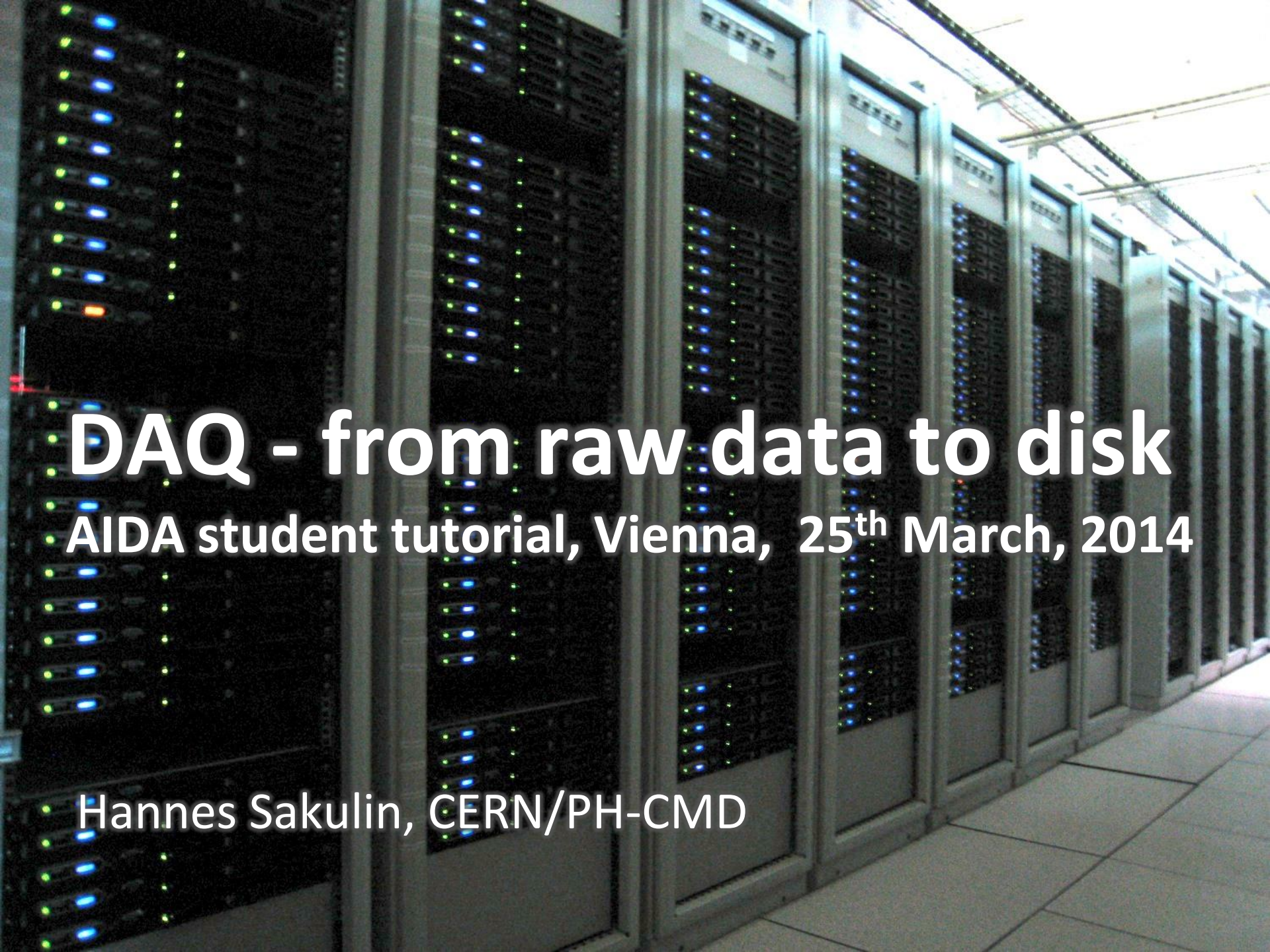


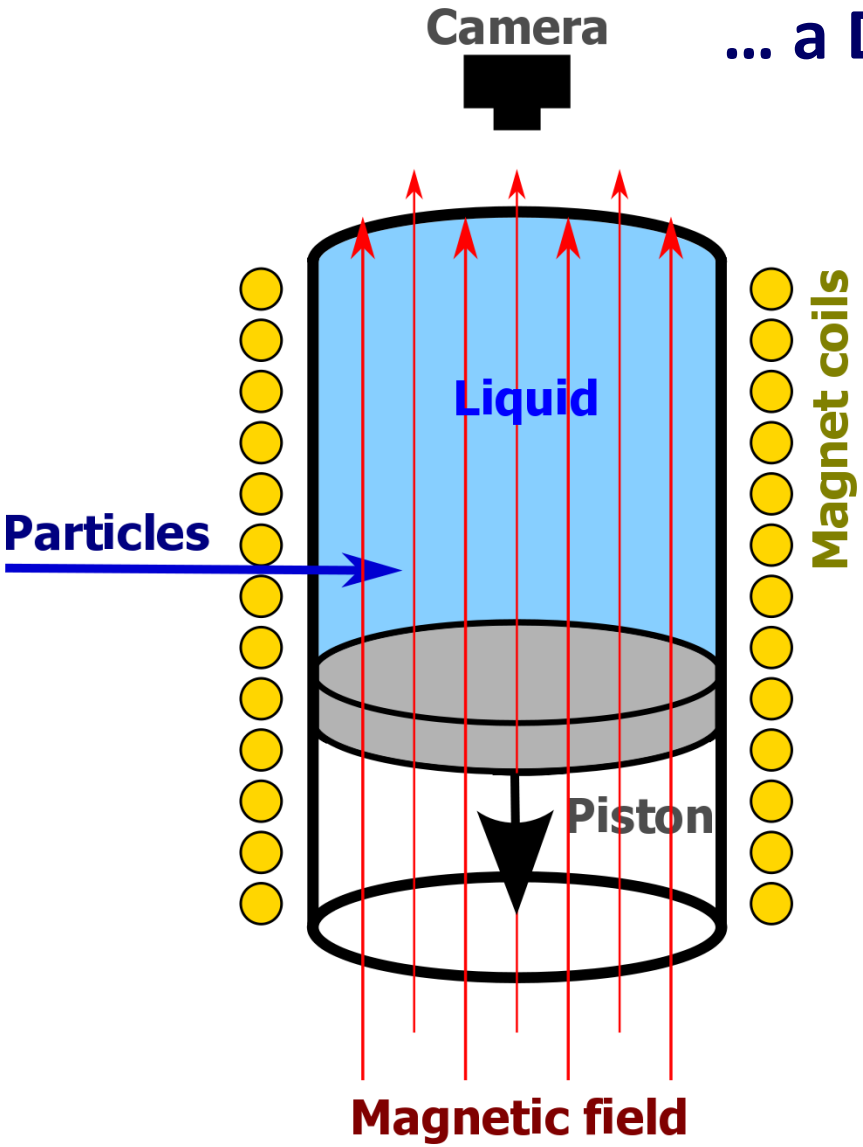
DAQ - from raw data to disk

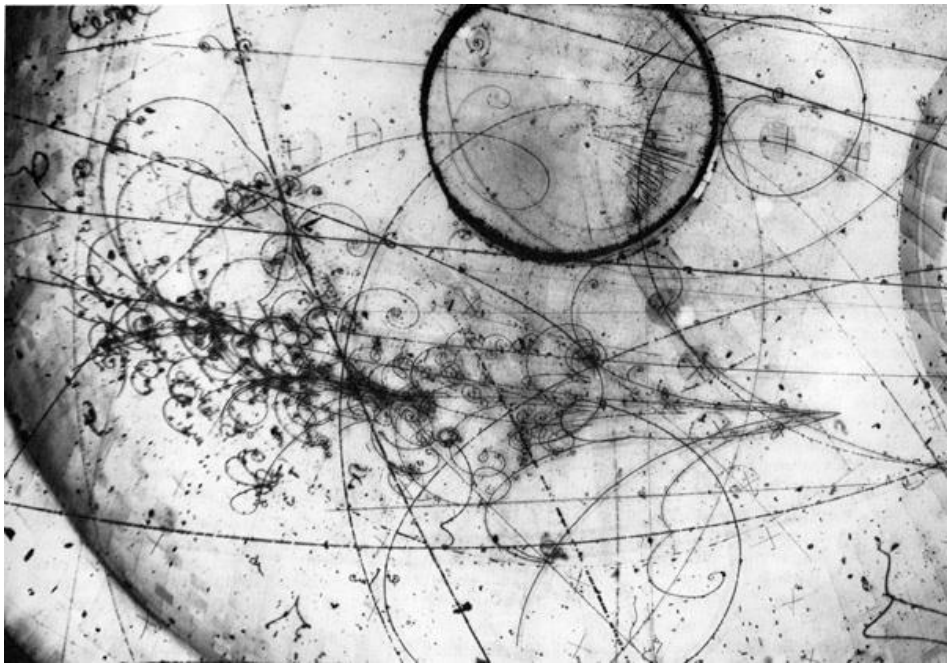
AIDA student tutorial, Vienna, 25th March, 2014

Hannes Sakulin, CERN/PH-CMD

Once upon a time…

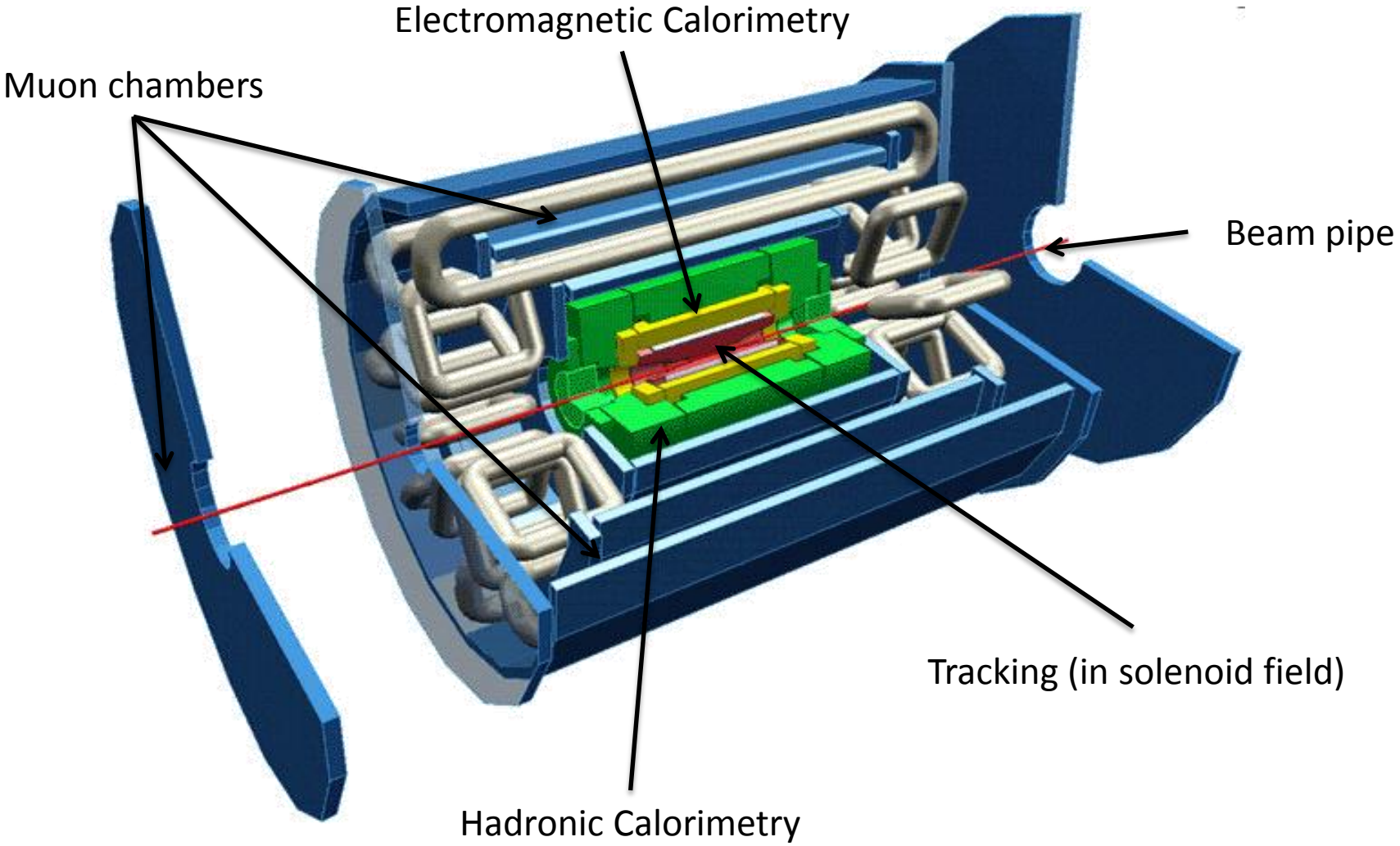
… a DAQ system looked like this





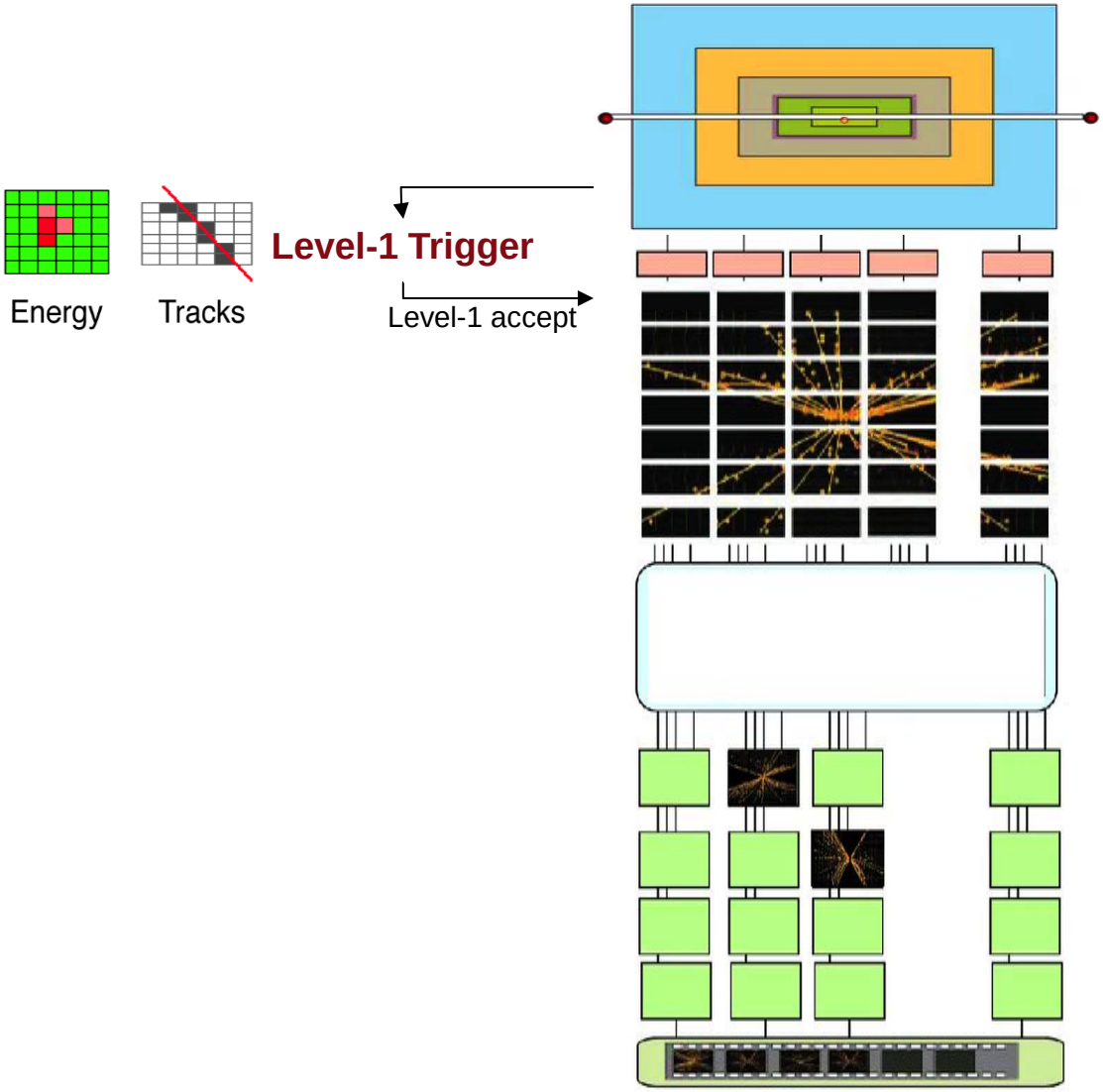
And event selection was done like that …

How do we "read" a modern experiment ?

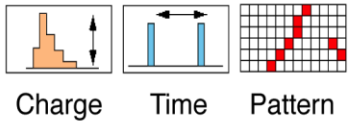


O(100 x 10^6) detector channels

A typical DAQ system



0) Detectors



1) Front-end systems
Analog-digital conversion
Detector readout link
Feature extraction

2) DAQ readout Links
Interface from custom to commercial electronics

3) Event Building
Network

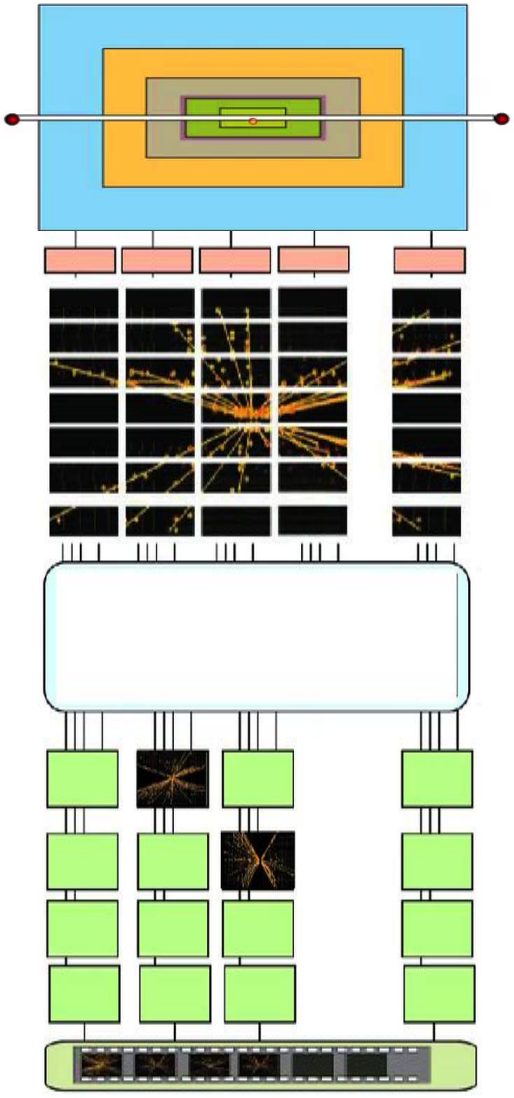
4) Event Filter (High Level Trigger)
Thousands of CPU cores

5) Local mass storage

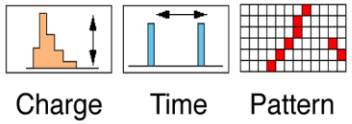
Why do we need a trigger ?

Beam crossing (bx) frequency given by the accelerator
LHC: 40 MHz

Maximum readout rate limited by
- **B/W of readout links**
- **B/W of DAQ links**
- **Computing Infrastructure Networks and Processing Power**



0) Detectors



1) Front-end systems
Analog-digital conversion
Feature extraction

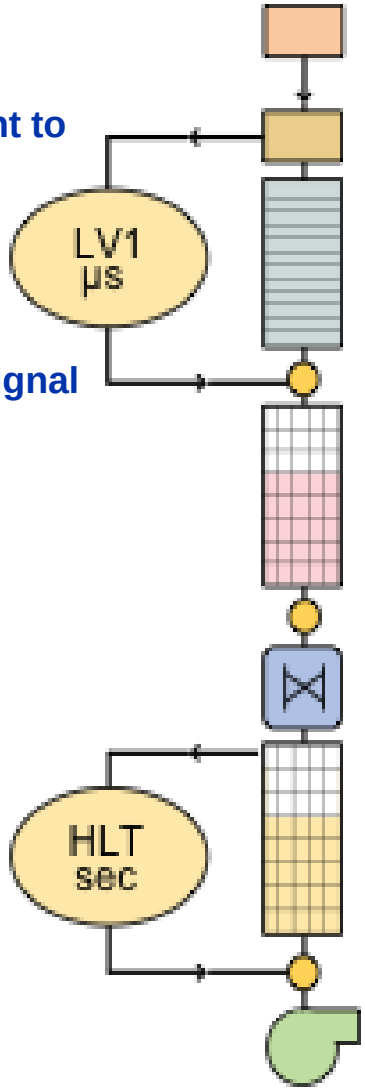
2) DAQ readout Links
Interface from custom to commercial electronics

3) Event Building
Network

4) Event Filter (High Level Trigger)
Thousands of CPU cores

5) Local mass storage

Trigger and DAQ interaction



Coarse data sent to trigger

Pipelined electronics process every bx

Level-1 accept signal

Full data stored in a pipeline

Accepted event fragments stored in de-randomizer buffer

Central DAQ reads accepted fragments ..

…. and builds events

High Level trigger processing (software)

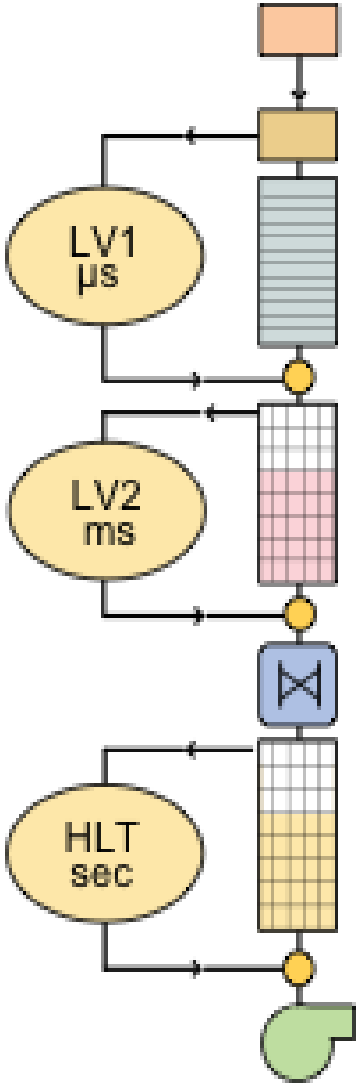
Full events buffer during high-level trigger processing

At beam crossing (bx) frequency
For the LHC: 40 MHz

Level-1 accept rate
= event building rate
CMS: 100 kHz
LHCb: 1MHz

Rate to disk:
few 100 Hz
up to kHz

Trigger and DAQ interaction



Beam crossing (bx) frequency
For the LHC: 40 MHz

Level-1 accept rate
ATLAS: 65 KHz

Some experiments have multiple trigger levels

Level-2 accept rate = event building rate
ATLAS: 6 KHz

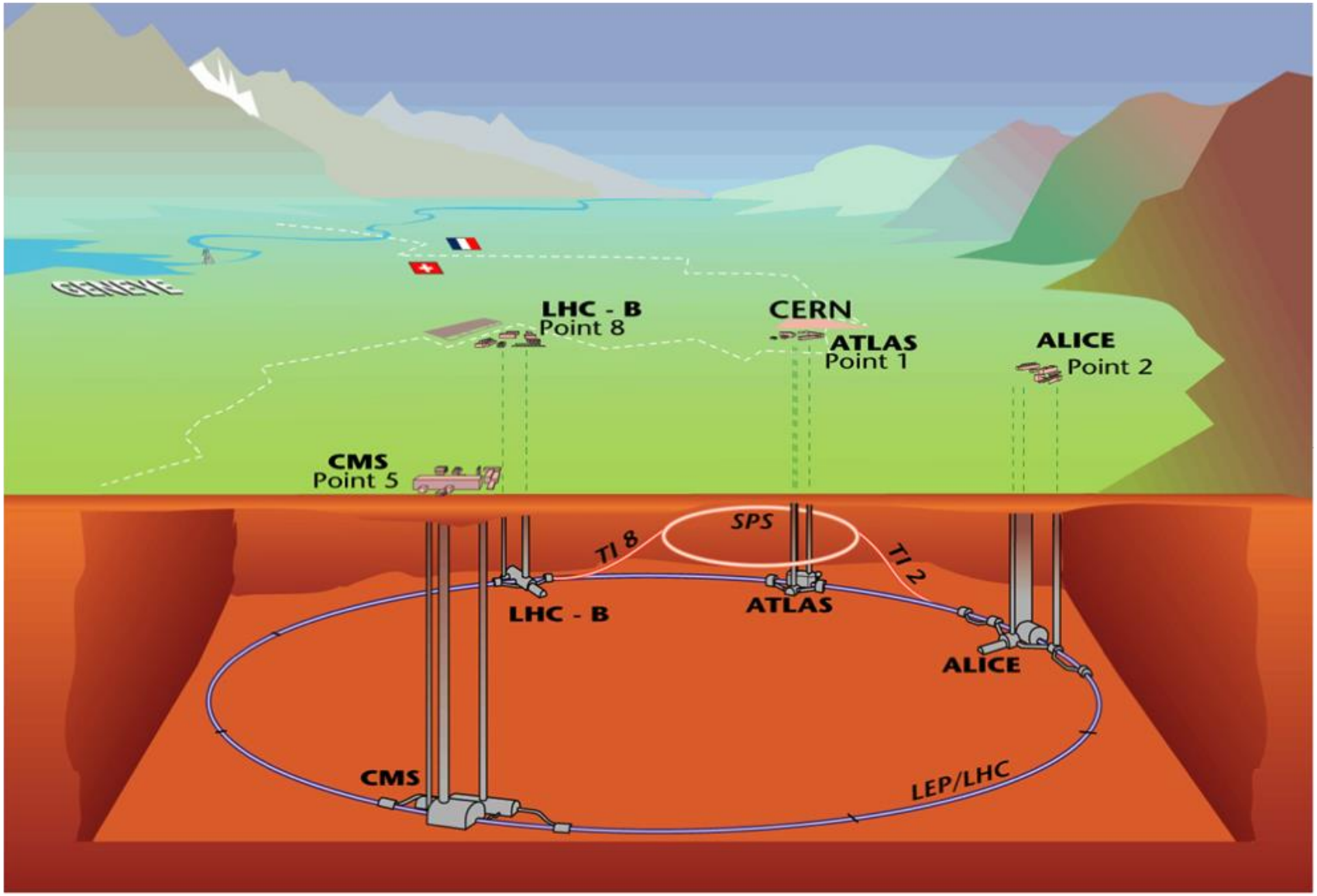
Rate to disk:
few 100 Hz
up to kHz

Disclaimer

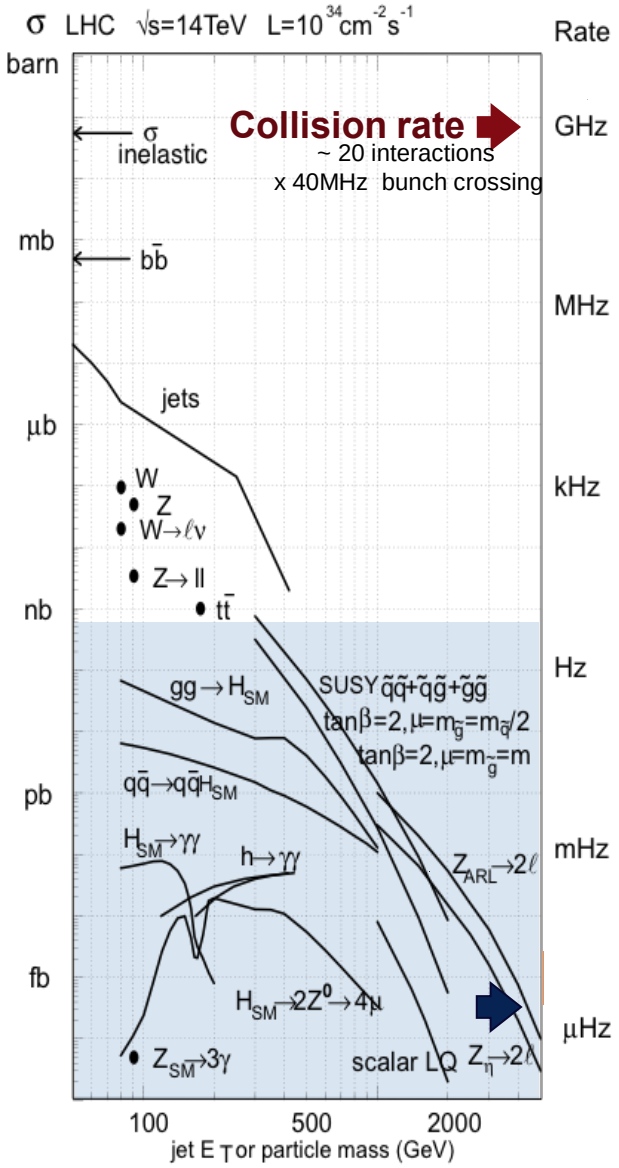
- This talk is based on the DAQ system of the LHC experiments. Many concepts are nevertheless generic.

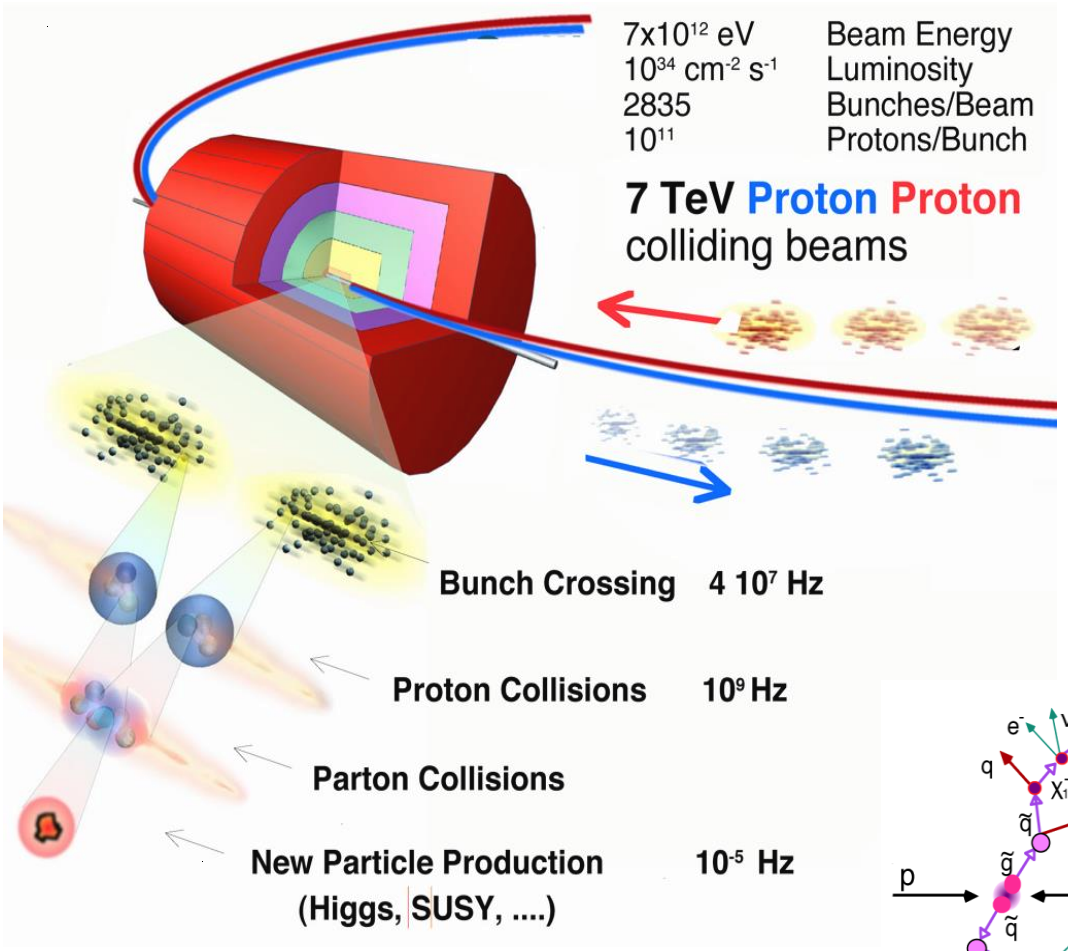
- Thanks to all my colleagues who contributed material to this talk (actively and passively), especially to Christoph Schwick and Sergio Cittolin.

DAQ at the LHC



Proton-proton collisions at the LHC. Searching issue

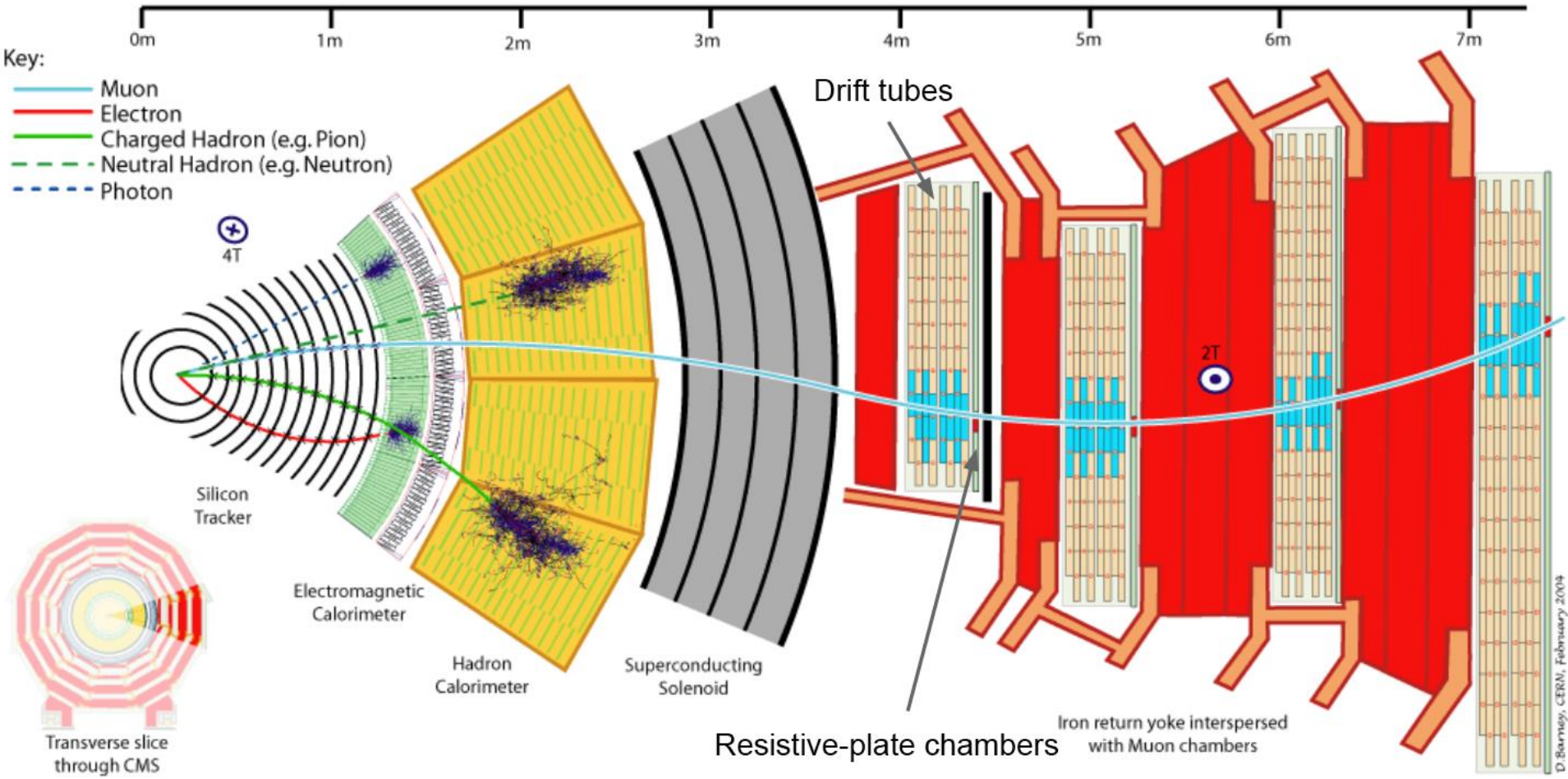




Collision Rate: ~10^9 Hz. Event Selection: ~1/10^{13}

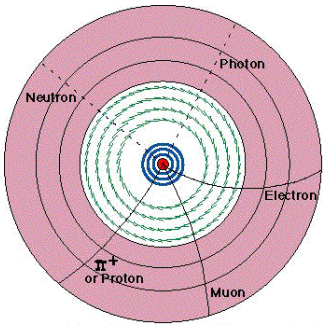
0) Detectors

Typical detector layout

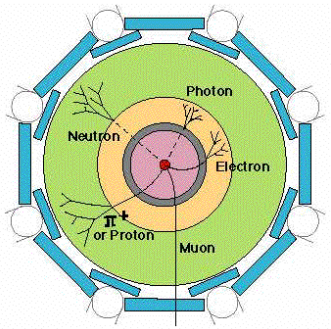


Combination of detectors allows particles to be identified

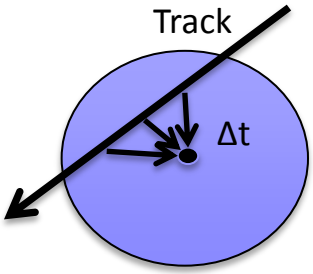
Detectors are a science by themselves



- Tracking detectors
  - Charge collected in silicon semiconductor
  - Precise position measurement – high number of channels



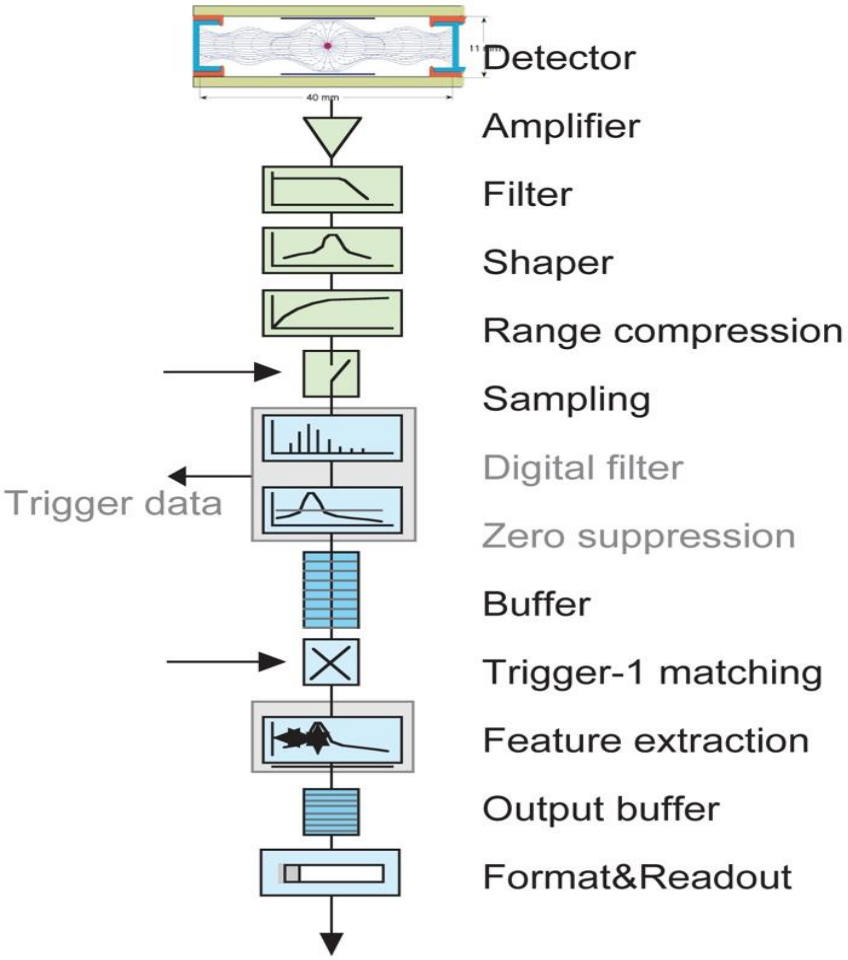
- Calorimeters
  - Scintillation light collected by Photo-Detectors
  - Use amplitude to measure deposited energy



- Muon detectors
  - Drift chamber. Electrons drift to central wire and cause avalanche.
  - Timing determines position

1) Front-end systems

Typical front-end electronics



Signal
typically small current pulse

Analog electronics
depending on detector type

Analog/Digital conversion

Digital Signal Processing (DSP)

Front-end pipeline buffer
Store data while trigger is computing its decision

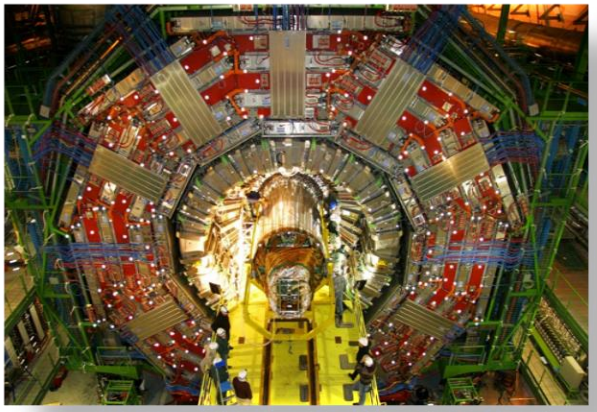
De-randomizing buffer
Store data until read out by DAQ

Readout-link to central DAQ

Getting the data out of the detector

10K – 100M rad

Max 4T



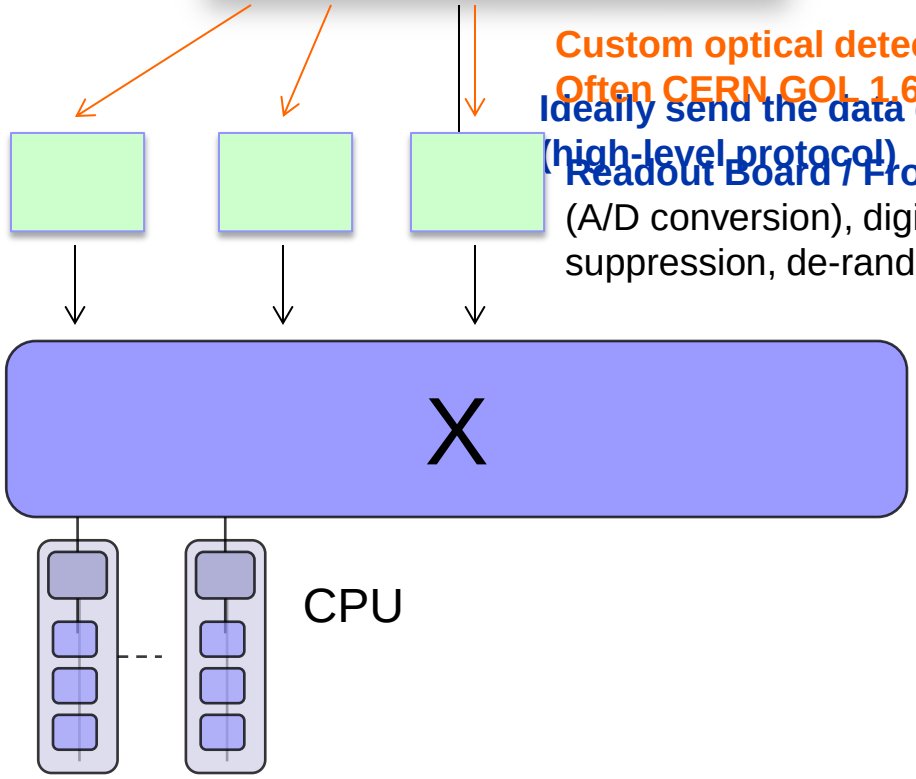
~zero mass

Gigabit Optical Link (GOL)
- ASIC
- Radiation hard
- designed at CERN

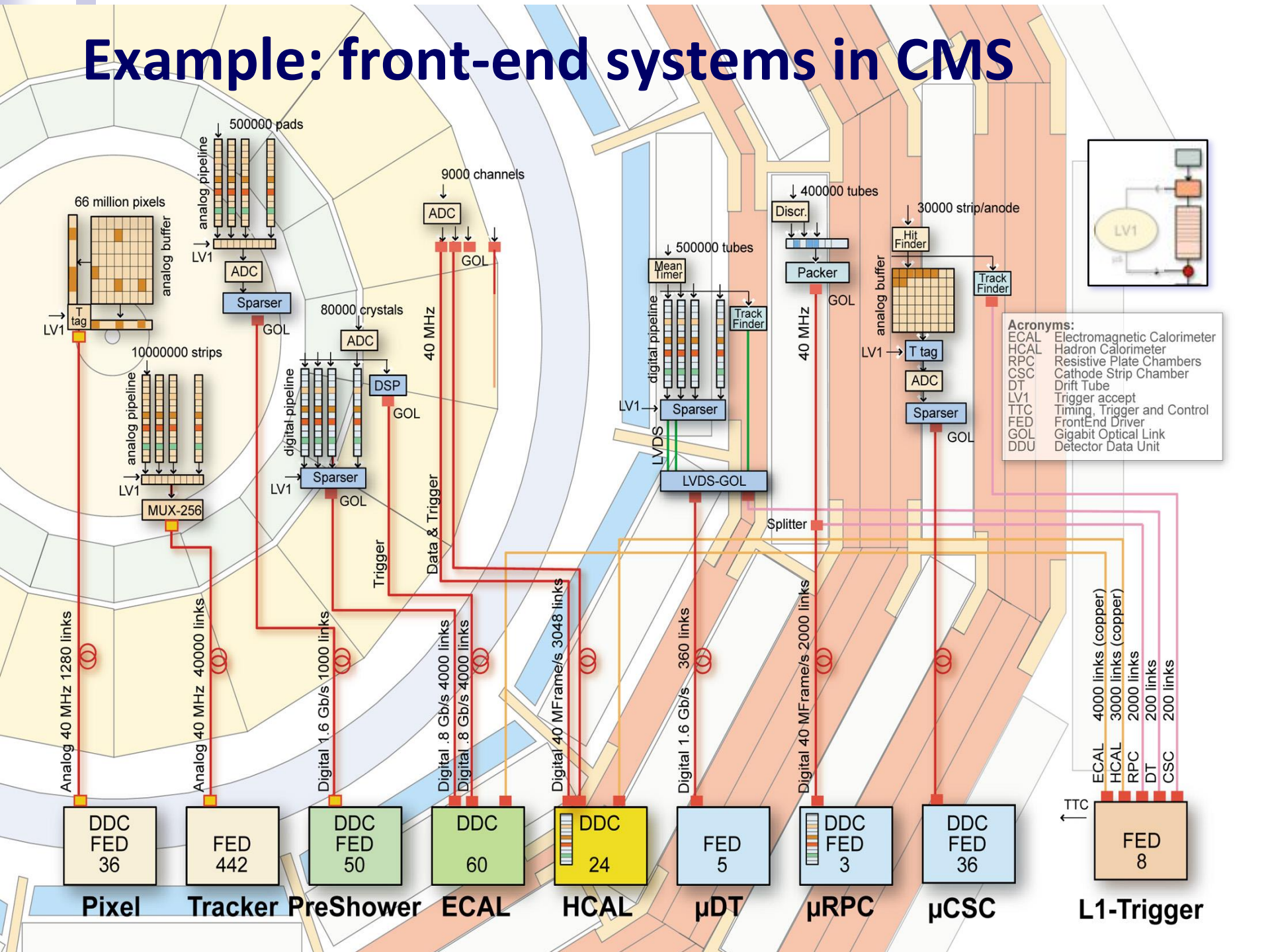
Custom optical detector-readout link.
Often CERN GOL 1.6 Gb/s

Ideally send the data directly to the event builder (high-level protocol)

Readout Board / Frontend Driver / Readout Driver
(A/D conversion), digital signal processing, zero suppression, de-randomizer buffer

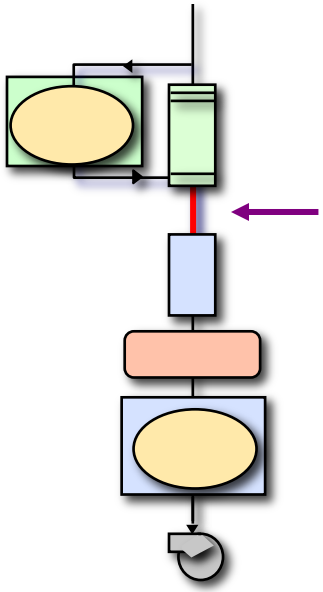


Example: front-end systems in CMS

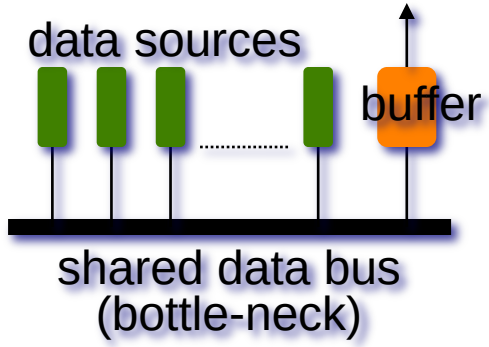


2) DAQ readout Links

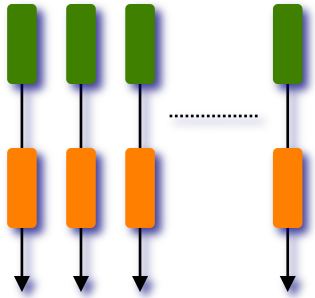
Data Flow: Data Readout



- Former times: Use of bus-systems
  - VME or Fastbus
  - Parallel data transfer (typical: 32 bit) on shared bus
  - One source at a time can use the bus



- Today: Point to point links
  - Optical or electrical
  - Data serialized
  - Custom or standard protocols
  - All sources can send data simultaneously



- Compare trends in industry market:
  - 198x: ISA, SCSI(1979),IDE, parallel port, VME(1982)
  - 199x: PCI(1990, 66MHz 1995), USB(1996), FireWire(1995)
  - 200x: USB2, FireWire 800, PCIexpress, Infiniband, GbE, 10GbE

Example readout link : ATLAS

High-Speed Optical Link for Atlas (HOLA) mezzanine board plugged onto Readout Driver electronics.

Detector electronics push 32-bit parallel data + header + trailer

Optical fiber
2 Gb/s

Readout-Buffer Input (ROBIN) PCI-x card plugged into a PC

Receiving up to 3 links

Data transfer to PC memory via DMA

Getting the data into the PC

Problem:

Read data into PC with high bandwidth and low CPU load

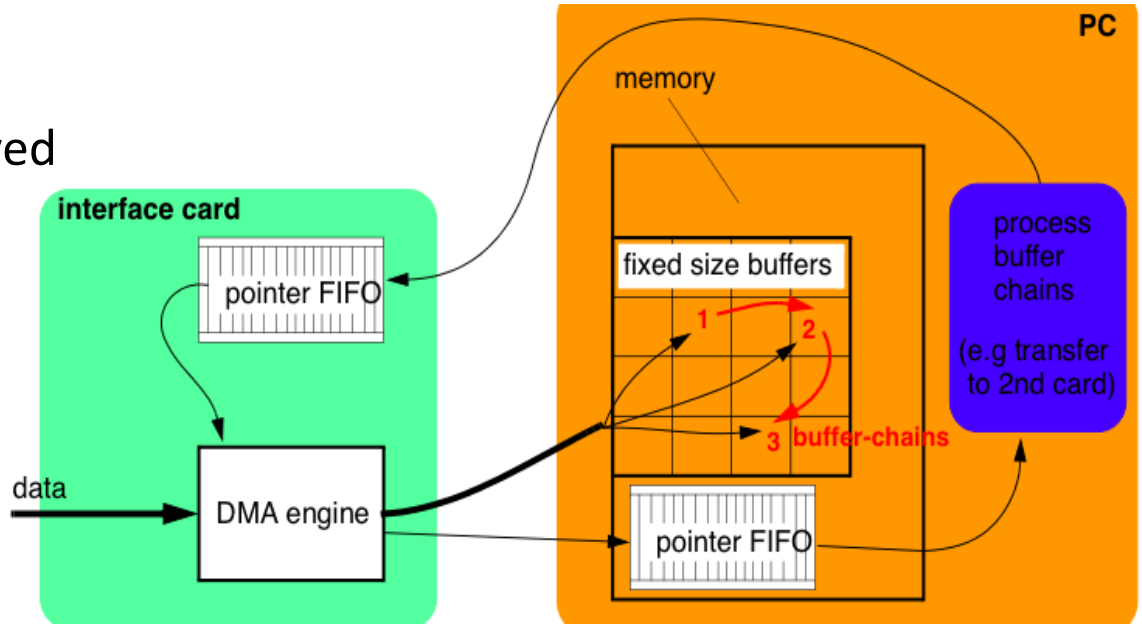
Note: copying data costs a lot of CPU time!

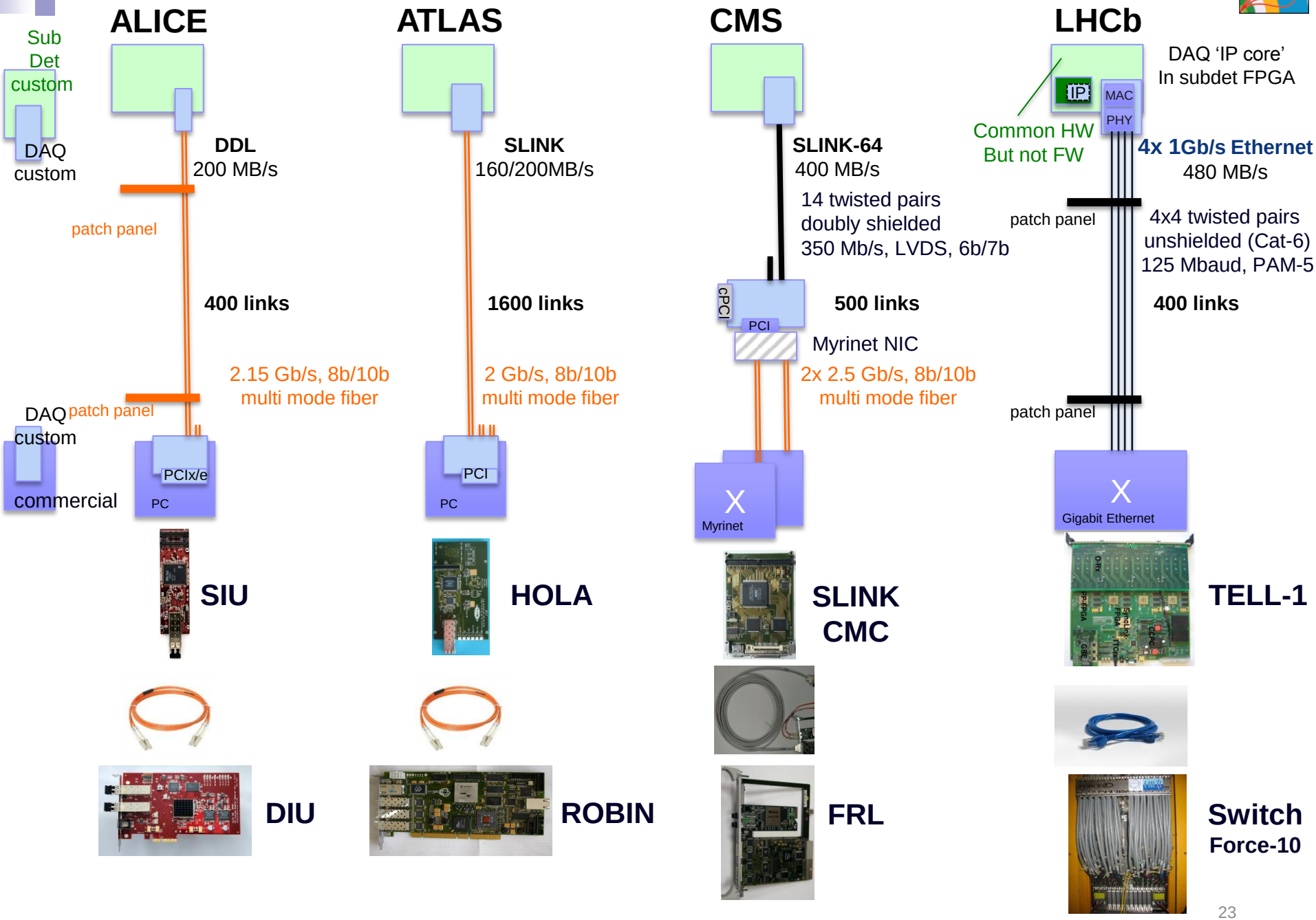
Solution: Buffer-Loaning

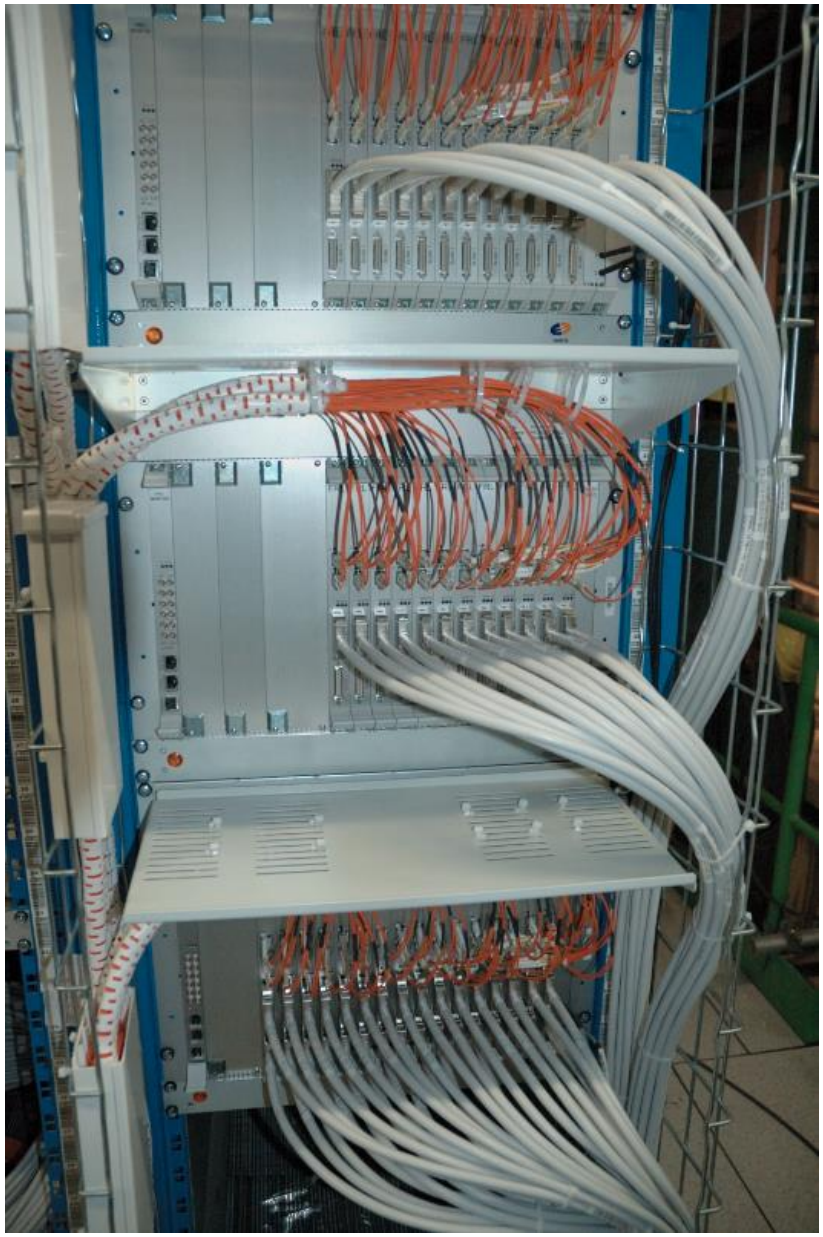
- Hardware shuffles data via DMA (Direct Memory Access) engines
- Software maintains tables of buffer-chains

Advantage:

- No CPU copy involved




ALICE
ATLAS
CMS
LHCb
Sub Det custom
DAQ custom
DDL
200 MB/s
patch panel
400 links
2.15 Gb/s, 8b/10b
multi mode fiber
DAQ custom
patch panel
commercial
PCIx/e
PC
SIU
DIU
SLINK
160/200MB/s
1600 links
2 Gb/s, 8b/10b
multi mode fiber
PCI
PC
HOLA
ROBIN
SLINK-64
400 MB/s
14 twisted pairs
doubly shielded
350 Mb/s, LVDS, 6b/7b
500 links
cPCI
PCI
Myrinet NIC
2x 2.5 Gb/s, 8b/10b
multi mode fiber
X
Myrinet
SLINK
CMC
FRL
DAQ 'IP core'
In subdet FPGA
IP
MAC
PHY
Common HW
But not FW
4x 1Gb/s Ethernet
480 MB/s
patch panel
4x4 twisted pairs
unshielded (Cat-6)
125 Mbaud, PAM-5
400 links
patch panel
X
Gigabit Ethernet
TELL-1
Switch
Force-10



CMS

SLINK-64
400 MB/s

14 twisted pairs
doubly shielded
350 Mb/s, LVDS, 6b/7b

cPCI
PCI
Myrinet NIC

2x 2.5 Gb/s, 8b/10b
multi mode fiber

X
Myrinet

SLINK
CMC

FRL

LHCb

DAQ 'IP core'
In subdet FPGA

IP
MAC
PHY

Common HW
But not FW

4x 1Gb/s Ethernet
480 MB/s

patch panel

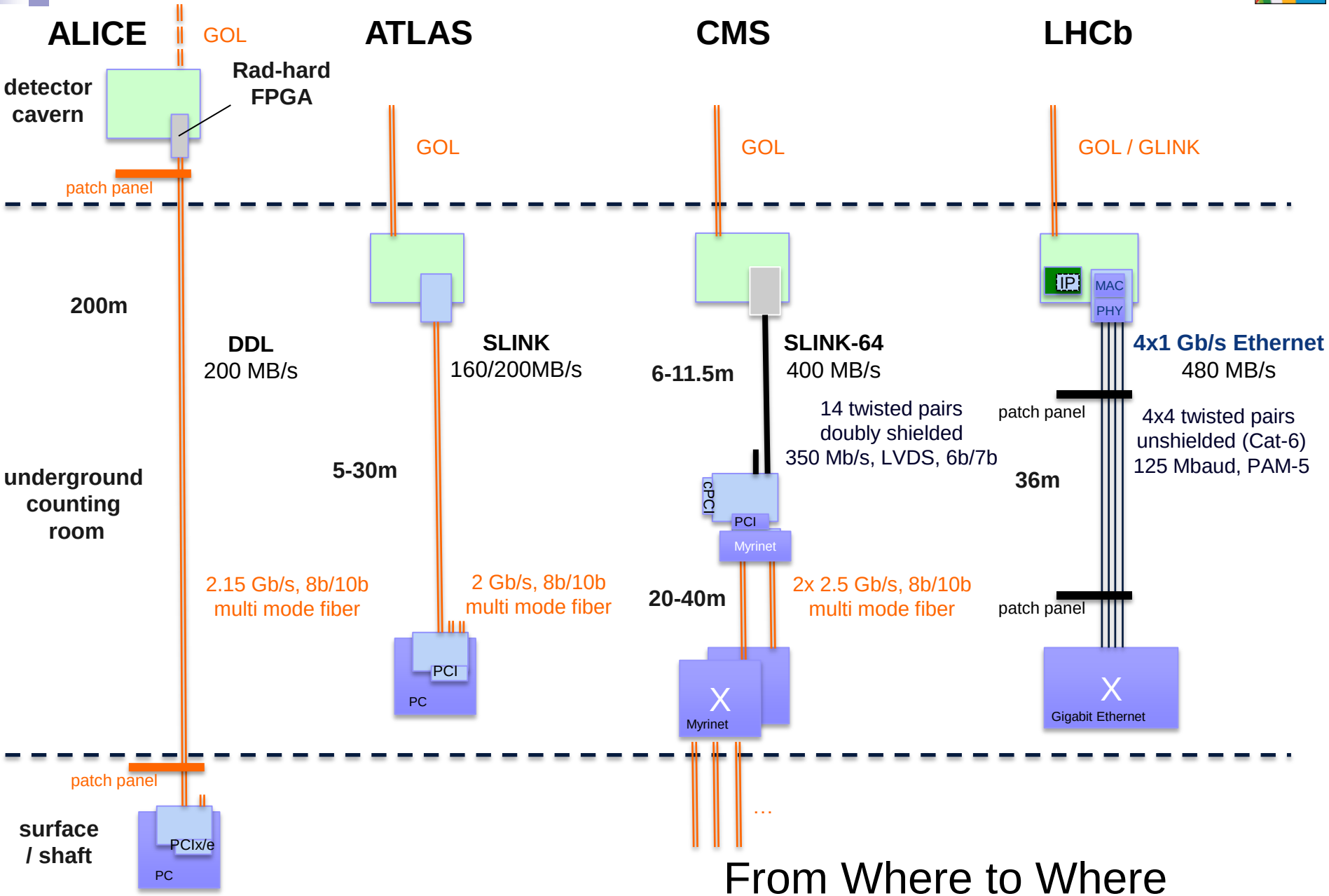
4x4 twisted pairs
unshielded (Cat-6)
125 Mbaud, PAM-5

patch panel

X
Gigabit Ethernet

TELL-1

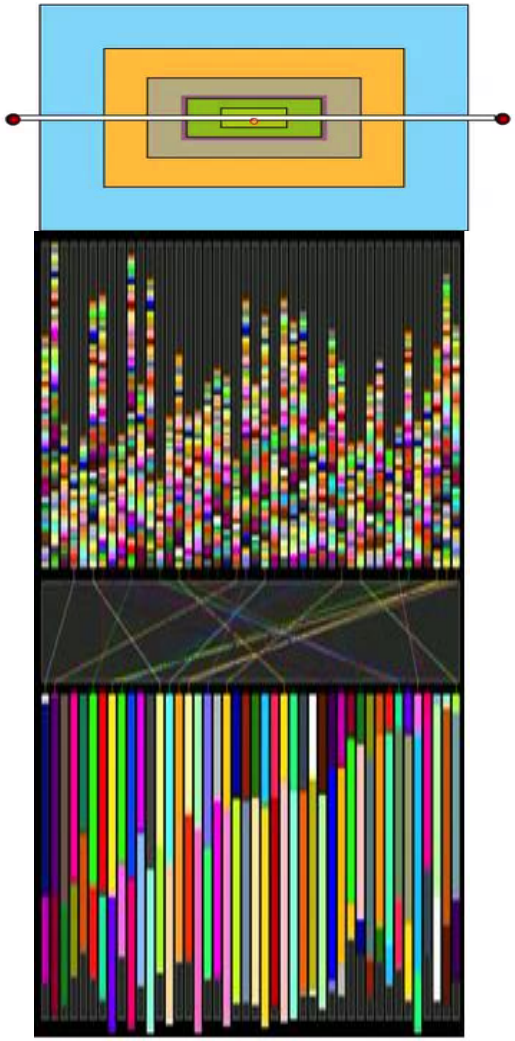
Switch
Force-10



From Where to Where

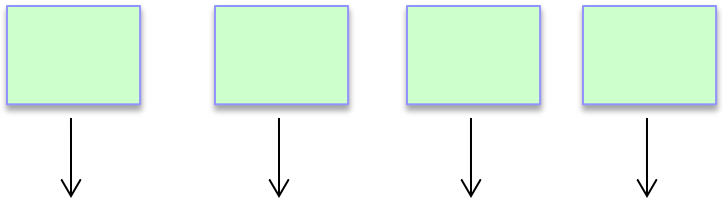
3) Event-Building

Event Building



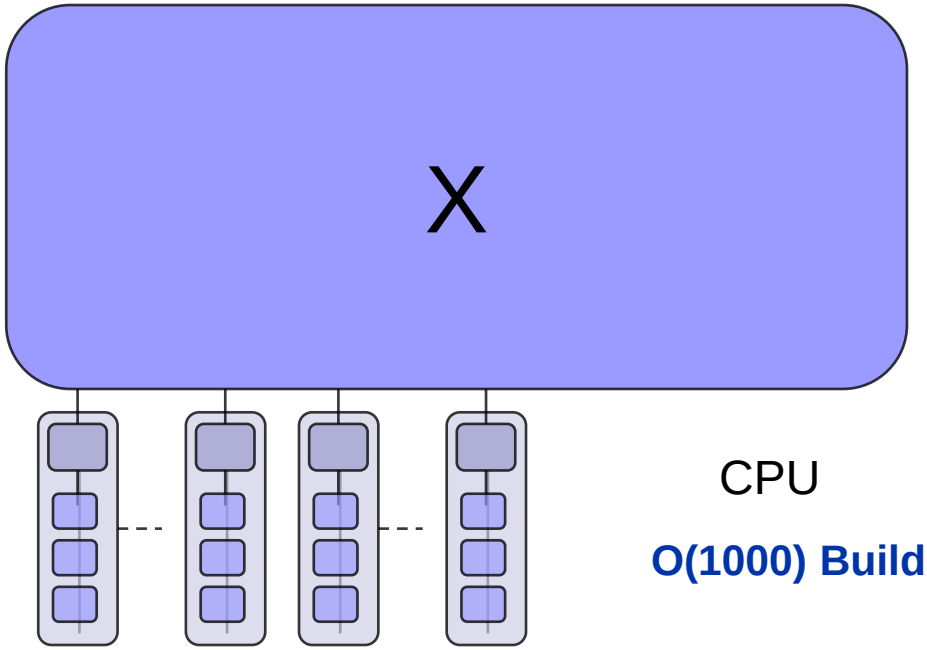
3) Event Building
Network

Principle of event building



Readout Board / Frontend Driver / Readout Driver

O(1000) data sources

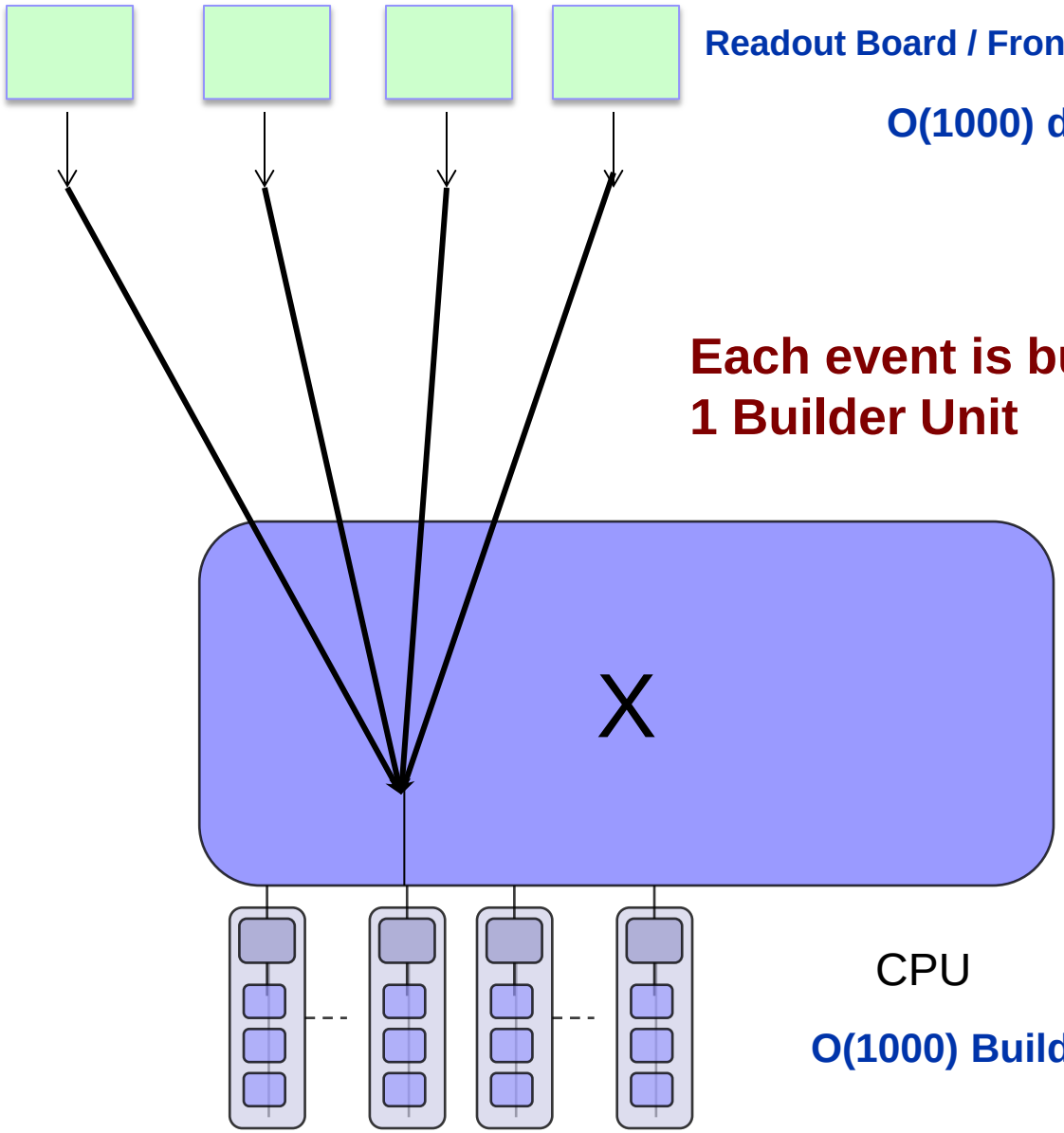


Network switch

CPU

O(1000) Builder / Processing nodes

Principle of event building



Readout Board / Frontend Driver / Readout Driver

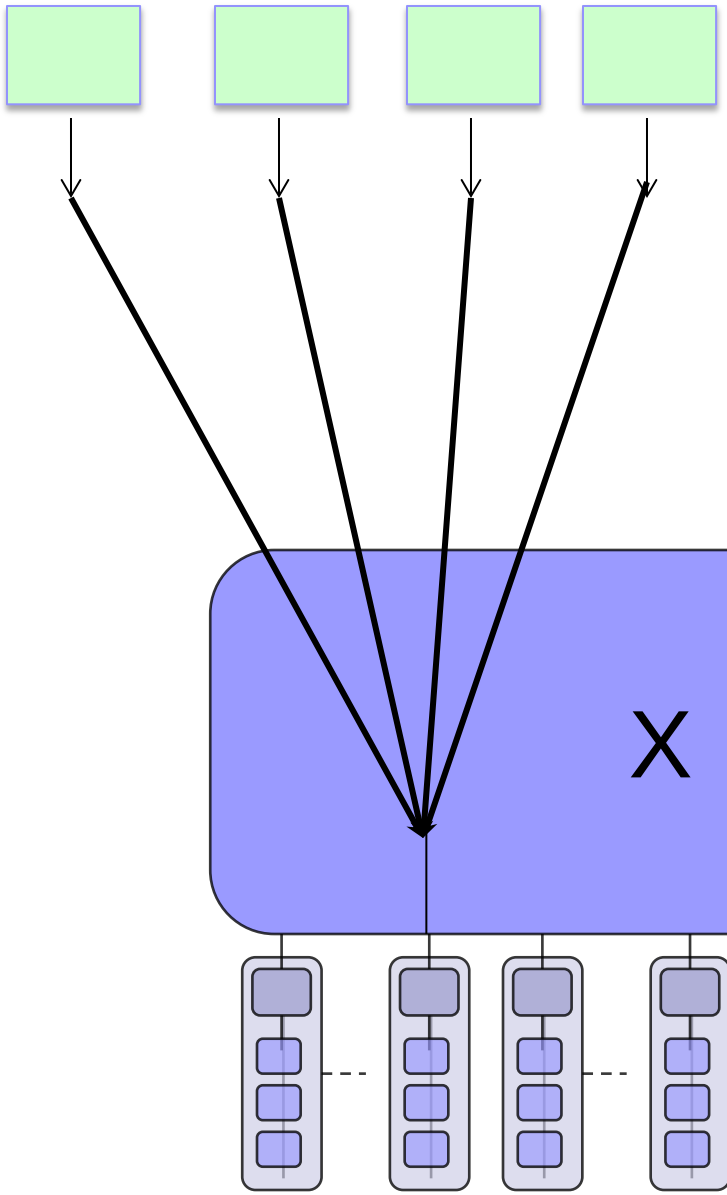
O(1000) data sources

Each event is built by exactly 1 Builder Unit

O(1000) Builder / Processing nodes

Congestion

Building the events



Readout Board / Frontend Driver / Readout Driver

O(1000) data sources

EVB traffic

all sources send to
the same destination
(almost) concurrently.

Switches typically are not built for this traffic pattern. Congestion.

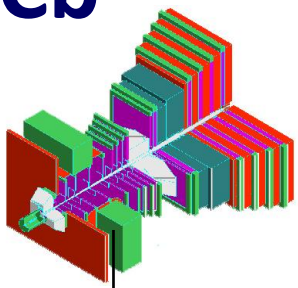
Extent of the effect depends on many factors: Buffer space, event size, event building rate, …

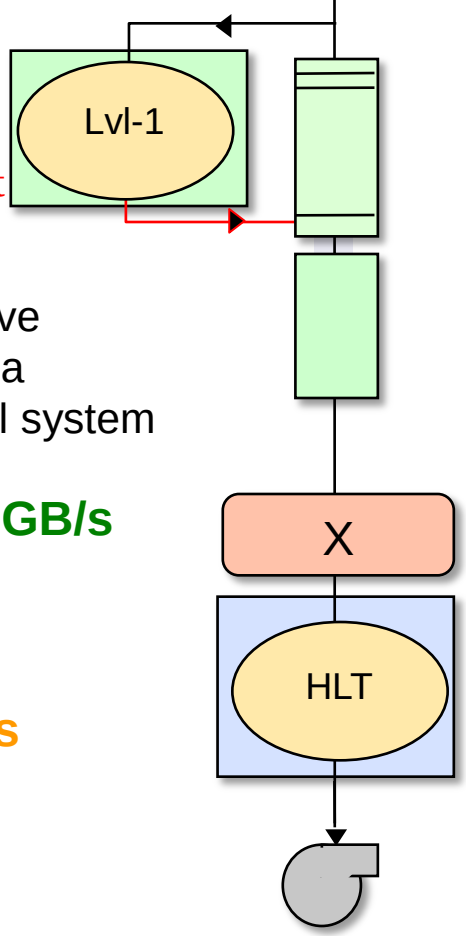
CPU

O(1000) Builder / Processing nodes

Data Flow: LHCb

custom hardware
PC
network switch





40 MHz

front end pipeline

1MHz(*)

* Effective event building rate is an order of magnitude lower as event building is done for Multi-Event-Packets (MEPs)

readout buffer

readout link

Event Size: 35 kB

readout/EVB network

HLT processing farm

2 kHz

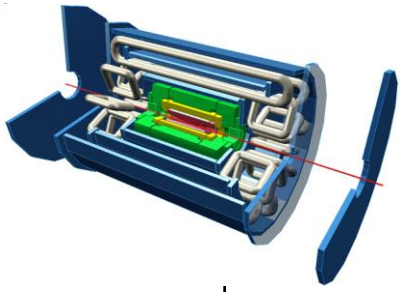
Push-design:
Readout buffers receive destination address via Timing/Trigger/Control system

Bandwidth: 40 (60)GB/s

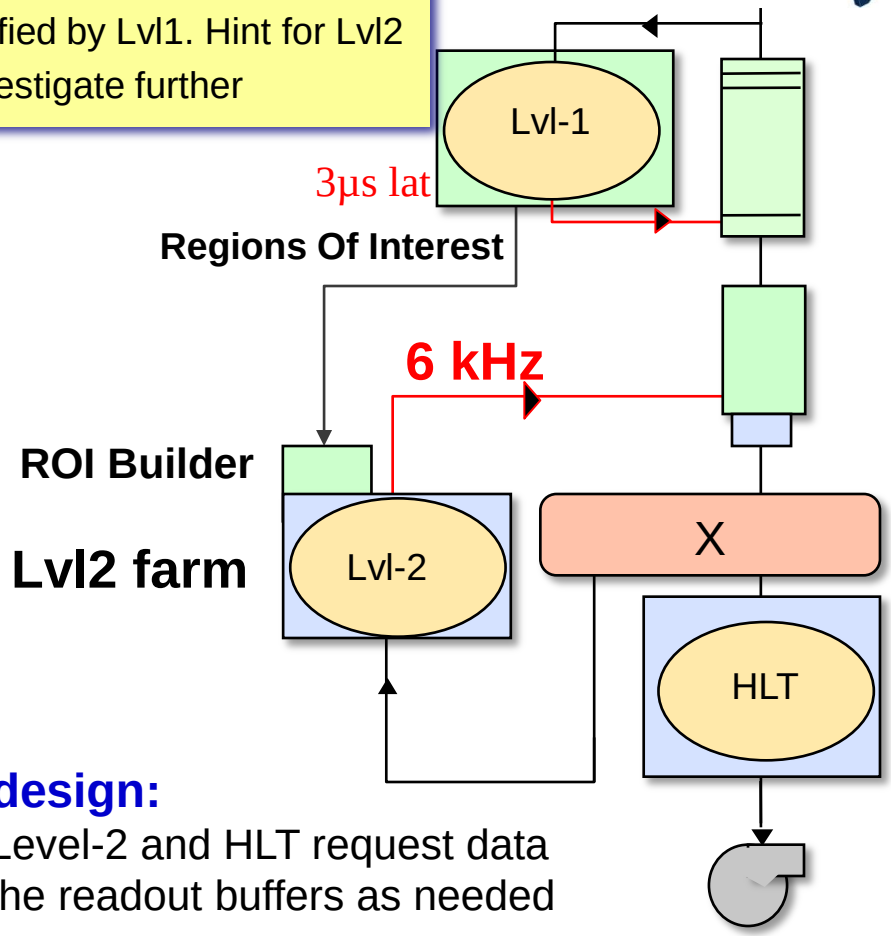
Technology:
Ethernet
1 Gb/s & 10 Gb/s

Data Flow: ATLAS

custom hardware
PC
network switch



Region Of Interest (ROI):
Identified by Lvl1. Hint for Lvl2 to investigate further



40 MHz

front end pipeline

65 kHz
readout link

readout buffer

Event Size: 1.5 MB

event builder

Bandwidth: 5.25 (10) GB/s

HLT farm

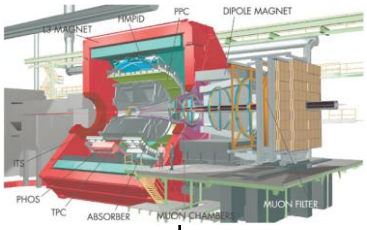
200 Hz

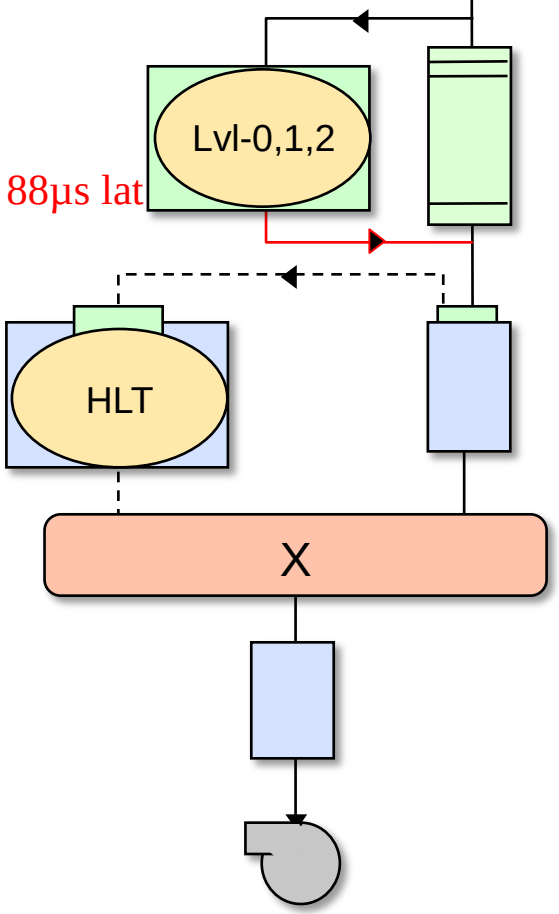
Technology:
Ethernet
1 Gb/s & 10 Gb/s

Pull design:
Both Level-2 and HLT request data from the readout buffers as needed

Data Flow: ALICE

custom hardware
PC
network switch





HLT performs data compression by factor 4-5

HLT farm

Push design:
Push over TCP/IP

front end pipeline

500 Hz
readout link

Event Size: 70 MB

readout buffer

100 Hz

Bandwidth: 2GB/s

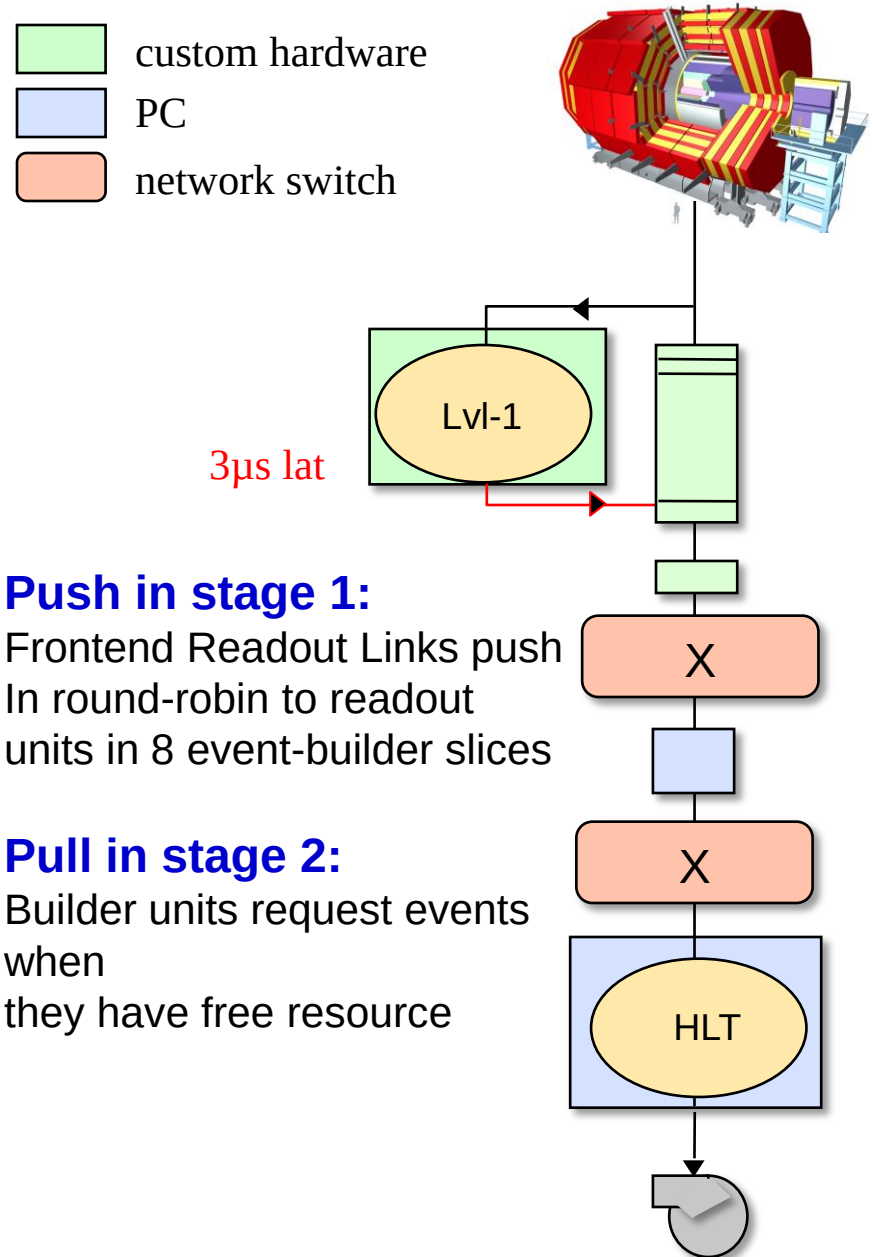
event builder

Technology: Ethernet
1 Gb/s & 10 Gb/s

event buffer

100 Hz

Data Flow: CMS



Push in stage 1:
Frontend Readout Links push
In round-robin to readout
units in 8 event-builder slices

Pull in stage 2:
Builder units request events
when
they have free resource

40 MHz

front end pipeline

100 kHz

readout link

Event Size: 1 MB

event builder network: stage 1

readout buffer

Bandwidth: 100 GB/s
Technology: Myrinet

event builder network: stage 2

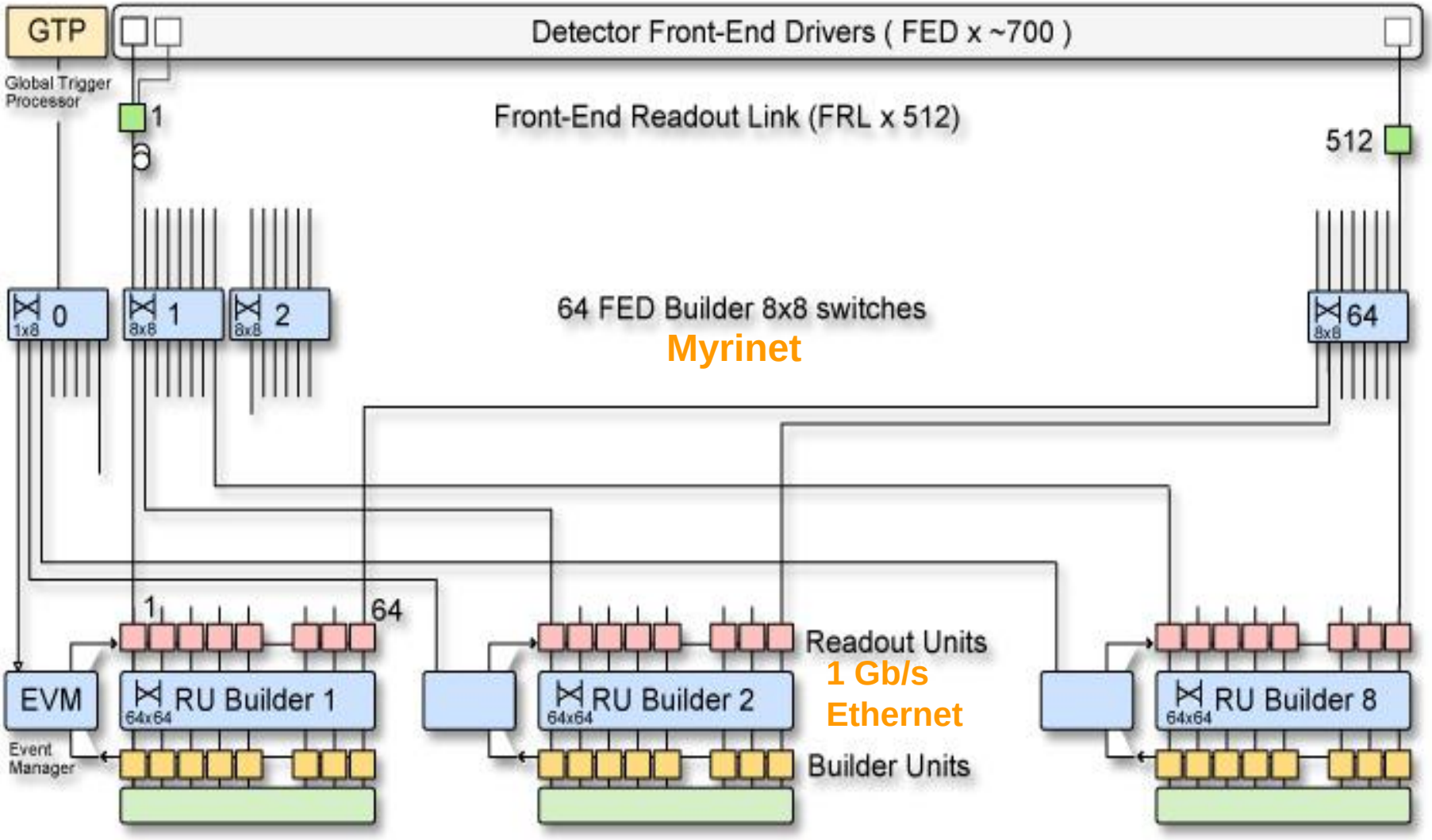
8 x 12.5 kHz

Bandwidth: 8 x 12.5 GB/s
Technology: 1 Gb/s Ethernet

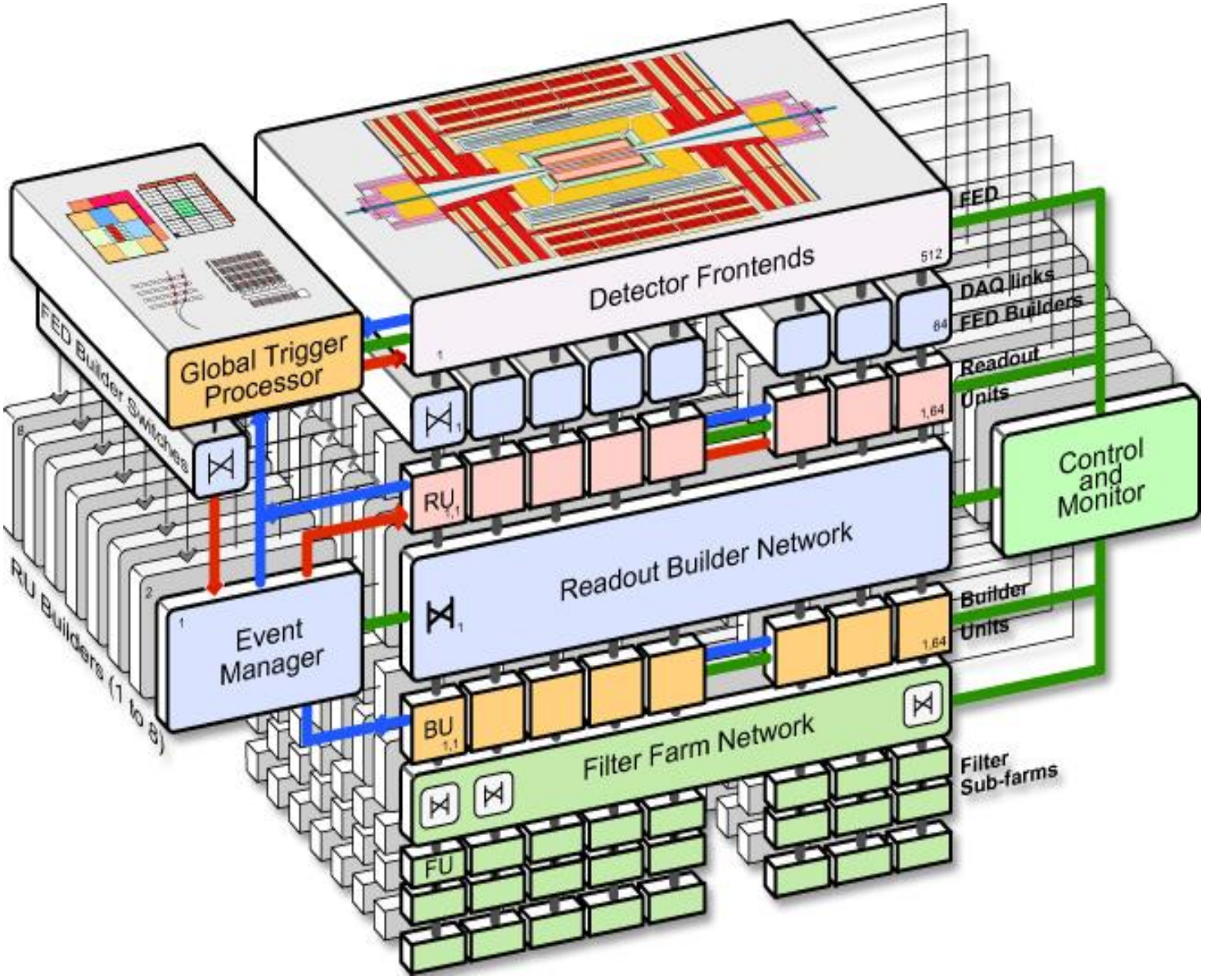
processing farm

300 Hz

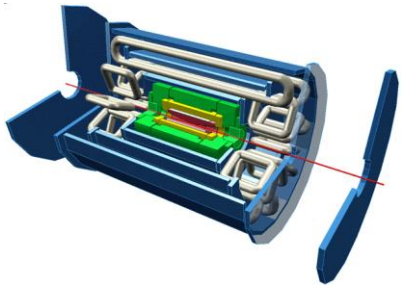
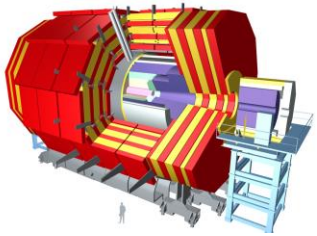
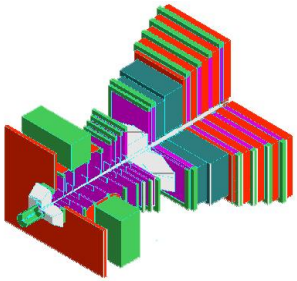
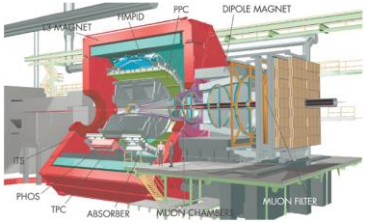
EVB CMS: 2 stages



CMS: 3D - EVB

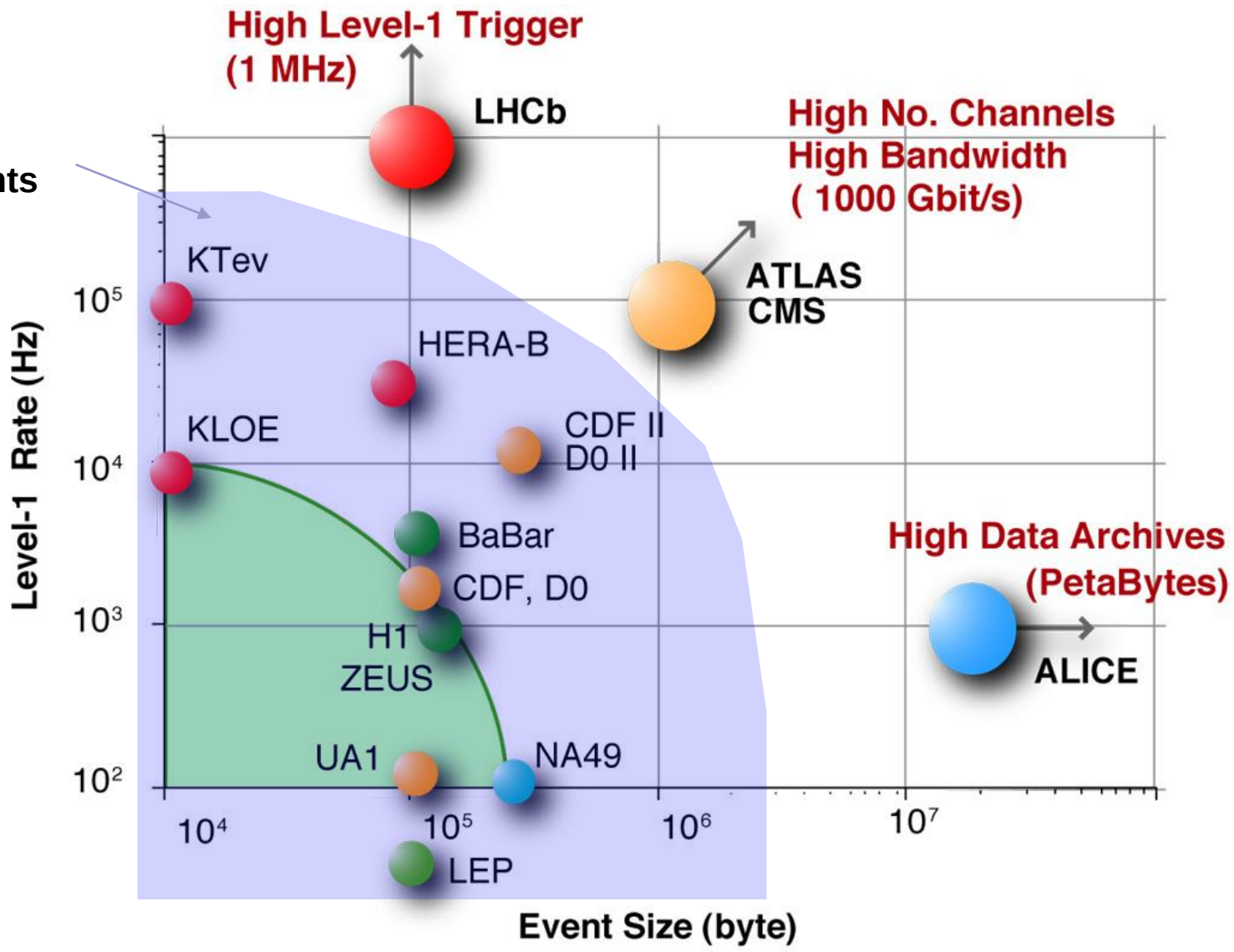

Detector Frontends
512
FED
DAQ links
FED Builders
Readout
Units
64
1,64
Global Trigger Processor
FED Builder Switches
RU 1,1
Control and Monitor
Readout Builder Network
Event Manager
RU Builders (1 to 8)
Builder
Units
1,64
BU 1,1
Filter Farm Network
Filter Sub-farms
FU

Trigger/DAQ parameters (Run 1)

	No.Levels	Lvl 0,1,2	Event	Evt Build.	HLT Out (design)
	Trigger	Rate (Hz)	Size (Byte)	Bandw.(GB/s)	MB/s (Event/s)
	3	LV-1 10^5 LV-2 $6x10^3$	1.5 MB	5.25	300 (200)
	2	LV-1 10^5	1.0 MB Pb-Pb 1500MB/s	100	300 (200)
	2	LV-0 10^6	30 kB	40	60 (2 kHz)
	4	Pb-Pb 500	70 MB	2	1250 (100)

High Level Trigger

LHC experiments: Lvl 1 rate vs size



4) High-Level Trigger

High-level trigger

- Reconstruct event data – often with full off-line software (minus calibration)
- To save time: multiple levels in software
  - E.g. First look only at muon and calorimeter data
  - Only if this looks interesting, spend time on reconstructing tracks in the trackers
- Issue of coupling of Online Software and Offline Software
- Multi-core machines
  - Run 1 or 2 copies of filter process per core
  - Processes forked from a mother-process to benefit from Copy-on-Write

High-level trigger farm sizes

Number of..	Boxes	CPU cores	Filter procs	Logical Grouping
ALICE	~ 200	~ 5000 (1)	~ 3000	
ATLAS	~ 1600	~ 17000	1 per core (2)	49 Racks
CMS	~ 1600	~ 16000	~ 35000 (2)	O(20)BUs in 8 slices
LHCb	~ 1600	~ 16000	~ 30000 (2)	57 Racks

(1) 2300 CPU cores + 54 FPGA + 64 GPU cards (estimated to 100-200% of the CPU)
(2) Overcommitment if hyper-threading is supported by worker node

5) Storage to disk

Local mass-storage

- Task
  - store a few days worth of data on-site.
  - handle peak recording rates (higher than network bandwidth to outside)
  - keep data safe
- Implementation
  - Disk arrays (Storage Area Network)
  - Mostly fiber channel (ATLAS: direct)

	ATLAS	CMS	ALICE	LHCb
GB/s max	1.6	0.8 pp 2 HI	2.2	0.25
space	160 TB	225 TB	610 TB	120 TB
raid	RAID-5	RAID-6	RAID-6	RAID-6
# writers	9	16	80+	6

ALICE Storage System

Run Control and Monitoring

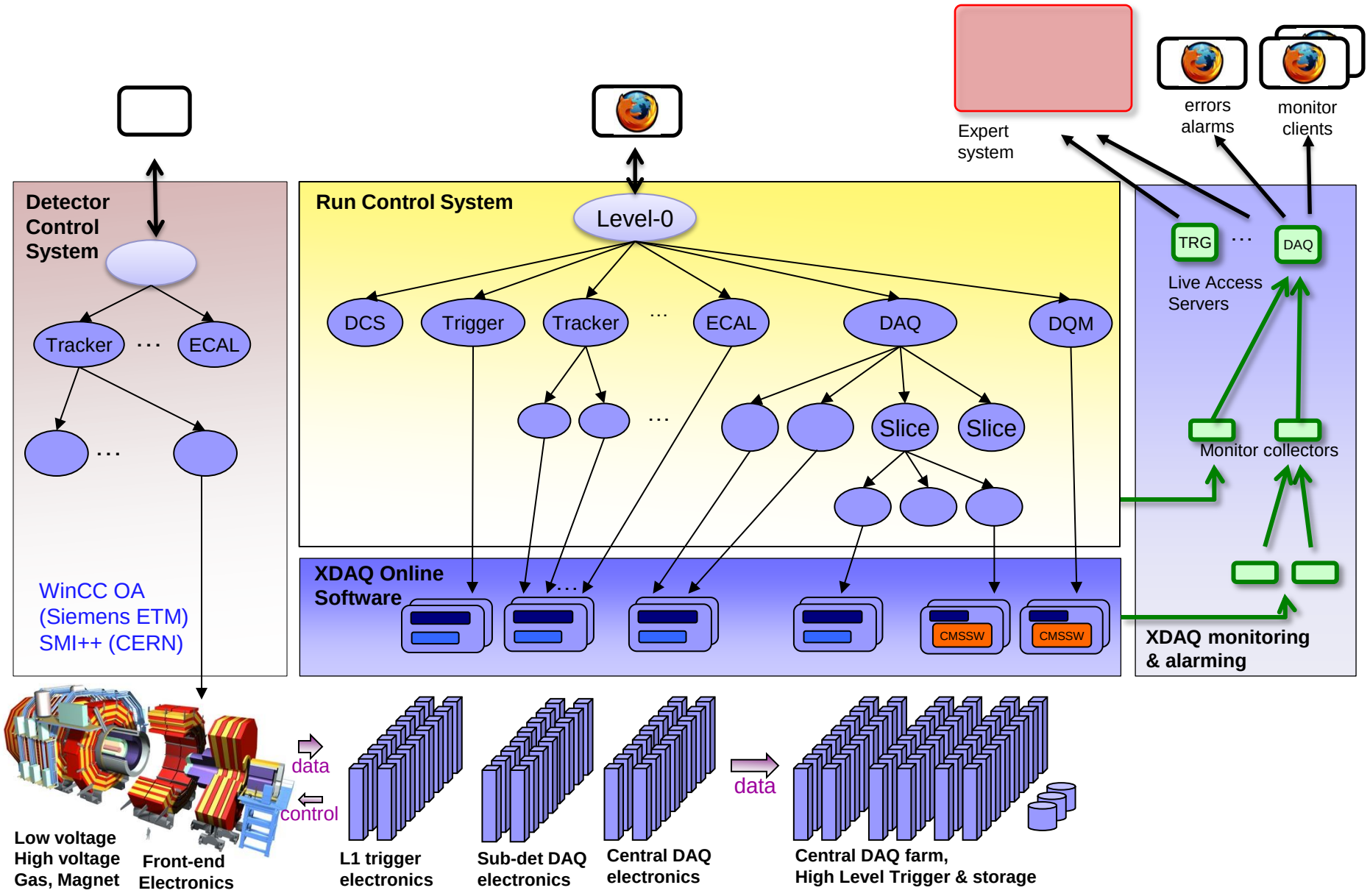
Run control tasks

- Start, configure and control O(10000) processes on farms of O(1000) machines
- Parallelize or execute in the right order according to dependencies
- Configure and monitor frontend electronics (millions of channels)
- Database access
- Graphical user interface to be used by non-expert shifters (integrated guidance)
- Automation

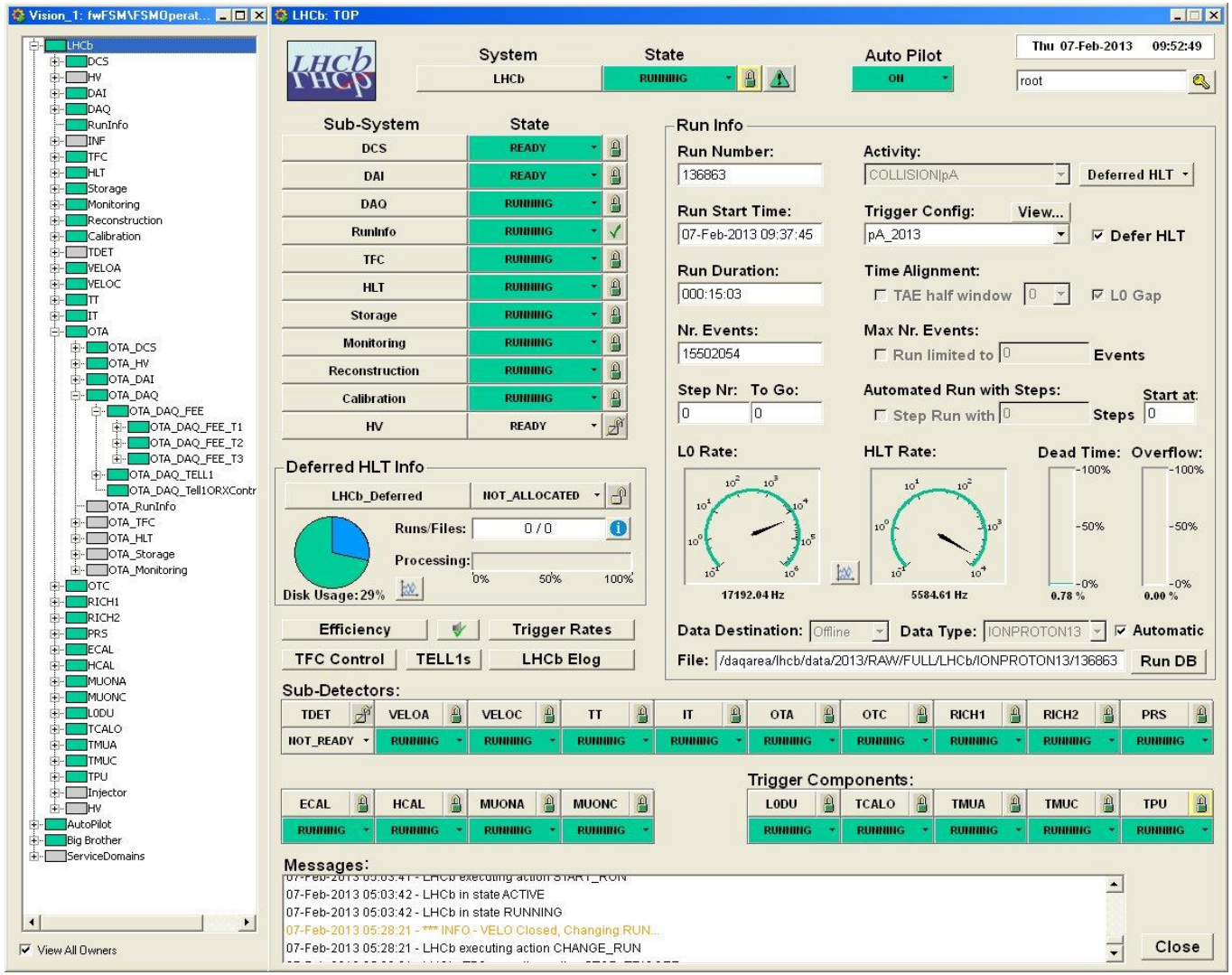
Technologies used

- Communication
  - CORBA (ATLAS)
  - HTTP/SOAP (CMS)
  - DIM (LHCb, ALICE)
- Behavior and Automation
  - SMI++ (Alice, LHCb)
  - CLIPS (ATLAS)
  - RCMS Web Applications / Java (CMS)
- Process control
  - Based on XDAQ online software, CORBA
  - FMC/PVSS

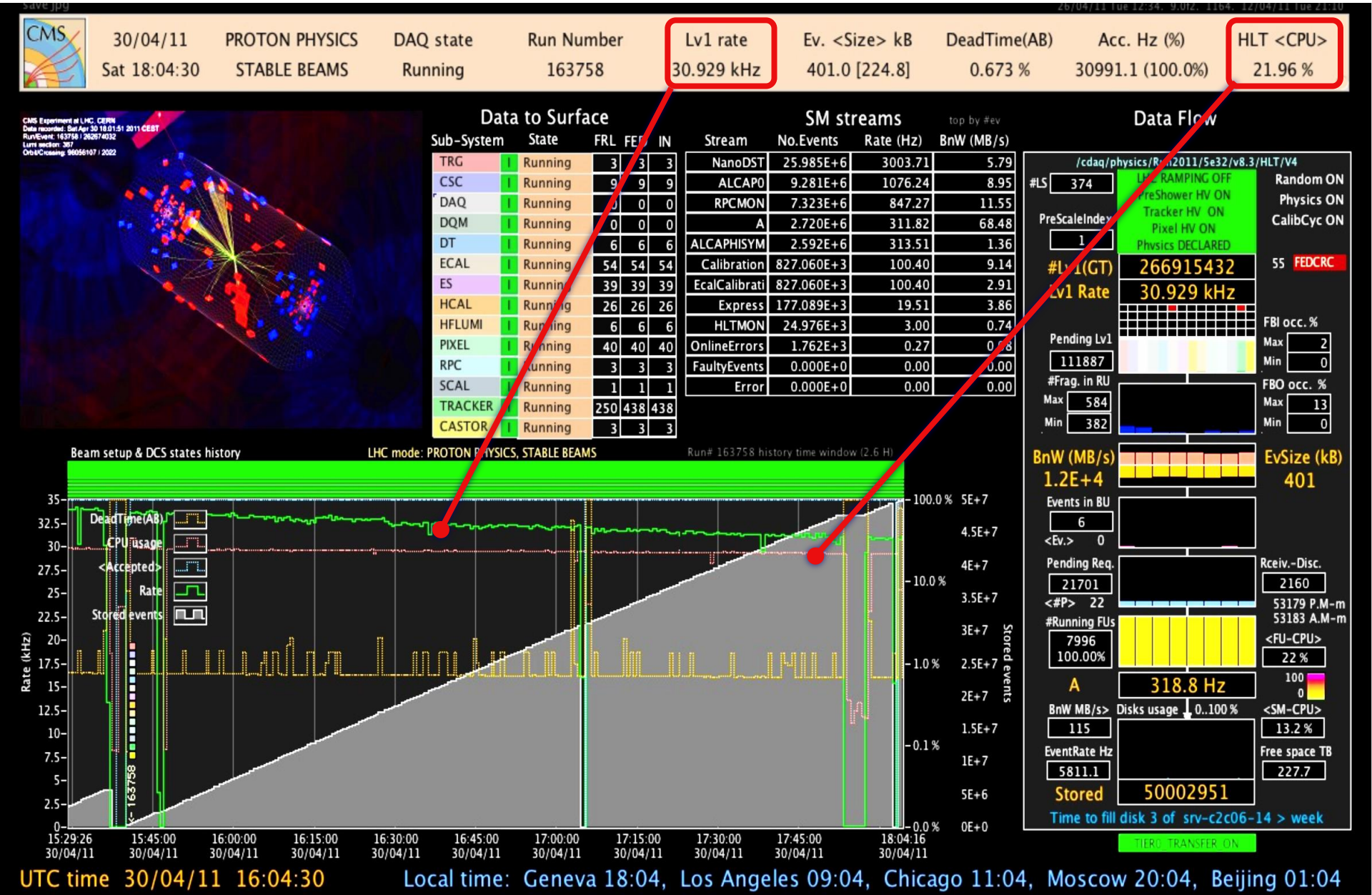
Run Control / Monitoring /Online software (CMS)



Run Control GUI (LHCb)



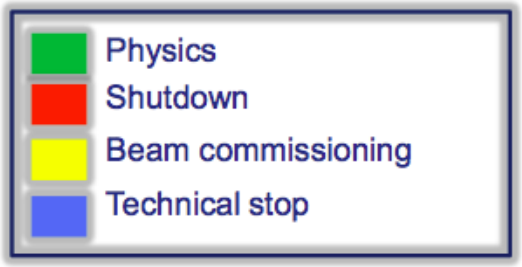
Monitoring (CMS)

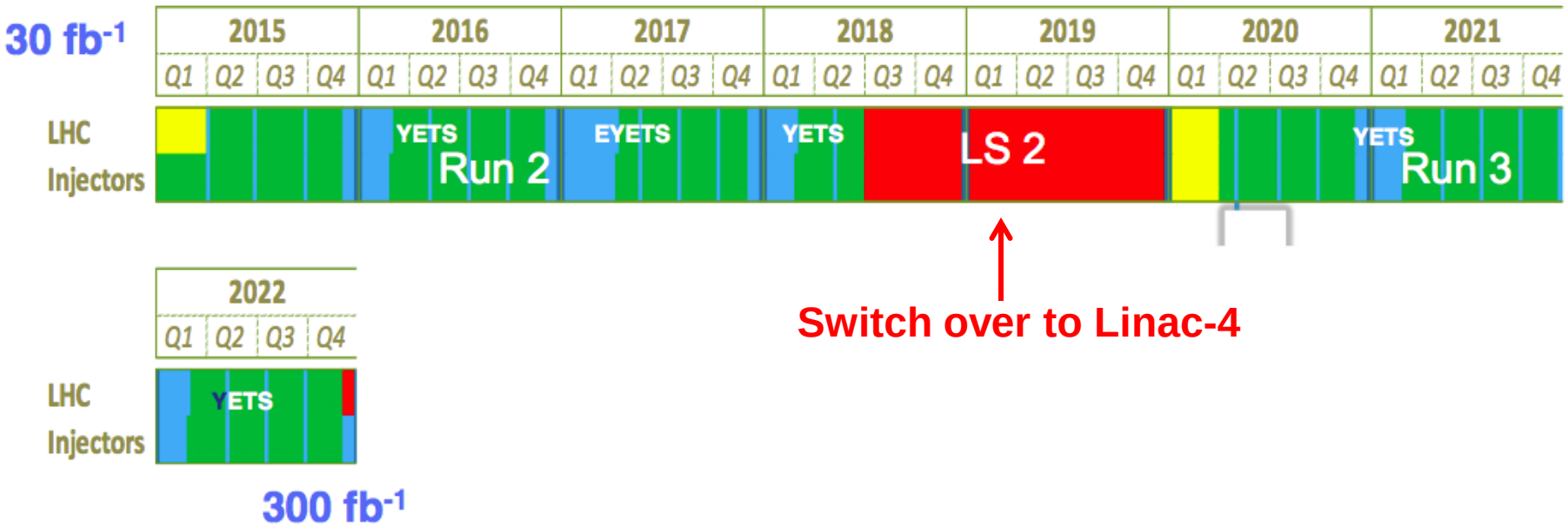


Upgrades.
around the corner
and further away

LHC plans

Run-2 : 13 TeV center-of-mass energy
Targeting 40 fb^{-1} / year
higher pile-up





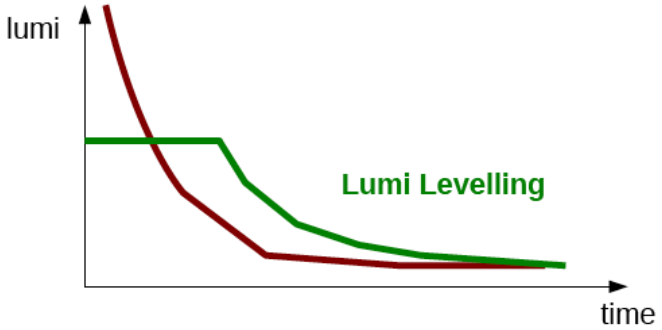
Switch over to Linac-4

(Extended) Year End Technical Stop: (E)YETS

LS3 : HL-LHC installation

Upgrade of the LHC

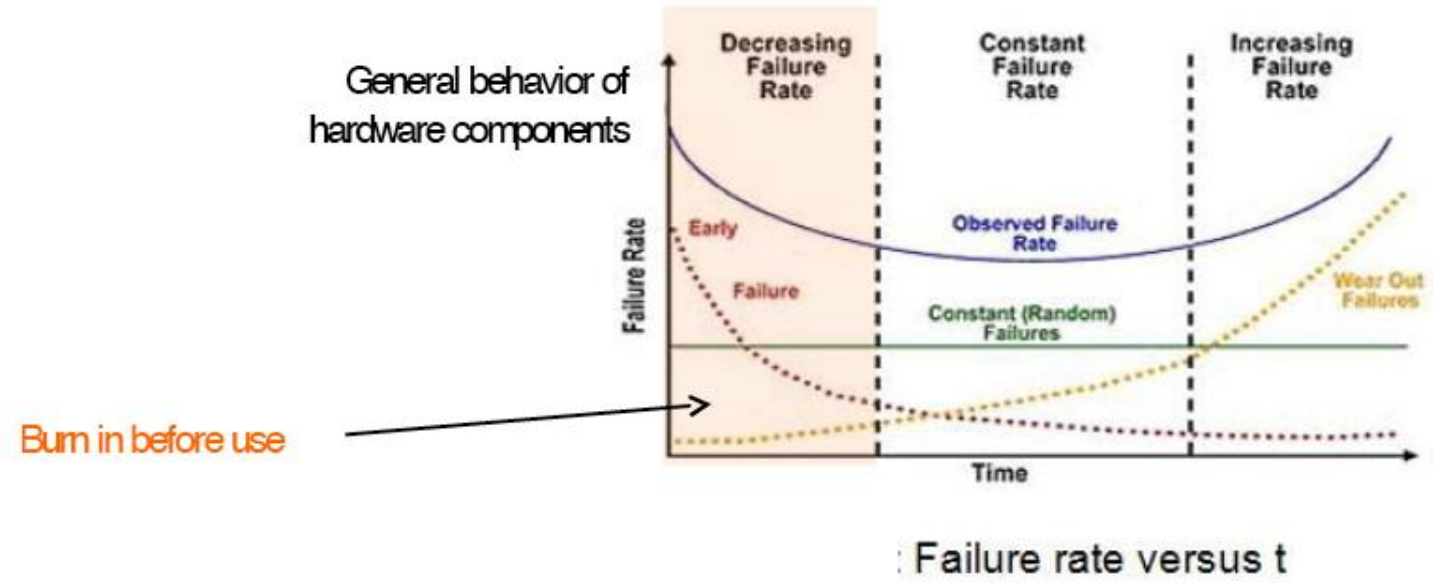
- Event sizes increase with the number of underlying events
  - LHC experiments were designed for a pile-up of 24 (with some margin)
  - In 2012 we had a pile-up of > 30
  - After LS1 we must be ready for a pile-up of 50
    - LHC may do “lumi-levelling” to keep pile-up below 50
    - Beams separated at the start of the fill



Plan for startup after LS1 (2015) →

BX spacing [ns]	Beam current [x10^{11} e]	Emittance [µm]	Peak Lumi [x $10^{34}cm^{-2}s^{-1}$]	Pileup
25	1.15	3.5	0.92	21
25	1.15	1.9	1.6	43
50	1.6	2.3	0.9-1.7	40-76
50	1.6	1.6	2.2	108

Hardware life-time … another reason to upgrade

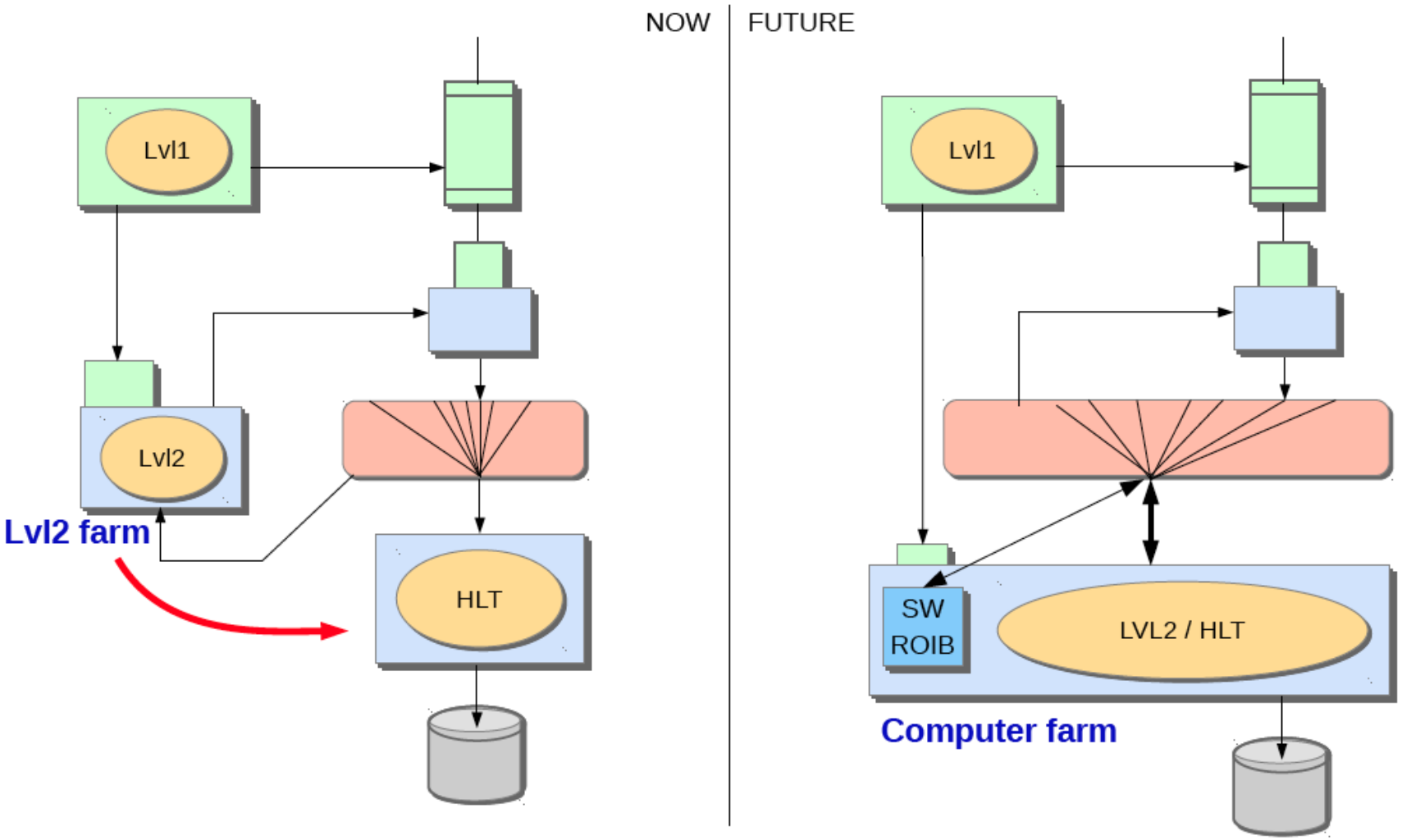


: Failure rate versus t

- System reliability decreases after a few years
  - It makes sense to replace PCs and networking equipment every ~5 years
  - Can then also benefit from advances in technology
    (Get about a factor 10 higher performance according to Moore's Law)
- Custom hardware is usually kept much longer
  - At some point it also starts becoming unreliable …

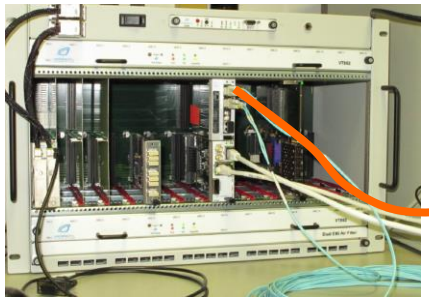
Upgrade highlights of the LHC experiments

ATLAS DAQ upgrade (for Run2)



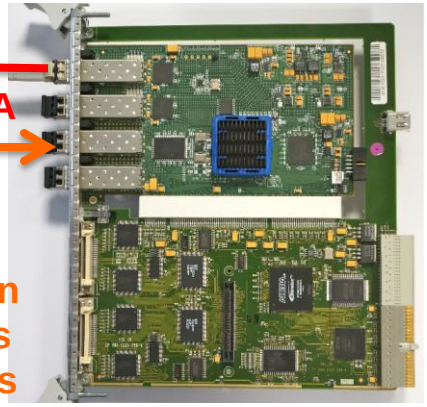
One integrated farm for Level-2 and High-Level Trigger.
Region-of-Interest principle is kept, but ROIs are built in software.

CMS DAQ-2
(for Run 2)



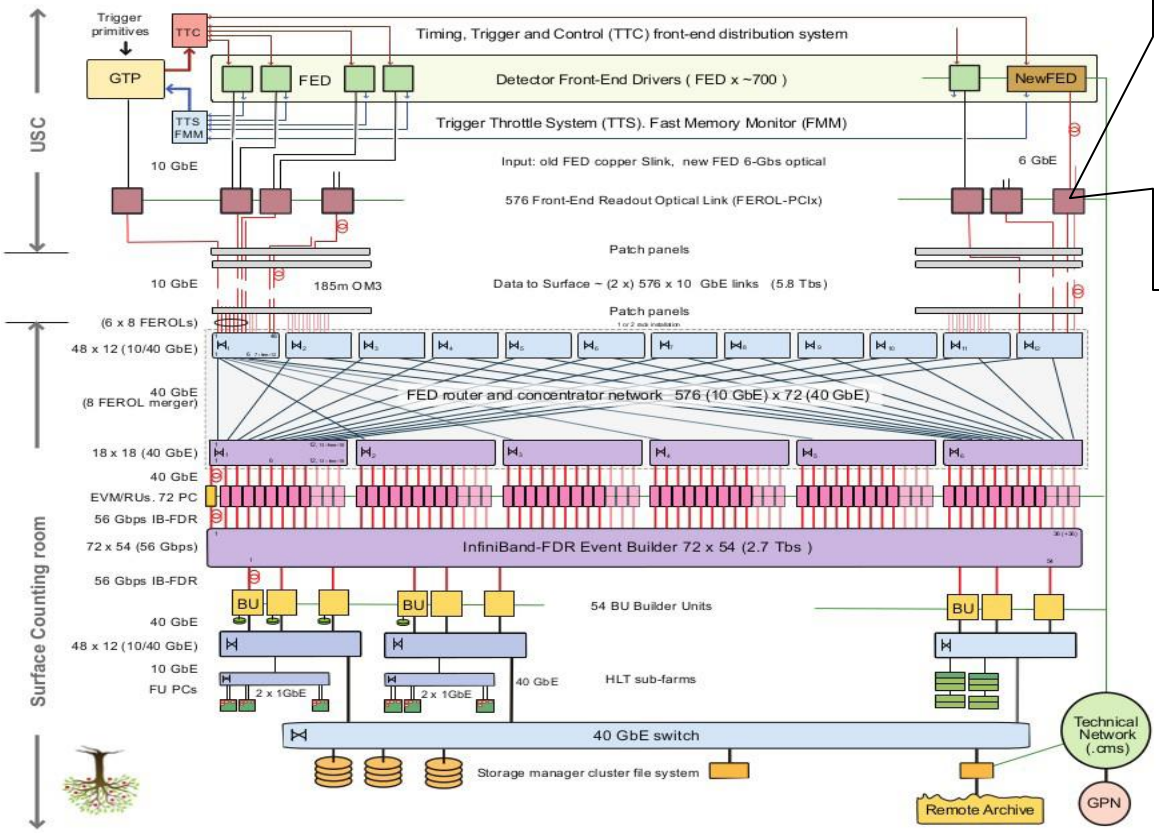
New Front-end drivers based on Micro-TCA

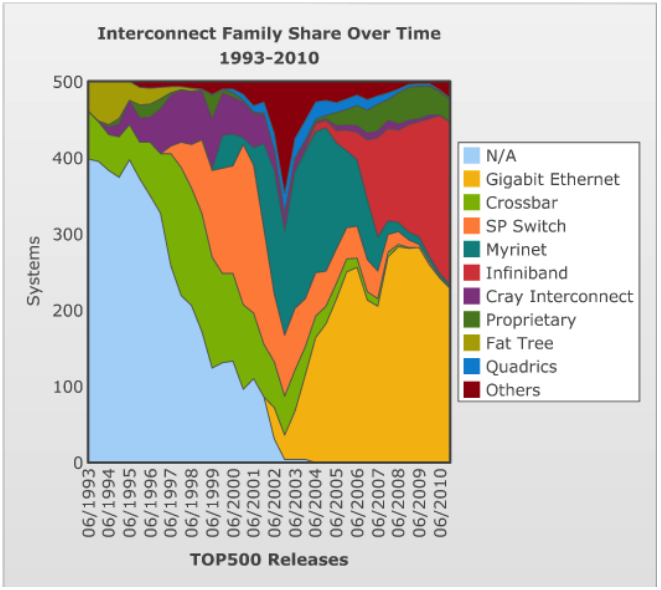
10 Gb/s simplified TCP/IP from FPGA



Optical SLINK-express in 2 x 6 Gb/s or 10 Gb/s

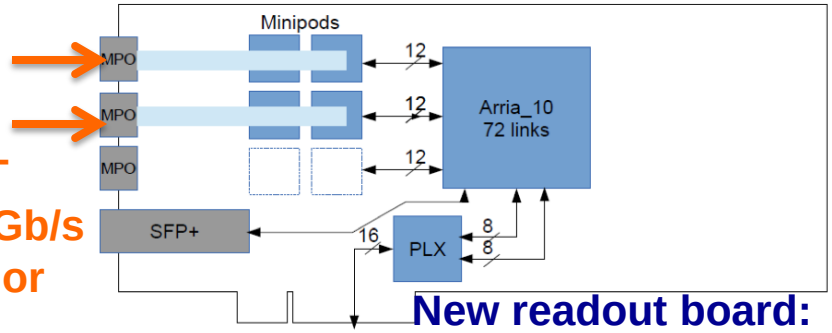
Myrinet NIC replaced by custom 10 Gb/s Ethernet NIC ("FEROL")





2-stage event builder based on 10/40 Gb/s Ethernet and 54 Gb/s FDR Infiniband being deployed this spring

LHCb to be read out at up to 40 MHz (for Run 3)?

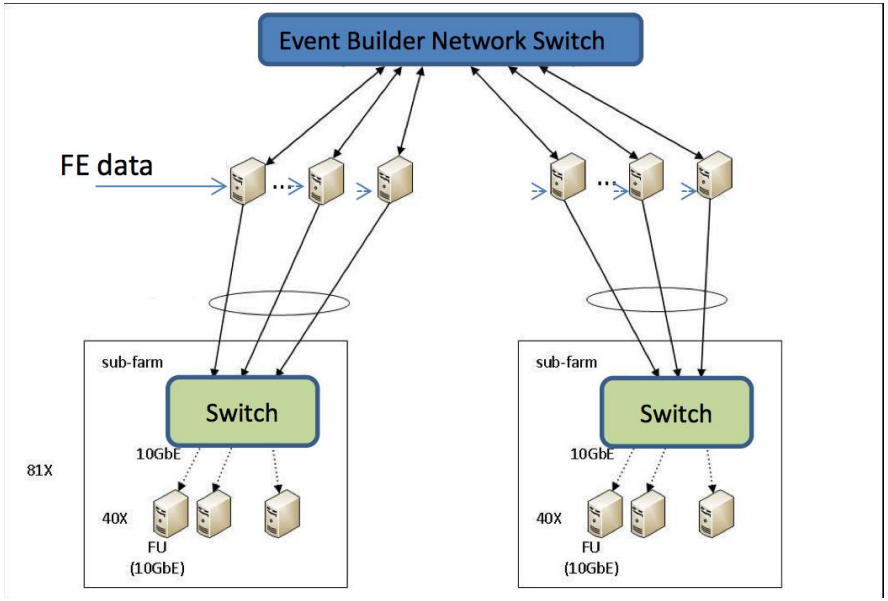


GBT
4.8 Gb/s
x24 or
X36
from detector

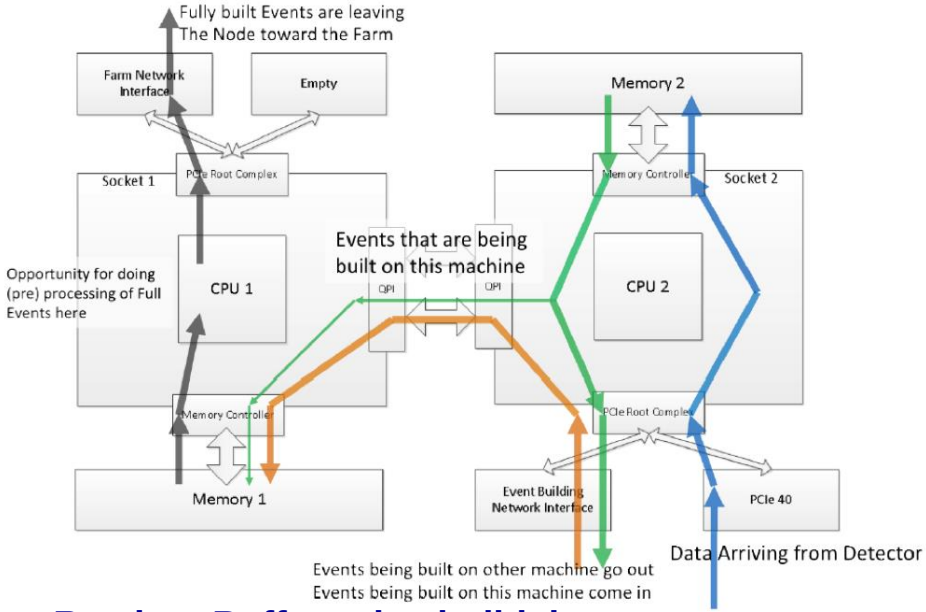
PCIe x 16

New readout board:
PCIe-40

Investigating DAQ in a container



Readout Buffers connected to event-builder and Filter-farm networks

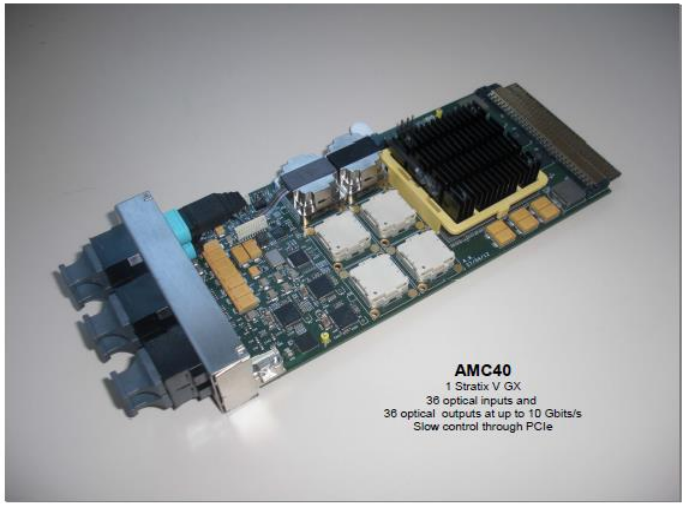


Readout Buffers also build the events

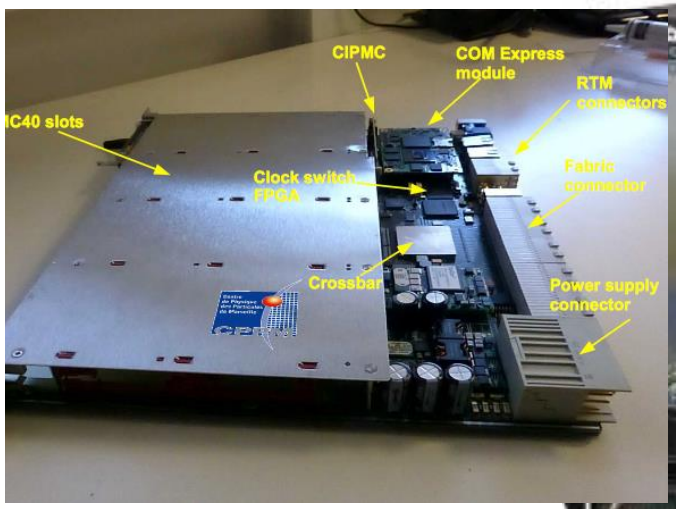
ALICE (for Run3)

- Read-out all Pb-Pb interactions at 50 kHz (= interaction rate). 1 TB/s event building.
- Continuous read-out of Time-Projection-Chamber
- No on-line filtering. Just on-line data reconstruction (clustering)
- Common cluster for Online + Offline - estimated 250000 cores

Investigating a Common Readout Unit based on Advanced TCA:
(also considering PCI-e)



Tell-40 AMC card
(originally developed by LHCb)
in: GBT x 24
Out: DDL3xn (commercial 10 Gbs)



ATCA carrier
carrying 4 Tell-40

10 ATCA shelves would be needed

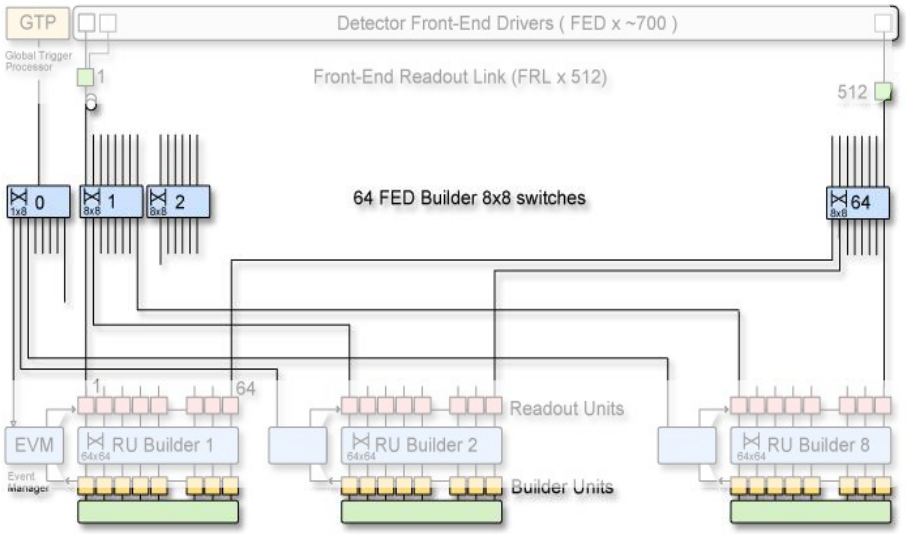
Summary

- DAQ is an interesting part of a HEP experiment
  - No DAQ system alike
- Exciting times ahead for the LHC DAQ systems
  - Upgrade projects in all experiments
- If you want to learn more …
  - ISOTDAQ school
    - Feb 2013, Budapest http://indico.cern.ch/event/274473/other-view?view=standard
    - Feb 2014, Rio (?)
  - DAQ@LHC workshop (March 2013)
    - http://indico.cern.ch/event/217480/other-view?view=standard

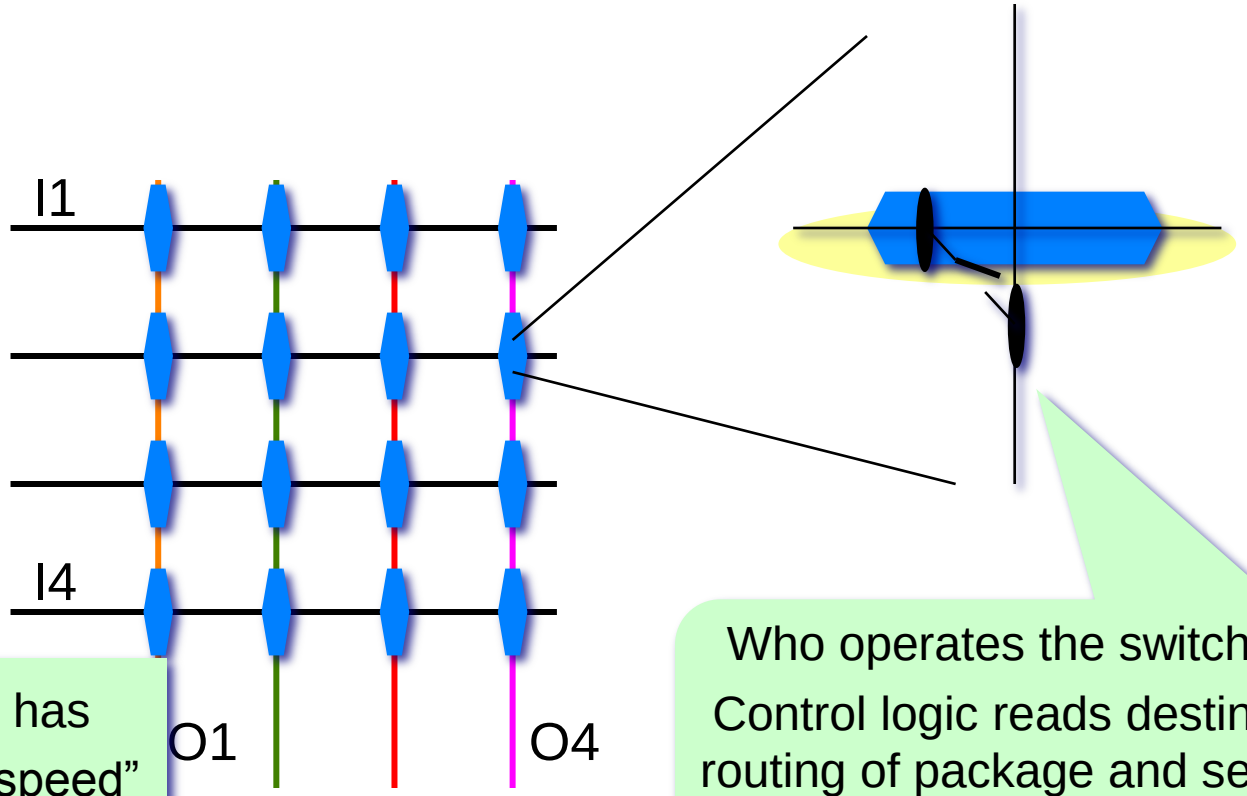
Backup

Example of network performance CMS Stage1 FED-Builder

- FED Builder functionality
  - Receives event fragments from approx. 8 Readout Links (FRLs).
  - FRL fragments are merged into "super-fragments" at the destination (Readout Unit).
- FED Builder implementation
  - Requirements:
    - Sustained throughput of 200MB/s for every data source (500 in total).
    - Input interfaces to FPGA (in FRL) -> protocol must be simple.
  - Chosen network technology: Myrinet
    - NICs (Network Interface Cards) with 2x2.5 Gb/s optical links (≈ 2x250 MB/s)
    - Full duplex with flow control (no packet loss).
    - NIC cards contain RISC processor. Development system available. Can be easily interfaced to FPGAs (custom electronics: receiving part of readout links)
    - Switches based on cross bars (predictable, understandable behavior).
    - Low cost!



Switch implementation: crossbar

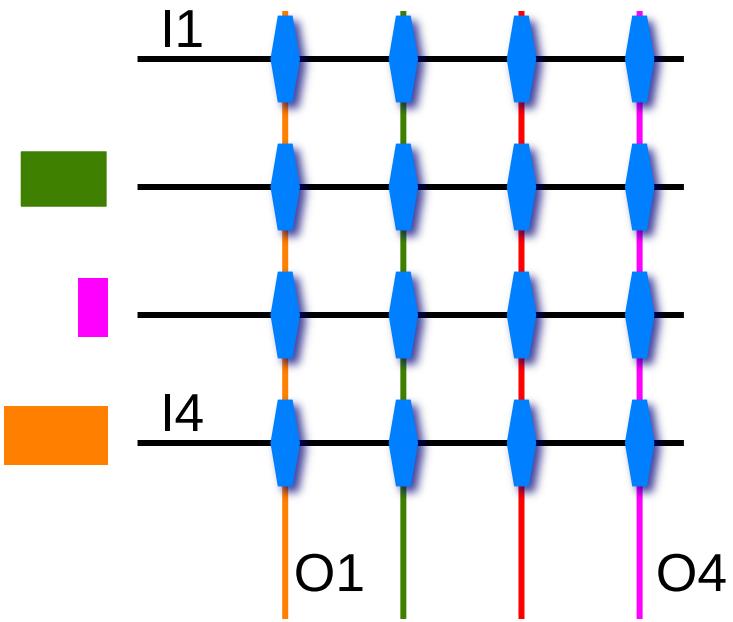


Every input / output has
A given max. “wire-speed”
(e.g. 2Gbits/sec).
Internal connections are
much faster!

Who operates the switches ?

Control logic reads destination routing of package and sets the switches appropriately.

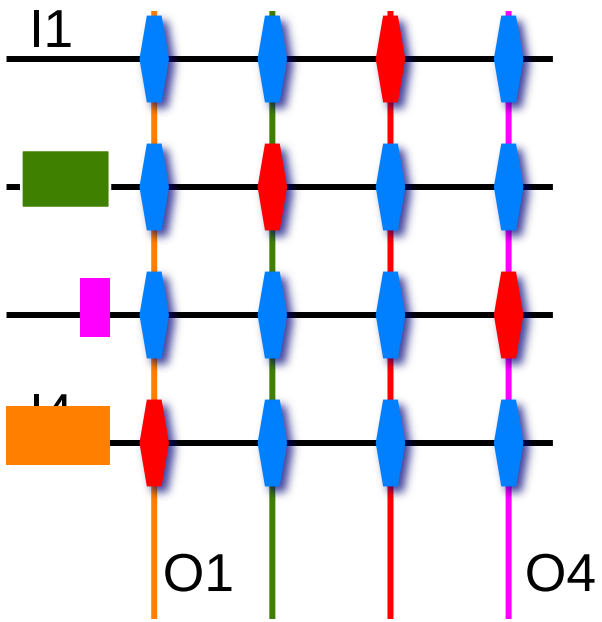
Switch implementation: crossbar



Paradise scenario:

All inputs want to send data to different destinations

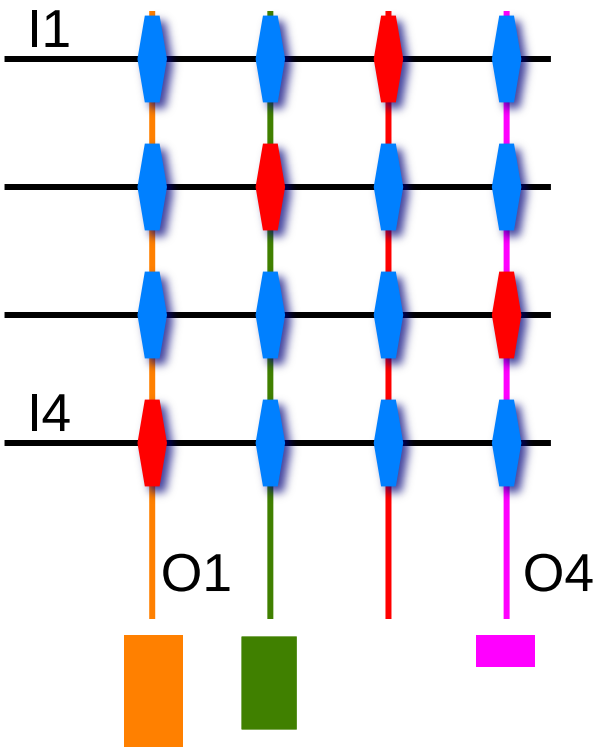
Switch implementation: crossbar



Paradise scenario:

No congestion, since every data package finds a free path through the switch.

Switch implementation: crossbar

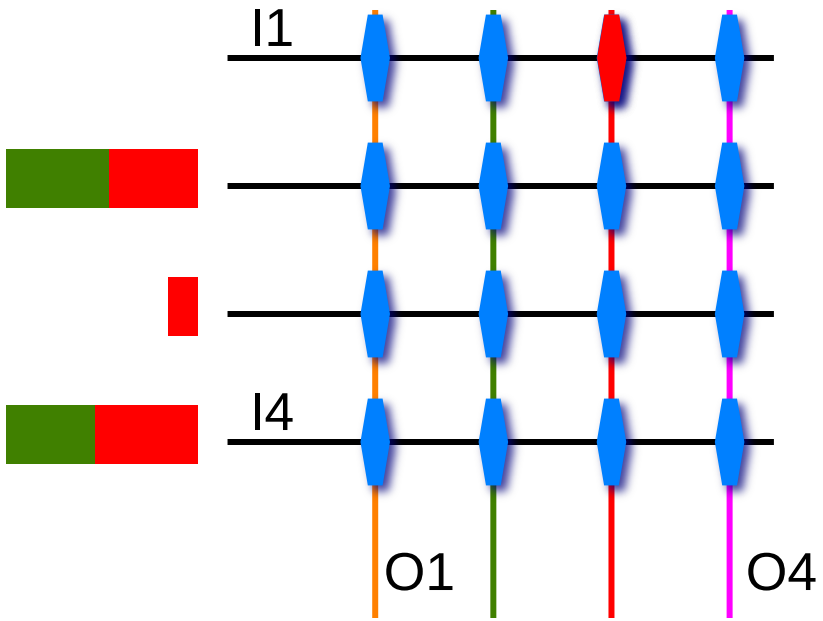


Paradise scenario:

Data traffic performs with "wire speed" of switch

Switch implementation

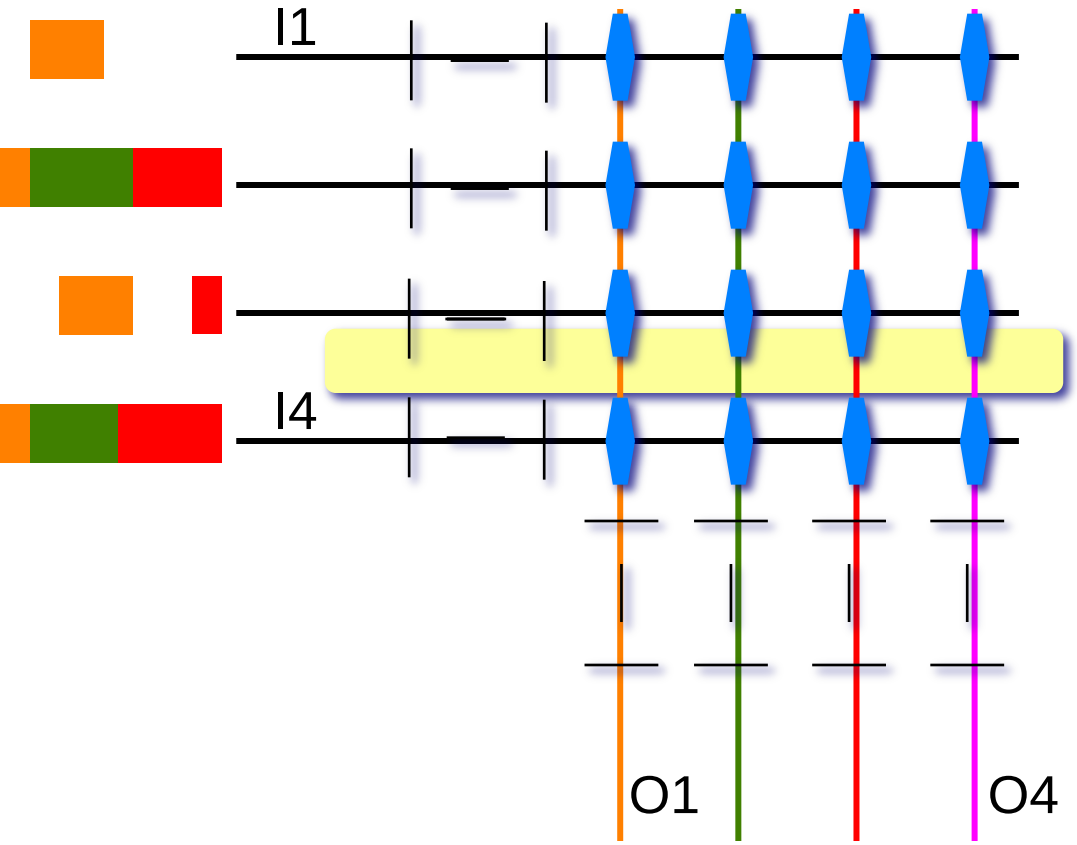
Crossbar switch: **Congestion in EVB traffic**



Only one packet at a time can be routed to the destination.
"Head of line" blocking

Switch implementation

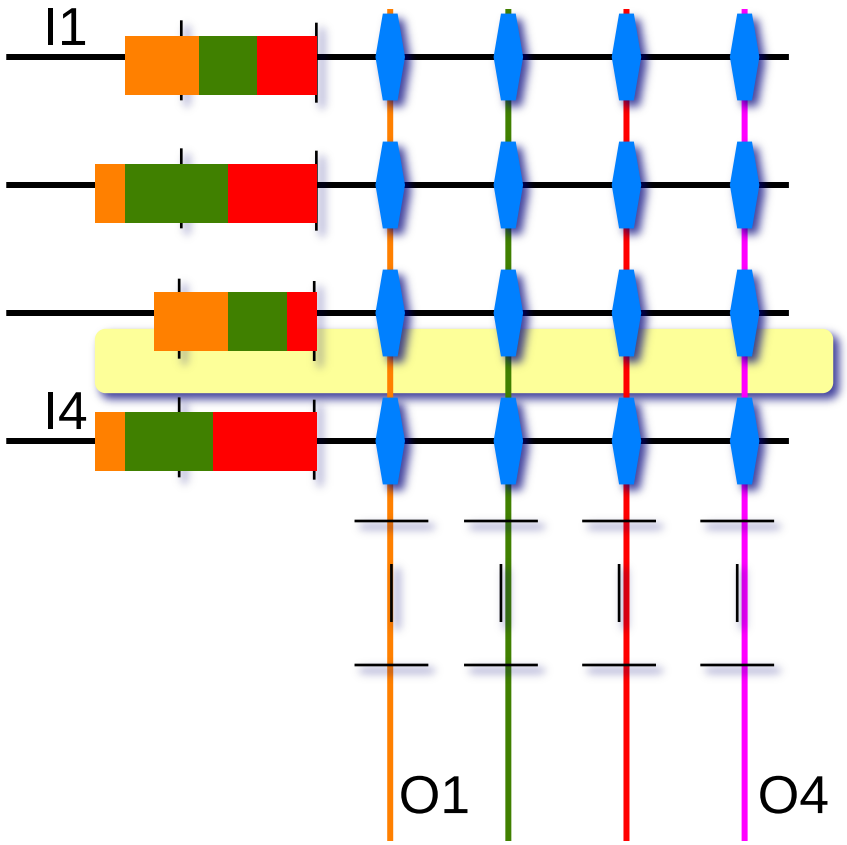
Crossbar switch: **Improvement : additional input FIFOs**



Fifos can “absorb” congestion …until they are full.

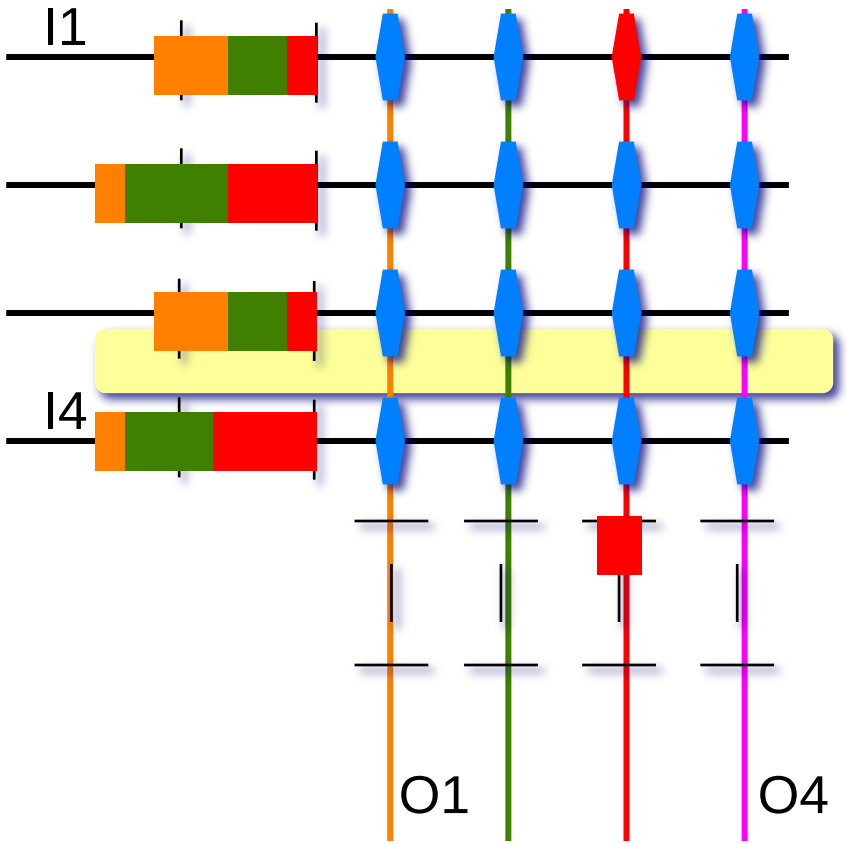
Switch implementation

Crossbar switch: **Improvement : additional input FIFOs**



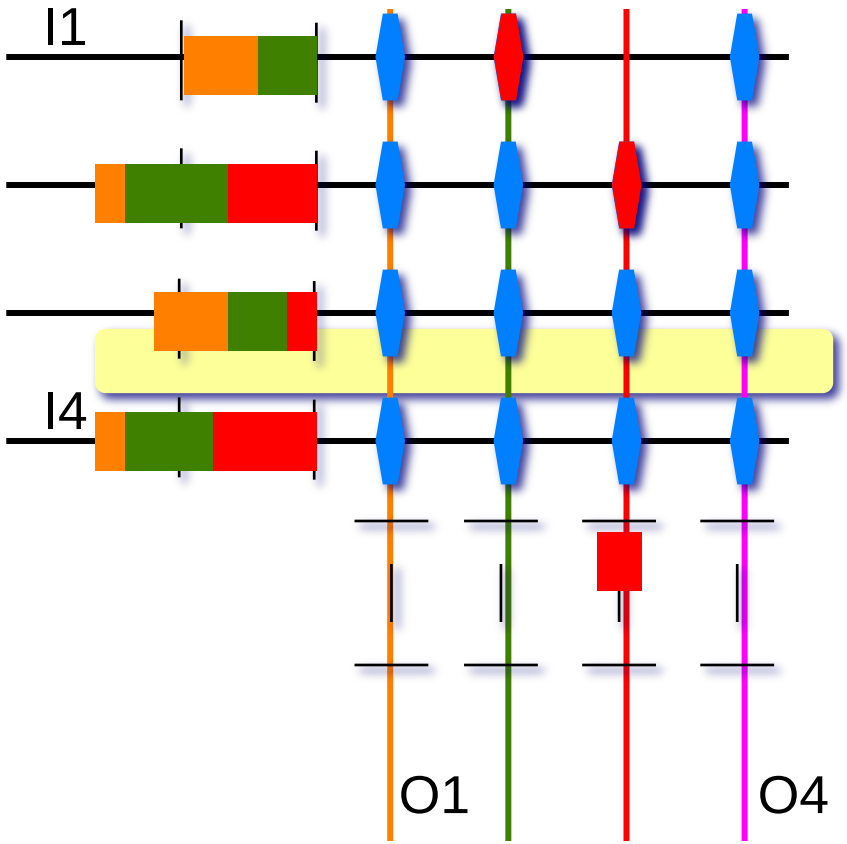
Switch implementation

Crossbar switch: **Improvement : additional input FIFOs**



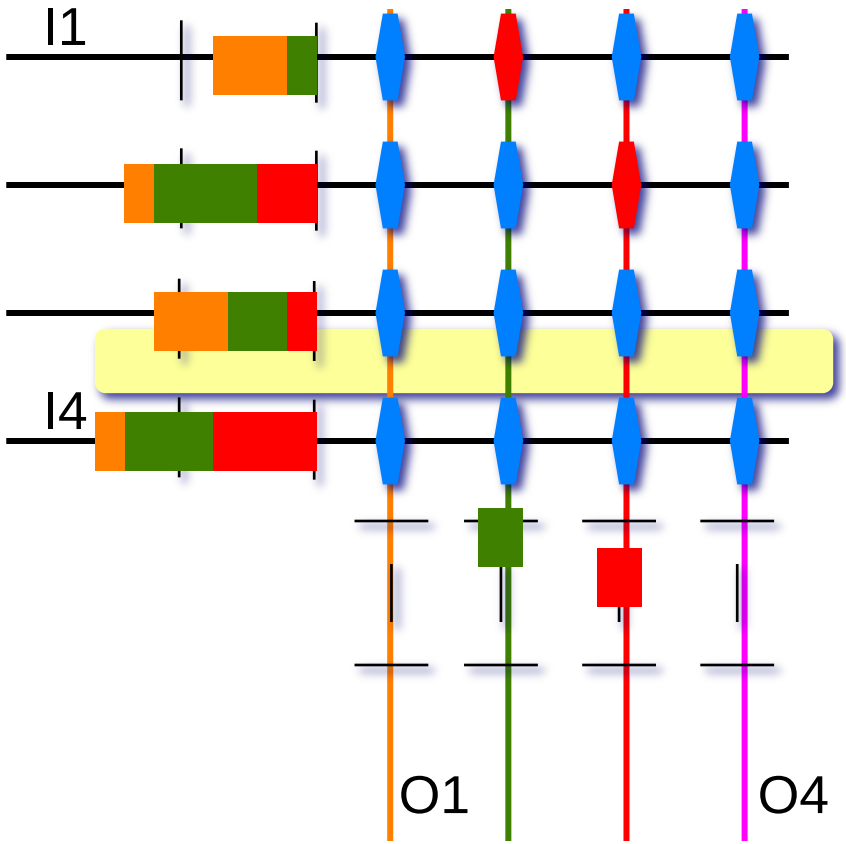
Switch implementation

Crossbar switch: **Improvement : additional input FIFOs**



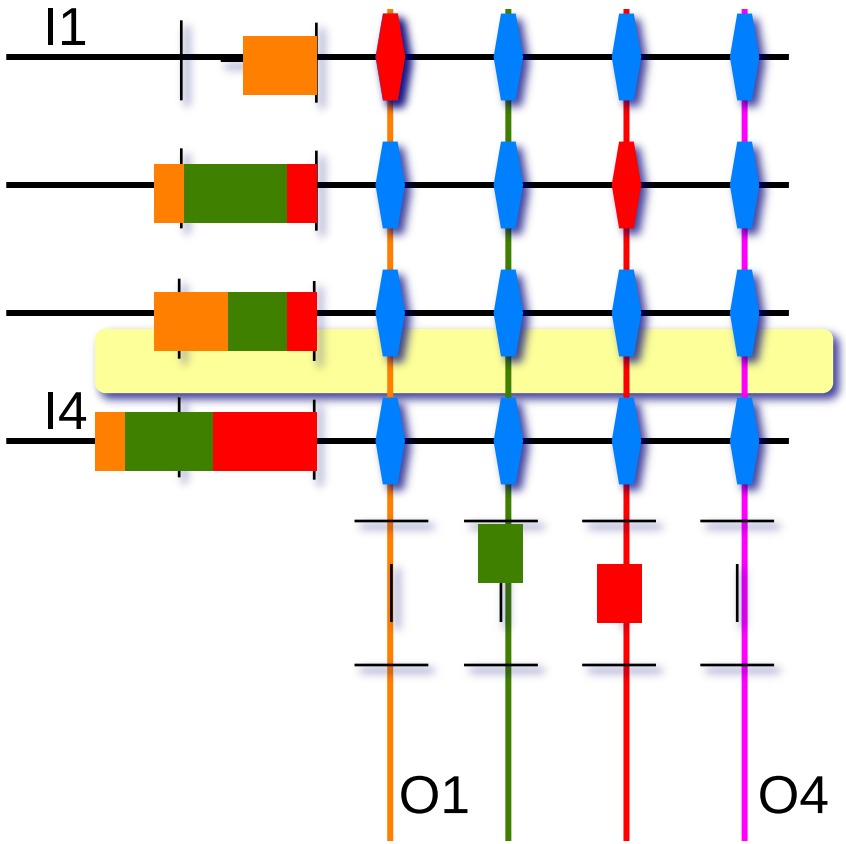
Switch implementation

Crossbar switch: **Improvement : additional input FIFOs**



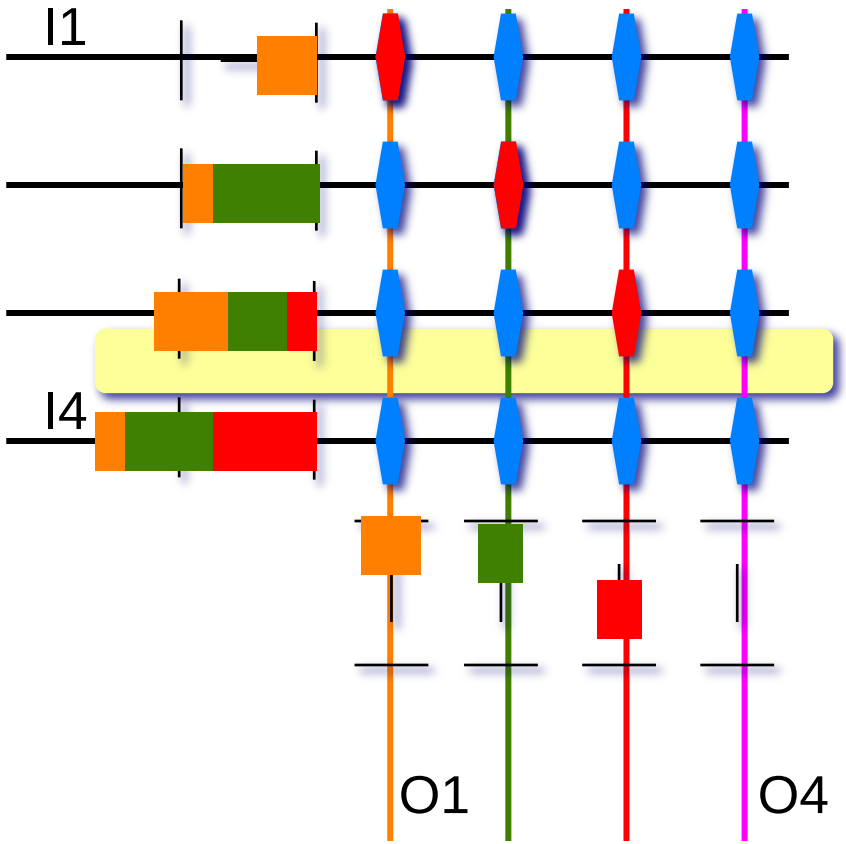
Switch implementation

Crossbar switch: **Improvement : additional input FIFOs**



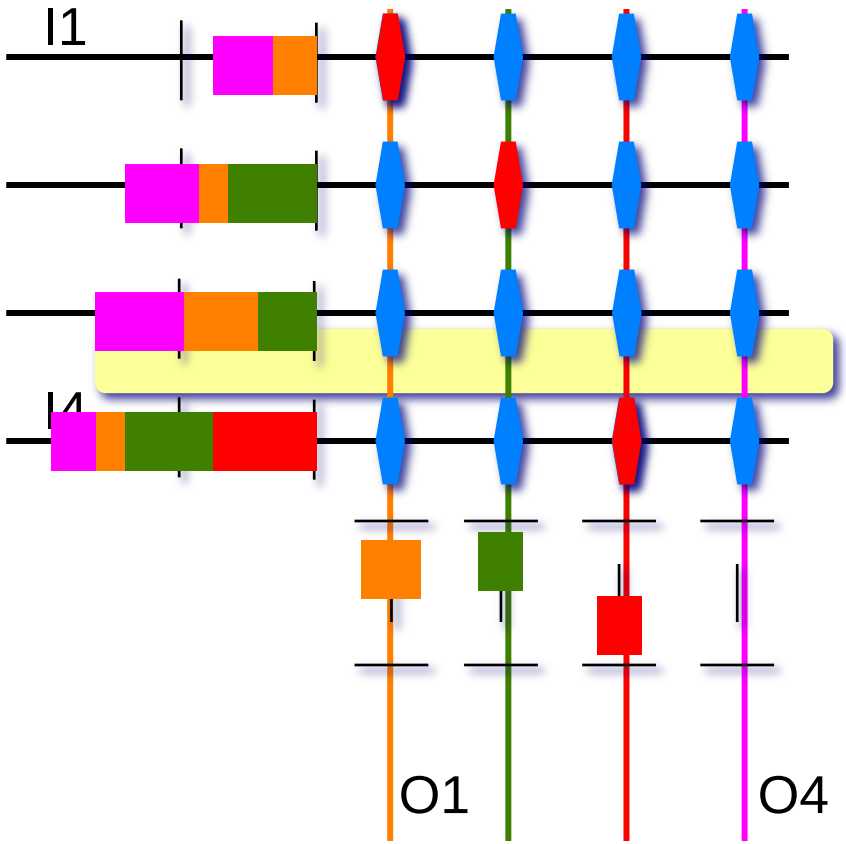
Switch implementation

Crossbar switch: **Improvement : additional input FIFOs**



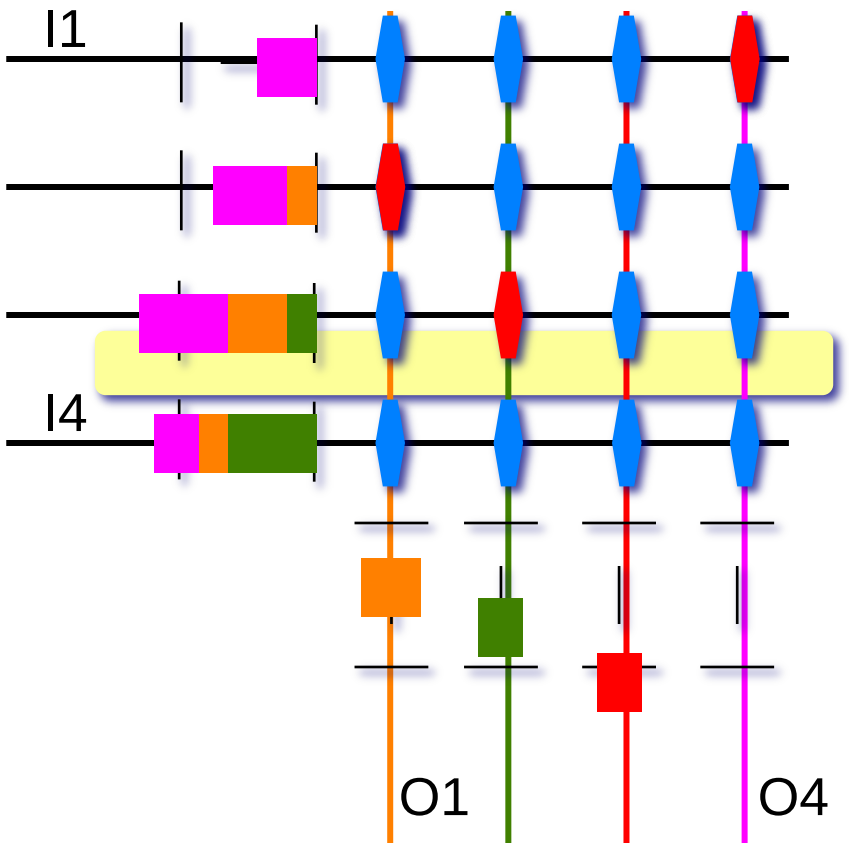
Switch implementation

Crossbar switch: **Improvement : additional input FIFOs**



Switch implementation

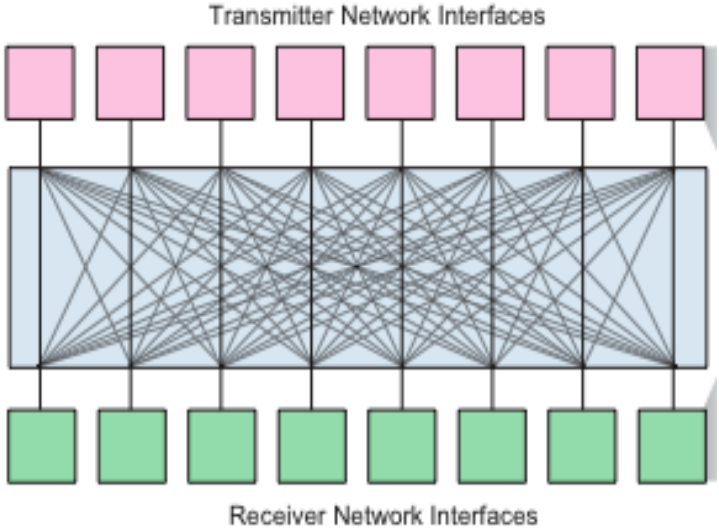
Crossbar switch: **Improvement : additional input FIFOs**



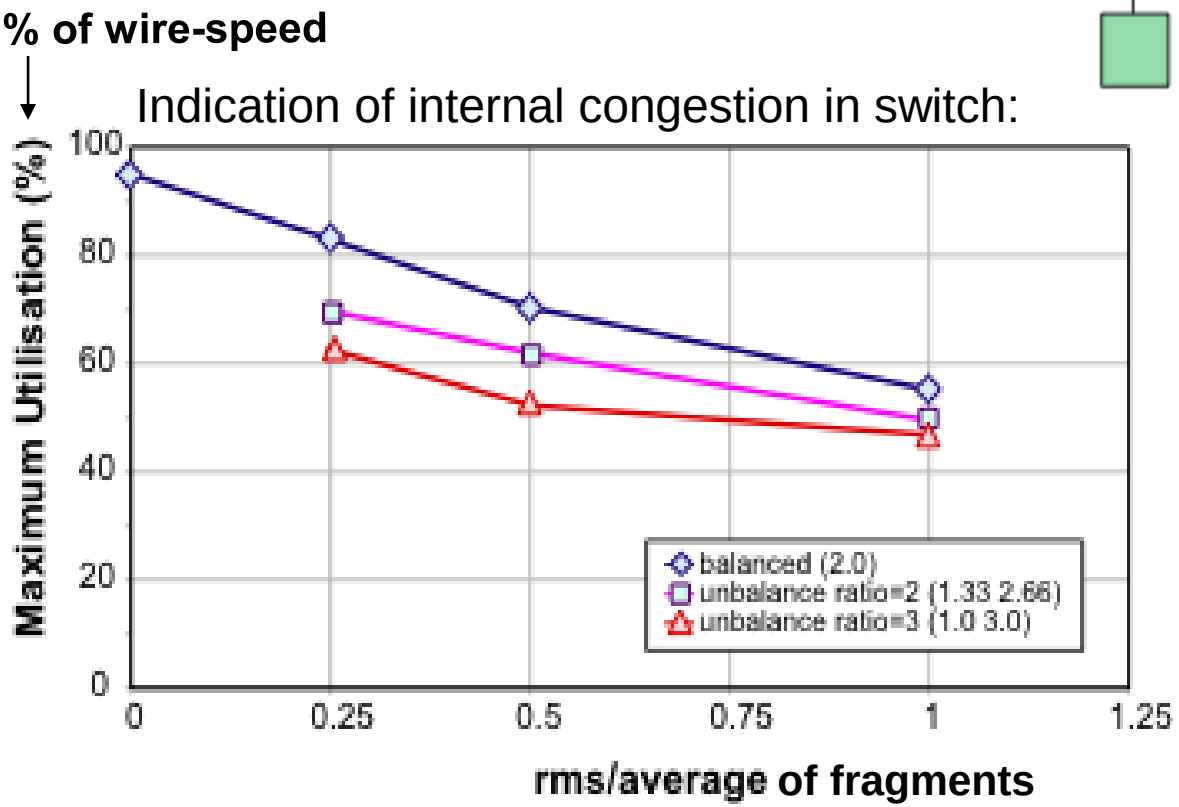
Still problematic:
Input Fifios can absorb data fluctuations until they are full.
How good it works depends on:
Fifos capacity
event size distribution
Internal speed of the switch
EVB traffic: blocking problem remains to some extend

Performance of "1 rail" FEDBuilder

Measurement configuration:
8 sources to 8 destinations



Indication of internal congestion in switch:



Measured switch utilization:
Blue: all inputs 2 kB avg
Magenta: 4 x 1.33 kB
4 x 2.66 kB
Red: 4 x 1 kB
4 x 3 kB

≈ 50 %

Conclusion: EVB traffic and switches

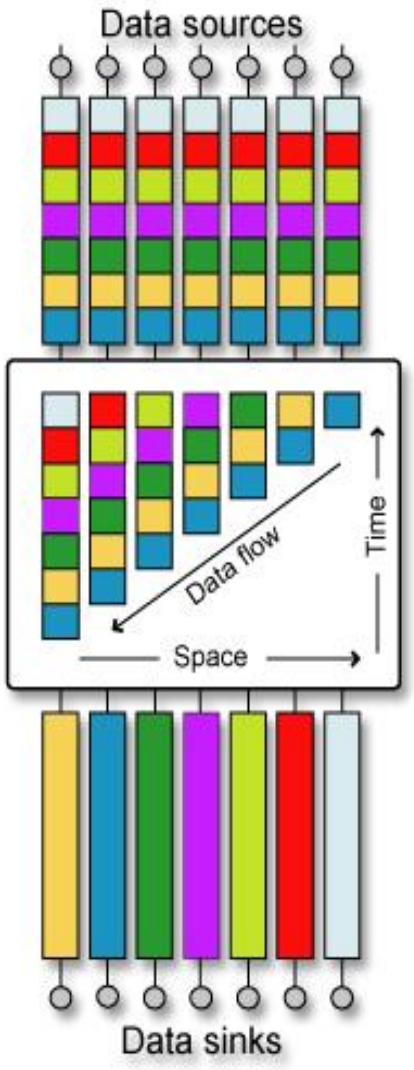
- EVB network traffic is particularly hard for switches
  - The traffic pattern is such that it leads to congestion in the switch.
  - The switch either "blocks" (= packets at input have to "wait") or throws away data packets (Ethernet switches)

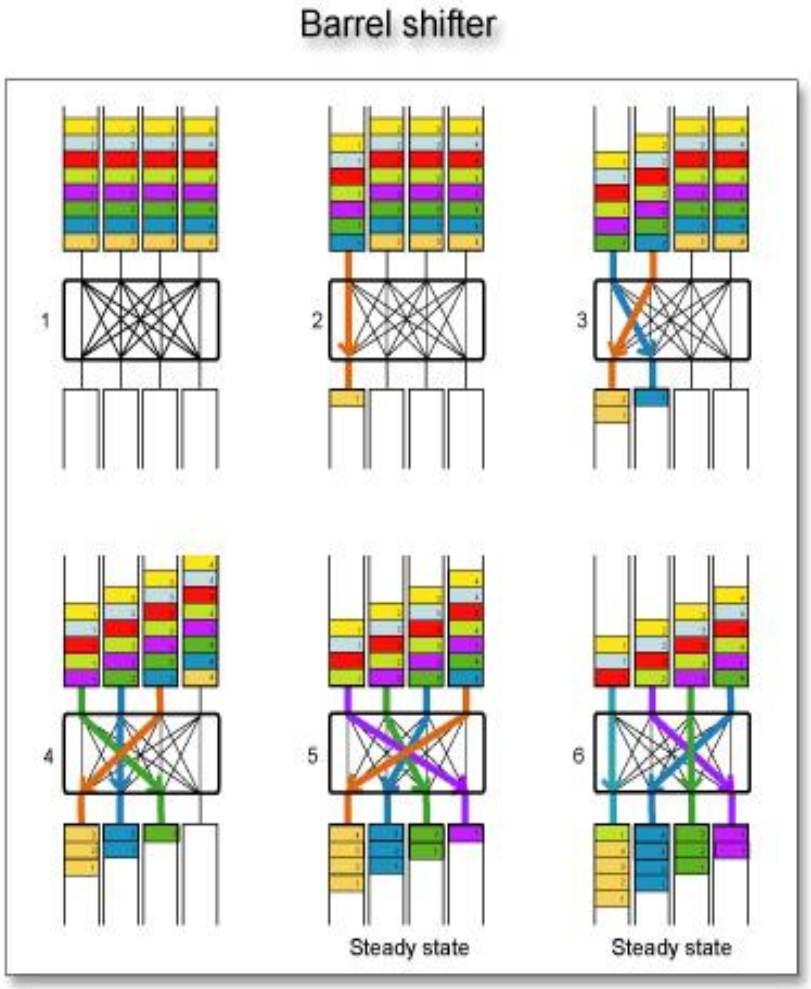
How to deal with this ???

→ 2 possible solutions

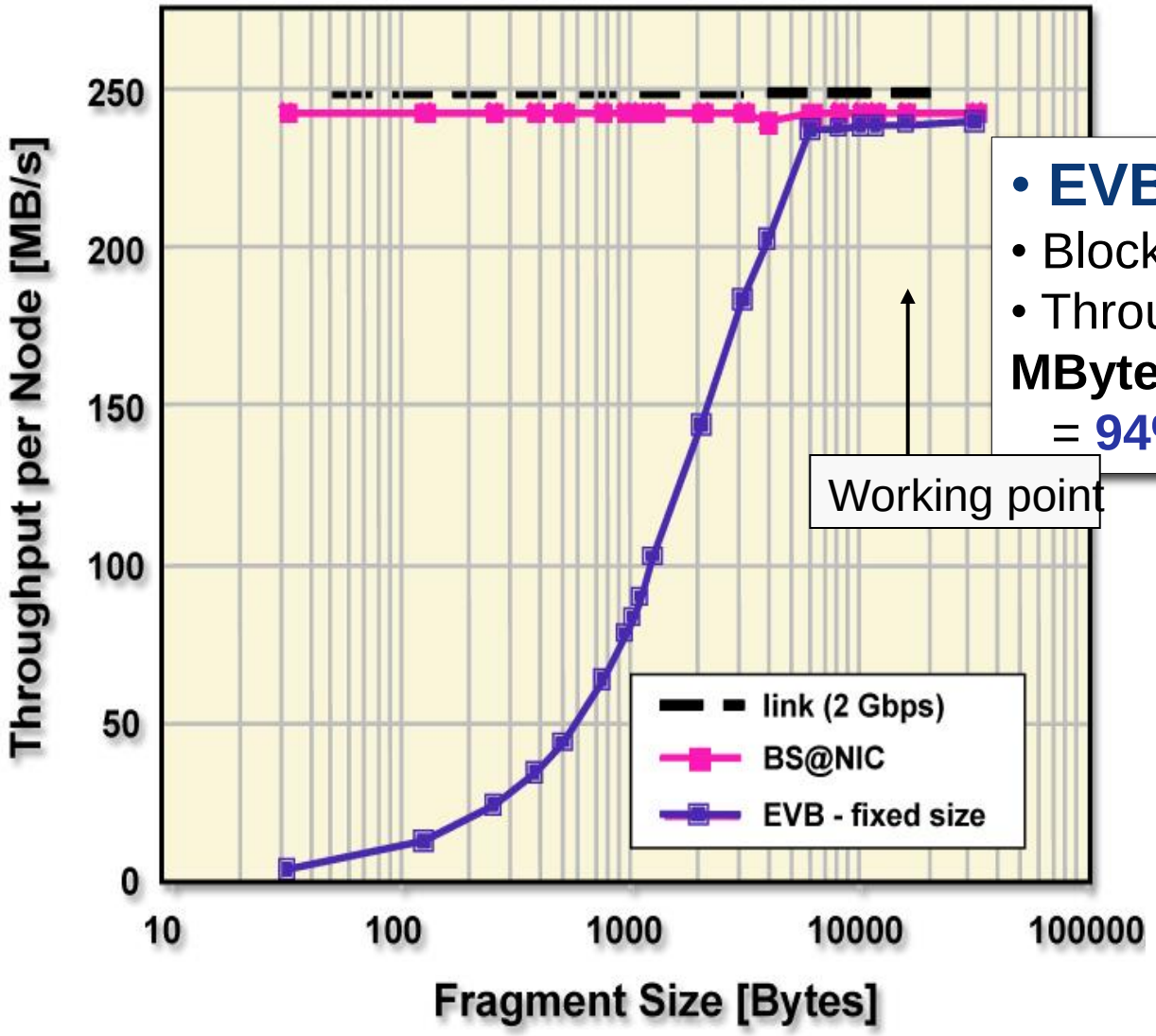
1st : the clever solution: traffic shaping

Example: Barrel Shifter





Barrel Shifter: Measured Performance

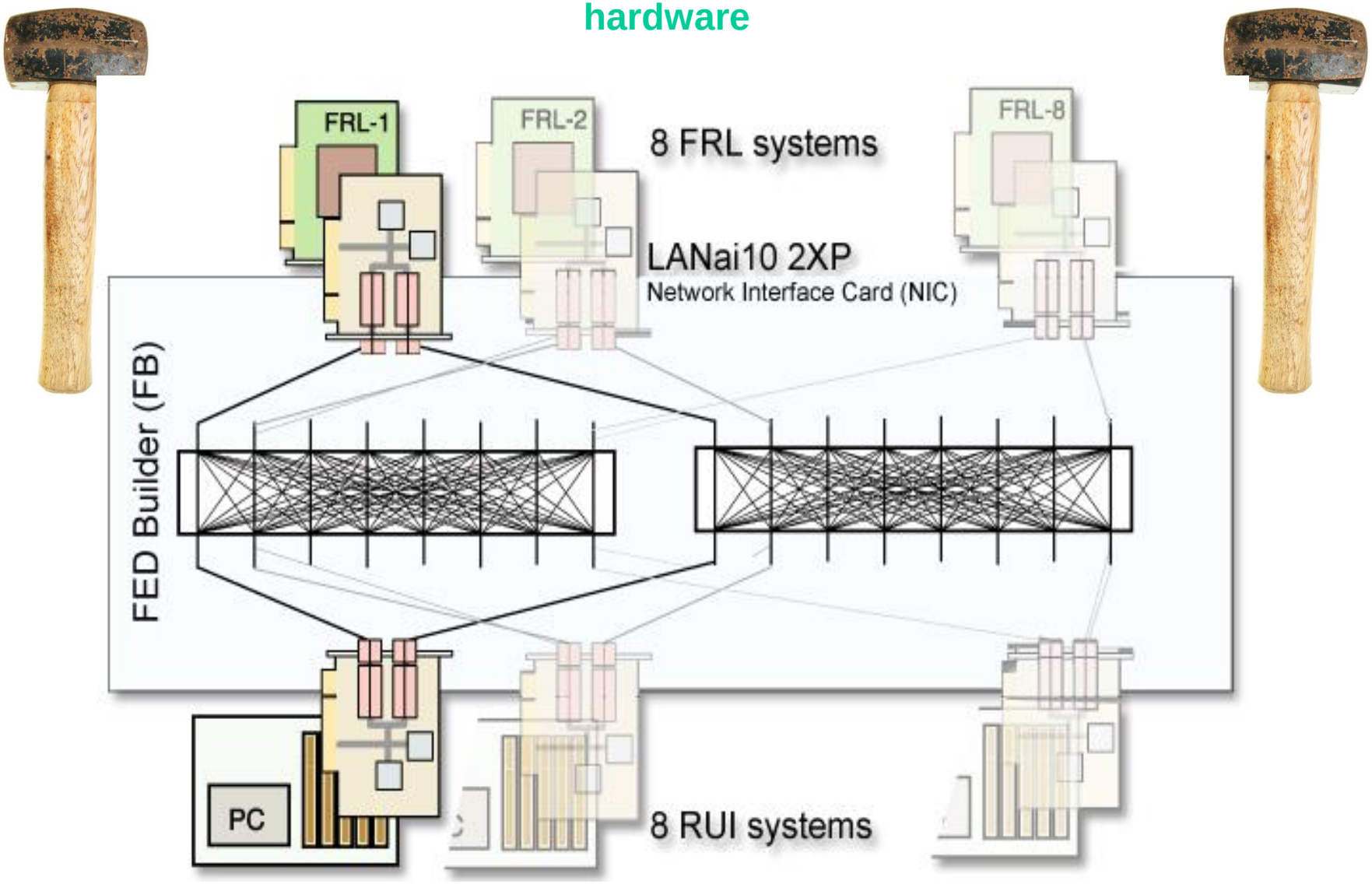


- **EVB - Demo 32x32**
- Blocksize 4kB
- Throughput at **234 MByte/s**
  = **94% of link Bandwidth**

Measurement 2003
(still valid)

2nd Solution: "take the hammer"

Over-dimension the system: buy twice as much hardware



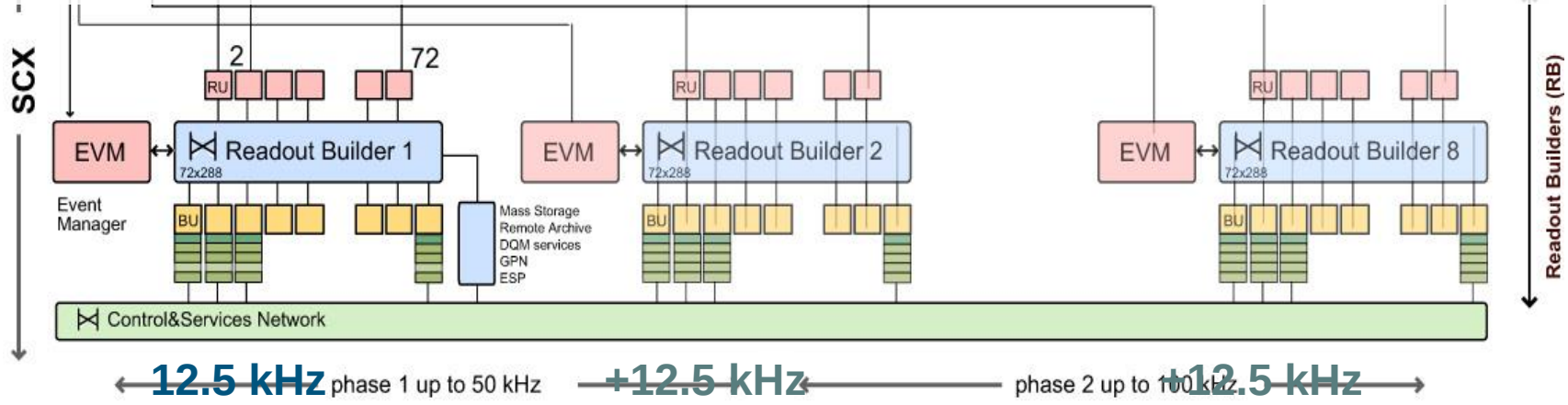
What did CMS do???

☺ Of course: we took the hammer ☺

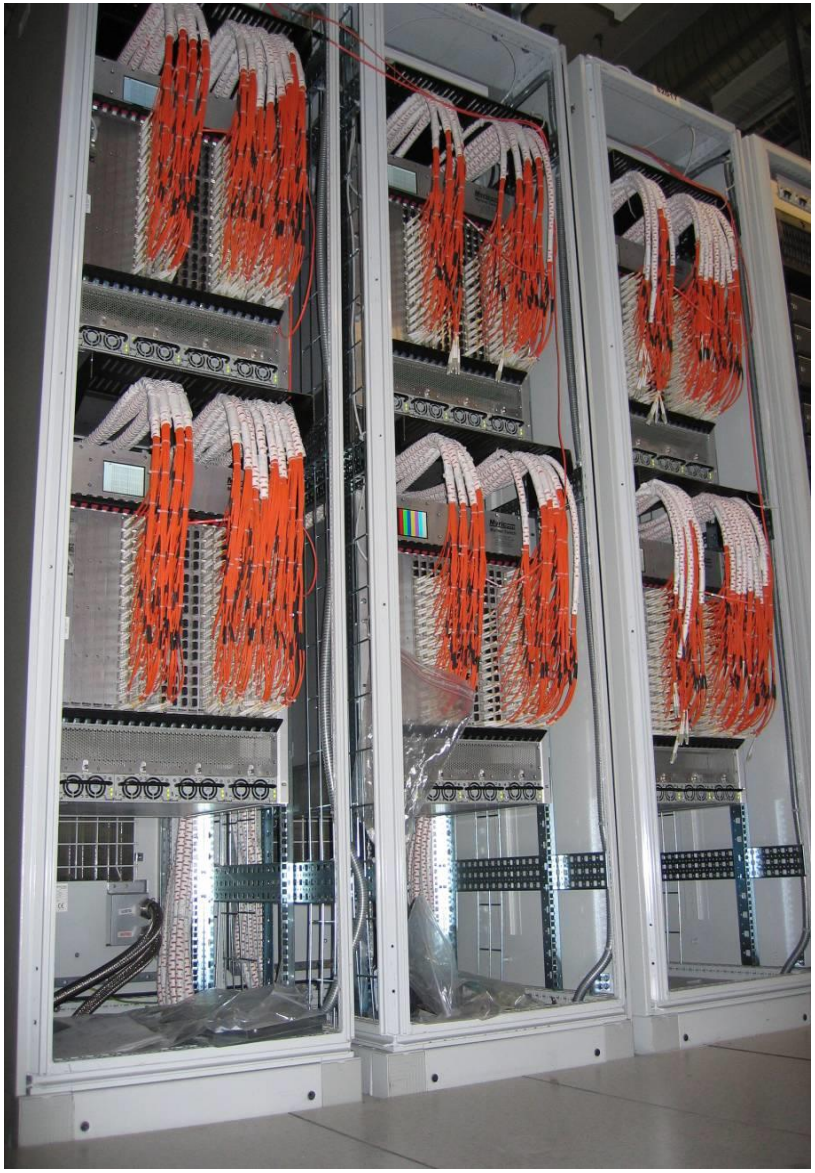
- Advantages:
  - Much less development work
  - No dependence on internal working of the switch
  - Much less maintenance work
  - Most important: **redundancy**
    - If one rail fails: continue to run with one leg (→ less performance but still taking data !!!!)

2nd stage Event Builder: "bread and butter"

- Technology: Gigabit Ethernet
  - One large switch can do the job.
- The Builder Unit PCs run also the HLT programs
  - Better usage of available CPU power.
  - There are more BU/HLT PCs than Readout Units connected to each RU-Builder



Event Builder Components

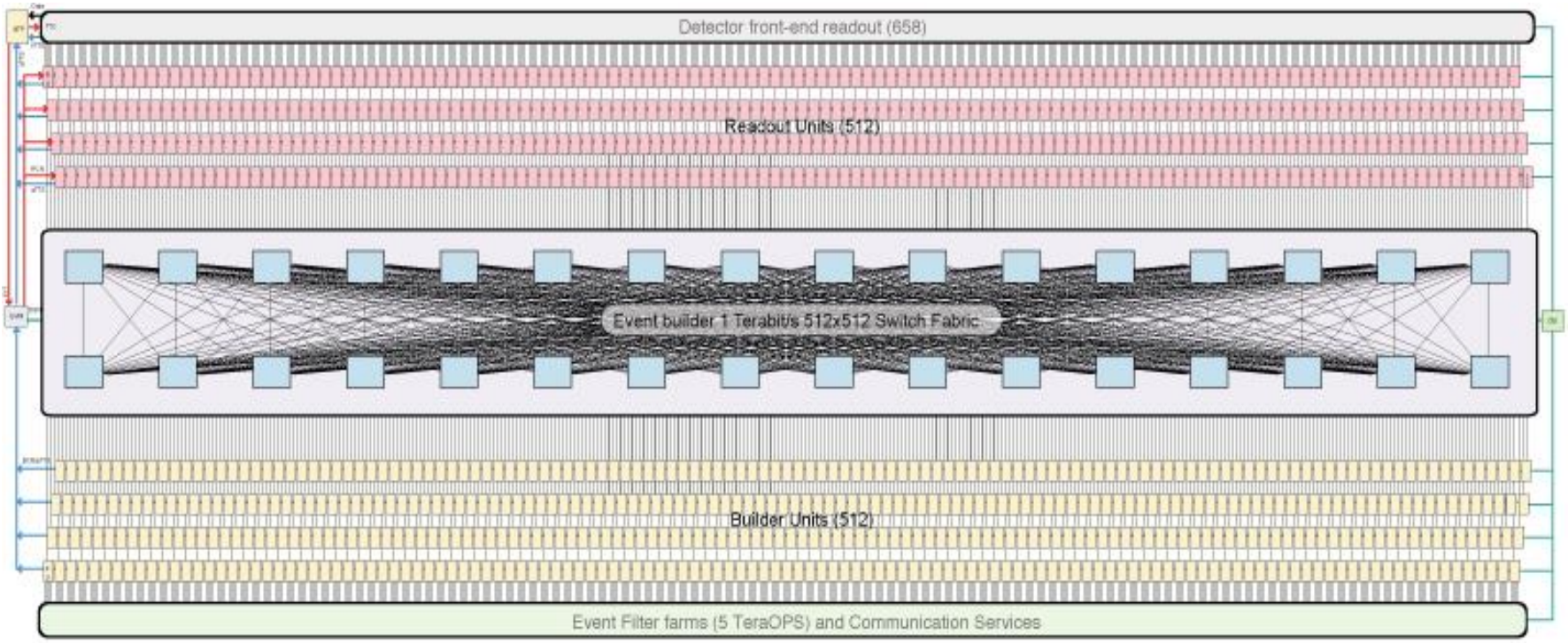


Half of the CMS FED Builder
One half of the FEDBuilder is installed close to the experiment in the underground.
The other half is on the surface close to the RU-Builder and the Filter Farm implementing the HLT.
The FEDBuilder is used to transport the data to the surface.

Event Builder Components

The RU-Builder Switch (for 2 slices)

EVB example: CMS



Level-1 maximum trigger rate	100 kHz	No. Readout systems	≈512
Average event size	1 Mbyte	No. Filter Subfarms	≈512 x n
Builder network	1 Terabit/s	No. (C&D) network ports	≈10000
Event filter computing power	5 10^6 MIPS	No. programmable units	≈10000
Event flow control	≈ 10^6 Mssg/s	System dead time	≈ %

Achronyms

TPG Trigger Primitive Generator
RTP Regional Trigger Processor
LTI Level-1 Trigger Processor
GTP Global Trigger Processor
TTC Timing, Trigger and Control
sTTS synchronous Trigger Throttle System
aTTS asynchronous Trigger Throttle System
FES FrontEnd System
FED FrontEnd Driver
FEC FrontEnd Controller
DSS Data to Surface
FRL FrontEnd Readout Link
RU Readout Unit
BU Builder Unit
FS Filter Subfarm
EVM Event Manager
RM Readout Manager
BM Builder Manager
EVB Event Builder
RCN Readout Control Network
BCN Builder Control Network
CSN Computing Service Network
DCN Detector Control network
DSN DAQ Service Network
DCS Detector Control System
RCS Run Control System