# LHCb Trigger and Online TDR Online section

Niko Neufeld

LHCC Detector Upgrade Review, June 3rd 2014

# LHCb Online & Trigger TDR Online section

- 9 sections
  - System design
  - Long distance cabling
  - Readout board
  - Timing and fast control
  - Event building
  - Event filter farm
  - Experiment  control system
  - Infrastructure
  - Project Organisation

Readout System

# Recap - requirements

- Event rate 40 MHz – event building at bunch-crossing rate
- Mean nominal event size 100 kBytes
- Readout board bandwidth up to 100 Gbits/s
- Supports full software trigger (HLT and LLT)
- Uniform Experiment Control System (ECS) for control, monitoring and configuration of all aspects of the experiment
- Delivery of synchronous and fast commands and clock: the Timing and Fast Control (TFC)
- Can house up to 4000 CPU nodes (i.e. power-cooling and rack-space for 4000 dual-xeon servers)

# and of course the Online system should…

- …allow operation by a minimal crew (2 shifters supported by on-call experts)
- …be scalable, reliable, maintainable and - affordable

# Current and future DAQ

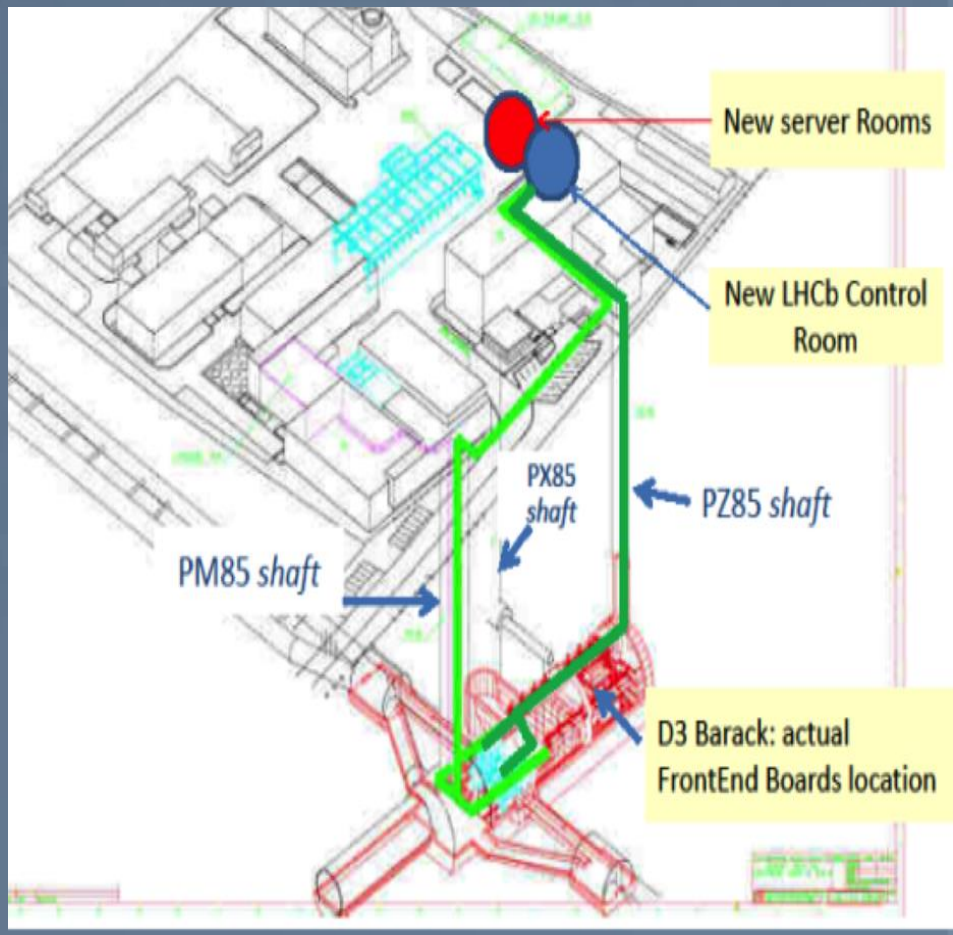
| | LHCb Run1 | LHCb Run 3 |
|---|---|---|
| Max. inst. luminosity | 4 x 10^32 | 2 x 10^33 |
| Event-size (mean – zero-suppressed) [kB] | ~ 60 (L0 accepted) | ~ 100 |
| Event-building rate [MHz] | 1 | 40 |
| # read-out boards | 313 | 400 - 500 |
| link speed from detector [Gbit/s] | 1.2 | 4.5 |
| output data-rate / read-out board [Gbit/s] | 4 | 100 |
| # detector-links / readout-board | up to 24 | up to 48 |
| # farm-nodes | 1500 | 1000 - 4000 |
| # links 100 Gbit/s (from event-builder PCs) | n/a | 400 - 500 |
| final output rate to tape [kHz] | 5 | 20 |
| final bandwidth to tape [MB/s] | 300 | 2000 |

# Cost optimisation

- ECS and TFC costs depend mostly on the number of detector components
- Event-filter farm cost depends on Moore's law and our capability (and imagination!) to exploit existing and emerging technologies:
  - multi-cores, non-x86, novel memory technologies etc…
- DAQ cost is driven by
  - number and type of interconnects
    - shorter → cheaper
    - faster → cheaper per unit of data transported
  - price of switching technology
    - telecom (feature-rich and expensive) vs data-centre (high-volume and inexpensive)

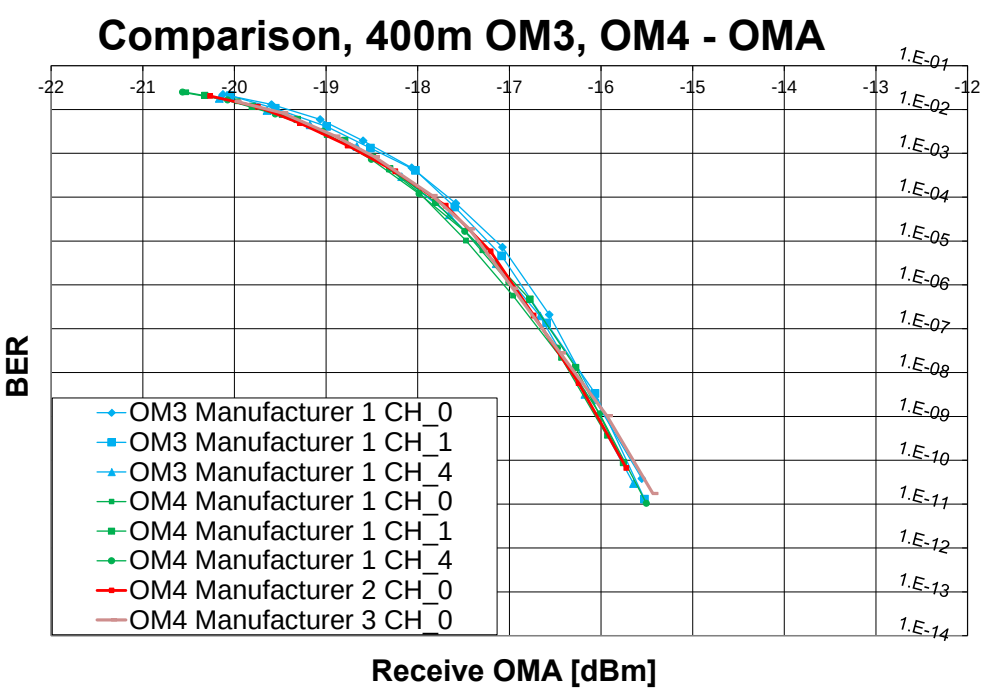
# Long-distance optical fibres

- 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

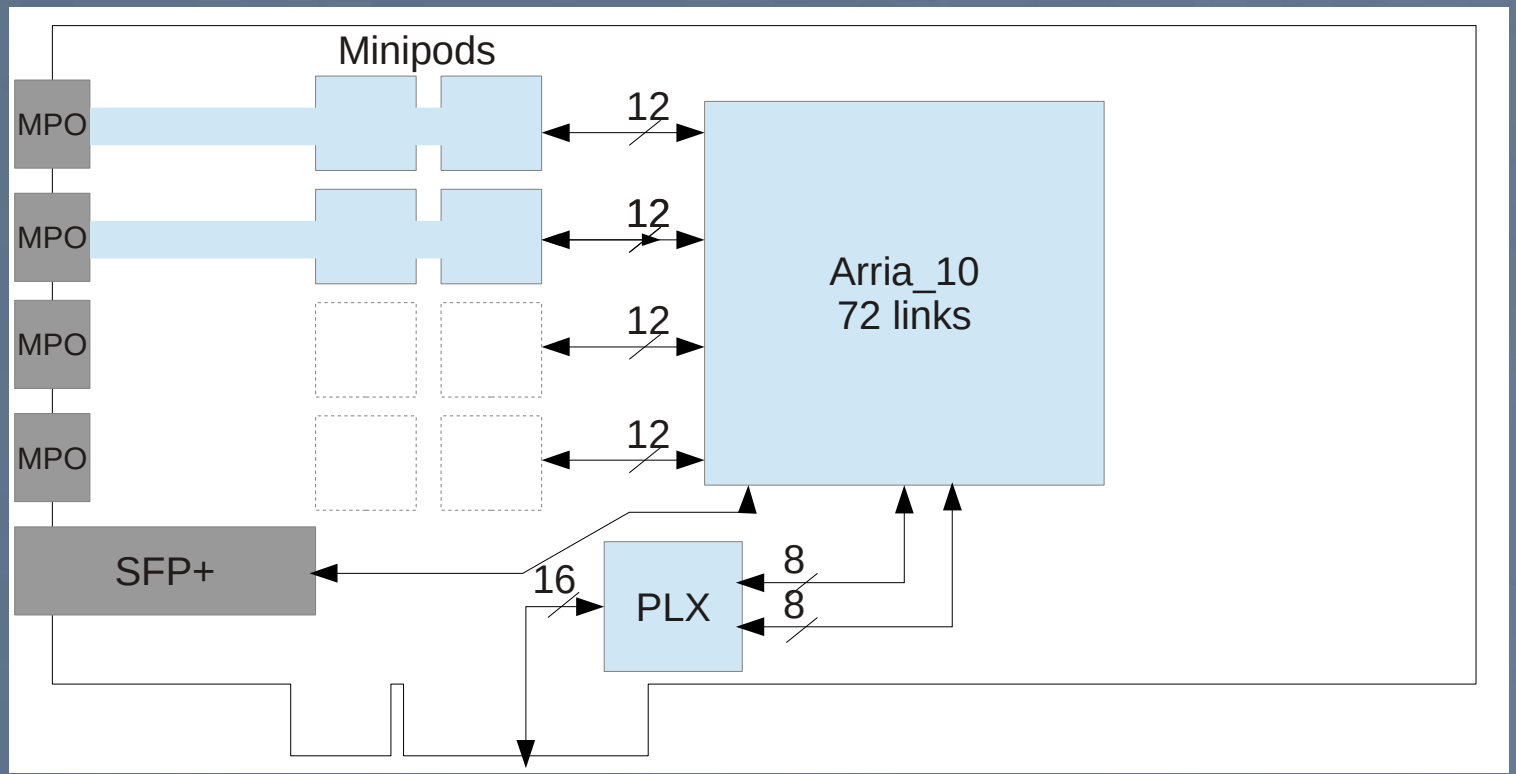


# Long distance versatile link tests

- 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



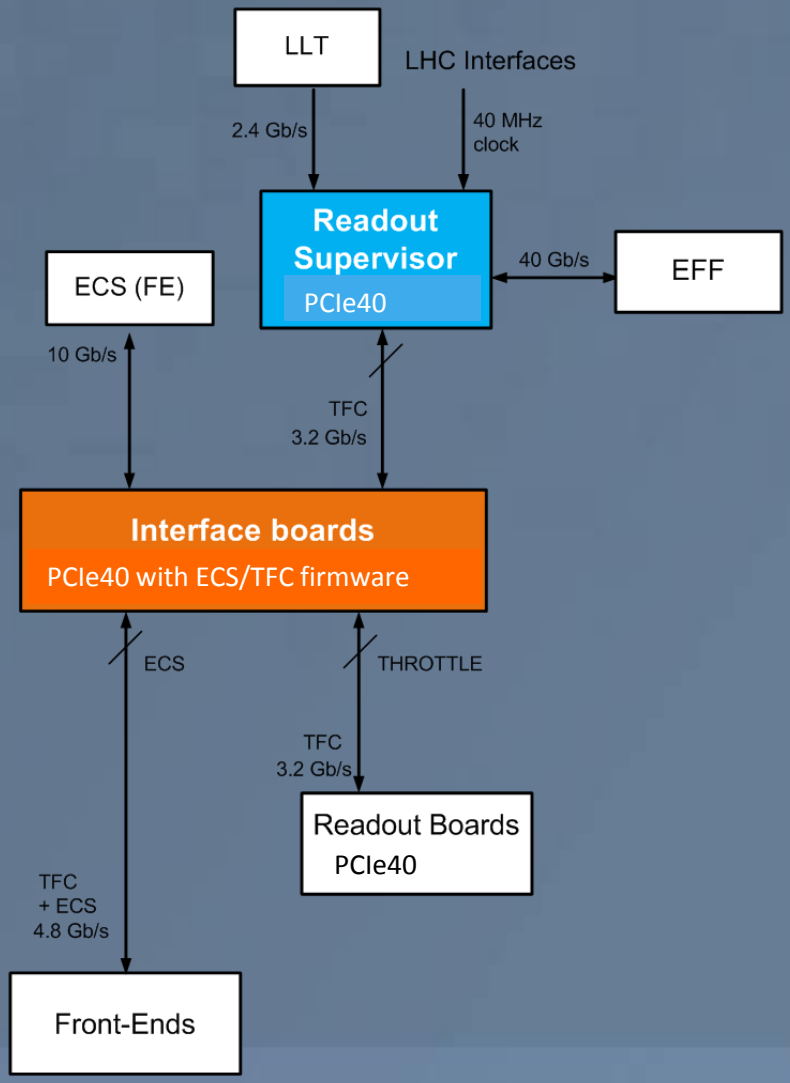
# The unique custom board: PCIe40



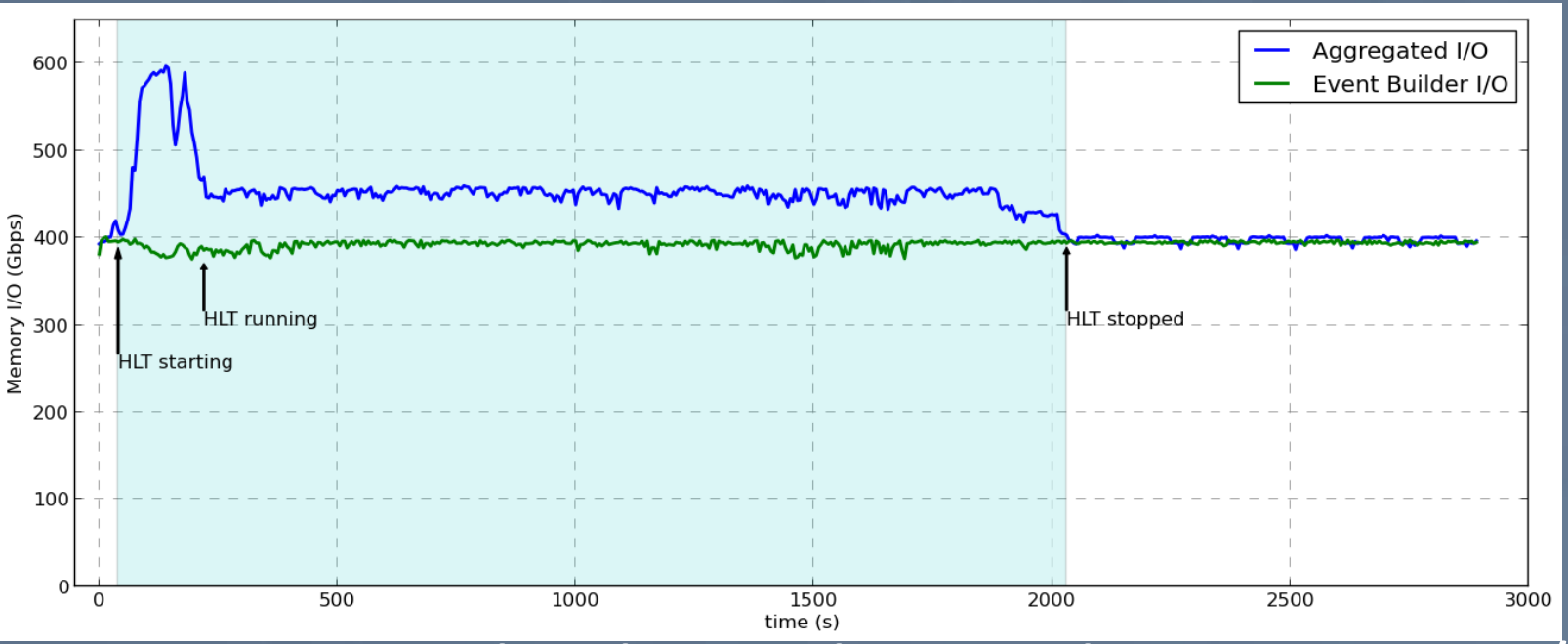

- Up to 48 bi-directional optical I/Os (GBT)
- Up to 100 Gbit/s I/O to the PC
- PCIe format removes need for external high-speed optical links
- Universal building block for DAQ, ECS and TFC

# Timing and Fast Control

- Main elements remain:
  - centralized flow-control "throttle"
  - support for event-management
  - distribution of all fast, synchronous commands
  - interface to relevant LHC data
- Technology upgraded from TTC to GBT
- Hardware platform is the PCIe40 with custom firmware

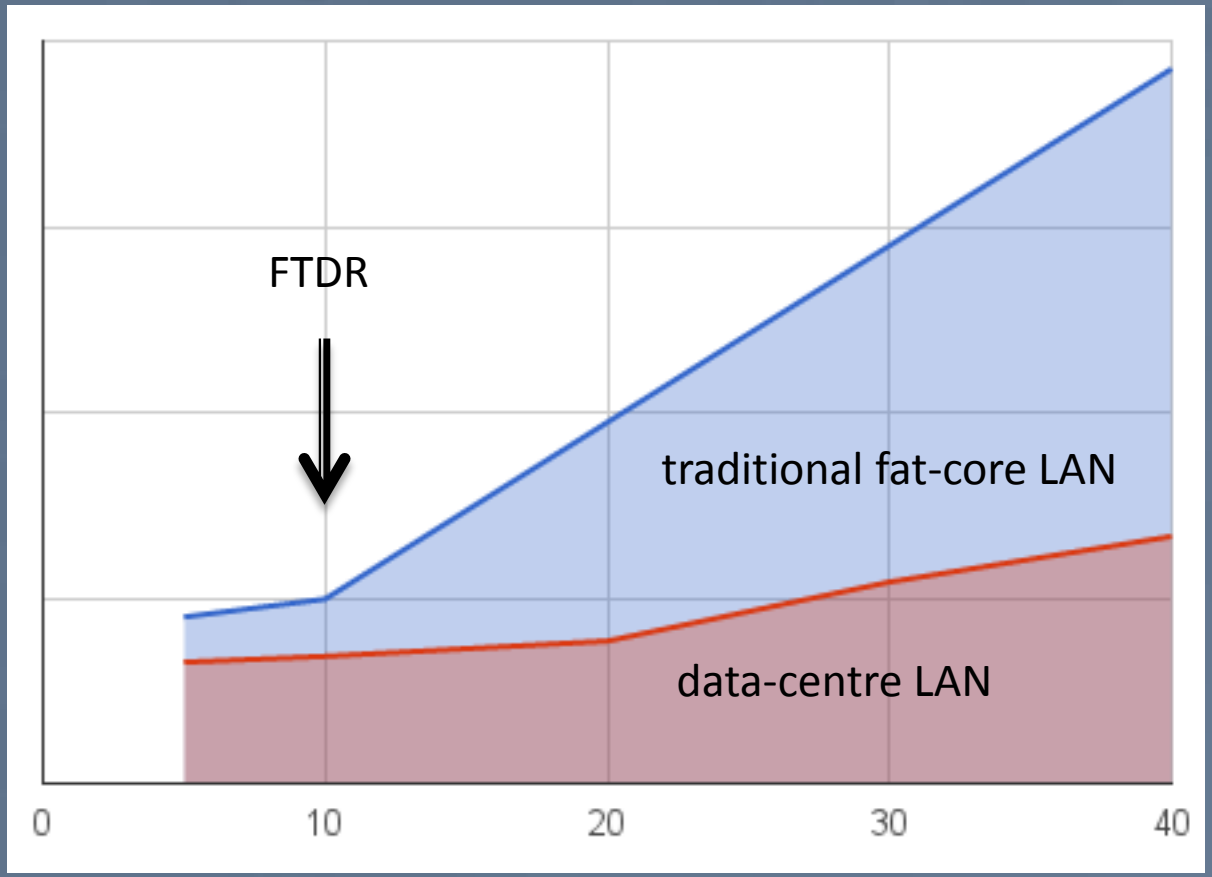


# Event-builder PC performance



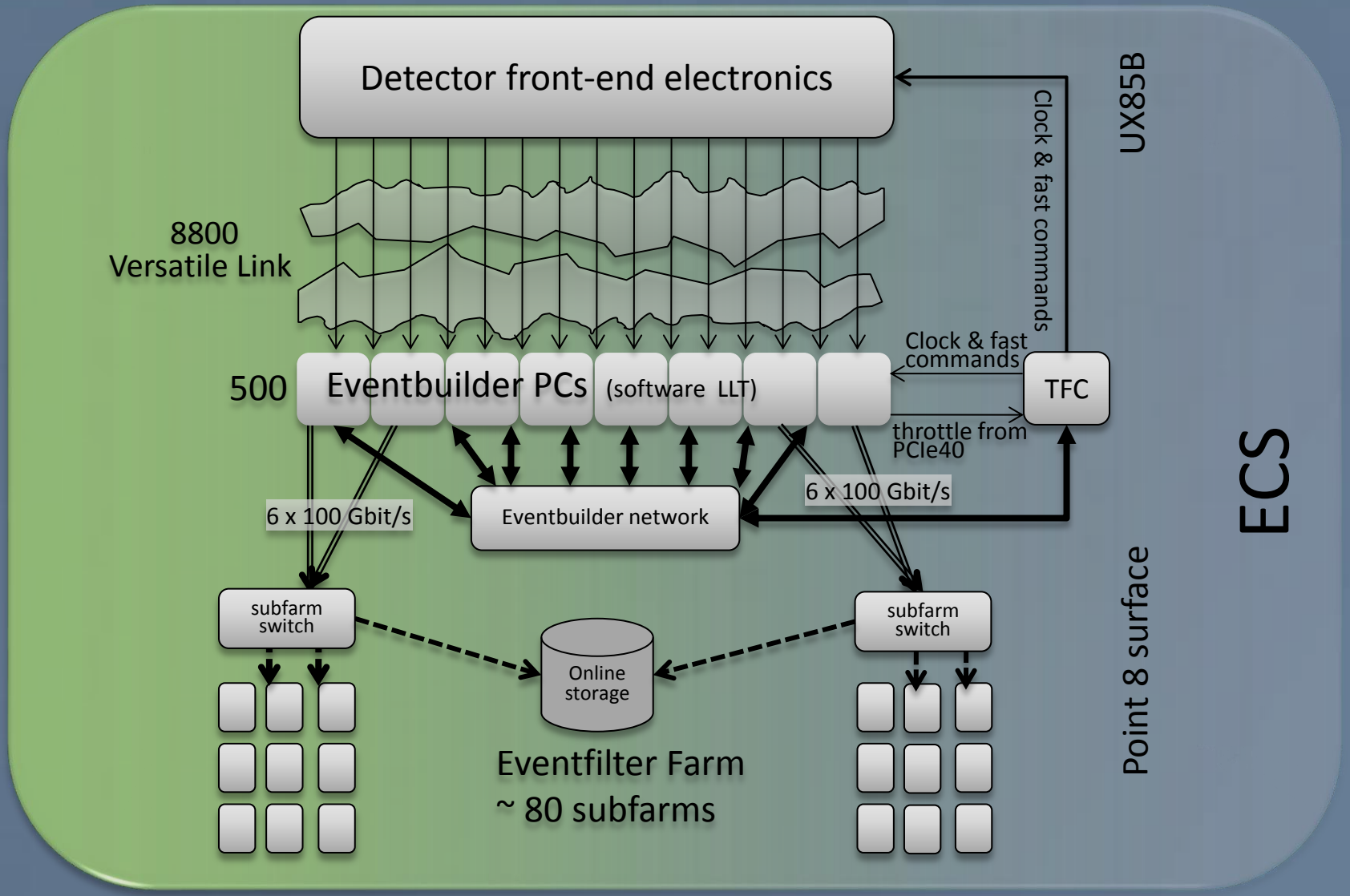

- Data-centre technology relies on the enormous I/O and memory performance of modern server PCs
- Hardware assisted networking leaves most of CPU performance available for software trigger, estimated to correspond to 2 ms / event in 2020

# Readout network with data-centre switching




- Cost of fat-core LAN based on 10 and 40G optical links
- Cost of data-centre LAN based on InfiniBand and direct attach cables

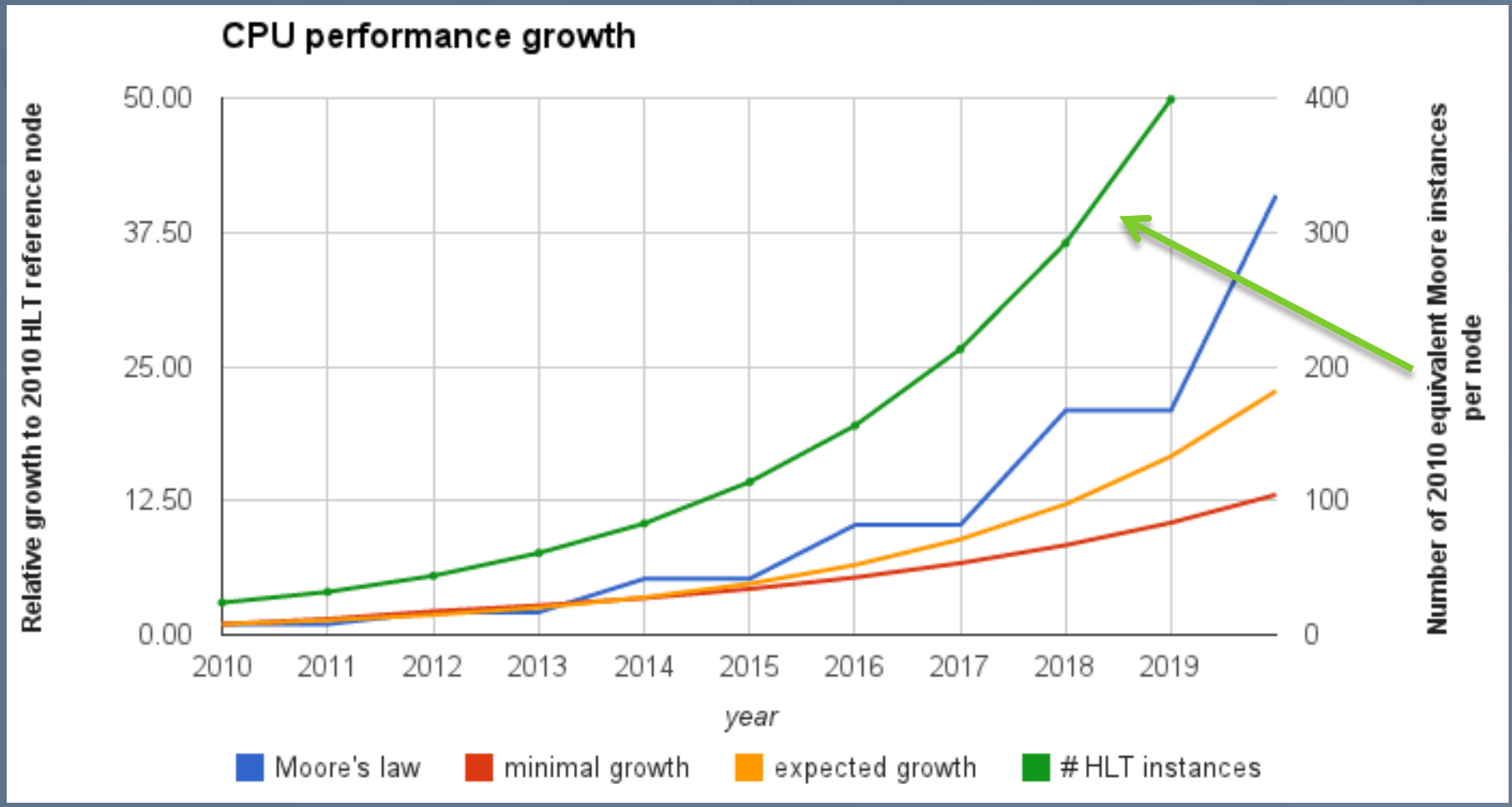
# Readout Architecture



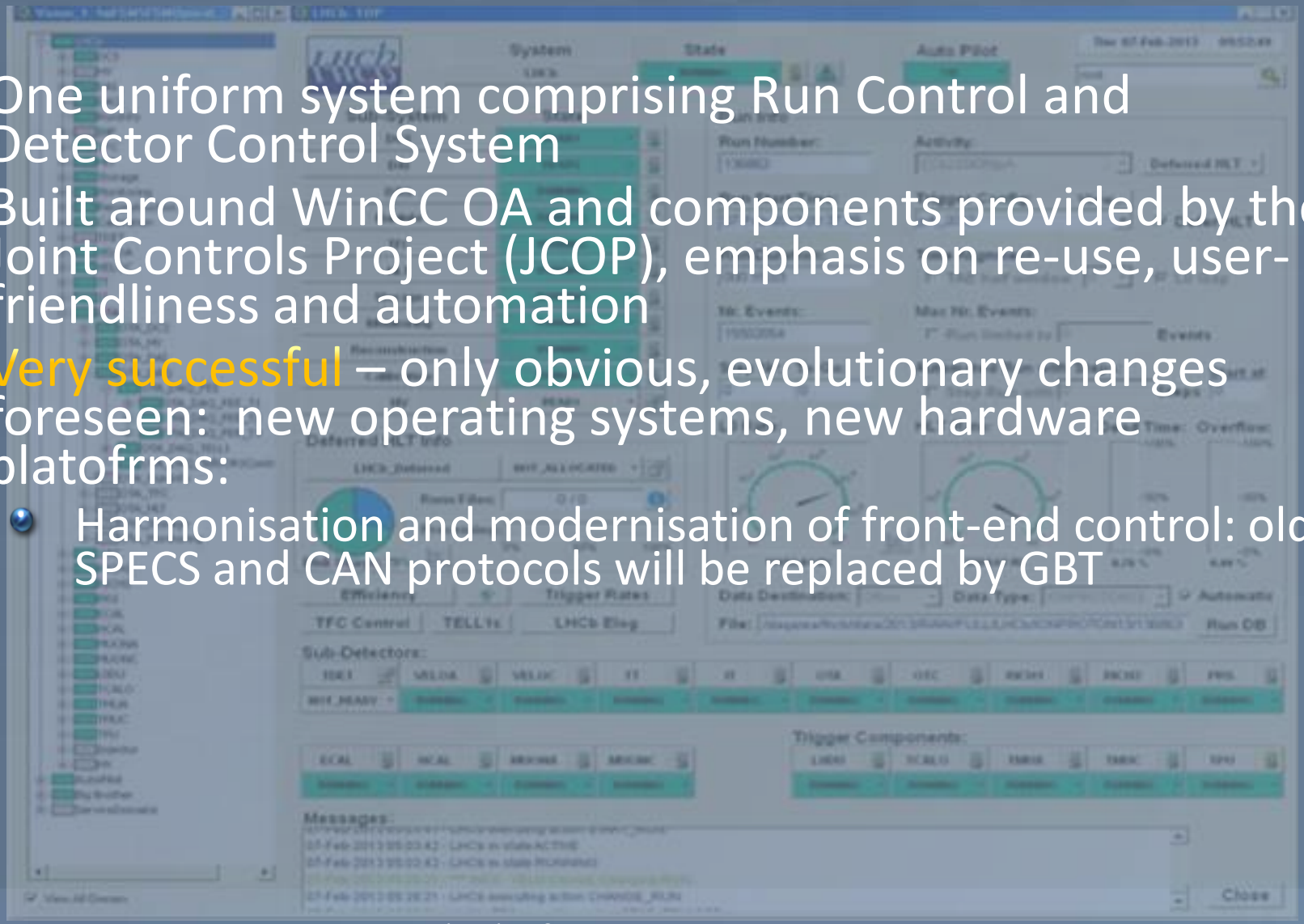
# Event Filter Farm

- Scalable architecture, resources can be added on the fly (also during physics) as necessary
- Can accommodate any technology which can be attached to a local area network
- Will be optimized for running the software triggers → less costly than general purpose facility
- Containerized data-centre allows to decide on optimal structure (mechanics, powering, etc… as late as possible)
- Mitigate against slippage of "Moore's law" by vigorous R&D to optimally use new and emerging technologies
- 1000 nodes of the assumed 2020 type allow 13 ms for the HLT

# Event-filter farm

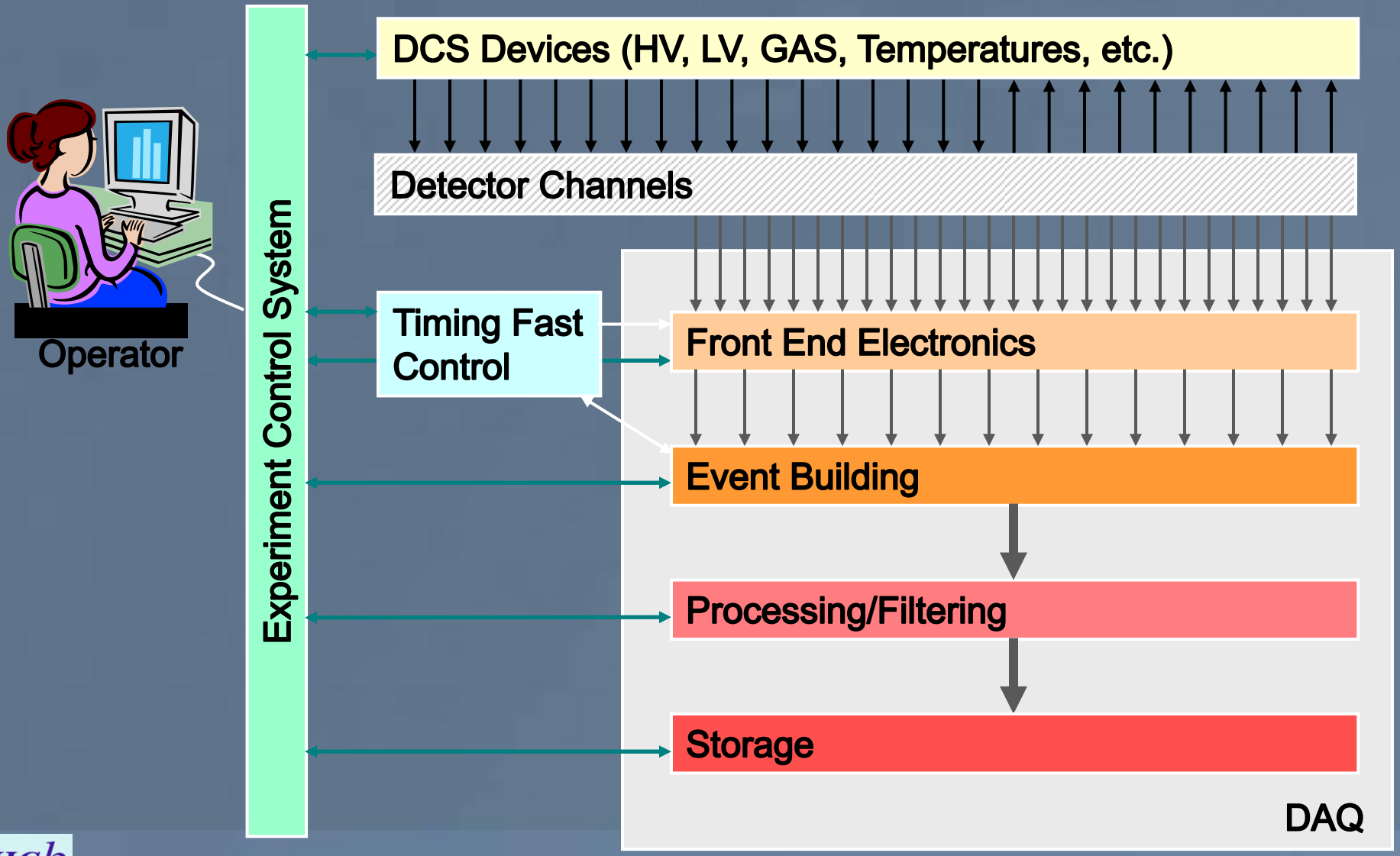


# Experiment Control System



- One uniform system comprising Run Control and Detector Control System
- Built around WinCC OA and components provided by the Joint Controls Project (JCOP), emphasis on re-use, user-friendliness and automation
- Very successful – only obvious, evolutionary changes foreseen: new operating systems, new hardware platofrms:
  - Harmonisation and modernisation of front-end control: old SPECS and CAN protocols will be replaced by GBT

# The Experiment Control System



# Infrastructure

- Infrastructure is a catch-all for ECS- and storage-networks (dedicated local area networks), IT infrastructure, storage, control-room equipment etc…
- Technically and – at least in the Online also financially – there is no problem to store 20, 50 or 100 kHz
  - a modern hard-disk can store about 150 MB/s → in principle 100 kHz require only 66 disks in parallel

# Project Organisation

- Long distance versatile link: CERN
- Readout board: CPPM, Bologna
- Firmware: LAPP, CPPM, Bologna, CERN
- Event-builder: Bologna, CERN
- (central) ECS and TFC, storage, online infrastructure: CERN

# Schedule

- Assume start of data-taking in 2020
  - System for SD-tests ready whenever needed → minimal investment
- 2013 – 16: technology following (PC and network)
- 2014 PCIe40 design and prototyping
- 2015 -16
  - PCIe40 Pre-series (6 months), followed by series production (18 monhts)
  - Large scale network tests
- 2017: tender preparations
- 2018: Acquisition of minimal system (full detector connectivity)
  - Acquisition of modular data-center
- 2019: Acquisition and Commissioning of full system
  - starting with network
  - farm as needed
- 2020: ready for data-taking

# Summary

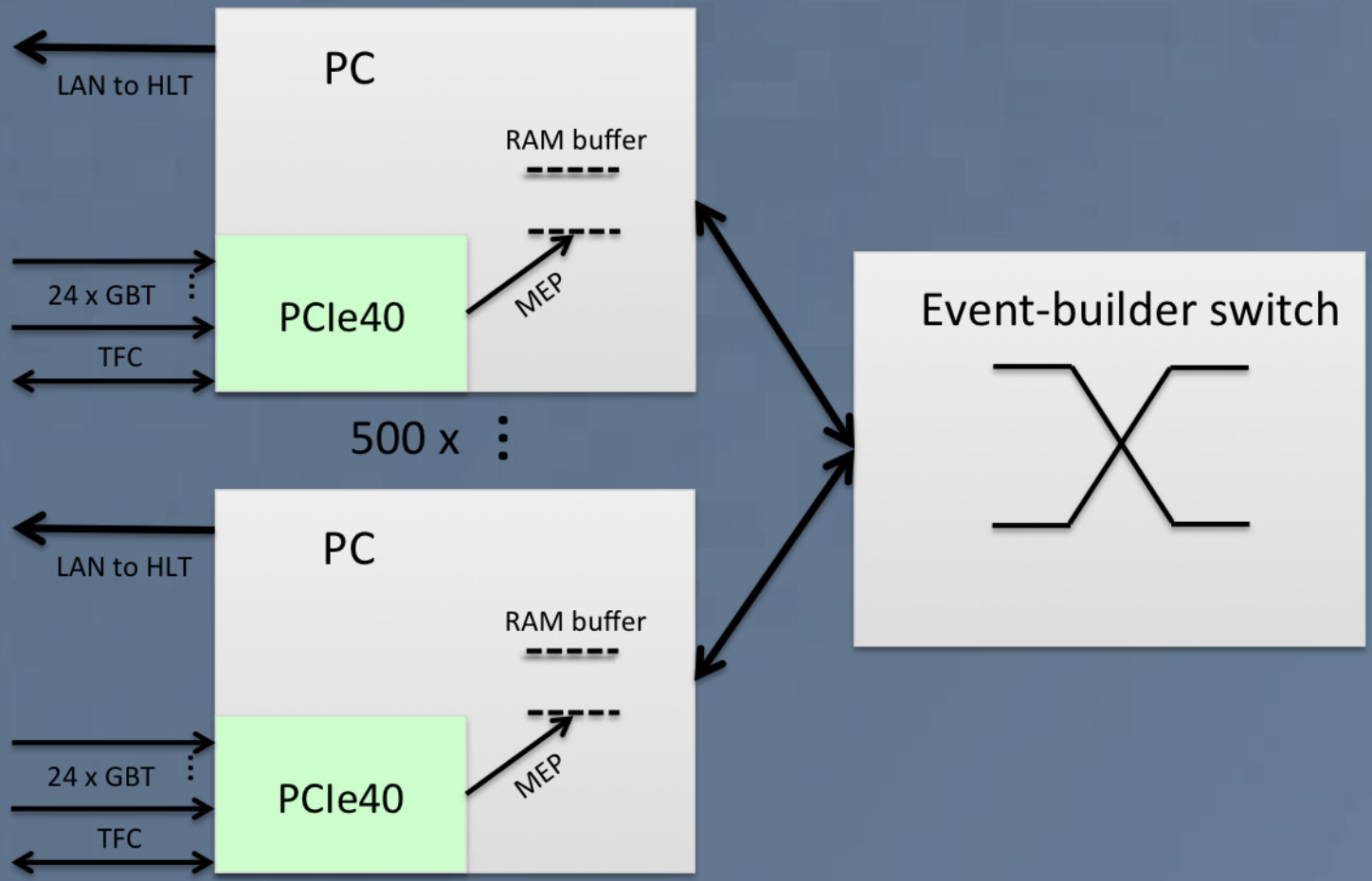
- The new LHCb Online system preserves the qualities of the existing system:
  - emphasis on simplicity ,COTS, architecture-centred design, common solutions
  - uniform experiment control
  - common solutions and standards: JCOP, Versatile Link, …
- The new LHCb Online system enables *trigger-free* read-out with full event-building at bunch-crossing rate
- A single custom made-board for both DAQ and ECS
- Unprecedented DAQ performance is cost-effectively achieved by
  - minimizing distances between online components
  - exploiting mass-market technologies (PCs, data-centre networks) driven by big-data applications

# More material

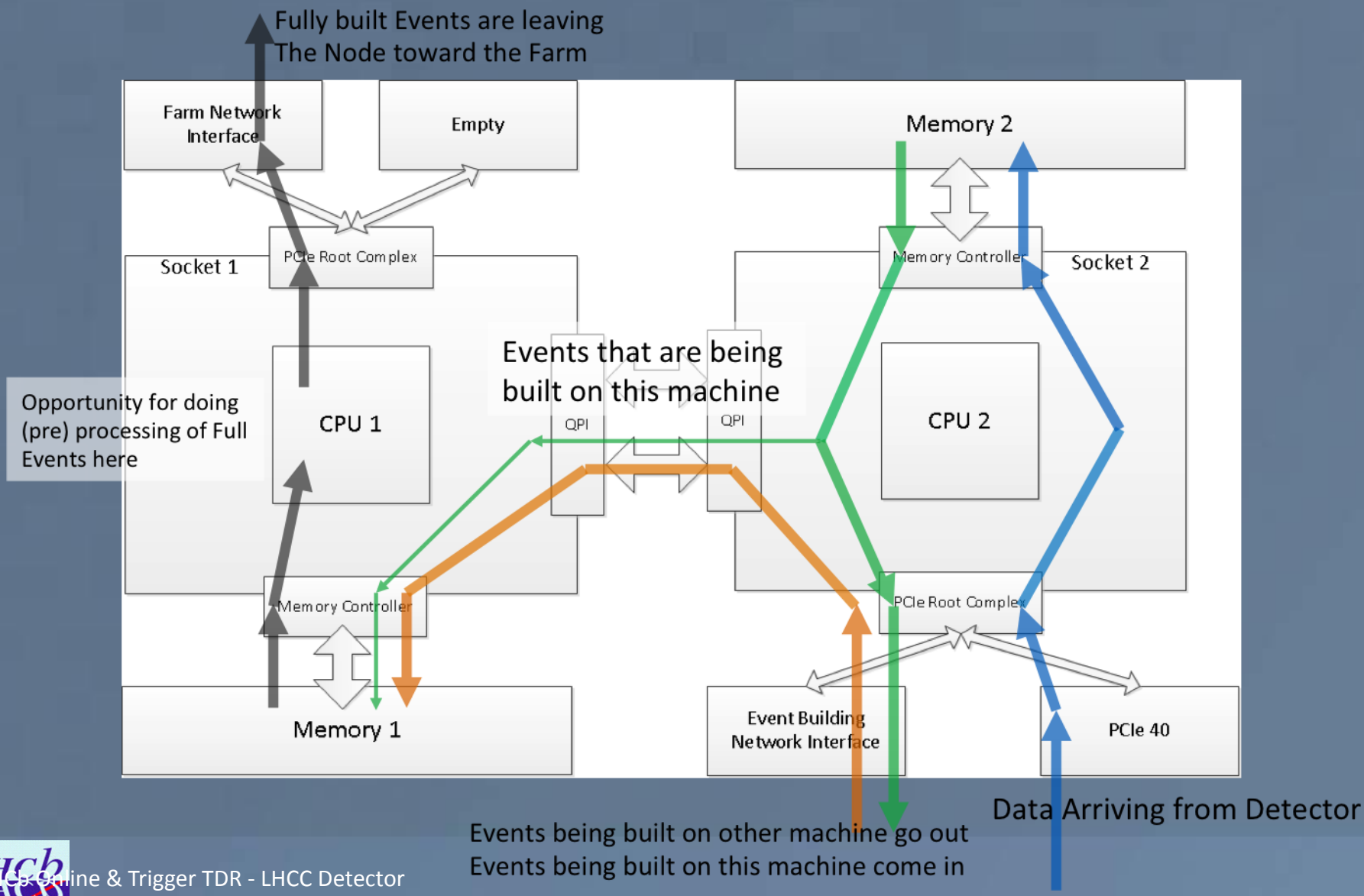
# Event-size

- The nominal event-size is 100 kB
- For about 8800 versatile links in wide-mode (4.5 Gbit/s) and 30 MHz non-empty crossings this corresponds to a link load of about 80%
  - Empty crossing-events should significantly smaller
- From occupancy in Monte Carlo one gets a slightly higher but compatible number (within 10%), however encoding is not yet optimised (overheads from protocol headers etc…)
- These numbers do not take into account reduction due to removal of duplicate numbers before event-building or final-storage (e.g. BCID)
- We think that 100 kB is a reasonably safe assumption: note that the theoretical maximum, assuming no empty crossings is 131 kB (@ 30 MHz).
- Event-building network is dimensioned for this worst case (+ a few % for DAQ protocol overheads)

# Event-builder data-flow: external



# Event-builder data-flow: internal



# R&D

- Reviewers recommend to do intense R&D not only on many-cores but also on
  - vectorization (data-parallelism)
  - memory-bandwidth optimisation
  - non-x86 CPU architectures (POWER, ARM)

| Country | Institute(s) |
|---|---|
| Germany | TU Dortmund |
| Italy | University and INFN Padova |
| Netherlands | NIKHEF, University of Groningen |
| Spain | University of Barcelona (with associate La Salle, Universitat Ramon Llull) |