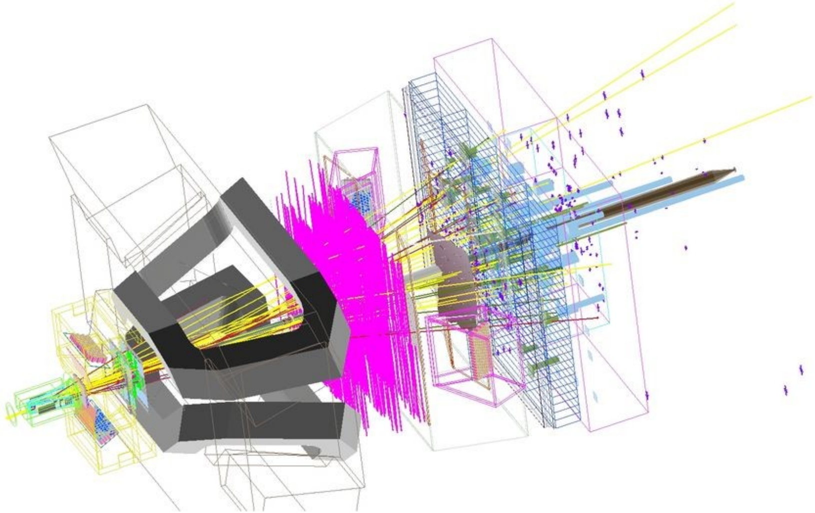
Back-end systems evolution in high energy physics
Example of LHCb Detector



J.P. Cachemiche,
Centre de Physique des Particules de Marseille

Outline

- The LHCb experiment
- Current architecture
- Upgrade
- Triggerless readout concept
- Initial and final readout architecture

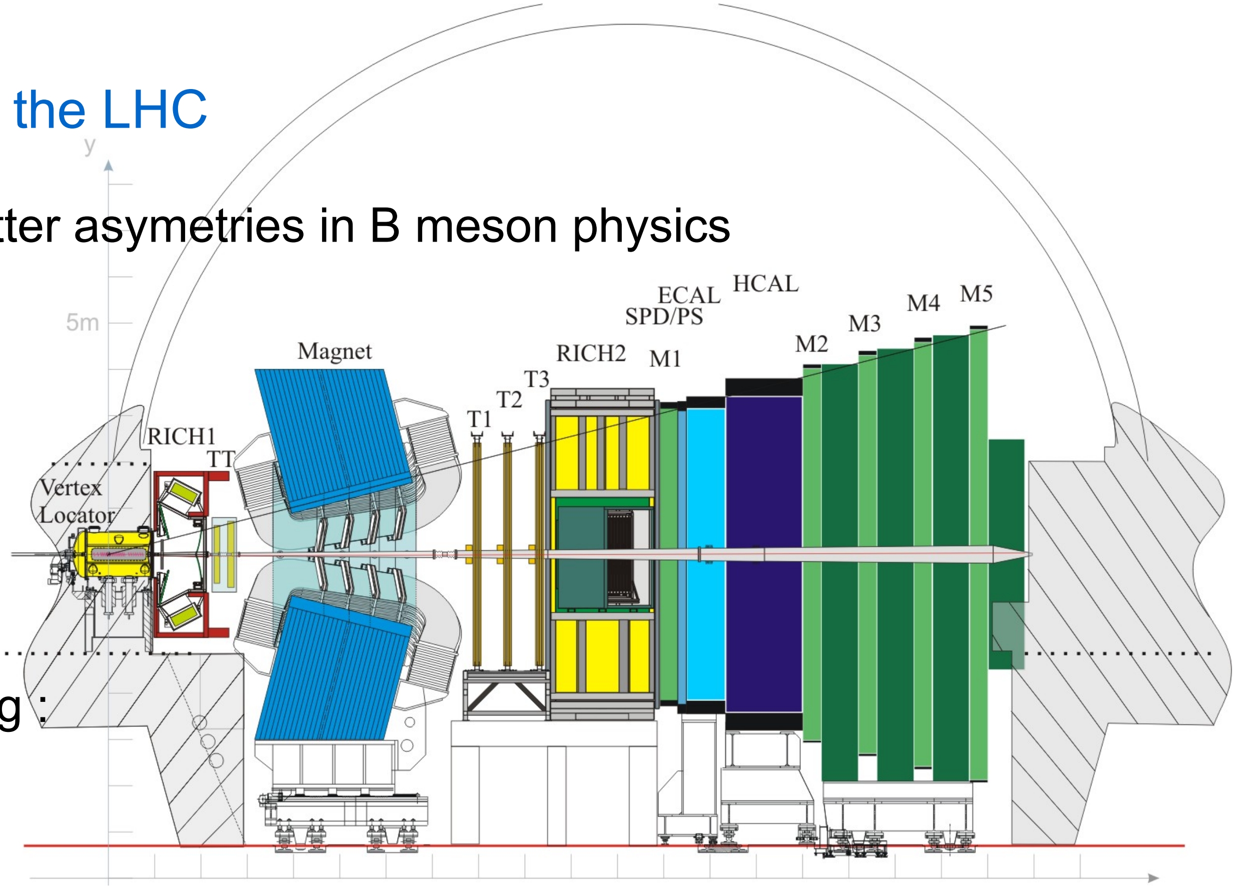
The LHCb detector

One of the 4 experiments on the LHC

- Study of mater/anti-matter asymetries in B meson physics

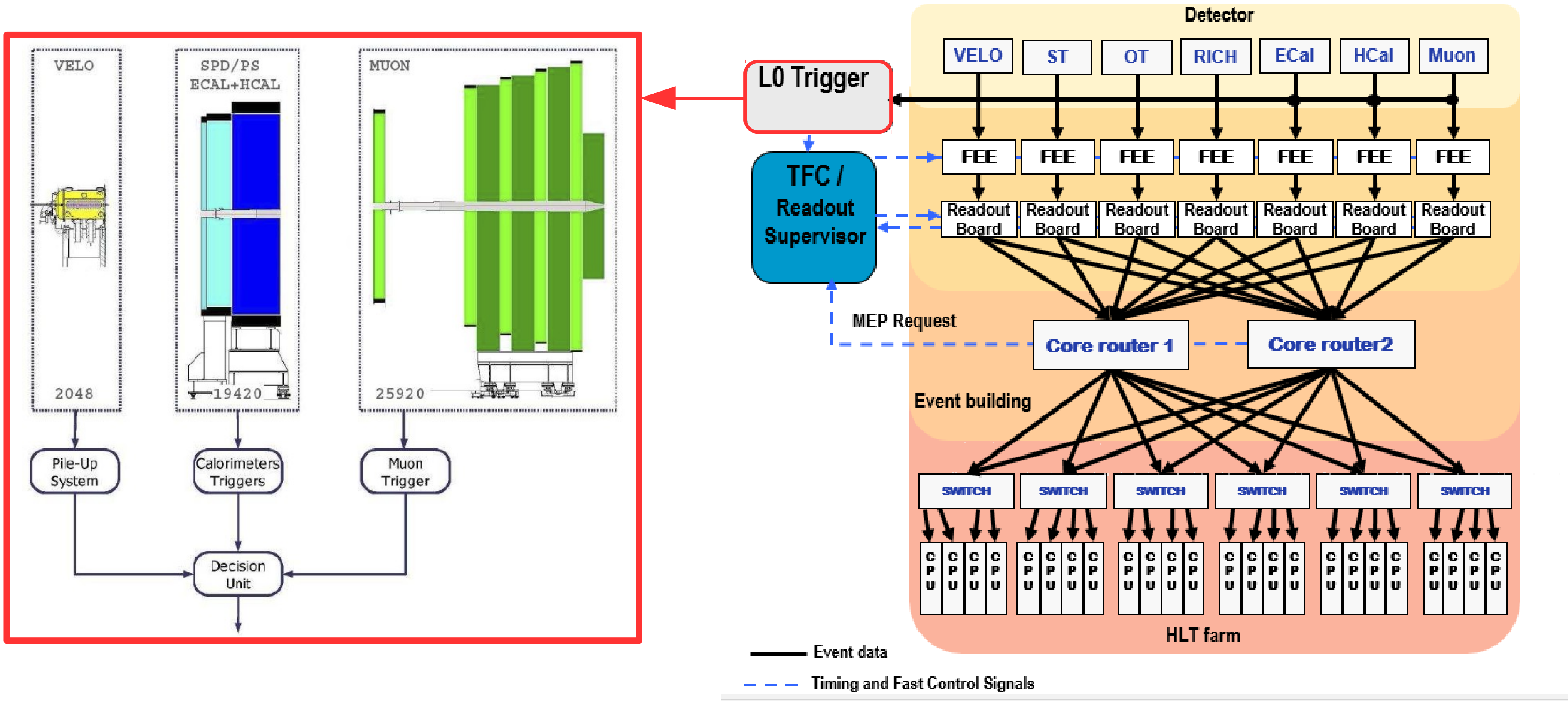
Few numbers

- ~1 million channels
- Sampled at 40 MHz
- Output rate after filtering : 5 kHz



Sub-detector	Vertex	Pileup	RICH1&2	IT + TT	Outer tracke	SPD	Preshower	Ecal	Hcal	Muon
Number of channels:	172k	8k	200k + 295k	129k 180k	54k	6k	6k	6k	1.5k	125k physical 26k logic

Current architecture



Hardware trigger to decrease data flow from 40 to 1 MHz

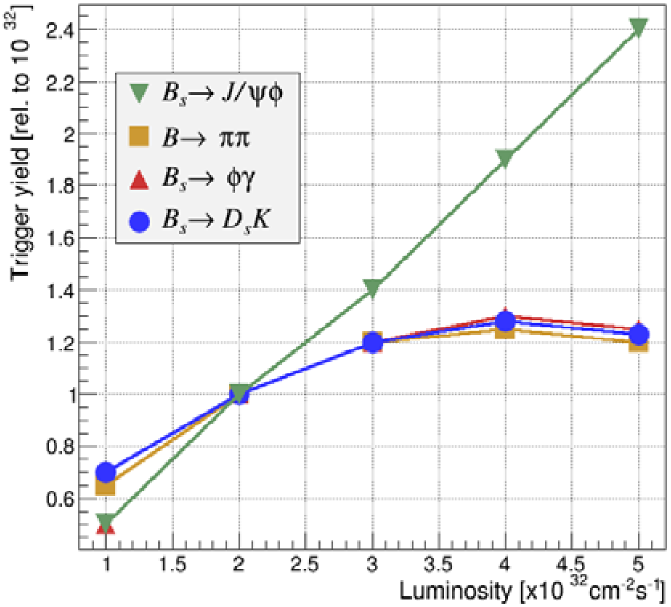
Why do we upgrade ?

Motivation

- Maximum luminosity in next 5 years: 5 fb-1
- At current rythme, statistical precision of measurements varies very slowly
- By increasing the luminosity from 2×10^{32} to 10^{33} $cm^{-2}s^{-1}$
  - ➔ Reach a cumulated luminosity > 50 fb-1

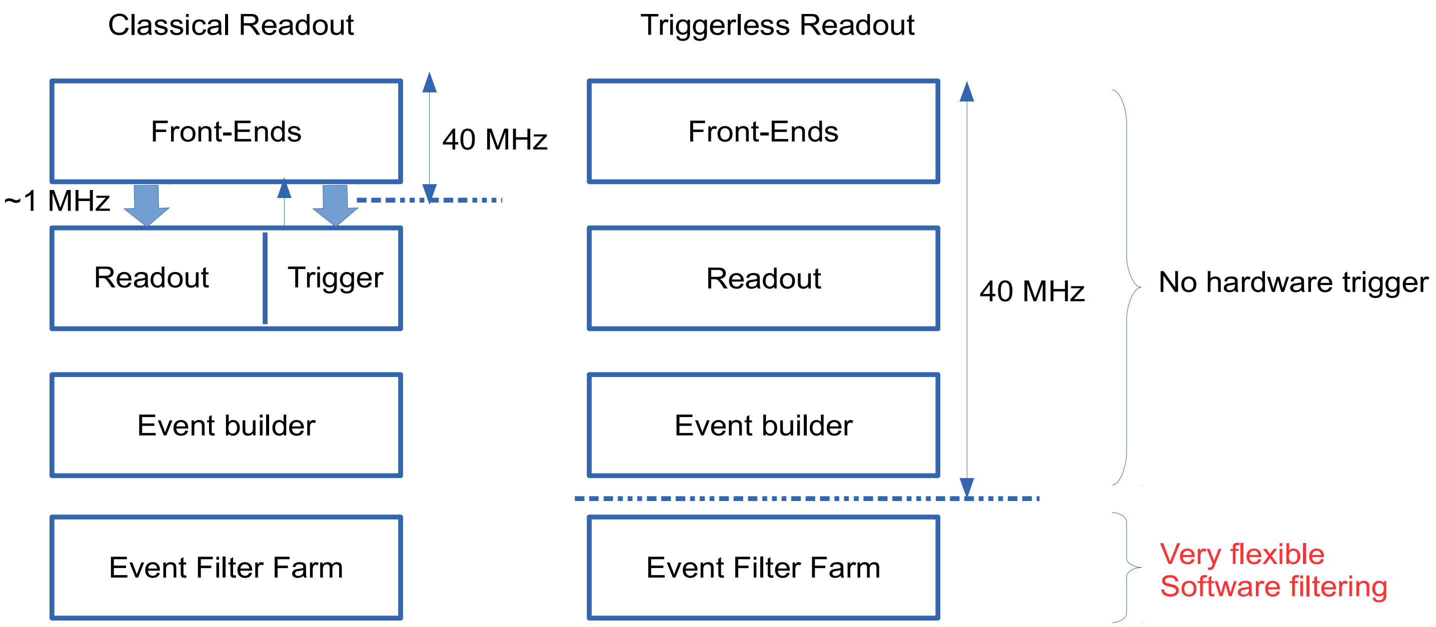
But …

- Saturation of hardware trigger on hadronics channels
  - ➔ Architecture change required



Solution

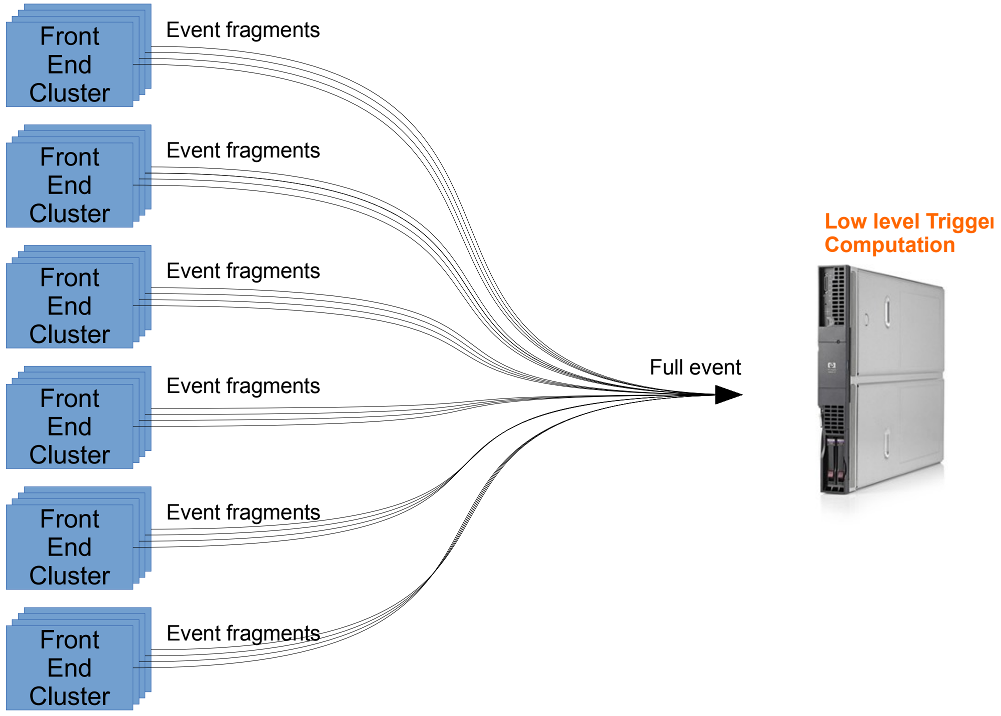
LHCb opted for a « triggerless » approach



- Previous systems were able to process events at 1 Mhz
- Current system implemented in 2009, upgrade planned for 2019
- Mooresʼs law predictions should allow to envisage filtering of events by software at 40 MHz
    - ➔ More powerful algorithmes
    - ➔ More flexible

Triggerless readout principle

- All event fragments of a same collision must be routed towards a single CPU
- Fragments of the next one routed to another CPU
- Etc ..
- Event building is required for **each** collision.
  - → Huge bandwidth

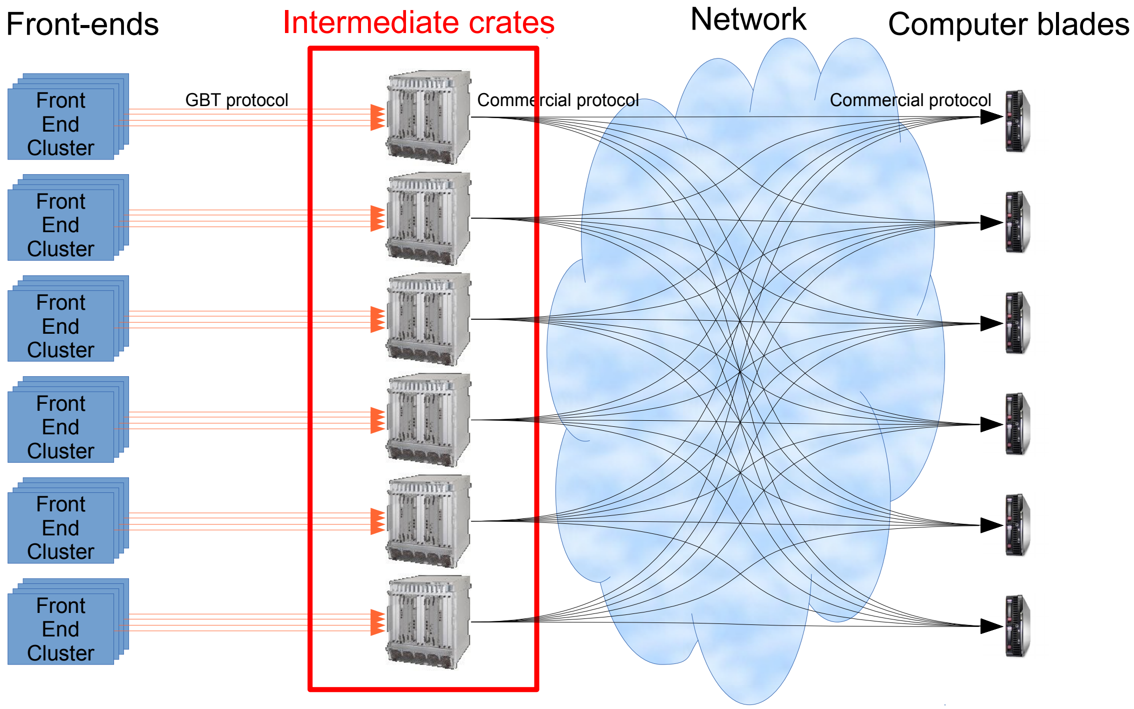


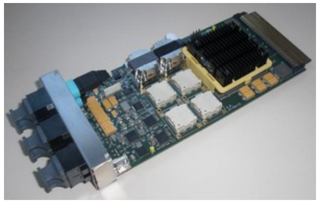
Initial architecture

- First level of concatenation achieved in ATCA crates

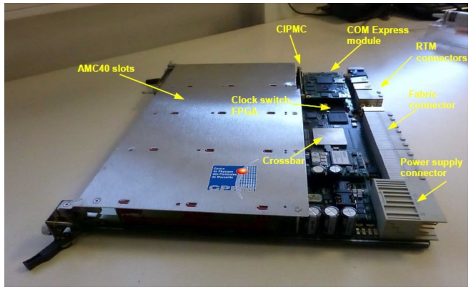
- All event fragments routed through a switch network to a single computer

- Key issues :
  - Requires hardware IPs for either 10GbE links, Infiniband, Omnipath or else
  - Collision management



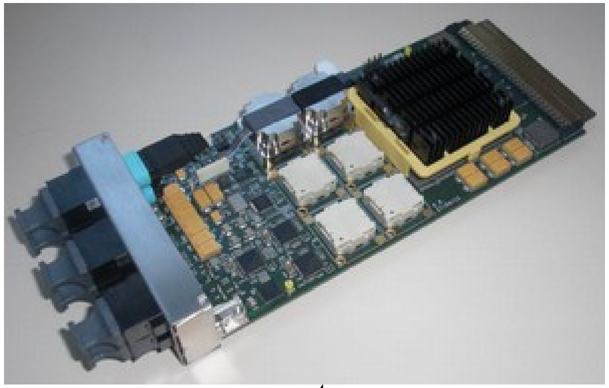


AMC40 board



ATCA40 board

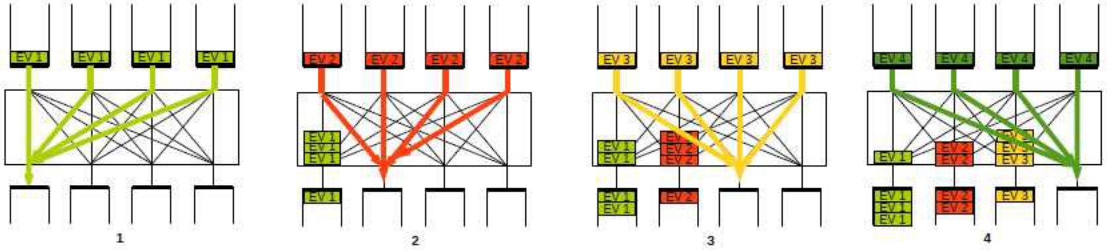
Collisions handling



2 methods to make data converge toward a single CPU:
both require memory

- **Push** : requires expensive switches with **memory** inside
- **Pull** : requires **memory** on already very dense back-end boards

Push mode

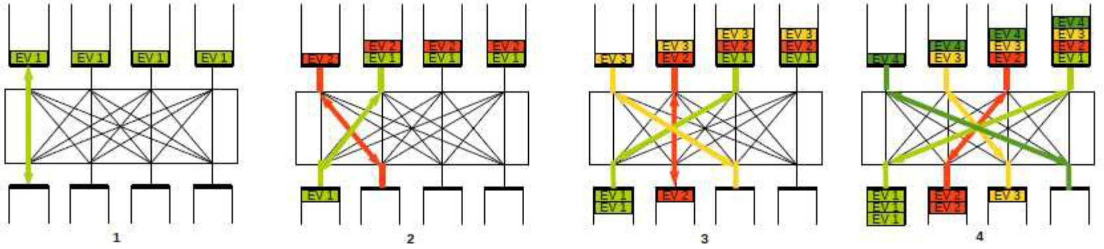


Back-end electronics

Switches

CPU blades

Pull mode



Back-end electronics

Switches

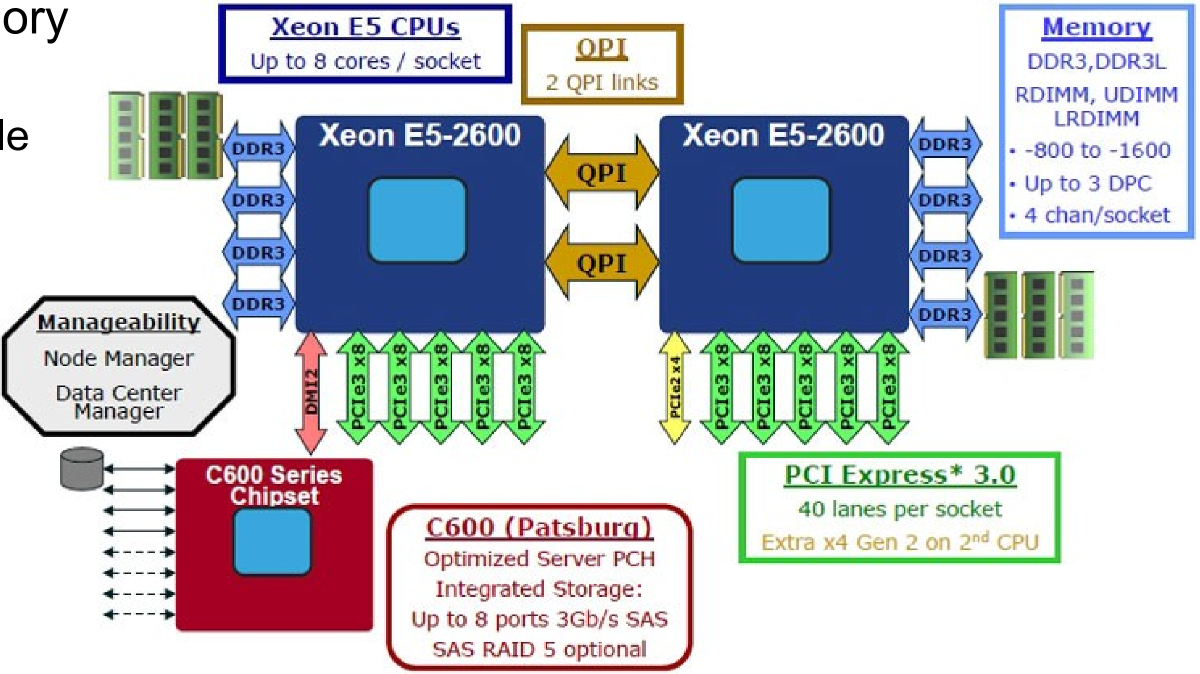
CPU blades

Architecture evolution

Implementation of the event builder directly in the farms

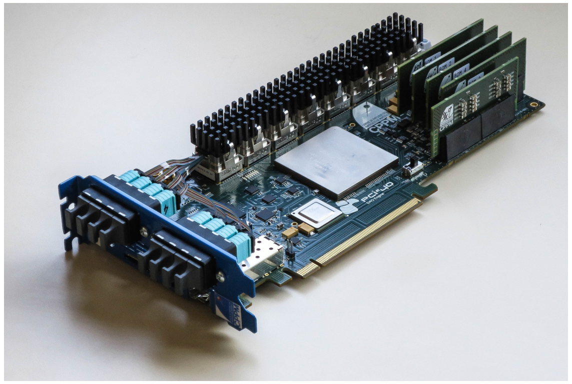
Allowed by recent architectures of Intel chips :

- Very large bandwidths inside the PC
- Independent paths to memory
- Multicore CPUs
  - ➨ Powerful enough to handle both Event building and Software Trigger
  - ➨ Solves the memory issue

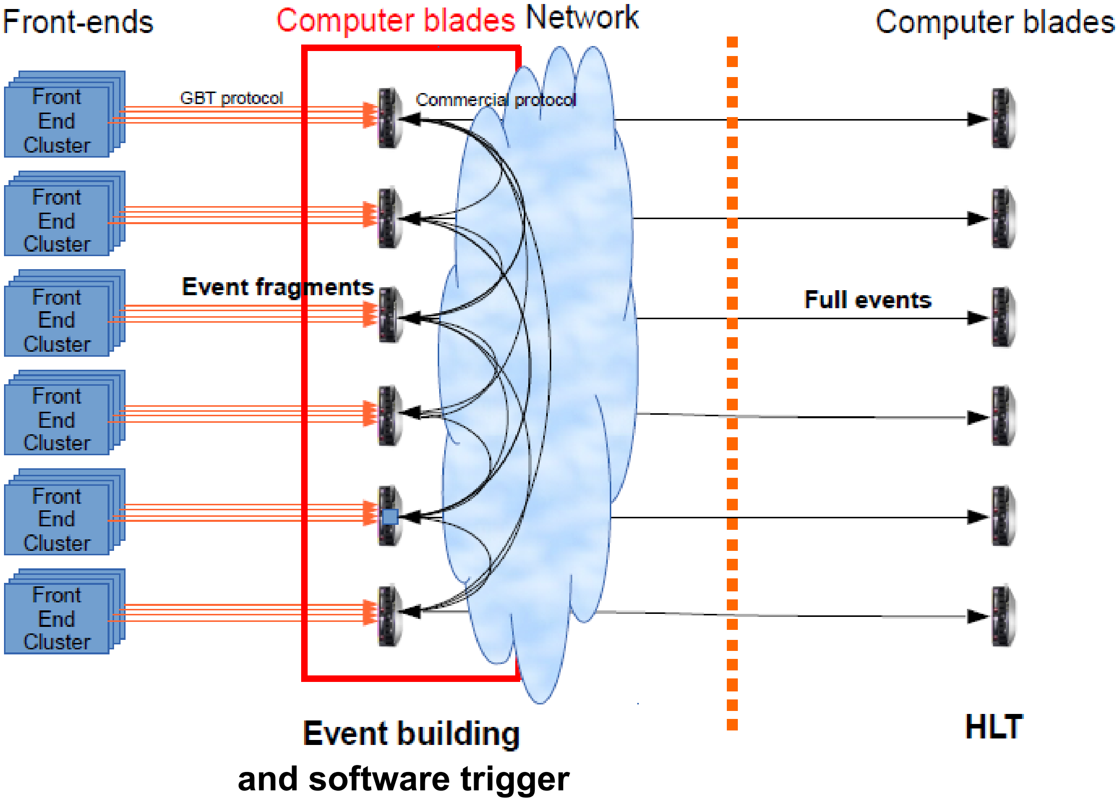


New readout/trigger scheme

Move back-end cards into already existing CPU blades



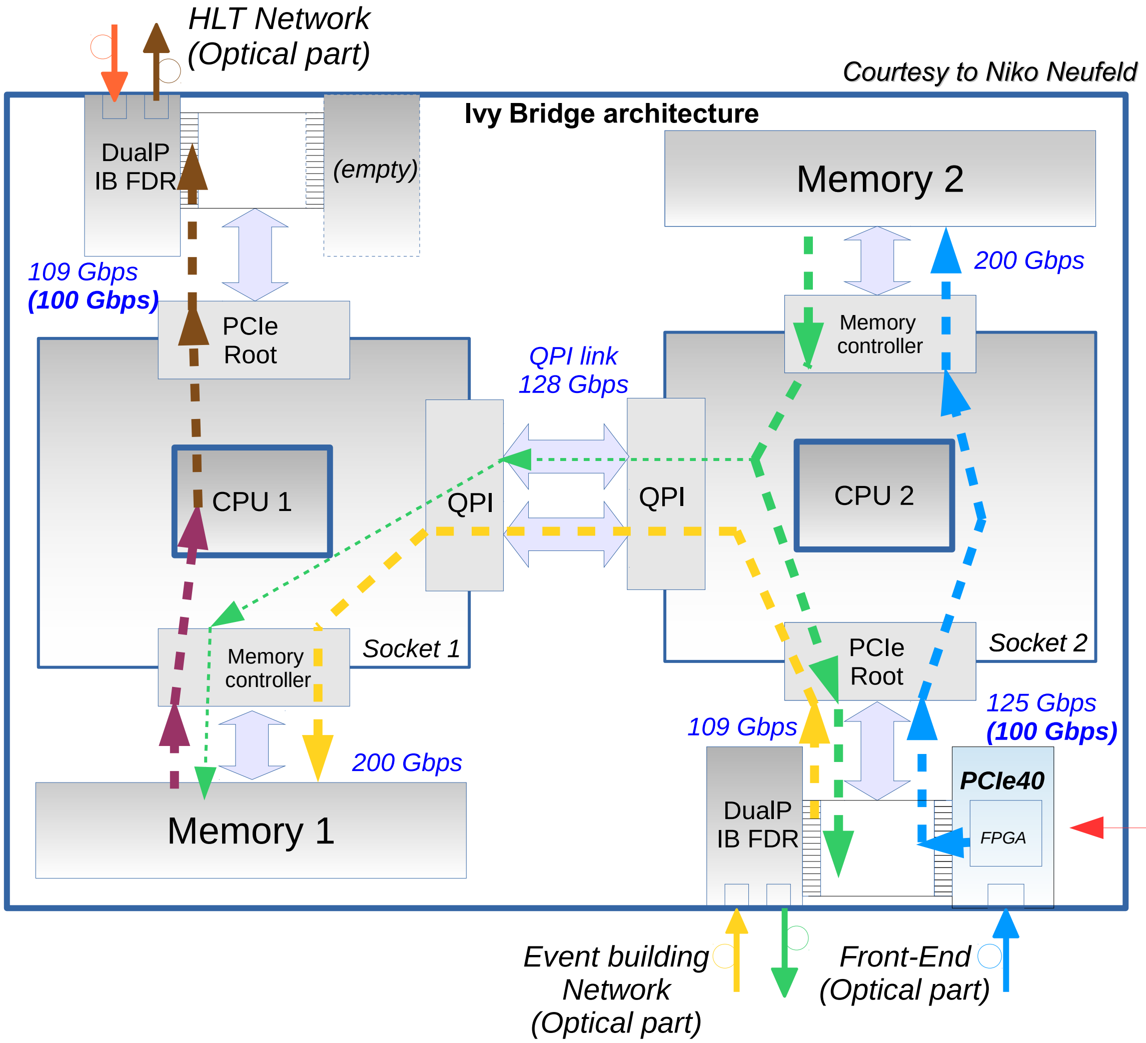
PCIe40 board



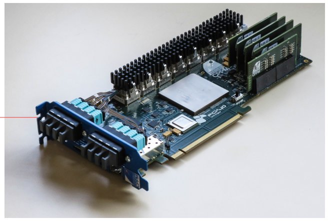
Advantages

- Large memory in the CPU → simpler acquisition board, cheaper switches
- Possibility to run LLT in the Event Building CPU blades
- No more intermediate crates
- Less optical links
- Network type can be changed easily
- Scalable

Deeper in the computer



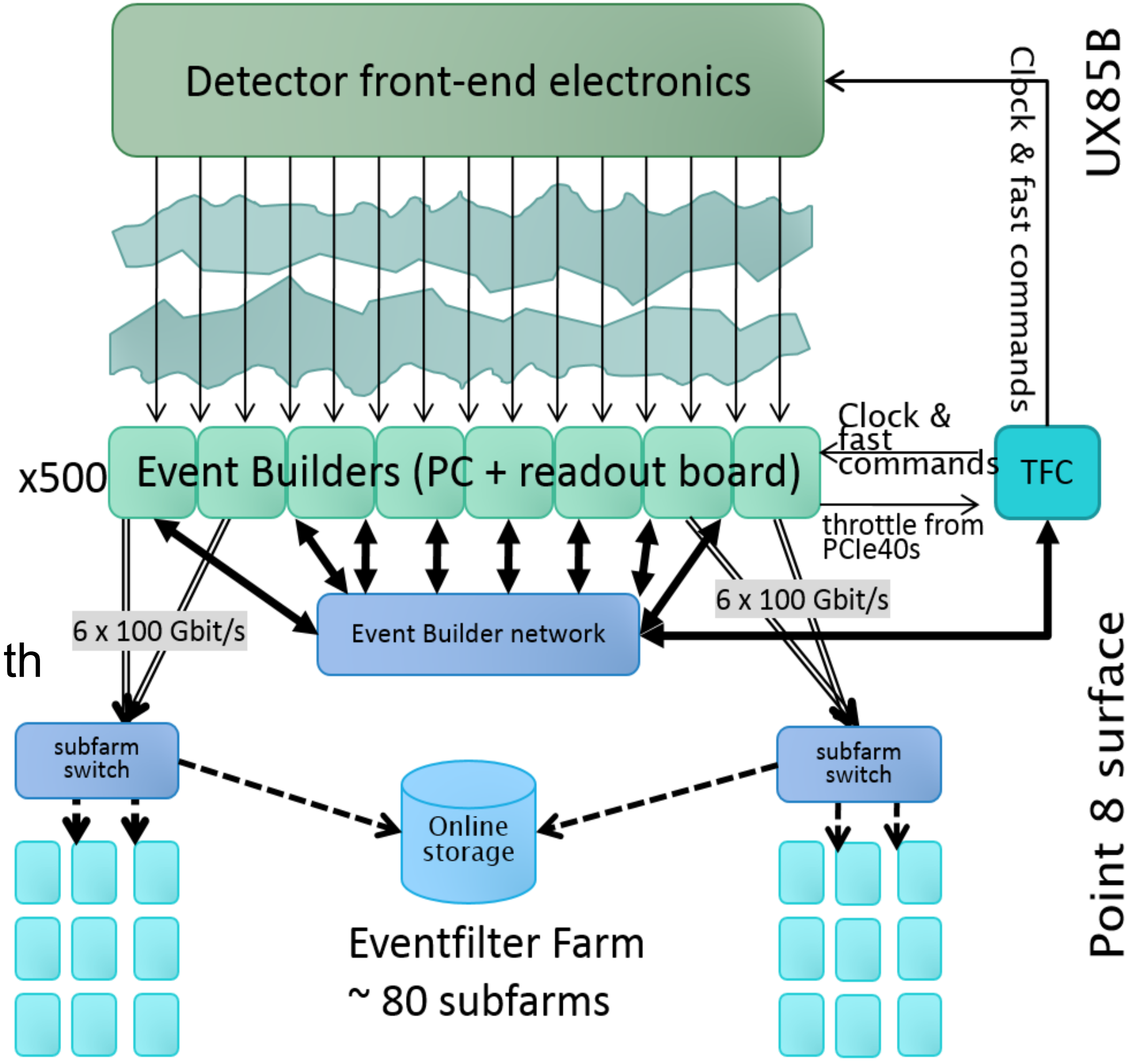
Courtesy to Niko Neufeld CERN



- *80 % of processing power still available for Low Level Trigger computation*

Final architecture

- Readout located on surface
  - Distance between FE and RO : ~350m
- ~12000 optical links
- ~ 500 readout boards
- ~24 to 48 links on each board
- ~100 kbytes per event
- ~32 Tb/s aggregate bandwidth



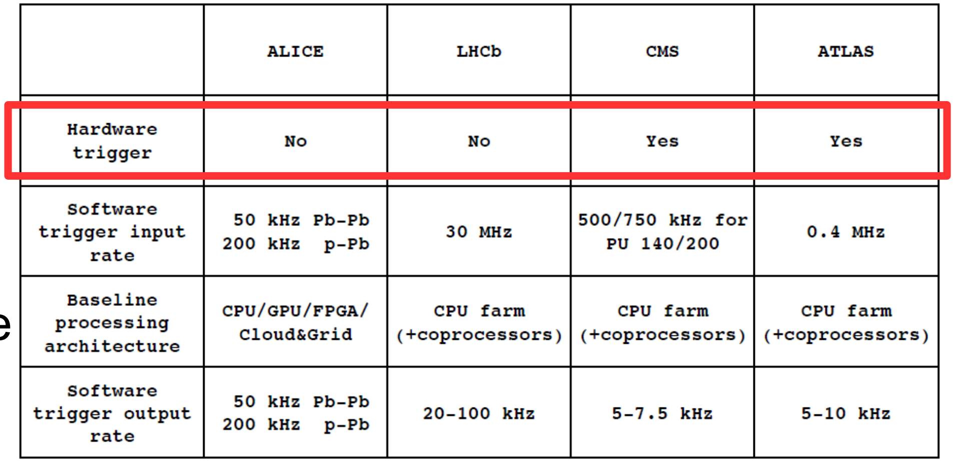
Conclusion

Promizing architecture

Is this a general trend ?

- More and more hardware features migrate towards software

- Half of the 4 experiments on LHC will use a software trigger



	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

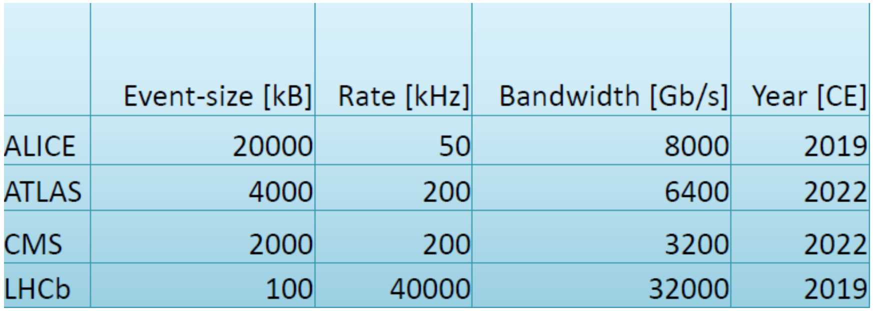
Future DAQ key numbers in the LHC V.V Gligorov, CERN

But ...

- Requires a lot of bandwidth

- Requires also a full redesign of previously used algorithmes due to Moore's law inflexion

➨ It will be critical to fully exploit multi-core architectures



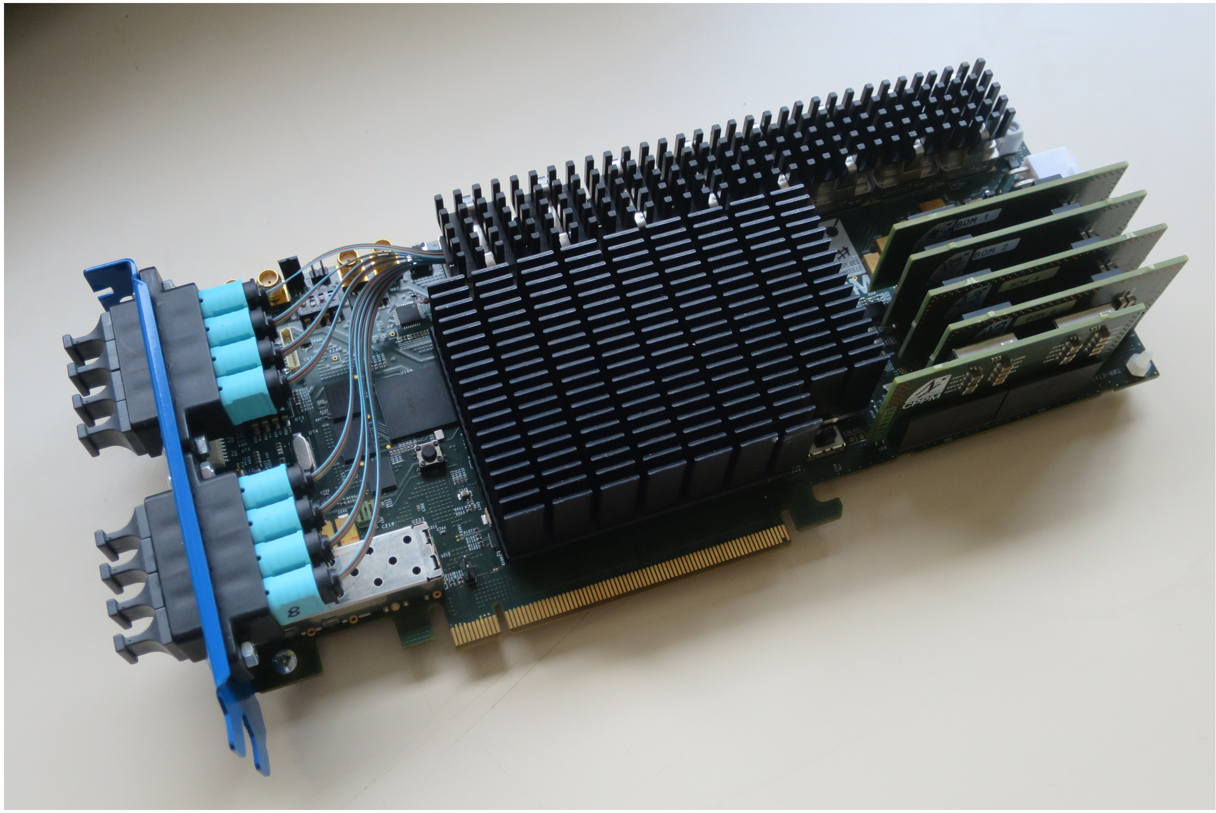
	Event-size [kB]	Rate [kHz]	Bandwidth [Gb/s]	Year [CE]
ALICE	20000	50	8000	2019
ATLAS	4000	200	6400	2022
CMS	2000	200	3200	2022
LHCb	100	40000	32000	2019

Future DAQ expected bandwidth in the LHC Niko Neufeld, CERN

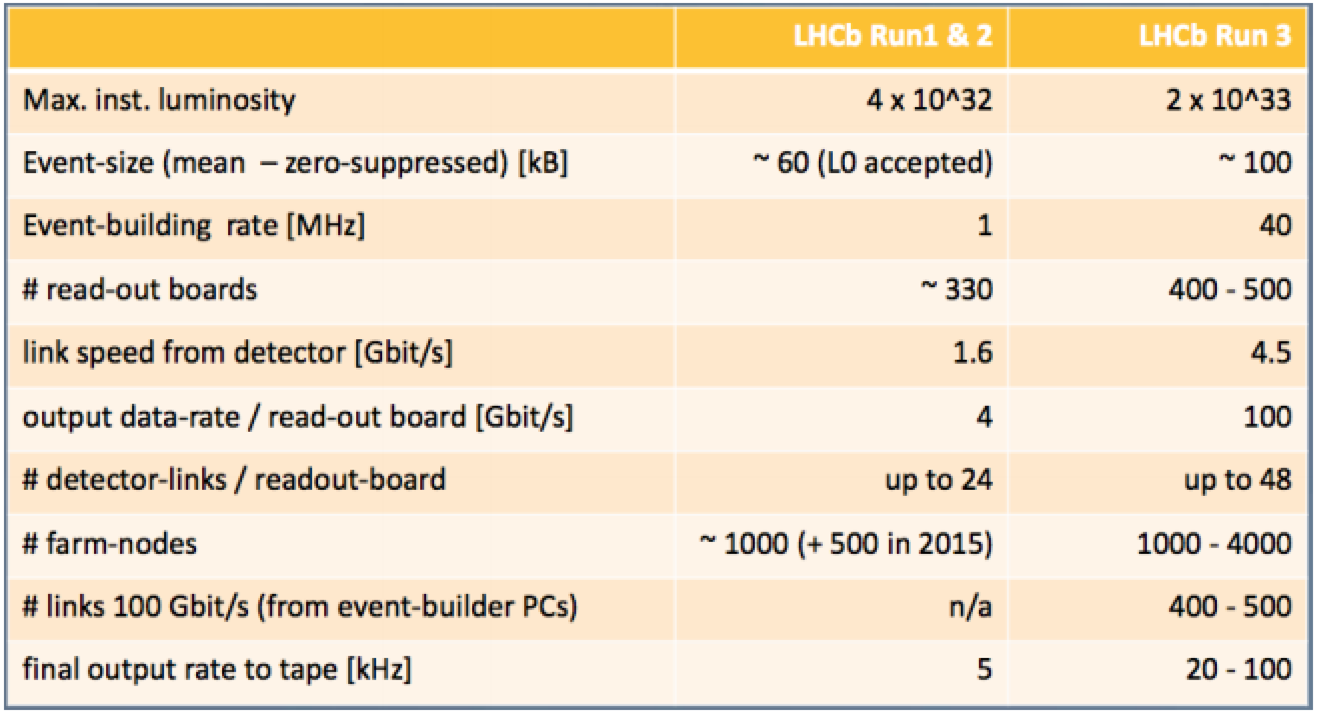
More ...

The PCIe40 board

- 48 bidirectional links at 10 Gbits for acquisition and control
- 2 bidirectional links at 10 Gbits/s for time distribution
- >100 Gbits/s PCIe Gen3 x 16
- 1.1 Million logic element Arria10 FPGA
- 150W power consumption



LHCb upgrade vs current



	LHCb Run1 & 2	LHCb Run 3
Max. inst. luminosity	4 x 10^32	2 x 10^33
Event-size (mean – zero-suppressed) [kB]	~ 60 (L0 accepted)	~ 100
Event-building rate [MHz]	1	40
# read-out boards	~ 330	400 - 500
link speed from detector [Gbit/s]	1.6	4.5
output data-rate / read-out board [Gbit/s]	4	100
# detector-links / readout-board	up to 24	up to 48
# farm-nodes	~ 1000 (+ 500 in 2015)	1000 - 4000
# links 100 Gbit/s (from event-builder PCs)	n/a	400 - 500
final output rate to tape [kHz]	5	20 - 100