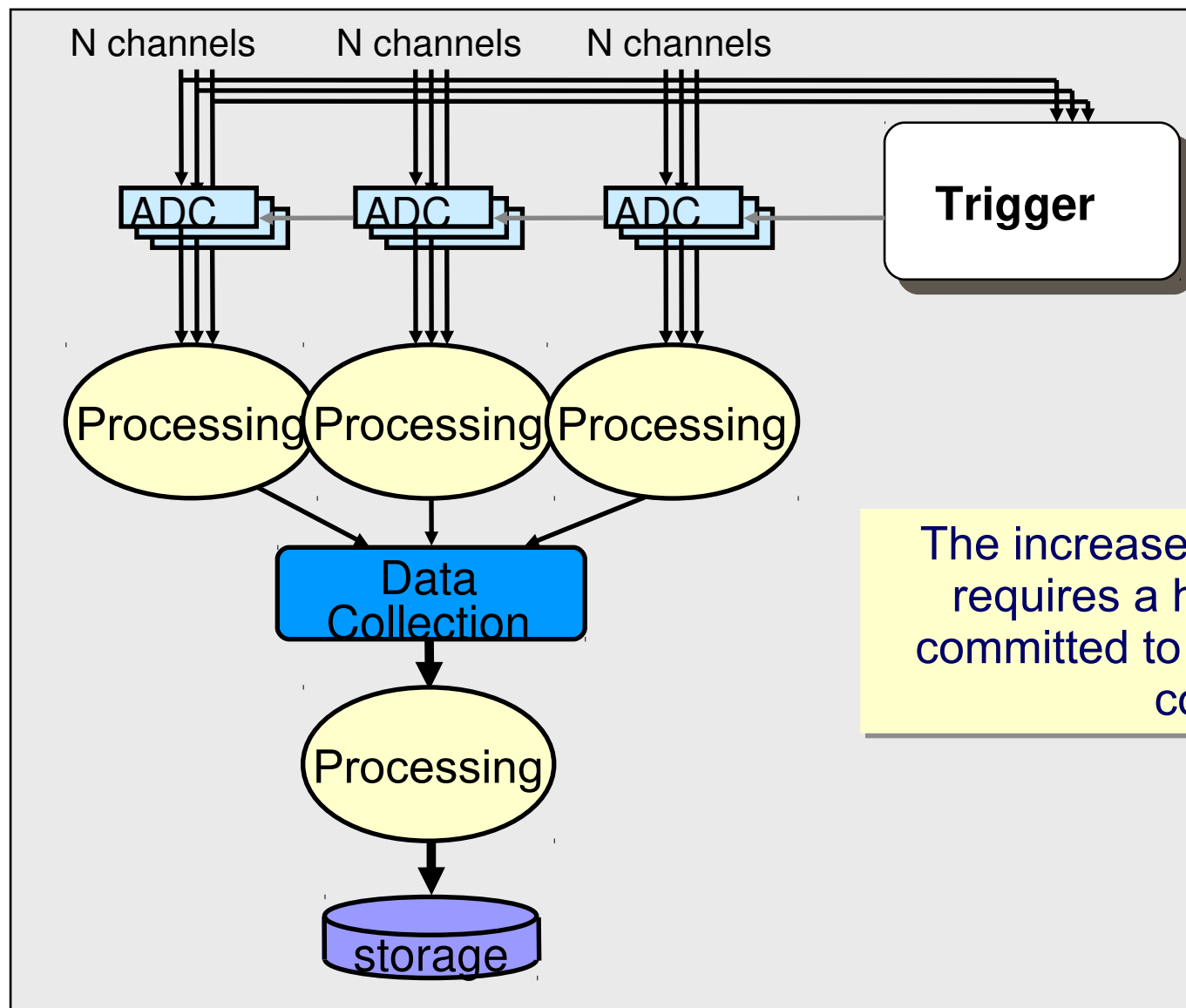
→ Data must be transported from level to level





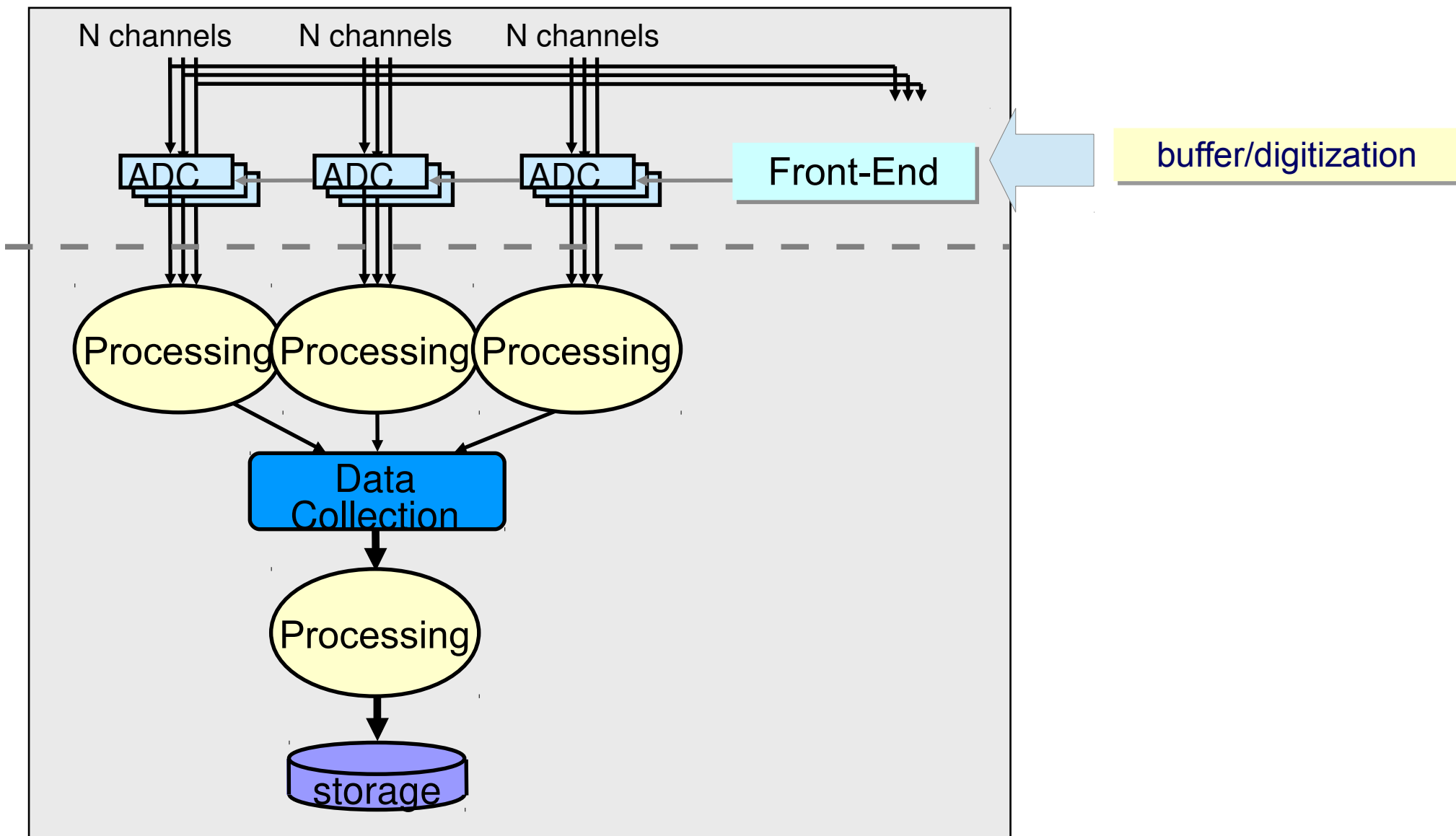
DAQ Scaling up

Basic DAQ: more channels

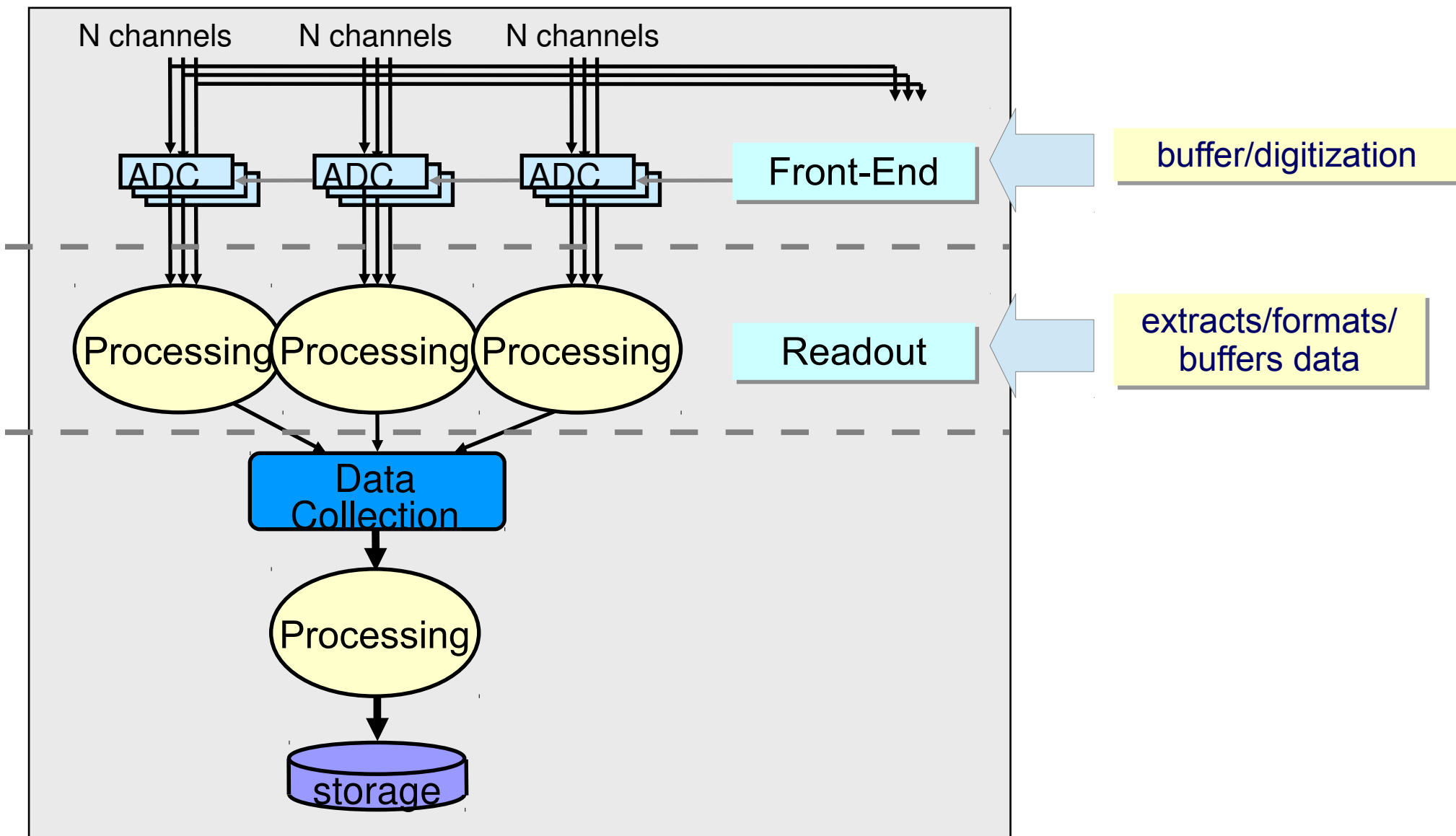


The increased number of channels requires a hierarchical structure committed to the data handling and conveyance

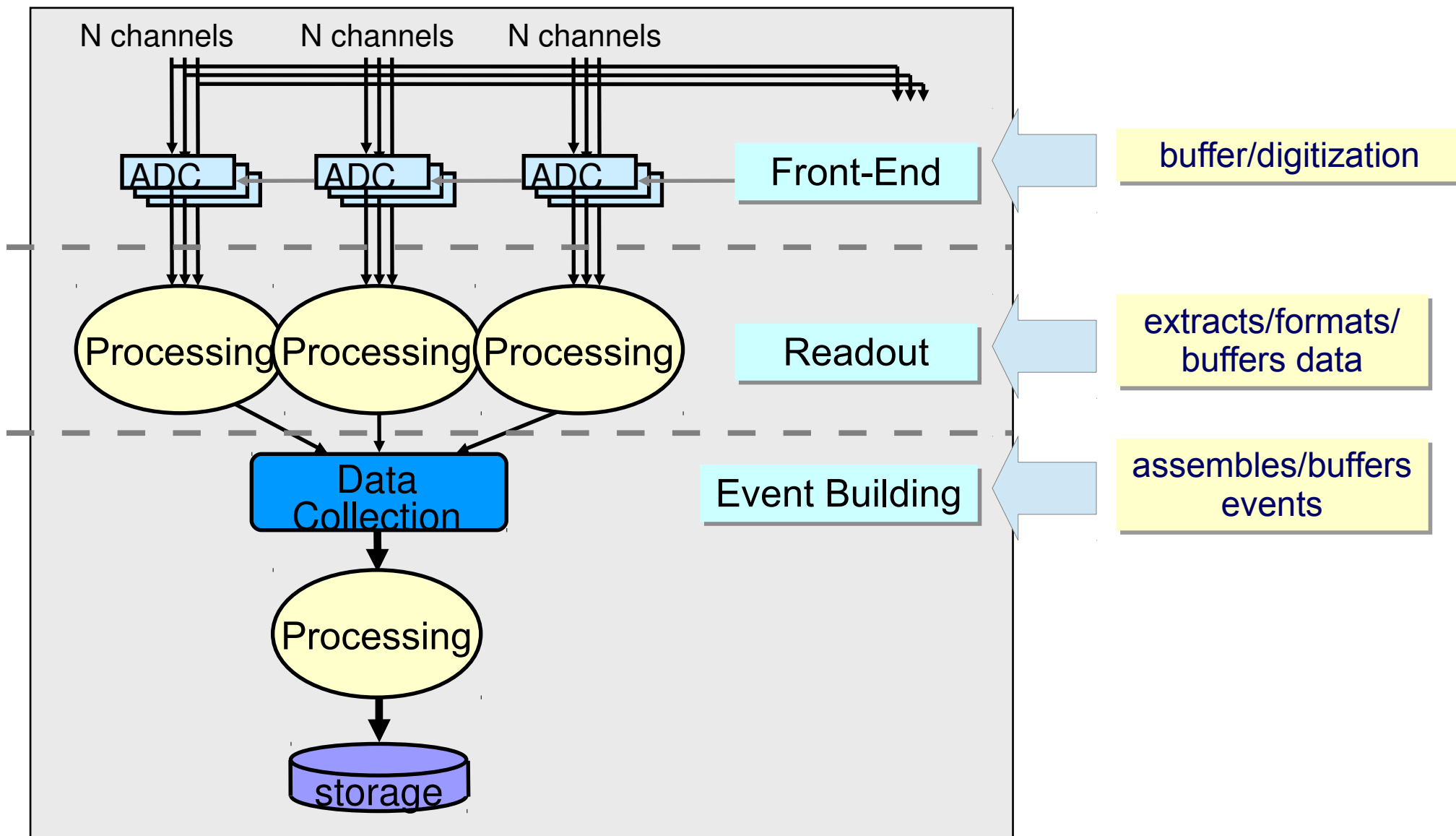
Large DAQ: Constituents



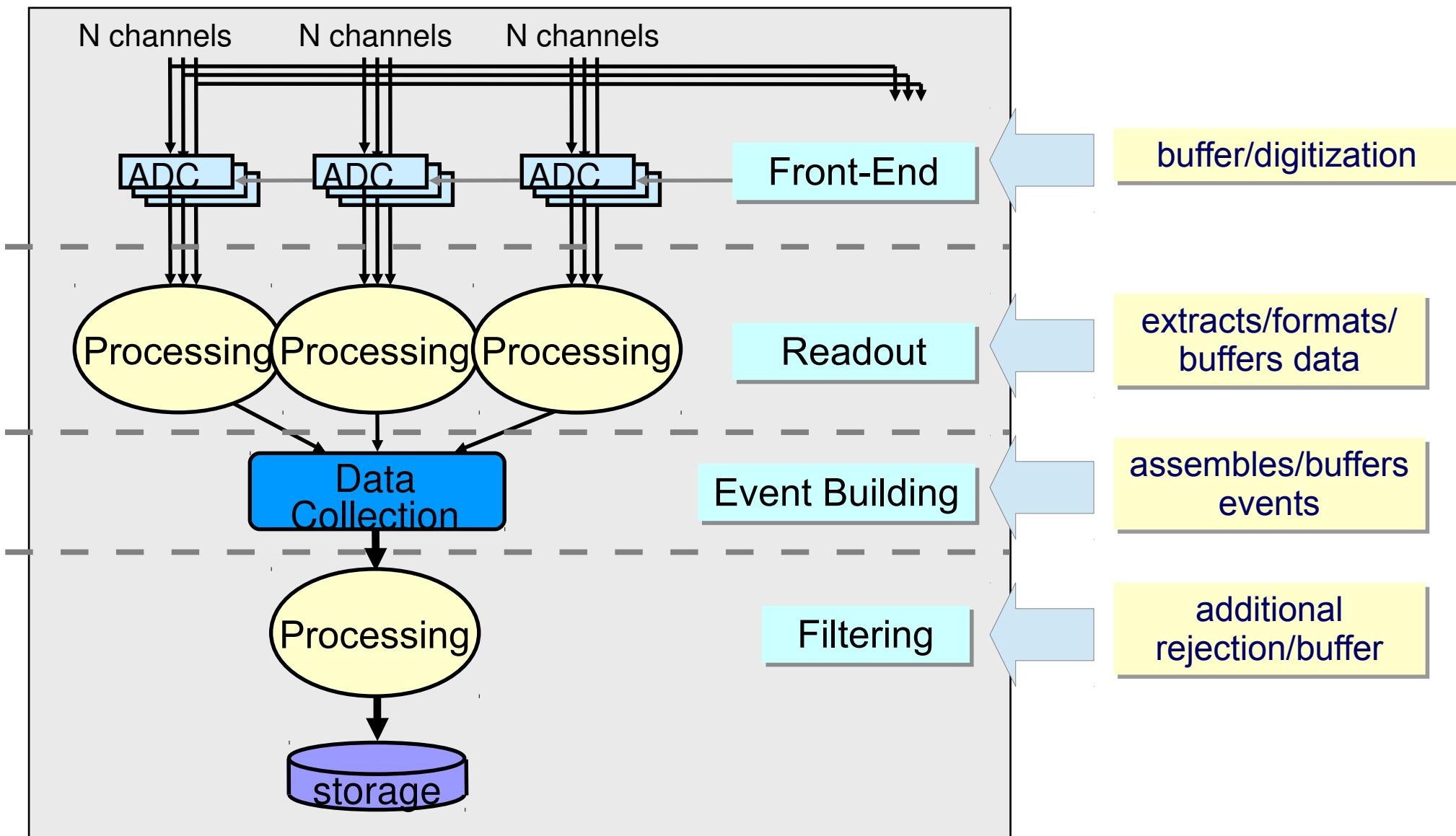
Large DAQ: Constituents



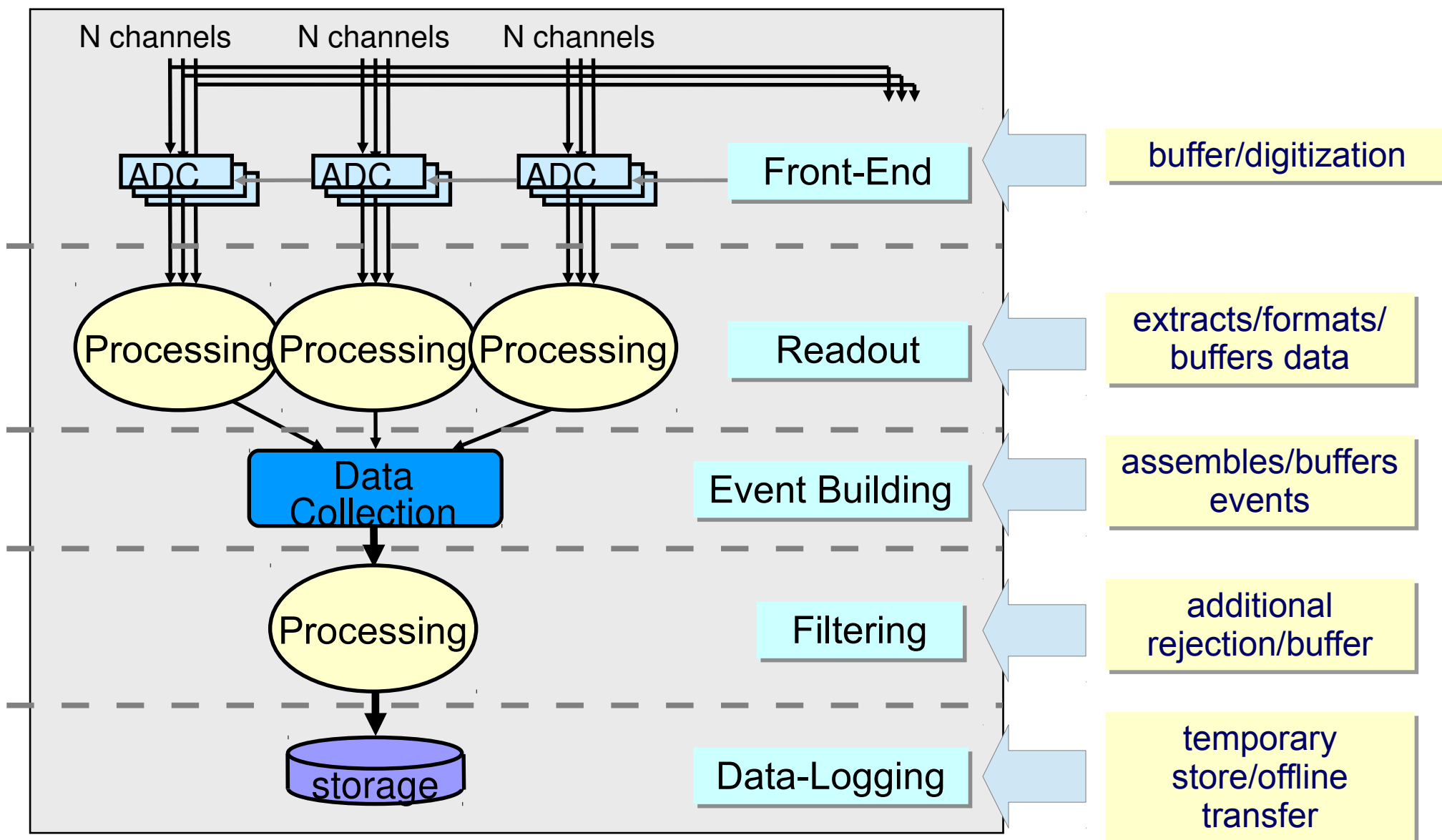
Large DAQ: Constituents



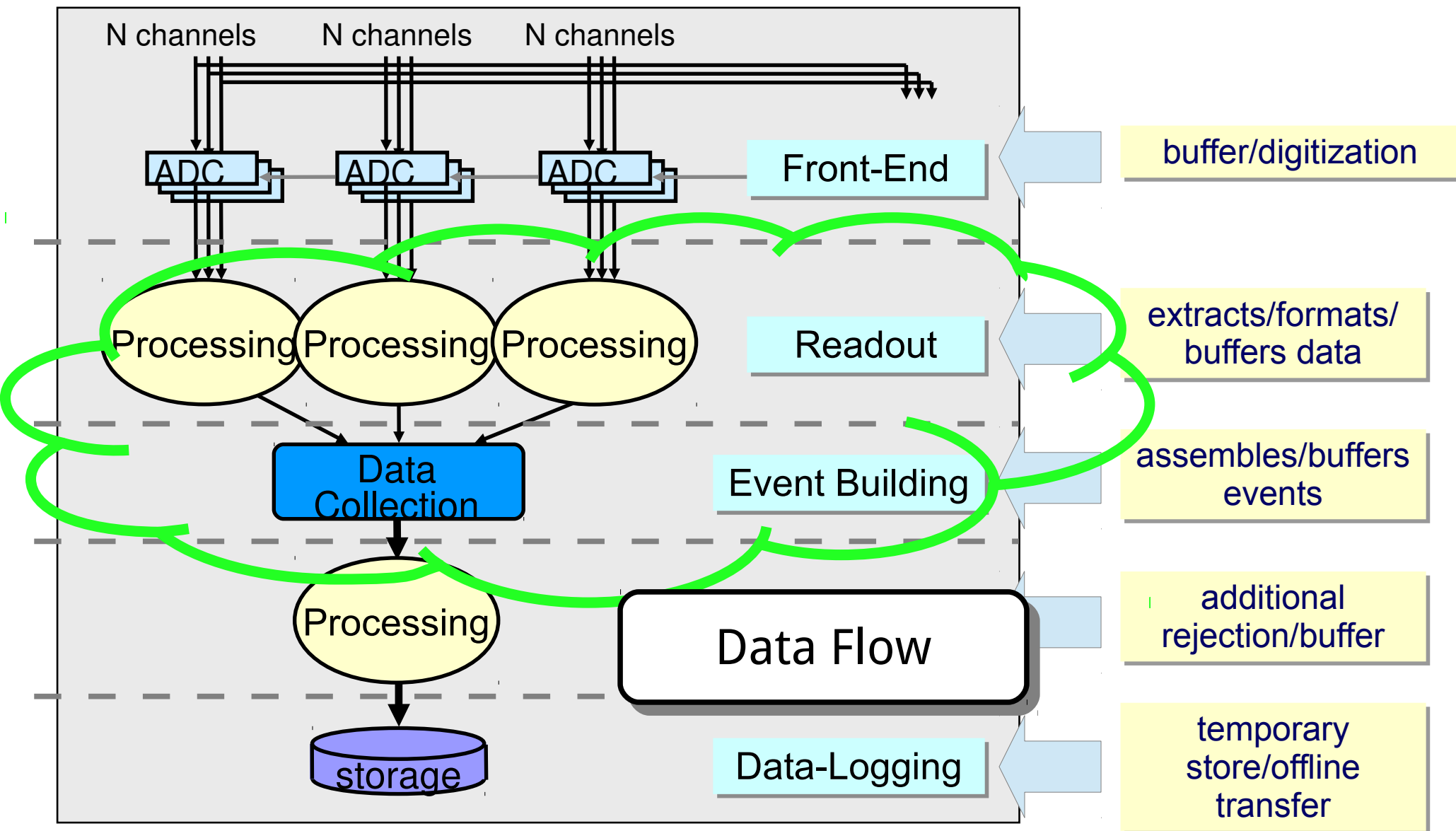
Large DAQ: Constituents



Large DAQ: Constituents



Large DAQ: Constituents





DAQ@LHC

→ LHC experiments have $O(10^7)$ channels operating at 40 MHz (25 ns) → **40 TB/s**

→ In addition, interesting phenomena are **extremely rare**

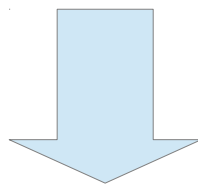
$$\sigma_H / \sigma_{Tot} \sim O(10^{-13})$$



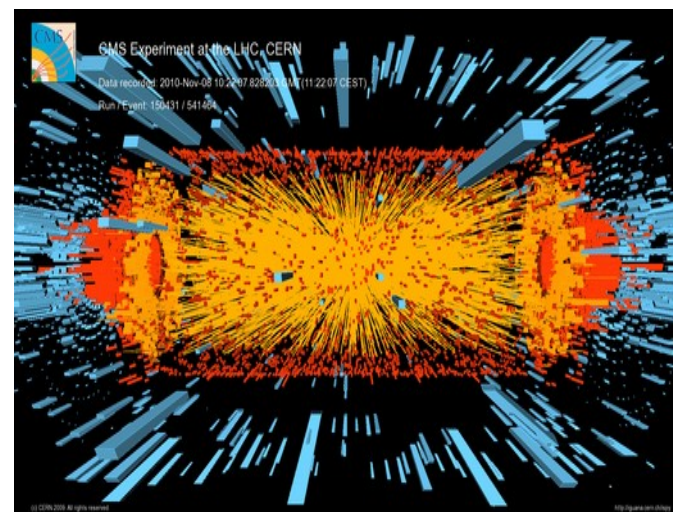
→ Events are complex

- significant number of overlapping collisions (pile-up μ)

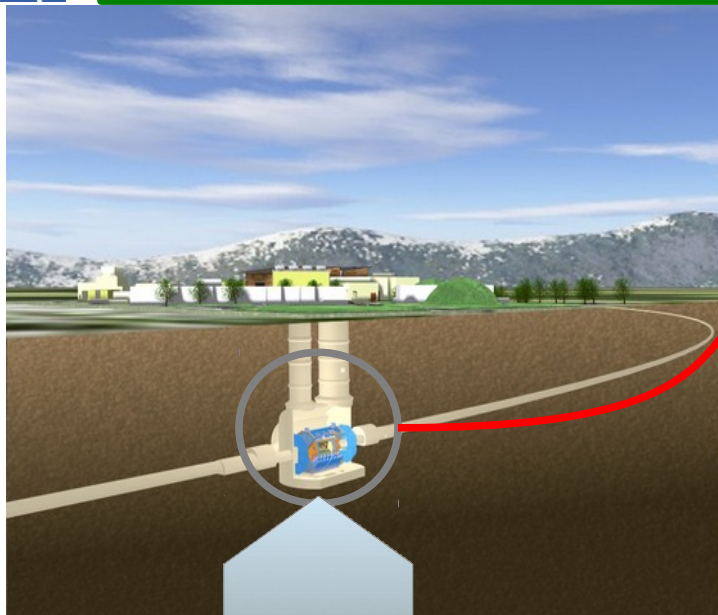
→ Experiments are large ($O(10\text{ m})$)



Multi-level trigger system and ...

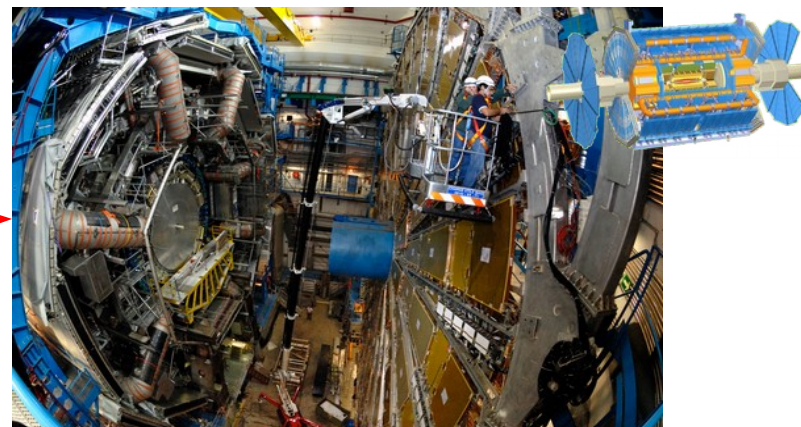


Trigger & DAQ Challenges at the LHC



Challenging environment and requirements

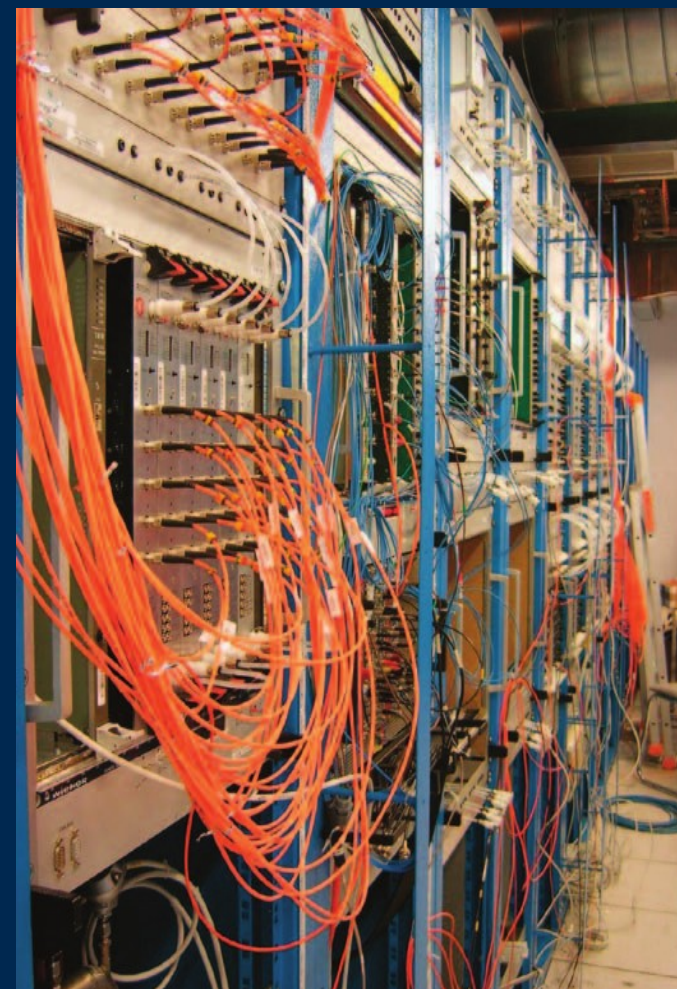
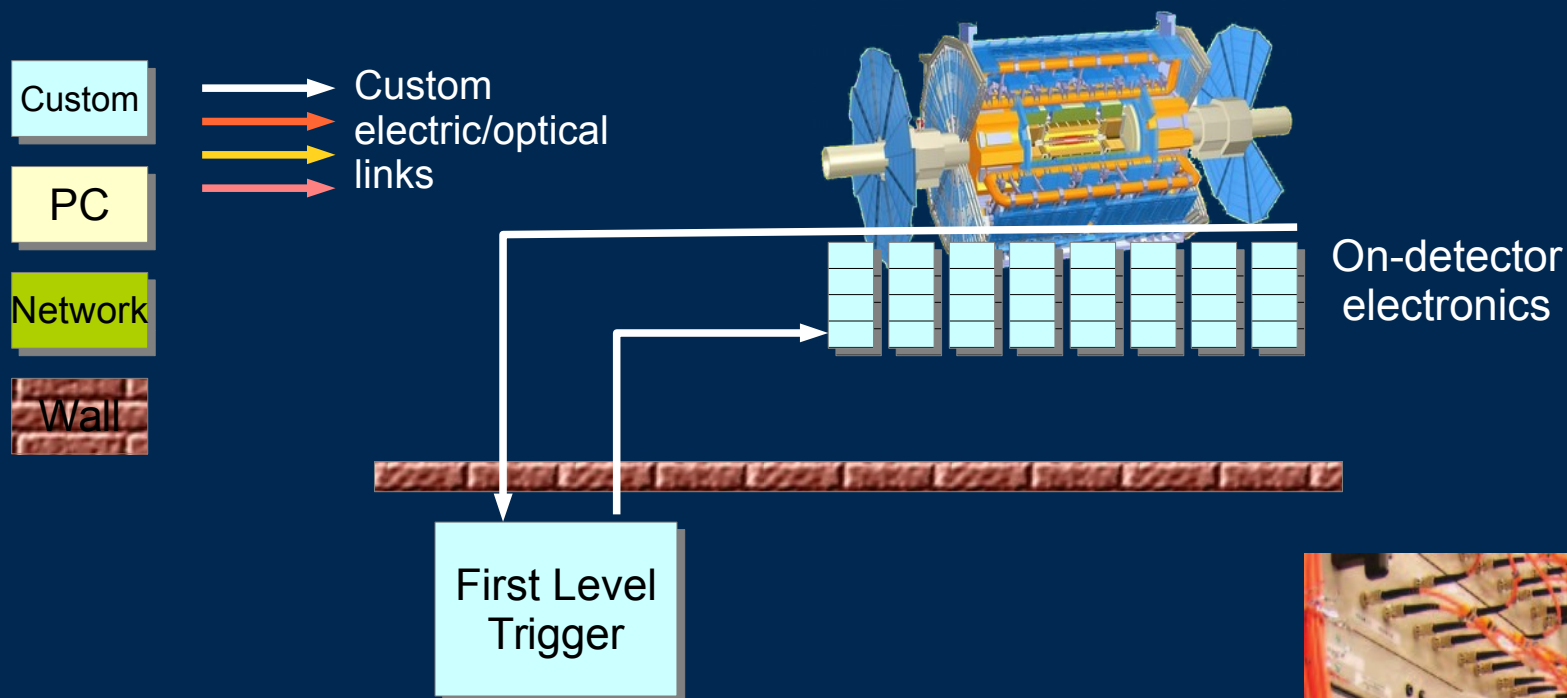
- underground cavern
- **space** constraints
- **power** consumption constraints
- high **radiation** levels
- desire to **limit non-active material**
- **magnetic field**



LHC Experiment

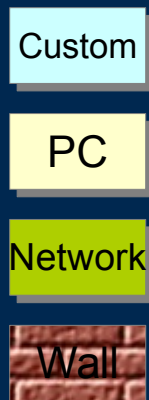
- Different particle detection technologies organized in layers
- on-detector **custom electronics** forms and **digitize** signals
- operates at the LHC rate of **40 MHz**

	ATLAS
Length (m)	46
Diameter (m)	25
Weight (t)	7000
Number of electr. channels	$100 \cdot 10^6$

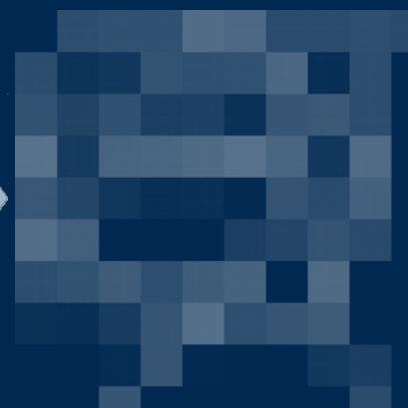
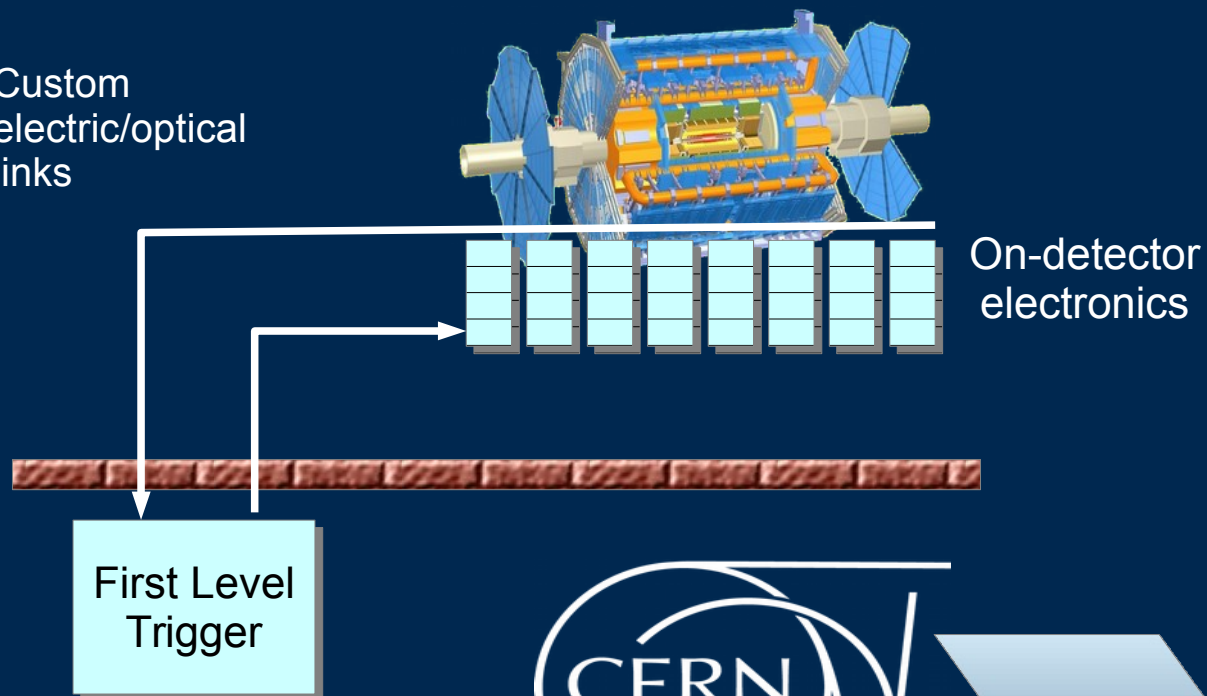


First Level Trigger

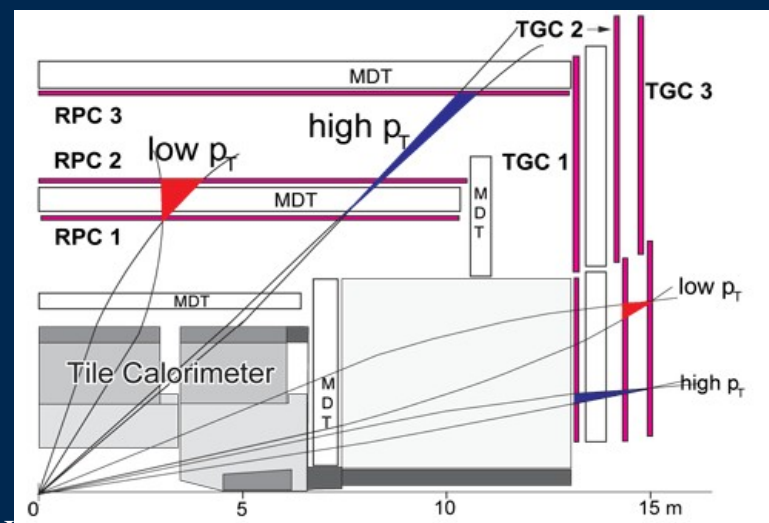
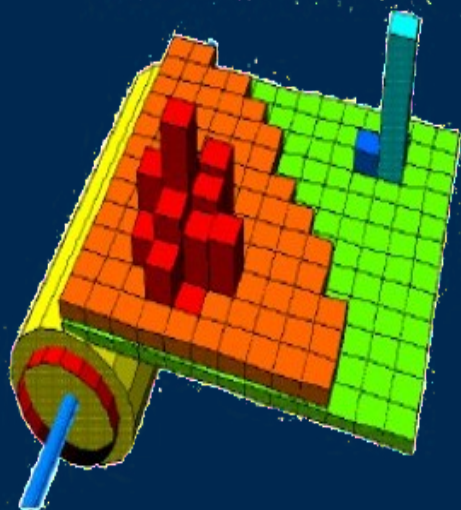
- Operates at 40 MHz
- Uses **reduced granularity** information from detector
- Simple, hardware-friendly algorithms
- Small, **deterministic decision time** ($\sim \mu\text{s}$)
- Implemented with **custom electronics: ASIC/FPGA**
 - massively parallel through locality
- Reduces the rate to **100 kHz**



Custom
PC
Network
Wall

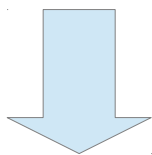


First Level Trigger

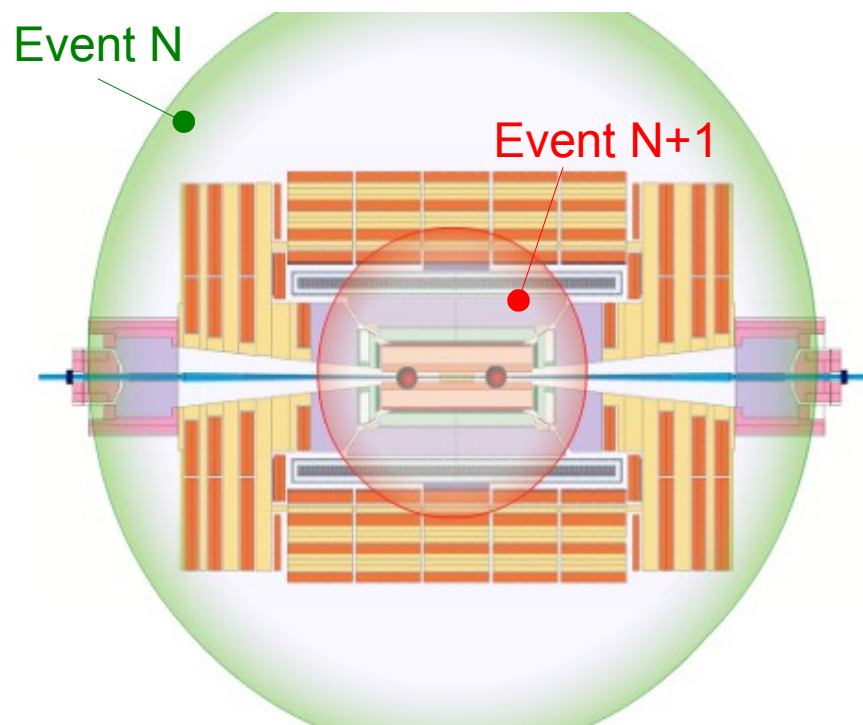


→ Particle time of flight $\gg 25$ ns

→ Cable delays $\gg 25$ ns

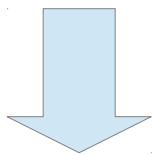


Dedicated synchronization, timing and signal distribution facilities

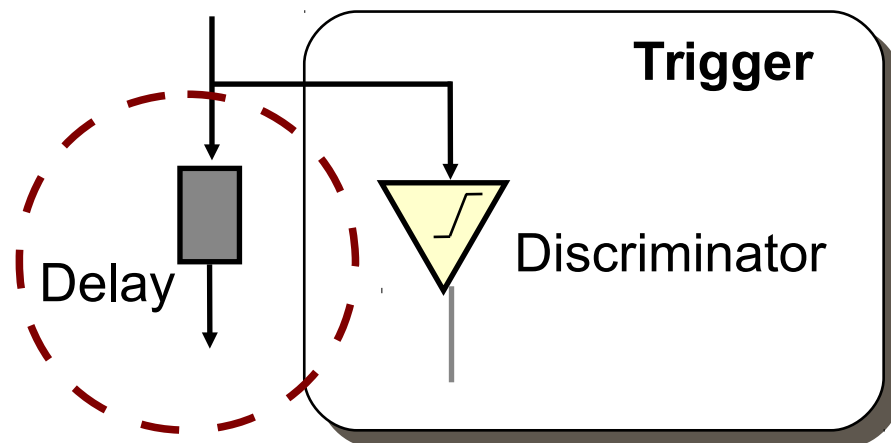


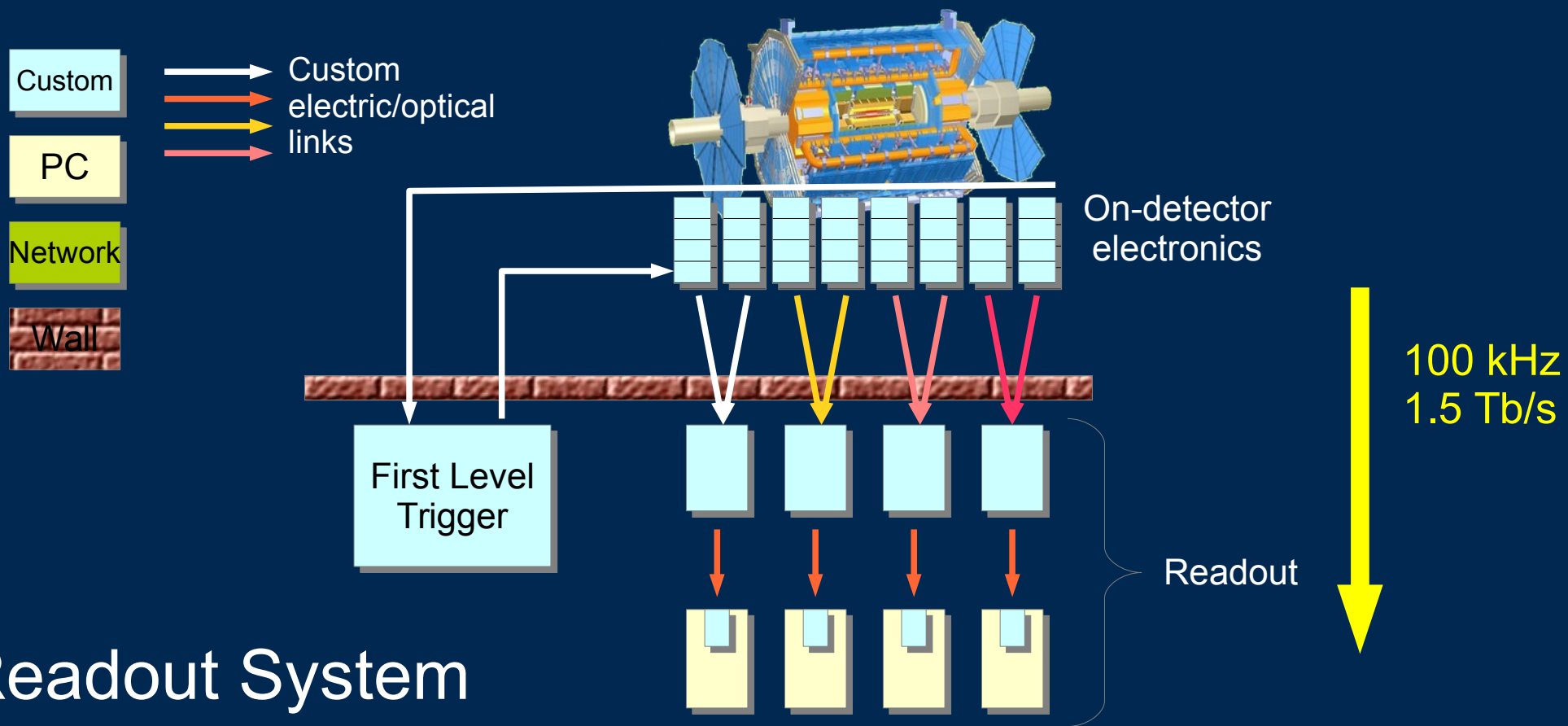
→ Typical L1 decision latency is $O(\mu\text{s})$

- dominated by signal propagation in cables



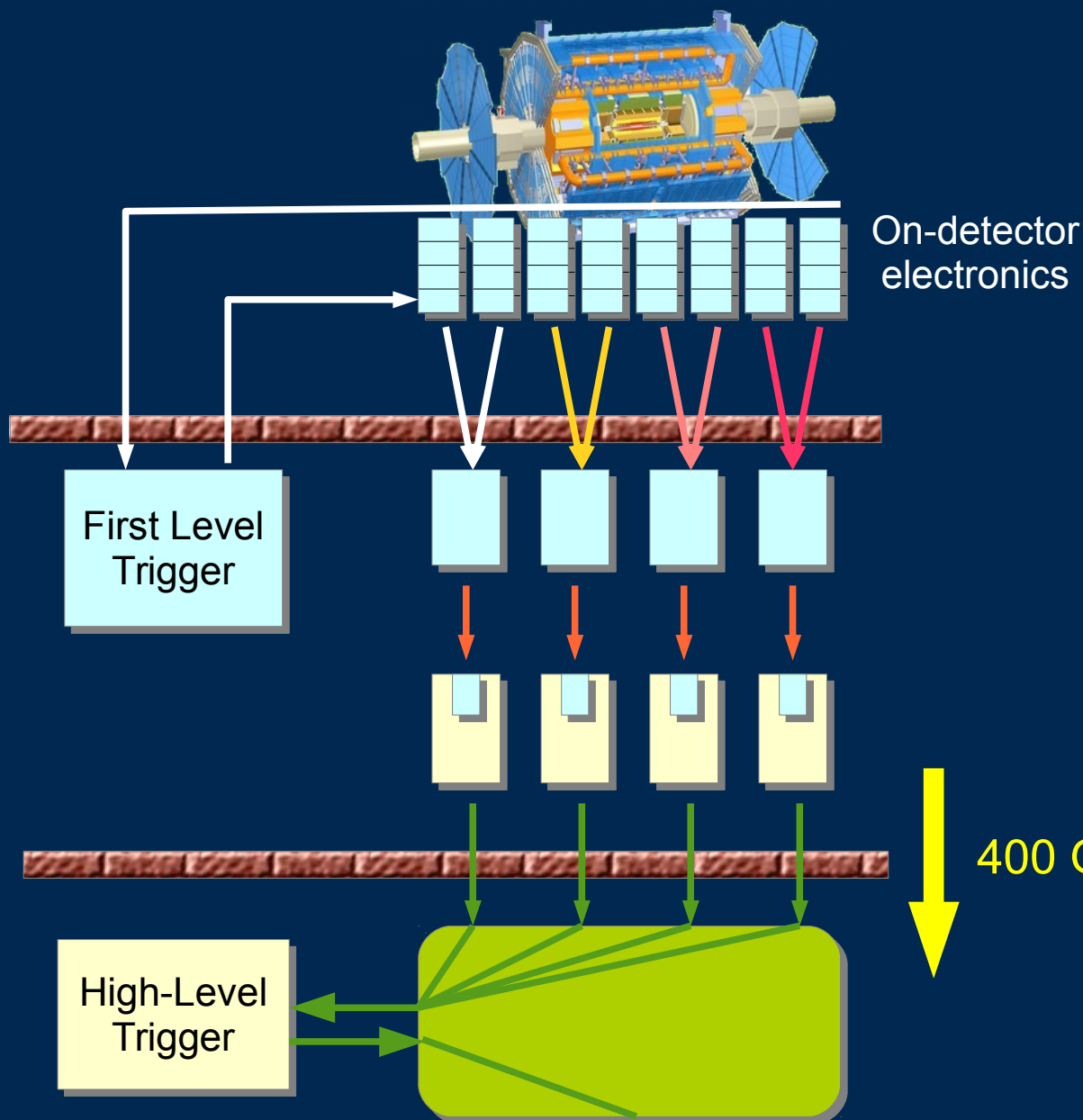
Digital/analog custom front-end pipelines store information during L1 trigger decision





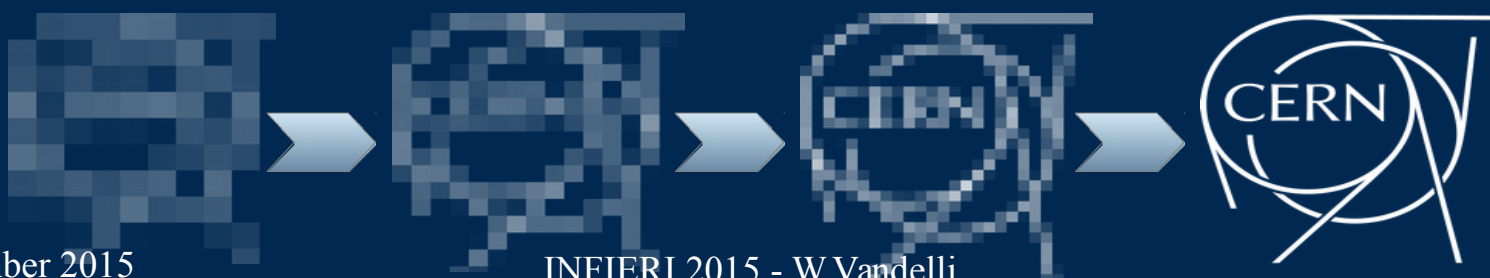
Readout System

- Data from events selected by First Level Trigger are pushed by on-detector electronics **over detector-depend links**
- Back-end custom electronics
 - adapt from specific link technology to common (custom) link technology
 - format data to a common standard
- Located in a **service cavern adjacent to the experiment**
- Readout PCs
 - buffer data
 - **connect 2000 custom links to a 10 GbE network**



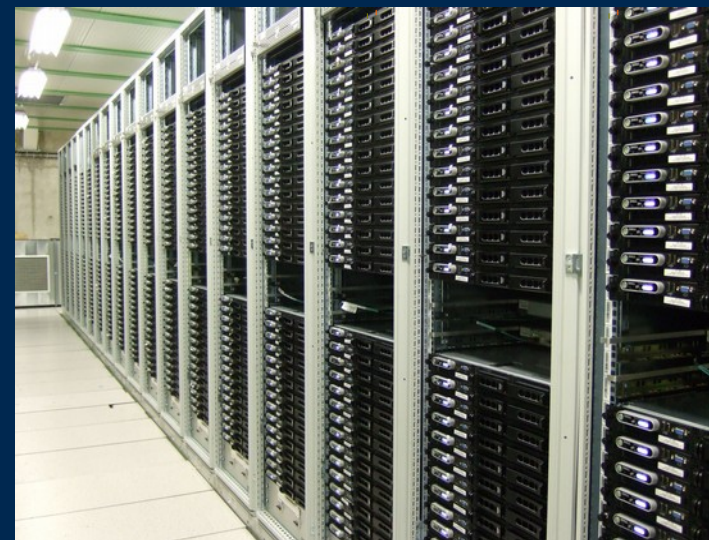
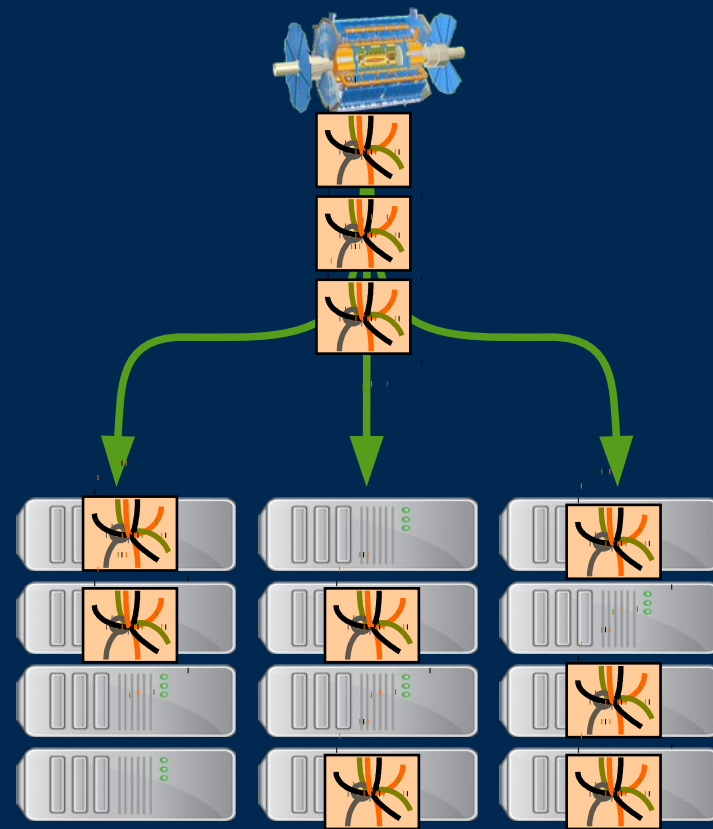
High Level Trigger

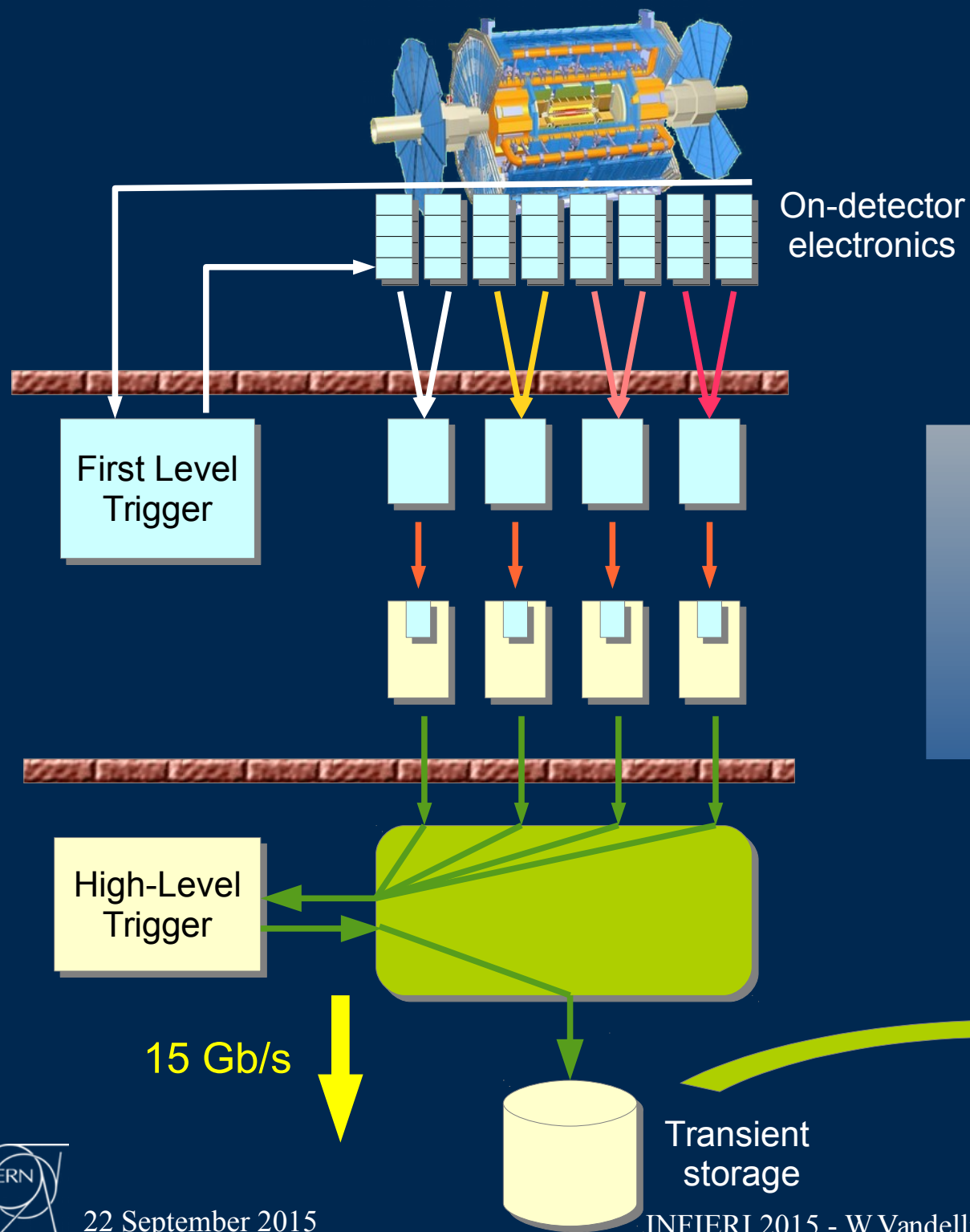
- Event rate is reduced further by SW algorithms
- Incremental data fetching and analysis
 - 400 Gb/s
 - average processing time ~200 ms
- From 100 kHz to 1 kHz



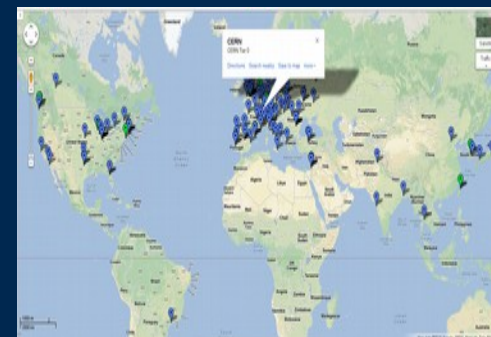
Data-Flow and High-Level Trigger: COTS domain

- High-Level Trigger is a **large computing farm**
 - several 10000 cores
 - fully based on COTS technologies
- Necessarily located in a **surface computing room**
- Primary **parallelism and scaling** scheme **through event distribution**
 - different **events are independent**
- Complemented by Readout and Data-Flow infrastructures
 - **convey data from the detector** (underground) to the computing farm
 - **equalizes farm usage** and implements buffering
- Maximize throughput
 - no inter-node communications





Select events are stored in a $\frac{1}{2}$ PB **transient storage system** and asynchronously moved to the off-line mass storage, data analysis and data distribution facilities



System distributed in space



radiation



power/cooling
limitations



space/material
limitations



magnetic field

Experimental cavern

Service cavern



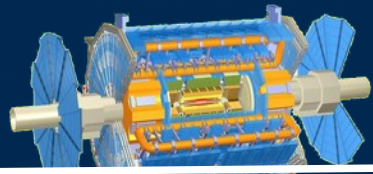
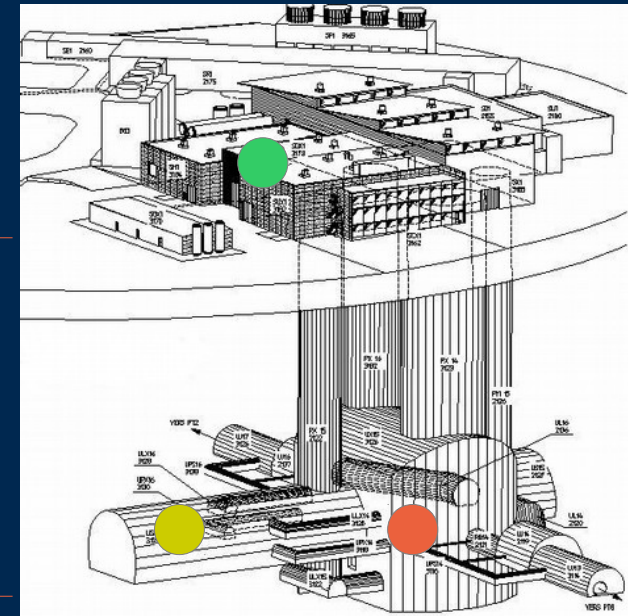
power/cooling
limitations



space
limitations

Surface

100 m



On-detector
electronics

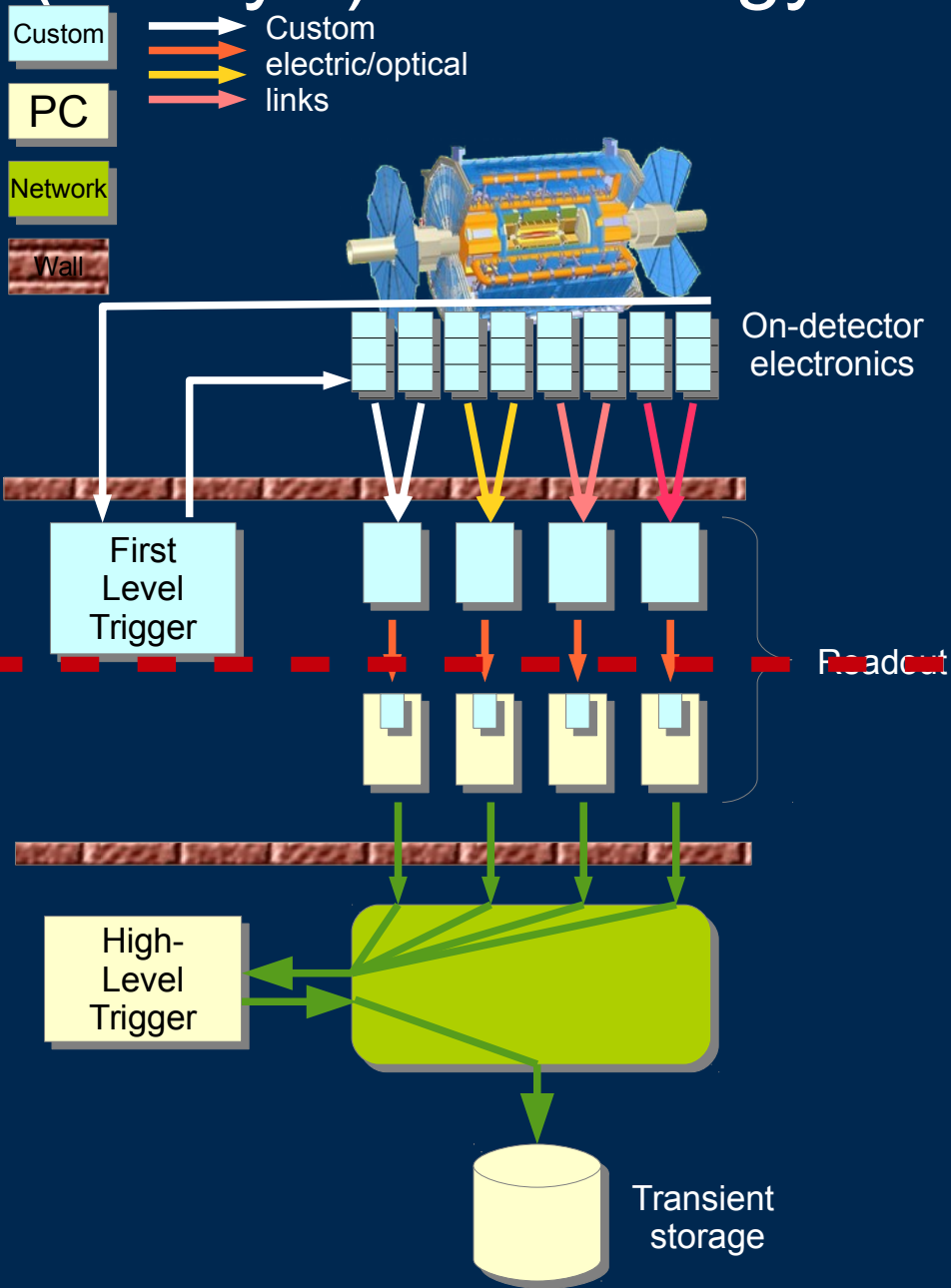
First
Level
Trigger

Readout

High-
Level
Trigger

Transient
storage

(Today's) Technology domains



- Custom electronics and serial links
- Best performance per cost for the different conditions

- e.g. inner detector has tighter requirements than muon spectrometer

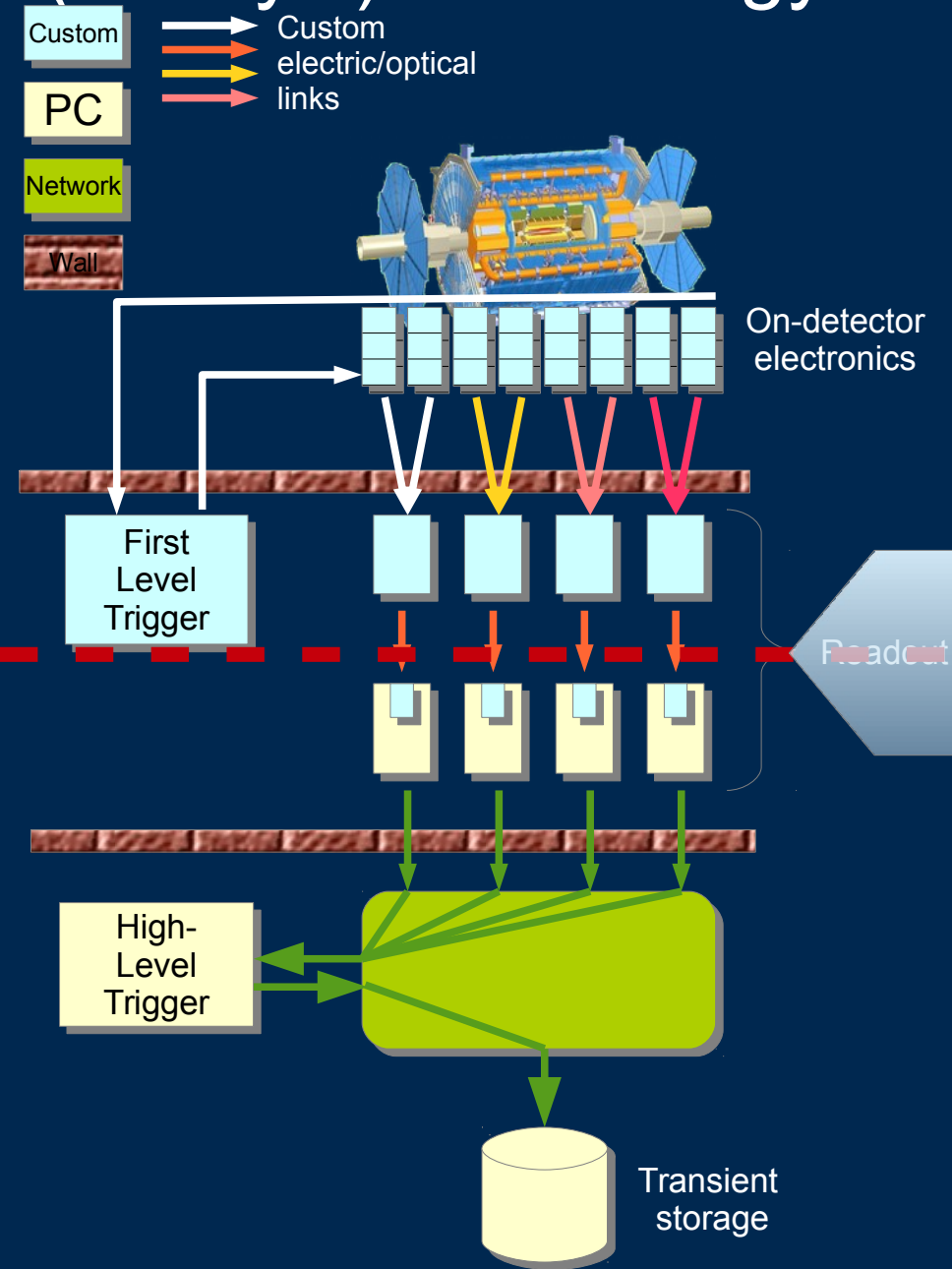
- Commercial Off-the-Shelf (COTS) Computing and Networking

- Software is more flexible than firmware

- less steep learning curve

- Easier to maintain, replace, mix and match

(Today's) Technology domains



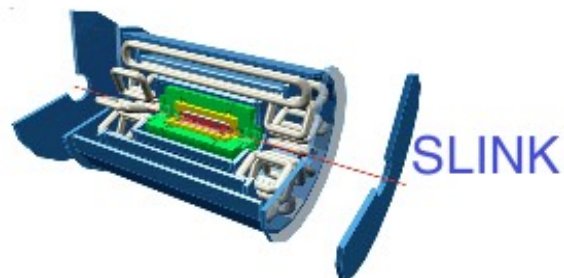
- Custom electronics and serial links
- Best performance per cost for the different conditions
 - e.g. inner detector has tighter

Two domains are linked by common (still custom) serial link technology

Electronics could not implement high-level protocols needed by a network (e.g. TCP/IP)

- Computing and Networking
- Software is more flexible than firmware
 - less steep learning curve
- Easier to maintain, replace, mix and match

Read-out links at the LHC (in Run 1)



Optical: 160 MB/s ≈ 1600 Links
Receiver card interfaces to PC.

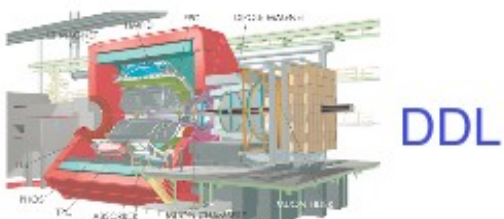
Flow Control

Yes



LVDS: 400 MB/s (max. 15m) ≈ 500 links
(FE on average: 200 MB/s to readout buffer)
Receiver card interfaces to commercial NIC
(Network Interface Card)

yes



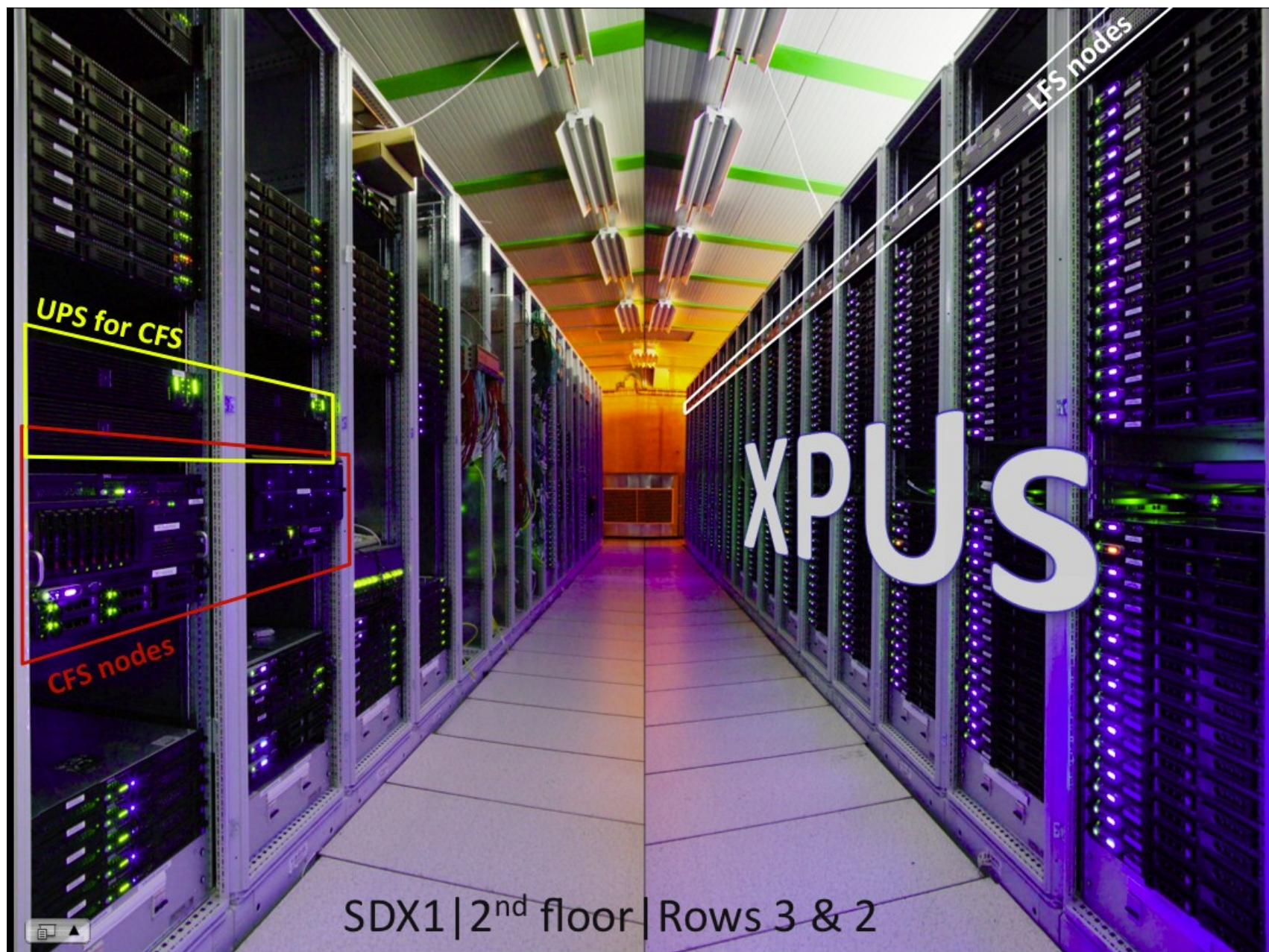
Optical 200 MB/s ≈ 500 links
Half duplex: Controls FE (commands,
Pedestals, Calibration data)
Receiver card interfaces to PC

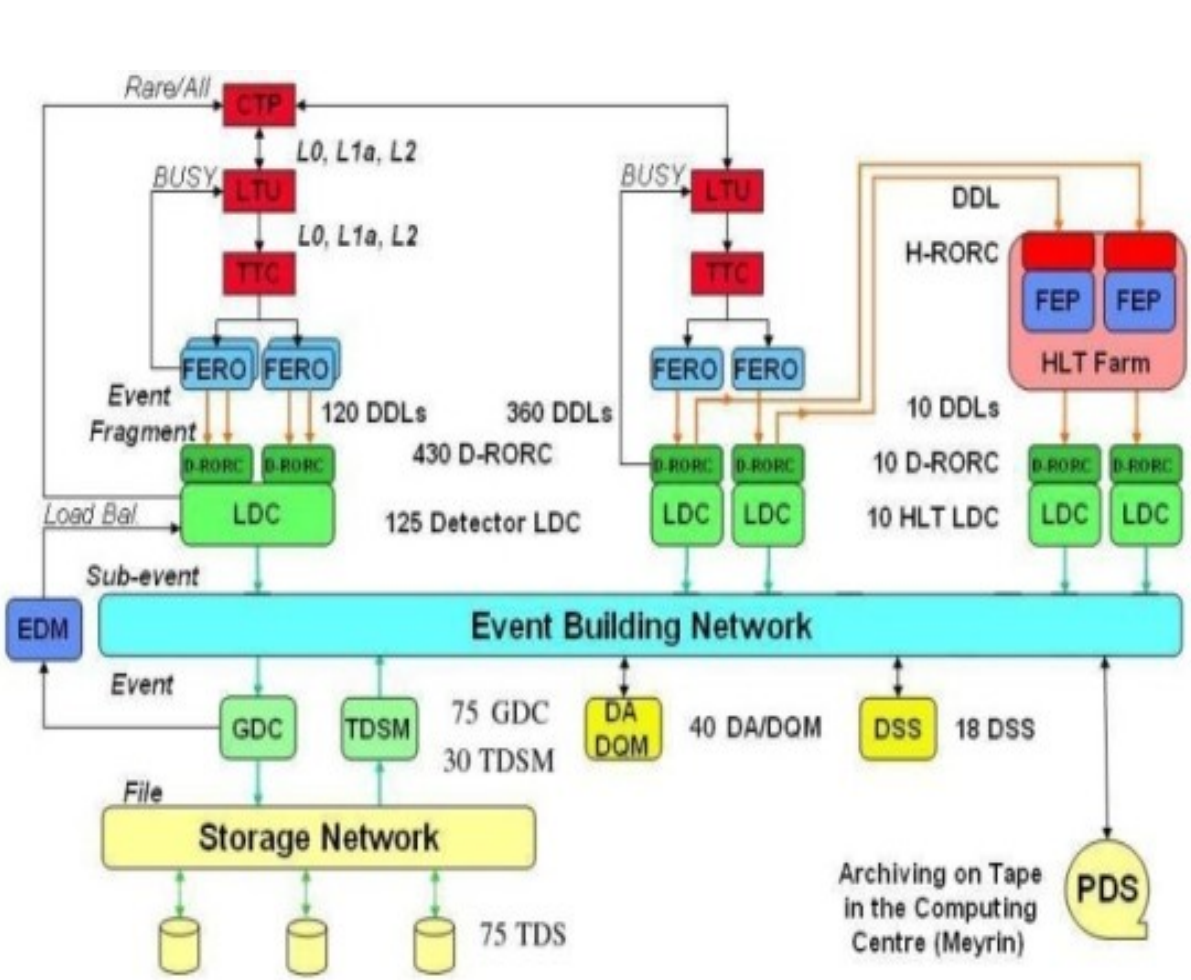
yes



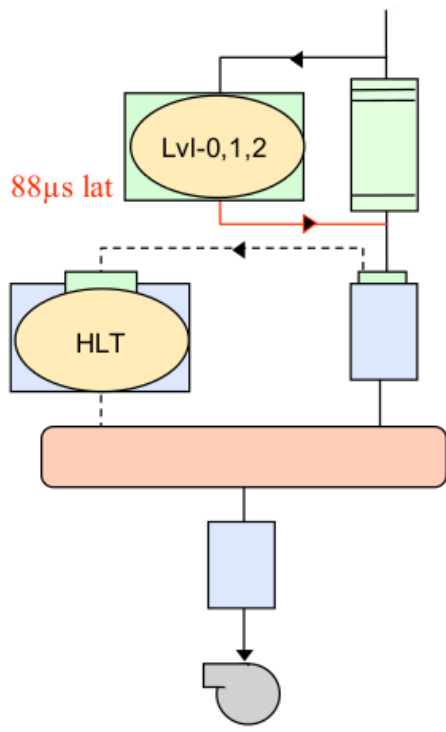
Copper quad GbE Link ≈ 400 links
Protocol: IPv4 (direct connection to GbE switch)
Forms "Multi Event Fragments"
Implements readout buffer

no

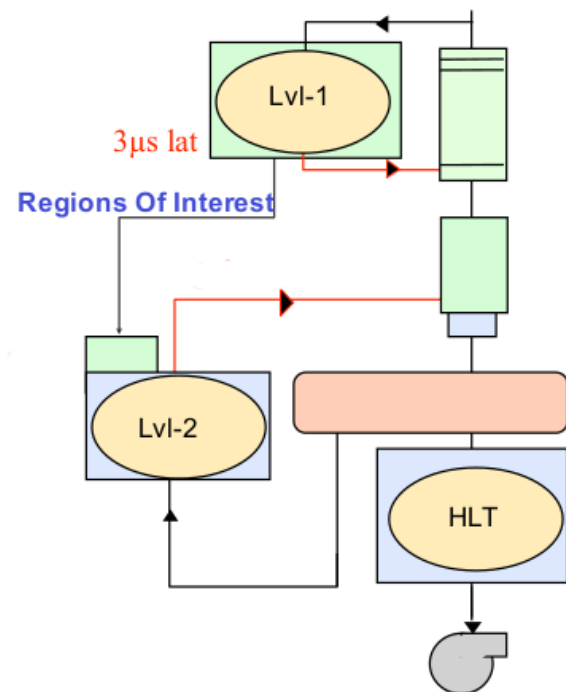
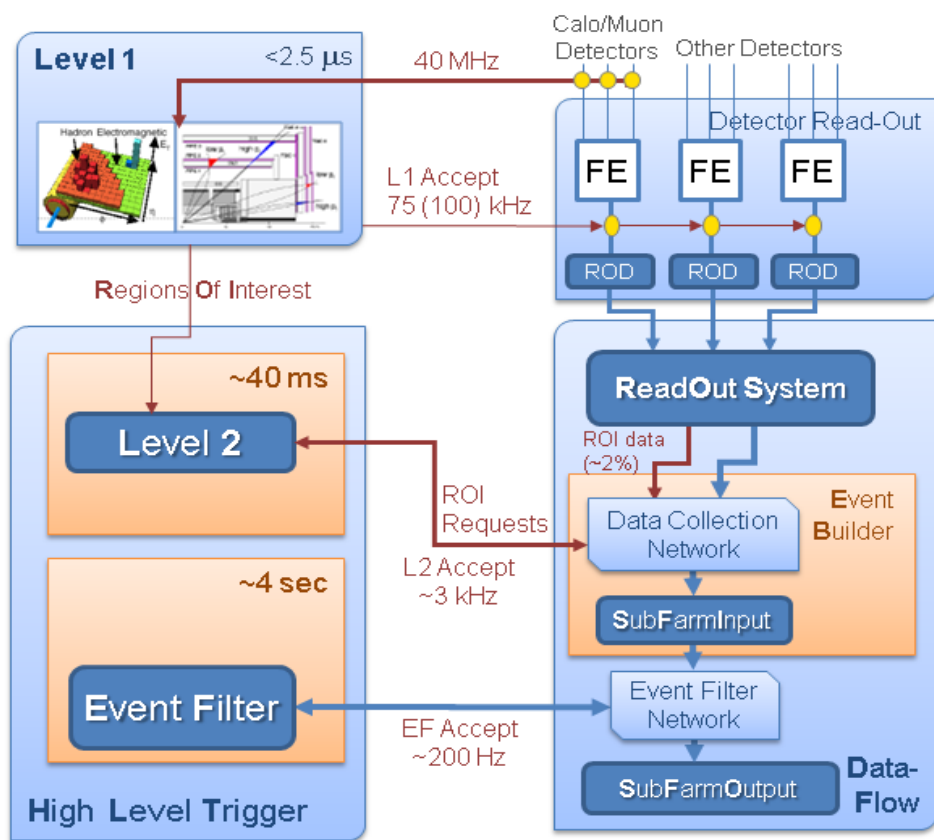


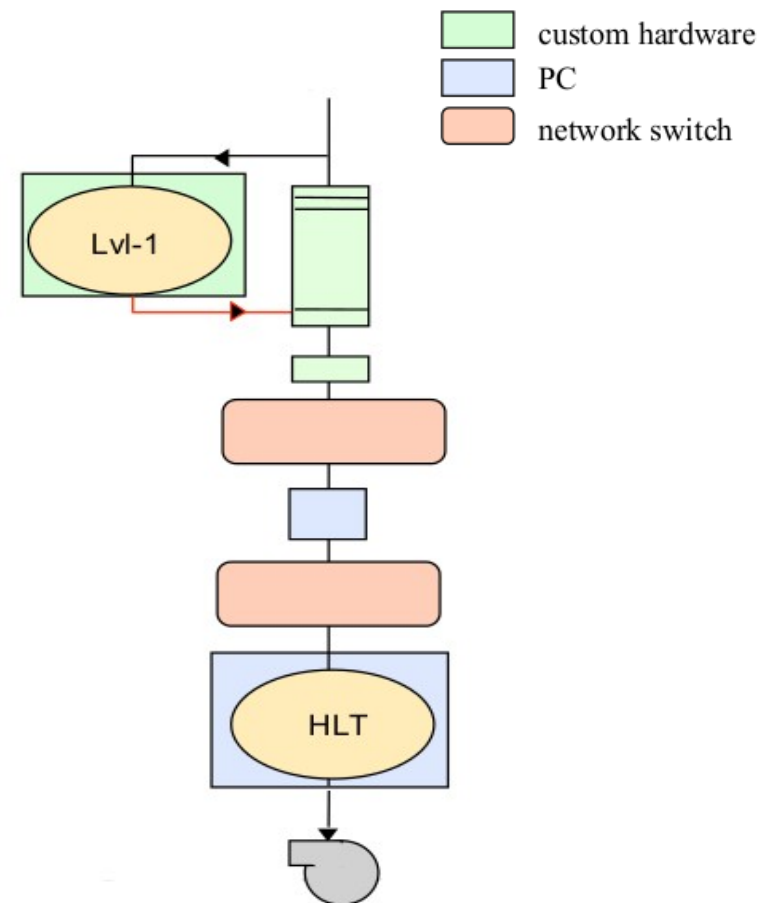
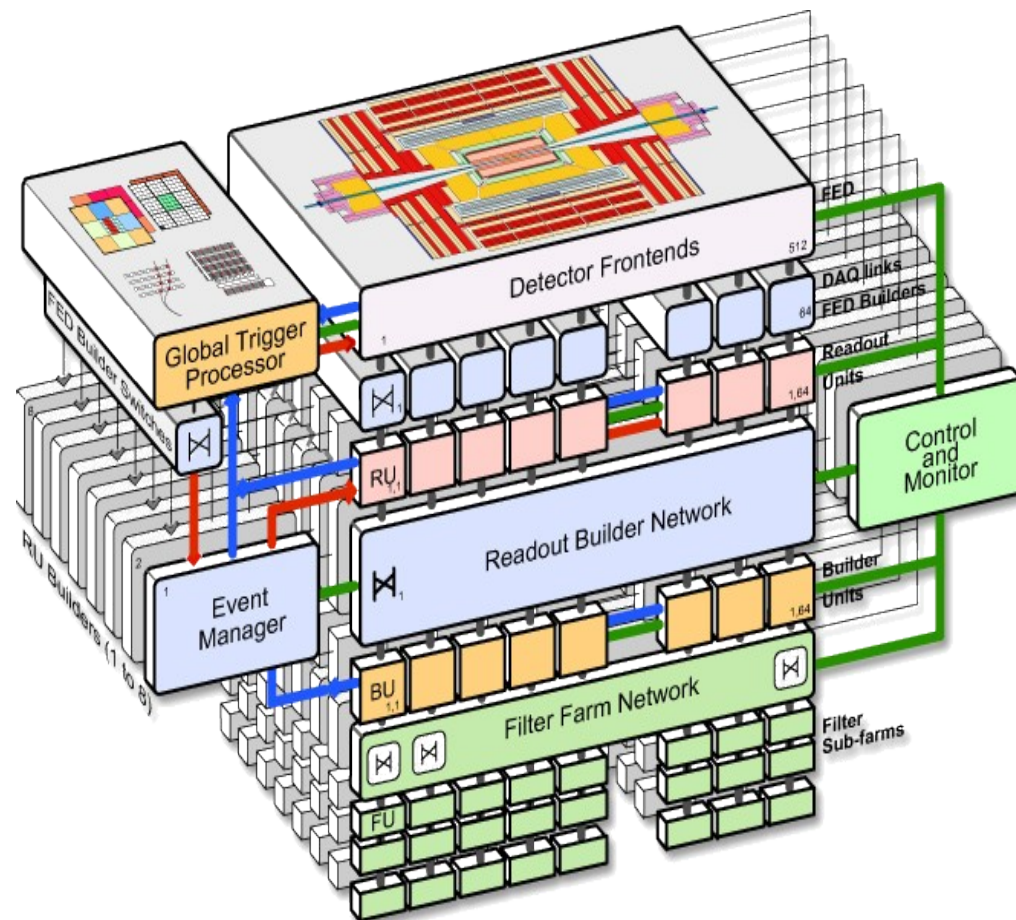


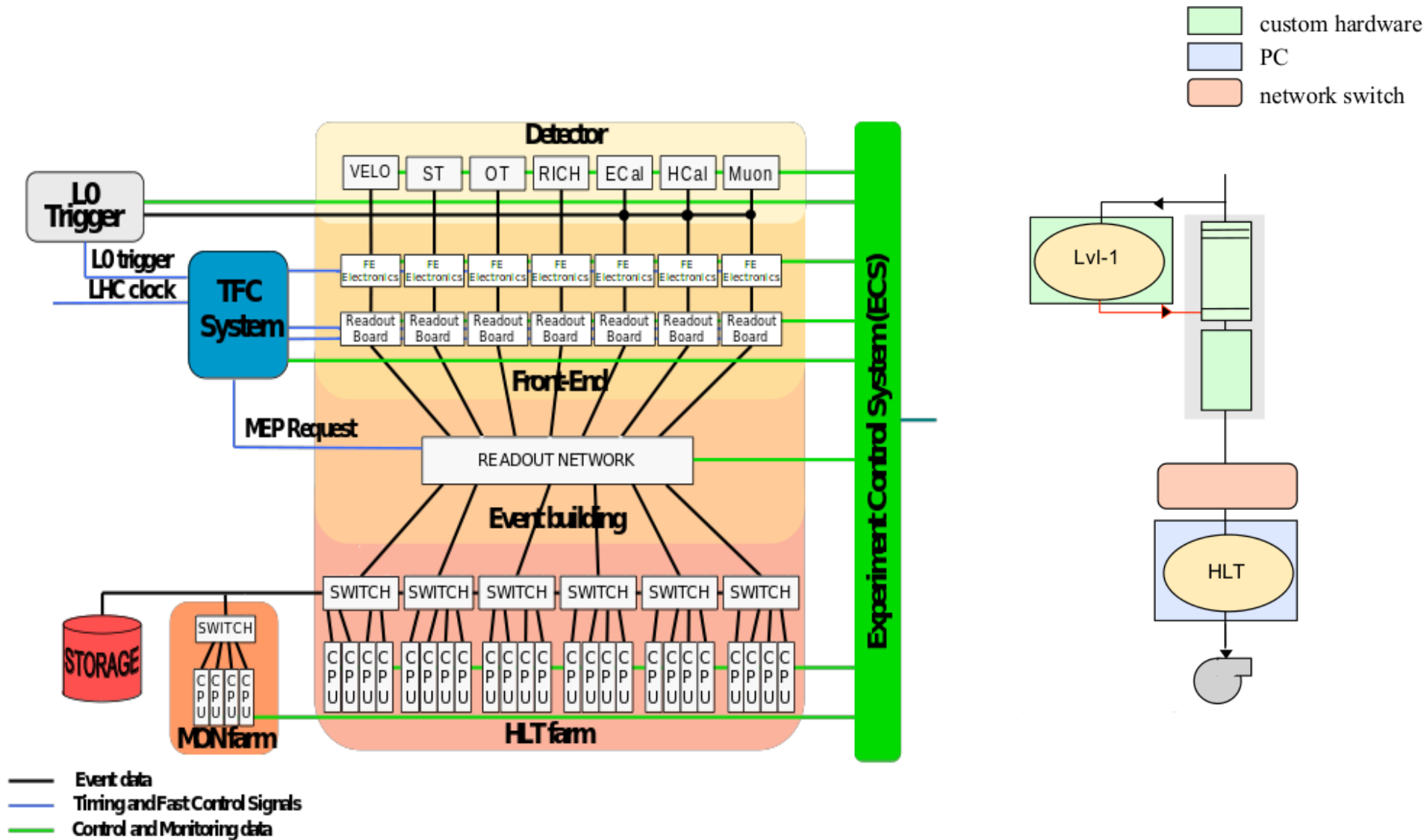
- custom hardware
- PC
- network switch



- custom hardware
- PC
- network switch



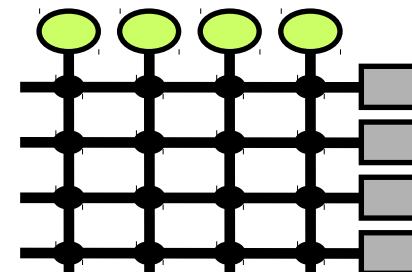






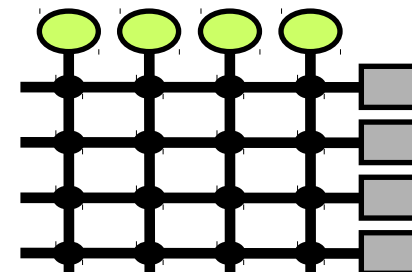
Networking

- ➔ Examples: Ethernet, Telephone, Infiniband, ...
- ➔ All devices are **equal**
- ➔ Devices **communicate directly** with each other
 - no arbitration, simultaneous communications
- ➔ Device communicate by sending messages
- ➔ In switched network, **switches** move messages between sources and destinations
 - find the right path
 - handle “congestion” (two messages with the same destination at the same time)



- Examples: Ethernet, Telephone, Infiniband, ...
- All devices are **equal**
- Devices **communicate directly** with each other
 - no arbitration, simultaneous communications
- Device communicate by sending messages
- In switched network, **switches** move messages between sources and destinations
 - find the right path
 - handle “congestion” (two messages with the same destination at the same time)

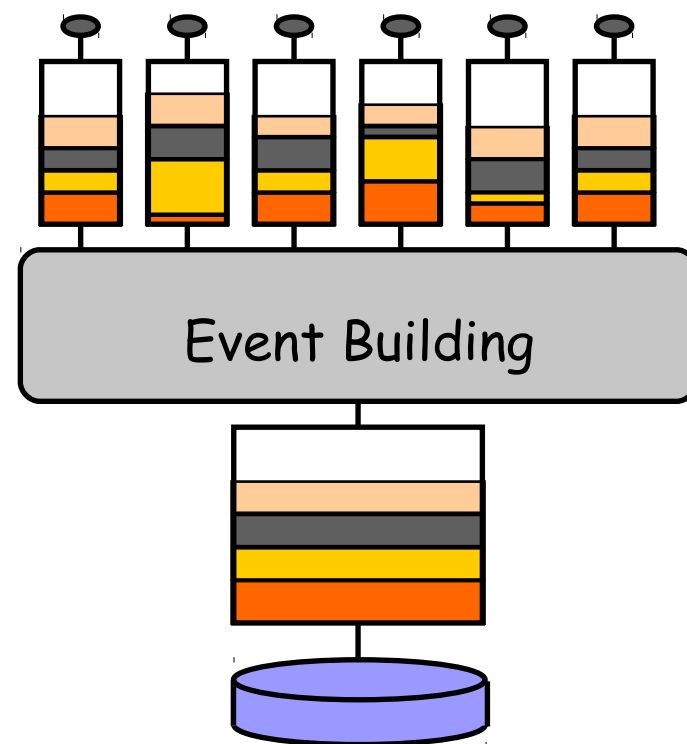
Thanks to these characteristics, **networks do scale** well. They are the backbones of LHC DAQ systems



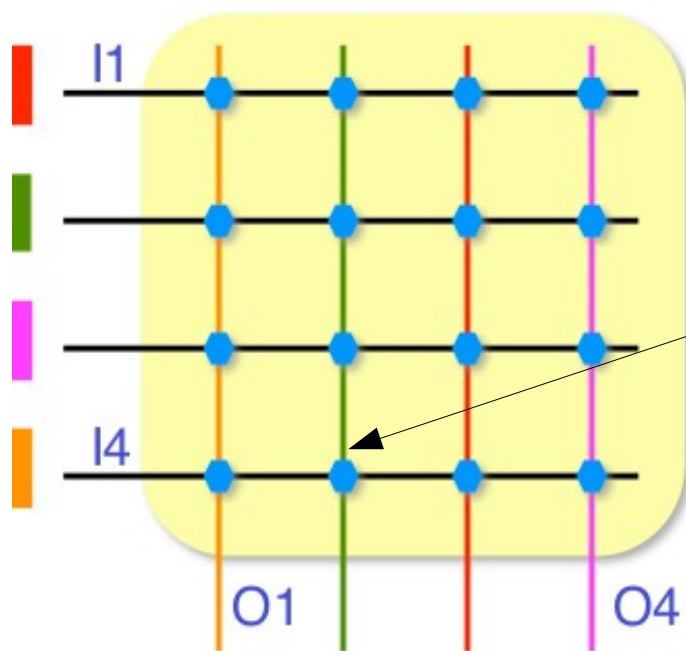
→ Event Building: collection and formatting of all the data elements of an event into a single unit

- normally last step before high-level trigger or storage
- can be implemented on buses, can use custom interconnects, can be based on (Ethernet) **network**

→ Network-based EB is choice of all LHC experiments and a case study for networking in DAQ

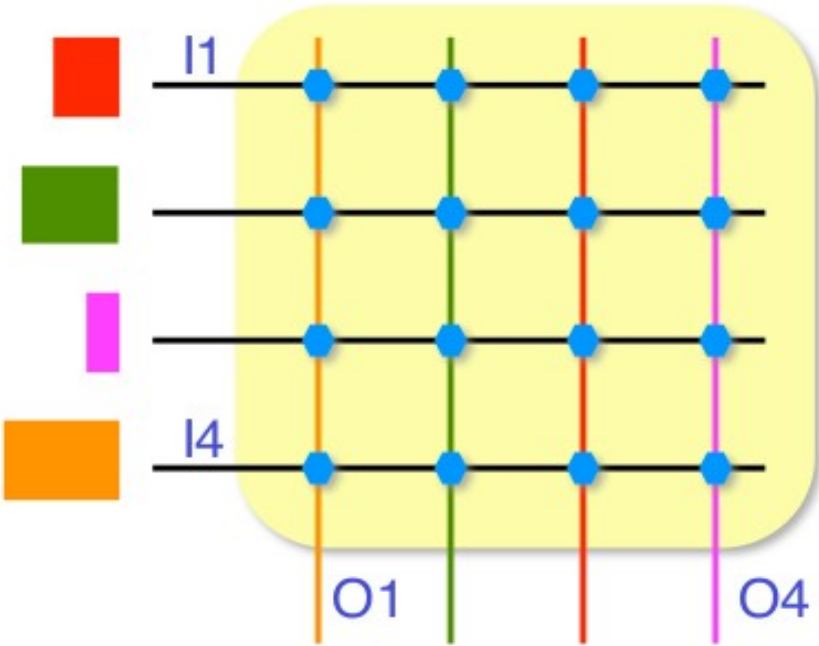


Network switch: crossbar

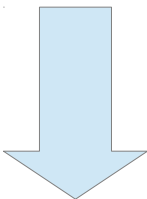


- ➔ Each input port can potentially be connected to each output port
- ➔ At any given time, only one input port can be connected to a given output port
- ➔ Different output ports can be reached concurrently by different input ports

Network switch: crossbar



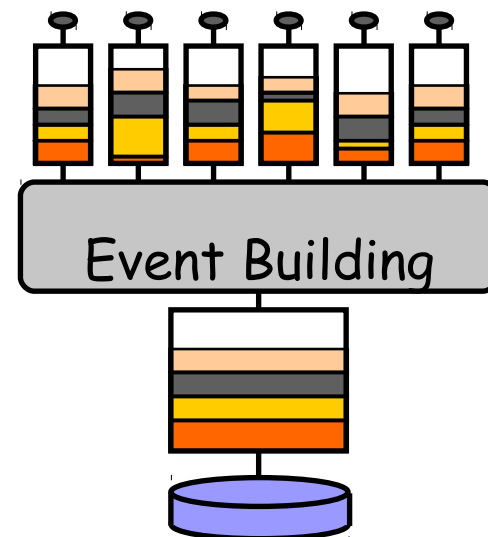
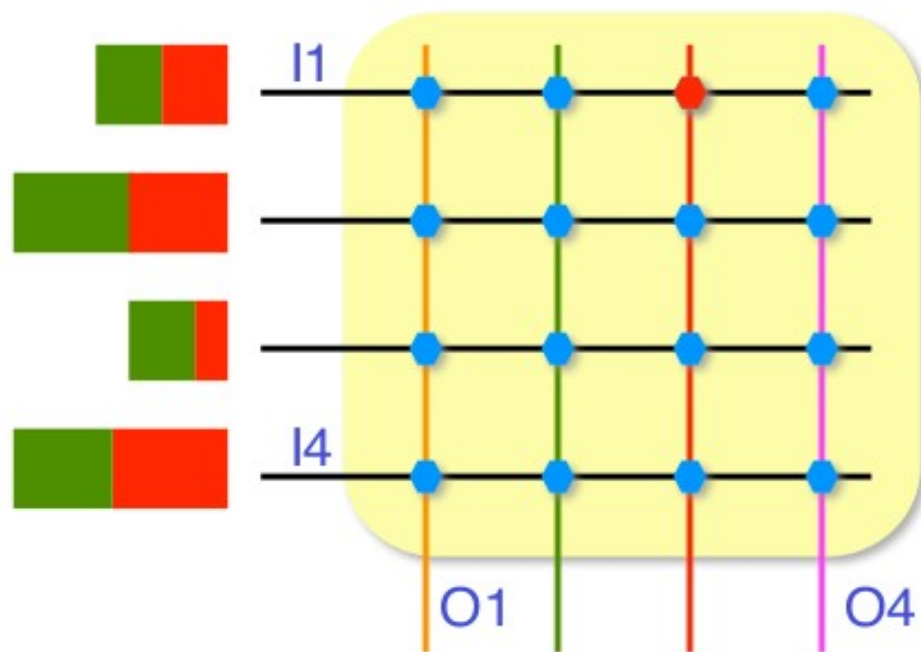
→ Ideal situation → all inputs send data to different outputs



No interference (Congestion)

All input ports send data concurrently

Crossbar switch: event building



→ EB workload implies converging data flow

- all inputs want to send to same destination **at the same time**

→ "Head of line blocking"

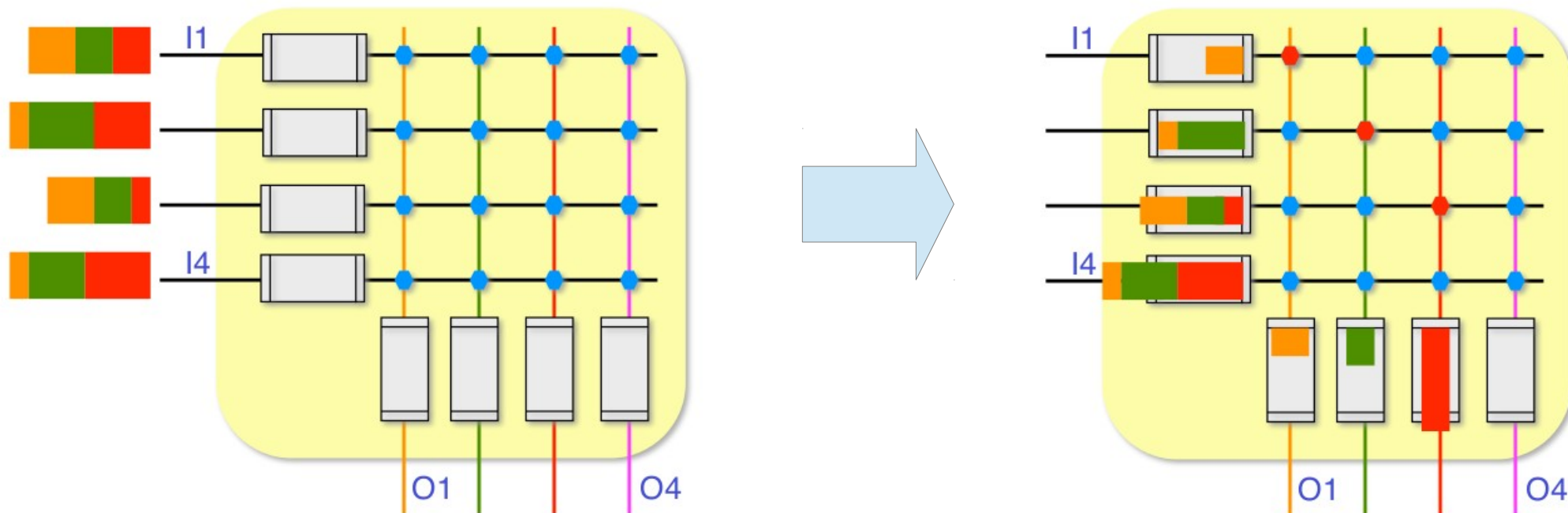
- congestion



→ Well know phenomena ..

→ Differently from road traffic, Ethernet HW is allow to “drop” packets

- Higher level protocols have to take care of re-sending
- Possibly important performance impacts



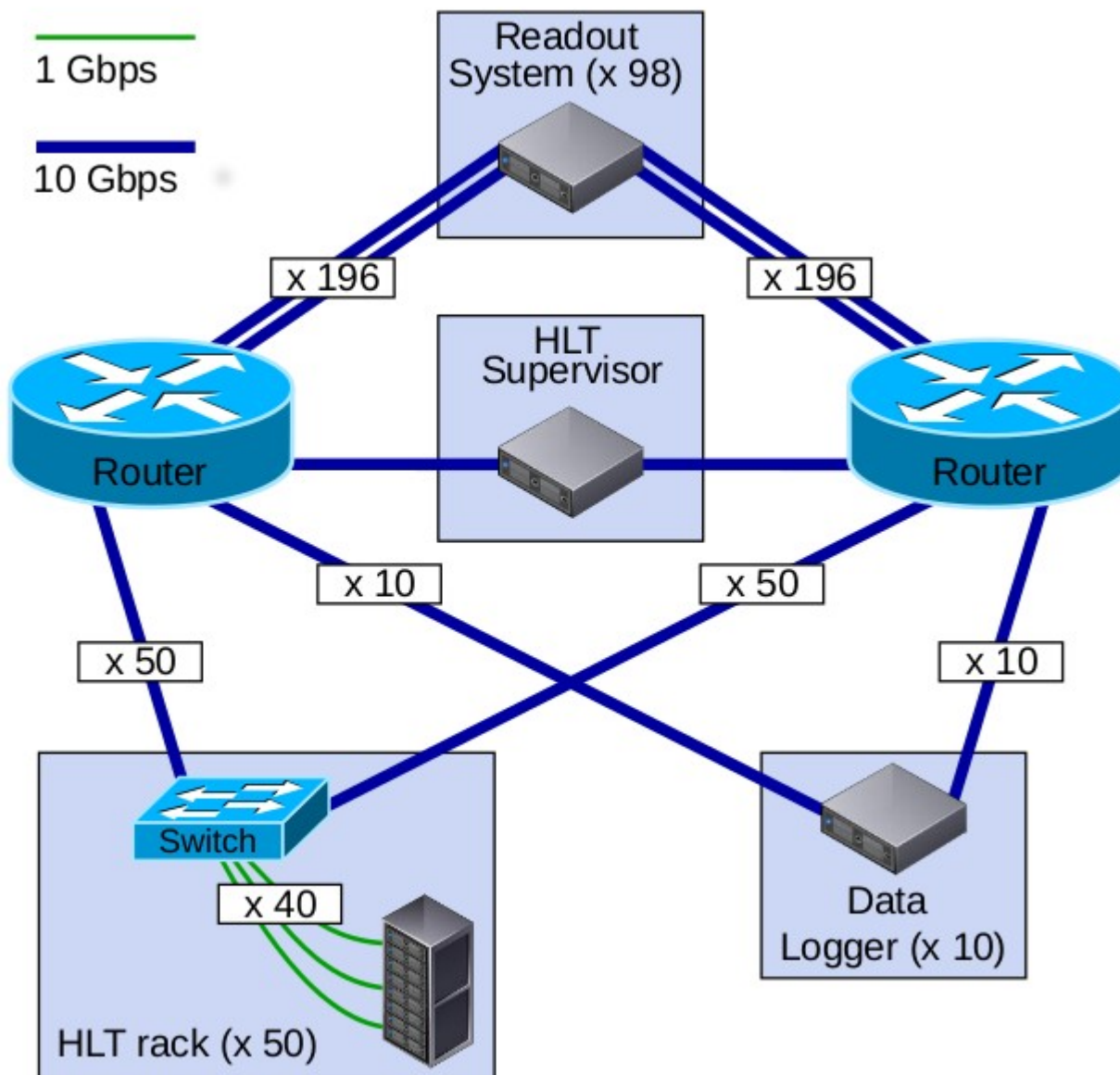
➔ Adding input and output FIFO dramatically improve the EB pattern handling

➔ EB workload anyway problematic

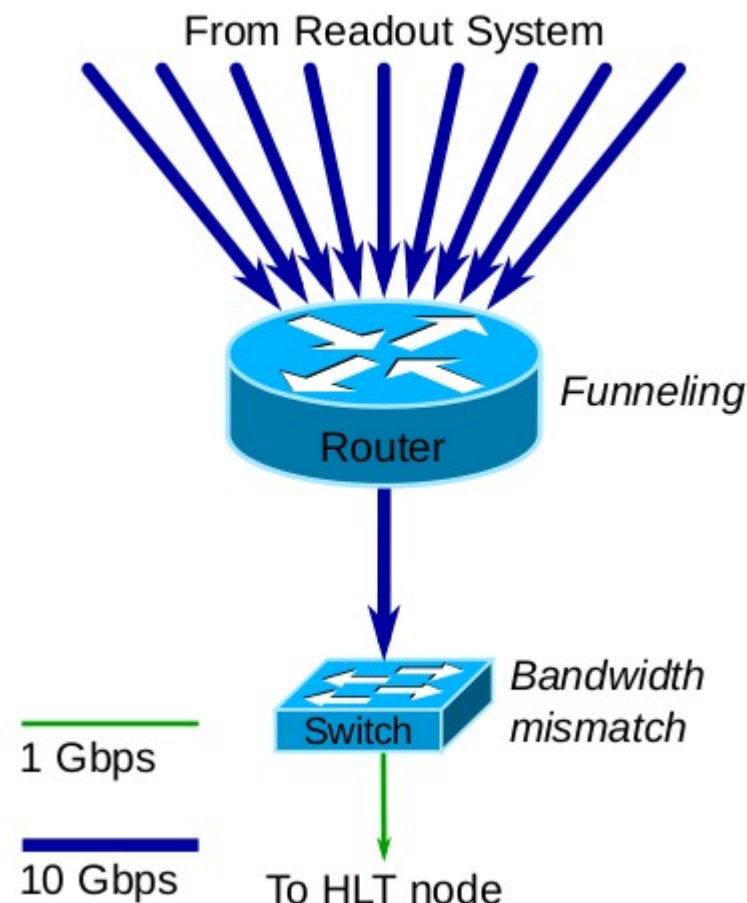
- FIFO size is limited, variable data size
- limited internal switching speed

Traffic shaping
or
Network over-sizing

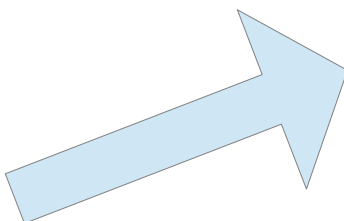
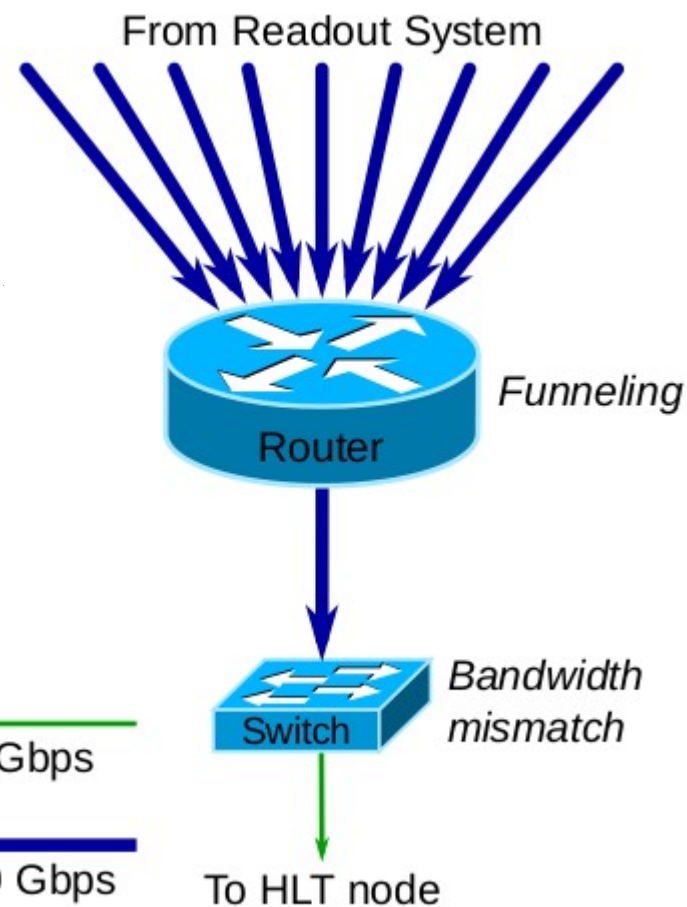
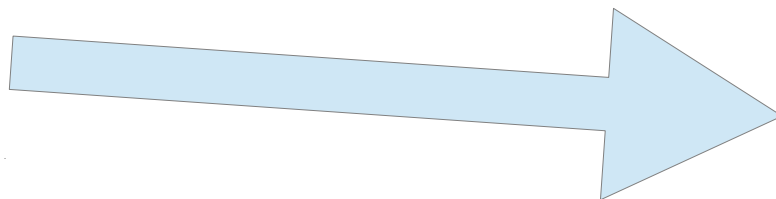
ATLAS Data Network Topology



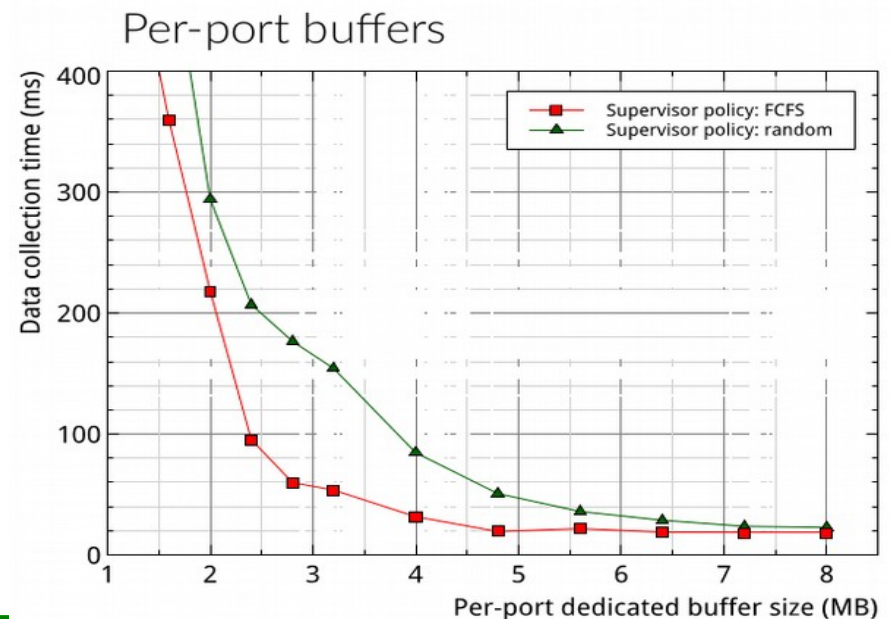
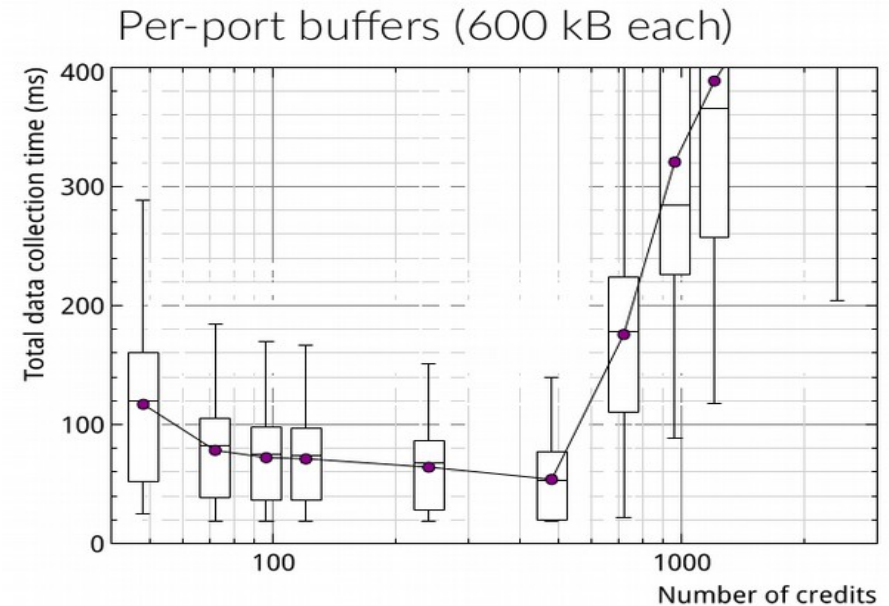
- ➔ Present ATLAS network topology two possible sources of congestion
- ➔ Funnelling: multiple links may overload output link
 - Brute force → central routers are large (and expensive) Carrier-class Internet-scale devices with massive buffering and switching capabilities
- ➔ Bandwidth mismatch: a faster input link may instantaneously overload a slower output link
 - Traffic shaping → control maximum burst size with respect to the switch buffer size



Event Building in ATLAS



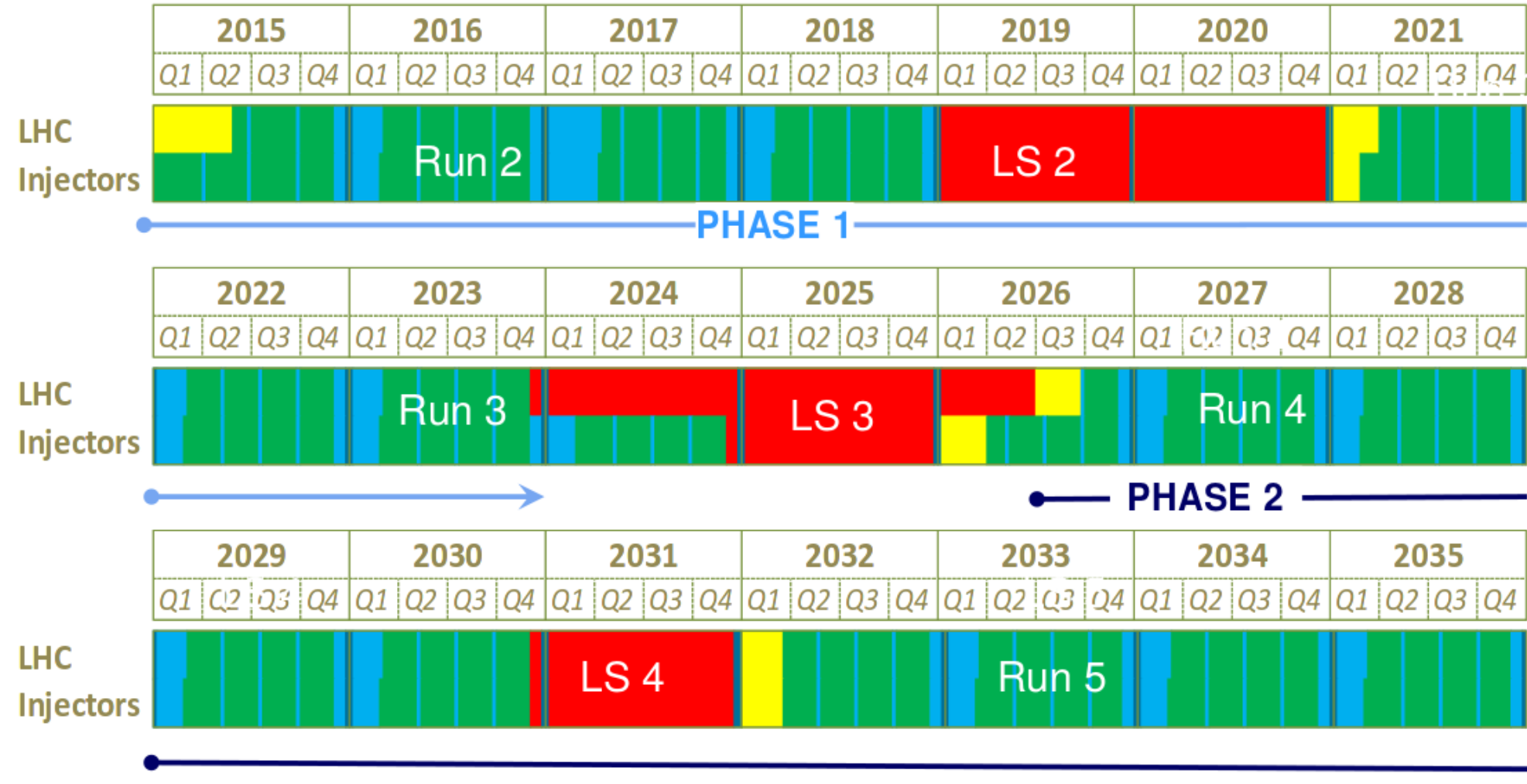
- ➔ Credit mechanism at application level to control the burst size
 - more credits → more concurrent responses
- ➔ Quality metric is collection time → time to fetch data
- ➔ If one credit corresponds to 1 kB response size
 - above the switch buffer size, packet drops happen
- ➔ Interesting to study what buffer size would allow no traffic shaping
 - simulations calibrated to reproduce the above measurements

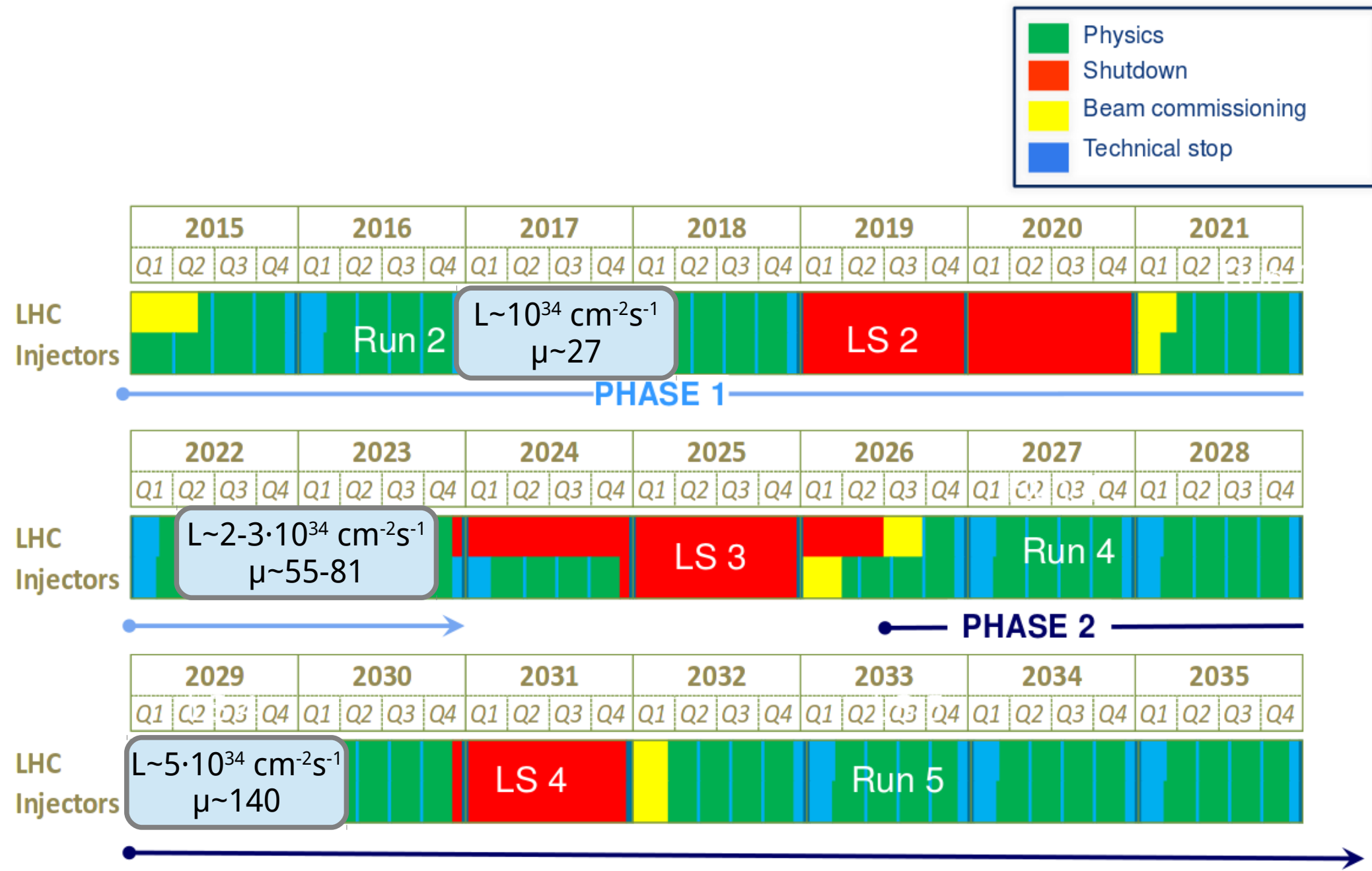




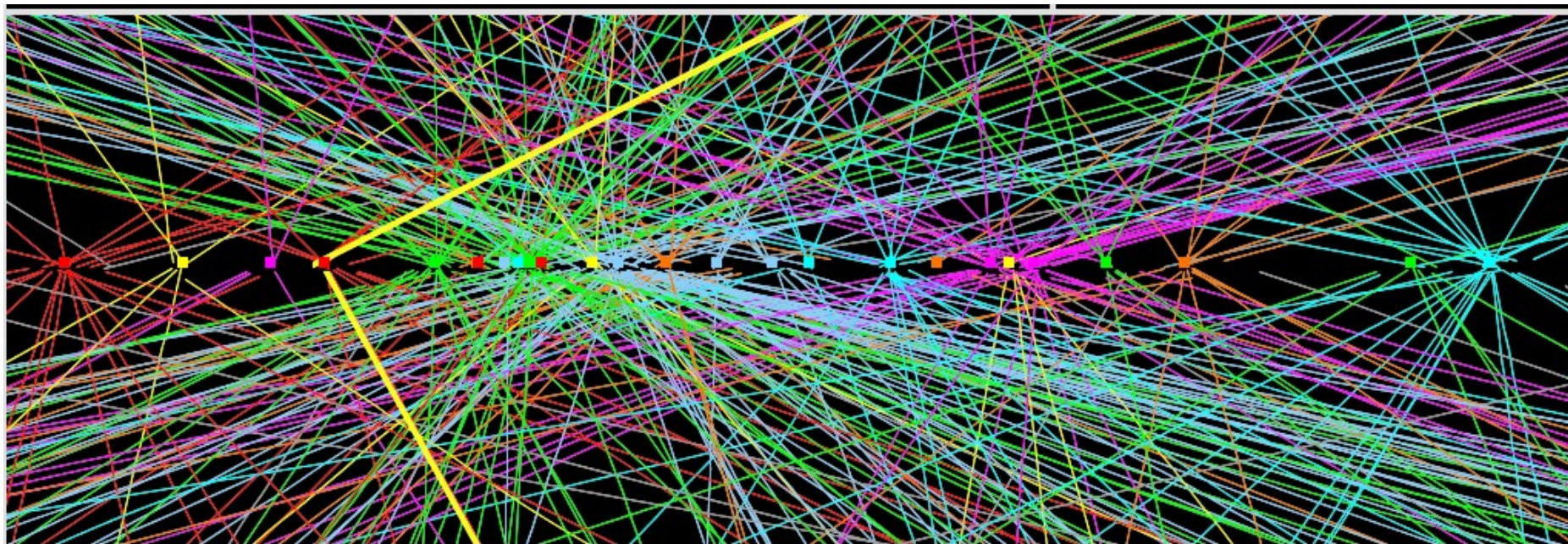
LHC Upgrades

LHC Upgrade Programme





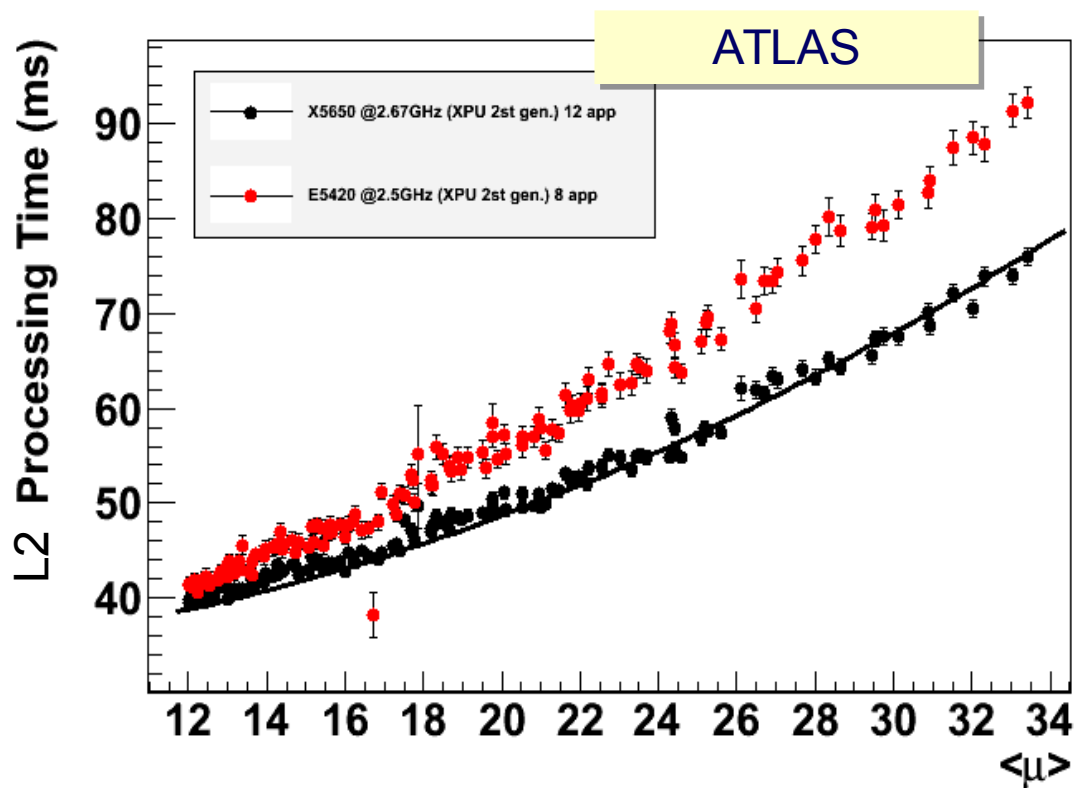
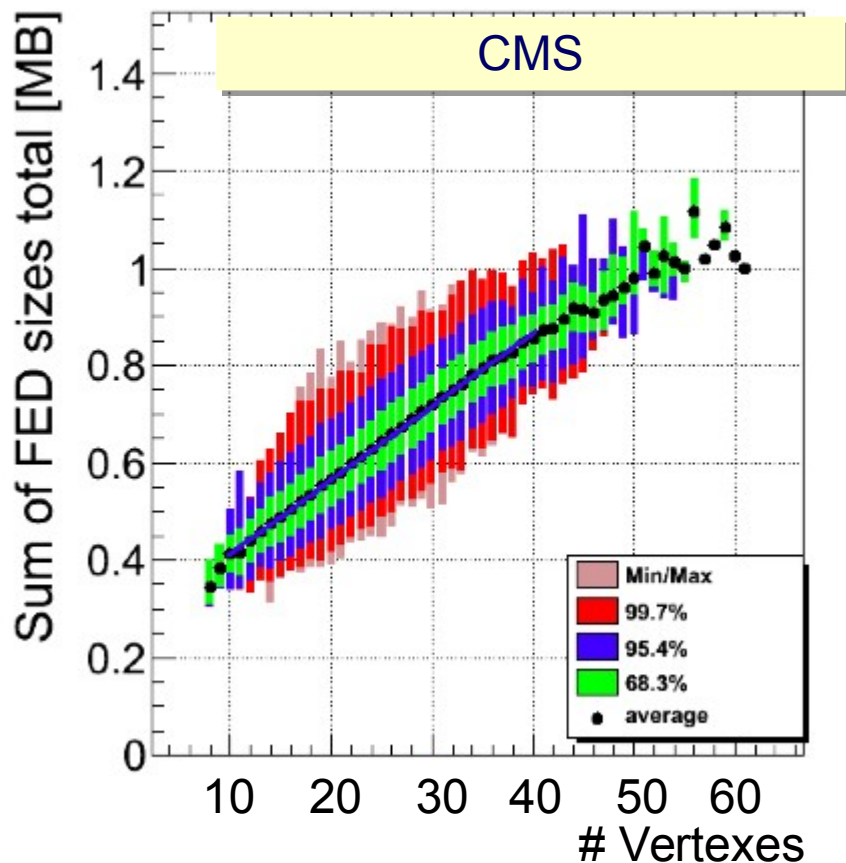
“Pile up”: collision multiplicity



- ➔ One “event” in LHC is the superimposition of many, almost concurrent, proton-proton collisions
 - pile up is the number of overlapping collisions in one event
- ➔ LHC upgrade programme increases the “brightness” of the accelerator increasing the pile-up
 - faster statistic collection, BUT non-linear increase in event complexity

→ Increased pile-up

- larger data size → bandwidth and storage
- more complex events → increased computing needs, reduced trigger efficiency and rejection power and increased acceptance rates



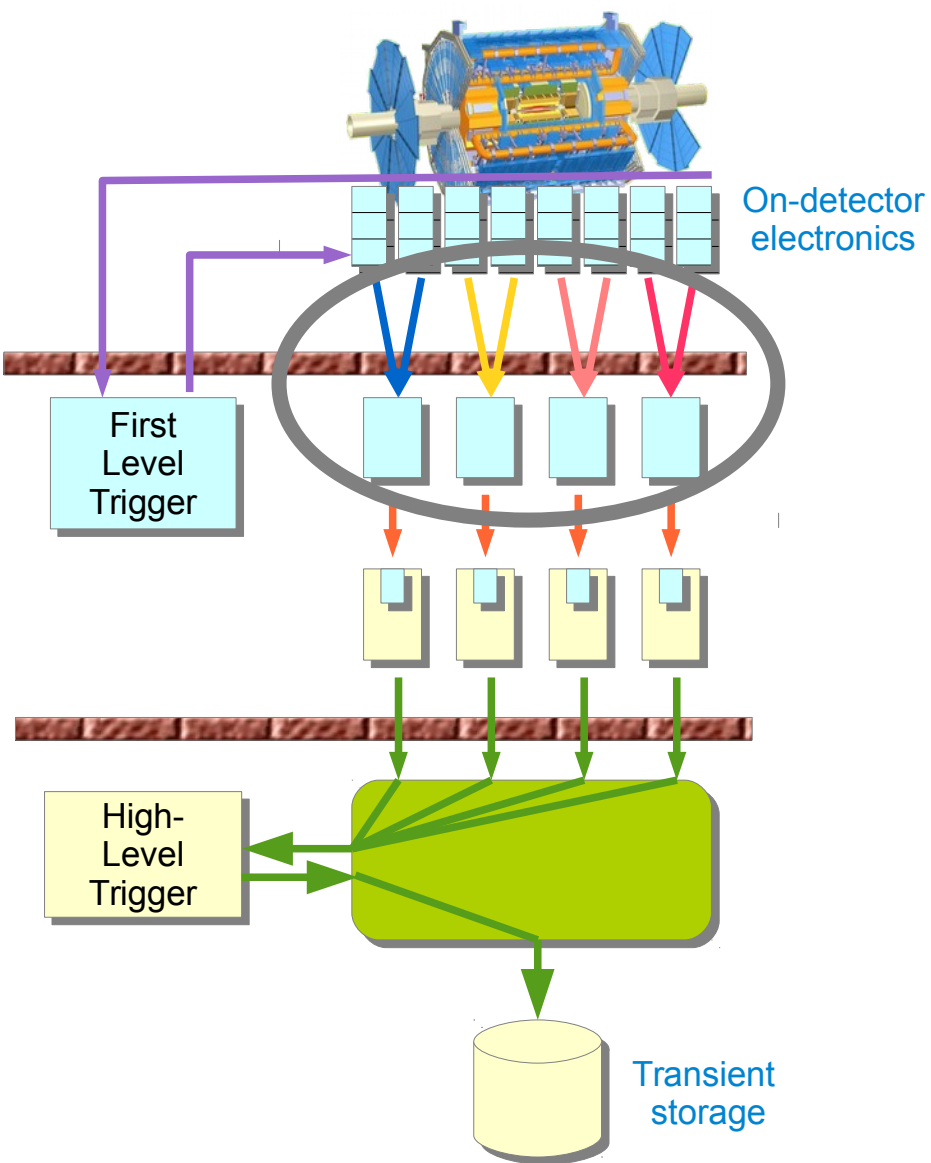
HL-LHC DAQ Matrix

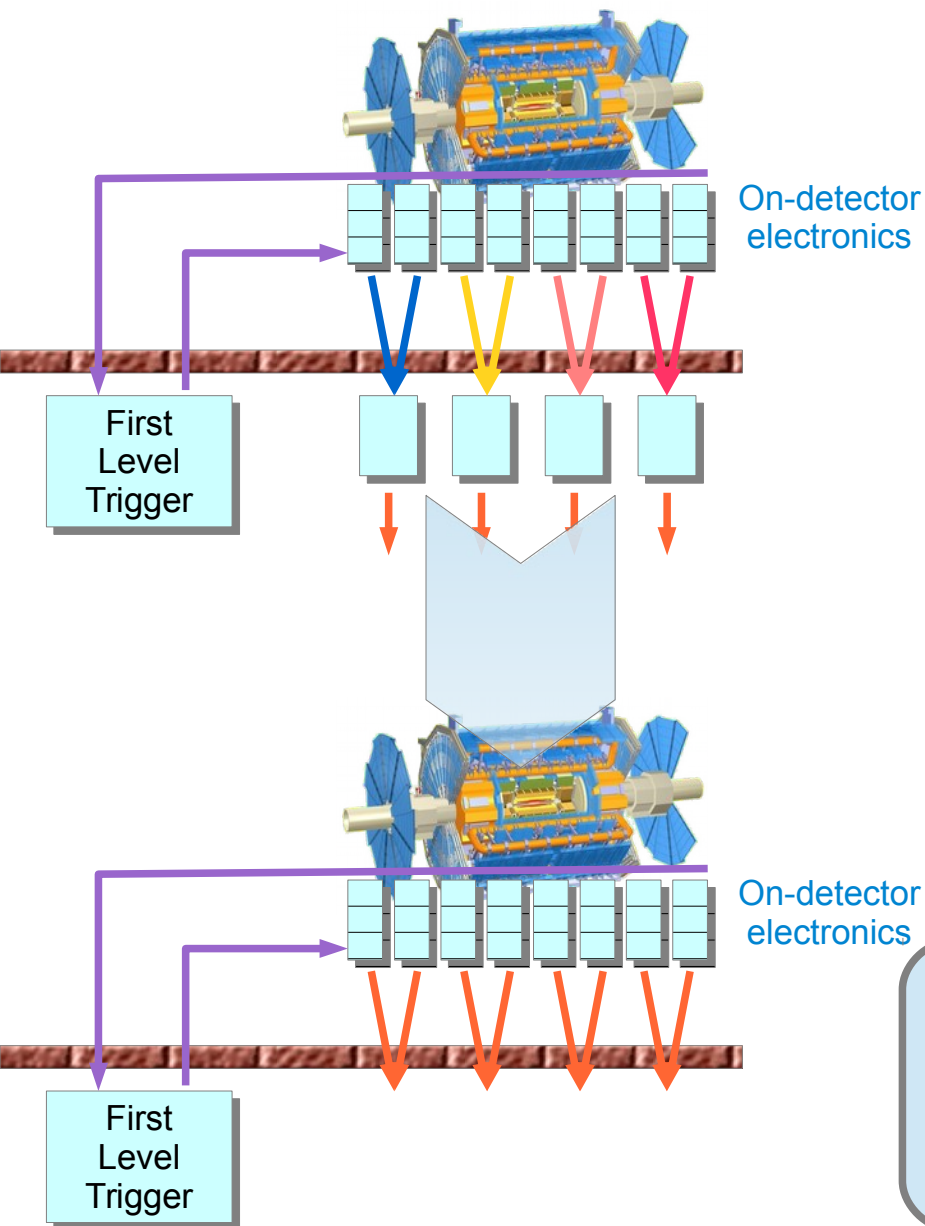
	ALICE	LHCb	CMS	ATLAS
Hardware trigger	No	No	Yes	Yes
Software trigger input rate	50 kHz Pb-Pb 200 kHz p-Pb	30 MHz	500/750 kHz for PU 140/200	0.4 MHz
Baseline processing architecture	CPU/GPU/FPGA/ Cloud&Grid	CPU farm (+coprocessors)	CPU farm (+coprocessors)	CPU farm (+coprocessors)
Software trigger output rate	50 kHz Pb-Pb 200 kHz p-Pb	20-100 kHz	5-7.5 kHz	5-10 kHz

→ Back-end electronics function is to adapt between two class of serial links

- specialized detector links
- common readout links

→ No main DAQ functionalities, just technology proxying





→ Common detector link technology would allow higher commonality downstream

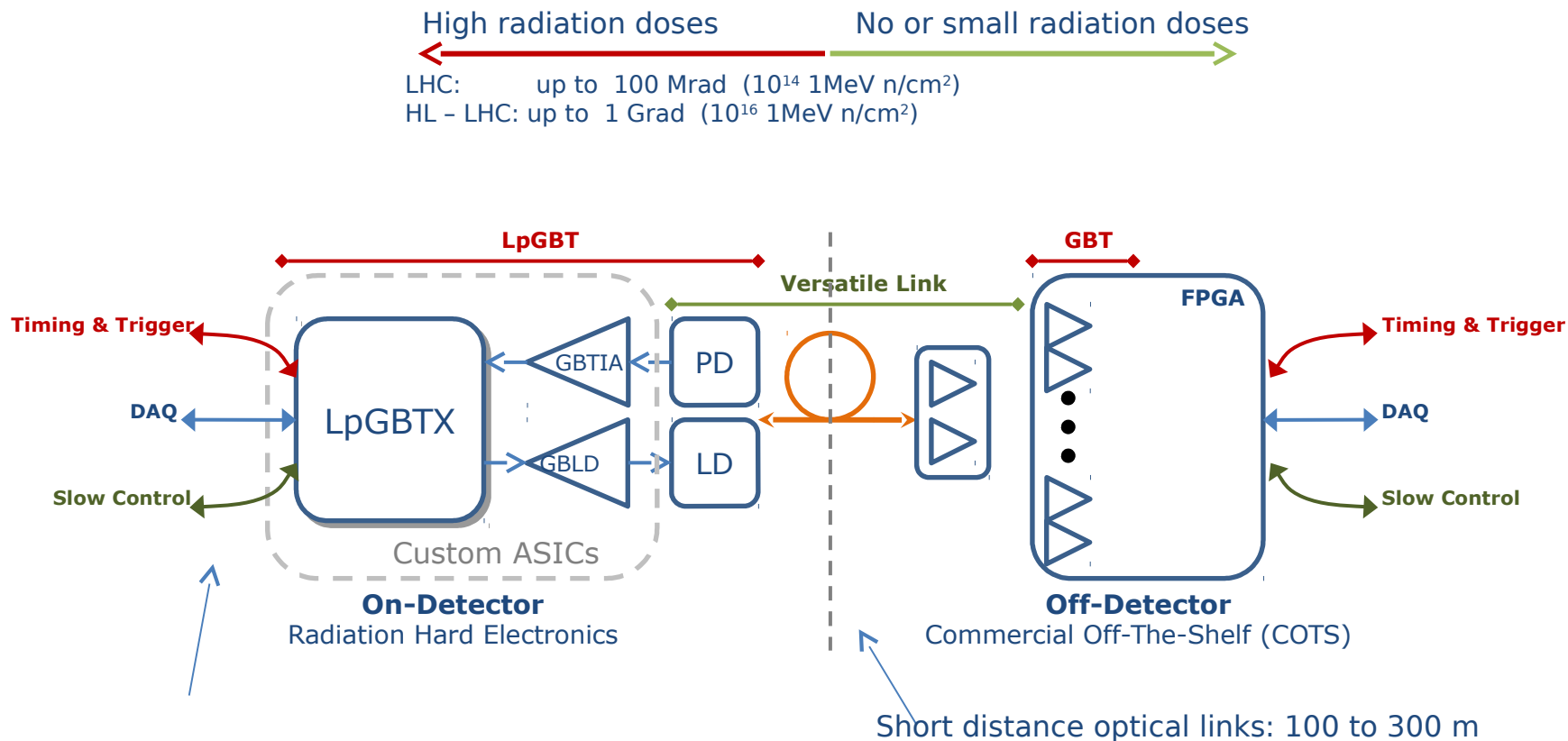
→ Must meet requirements of all detectors

- from radiation hardness to power dissipation and cost

→ HL-LHC will require replacement of most detector links for bandwidth reason

Clear occasion for a paradigm change →
Dedicated CERN projects to develop the
needed technology
GigaBit Transceiver (GBT) & Versatile Link
(VL)

GBT & VL



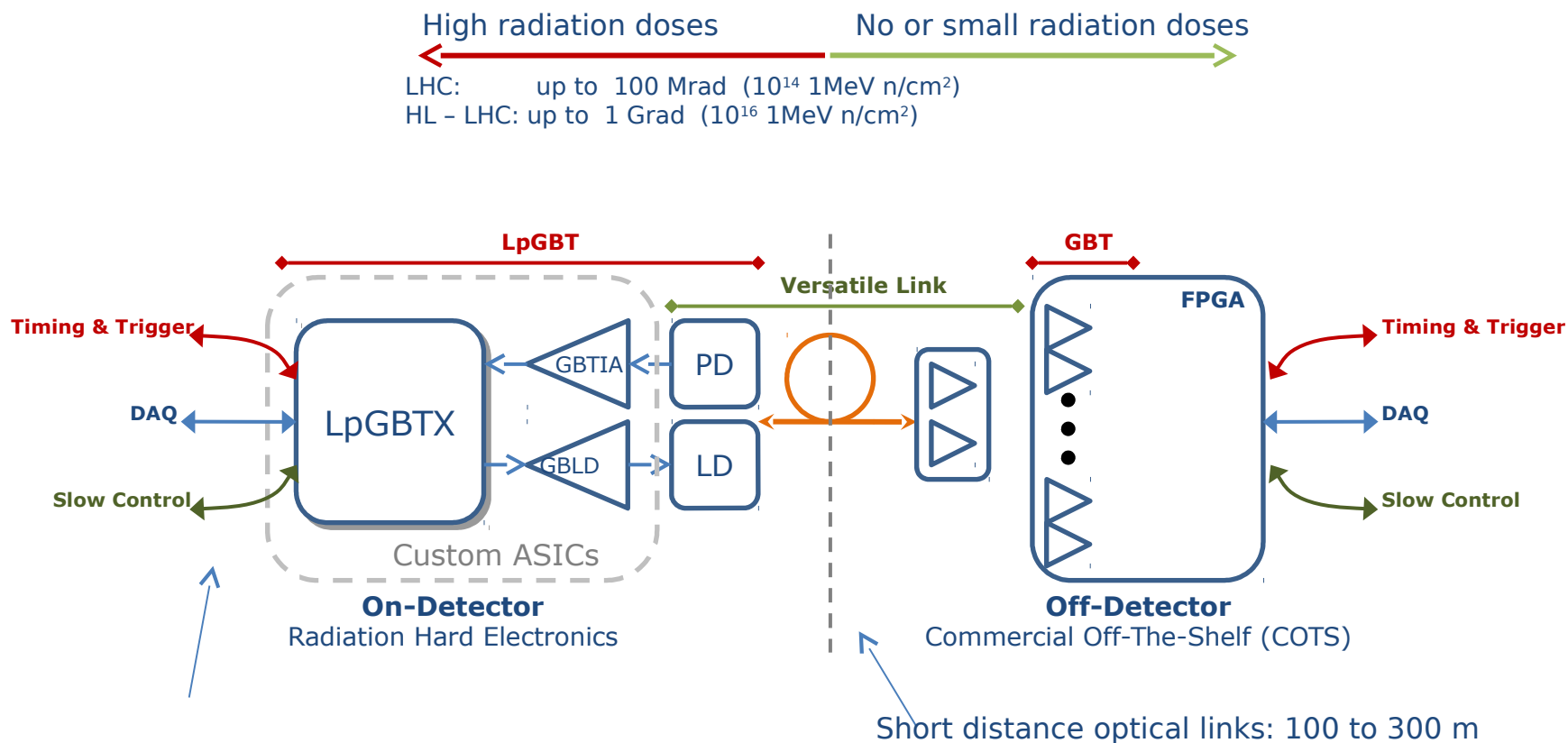
→ Asymmetric technology

- radiation hard components (ASIC) on the detector, COTS (FPGA) on the receiving end
- 500 mW power consumption (on-detector)
- up to 10.24 Gb/s uplink – 2.56 Gb/s downlink

→ Logical, fixed bandwidth sharing per link allows multiplexing of different information

- unique detector interface: data, configuration, trigger, slow control ...

GBT & VL



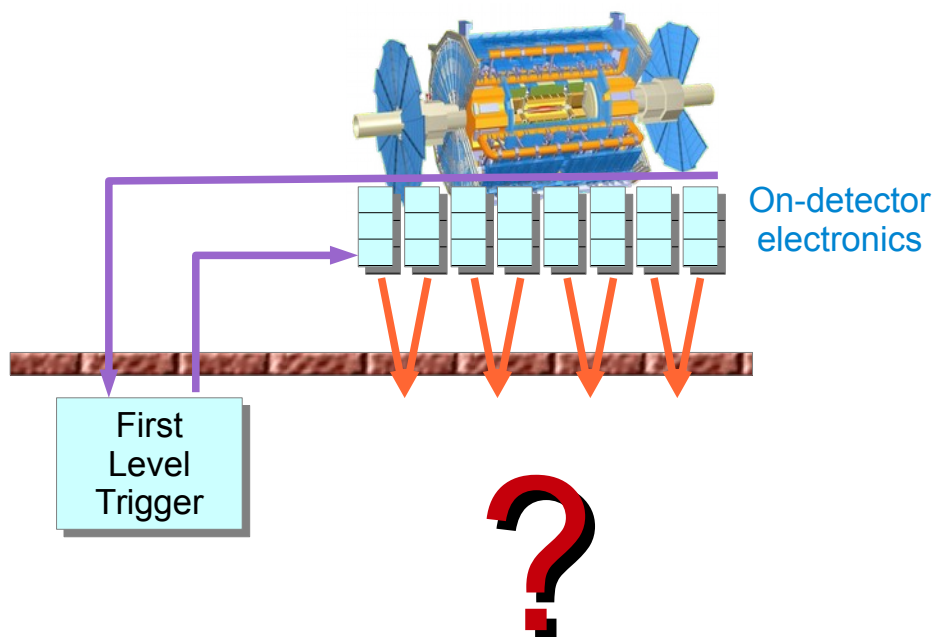
→ Asymmetric technology

- radiation hard components (ASIC) on the detector, COTS (FPGA) on the receiving end
- 500 mW power consumption (on-detector)
- up to 10.24 Gb/s uplink – 2.56 Gb/s downlink

→ Logical, fixed bandwidth sharing per link allows

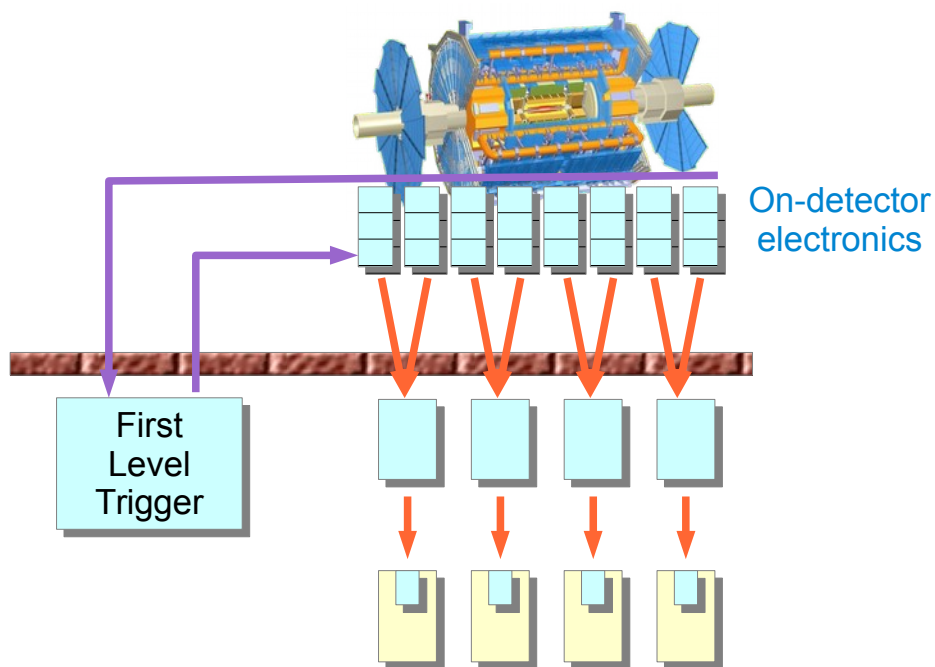
- unique detector interface: data, configuration, trigger

GBT deployment planned for all LHC experiments



→ Detector is completely interfaced by common technology

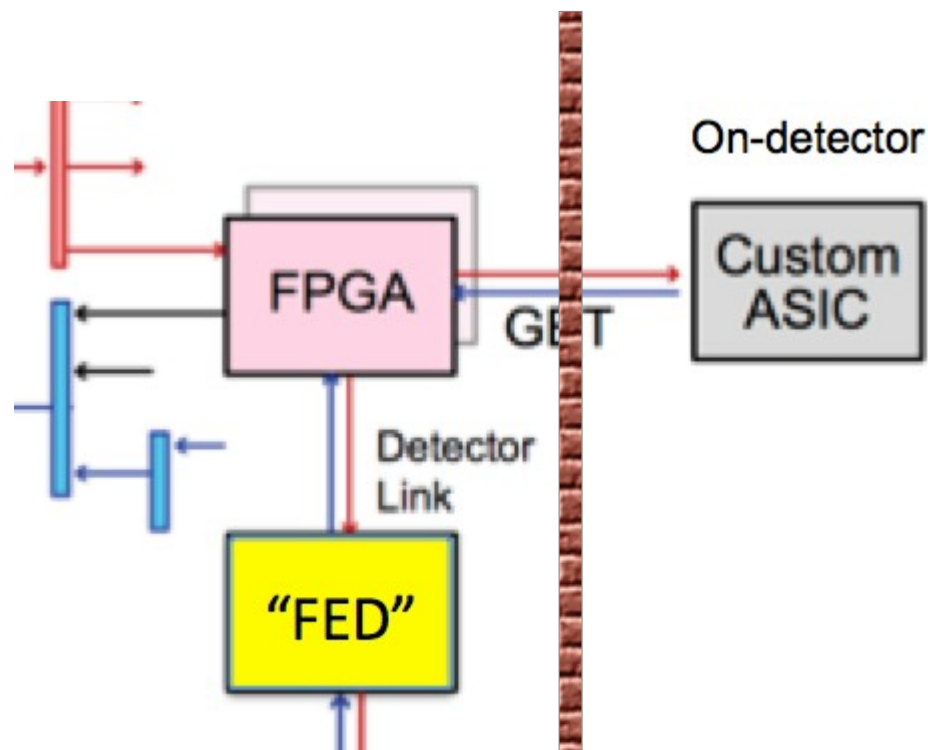
→ How to read it out?

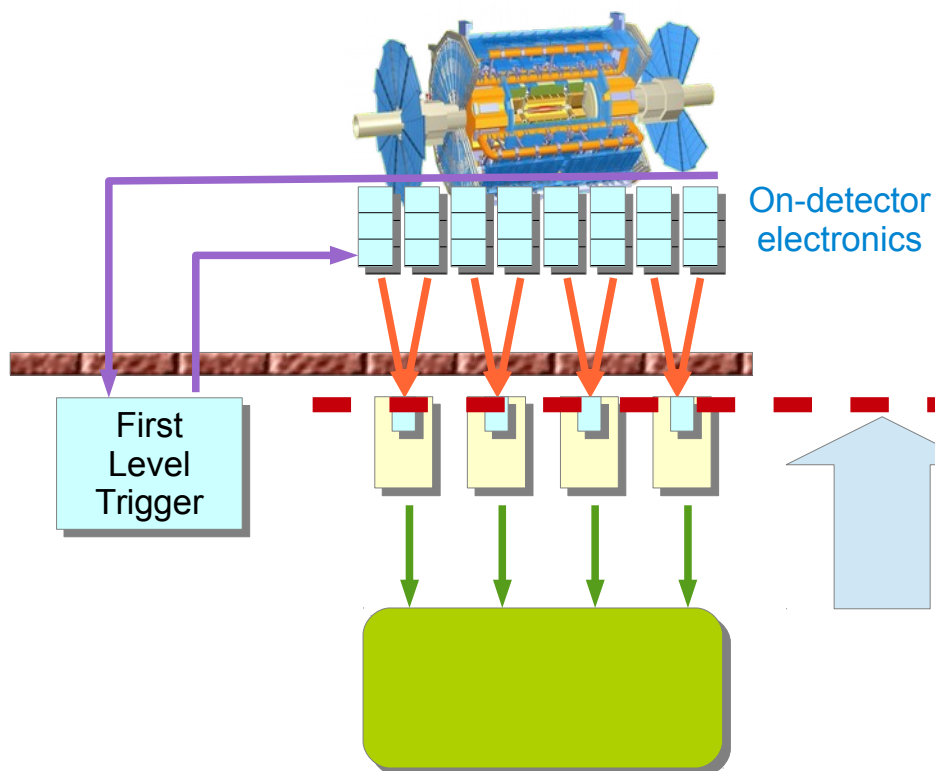


→ Detector is completely interfaced by common technology

→ Keep same scheme two-link scheme

- now with common, still custom, back-end electronics





→ Detector is completely interfaced by common technology

→ Push COTS domain closer to the detector

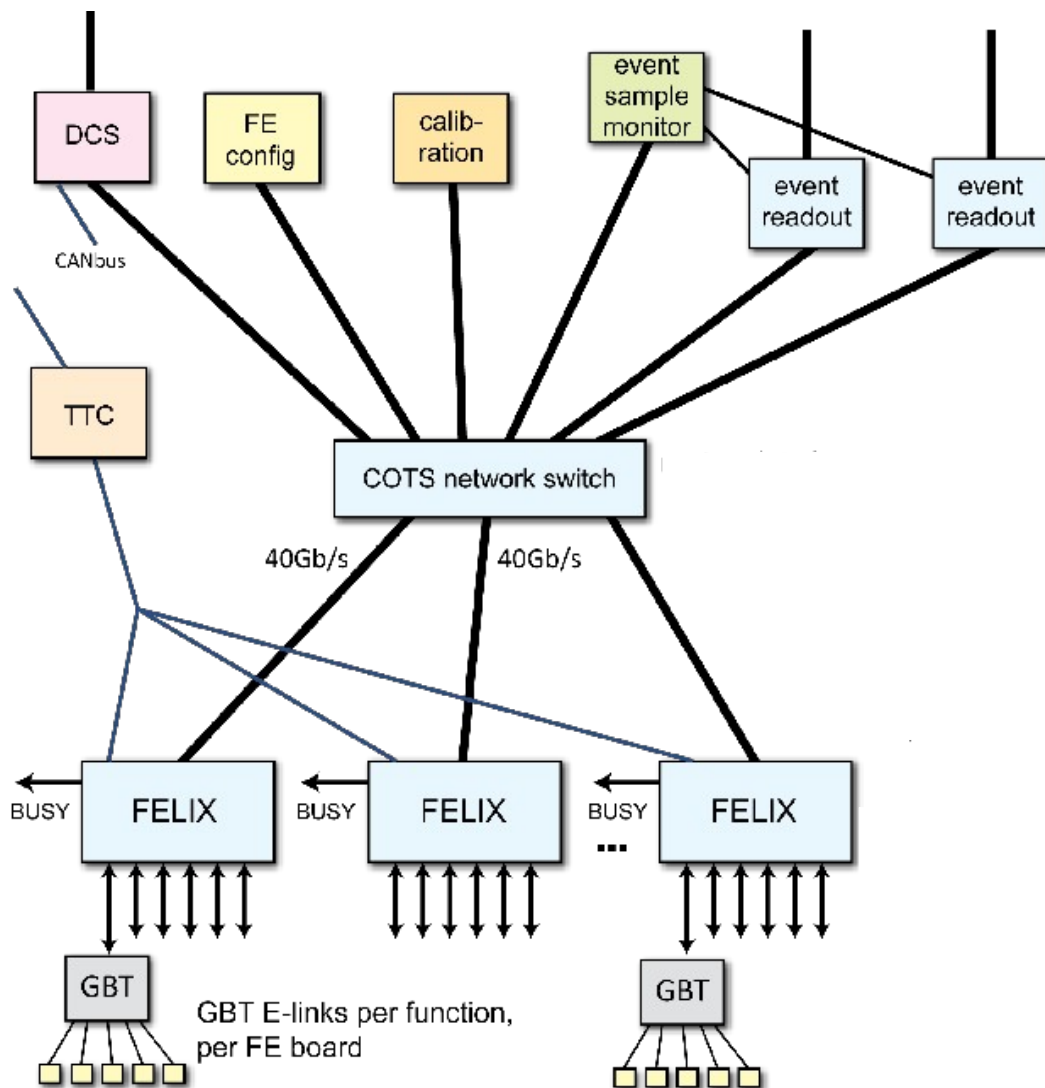
→ Interfacing detector serial links directly a switched network allows to

- move many functionalities in software
- reduce single point of failures
- simpler load balancing and maintenance

→ LHC experiments ~10000 GBT links

- need high density heterogeneous data router

Front-end Link Exchange (FELIX)



→ ATLAS project for interfacing GBT links

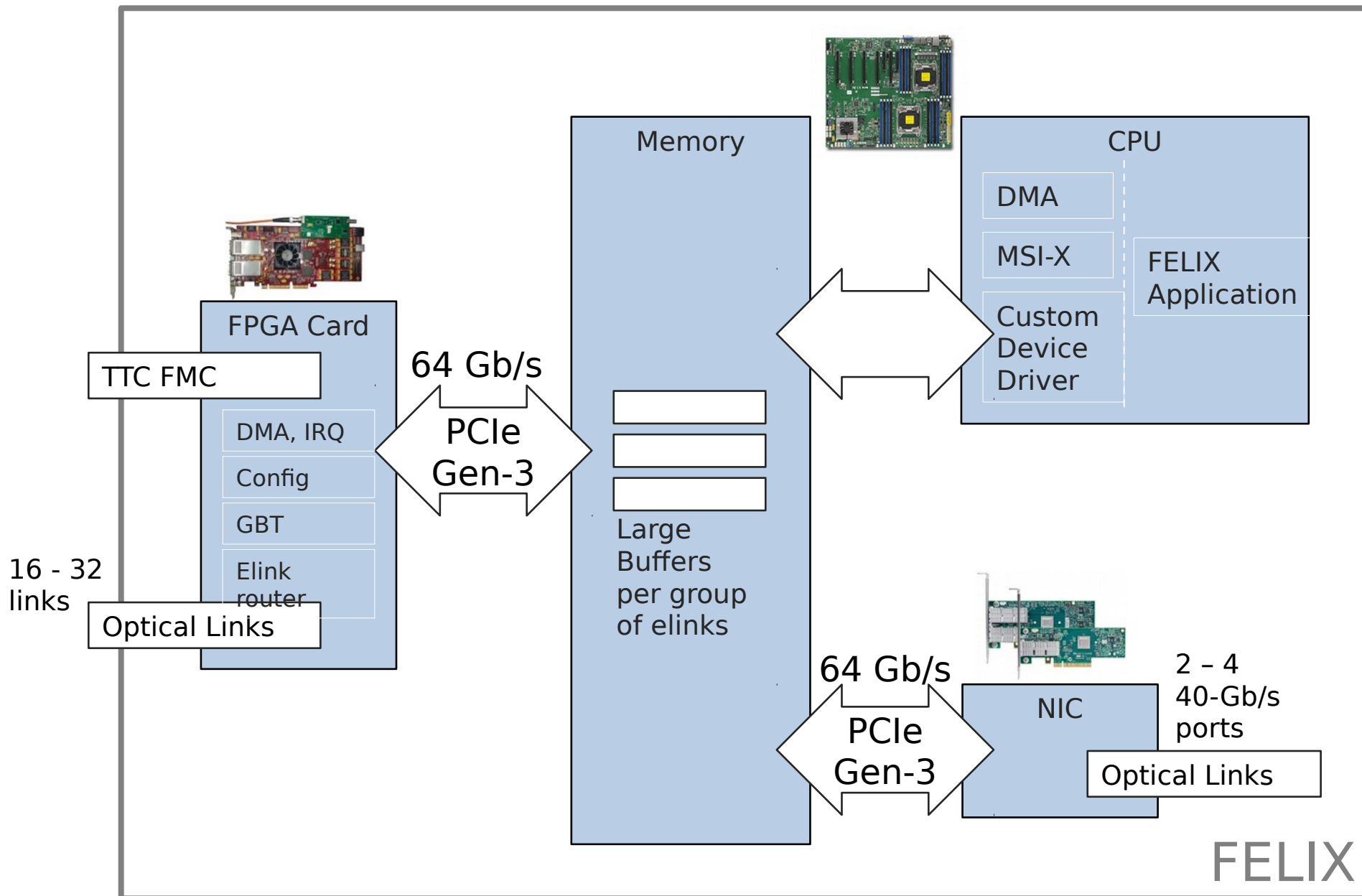
→ **Stateless, configurable** data routing device

- route data by **streams or type**
- propagate commands
- data duplication and sampling for monitoring

→ Handling of high-level switched protocol

- Infiniband/Ethernet/...
- QoS for different traffic types

FELIX Prototype



→ HL-LHC DAQ system will need network throughput of ~20 Tb/s

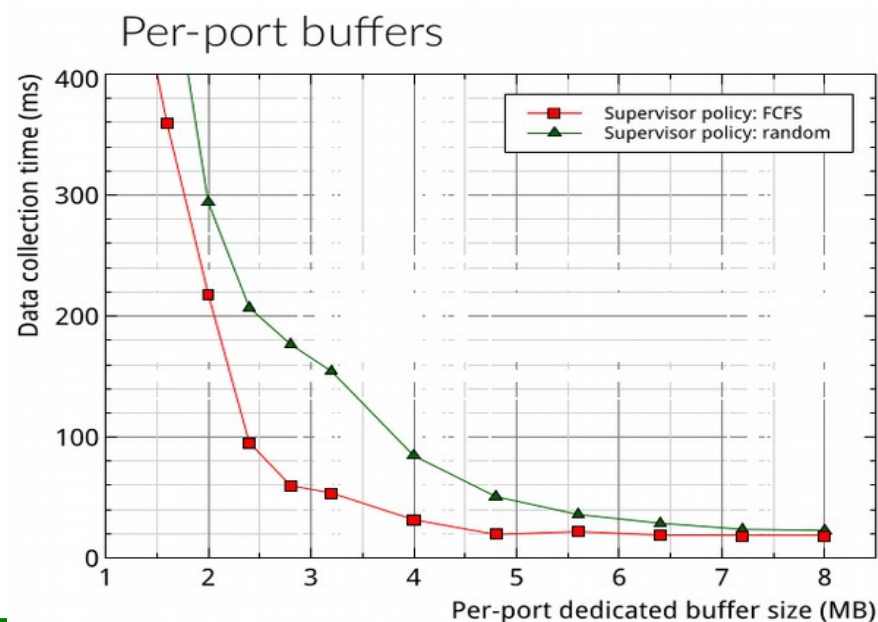
- currently ~1 Tb/s

→ Ethernet is the current champion in networking

- 100GbE available, 400 GbE in preparation
- 20 Tb/s → 200 100GbE, 50 400 GbE

→ Event Building pattern is not easy on network

→ Ideally want deep buffers to avoid application level solutions



→ Currently two major classes of Ethernet network devices

→ Carrier

- deep-buffer, high-density, flexible firmware (FPGA, network processors), \$\$\$/port

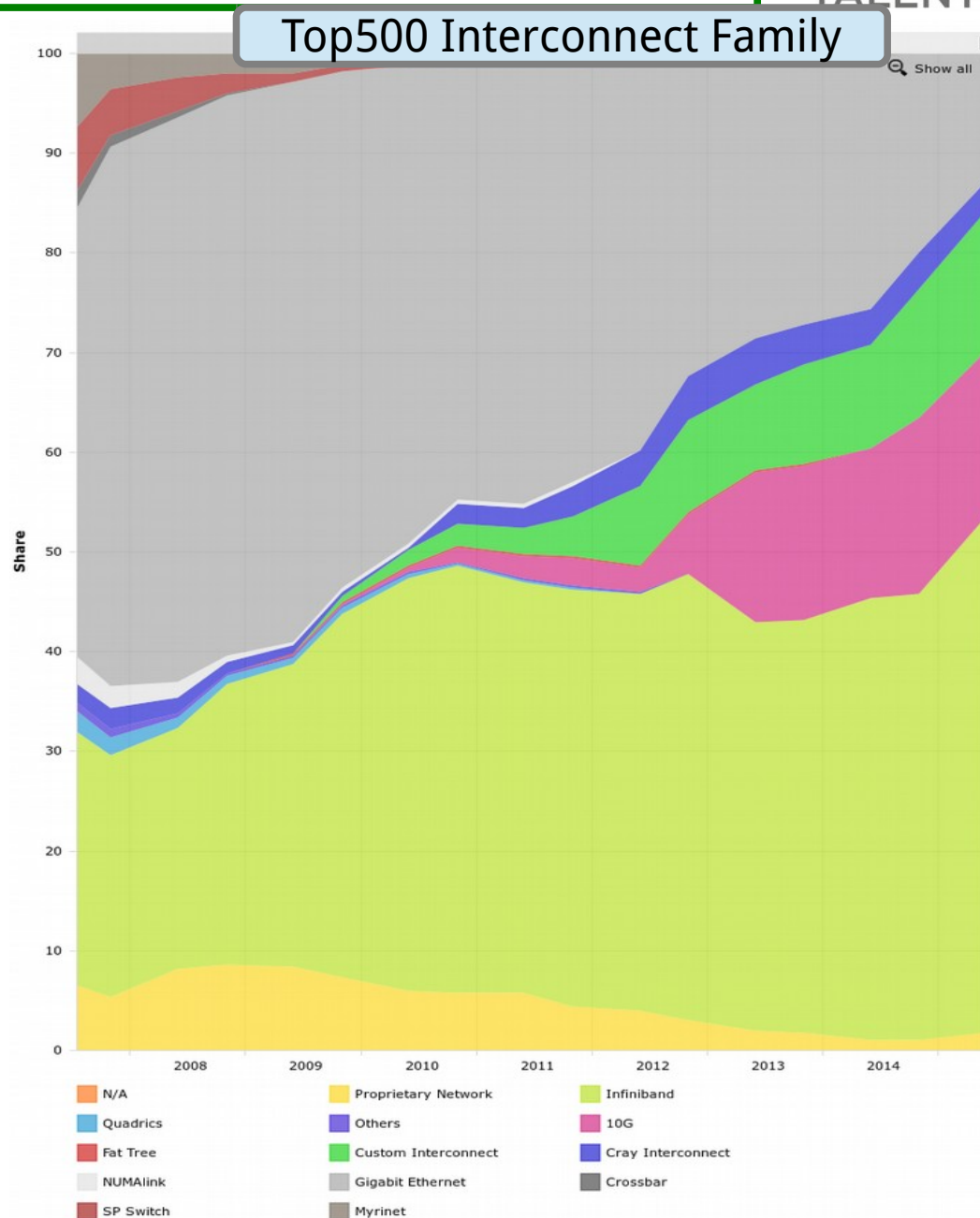
→ Data-centre (Top-of-Rack TOR)

- shallow buffer, ASIC based, ultra-high density, focused on layer 2 and simple layer 3 features, very low latency, \$/port

Speed	Carrier [USD / port]	TOR [USD / port]
10GbE	400 - 1000	200 - 250
40GbE	1000 - 4000	500 - 900

Other network technologies

- ➔ Infiniband is leader in HPC and Top500
- ➔ Loss-less network using specific hardware and software stack
 - Single-vendor
- ➔ Data-Center Bridging (DCB)
 - aka loss-less Ethernet
- ➔ Umbrella for a zoo of IEEE standards
- ➔ Looking forward to official release of new Intel interconnect
 - Omni-Path, expected end of 2015

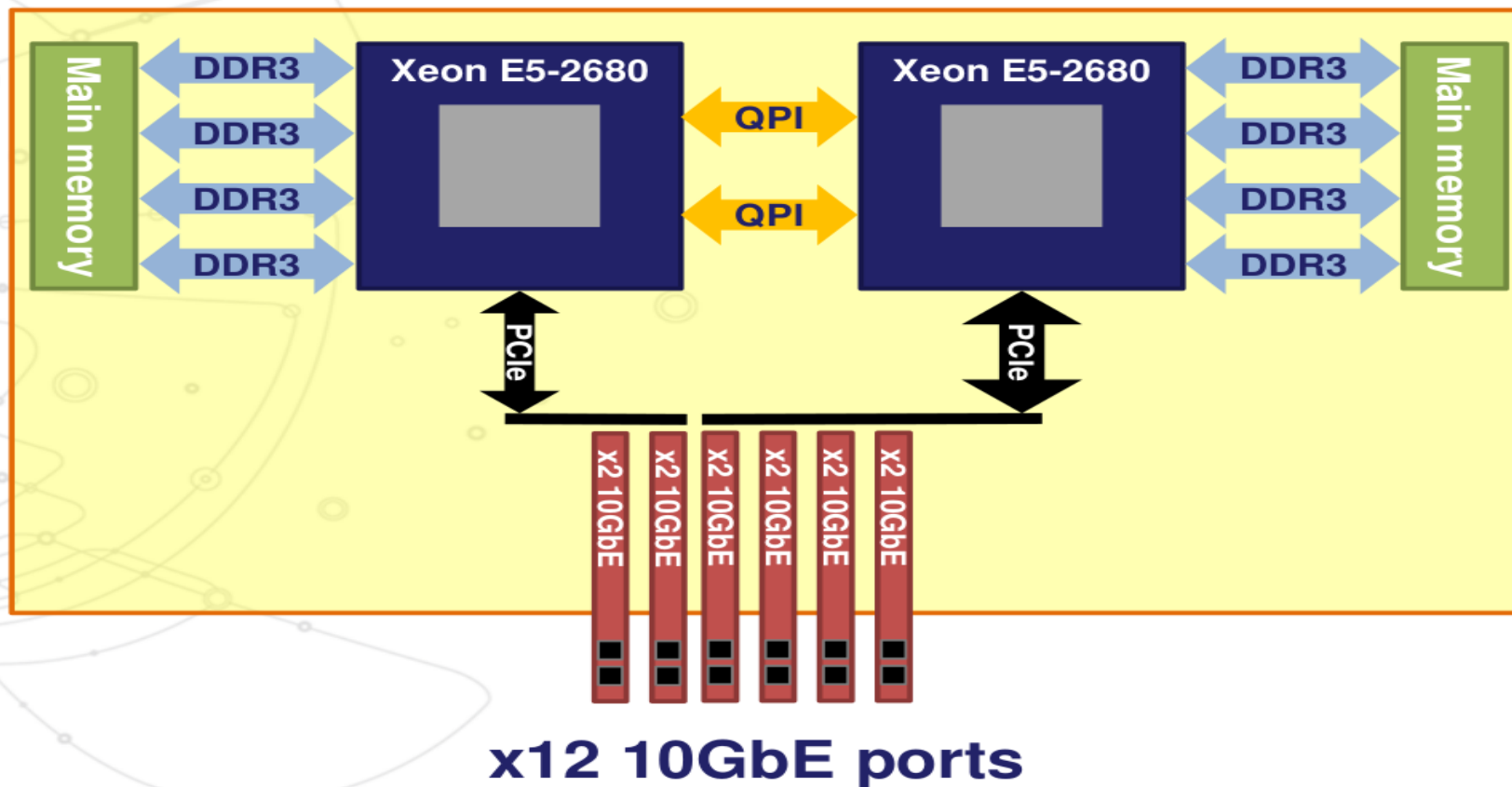


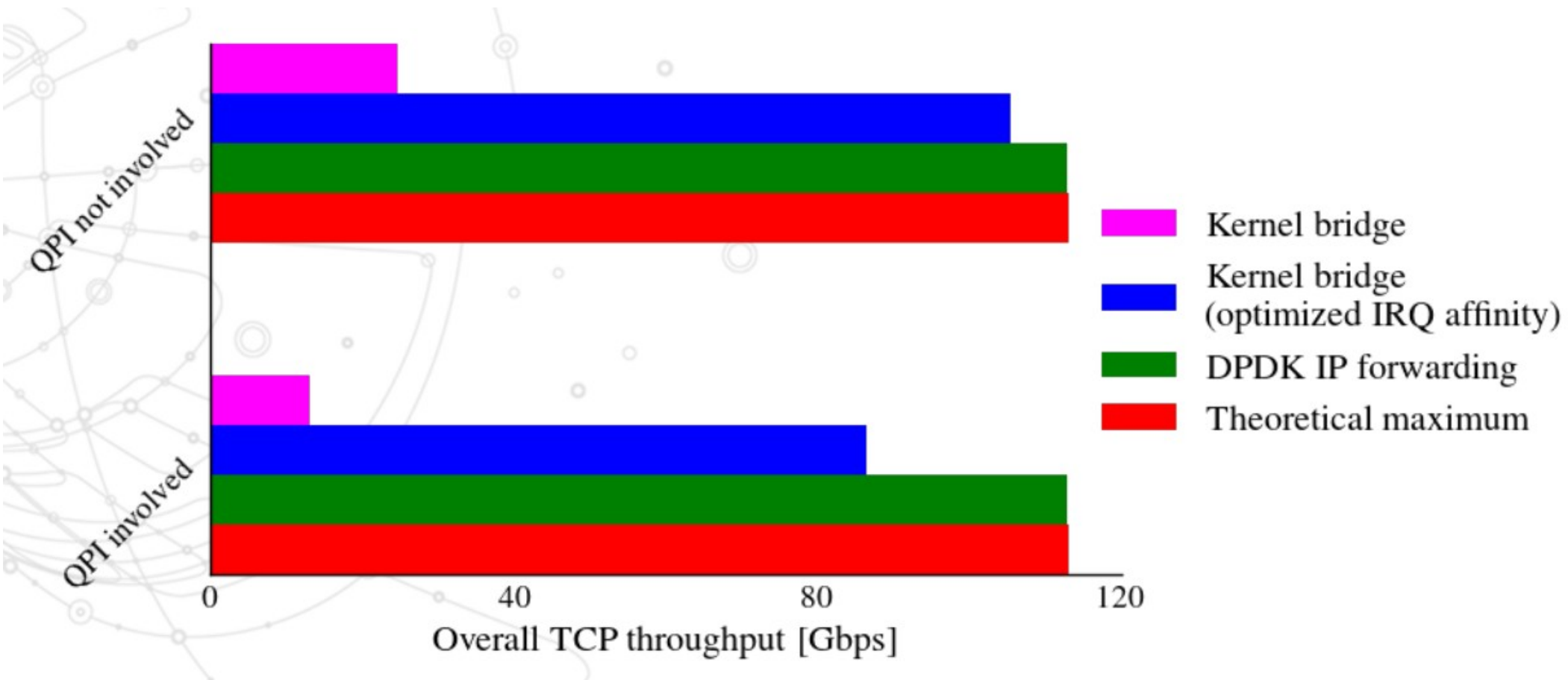
Software Switching

- Ultimately network “devices” with large buffers would be the simplest solution for the DAQ problem
- How to make these affordable?
- Re-use other affordable elements

→ Intel DPDK (Data-Plane Development Kit)

- fast packet processing library → allow building PC-based network switches

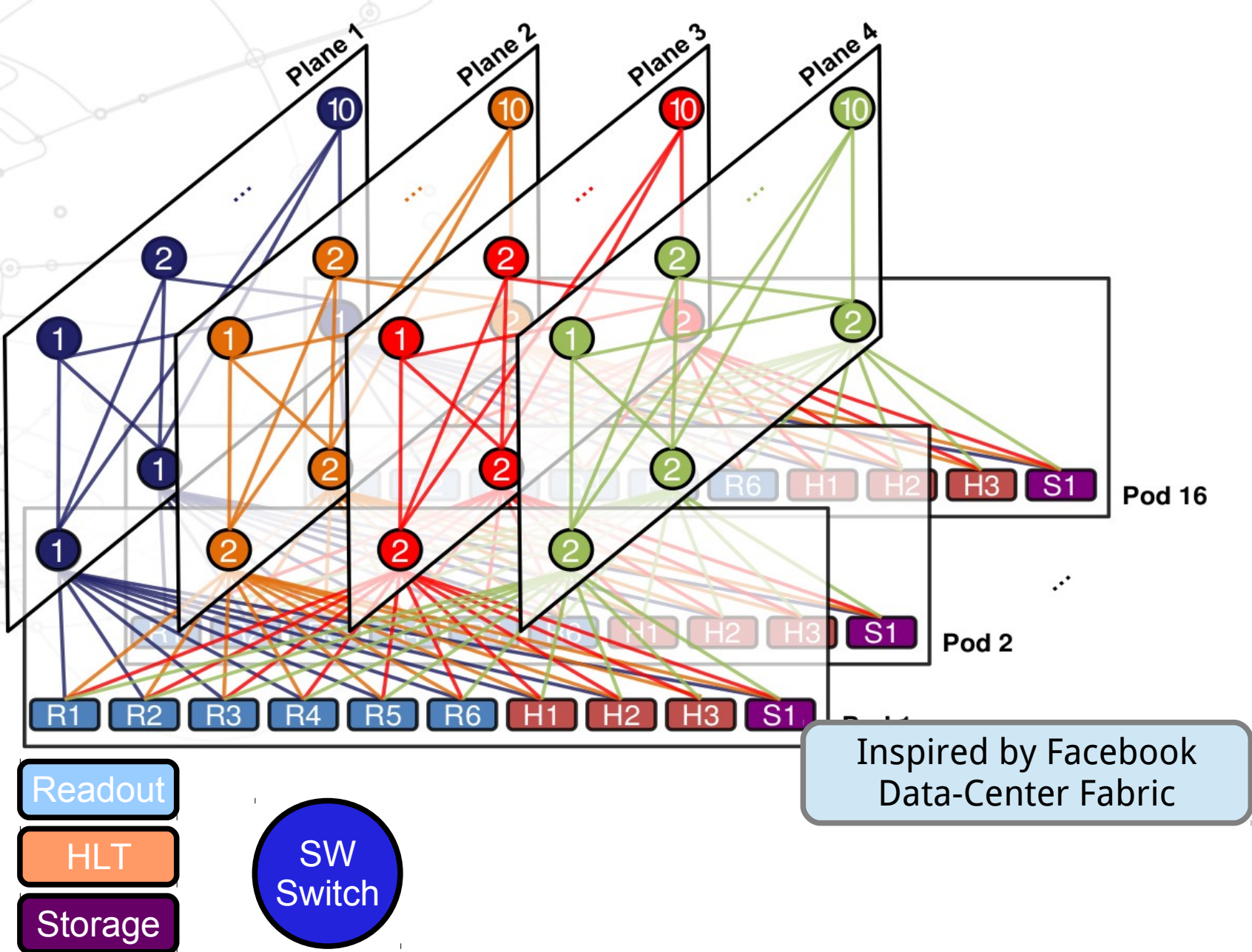




➔ Main limitation of this approach is density

- limited number of ports per “switch”
- scaling to a LHC-size network requires to re-think the topology

Possible DAQ System



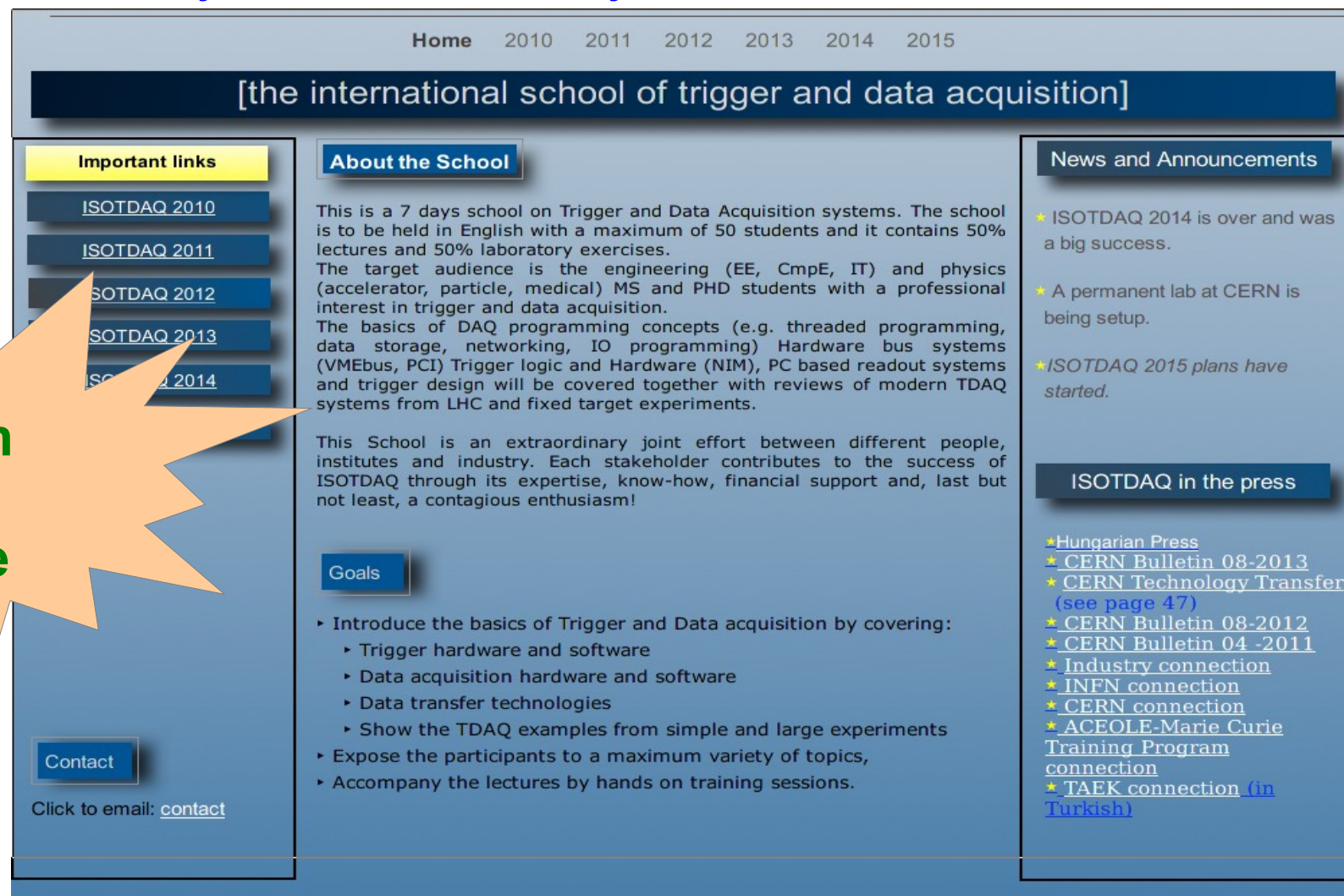


Almost The End

What's next?

- ➔ If you are technology-oriented
- ➔ If you found these topics interesting
- ➔ If you look forward to the challenges
- ➔ If you like to be in the centre of the action
- ➔ Your chance of hearing much more and learn through practice ...

➔ Sixth edition of the **International School of Trigger and Data Acquisition** will be held in February 2016 and hosted by Weizmann Institute



Home 2010 2011 2012 2013 2014 2015

[the international school of trigger and data acquisition]

Important links

- [ISOTDAQ 2010](#)
- [ISOTDAQ 2011](#)
- [ISOTDAQ 2012](#)
- [ISOTDAQ 2013](#)
- [ISOTDAQ 2014](#)

About the School

This is a 7 days school on Trigger and Data Acquisition systems. The school is to be held in English with a maximum of 50 students and it contains 50% lectures and 50% laboratory exercises. The target audience is the engineering (EE, CmpE, IT) and physics (accelerator, particle, medical) MS and PHD students with a professional interest in trigger and data acquisition. The basics of DAQ programming concepts (e.g. threaded programming, data storage, networking, IO programming) Hardware bus systems (VMEbus, PCI) Trigger logic and Hardware (NIM), PC based readout systems and trigger design will be covered together with reviews of modern TDAQ systems from LHC and fixed target experiments.

This School is an extraordinary joint effort between different people, institutes and industry. Each stakeholder contributes to the success of ISOTDAQ through its expertise, know-how, financial support and, last but not least, a contagious enthusiasm!

Goals

- Introduce the basics of Trigger and Data acquisition by covering:
 - Trigger hardware and software
 - Data acquisition hardware and software
 - Data transfer technologies
 - Show the TDAQ examples from simple and large experiments
- Expose the participants to a maximum variety of topics,
- Accompany the lectures by hands on training sessions.

News and Announcements

- ★ ISOTDAQ 2014 is over and was a big success.
- ★ A permanent lab at CERN is being setup.
- ★ ISOTDAQ 2015 plans have started.

ISOTDAQ in the press

- ★ Hungarian Press
- ★ [CERN Bulletin 08-2013](#)
- ★ [CERN Technology Transfer \(see page 47\)](#)
- ★ [CERN Bulletin 08-2012](#)
- ★ [CERN Bulletin 04 -2011](#)
- ★ [Industry connection](#)
- ★ [INFN connection](#)
- ★ [CERN connection](#)
- ★ [ACEOLE-Marie Curie Training Program connection](#)
- ★ [TAEK connection \(in Turkish\)](#)

Contact

Click to email: [contact](#)

Watch
this
space

<http://isotdaq.web.cern.ch/isotdaq/isotdaq/Home.html>



The End

W.Vandelli CERN/PH-ADT
Wainer.Vandelli@cern.ch