Event building and reconstruction at 30 MHz using a CPU farm

Umberto Marconi
INFN Bologna

On behalf of the LHCb Collaboration
WIT2014, Philadelphia
The LHCb experiment

• LHCb is a high-precision experiment devoted to the search for New Physics beyond the Standard Model:
  – By studying CP violation and rare decays in the b and c-quark sectors;
  – Searching for deviations from the SM due to virtual contributions of new heavy particles in loop diagrams.

• Past and running experiments have shown that:
  – Flavour changing processes are consistent with the CKM mechanism;
  – Large sources of flavour symmetry breaking are excluded at the TeV scale;
  – The flavour structure of the NP, if it exists, would be very peculiar at the TeV scale (MFV).
The LHCb Detector
The LHCb detector

- Decay time resolution $\Delta t$: 30-50 fs
- $\Delta p/p = 0.4-0.6\%$
- $\Delta m = 10-20\text{ MeV}/c^2$
- Muon ID: $\varepsilon(\mu/\mu) = 95\%$, mis-ID $\varepsilon(\pi/\mu) \sim 1\%$
- RICH ID $\pi/K$: $\varepsilon(K/K) = 95\%$, mis-ID $\varepsilon(K/\pi) \sim 5\%$
- CALO ID: $\varepsilon(e/e) = 90\%$, mis-ID $\sim 5\%$

B field polarity reversed regularly

"The LHCb Detector at CERN", The LHCb Collaboration, JINST 3, S08005 (2008)
Instantaneous Luminosity

- Instantaneous luminosity leveling at $4. \times 10^{32}$ cm$^{-2}$ s$^{-1}$, ±3% around the target value

**PROTON PHYSICS: STABLE BEAMS**

|---------------|----------|----------------|----------------|

- LHCb was designed to operate with a single collision per bunch crossing, running at an instantaneous luminosity of $2 \times 10^{32}$ cm$^{-2}$ 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 $4 \times 10^{32}$ cm$^{-2}$ 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 crossing in 2012 raised up to $\mu > 2.5$ ($\mu$: average n. of visible interactions)
LHCb events

$B^0 \rightarrow K^- \pi^+$
decay length: 19.282 mm
tau: 5.727 ps
$p_T$: 5081.448 MeV/c
mass: $(5370.595 \pm 15.413)$ MeV/c$^2$
The present LHCb Trigger

- **The Level-0 trigger** based on the signals from ECAL, HCAL and MUON detectors read at 40 MHz, operates on custom electronics, with a maximum output rate limited to 1.1 MHz.
  - Fully pipelined, constant latency of about 4 µs.
  - Bandwidth to the HLT ~ 4 Tb/s, GOL serdes, optical links.
  - High $p_T$ muon (1.4 GeV) or di-muon.
  - High $p_T$ local cluster in HCAL (3.5 GeV) or ECAL (2.5 GeV)
  - 25% of the events are deferred: temporarily stored on disk and processed with the HLT farm during the inter-fills.

- **HLT** is a software trigger.
  - Reconstruct VELO tracks and primary vertices
  - Select events with at least one track matching $p$, $p_T$, impact parameter and track quality cuts.
  - At around 50 kHz performs inclusive or exclusive selections of the events.
  - Full track reconstruction, without particle-identification.
  - Total accept rate to disk for offline analysis is 5 kHz.
LHCb DAQ today

Push-protocol with centralized flow-control

Readout boards: 313 x TELL1

<table>
<thead>
<tr>
<th># links (UTP Cat 6)</th>
<th>~ 3000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event-size (total – zero-suppressed)</td>
<td>65 kB</td>
</tr>
<tr>
<td>Read-out rate</td>
<td>1 MHz</td>
</tr>
<tr>
<td># read-out boards</td>
<td>313</td>
</tr>
<tr>
<td>output bw / read-out board</td>
<td>up to 4 Gbit/s (4 Ethernet links)</td>
</tr>
<tr>
<td># farm-nodes</td>
<td>1500 (up to 2000)</td>
</tr>
<tr>
<td>max. input bw / farm-node</td>
<td>1 Gbit/s</td>
</tr>
<tr>
<td># core-routers</td>
<td>2</td>
</tr>
<tr>
<td># edge routers</td>
<td>56</td>
</tr>
</tbody>
</table>
• Due to the available bandwidth and the limited discrimination power of the hadronic L0 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.
The LHCb upgrade

- Readout the whole detector at 40 MHz.
- **Trigger-less data acquisition system**, running at 40 MHz (~30 MHz are non empty crossings)
  
  - Use a 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 $\mu \approx 5.2$
  - Increase the yield in the decays with muons by a factor five and the yield of the hadronic channels by a factor ten.
- 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 upgrade shall take place during the Long Shutdown 2 (LS2) in 2018.
LHCb upgrade: consequences

• 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. \( \times 10^{32} \text{ cm}^{-2} \text{ s}^{-1} \)
LHCb upgrade: TDRs

• **Letter of Intent for the LHCb Upgrade.**

• **Framework TDR for the LHCb Upgrade : Technical Design Report**

• **LHCb VELO Upgrade Technical Design Report**

• **LHCb PID Upgrade Technical Design Report**
  CERN-LHCC-2013-022 ; LHCB-TDR-014. - 2013.

• **LHCb Tracker Upgrade Technical Design Report**
  CERN-LHCC-2014-001; LHCB-TDR-015. – 2014

• Online and Trigger TDR in preparation.
LHCb DAQ upgrade: First idea

- 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 GbEthernet.
First idea (II)

- Hardware LLT to provide a reduction in the rate of input events to be processed by the Event Filter Farm (EFF), especially in the early phase of data taking after LS2.

LLT performances at $L=2\times10^{33} \text{ cm}^{-2}\text{s}^{-1}$

Hadron Trigger

At 15 MHz, efficiency between 65% and 80% for hadronic channels, 85%-95% for muons
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 because data-centre interconnects can be used on the PC, which are 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 network adapter cards and drivers available as pluggable modules.
DAQ present view (II)

Event builder
High speed network

16-lane PCIe-3
RU/BU
RU: PCIe
BU: 2x FDR
FDR: 2x 56GbE

16-lane PCIe-3

HLT Event filter

sub-farm
SX1036
(10x 56GbE + 52x 10GbE)

10GbE
52x
FU (10GbE)

sub-farm
SX1036
(10x 56GbE + 52x 10GbE)

10GbE
52x
FU (10GbE)
PCIe Gen3 based readout

- A main FPGA manages the input streams and transmits data to the event-builder PC by using DMA over PCIe Gen3.
- The readout version of the board uses two de-serializers.
- The same board can be used to clock and control distribution.

24 optical links from the FEE

MPO → Minipods → Arria_10
12 links
72 links

SFP+ → PLX
16 links

16-lane PCIe-3 edge connector to the motherboard of the host PC
The PCIe-3 DMA test setup

- ALTERA development board, Stratix V GX FPGA, model 5SGXEA7K2F40C2N

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

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

DMA maximum transfer rate ~ 56 Gb/s

<table>
<thead>
<tr>
<th>n. of descriptor (4096 dw)</th>
<th>bandwidth (Gb/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>54.32</td>
</tr>
<tr>
<td>2</td>
<td>54.04</td>
</tr>
<tr>
<td>4</td>
<td>54.76</td>
</tr>
<tr>
<td>8</td>
<td>55.63</td>
</tr>
<tr>
<td>16</td>
<td>55.80</td>
</tr>
<tr>
<td>32</td>
<td>55.70</td>
</tr>
<tr>
<td>64</td>
<td>55.71</td>
</tr>
<tr>
<td>127</td>
<td>55.78</td>
</tr>
</tbody>
</table>
Event builder fluxes: 400 Gb/s

Opportunity for doing pre-processing of the full event

Memory throughput 200 Gb/s

DDR3 40-50 GB/s
Half duplex

Socket 1

Memory Controller

CPU 1

Network Interface

QPI 2x16 GB/s
Full duplex

Accelerator

Events to be assembled on this machine

from the FEE 128 Gb/s
to the event builder

to the HLT

Presently it could be a dual FDR – 110 Gb/s

... it works!

WIT2014
EVB performance

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

CPU consumption

Memory I/O bandwidth

- 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.

Memory consumption

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.
Software LLT

• 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 then 1ms:

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

processing time budget = \(\frac{1}{\text{event rate}} \times \text{nodes} \times \text{cores per node} \times \text{task per node}\)
scaling and cost

- **Unidirectional**: scaling the present LHCb architecture to 40 MHz, use of intermediate crates, AMC board and cables, 10 and 40 GbEthernet. **Cost to operate at 40 MHz: 8.9 MHCF** (cost saving due to the not needed ATCA crate not included).

- **Bidirectional**: PCIe and InfiniBand proposed approach. **Cost to operate at 40 MHz: 3.8 MHCF**
LHCb upgrade: HLT farm

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

IBM Power8 CPU, http://www.extremetech.com/computing/181102-ibm-power8-openpower-x86-server-monopoly
Conclusions

- The concept of the LHCb experiment has been definitely proved: exploiting a forward spectrometer at a hadron collider to perform a dedicated experiment for heavy flavour physics.
  - Many world leading results and many more to come with the 3.2 fb\(^{-1}\) full data set collected.
- LHCb plans the upgrade, to be installed in 2018: the upgrade is an essential next step forward for flavour physics.
- The DAQ system we envisage for the upgrade will allow us to feed HLT trigger and exploit its capabilities at 40 MHz.
- TDRs are all ready, but the “Trigger and DAQ”, on the way to come: it is planned for the next June.
Spare material
### LHCb upgrade: statistical sensitivity to key observables

<table>
<thead>
<tr>
<th>Type</th>
<th>Observable</th>
<th>Current precision</th>
<th>LHCb 2018</th>
<th>Upgrade (50 fb⁻¹)</th>
<th>Theory uncertainty</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B^0$ mixing</td>
<td>$2\beta_s(B^0 \rightarrow J/\psi \phi)$</td>
<td>0.10 [139]</td>
<td>0.025</td>
<td>0.008</td>
<td>~0.003</td>
</tr>
<tr>
<td></td>
<td>$2\beta_s(B^0 \rightarrow J/\psi f_0(980))$</td>
<td>0.17 [219]</td>
<td>0.045</td>
<td>0.014</td>
<td>~0.01</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{el}^r$</td>
<td>$6.4 \times 10^{-3}$ [44]</td>
<td>$0.6 \times 10^{-3}$</td>
<td>$0.2 \times 10^{-3}$</td>
<td>$0.03 \times 10^{-3}$</td>
</tr>
<tr>
<td>Gluonic penguins</td>
<td>$2\beta_s^{\text{eff}}(B_s^0 \rightarrow \phi \phi)$</td>
<td>–</td>
<td>0.17</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>$2\beta_s^{\text{eff}}(B_s^0 \rightarrow K^{*0}\bar{K}^{*0})$</td>
<td>–</td>
<td>0.13</td>
<td>0.02</td>
<td>&lt; 0.02</td>
</tr>
<tr>
<td></td>
<td>$2\beta_s^{\text{eff}}(B_s^0 \rightarrow \phi K_s^0)$</td>
<td>0.17 [44]</td>
<td>0.05</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>Right-handed currents</td>
<td>$2\beta_s^{\text{eff}}(B_s^0 \rightarrow \phi \gamma)$</td>
<td>–</td>
<td>0.09</td>
<td>0.02</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td></td>
<td>$\tau^{\text{eff}}(B_s^0 \rightarrow \phi \gamma)/\tau_{B_s^0}$</td>
<td>–</td>
<td>5 %</td>
<td>1 %</td>
<td>0.2 %</td>
</tr>
<tr>
<td>Electroweak penguins</td>
<td>$S_0(B^0 \rightarrow K^{*0}\mu^+\mu^-; 1 &lt; q^2 &lt; 6 \text{ GeV}^2/c^4)$</td>
<td>0.08 [68]</td>
<td>0.025</td>
<td>0.008</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>$s_0 A_{FB}(B^0 \rightarrow K^{*0}\mu^+\mu^-)$</td>
<td>25 % [68]</td>
<td>6 %</td>
<td>2 %</td>
<td>7 %</td>
</tr>
<tr>
<td></td>
<td>$A_1(K\mu^+\mu^-; 1 &lt; q^2 &lt; 6 \text{ GeV}^2/c^4)$</td>
<td>0.25 [77]</td>
<td>0.08</td>
<td>0.025</td>
<td>~0.02</td>
</tr>
<tr>
<td></td>
<td>$B(B^+ \rightarrow \pi^+\mu^+\mu^-)/B(B^+ \rightarrow K^{+}\mu^+\mu^-)$</td>
<td>25 % [86]</td>
<td>8 %</td>
<td>2.5 %</td>
<td>~10 %</td>
</tr>
<tr>
<td>Higgs penguins</td>
<td>$B(B^0 \rightarrow \mu^+\mu^-)$</td>
<td>$1.5 \times 10^{-9}$ [13]</td>
<td>$0.5 \times 10^{-9}$</td>
<td>$0.15 \times 10^{-9}$</td>
<td>$0.3 \times 10^{-9}$</td>
</tr>
<tr>
<td></td>
<td>$B(B^0 \rightarrow \mu^+\mu^-)/B(B^0 \rightarrow \mu^+\mu^-)$</td>
<td>–</td>
<td>~100 %</td>
<td>~35 %</td>
<td>~5 %</td>
</tr>
<tr>
<td>Unitarity triangle angles</td>
<td>$\gamma(B \rightarrow D^{(<em>)}K^{(</em>)})$</td>
<td>~10–12° [252, 266]</td>
<td>4°</td>
<td>0.9°</td>
<td>negligible</td>
</tr>
<tr>
<td></td>
<td>$\gamma(B_s^0 \rightarrow D_s K)$</td>
<td>–</td>
<td>11°</td>
<td>2.0°</td>
<td>negligible</td>
</tr>
<tr>
<td></td>
<td>$\beta(B^0 \rightarrow J/\psi K_s^0)$</td>
<td>0.8° [44]</td>
<td>0.6°</td>
<td>0.2°</td>
<td>negligible</td>
</tr>
<tr>
<td>Charm CP violation</td>
<td>$A_\Gamma$</td>
<td>$2.3 \times 10^{-3}$ [44]</td>
<td>$0.40 \times 10^{-3}$</td>
<td>$0.07 \times 10^{-3}$</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>$\Delta A_{CP}$</td>
<td>$2.1 \times 10^{-3}$ [18]</td>
<td>$0.65 \times 10^{-3}$</td>
<td>$0.12 \times 10^{-3}$</td>
<td>–</td>
</tr>
</tbody>
</table>
# Samples

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ATLAS &amp; CMS</strong></td>
<td>25 fb(^{-1})</td>
<td>100 fb(^{-1})</td>
<td>300 fb(^{-1})</td>
<td>(\rightarrow)</td>
<td>3000 fb(^{-1})</td>
</tr>
<tr>
<td><strong>LHCb</strong></td>
<td>3 fb(^{-1})</td>
<td>8 fb(^{-1})</td>
<td>23 fb(^{-1})</td>
<td>46 fb(^{-1})</td>
<td>100 fb(^{-1})</td>
</tr>
<tr>
<td><strong>Belle II</strong></td>
<td>-</td>
<td>0.5 ab(^{-1})</td>
<td>25 ab(^{-1})</td>
<td>50 ab(^{-1})</td>
<td>-</td>
</tr>
</tbody>
</table>

![Graphs showing data trends](attachment:image.png)

\(\sigma_{\text{theory}} \approx -5\%\)

\(\sigma_{\Phi}\) and \(\sigma_{\Phi}^{\Phi}\) can be quantified with data.

WIT2014 28
**Evolution of LHC luminosity**

- **bunch spacing**: 50 ns (LS1), 25 ns
- **$E_{CM}$**: 7 TeV, 8 TeV, 13 TeV
- **$f \mathcal{L}$**: 3 fb$^{-1}$, > 5 fb$^{-1}$
- **# of bunch crossings**: up to 1262, ~ 2622 (nominal)
- **$\mathcal{L} (cm^2 s^{-1})$**: $4 \cdot 10^{32}$

**LHCb up to 2018 → ~ 8-10 fb$^{-1}$:**
- find or rule-out large sources of flavour symmetry breaking at the TeV scale

**LHCb upgrade → ≥ 50 fb$^{-1}$:**
- increase precision on quark flavour physics observables
- aim at experimental sensitivities comparable to theoretical uncertainties
- reinforce LHCb as a general purpose forward detector

LHCb upgrade 50 fb$^{-1}$ in less than 10 years
Long distance fibres (II)

- 4.8 Gbit/s signal produced on the detector by Versatile Link transmitters.
- VTTx to MiniPod for data acquisition.
- MiniPod to VTRx for control, configuration.

- 144 fibres per cable. A total of 120 such cables.
- 3 patch panels (breakpoints) foreseen: expected attenuation ~ 3dB
Optical fibres studies

Transmission test at 4.8 Gb/s

Figure 3: Schematic of Setup B, used for measuring BER vs. OMA for the DAQ direction.

Measurement of BER vs. receive OMA(*) on different OM3 and OM4 fibres.

(*) optical modulation amplitude

The target value of 3 dB on OM3 using the Versatile Link and MP is reachable.
## DAQ numbers

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GBT links simplex (DAQ)</td>
<td>9000</td>
</tr>
<tr>
<td>GBT duplex (ECS/TFC)</td>
<td>2400</td>
</tr>
<tr>
<td>PCIe40 (DAQ)</td>
<td>400</td>
</tr>
<tr>
<td>SOL40 (ECS/TFC)</td>
<td>66</td>
</tr>
<tr>
<td>Estimated total mean event-size</td>
<td>100 kB</td>
</tr>
<tr>
<td>100G links for event-building network @ 40 MHz</td>
<td>400</td>
</tr>
<tr>
<td>Dual-socket servers for High Level Trigger</td>
<td>2000 – 4000</td>
</tr>
</tbody>
</table>
TPU: HLT assisted tracking

• “A specialized processor for track reconstruction at the LHC crossing rate”. Presented by G. Punzi, INSTR 2014
• https://indico.inp.nsk.su/contributionDisplay.pycontribId=129&sessionId=6&confId=0
• “We have shown with a realistic detector arrangement that it is possible to reconstruct tracks and measure their parameters very well with a “brain inspired” cell-matrix method.”

A “cellular” tracking algorithm

Inspired by mechanism of visual receptive fields [D.H. Hubel, T.N. Wiesel, J. Physiol. 148 (1959) 574].

Implementation

• Use modern, large FPGA devices.
  – Large I/O capabilities: now O(Tb/s) with optical links!
  – Large internal bandwidth – a must!
  – Fully flexible, easy to program and simulate
  – Sleep Moore’s slope, and easy to upgrade
  – Highly reliable, easy to maintain and update
  – Industry's method of choice for complex project with a small number of pieces (CT scanners, high-end radars...)
• We used Altera’s Stratix V
  – Same device used elsewhere in LHCb readout system.
Fast Control

Clock distribution and phase/latency control

- LATENCY
  - Alignment with Bunch crossing (BXID)
- FINE PHASE
  - Alignment with best sampling point

3 types of links to be qualified:
- processors specified (red boxes)
- ultimately, qualification (and adaptation) to be done with real hardware
- Already tested on Strata IV to OR (see JPs)

General considerations and conclusions

From TFC point of view:
- the choice of PCIe does not pose any problems in terms of hardware, functional and firmware implementation (of TFC). ATCA was already demonstrated to be adequate.
  - Additional number of cards in PCIe and additional number of fibers.
  - No crates. Distributed system + scalability at its best.
  - No need to test backplane if going for PCIe. One solution fits all.
  - No need of a motherboard if going for PCIe. Commercial PCs.
  - Output bus is flexible and bigger bandwidth for ECS to FE in PCIe.
  - PCIe seems more flexible in terms of future development and optimizations.
  - Fully 40 MHz streamlined solution (data center-like).

My personal point of view:
- ATCA is a more compact solution, perfect for a triggered system (with LLT)
- PCIe is a more distributed and streamlined solution, perfect for a trigger-less system
  - Technology at the service of development (and efficiency)!
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 10GbEthernet approx.
40% higher than cost of 40Gb/s InfiniBand
Network Topology

- **Fat-Tree** (Constant Bisectional Bandwidth).
  - The fat tree topology maintains identical bandwidth at each level of the network.
  - The Fat Tree topologies do not scale linearly with the cluster size. Cabling and switching become increasingly difficult and expensive as cluster size grows, with very large core switches required for larger clusters.
- In the simple binary tree, the number of links and thus the aggregate bandwidth is reduced by half at each stage of the network.


Note that the links are bidirectional so the notion of upstream and downstream describes the direction of the interconnect topology towards or away from the shortest path to an end node, rather than the actual data flow.
The distance to cover with 850 nm OM multimode optical cables, from underground to the surface, is **300 m**.
The ATCA-AMC solution

ATCA carrier board and the AMC module

First AMC module prototype

AMC module input stage test results

• AMC first prototype based on Altera Stratix V FPGA.
• Input stage emulating the GBT protocol at 4.8 Gb/s over 400m long OM4 optical link.
  – BER < 10^{-16}
  – Attenuation margin better than 10 dB
• Buffering requires additional dedicated external memory.
• Implementation of a network standard protocol to the farm into the FPGA problematic with this approach.
Test of PLX bridge

- Long-term test using GTX690 card / PLX 8747 bridge.
- Zero impact of using a bridge and two independent PCIe targets pushing data into a PC. Consistently around 110 Gb/s over long-term.
- No load balancing issues between the two competing links observed.
- Details at: https://lbonupgrade.cern.ch/wiki/index.php/I/O_performance_of_PC_servers#Upgrade_to_GTX690
Memory architecture

**GT/s = Gigatransfers per second**
**GB/s = Gigabytes per second**

**DDR3 memory channel with up to 3DPC**
- 1866, 1600, 1333, 1066 or 800 MHz
- 14.9, 12.8, 10.6, 8.5 or 6.4 GB/s (half duplex) depending on CPU and DIMM type, DPC value and BIOS setting

**Bank 3**
- if used, max frequency is 1066 MHz

**Bank 2**
- max frequency 1866 MHz

**Bank 1**
- max frequency 1866 MHz

**DIMM 3A**
**DIMM 3B**
**DIMM 3C**
**DIMM 3D**

**DIMM 2A**
**DIMM 2B**
**DIMM 2C**
**DIMM 2D**

**DIMM 1A**
**DIMM 1B**
**DIMM 1C**
**DIMM 1D**

**DIMM 3E**
**DIMM 3F**
**DIMM 3G**
**DIMM 3H**

**DIMM 2E**
**DIMM 2F**
**DIMM 2G**
**DIMM 2H**

**DIMM 1E**
**DIMM 1F**
**DIMM 1G**
**DIMM 1H**

**Channel A**
**Channel B**
**Channel C**
**Channel D**

**Channel E**
**Channel F**
**Channel G**
**Channel H**

**DMI2 connectivity to Intel C600 series chipset**

**PCle GEN3 connectivity up to 40 lanes per CPU allocation depending on server model**

Two bidirectional QPI links, each:
- 8.0 or 7.2 or 6.4 GT/s
- 16.0 or 14.4 or 12.8 GB/s (full duplex) depending on CPU model

Interleaving
InfiniBand provides the Remote DMA mechanism for data transfer through a channel.

- InfiniBand dual FDR adapter card
- InfiniBand switch
- Full-duplex
- 16-lane PCIe-3
- PC motherboard

EDR coming soon ...

FDR is 14.0625 Gbit/s per lane
EDR is 25.78125 Gbit/s per lane.
PC event-builder tests

- Test with a Ivy Bridge Intel dual CPU, 12 cores, PC.
- InfiniBand FDR connections
- OpenMPI-based application