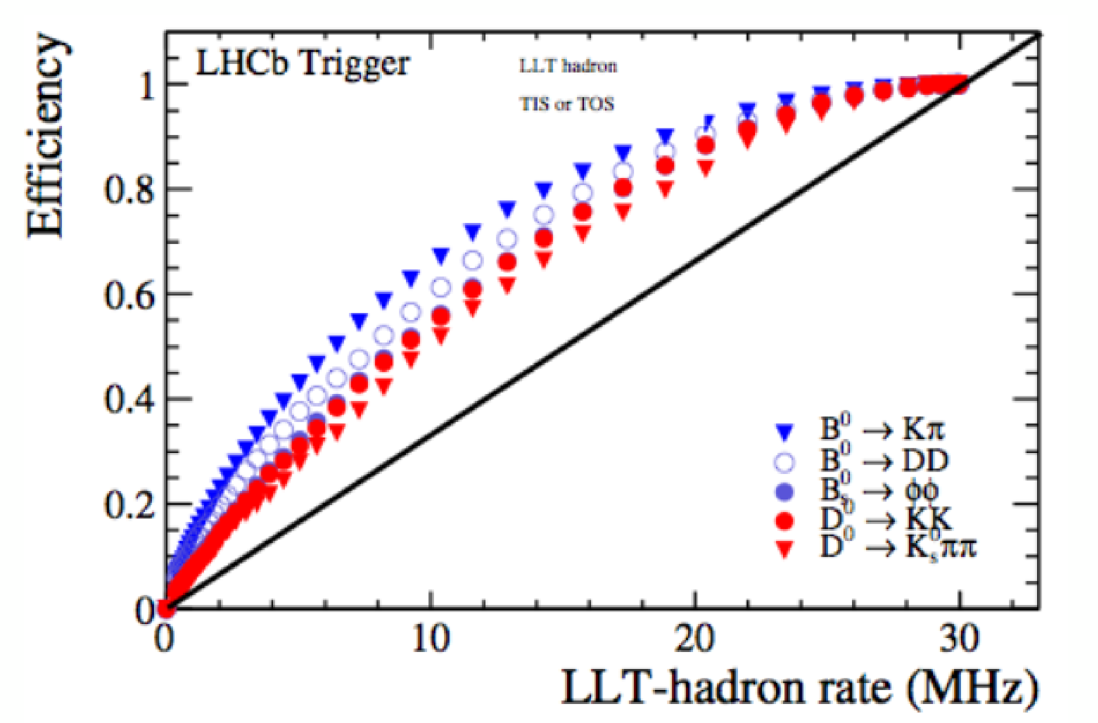
LHCb DAQ for Run3 and beyond

Niko Neufeld
HL-LHC Trigger, Online and Offline Computing Working Group Topical Workshop
Sep 5th 2014

LHCb after LS2

- Substantial increase in physics reach only possible with massive increase in read-out rate
- Geometry (spectrometer) and comparatively small event-size make it possible – and the easiest solution – to run trigger-free, reading every bunch-crossing
- Note:
  - Any increase beyond 1 MHz requires change of all front-end electronics
  - To keep data-size reasonable, all detectors must zero-suppress at the front-end



Recap - requirements

- Event rate 40 MHz
  - of which ~ 30 MHz have protons
- Mean nominal event size 100 kBytes
- Readout board bandwidth up to 100 Gbits/s
  - to match DAQ links of 2018
- CPU nodes up to 4000
  - actual requirements are probably less, but provide for sufficient power, cooling and connectivity to accommodate a wide range of implementations
- Output rate to permanent storage 20 to 100 kHz

In one number…

8800 (# VL) * 4.48 Gbit/s (wide mode) →

40 Tbit/s

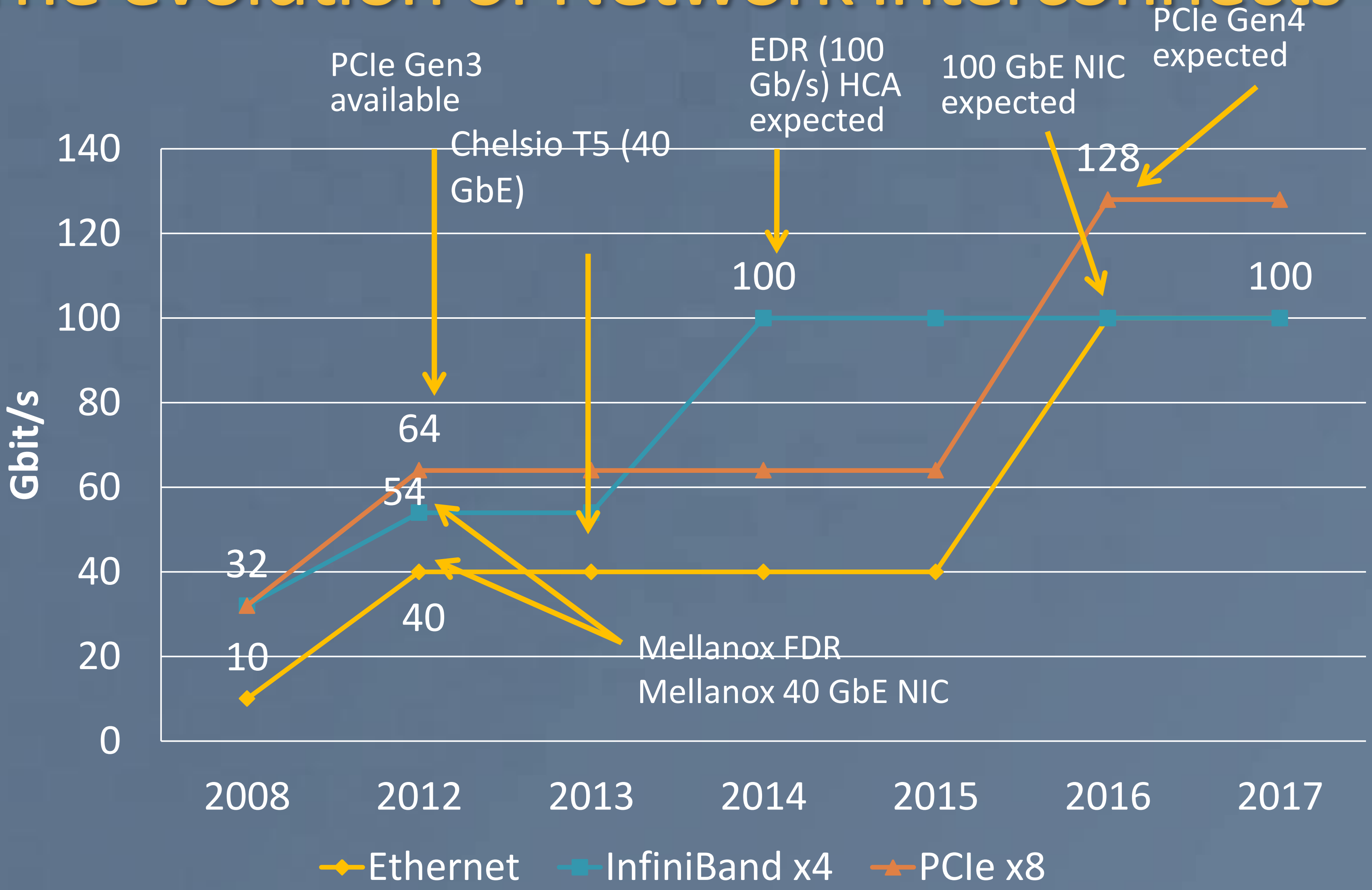
40 Tbit/s DAQ in practice

- By 2018 100 Gbit/s technologies will be well established in the data-centre. Currently we see three candidates:
  - 100 G Ethernet (data-centre links probably 2015)
  - InfiniBand EDR (available end of 2014)
  - Intel OmniScale Fabric (available ~ 2015)
- The event-builder will use 100 Gbit/s links.
- Add 20% safety margin for protocol overheads etc... → need 500 100 Gbit/s links
- Start study with InfiniBand (because it’s already available)

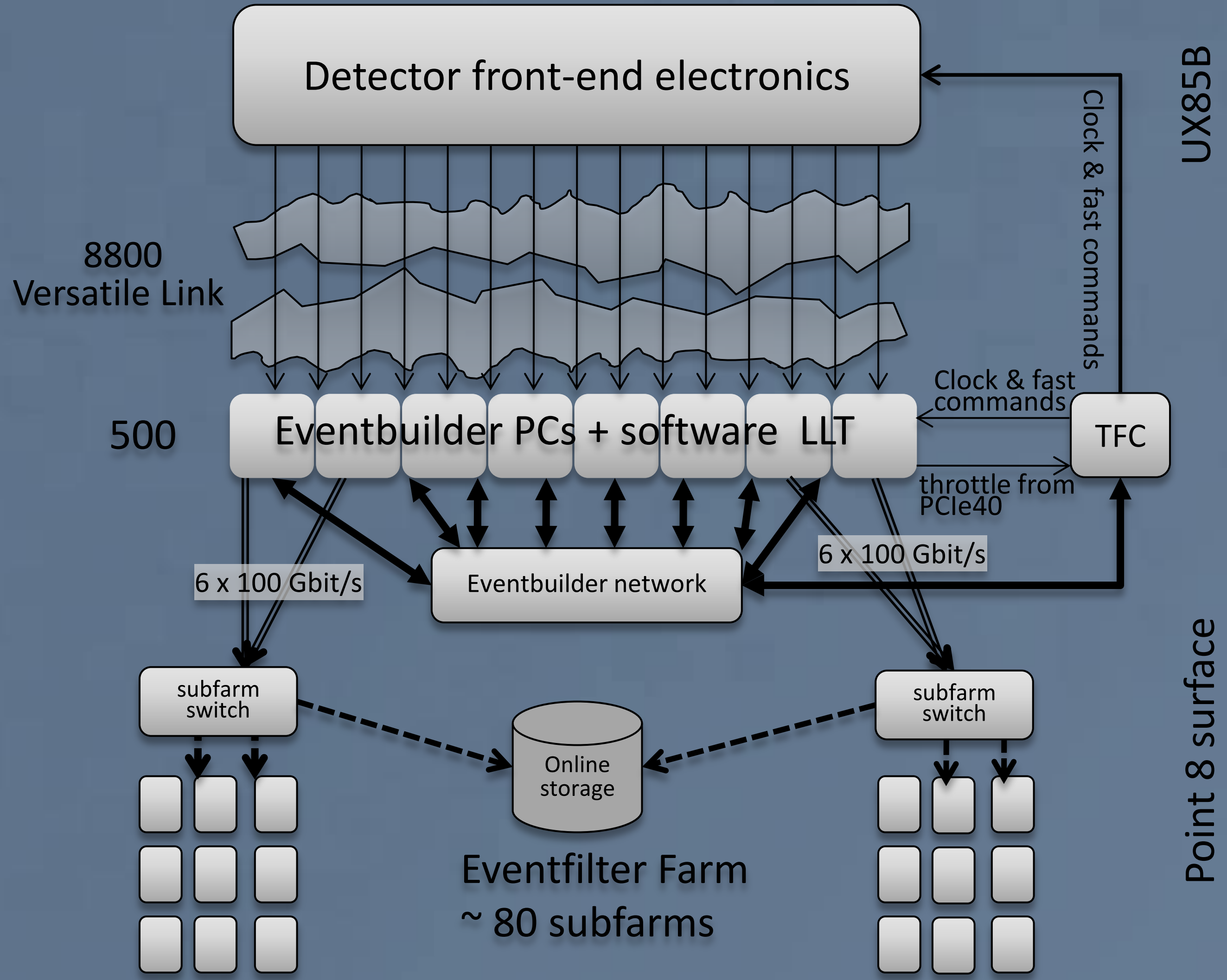
Architecture considerations

- Want to be able to use data-centre switches → need lots of buffering in event-builder units
- Want to decide on network technology and manufacturer as late as possible → use COTS network interfaces (i.e. PC)
- Trigger processing in any imaginable “compute unit” will be CPU-bound not I/O-bound → optimal match by combination of “high-speed” network (event-building) and “low-speed” network (event-filtering)
- Keep distances short → minimize cost of individual links

The evolution of Network Interconnects



Readout Architecture

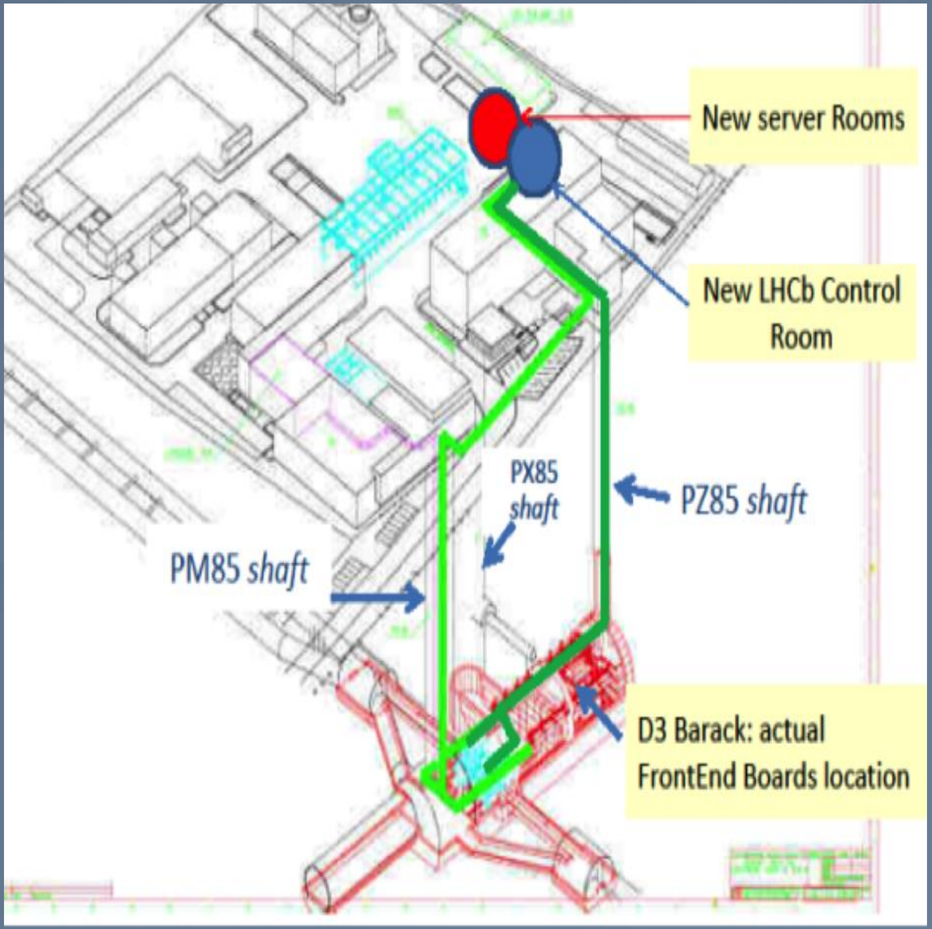


Challenges

- 200 Gbit/s full duplex in PC (including opportunistic use of idle CPU resources)
- 100 Gbit/s FPGA receiver card
- 300 m operation of Versatile Link
- 40 Tbit/s event-building network

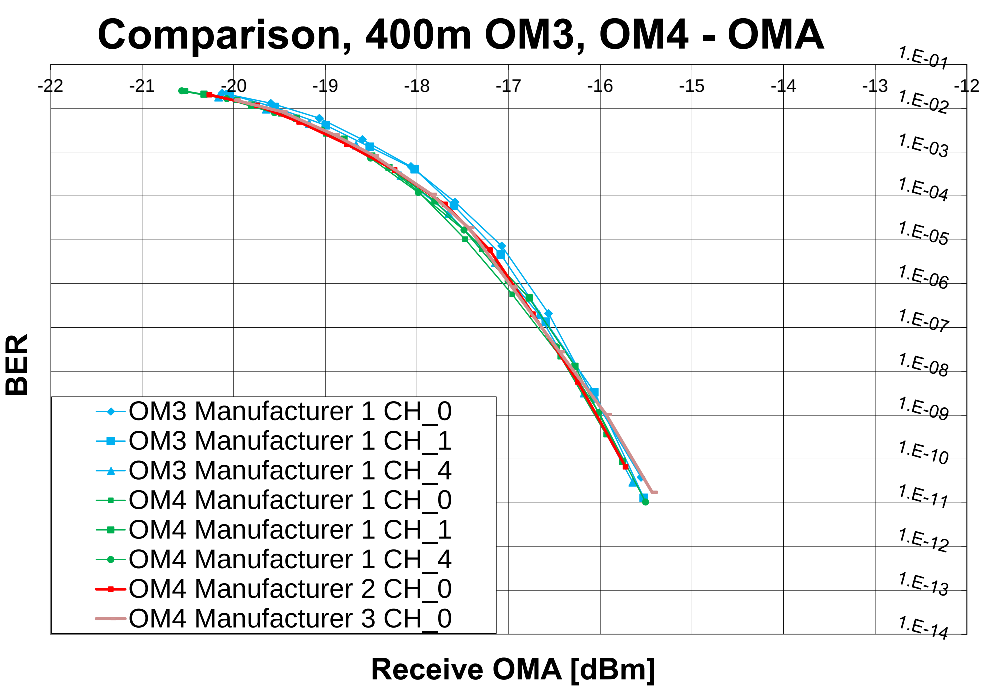
Long-distance optical fibres

- Most compact system achieved by locating all Online components in a single location
- Power, space and cooling constraints allow such an arrangement only on the surface: containerized data-centre
- Versatile links connecting detector to readout-boards need to cover 300 m
- Test installation will start tomorrow 5/9/14 in collaboration and with the help of EN/MEF and EN/EL

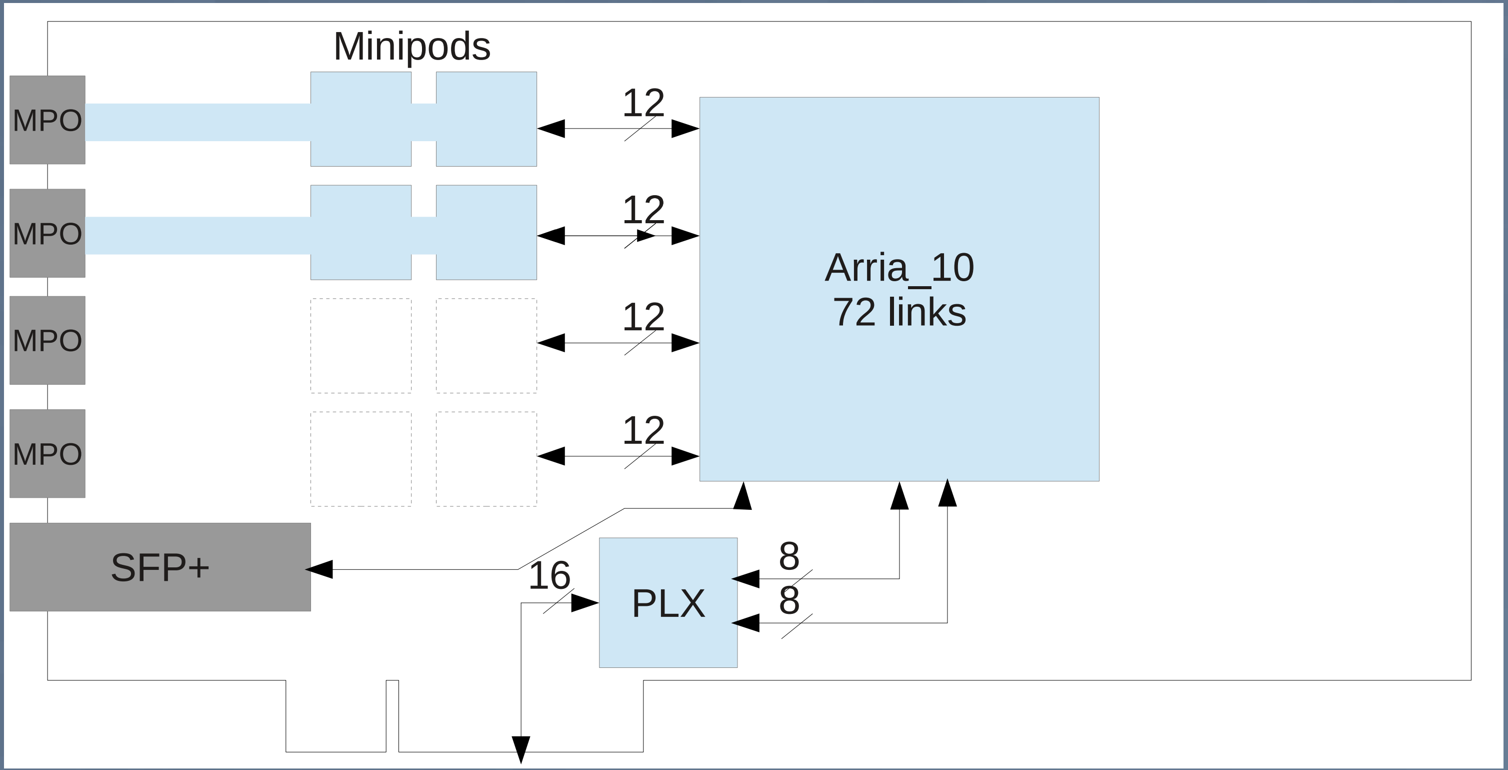


Long distance versatile link lab tests

- Various optical fibres tested show good optical power margin and very low bit error rates
- For critical ECS and TFC signals Forward Error Correction (standard option in GBT) gives additional margin
- On DAQ links expect < 0.25 bit errors / day / link in 24/7 operation



PCIe40



- Up to 48 bi-directional optical I/Os (VL)
- Up to 100 Gbit/s I/O to the PC (PCIe Gen3 x 16 card)
- Designed by CPP Marseille. Firmware and production support by INFN Bologna, LAPP and CERN
- Universal building block for DAQ, ECS and TFC

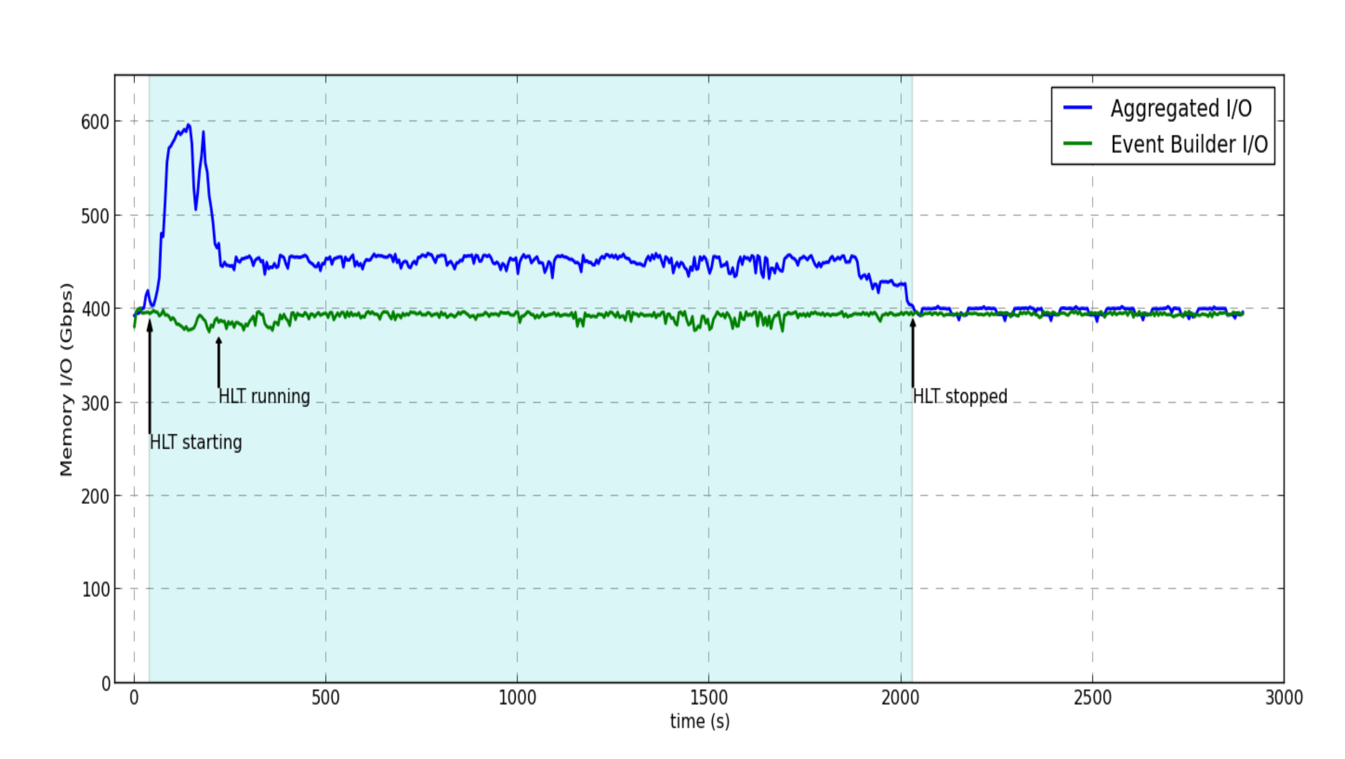
Latency measurements

Message Size[byte]	lat[µs]
256	4.3
512	4.7
1024	5.2
2048	7.1
4096	9.1
8192	13.3
16384	17.2
32768	22.9
65536	33.7

- Latency measurements for single threaded client/server
- Average value over 100 repetitions
- Low latencies are good for RDMA / pull protocols

measurements by A. Falabella et al. (UNIBO & INFN) on Qlogic IB

Performance results – eventbuilder PC



400 Gbps stable on I/O

- Opportunistic CPU usage on event-builder nodes possible
- Can be used for High Level Trigger and/or Low Level Trigger

measurements by D. Campora et al. (CERN) on Mellanox IB

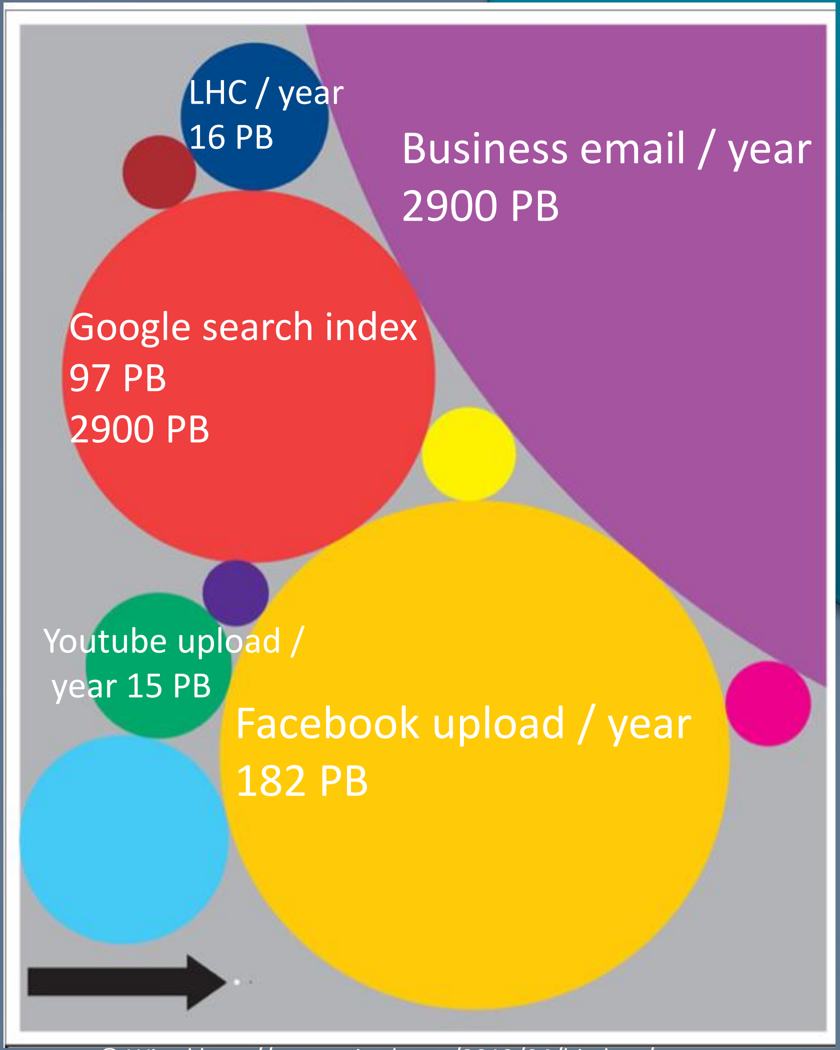
Network building & testing

- Core network will require a 500 port 100 Gbit/s device → this will be available
  - Internally probably a Clos (like) topology → need to carefully verify blocking factors and protocol
- Large scale tests require large system
  - Can test opportunistically in HPC sites

Current and future DAQ

	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

Talking about BIG DATA



© Wired http://www.wired.com/2013/04/bigdata/

Data processed by the LHCb software trigger per year from 2021

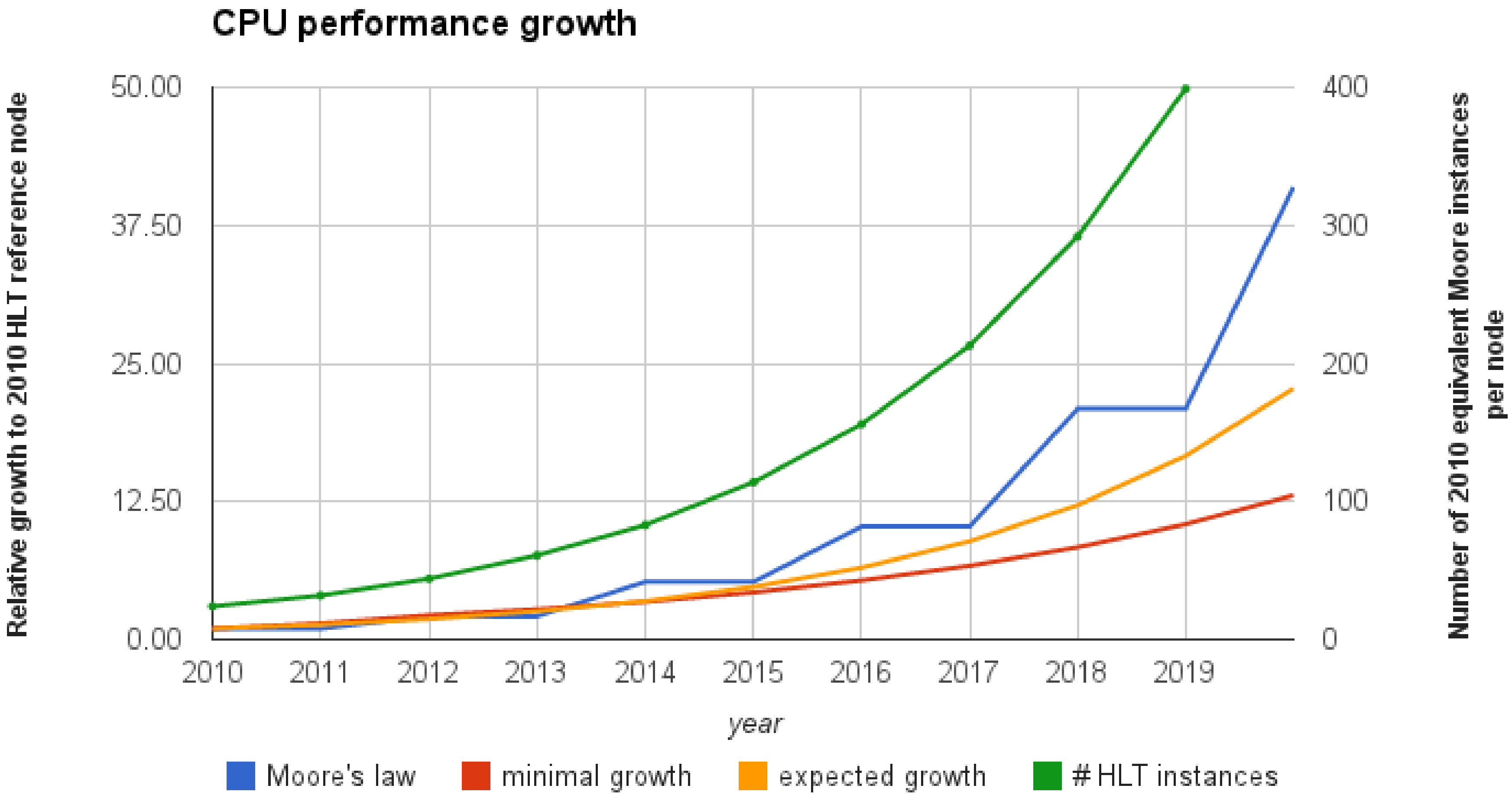
19000 PB

Summary

- The trigger-free readout of the LHCb detector requires
  - new, zero-suppressing front-end electronics
  - a 40 Tbit/s DAQ system
- This will be realized by
  - a single, high performance, custom-designed FPGA card (PCIe40)
  - A PC based event-builder using 100 Gbit/s technology and data centre-switches
- We are confident that all inherent challenges can be met at a reasonable cost

More material

Event-filter farm



Cost

Cost of the Online System	Cost [kCHF]
Event builder (network and PCs)	3600
Optical Fibres	1700
Controls network	905
Controls system (ECS)	930
Event-filter farm	2800
Infrastructure	775
Timing and Fast Control (TFC)	500

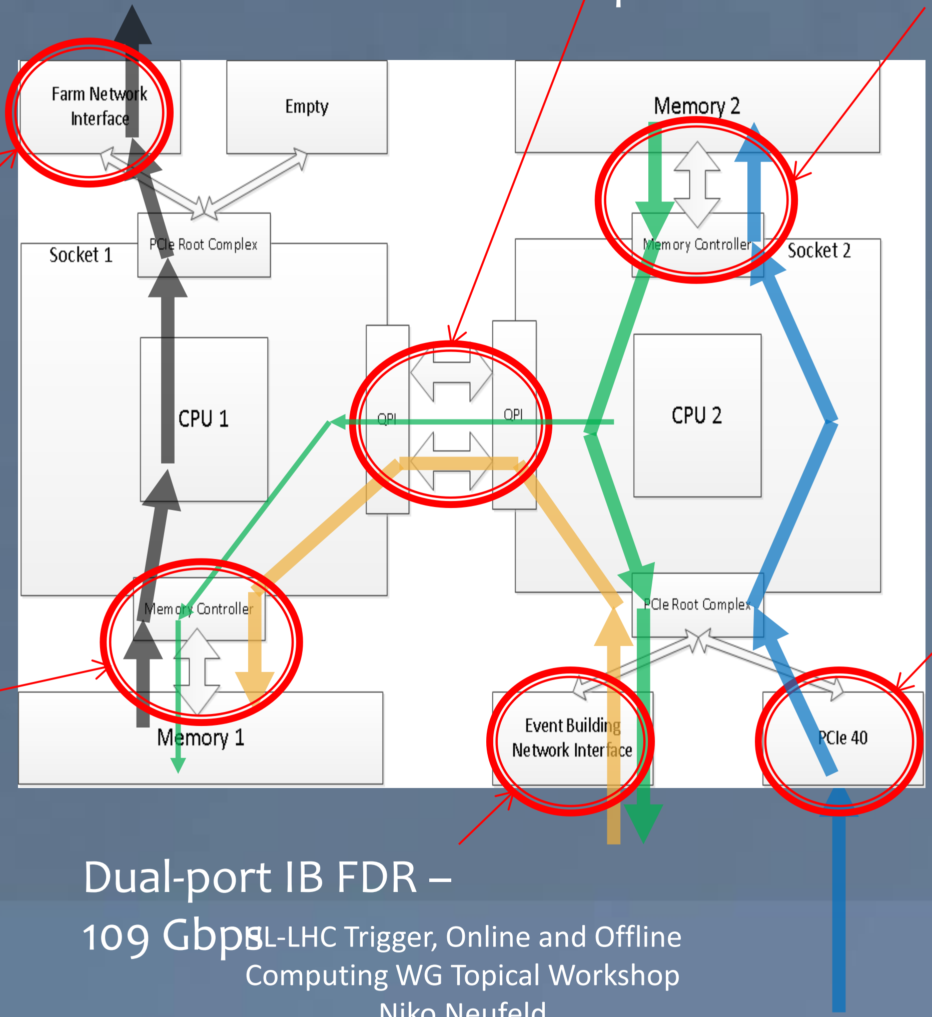
Performance tests

I/O Setup (2)

QPI link - 128 Gbps

Memory throughput – 200 Gbps

Dual-port IB FDR – 109 Gbps



Memory throughput – 200 Gbps

GPUgen – 110 Gbps

Dual-port IB FDR – 109 Gbps