Development of a High-Throughput Tracking Processor on FPGA boards

Riccardo Cenci, Federico Lazzari, Pietro Marino, Luciano F. Ristori, Franco Spinella, Michael J. Morello, Giovanni Punzi, Simone Stracka, John Walsh

UNIVERSITÀ DI PISA, INFN PISA, SCUOLA NORMALE SUPERIORE, FERMILAB

Topical Workshop on Electronics for Particle Physics 2017

Sep 12th, 2017

Why Real-Time Reconstruction?

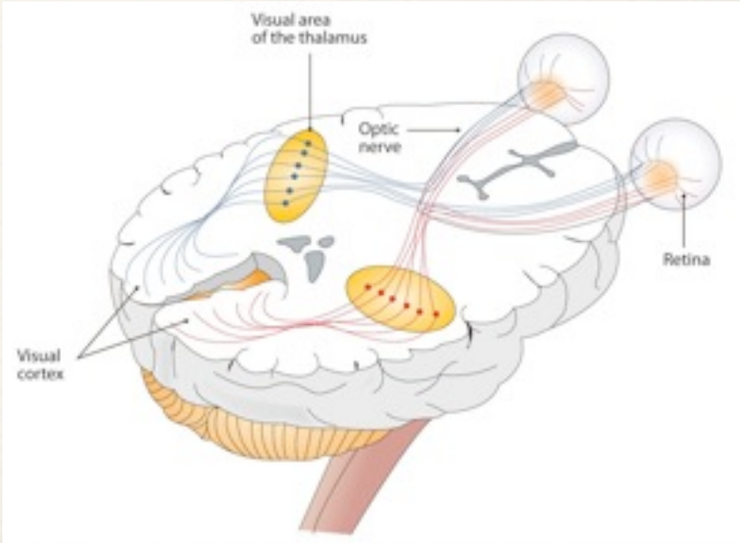
- Experiments at forthcoming **HL-LHC** pose major challenges regarding trigger and data processing:
  - **Physics complexity is growing** (luminosity, high event pile-up, higher precision measurements) and data scales by more than one order of magnitude
  - **Ready-to-go advancements in electronics technology are slowing down** (compare Moore's law and clock speed increase to multi-core CPU's that require special software)
- **Real-time reconstruction of charged particle trajectories at lower trigger level** can help selecting events and reducing data size, but not easy due large combinatorial problems, that require higher parallelization to be solved within typical latencies
  - Customized architecture based on detector features, with larger development time (firmware has to be included here)
  - Commercial solutions are less usable, also due to the required parallelization
- Ideal goal: **"detector-embedded" reconstruction**, that gives also the option of throwing away raw data, saving even more bandwidth and disk space

Pattern Recognition and Mammal Vision

- Real-time reconstruction using patter recognition has been done before, but matching HL-LHC requirements (factor 80 above today's systems)

Name	Technology	Experiment	Year	Event Rate	Clock	Cycles/event	Latency
XFT	FPGA	CDF-L0	2000	2.5 MHz	200 MHz	80	<4μs
SVT	AM	CDF-L2	2000	30 kHz	40 MHz	~1600	<20μs
FTK	AM	ATLAS-L2	2015	100 kHz	~200 MHz	~2000	O(10μs)
?	**?**	**<LHC>-L0**	**~2020**	**40 MHz**	**~1GHz**	**~25**	**few μs**
Vision	**Neural**	**Brain**	**old**	**~40 Hz**	**~1 kHz**	**~25**	**<100 ms**

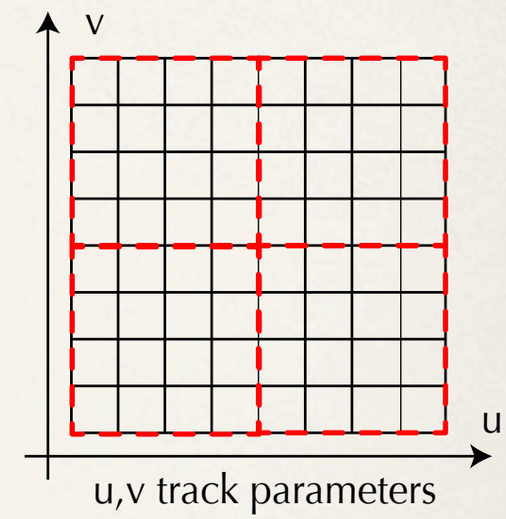
- Inspiration from the vision, why can it do this better?
  - Patterns (neurons) fed only with data relevant for them, reducing internal bandwidth
  - Number of patterns is reduced due to interpolation of analog response
- **Our final goal is implement a tracking trigger at a HL-LHC experiment based on this approach**

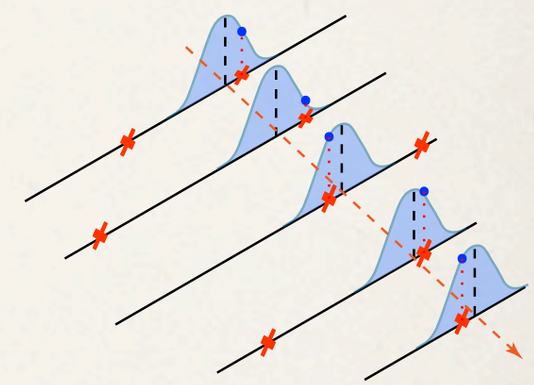


The “Artificial Retina” algorithm

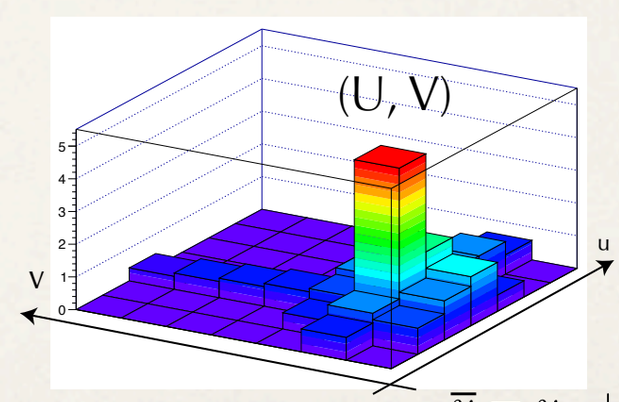
The "Artificial Retina" Algorithm

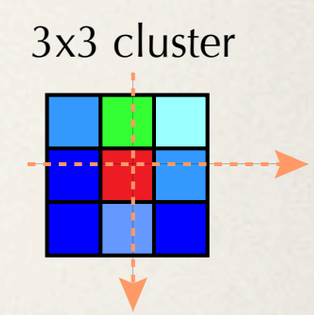
- Original idea in 2000 [NIM A453 (2000) 425-429]: **fully-parallelized pattern matching architecture** inspired by biological analogy with mammal vision
- **Parallelization at two levels**:
  - Hits delivered to a reduced number of cells that process them simultaneously
  - Separate events can be processed if timestamp is associated to hits and cells have multiple registers for accumulating weights
- **Advantages: fully parallelized, intelligent distribution of internal bandwidth, suitable for FPGA implementation (low latency, power efficient)**





- Similar to:
  - Hough transform, but computationally simpler (higher dimensions) and non-integer weights
  - Associative memories for pattern matching, but analog responses using cells interpolation, implying similar or better resolution with lower number of stored patterns

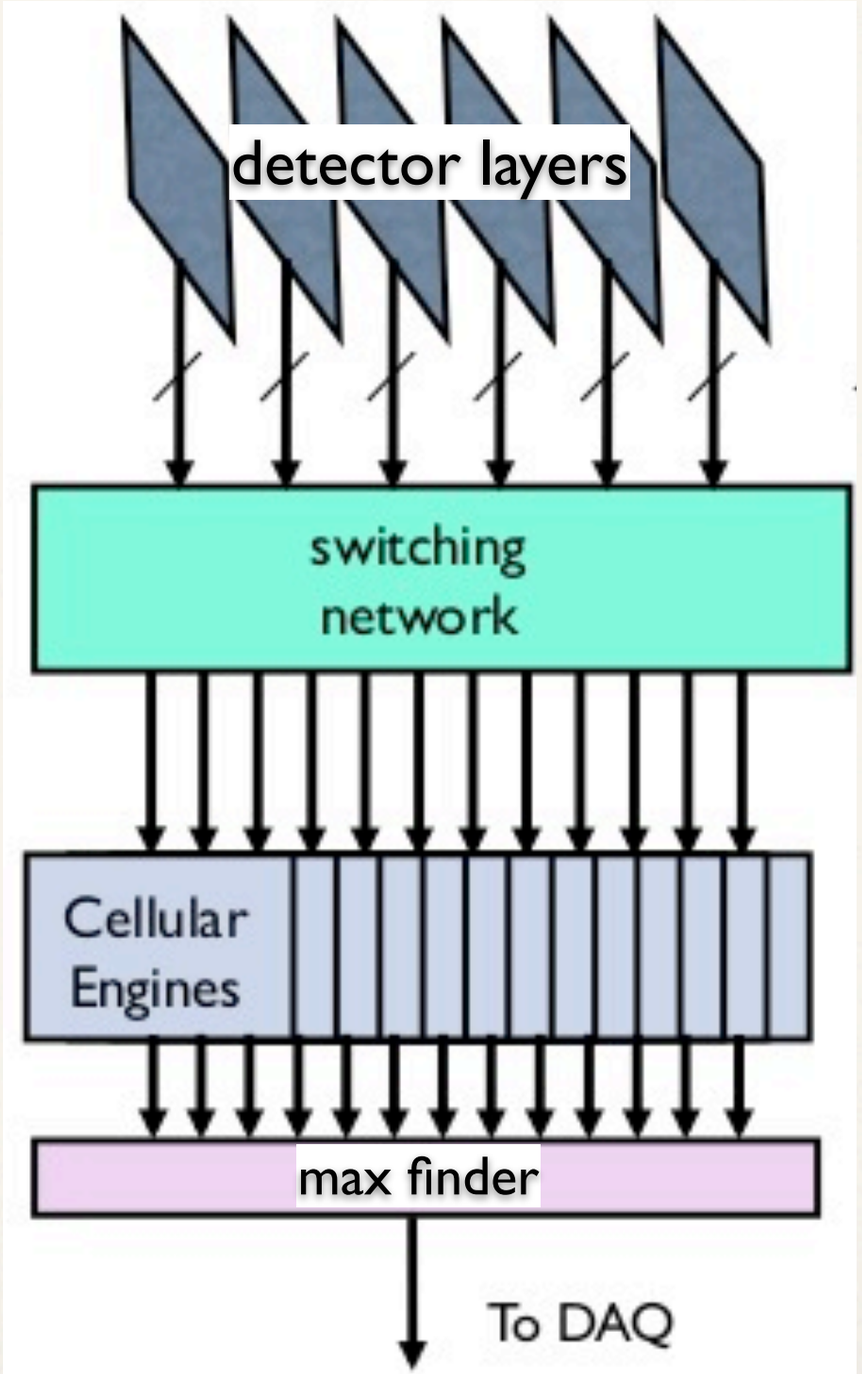


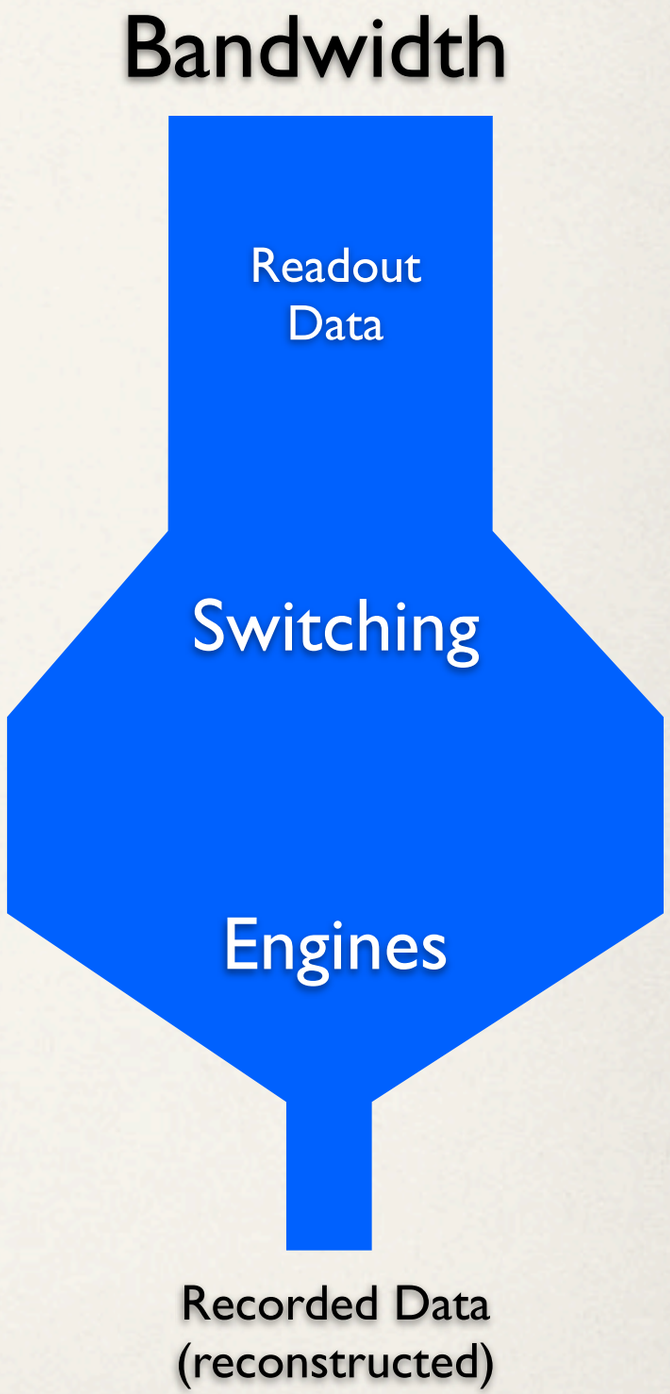


$$\overline{u} = u_0 + \frac{\sum_{ij} iR_{i,j}}{\sum_{ij} R_{i,j}} \quad \begin{matrix} i=U-1 \\ j=V-1 \end{matrix}$$

The Architecture and The Bandwidth

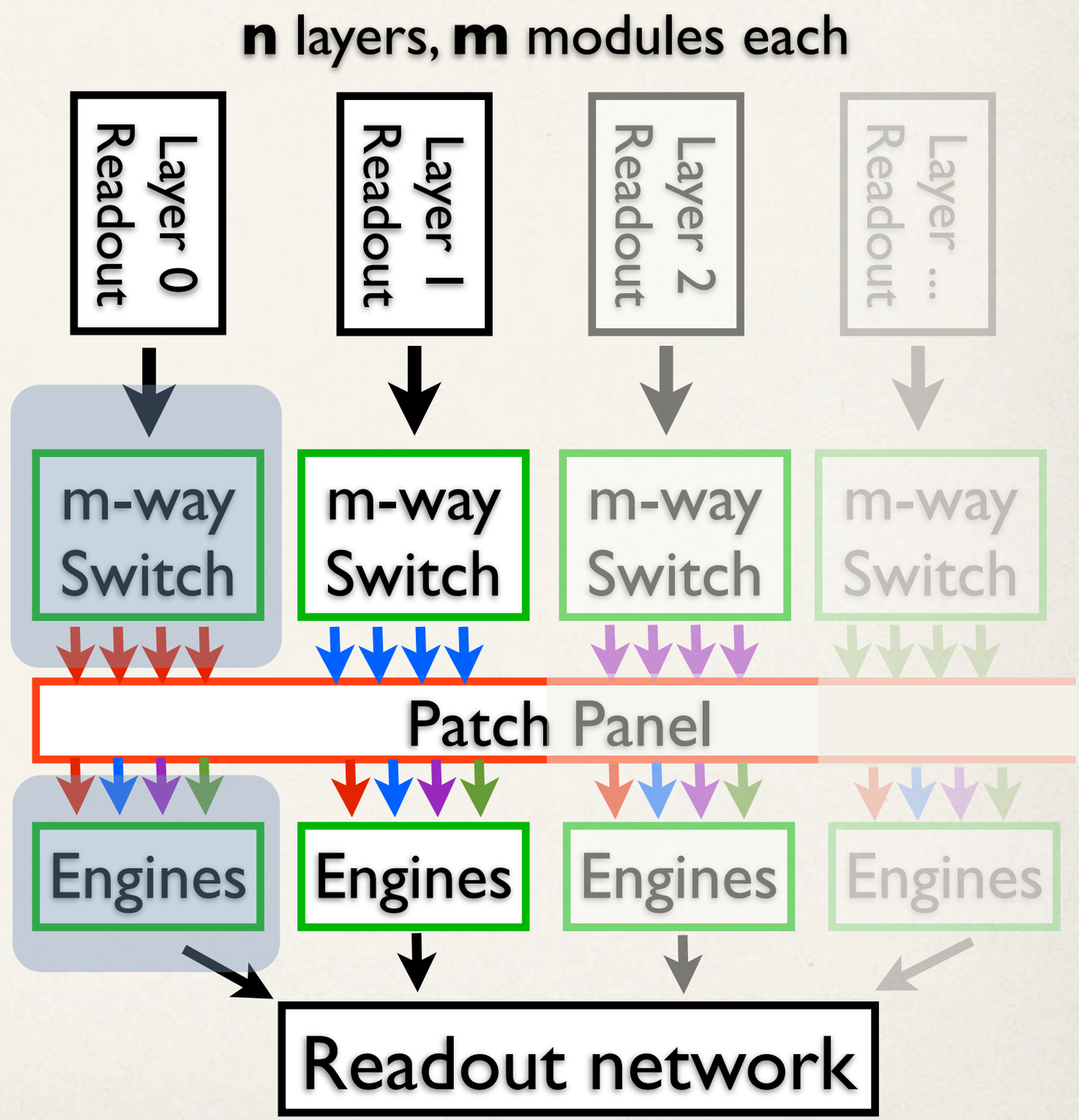
- A **switching network** redirects hits from readout to cell processor units (**engines**), minimizing delivered data
- **Unusual bandwidth profile: it increases after the switching network, and shrinks to a value lower than input back after matched patterns are found**
- Cells grid divided into **blocks**, each implemented on one device with large internal bandwidth (**FPGA**, see below)





System Scalability

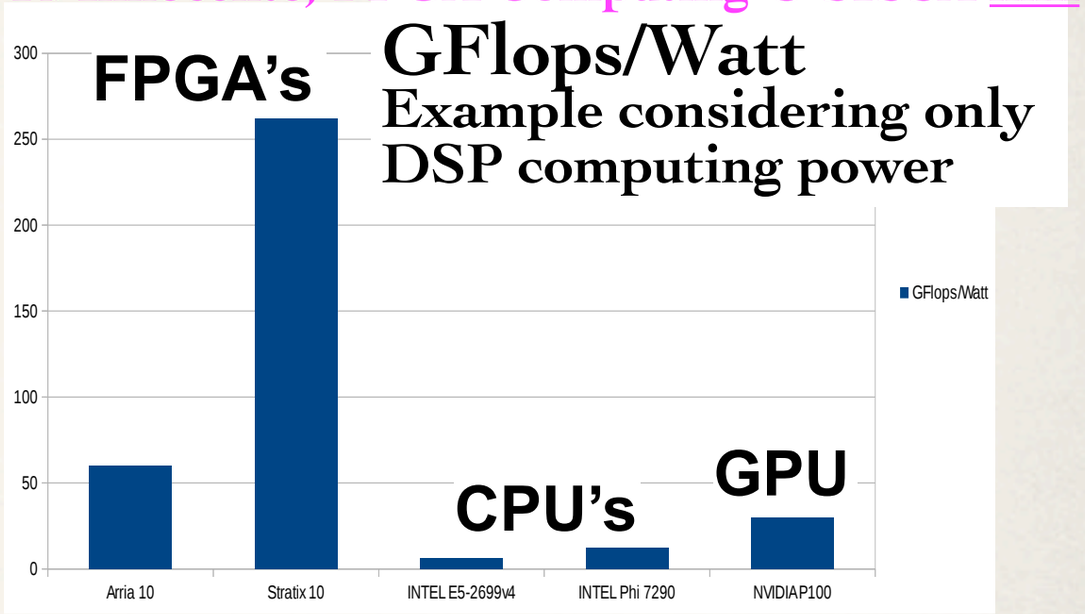
- System feasibility can be predicted after implementing and testing its basic units
- **Switch**: *m* inputs from readout and *m* outputs for each of the *n* switches
- **Engines matrix**: *n* inputs for each of *m* matrices (output bandwidth is not critical)
- The patch panel is a standard COTS component with *nxm* inputs&outputs
- **Speed depends on speed of basic units and of single line I/O**
- **Scalability depends on number of high-speed I/O**



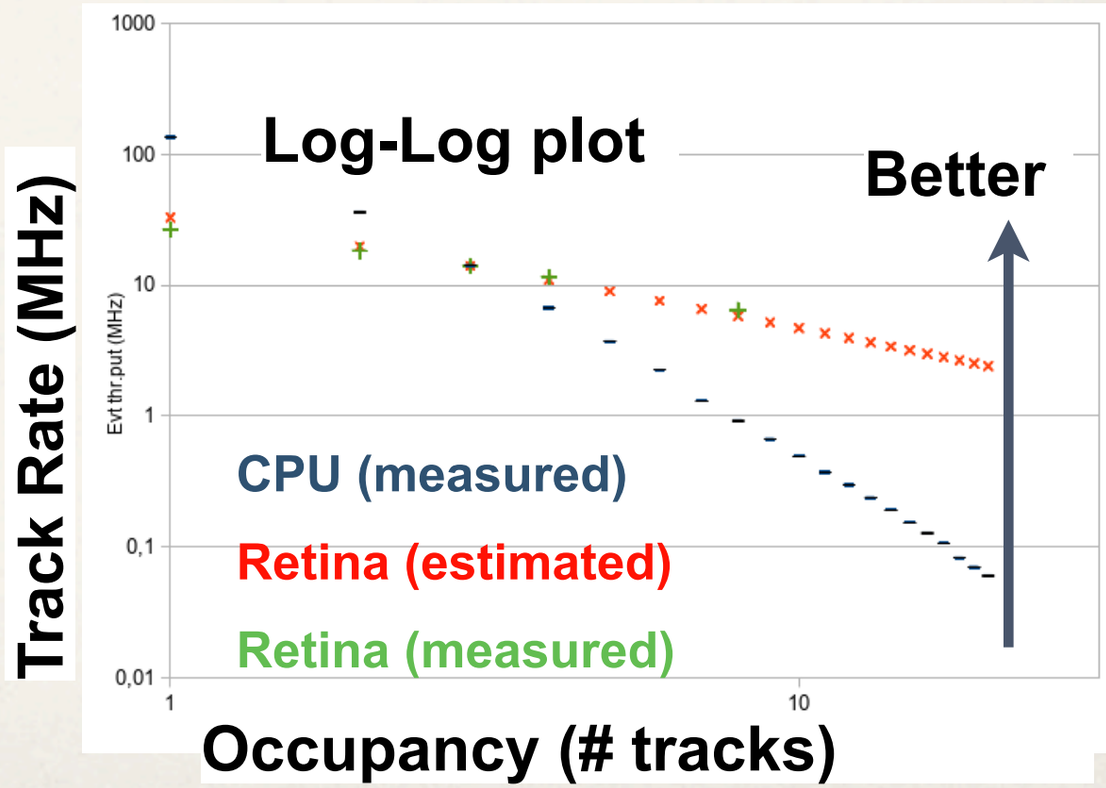
FPGA and Performance Scaling

- FPGA's are currently the **best compromise** between flexibility and computing power, plus largest bandwidth on a single chip
- **FPGA's performances are still growing**
- For last-chance improvement and mass production, designs can be easily transferred to ASIC

R. Innocente, FPGA Computing @ SISSA link



- Scaling of tracking performances for the simplest 3-layer tracker: **Retina vs CPU** (cut on chi2 of all the combinations of 3 hits)
- Absolute scale can change, but slope is significantly different, showing that **best architecture depends on occupancy conditions**

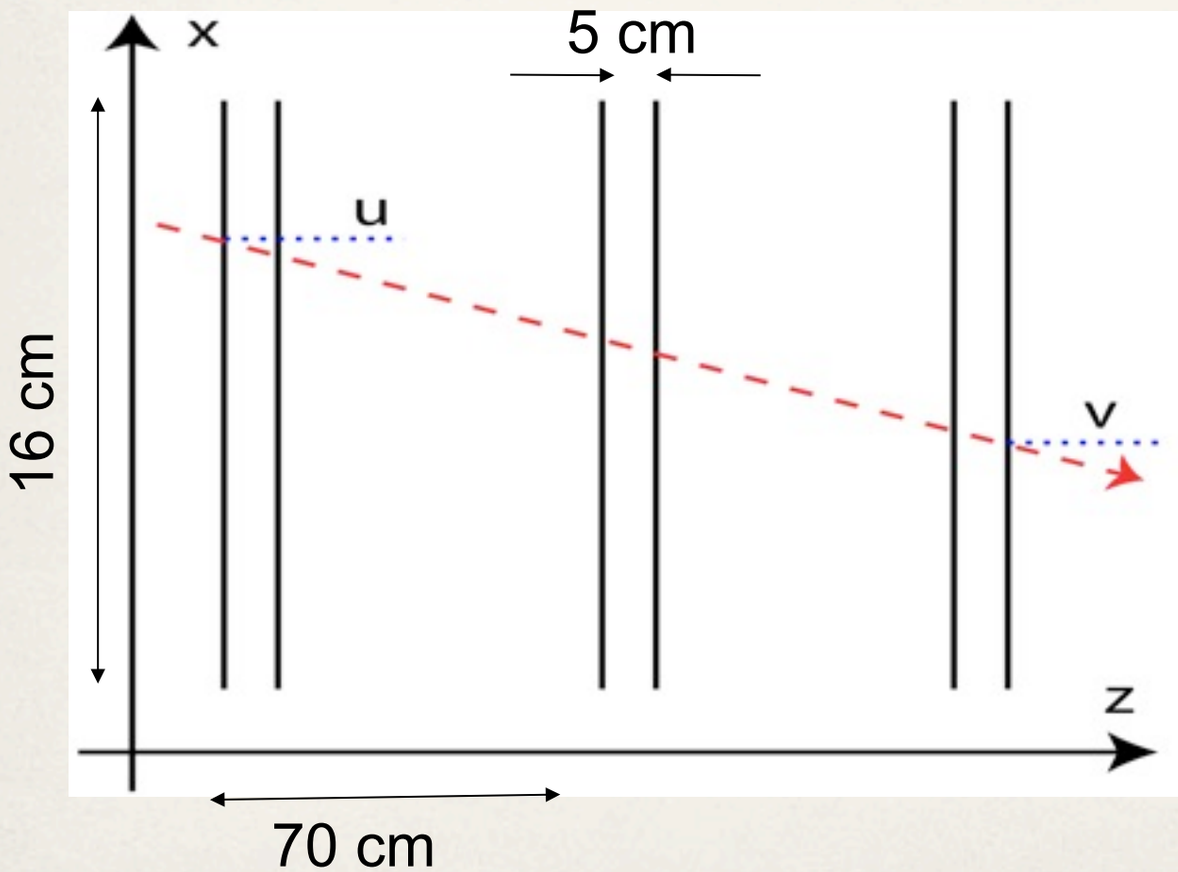


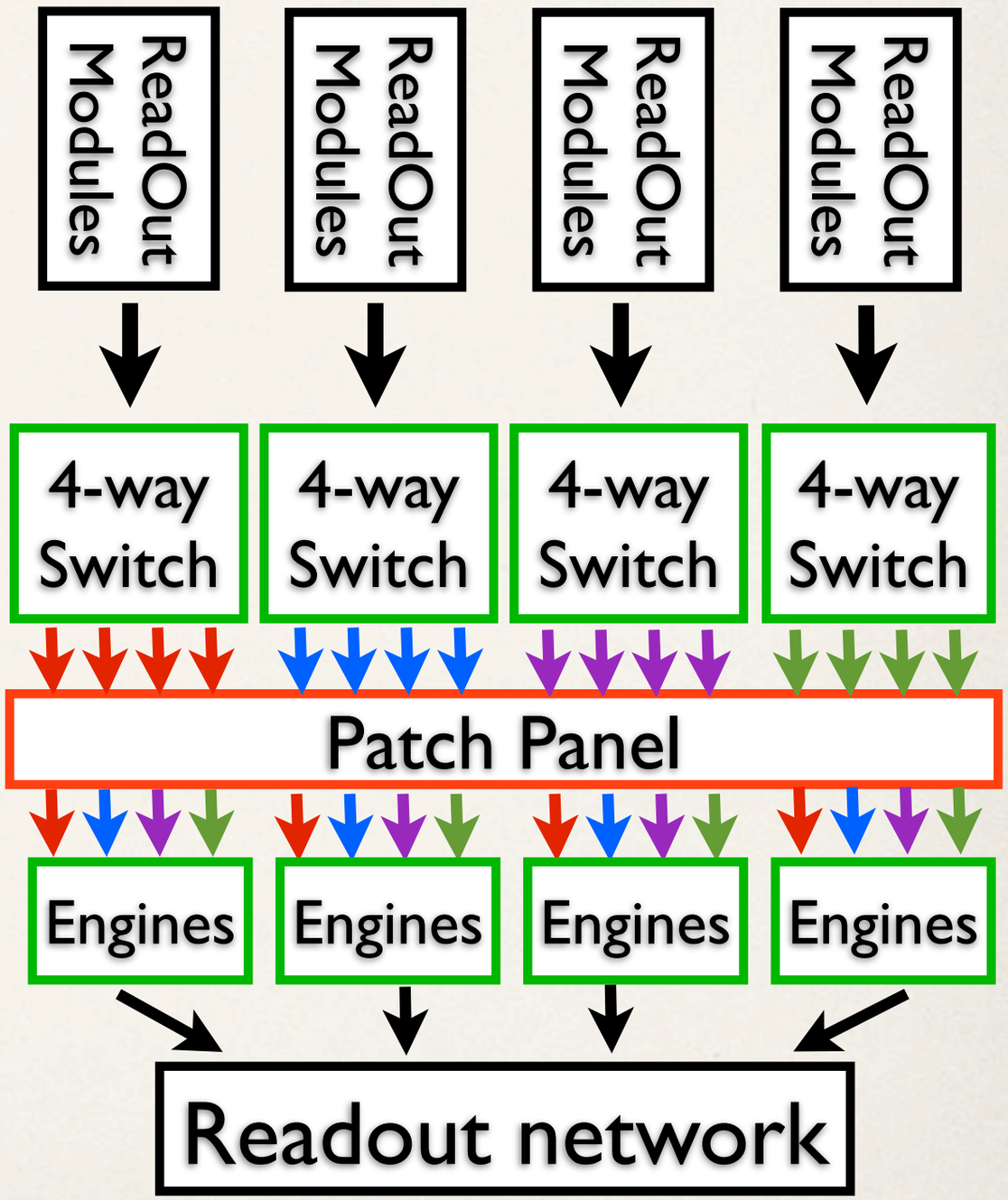
Functional Prototype

Goal: test the logic functionality
of the system when applied to a simple tracker

Retina Applied to 2D Tracking

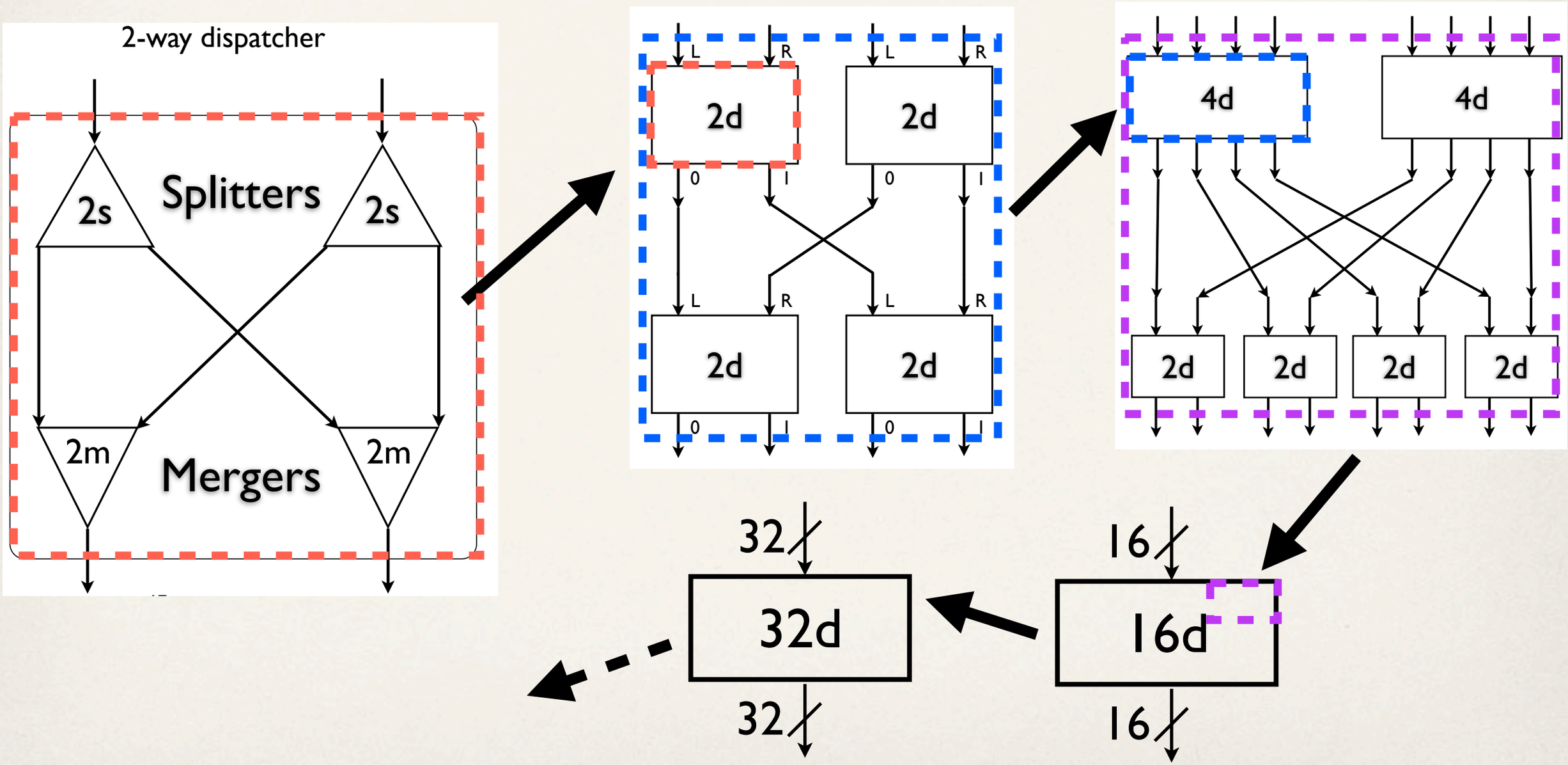
- In 2015 we started the **"Retina" project**, a 3-year R&D program supported by INFN-CNS5 (Technological Research Division)
- First prototype, configured for **2D straight tracks in 6-layer silicon strip detector**





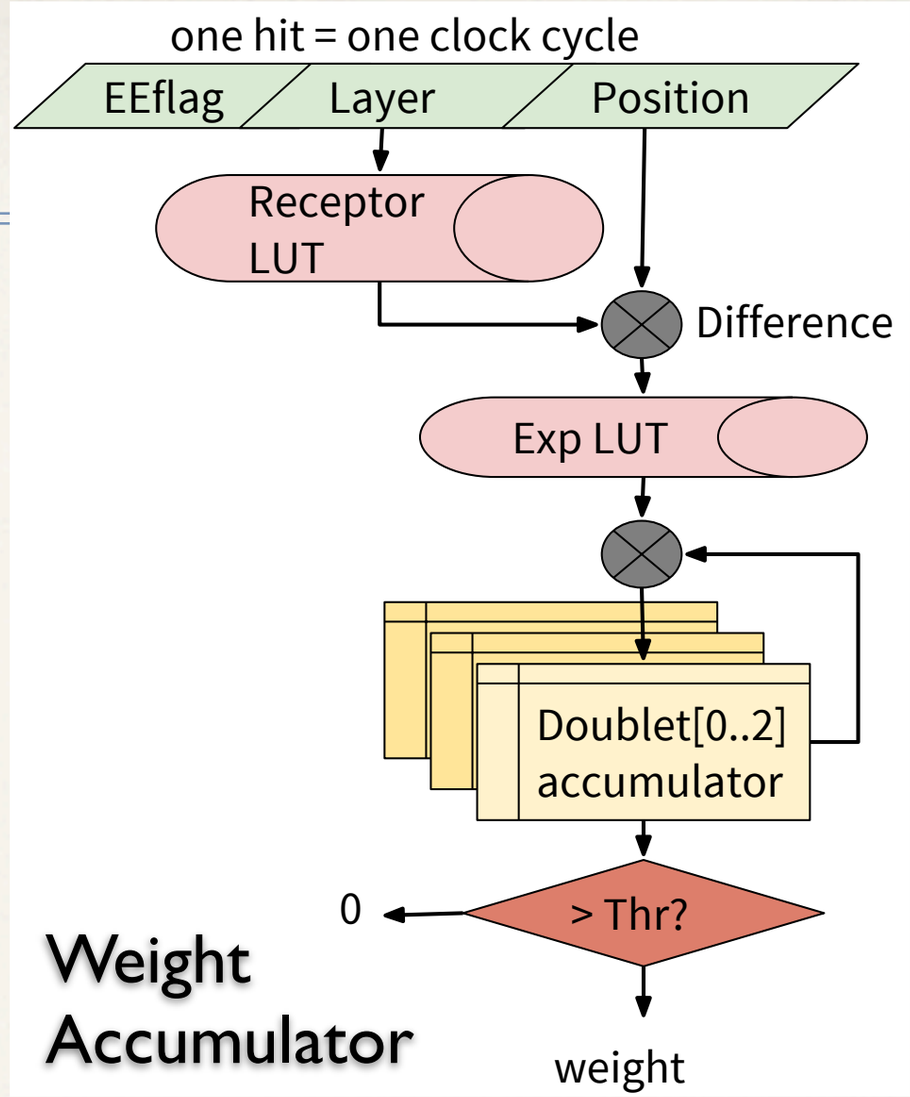
Basic Blocks: the Switch

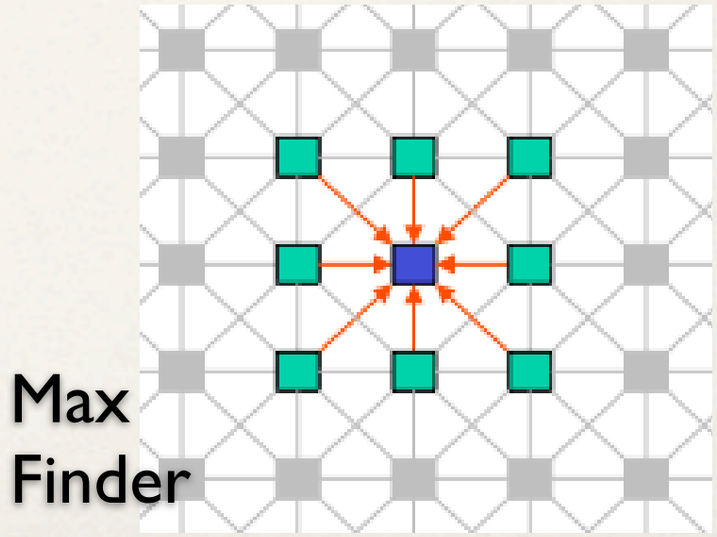
- Pipelined and modular implementation to increase the throughput
- Latency proportional to log2 of #inputs/outputs



Basic Blocks: the Engine

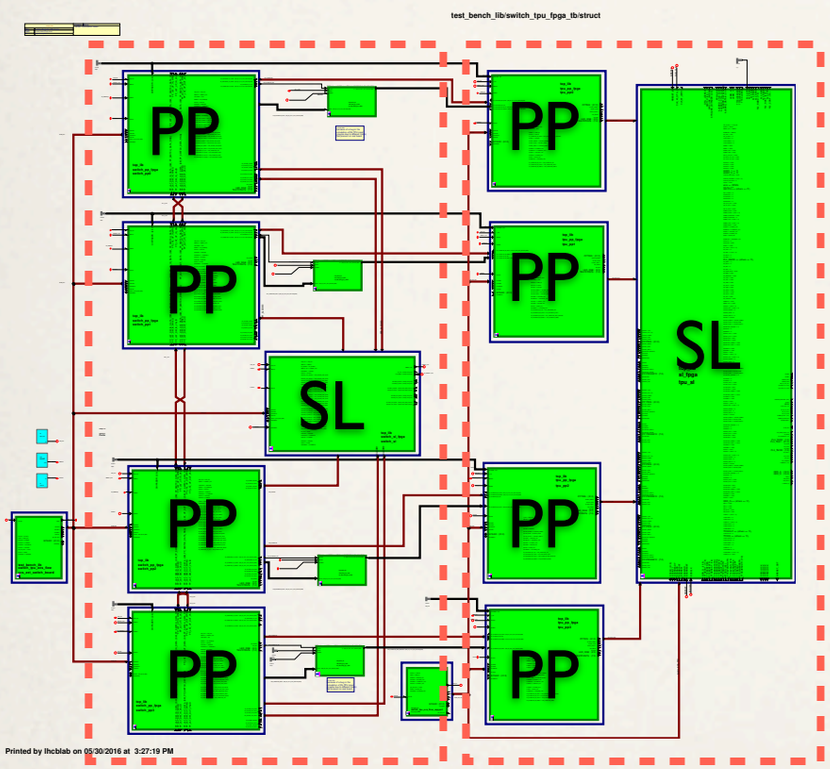
- Function: accumulate weights and find local maxima
- Fully pipelined (one hit = one clock cycle) and parallelized
- **Weight accumulator component:**
  - Retrieve weight from LUT based on layer +receptor and position
  - One accumulator for each layer doublets, output only if above threshold, then summed together
- **Max finder component:**
  - Check if maximum over a 3x3 cluster
  - Reduce data from 9 values to 3 sending out only summations needed to compute centroid
- Data sent out through a priority encoder



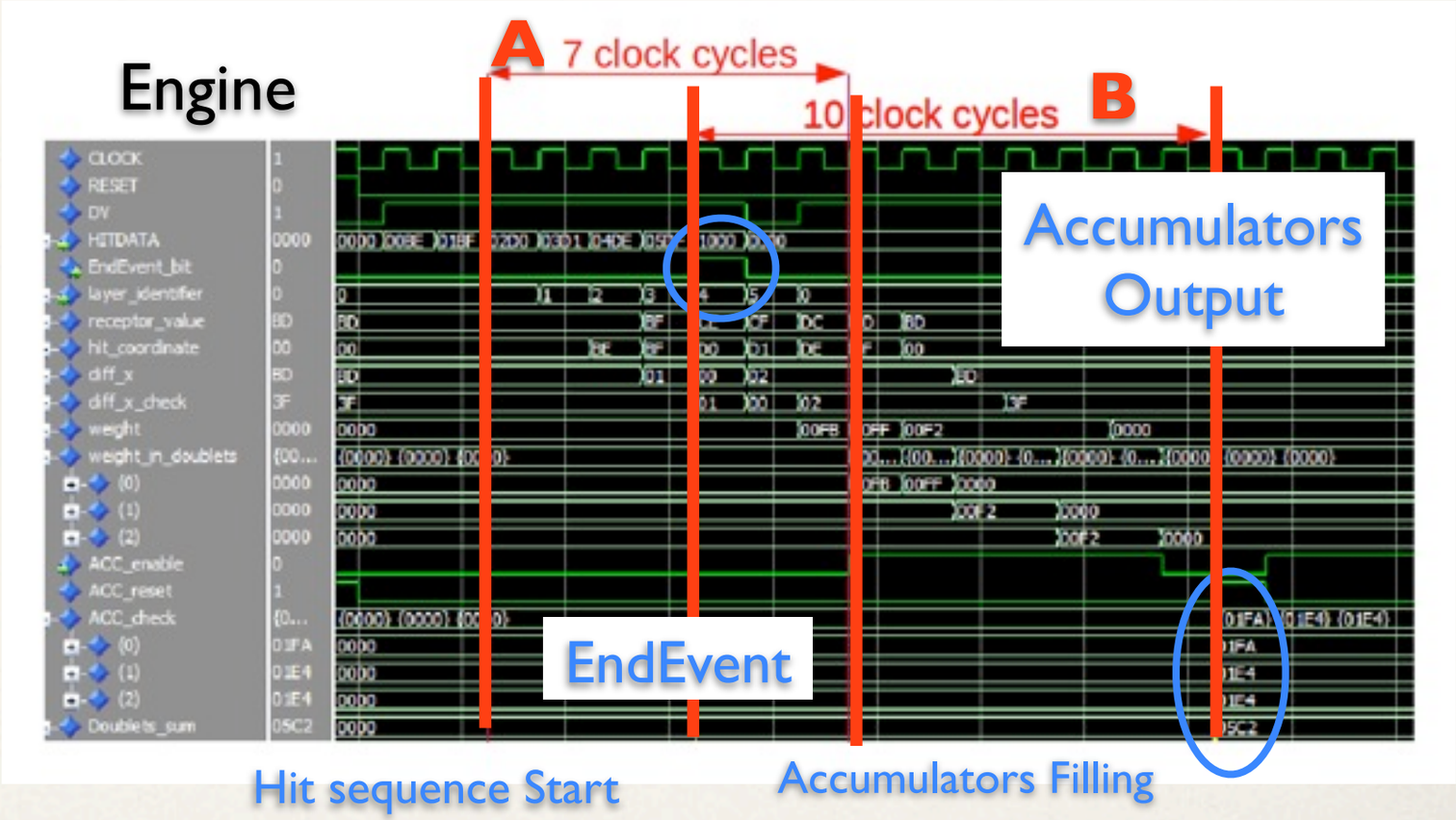


Low-level Simulation (ModelSim)

- The whole system has been simulated at low-level to verify the logic
- Latencies can be estimated and optimized using this simulation
- For the engines latency is made by two parts: data input latency (depends on the # of hits) and accumulators output latency (fixed starting from end event data)



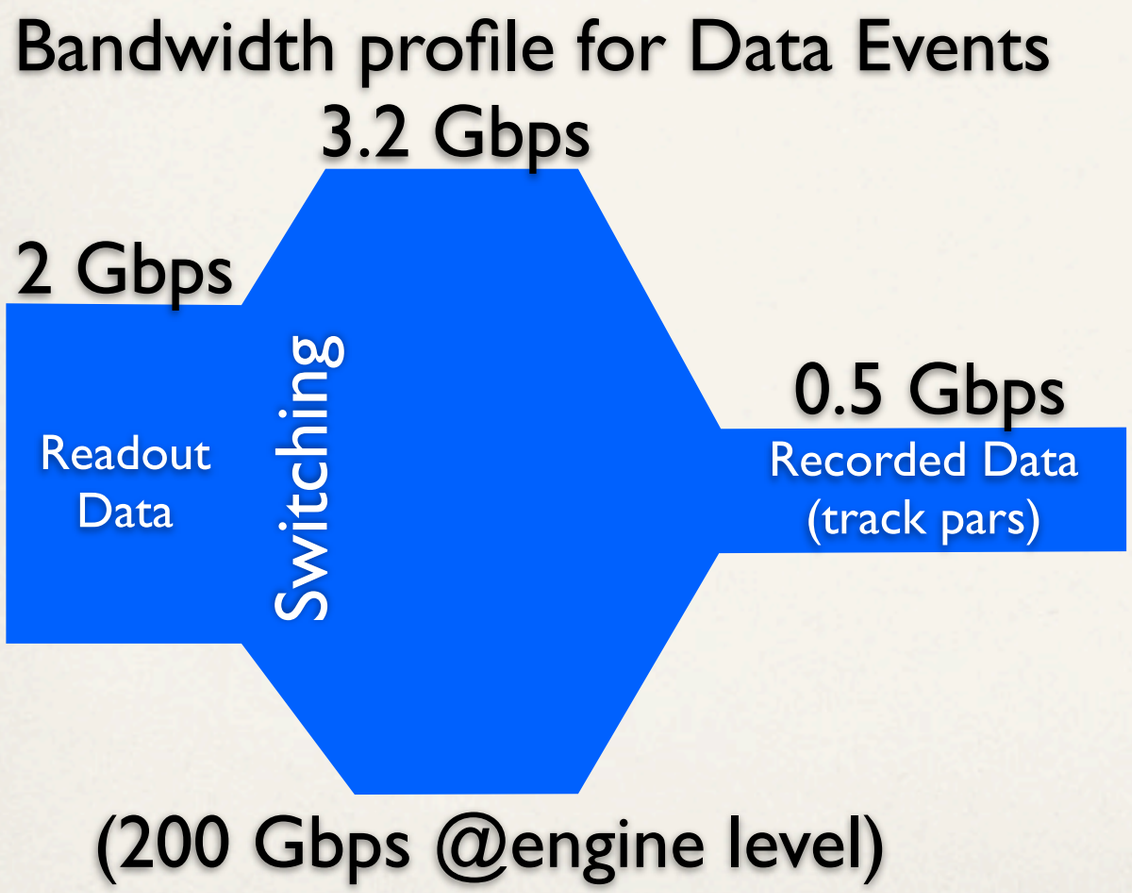
Switch board Engine board

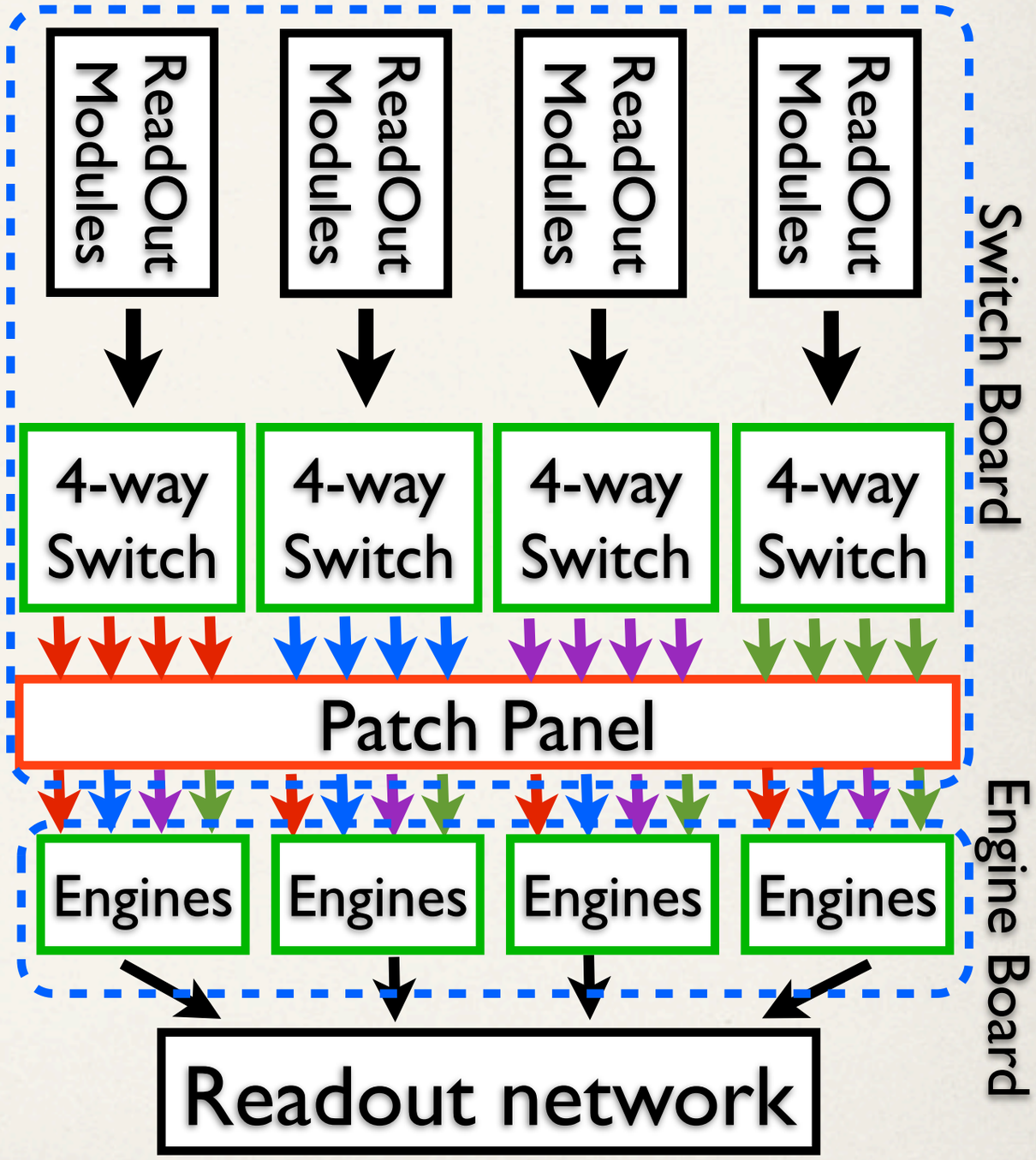


A Time between hit delivery and accumulator update
B Time between end sequence and accumulator output

Results: Functional Prototype

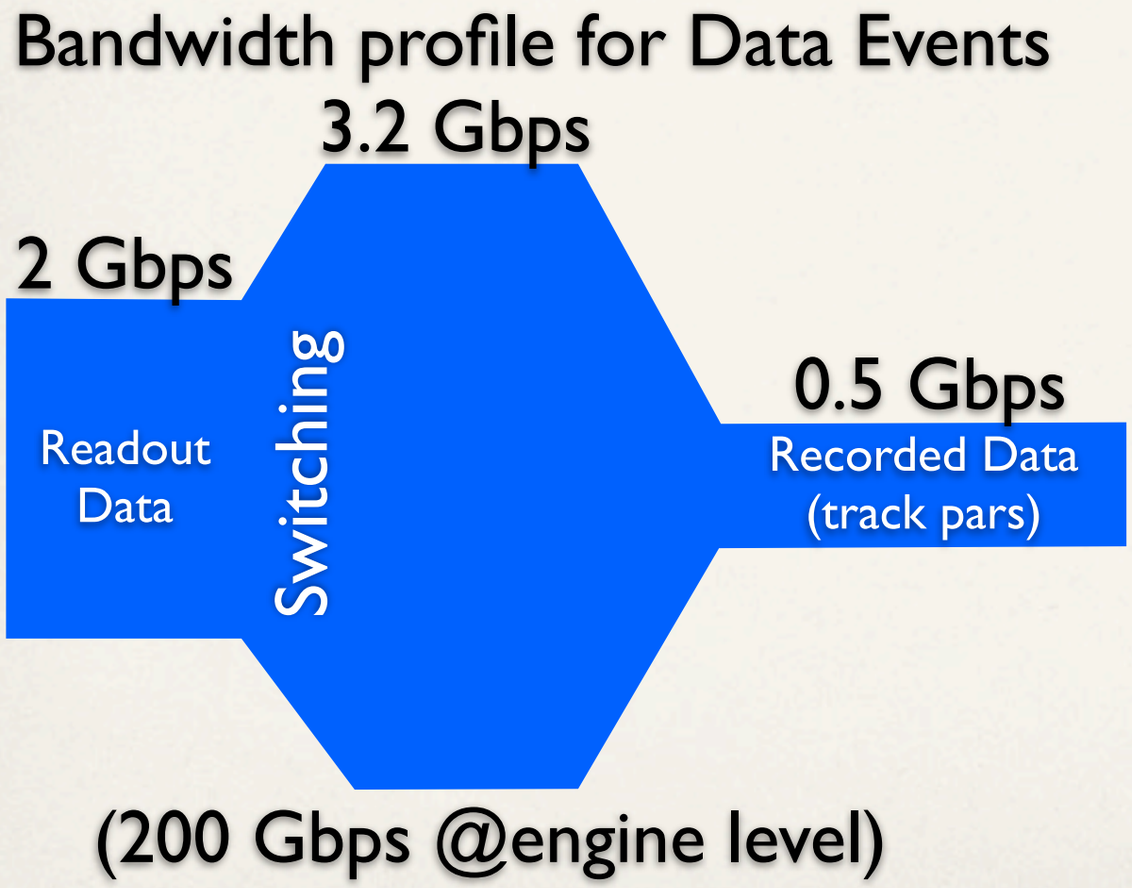
- Early prototype configured for **2D straight lines in 6-layer silicon strip detector**, similar to LHCb IT
- **O(2 MHz) evt rate** using a DAQ board (**Tel62**) equipped with Stratix III FPGA's (65 nm)





Results: Functional Prototype

- Early prototype configured for **2D straight lines in 6-layer silicon strip detector**, similar to LHCb IT
- **O(2 MHz) evt rate** using a DAQ board (**Tel62**) equipped with Stratix III FPGA's (65 nm)

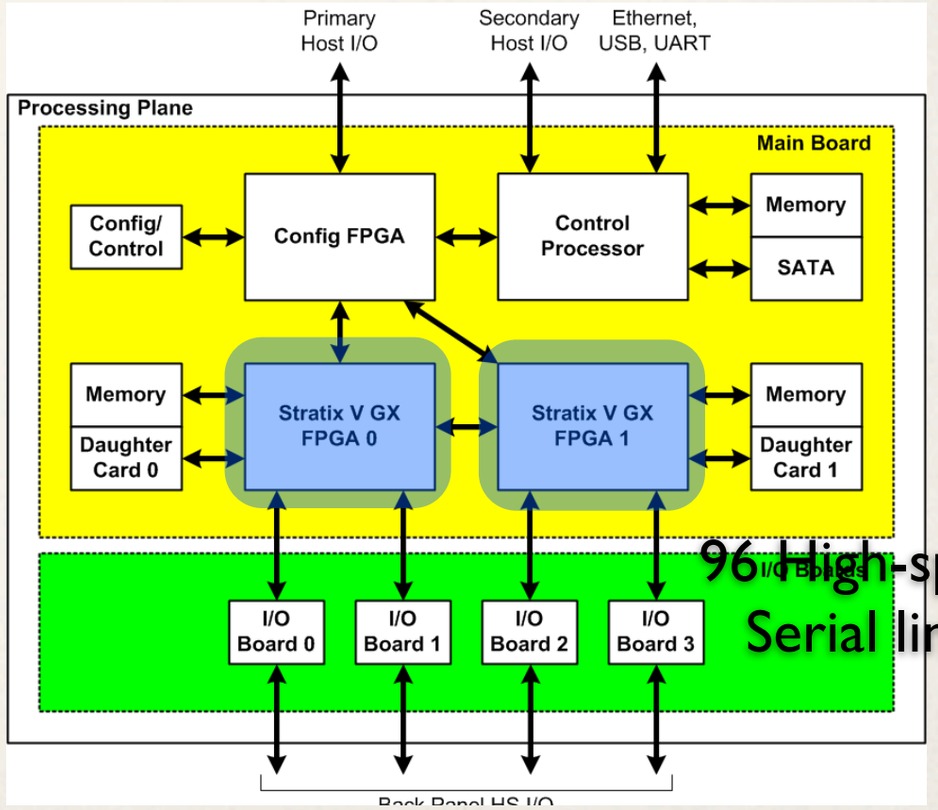


Full-speed Prototype

Goal: test the speed/latency performances for the basic components when implemented on modern devices

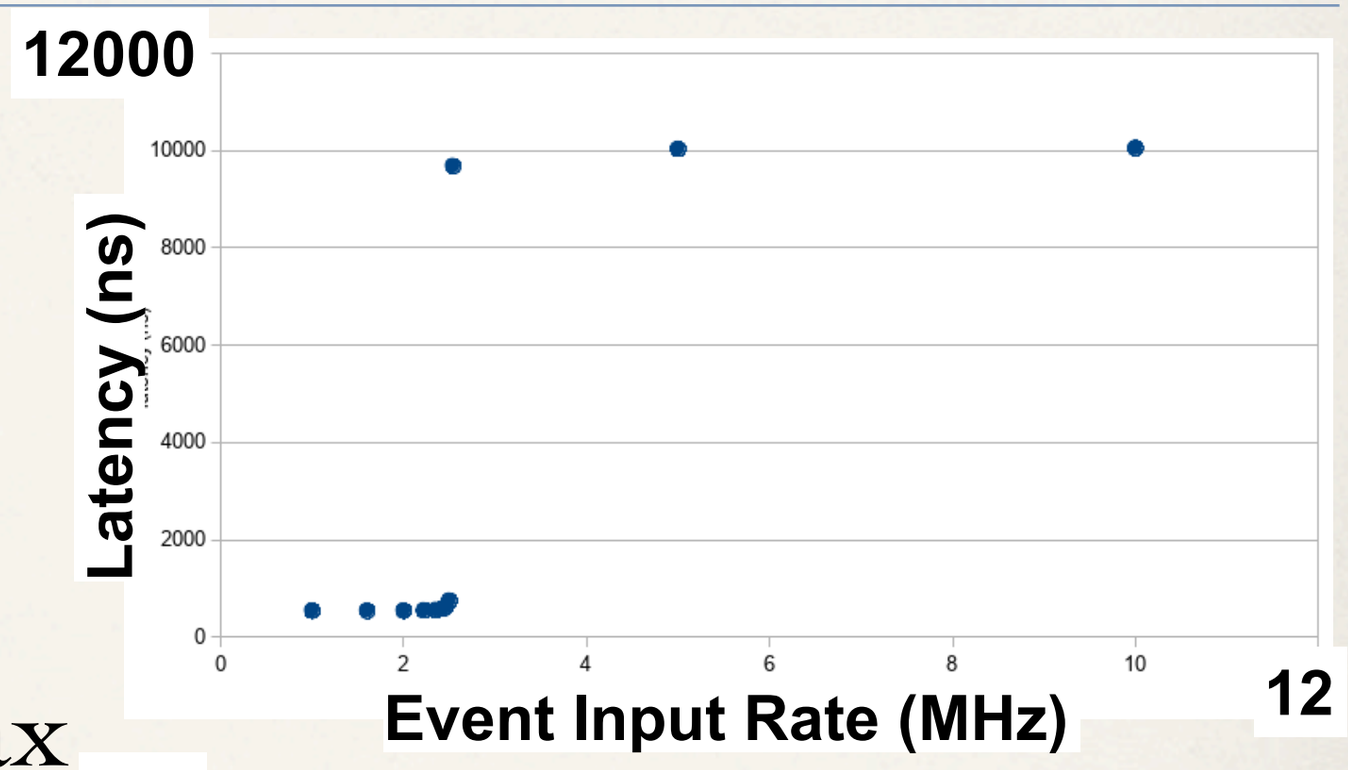
Porting to Faster FPGA: Stratix V

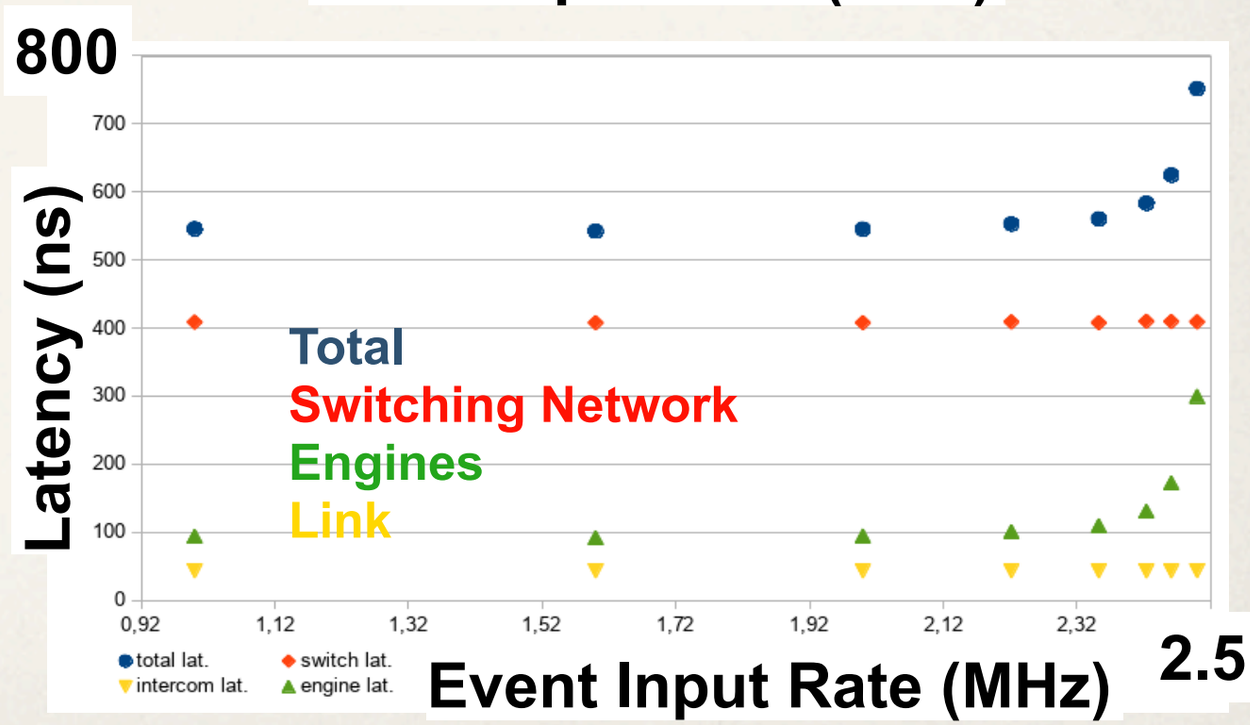
- Board from DiniGroup with 2 Stratix-V (28 nm, 1M logic elements, ~1 Tb/s total available bandwidth on optical fibers)
- Many configurations achievable, including implementing the first prototype system in one chip
- Improvements:
  - **Higher clock frequency (x2 gain)** with 28nm chips (now available 20 and 14nm)
  - **Faster I/O lines (x2)**, or also everything in one chip
  - **Separated input lines for layers (x2)**, weights are summed later further down in the pipeline



Latency

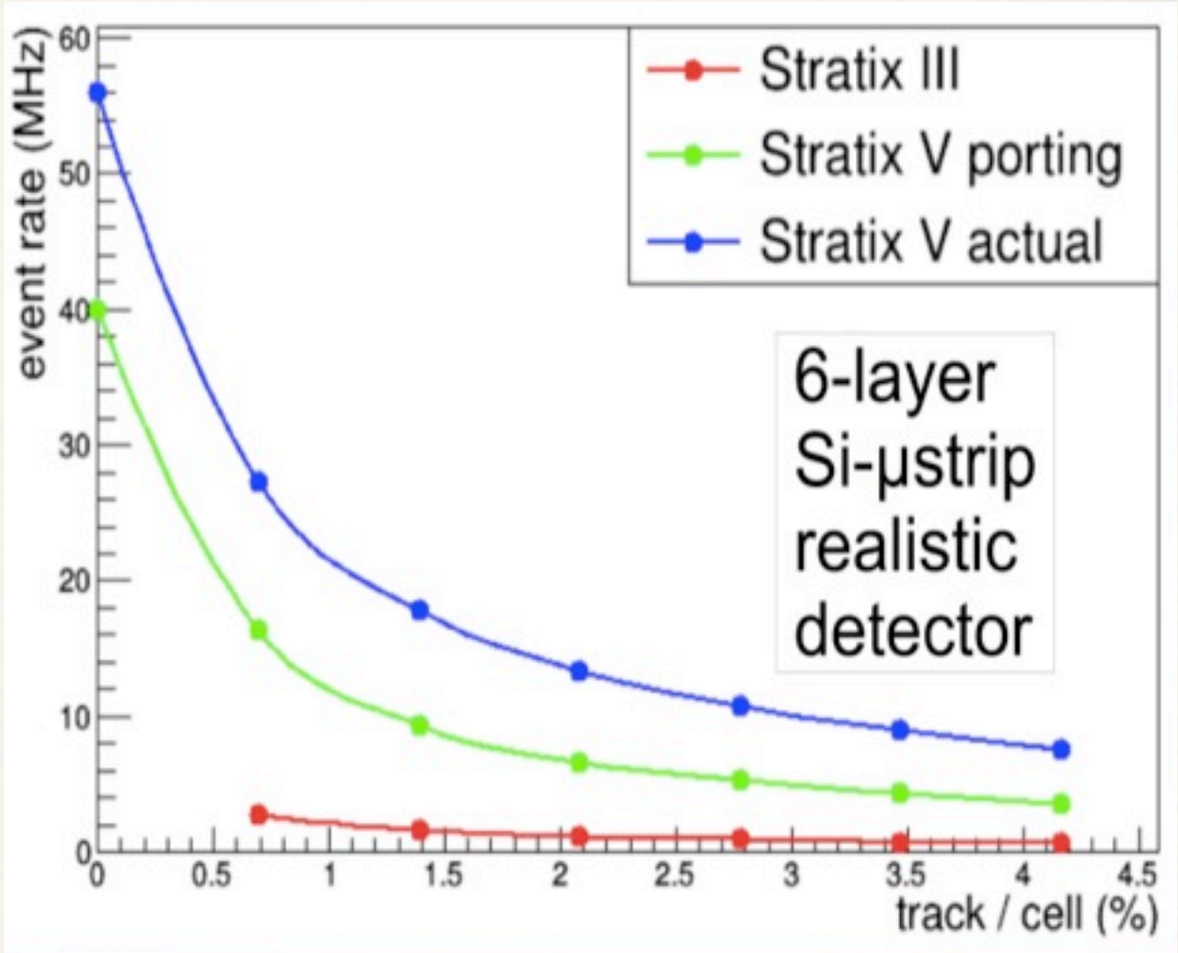
- Direct measurement on the board, while sending hits at a specific rate: **~550ns**
- Latency explodes when we cross the max frequency the system can sustain
- Main contribution from the switching network





Results: Functional Prototype

- Prototype achieved a track rate of **~20 MHz** for occupancy of 1 track every 100 cells
- **Short latency <0.5 us** facilitates embedding in the DAQ system
- Cost estimates for extrapolation to larger systems
  - **Hardware cost <0.1 euro/kHz** of tracks (current chip, expected to decrease further)
  - **Power cost 0.2 mW/kHz** of tracks (very low)

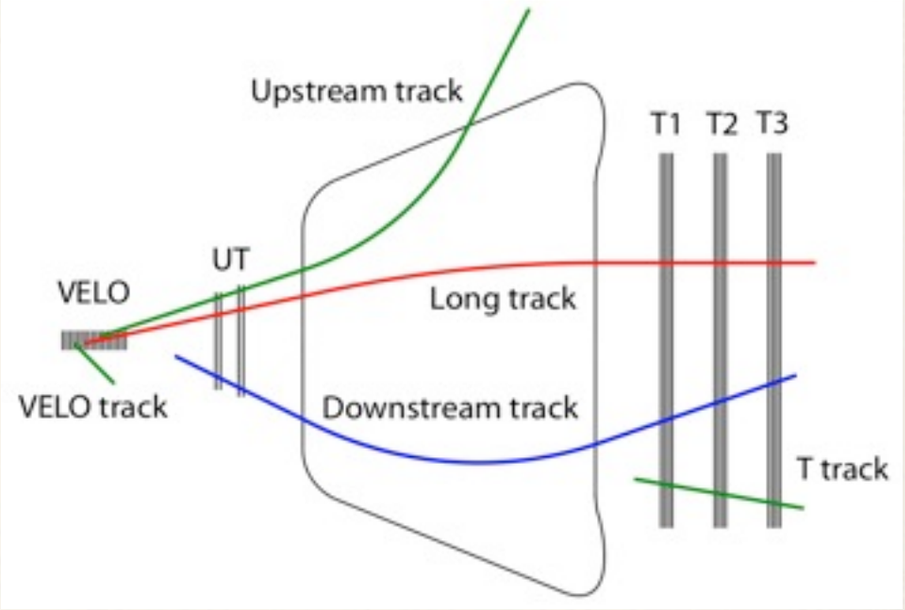


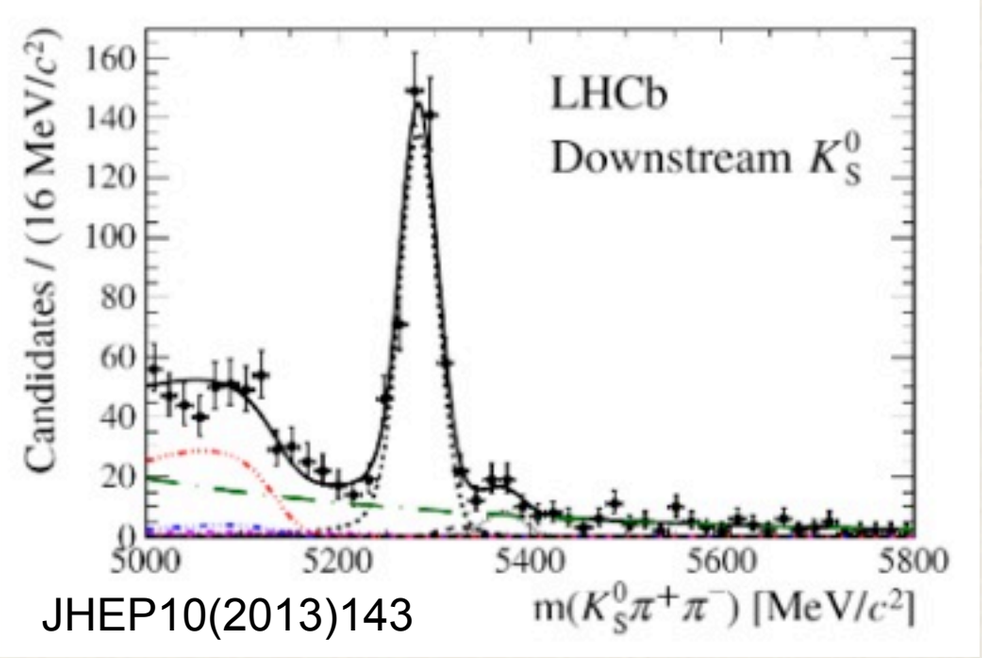
Application to a Real Case

Goal: challenges, motivations and evaluation of tracking performances with occupancy expected at HL-LHC

Physics Case

- For LHCb Upgrade only "**long**" tracks available at trigger level
- **Including Downstream tracks will allow to increase the acceptance of long-lived particles**, but challenging because of much higher combinatorial (no pixel)
- **Long-lived Neutrals, x2** improvement for KS, expected greater gain for Lambda's
  - Charmless decays of B and Bs, CP eigenstates, hadronic charm decays, very rare decays of KS (e.g. KS -> μμ), precision study of the properties of strange baryons
- **Displaced Vertices,** predicted long-lived particles: dark sector bosons, ALPs / Higgs, majorana neutrinos → increase the sensitivity at long lifetimes

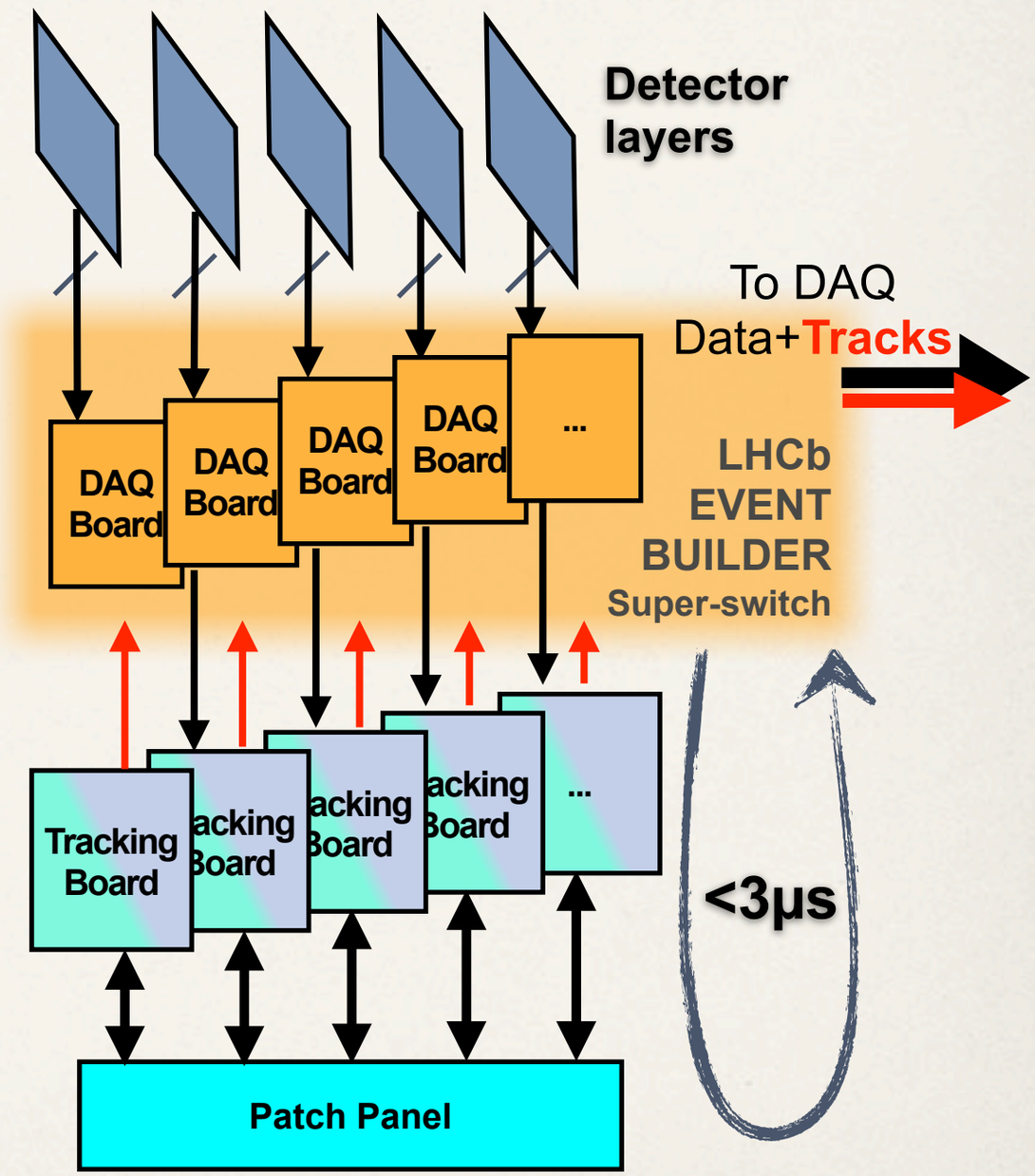




JHEP10(2013)143

Proposed Approach

- We propose to build a **downstream tracking unit that can be integrated in the DAQ architecture and act as an "embedded track-detector", using standard commercial PCIe FPGA boards**
  - Goal: provide primitives for event reconstruction immediately available to event builder and HLT (the trends of migrating reconstruction to early stages)
- Note: distributed event builder with high-speed network force us to change architecture (transparent integration)
  - This approach has been included in the recent **Expression of Interest for Run5** presented to LHCC [CERN-LHCC-2017-003]



Configuration and Output

- Adapting prototype to **one quadrant of axial T-stations**, 6 layers of scintillating fibers (**2D SciFi**)
- Estimated from LHCb MC for Upgrade **~ 50 reconstructable tracks per quadrant**
- Assuming the chip occupancy for our prototype and max number of chip, we can implement **20k cells per quadrant**
- **High number of ghost, expected because we are using only x-layers**

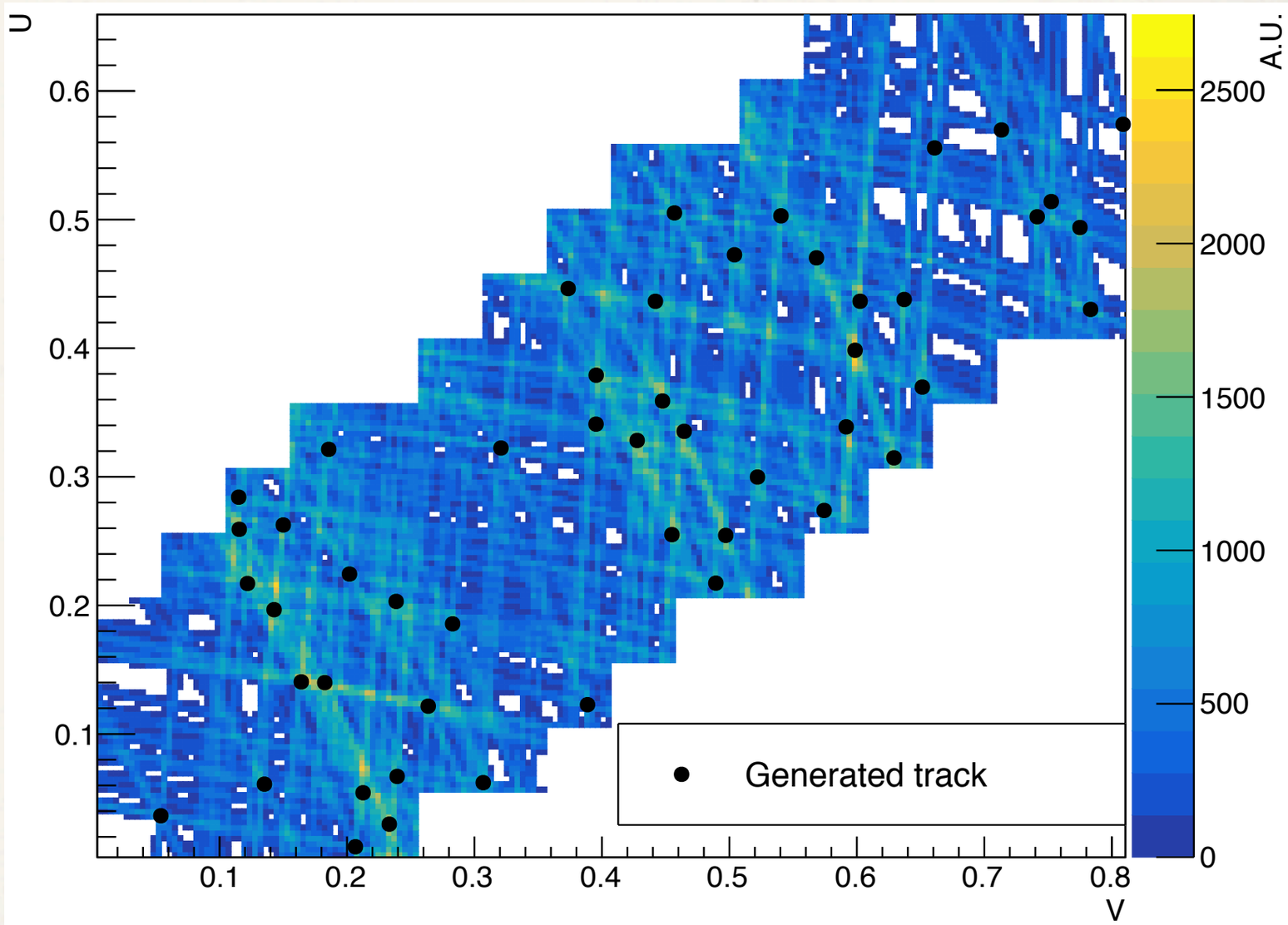


Figure: Weights accumulators in each cell for one event with **50 tracks**.

Configuration and Output

- Adapting prototype to **one quadrant of axial T-stations**, 6 layers of scintillating fibers (**2D SciFi**)
- Estimated from LHCb MC for Upgrade **~ 50 reconstructable tracks per quadrant**
- Assuming the chip occupancy for our prototype and max number of chip, we can implement **20k cells per quadrant**
- **High number of ghost, expected because we are using only x-layers**

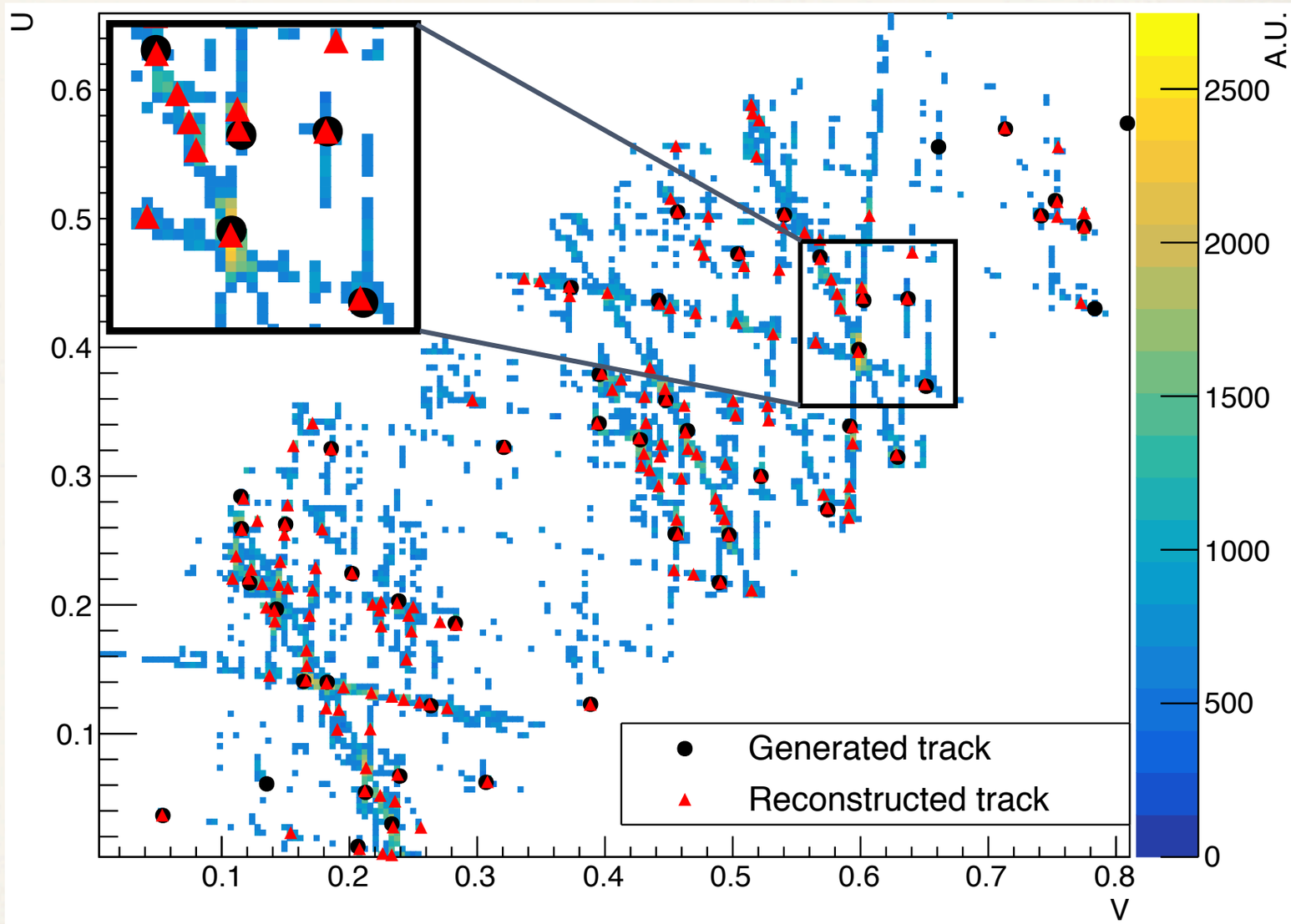
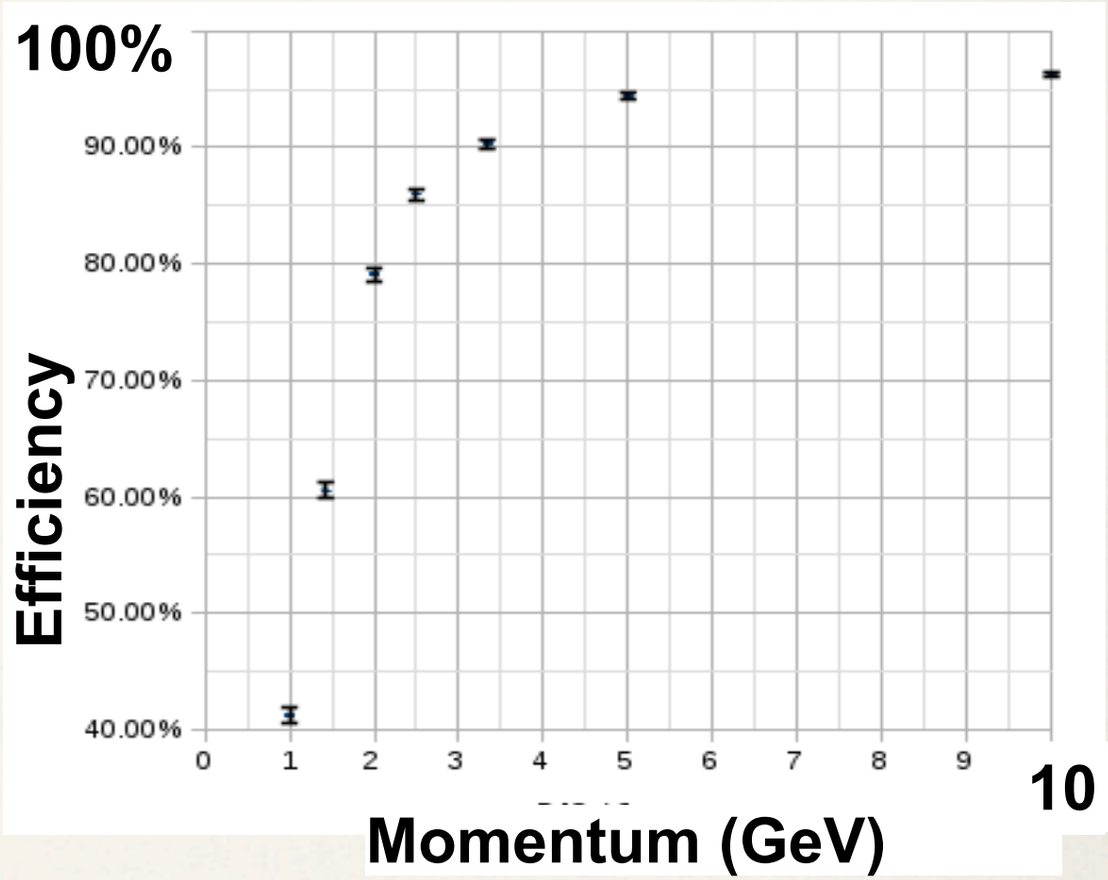


Figure: Weights accumulators in each cell for one event with **50 tracks**.

Results: Application to 2D SciFi

- Test using tracks and hits generated with toy MC (no fringe magnetic field, multiple scattering, or noise): **efficiency 95%, ghost 48%**
  - Estimated performances are very **similar to Offline software reconstruction** under the same conditions
  - Ghost rate is but is expected using axial layers
- Test using tracks and hits simulated with **LHCb full simulation**, no noise: see plot
  - No magnetic bending for patterns
  - **Good efficiency down to ~3 GeV**



Future Improvements and Ideas

- Implement the 2D SciFi configuration in the board to measure speed and latency with realistic events
- Demonstrate that integration with LHCb TDAQ system is feasible
- Simultaneous processing of events, adding hit timestamp
- Add information from stereo layers to reduce ghosts (4 pars phase-space), and also UT hits (5 pars)

Conclusions

- **Real-time reconstruction** of tracks at lower trigger level can help a lot when selecting events, specially in the HL-LHC environment, but it is challenging due to very high combinatorial
- We demonstrated that the **"artificial Retina" algorithm is a feasible approach for a tracking trigger** at low-level for HL-LHC
  - Using FPGA, speed (tenths of MHz) and latency (few μs) are in the right order of magnitude
  - Two prototypes developed by the Retina project gave us also the chance to develop the basic blocks for a larger system
- Therefore we started the **development of a processor for reconstructing "downstream" tracks at LHCb in Run4**
  - Study of performances using high-level simulation of the tracking system shows **performances similar wrt offline tracking** (axial layers only)
  - We are less than a factor 2 away from required performances using FPGA from the previous generation

Backup

References (1)

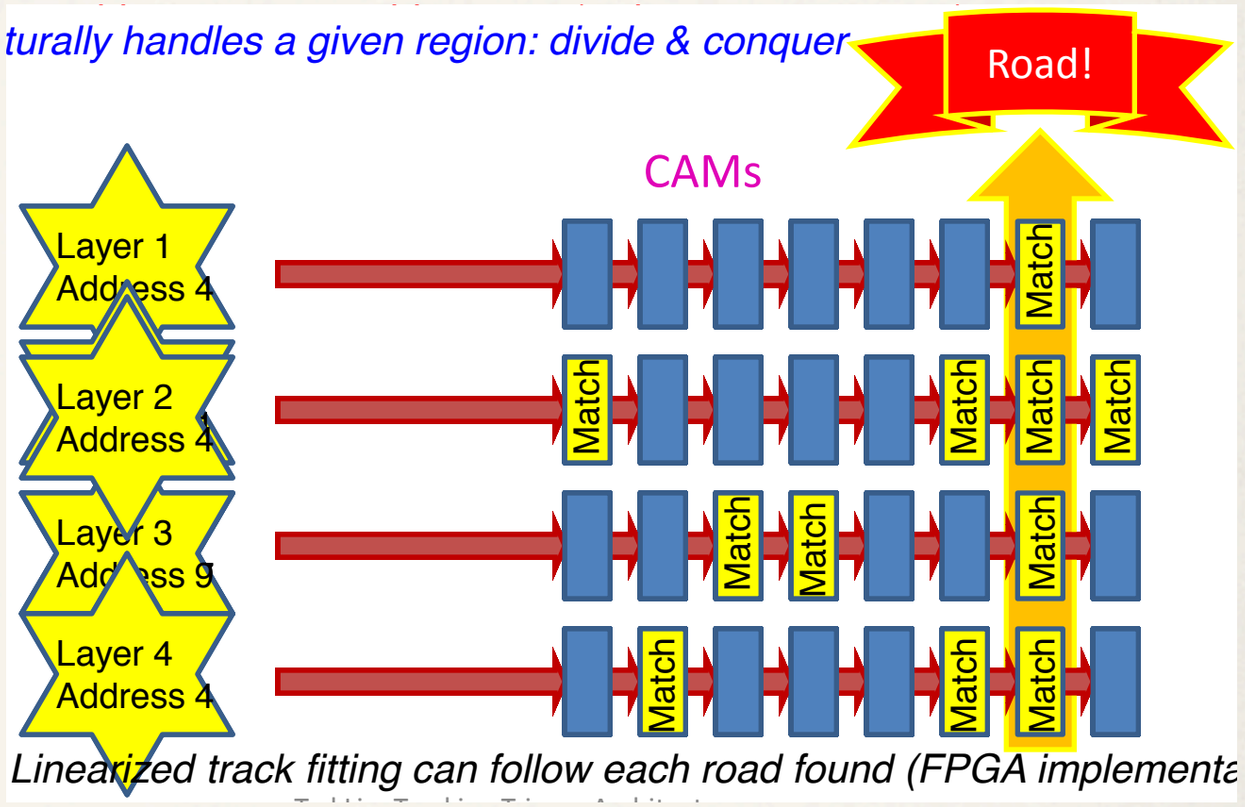
- TTFU Elba Stracka, 2017
- S. Stracka, CHEP16
- M. J. Morello, IEEE-NSS 2016
- R. Cenci, Seminar @ INFN Pisa, May 2016
- S. Stracka, First results with an "Artificial Retina" processor prototype, The International Conference on Modern Circuits and Systems Technologies (MOCAST), May 2016 Thessaloniki, Greece, http://mocast.physics.auth.gr/
- R. Cenci, First prototype of an "Artificial Retina" Processor for Track Reconstruction, Talk at Connecting The Dots Workshop 2016, https://indico.hephy.oeaw.ac.at/event/86/session/0/contribution/2/material/slides/0.pdf
- N. Neri, VCI2016
- G. Punzi, The Retina Algorithm, Talk at DataScience@LHC 2015 Workshop, http://indico.cern.ch/event/395374/session/6/contribution/35/attachments/1186122/1719654/DataScience-Punzi.pdf
- G. Punzi, Real-time computing, neural systems, and future challenges in HEP, XII Seminar on Software for Nuclear, Subnuclear and Applied Physics, May 2015, Alghero https://agenda.infn.it/getFile.py/access?contribId=12&resId=0&materialId=slides&confId=8781
- M. Petruzzo, Poster at Frontier Detectors for Frontier Physics - 13th Pisa Meeting on Advanced Detectors
- R. Cenci, An "artificial retina" processor for track reconstruction at the full LHC crossing rate, Nucl. Instrum. Meth. A (accepted manuscript, in press) Poster at Frontier Detectors for Frontier Physics - 13th Pisa Meeting on Advanced Detectors, http://dx.doi.org/10.1016/j.nima.2015.10.048
- N. Neri, First results of the silicon telescope using an 'artificial retina' for fast track finding, Talk at ANIMMA15 conference, http://www.ipfn.ist.utl.pt/ANIMMA2015/
- S. Stracka, A specialized processor for track reconstruction at the LHC crossing rate, Talk at Connecting The Dots Workshop 2015, https://indico.physics.lbl.gov/indico/getFile.py/access?contribId=14&sessionId=2&resId=0&materialId=slides&confId=149
- R. Cenci, Artificial retina processor for track reconstruction, Talk at Connecting The Dots Workshop 2015, https://indico.physics.lbl.gov/indico/getFile.py/access?contribId=2&sessionId=9&resId=0&materialId=slides&confId=149
- A. Abba et al., Progress Towards the First Prototype of a Silicon Tracker Using an 'Artificial Retina' for Fast Track Finding, Poster at TWEPP14, https://indico.cern.ch/event/299180/session/7/contribution/64
- A. Piucci, Reconstruction of tracks in real time at high luminosity environment at LHC, Master thesis, https://etd.adm.unipi.it/theses/available/etd-06242014-055001/.

References (2)

- D. Ninci, Real-time track reconstruction with FPGA at LHC, https://etd.adm.unipi.it/theses/available/etd-11302014-212637/.
- F. Spinella et al., The TEL62: A real-time board for the NA62 Trigger and Data Acquisition. Data flow and firmware design, IEEE Nucl. Sci. Symp. Conf. Rec., 1 (2014).
- A. Abba et al., The artificial retina for track reconstruction at the LHC crossing rate, arXiv:1411.1281 [ICHEP 2014], https://inspirehep.net/record/1326137.
- N. Neri, First prototype of a silicon tracker using an 'artificial retina' for fast track finding , PoS TIPP2014 (2014) 199 [TIPP2014], https://inspirehep.net/record/1315951.
- A. Abba, The artificial retina processor for track reconstruction at the LHC crossing rate, JINST 10 (2015) 03, C03018 [WIT2014], https://inspirehep.net/record/1315154.
- A. Abba et al, Simulation and performance of an artificial retina for 40 MHz track reconstruction, JINST 03-10 (C03008) [WIT2014], https://inspirehep.net/record/1314984.
- A. Abba et al., A Specialized Processor for Track Reconstruction at the LHC Crossing Rate, JINST 9 (C09001) 2014 [INSTR14], https://inspirehep.net/record/1303542.
- A. Abba et al., The Readout Architecture for the Retina-Based Cosmic Ray Telescope, Real Time Conference (RT), 2014 19th IEEE-NPSS, [IEEE-RT 2014] http://dx.doi.org/10.1109/RTC.2014.7097516.
- A. Abba, et al., A retina-based cosmic rays telescope, Real Time Conference (RT), 2014 19th IEEE-NPSS, [IEEE-RT2014], http://dx.doi.org/10.1109/RTC.2014.7097515, https://inspirehep.net/record/1367442.
- A. Abba et al., A specialized track processor for the LHCb upgrade, CERNa-LHCb-PUB-2014-026 https://cds.cern.ch/record/1667587.
- M. M. Del Viva, G. Punzi, D. Benedetti, Information and Perception of Meaningful Patterns [PDF]
- M. M. Del Viva, G. Punzi, The brain as a trigger system, http://arxiv.org/abs/1410.5123
- L. Ristori, An artificial retina for fast track finding, NIM A 453 (425-429), http://inspirehep.net/record/539203

Pattern recognition

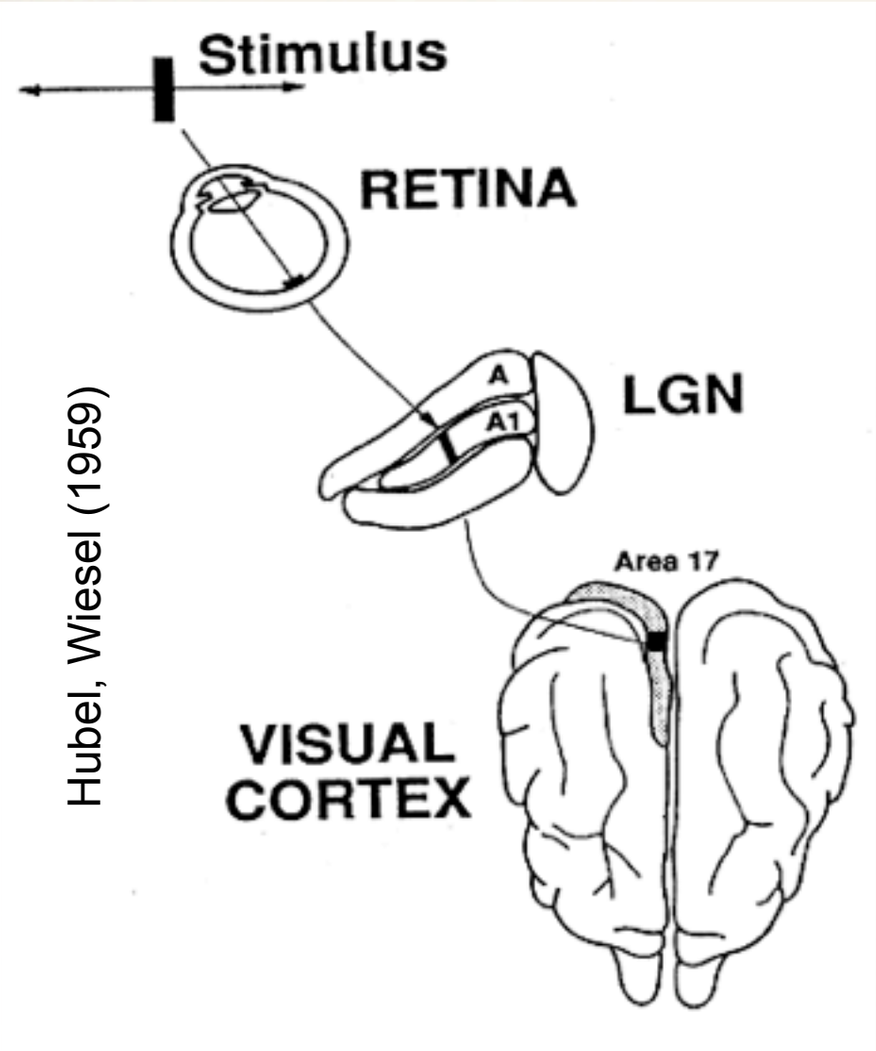
- The fastest approach to tracking implemented in a real experiment is direct matching to a bank of stored templates: **Associative Memory** (SVT@CDF)
  - No combinatorics, comparison in parallel, but patterns are still sequential in AM cell
  - Same approach will be used for Atlas L2 trigger (FTK) and CMS Phase-2
- But requirements for L0 at **HL-LHC** are not matched by a **factor ~80, is it impossible then?**



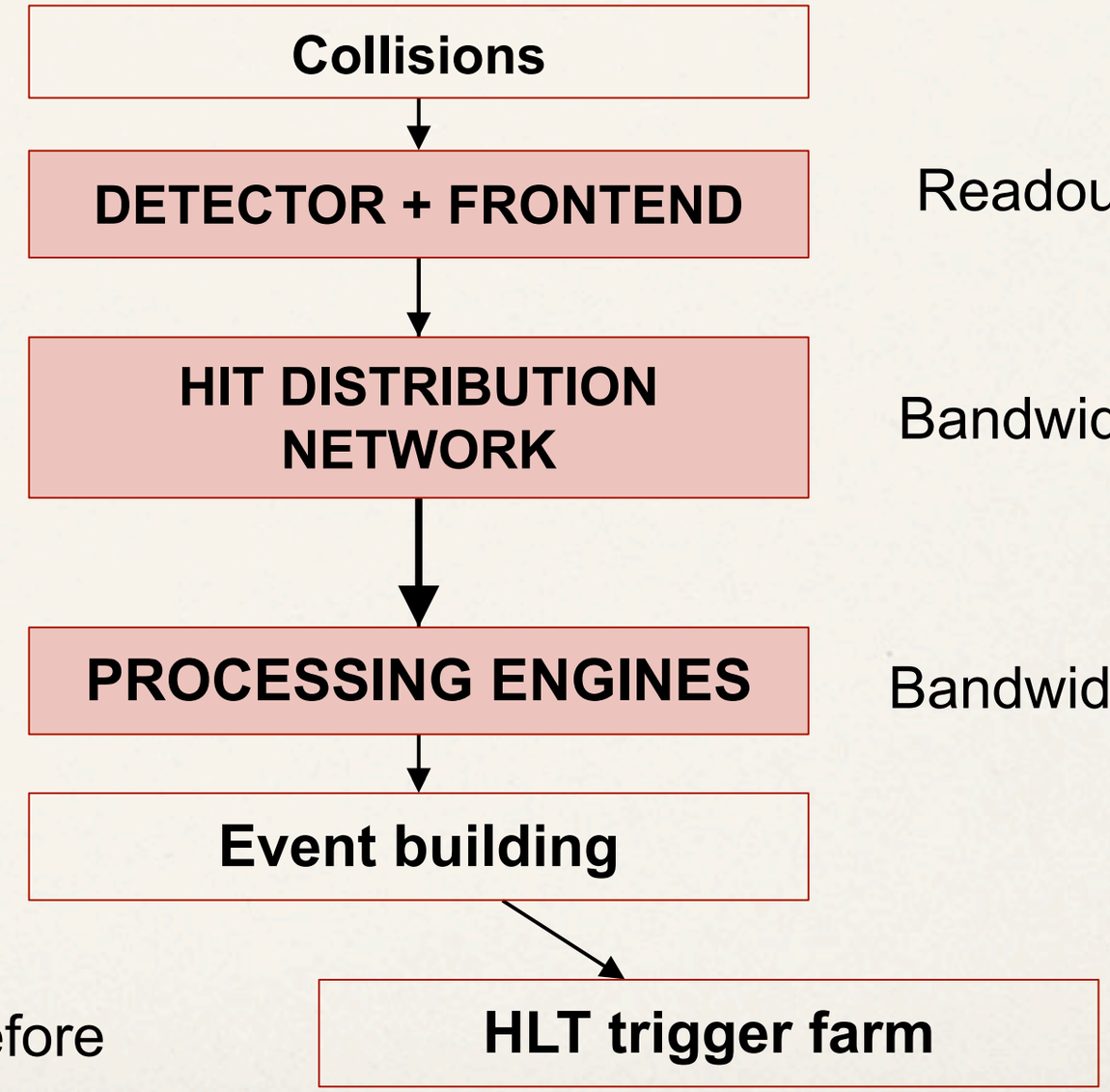
Name	Technology	Experiment	Year	Event Rate	Clock	Cycles/event	Latency
XFT	FPGA	CDF-L0	2000	2.5 MHz	200 MHz	80	<4μs
SVT	AM	CDF-L2	2000	30 kHz	40 MHz	~1600	<20μs
FTK	AM	ATLAS-L2	2015	100 kHz	~200 MHz	~2000	O(10μs)
?	**?**	**<LHC>-L0**	**~2020**	**40 MHz**	**~1GHz**	**~25**	**few μs**

A biologically inspired architecture

- Rely on **Retina algorithm**, whose architectural choices are targeted to the integration of tracking and DAQ



Conceived for parallelism: process before event building



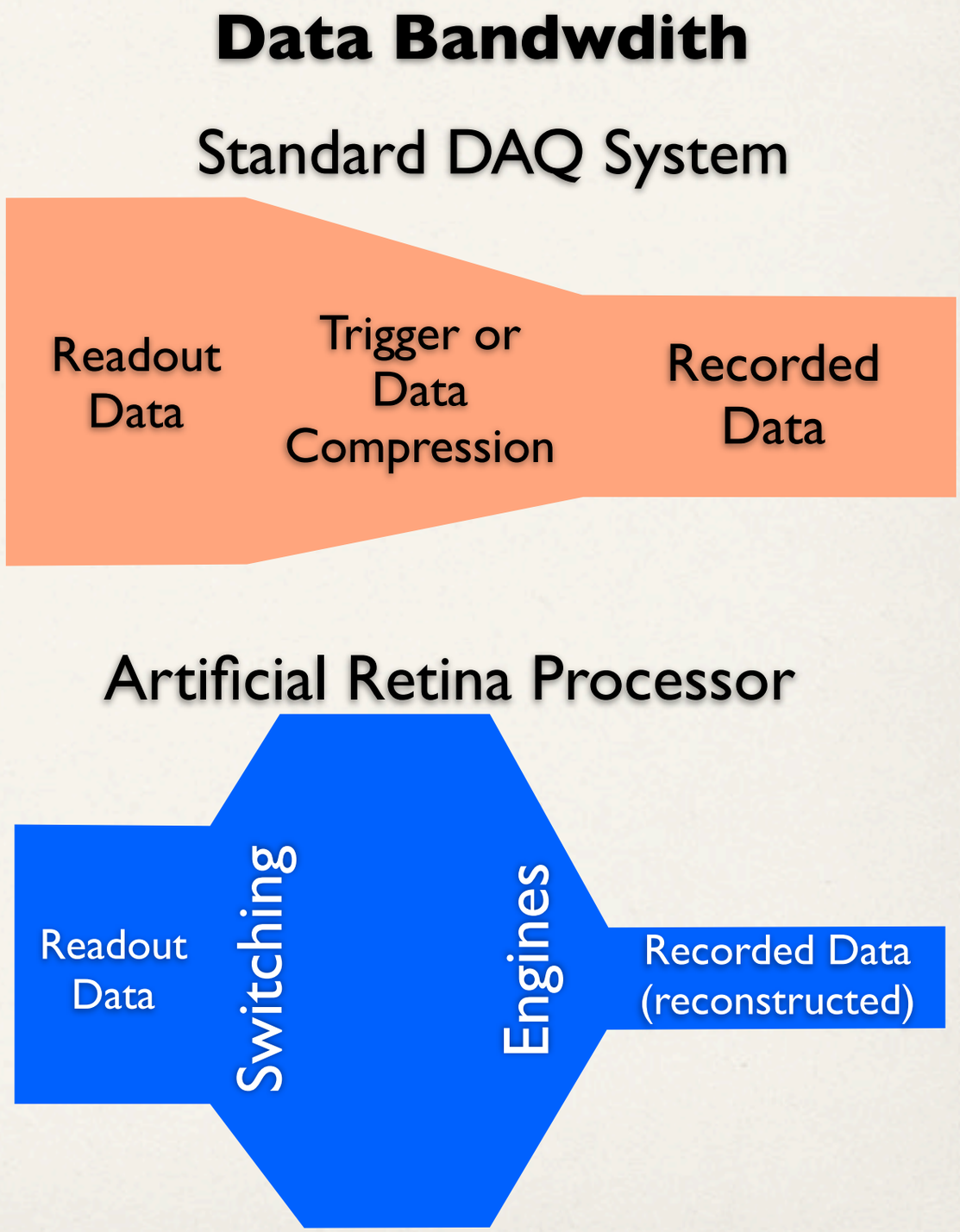
Readout bandwidth

Bandwidth increases

Bandwidth decreases

The data bandwidth

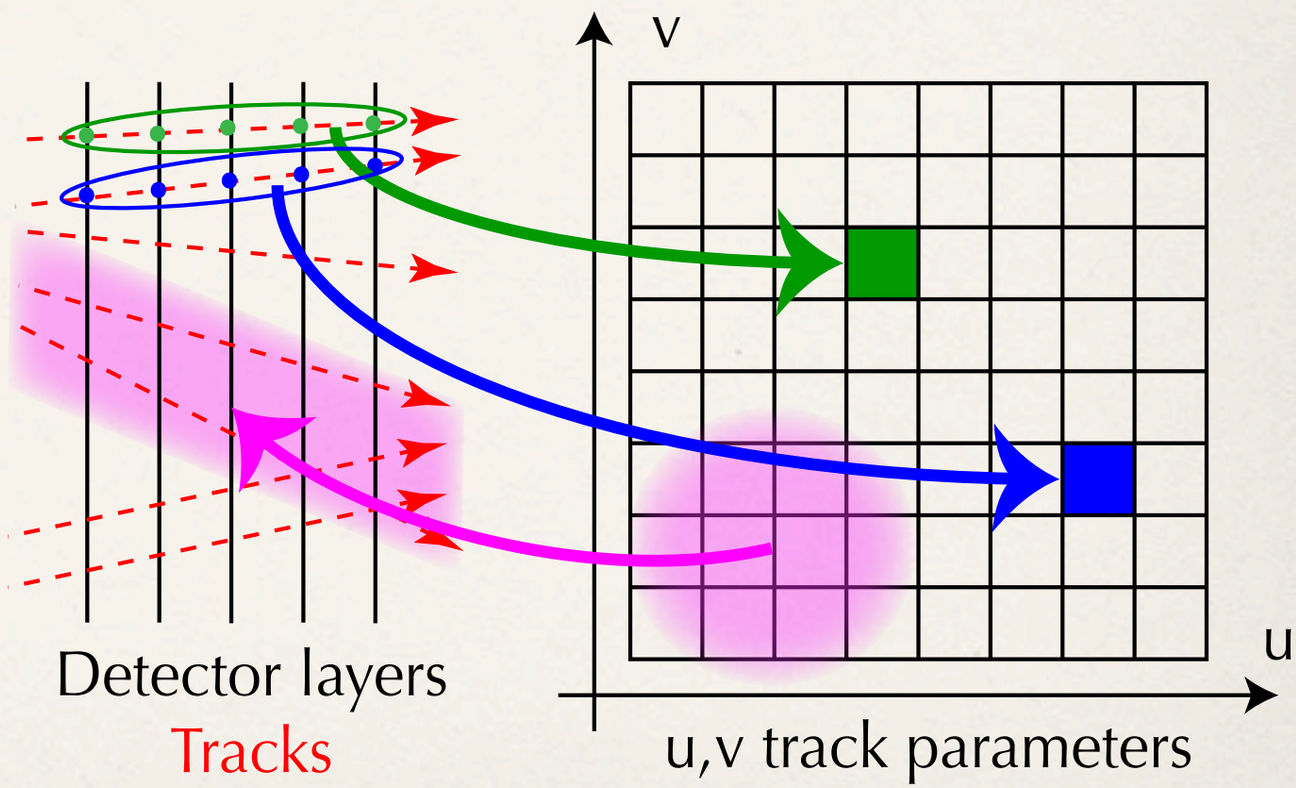
- DAQ and trigger electronics work always reducing the data bandwidth, like a sort of funnel
- Switching duplicates hits, increasing the bandwidth
- Engines send out only the reconstructed tracks, shrinking down the bandwidth to a value lower than before switching
- Curiously we have evidence of similar process used by the brain for visual data



A "neural-like" tracking algorithm (1)

- Algorithm proposed by Luciano Ristori [NIM A453 (2000) 425-429] inspired to visual apparatus of mammals (from here the name **Artificial Retina**). Similarities with:
  - Hough transform until 2D, but computationally simpler with more dimensions
  - Associative memories for pattern matching, but analog responses using cells interpolation, implying similar or better resolution with lower number of stored patterns

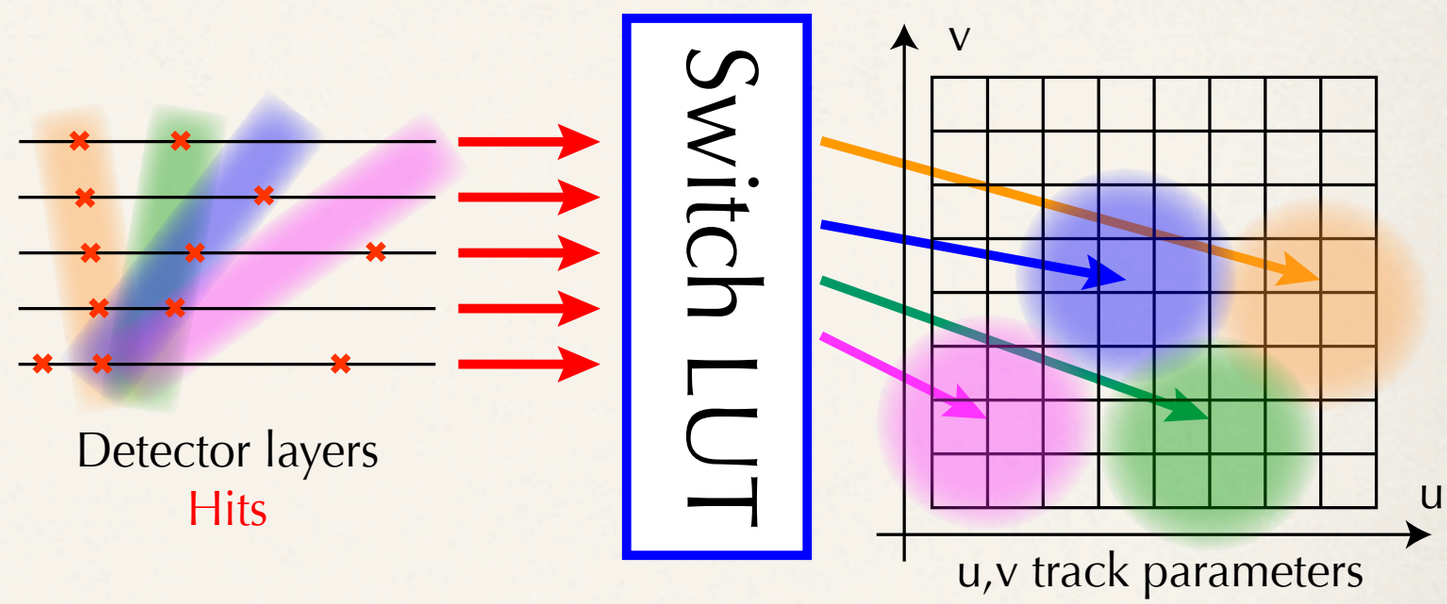
- **Configuration phase** (common PC):
  1. Discretize space of track parameters (**cells**)
  2. Mapping 1: generate track intersections with detector planes (**receptors**) and connect them to cells
  3. Mapping 2: assuming contiguous cells corresponding to slightly different tracks, we connect cluster of cells to areas of detector readout



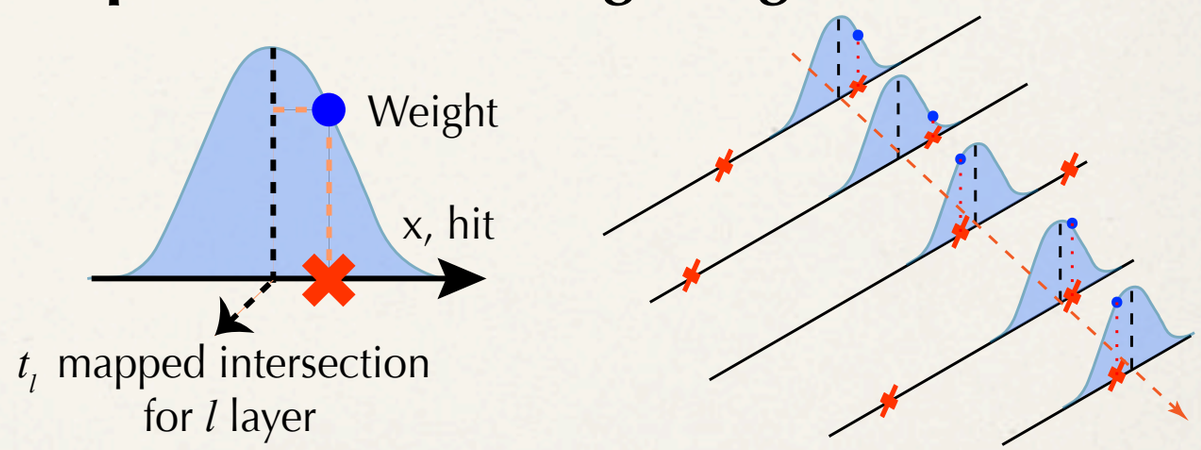
A "neural-like" tracking algorithm (2)

- **Track Processing** (running in real-time on high-speed device)
- Step 1: detector hits are distributed only to a reduced number of cells according the mapping 1 (LUT)
- Step 2: a logic unit (engine) for each cell accumulates a Gaussian weight proportional to the distance with the receptors
  - executed in parallel for each cell
  - σ is an algorithm parameter, adjusted to optimize the sharpness of the response

Step 1: Hits delivery through a switch



Step 2: Accumulating weights (each cell)



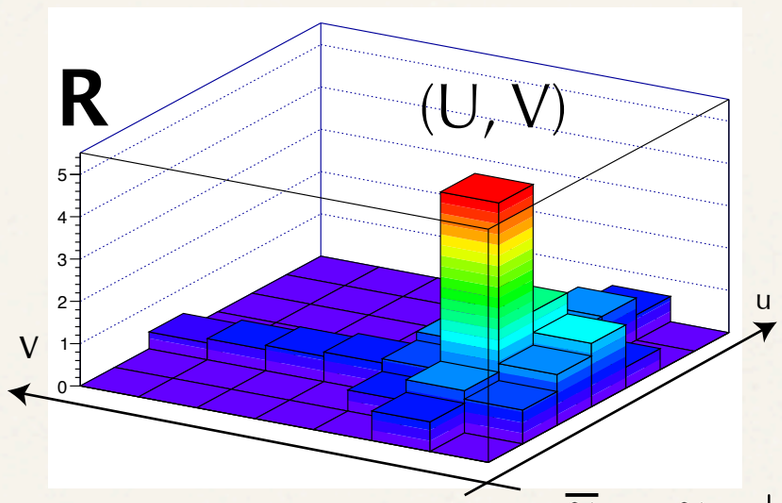
$$R = \sum^{\text{hits}} e^{-\frac{(x_l - t_l)^2}{2\sigma}}$$

R is close to N (# of layers) only if we have a set of hits near the mapped track

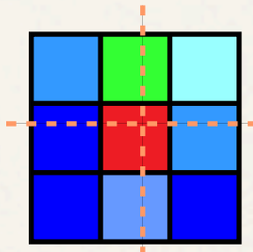
A "neural-like" tracking algorithm (3)

- Step 3: tracks are identified as local maxima of accumulated weights, above a certain threshold, over the cells grid
  - High granularity not required, if centroid is computed over 3x3 cluster
  - Immediate track parameters estimate

Step 3: Find the local maxima and compute centroid



3x3 cluster

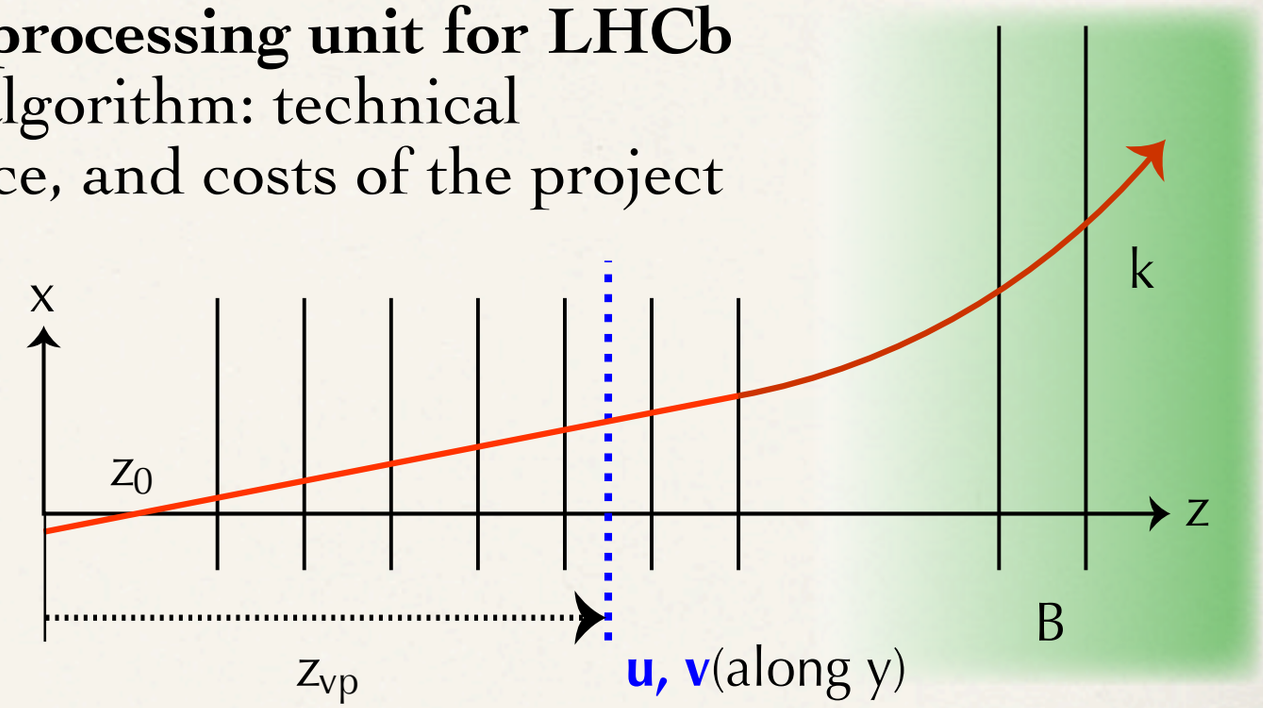


$$\overline{u} = u_0 + \frac{\sum_{ij} iR_{i,j}}{\sum_{ij} R_{i,j}} \qquad \overline{v} = v_0 + \frac{\sum_{ij} jR_{i,j}}{\sum_{ij} R_{i,j}} \qquad \begin{matrix} i=U-1,U,U+1 \\ j=V-1,V,V+1 \end{matrix}$$

- **Parallelization at two levels**:
  - Hits delivered to a reduced number of cells that process them simultaneously
  - Separate events can be processed if timestamp is associated to hits and cells have multiple registers for accumulating weights

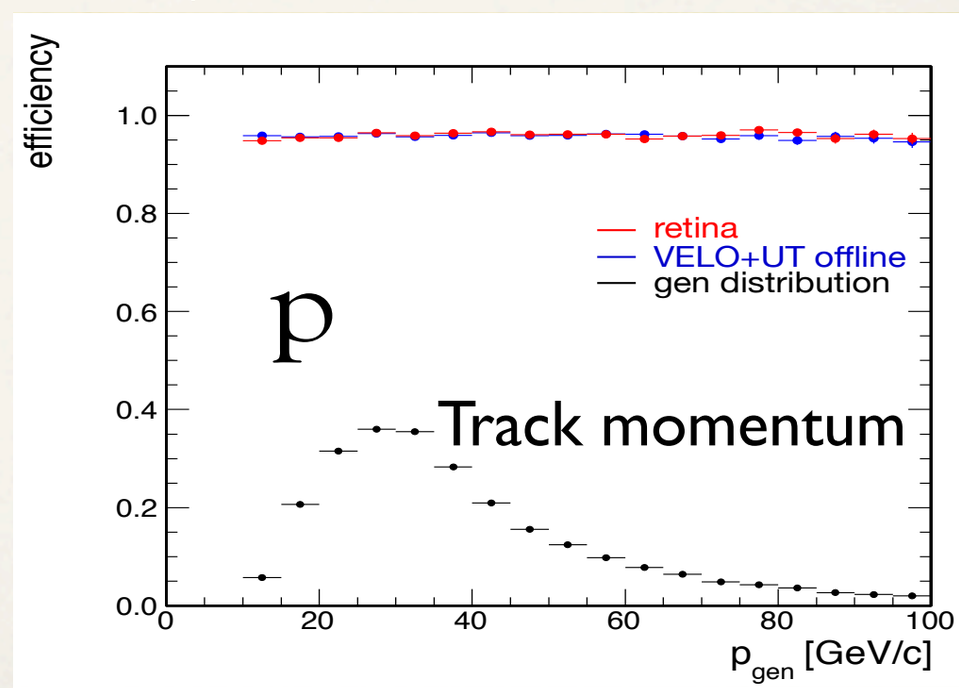
Early Study (2014)

- Study supported by CSN1 for a **track processing unit for LHCb Upgrade** based on the artificial retina algorithm: technical implementation, simulation, performance, and costs of the project



- Input data from pixel and strip detectors
- Track parametrized using 5 variables: u, v, z_0, d (impact parameter), k (curvature)
- Efficiency and resolution similar to offline reconstruction
- Feasibility using FPGA's: OK

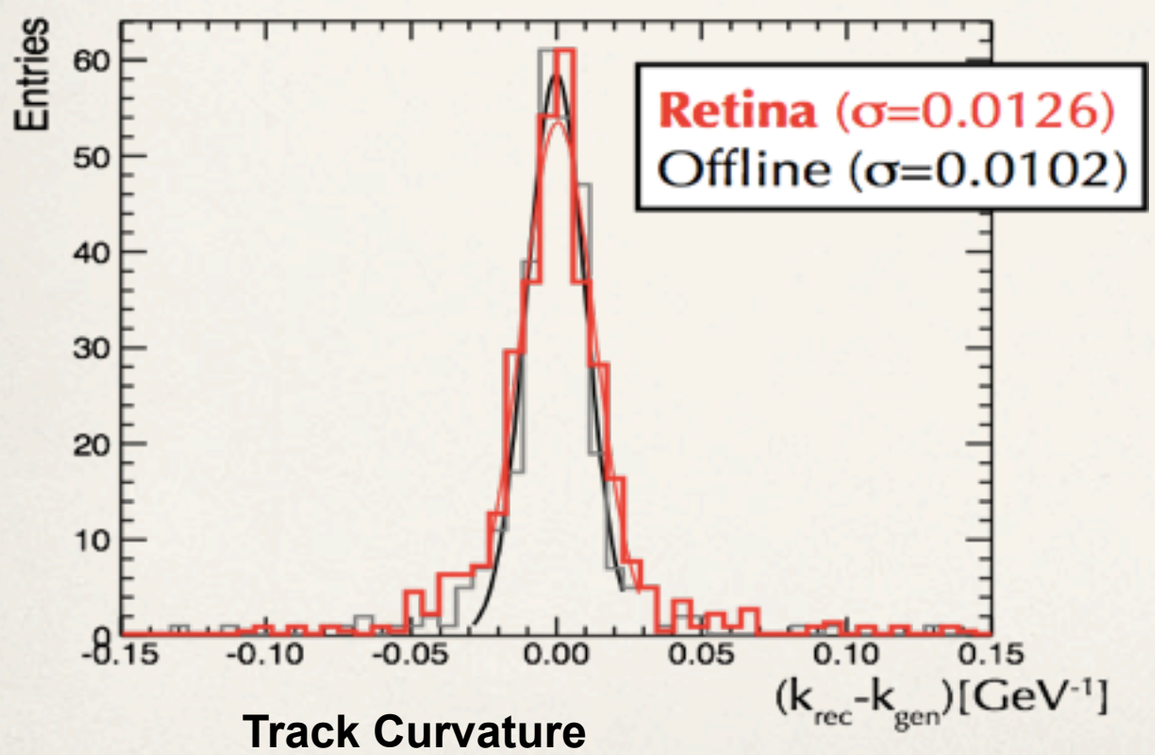
CERN-LHCb-PUB-2014-026
https://cds.cern.ch/record/1667587

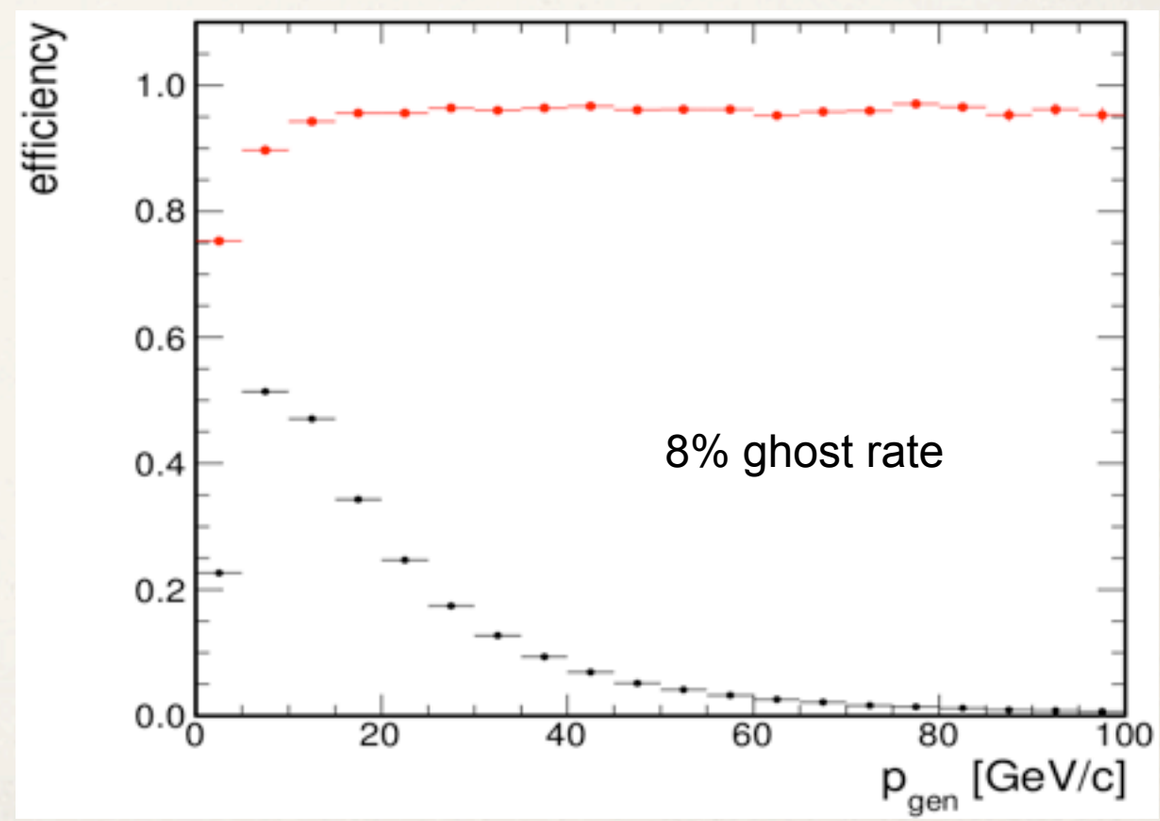


Early Study (2014)

- **Performances are similar to offline with a feasible number of cells**, e.g., with a pixel detector and 50k cells (50 Stratix V FPGA's)

- Estimated O(100) MHz tracks / FPGA at a reasonable cost

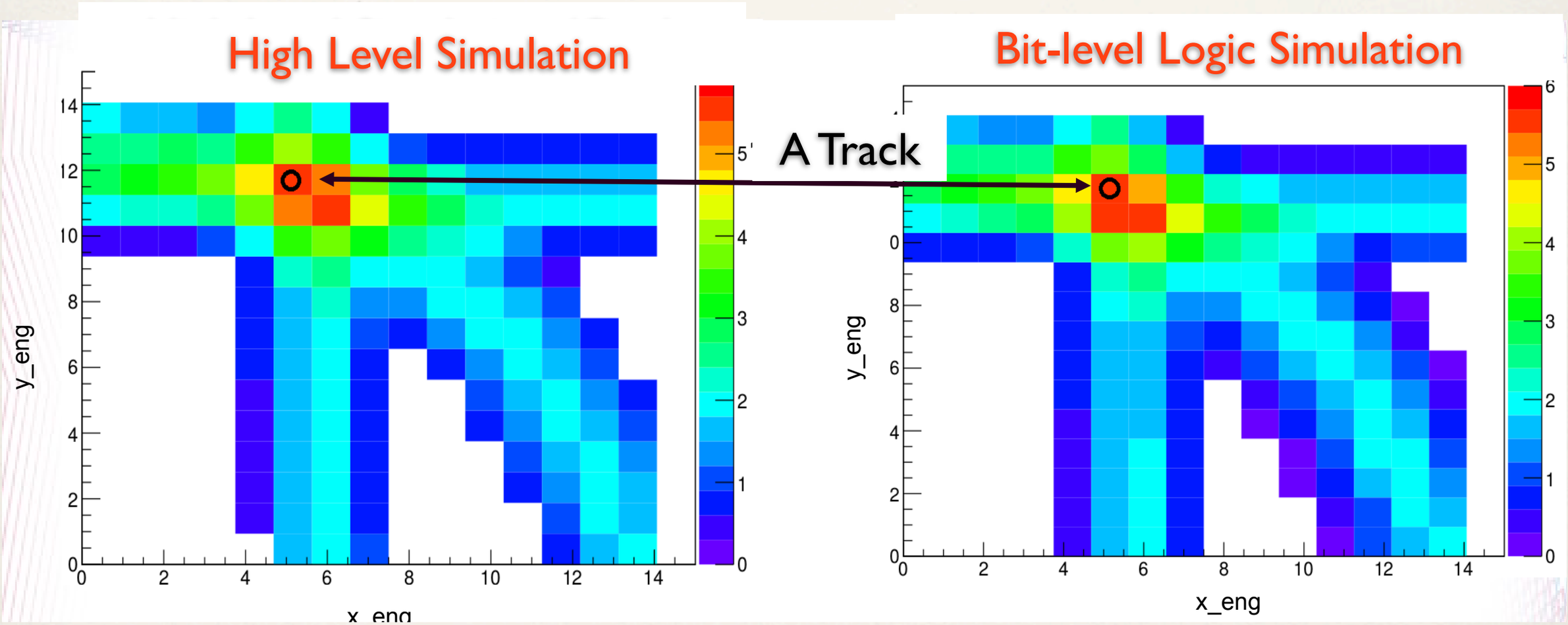




Abba et al., JINST 10 (2015) 03, C03008

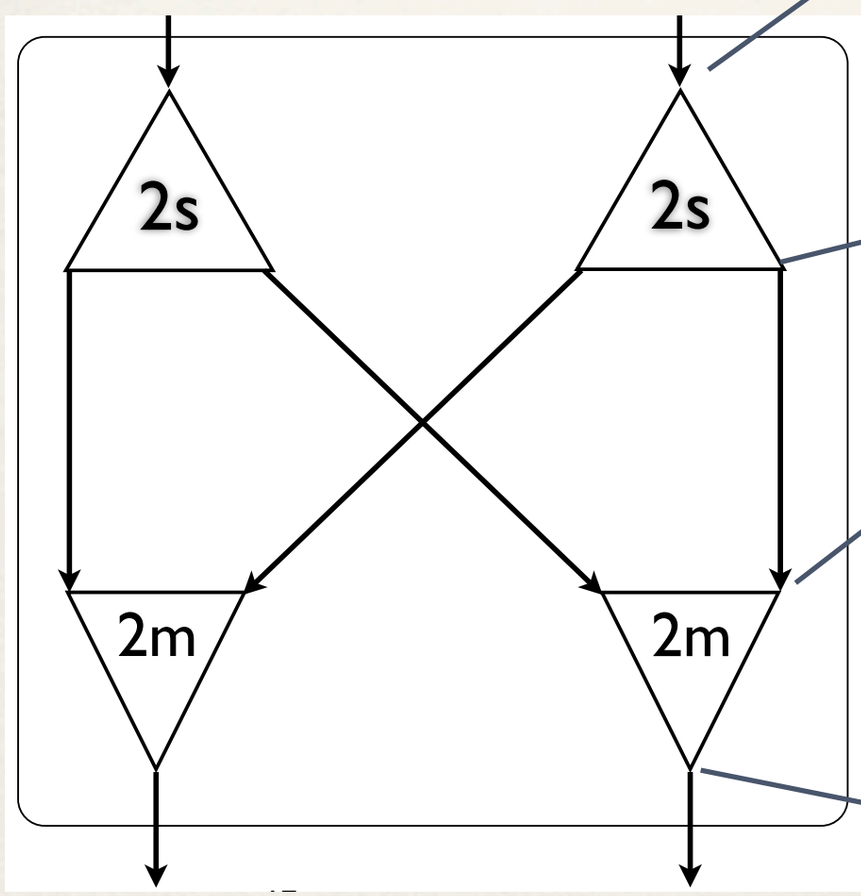
Engine Simulation

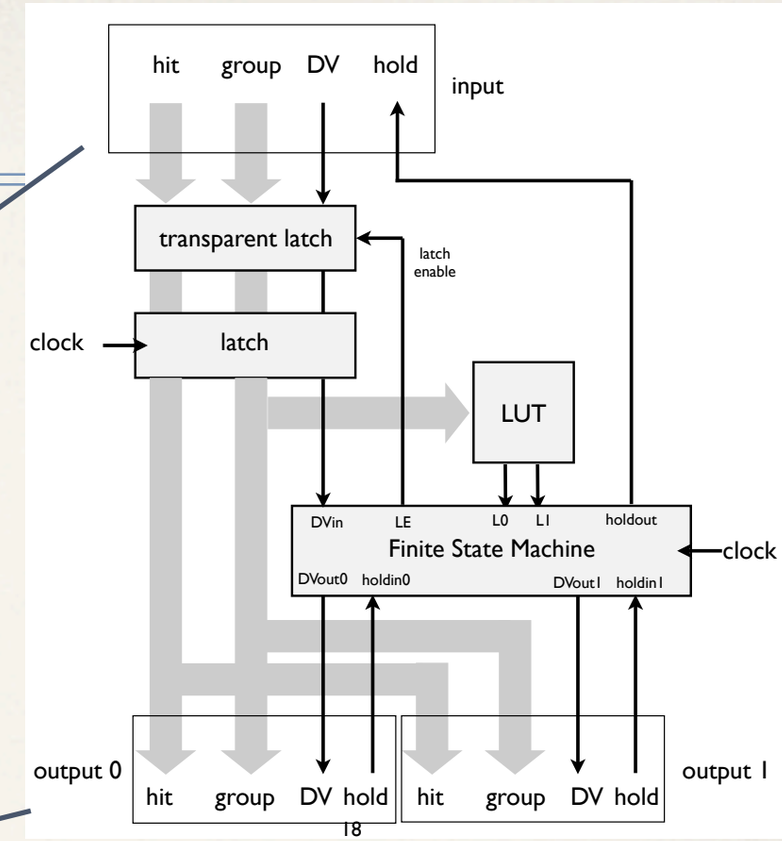
- Accumulators value on the cells matrix:
  - C++ simulation output and logic simulation output (ModelSim) show negligible differences, due to integer calculation inside the FPGA

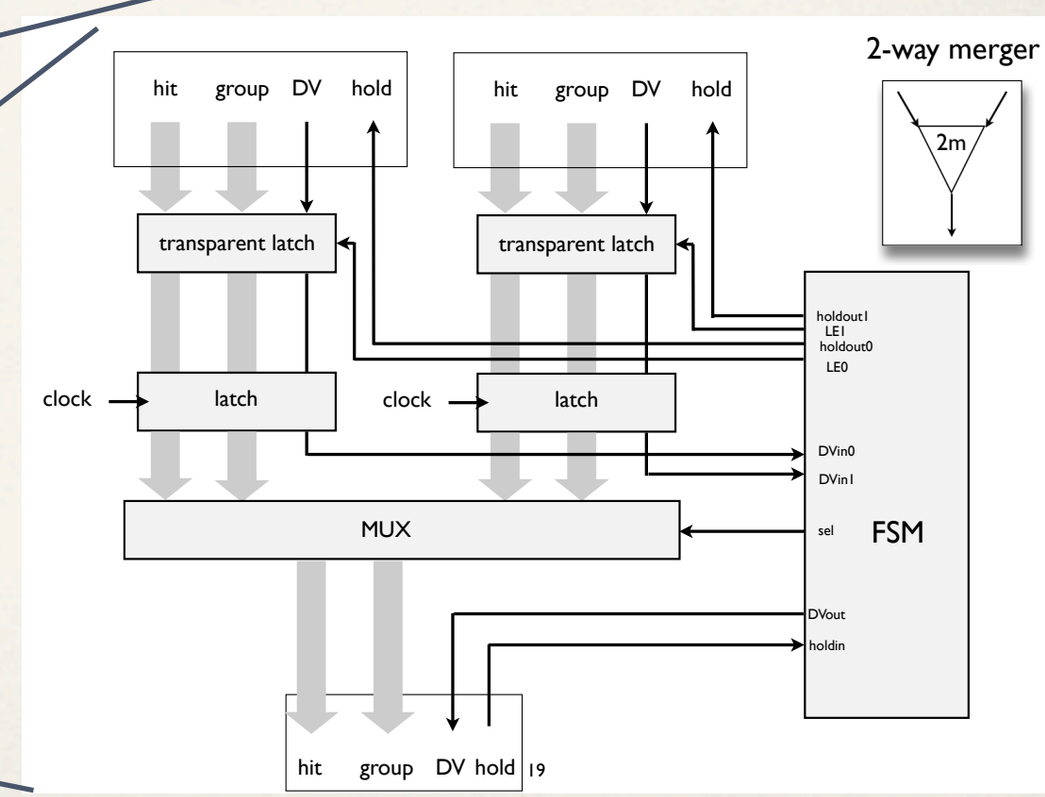


Basic Dispatcher

- **Splitter**: latch, LUT, control fsm
- **Merger**: 2 latches, mux, control fsm

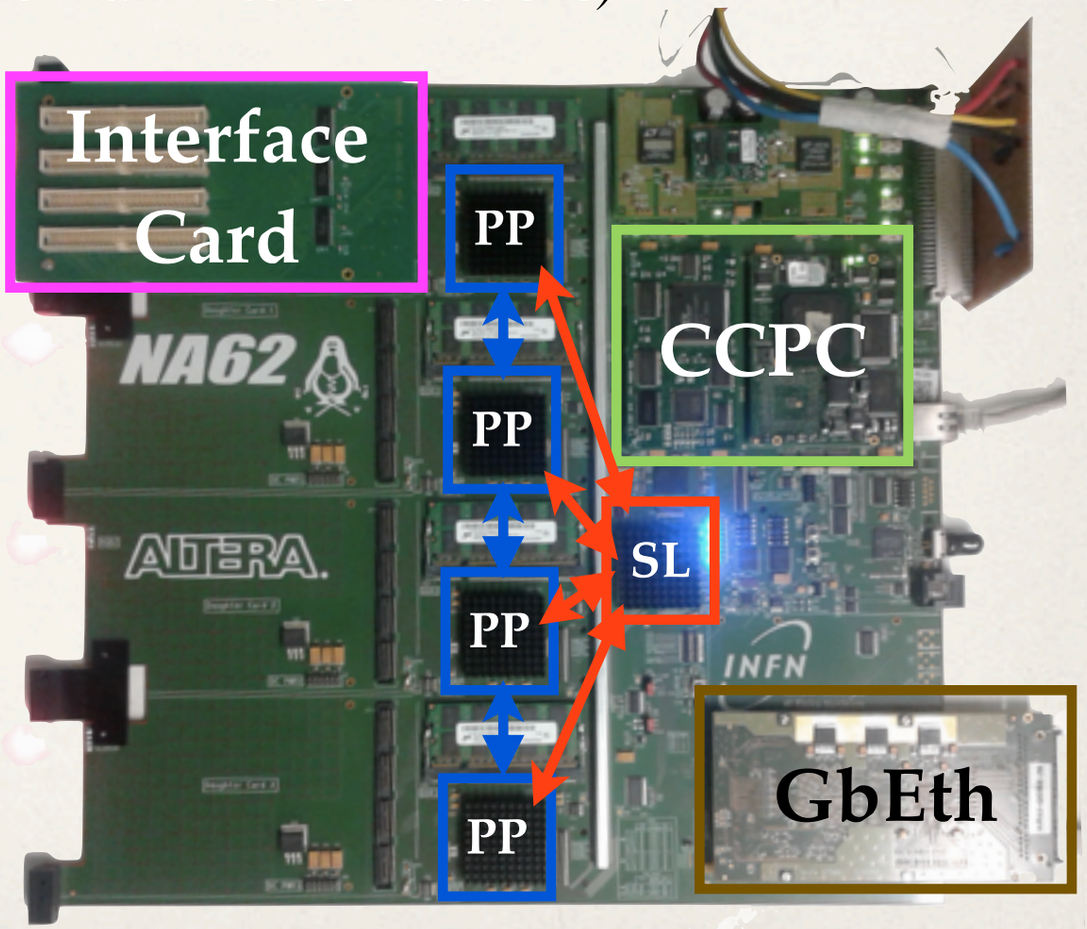






The Tel62 board

- **DAQ board already used in a HEP experiment to keep the prototype for embedded reconstruction as realistic as possible**
- Thanks to our NA62 colleagues for letting use the board and the firmware/software
- Five Altera Stratix III FPGA (200k LE): 4 chips to process data (**PP**), 1 chip to control (**SL**)
- Clock: 40 MHz (main), 160 MHz (inside chips and main interconnections)
- Fast interconnections between SL and PP's (also lateral between PP's themselves)
- Other features:
  - 2-GB DDR2 memory per PP chip
  - **Onboard mini PC (CCPC)**, slow control
  - **Output mezzanine with 4x 1Gb Ethernet link**
- We design an additional **interface card** to connect two boards together at high speed

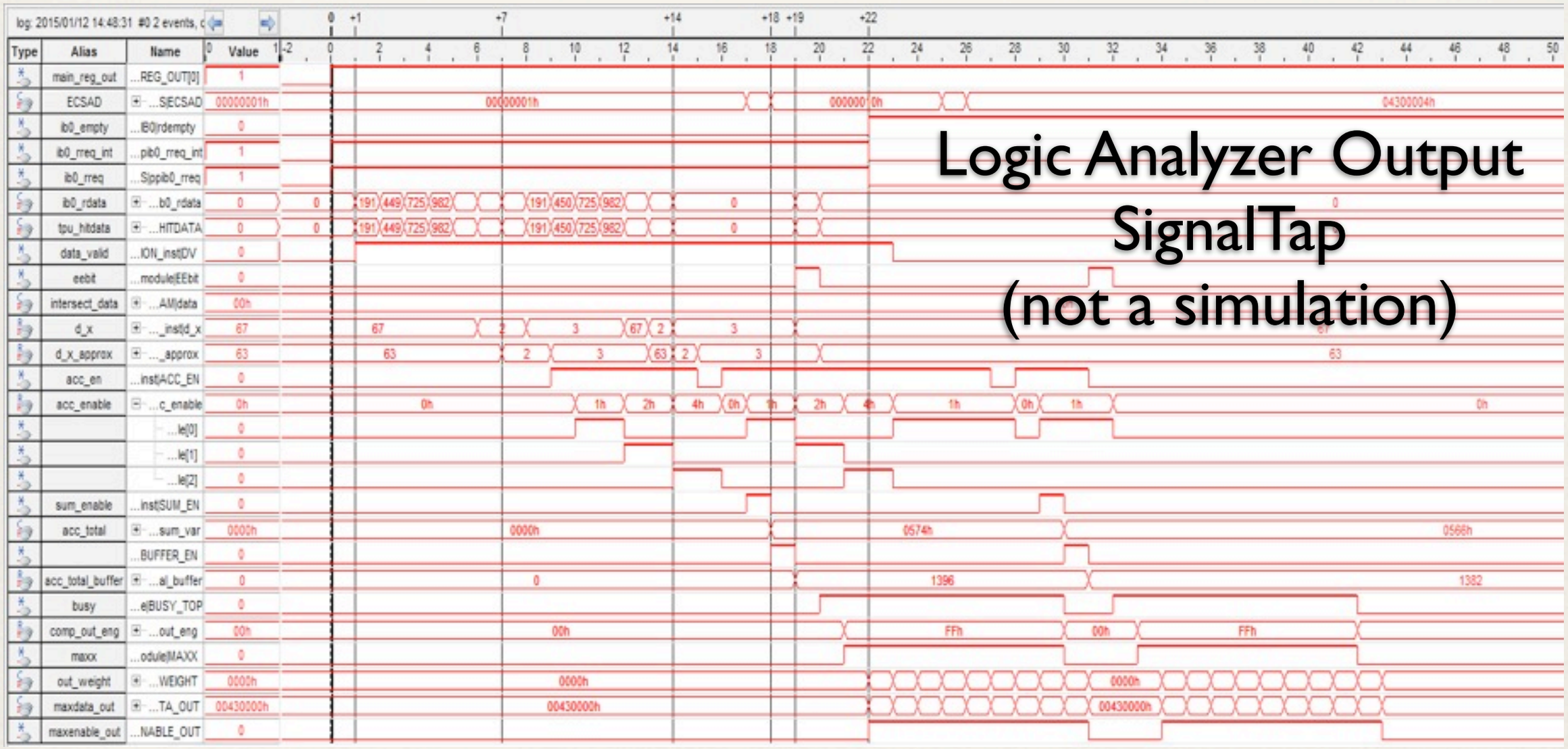


B Angelucci *et al* 2012 *JINST* 7 C02046

Tel62 crate

Test with internal logic analyzer

- Engine running at 160 MHz



Hardware Details

- Typical event rate is **40 MHz**, typical bandwidth **>10 Tbps**
- Need a board with:
  - High clock frequency (~0.5GHz)
  - Large FPGA (~1M logic elements)
  - >10 high-speed serial links (10 Gbps)
  - Ready to buy
- **PCIe40 board** (Altera Arria 10, 48 I/O links @ 10Gbps, LHCb DAQ) or similar commercial boards
- **Required at least 2 boards (DAQ & Switch/Engine) on 2 PC's**, plus small patch panel for a minimal prototype of two nodes attached to the Event Builder system