The LHCb DAQ system

Domenico Galli
Alma Mater Studiorum – Università di Bologna
e INFN, Sezione di Bologna

SuperB Computing Workshop
Frascati, 17th December 2008

Outline

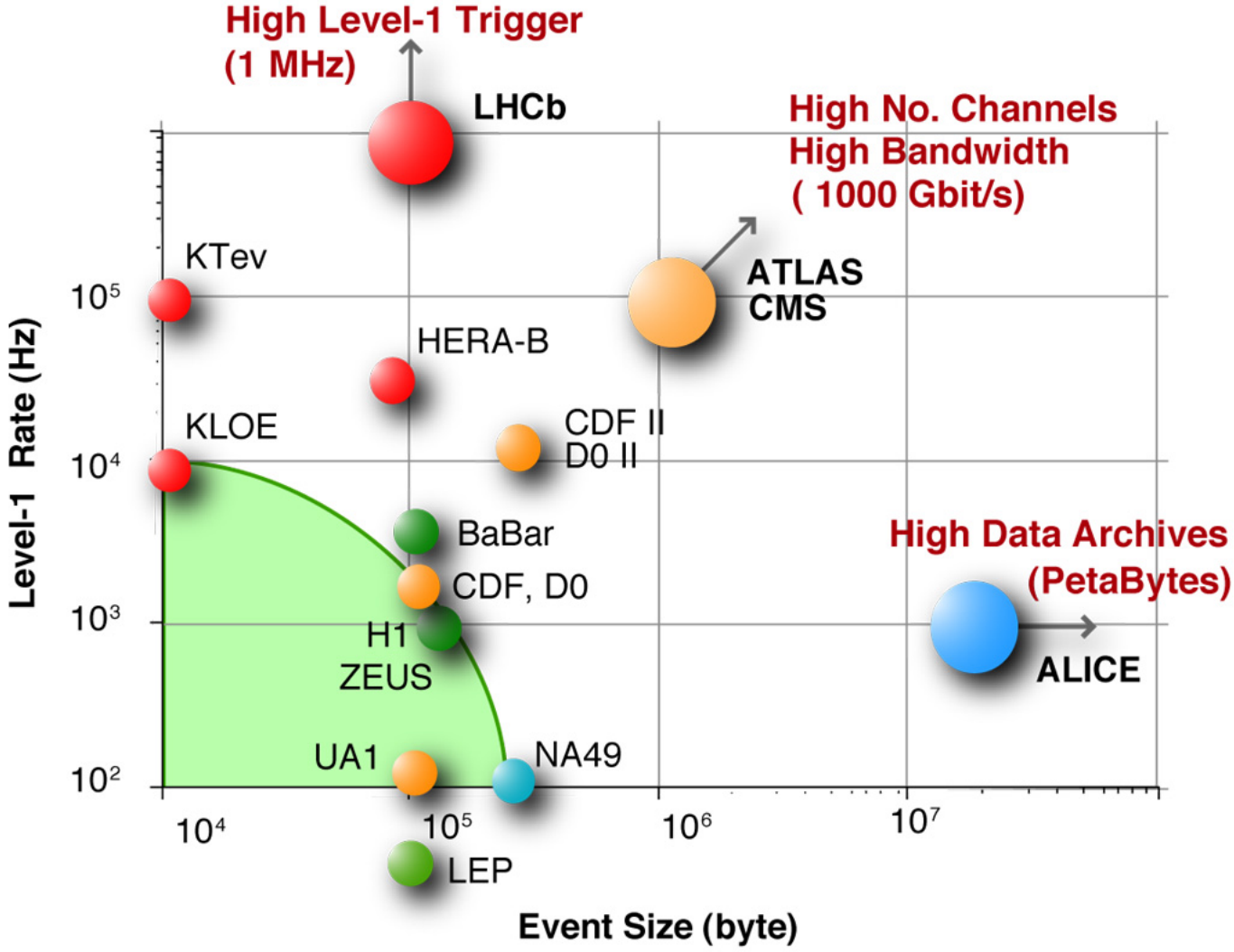
- LHCb feature:
  - LHCb rates.
  - Comparison with other experiments.
- LHCb DAQ/trigger overview:
  - Layout.
  - Key design features.
  - Data flow.
  - Components:
    - Timing and fast control.
    - Readout boards.
    - Readout network.
    - Event filter farm.
- Beyond LHCb:
  - 10 Gb/s technologies.
  - Readout boards.

LHCb Rates

- Experiment raw data:
  - LHC bunch crossing rate: **40 MHz**.
  - LHCb luminosity: **2 x 10^{32} cm^{-2} s^{-1}**:
    - **Single interaction** preferred to better match the B decay to its production vertex.
  - Visible cross section: **70 mb**.
    - At least 2 tracks in the acceptance.
  - Event rate: **10 MHz** (events with at least one interaction).
  - Event size: **35 KiB/event**.
    - **330 GiB/s**
- Events of interest:
  - Beauty production cross section: **500 μb**.
  - **100 kHz** beauty pairs.
  - Branching fraction of the events of interest: **10^{-6} ÷ 10^{-5}**.
  - **~10 Hz** of events of interest (all HLT channels).
- Events to be written to tape:
  - **200 Hz** of **exclusive B meson** decay modes.
  - **1.8 kHz** of **inclusive b-decays** and calibration signals.
    - **68 MiB/s**.

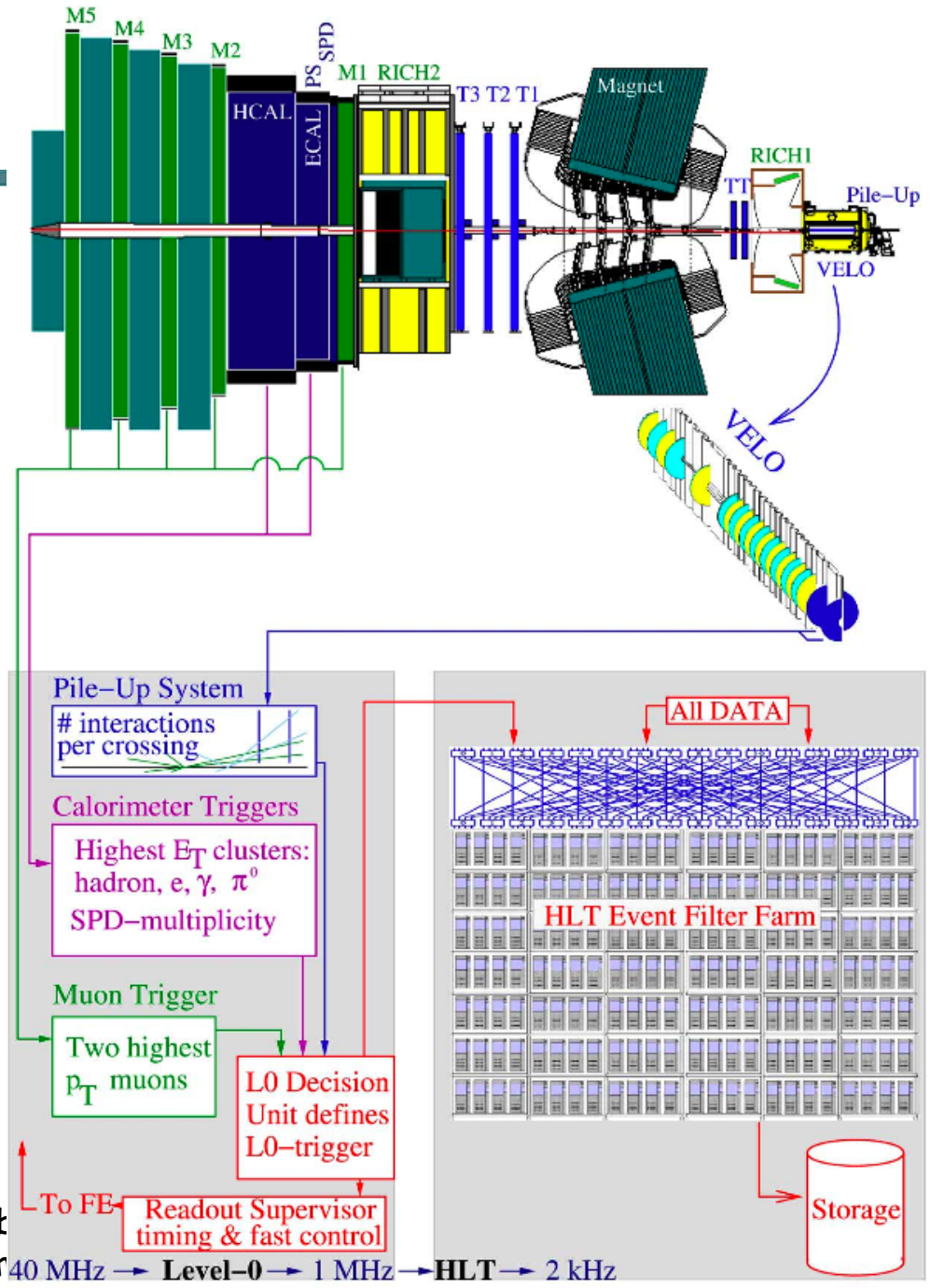
INFN Istituto Nazionale di Fisica Nucleare

Trigger Rate and Event Size Trends

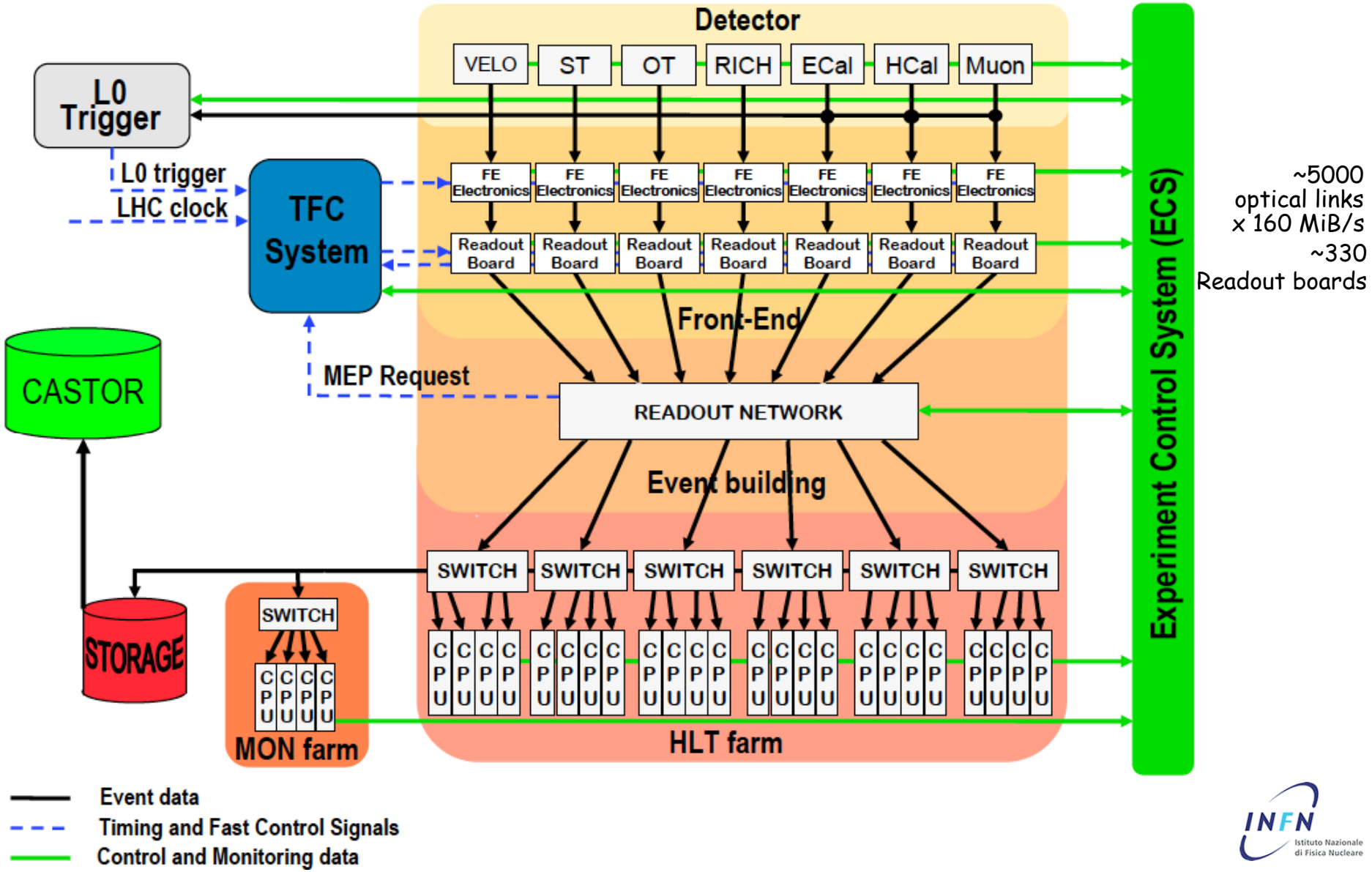


LHCb Trigger Overview

- **2 stage** trigger system:
  - **Level-0**: synchronous in hardware;
    40 MHz → 1 MHz.
  - **HLT**: software on CPU-farm; 1 MHz → 2 kHz.
- **Front-end Electronics**:
  - interfaced to Readout Network.
- **Readout network**:
  - **Gigabit Ethernet** LAN.
  - Full readout at **1.1 MHz**.
  - Throughput: **60 GiB/s**.
- **Event filter farm**:
  - ~ 2200 1 U servers.
  - ~ 18000 CPU cores.



The LHCb DAQ/Trigger Layout



LHCb DAQ: Key Design Features

- Event data **read-out** from the detector via **~5000 optical links** (~100 m long):
  - Maximum raw bandwidth: **160 MiB/s per link**.
- Optical links are fed into **~330 read-out boards**.
  - **TELL1**, custom electronics.
  - Zero suppression, time de-randomization, etc.
- On average **every 1 μs** new data become available at each of the ~330 read-out boards (**~100 B/board**):
- Data from **several** 1 μs cycles ("triggers") are **concatenated** into **one IP datagram** (MEP, **multi-event packet**):
  - Reduce the data **overhead**.
  - Reduce datagram rate:
    - optimize the **network efficiency**.
- IP packets are pushed over **1000 BaseT** Gigabit Ethernet **links**:
  - **Short** distances allow using 1000 **BaseT** throughout.
  - **Large** Ethernet/IP network (**3000 GbE ports**).

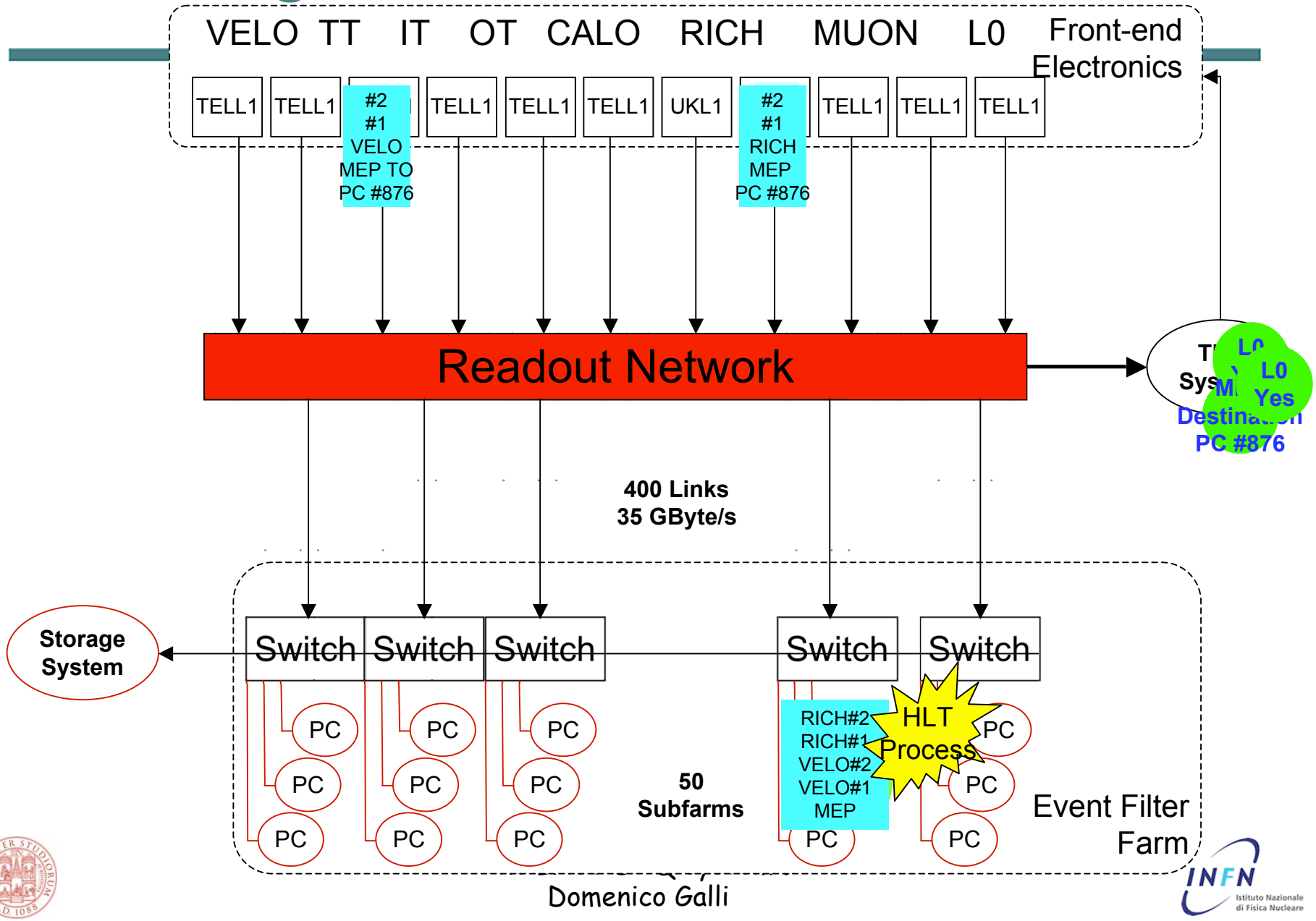
INFN Istituto Nazionale di Fisica Nucleare

LHCb DAQ: Key Design Features (II)

- Destination IP AND Ethernet addresses is synchronously and centrally assigned to all TELL1s:
  - Via a custom optical network (TFC, Timing and Fast Control):
    - The LHCb implementation of the standard CERN TTC (Timing and Trigger Control) system.
    - B-channel used to TTC broadcast the destination IP to all TELL1s (A-channel used to TTC broadcast the low-latency L0 accept/reject signal).
- For each trigger a PC-server must receive IP packets from all TELL1 boards:
  - In order to perform event-building.
- The TFC System is also used to implement dynamic load-balancing.
  - The farm nodes announce their availability to the TFC.
- LHCb DAQ uses a push protocol:
  - Global Throttle Mechanism implemented by TFC.

INFN Istituto Nazionale di Fisica Nucleare

Following the data-flow



DAQ Components

- **Timing and Fast Control**
  - **Custom** build (**TFC**).
    - Standard CERN TTC interface.
- **Readout boards**
  - **Custom** build (**TELL1**).
    - 4 x Gigabit Ethernet interfaces.
- **Readout network**
  - **COTS** (commercial of the shelf) components
    - **Core switch** (1260 port, 5Tb/s backplane). **High End COTS**.
    - **Edge switches** (1 x 44 server rack, 12 GbE uplink, 50 racks).
- **Farm PCs (~2200 x 16 cores = 18000 cores)**
  - **COTS** components (twin 2 x 4 core 1u PCs = 16 cores / 1u).
  - **Open Source OS** (SLC Linux)

INFN Istituto Nazionale di Fisica Nucleare

Rightsizing the HLT Farm

- The HLT trigger in **not** designed in order to have a **fixed latency**.
  - Being the last filter stage.
- We can therefore talk over average values.
- The **average time spent for the selection algorithm**, $\langle T_s \rangle$, must be **less** than the **average period which separates the input of two following events in the same trigger node**, $N_{\mathrm{cpu}}/\nu_{input}$, i.e.:
  $\langle T_s \rangle \leq N_{\mathrm{cpu}}/\nu_{input}$.
- So must be:
  $N_{\mathrm{cpu}} \geq \langle T_s \rangle \, \nu_{input}$.
- E.g.: $\langle T_s \rangle \sim 2$ ms and $\nu_{input} = 1$ MHz
  - $N_{\mathrm{cpu}} \geq 2000$.

INFN Istituto Nazionale di Fisica Nucleare

Farm Operating System

- Intel-compatible CPUs used on HLT farm can run a number of OSs:
  - **Linux**, **Solaris**, **Lynx-OS**, **FreeBSD**, **Mac-OS**, **MS Windows**, etc.
  - **Linux** seems to have won out.
  - Exception:
    - BaBar uses **Solaris**.
    - DØ (Fermilab) was using **MS Windows** on its L3 Farm (**but** switched to Linux in Run II).
    - CDF uses **VxWorks** for a specific task (data transfer from VME readout board to ATM network).
- **LHCb** has chosen **SLC Linux** as HLT OS.

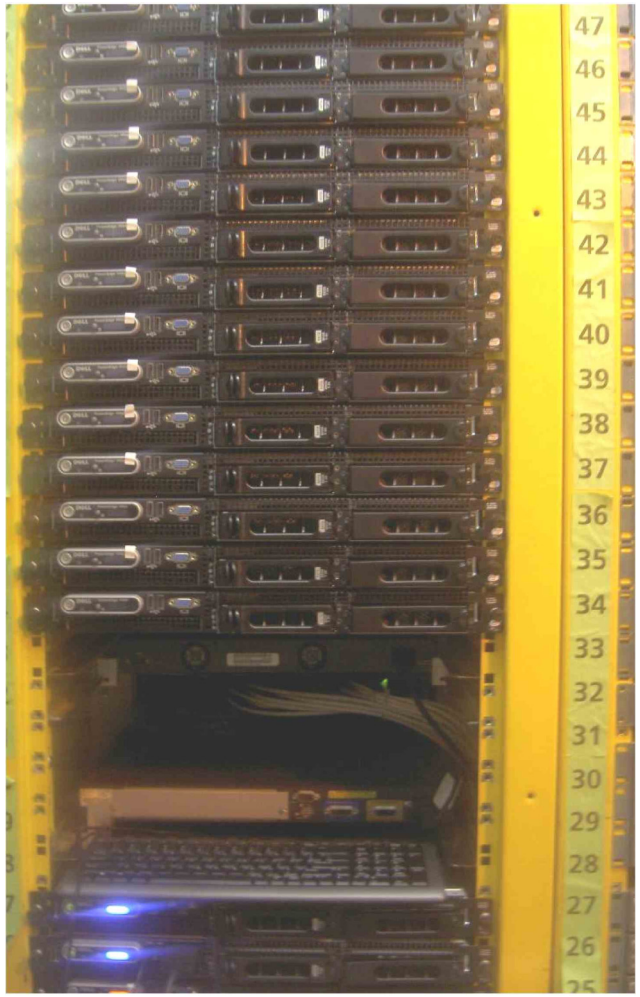
INFN Istituto Nazionale di Fisica Nucleare

Farm Worker Nodes

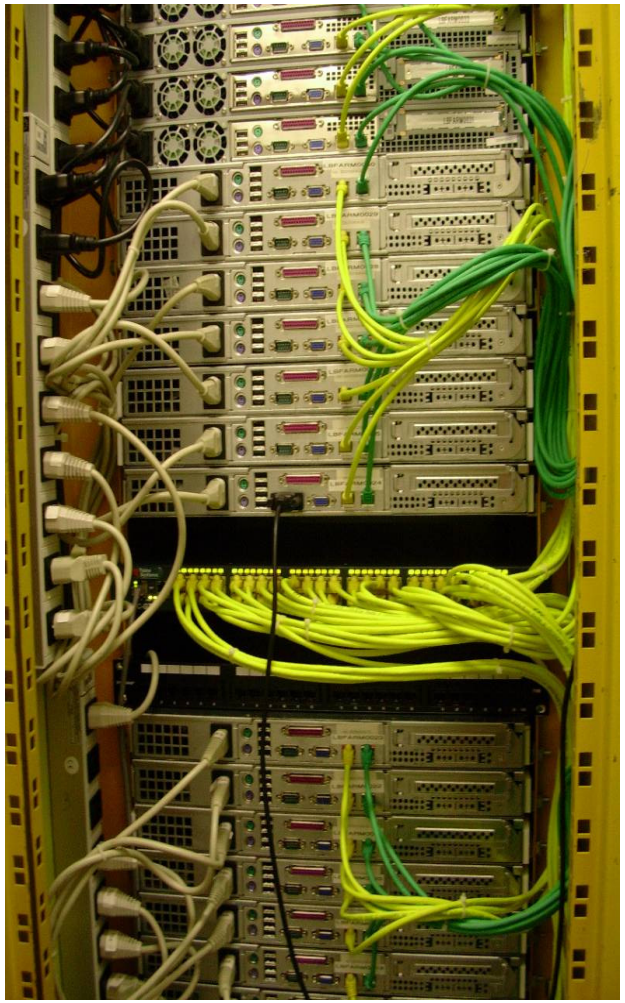
- **Each worker node** (PC server) in the EFF runs:
  - 1 **event-builder** process:
    - Assemblies the MEPs and distribute the events among the trigger processes.
  - 8 **HLT processes** typically (as many as the number of CPU cores).
  - 1 **data writer**:
    - Sends accepted events to the **streaming** and **formatting** layers.
- Data are **not moved** inside a PC:
  - Data are kept in a **shared memory** area.
  - **Descriptors** to the events are **passed** between processes.
    - By means of a shared buffer library
- **Selection algorithm** reduce the event rate:
  - **1 MHz ➔ 2-5 kHz**.

INFN Istituto Nazionale di Fisica Nucleare

Farm CPU Racks



- 2200 1U rack-mounted boxes.
- Operated disk-less.
- 2 x 1000Base-T interfaces, to keep separate:
  - Data;
  - ECS (Experimental control system).



High Speed Data Link Technology

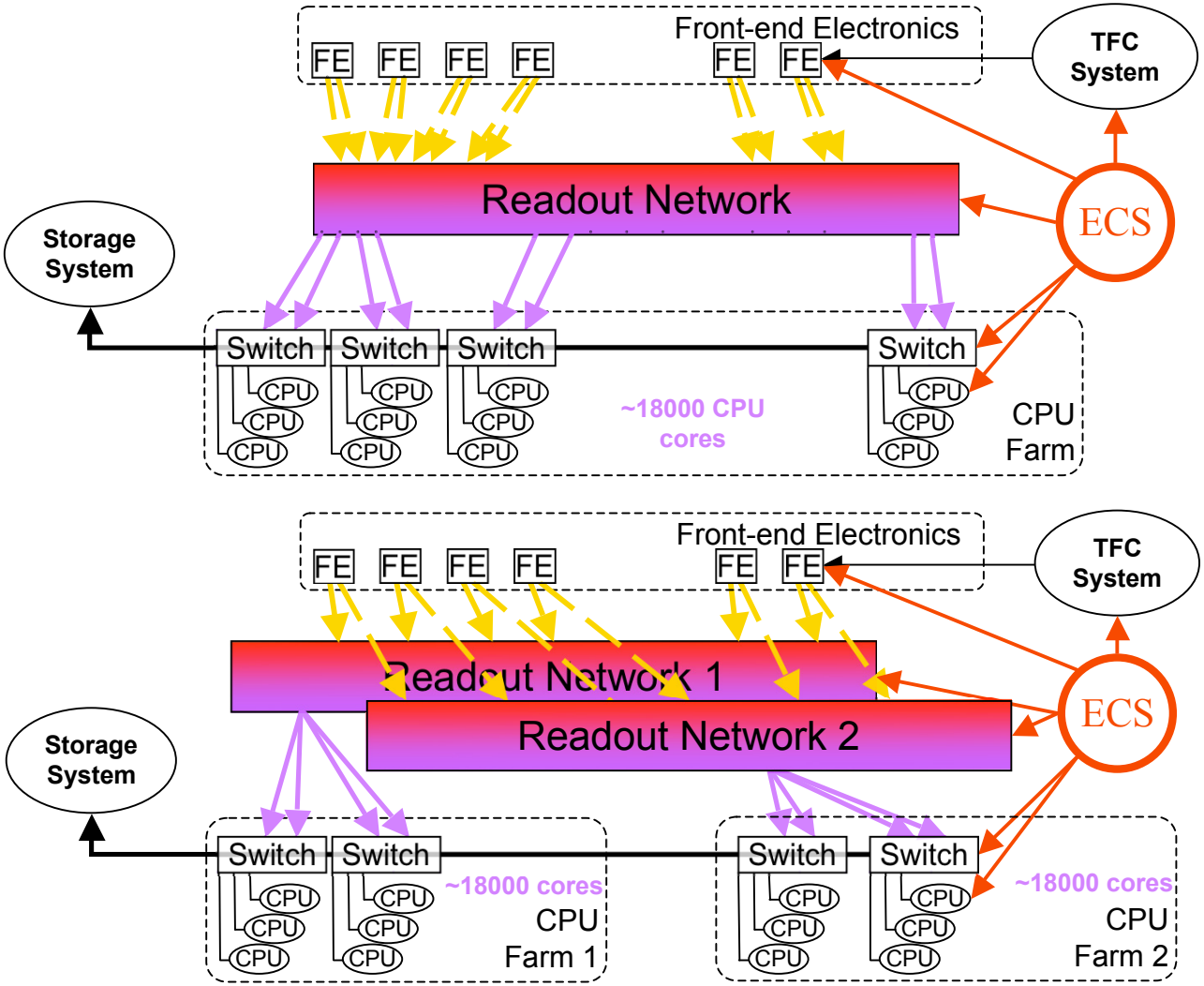
- **Trend toward COTS technologies**:
  - HERA-B:
    - **Shark link** (proprietary, by Analog Devices) until level 2, than **Fast Ethernet**.
  - BaBar:
    - **Fast Ethernet**.
  - DØ:
    - **Fast Ethernet** / **Gigabit Ethernet**.
  - CDF:
    - **ATM** / **SCRAMnet** (proprietary, by Systran, low latency replicated non-coherent shared memory network).
  - CMS:
    - **Myrinet** (proprietary, Myricom) / **Gigabit Ethernet**.
  - Atlas / LHCb / Alice:
    - **Gigabit Ethernet**.
  - Possible new experiments:
    - **10-Gigabit Ethernet** (soon also on copper), **10-40-Gigabit infiniBand**, **100-Gigabit Ethernet**.

LHCb HLT Core Switch (also in CMS)

Force10 E1200 equipment

- Port densities:
  - 14 slots for line-cards
  - Biggest port density is 90 1000Base-T ports per line-card (90/48 over-committed)
  - 14 x 90 = **1260 1000Base-T ports**.
- Switching Fabric
  - Switching capacity is
    - Raw: ~1.6 Tb/s,
    - Usable: ~1.2 Tb/s (140 GiB/s),
    - Backplane capacity: **~5 Tb/s**.

Scaling (Doubling) the System

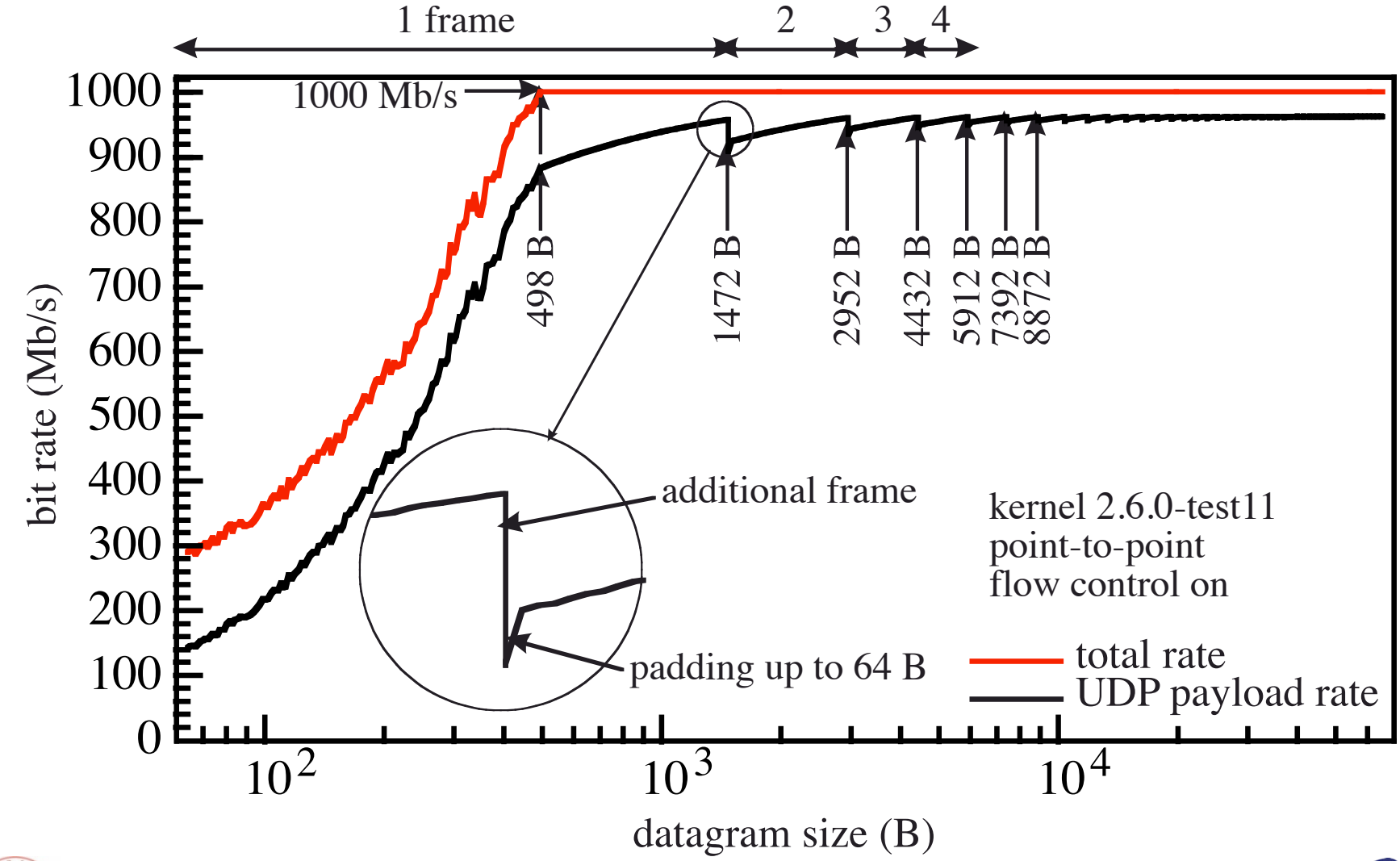


Which Protocol to Move Data through the Read-Out Network?

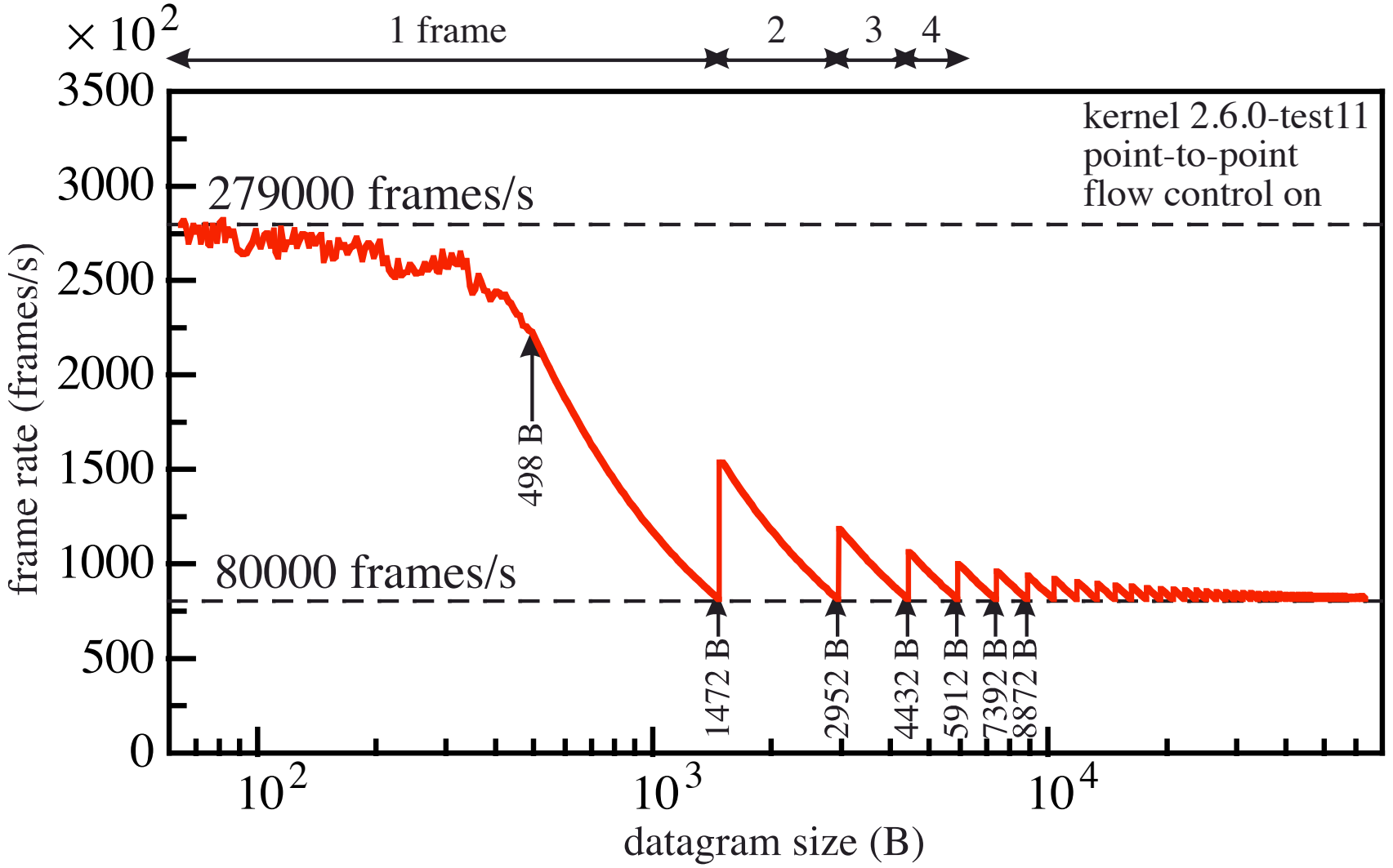
- Why not **TCP**?
  - To **avoid** mechanisms which slow down data transmission (**slow start**, **congestion avoidance**).
  - **Reliability** mechanisms (fast retransmission, fast recovery) are **useless** due to latency constraints:
    - If a fragment of an event is dropped by the network we prefer to **get the next** event **rather than retransmit** the same event.
  - Keep the **implementation** in the FPGA based readout board **simple**.
- Why not **UDP**?
  - In our application we have **no use** for the UDP **port** numbers,
  - UDP **checksum redundant** with the **Ethernet CRC** (Cyclic Redundancy Check) information in a switched network.
    - Also, the UDP checksum is performed by the CPU (at least for fragmented datagrams), as opposed to the **Ethernet CRC done by the MAC** and so uses up additional resources.
- Why **IP**?
  - Datagram **fragmentation** is well defined by the standard.

INFN Istituto Nazionale di Fisica Nucleare

Gigabit Ethernet IP Transfer Rate (using PCI-X)



Gigabit Ethernet Frame Transfer Rate (using PCI-X)



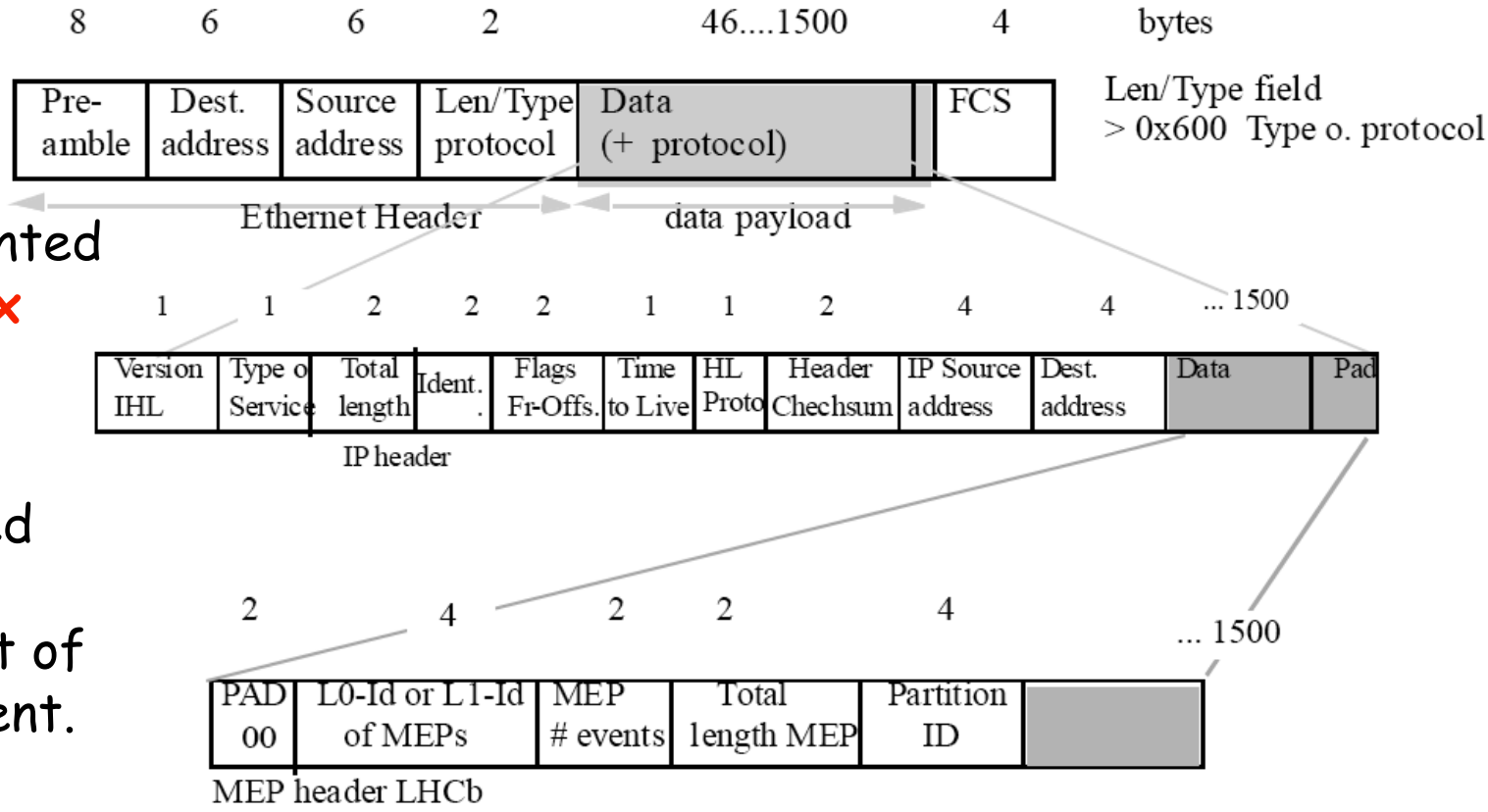
Why a New Transport Protocol?

- The optimal Ethernet payload/overhead ratio, is achieved when the IP datagram fills completely the 1500 B Ethernet payload.
- Moreover the Gigabit Ethernet throughput drops for small frame size.
- However, each Tell-1 board can send only data-fragments pertaining to the associated sub-detector element, which usually is much smaller (~100 B).
- In order to optimize the payload/overhead ratio, fragments from multiple (~15) events have to be aggregated (MEP, Multi Event Packet) into a single IP datagram.
- MEP is a LHCb custom OSI-level 4 (transport) protocol.
  - OSI-level 3 (network) is IP;
  - OSI-level 2 (datalink) is Ethernet.

INFN Istituto Nazionale di Fisica Nucleare

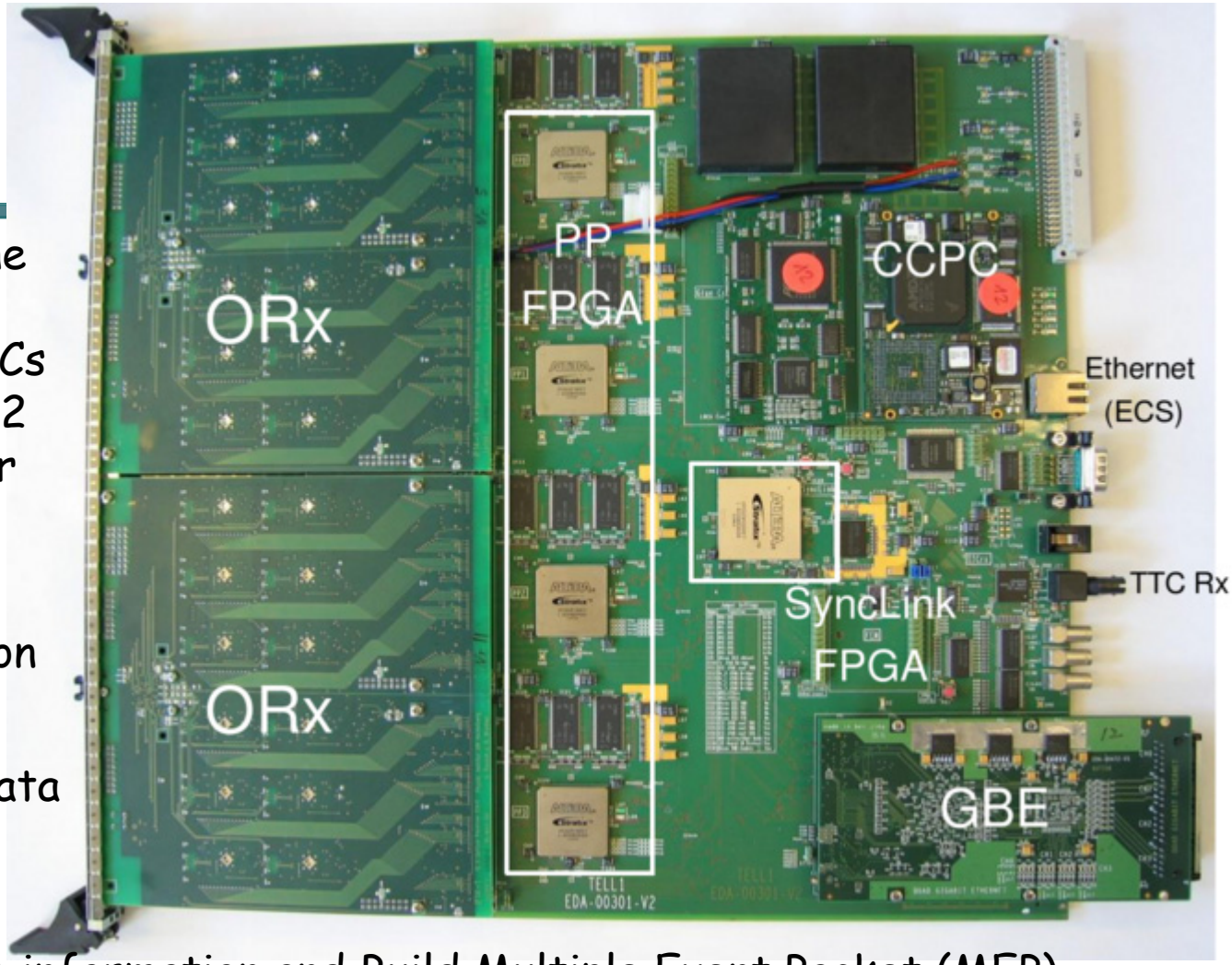
The LHCb MEP Protocol over IPv4

- Custom protocol
- Implemented as a **Linux Kernel module**.
- Optimized for the transport of Multi Event.



To be created in FPGA
And transmitted to GBE

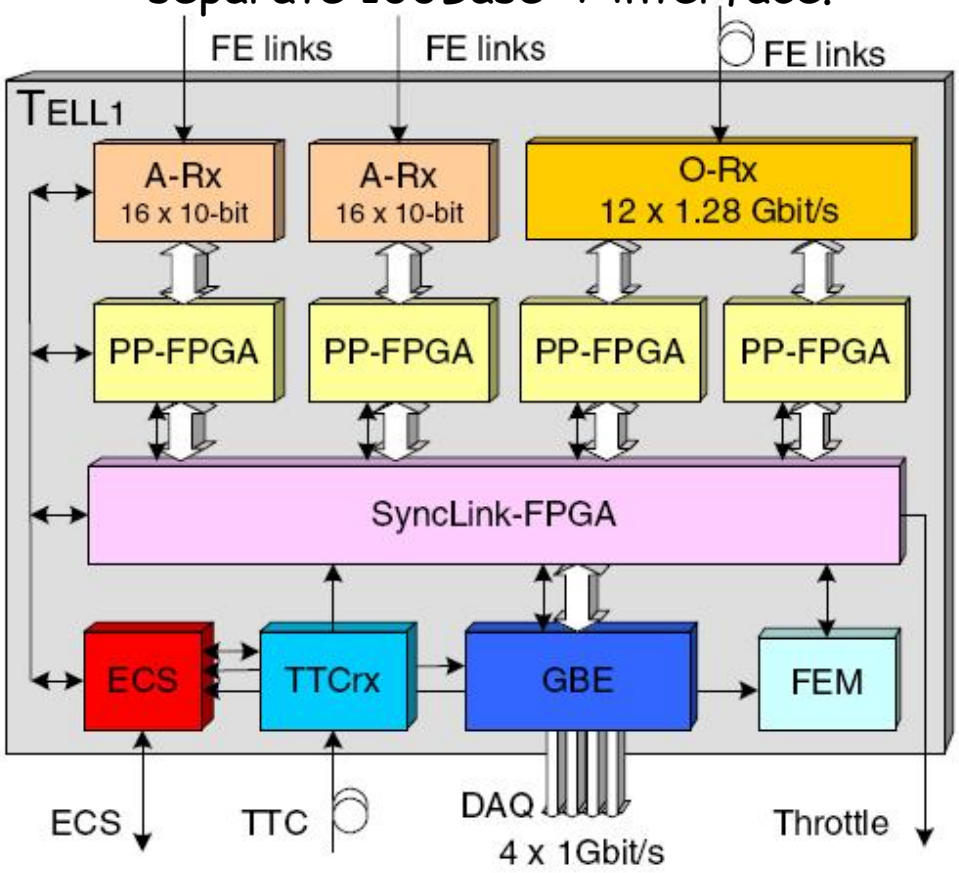
TELL1 Boards

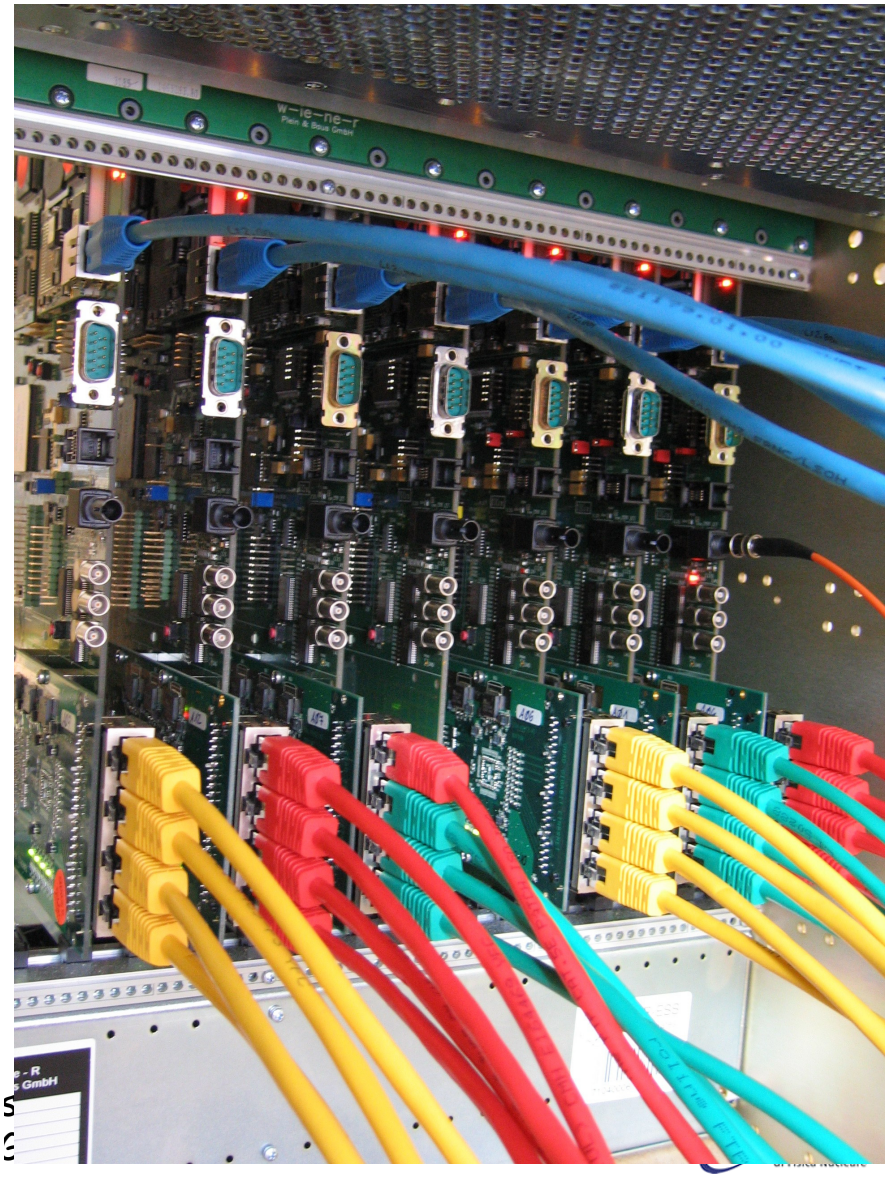


- Optical mezzanine (ORx): Optical fibers from CROCs [only up to 8 on 12 links are used per ORx]
- Credit Card PC (CCPC): Connection with ECS
- PP-FPGA: User data compression
- Synclink FPGA: Gathers PP-FPGA information and Build Multiple Event Packet (MEP)
- Giga Bit Ethernet (GBE): Transmits MEP to Event Builder
- TTC Receiver: Clock, L0 Trigger, Reset (L0evID – BX-ID) and DAQ IP destination address

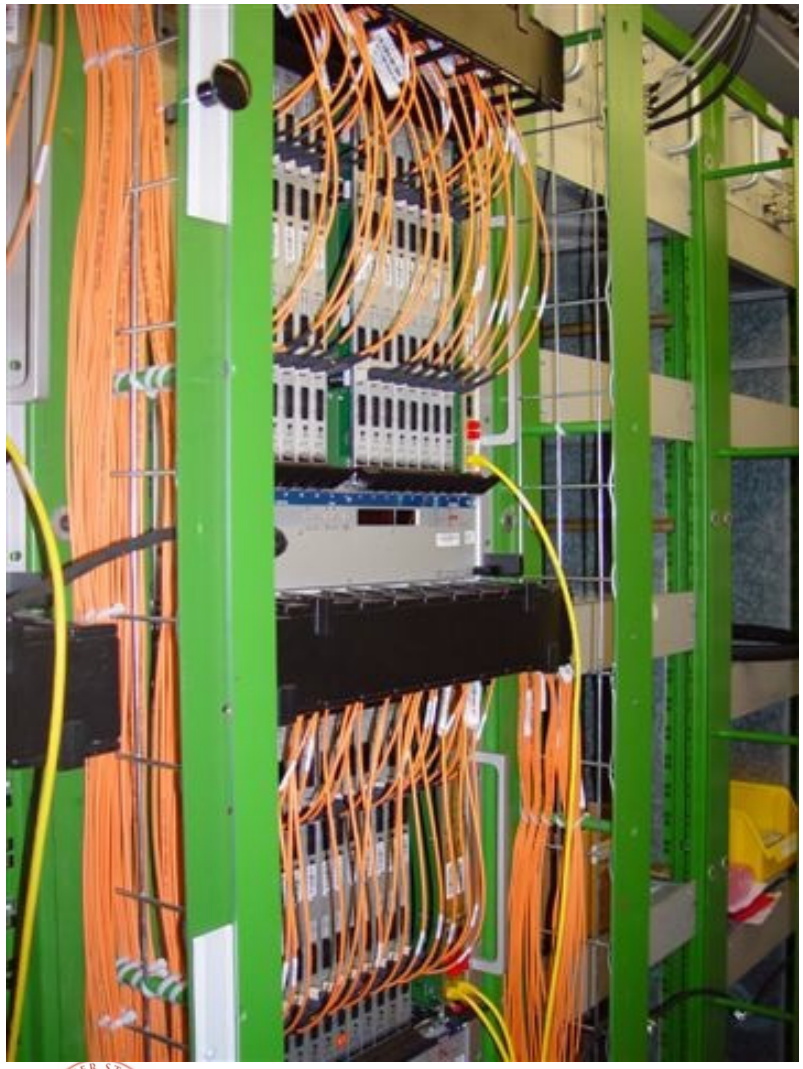
TELL1 Boards (II)

- Input: 24 x 1.6 Gb/s optical link or 64 x analog copper links.
- Output: 4 x 1000Base-T.
- ECS: Credit card PC (Linux) with separate 100Base-T interface.

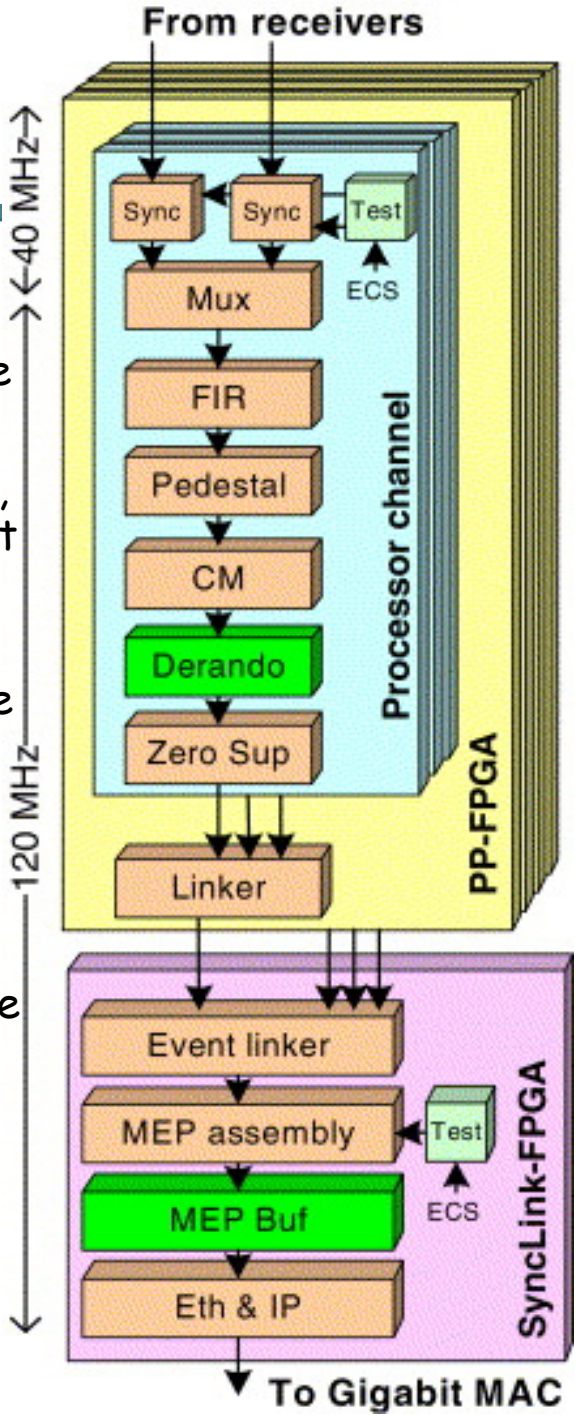




TELL1 Boards (III)

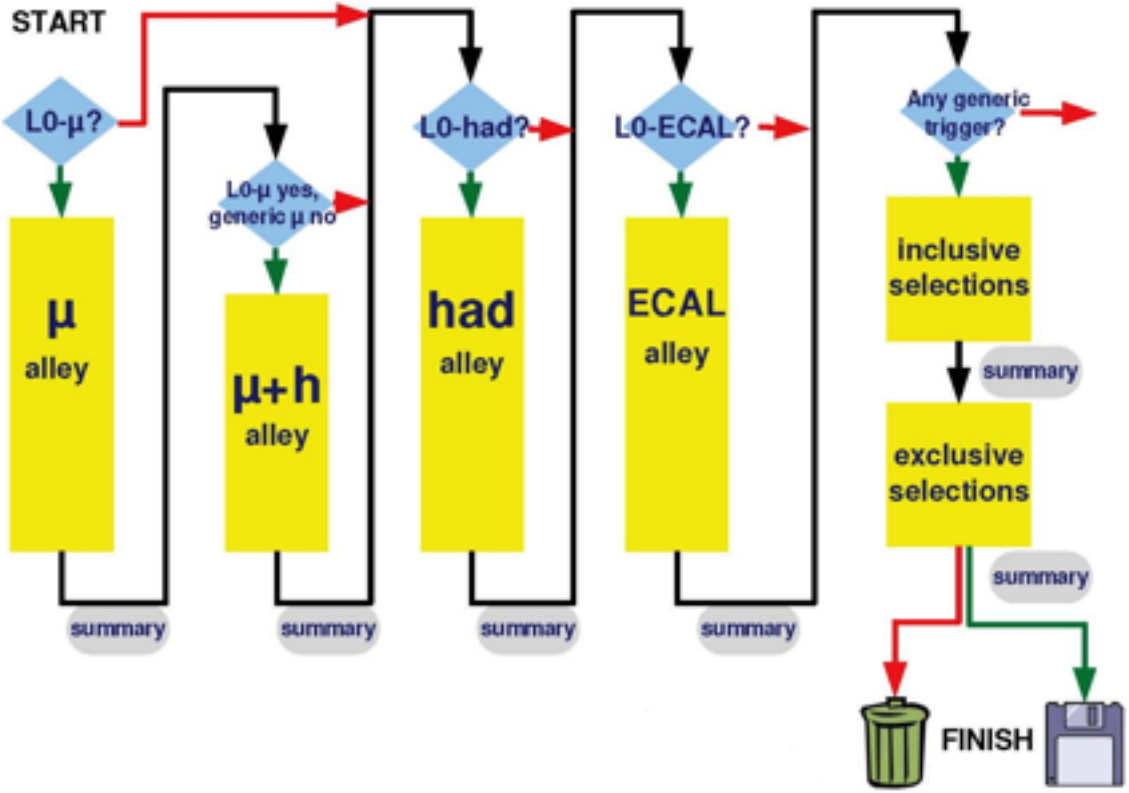


- FIR: Finite Impulse Response filter.
- CM: Common Mode noise corrections.
- After **zero suppression**, the length of each event is variable.
- **Derandomizing** buffers are employed to average the data rate and the data processing time.
- To prevent any overflows, each buffer can generate a **throttle** signal that is sent to the readout supervisor.
- The **readout supervisor** suspends the trigger signal until the buffers have recovered.



LHCb HLT Strategy

- "Trigger Alleys"
  - to exploit and refine, L0 trigger after a **partial reconstruction**,
  - **immediately reject** L0 misinterpretation.
- Inclusive selections (D_s, D^*, ϕ and μ).
- Exclusive selections: ~20 channels.



Beyond LHCb: 10 Gb/s Technologies

- **Ethernet**:
  - **10 Gb/s** well established
    - Various **optical** standards, short range **copper** (**CX4**), long range **copper** over **UTP CAT6A** standardised), widely used as aggregation technology.
  - Begins to conquer MAN and WAN market (succeeding SONET).
  - Large market share, vendor independent IEEE **standard** (802.3x).
  - Very active R&D on **100 Gigabit/s** and 40 Gigabit/s (will probably die).
- **Myrinet**:
  - Popular cluster-interconnect technology, **low latency**.
  - **10 Gb/s standard** (**optical** and **copper** (CX4) exist)
  - **Single vendor** (Myricom).
- **InfiniBand**:
  - Cluster interconnect technology, **low latency**.
  - **10 Gb/s** and **20 Gb/s** standards (**optical** and **copper**).
  - **Open** industry **standard**, **several vendors** (OEMs) but **very few chipmakers** (Mellanox).
  - Powerful protocol/software stack (reliable/unreliable datagrams, QoS, out-of-band messages etc…).

INFN Istituto Nazionale di Fisica Nucleare

10 Gb/s Technologies: Ethernet

- Ethernet still allows **simple FIFO-like interface** (like in TELL1 cards):
  - However due to the (ridiculously) **small frame size** use of a **higher level protocol** (e.g. IP) is **mandatory**.
- **Prices** per router port are **dropping** quickly:
  - But still **more expensive** than InfiniBand.
  - **Copper** standard exist, requires cabling **Cat 6A**.
  - High power consumption **optical** still **very expensive**.
- **Various** NIC cards exist.
- Not yet on the **mainboard**, but only a question of time – at least for high end servers.

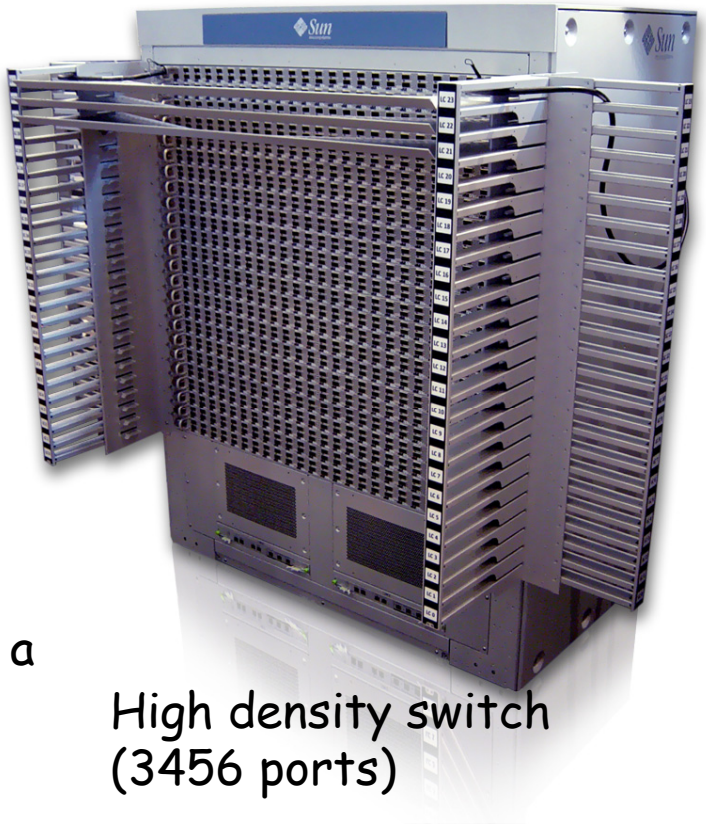
INFN Istituto Nazionale di Fisica Nucleare

10 Gb/s Technologies: InfiniBand

- On the **source side** (TELL10):
  - Nallatech plug-in card.
  - On-board Xilinx Virtex-II Pro FPGA.
  - Up to 20k logic cells of programmable logic per module.
  - Up to 88 Block RAMs and 88 embedded multipliers per module.
  - 2x InfiniBand ™ I/O links.
  - 2x RocketIO serial links.

10 Gb/s Technologies: InfiniBand (II)

- InfiniBand **switches**:
  - **10 Gb/s** standard.
  - **20 Gb/s** (DDR, double data rate), **30 Gb/s** (TDR) and **40 Gb/s** (QDR) also available.
- Key features:
  - Up to **3456** InfiniBand 4x **ports** in a single chassis.
  - Up to **110 Tb/s** of switching capacity in a single system.
  - 1/2/4 **Fibre Channel** and 10 Gb/s **Ethernet interface** options.
  - Fully **redundant** power, cooling and logic components.



High density switch (3456 ports)

Edge switch (24 ports)

10 Gb/s Technologies: InfiniBand (III)

- InfiniBand for the **Servers**.
  - Quite a few **InfiniBand Adapter Cards** (HCA) exist.
  - Few main-boards exist with **onboard** InfiniBand adapter:
    - Availability of onboard Gigabit Ethernet NICs makes (copper UTP) Gigabit Ethernet essentially zero-cost on the servers.
    - Physical signaling is compatible (Myricom makes dual personality cards!), there are rumours that Intel will bring out a chipset with both options
  - InfiniBand on the server might have **performance** advantages.
  - **Dual-personality switches** exist: they act as an InfiniBand to Ethernet / IP **bridge**.

INFN Istituto Nazionale di Fisica Nucleare

10 Gb/s Technologies: InfiniBand (IV)

- **Potential advantages** of InfiniBand:
  - **Low latency** & **reliable** datagrams:
    - Implement **pull** protocol ➔ could result in much more efficient usage of network bandwidth (currently we can use only ~ 20% of the theoretically available bandwidth).
    - Implement **load balancing** and **destination assignment** ➔ no need for custom ("TTC"-like) network for this purpose.
  - **Cost per switch-port** much lower than in Ethernet (requires much **less buffer** per port / very high-speed buffer memory is very expensive).
  - Even- building using **remote DMA** could result in much lower CPU "wasted" for data movement.
    - Currently we cannot handle more than ~ 300 MiB/s per server (for illustration this means that for a 10 MHz readout at current LHCb event size 35 KiB we would need 1000 servers just for the data formatting, checking and moving.

INFN Istituto Nazionale di Fisica Nucleare

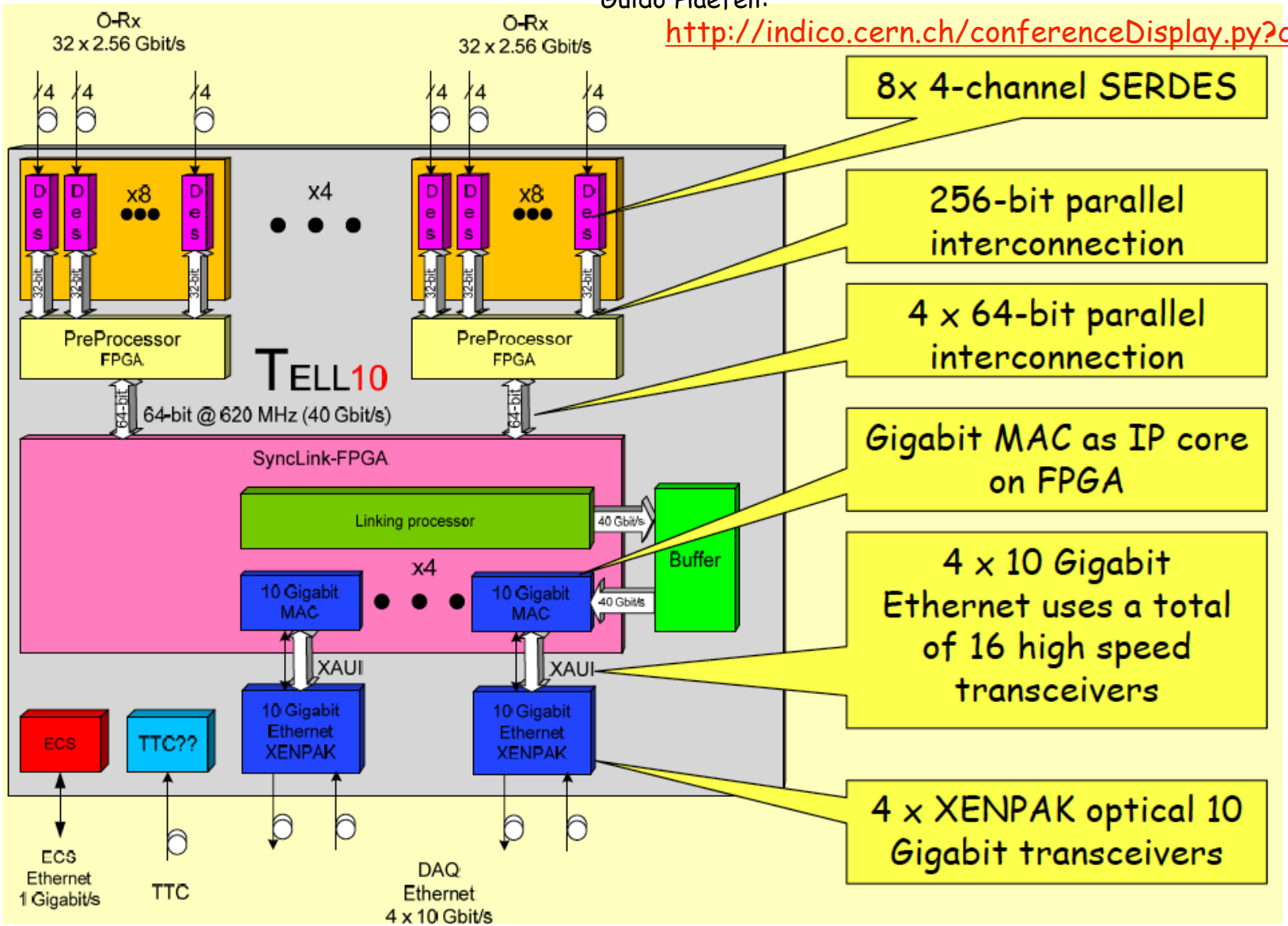
Beyond LHCb.
TELL1 ➔ TELL5, TELL10 and TELL40

- EPFL, Lausanne is designing the boards TELL5, TELL10 and TELL40:
  - In bandwidth 5x, 10x and 40x with respect to the current TELL1.
- Input/output data bandwidth:
  - TELL1: 30 Gb/s ➔ 4 Gb/s.
  - TELL10: 328 Gb/s ➔ 40 Gb/s.
  - TELL40: 1311 Gb/s ➔ 140 Gb/s.
- Same tasks as current TELL1:
  - Synchronization to TTC.
  - Data pre-processing to achieve data compression by a factor 10!
  - Data buffering for zero suppression and DAQ interface.
- See presentation of Guido Haefeli at
  - http://indico.cern.ch/conferenceDisplay.py?confId=8351

INFN Istituto Nazionale di Fisica Nucleare

TELL10 Outline with Current Technology

Guido Haefeli:
http://indico.cern.ch/conferenceDisplay.py?confId=8351



INFN Istituto Nazionale di Fisica Nucleare

Summary

- LHCb experiments has a **2 stages** trigger system.
- The **input rate of the software trigger** is the **highest** among the LHC experiments (**1.1 MHz**).
- The readout network uses **(1) Gigabit Ethernet** Technology for a farm of **2200 1u boxes** (**18000 CPU cores**).
- DAQ uses a **push** protocol, with **global throttle** mechanism.
- **Readout boards** (TELL1) send data to **4 x Gigabit Ethernet interfaces**.
- Data fragments from readout boards packed in ~15 event groups (**MEP**s).
- **Node destination** of a MEP **broadcasted** by the **TFC** to all the readout boards.
- **TFC** implements **dynamic load balancing** and **throttle**.
- The system is **scalable** and can be ported to a **higher speed COTS link technology**.

INFN Istituto Nazionale di Fisica Nucleare