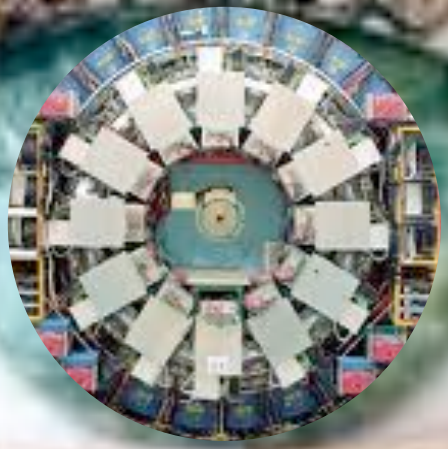
# The artificial retina processor for tracking at 40 MHz



**A. Abba, F. Bedeschi, F. Caponio, M. Citterio, 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)**

**TIPP2014, Amsterdam – June 5, 2014**

# Outline

Complete feasibility demonstration of offline-quality track reconstruction at the full LHC crossing rate in a real HEP experiment

- Concept
- Hardware implementation
- System size and timing performance
- Tracking performance

# 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 |

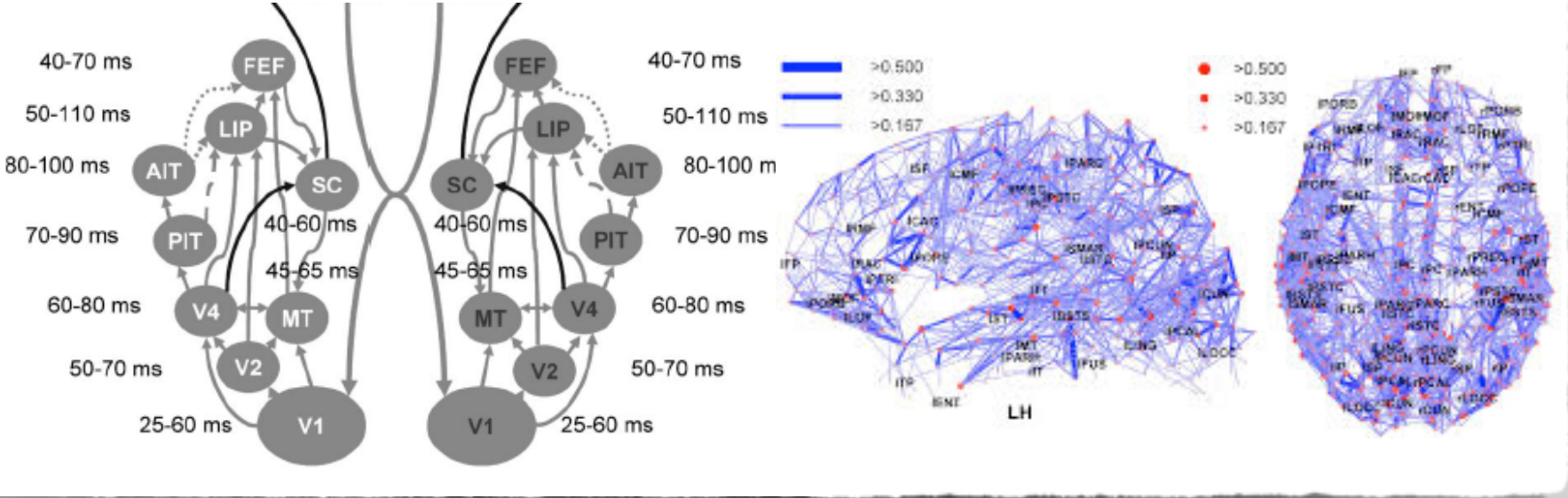
Perform tracking synchronous with LHC collisions appears daunting.

Past and current realizations of similar devices for complex tracking call 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 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)

Can these features be engineered into a viable tracking system?

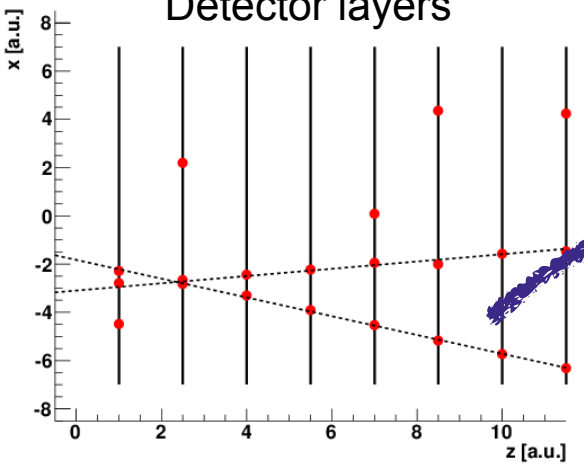
# Retina algorithm

NIM A453, 425 (2000)

D.H. Hubel and T.N. Wiesel, J. Physiol, 148 (1959) 574

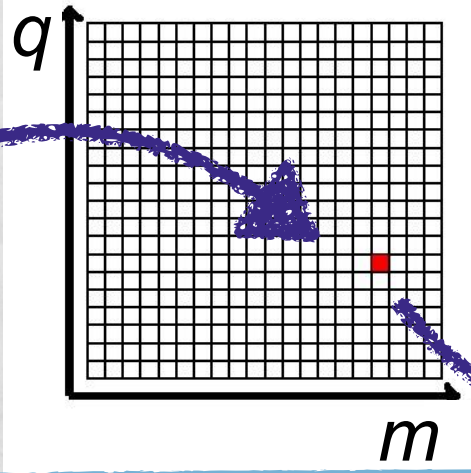
hit

Detector layers

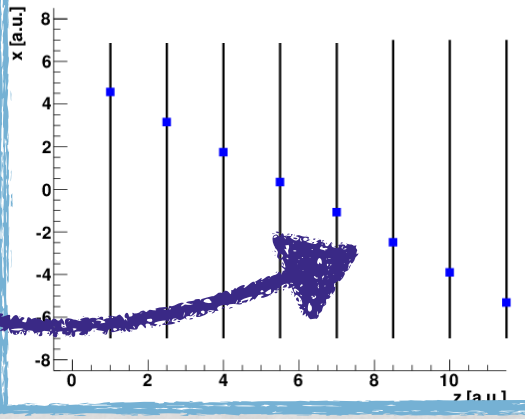



$$x = mz + q$$

Discretize space of track parameters in *cellular units*




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



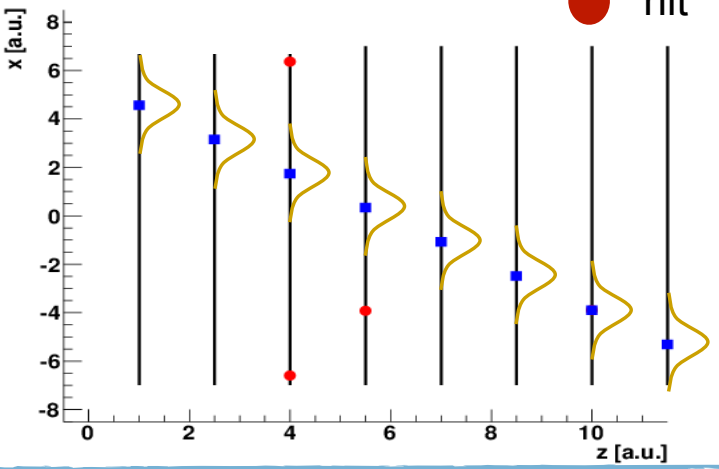

In a detector layer, the distance *s* between the hit and the receptor is used to compute the contribution of that hit to the excitation of the cellular unit.

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

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

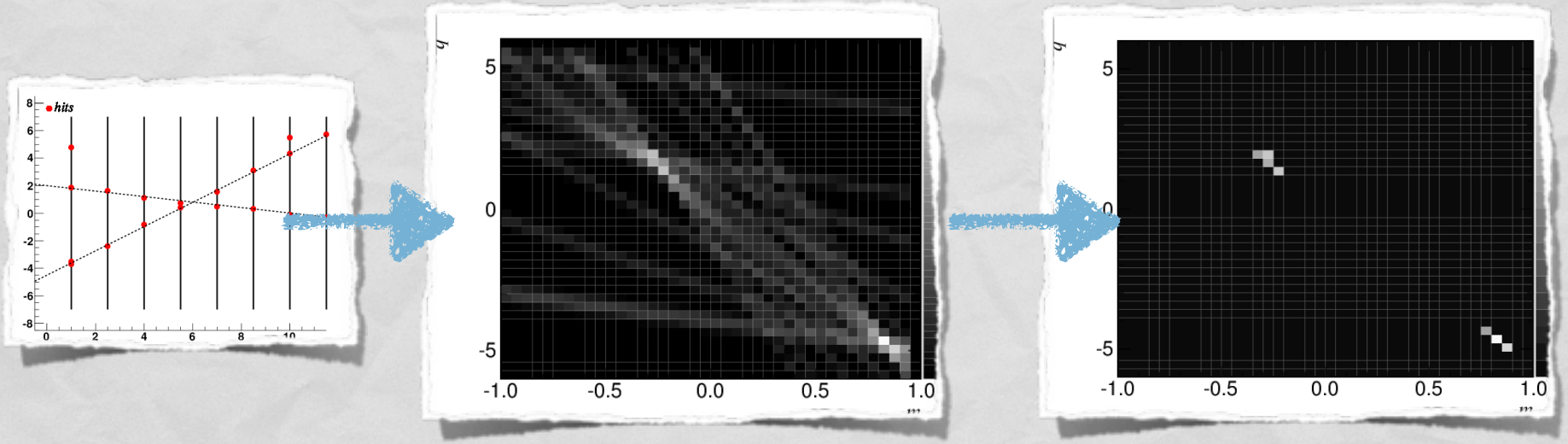
receptor

hit



# Response

The union of responses $R_{ij}$ on all cells is 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 really new.

Designed and proved conceptually feasible in a toy 2D tracker 15 years ago, but unviable for 1990s electronics.

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

However, a few important original features:

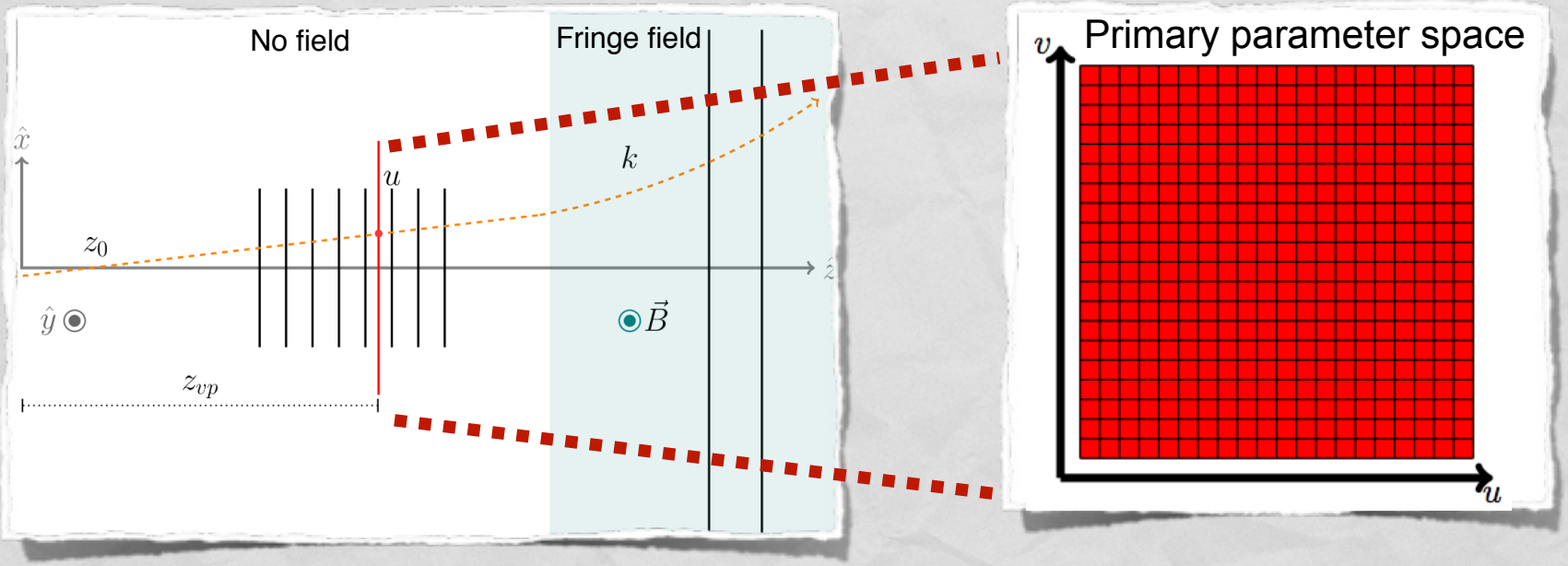
- Not just yes or now response. Signal on each cell is a smooth function of hit positions and used as a weight to interpolate track parameters with better resolution than grid step.
- "Neural" communication btw nodes allows strong parallelism.

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

# Layout

**Geometry impacts significantly the implementation**

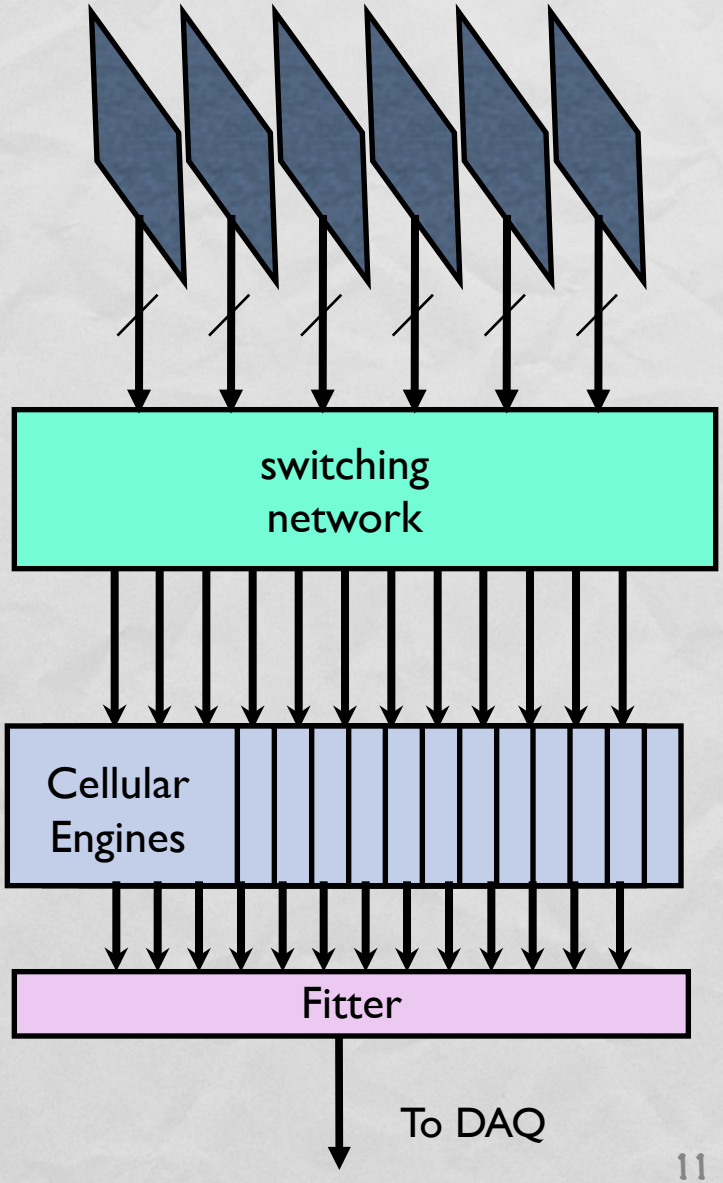
Forward spectrometer with pixel and strip detectors




- 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

# Hardware

# 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

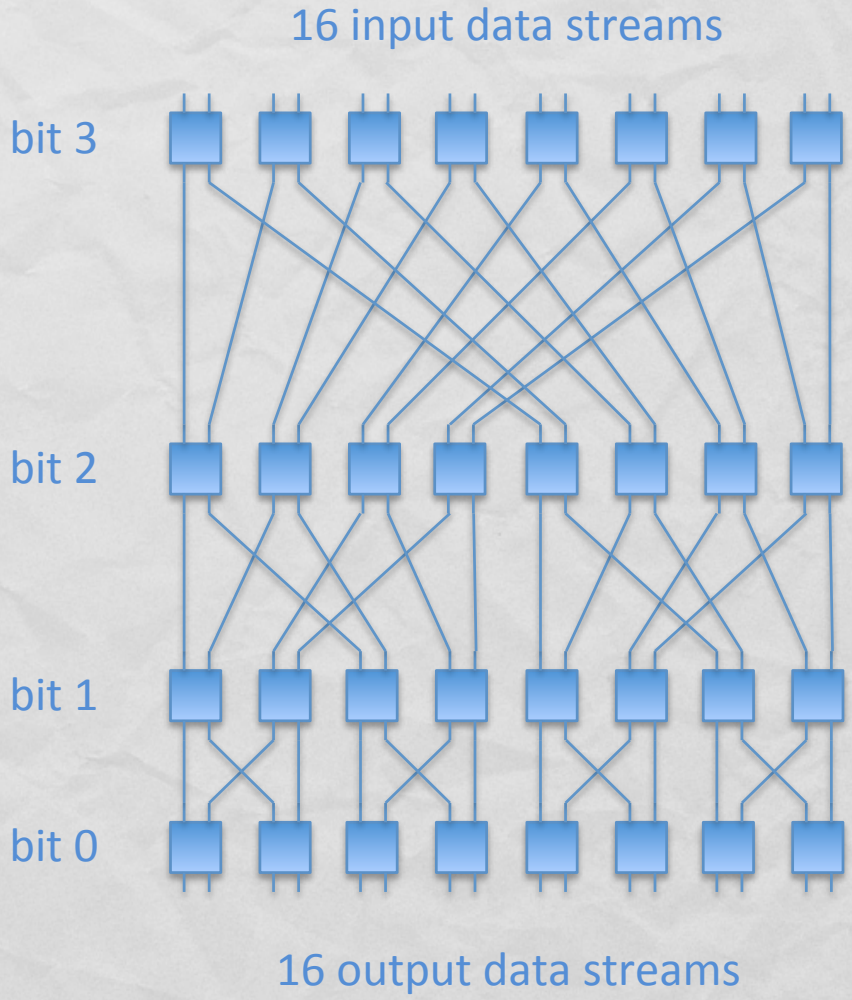
# Switching network

Each input can get to any output.

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

Each hit comes with a “zip-code”

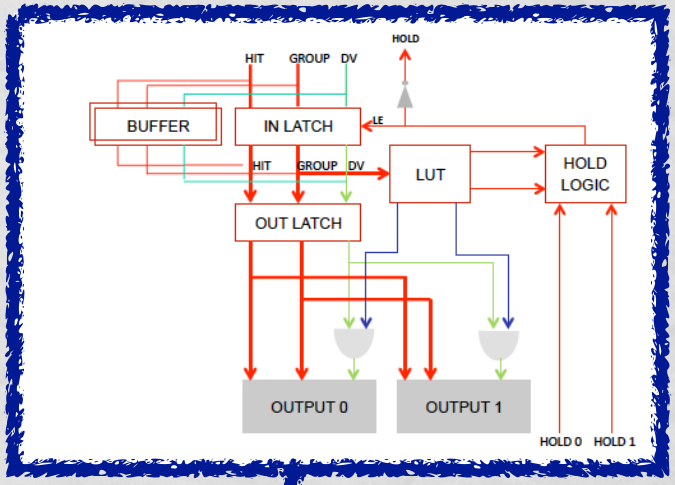
The switching network “knows” where to deliver it, according to embedded programmable maps distributed over the nodes.

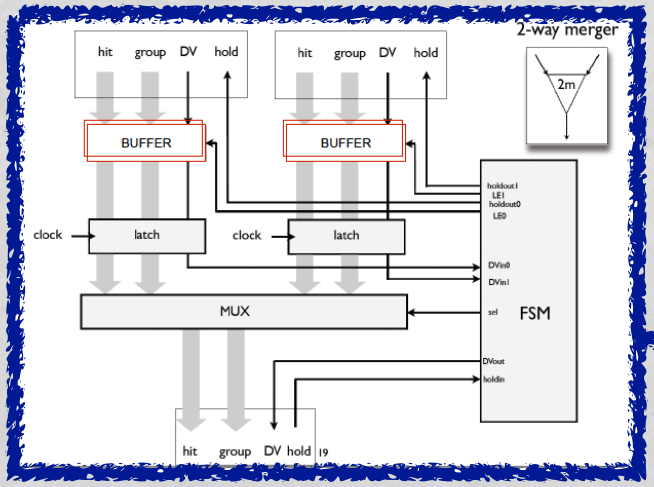


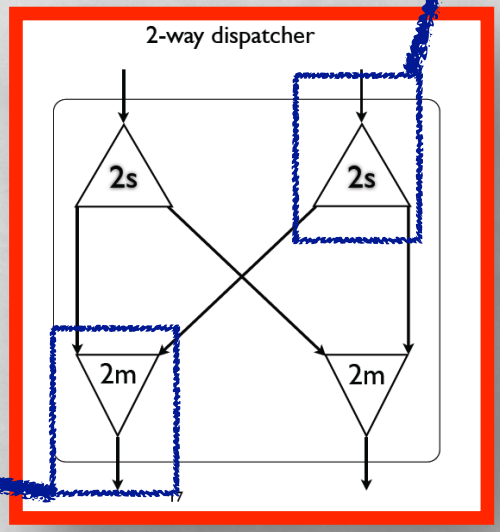
# Switching basic unit

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.



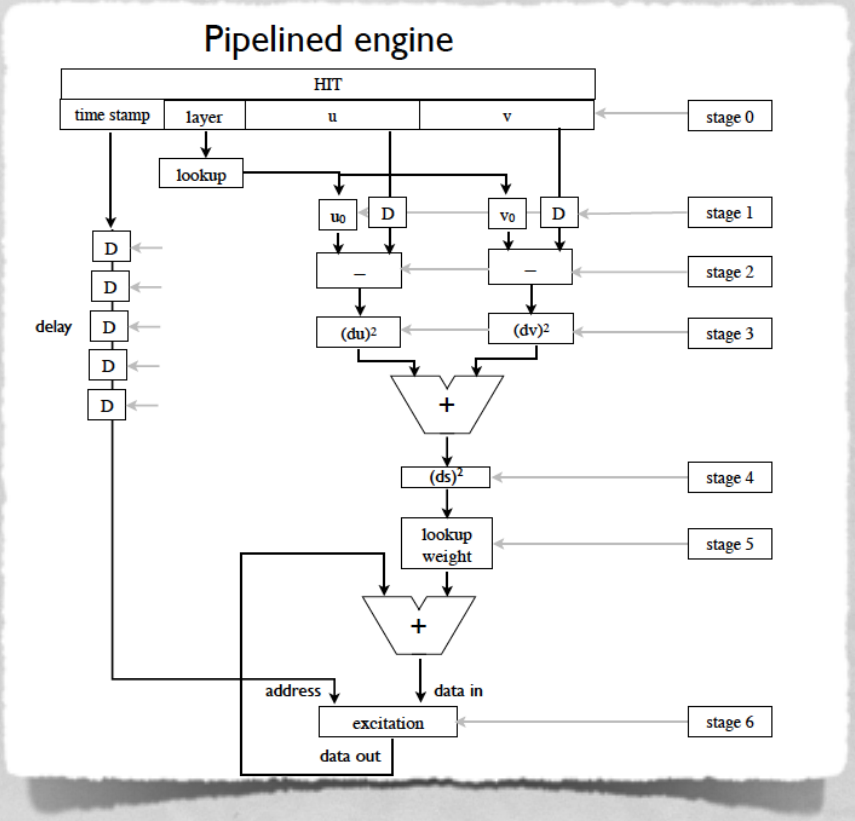



# The excitation engine

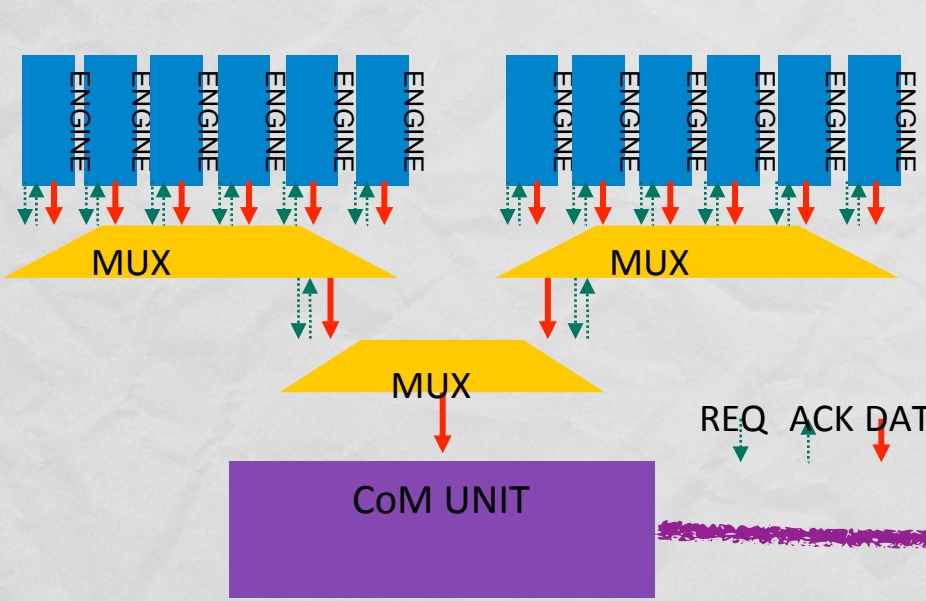
Logic module of the cell. Implemented as a clocked pipeline

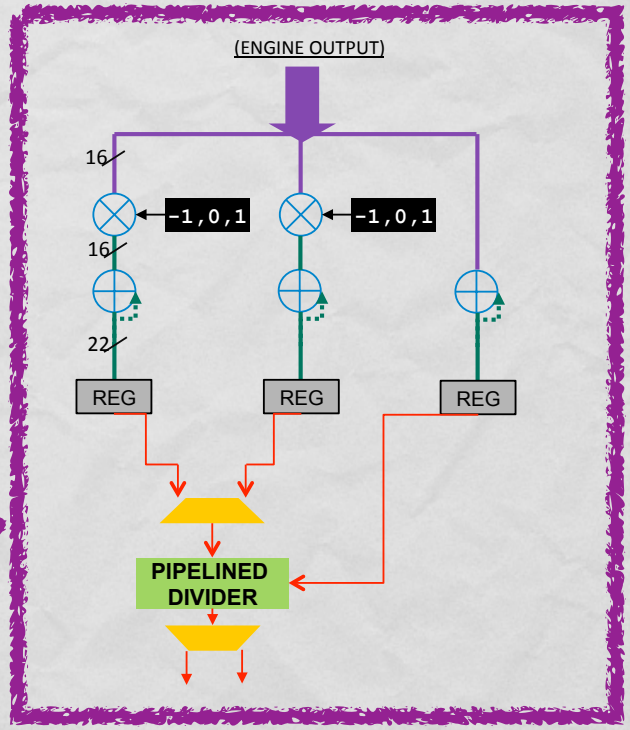
- Layer ID determines the coordinates of the receptor center to be subtracted from hits' coordinates.
- Outcome squared and summed. The result R is rounded
- A weight function common to all engines mapped in a LUT
- Rounded result is used as address to the LUT.
- LUT outputs accumulated for each hit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




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

# Clustering

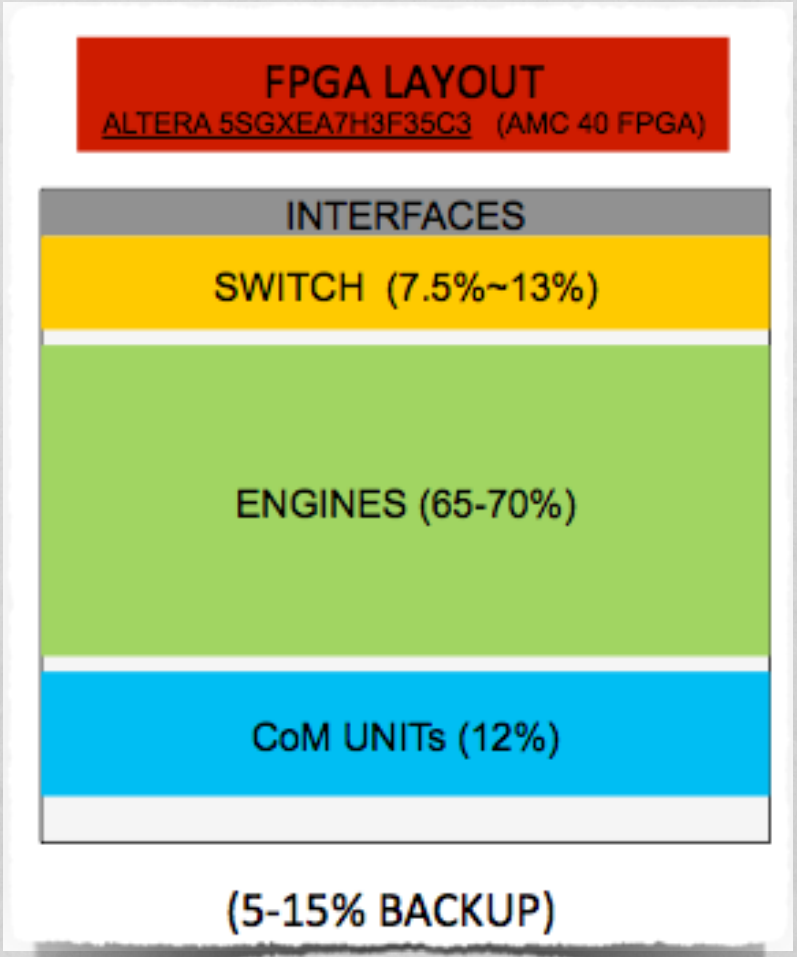



Data reduction within the engine allows O(10) reduction to keep up with data flow.

2nd stage: local clustering (center of excitation) in parallel and queuing results to output

# Placing


All main components implemented in VHDL and placed on the FPGA

Input data rate of few Tb/s

Fit O(1000) engines per chip.

Typical tracking system implemented with O(100) chips

# Timing

Upgrade LHCb-like benchmark: O(1000) hits over O(10) layers to reconstruct O(100) tracks every 25 ns. A few Tb/s distributed over 45000 engines.

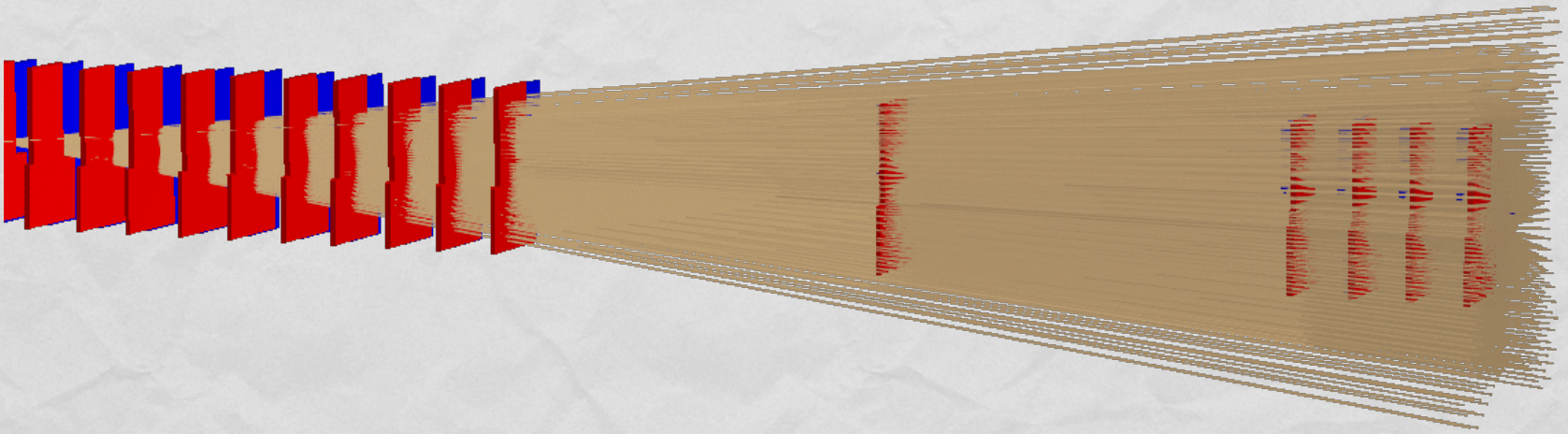
| Task | Latency (cycles) |
|---|---|
| Switch in readout board | 15 |
| Switch in TPU – dispatcher | 15 |
| Switch in TPU – fanout | 6 |
| Engine processing | 70 |
| Clustering | 11 |
| Output data | 10 |
| Total | $< 150$ |

Total latency about 125 clock cycles at 350 MHz – less than 1µs.

Irrelevant compared with other latencies already present in DAQ.

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

# Tracking performance



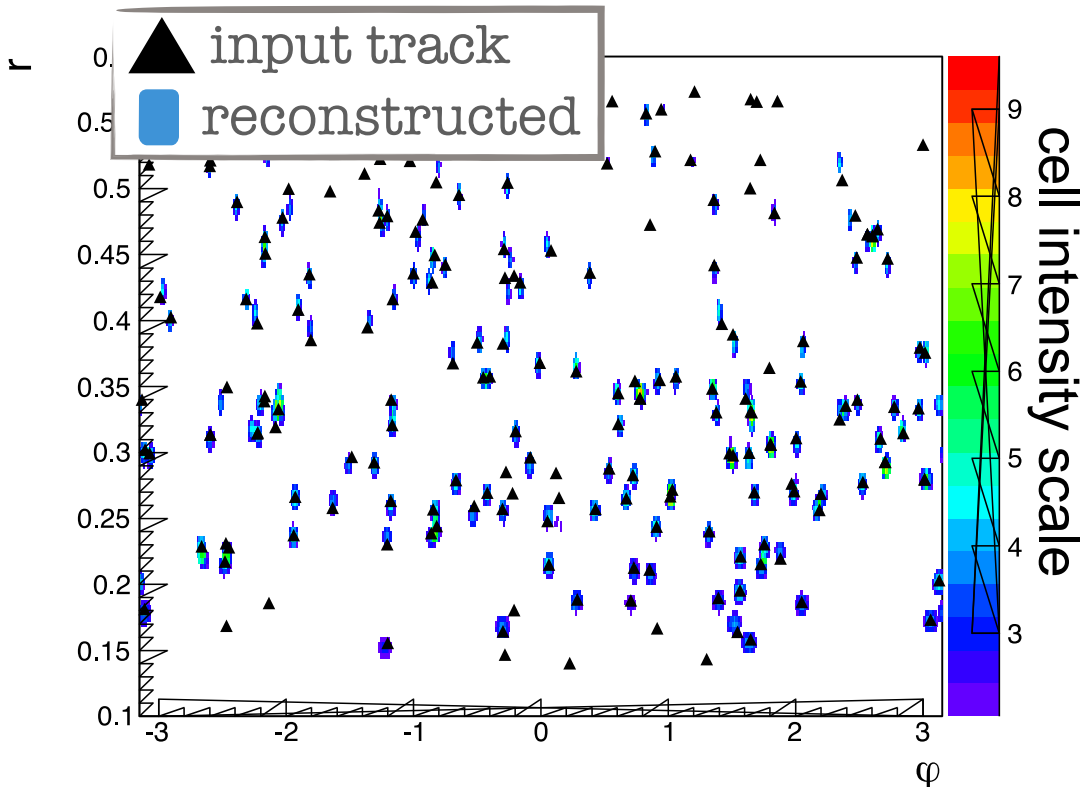
# The real thing

Pythia 8: generic 14 TeV pp collisions (LHCb in 2020)

L = 2 (or 3)×$10^{33}$ Hz/cm$^2$ corresponding to 7.6 (or 11.4) interactions/x-ing.

Detector geometry, evt. topology, occupancies, noise etc. from standard LHCb upgrade simulation

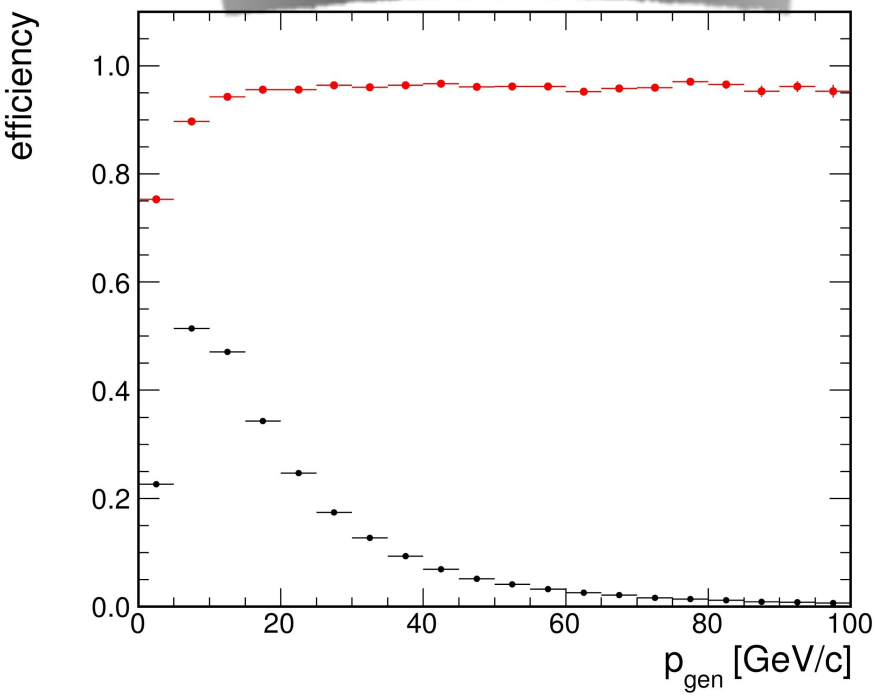
8 parallel pixel layers (no field, 15 μm resolution) plus 2 μstrip layers in the fringe field of the magnet (0.1T, 50 μm resolution)



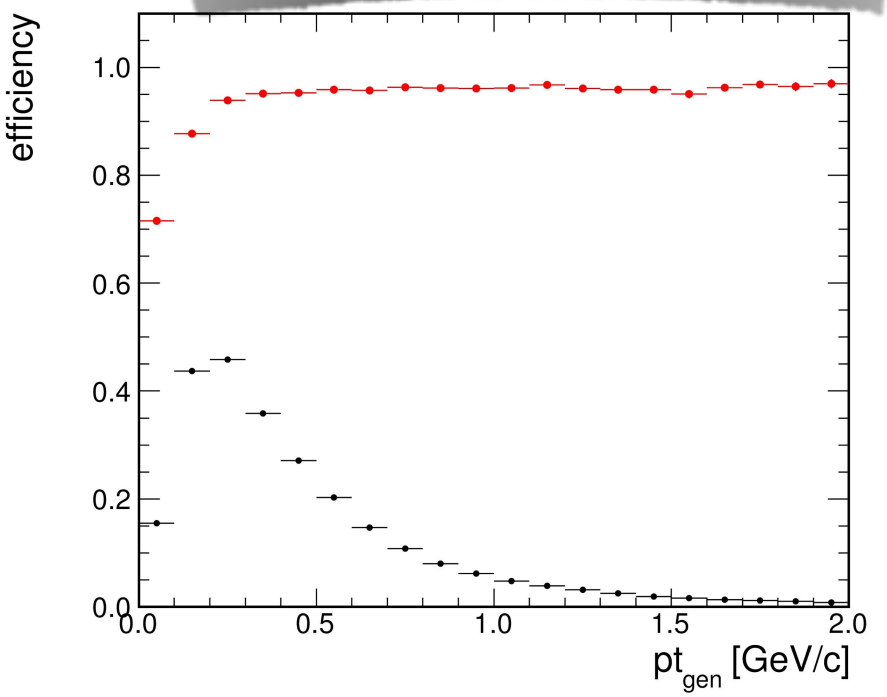

| | $2 \times 10^{33}\text{cm}^{-2}\text{s}^{-1}$ | $3 \times 10^{33}\text{cm}^{-2}\text{s}^{-1}$ |
|---|---|---|
| Number of hits | 880 | 1220 |
| Number of clusters (over thld) | 121 | 223 |
| Number of hits per engine | 1.3 | 1.95 |
| Ghost rate | 0.08 | 0.12 |

# Efficiency - momentum

Total momentum



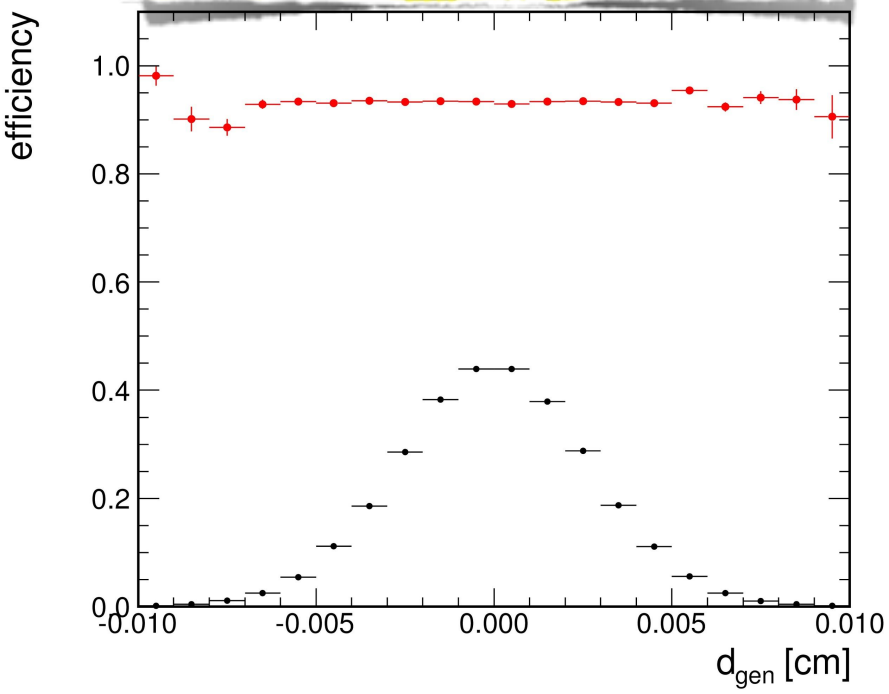

Transverse momentum



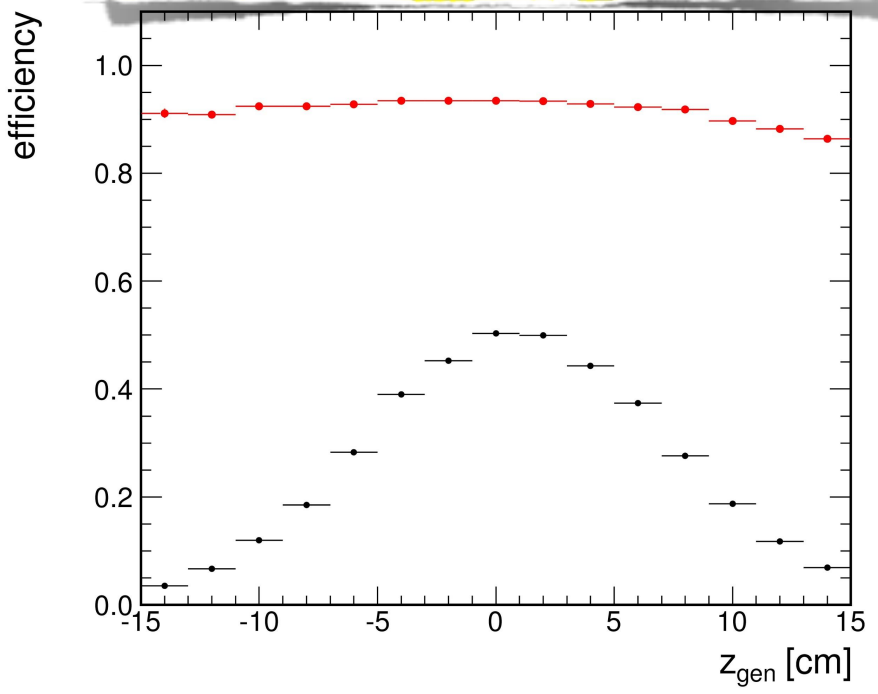

About 95% efficiency and uniform response. Comparable with offline reconstruction.

# Efficiency - origin

Transverse displacement

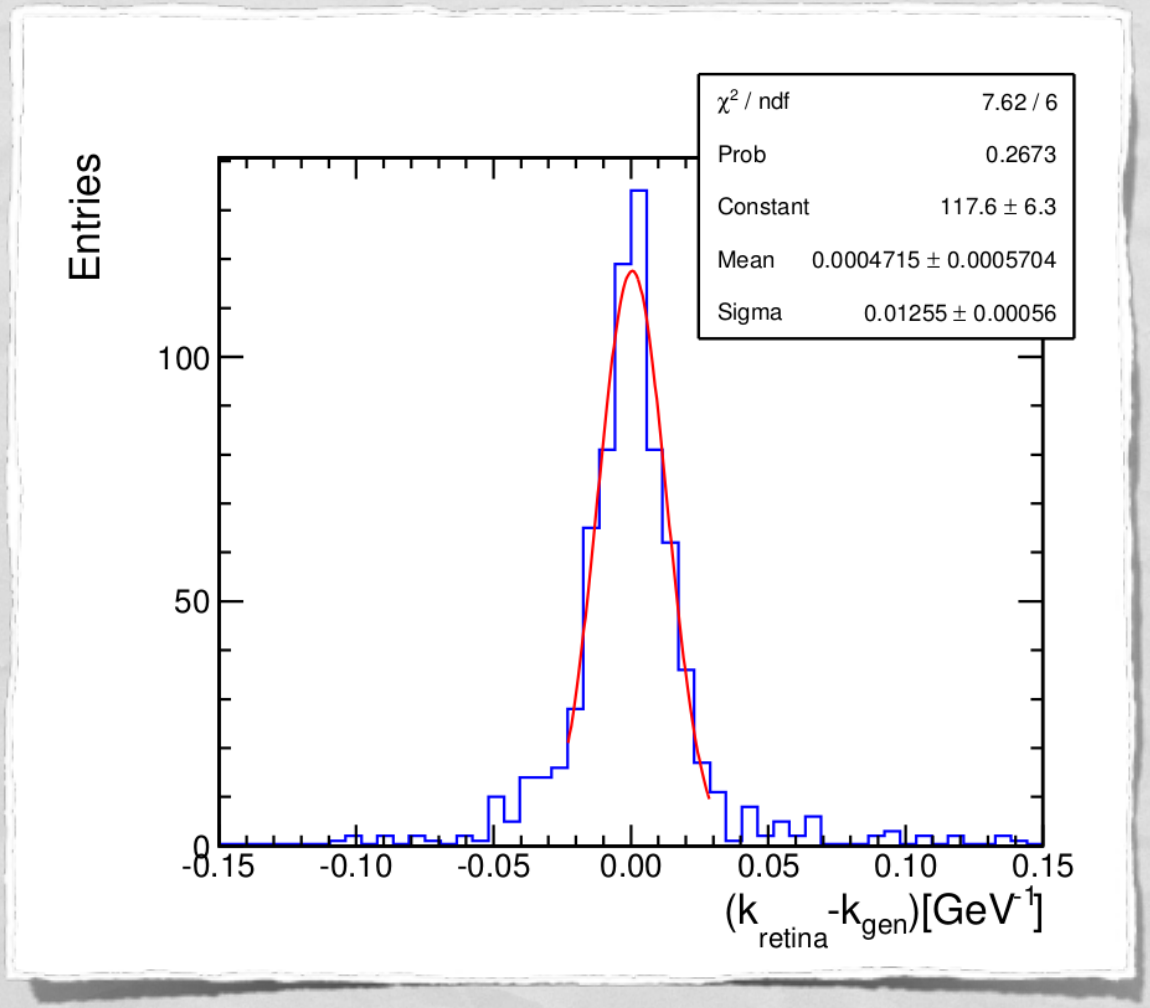



Longitudinal displacement




About 95% efficiency and uniform response. Comparable with offline reconstruction.

# Curvature resolution




Just 25% less than offline

# Summary

- Reconstruction of tracks at full rate of high-luminosity LHC achieved with an algorithm inspired by the vision process as it happens in mammals' brain.
- Implemented a realistic model suited for pixels and based on a real HEP detector.
- Detailed design of the device's architecture and simulation in realistic, LHCb-like experimental conditions.
- Get offline-like tracks at 40 MHz with sub-μs latency.

Effectively, an additional detector that outputs directly tracks

Next:

- Demonstrator using TELL62 readout boards.
- Test on a telescope of planar layers (see N. Neri talk).

# Further information

- L. Ristori www.sciencedirect.com/science/article/pii/S0168900200006768
- Public document LHC-PUB-2014-026 http://cds.cern.ch/record/1667587?ln=en
- G. Punzi at INSTR2014 https://indico.inp.nsk.su/getFile.py/access?contribId=129&sessionId=6&resId=0&materialId=slides&confId=0
- D. Tonelli at WIT2014 https://indico.cern.ch/event/293354/contribution/27/material/slides/0.pdf
- P. Marino at WIT2014 https://indico.cern.ch/event/293354/contribution/24/material/slides/0.pdf
- A. Abba at RT2014 No link yet, sorry.
- M.M. Del Viva at TIPP2014 http://indico.cern.ch/event/192695/session/12/contribution/379/material/slides/1.pdf
- N. Neri at TIPP2014 http://indico.cern.ch/event/192695/session/2/contribution/272/material/slides/0.pdf

# The end

# 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*

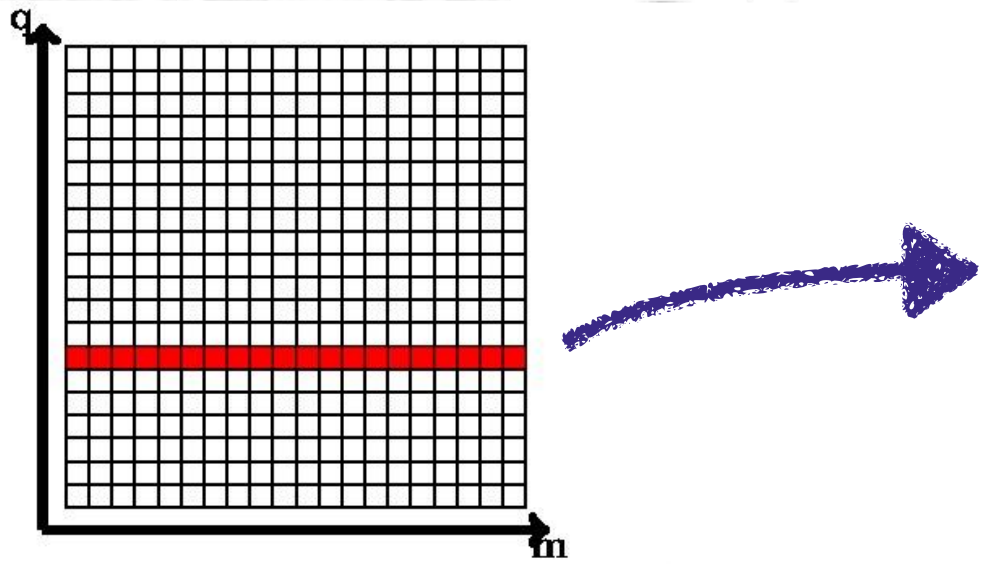
Accepted 21 June 2000

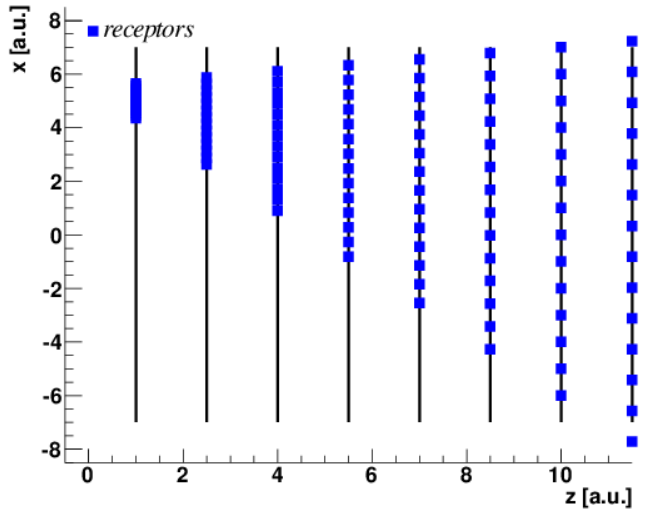
**Abstract**

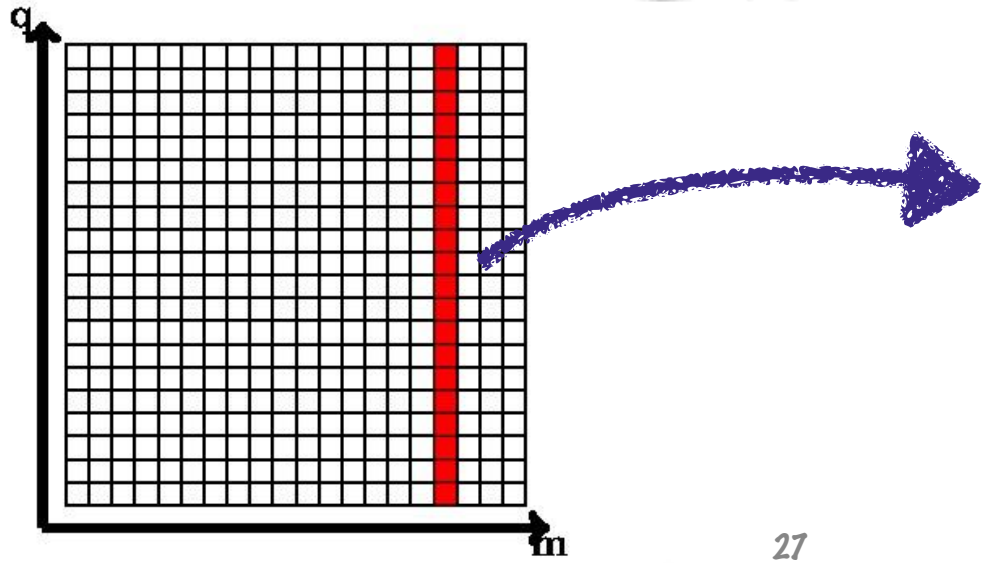
A new approach is proposed for fast track finding in position-sensitive detectors. The basic working principle is modeled on what is widely believed to be the low-level mechanism used by the eye to recognize straight edges. A number of receptors are tuned such that each one responds to a different range 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. 

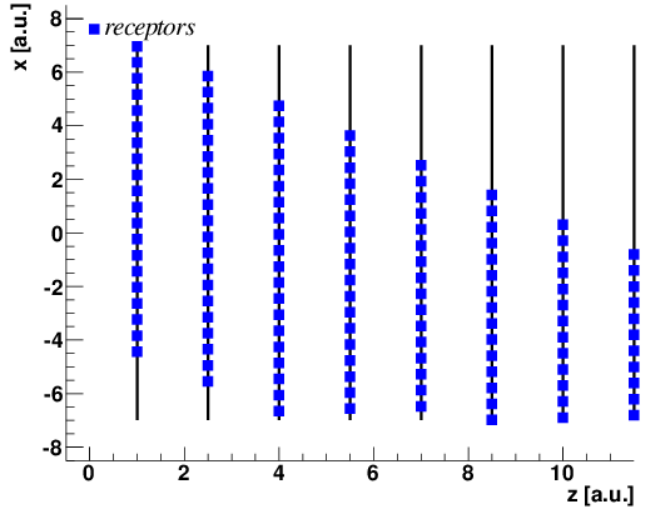
Inspired by mechanism of visual receptive fields D.H. Hubel and T.N. Wiesel, J. Physiol, 148 (1959) 574

# Track mapping



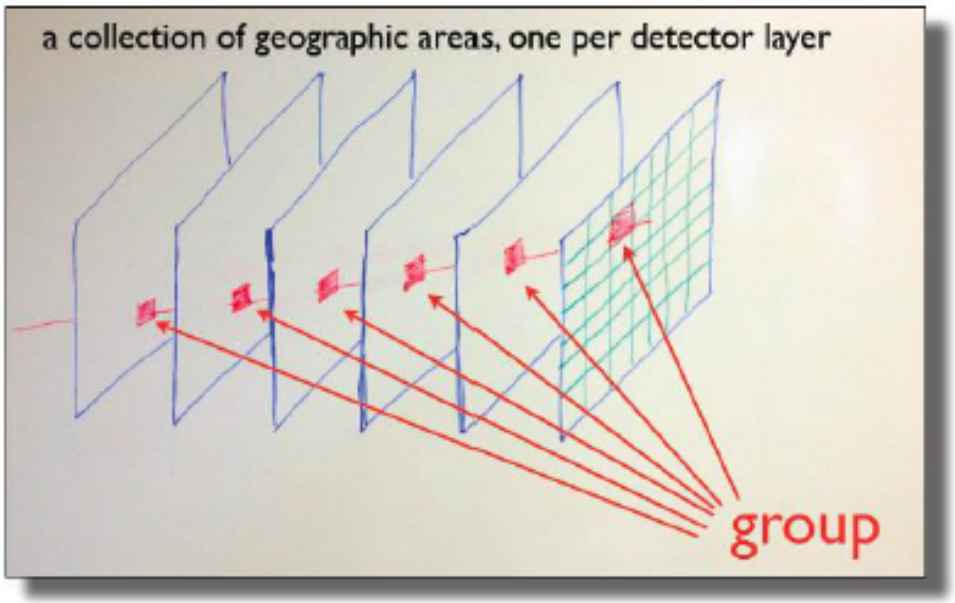



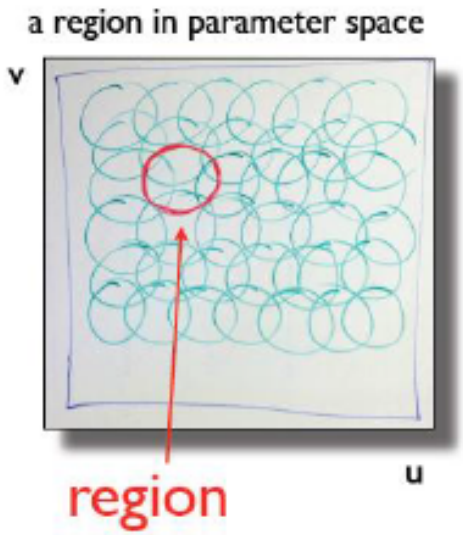



# Switching concept

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



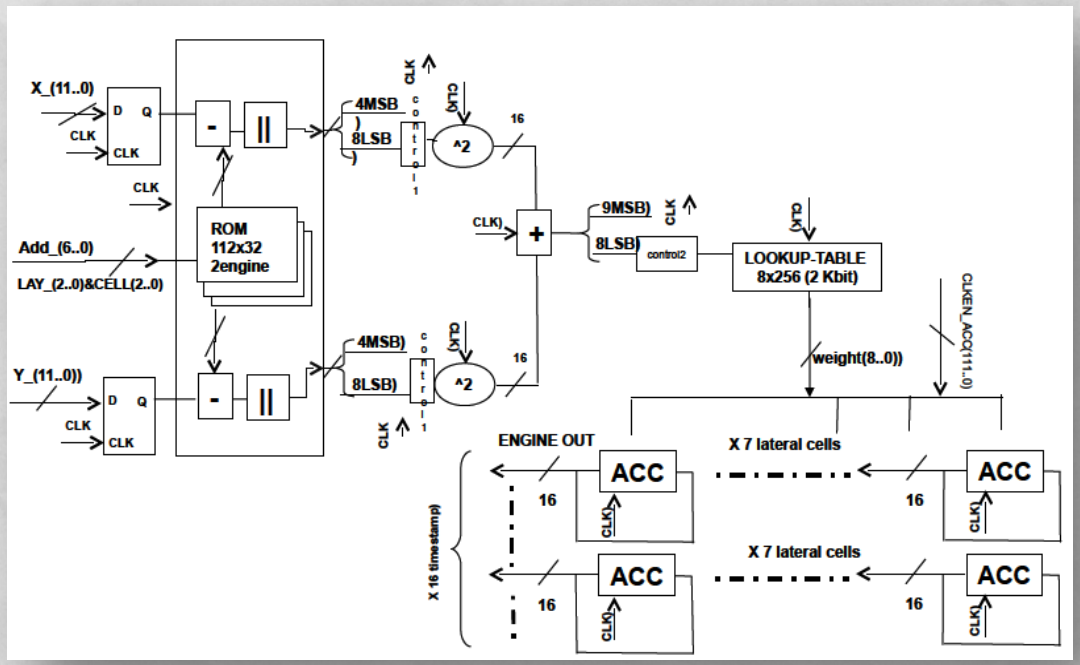




Each hit is only delivered to the cells that are likely to be significantly excited by that hit.

# The engine

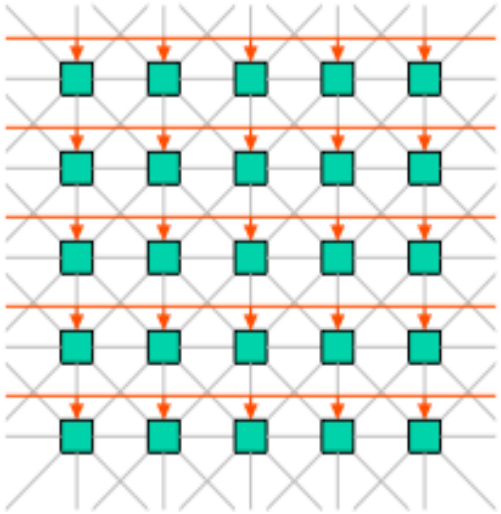
Logic module of the cell. Implemented as a clocked pipeline

- 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

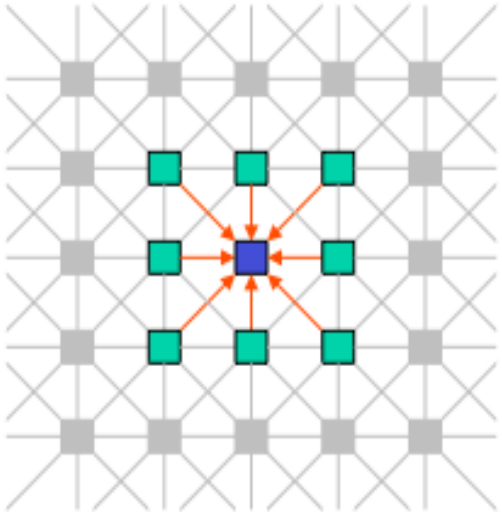



Each hit is cycled multiple times to compute excitation in lateral cells.

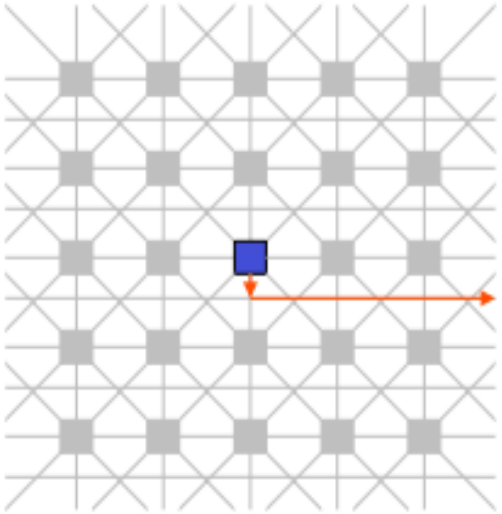
# Overview



INPUT
all cells in parallel



CLUSTER FIND
all cells in parallel

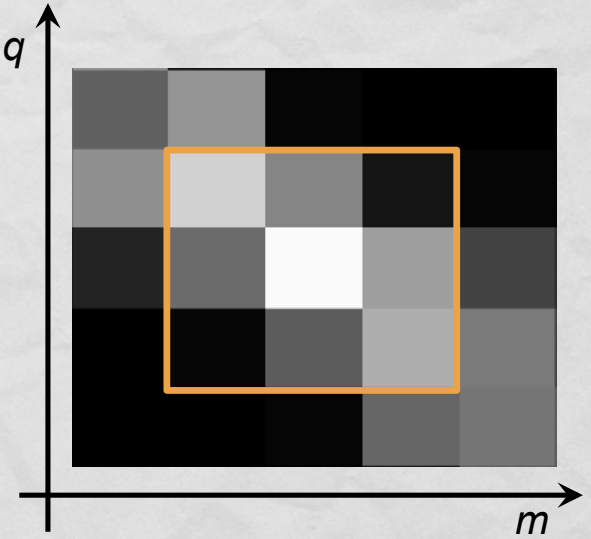


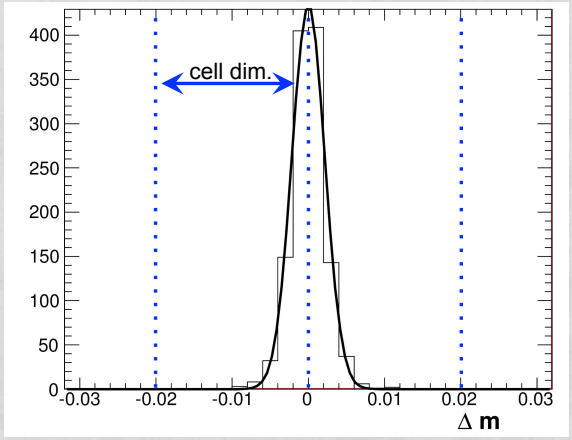
OUTPUT
sequential

# Primary track parameters

Once local maxima are found, primary track parameters can be estimated through excitation centroid of nearest cells

$$m = \frac{\sum_{ij} m_i w_{ij}}{\sum_{ij} w_{ij}} \qquad q = \frac{\sum_{ij} q_j w_{ij}}{\sum_{ij} w_{ij}}$$

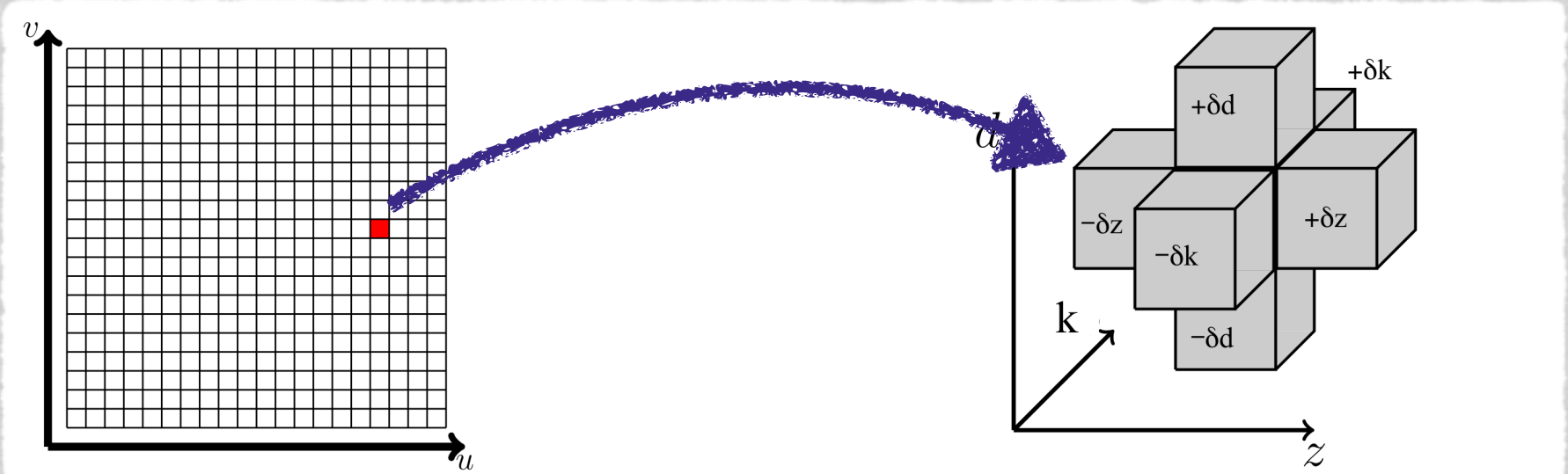



8

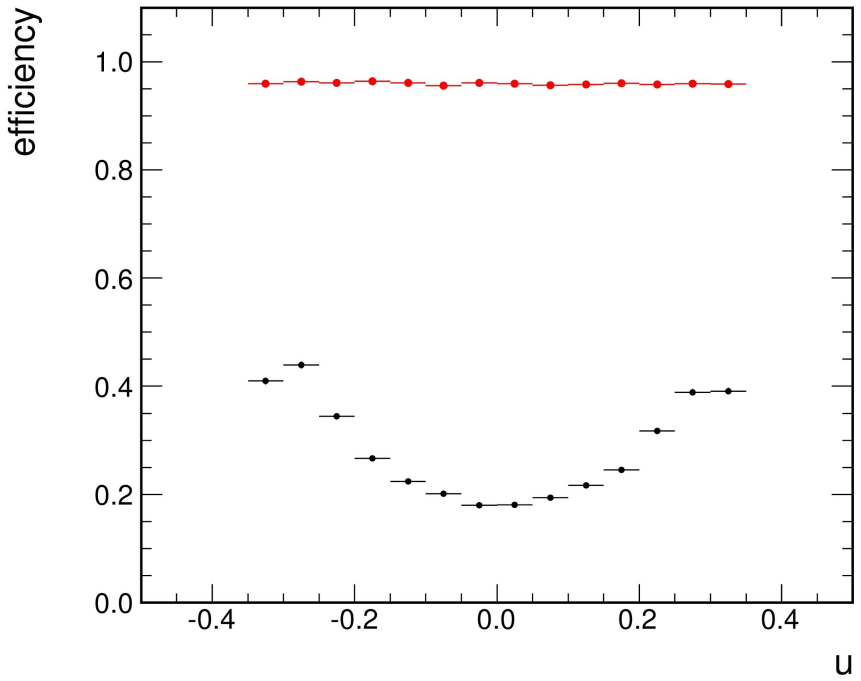
# Other track parameters

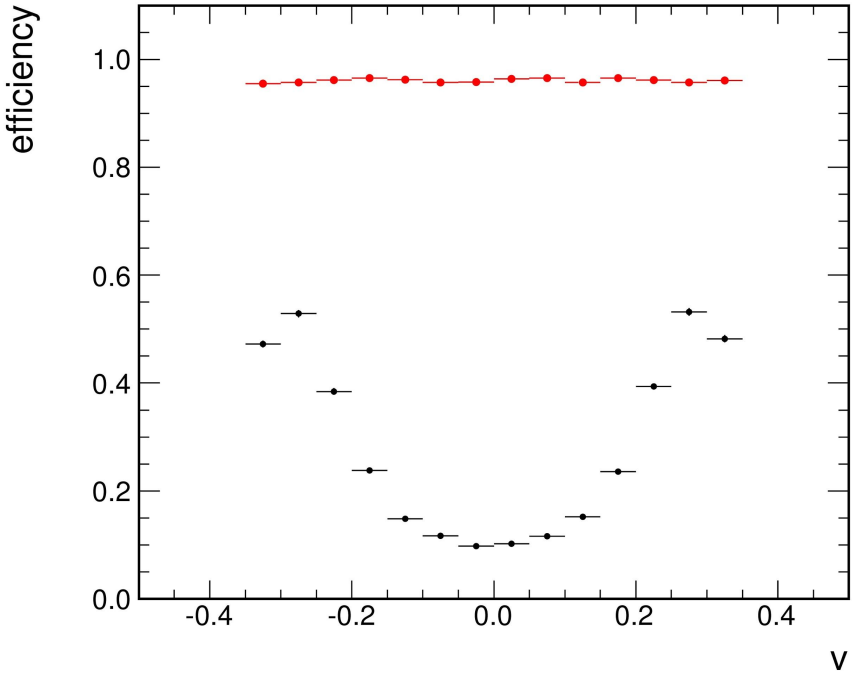
Once primary parameters are determined, other parameters found by “unfolding” the primary cell into full dimensionality through addition of lateral cells and interpolating their response




$$d_{\text{rec}} = \frac{w_{\delta d} - w_{-\delta d}}{w_{\delta d} + w_{d=0} + w_{-\delta d}} \cdot \delta d$$

# Efficiency - primary plane



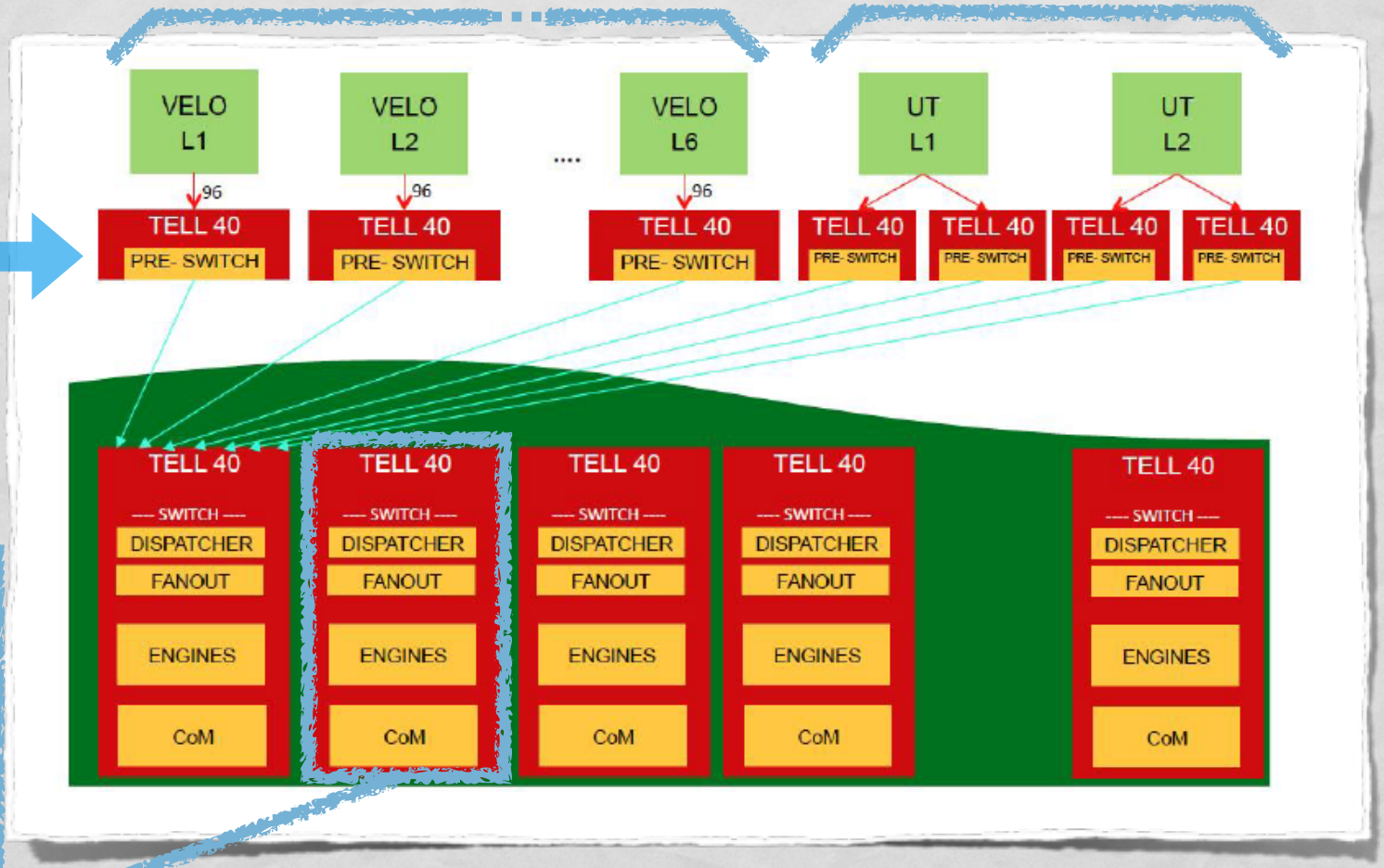




About 95% efficiency and uniform response. Comparable with offline reconstruction.

# Fit in LHCb's DAQ*

8 layers of pixel vertex detector  2 layers of strip detector

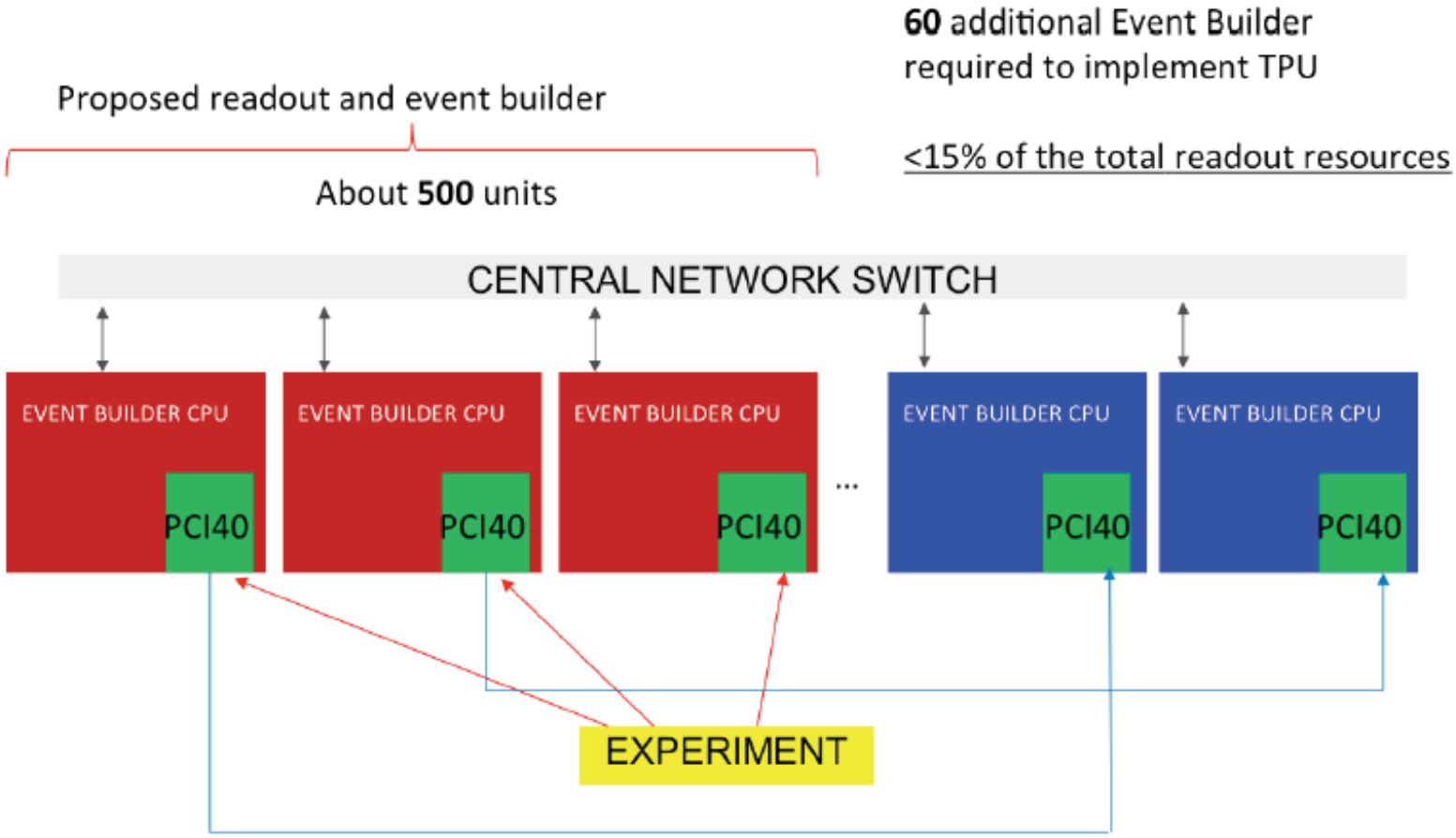
Front end




A solid angle "projective tower" of multiple detector layers

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

# Fit in LHCb's DAQ



# Field map