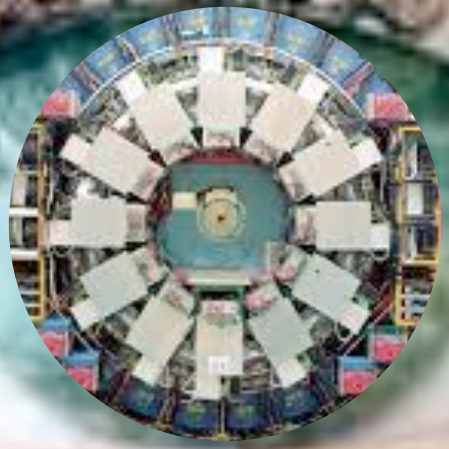
The artificial retina processor for tracking at 40 MHz



A. Abba, F. Bedeschi, F. Caponio, M. Citterio, A. Cusimano, A. Geraci, P. Marino, M.J. Morello, N. Neri, D. Ninci, M. Petruzzo, A. Piucci, G. Punzi, L. Ristori, F. Spinella, S. Stracka, D. Tonelli

(Pisa/Milano/FNAL/CERN)

WIT2014, Philadelphia, May 15, 2014

The issue

	Technol.	Experim.	Year	Rate	Clock	Cycles/ evt	Latency
SVT	AM	CDF-L1	2000	0.03 MHz	40 MHz	≈ 1600	< 20μs
FTK	AM	ATLAS-L1	2014	0.1 MHz	≈ 200 MHz	≈ 2000	O(10) μs
?	?	LHC-L0	2020	40 MHz	≈1 GHz	25	few μs

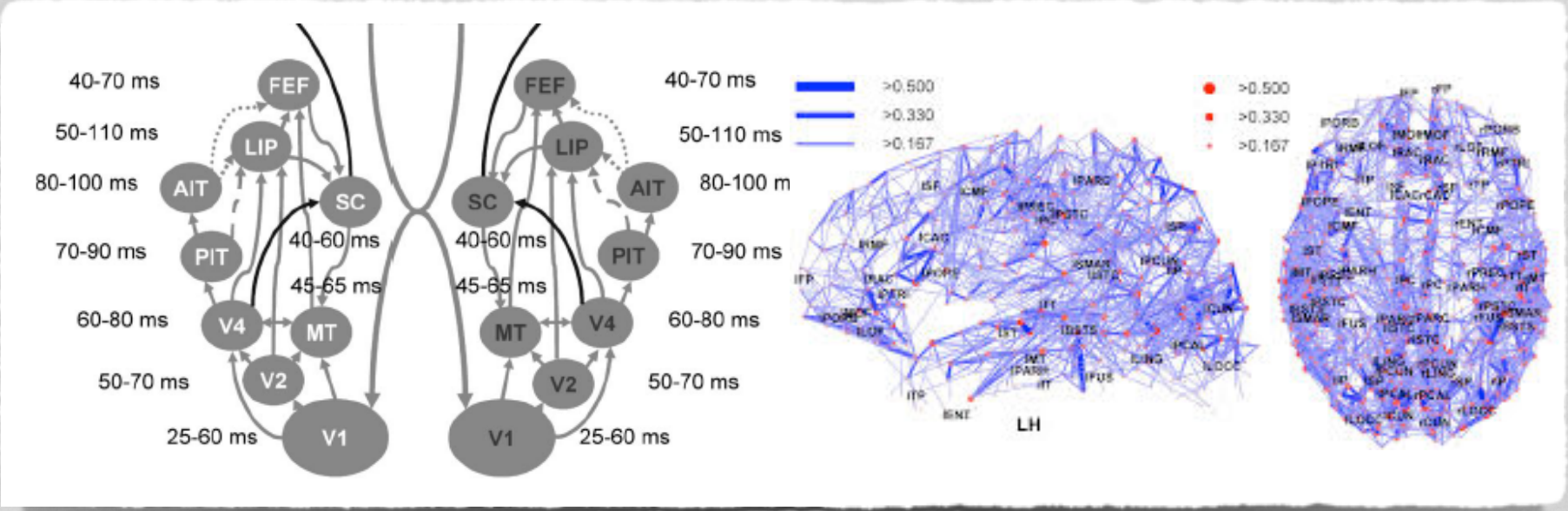
Perform tracking synchronous with LHC collisions appears daunting.

Any complex tracking calls for O(1000) clock cycles per event

No known example of a system capable of nontrivial pattern recognition in O(25) time units.

Well...

..maybe one can think of one example...



Early visual areas in the human brain produce a recognizable sketch of the image in about 30 ms.

Maximum neuron firing frequency is about 1 kHz ==> 30 time units

Far fetched? Experimental evidence that V1 functionality can be quantitatively modelled as a trigger. MM Del Viva, G. Punzi et al., D PloS one (2013)

How?

What makes the brain algorithm special?

Parallelism. But AM-based devices have a very parallel architecture as well.

Differences:

- Detector hit readout into the AM still proceeds serially. The visual system does not seem to have such serialization thus gaining processing power through connectivity.
- AM matches patterns against fixed templates whereas the brain interpolates among analog responses. Saves lots of internal storage. Makes it easier to handle “missing layers”

Can these features be engineered into a viable tracking system?

The algorithm

NIM A453, 425 (2000)

An artificial retina for fast track finding

Luciano Ristori

INFN, Sezione di Pisa, Via Livornese 1291, I-56010 S. Piero a Grado, Pisa, Italy

Accepted 21 June 2000

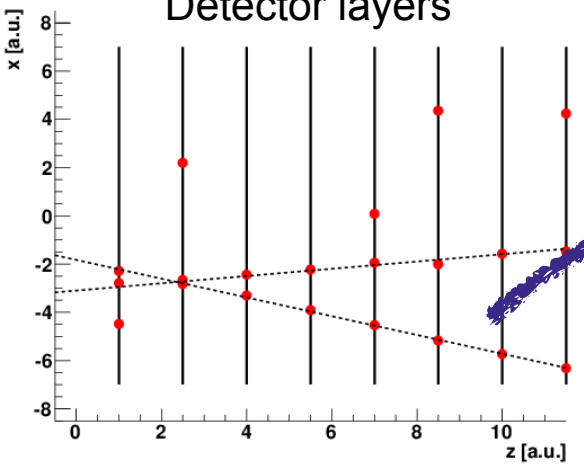
Abstract

A new approach is proposed for fast track finding in position-sensitive detectors. The basic working principle is modeled on what is widely believed to be the low-level mechanism used by the eye to recognize straight edges. A number of receptors are tuned such that each one responds to a different range of track orientations, each track actually fires several receptors and an estimate of the orientation is obtained through interpolation. The feasibility of a practical device based on this principle and its possible implementation using currently available digital logic is discussed.

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

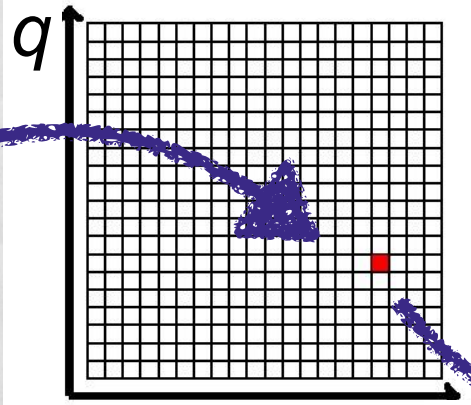
hit

Detector layers

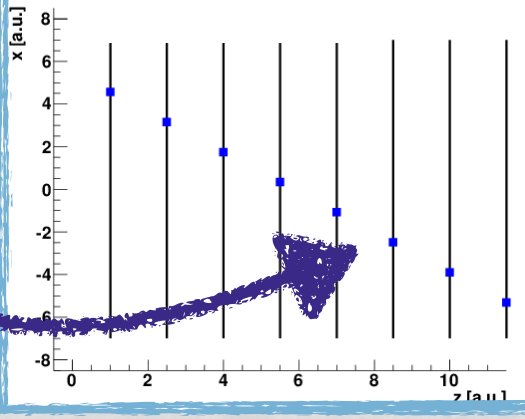


$$x = mz + q$$

Discretize space of track parameters in *cellular units*



Each unit corresponds to a track that intersects each layer in *receptors*



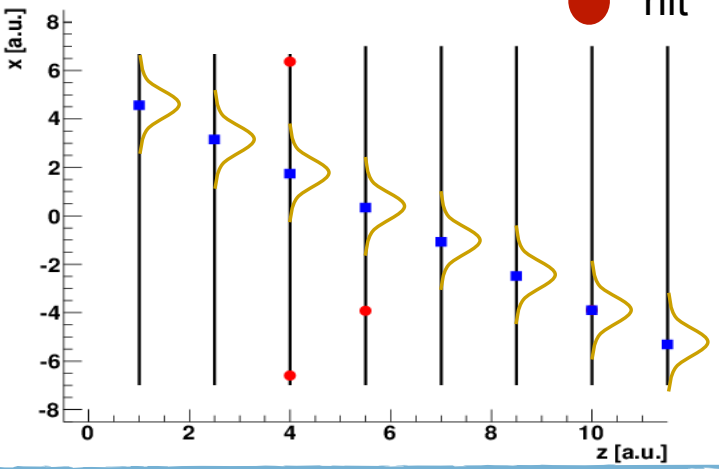
In a detector layer, the distance *s* between the hit and the receptor is used to compute the contribution of that hit to the excitation of the cellular unit.

Then sum over all hits, all layers, to have the excitation of one cell (ij)

$$R_{ij} = \sum_{k,r} \exp\left(-\frac{s^2_{ijkr}}{2\sigma^2} \right)$$

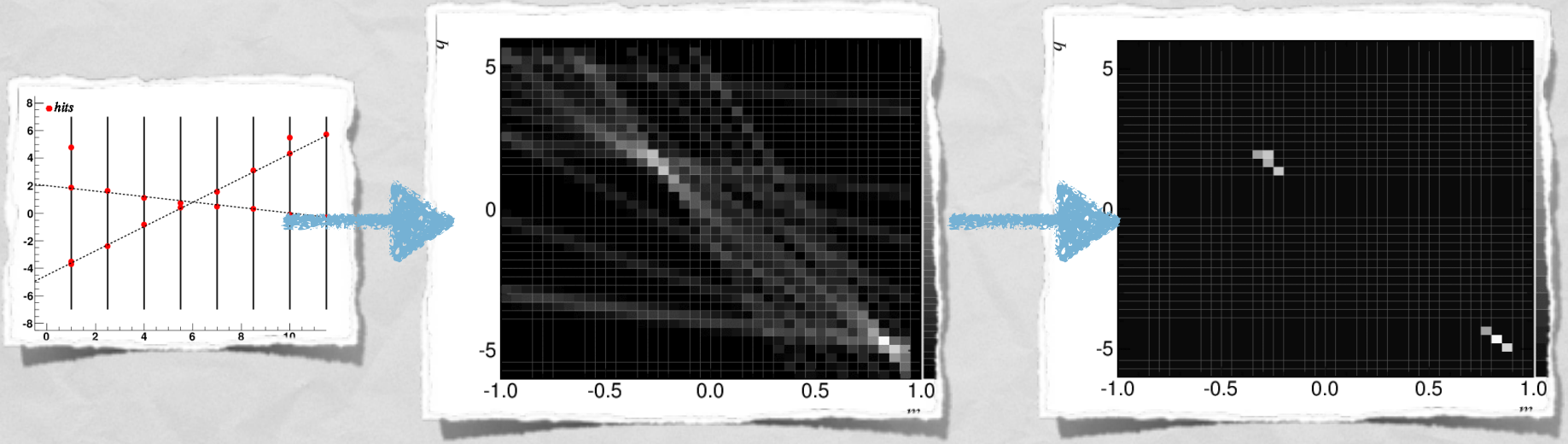
receptor

hit



“Retina” response

The response Rij of all the cells yields the response of the retina



Detector
(physical space)

Retina
(parameter space)

Zero suppressed

A track is identified by a local excitation-cluster.
Parameters determined accurately interpolating nearby cells

Comments

Not new, really. Designed and proved conceptually feasible in a toy 2D tracker 15 years ago, but unviable for 90s electronics.

Concept closely related to Hough transf. P.V.C. Hough Conf. Proc. C590914, 54 (1959)

However, a few important original features.

- Not just yes/now response. Each cell receives a signal that is a smooth function of hit positions. Used as weight to interpolate track parameters with better resolution than grid step
- Neural communication btw nodes allows massive parallelism.

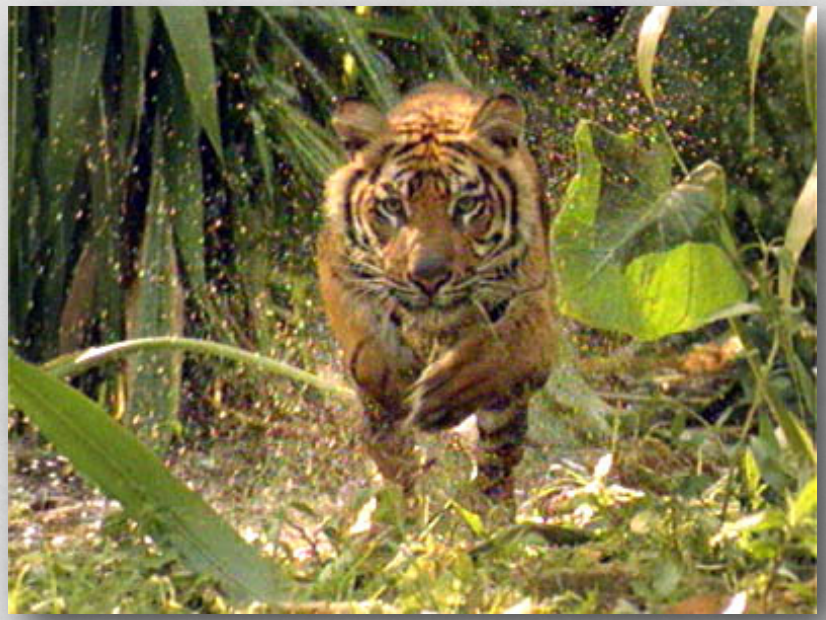
I am going to shown a realistic implementation on a realistic pixel detector, with existing electronic components.

Implementation challenges

LHCb-like scenario as benchmark: O(1000) pixel+strip hits distributed over O(10) layers to reconstruct O(100) tracks Every 25 ns.

A few Tb/s data flow.

- Switching: route each detector hit to those cells for which that hit is relevant (possibly only those)
- Pattern recognition: identify clusters of excited cells to distinguish genuine tracks from random combinations of hits.

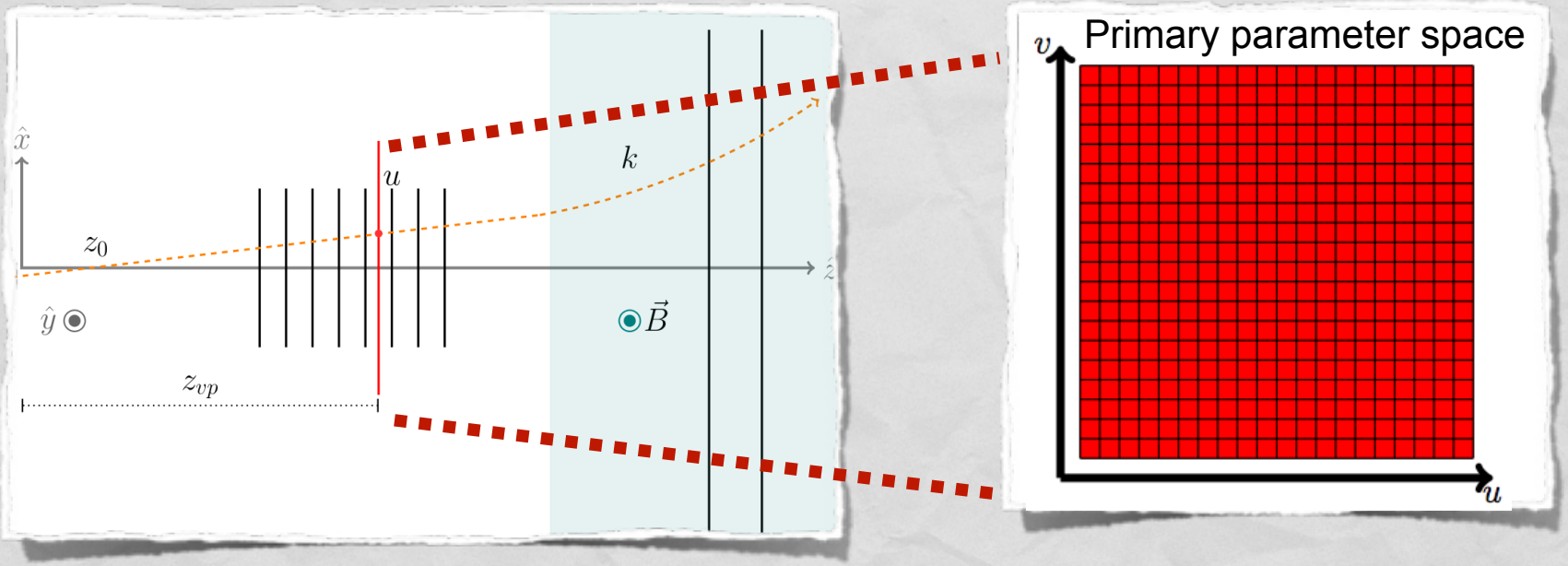


Lots of data, little time

Layout

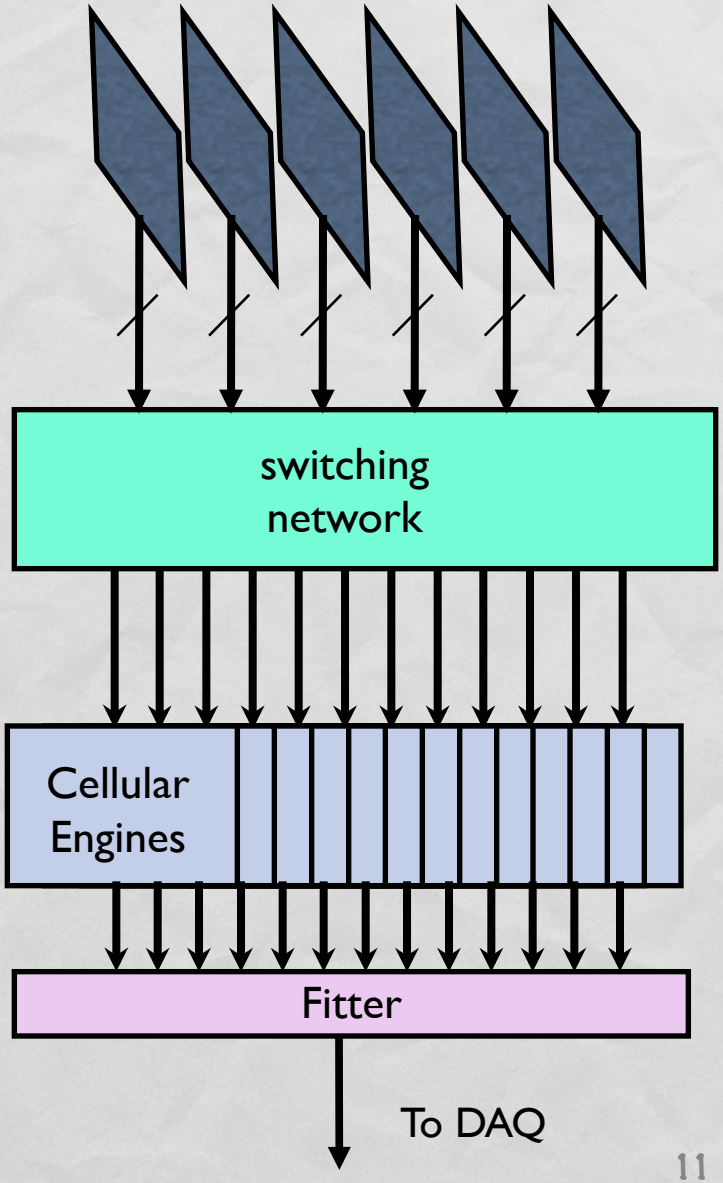
Geometry has significant impact on implementation

LHCb-like forward spectrometer with pixel and strip detectors



- Do tracking in a volume where B field is weak. Approximate tracks to straight lines originating from a single point. This identifies a primary 2D-plane to perform pattern recognition.
- Treat momentum and origin of the track as perturbations

Architecture



Tracking layers

Separate trigger-DAQ path

Custom switching network
delivers hits to appropriate cells

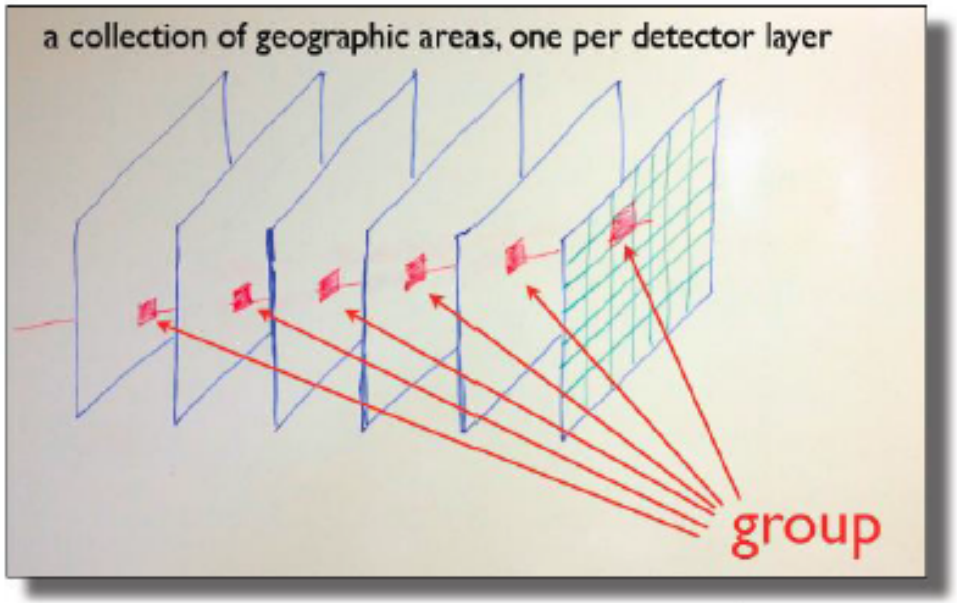
Data organized
by cell coordinates

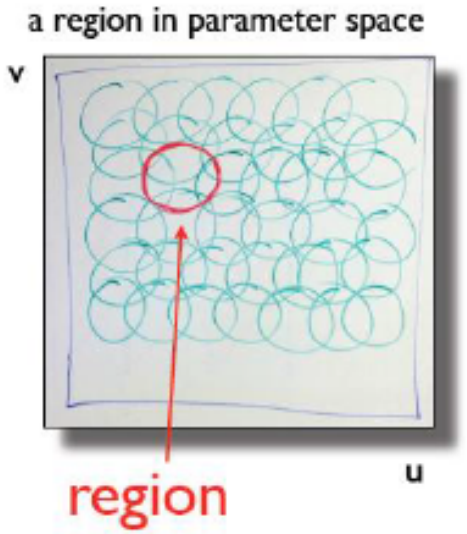
Blocks of cellular
processors

Track finding and
parameter determination

Switching concept

Compact regions in detector layers map into compact regions in parameter space, which have limited overlap with one another





Each hit can only belong to one group.

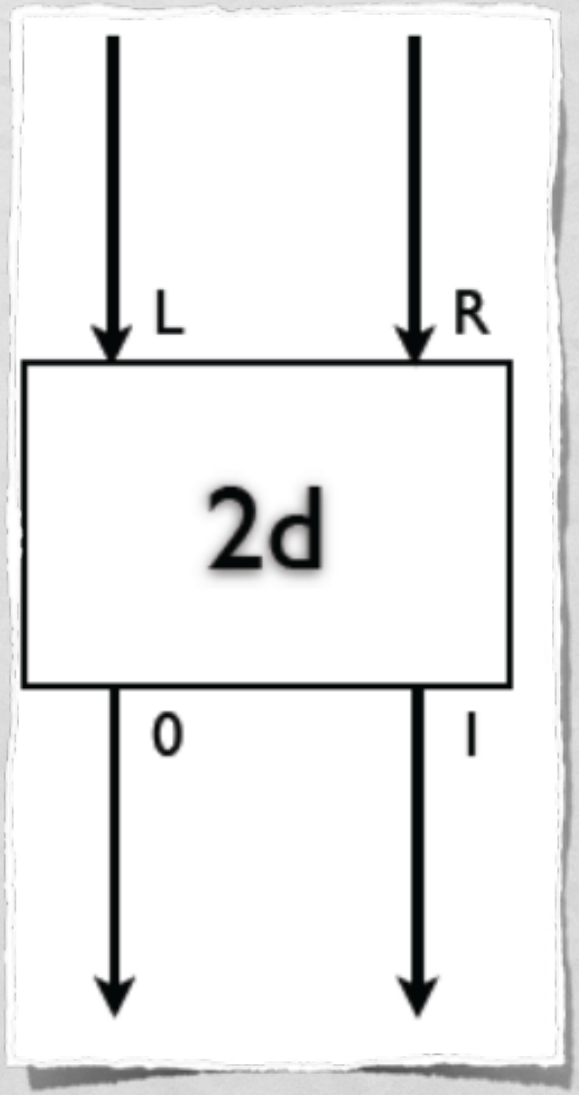
Each hit is only delivered to the union of all cells affected by all the hits in that group: the region associated with that group.

Switching basic unit

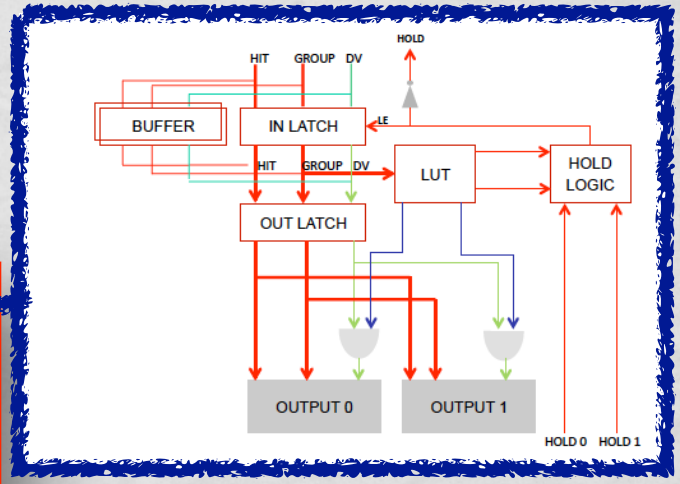
Information is carried by hits: 41-bits word encoding hit coordinates, layer ID, timestamp...

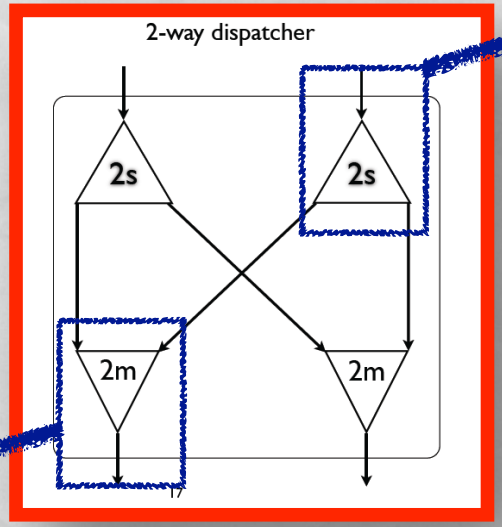
Two-way dispatcher

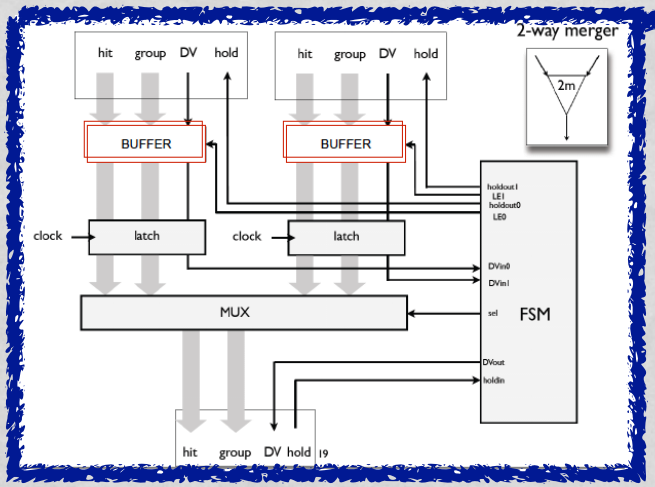
- Merges left and right inputs.
- Dispatches to one or both outputs according to a look-up table addressed by the hit's group #.
- If a stall happens downstream inputs may be held.



Look inside..


HOLD
HIT
GROUP
DV
BUFFER
IN LATCH
LE
HIT
GROUP
DV
LUT
HOLD LOGIC
OUT LATCH
OUTPUT 0
OUTPUT 1
HOLD 0
HOLD 1


2-way dispatcher
2s
2s
2m
2m


2-way merger
hit
group
DV
hold
hit
group
DV
hold
2m
BUFFER
BUFFER
clock
latch
clock
latch
MUX
FSM
sel
hit
group
DV
hold

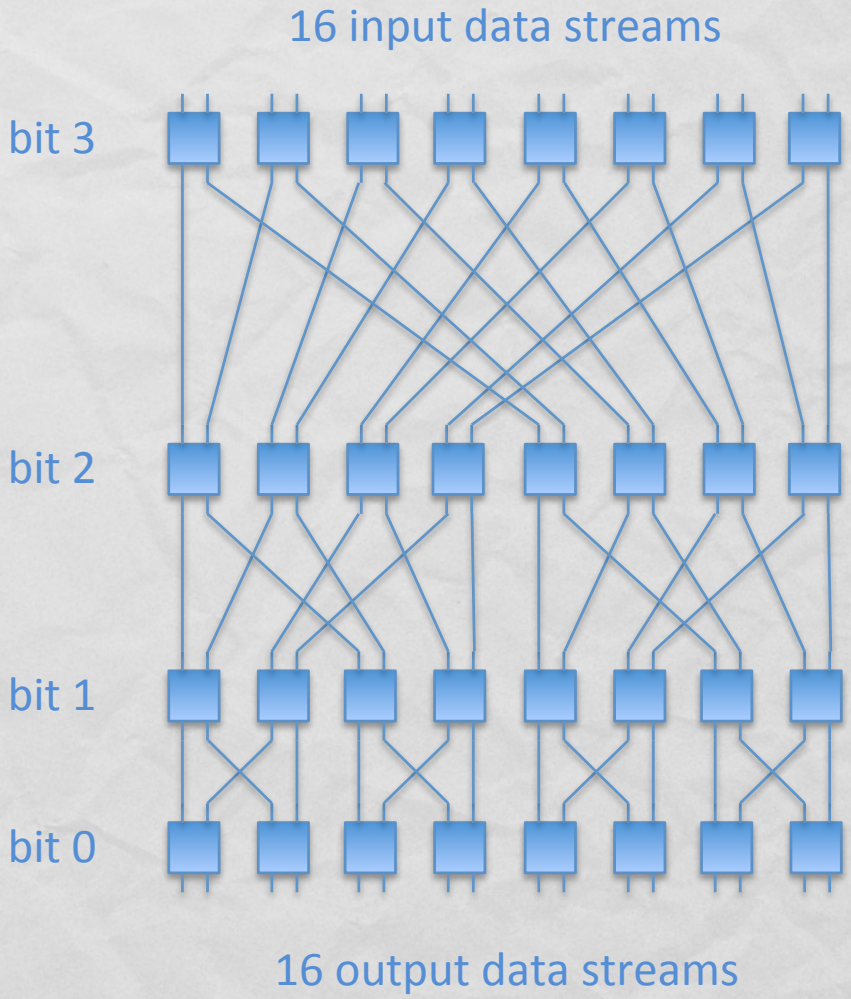
Network

Combination of dispatchers builds whole network.

N×N requires $(N/2)*\log_2(N)$ elements

Each hit comes with a “zip-code”

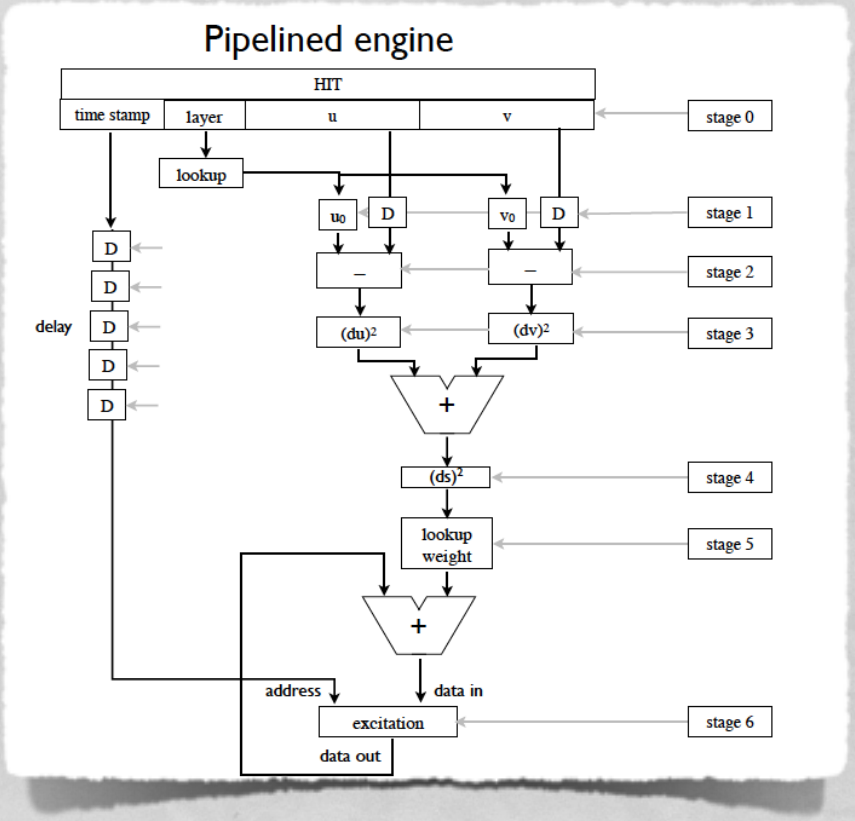
The switching network “knows” where to deliver it, according to programmable maps distributed over the nodes. embedded



The engine

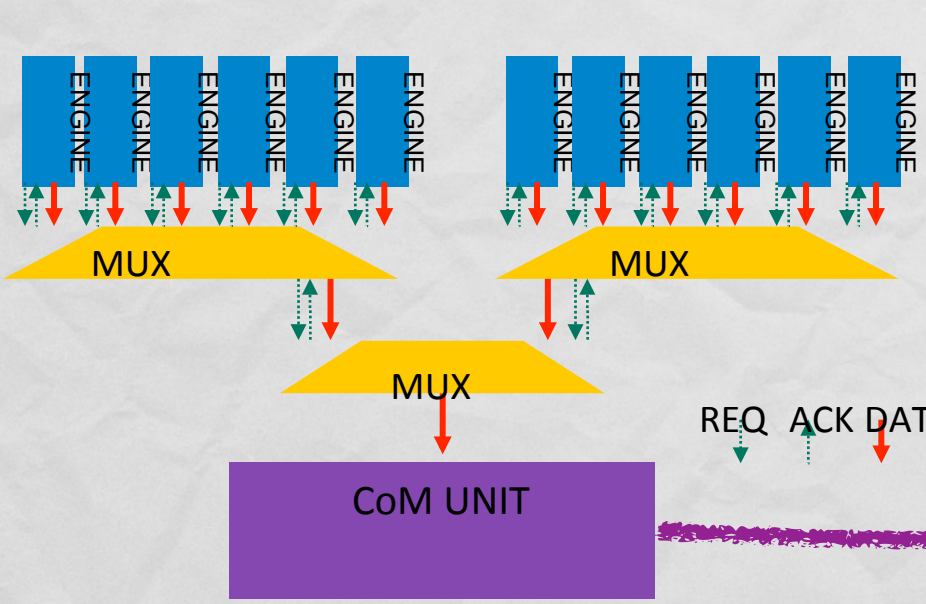
Logic module of the cell. Implemented as a clocked pipeline

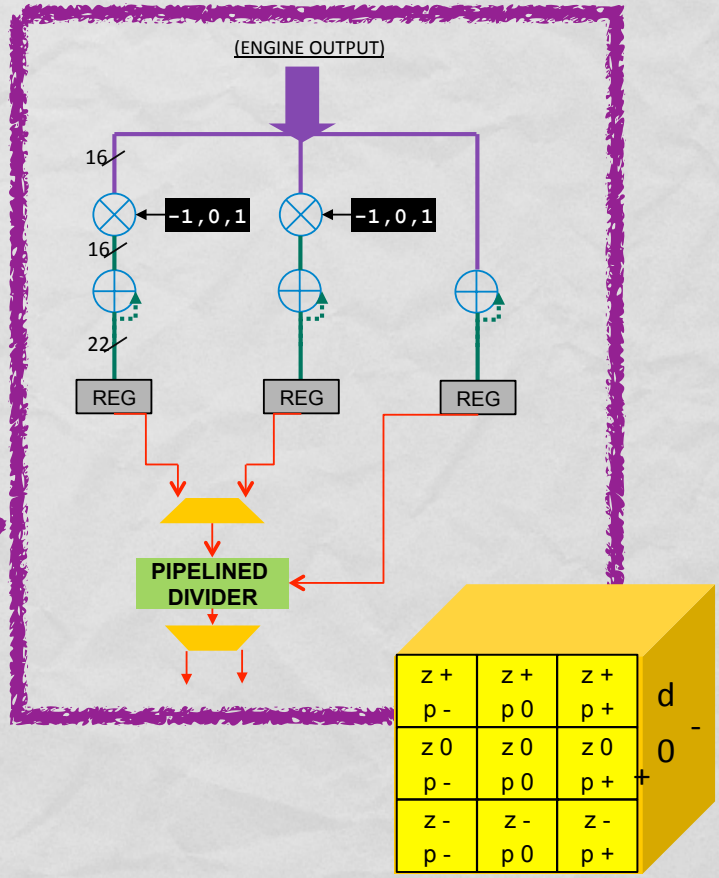
- Layer ID determines the coordinates of the receptor center to be subtracted from hits' coordinates.
- Outcome squared and summed. The result R is rounded
- A weight function common to all engines mapped in a LUT
- Rounded result is used as address to the LUT.
- LUT outputs accumulated for each hit of the event



Each hit is cycled multiple times to compute excitation in lateral cells.

Clustering





Data reduction within the engine allows gaining a O(10) reduction to keep up with data flow.

Second stage: performs local clustering (center of excitation) in parallel and queues results to output

Implementation

High-end FPGA devices. Combined advantage of high bandwidth and being the standard chip for LHCb Run II DAQ.

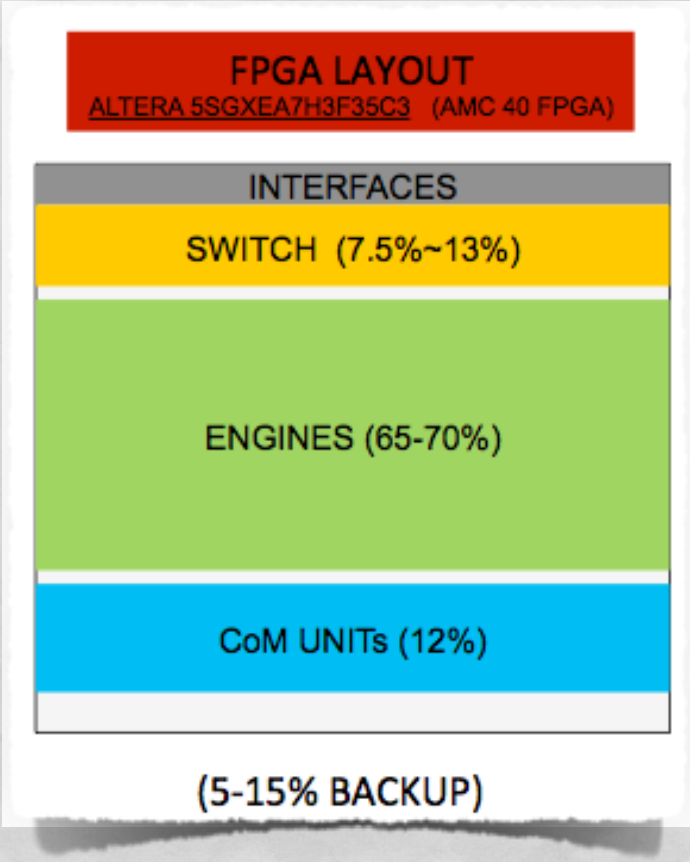
Placing

All main components implemented in VHDL and placed on the FPGA

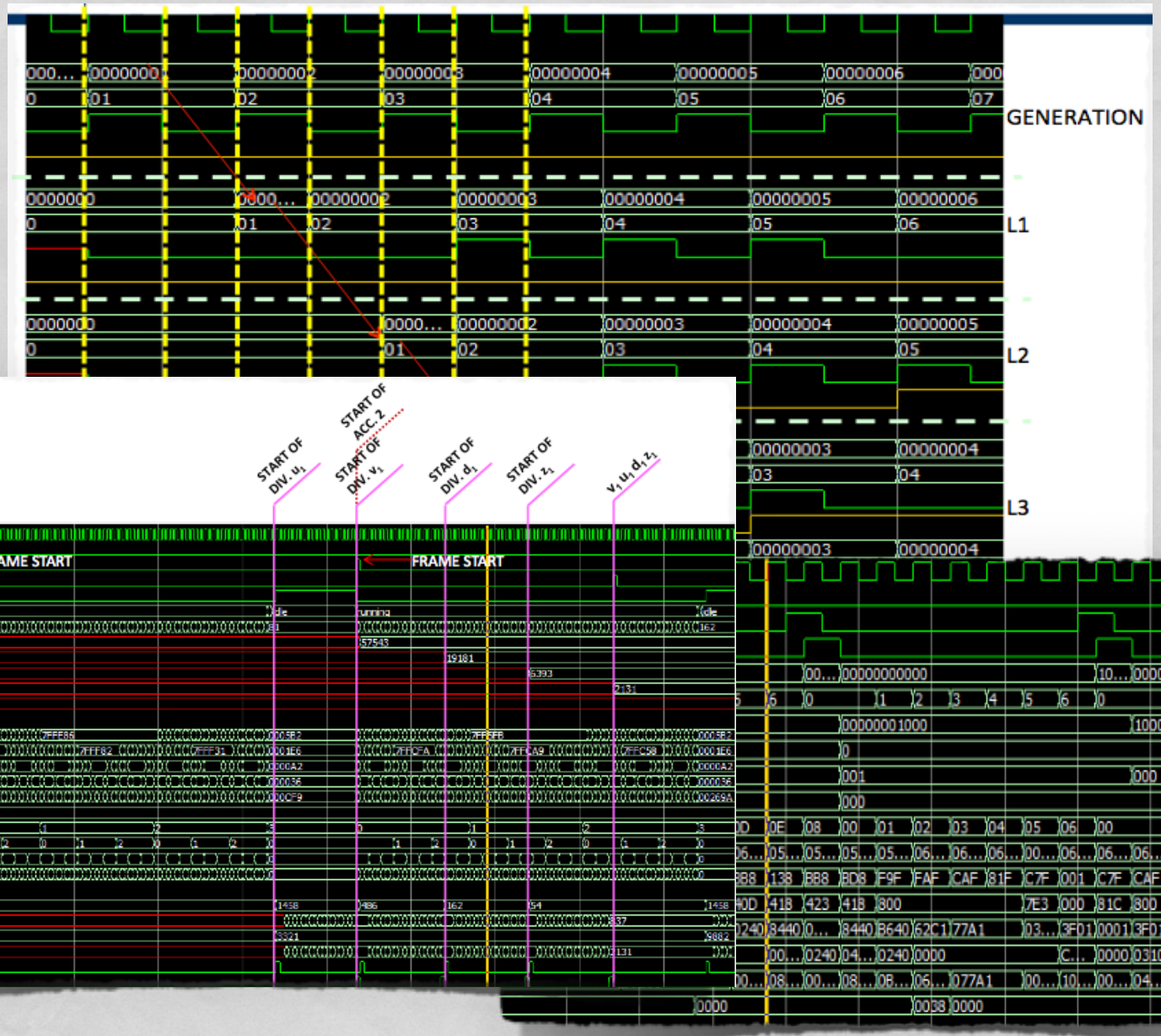
Input data rate of few Tb/s

Can fit O(1000) engines per chip.
Exact figure depends on specific choice on details (time ordering of pixel data)

Typical tracking system can be built with O(100) chips.



Logic simulation



Timing

Task	Latency (cycles)
Switch in readout board	15
Switch in TPU – dispatcher	15
Switch in TPU – fanout	6
Engine processing	70
Clustering	11
Output data	10
Total	< 150

Total latency about 125 clock cycles at 350 MHz.

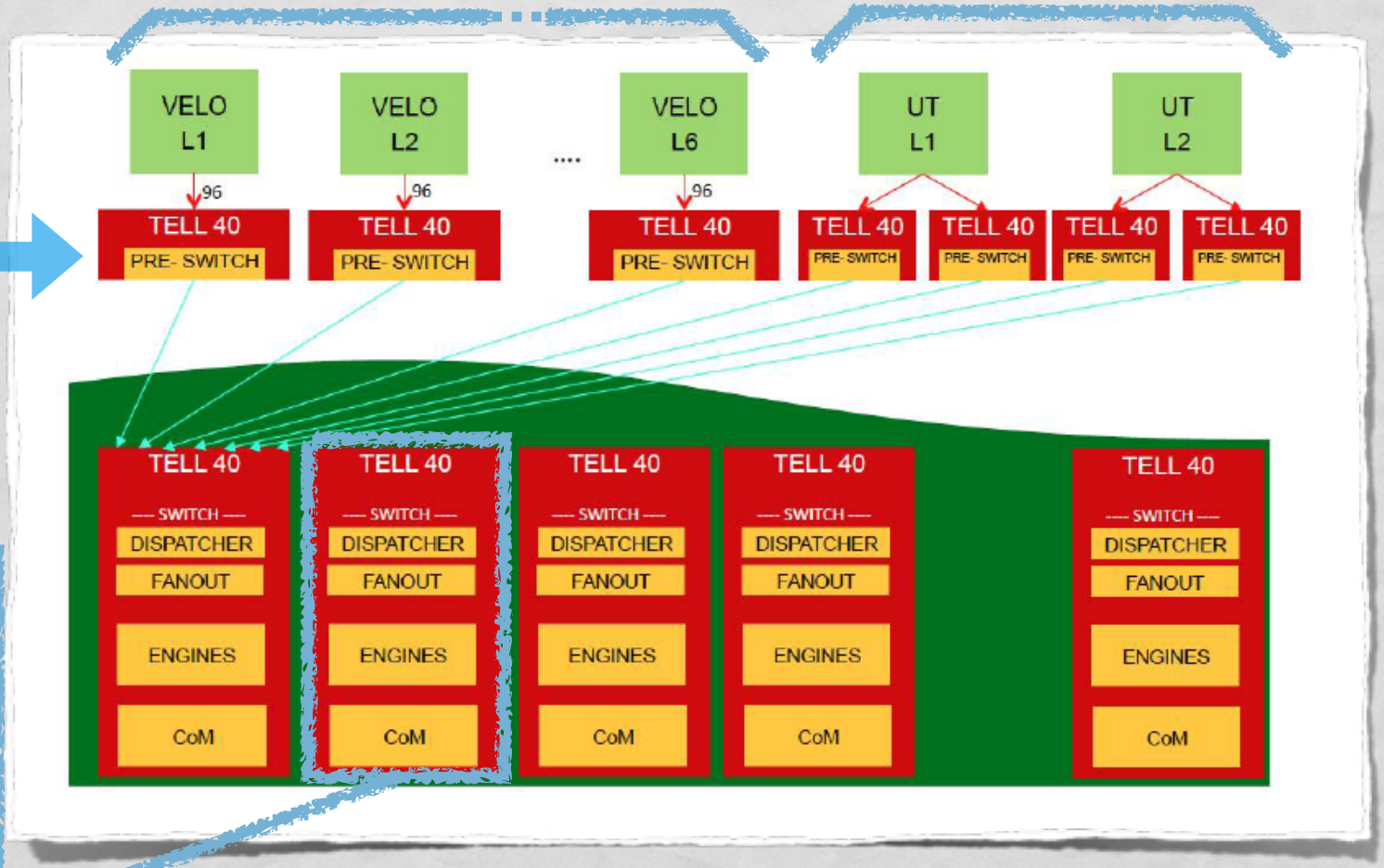
Much less than 1 microsecond – likely irrelevant compared with other latencies already present in DAQ.

Device effectively appears to the DAQ as just another detector that outputs tracks.

Fit in LHCb's DAQ*

8 layers of pixel vertex detector 2 layers of strip detector

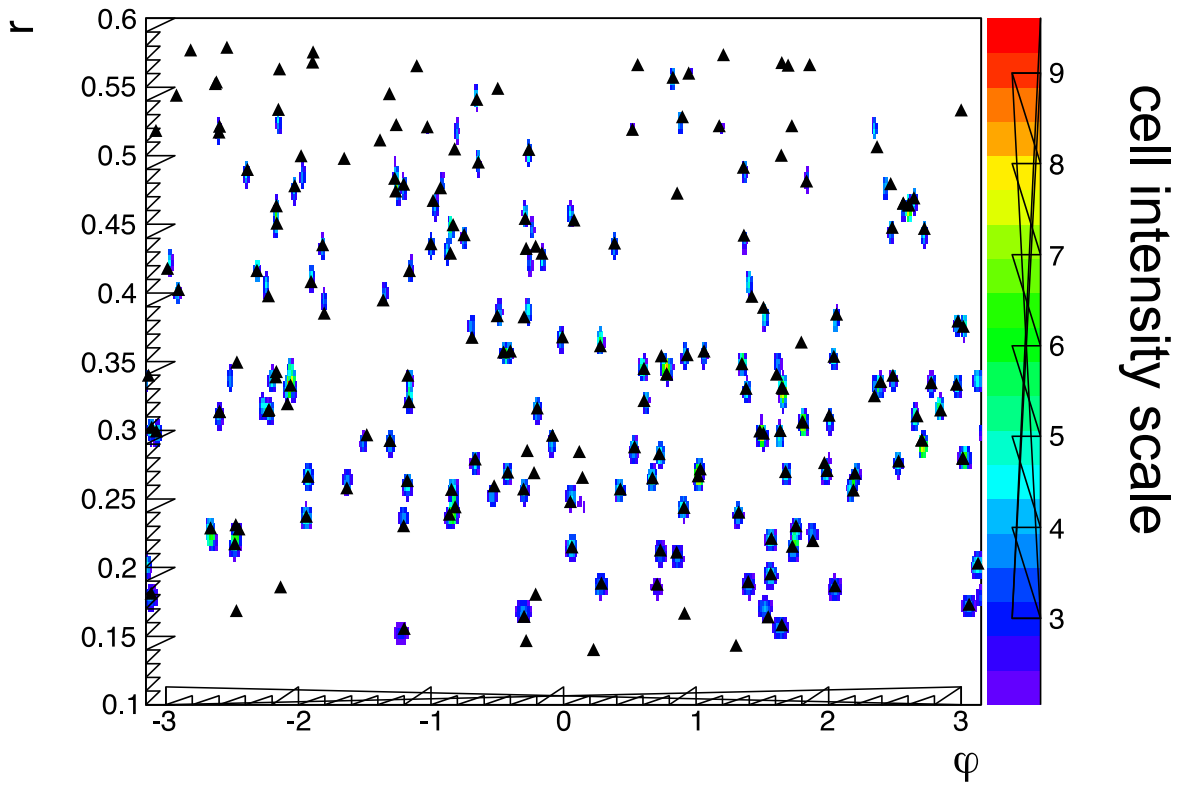
Front end



A solid angle "projective tower" of multiple detector layers

* Current LHCb DAQ evolved towards replacing TELL40 with PCIe40 (see backup)

Is this approach actually effective in a realistic scenario?



input track

reconstructed

...see next talk, by P. Marino

Summary

- Real-time reconstruction of tracks at full rate of high-luminosity LHC is achievable with an algorithm inspired by the vision process as it happens in mammals' brain.
- Implemented a realistic model suited for pixel detectors.
- Detailed design of the device's architecture and detailed simulation in realistic, LHCb-like experimental conditions.
- Reconstruct tracks at 40 MHz with submicrosecond latency.

Effectively an additional detector that outputs directly tracks

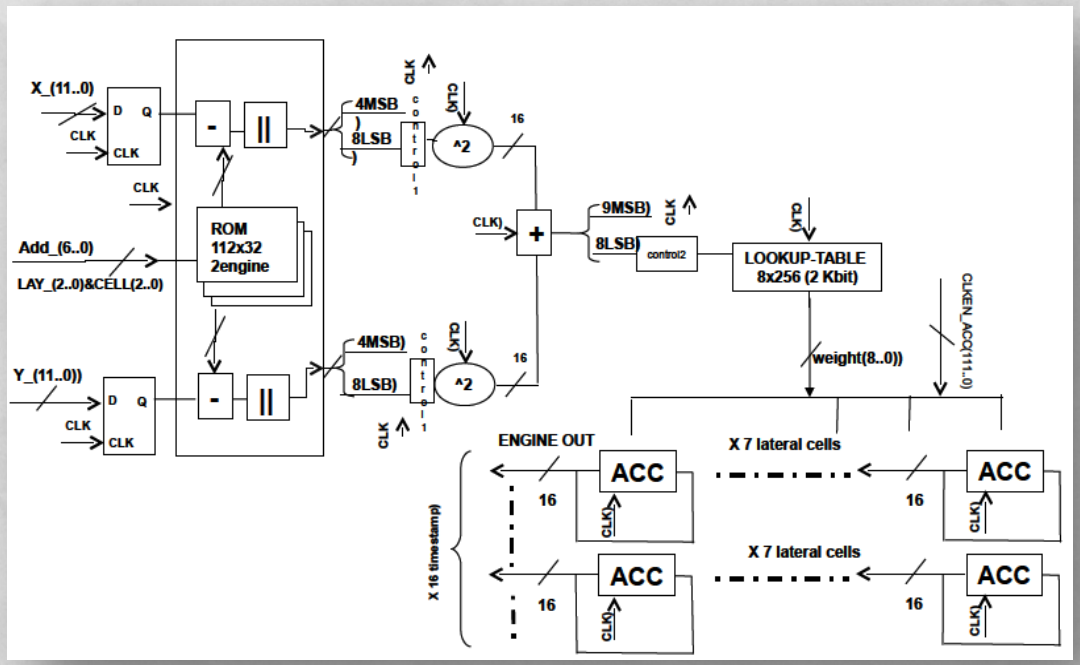
Next: hardware demonstrator based on existing readout boards from NA62's experiment.

The end

The engine

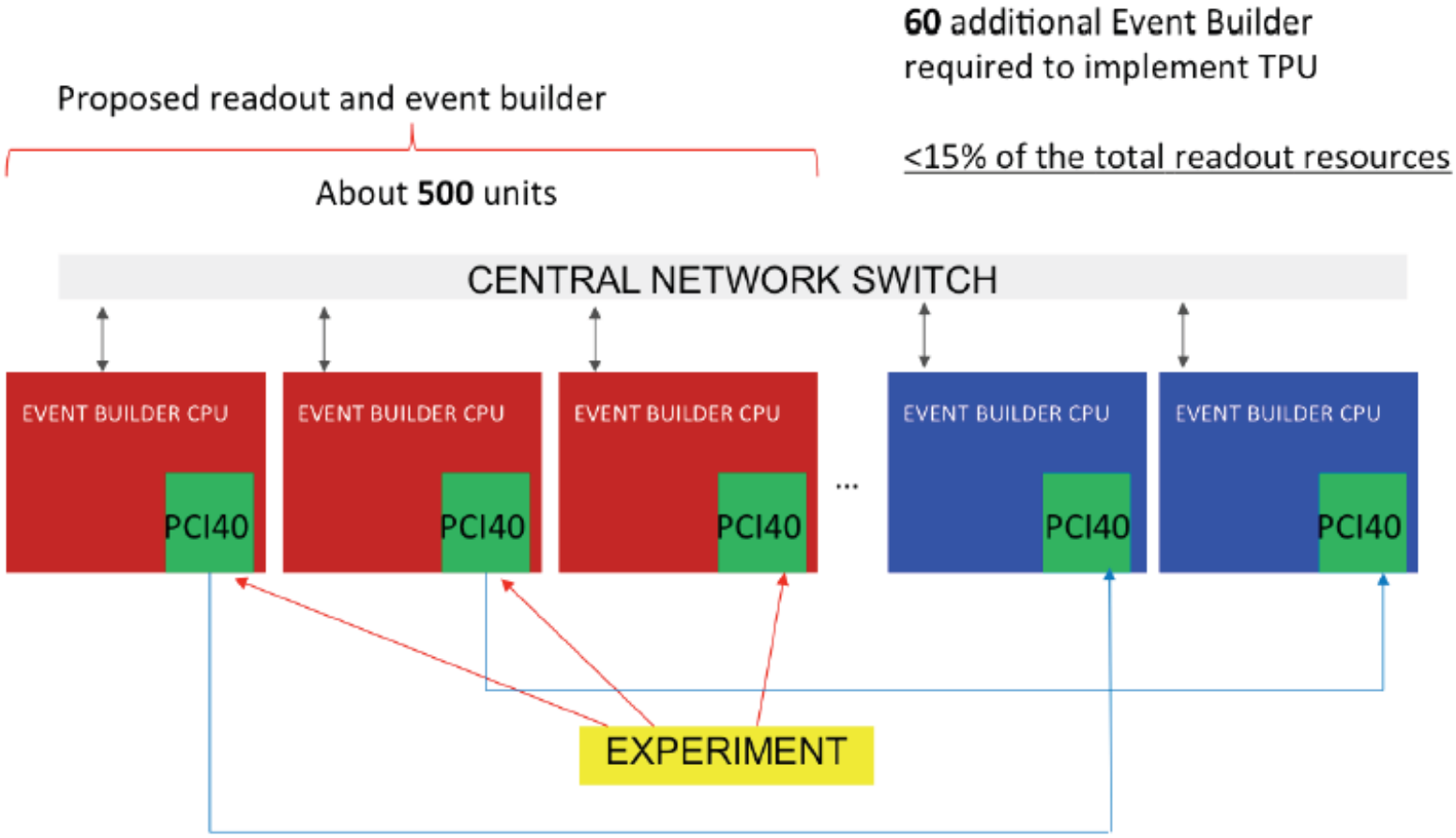
Logic module of the cell. Implemented as a clocked pipeline

- Layer ID determines the coordinates of the receptor center to be subtracted from hits' coordinates.
- Outcome squared, summed, yielding and the result R is rounded
- A sigma function common to all engines mapped in a LUT
- Rounded result is used as address to the LUT.
- LUT outputs accumulated for each hit of the event

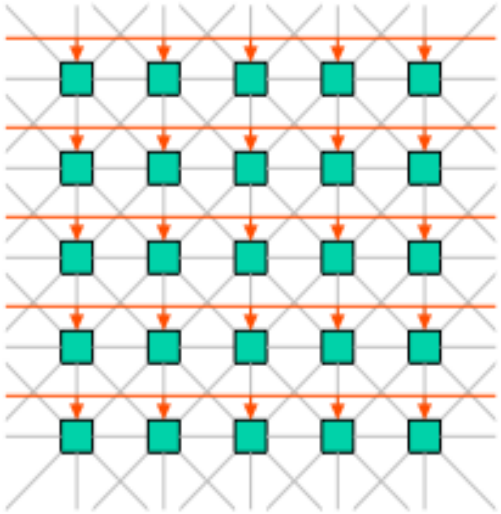


Each hit is cycled multiple times to compute excitation in lateral cells.

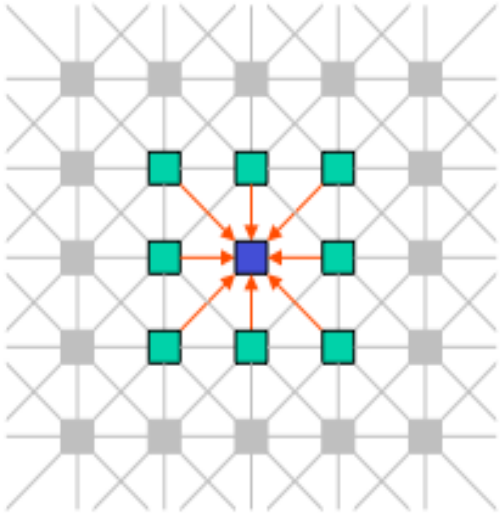
Fit in LHCb's DAQ



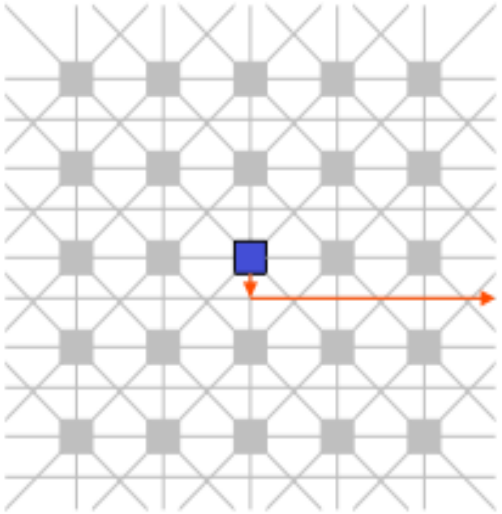
Overview



INPUT
all cells in parallel



CLUSTER FIND
all cells in parallel



OUTPUT
sequential