UNIVERSITÀ DI PISA

FPGA arrays for fast real-time reconstruction and simulation

Giovanni Punzi
giovanni.punzi@pi.infn.it
University & INFN-Pisa

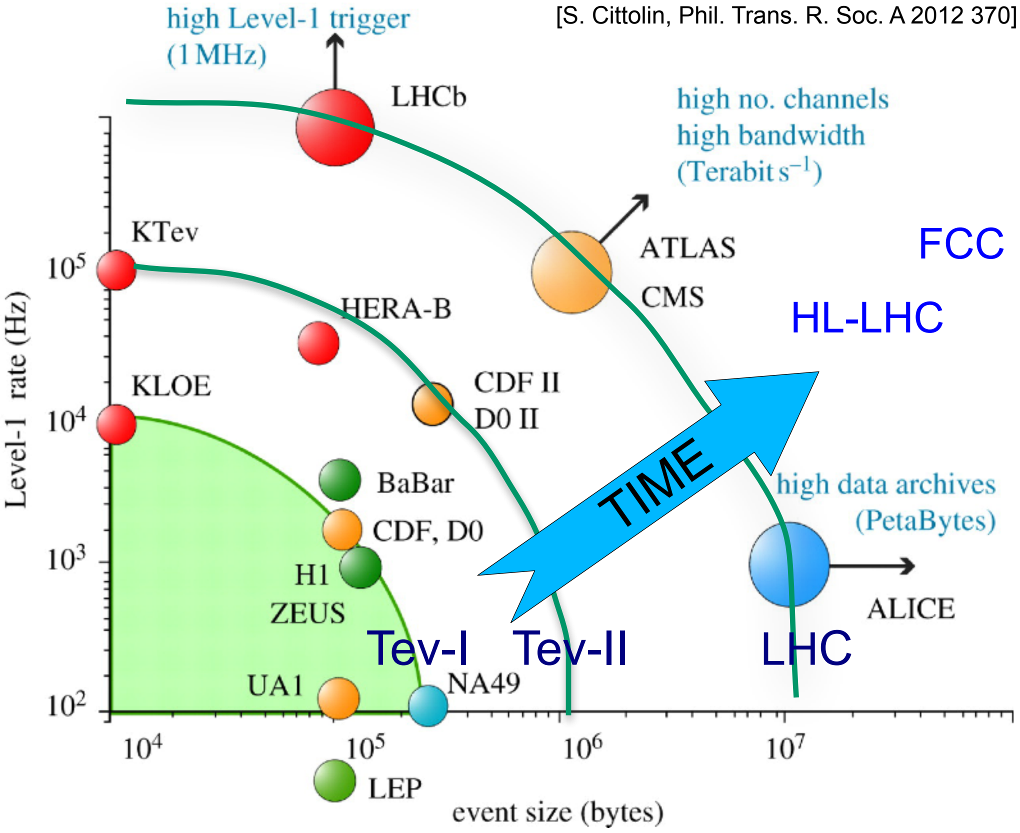
ICSC kickoff meeting
October 13, 2022

Foreword

- Growth of HEP **online data processing** has exploited the progress in consumer CPUs for many years.
- Today's slowdown of Moore's law, and physics demands for more ***precision measurements*** push for new, stronger solutions.
  -> **heterogeneous computing:** customize solutions to problems

-> boost computing power
-> but more need for infrastructure

"use the CN infrastructure as enabling technology"



[S. Cittolin, Phil. Trans. R. Soc. A 2012 370]

This talk: *status and plans of a development line for FPGA-based real-time computing for high-rate experiments, first presented in 2014 at INFN WhatNext* and 1st FHC workshop.
[See this webpage for an up-to-date list of references]

Why FPGAs

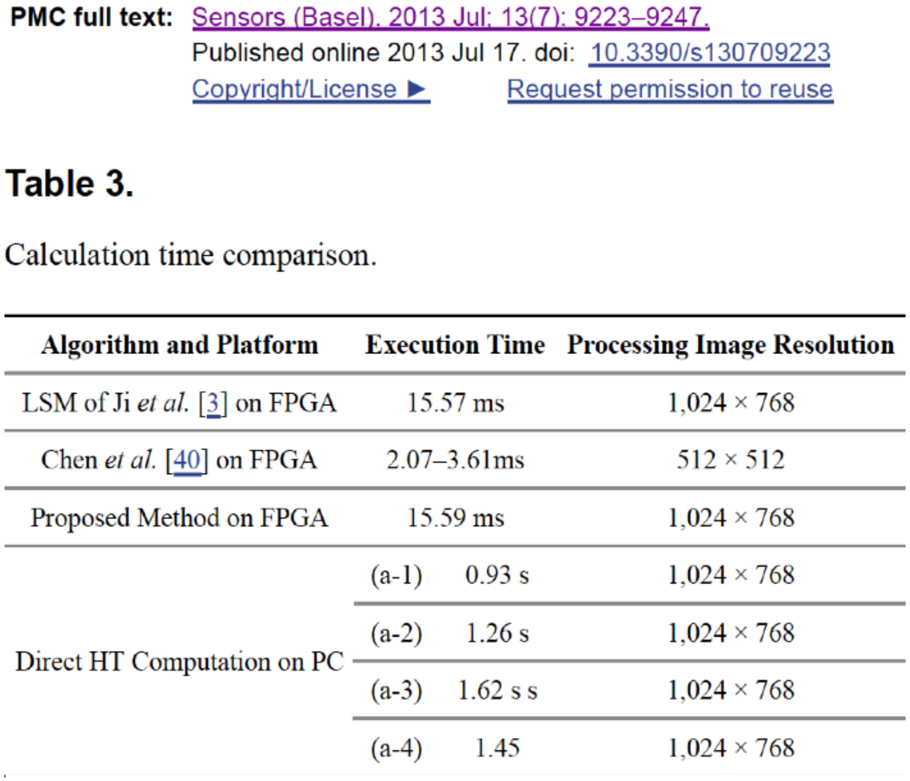
- State-of-the-art FPGA devices are a promising medium:
  - Large I/O capabilities: now O(Tb/s) with optical links.
  - Large internal bandwidth
  - **Low power** consumption → critical in the current era
  - Distributed computing resources: DSP slices, SoC...
  - Highly reliable, easy to maintain and update
  - Fully flexible, easy (!) to program and simulate in software
  - Steep Moore's slope, easily upgradable

→ Industry's method of choice for complex projects for small productions (CT scanners, high-end radars...), low-latency (finance, military)

→ Promising for HEP (small productions, flexibility, real-time)

Already in use for simple tasks: new frontier is <u>powerful applications</u>

What could you gain ? Experience outside HEP

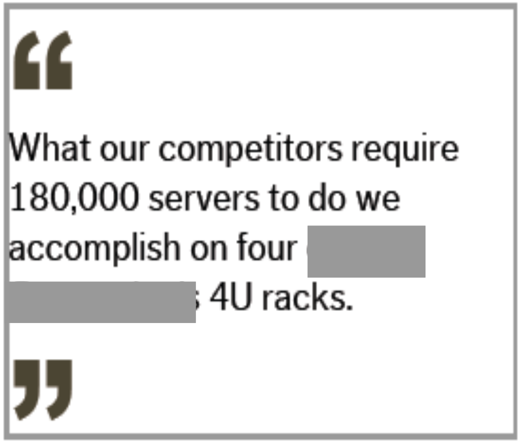


PMC full text: Sensors (Basel). 2013 Jul; 13(7): 9223–9247.
Published online 2013 Jul 17. doi: 10.3390/s130709223
Copyright/License ▶ Request permission to reuse

Table 3.

Calculation time comparison.

Algorithm and Platform	Execution Time	Processing Image Resolution
LSM of Ji *et al.* [3] on FPGA	15.57 ms	1,024 × 768
Chen *et al.* [40] on FPGA	2.07–3.61ms	512 × 512
Proposed Method on FPGA	15.59 ms	1,024 × 768
Direct HT Computation on PC	(a-1) 0.93 s	1,024 × 768
	(a-2) 1.26 s	1,024 × 768
	(a-3) 1.62 s s	1,024 × 768
	(a-4) 1.45	1,024 × 768



> What our competitors require 180,000 servers to do we accomplish on four [redacted] 4U racks.

Speedup factors of 70÷500 regularly obtained in vision, military, finance applications

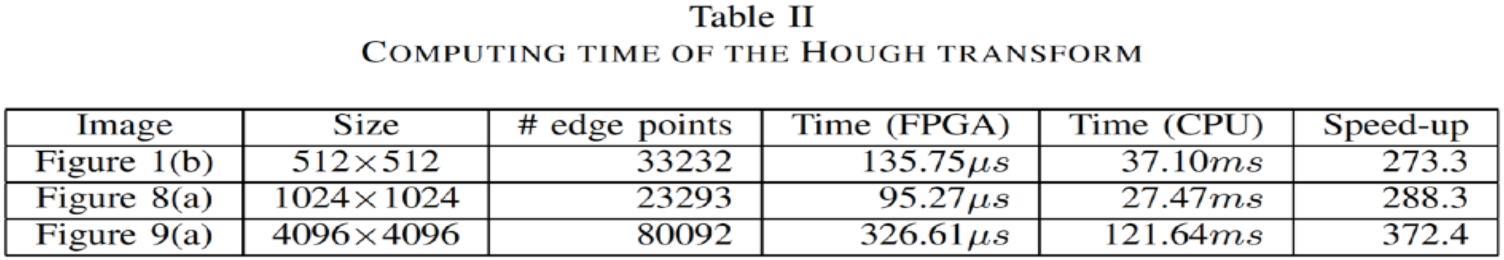


Table II
COMPUTING TIME OF THE HOUGH TRANSFORM

Image	Size	# edge points	Time (FPGA)	Time (CPU)	Speed-up
Figure 1(b)	512×512	33232	135.75μs	37.10ms	273.3
Figure 8(a)	1024×1024	23293	95.27μs	27.47ms	288.3
Figure 9(a)	4096×4096	80092	326.61μs	121.64ms	372.4

-> additional feature of growing importance: "greener" computing

An architecture for tracking in REAL TIME

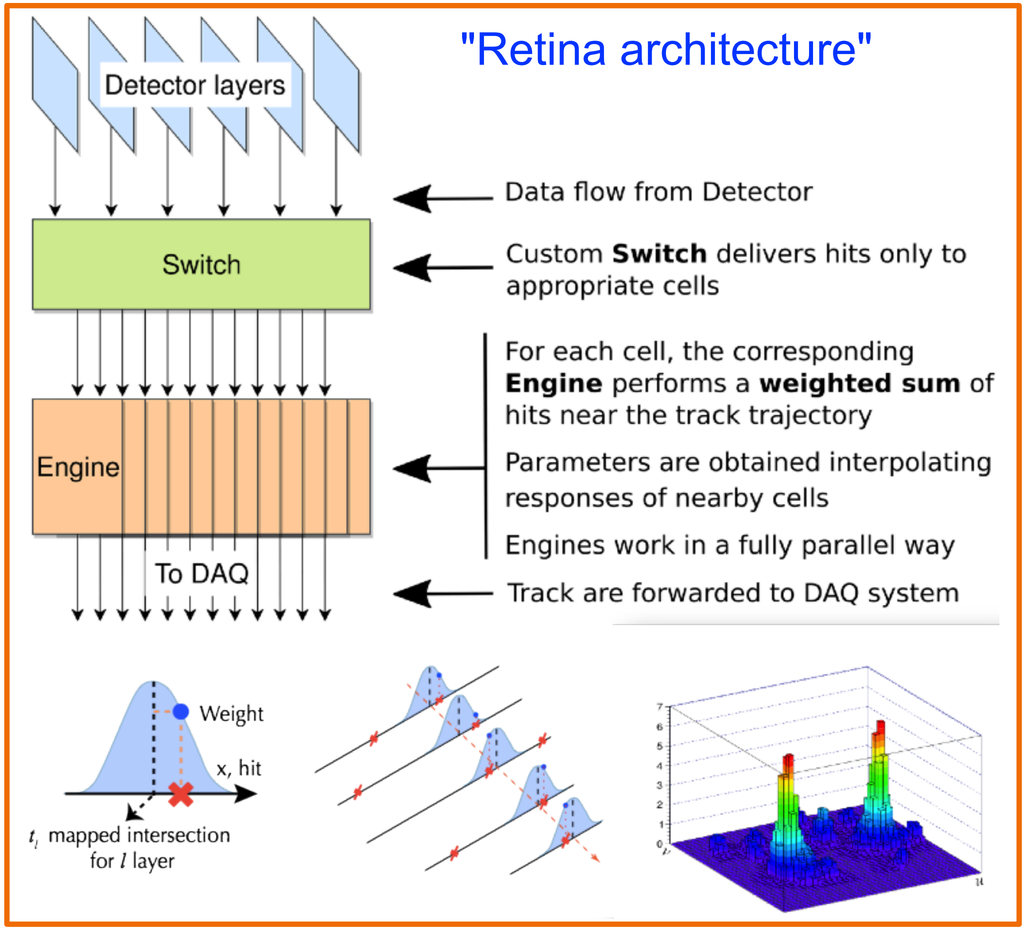
With a highly-parallel data-flow architecture, very fast tracking is possible: [NIM A453(2000),425]

Ingredients:

1) large-bandwidth, fast custom switch (perform similar function to a "*Hough-transform*")

2) an array of cellular processors (engines) working in parallel to perform a vision-like neural algorithm

Everything programmed in FPGA **at low-level (VHDL)** to achieve maximum performance



"Retina architecture"

Data flow from Detector

Custom **Switch** delivers hits only to appropriate cells

For each cell, the corresponding **Engine** performs a **weighted sum** of hits near the track trajectory

Parameters are obtained interpolating responses of nearby cells

Engines work in a fully parallel way

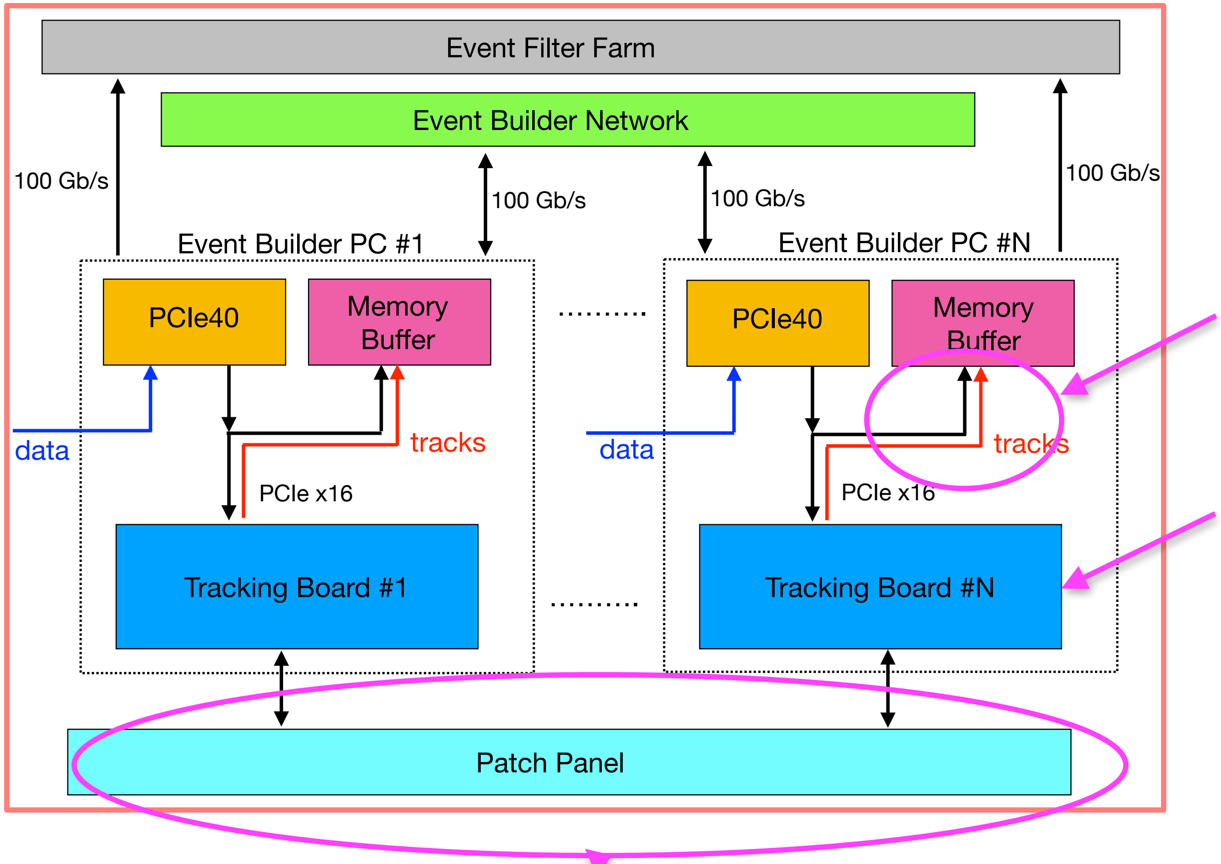
Track are forwarded to DAQ system

INFN R&D project ("RETINA" 2015-2017) demonstrated technical feasibility of 30 MHz tracking already with previous-generation FPGAs [PoS(TWEPP-17) 136].

Currently being developed as a LHCb project.

Demonstrated to produce quality tracking in simulation. implementation ongoing.

From monolithic to an array of (PCIe) cards



Tracking cards ADD extra "RAW Data" to event

Match every readout card (PCIe40) with a tracking card

Architecture for Real-Time data processing, embedded in array of commercial servers:

- Tracking cards communicating via **fast optical network**
- ***Very low latency*** (<1µs), makes all boards appear as a single device to the EB
- Architecture experimentally demonstrated in a testbed setting [F.Lazzari, CDOT2020]

Comparison to Microsoft's CATAPULT architecture

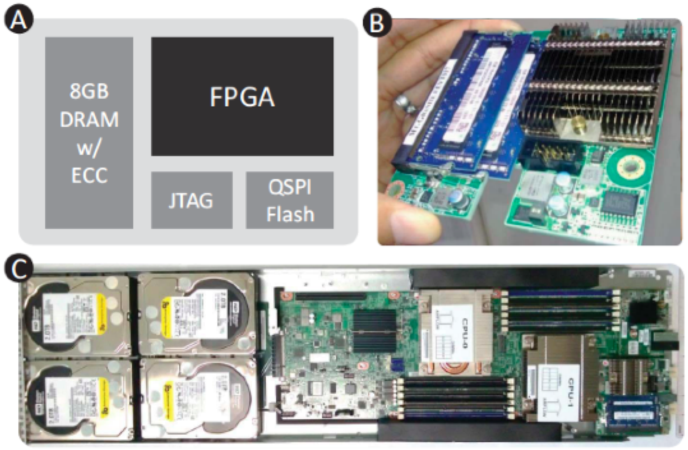


Figure 1: (a) A block diagram of the FPGA board. (b) A picture of the manufactured board. (c) A diagram of the 1 U, half-width server that hosts the FPGA board. The air flows from the left to the right, leaving the FPGA in the exhaust of both CPUs.

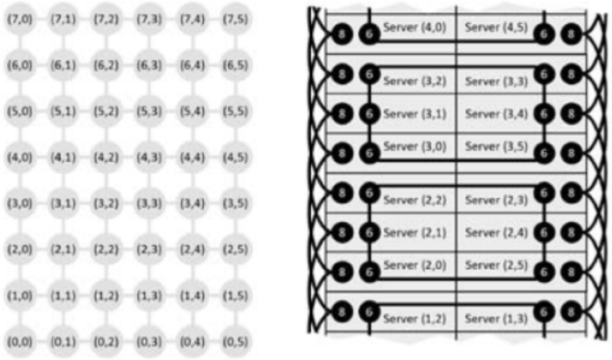


Figure 2: The logical mapping of the torus network, and the physical wiring on a pod of 2 x 24 servers.

Proceeding of the 41st ISCA - ISBN 978-1-4799-4394-4 p. 13-24

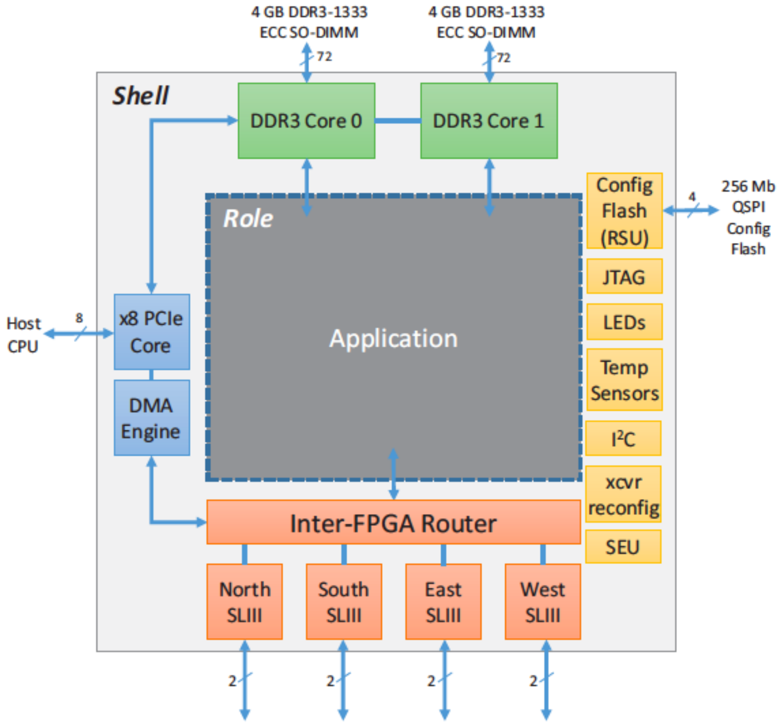
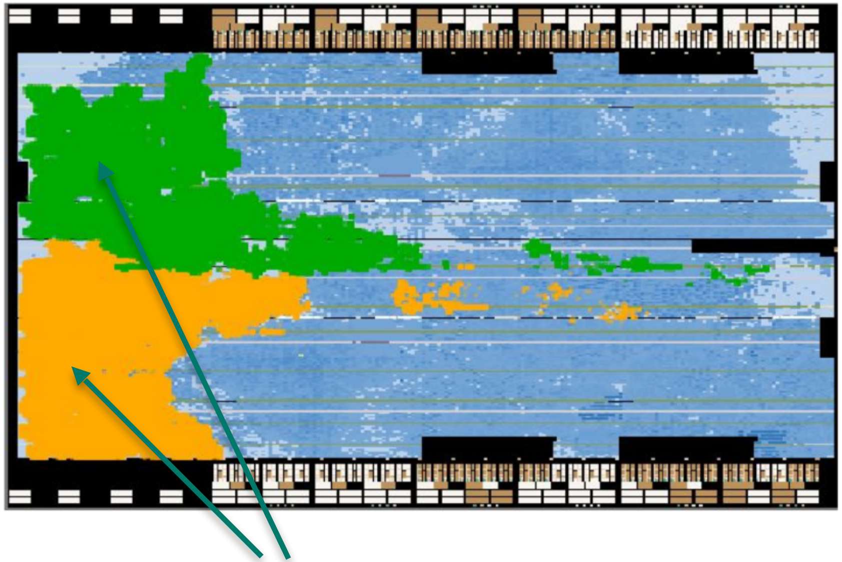


Figure 3: Components of the Shell Architecture.

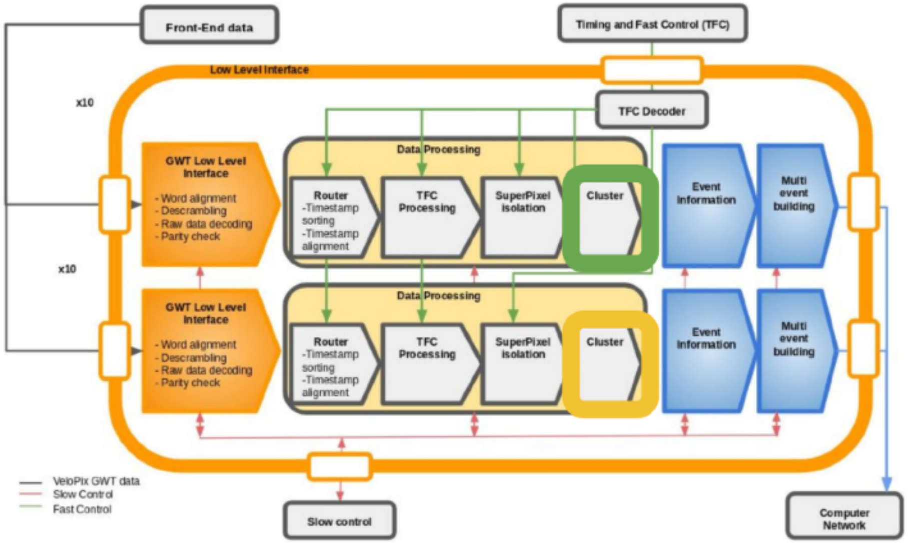
- Interesting structural analogy with independently-developed 'Catapult' system
- Distributed, inter-connected FPGA boards (powering *Bing* in clouds)
- Larger latencies, but similar issues

First real-life application integrated in LHCb Readout

- LHCb groups: INFN-PI + SNS + INFN-FE + INFN-CA
- We designed and embedded our own cluster-finding firmware within the VELO DAQ readout board. Real-time hit-finding in pixel detector @30M events/s. Latency <<1µs
- Saves >11 % of computing power of mighty LHCb-HLT1 farm using little FPGA space
- Same performance for resolution, background etc.
- **Now in final commissioning, physics data being acquired.**
- Comparison of electrical power shows a factor of about **~50x saving**.
- First layer within the bigger project of a real-time embedded tracking system



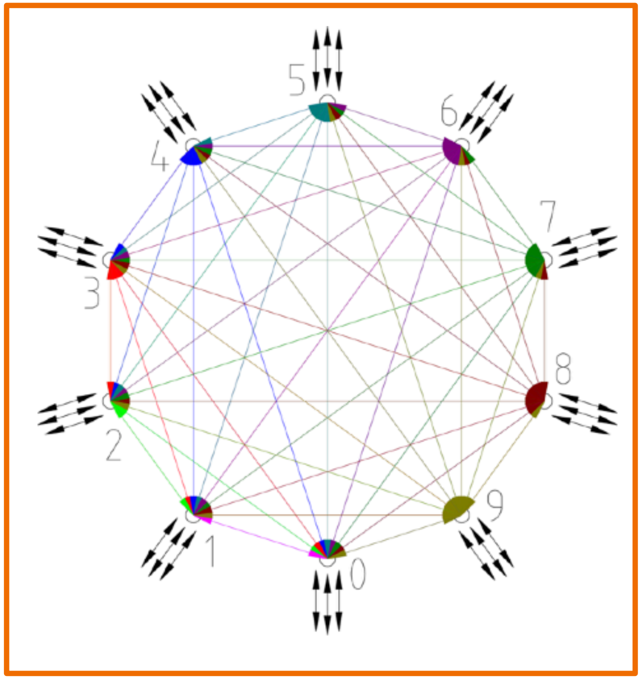
Clustering placement in VELO DAQ FPGA chip

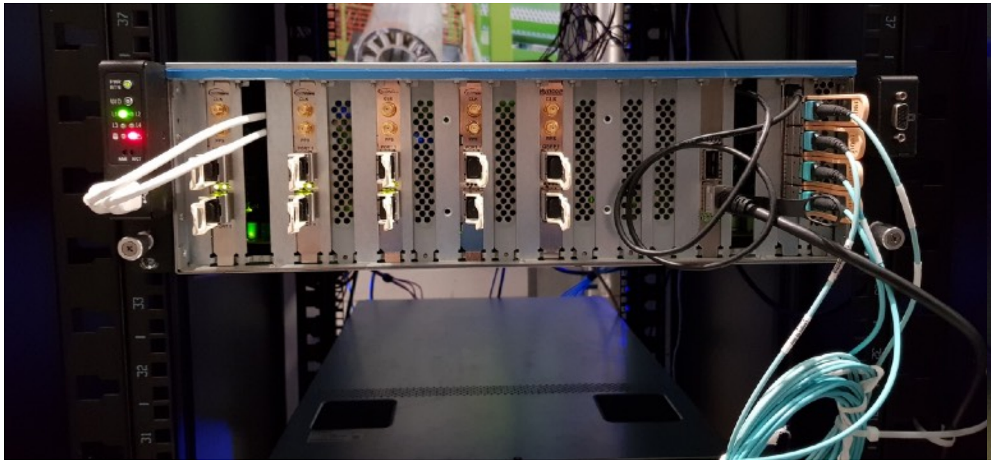


VELO firmware block diagram

Status of tracking demonstrator in LHCb testbed

- First complete system assembled with optical network and tracking engines - capable of 6-layer track reconstruction in the VELO pixel detector of LHCb
  - FPGA boards w/ Intel Stratix-10 FPGA 2.8 MLE, PCIe x16, 0.5 Tb/s each.
  - Each board gets hits from different layers, shares with other FPGAs through optical patch panel (26 Gb/s links)
  - Input from internally stored data in loop (will be PCIe)
  - Currently performing dummy processing, detailed data checking: test at full speed continuously for 2 months (~1500 h) - $7 \cdot 10^{16}$ bit tested (~10 PB) with 0 errors - **BER < $4.2 \cdot 10^{-17}$ (CL = 95%) - extremely reliable.**
  - **Now starting to get real LHCb data in parasitic mode**





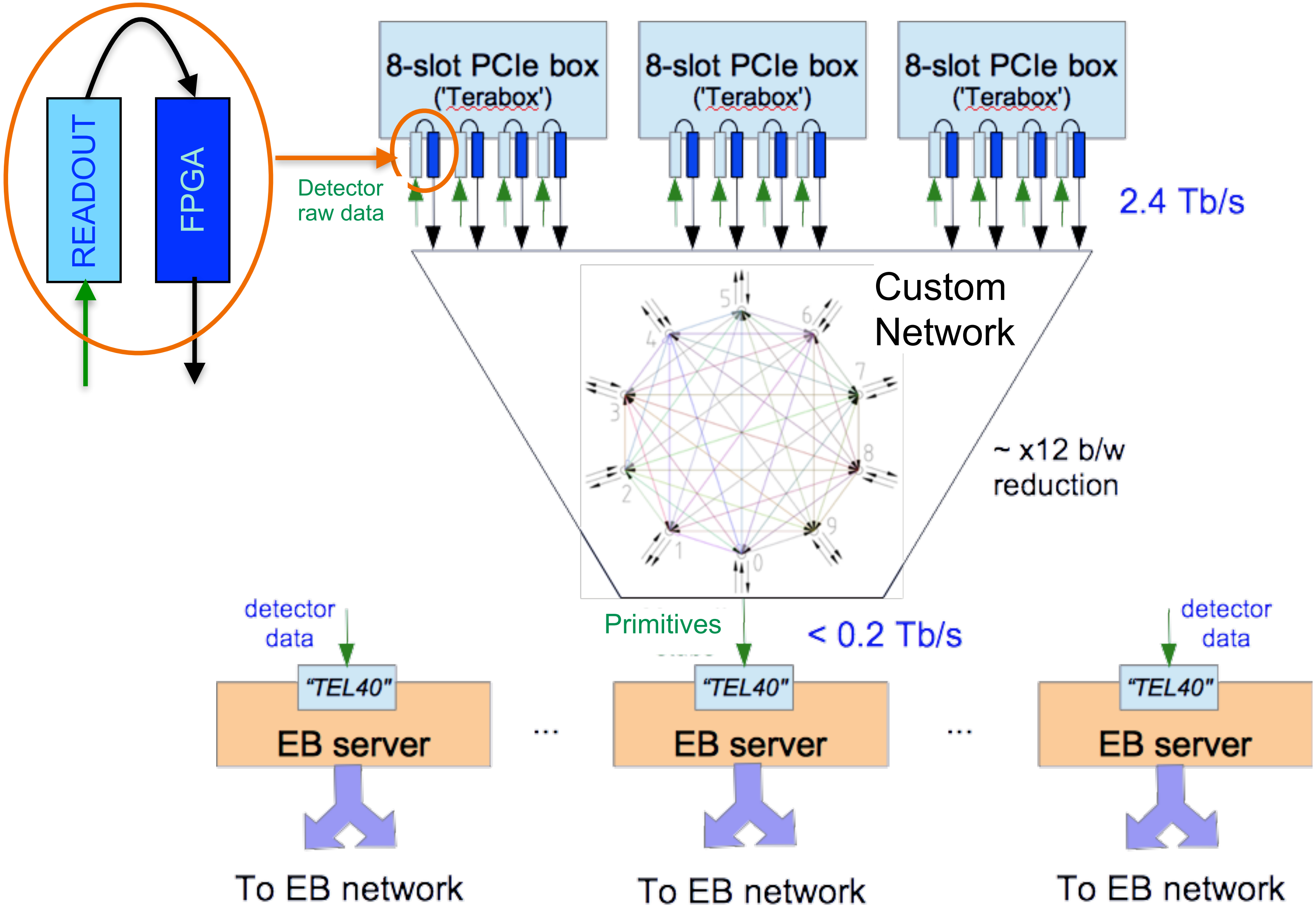
Lesson learned: Development Infrastructure

- Development/test board acquired by INFN under CSN5 RETINA project
- 2 Stratix-V (1MLE, high speed grade) **1.2 Tb/s bidirectional bandwidth** up to 700 MHz clock
- On-board CPU, ample DDR memory, 96 inter-FPGA LVDS connections
- 96 high-speed SerDes I/O (12 Gb/s)
- With optical links, buffer memories, disks, CPU rack etc, for high-rate tests

- Utilizzo sia per girare i **prototipi degli algoritmi** e per testare i **prototipi hardware**
- Disponibilita' di infrastrutture come queste critica per il successo di progetti FPGA
- Costose per singoli progetti, ma ben gestibili come infrastrutture in sharing **-> CN1**

Plan for next 3 years

Embedded real-time reconstruction for future DAQ



- Pre-Evaluate primitives (tracks, calo clusters, muon stubs....) for input to EB
  => order-of-magnitude reduction of the data-flow and processing workload.

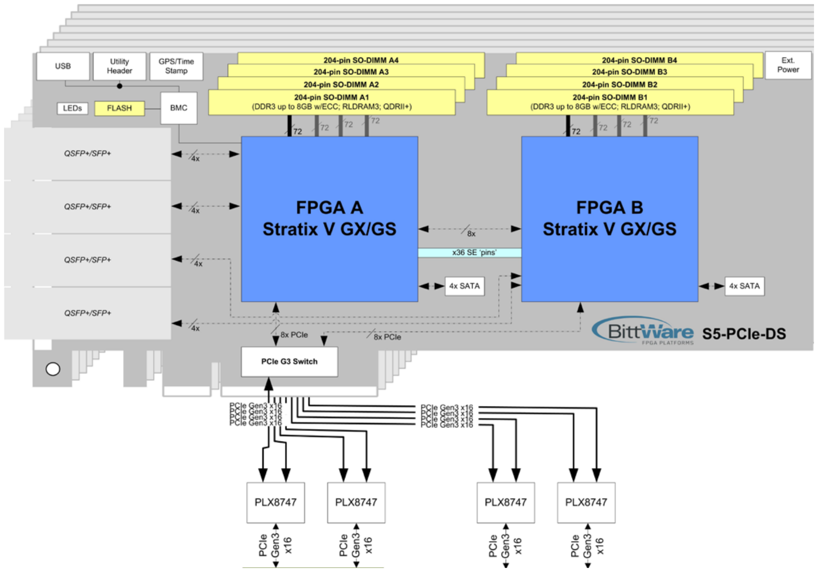
Concrete proposal under development at LHCb, handling ~200 Tb/s

Example of COTS support for FPGA-dense installations

TeraBox High Performance Reconfigurable Computing Platform

Multi-FPGA System for Tera Class High Performance Computing & Network Processing

- 24 TeraFLOPS processing: 16x Intel Arria 10 or Stratix V FPGAs
  - Up to 18 million logic elements (Arria 10 GX)
  - Up to 62,000 multipliers (Stratix V GS)
- 1.28 Terabits/sec I/O
  - 128x 10GbE, 32x 40GbE, or 32x QDR Infiniband
- 6.5 Terabits/sec memory bandwidth
  - Up to 64 banks DDR3-1600 (512 GBytes)
  - DDR4, QDRII+, and RLDRAM3 memory options
- 4U or 5U Rackmount PCIe system (server, industrial, or expansion)



- A Gen-4 version would have 2x bandwidth

The other side of reconstruction: simulation

- Developing and testing a large DAQ system with advanced capability requires an advanced system for **large-scale simulation**
- **Plan to develop FPGA-based simulation as well as reconstruction, based on the same architecture.**
- Move from CPU simulation -> FPGA reconstruction to FPGA simulation -> FPGA reconstruction (two halves of the same system).
- Large, complex system feeding the prototypes of reco system, reproducing the variability of real detector (e.g. alignment changes)
- Allows advanced, realistic testing of HEP real-time reconstruction systems (not only on FPGA, but also, e.g. on GPUs)

- The same system can serve more general purposes, and stimulate the development of know-how in the wider field of "advanced simulation"

How can ICSC support the development of FPGA solutions

“design and test novel computing oriented solutions”

1. CPU time use on a main CPU-based facility
2. A few specialized fast workstations (not from large number of cores)
    1. Simulation of FPGA designs
    2. Place-and-Route of complex designs
3. FPGA arrays (aim at >1 MLE/FPGA)
    1. Preferably in high density format, via PCIe (Gen3 or superior)

       Interfaced via regular CPU servers, or specialized structures
    2. Extensive optical network, saturating all available I/Os

       At least 24Gb/s lines, bandwidths O(Tb/s) each card
4. Ecosystem support: Software licences, specialized personnel...

Questions ?

BACKUP

Dettagli sui parametri di sistema

- N schede FPGA commerciali, dotate di interfaccia PCIe + connettori ottici sul Front Panel
    - Il mercato evolve per cui dipende quando uno le compra, comunque mi aspetterei almeno 1MLE/FPGA, una Gen4 x16 (200Gbs) e 2 QSFP-DD (200-400 Gbs x2)
    - Come modelli io mi sono trovato bene con i chip Intel, ma non ho particolari pregiudizi. Es. prezzi: l'ultima scheda di questo tipo che ho comprato era 6kE, ma ad acquisto singolo, costerebb naturalmente meno in acquisti in blocco.
- Una serie di box rack-mounted per ospitare le suddette - ad esempio, Terabox o similari (costa una modesta aggiunta a quello delle schede)
- Uno switch abbastanza potente, in grado di gestire la parte ottica di *tutte* le schede suddette (via Ethernet/Infiniband ?) (questo non ho sottomano una stima di prezzo, ma non dovrebbe essere una indagine difficile)
- Dei PC dedicati alla compilazione e simulazione del firmware. In linea di principio in un centro di calcolo queste risorse abbondano, ma e' di grande vantaggio avere macchine *dedicate* con grossi quantitativi di RAM, perche' i piazzamenti durano molte ore. (costo O(10-20 kE))
- Da non dimenticare: licenze per il piazzamento e la simulazione, e che siano mantenute costantemente aggiornate per tutta la vita del centro.
- Supporto tecnico: una ragionevole frazione di FTE di personale in grado di manutenere il sistema (sia all'hardware che al software)

Il costo totale e' naturalmente scalabile con la scelta di N - uno puo' pensare da N=20 a N=50 come range ragionevole (che puo' anche essere messo su gradualmente naturalmente, se torna meglio).