

Trigger and Data Acquisition Systems In High Energy Physics: The Challenge of the LHCb Upgrade

Domenico Galli

Università di Bologna e INFN, Sezione di Bologna

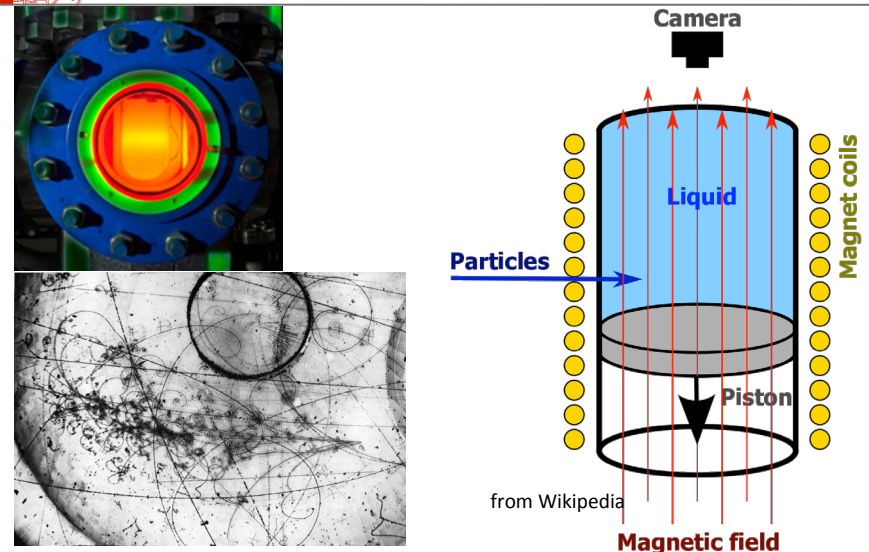
INFN Bologna, "Aperitivi Scientifici", November 27, 2015

- **Trigger:**
 - Multi-level trigger, Trigger trend;
- **Readout and Signal Processing:**
 - Pulse shaping, Range compression, Digitisation, Zero suppression;
- **Data Acquisition and Event Building:**
 - Basic DAQ, Collider DAQ, Event Building;
- **Modules and Data Bus** for DAQ systems:
 - NIM, CAMAC, FastBus, VME, PCI, PCIE, ATCA/ μ TCA;
- **Network based DAQ:**
 - Ethernet, InfinBand;
- **The Challenge of the DAQ for the LHCb Upgrade:**
 - 30 MHz Rate;
 - 32 Tb/s Aggregate Throughput.

Top and Bottom Halves of Data Processing in HEP

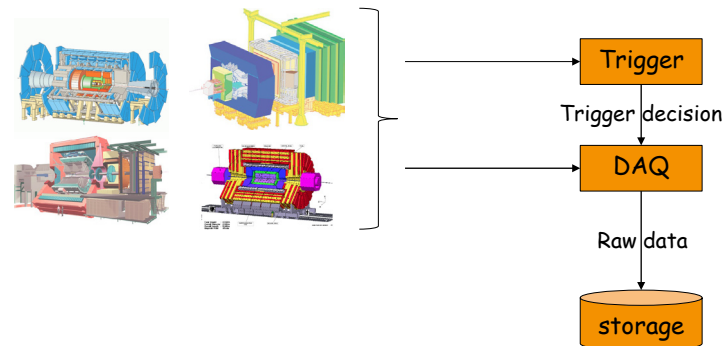
- **Top Half:**
 - Processing of particle interaction events is performed on **data stored on disks/tapes** of Tier-0/Tier-X computer centre:
 - Event **tracks** are **reconstructed**;
 - Event **kinematics** is **reconstructed**;
 - Useless event data are **stripped**;
 - **Full event** is **reconstructed** and **tagged**;
 - **Mass analysis/Group analysis** is performed;
 - **Local (user, n-tuple) analysis** is performed.
- **Bottom Half:**
 - Who **brought** the data to the **Tier-0**?
 - How are they **driven** from the detector to the storage?
 - How data are **selected**?
 - How data coming from different sub-detector are **merged together**?

Once Upon a Time...



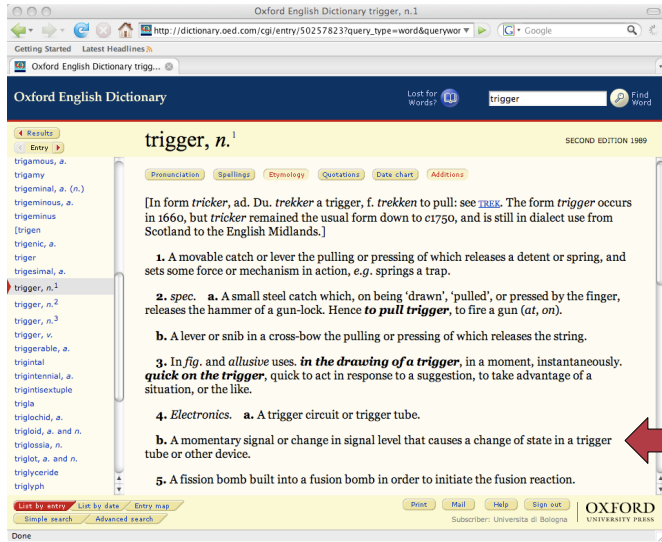



- How are **LHC experiment read** instead?
- Overall the main role of **Trigger + DAQ** is:
 - To **process** the signals generated in a detector;
 - **Saving the interesting information** on a permanent storage.





- **Trigger:**
 - Either **selects** interesting events (**signal**) or **rejects** useless ones (**background**), in **real time**:
 - I.e. with minimal controlled **latency**:
 - Time it takes to **form** and **distribute** its **decision**.
- **DAQ (Data Acquisition):**
 - Gathers data produced by detectors: **Readout**;
 - Feeds several **trigger** levels;
 - Forms complete events from fragments: **Event Building**;
 - Stores event data: **Data Logging**;
- **ECS (Experiment Control System):**
 - Provides **Run Control**, **Configuration** and **Monitoring** facilities.

Trigger

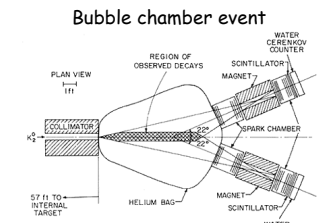
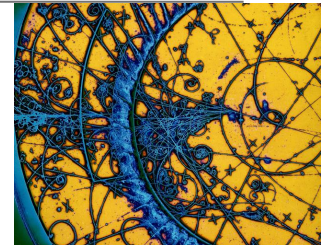


- A HEP experiment can collect **hundreds of EiB** (2^{60} B) of data **in a year**:
 - Only a small subset ($\sim 1/10^6$ events) of primary physics interest.
- **Tape I/O** lags behind many other computer components:
 - This problem can be overcome writing in parallel to many tapes;
- Storage media could run over **tens of G€/year**. 
- How much **CPU power** for post-triggering reconstruction of all the events?
- **Cannot save all raw data all the time.**
- Eliminate useless background **as early as possible**:
 - In order to save resources to process interesting events.

- **Experiment raw data**:
 - LHC bunch crossing rate: **40 MHz**.
 - Event rate: **10 MHz** (events with at least one interaction).
 - Event size: **35 KiB/event**.
 - **330 GiB/s = 3.1 EiB/y** = 3.1×2^{60} B/y.
 - **3.3 G€/year** of tapes. 
- **Events of interest**:
 - **100 kHz** beauty pairs.
 - Branching fraction of the events of interest: **[10^{-6} , 10^{-5}]**.
 - **~ 10 Hz** of events of interest.
- **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 = 650 TiB/y** = 650×2^{40} B/y.
 - **65 k€/year** of tapes.

30% duty cycle.
100 €/TiB tape media.

- **Bubble chambers**:
 - DAQ: stereo photograph.
 - Low level trigger: piston expansion.
 - High level trigger: humans (scanners).
- **Early fixed target** experiments:
 - **Merely hardware** implementation.
 - Very **simple** calculation.
 - **Raw** discrimination.
 - Large **dead-time** possible during readout.
 - DAQ came after the trigger.
- **Nowadays** HEP experiments:
 - **Multi-level**.
 - **Pipelines** to pull down dead-time ($< 5\%$).
 - Hardware **look-up tables** for fast calculation.
 - **Software** implementation of higher level.



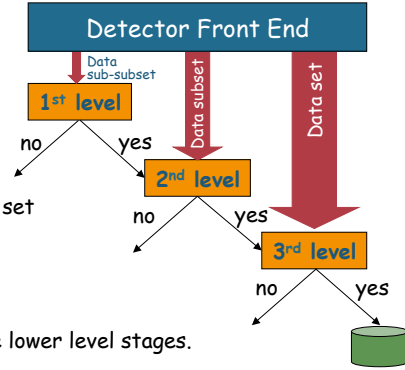
Cronin-Fitch experiment, 1964

- Design **constraints**:
 - Detector input event rate.
 - Tape output event rate.
 - Available **CPU power** and **network bandwidth**.

- **Finer selection** requires:
 - **more data**;
 - **more computing time**.

Multi-Level Trigger Systems:

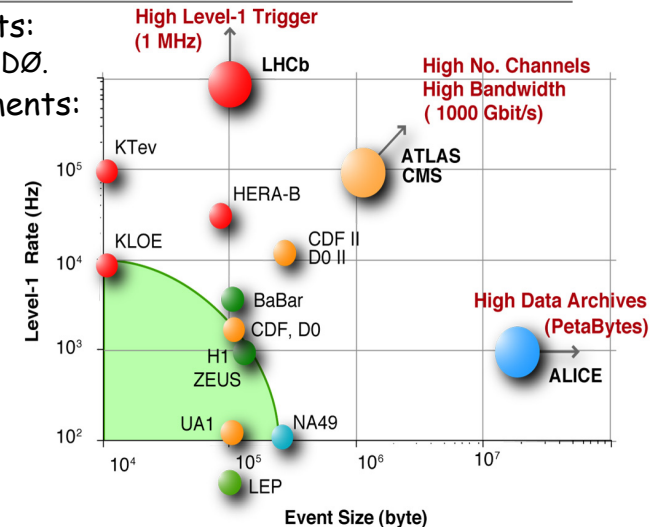
- **First level**:
 - Most coarse.
 - Uses a **small subset** of the whole data set of **all the events**.
- **Last level**:
 - Finest.
 - Uses the **whole data set** of the **only events** passed through the lower level stages.
- **More trigger levels, less data flow**.



- **1st level**:
 - Cut out **simple, high rate** background.
 - Feature extraction from **single sub-detector**.
 - **Hardware** implementation (custom electronics, **ASIC, FPGA**).
 - High speed (**LUT**) and dead-timeless (**pipelines**).
- **2nd level**:
 - **Matches** of data from several sub-detectors.
 - **ROI** (Region Of Interest) trigger: HERA-B, ATLAS.
 - Uses previous trigger info to decide what regions of the detector it is interested in.
 - **Hybrid hardware/software**.
 - Sometimes **DSP** (INMOS Transputer in ZEUSS, Analog Devices Sharc in HERA-B).
- **3rd level**:
 - **Full event reconstruction**.
 - **Software** (large CPU farm).

- **1st level**:
 - **Disappearing** in lepton accelerators (ILC) and in LHCb Experiment.
 - Broad usage of programmable units (**FPGA**).
 - Can implement **algorithms** that once could be implemented only in software.
 - Line between hardware and software is **blurring**.
 - But data access limited to **local** (sub-detector) data.
 - Sometimes use of **Neural Network**:
 - Chips: CNAPS (H1), ETANN (CDF).
 - Higher level trigger decision are **migrating** to lower levels.
- **2nd level**:
 - **Disappearing** (but **software ROI** in Atlas).
- **3rd level (High Level Trigger, HLT)**:
 - Working at **higher and higher rates** (large CPU farm).
 - Pushed **further and further up** into the DAQ chain.

- **Past experiments**:
 - UA1, LEP, CDF, DØ.
- **Recent Experiments**:
 - KLOE, NA49, H1, Zeus, CDF II, DØ II.
- **Present experiments**:
 - Alice, Atlas, CMS, LHCb



- Software trigger advantages:
 - **Flexibility:**
 - Selection rules for the events can be changed simply by modifying a software code.
 - **Scalability:**
 - Processed **event rate** can be increased simply by increasing the farm size (number of PCs) and the port number of the network switch .
 - **Cost:**
 - **Commodity components** used in software triggers are very cheap.
 - Allows to profit from the rapid **price drop** in commodity components (PC, Ethernet cables and switches, etc.).
 - **Maintainability:**
 - **Widespread** commodity **interfaces** will continue to be available on the market (with increased performance).
 - **Upgradeability:**
 - Can profit from the **rapid development** undergone by commodity components.
- Drawbacks:
 - **Variable latency.** Difficult to be used for other but the **last trigger level**;
 - **DAQ/EB throughput** must be suitable for the trigger.

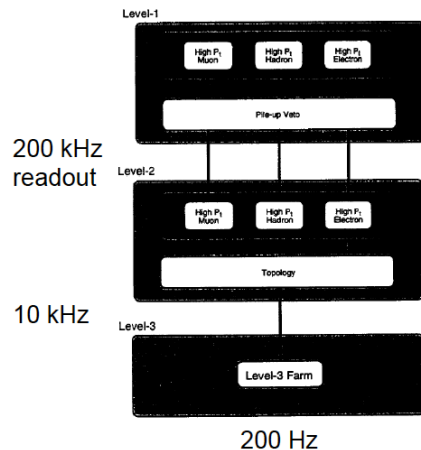
- The software trigger is not designed in order to have a fixed latency.
- We can therefore talk over average values.
- The **average time spent for the selection algorithm**, $\langle T_s \rangle$, must be **less than the average period which separates the input of two following events in the same trigger node**, N_{cpu} / ν_{input} , i.e.:

$$\langle T_s \rangle \leq N_{cpu} / \nu_{input}$$
- So must be:

$$N_{cpu} \geq \langle T_s \rangle \cdot \nu_{input}$$
- In the LHCb case, $\langle T_s \rangle \approx 2 \text{ ms}$ and $\nu_{input} = 1 \text{ MHz}$ so must be $N_{cpu} \geq 2000$.

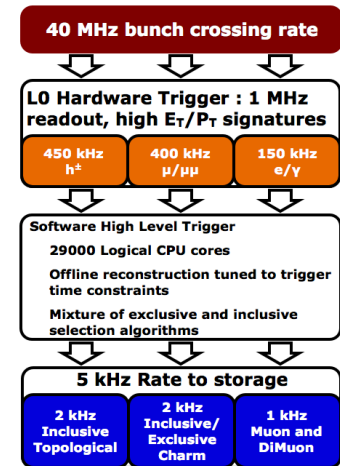
LHCb TDR (2003)

Input lumi = $1.5 \times 10^{32} \text{ cm}^{-2}\text{s}^{-1}$

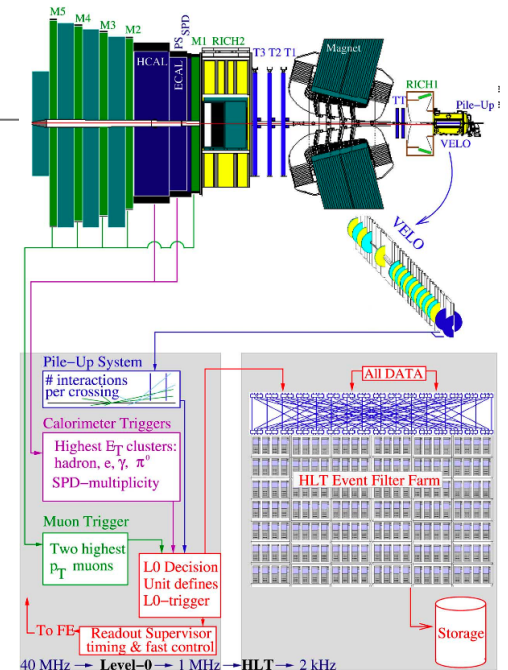


LHCb Run (2010-2012)

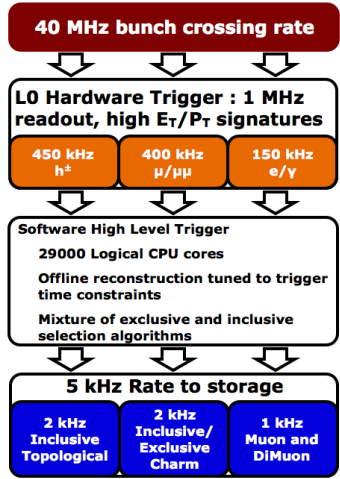
Input lumi = $1.5 \times 10^{32} \text{ cm}^{-2}\text{s}^{-1}$



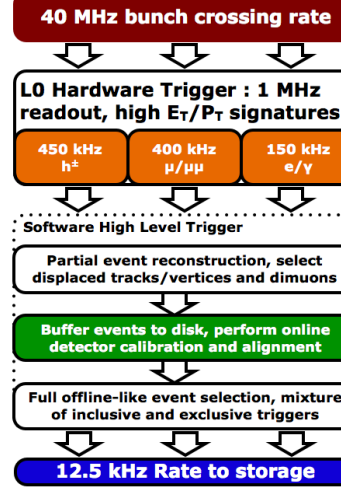
- **2 stages:**
 - **Level-0:** synchronous, hardware + FPGA; 40 MHz \rightarrow 1 MHz.
 - **HLT:** software, PC farm: 1 MHz \rightarrow 2 kHz.
- **Front-End Electronics:**
 - Interfaced to Read-out Network.
- **Read-Out Network:**
 - **Gigabit Ethernet LAN.**
 - Read-out @ **1.1 MHz.**
 - Aggregate throughput: **60 GiB/s.**



LHCb Run (2010-2012)
Input lumi = $1.5 \times 10^{32} \text{ cm}^{-2}\text{s}^{-1}$

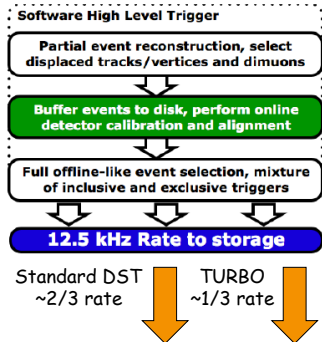


LHCb 2015 Trigger Diagram

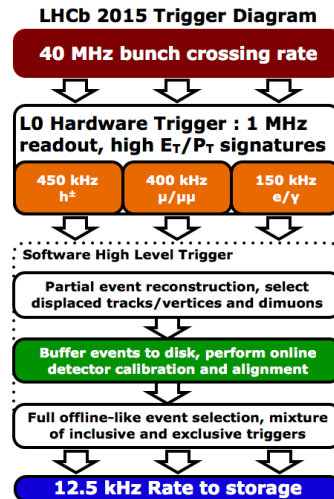


- The events are written to **local disks** (~4 PiB available) while **calibration** and **alignment** is performed.
 - Only when this is **satisfactory** the **second stage of HLT** executed.
- This step important, as **better alignment** provides **better signal discrimination**.
- Also means we can **trust** information we would not use otherwise in HLT (i.e. from **RICH**).

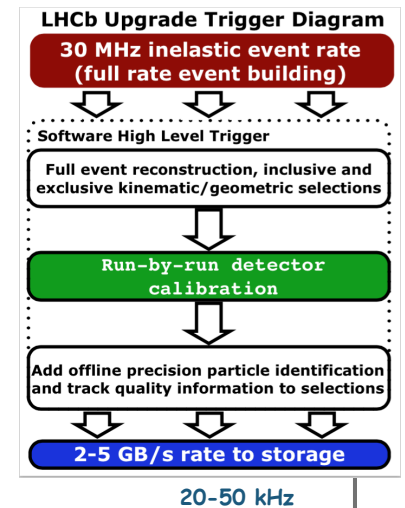
- TURBO Stream:**
 - Corollary of new procedure is that **recorded data** are already **suitable for physics analysis**.
 - Cautiously** start to exploit this feature for **high yield studies** - establish the 'TURBO' stream:
 - Store candidates as found by the HLT;
 - Discard** most of **raw data**;
 - Hence **reduce storage** by ~90%;
 - No need** for **offline processing**;
 - Data immediately **ready for analysis**.
 - TURBO used for first run-2 physics results.



Input lumi = $4 \times 10^{32} \text{ cm}^{-2}\text{s}^{-1}$

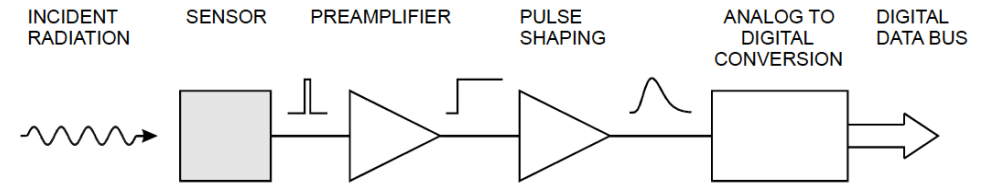


Input lumi = $20 \times 10^{32} \text{ cm}^{-2}\text{s}^{-1}$



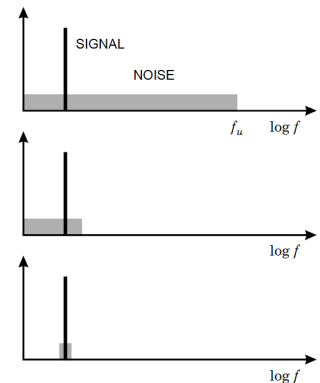
Readout

- Typically, the detector signal undergoes the following **analog processing** before being digitized:
 - **Amplification;**
 - **Filtering;**
 - **Pulse shaping;**
 - **Baseline restoration;**
 - **Range compression;**
 - **Pedestal subtraction;**
 - Etc.



- **Thermal noise:**
 - Created by **velocity fluctuations** of charge carriers in a conductor;
- **Shot noise:**
 - Created by **fluctuations in the number** of charge carriers (e.g. tunneling events in a semi-conductor diode);
 - Proportional to the total average current.

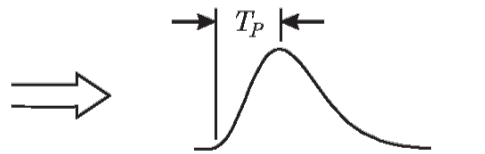
- Thermal noise and shot noise are **both white noise.**
 - Noise **power density per unit bandwidth** is constant:
 - **Larger bandwidth** → **larger noise.**
- S/N ratio increases as **noise bandwidth is reduced:**
 - Until signal components are attenuated significantly.



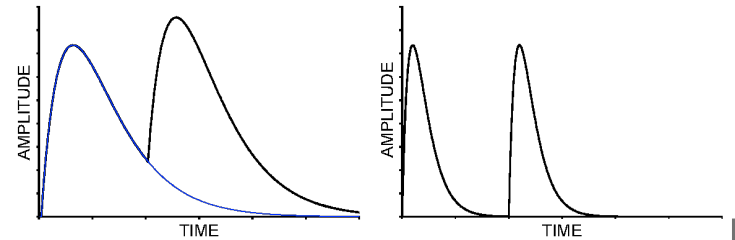
- Typically, the **pulse shaper** transforms a **narrow** detector current pulse to a **broader** pulse:
 - In order to **increase rising time**;
 - To **reduce bandwidth**;
 - To **reduce electronic noise**;
- With a **gradually rounded maximum** at the peaking time TP:
 - To **facilitate measurement** of the **peak amplitude**;

SENSOR PULSE

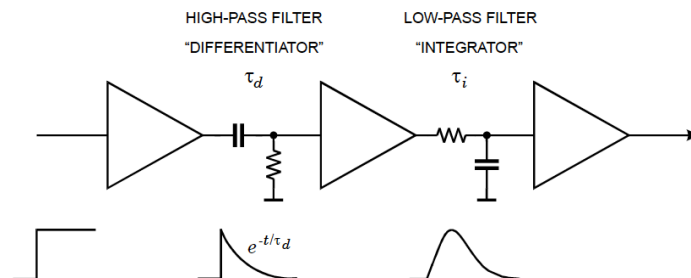
SHAPER OUTPUT



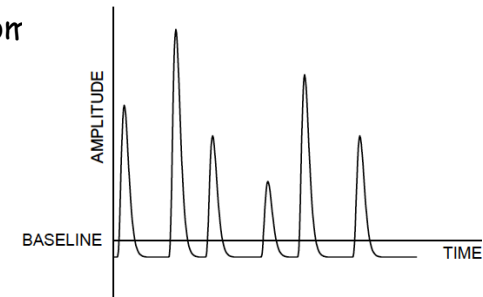
- Broad pulses reduce the temporal spacing between consecutive pulses;
- Need to limit the effect of "**pile-up**":
 - Pulses **not too broad**;
- As usual in life: a **compromise**.



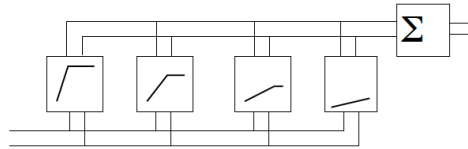
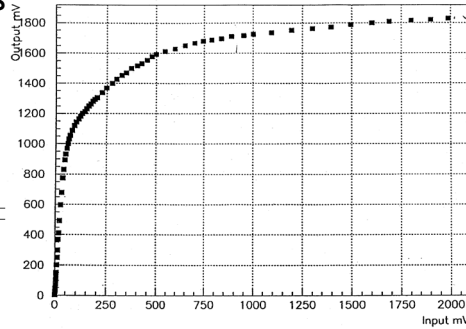
- Example: **CR-RC Shaper**:
 - In this case made out of CR (differentiator) and RC (integrator) filters;
 - Key elements:
 - **Lower frequency bound** (related to **pulse duration**);
 - **Upper frequency bound** (related to **rise time**).



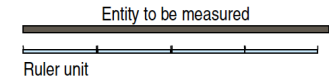
- **Any series capacitor** in a system **prevents transmission of a DC component**.
- A sequence of unipolar pulses has a DC component that depends on the duty factor, i.e. the event rate.
- The **baseline shifts** to make the **overall transmitted charge equal zero**.
- **Random rates** lead to random fluctuations of the baseline shift:
 - **Spectral broadening**;
- Need baseline restorer.



- A non-linear transformation:
 - **Compressing** the signal according to an appropriate **piecewise linear transfer function**;
 - Producing an output in the **range best suited for digitization** circuit.
- Typically **sum** of the outputs of several linear amplifiers with **different gain and upper cutoff**.



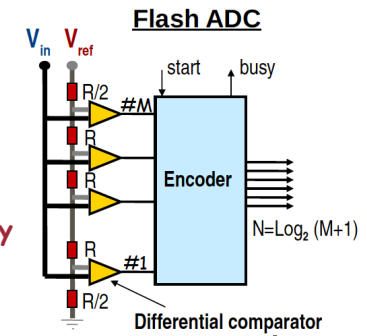
- **Digitization:**
 - Encoding an analog value into a **binary representation**;
 - By **comparing** entity with a **ruler**.
- **Flash ADC** is the simplest and fastest implementation:



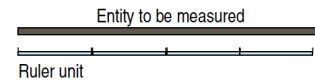
- **M comparisons in parallel**;
- Input voltage V_{in} compared with M fractions of a reference voltage V_{ref} :

$$\frac{1}{2} \frac{V_{ref}}{M}, \dots, \left(M - \frac{1}{2}\right) \frac{V_{ref}}{M}$$

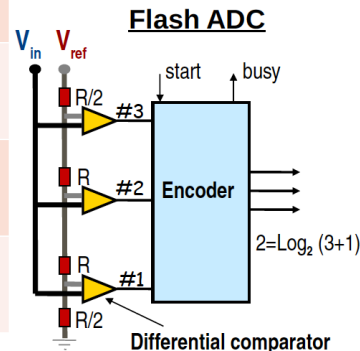
- Result is **encoded into a compact binary form** of N bits:
 $N = \log_2(M + 1)$



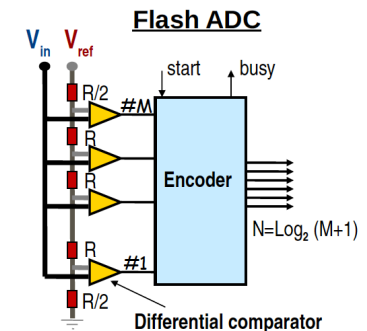
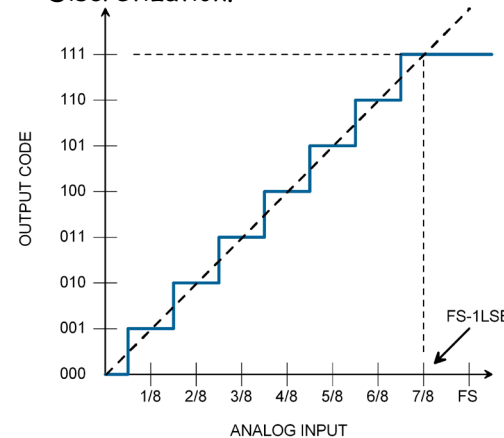
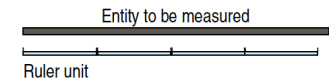
- **E.g.:** $M = 3, N = 2$.



	Comparison results	Encoded form
$\frac{V_{in}}{V_{ref}} \in \left[0, \frac{1}{6}\right]$	0-0-0	00
$\frac{V_{in}}{V_{ref}} \in \left[\frac{1}{6}, \frac{3}{6}\right]$	0-0-1	01
$\frac{V_{in}}{V_{ref}} \in \left[\frac{3}{6}, \frac{5}{6}\right]$	0-1-1	10
$\frac{V_{in}}{V_{ref}} \in \left[\frac{5}{6}, \infty\right]$	1-1-1	11

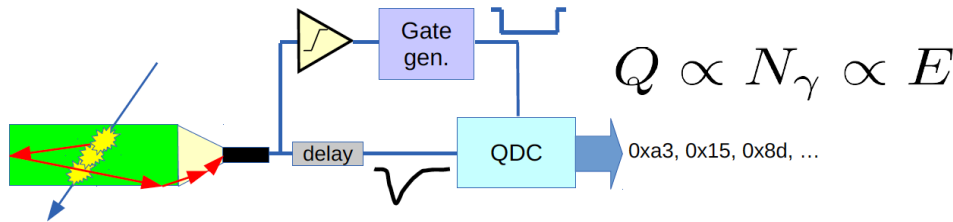
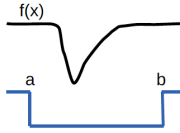


- **ADC transfer function:**
 - Output code vs analog input.
 - Discretization.

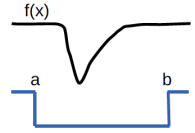


- **QDC** (Charge to Digital Converter).
 - Often we have a **current** and we are **interested** in the **total charge**;
 - Essentially an **integration step** followed by an **ADC**;
 - Integration require limits: **gate**.

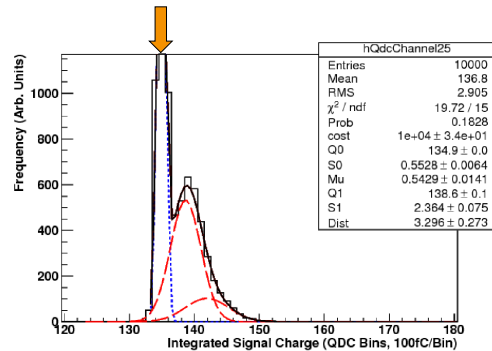
$$I = \int_a^b f(x) dx$$



- **Relative timing** between **signal** and **gate** is important:
 - Delay tuning.
- Gate should be **large enough** to contain the **full pulse** and to accommodate for the **jitter**:
 - Fluctuations are always present.
- Gate should **not be too large**:
 - Increases the noise level;
 - Increase dead time.

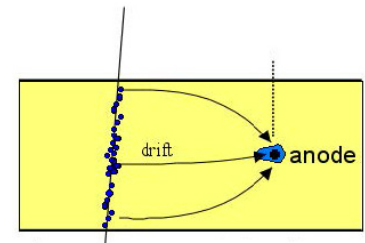


- **Pedestal**:
 - Due to **PMT dark current** (thermionic emission), **thermal noise**, etc.;
 - The same noise enters the physics measurements and contributes with an offset to the distribution;
 - Can be measured with an **out-of-phase trigger**.
- The result of a pedestal measurement has to be **subtracted** from the charge measurements.



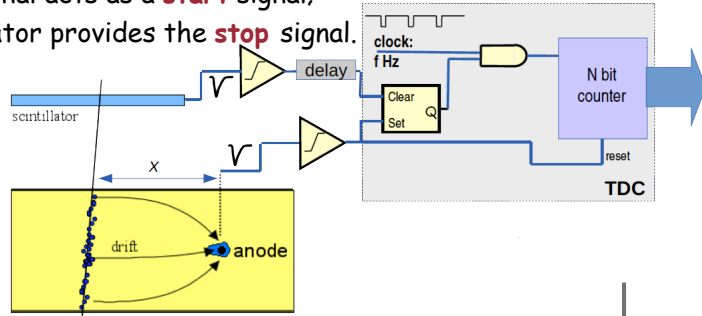
- Example: measure the **position** of a particle in a **wire (drift) chamber**.
- The ionization electrons created by the passage of the particle will take a time Δt to reach the anode wire:
 - Transit time is normally **negligible** with respect to Δt ;
 - If we consider a **constant drift speed** v_D (e.g.: $50 \mu\text{m/ns}$), then position is:

$$x = v_D \cdot \Delta t$$



- Wire chamber alone is not sufficient:
 - We need a triggering system (e.g. a scintillator slab).
- We can measure the **time offset** between the two signals using a **N -bit digital counter** driven by a **clock of frequency f** :

- The wire signal acts as a **start** signal;
- The scintillator provides the **stop** signal.
- This device is a **TDC**.



- Why **spend bandwidth** sending **data** that is **zero for the majority of the time?**
- Perform **zero-suppression** and only send data with non-zero content:
 - **Requires to identify the data** with a **channel number** and/or a **time-stamp**;
 - We do not want to loose information of interest so this must be done with **great care** taking into account:
 - Pedestals,
 - Baseline variations,
 - Noise.
 - Not worth it for occupancies above ~10%.
- Alternative: **data compression** Huffman encoding and alike:
 - **Slow, Power-intensive.**

Channel ID
Time tag
Measurement
Channel ID
Time tag
Measurement
Channel ID
Time tag
Measurement
Channel ID
Time tag
Measurement

- In the **LHCb upgrade** all **Front End** electronics **transmit data continuously** at **40 MHz** to the Readout Boards.
- **Very large number of optical links** needed between the FE and the new Readout Boards.
- Almost a factor of ten could be gained by sending **zero-suppressed data** already at the FE:
 - **Reducing the number of optical links** from ~80000 to ~10000.
 - The zero-suppression will be **performed in radiation-tolerant FE chips**.
 - A possible consequence of zero-suppression is a **varying latency** of data transmission.

Data Acquisition

- E.g.: Measure temperature at a fixed frequency:

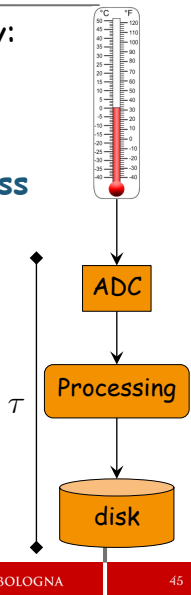
- **FEE**: ADC performs analog to digital conversion;
- **DAQ**: CPU does ADC readout and disk write.

- System limited by the **time τ needed to process an event**:

- ADC conversion + CPU processing + storage.

- The maximum sustainable DAQ rate is the inverse of τ , e.g.:

$$\tau = 1 \text{ ms} \Rightarrow \nu = \frac{1}{\tau} = 1 \text{ kHz}$$

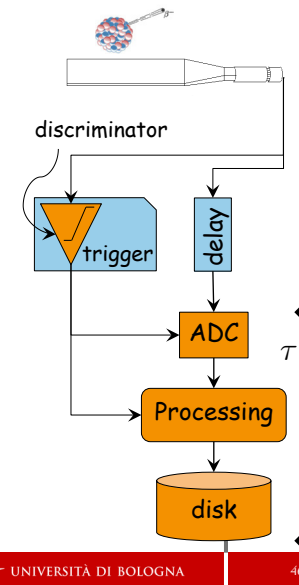
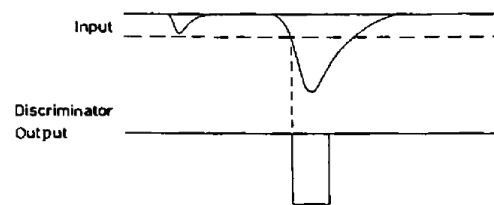


- E.g.: Beta decay frequency:

- Events **asynchronous** and **unpredictable**;

- A **physics trigger** is needed:

- **Delay**: compensates for trigger latency;
- **Discriminator**: generate an output signal when the amplitude of input signal exceeds a certain threshold.

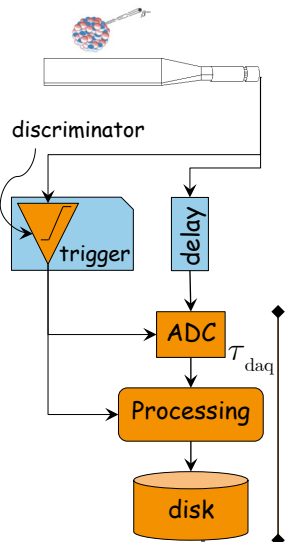
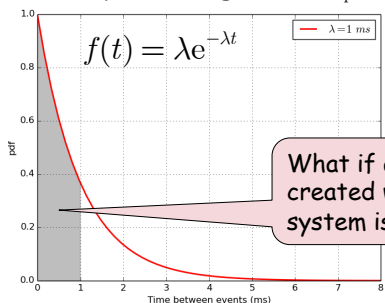


- The process is **poissonian**:

- Fluctuations in time between events;

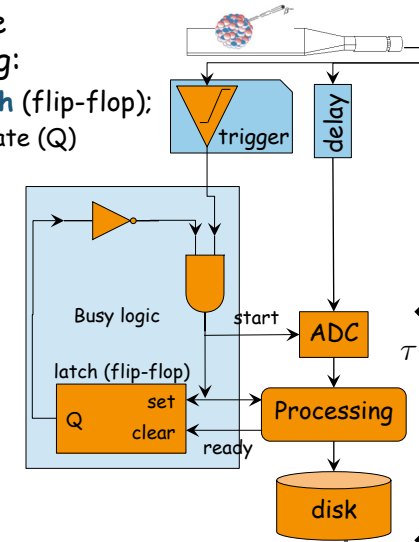
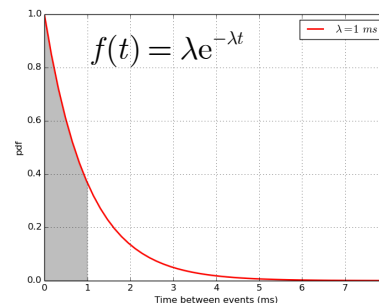
- Let's assume for example:

- A process rate $\nu_{ph} = 1 \text{ kHz}$, i.e. $\lambda = 1 \text{ ms}$;
- A processing time $\tau_{daq} = 1 \text{ ms}$.



- Busy logic** avoids triggers while the system is busy in processing:

- E.g.: using an **AND port** and a **latch (flip-flop)**:
- A bistable circuit that changes state (Q) by signals applied to the control inputs (SET, CLEAR).



- **Definitions:**
 - Average **rate** of **physical** events (input): ν_{ph} ;
 - Average **rate** of **DAQ** (output): ν_{daq} ;
 - **Dead time**, the time the system requires to process an event, without being able to handle other triggers: τ ;
 - Probability that DAQ is busy: $P(\text{busy}) = \nu_{daq} \cdot \tau_{daq}$;
 - Probability that DAQ is free: $P(\text{free}) = 1 - \nu_{daq} \cdot \tau_{daq}$;

• **Therefore:**

$$\nu_{daq} = \nu_{ph} \cdot P(\text{free}) = \nu_{ph} (1 - \nu_{daq} \cdot \tau_{daq})$$

$$\nu_{daq} = \frac{\nu_{ph}}{1 + \nu_{ph} \cdot \tau_{daq}} < \nu_{ph}$$

$$\varepsilon = \frac{N_{\text{saved}}}{N_{\text{tot}}} = \frac{\nu_{daq}}{\nu_{ph}} = \frac{1}{1 + \nu_{ph} \cdot \tau_{daq}} < 100\%$$

- **Due to stochastic fluctuations:**
 - **DAQ rate always less than physics rate;**

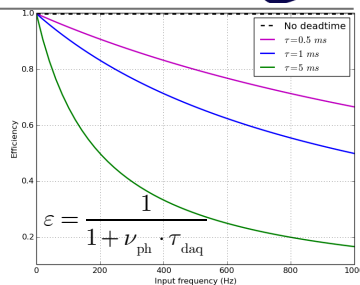
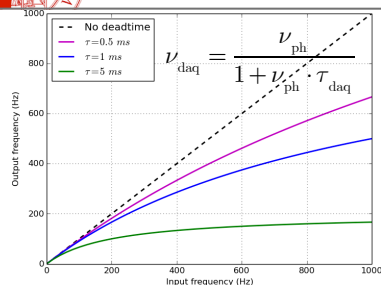
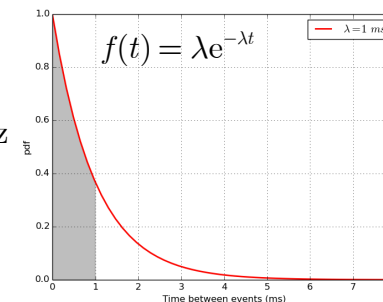
$$\nu_{daq} = \frac{\nu_{ph}}{1 + \nu_{ph} \cdot \tau_{daq}} < \nu_{ph}$$

- **Efficiency always less than 1;**

$$\varepsilon = \frac{1}{1 + \nu_{ph} \cdot \tau_{daq}} < 100\%$$

• **Example:**

$$\begin{cases} \nu_{ph} = 1 \text{ kHz} \\ \tau_{daq} = 1 \text{ ms} \end{cases} \Rightarrow \begin{cases} \nu_{daq} = 500 \text{ Hz} \\ \varepsilon = 50\% \end{cases}$$



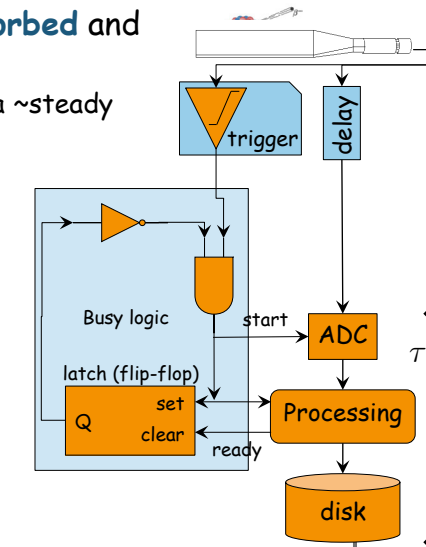
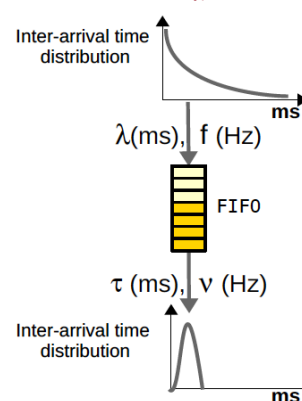
$$\varepsilon \approx 100\% (\nu_{daq} \approx \nu_{ph}) \Rightarrow \nu_{ph} \cdot \tau_{daq} \ll 1 \Rightarrow \tau_{daq} \ll \lambda_{ph}$$

$$\begin{cases} \varepsilon \approx 99\% \\ \nu_{ph} = 1 \text{ kHz} \end{cases} \Rightarrow \tau_{daq} < 0.01 \text{ ms} \Rightarrow \frac{1}{\tau_{daq}} > 100 \text{ kHz}$$

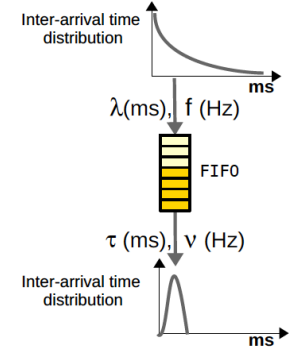
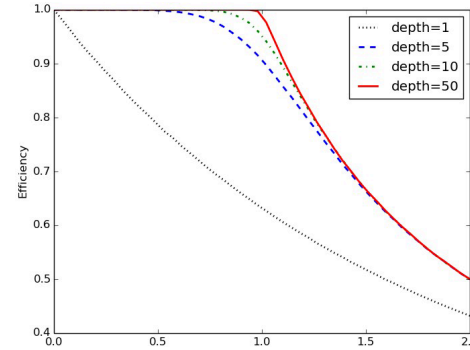
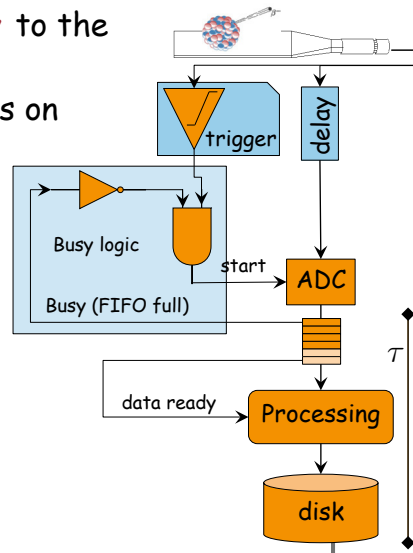
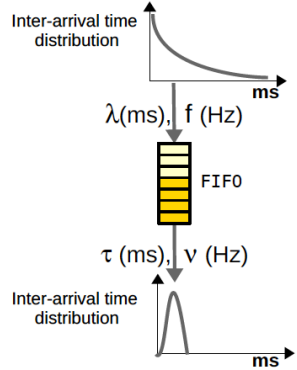
- To cope with the **input signal fluctuations**, we have to **over-design** our **DAQ system** by a factor 100!

- **Input fluctuations can be absorbed and smoothed by a queue:**

- A First In First Out can provide a ~steady and **de-randomised** output rate.

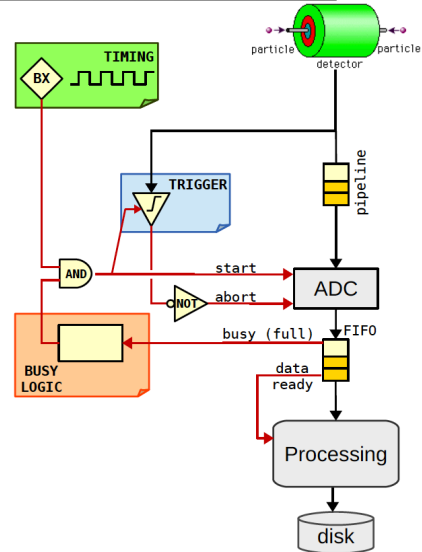


- It introduces **additional latency** to the data path;
- The effect of the queue depends on its **depth**.

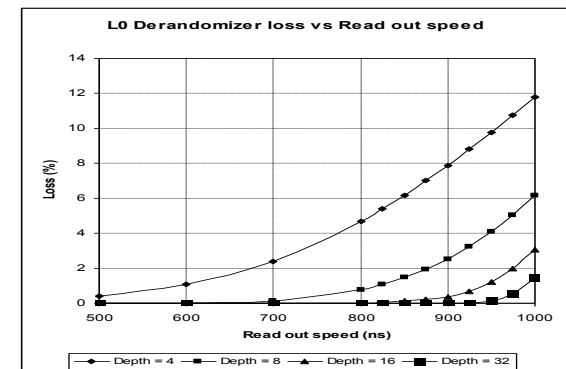


- **Efficiency vs traffic intensity** ($\rho = \tau / \lambda$) for different queue depths;
- Analytic calculation possible for very simple systems only
 - Otherwise Monte Carlo simulation is required

- Particle collisions are **synchronous**:
 - So, do we still need de-randomisation buffers?
- **Trigger** rejects uninteresting events:
 - **Good events are unpredictable**;
- Even if collisions are synchronous, the **time distribution of triggers is random**:
 - **De-randomization is still needed**.

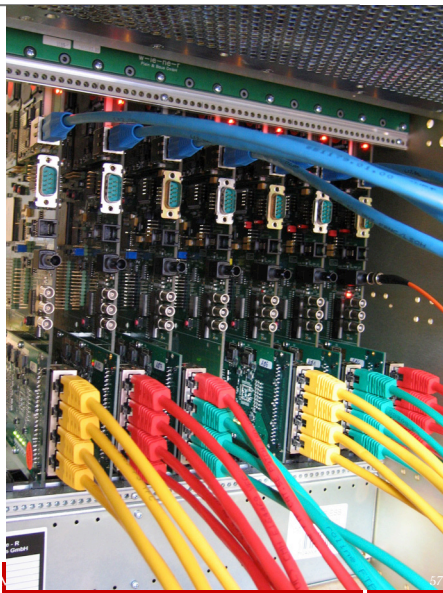
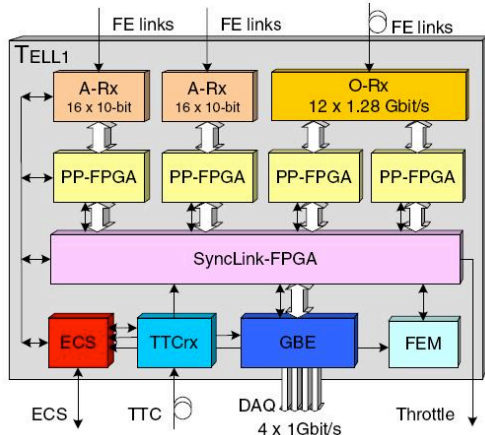


- Working point for the **LHCb** experiment:
 - Max readout time: **900 ns**;
 - Derandomiser depth: **16 events**;
 - **1 MHz** maximum trigger accept rate.



LHCb DAQ Custom Component: Tell1 Boards (the same for all sub-detectors)

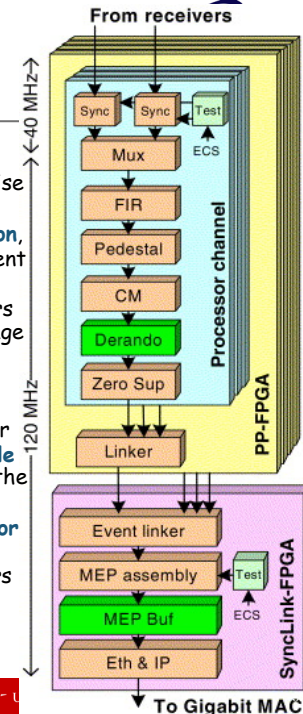
- Input: 24 x 1.6 Gb/s optical link or 64 x analog copper links.
- Output: 4 x 1000Base-T.
- ECS: Credit card PC (Linux) with separate 100Base-T interface.



Tell-1 Boards (II)



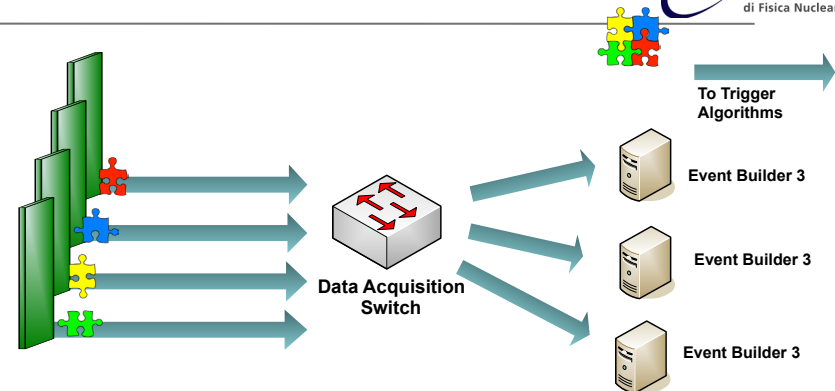
- FIR: Finite Impulse Response filter.
- CM: Common Mode noise corrections.
- After **zero suppression**, the length of each event is variable.
- **Derandomizing** buffers are employed to average the data rate and the data processing time.
- To prevent any overflows, each buffer can generate a **throttle**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



Event Building

- Most HEP detectors are read out through **multiple DAQ front-ends** (~10000 in LHCb Upgrade):
 - Each responsible for a **segment** of the full detector.
 - Event-fragments are digitized, pre-processed and **tagged** with a **unique, monotonically increasing number**.
- **Event building** is the process of **assembling** the many fragments of readout into a **single whole**:
 - **Getting** the **pieces** of the event **together**;
 - It often requires the **coordinated work** of **several computing nodes**;
 - Involved computing nodes need to **exchange data** each other:
 - This require a **data bus** or a **network**.

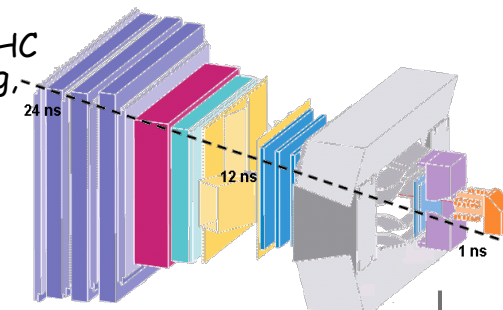
Event Building (II)



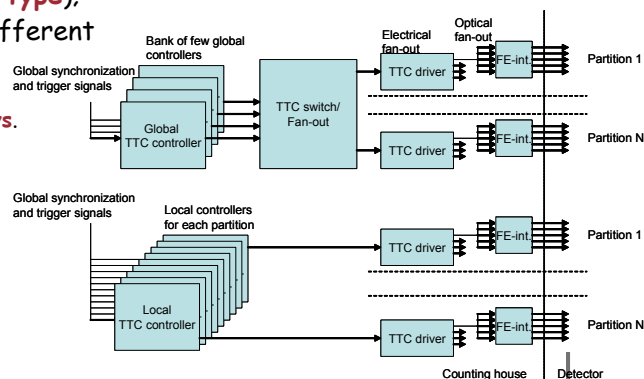
- 1 Event fragments are received from detector front-end
- 2 Event fragments are read out over a network to an event builder
- 3 Event builder assembles fragments into a complete event
- 4 Complete events are processed by trigger algorithms

- Usually **low-level trigger** is based on **local (sub-detector) data**:
 - Event fragments are sent to trigger electronics through **dedicated lines, before event building**;
- **High-level trigger** requires **all detector data**:
 - HLT is performed on **built events**.

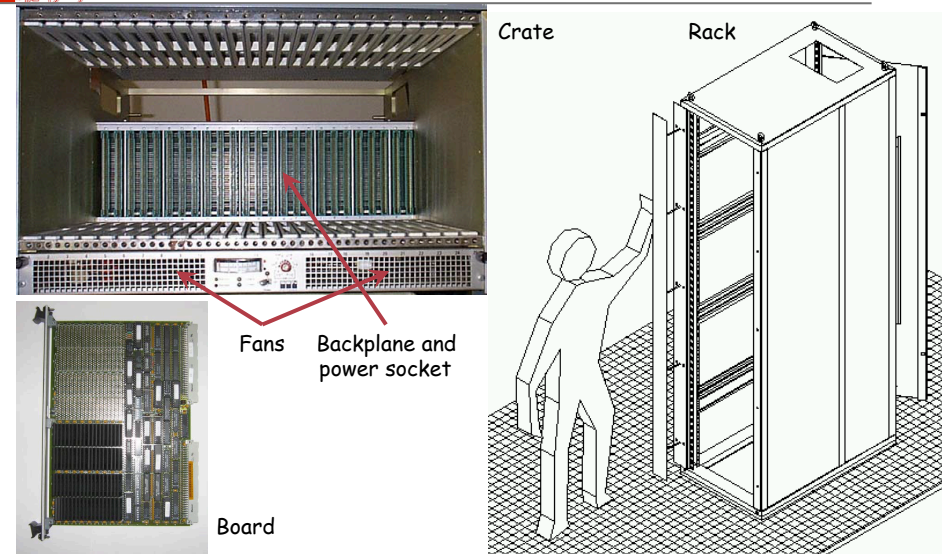
- An **event** is a snapshot of the values of all detector front-end electronics elements, which have their value caused by the same collision;
- A common clock signal must be provided to all detector elements:
 - Since the c is constant, the detectors are large and the electronics is fast, the **detector elements must be carefully time-aligned**.
- Common system for all LHC experiments **TTC** (Timing, Trigger and Control) based on radiation-hard opto-electronics.



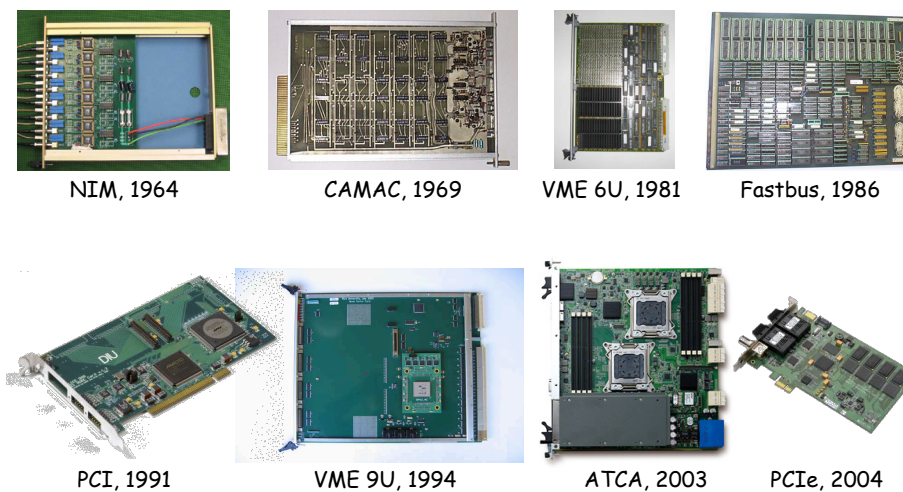
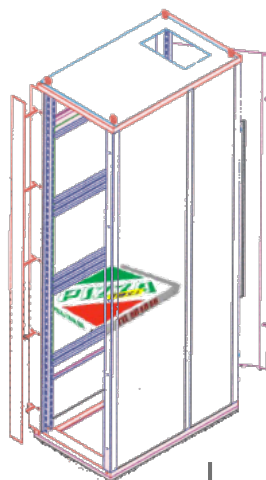
- Sampling clock with **low jitter**;
- Synch **reset**;
- **Synchronization** with **machine** bunch structure;
- Calibration;
- **Trigger** (with event **type**);
- Time align all the different sub-detectors and channels:
 - **Programmable delays**.



- **Modularizing DAQ electronics** helps in these respects:
 - Allows for the **re-use** of **generic modules** in different applications;
 - **Limiting** the **complexity** of **individual modules** increases their reliability and maintainability;
 - You can profit from **3rd party support** for common modules
 - Makes it easier to achieve **scalable designs**;
 - **Upgrades** (for performance or functionality) are less difficult.



- The width and height of racks and crates are measured in **US units: inches (in, '')** and **rack units (U)**:
 - 1 in = 2.54 cm;
 - 1 U = $1\frac{3}{4}$ in = 4.445 cm;
- The **width** of a "standard" rack is **19 in**;
- The **height** of a crate (also sub-rack) is measured in **Us** (typically, **42U**);
- Rack-mountable things, in particular computers, which are 1U high are often called **pizza-boxes**;
- At least in Europe, the **depth** is measured in mm;
- Gory details can be found in IEEE 1101.x (VME mechanics standard).

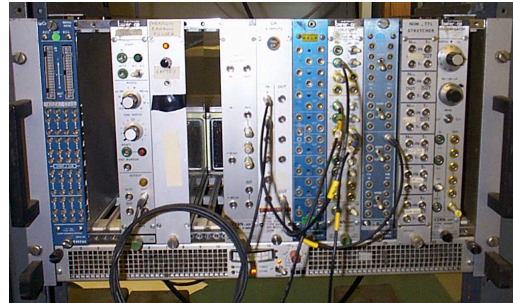


- **Not actually a data bus:**
 - No common backplane bus;
- **Backplane provides only powers** to functional modules;
- 250 x 193 mm board size:
 - 12 boards per crate maximum;
- Plug-and-play approach:
 - Does not need any software;
- **Front panel settings and cable connections;**
- Amplifiers, shapers, discriminators, delay units, etc.

BNC connectors

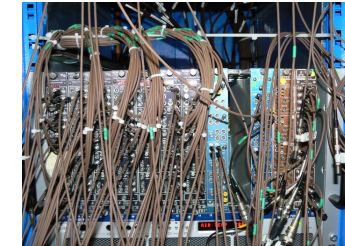
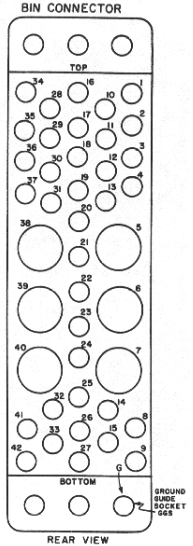


LEMO connectors



- NIM logic levels:
 - 0 = 0A (0V);
 - 1 = -12 to -32 (typical -16) mA at 50 Ω (-0.8V);
- NIM connector:
 - 42 pins in total;
 - 11 pins used for power (+/- 6, 12, 24V);
 - 2 logic pins (reset & gate).

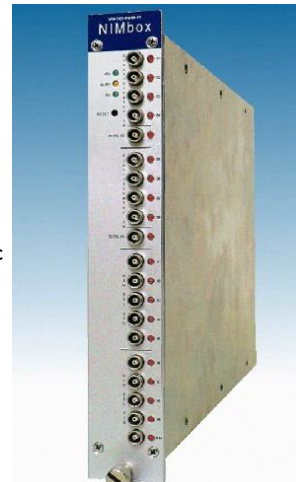
NIM backplane connector



- NIM is still alive.

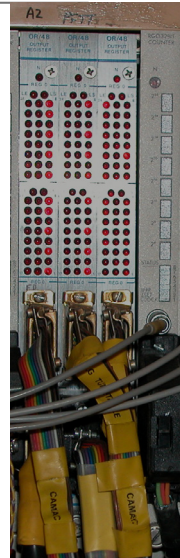


100 MS/s digitizer with optical read-out

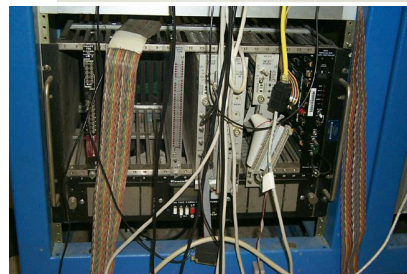
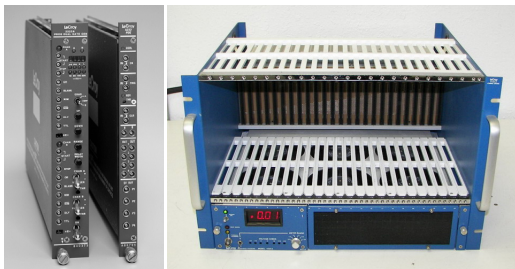


General purpose NIM module with programmable logic

- **CAMAC** was the **first successful databus interface** between **commercial computers** and **custom detector electronics**;
- **Most of physics experiments in late 60's -late 80's** were based on parallel CAMAC electronics, e.g.:
 - UA1 (interfaced to Apple MacIntosh Plus);
 - UA2 (interfaced to VAX 11/780).
- Several large distributed **accelerator control systems** (CERN, Fermilab) were based on serial CAMAC.



- **IEEE Standard 583-1975;**
- Up to **24 modules** in a crate;
- **1 crate controller:**
 - An interface to a computer or to other crates;
- Relatively slow:
 - **1 μ s** data exchange **cycle;**
 - **24 bit bus;**
 - **24 Mb/s.**
- Board size is relatively **small**, especially for multi-channel electronics.
- **Backplane interface** takes significant amount of board space.
- Not suitable for demanding analog requirements because of high **noise of digital circuits.**
- A lot still around.



IEEE 596 Parallel Highway Interface:

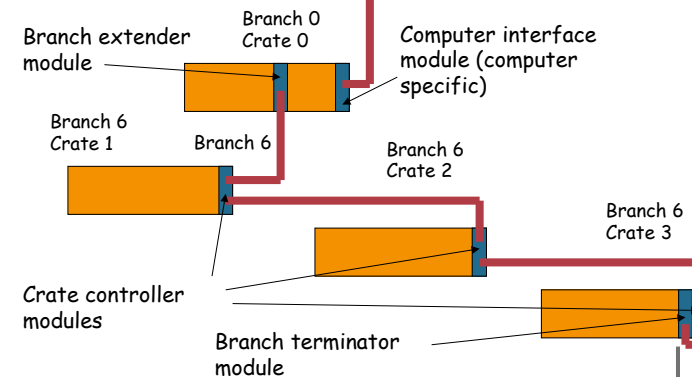
- Up to **7 crates;**
- **24-bit** wide data transfers @ **1 MHz;**
- **Twisted pair cable bus**, up to **15 m.**



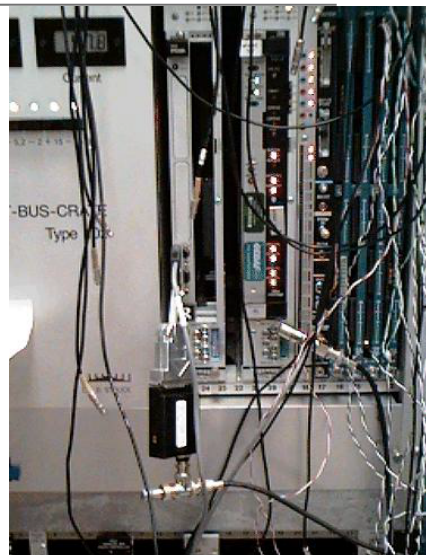
CAMAC interface plugged into computer bus



Computer with CAMAC driver loaded



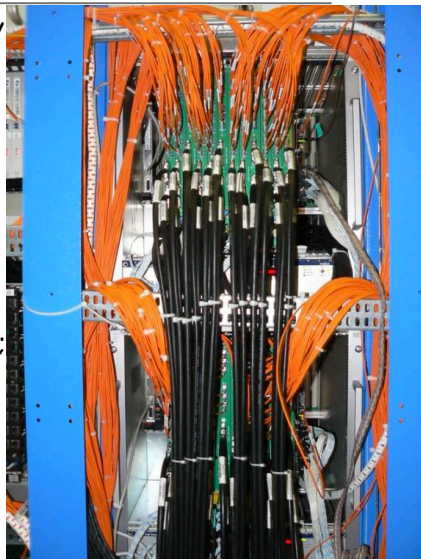
- Designed by **physicists for physicists.**
- Large form-factor, high channel densities.
- Widely used in late 80's -mid 90's in HEP community:
 - **CDF Run I** (1985-1995): ~1500 modules in ~150 crates
 - All four experiments at the **LEP e+e-** collider at CERN;
 - **DELHI, ALEPH, L3, OPAL** in 1989-2000: ~700 crates;
 - At **SLAC, DESY** and other smaller experiments;



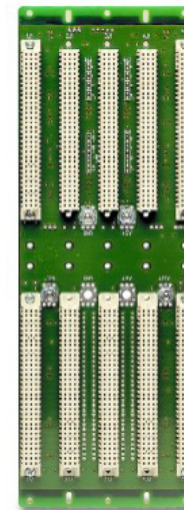
- **ANSI/IEEE 960-1986 Standard:**
 - Large 367 x 400 mm board;
 - 26 slots in crate;
 - ECL backplane signals;
 - Asynchronous transfers;
 - **32-bit data / 32-bit address;**
 - **100 ns cycle** (10 MHz);
 - Up to **320 Mb/s bandwidth;**
 - Multiprocessing;
 - Block transfers;
 - Sparse data scanning;
 - Well defined control and status registers.
- **But:**
 - **High power consumption;**
 - **Poor backplane** connector;
 - Complex **inter-crate interface;**
 - **Weak industry support.**



- VME standard was proposed in 1981 by **Motorola, Mostek and Signetics**;
- Processor independent, but **signal set** has its roots in **MC 68000 CPU**;
- **Open architecture**;
- VME International Trade Association (VITA) remains the driving force;
- **Large number of commercial products** (used heavily in the military);
- **32/64 bit bus** (320/640 Mb/s);
- Currently there are **more than 1000 VMEbus systems at CERN** (accelerator and experiments).

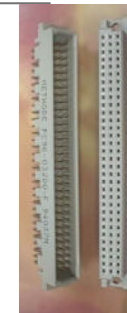


Mandatory bus for all 3U and 6U boards

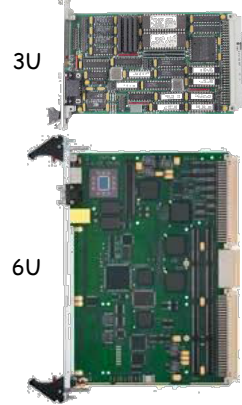


Extension bus for A32D32 6U boards

- 16-bit Data Bus
- 24-bit Address Bus
- 6-bit Address Modifier Bus
- 7 Interrupts
- Arbitration Bus
- Clock, Control and Status signals
- +5V, +12V, -12V powers, 5 GND lines



- 16-bit Data Bus Extension
- 8-bit Address Bus Extension
- Additional +5V (3 lines) and GND (4 lines)
- 64 unbusse User I/O lines



9U

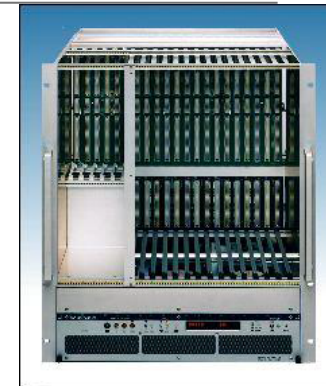


160 mm

340 mm



21 slot 6U crate

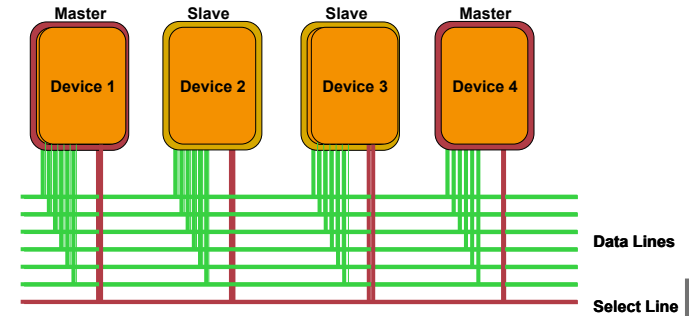


21 slot 9U crate (with 6U section)

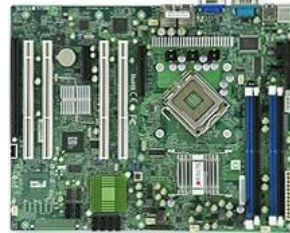
Classes of modules (logical)

- **Master:**
 - A module that can initiate data transfers;
- **Slave:**
 - A module that responds to a master;
- **Interrupter:**
 - A module that can send an interrupt (usually a slave);
- **Interrupt handler:**
 - A module that can receive (and handle) interrupts (usually a Single Board Computer);
- **Arbiter:**
 - A piece of electronics (usually included in the SBC) that arbitrates bus access and monitors the status of the bus.
 - It should always be installed in slot 1 of the VMEbus crate if interrupts are used.

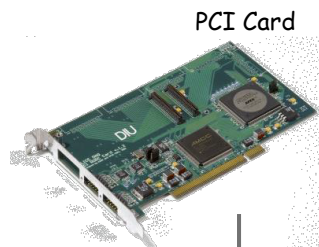
- A bus connects two or more devices and allows them to communicate;
- The bus is **shared** between all devices on the bus:
 - Arbitration is required;
- Devices can be **masters** or **slaves** (some can be both);
- Devices can be uniquely identified ("**addressed**") on the bus.



- **Local computer bus** for attaching hardware devices in a **computer**.
- First standardized in 1991:
 - Replaced the older **ISA/EISA/MCA** cards;
 - Initially intended for PC cards;
 - Later spin-offs: **CompactPCI, PXI, PMC**.

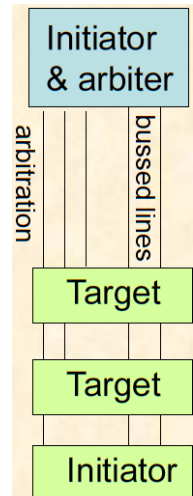


PCI Motherboard



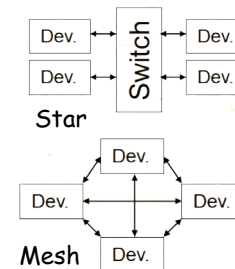
PCI Card

- **Main features:**
 - Synchronous timing (but wait cycles possible);
 - **Clock** rates:
 - Initially **33 MHz**. Later: **66 MHz**, (**PCI-X: 100 and 133 MHz**);
 - **Bus width:**
 - Initially **32 bit**. Later: **64 bit**;
 - Signaling voltage:
 - Initially **5 V**. Later **3.3 V**;
 - Bus topology:
 - **1 to 8 slots per bus**;
 - Busses can be connected to form a **tree**;
 - **Address** and **data** as well as most protocol lines are **shared** by all devices;
 - The lines used for **arbitration** are connected **point-to-point**;
 - The routing of the **interrupt** request lines is more complicated...
 - A system can consist of several Initiators (master) and Targets (slave) but **only one Initiator (master) can receive interrupts**.

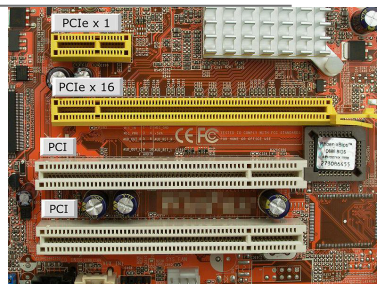


- What is wrong about "parallel"?
 - You need **lots of pins** on the **chips** and **wires** on the **PCBs** (printed circuit boards):
 - Control, data and address lines;
 - The **skew** (difference in arrival time of simultaneously transmitted bits) between lines **limits the maximum speed**;
- What is wrong about "bus"?
 - A bus is **shared between all devices** (each new active device **slows every other device down**);
 - **Speed** is a function of the **length (impedance)** of the lines;
 - Bus-frequency (number of elementary operations per second) can be increased, but decreases the **maximum physical bus-length**;
 - **Number of devices** and **physical bus-length** is limited;
 - Communication is **limited to one master/slave pair at a time**.
- Buses are typically **useful for systems < 1 GB/s**:
 - **Not useful for DAQ at LHC**.

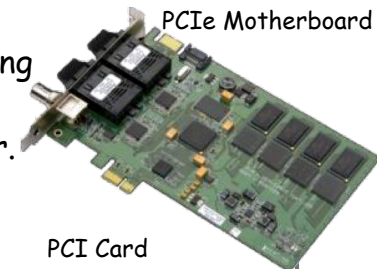
- **Parallel Buses Are Dead!**
 - RTC magazine, September **2006**, Ben Sharfi CEO, General Micro Systems;
- **Switched Serial Link:**
 - Packet switching;
 - Star or mesh topology;
- Examples:
 - **PCIe** (PCI Express);
 - **InfiniBand**;
 - **Ethernet**;
 - **Serial ATA**;
 - **Fiber Channel**;
 - Etc.



- **Not a bus any more:**
 - But a **point-to-point link**;
- Data not transferred on parallel lines but on one or **several serial lanes**:
 - **Lane**: One pair of LVDS lines per direction;
 - Devices can support up to 32 lanes;
- Protocol at the link layer has nothing to do with protocol of parallel PCI;
- Fully transparent at the S/W layer.



PCIe Motherboard

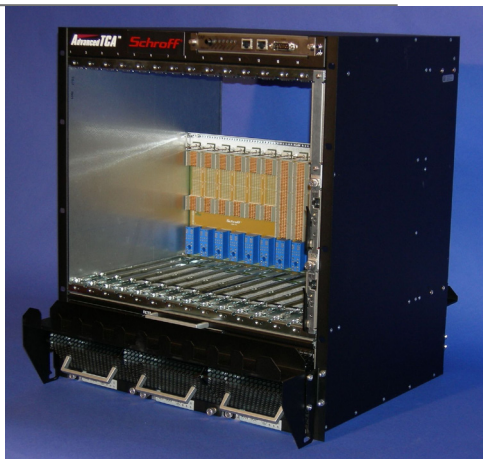


PCI Card

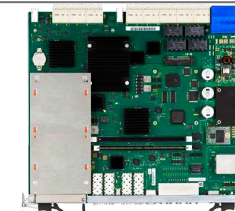
- **Clock rate:**
 - 2.5 GHz (PCIe1.0);
 - 5 GHz (PCIe 2.0);
 - 8 GHz, (PCIe 3.0);
 - 16 GHz (PCIe 4.0-draft);
- **Encoding:**
 - 8b/10b;
 - 128b/130b (PCIe3.0);
- **Raw transfer rate per lane:**
 - 250 MB/s, 2 Gb/s (PCIe 1.0);
 - 500 MB/s, 4 Gb/s (PCIe 2.0);
 - 985 MB/s, 7.88 Gb/s (PCIe 3.0);
 - 1969 MB/s, 15.7 Gb/s (PCIe 4.0-draft);
- Devices can support **up to 32 lanes**;

The Basic Idea:

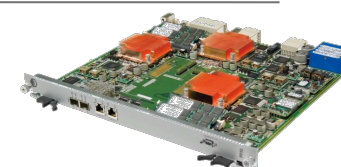
- Telecom companies are using **proprietary electronics**;
- Let's design a **standard** for them **from scratch**;
- It has to have all the **features** telecom companies need:
 - **High availability** (99.999%);
 - **Redundancy** at all levels;
 - Very high **data throughput**;
 - Sophisticated **remote monitoring** and **control**;



ATCA Shelf



Switch blade



Payload card

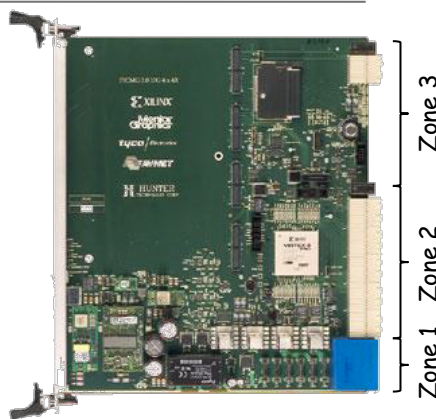


AMC (Advanced Mezzanine Card) carrier



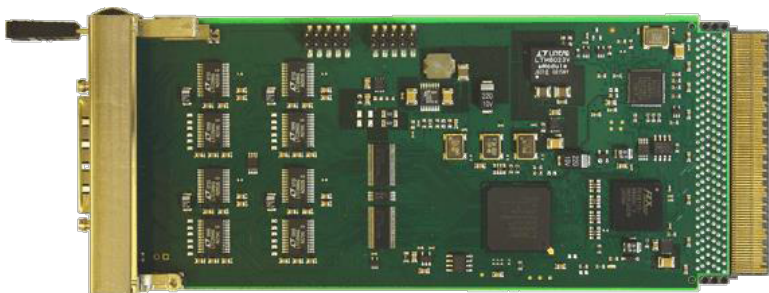
Rear Transition Module

- More of a system than a board standard;
- Started in 2001 by **~100 companies**;
- One form factor:
 - Front: **8U** x 280 mm x 30.48 mm (**14 slots** per 19" crate);
 - Rear: 8U x 60 mm (5W);
- Supply voltage: **-48 V**:
 - DC-DC conversion each on-board;
- Power limit: **200 W** (400-600-800 W) per card;
- Connectors:
 - Zone 1: One connector for power & shelf management;
 - Zone 2: 1-5 ZD connectors for data transfer;
 - Zone 3: User defined connector for rear I/O.

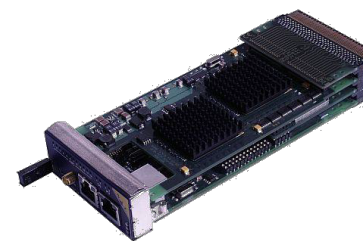


- The ATCA **backplane** provides **point-to-point connections** between the boards and **does not use a data bus**.
- **Zone-2** provide the connections to the **Base Interface** and **Fabric Interface**.
- The **Base Interface** can only be 10BASE-T, 100BASE-TX, or 1000BASE-T Ethernet:
 - Since all boards and hubs are required to support one of these interfaces there is always a network connection to the boards.
- All **Fabric connections** use **point-to-point 100 Ω differential signals**:
 - Zone-2 is called "**Fabric Agnostic**" which means that any Fabric that can use 100 Ω differential signals can be used with an ATCA backplane.
- The Fabric is commonly **Gigabit Ethernet**, but can also be **Fibre Channel**, **40-Gigabit Ethernet**, **InfiniBand**, **PCI Express**, etc.

- **ATCA blades** are big.
- Small **mezzanine modules** could be helpful to modularize their functionality:
 - PMC/XMC mezzanines are not hot-swappable;
 - Let's design a new type of mezzanine for ATCA



- AMC mezzanines are great but ATCA is a heavy standard and the H/W is expensive:
 - Let's define a standard that **allows for using AMCs directly in a shelf**;
 - i.e. **Promote** the AMC from "mezzanine" to "module".



MicroTCA.4 Based
40GbE Ready Platform

- LHC and experiments at CERN:
 - Still many VMEbus and PCI based;
 - **CMS**: Several **μTCA** systems in operation;
 - **ATLAS**: **ATCA** proposed as VMEbus replacement, many R&D projects;
 - **LHCb**: first favored **ATCA** then decided to go for **PCs**;
 - **ALICE**: Still planning to use **ATCA**;
- Control systems of new accelerators:
 - **μTCA** everywhere;
 - XFEL@DESY, SCLS@SLAC, FAIR@GSI.

Network Based DAQ

- What defines **large**?
 - The **number of channels**: for LHC experiments $O(10^7)$ channels:
 - A (digitized) channel can be **between 1 and 14 bits**;
 - The **rate**: for LHC experiments everything happens at **40.08 MHz**, the LHC bunch crossing frequency:
 - This corresponds to 24.9500998 ns or **25 ns among events**;
- HEP experiments usually consist of **many different sub-detectors**:
 - Tracking, calorimetry, particle-ID, muon-detectors.

Network Based DAQ

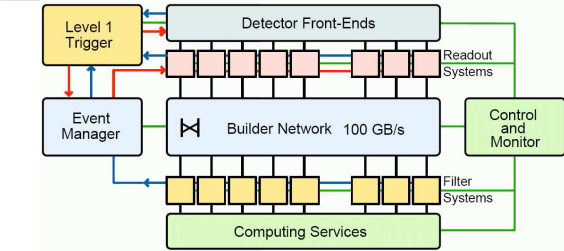
- In large (HEP) experiments we typically have **thousands of devices** to read, which are sometimes **very far from each other**:
 - Buses **can not** do that;
- Network technology solves the **scalability** issues of buses:
 - In a network **devices are equal** ("peers");
 - In a network **devices communicate directly with each other**:
 - **No arbitration** necessary;
 - **Bandwidth guaranteed**;
 - **Data and control** use the **same path**:
 - **Much fewer lines** (e.g. in traditional Ethernet only two)
 - At the signaling level buses tend to use parallel copper lines. Network technologies can be also optical, wire-less and are typically (differential) serial.

High Speed Data Link Technology

- HERA-B:
 - **Shark link** (proprietary, by Analog Devices) until level 2, than **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**.
- LHCb Upgrade:
 - **InfiniBand, 100-Gigabit Ethernet**.

- Send everything, ask questions later:

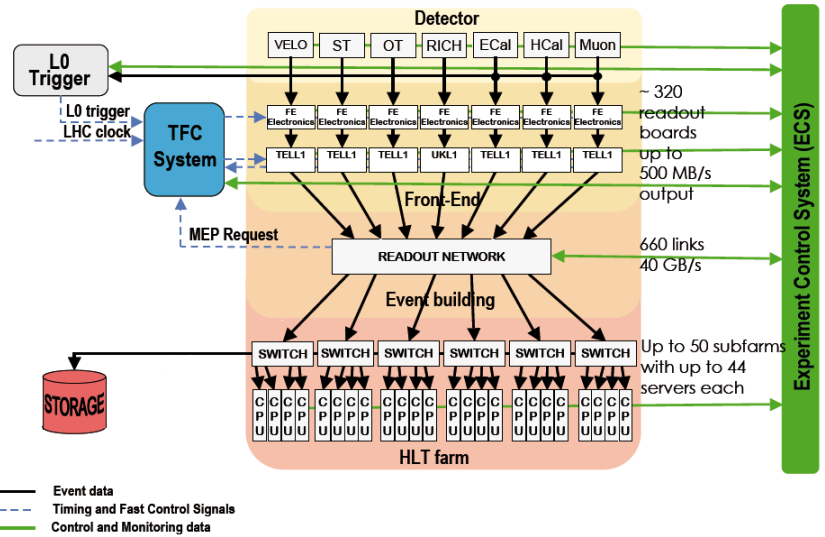
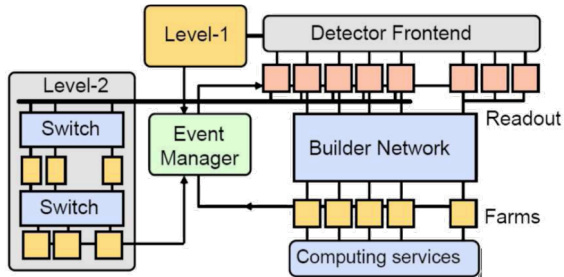
- ALICE, CMS, LHCb:



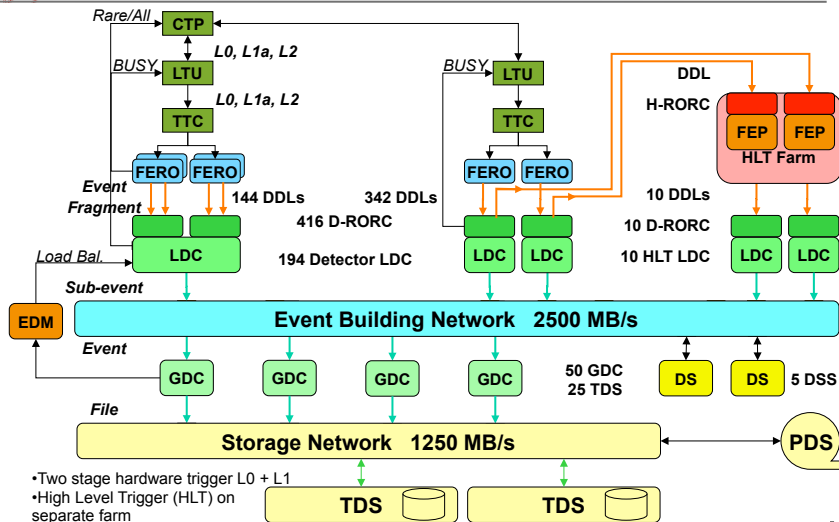
- Send a part first, get better question:

- Send everything only if interesting:

- ATLAS.

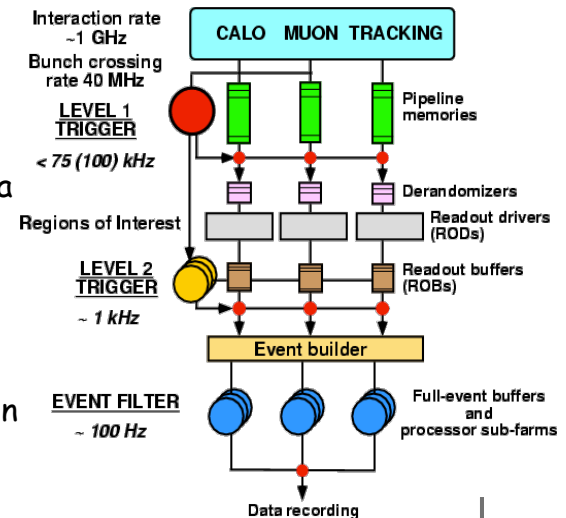


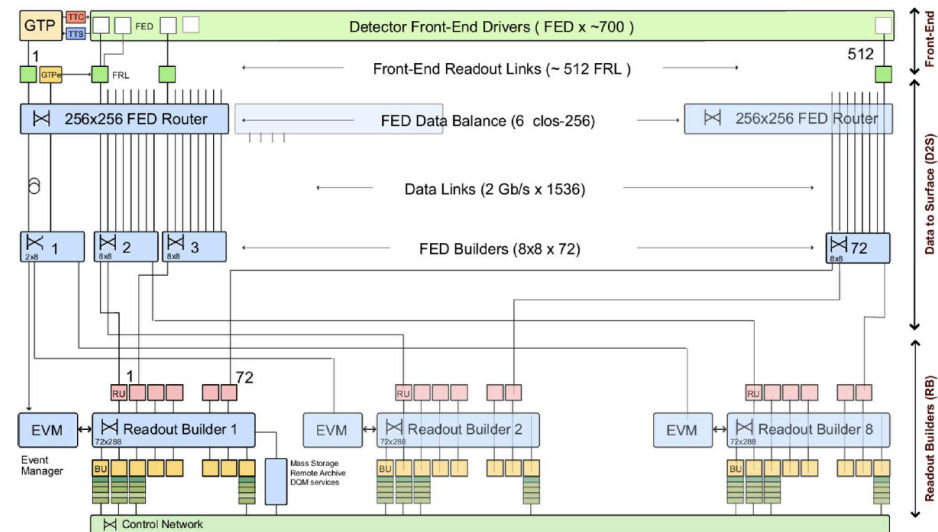
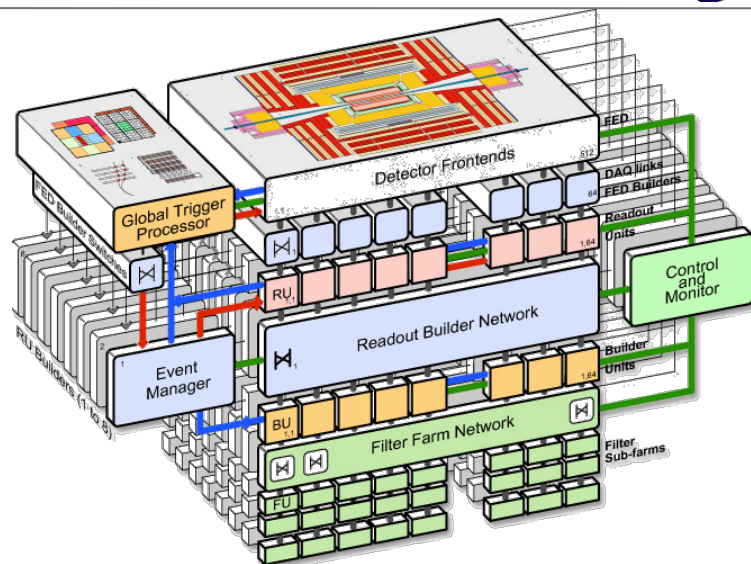
— Event data
 - - - Timing and Fast Control Signals
 — Control and Monitoring data



• Two stage hardware trigger L0 + L1
 • High Level Trigger (HLT) on separate farm

- L1 selects events at 100 kHz and defines regions of interest;
- L2 pulls data from the region of interest and processes the data in a farm of processors L2 accepts data at ~1 kHz;
- Event Filter reads the entire detector (pull), processes the events in a farm and accepts at 100 Hz;



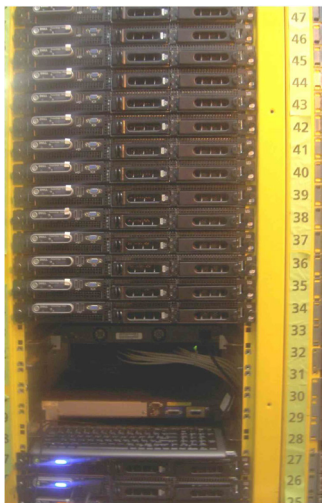


	No. Levels Trigger	Level-0,1,2 Rate (Hz)	Event Size (Byte)	Readout Bandw.(GB/s)	HLT Out MB/s (Event/s)
ALICE	4	Pb-Pb 500 p-p 10 ³	5x10 ⁷ 2x10 ⁶	25	1250 (10 ²) 200 (10 ²)
ATLAS	3	LV-1 10 ⁵ LV-2 3x10 ³	1.5x10 ⁶	4.5	300 (2x10 ²)
CMS	2	LV-1 10 ⁵	10 ⁶	100	~1000 (10 ²)
LHCb	2	LV-0 10 ⁶	3.5x10 ⁴	35	70 (2x10 ³)

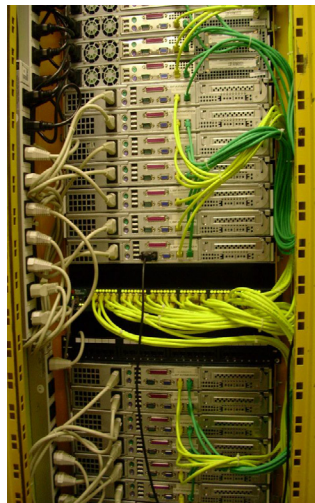
Force10 E1200 equipment

- 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.**

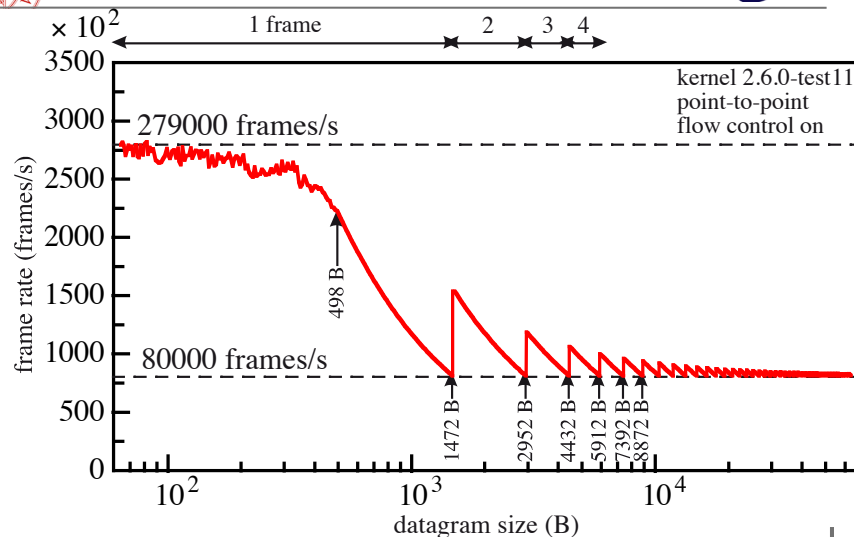
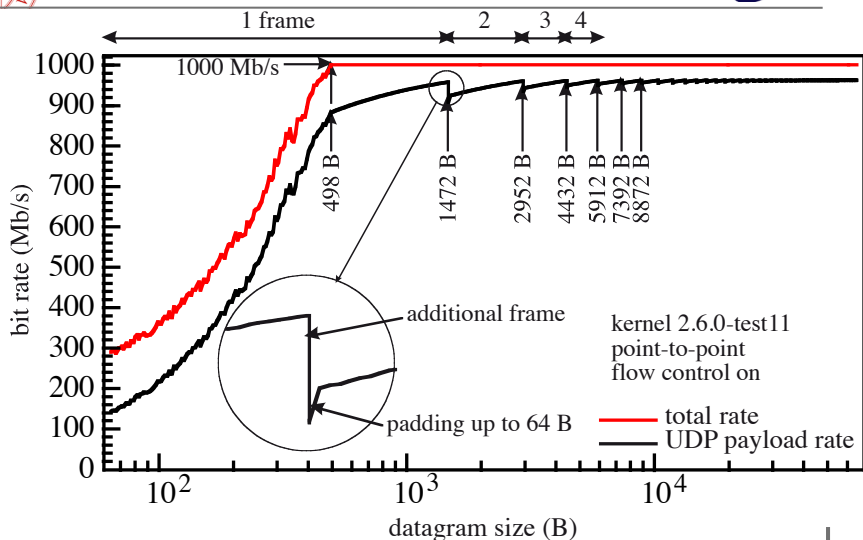




- 1800 1U rack-mounted boxes.
- 2 x 1000Base-T interfaces, to keep separate:
 - Data;
 - ECS (Experimental control system).



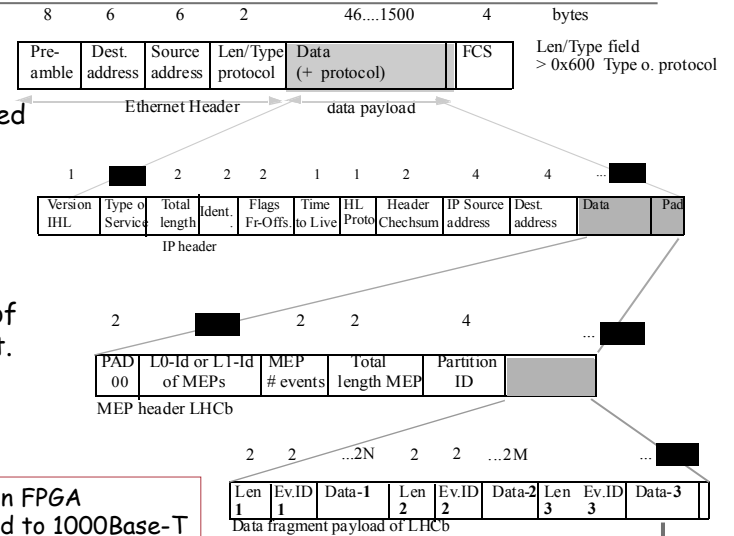
- Why not **TCP**?
 - To **avoid** mechanisms which slow down data transmission (**slow start**, **congestion avoidance**).
 - **Reliability** mechanisms (fast retransmission, fast recovery) are **useless** due to latency constraints:
 - If a fragment of an event is dropped by the network we prefer to **get the next event rather than retransmit** the same event.
- 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.



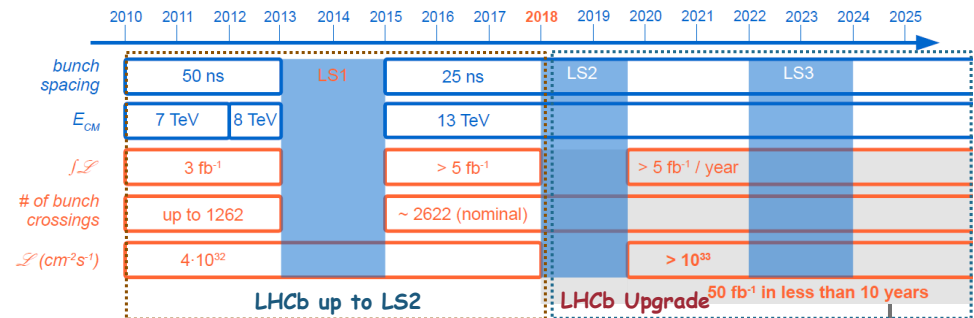
- The **optimal Ethernet payload/overhead ratio**, is achieved when the IP datagram **fills completely the 1500 B Ethernet payload**.
- Moreover the Gigabit Ethernet throughput drops for small frame size.
- However, each Tell-1 board can send only **data-fragments pertaining** to the associated sub-detector element, which usually is much smaller.
- In order to **optimize the payload/overhead ratio**, fragments from multiple (~20) 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.

The LHCb Upgrade

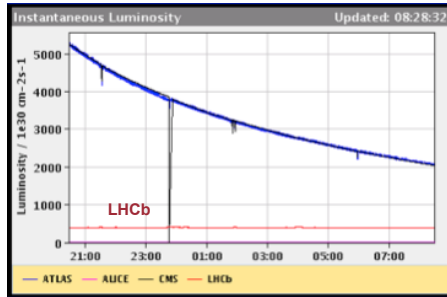
- Custom protocol
- Implemented as a **Linux Kernel module**.
- Optimized for the transport of Multi Event.



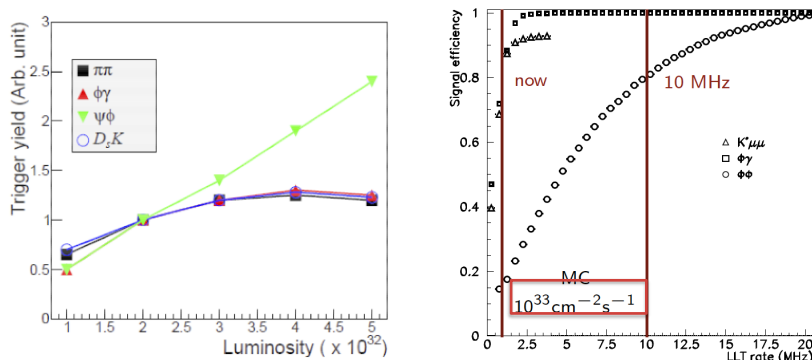
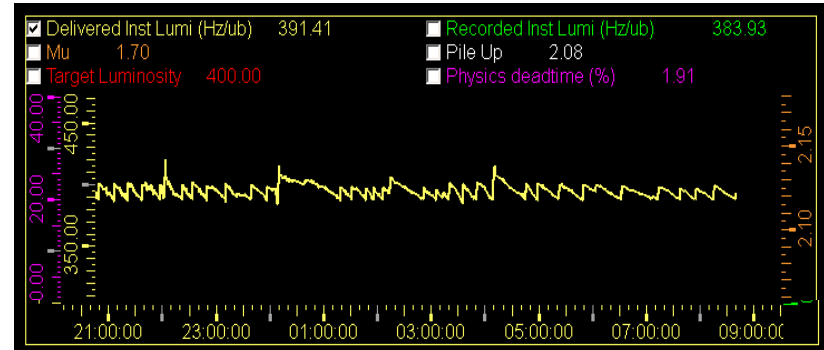
- Shall take place during the **Long Shutdown 2 (LS2)**
 - In **2018**.



- Instantaneous luminosity leveling at $4 \times 10^{32} \text{ cm}^{-2} \text{ s}^{-1}$, $\pm 3\%$ around the target value.
- LHCb was **designed** to operate with a **single collision per bunch crossing**, running at a instantaneous luminosity of $2 \times 10^{32} \text{ cm}^{-2} \text{ s}^{-1}$ (assuming about 2700 circulating bunches):
 - At the time of design there were **worries** about possible **ambiguities** in assigning the B decay vertex to the proper primary vertex among many.
- Soon LHCb realized that running at **higher multiplicities** would have been **possible**. In 2012 we run at $4 \times 10^{32} \text{ cm}^{-2} \text{ s}^{-1}$ with only 1262 colliding bunches:
 - 50 ns separation between bunches while the nominal 25 ns (will available by 2015).
 - 4 times more collisions** per crossing than planned in the design.
 - The average number of visible collisions per bunch crossing in 2012 raised up to $\mu > 2.5$.
 - $\mu \sim 5$ feasible but...

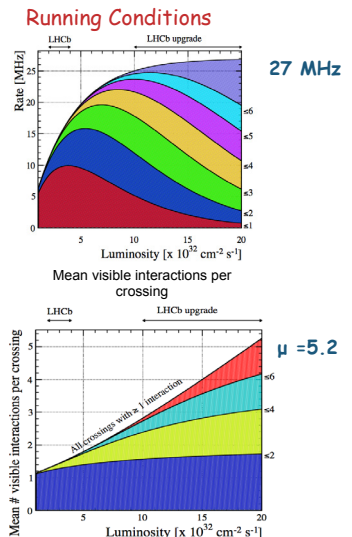


- At present conditions, if we **increase** the luminosity:
 - Trigger yield of hadronic events **saturates**;
 - The p_T **cut** should be **raised** to remain within the 1 MHz LO output rate;
 - There would be **not** a real **gain**.



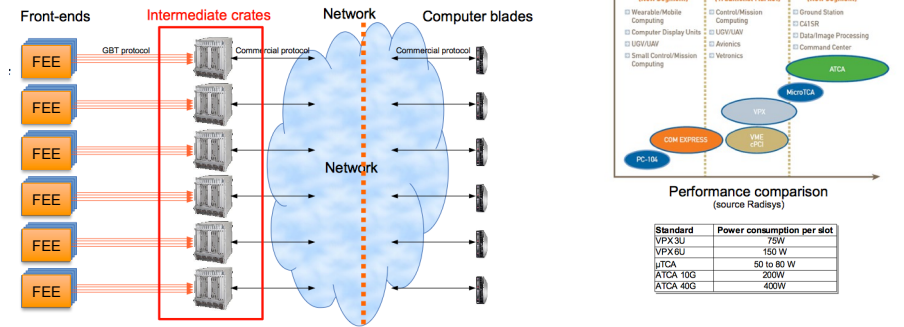
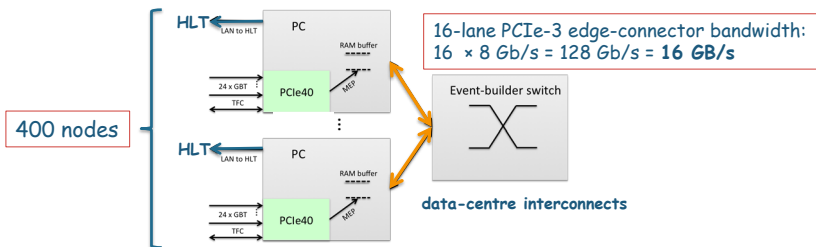
- Due to **the available bandwidth** and the **limited discrimination power** of the hadronic LO trigger, LHCb experiences the **saturation of the trigger yield** on the **hadronic channels** around $4 \times 10^{32} \text{ cm}^{-2} \text{ s}^{-1}$.
- Increasing the first level trigger rate** considerably increases the efficiency on the hadronic channels.

- Readout** the whole detector at **40 MHz**.
- Trigger-less data acquisition system**, running at 40 MHz (~30 MHz are non empty crossings):
 - Use a **(Software) Low Level Trigger** as a **throttle** mechanism, while progressively increasing the power of the event filter farm to run the HLT up to 40 MHz.
- We have foreseen to reach $20 \times 10^{32} \text{ cm}^{-2} \text{ s}^{-1}$ and therefore to prepare the sub-detectors on this purpose:
 - pp interaction rate 27 MHz**.
 - At $20 \times 10^{32} \text{ cm}^{-2} \text{ s}^{-1}$ pile up $\mu = 5.2$
 - Increase the yield in the decays with muons by a **factor 5** and the yield of the hadronic channels by a **factor 10**.
- Collect **50 fb⁻¹** of data over ten years.
 - 8 fb⁻¹** is the integrated luminosity target, to reach by 2018 with the present detector;
 - 3.2 fb⁻¹** collected so far.

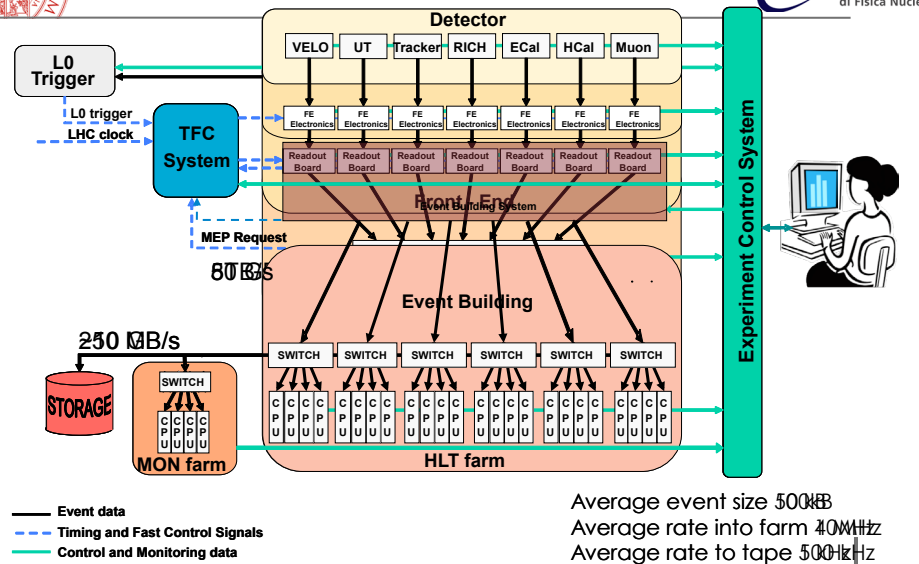


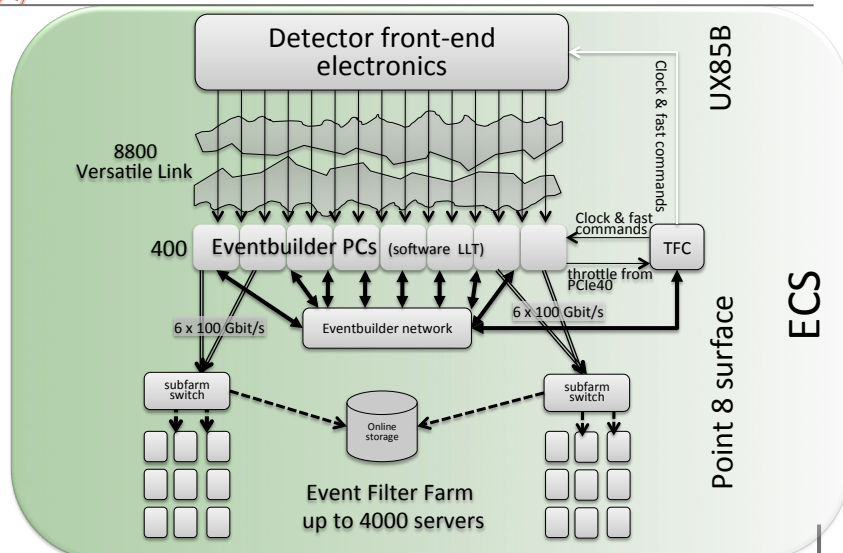
- The **detector front-end electronics** has to be entirely **rebuilt**, because of the current readout speed is limited to 1 MHz.
 - Synchronous readout, no trigger.**
 - No more buffering** in the front-end electronics boards.
 - Zero suppression and data formatting before transmission** to optimize the number of required links.
 - Average event size 100 kB
 - Three times the optical links as currently** to get the required bandwidth, needed to transfer data from the front-end to the read-out boards at 40 MHz.
 - GBT links simplex (DAQ) 9000, GBT duplex (ECS/TFC) 2400
- New HLT farm and network** to be built by exploiting new LAN technologies and powerful many-core processors.
- Rebuild** the current sub-detectors equipped with embedded front-end chips:
 - Silicon strip detectors: VELO, TT, IT
 - RICH photo-detectors: front-end chip inside the HPD.
- Consolidate sub-detectors to let them stand the foreseen luminosity of $20 \times 10^{32} \text{ cm}^{-2} \text{ s}^{-1}$.

- Use **PCIe Generation 3** as communication protocol to **inject data from the FEE directly into the event-builder PC**.
- A **much cheaper** event-builder network:
 - Data-centre interconnects** can be used on the PC:
 - Not realistically implementable on an FPGA (large software stack, lack of soft IP cores,...)
- Moreover PC provides: **huge memory for buffering, OS and libraries**.
- Up to date NIC and drivers available as pluggable modules.



- Intermediate layer** of electronics boards arranged in crates to decouple FEE and PC farm: for buffering and data format conversion.
- The optimal solution with this approach: **ATCA, µTCA** crates, **ATCA** carrier board hosting AMC standard mezzanine boards.
- AMC boards equipped with FPGAs to **de-serialize** the input streams and transmit event-fragments to the farm, using a standard network protocol, using **10 Gb Ethernet**.



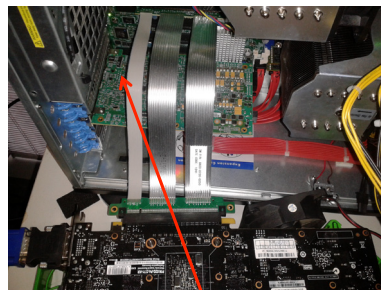


The LHCb Upgrade PCI-e Gen 3 Tests

Electronics Front-End → Data-Centre Interconnect

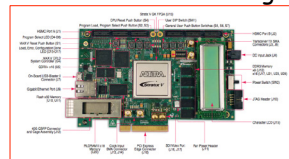
The PCIe-Gen3 DMA Test Setup

- ALTERA evaluation board, Stratix V GX FPGA



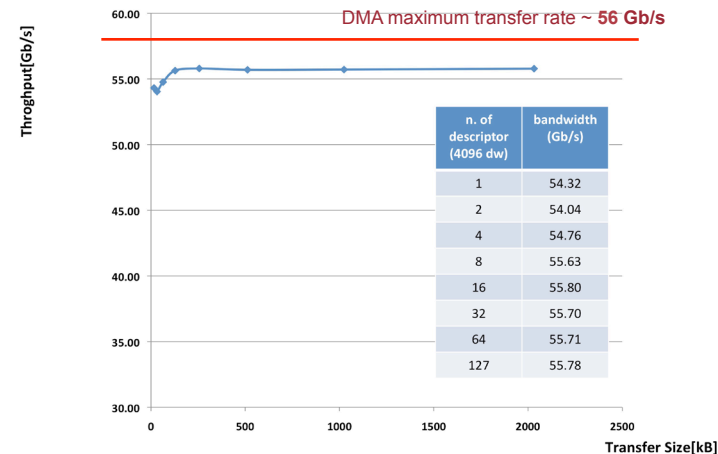
The FPGA provides 8-lane PCIe-3 hard IP blocks and DMA engines.

GPU used to test 16-lane PCIe-3 data transfer between the device and the host memory



DMA PCIe-Gen3 Effective Bandwidth

DMA over 8-lane PCIe-3 hard IP blocks ALTERA Stratix V



- A main FPGA manages the input streams and transmits data to the event-builder PC by using **DMA over PCIe Gen3**.

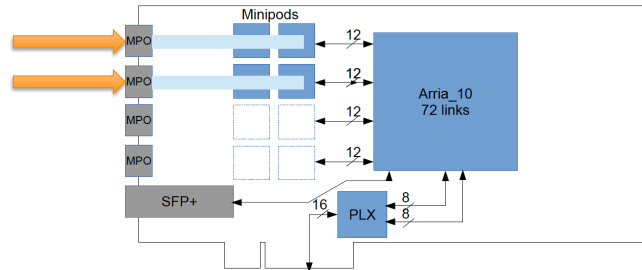
Nominal configuration:

1 bidir link for TFC

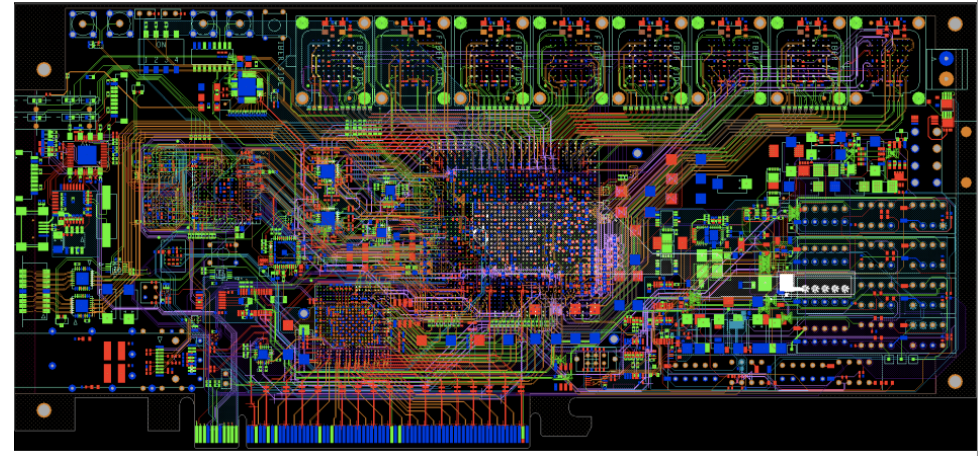
24 GBT inputs → limited by PCIe output bandwidth

- PCIe GEN3 x16 = 110 Gbits/s
- 24 GBT wide bus = 107 Gbits/s

Up to 48 bidir links available on board for low luminosity sub detectors → decrease the costs



PCIe layout



The LHCb Upgrade

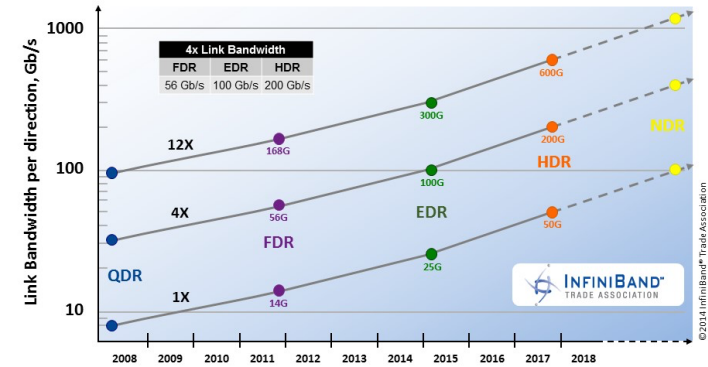
InfiniBand Tests

Event Builder Network

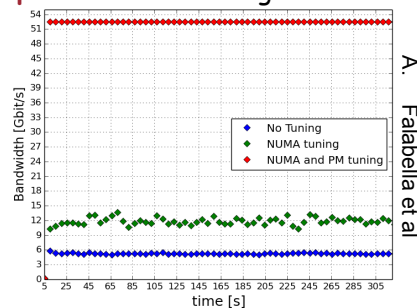
InfiniBand vs Ethernet

- **Guaranteed delivery**. Credit based flow control:
 - Ethernet: Best effort delivery. Any device may drop packets;
- Hardware based **re-transmission**:
 - Relies on TCP/IP to correct any errors;
- **Dropped packets** prevented by **congestion management**:
 - Subject to micro-bursts;
- **Cut through** design with late packet invalidation:
 - Store and forward. Cut-through usually limited to local cluster;
- **RDMA** baked into standard and proven by interoperability testing:
 - Standardization around compatible RDMA NICs only now starting;
 - Need same NICs are both ends;
- **Trunking** is built into the architecture:
 - Trunking is an add-on, multiple standards an extensions;

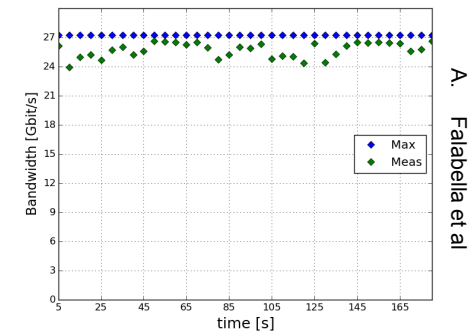
- **All links are used:**
 - Spanning Tree creates idle links;
- **Must use QoS** when sharing with different applications:
 - Now adding congestion management for FCoE but standards still developing;
- Supports **storage** today;
- **Green field design** which applied lessons learnt from previous generation interconnects:
 - Carries legacy from it's origins as a CSMA/CD media;
- Legacy protocol support with IPoIB, SRP, vNICs and vHBAs;
- Provisioned **port cost** for 10 Gb Ethernet approx. **40% higher** than cost of 40 Gb/s InfiniBand.



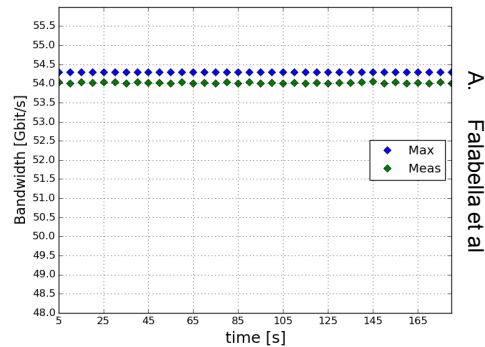
- Performances tests performed at **CNAF**.
- **PCIe Gen 3, 16 lanes** needed:
 - Any previous version of the PCI bus represents a bottleneck for the network traffic;
- Exploiting the best performances required some **tuning**:
 - Disable node interleaving and **bind processes** according to **NUMA** topology;
 - **Disable power saving** modes and CPU frequency selection:
 - **PM** and **frequency switching** are latency sources.



- **Ib QDR** (Quad Data Rate):
 - Point-to-point bandwidth with RDMA write semantic (similar results for send semantic);
 - QLogic : QLE7340, Single port **32 Gbit/s** (QDR);
 - Unidirectional throughput: **27.2 Gbit/s**;
 - Encoding 8b/10b.



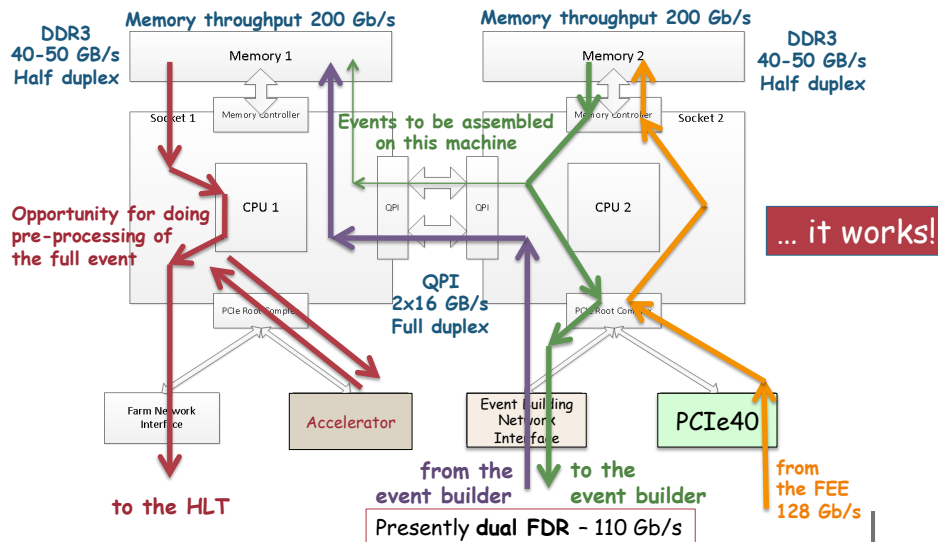
- Ib **FDR** (Fourteen Data Rate):
 - Point-to-point bandwidth with RDMA write semantic (similar results for send semantic);
 - Mellanox : MCB194A-FCAT, Dual port, **56 Gbit/s** (FDR);
 - Unidirectional throughput: **54.3 Gbit/s** (per port);
 - Encoding 64b/66b.



The LHCb Upgrade Event Builder Tests

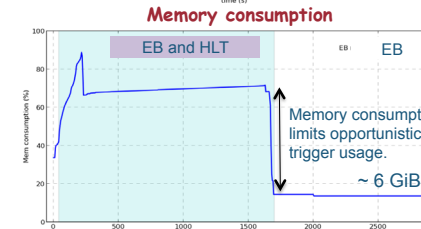
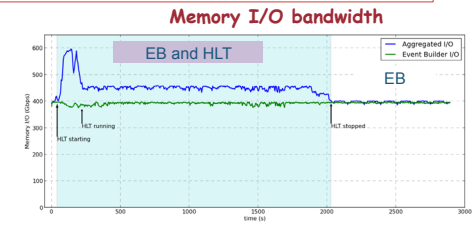
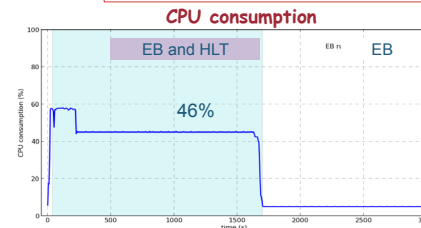
CPU NUMA Architectures, Event Builder Network

Event Builder Fluxes: 400 Gb/s



Event Builder CPU Performance

At about 400 Gb/s more than 80% of the CPU resources are free

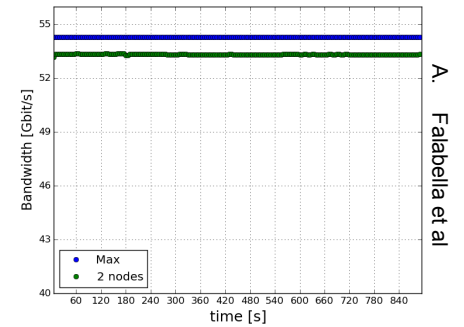


The CPUs used in the test are Intel E5-2670 v2 with a C610 chipset. The servers are equipped with 1866 MHz DDR3 memory in optimal configuration. Hyper-threading has been enabled.

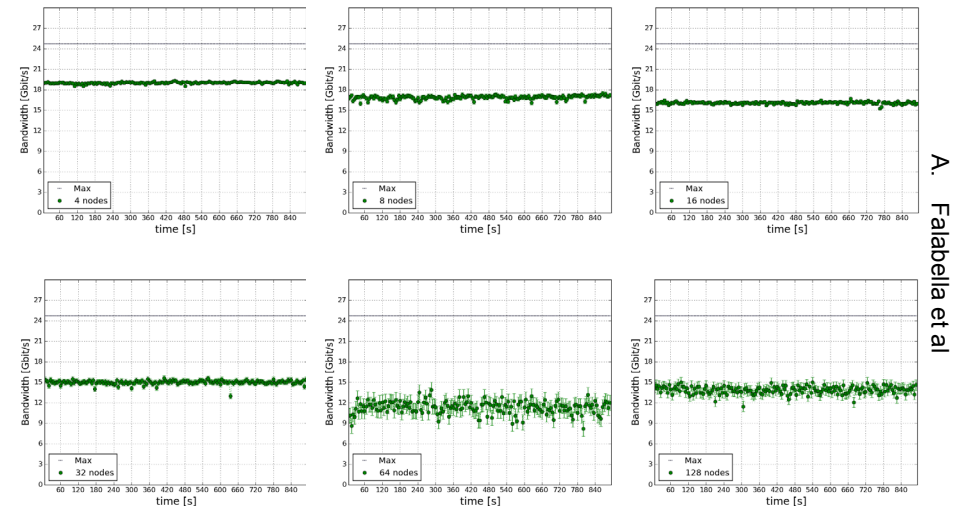
- PC sustains the event building at 100 Gb/s today.
- The Event Builder performs stably at 400 Gb/s
- Aggregated CPU utilization of EB application and trigger 46%
- We currently observe 50% free resources for opportunistic triggering on EB nodes: event builder execution requires about 6 logical core. Additional 18 instances of the HLT software running simultaneously.

- **LHCb-daqpipe** software:
 - Allows to test both **PULL** and **PUSH** protocols;
 - It implements several transport layer implementation: **IB verbs, TCP, UDP**;
- EB software tested on **test beds** of increasing size:
 - At CNAF with **2** Intel Xeon server connected back-to-back;
 - At Cern with **8** Intel Xeon cluster connected through an IB-switch;
 - On **128** nodes at the 512 nodes **Galileo cluster** at the **Cineca**.

- Measured bandwidth as seen by the builder units on **two nodes** equipped with Mellanox **FDR** (max bandwidth 54.3 Gbit/s considering the encoding);
- Duration of the tests: 15 minutes (average value reported).
- Bandwidth measured is on average **53.3 Gbit/s**:
 - **98%** of maximum allowed;
- PM disabled.



- Extensive test on the **CINECA Galileo TIER-1 cluster**.
 - **Nodes**: 516;
 - **Processors**: 2 8-core Intel Haswell 2.40 GHz per node;
 - **RAM**: 128 GB/node, 8 GB/core;
 - **Network**: Infiniband with **4x QDR** switches.
- **Limitations**:
 - Cluster is in production:
 - **Other processes** are polluting the network traffic;
 - **No control on power management** and **frequency switching**;
- The fragment composition is performed **correctly** up to a scale of 128 nodes:
 - Maximum allowed for the cluster batch system.



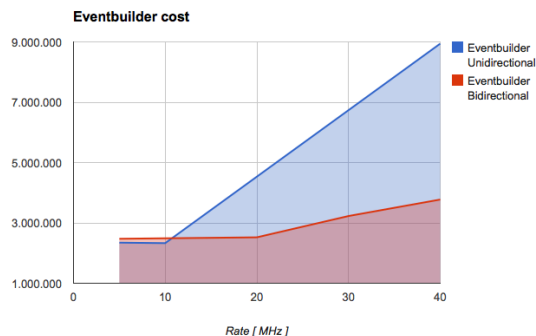
- **Throttle mechanism**, while progressively increasing the power of the EFF to run the HLT up to 40 MHz.
- The **LLT algorithms** can be executed in the event builder PC after the event building.
- Preliminary studies show that the LLT runs in **less than 1 ms**, if the CALO clusters are built in the FEE.
- Assuming 400 servers, 20 LLT processes running per PC, and a factor 8 for the CPU power from the Moore Law, the time budget available turns out to be **safely greater than 1 ms**:

$$\frac{1}{40\text{MHz}} \times 400 \times 20 \times 8 \approx 3.2 \text{ ms}$$

$$\text{processing time budget} = \frac{1}{\text{event rate}} \times \text{nodes} \times \text{cores per node} \times \text{task per node}$$

- Trigger-less system at **40 MHz**:
 - A selective, efficient and adaptable software trigger;
- Average **event size**: **100 kB**;
- Expected **data flux**: **32 Tb/s**;
- Total **HLT trigger process latency**: **~15 ms**:
 - Tracking time budget (VELO + Tracking + PV searches): 50%
 - Tracking finds **99%** of offline tracks with $p_T > 500 \text{ MeV}/c$
- Number of **running trigger process** required: **4×10^5** ;
- Number of **core/CPU** available in 2018: **~200**:
 - Intel tick-tock plan: 7 nm technology available by 2018-19, the number of core accordingly scales as $12 \times (32 \text{ nm} / 7 \text{ nm})^2 = 250$, equivalent 2010 cores.
- Number of **computing nodes** required: **~1000**.

- **Unidirectional**: scaling the present LHCb architecture to 40 MHz, use of intermediate crates, ATCA and AMC board and cables, 10 and 40 GbEthernet. **Cost to operate at 40 MHz: 8.9 MCHF**. The cost due to the ATCA crate has not been included.
- **Bidirectional**: PCIe and InfiniBand proposed approach. **Cost to operate at 40 MHz: 3.8 MCHF**.



- **INFN-Bologna**: Umberto Marconi, Domenico Galli, Vincenzo Vagnoni, Stefano Perazzini et al.;
- **Laboratorio di Elettronica INFN-Bologna**: Ignazio Lax, Gabriele Balbi et al.;
- **INFN-CNAF**: Antonio Falabella, Francesco Giacomini, Matteo Manzali et al.;
- **INFN-Padova**: Marco Bellato, Gianmaria Collazuol et al.;
- **CERN**: Niko Neufeld, Daniel Hugo Cámpora Pérez, Guoming Liu, Adam Otto, Flavio Pisani, et al.;
- Other...



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Prof. Domenico Galli

Dipartimento di Fisica and INFN

domenico.galli@unibo.it

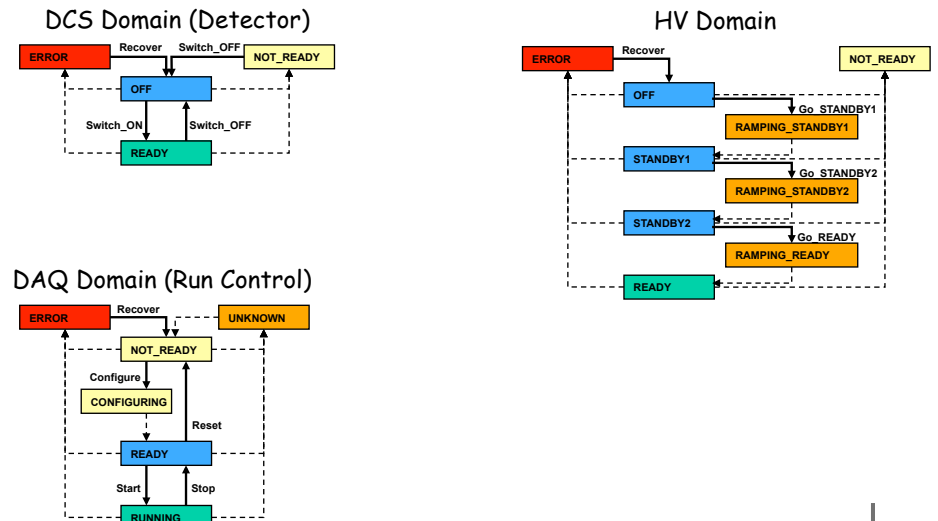
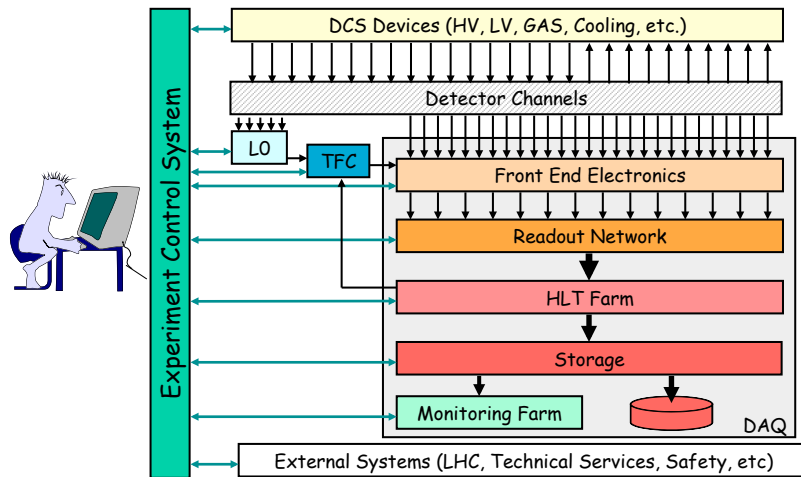
domenico.galli@bo.infn.it

<http://www.unibo.it/docenti/domenico.galli>

<http://lhcbweb2.bo.infn.it/bin/view/GalliDidattica>

- Besides Trigger and DAQ, the Online System include the ECS, in charge of the **Control** and **Monitoring** of:

- **Detector Operations** (ex Slow Controls):
 - GAS, HV, LV, temperatures...
- **Data Acquisition and Trigger**:
 - FE Electronics, Event building, EFF, etc.;
- **Experimental Infrastructure**:
 - Cooling, ventilation, electricity distribution, ... ;
- Interaction with the **outside world**:
 - Magnet, accelerator system, safety system, etc.;





Trigger Rate / Event Size Comparison



	Event Size	L1 Input Rate	L1 output Rate	L2 output Rate	L3 output Rate
KTeV	8 KiB		100 KHz 800 MiB/s	20 KHz 160 MiB/s	2 KHz 7 MiB/s
CDF	270 KiB		50 KHz 13 GiB/s	300Hz 80 MiB/s	80 Hz 23 MiB/s
DØ	250 KiB		10 KHz 2.5 GiB/s	1 KHz 250 MiB/s	70 Hz 13 MiB/s
BaBar	33 KiB (1200 L1)		2 KHz 2.4 GiB/s	None (65 MiB/s)	100 Hz 4 MiB/s
BTeV	50-80 KiB	800 GiB/s	80 KHz 8 GiB/s		4 KHz 200 MiB/s



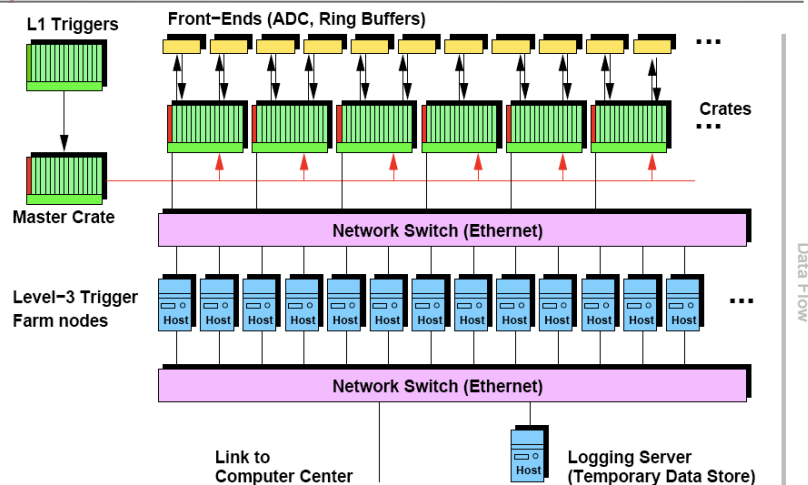
Trigger Rate / Event Size Comparison (II)



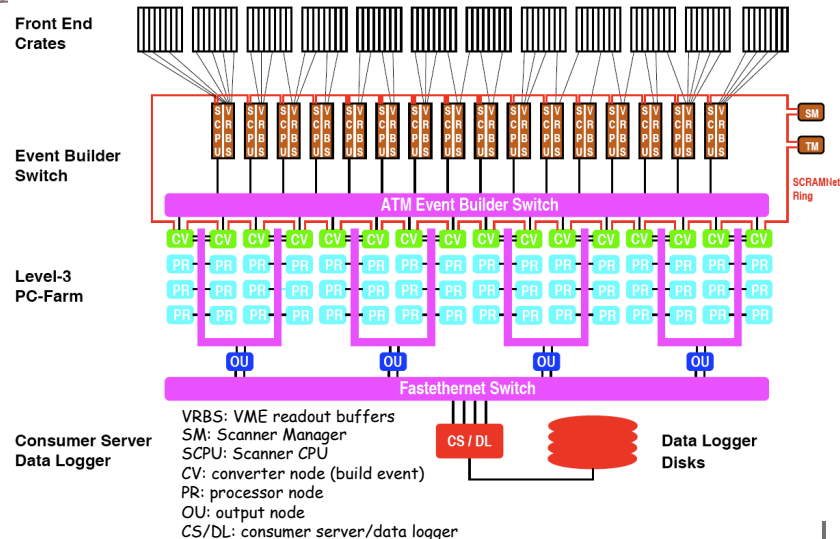
	Event Size	L1 Input Rate	L1 output Rate	L2 output Rate	L3 output Rate
Atlas	1-2 MiB		75 KHz 100 GiB/s	3 KHz 5 GiB/s	200 Hz 300 MiB/s
CMS	1 MiB		100 KHz 100 GiB/s		100 Hz 100 MiB/s
LHCb	35 KiB	1 MHz	1.1 MHz 60 MiB/s		2 KHz 68 MiB/s

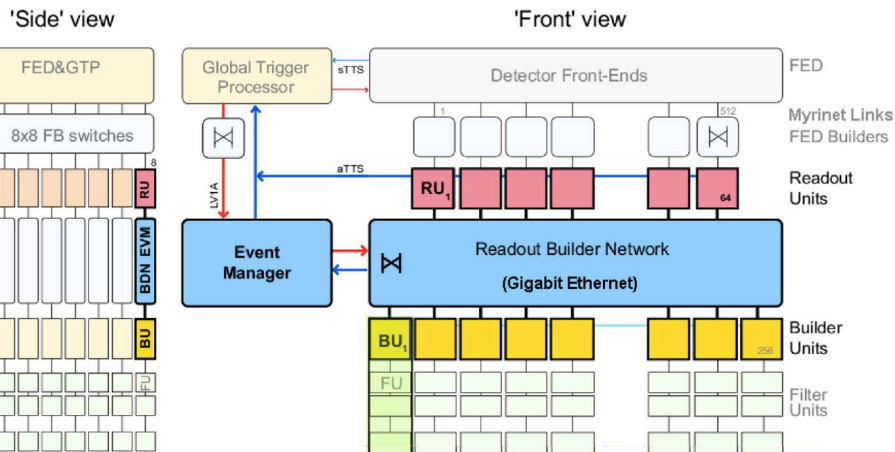
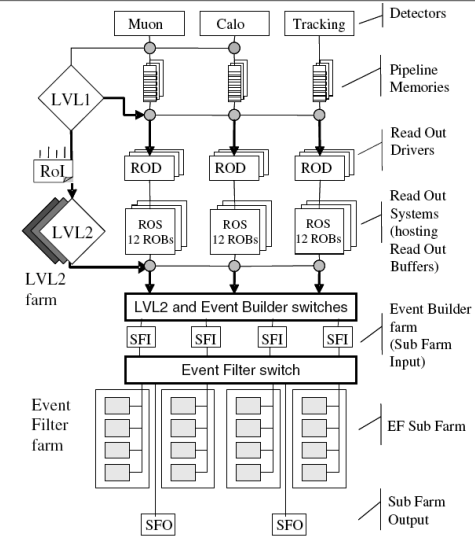
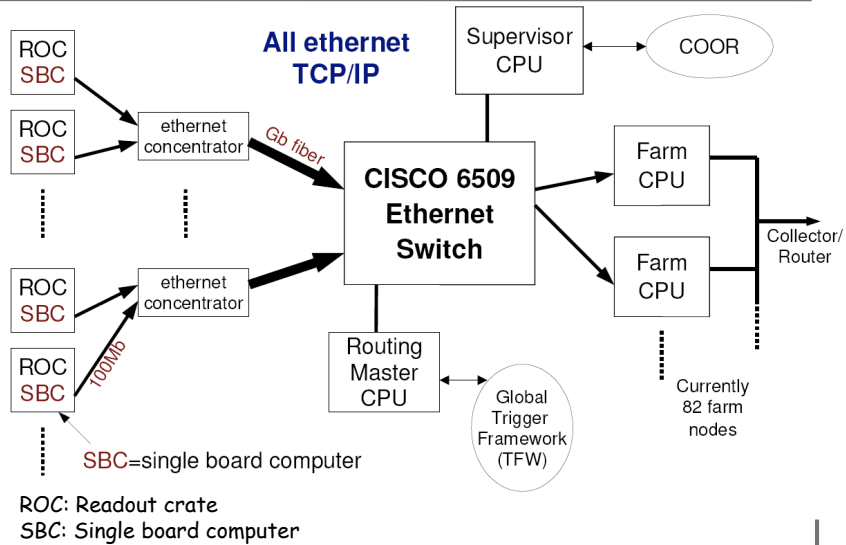


BaBar Trigger and DAQ System



CDF Trigger and DAQ System





1964 Cronin & Fitch CP Violation Experiment:

- K_L^0 mesons produced from 30 GeV protons bombarding Be target.
- Two arm spectrometer with Spark Chambers, Cerenkov counters and Trigger scintillators.
- Spark chambers require fast (~20 ns) HV pulse to develop spark, followed by triggering camera to photograph tracks.
- Trigger on coincidence of Scintillators and Water Cerenkov counters.
- Only one trigger level.
- Deadtime incurred while film advances.

Detector Layout of $K_L^0 \rightarrow \pi^+ \pi^-$ Experiment of Cronin and Fitch (1964)

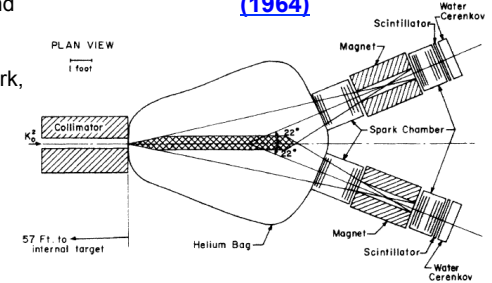


FIG. 1. Plan view of the detector arrangement. Christenson, Cronin, Fitch and Turlay PRL 13, 138 (1964)