Online Track and Vertex Reconstruction on GPUs for the Mu3e Experiment

Dorothea vom Bruch

March 7th 2017

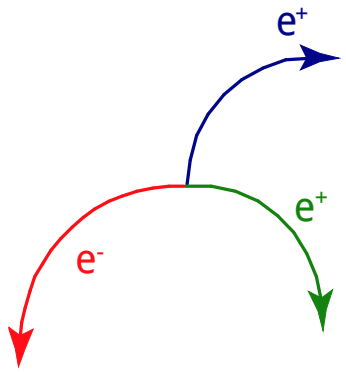
Connecting the Dots / Workshop on Intelligent Trackers 2017

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

JG|U

Mu3e Signal

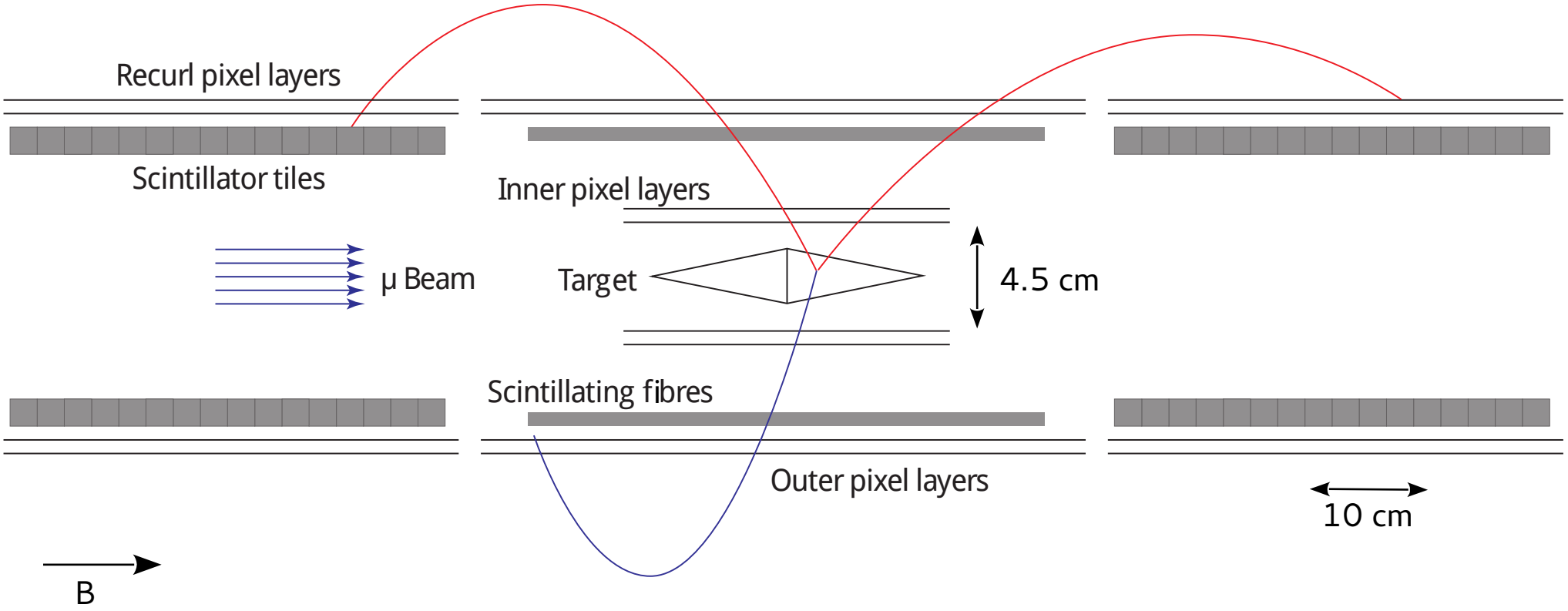
Search for charged lepton flavour-violating decay $\mu^+ \rightarrow e^+e^-e^+$ with a sensitivity in branching ratio better than 10^{-16}



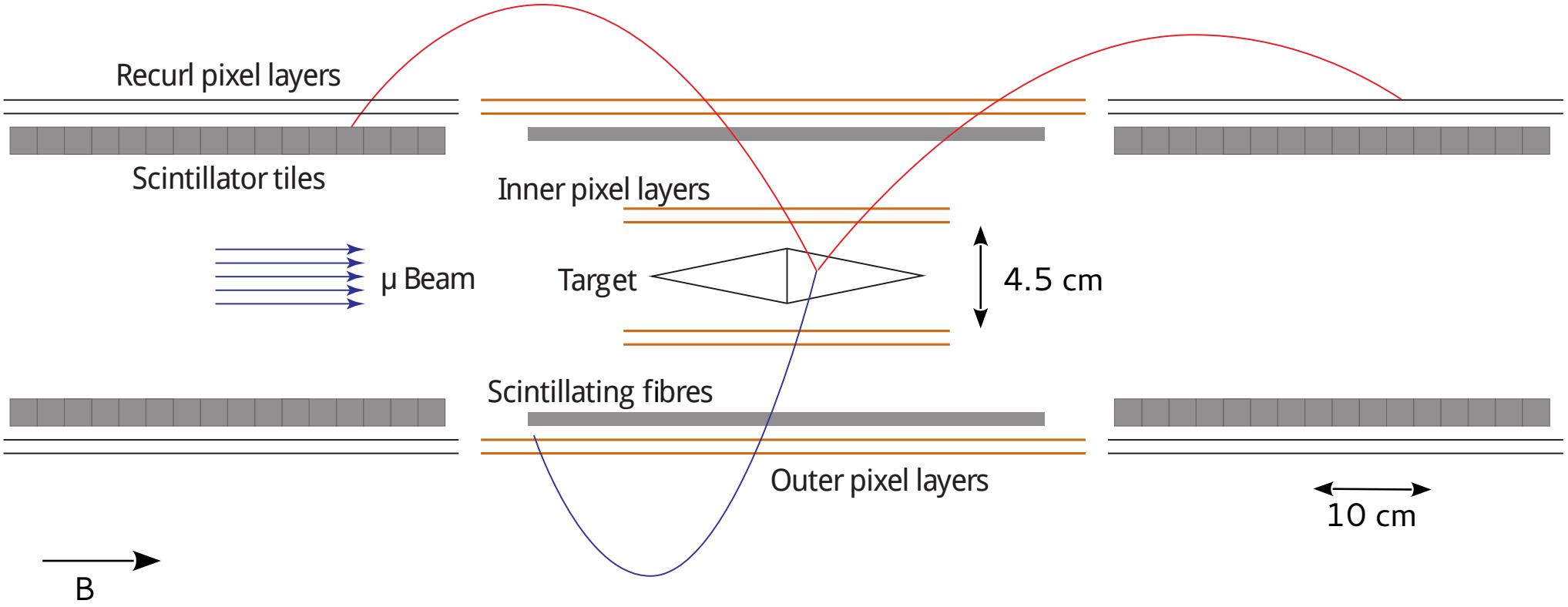
Signal

- Coincident in time
- Single vertex
- $\sum \vec{p}_i = 0$
- $E = m_\mu$

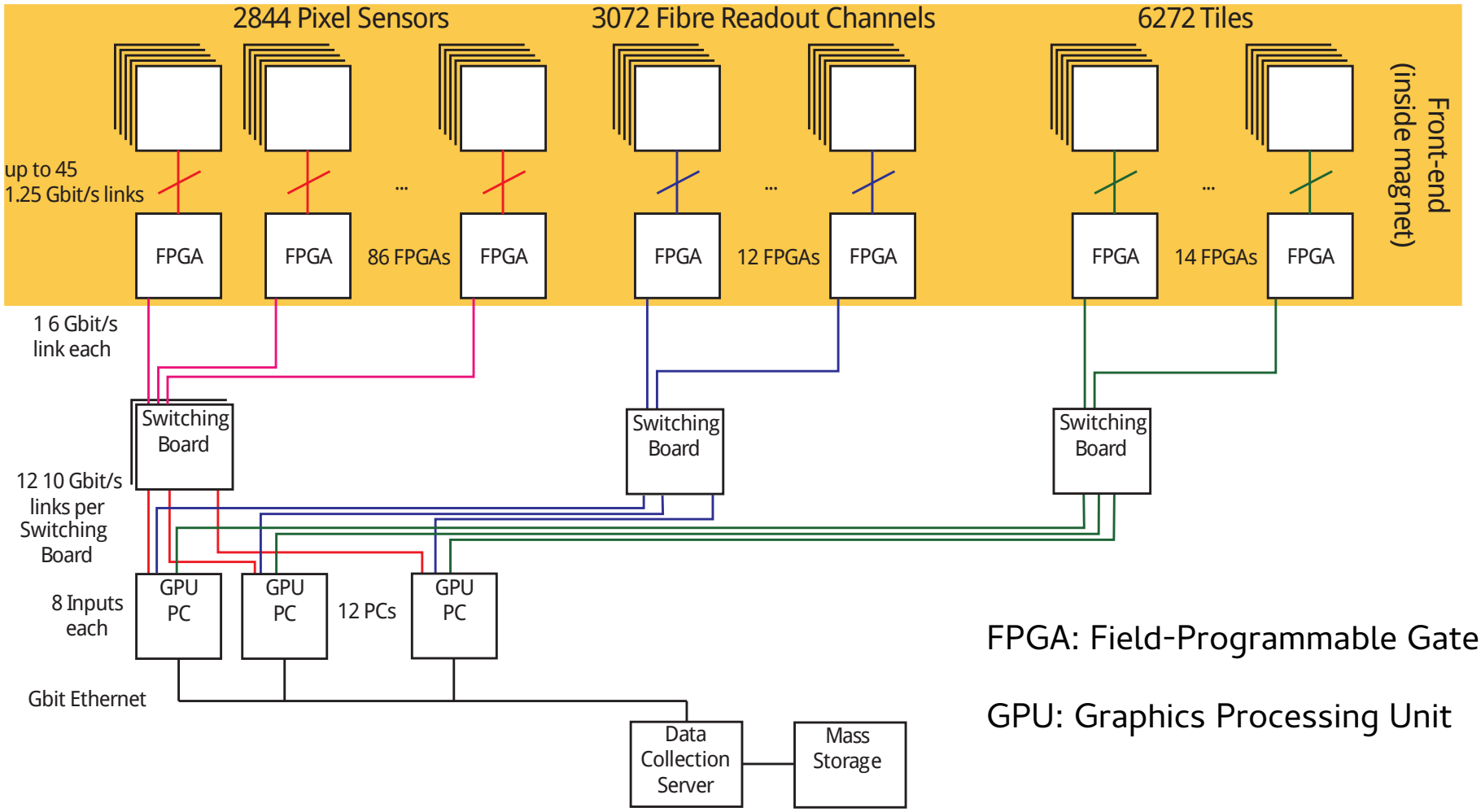
The Mu3e Detector



The Mu3e Detector



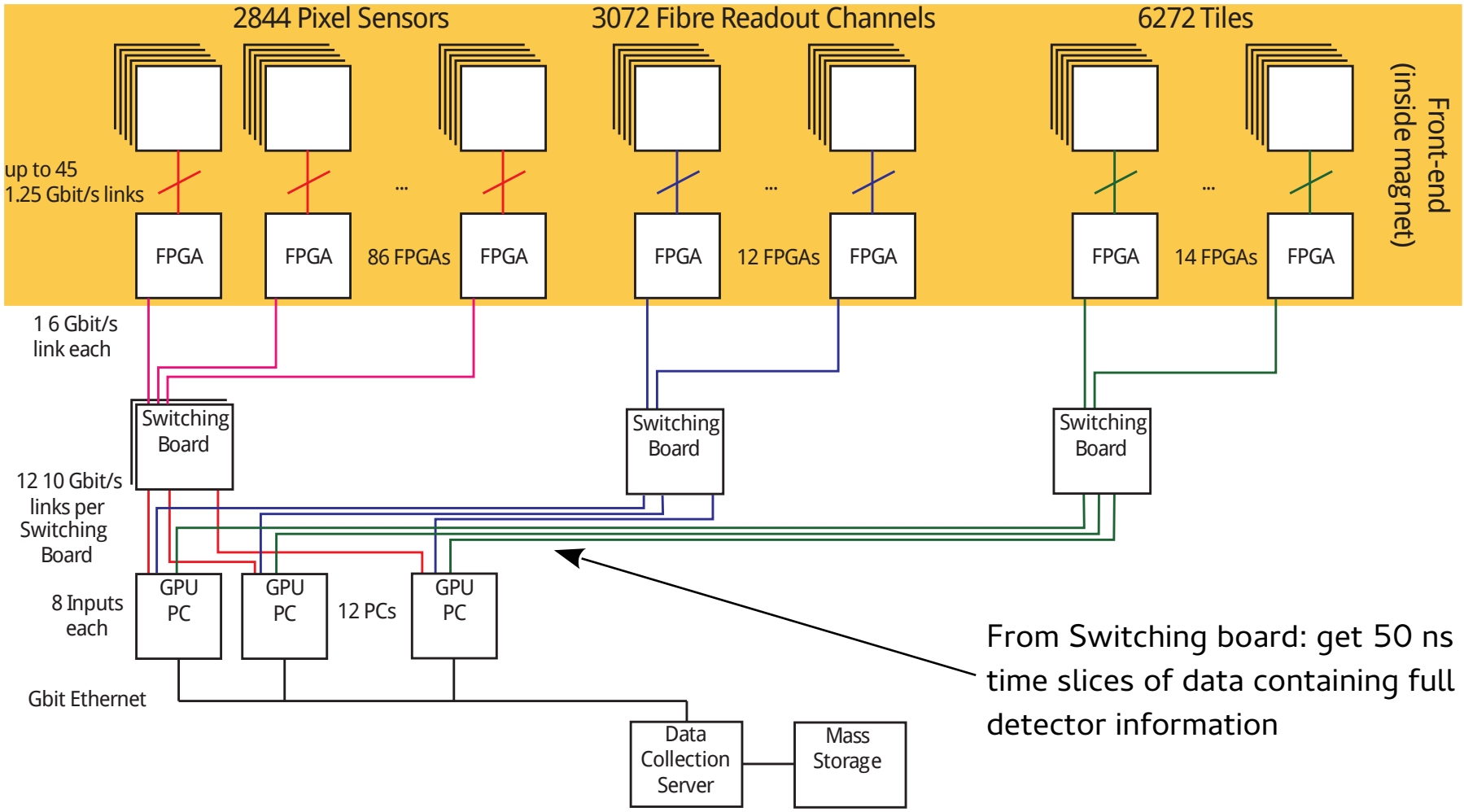
Readout Scheme



FPGA: Field-Programmable Gate Array

GPU: Graphics Processing Unit

Readout Scheme



Readout Rate

At a rate of 10^8 muons / s

Triggerless, zero-suppressed readout

	Data rate [Gbit / s]
Pixel detector	40
Fiber detector	20
Tile detector	negligible
Total	**~ 60**

Need factor ~80 reduction to reach 100 MB/s

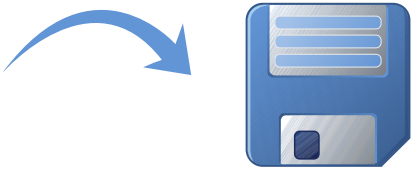
Readout Rate

At a rate of 10^8 muons / s

Triggerless, zero-suppressed readout

	Data rate [Gbit / s]
Pixel detector	40
Fiber detector	20
Tile detector	negligible
Total	**~ 60**

Need factor ~80 reduction to reach 100 MB/s

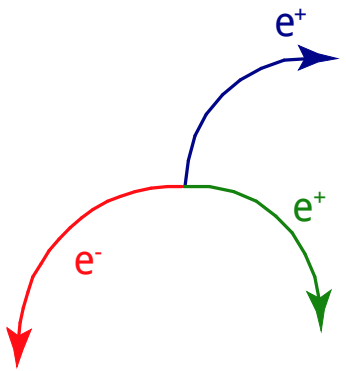


Selection Process

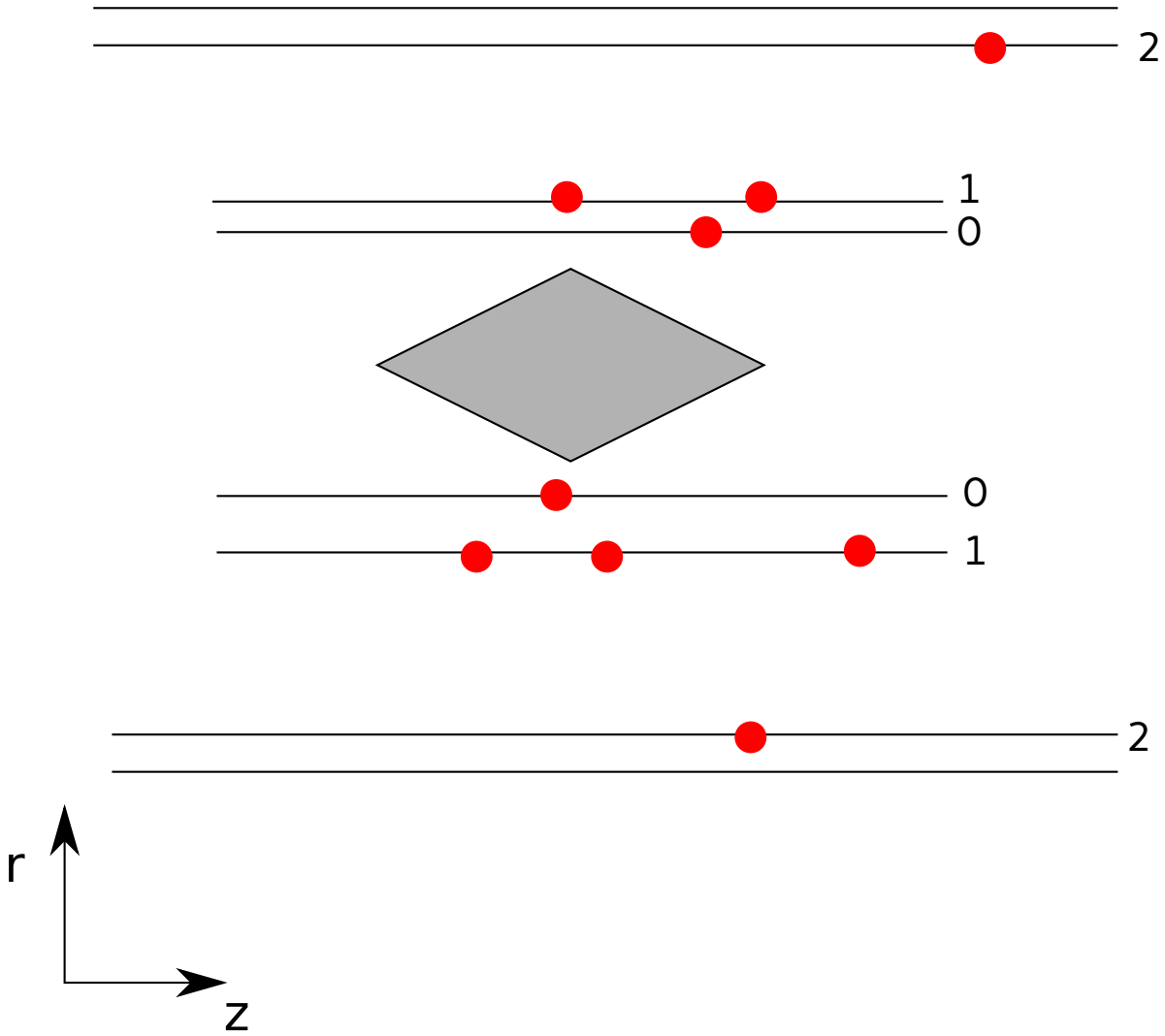
How do we find the three signal tracks?

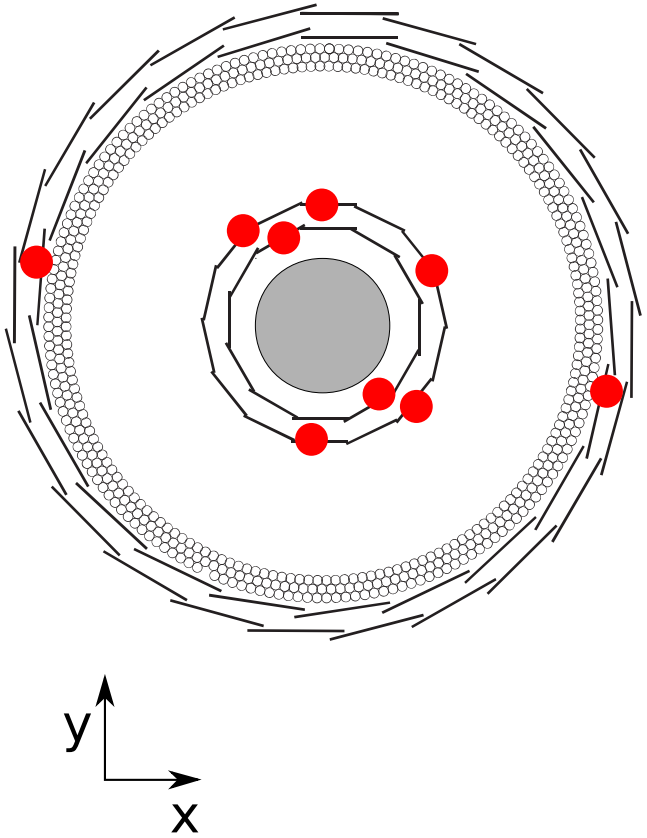
1) Track fitting

2) Vertex search

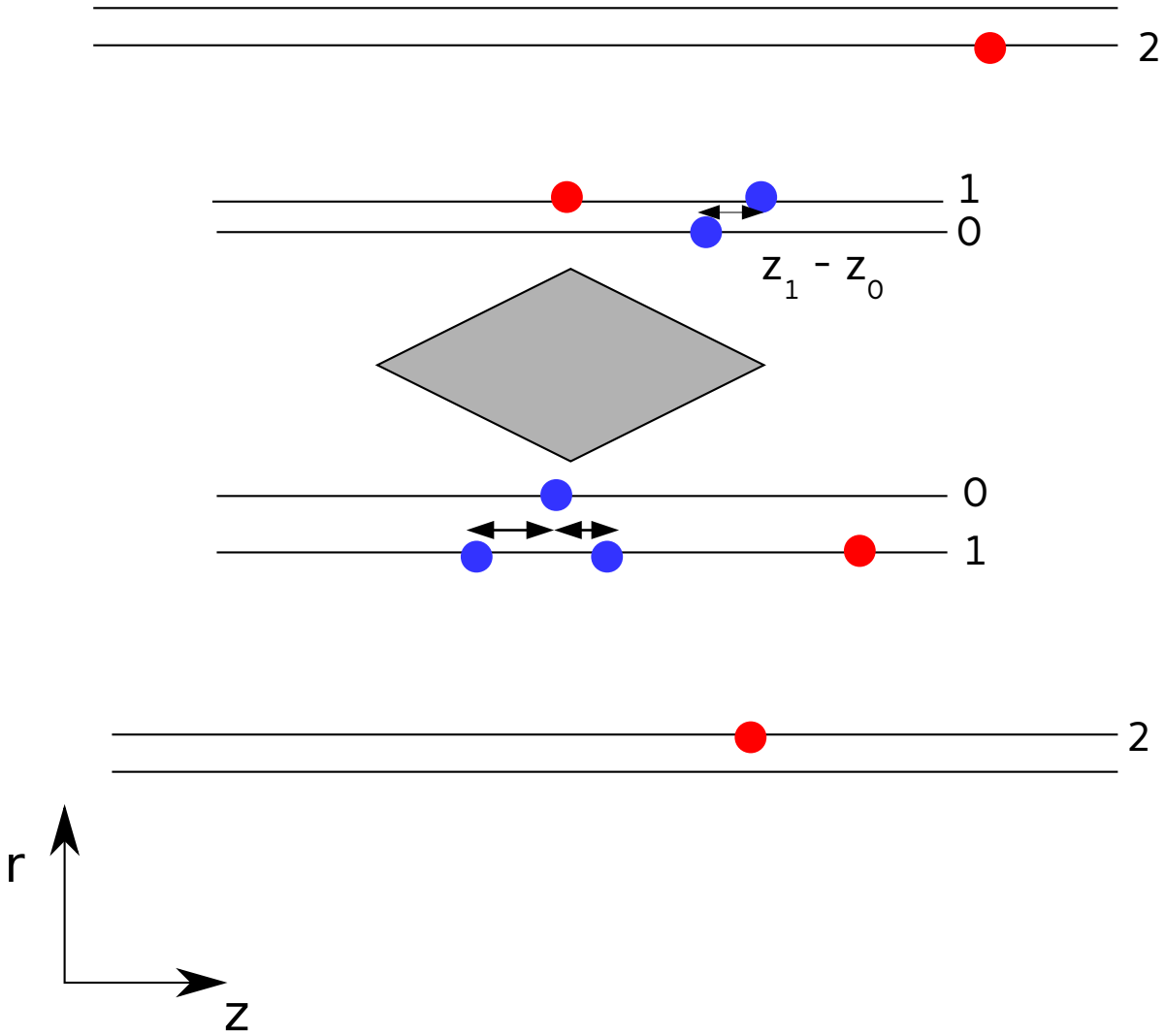


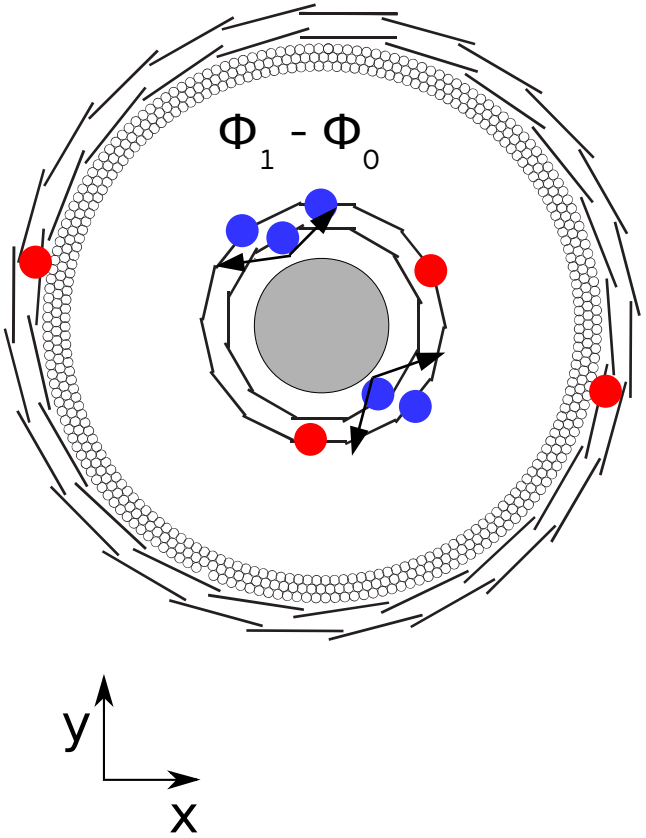
Geometrical Selection



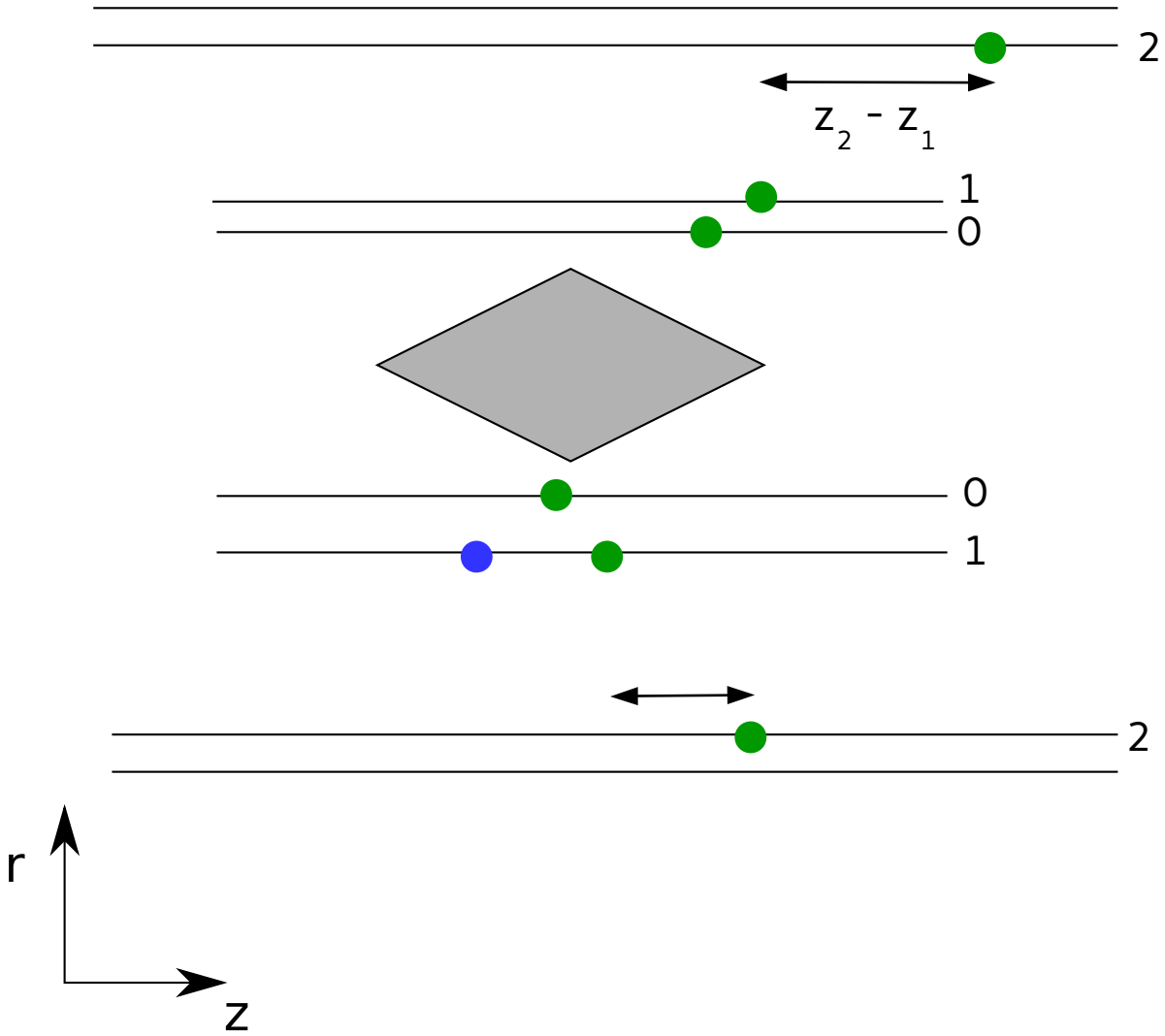


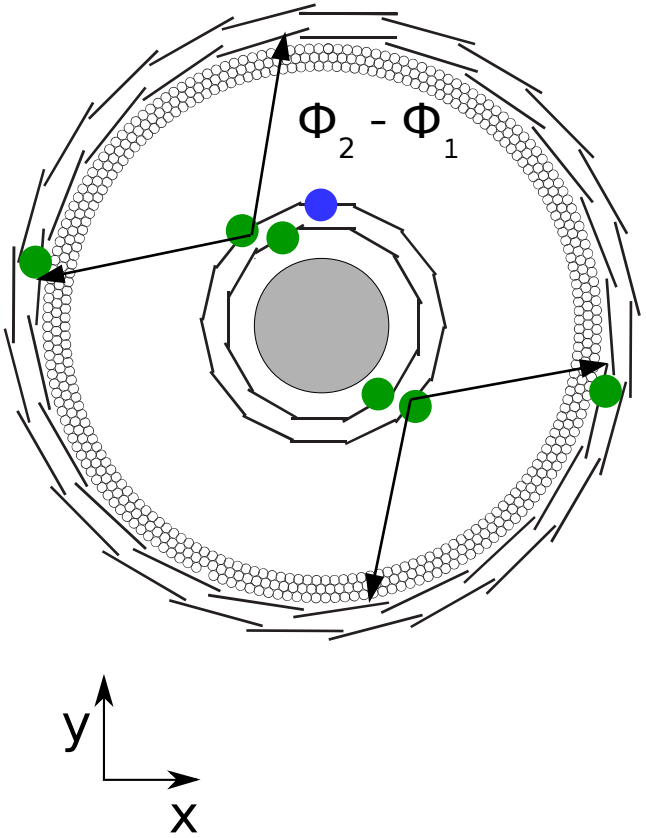
Geometrical Selection



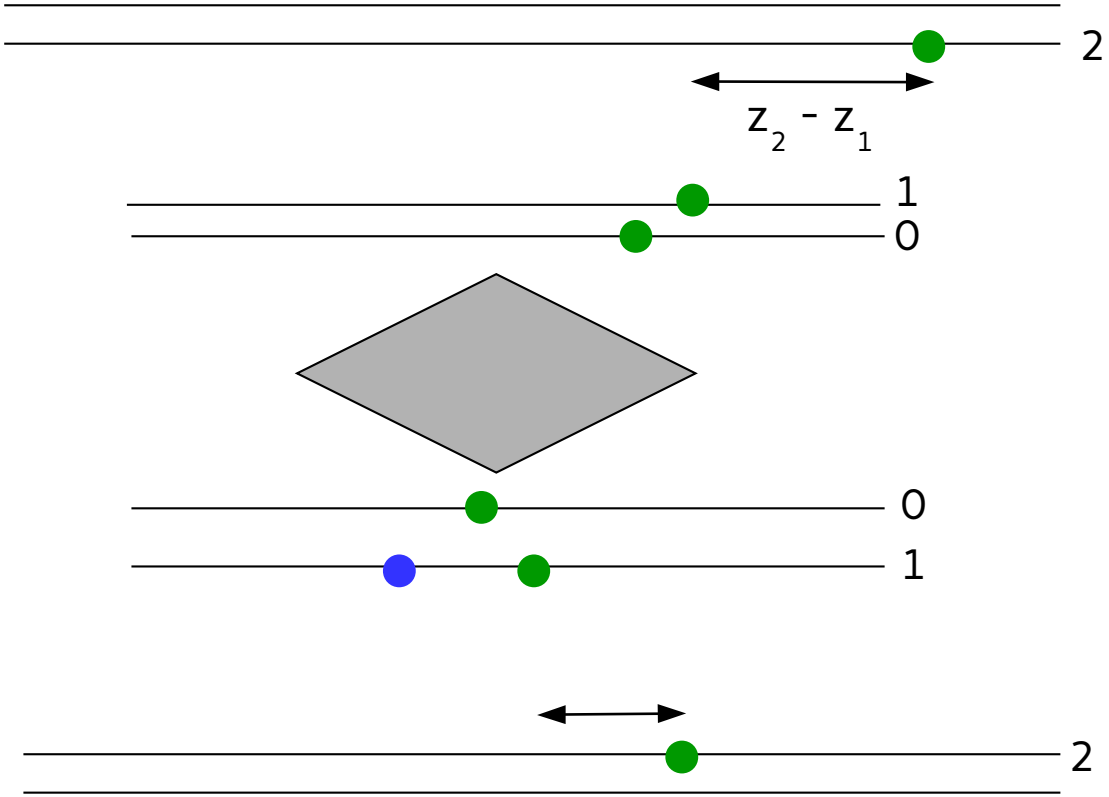


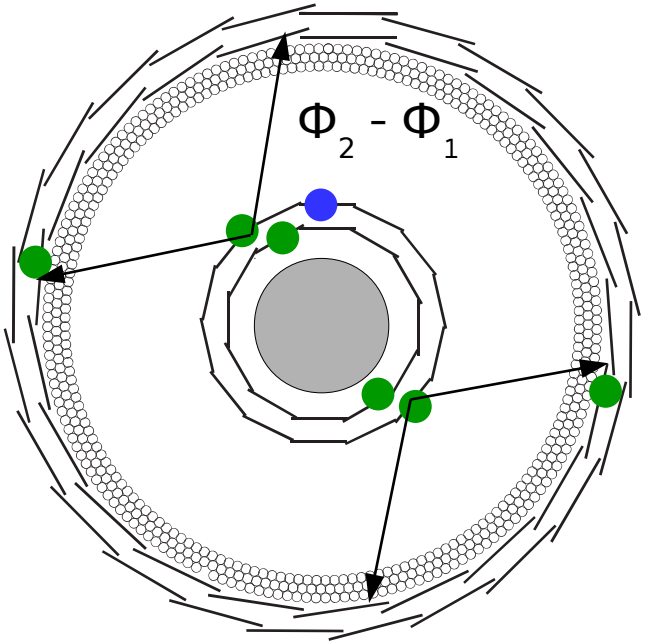
Geometrical Selection

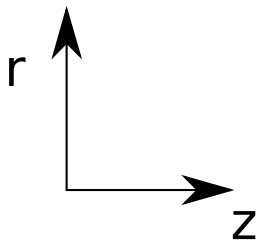


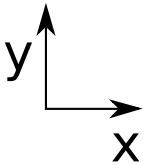


Geometrical Selection









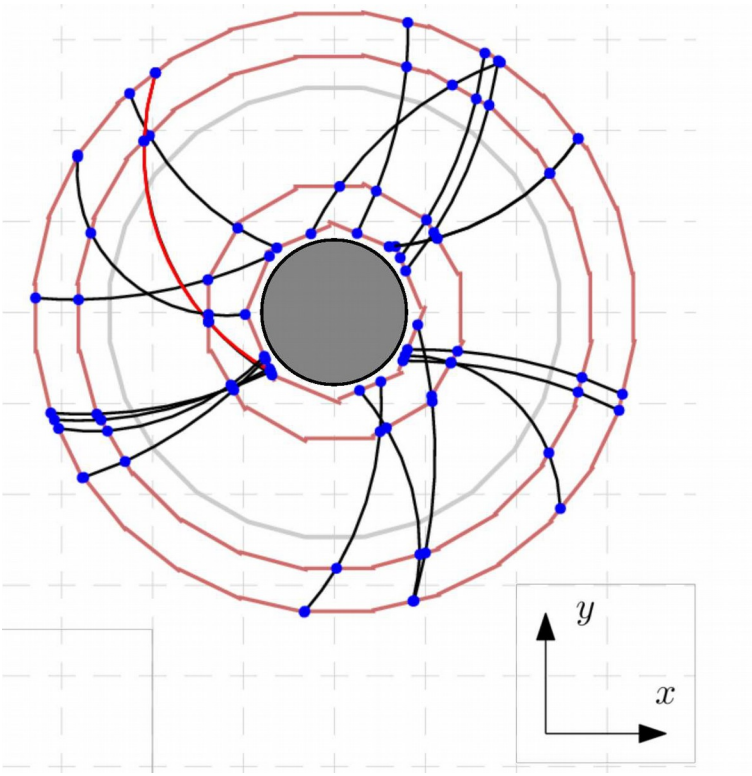
Reduce 3-hit combinations by factor 50

Fitting

- Use Multiple Scattering Fit (→ talk by A. Kozlinskiy)
- Fit hits in first three layers
- Propagate to 4th layer
- Select hit in 4th layer closest to propagated position
- Redo fit with a second triplet, cut on χ^2

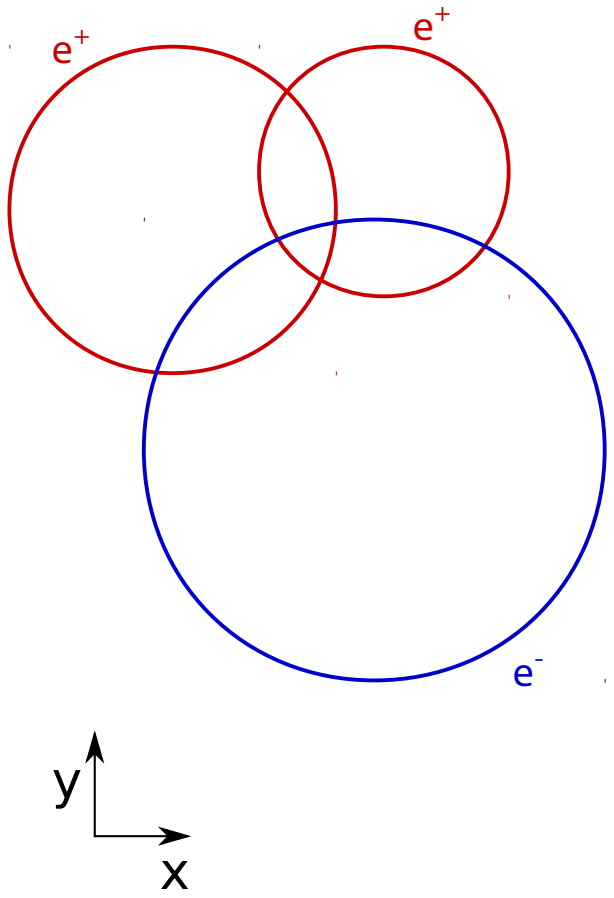
After all selections:

- 98.5 % of true 4-hit MC tracks selected
- 74 % of 4-hit tracks are true MC tracks



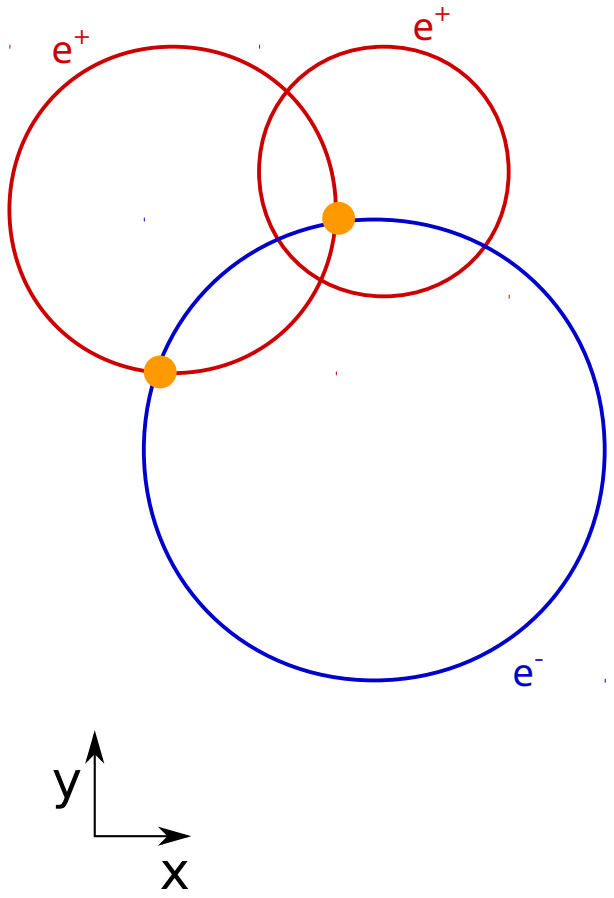
Vertex Estimate: XY-Plane

- Study each combination of two e^+, one e^-
- In xy-plane: find intersections of track circles
- Calculate weights of intersections based on uncertainties due to
  - multiple scattering
  - pixel size



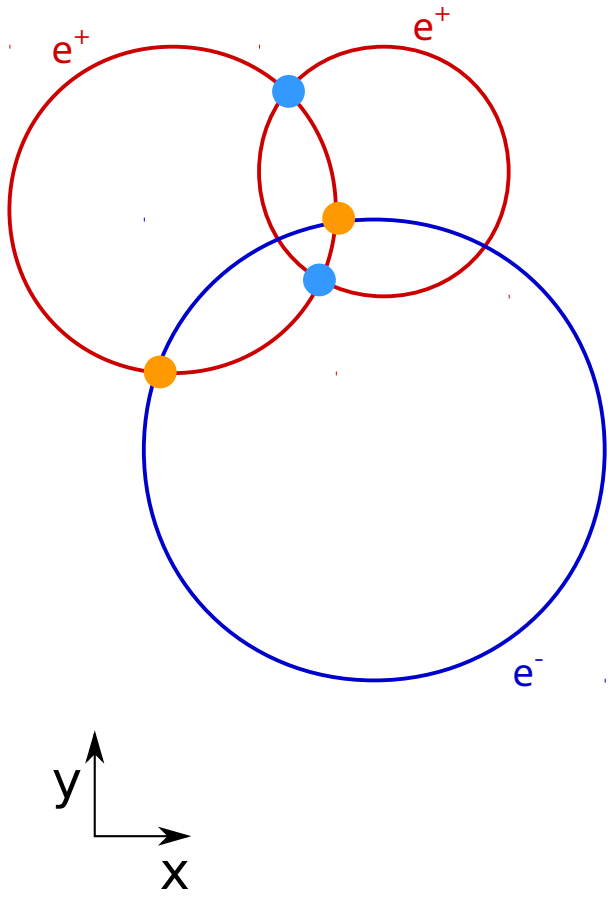
Vertex Estimate: XY-Plane

- Study each combination of two e^+, one e^-
- In xy-plane: find intersections of track circles
- Calculate weights of intersections based on uncertainties due to
  - multiple scattering
  - pixel size



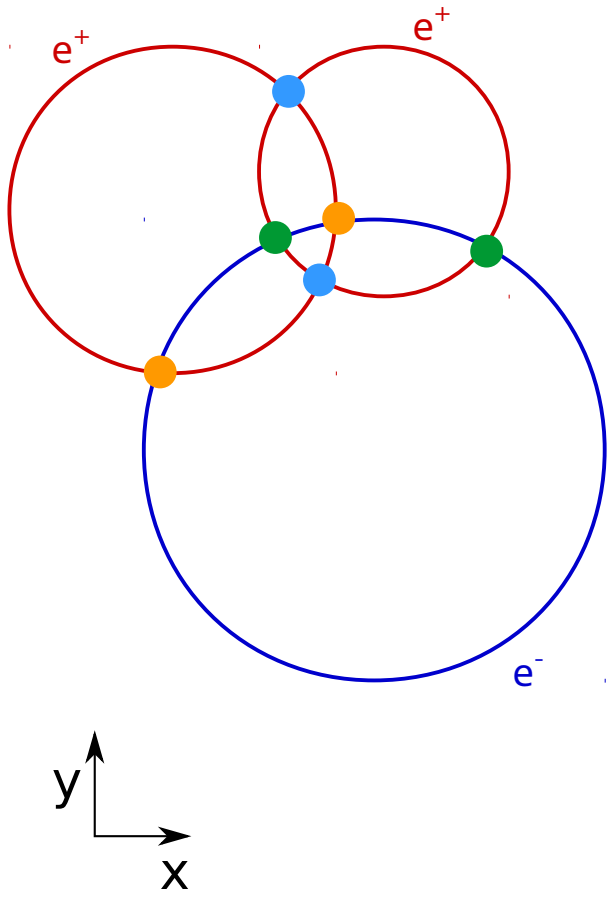
Vertex Estimate: XY-Plane

- Study each combination of two e^+, one e^-
- In xy-plane: find intersections of track circles
- Calculate weights of intersections based on uncertainties due to
  - multiple scattering
  - pixel size



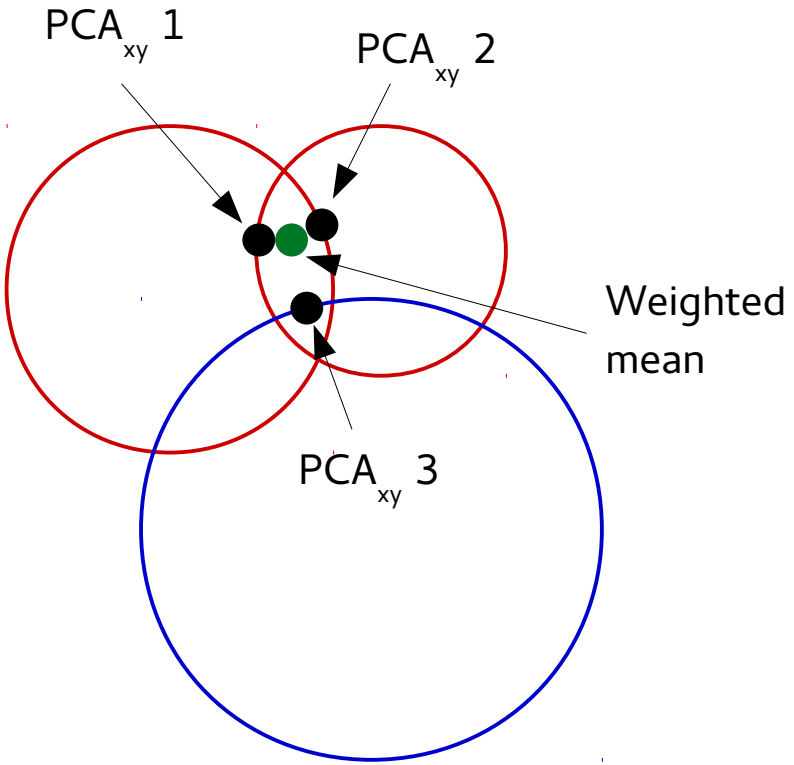
Vertex Estimate: XY-Plane

- Study each combination of two e^+, one e^-
- In xy-plane: find intersections of track circles
- Calculate weights of intersections based on uncertainties due to
  - multiple scattering
  - pixel size

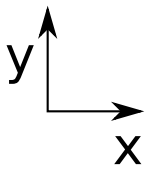


Vertex Estimate

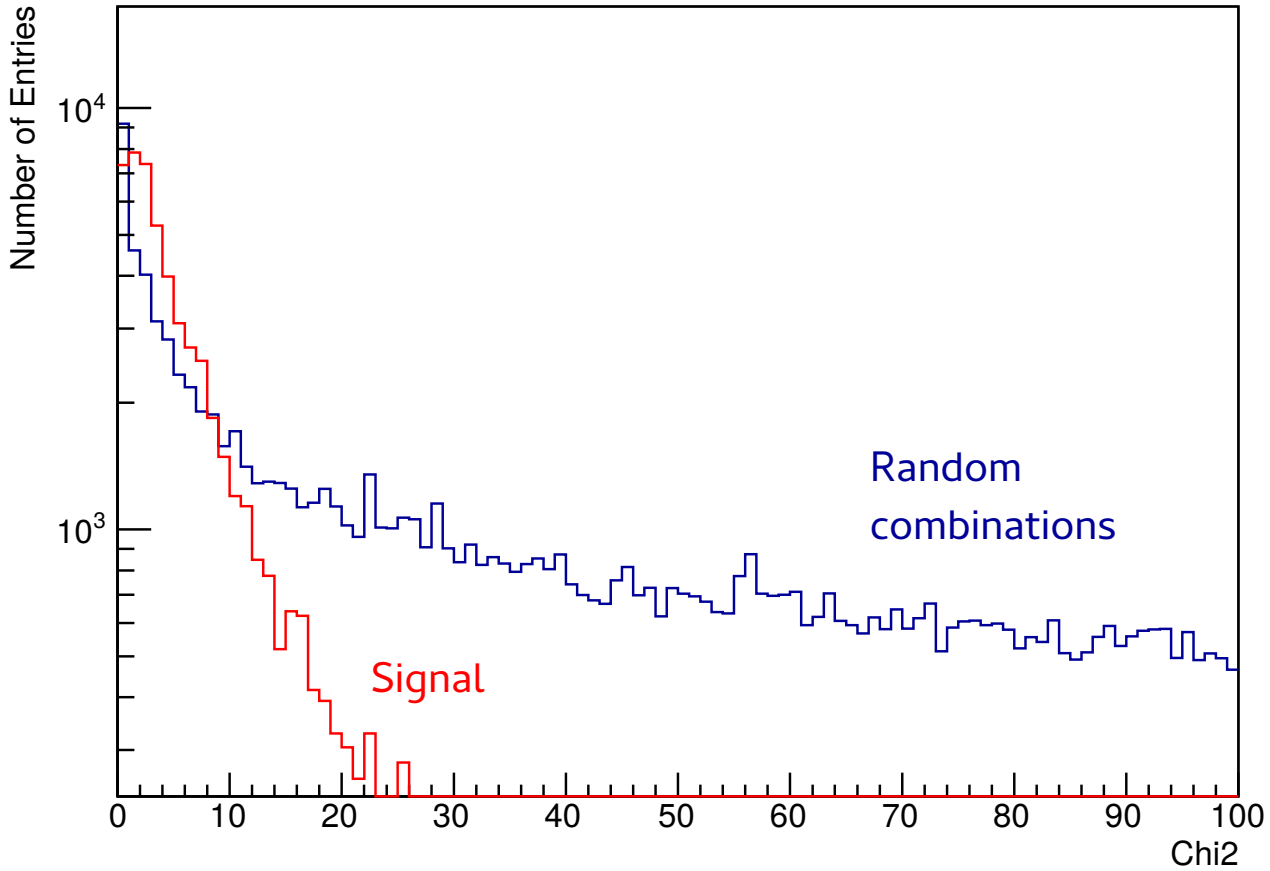
- Calculate weighted mean of intersections from three different tracks
- Find point of closest approach (PCA_{xy}) to weighted mean in xy-plane on each track
- Calculate z-position PCA_z and weight at PCA_{xy}
- Find weighted mean in z-coordinate $\bar{z}$
- Achieve vertex resolution of ~400 µm sigma



$$\chi^2 = \sum_{i=0}^{3} \frac{PCA_{xy,i} - \bar{xy}}{\sigma_{PCA_{xy,i}}} + \frac{PCA_{z,i} - \bar{z}}{\sigma_{PCA_{z,i}}}$$

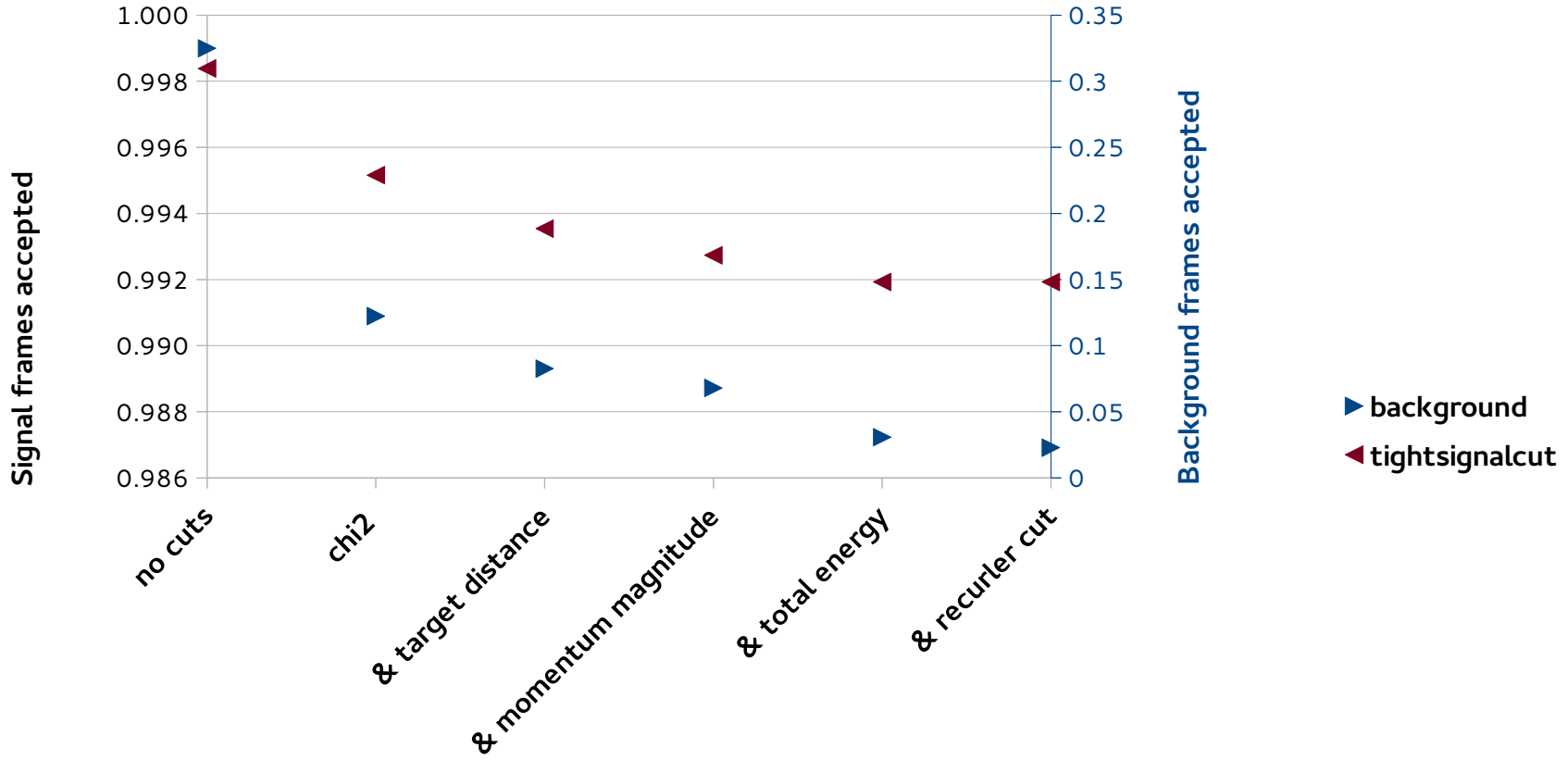


χ^2 Distribution



Cut Effects

Signal reference: full offline track reconstruction and offline vertex fit



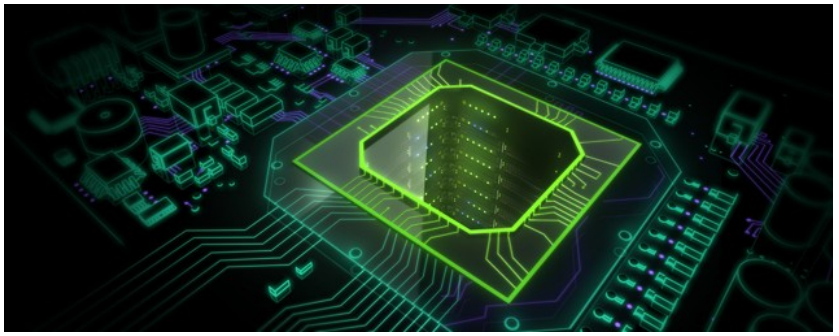
Fast Reconstruction on GPU

- Use time slices of 50 ns for track & vertex search

  → Process $20 \cdot 10^6$ time slices per second

- Plan for 12 filter farm PCs with one GPU each

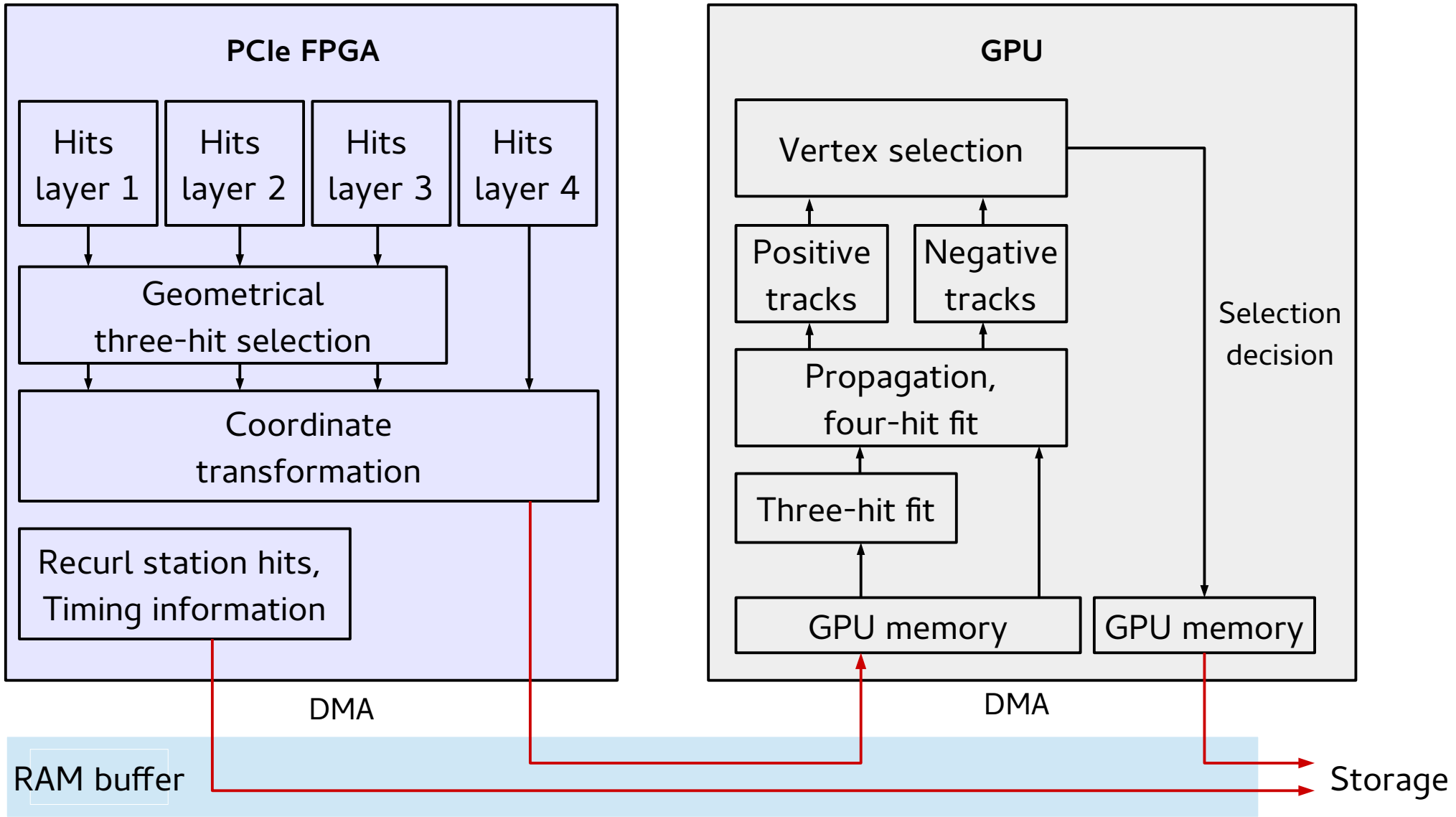
  → Process at least $1.7 \cdot 10^6$ time slices per second



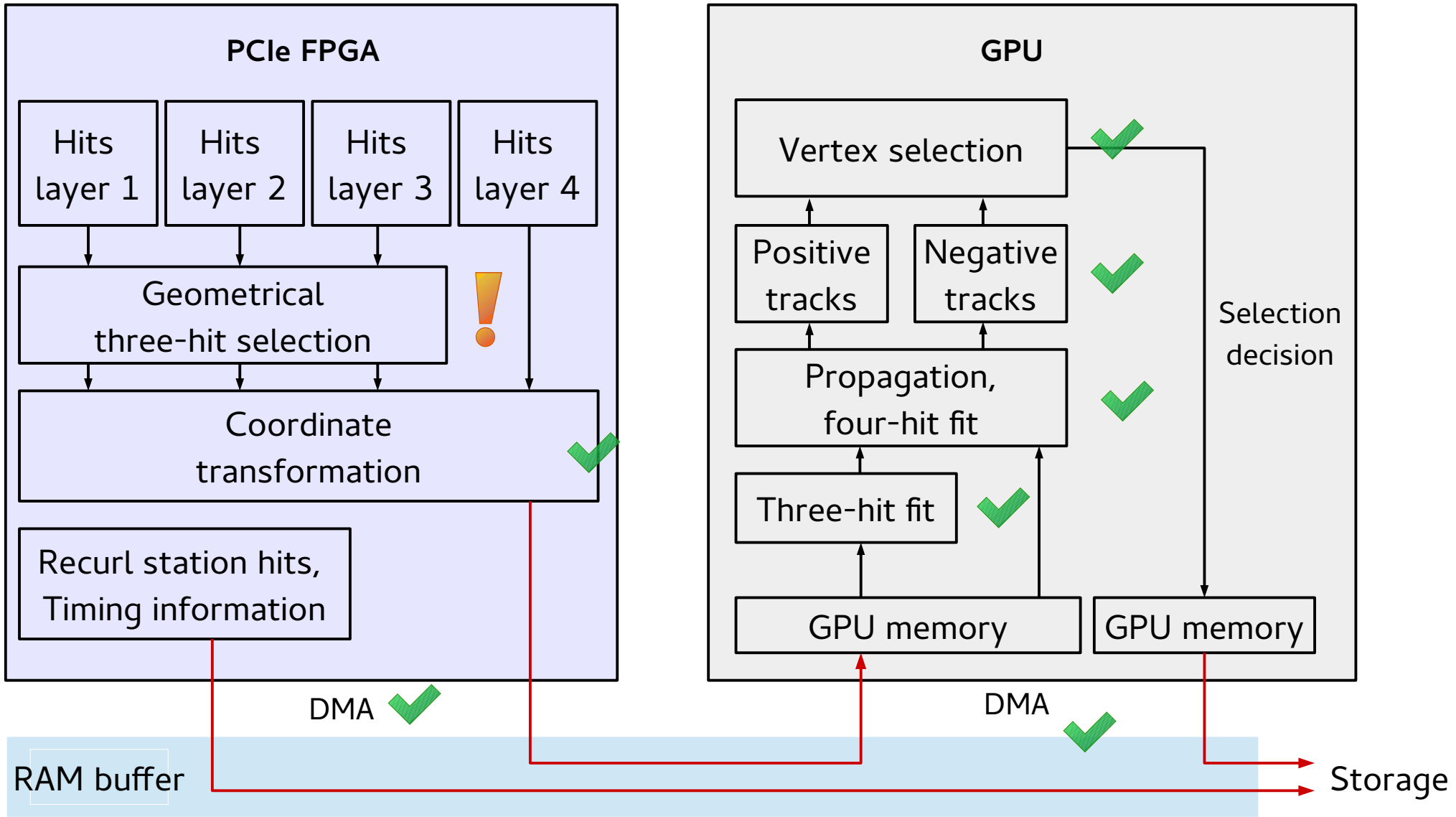
- Thousands of cores
- Optimal parallel performance
- Best suited for many floating-point operations / second

→ use GPUs

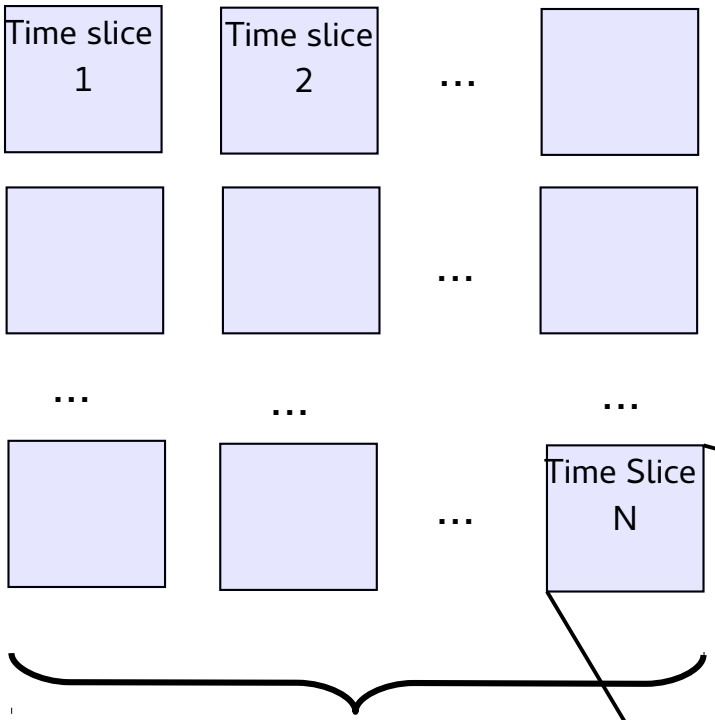
Selection on GPU



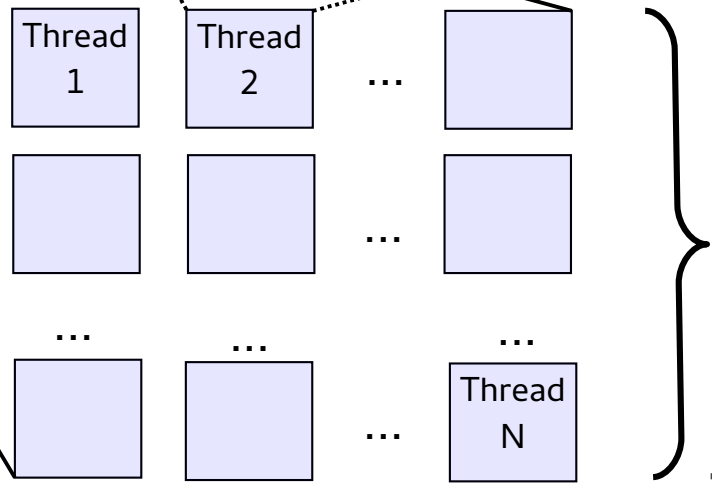
Selection on GPU



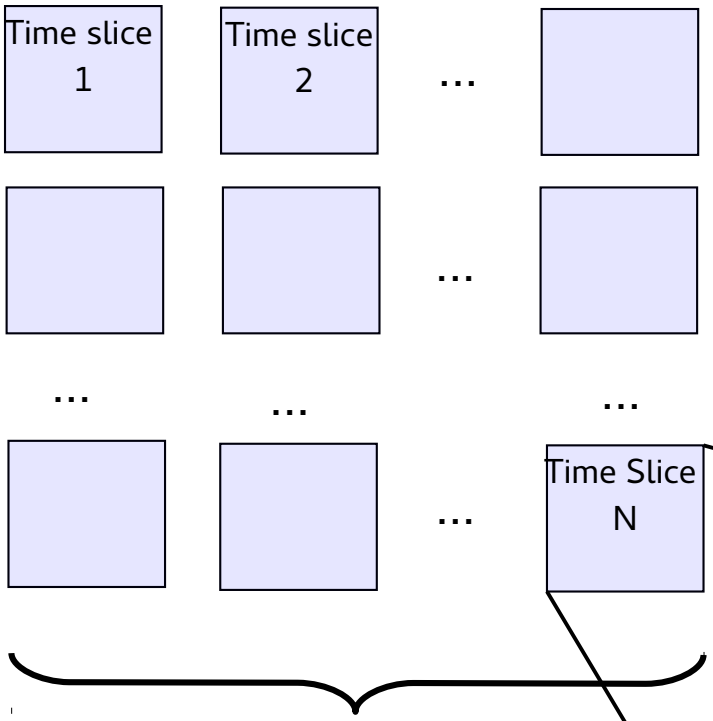
Parallelization Track Fit



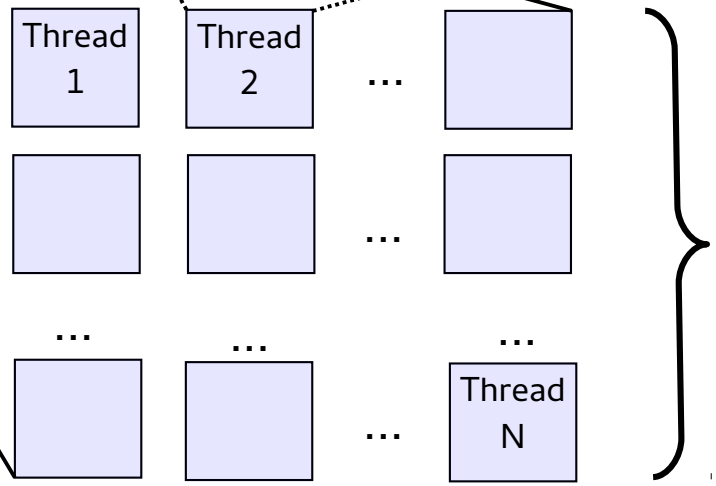
- Fit for one combination of three hits
- Propagation to 4th layer
- Loop over hits in 4th layer: check if hit exists in proximity of propagated track, re-fit
- Wait for all cores in one time slice to be done with previous steps



Parallelization Track Fit

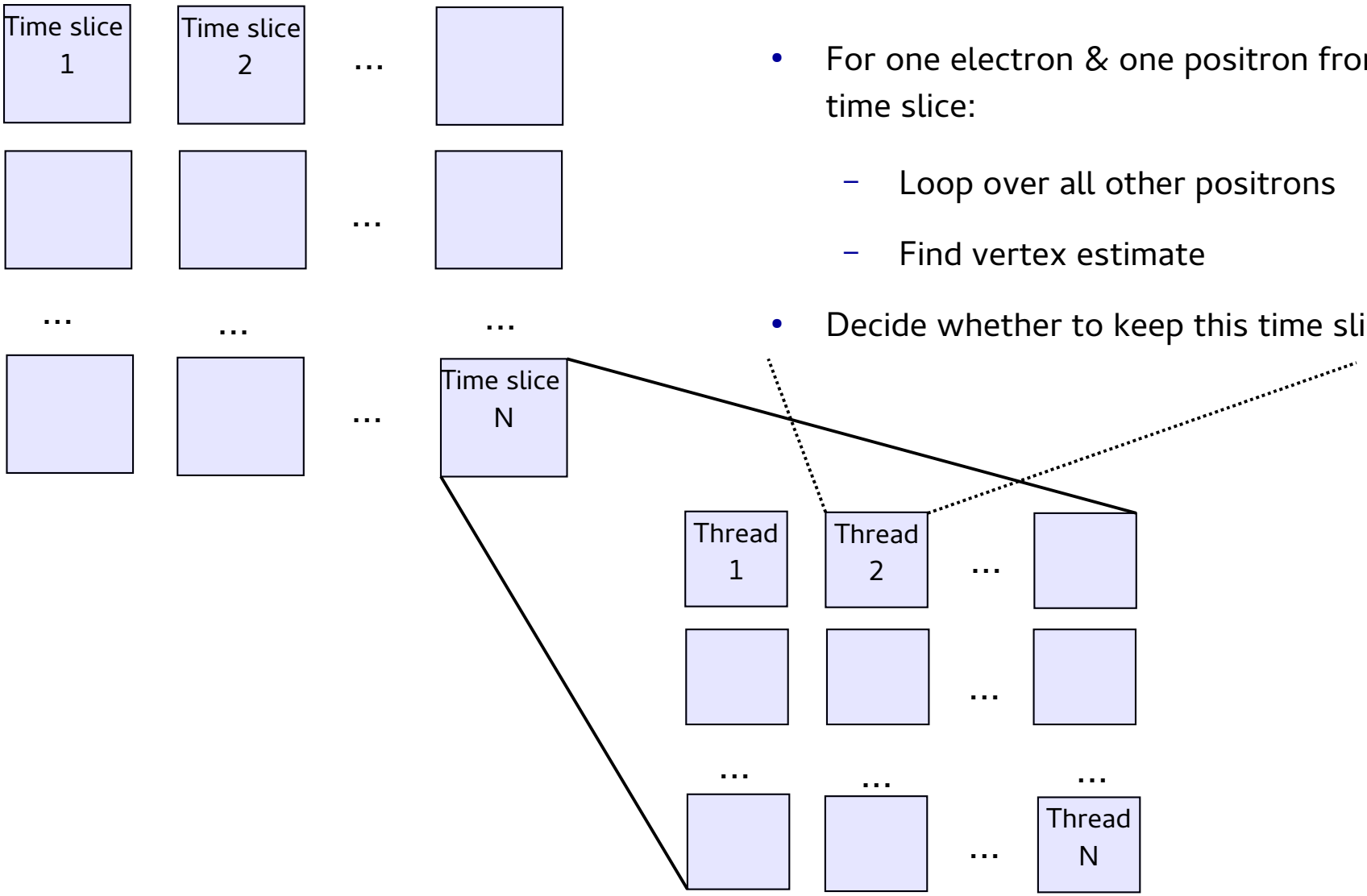


- Fit for one combination of three hits
- Propagation to 4th layer
- Loop over hits in 4th layer: check if hit exists in proximity of propagated track, re-fit
- Wait for all cores in one time slice to be done with previous steps



Total of 12.6 million threads to be distributed among 2560 cores

Parallelization Vertex Selection



- For one electron & one positron from this 50 ns time slice:
  - Loop over all other positrons
  - Find vertex estimate
- Decide whether to keep this time slice

Performance

Optimizations performed:

- Memory layout and access pattern
- Register usage
- Grid dimensions
- Approximations

Performance

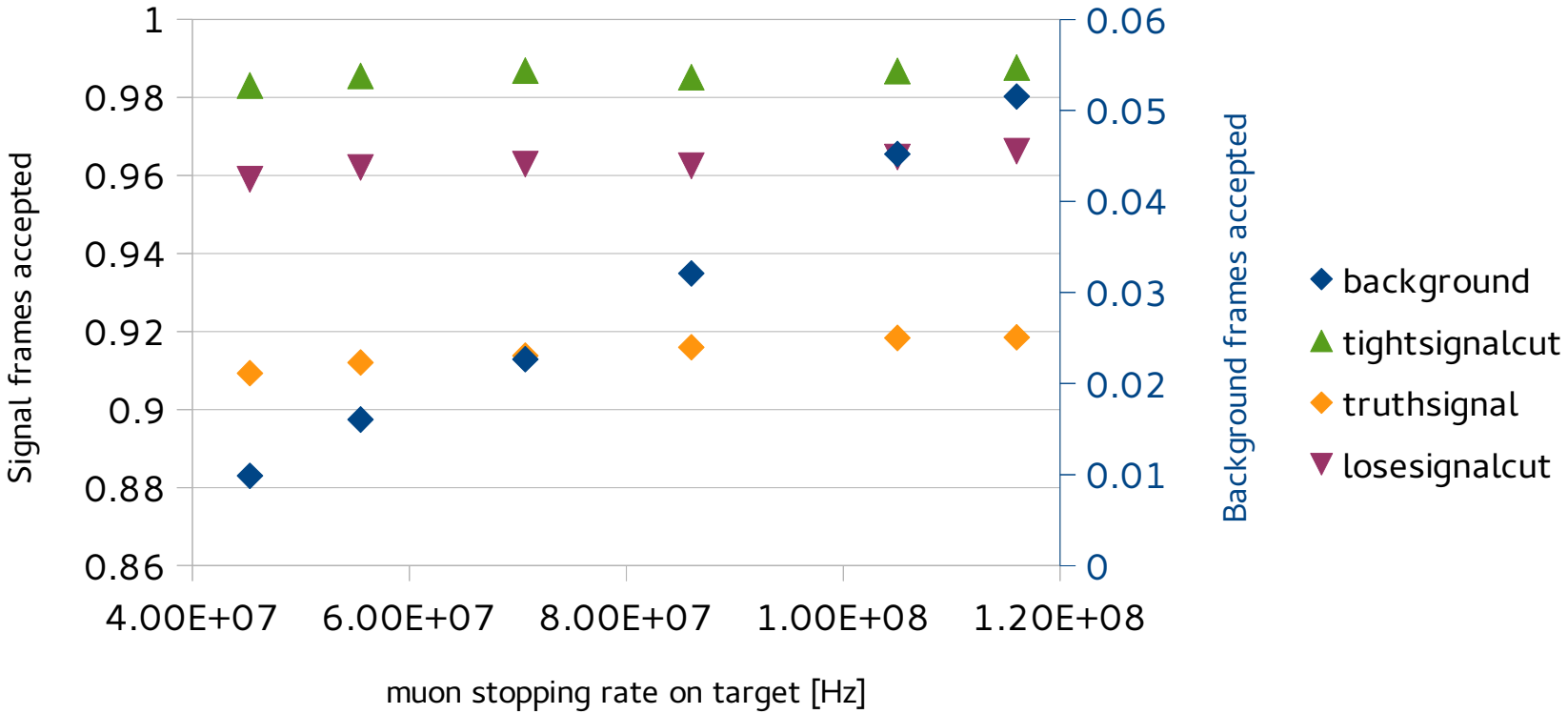
Optimizations performed:

- Memory layout and access pattern
- Register usage
- Grid dimensions
- Approximations

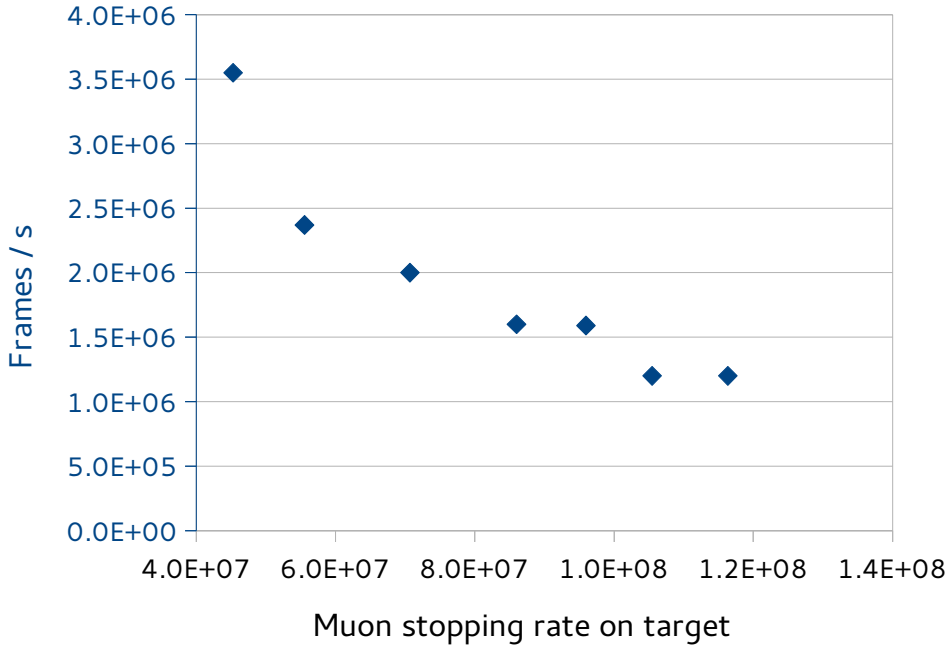
Currently process $2 \cdot 10^6$ time slices / s on one nvidia GTX 1080 at a muon stopping rate of $7 \cdot 10^7$ Hz

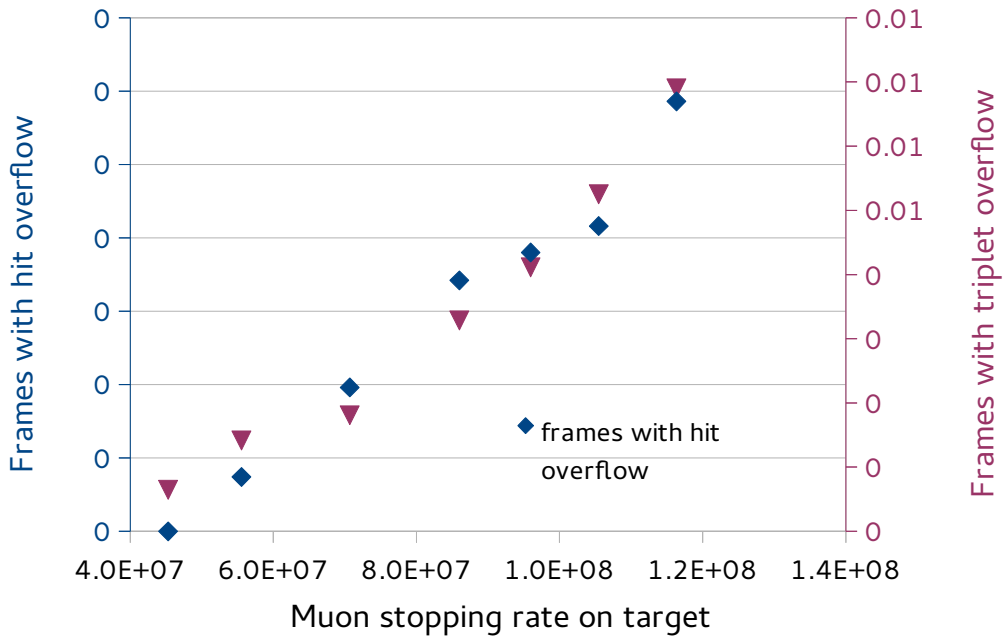
Backup

Muon Stopping Rate Study I



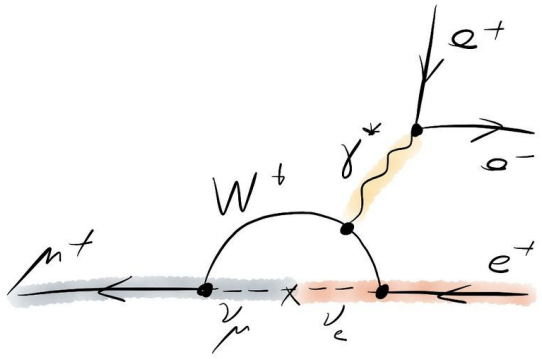
Muon Stopping Rate Study II



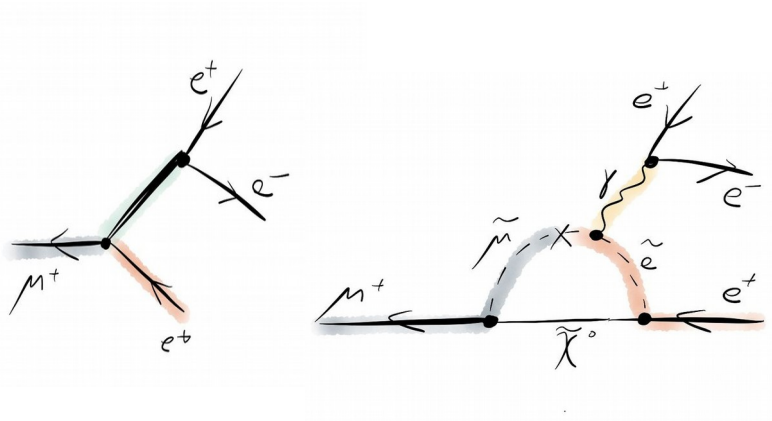


The Mu3e Experiment

Search for charged lepton flavour-violating decay $\mu^+ \rightarrow e^+e^-e^+$ with a sensitivity in branching ratio better than 10^{-16}



Branching ratio suppressed in Standard Model to below 10^{-54}

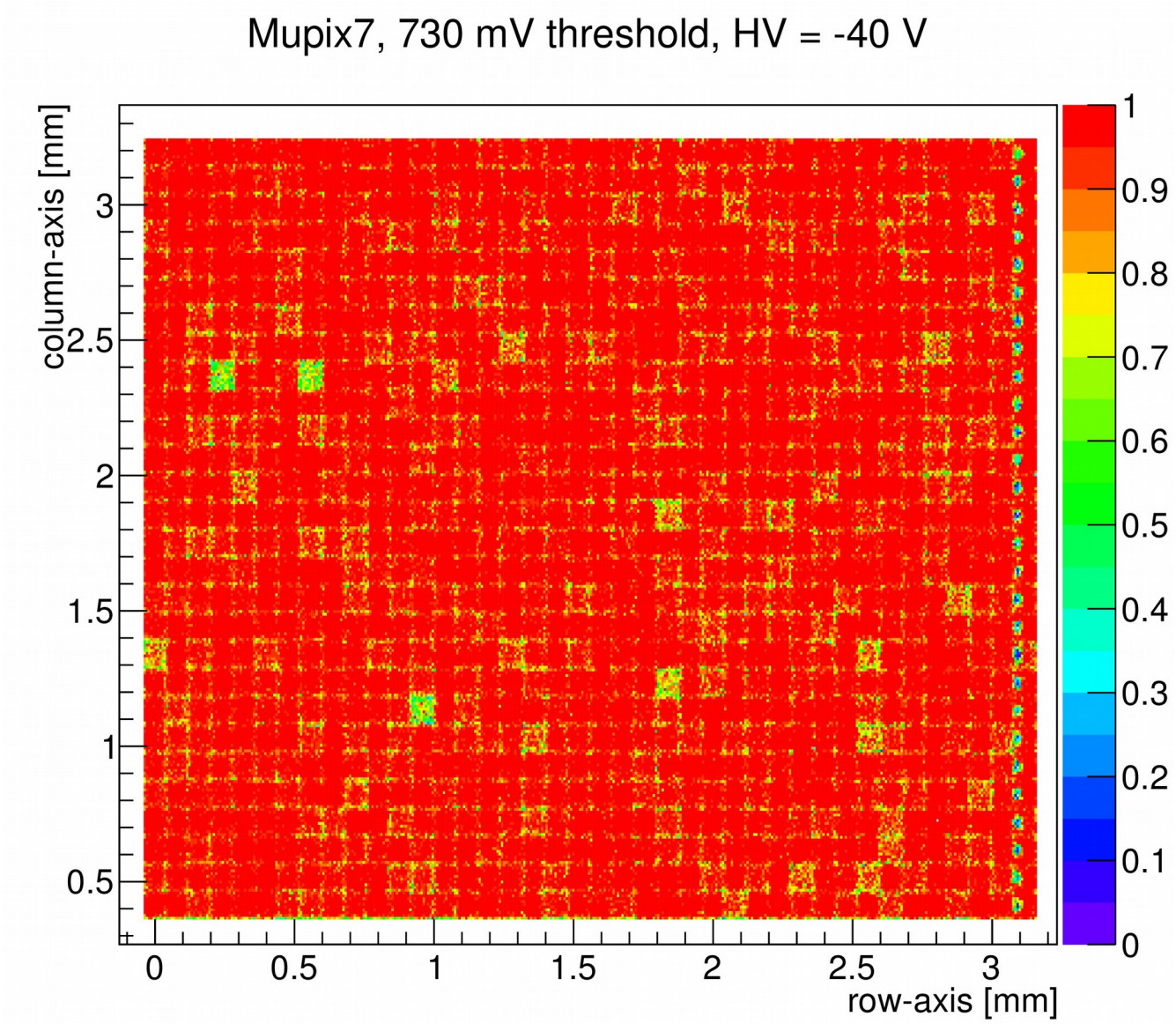


Any hint of signal ⟶ new physics

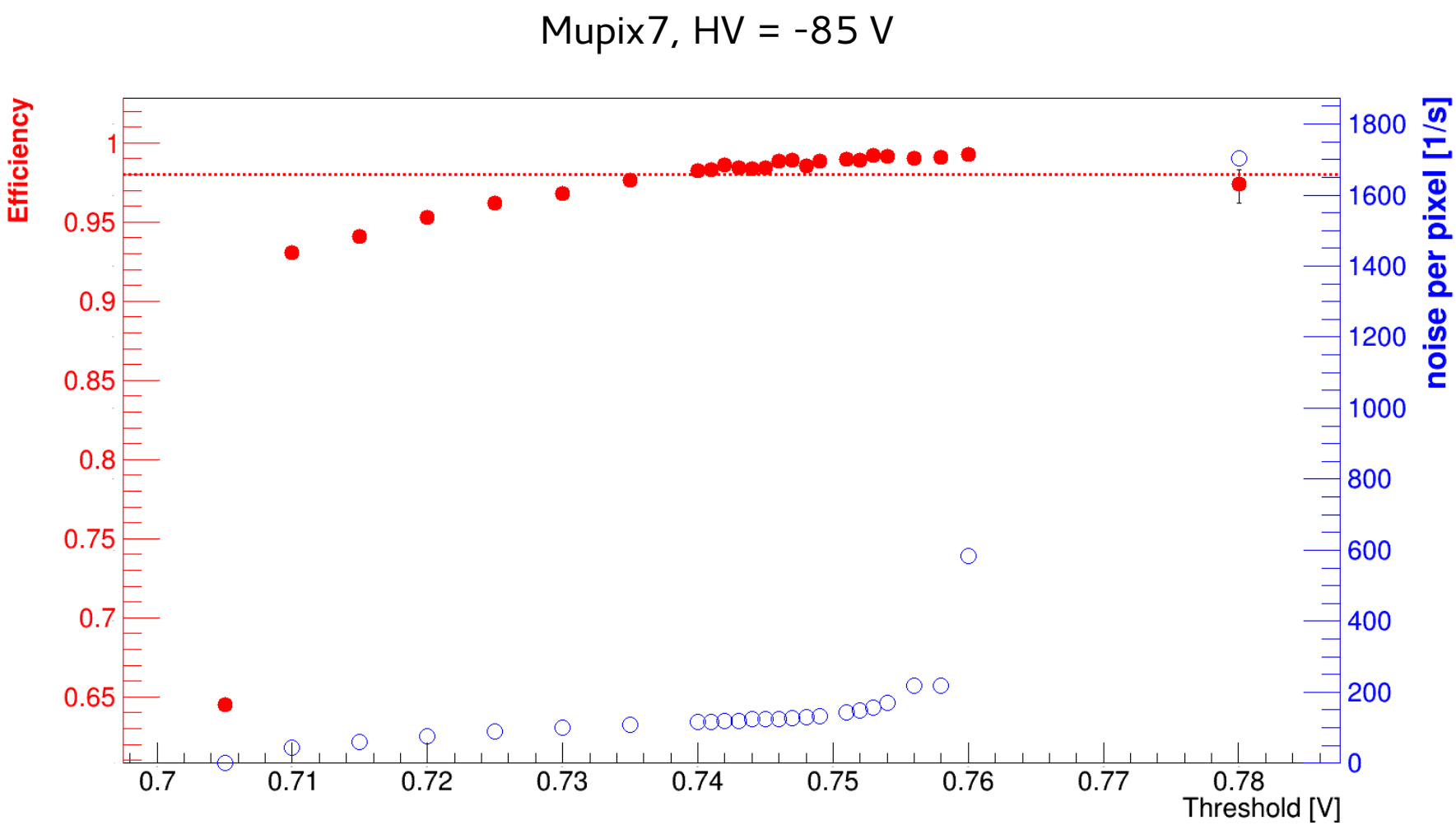
- Supersymmetry
- Grand unified models
- Extended Higgs sector
- ...

Current limit on branching ratio: 10^{-12} (SINDRUM, 1988)

Mupix7: Efficiency



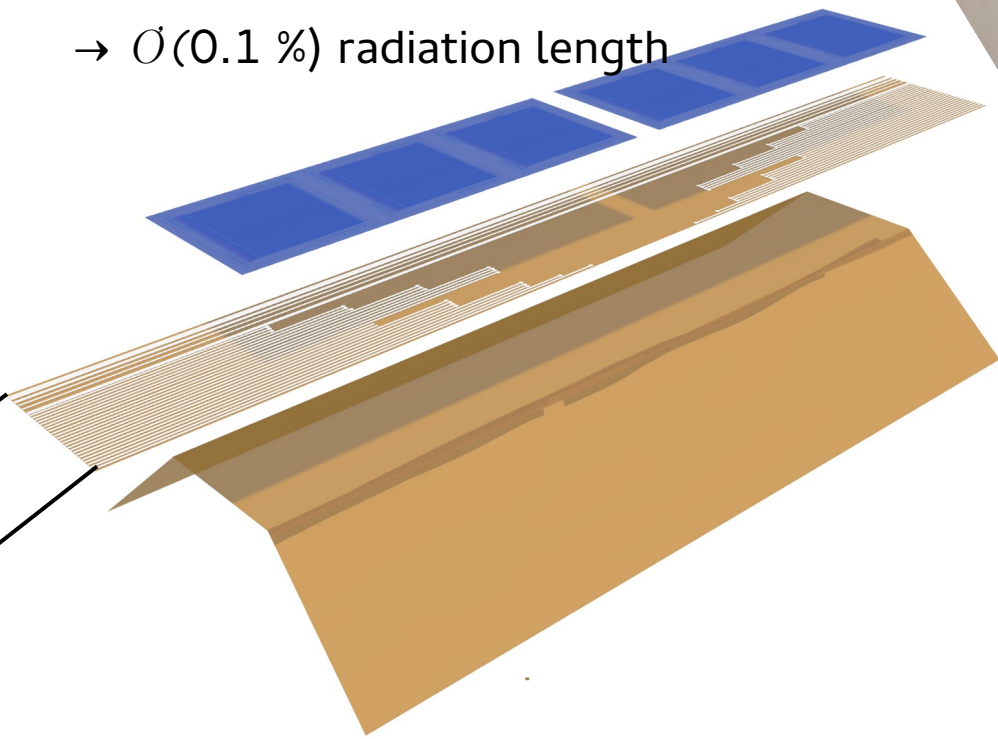
Mupix7: Efficiency



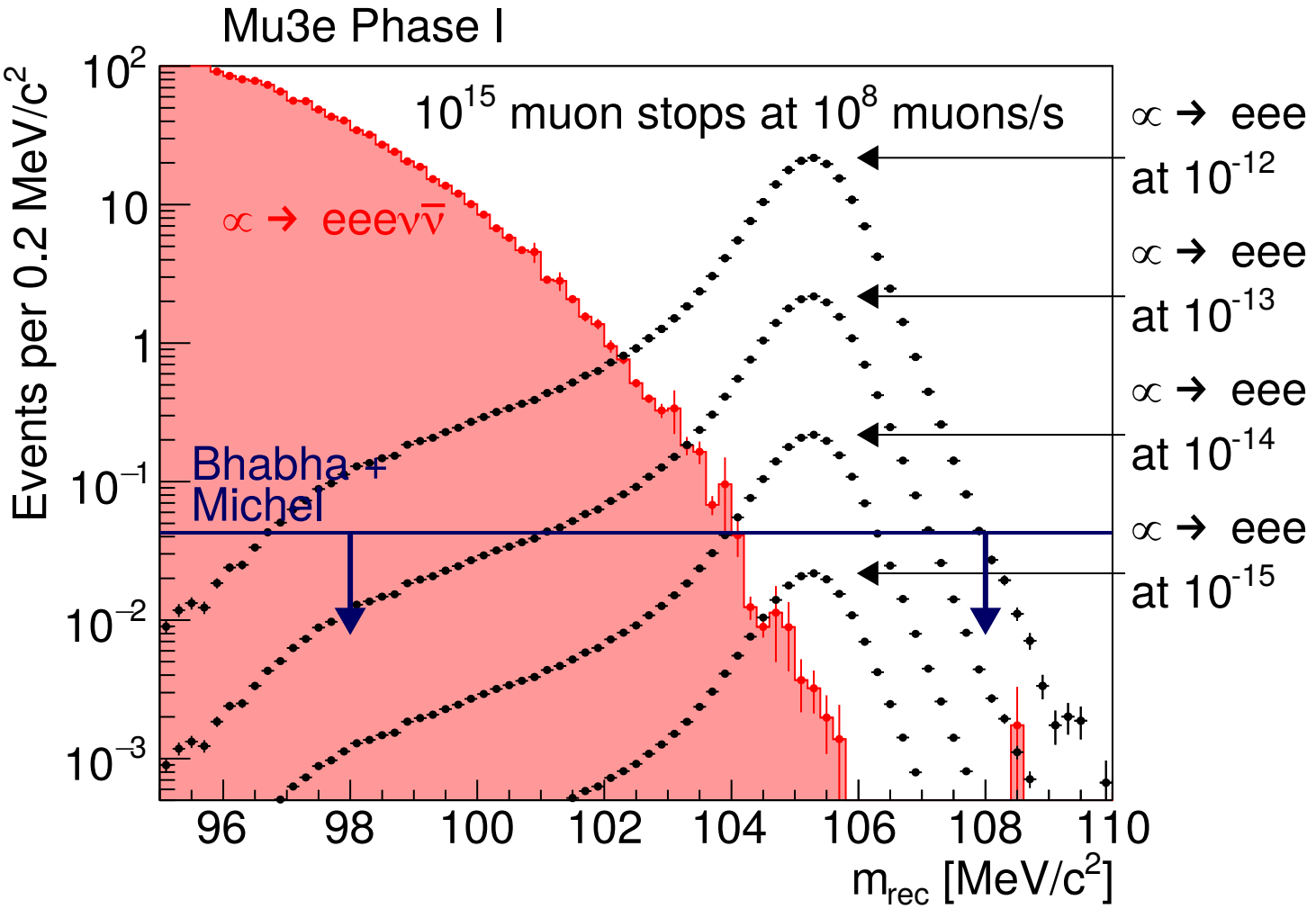
Mupix: Mechanics

- 50 µm silicon
- ~ 50 µm flexprint: Kapton, aluminum, copper
- 25 µm Kapton foil

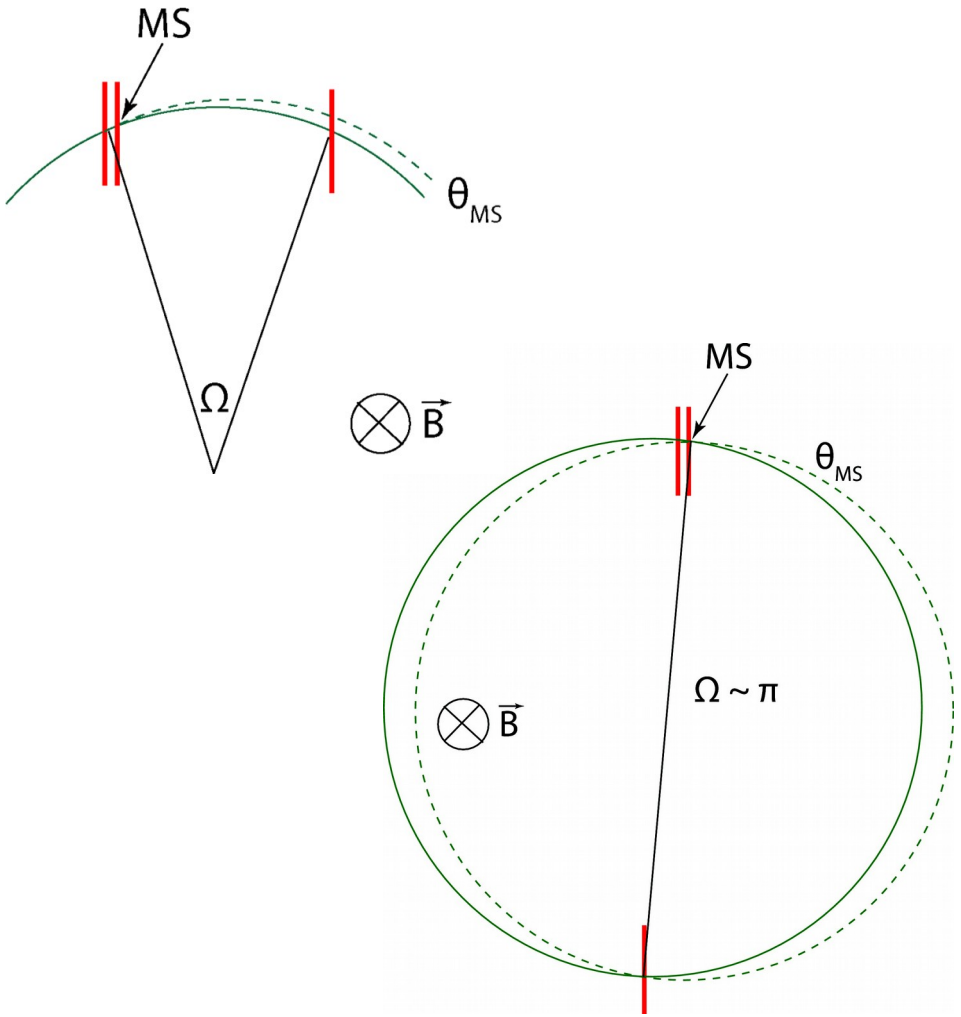
→ $\mathcal{O}$(0.1 %) radiation length



Sensitivity Study



Multiple Scattering



- Muons decay at rest

→ momentum < 53 MeV/c

- Momentum resolution to first order:

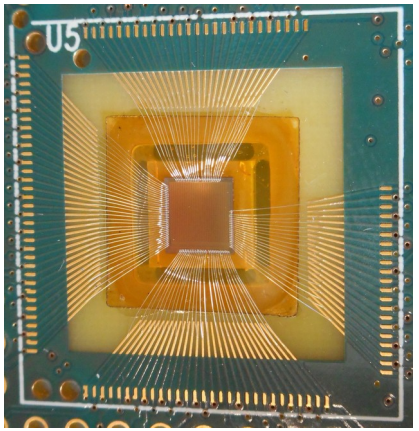
$$\sigma_p/p \sim \theta_{MS}/\Omega$$

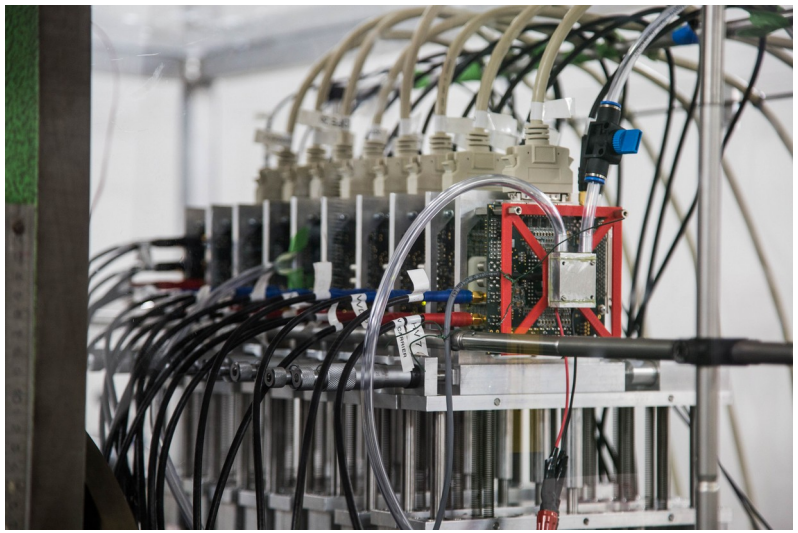
- Use recurling tracks for momentum measurement

→ **Minimize material budget**

Mupix Protoype

- Mupix7: latest prototype
  - Thinned to 50 µm
  - 32 x 40 pixel matrix
- Pixel size: 103 µm x 80 µm
  - 3.2 x 3.2 mm^2

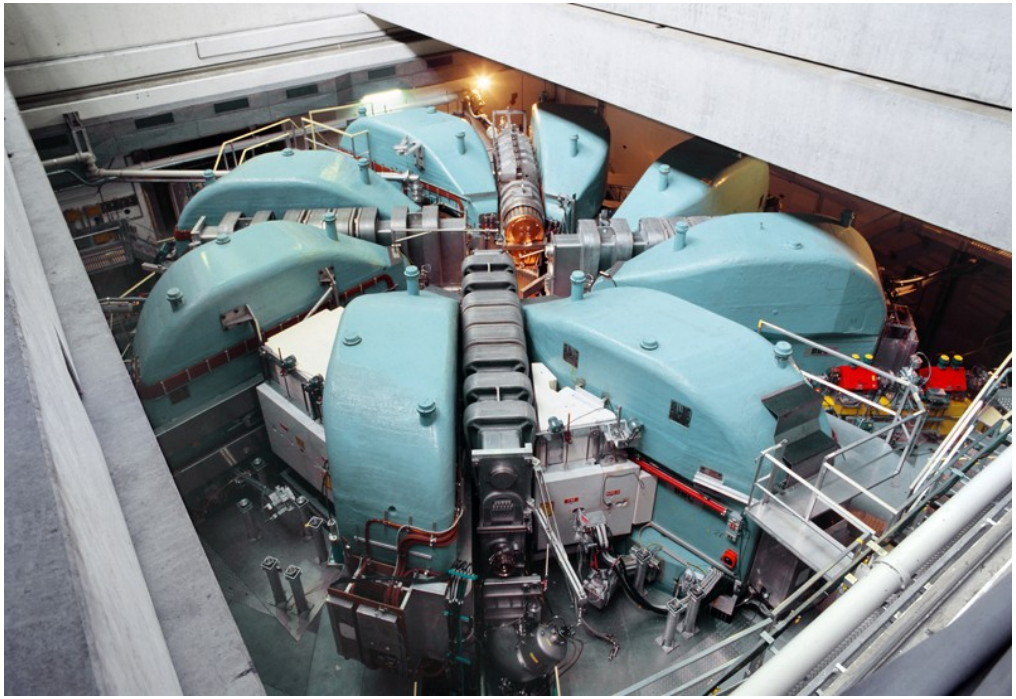




- Readout electronics on chip
- Fast LVDS link: 1.25 Gbit/s

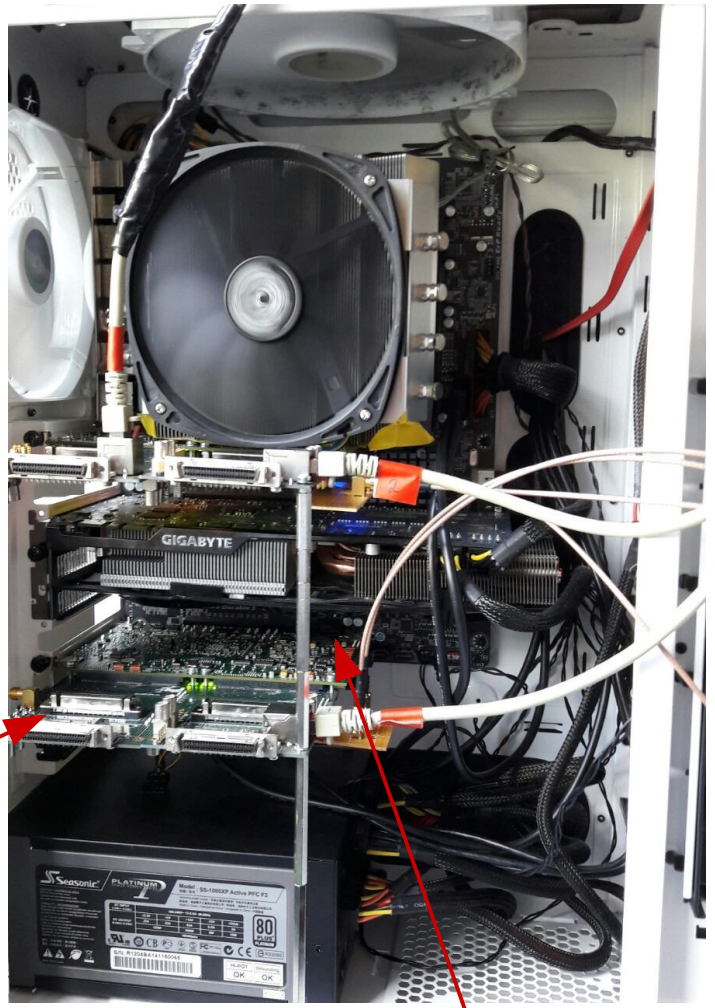
Muon Beam
@ Paul Scherrer Institute (PSI)

- 590 MeV cyclotron
- 2.2 mA proton beam
- Most powerful proton beam worldwide
- Target E: 28 MeV/c surface muons to πE5 beamline



Data Transfer

- Transfer data from FPGA to RAM via direct memory access (DMA)
- Tested at 1.5 GB/s: BER ≤ $4 \cdot 10^{-16}$ (at 95% confidence level)
  - Tested on beam test campaigns
- Will be used for readout of next MuPix prototype

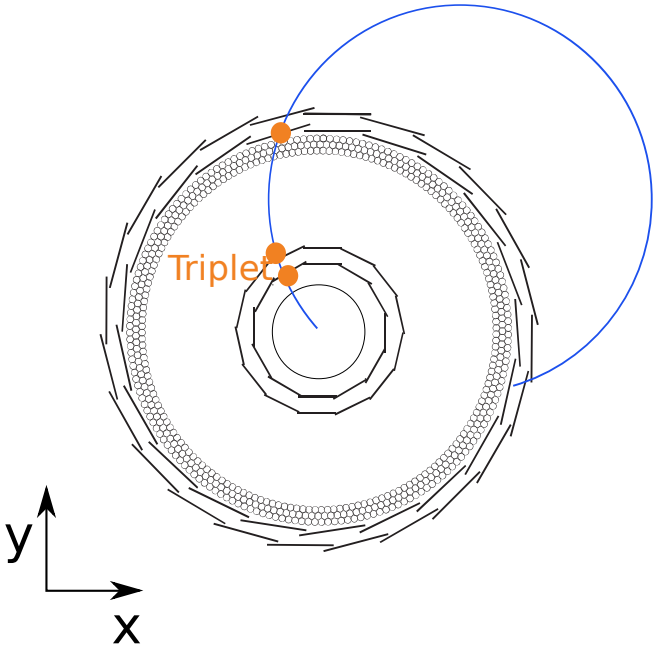


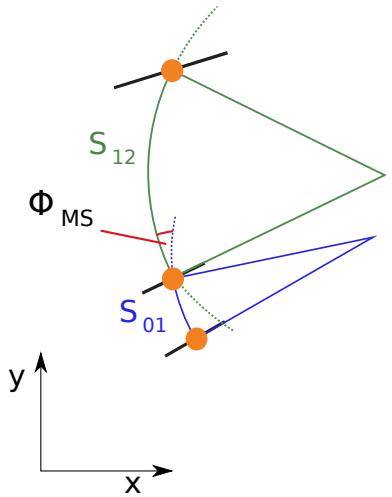
LVDS connector for data cable from MuPix chip

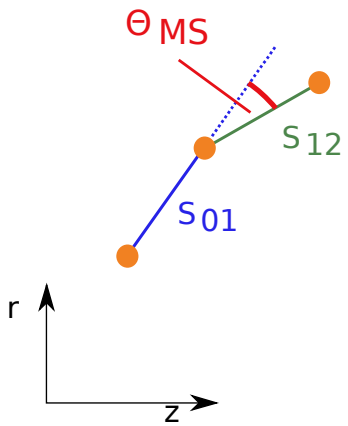
PCIe readout board

Multiple Scattering Fit

- Electrons: 12 – 53 MeV/c
- Resolution dominated by multiple Coulomb scattering
- Ignore hit uncertainty







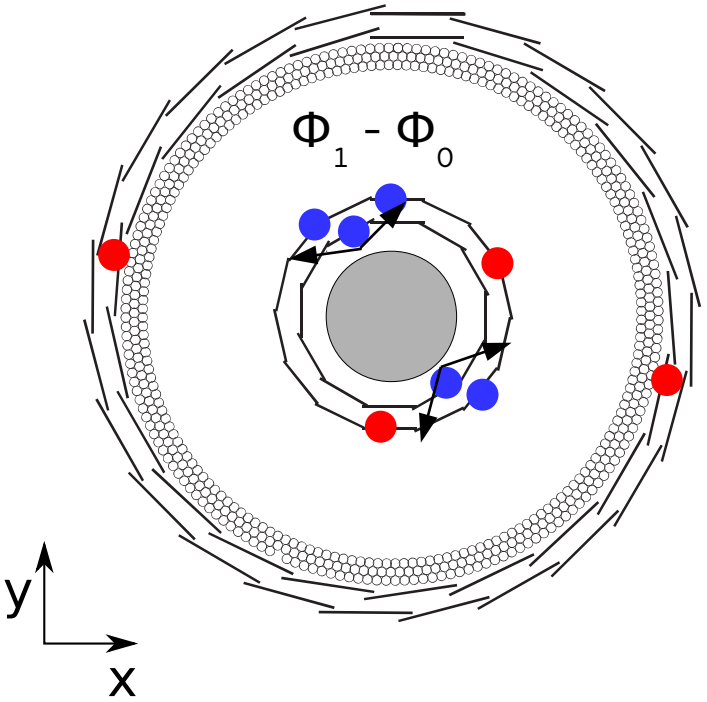
- Three consecutive hits: “triplet”
- Multiple scattering at middle hit of triplet
  - Minimize multiple scattering

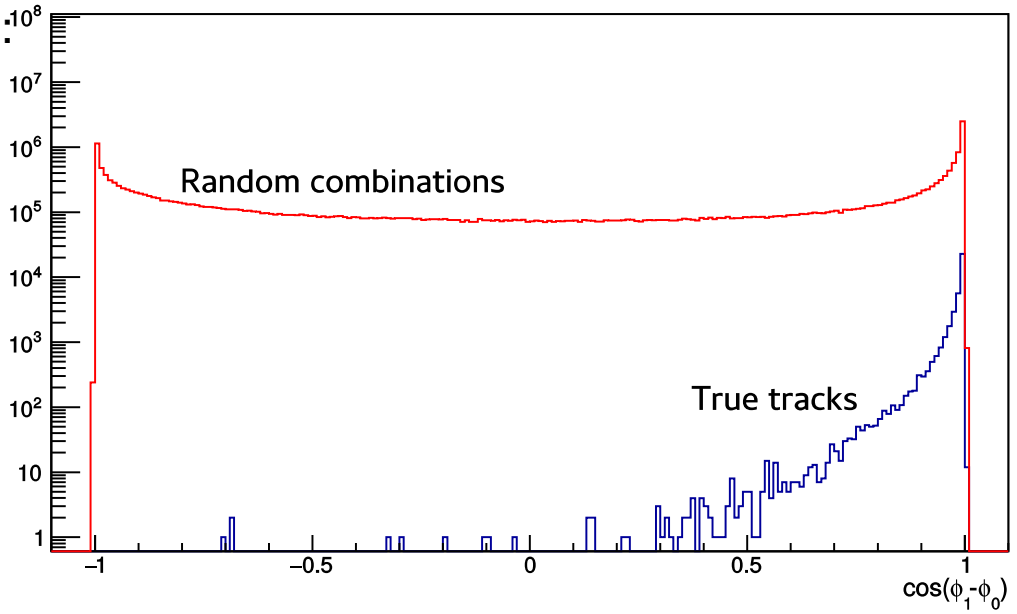
$$\chi^2 = \frac{\Phi_{MS}^2}{\sigma_{MS,\Phi}^2} + \frac{\theta_{MS}^2}{\sigma_{MS,\theta}^2}$$

Geometrical Selection

In subsequent layers, cut on:

- Z-difference of hits
- Φ-difference of hits

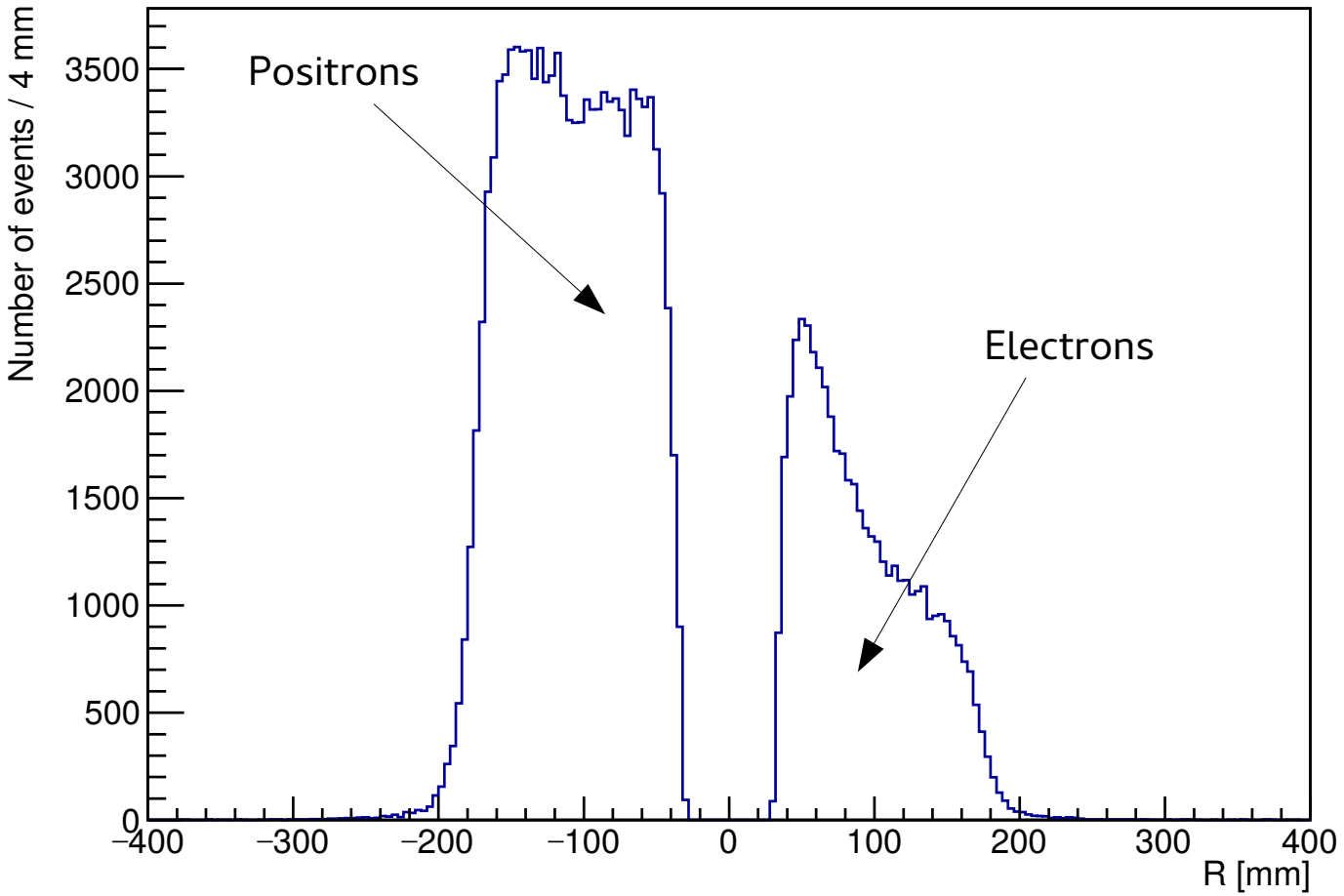




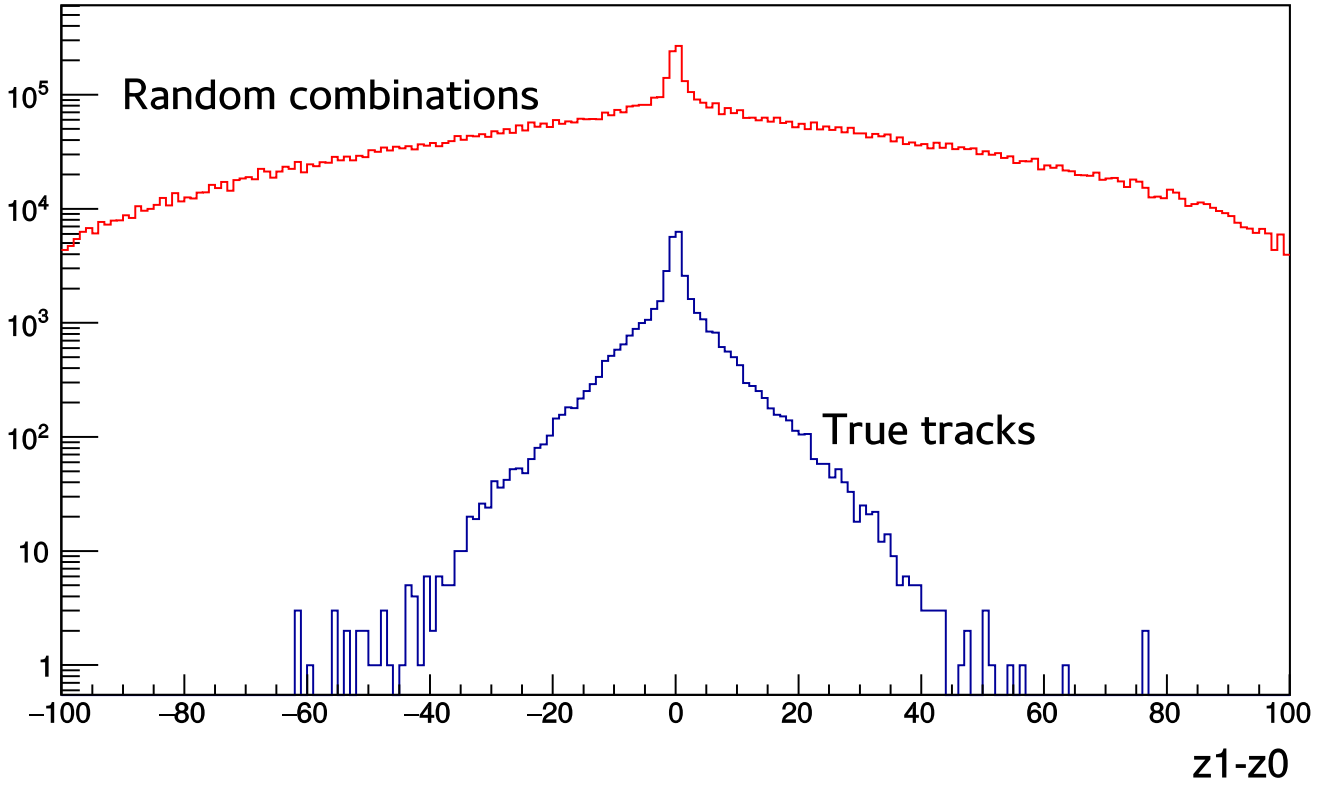
After all cuts:

Reduce 3-hit combinations by factor 50

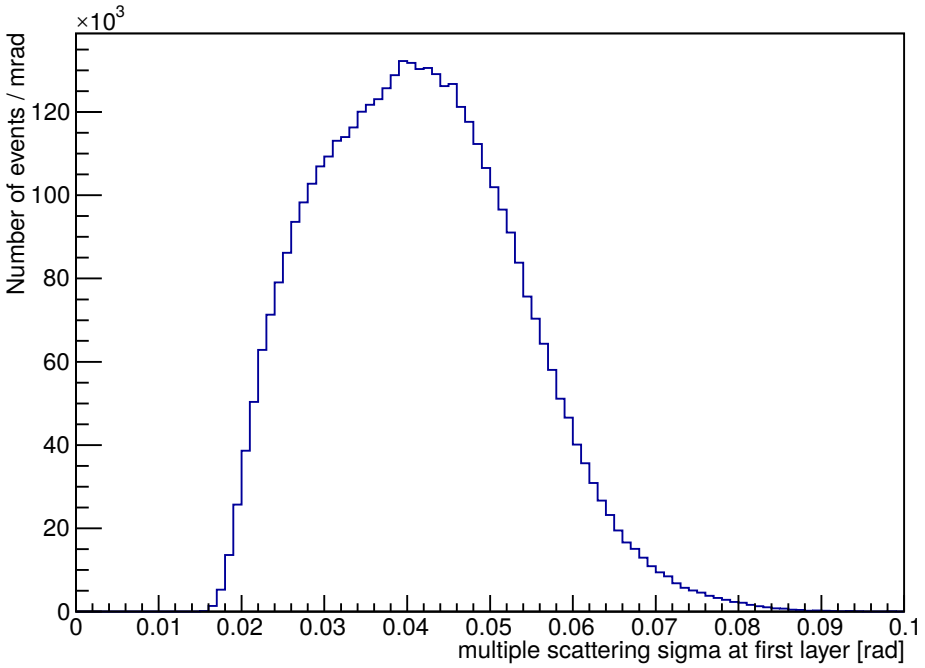
Radius Distribution

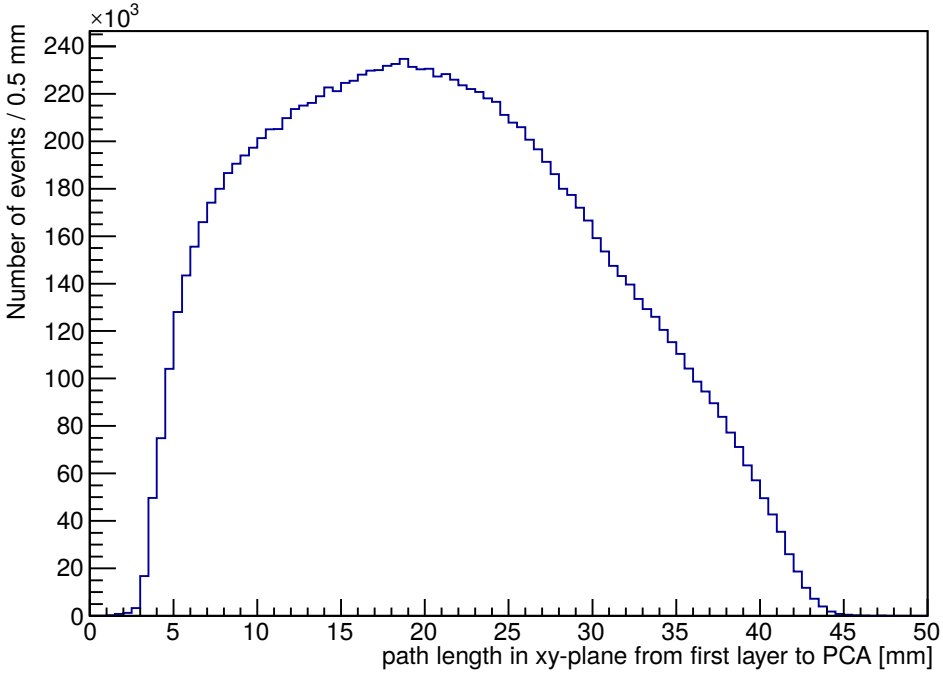


Z distance


Random combinations
True tracks
10^5
10^4
10^3
10^2
10
1
−100
−80
−60
−40
−20
0
20
40
60
80
100
z1-z0

Uncertainty at Intersection



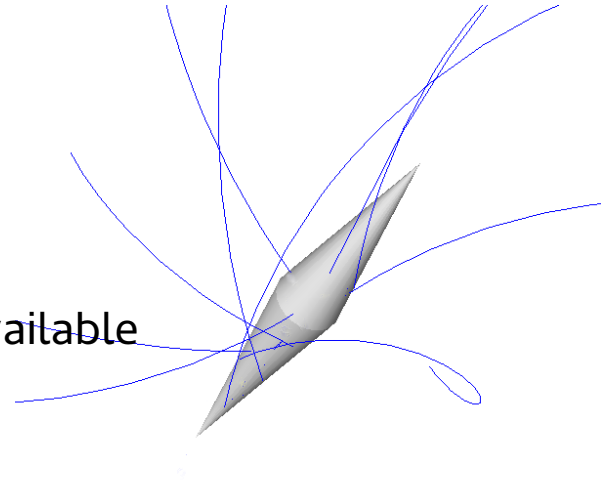


$$\sigma_{MS,PCA} = \sigma_{MS,\,first\ layer} \cdot s \approx 0.8\,mm$$

$$\sigma_{pixel} = 0.08\,mm/\sqrt{12} = 0.02\,mm$$

Take both into account when calculating weights

Offline Reconstruction Reference



- Full detector simulation is available
- For this study:
  - Simulated signal events with one signal decay / 50 ns frame
  - Simulated background events with ordinary muon decays
- Full offline reconstruction includes:
  - Track reconstruction with hits from all layers and recurl stations
    - Matching and linking of recurling track pieces
  - Linearised vertex fit for low momentum tracks in magnetic field

Selection on GPU

- Obtain 50 ns data slices on DAQ computer, so called frames
  - Need to process $20 \cdot 10^6$ frames / s
  - Will have about 10 DAQ computers
- → Process $2 \cdot 10^6$ frames / s on each computer

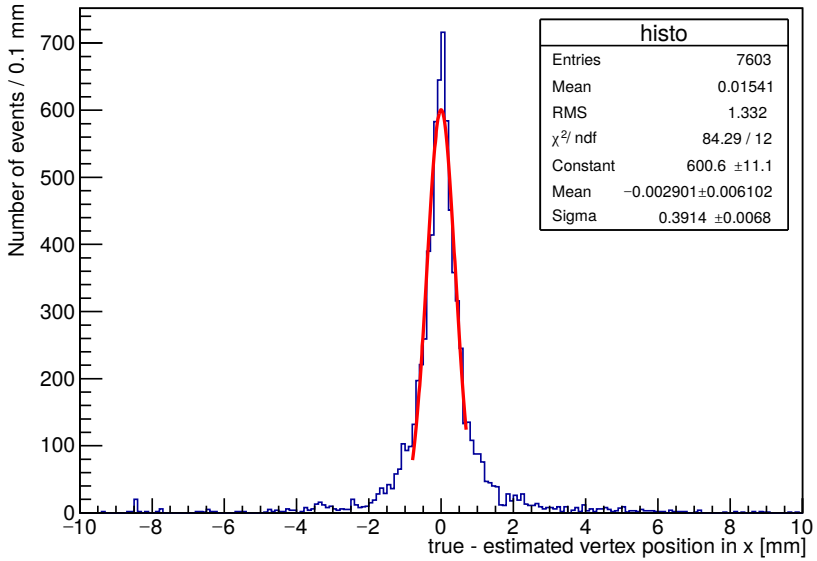
FPGA

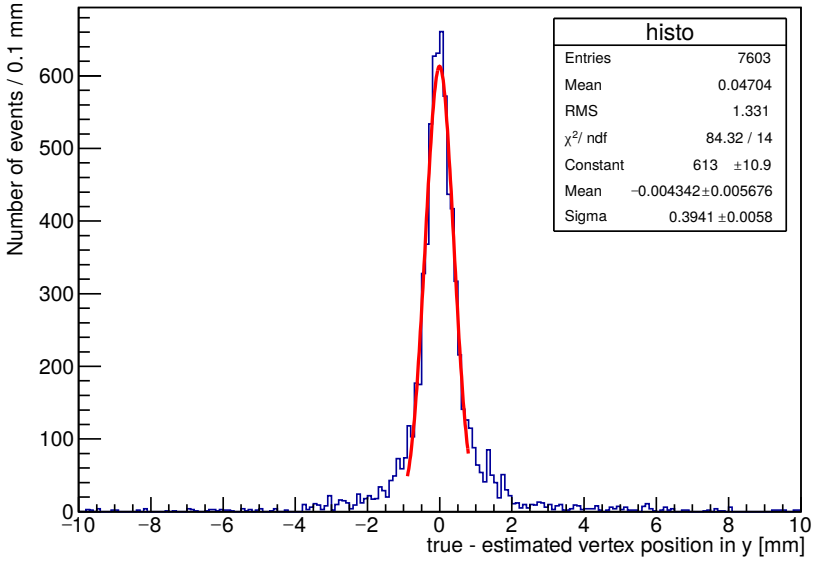
DMA

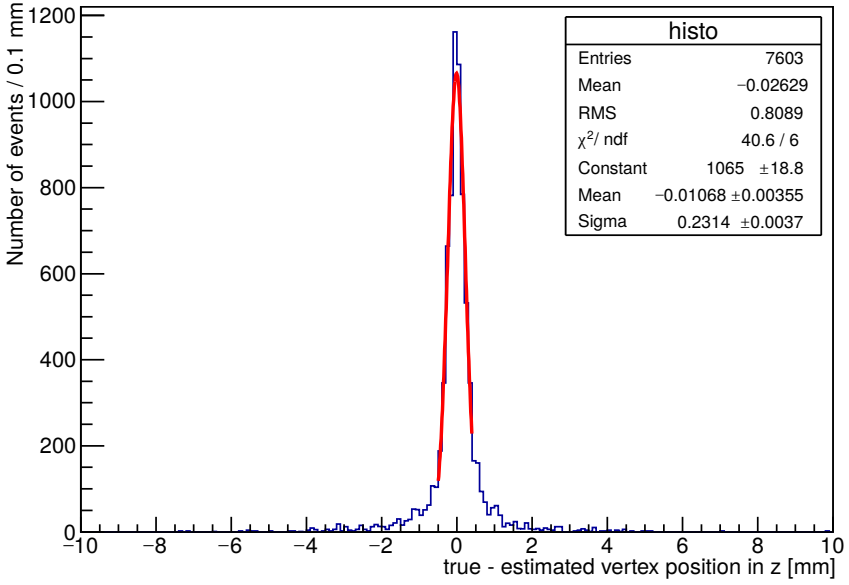
GPU

- Geometric selection cuts
  - Save hit positions of the three hits belonging to one triplet and hits in fourth layer
- Fits with three and four hits
  - Vertex selection
  - Save frame decision

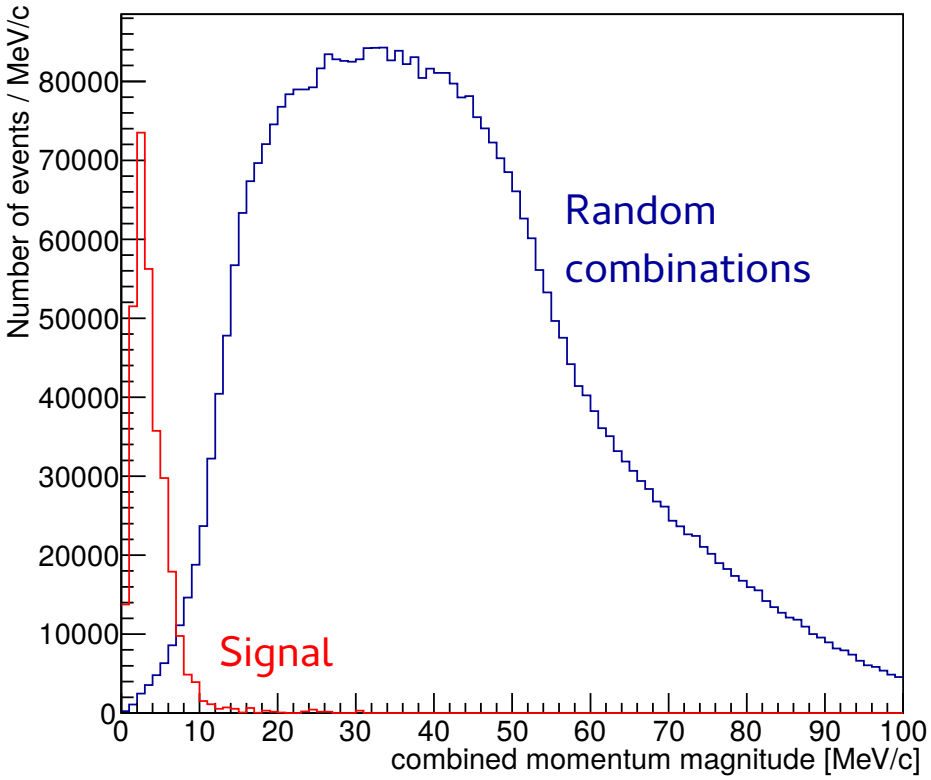
Vertex Position Distribution

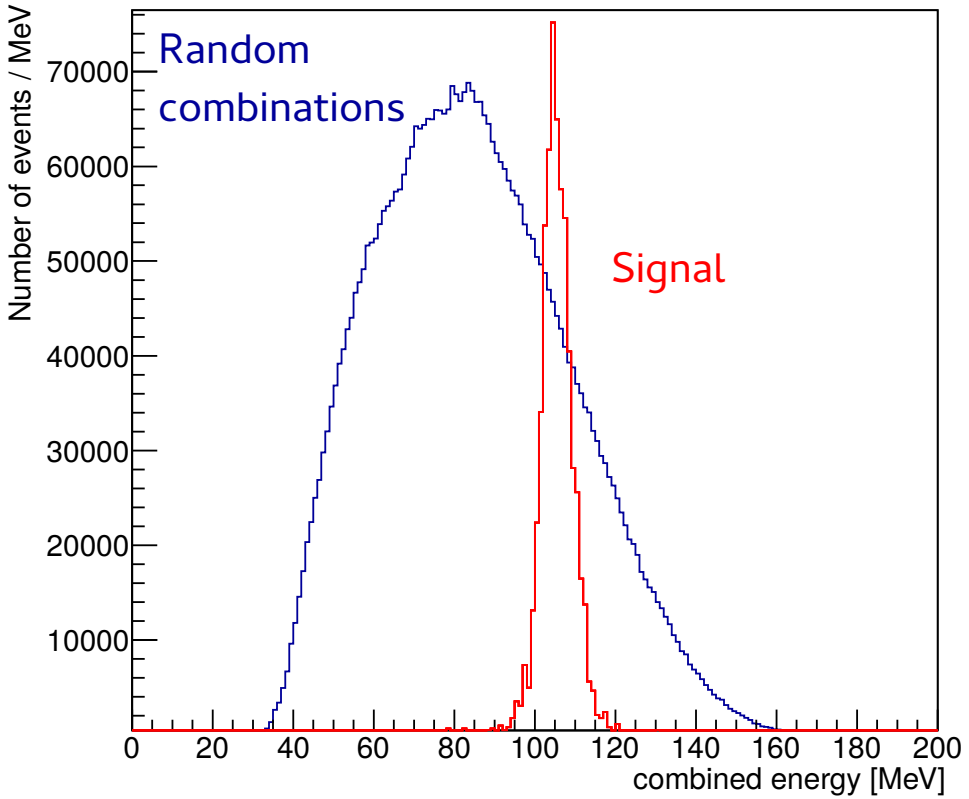




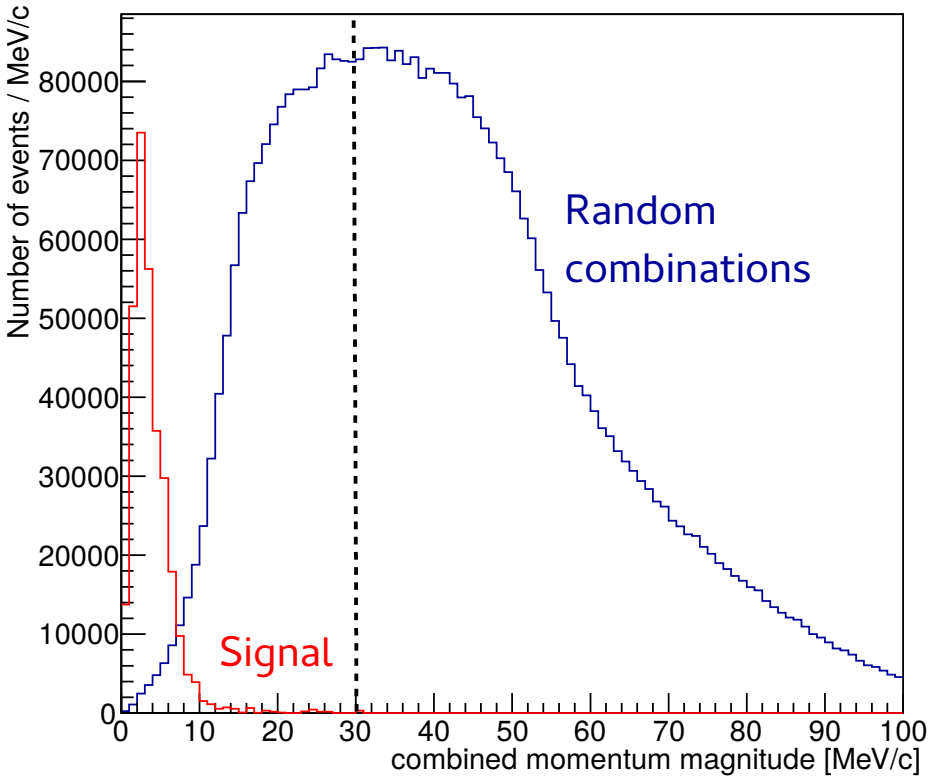


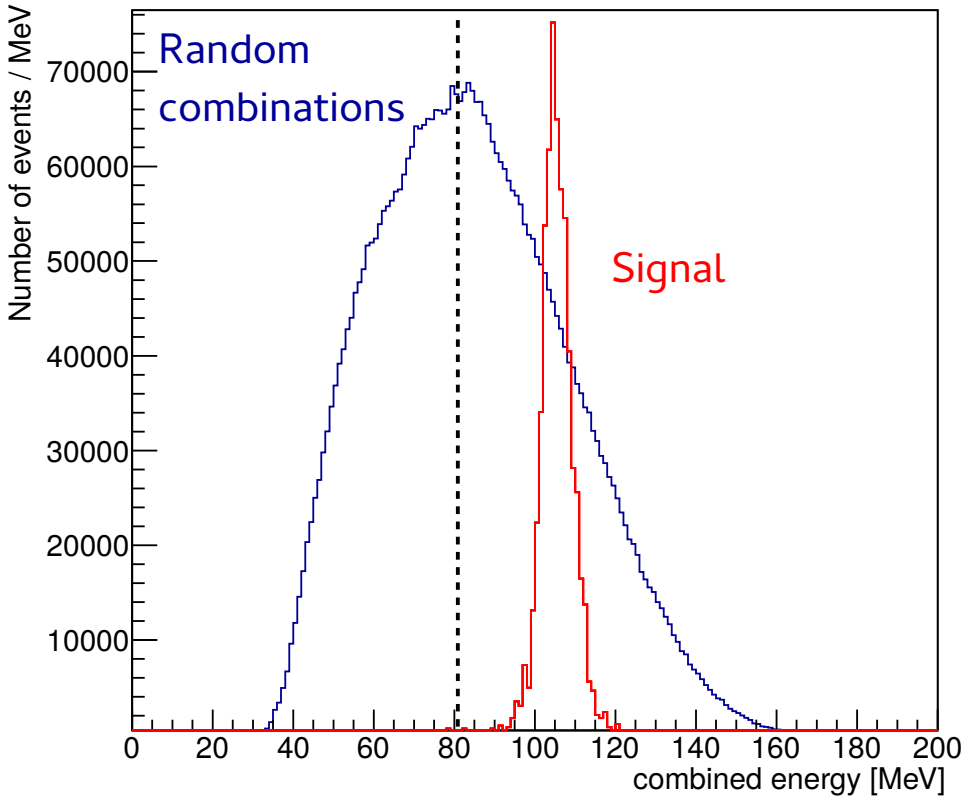
Combined Momentum and Energy



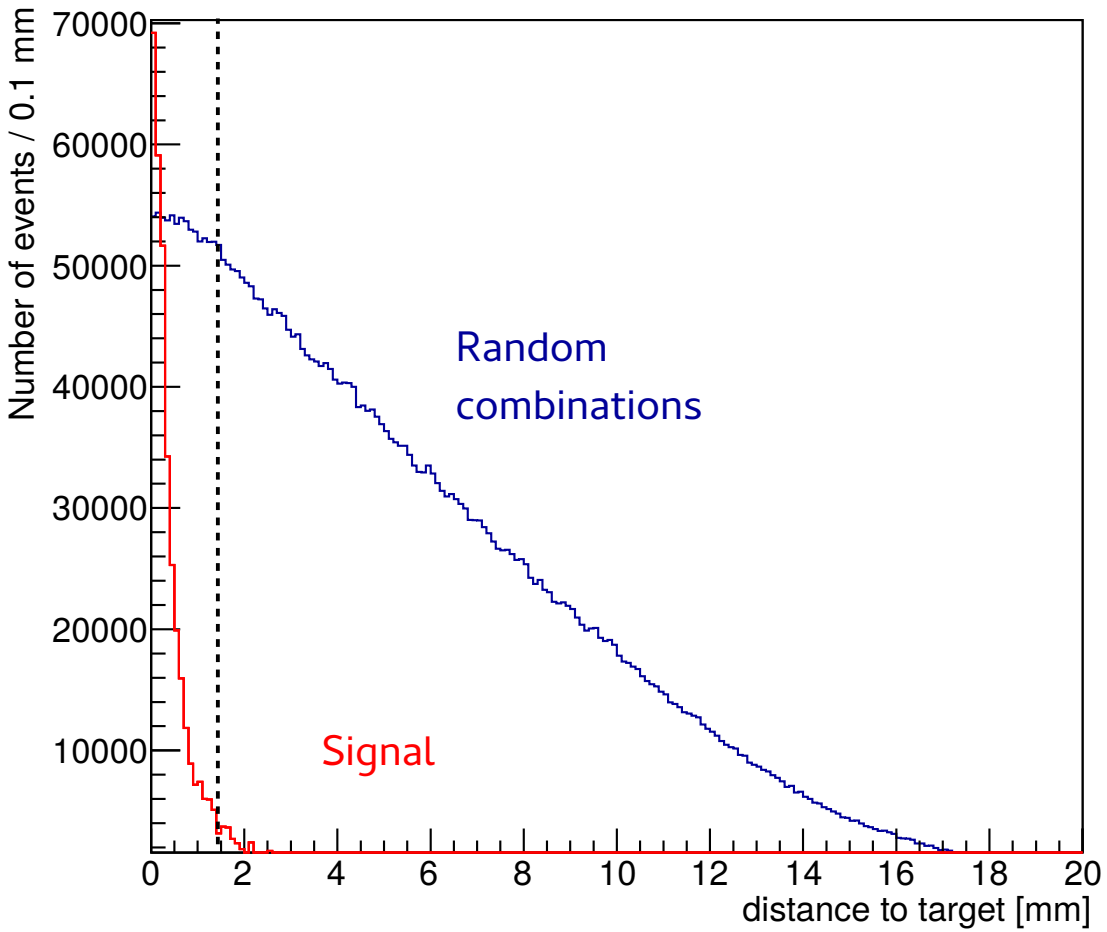


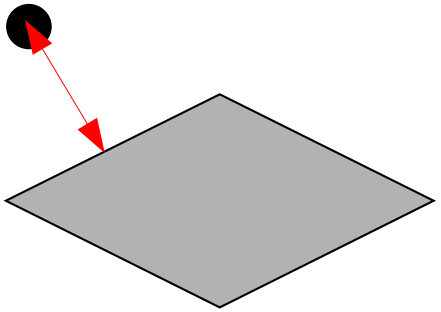
Combined Momentum and Energy



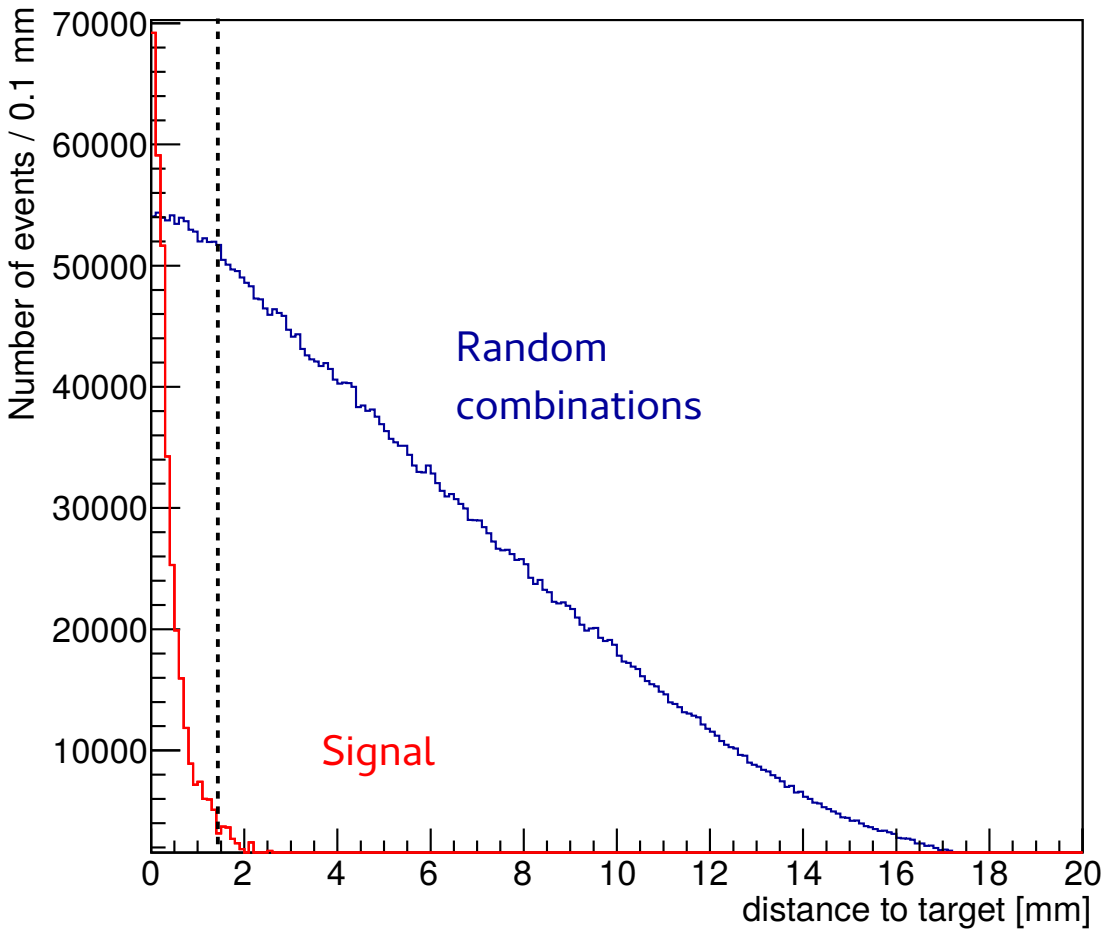


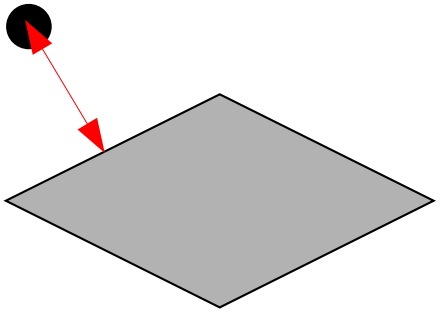
Distance to Target





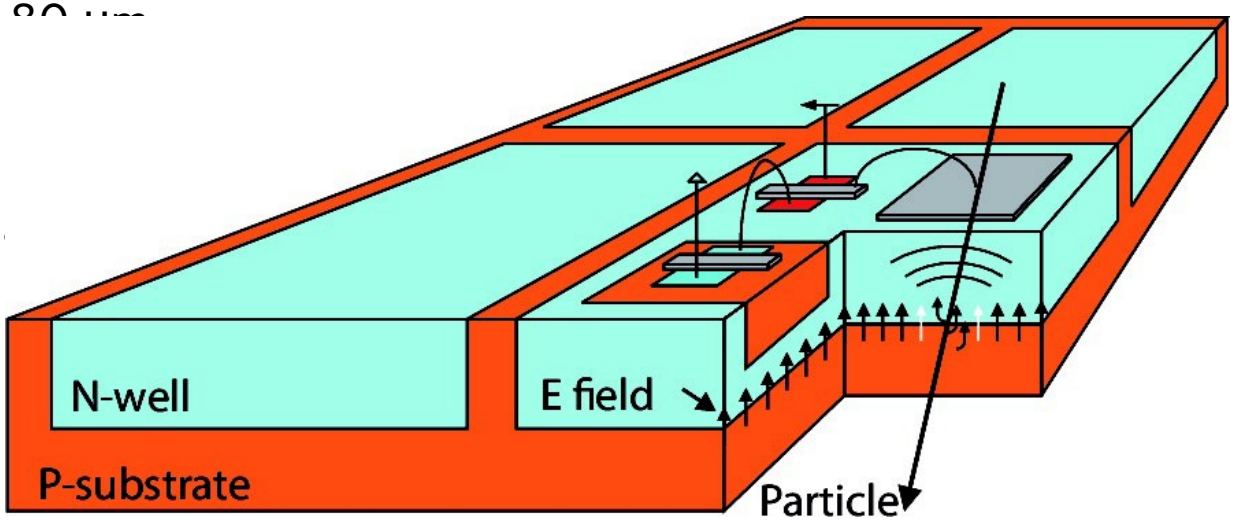
Distance to Target





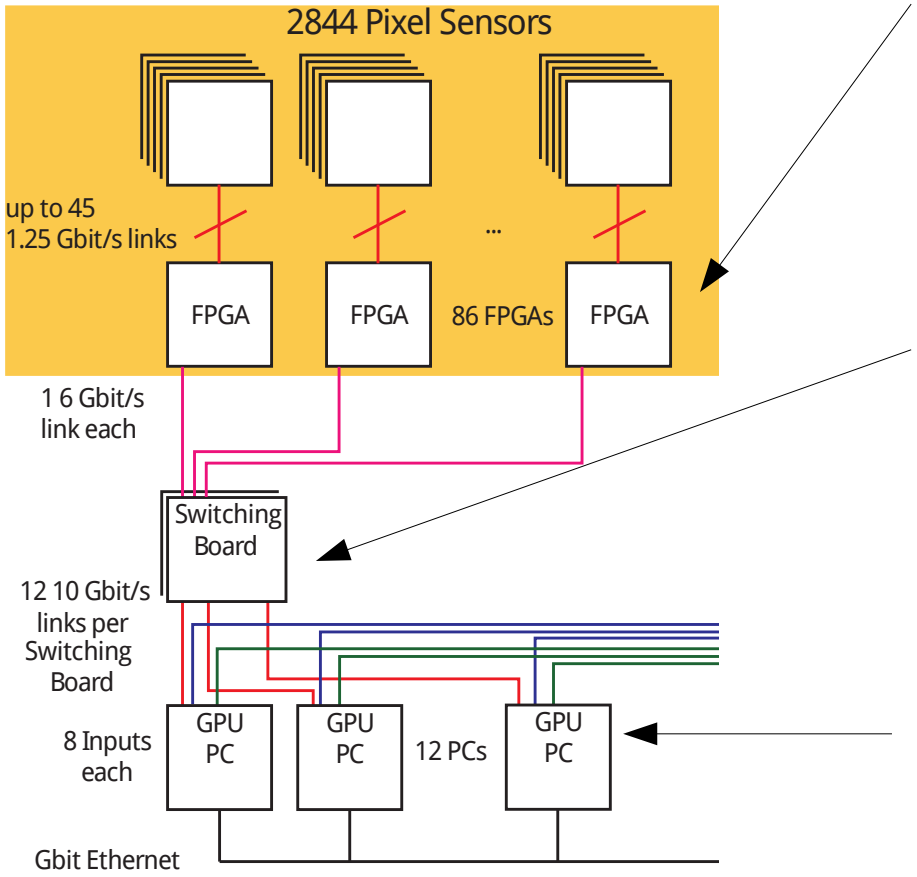
Pixel Detector

- High Voltage Monolithic Active Pixel Sensors (HV-MAPS)
  - Fast charge collection via drift
    - Thinned down to 50 µm
    - Pixel size: 80 µm x 80 µm
    - Chip size: 2 cm x
    - Thickness chip & re
    - $\mathcal{O}$(0.1 %) radiation



I. Peric, P. Fischer et al, NIM A 582 (2007) 876

Readout Scheme



Front-end board:

- Sort hits according to time stamps
- Send off via optical links

Switching board:

- Merge data from different detector regions
- Pack into 50 ns time slices
- Send off via optical links

PCIe board:

- First data selection
- Transfer data to RAM of PC via PCIe