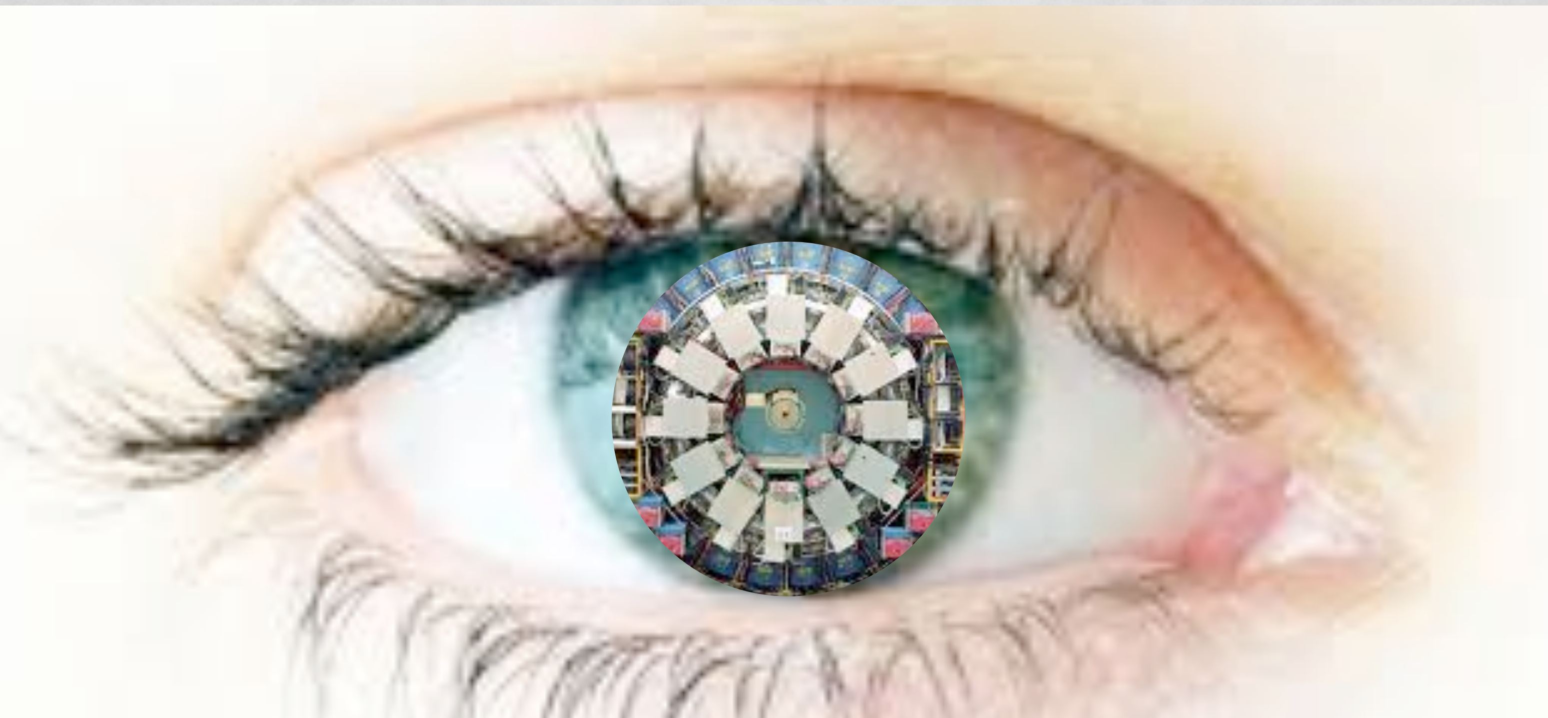
Tracking at 40 MHz



Diego Tonelli (CERN)

R&T seminar FNAL, May 13 2014

R&D project

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

(CERN/FNAL/Milano/Pisa)

The challenge

Tracks — among the better measured and information-rich objects available in high-energy collisions.

Often provide unique discrimination against backgrounds, especially in flavor physics (not only there).

Track information made available early can greatly enhance the reach in high rate environments.

But getting tracks in real time is hard.

Take high-lumi LHC: 40M pp interactions/s, each yielding O(100) charged particles. Can we reconstruct billions of tracks per sec. ?

- Massive combinatorial problem, calls for high parallelism
- Latency often an issue: calls for complex buffering

Collider folks have been attacking the problem since the early 80s.

Back in the days

Nuclear Instruments and Methods 217 (1983) 361–366
North-Holland Publishing Company

361

ON-LINE EVENT SELECTION WITH THE FAMP MICROPROC...

C. DAUM, H. DIJKSTRA, D. GOSMAN, C. HARDWICK, L. H...
G. de RIJK, H. TIECKE and L. WIGGERS
NIKHEF-H, Amsterdam, The Netherlands

In the experiment NA11 at the CERN SPS the MC68000 based FAM... trigger. It was used in a study of inclusive ϕ-meson production to s... ...ories and calculating the K^+K^- invariant mass. A trigger ra... ...ays per unit time by a factor of 8 has been achieve...

Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment

ELSEVIER

Volume 269, Issue 1, 1 June 1988, Pages 93–100

A fast hardware track-finder for the CDF central tracking chamber

G.W. Foster, J. Freeman, C. Newman-Holmes, J. Patrick

436

...ments and Methods in Physics Research A278 (1989) 436–440
North-Holland, Amsterdam

VLSI STRUCTURES FOR TRACK FINDING

Mauro DELL'ORSO
Dipartimento di Fisica, Università di Pisa, Piazza Torricelli 2, 56100 Pisa, Italy

Luciano RISTORI
INFN Sezione di Pisa, Via Vecchia Livornese 582a, 56010 S. Piero a G... ...aly

Received 24 October 1988

...designed to solve the track finding problem, ...ven for very high multiplicity events. This ...qual and each of them stores a number of ...om the detector while the detector is being

The AMchip: a full-custom CMOS VLSI associative memory for pattern recognition

S.R.Amendolia[+*], S.Galeotti[*], F.Morsani[*], D.Passuello[*], L.Ristori[#], G.Sciacca[#], N.Turini[x]
[*]*INFN-Pisa,* [#]*Univ. and INFN of Catania,* [+]*Univ. of Sassari,* [x]*Univ. and INFN of Bologna*
ITALY

Abstract

An Associative Memory full-custom CMOS VLSI

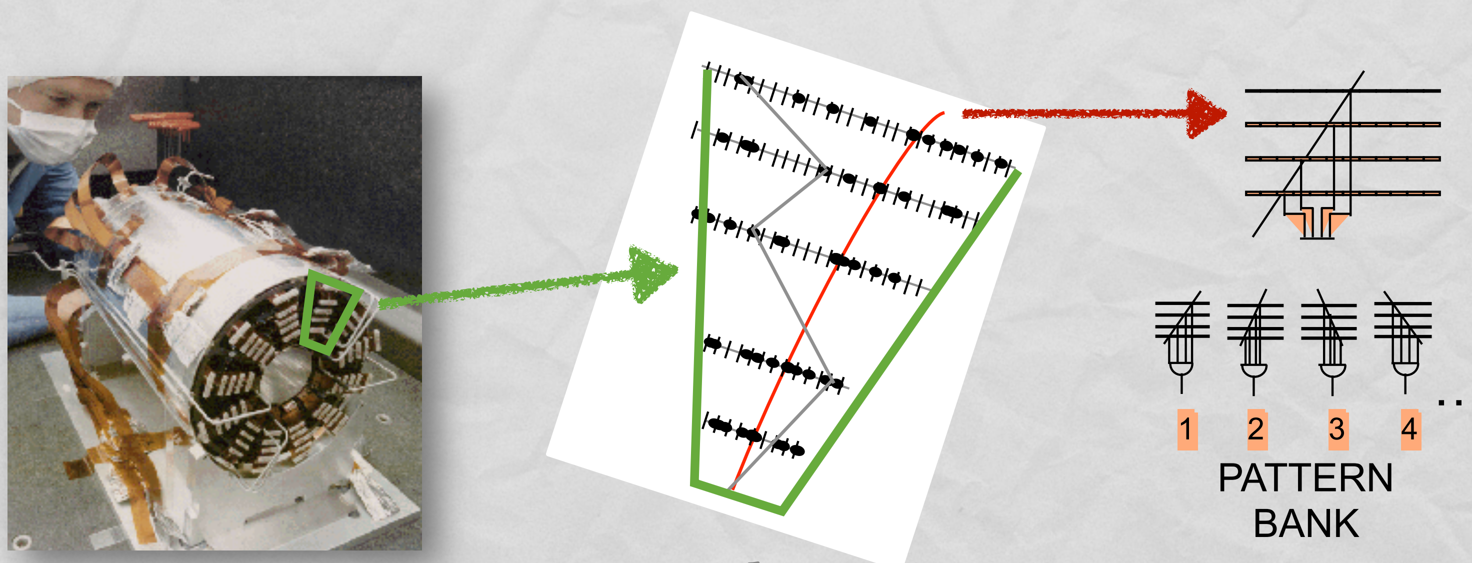
array consists of 128 rows and 5 columns for a total of 640 WORDs. Each row is called a PATTERN and each column is called a PLANE (see fig.1). The data bus is 60 bit wide so that

Pattern matching

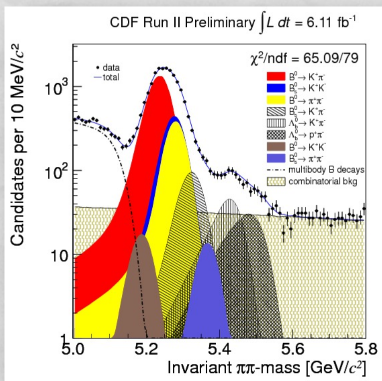
Pattern: a sequence of hits in the detector, represented by a set of coordinates.

A genuine particle trajectory is a specific sequence of hits.

Real hit coordinates are read out sequentially and compared in parallel to the set of all stored track patterns

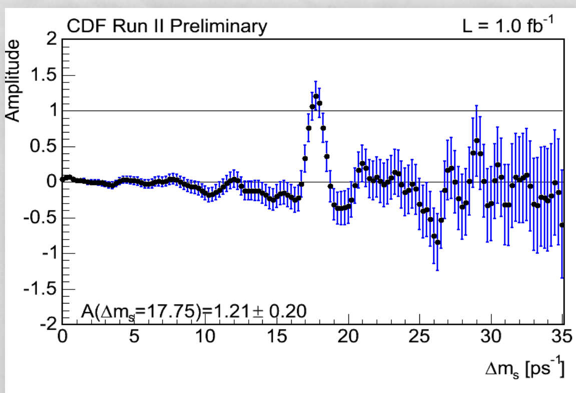


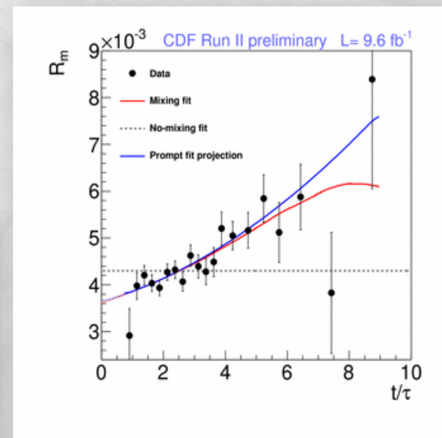
The payoff



Was motivated by this...

...ended up achieving something much more relevant..





plus, opened a whole new program, for free

"The great advances in science usually results from new tools rather than from new doctrines"

F. Dyson

Some numbers

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

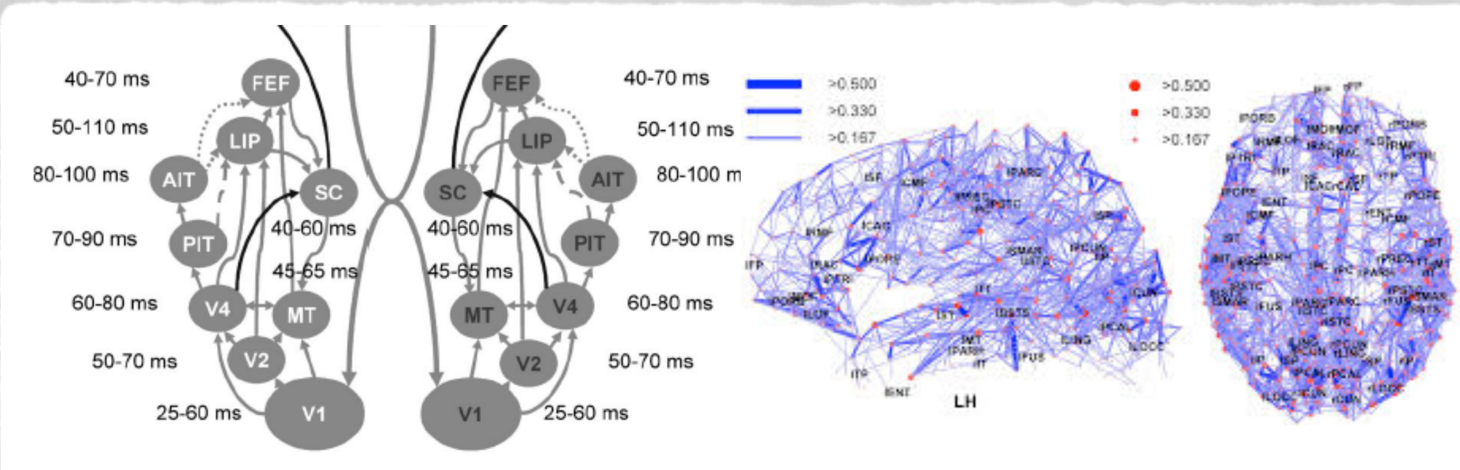
Perform tracking synchronous with LHC collisions appears daunting.

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, of course.

But SVT and FTK are based on Associative Memory, which are very parallel pattern-matching devices as well.

Key differences:

- Detector hit processing in 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

Retina cellular tracking

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

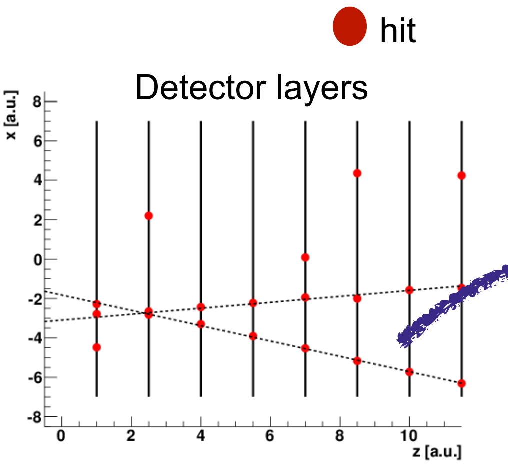
Abstract

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.

Acknowledgements

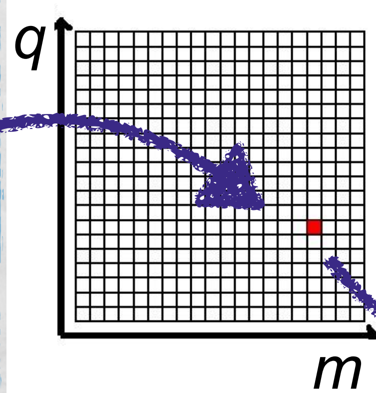
I wish to thank Giovanni Punzi for many useful discussions and Maria Michela Del Viva for directing me to the papers by Hubel and Wiesel [8,9] by which this work was inspired.

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

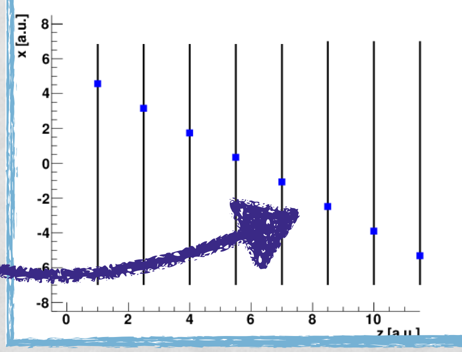


$$x = mz + q$$

Discretize space of track parameters in *cellular units*



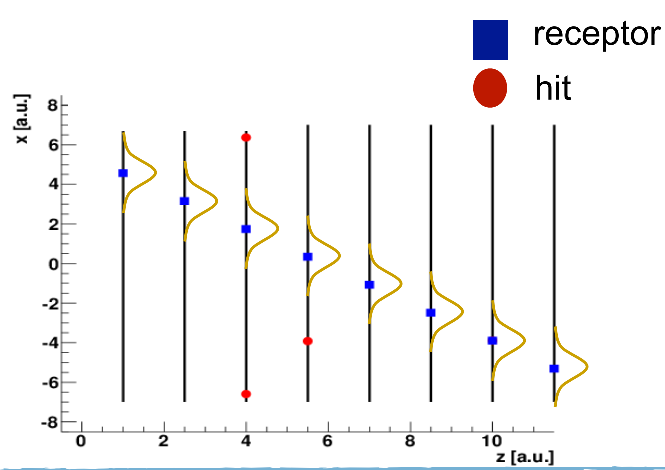
Each unit corresponds to a track that intersects each layer in *receptors*



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.

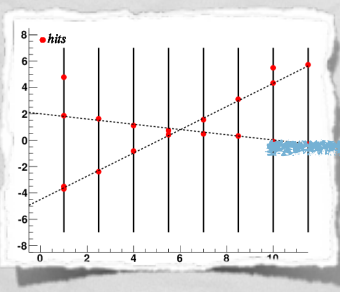
Then sum over all hits, all layers, to have the excitation of one cell (ij)

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

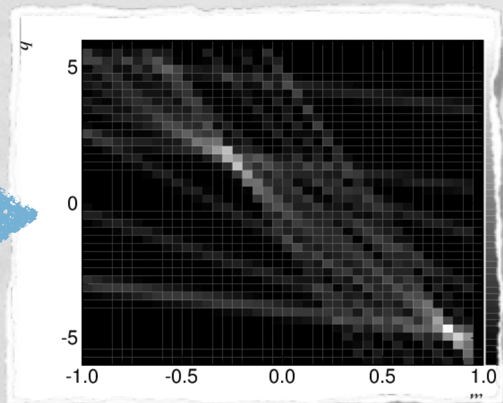


Retina response

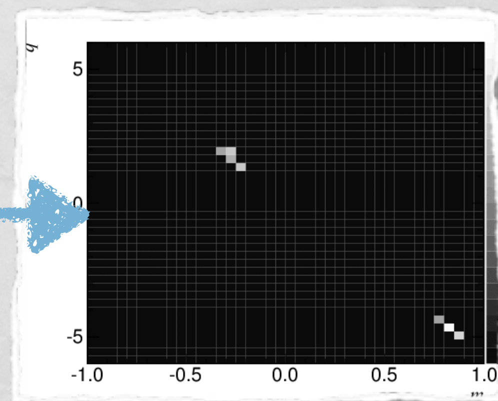
The response Rij of all the cells yields the response of the retina



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.

Core concept closely related to the Hough transform P.V.C. Hough Conf. Proc.C590914, 54 (1959)

However, a few crucial new 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.

Significant complexity leap in going from toy 2D to a realistic scenario

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

Implementation challenges

O(1000) hits on O(10) layers to reconstruct O(100) tracks.

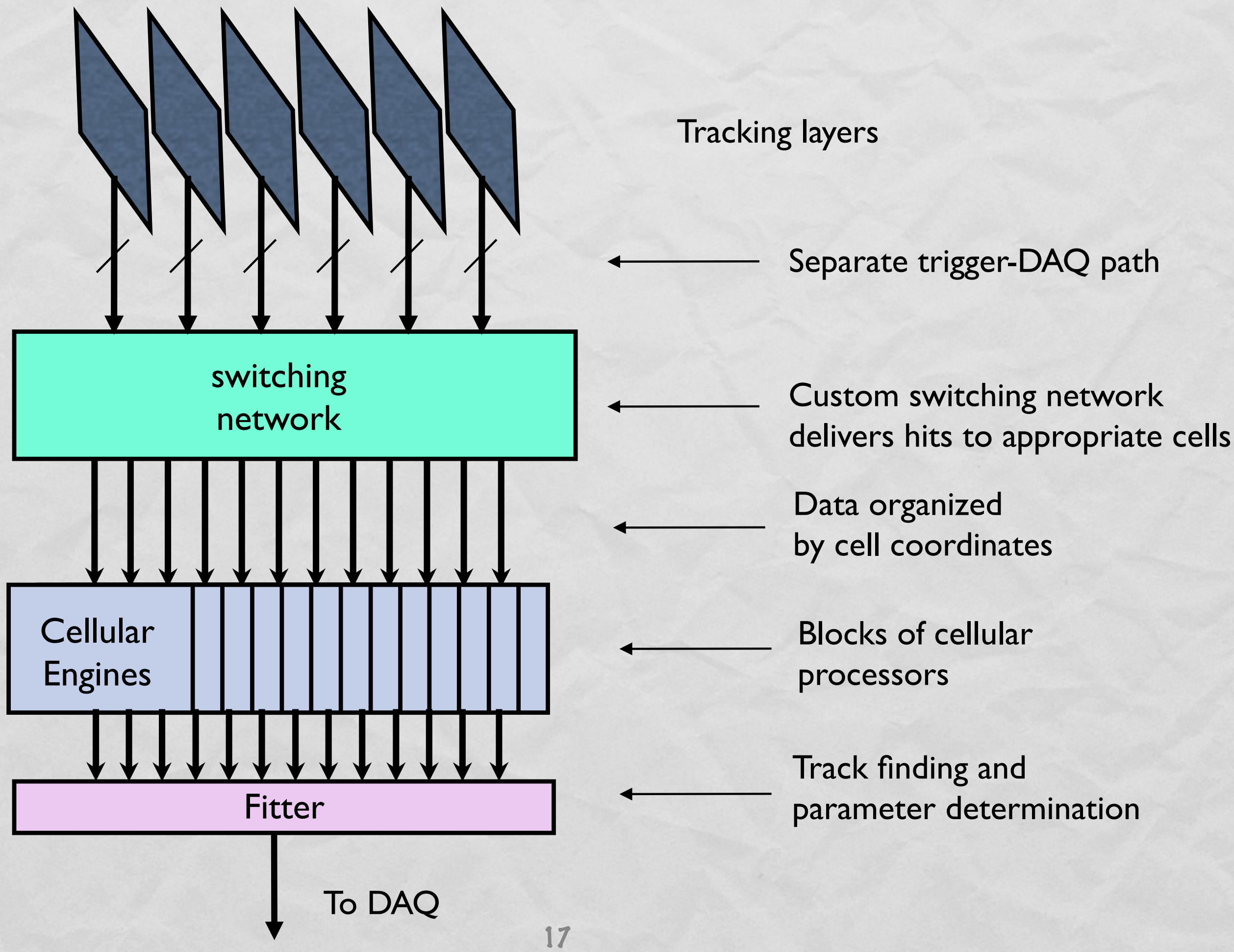
Every 25 ns.

- 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.

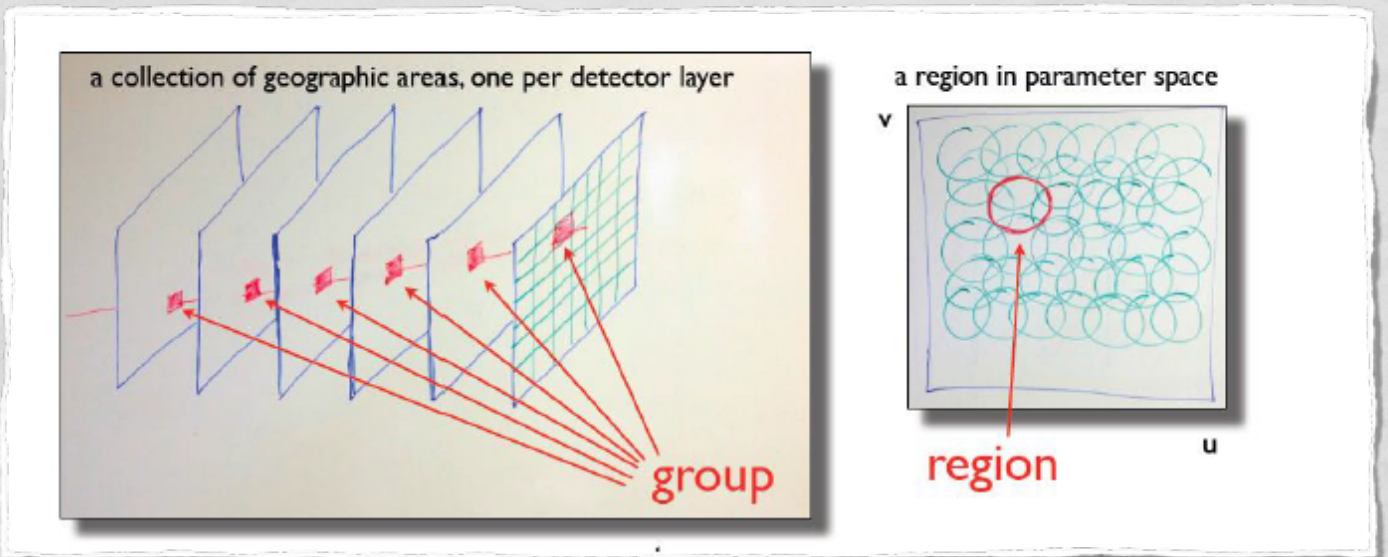
Lots of data, little time

Device logic

Architecture



Switching concept



Group: geographic area in each detector layer.

Each hit can only belong to one group.

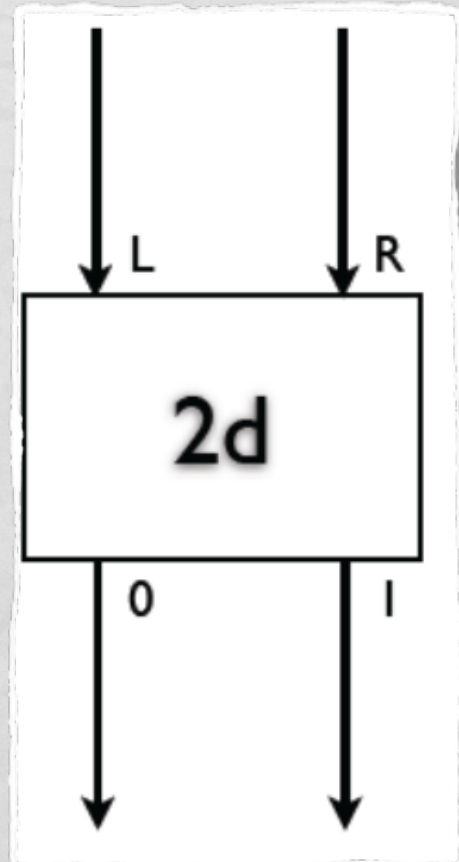
Each hit is only delivered to the union of all cells affected by all the hits in its group — the region associated with that group.

Switching basic unit

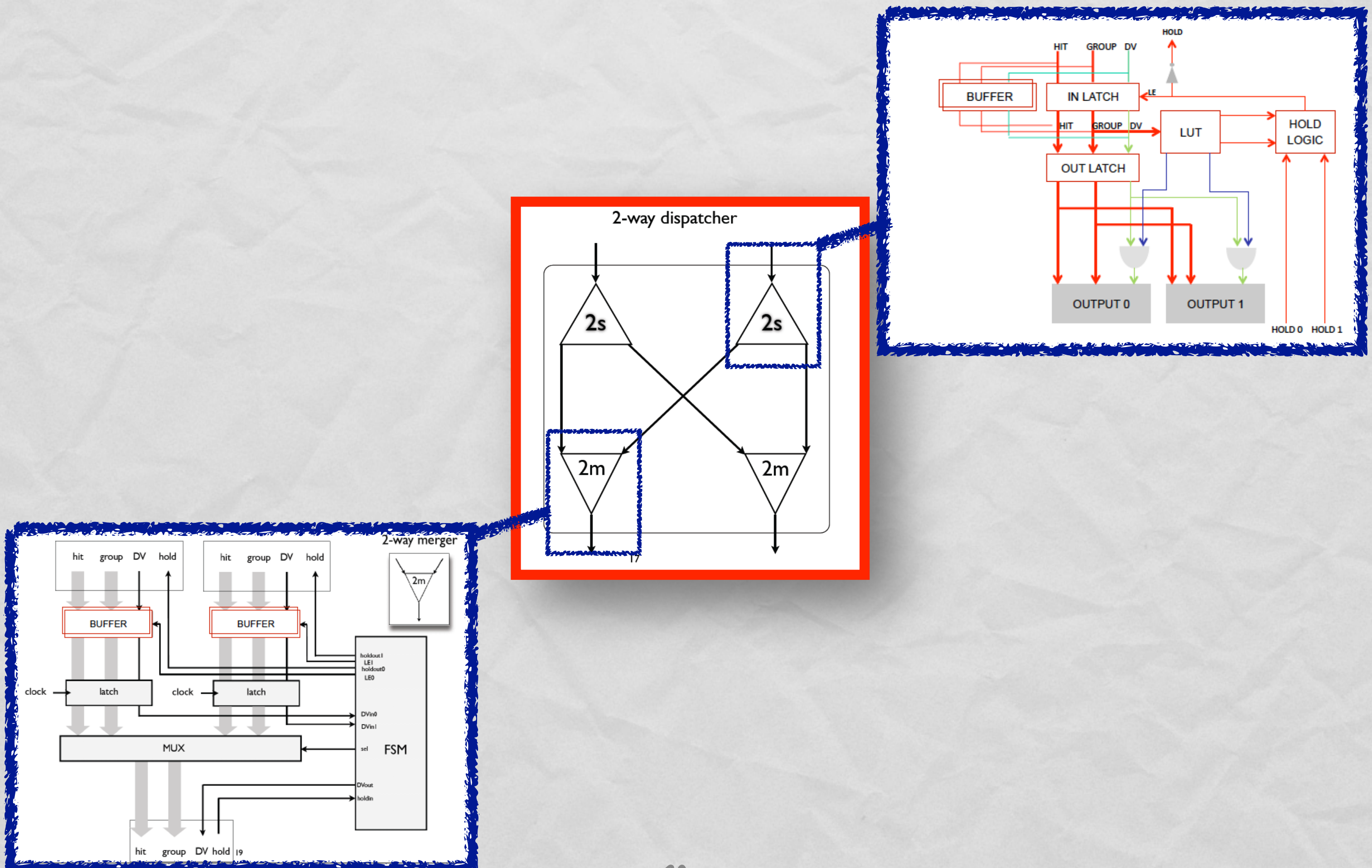
Information is carried by hits: 41-bits word containing the hit coordinates, layer ID, timestamp...

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 are held.



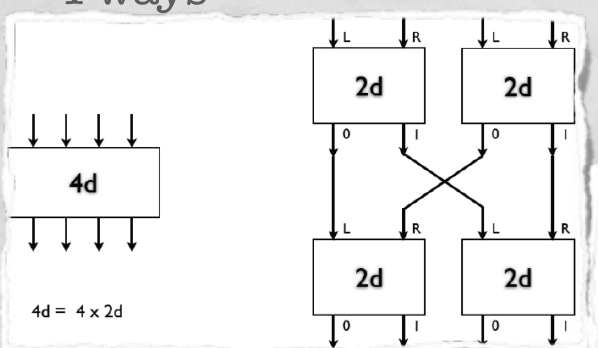
Look inside..


2-way dispatcher
2s
2s
2m
2m
HIT
GROUP
DV
HOLD
BUFFER
IN LATCH
LE
HIT
GROUP
DV
LUT
HOLD
LOGIC
OUT LATCH
OUTPUT 0
OUTPUT 1
HOLD 0
HOLD 1
2-way merger
hit
group
DV
hold
BUFFER
clock
latch
MUX
FSM
hit
group
DV
hold

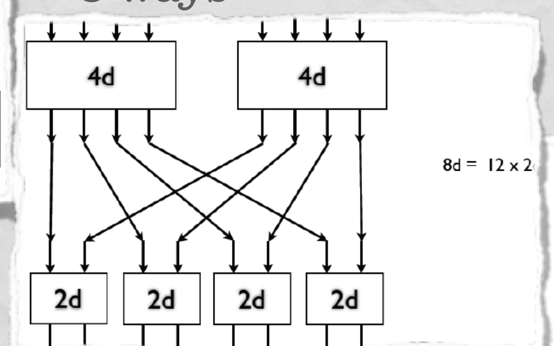
Network

Combination of dispatchers builds whole network.

4 ways



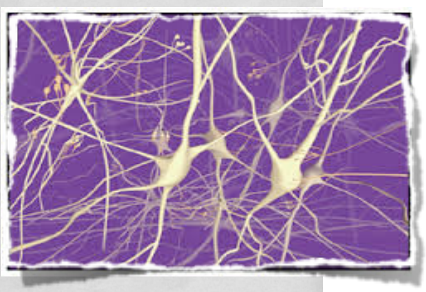
8 ways

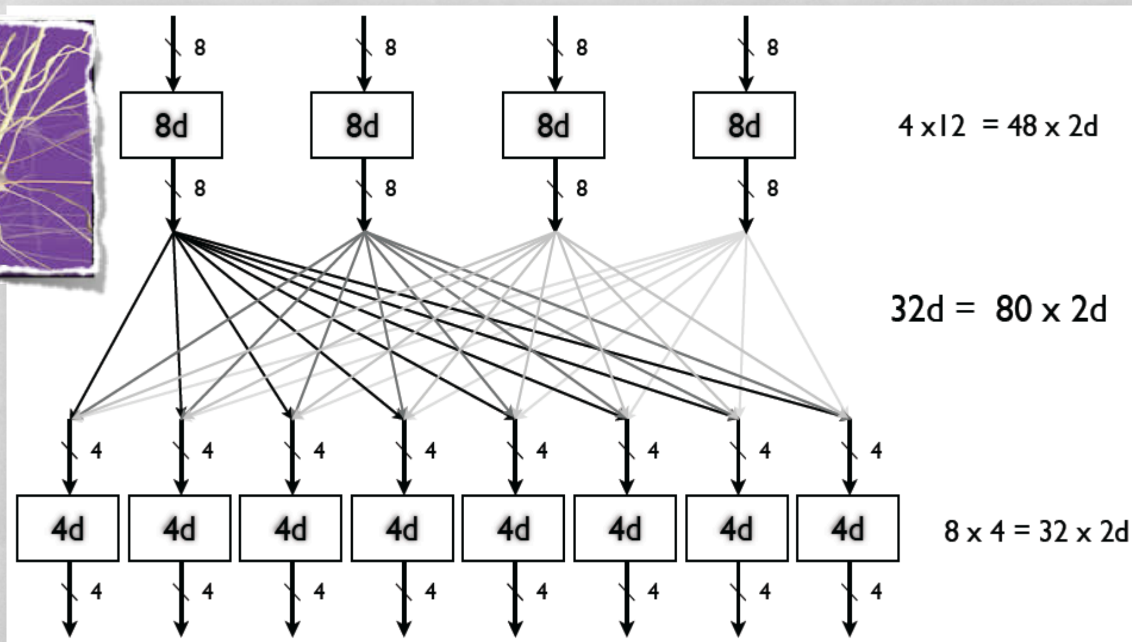


....

N×N requires $\log_2(N)N/2$ elements

Intelligent delivery





Each hit comes with a “zip-code”

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

Excitation concept

Each cell is defined as a logic module, the engine.

Layer ID determines the appropriate cartesian coordinates (center of the receptor) to be subtracted from hit coordinates

Outcome is squared, summed, and the result R is rounded by keeping the 8 least significant bits

A sigma function common to all engines is mapped into a LUT

The rounded result is used as address to the LUT.

Outputs of the LUT are accumulated for each hit of the event

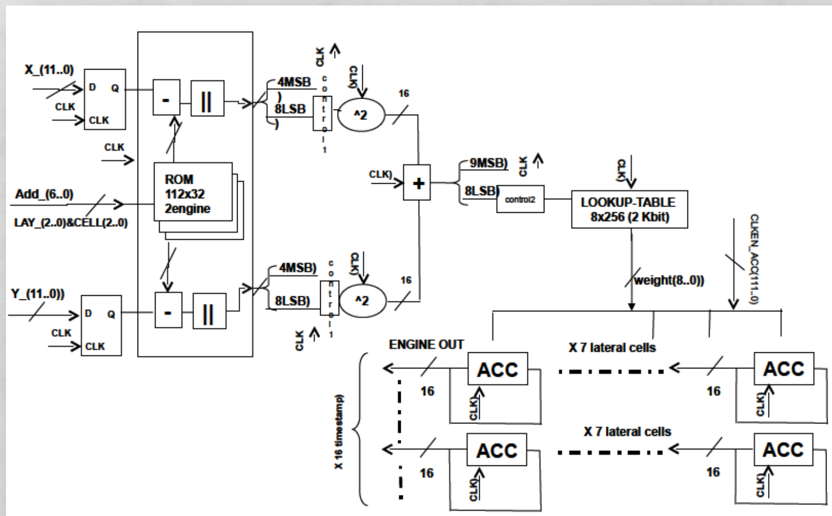
Each hit is cycled multiple times to compute excitation in lateral cells

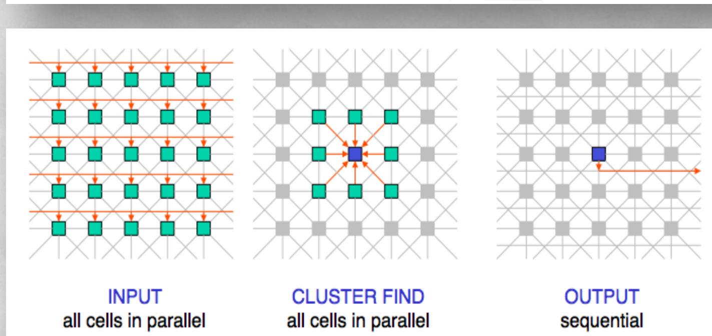
Retina cell implementation

Clocked pipeline.

Performs calculation of weights for a hit into a cell and deals with surrounding cells as well

Second stage performs local clustering in parallel and queues results to output



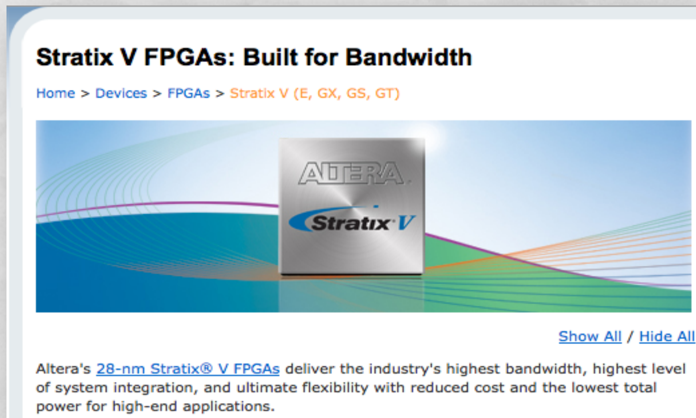


Implementation

High-end FPGA devices.

Less powerful than ASICs, but well suited to prototype the problem.

Fully reconfigurable, relatively easy to program and simulate



Popular choice for complex projects on small number of units (CT scanners, high-end radars)

Exploit progress driven by telecommunications:

- Large I/O capabilities. Now O(TB/s) with optical links
- Large internal bandwidth.

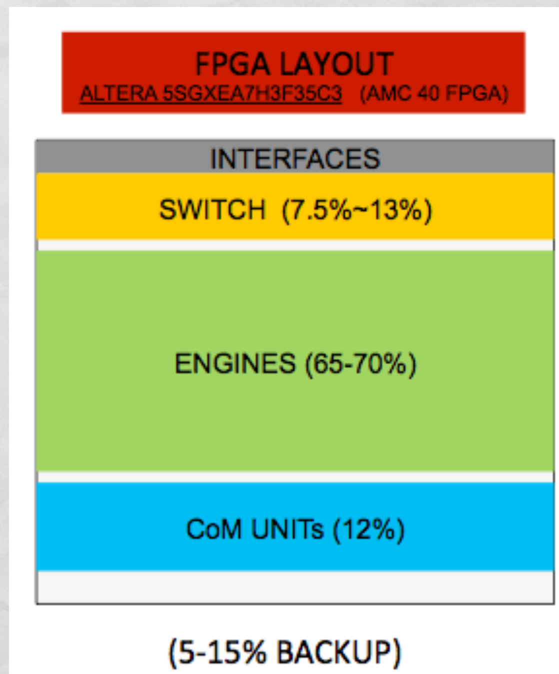
Placing

All main components implemented in VHDL and placed on the FPGA

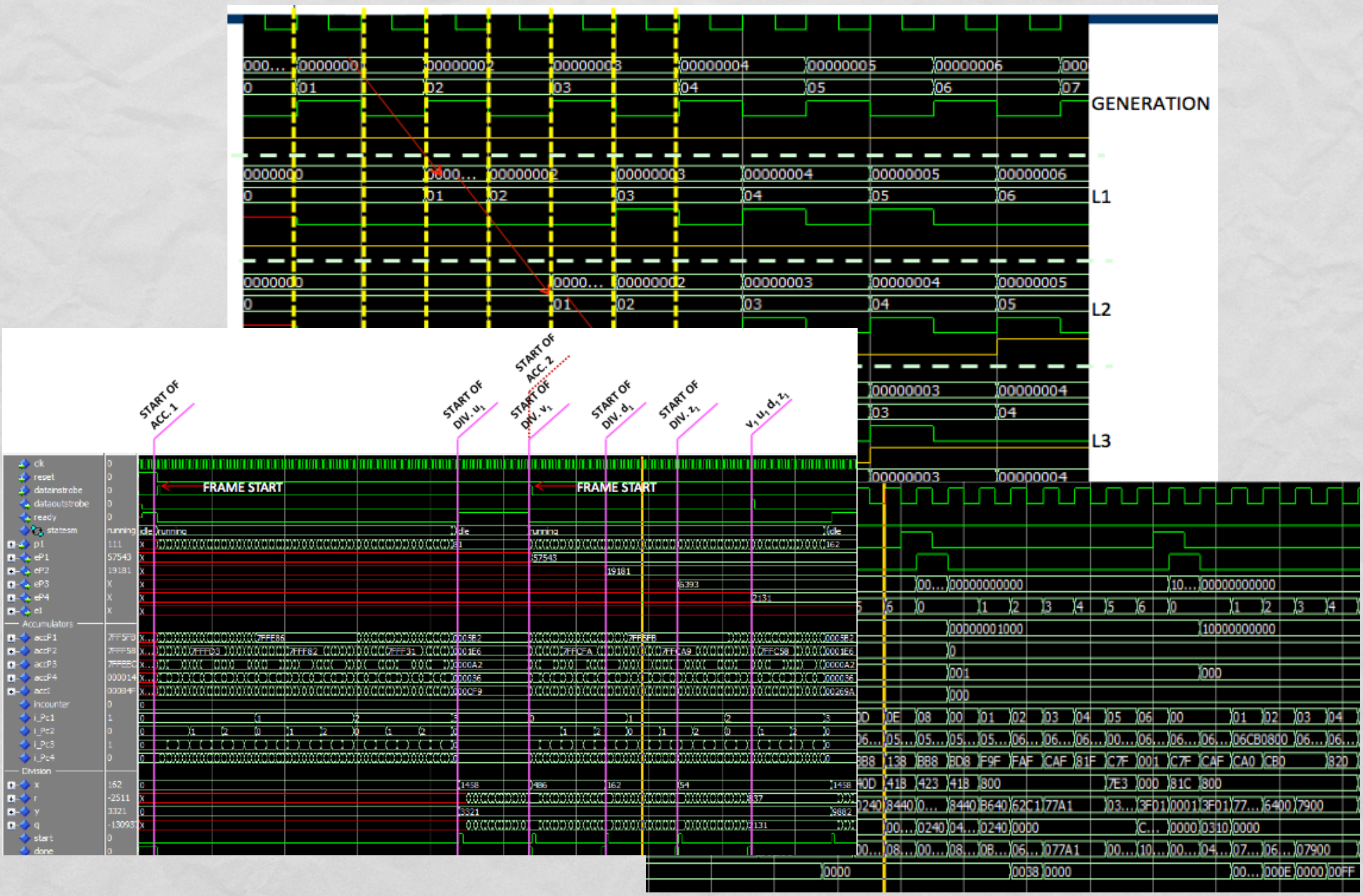
Can fit O(1000) engines per chip.

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

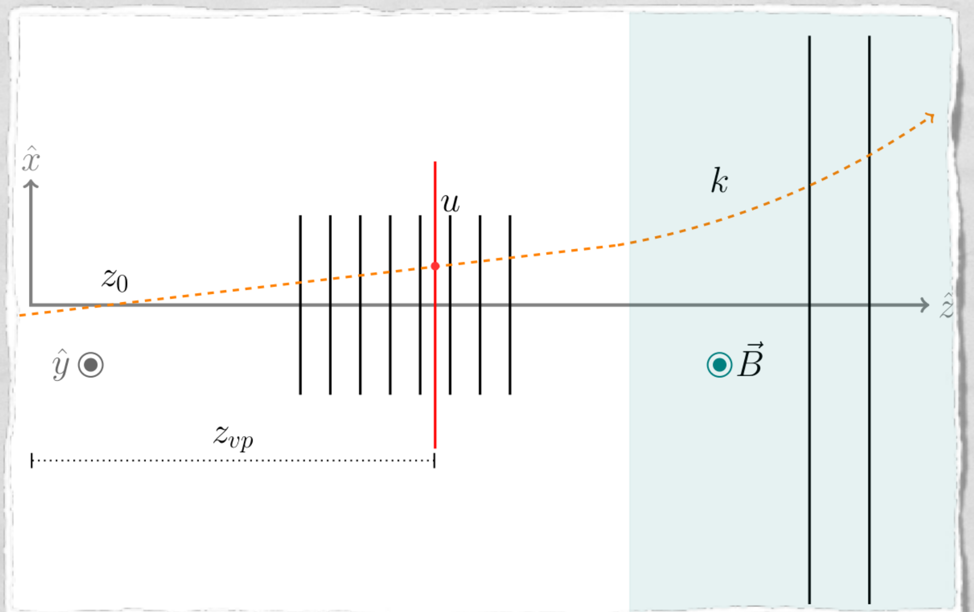
Total latency about 125 clock cycles at 350 MHz.

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.

Can we do it for real?

Basics

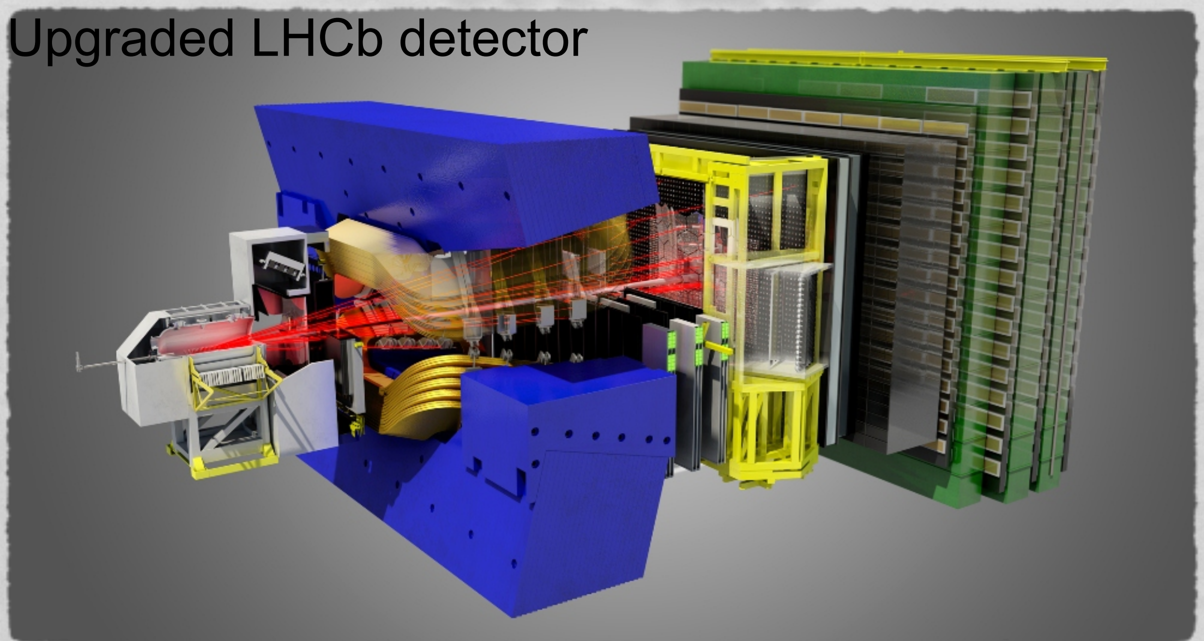


Array of silicon pixel detectors, each providing (x, y) at fixed z. Magnetic field kick.

Measure tracks in 3D — five parameters

No need to assume uniform B or ideal alignment.

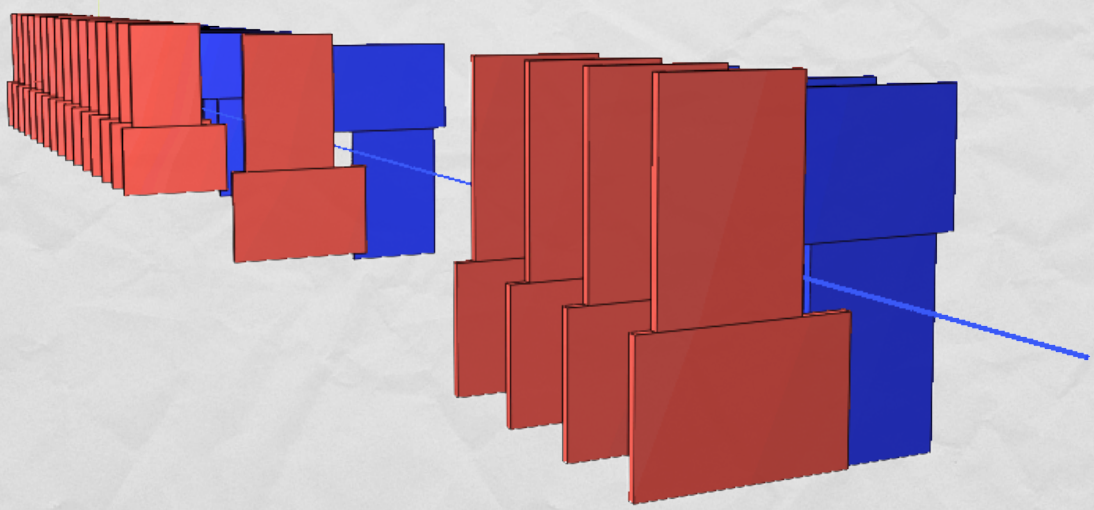
Use case



LHCb is a nicely-fitting application: flavor physics at high luminosity, heavily track-based

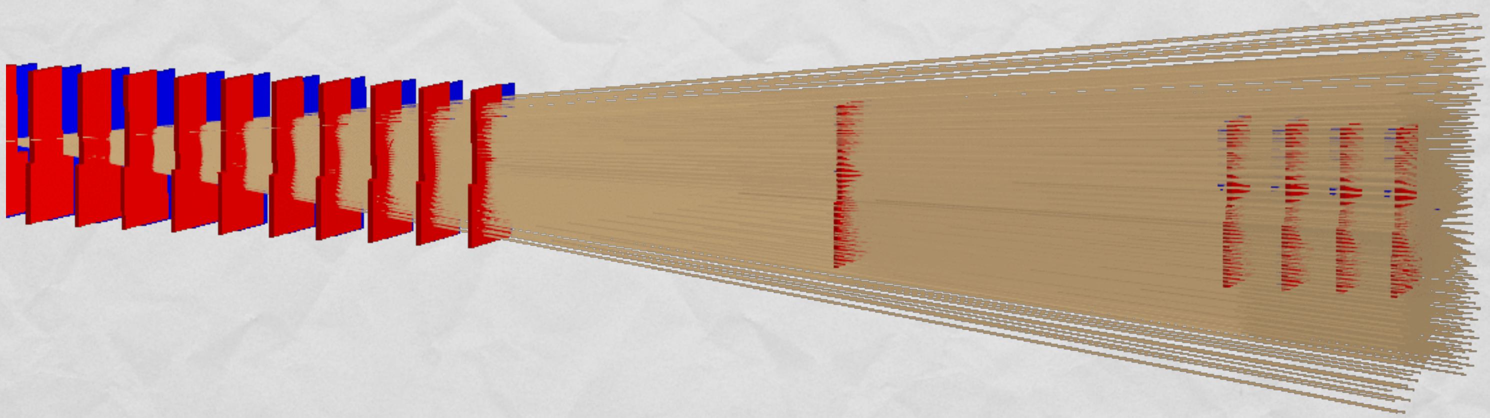
Will read out all events at 40 MHz in the upgrade (2020-)

Upgrade LHCb tracking

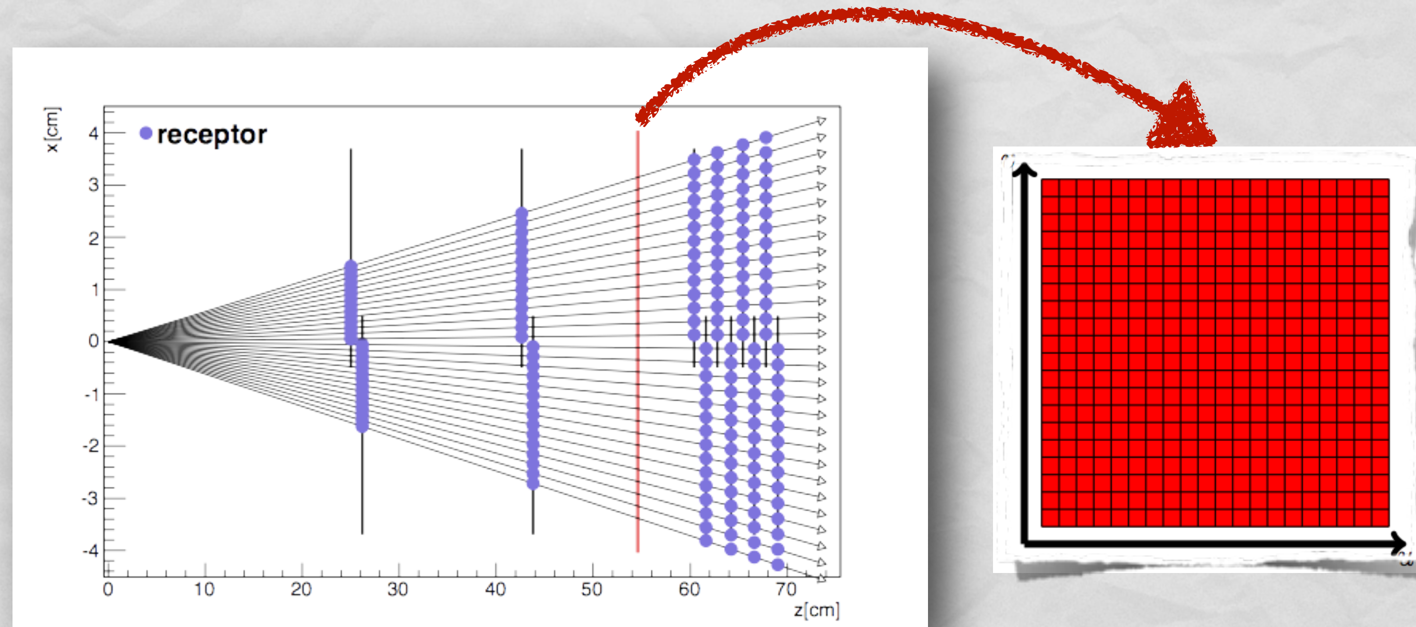


- Vertex detector based on silicon pixel technology
- Two layers of microstrips in the 0.05 T fringe field of the magnet to get some momentum sensitivity.

..more realistic



Cellular mapping



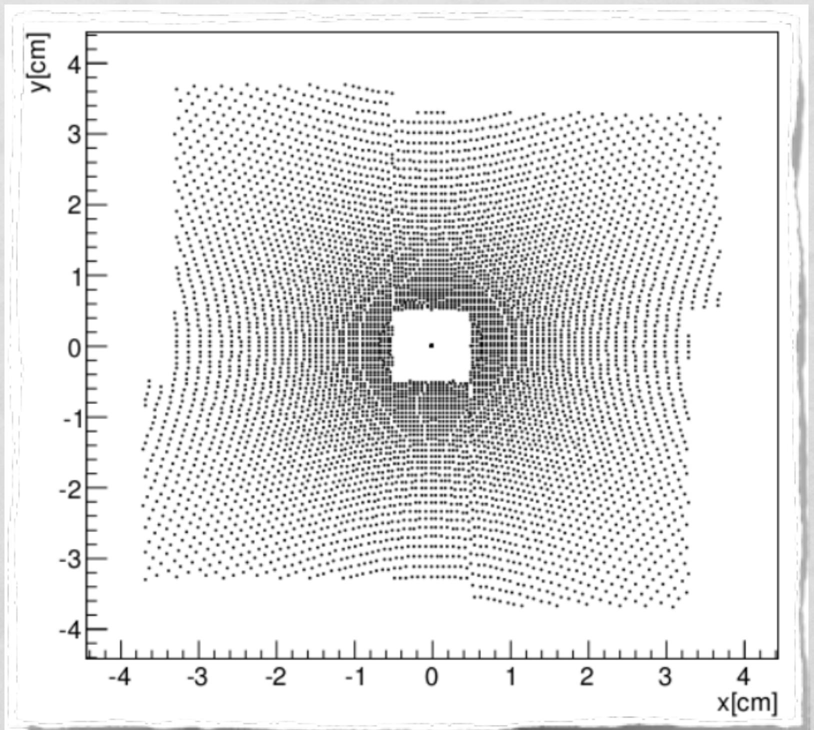
Intuitive parametrization: two “main” parameters given by intersection of the track on an arbitrary plane.

Map onto a 2D main grid.

Receptors layout

Intersections of “base tracks” on each layer gives a map of receptors

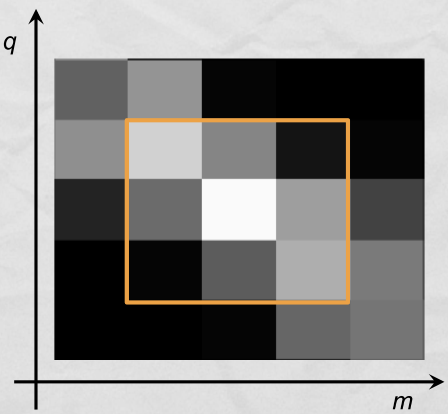
Optimal operation suggests nonuniform receptor distribution.



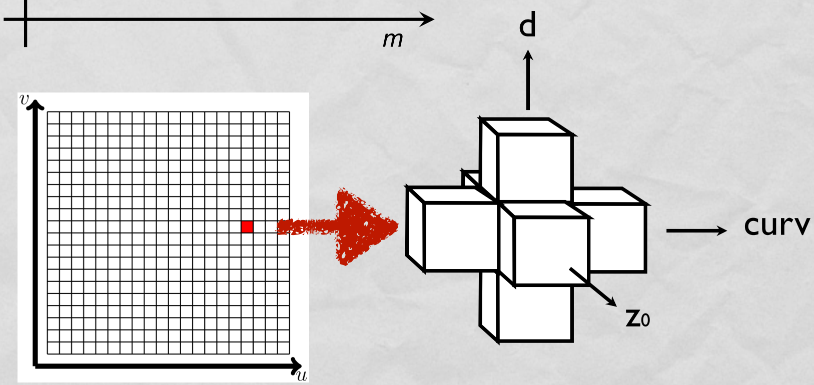
Once the pattern recognition is done, remaining track parameters implemented as a perturbative correction in a separate step.

Parameter determination

Principal parameters u,v directly from cluster centroid.



Determine other 3 parameters by interpolating response of "lateral cells"



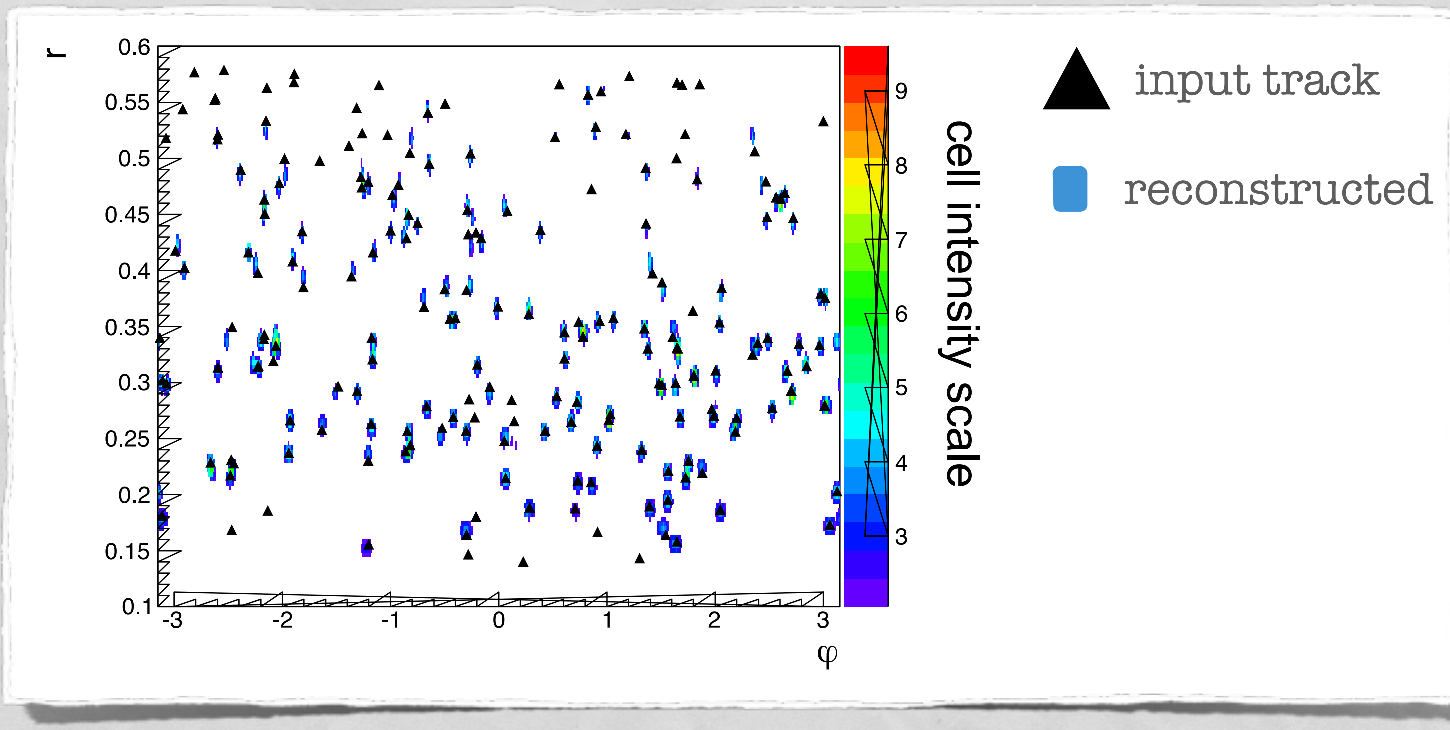
Other options (local linearized fit) possible.

Performance

Display

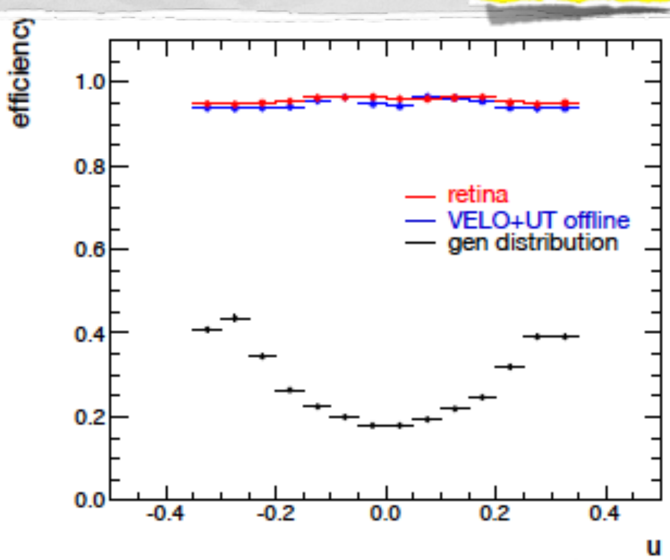
Standard LHCb simulation as input.

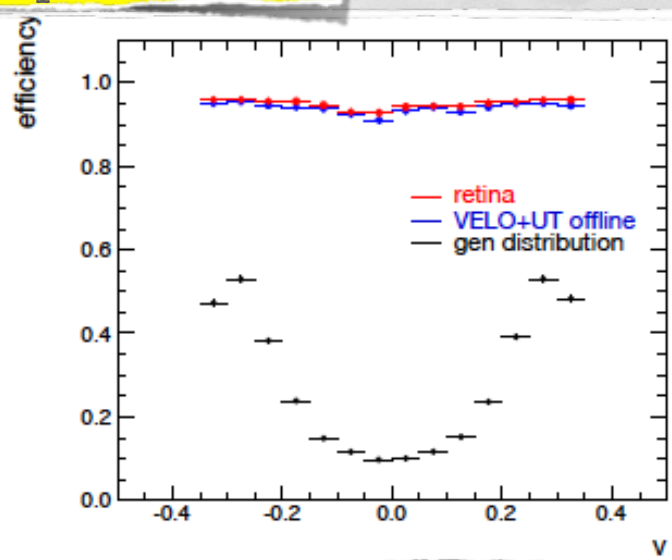
L=10^{33} cm^{-2} Hz (Poisson centered at 7.6 interactions per beam x-ing)



Efficiency

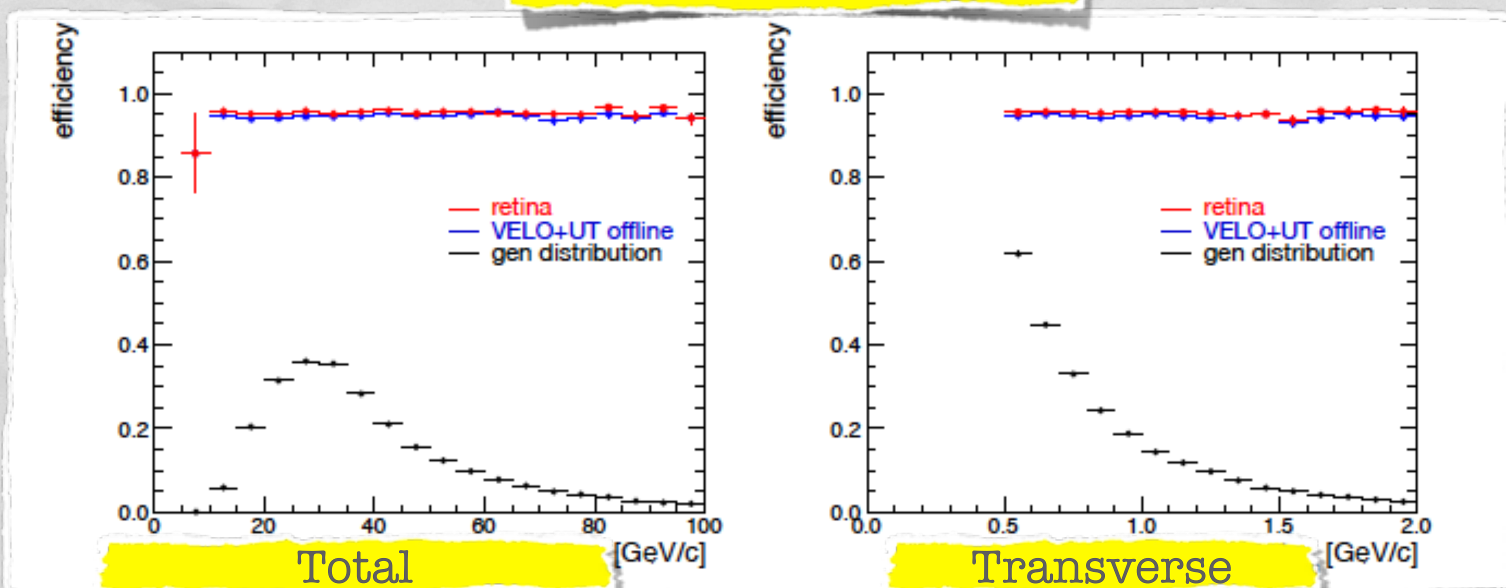
Cartesian parameters





Efficiency

Particle momentum

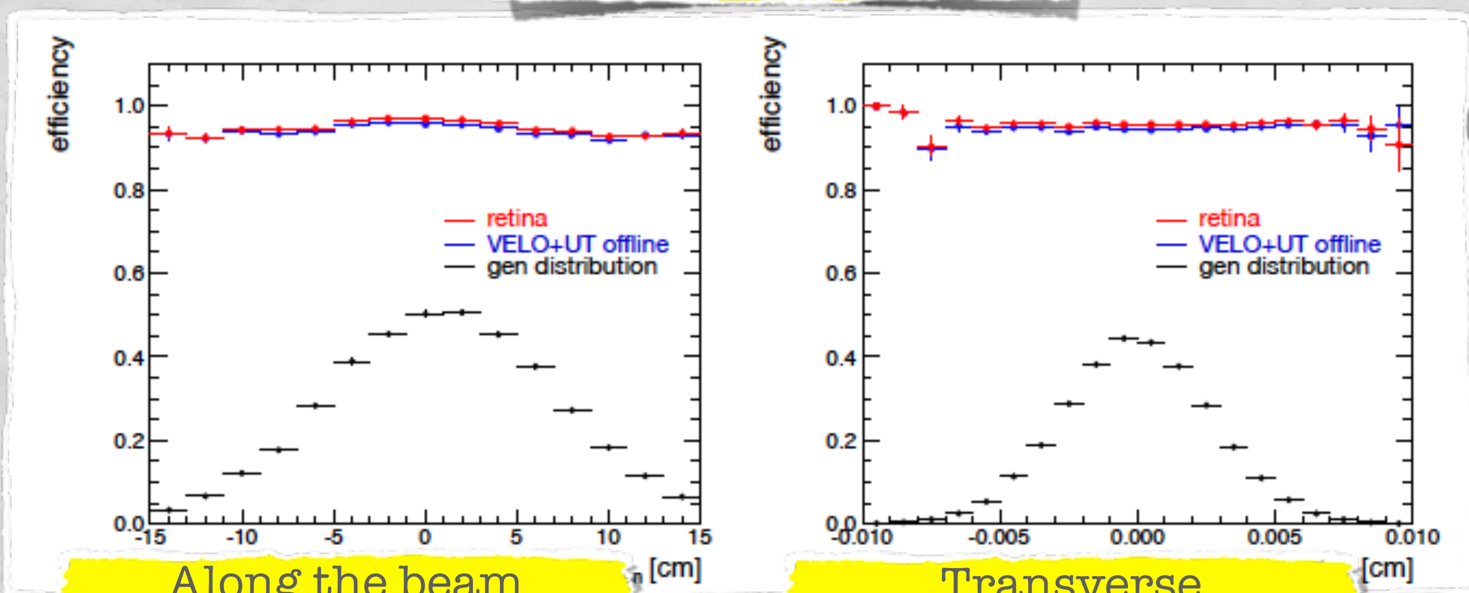


Total

Transverse

Efficiency

Particle origin



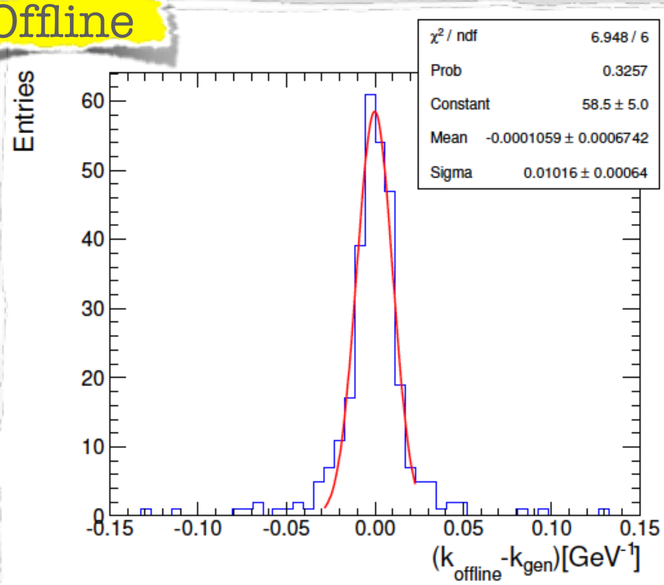
Along the beam

Transverse

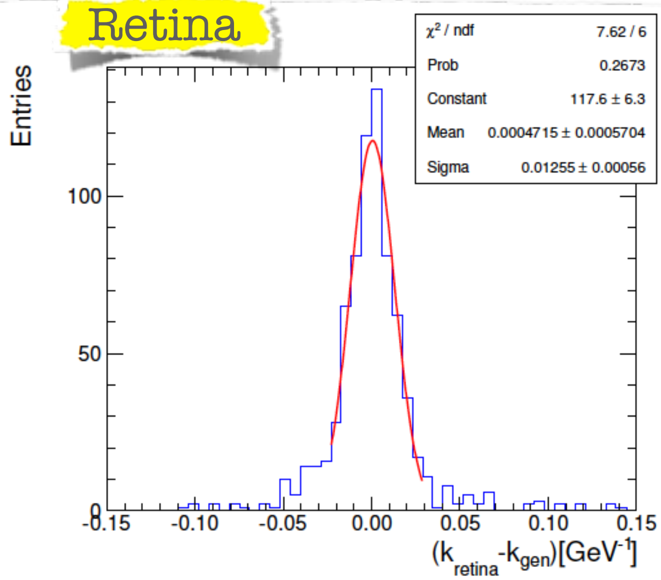
Performance

Track curvature resolution

Offline



Retina



Unbiased determination. Resolution comparable with offline

Summary

- ☐ Real-time reconstruction of charged particle trajectories, the key to fully exploit LHC potential, especially for quark-flavor physics.
- ☐ Current pattern-matching algorithms insufficient.
- ☐ Implemented a realistic model of a novel algorithm inspired by mammals vision and suitable for pixel detectors subject to very high track rates
- ☐ Designed in detail the architecture of the device and simulated it in realistic experimental conditions.
- ☐ Reconstruct tracks at 40 MHz with offline-like resolutions and efficiency. This is 400 times faster than any existing or foreseen device.

Effectively an additional detector that outputs directly tracks

The end

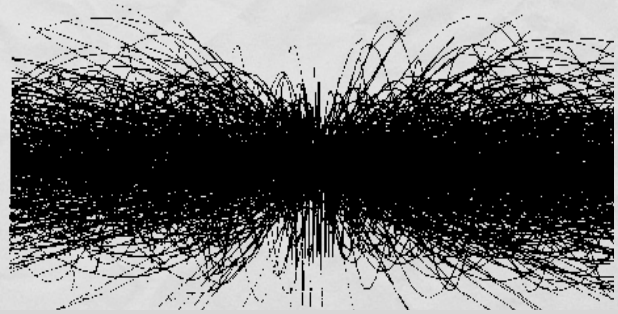
Data-reduction compared

Vision

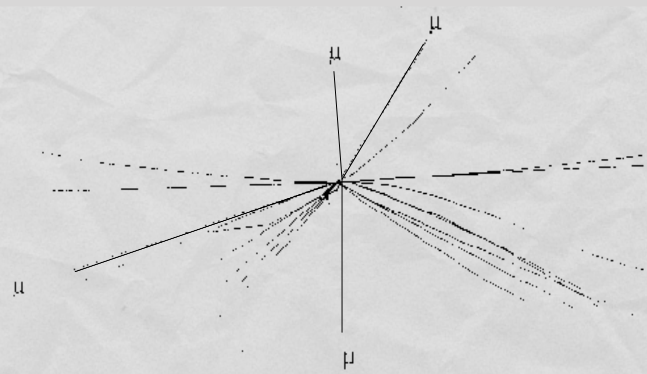
Visual system may achieve efficient reduction by creating a compact summary of the image based on few simple features Marr (1982)

HEP

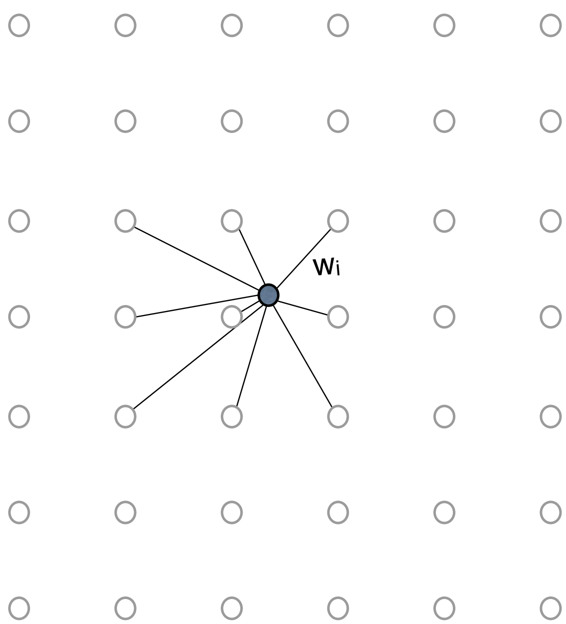
Full information



Relevant information

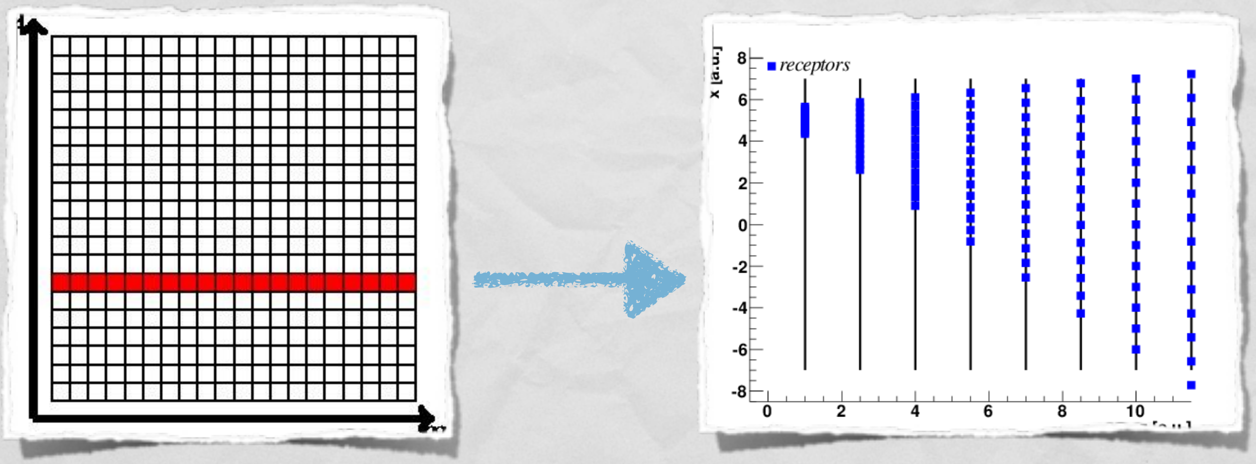


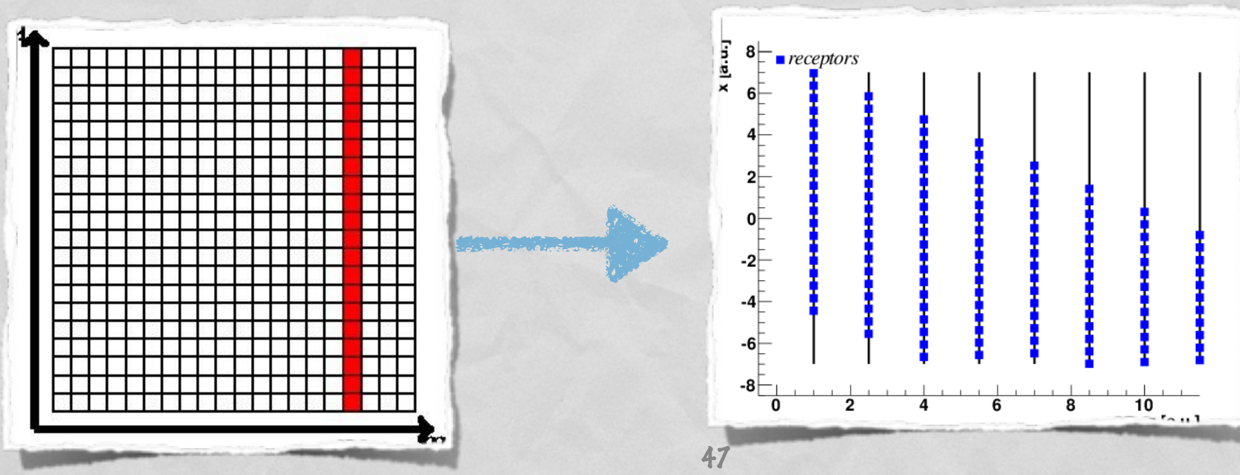
one-hit excitation



$$R = \sum w_i = \sum e^{-\frac{s_i}{2\sigma^2}}$$

Mapping


receptors
x [a.u.]
8
6
4
2
0
-2
-4
-6
-8
0
2
4
6
8
10


receptors
x [a.u.]
8
6
4
2
0
-2
-4
-6
-8
0
2
4
6
8
10