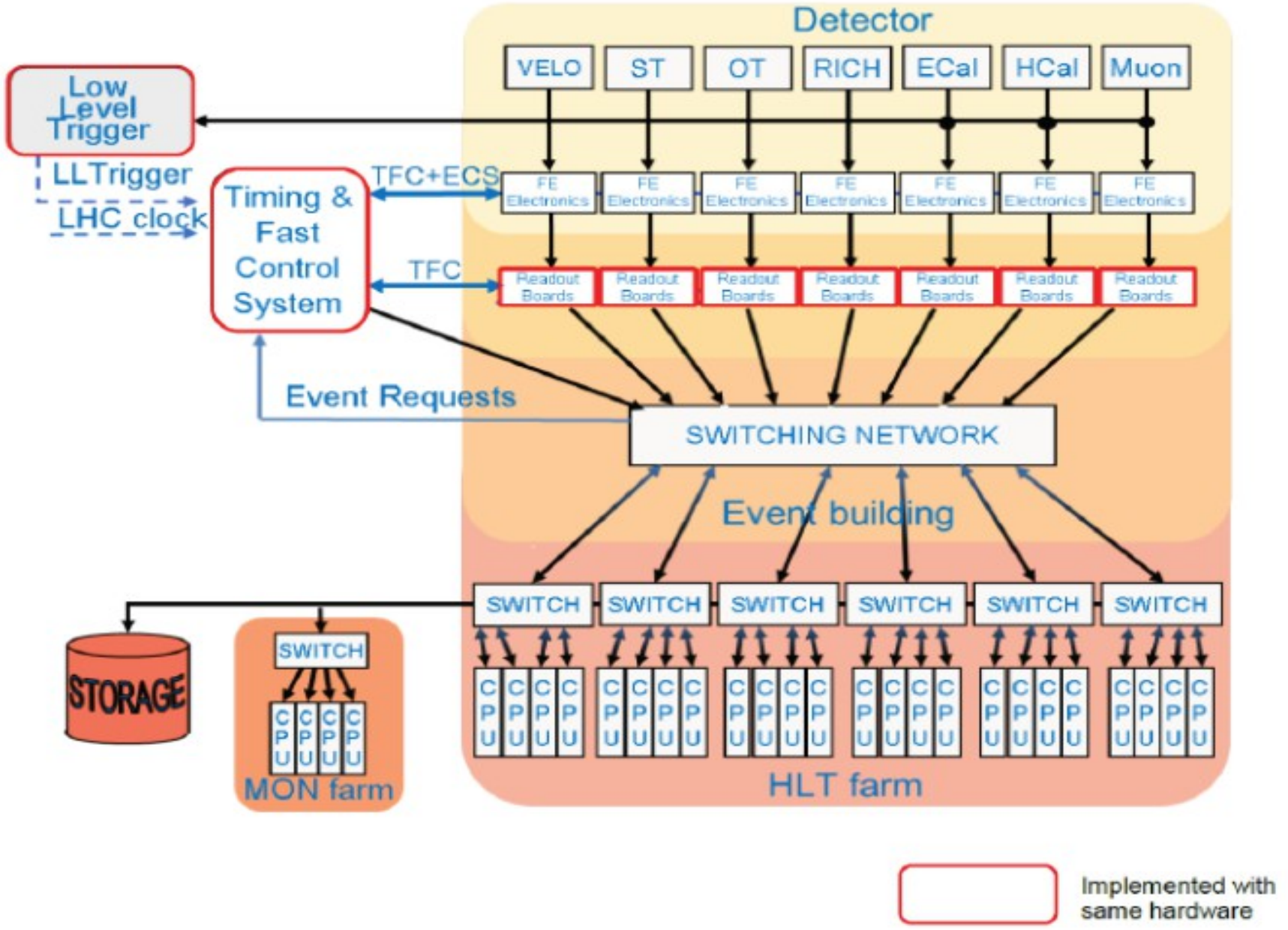
PCIe-gen3 links per LHCb-TDAQ upgrade (2018)

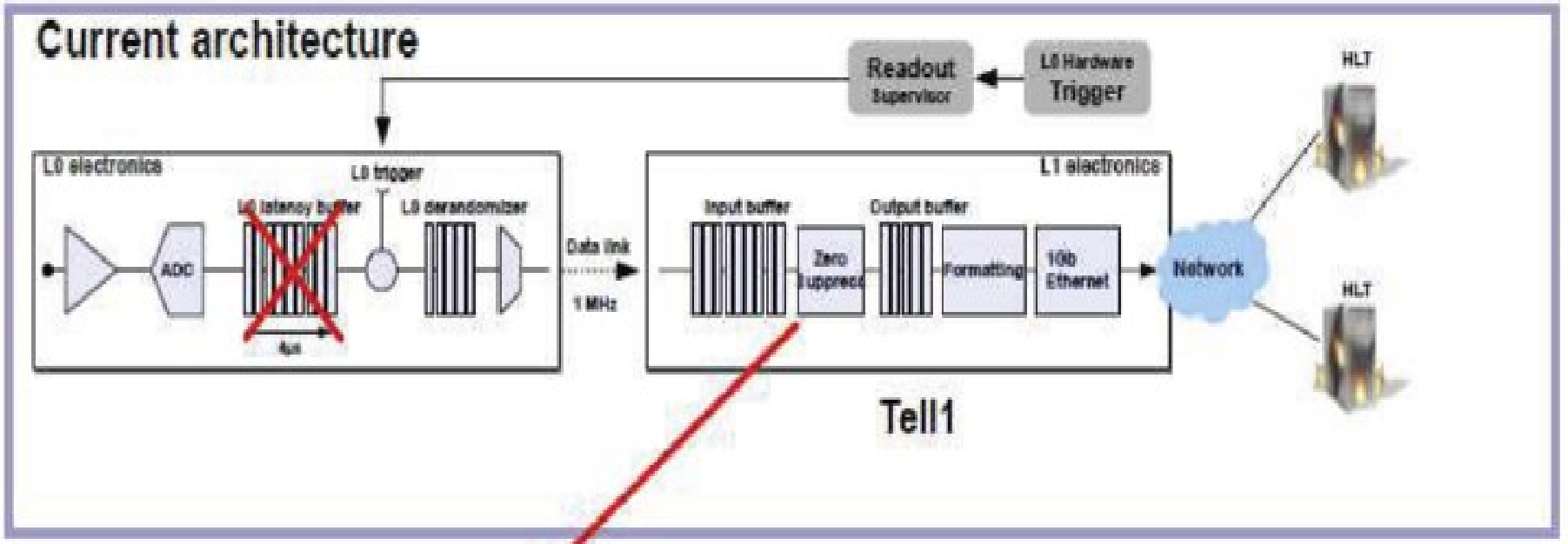
Slides copiate da talk Umberto Marconi (coord. TDAQ LHCb-Ita) per meeting con referee INFN (20/3/2013)

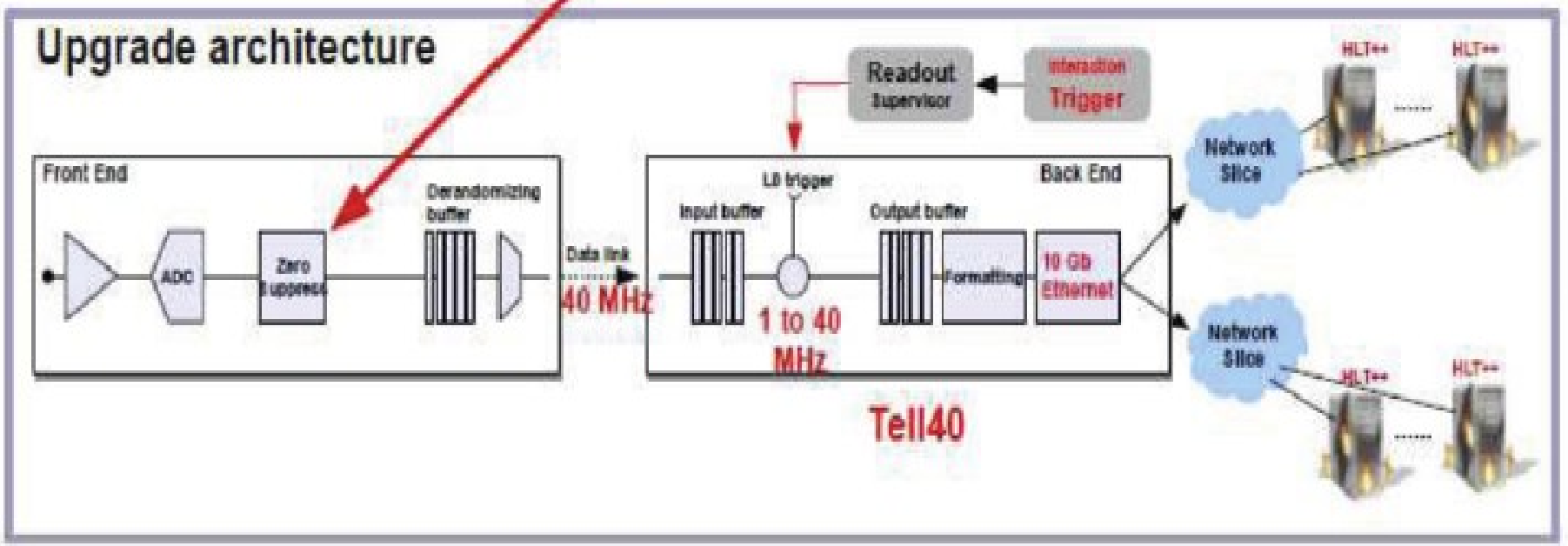
G.Collazuol

LHCb Data Flow (upgrade)

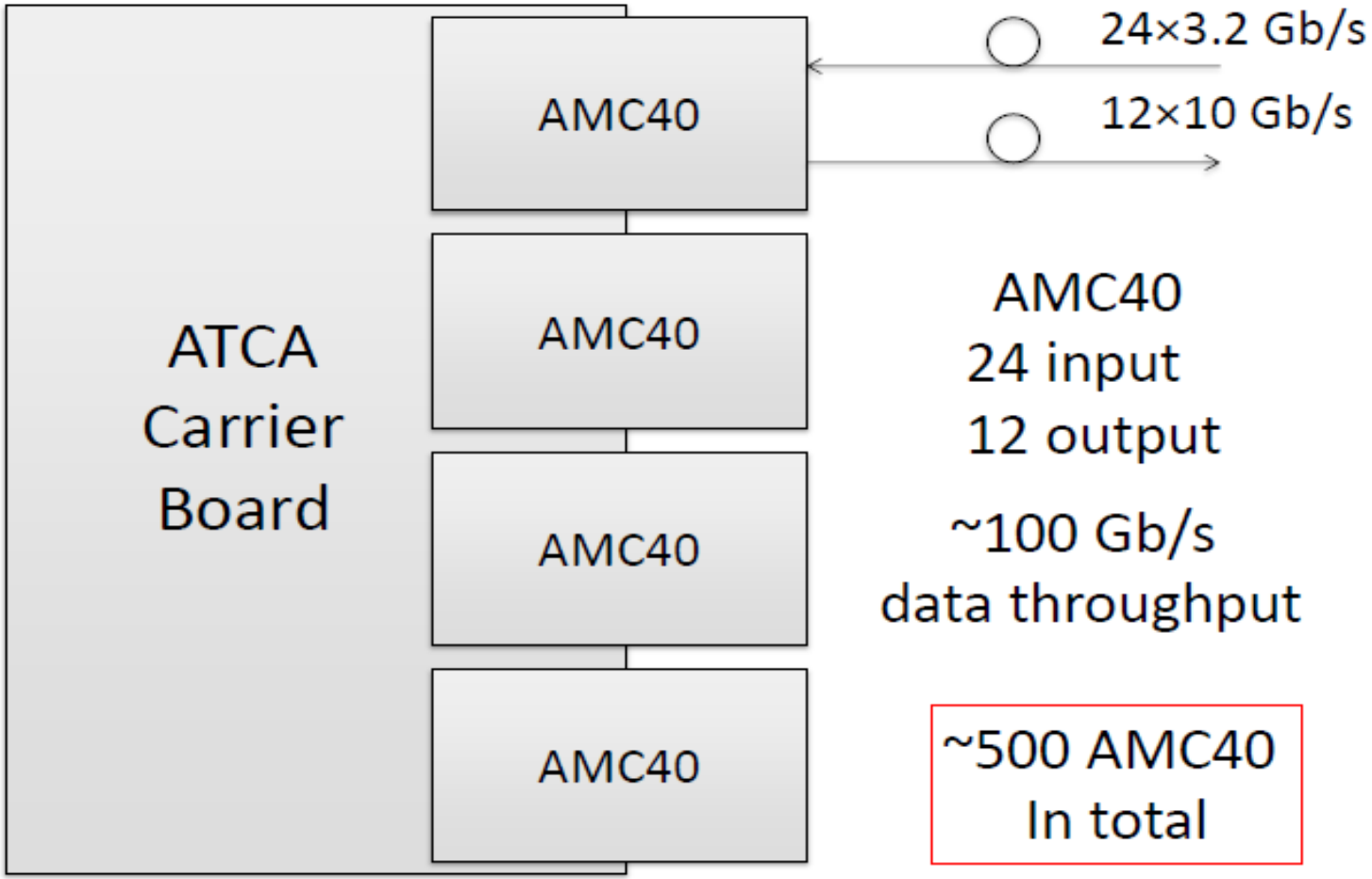


LHCb Data Flow





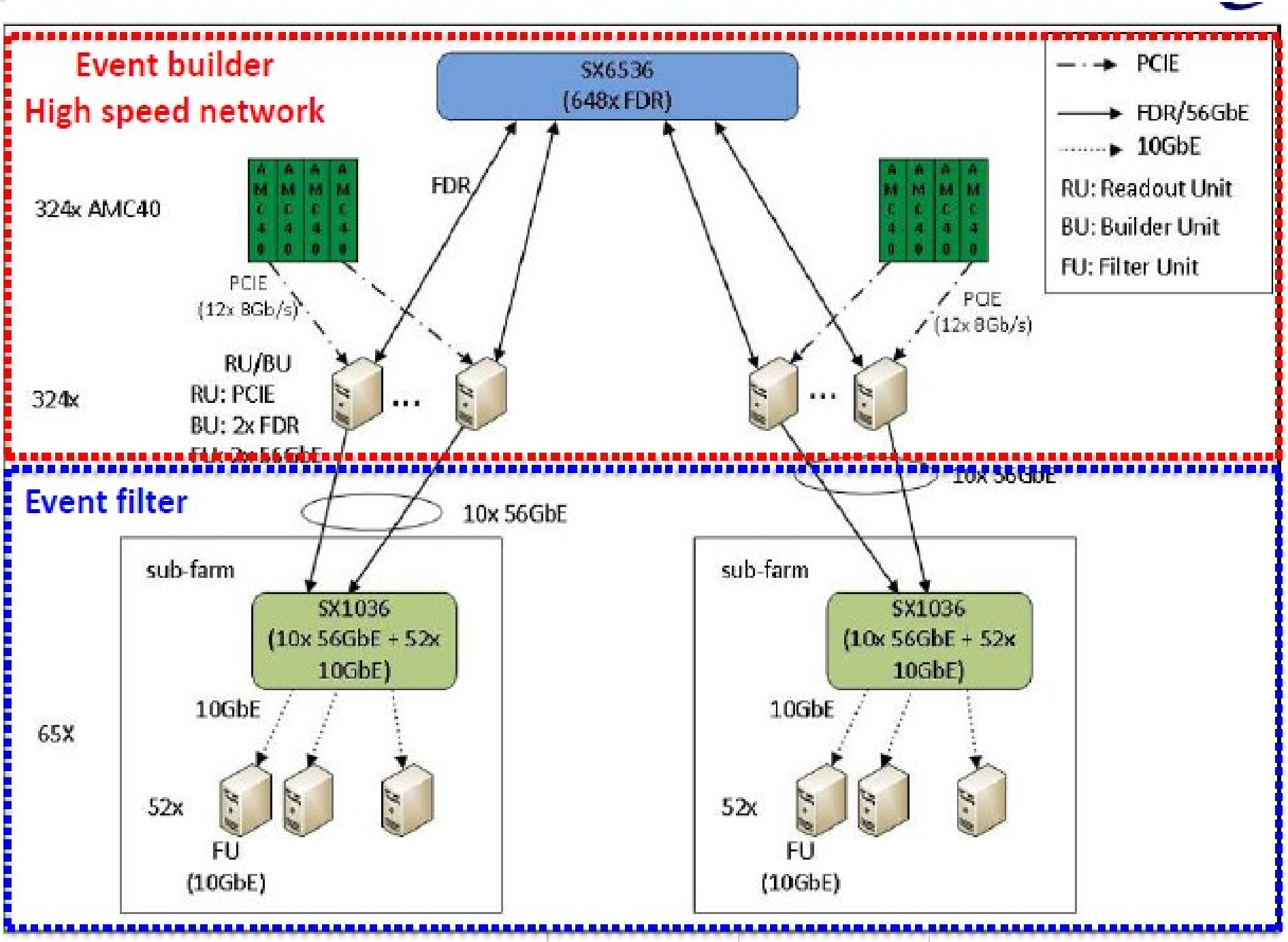
TELL40



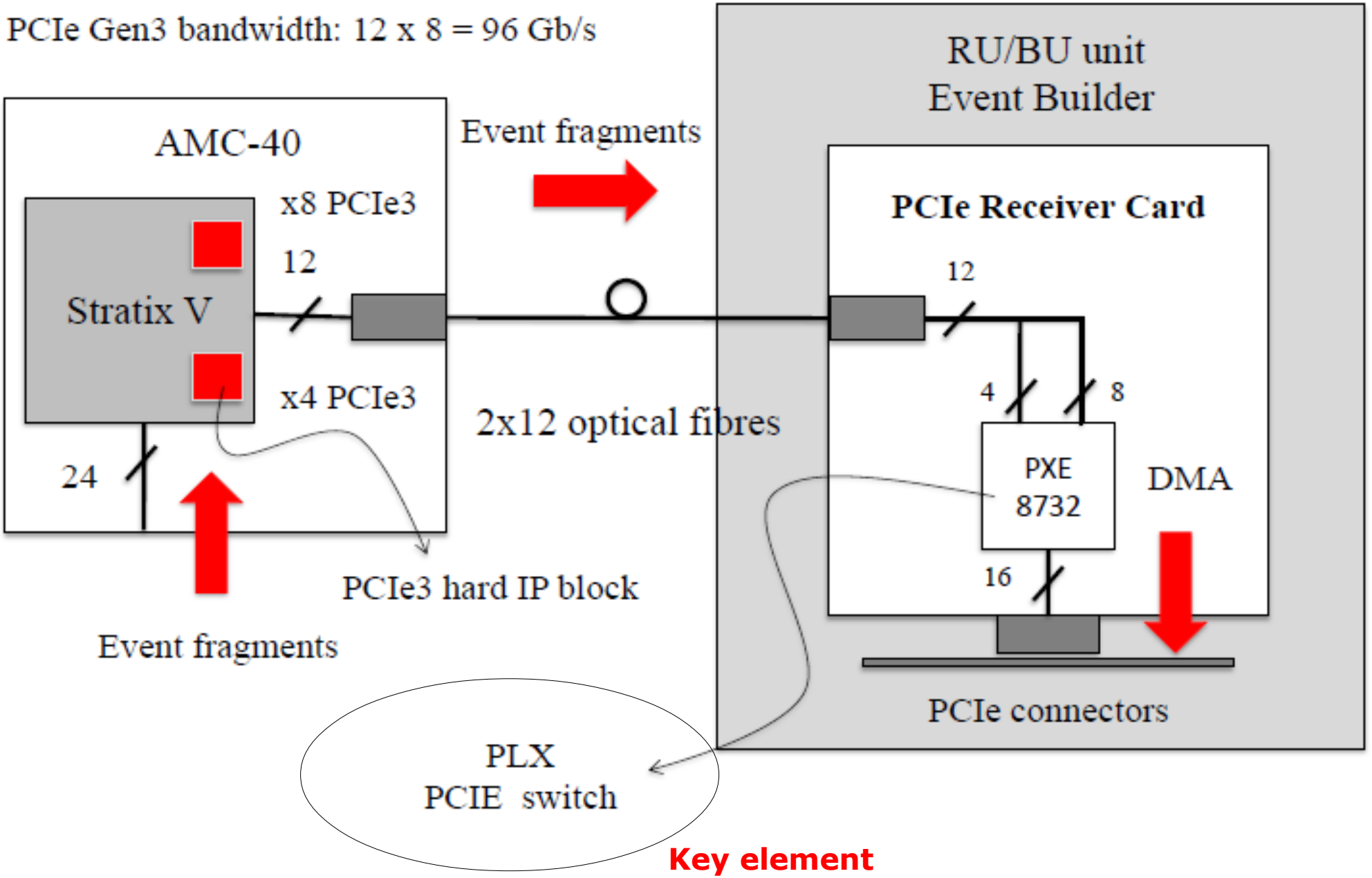
Contributo di LHCb-Bologna al progetto
- Link ottici (GBT) dal FE alle readout boards (TELL40)
- Sistema di readout (da FPGA sulle mezzanine delle AMC40 a HLT)

→ proposta per link AMC40-RU/BU PCs via PCIe-gen3 su fibra
(coivolgimento LHCb-Padova - M.Bellato e G.C)

PCIe-IB-ETH uniform cluster



Proposal for a PCIe-Gen3 extension



Proposal for a PCIe-gen3 extension

Abstract content
The architecture of the data acquisition for the LHCb upgrade is designed to allow for data transmission from the front-end electronics directly to the readout boards synchronously with the bunch crossing at the rate of 40 MHz. To connect the front-end electronics to the readout boards the upgraded detector will require order of 12000 GBT based (3.2 Gb/s radiation hard CERN serializers) optical links, for a corresponding aggregate throughput of about 38 Tb/s. The readout boards act as event buffers and the data format converters for the injection of the event fragments into the network of the High Level Trigger (HLT) computing farm. The connection between the readout boards and the HLT farm has to be designed to be capable to be seamlessly scaled up to the full readout of 40 MHz bunch-crossings. The data transfer rate will be tuned by means of a new Low Level Trigger (LLT) based on custom hardware, which will allow varying the HLT input frequency in a range between 10 to 40 MHz. A readout board consists of an ATCA compliant carrier-board, hosting up to four active AMC40-card pluggable modules (mezzanines). Each AMC40-card is equipped with a single powerful FPGA (likely a last generation Stratix V by ALTERA) used for establishing high-speed serial connections and for data processing. The AMC40-card as proposed today has 24 GBT input-links and 12 output-links. All the Stratix V FPGA serializers are 10 Gb/s. The 24 input-links deliver a maximum amount of user-data of 77 Gbit/s in the GBT standard mode. The baseline for the AMC40 foresees to implement a local area network protocol (LAN) directly in the FPGA. The candidate technologies considered so far are Ethernet and InfiniBand. An alternative solution for the read-out system is to send data from the FPGAs to the HLT farm via PCIe Gen3 bus extension/expansion (at the link-level PCIe does not look very different from the LAN protocols). Data in this approach would be pushed over a suitable physical link (optical fibre for instance) from the FPGA into a PCIe custom receiver card plugged to a HLT server motherboard. PCIe Gen3 would use 8 Gb/s on the serializers. The 12 output-links of the FPGA allows to set up two PCIe devices of varying lane-count (x4 and x8) for data transmission. The PCIe hard IP blocks available in the ALTERA FPGAs are very efficient: one 8-lane block uses less than 1% of the resources. The PCIe custom receiver card consists of an optical-to-electrical transducer plus a PCIe switch chip used to adapt to the PCIe slot of the HLT server. The main architectural advantage of using PCIe Gen3 is that the LAN protocol and link-technology can be left open until very late to profit from the most cost-effective industry technology available by the time of LS2.

- proposta da testare in 2013/14 per eventuale preproduzione 2015/16 e produzione 2016/17

- BONUS: se il flusso (12GBs) di dati per PC puo' venire scritto direttamente in memoria GPU invece che in RAM
  → LLT in PC e non HW

PCIe-gen3

A Demonstration of PCI Express Generation 3 over a Fiber Optical Link

PLX Technology and Avago Technologies

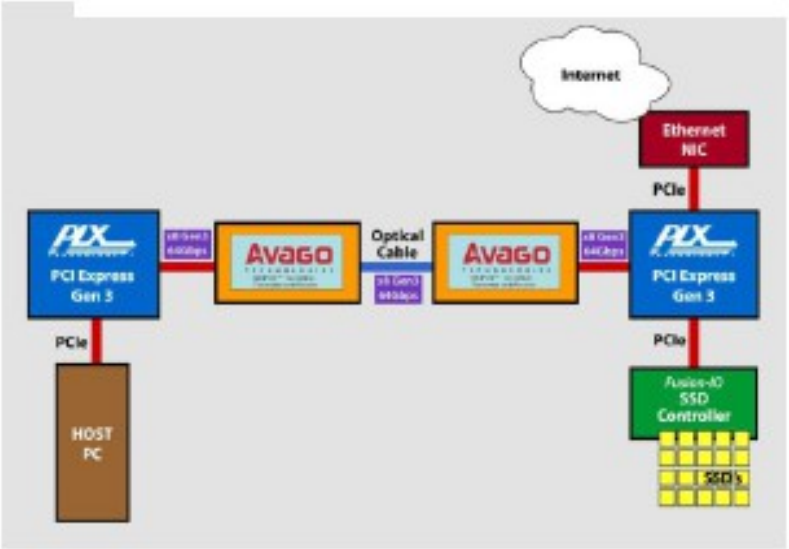
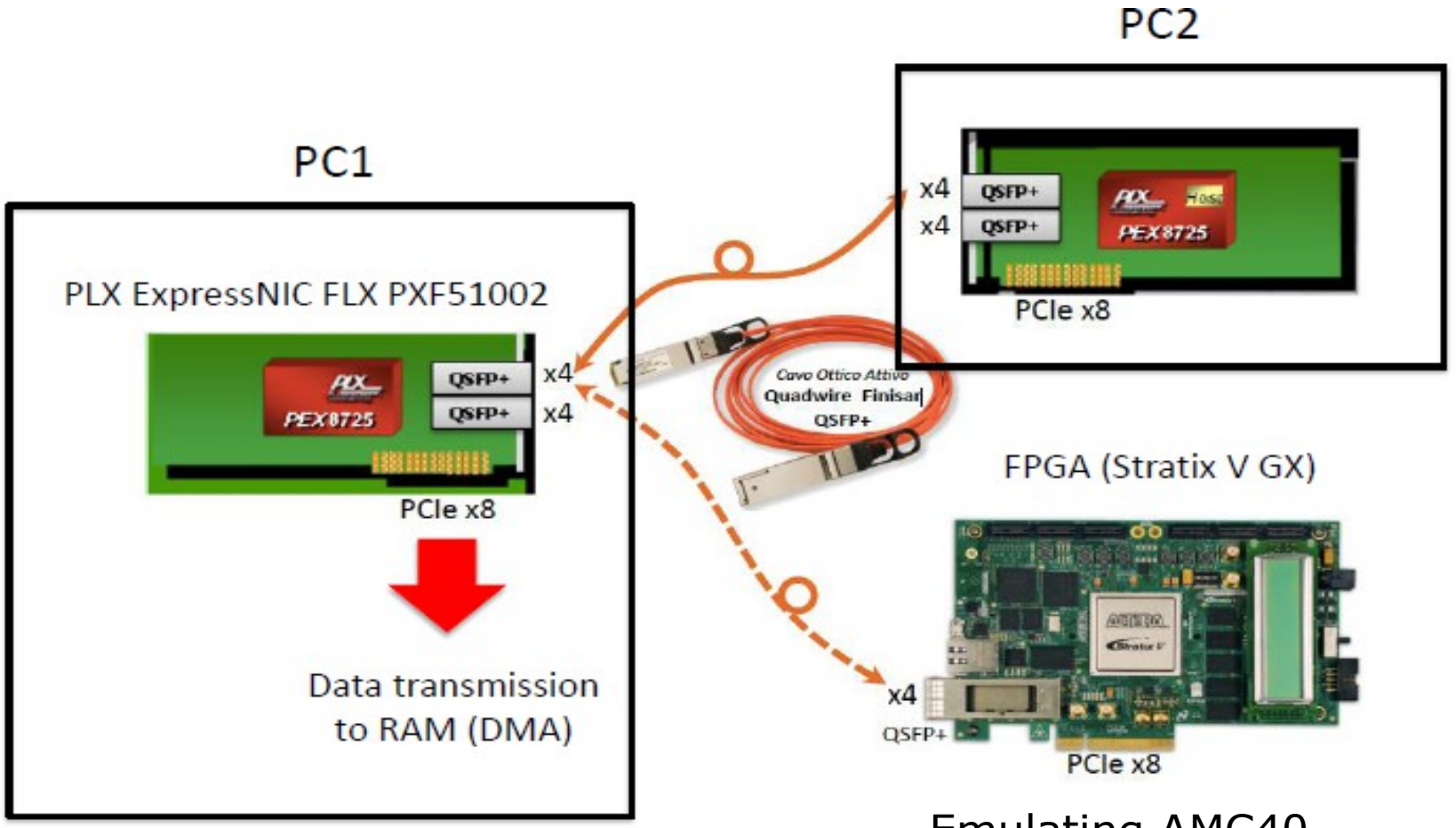


Figure 2. PEX8748 SI card with Avago Technologies MiniPOD™ adapter

http://www.plxtech.com/download/file/2346/224022

R&D test setup



Emulating AMC40

INFN Pd/LNL (M/Bellato) ha costruito DAQ di AGATA su links PCIe-gen2

- c'e' gia` hw setup per test di questo tipo
- c'e' anche expertise driver linux di basso livello necessari per gestire in modo efficiente lettura/scrittura dati da bus PCIe-g3

- test scrittura RDMA su GRAM ...

Proposal for a PCIe-Gen3 extension

- The PCIe receiver card. The most flexible solution will probably be a x16 card, which allows to present two devices of varying lane-count (x4 and x8).
- Optical and copper cables for the PCIe links.
- A high-speed, zero-copy device Linux driver working in tandem with a firmware on the AMC40 to push data into the PC-memory.
- Control and monitoring software for both the Linux driver and the firmware.
- Study of I/O throughput of modern PCIe platforms for the event-builder nodes. Dual-socket systems today offer 80 PCIe Gen3 lanes, this is a theoretical throughput of 1.2 Tb/s
- Measurement of the residual CPU power for event-filtering of the event-builder nodes.
- Establishment of system-limits and tunings to ensure reliable data-flow.
- Event-building software for uniform event-building network.

HLT GbE unidirectional cluster

LHCb Letter of Intent. It has been used for cost estimate in the FTDR

Unidirectional usage of the core ports is inefficient