Online Track Reconstruction on GPUs for the Mu3e Experiment

Dorothea vom Bruch

for the Mu3e Collaboration

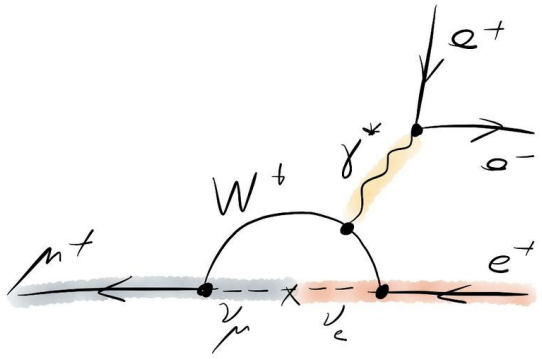
DPG Frühjahrstagung 2016, T42: Trigger und DAQ II

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

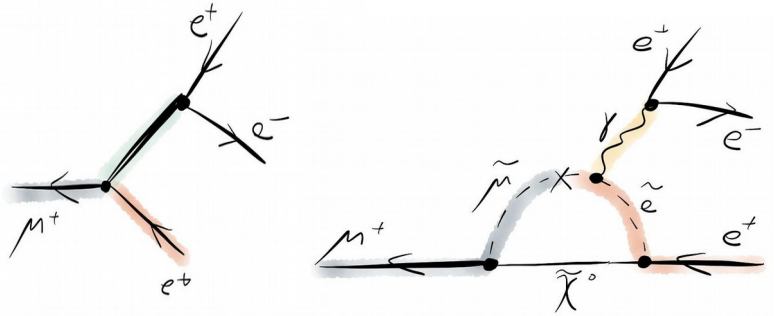
JG|U

The Mu3e Experiment

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



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

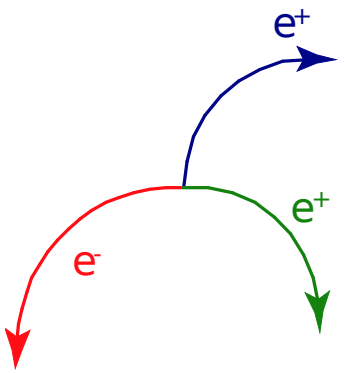


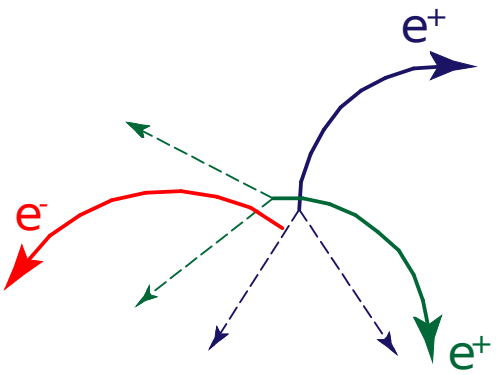
Any hint of signal ⟶ new physics

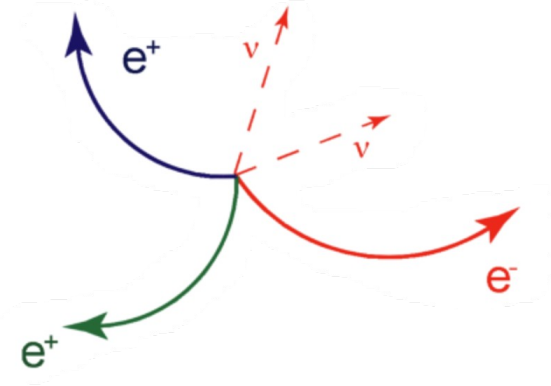
- Supersymmetry
- Grand unified models
- Extended Higgs sector
- ...

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

Signal versus Background







Signal

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

Random Combinations

- Not coincident in time
- No single vertex
- $\sum \vec{p}_i \neq 0$
- E $\neq m_\mu$

Internal Conversion

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

The Mu3e Detector

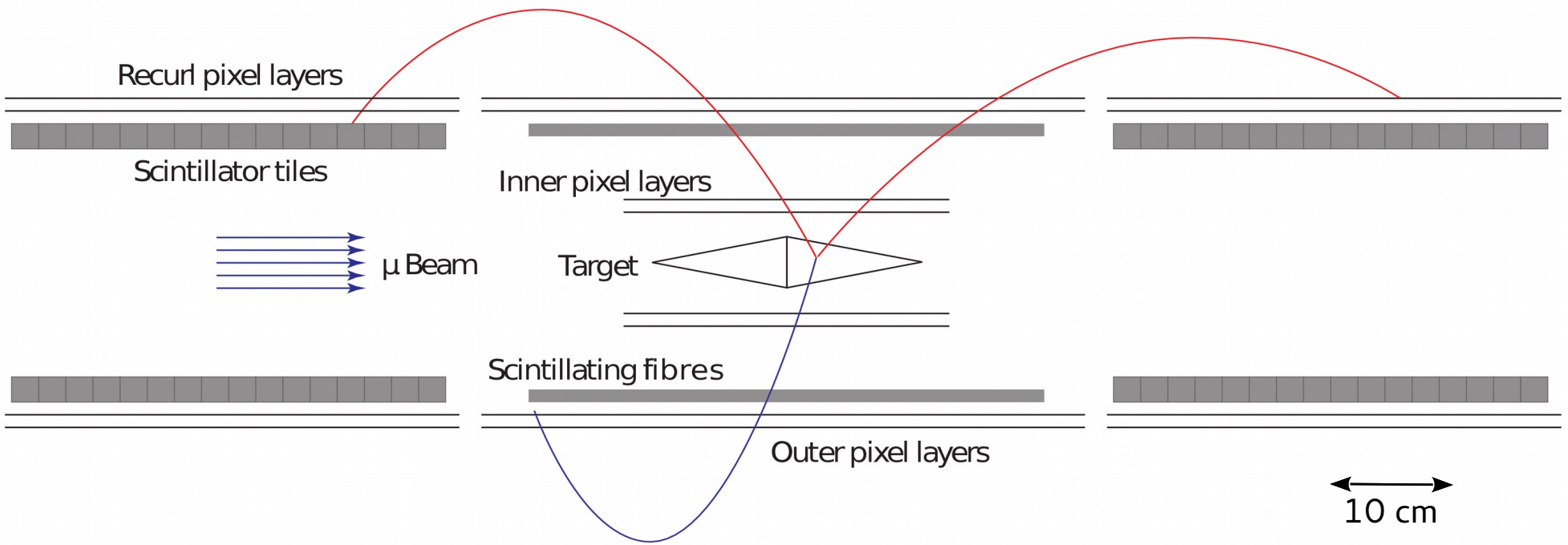
Requirements

- Excellent momentum resolution: < 0.5 MeV/c
- Good timing resolution: 100 ps for tiles, 1 ns for fibres, < 20 ns for pixels
- Good vertex resolution: 300 µm
- High rates: 10^8 -10^9 µ/s (Paul Scherrer Institute, Switzerland)

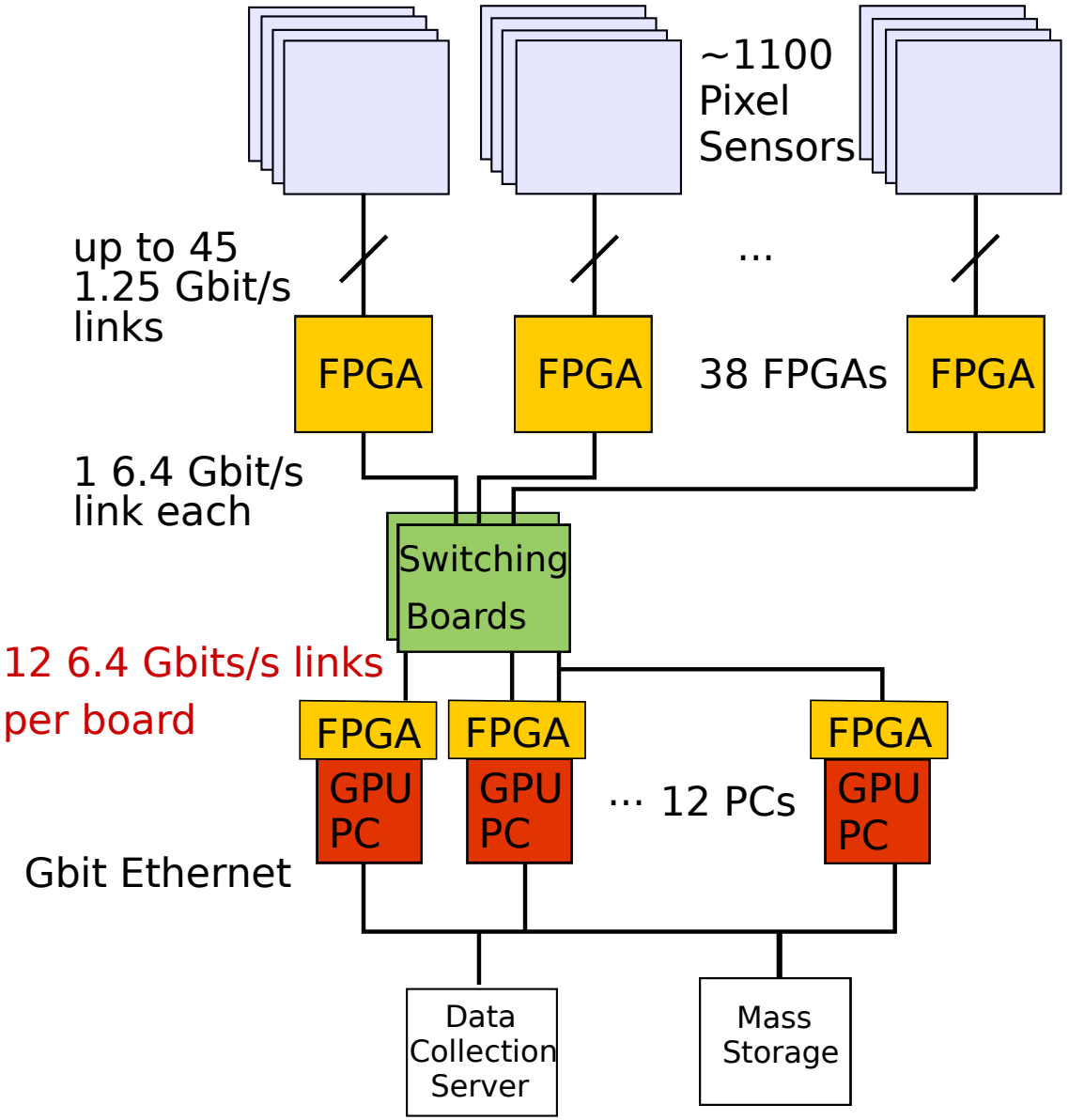
The Mu3e Detector

Requirements

- Excellent momentum resolution: < 0.5 MeV/c
- Good timing resolution: 100 ps for tiles, 1 ns for fibres, < 20 ns for pixels
- Good vertex resolution: 300 µm
- High rates: 10^8 -10^9 µ/s (Paul Scherrer Institute, Switzerland)



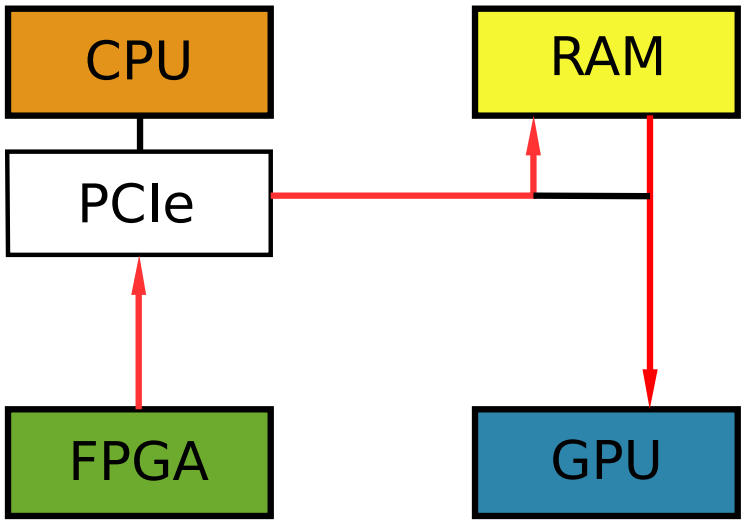
Readout Scheme



- Triggerless readout →

  50 Gbit/s data rate
- Online data reduction
- Track reconstruction and vertex fitting on Graphics Processing Units (GPUs)
- Reduction factor of ~1000

Fast Data Transfer

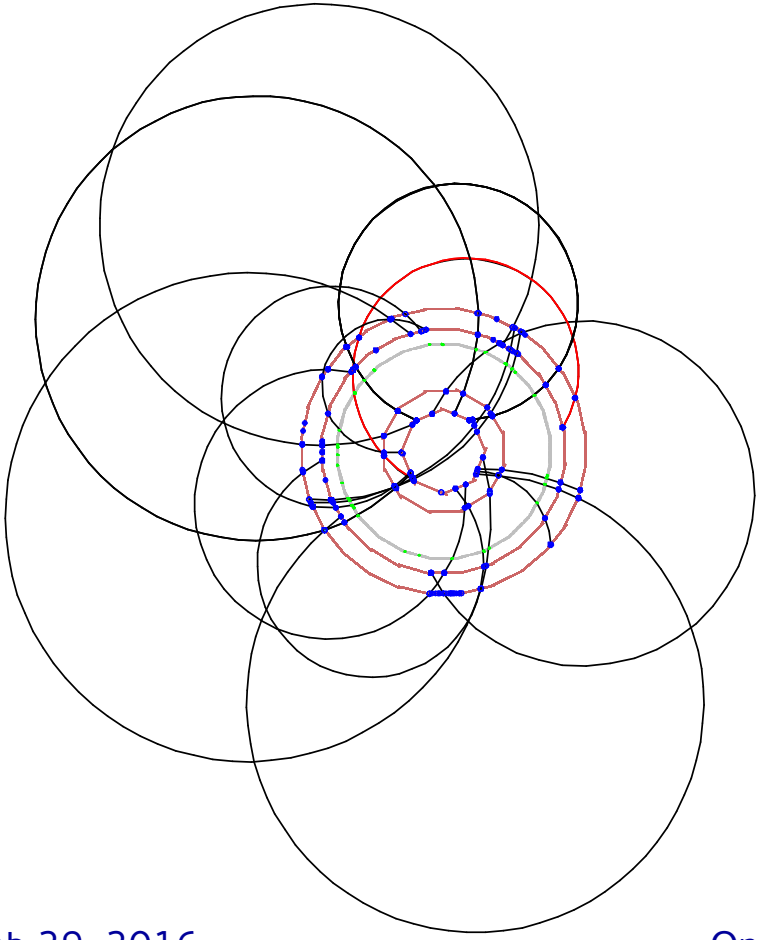


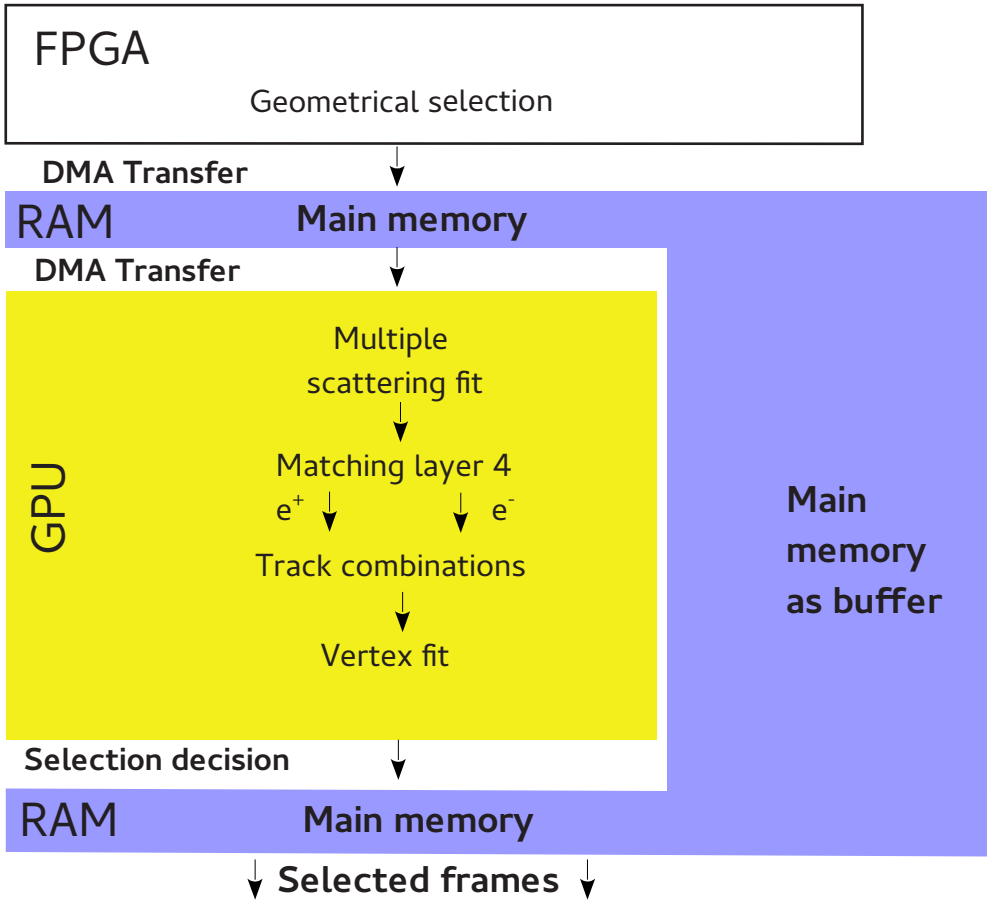
- Direct Memory Access to main memory
- Copy to GPU memory
- At 1.5 GB/s: measured bit error rate

  $< 4 \times 10^{-16}$

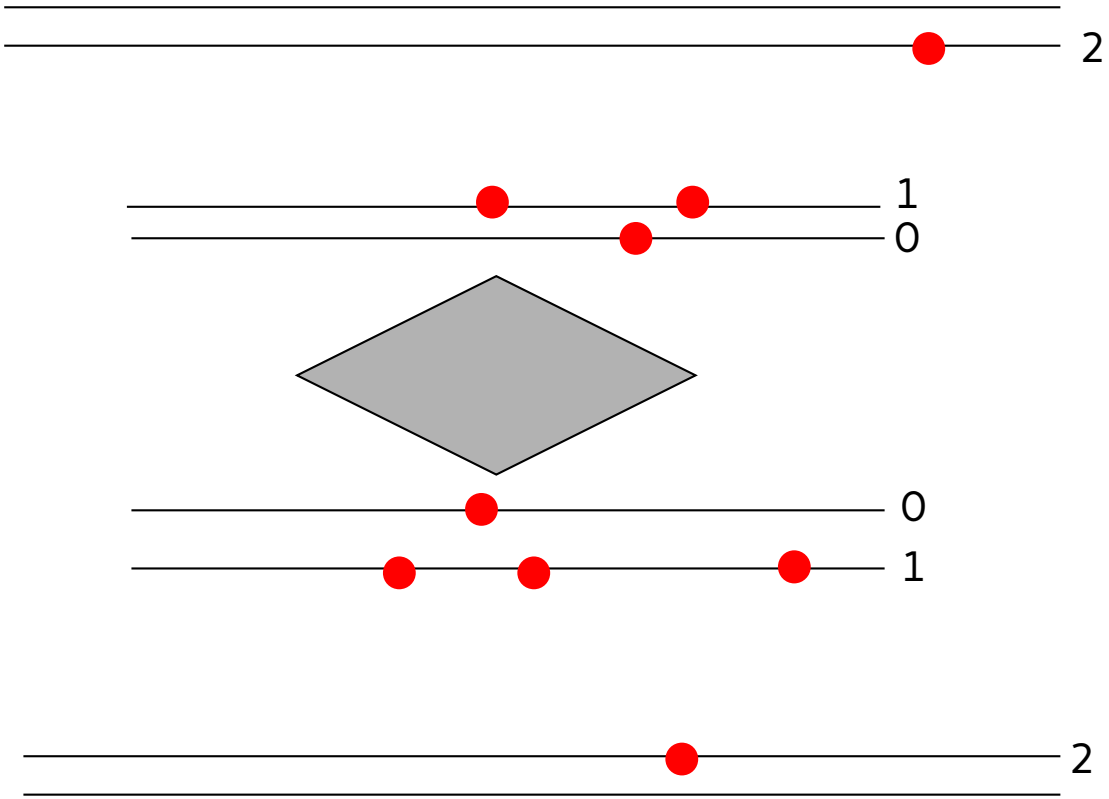
Online Reconstruction

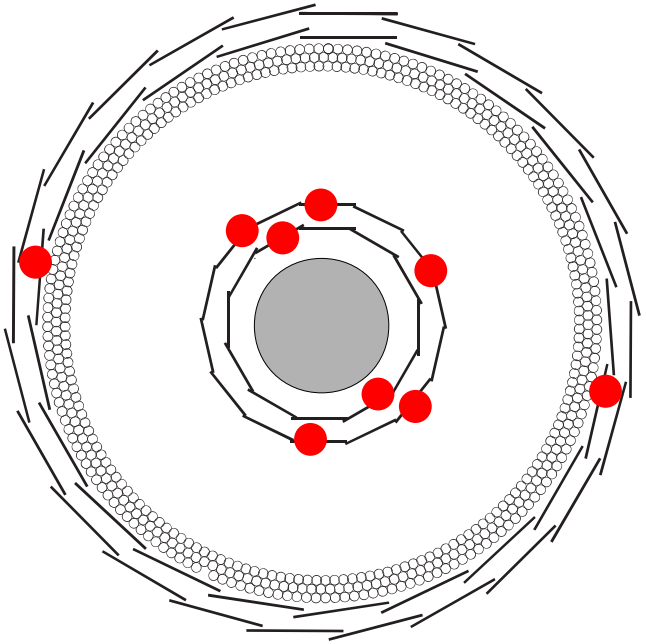
- Number of possible track candidates ~ n^3
- At 10^8 μ/s: ~10 hits / layer / 50 ns → $\mathcal{O}(10^3)$ combinations / 50 ns

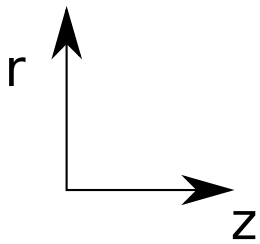


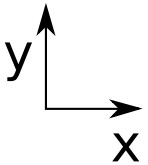


Geometrical Selection

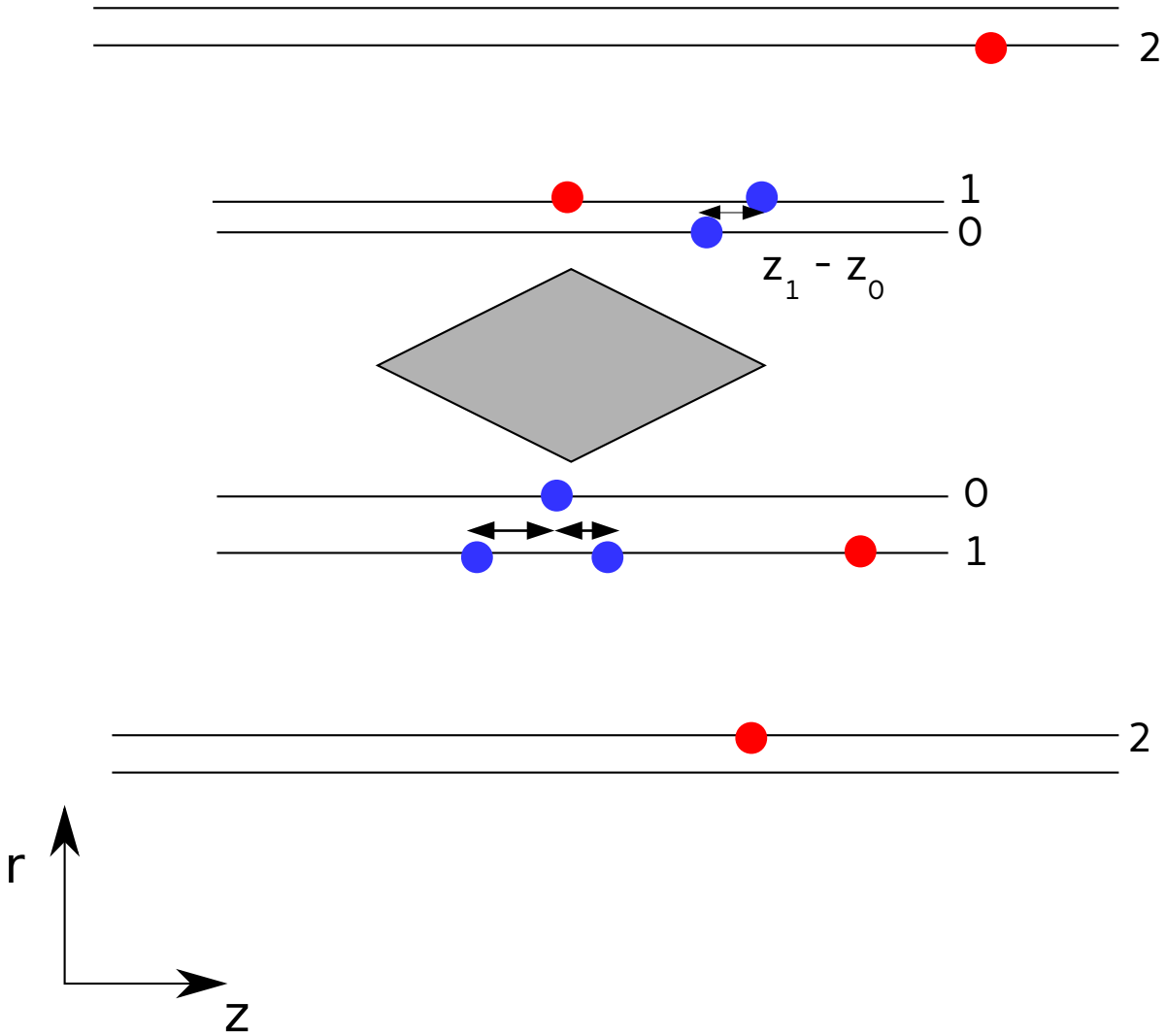


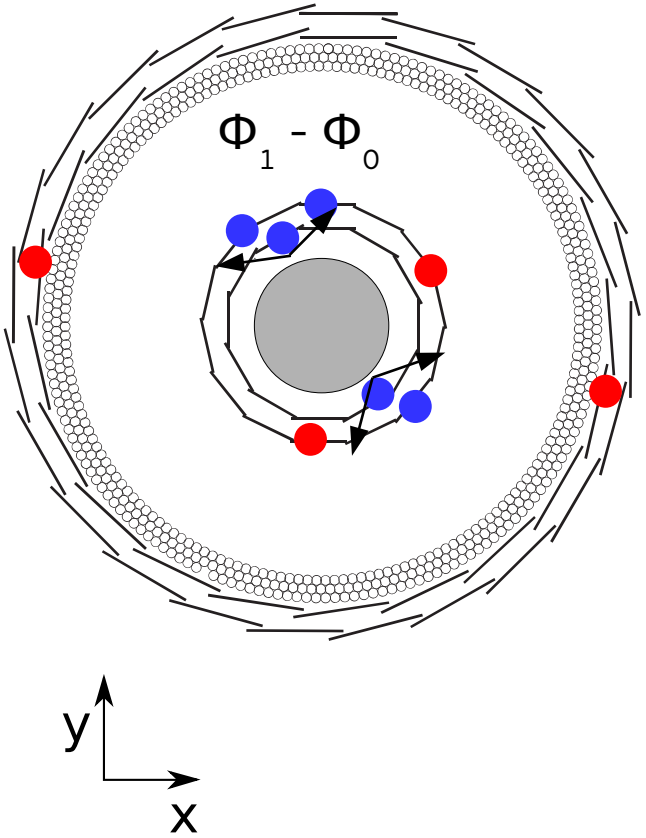




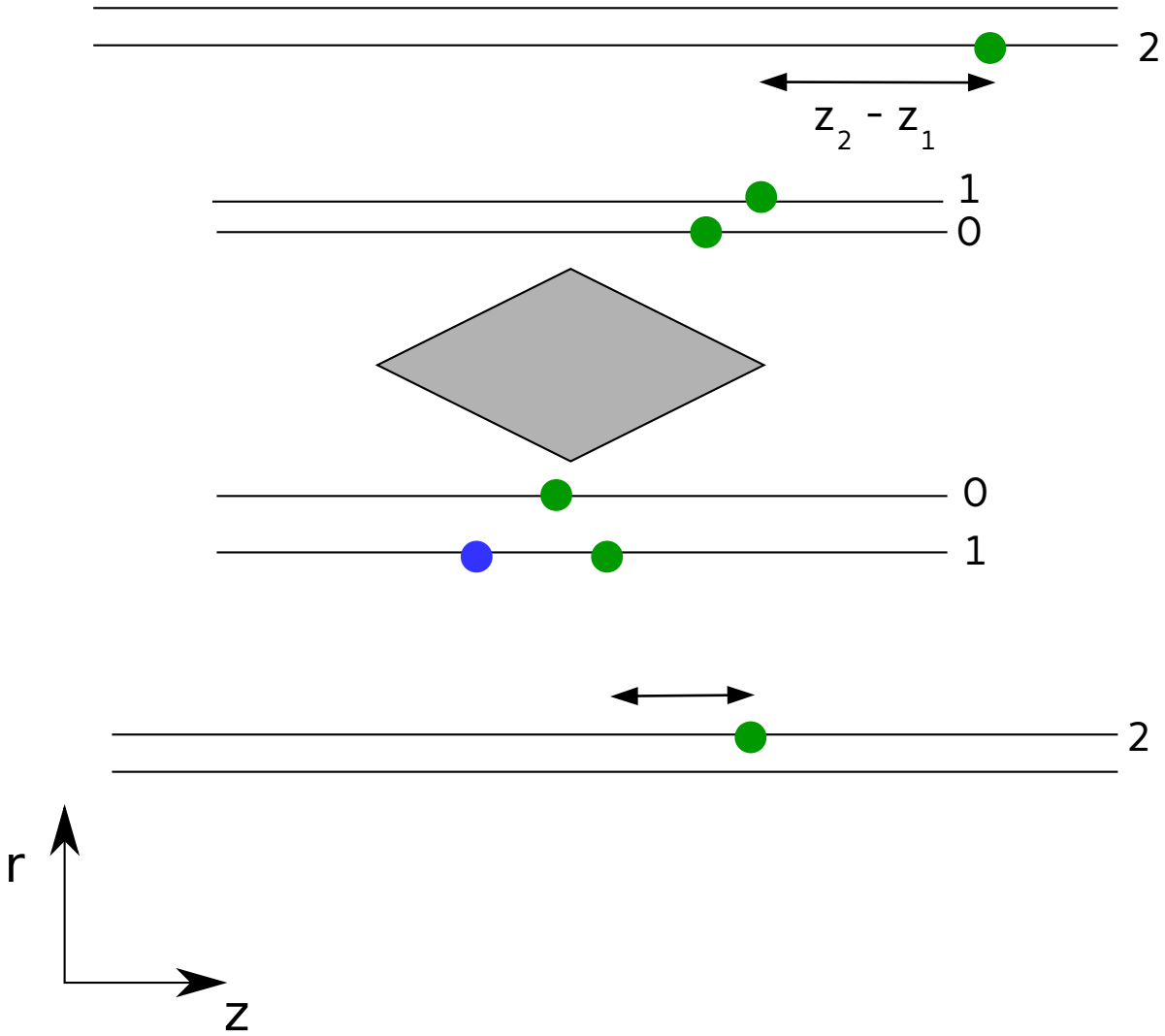


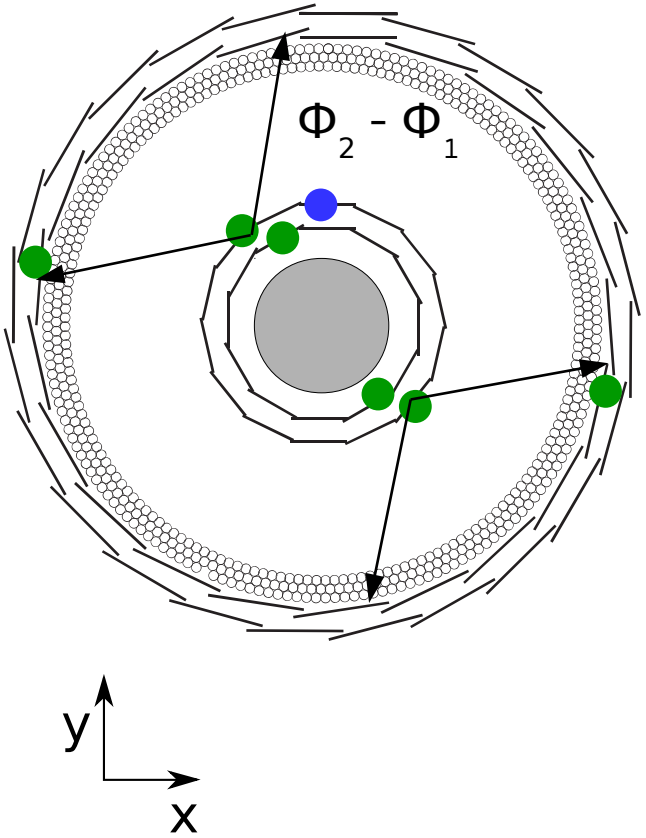
Geometrical Selection



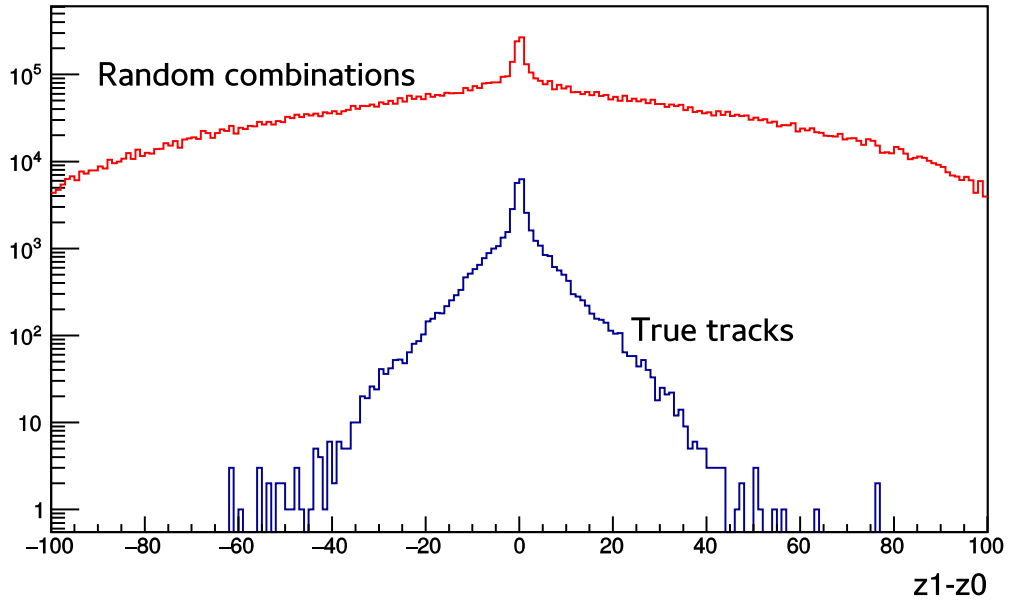


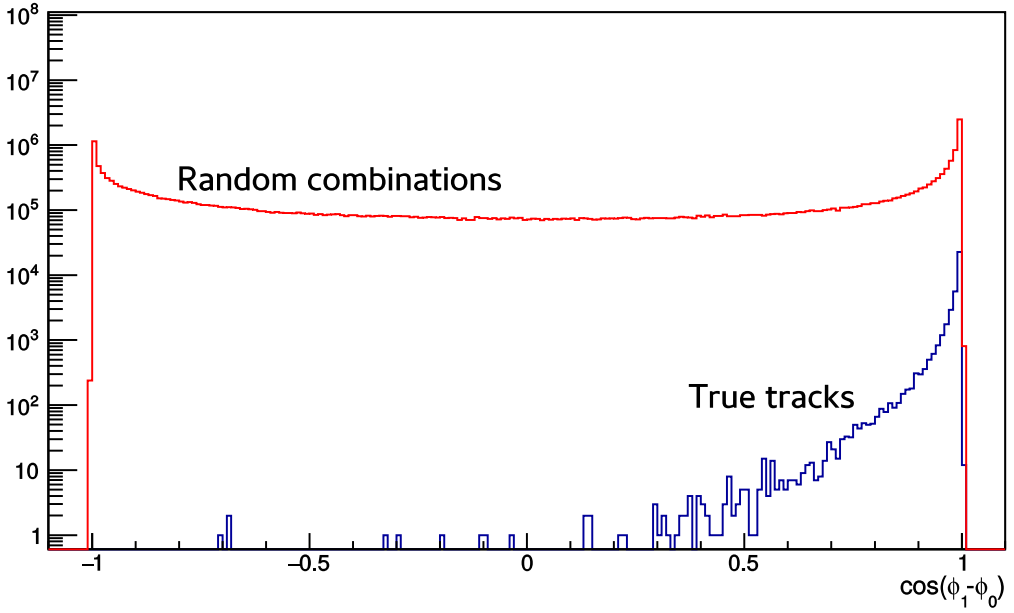
Geometrical Selection





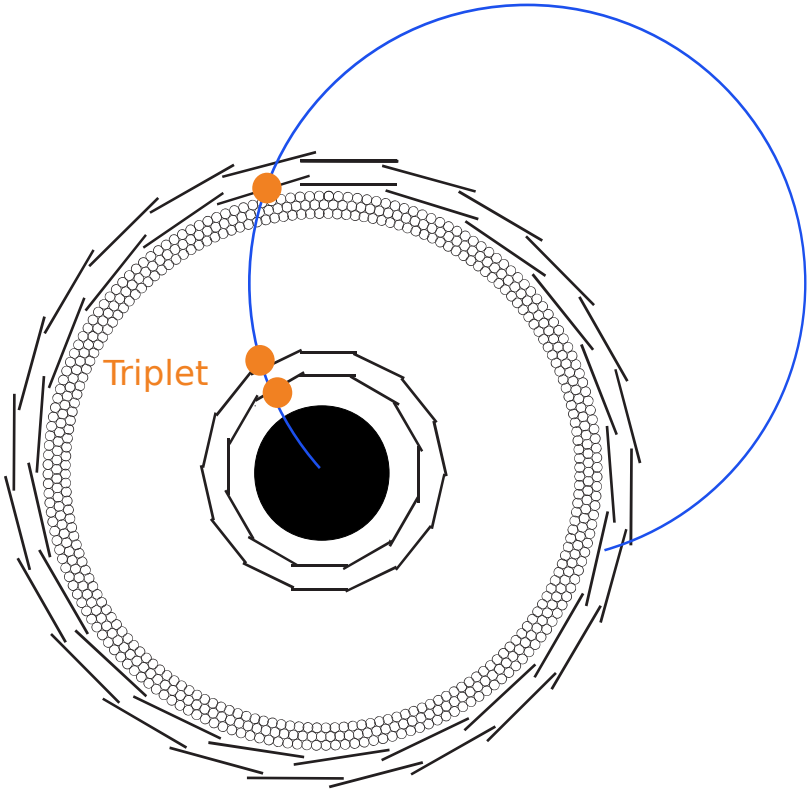
Geometrical Selection





Multiple Scattering Fit

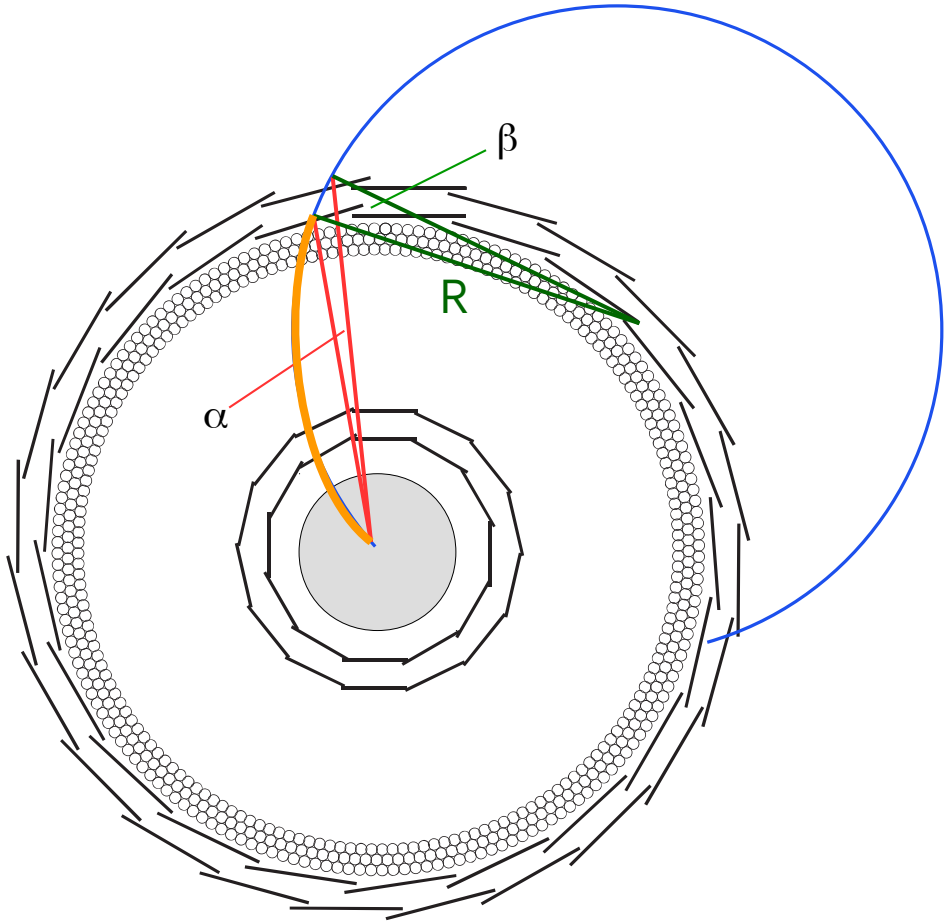
- Electrons: 12 – 53 MeV/c
- Resolution dominated by multiple Coulomb scattering



- Ignore hit uncertainty
- Describe track as sequence of hit triplets
- Multiple scattering at middle hit of triplet
- Minimize multiple scattering

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

Propagation to 4th Layer

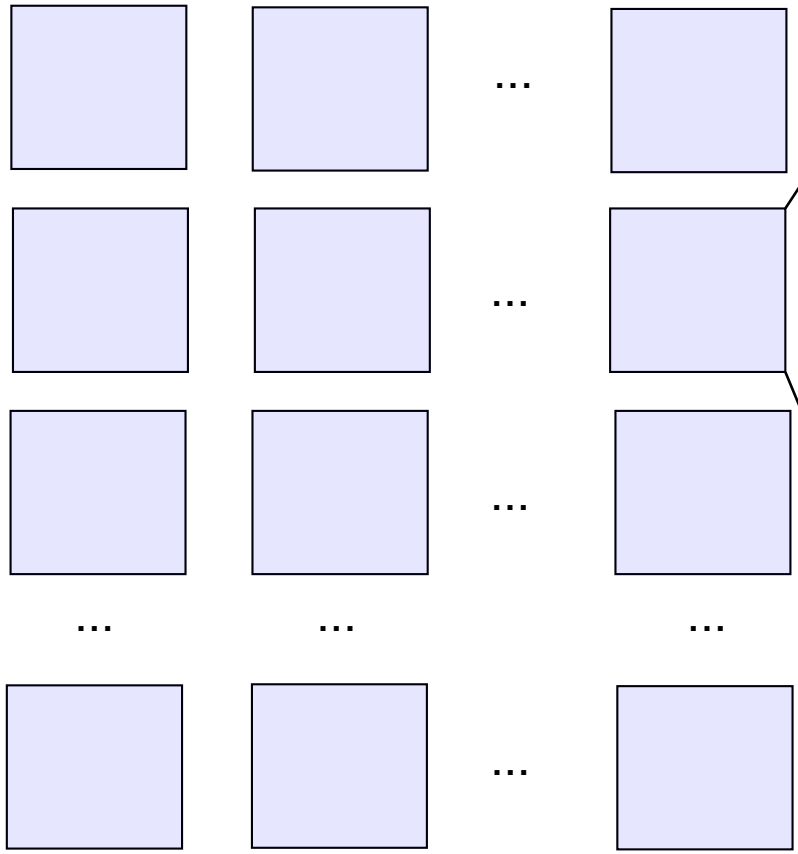


- Position of 4th layer known
- α : propagate in xy-plane
- β : propagate in z direction

After all selections:

- 98 % of true 4-hit tracks selected
- 65 % random combinations of 3 hits

Parallelization



- Fit for one combination of three hits
- Cut on χ^2
- Propagation to 4th layer
- Loop over hits in 4th layer: check if hit exists in proximity of propagated track

~ 2000 compute cores on GPU

Performance

10^8 muons / s	GTX680	GTX980
Fits / s	$2x10^7$	$3x10^7$
10^9 muons / s		
Fits / s	$9.7x10^9$	$1.6x10^{10}$

Pictures: pcmag.com, nvidia.com

Performance

10^8 muons / s	GTX680	GTX980
Fits / s	$2x10^7$	$3x10^7$
10^9 muons / s		
Fits / s	$9.7x10^9$	$1.6x10^{10}$

10^8 muons / s	Reduction factor	Triplets / s
Total		$2x10^{10}$
After geometrial selection	50	$4x10^8$
After multiple scattering fit	2	$2x10^8$
After propagation To 4th layer	2.5	$8x10^7$

@ 10^8 μ/s: $\mathcal{O}(10)$ DAQ computers are sufficient

Pictures: pcmag.com, nvidia.com

Next Steps

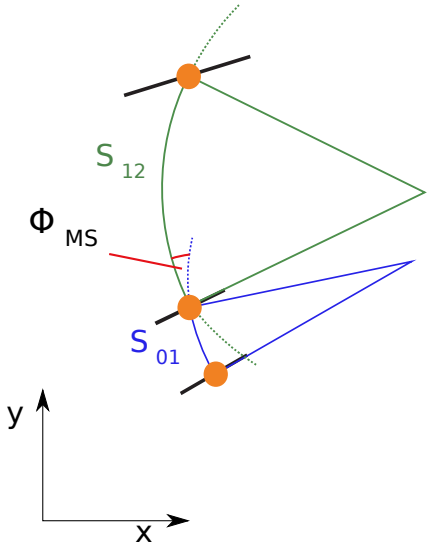
- Study, optimize vertex fit performance
- Simplify for GPU implementation
- Implement geometrical selection on FPGA
- Test whole chain of online selection

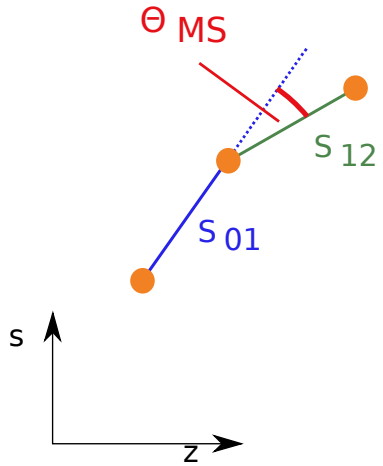
More Mu3e talks:

- Mu3e Experiment: T22.4&5, T42.7, T75.7, T98.1&5
- MuPix Telescope: T42.6, T99.5
- HV-MAPS / MuPix: T72.1-3

Backup Slides

Multiple Scattering Fit





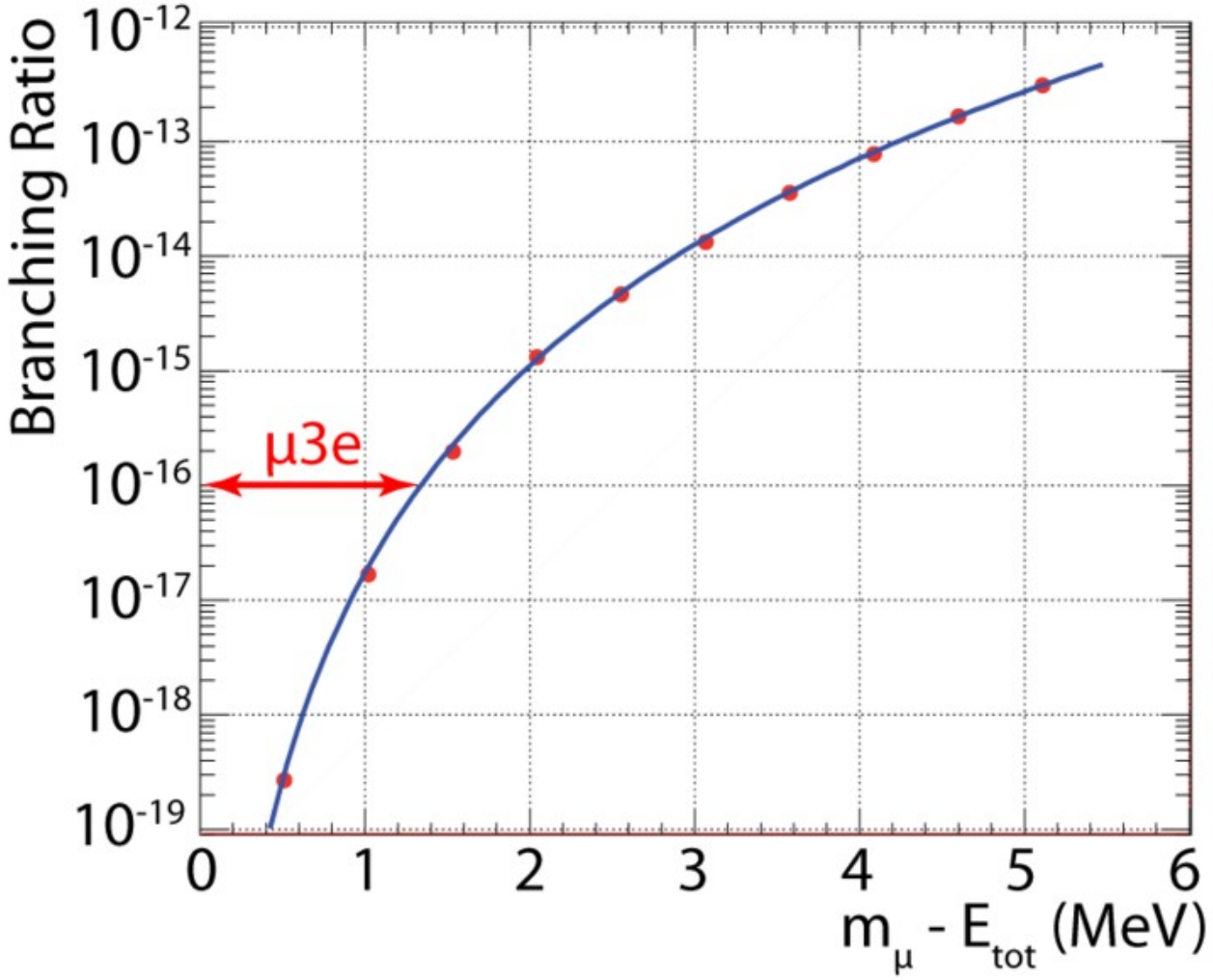
$$\chi^2 = \frac{\varphi^2_{MS}}{\sigma^2_{MS}} + \frac{\theta^2_{MS}}{\sigma^2_{MS}}$$

- R_{3D} from fit
- Sign of R_{3D}

  → track curvature
- Cut on fit success and χ^2

⇒ Reduce by factor 2

Required Momentum Resolution

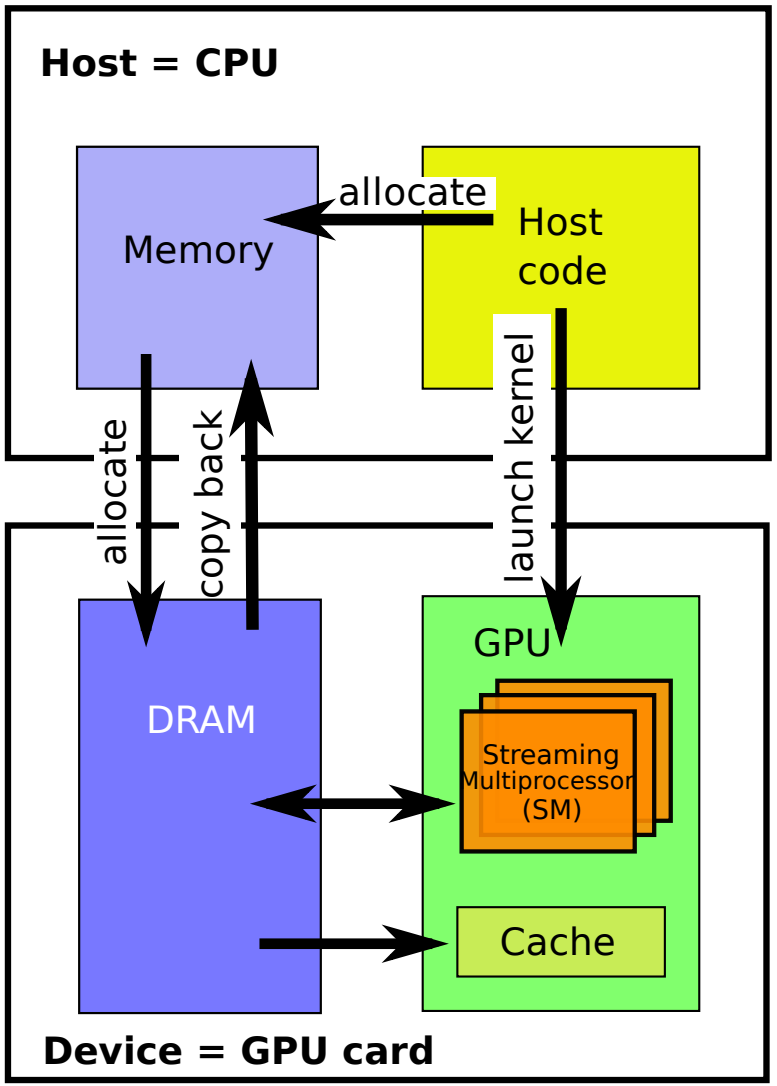


Graph: R. M. Djilkibaev, R. V. Konoplich, Phys.Rev.D79(2009)073004

Performance @ 10^9 muons/s

10^9 muons / s	Reduction factor	Triplets / s
Total		$2x10^{13}$
After geometrial selection	50	$4x10^{11}$
After multiple scattering fit	2	$2x10^{11}$
After propagation To 4th layer	2.6	$8x10^{10}$

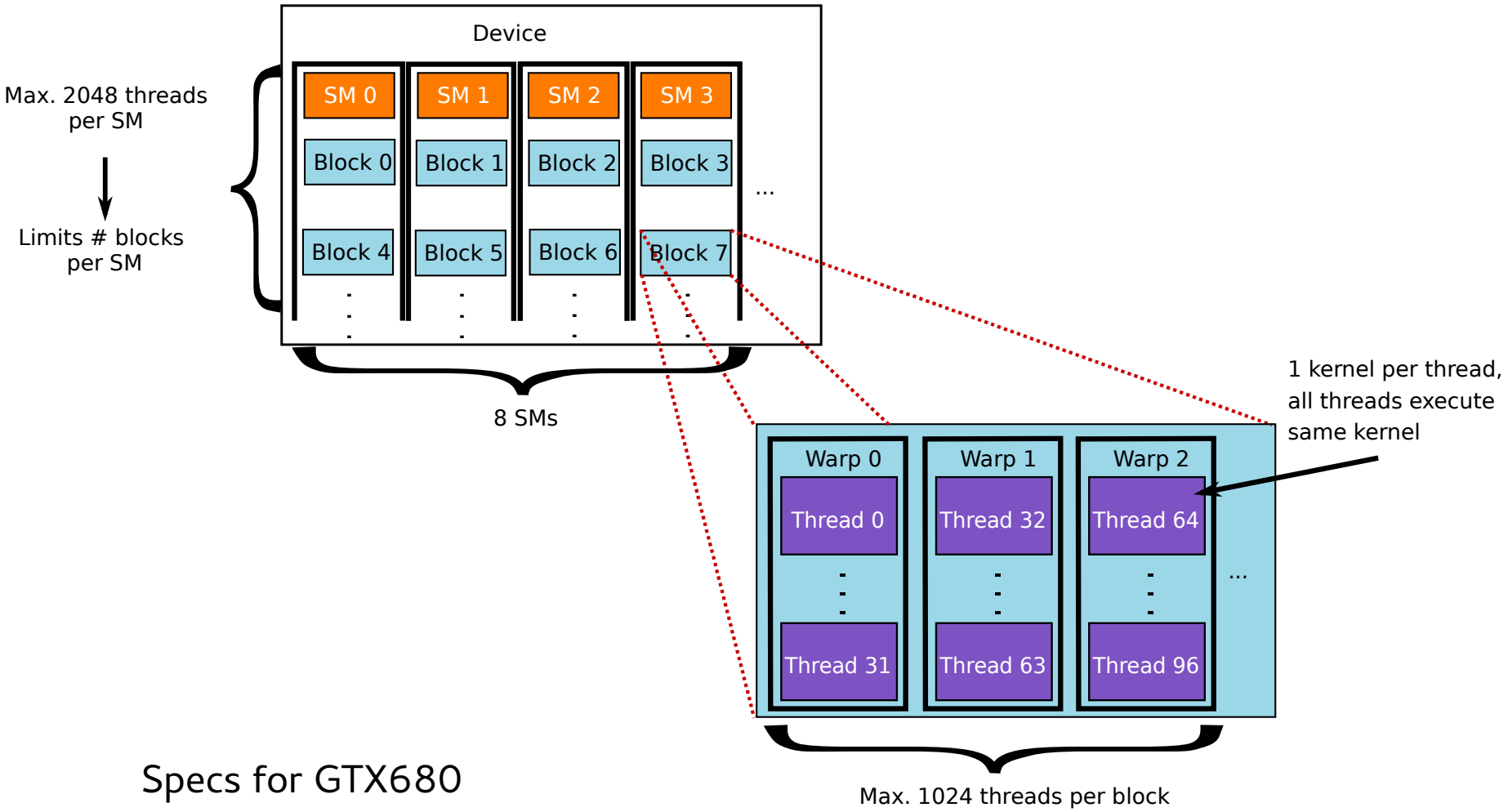
GPU Properties



- Highly parallel structure
- Process large blocks of data
- Nvidia: API extension to C:

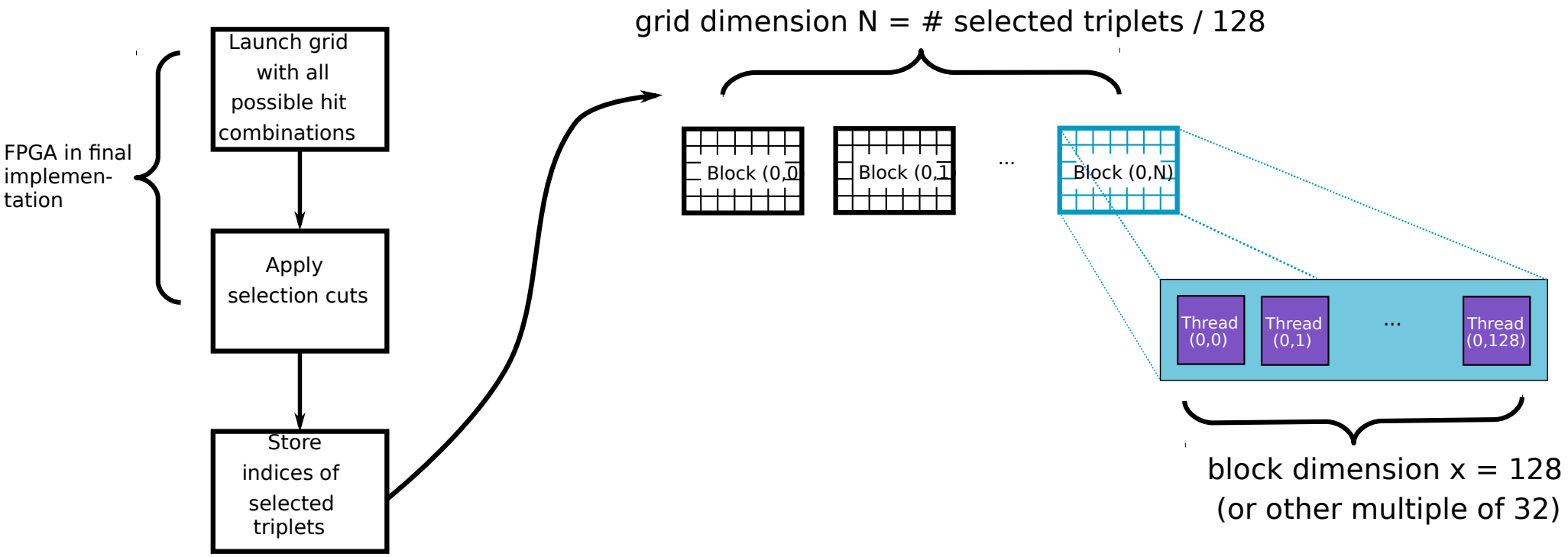
  CUDA (Compute Unified Device Architecture)

GPU Architecture



Specs for GTX680

Fitting Kernel

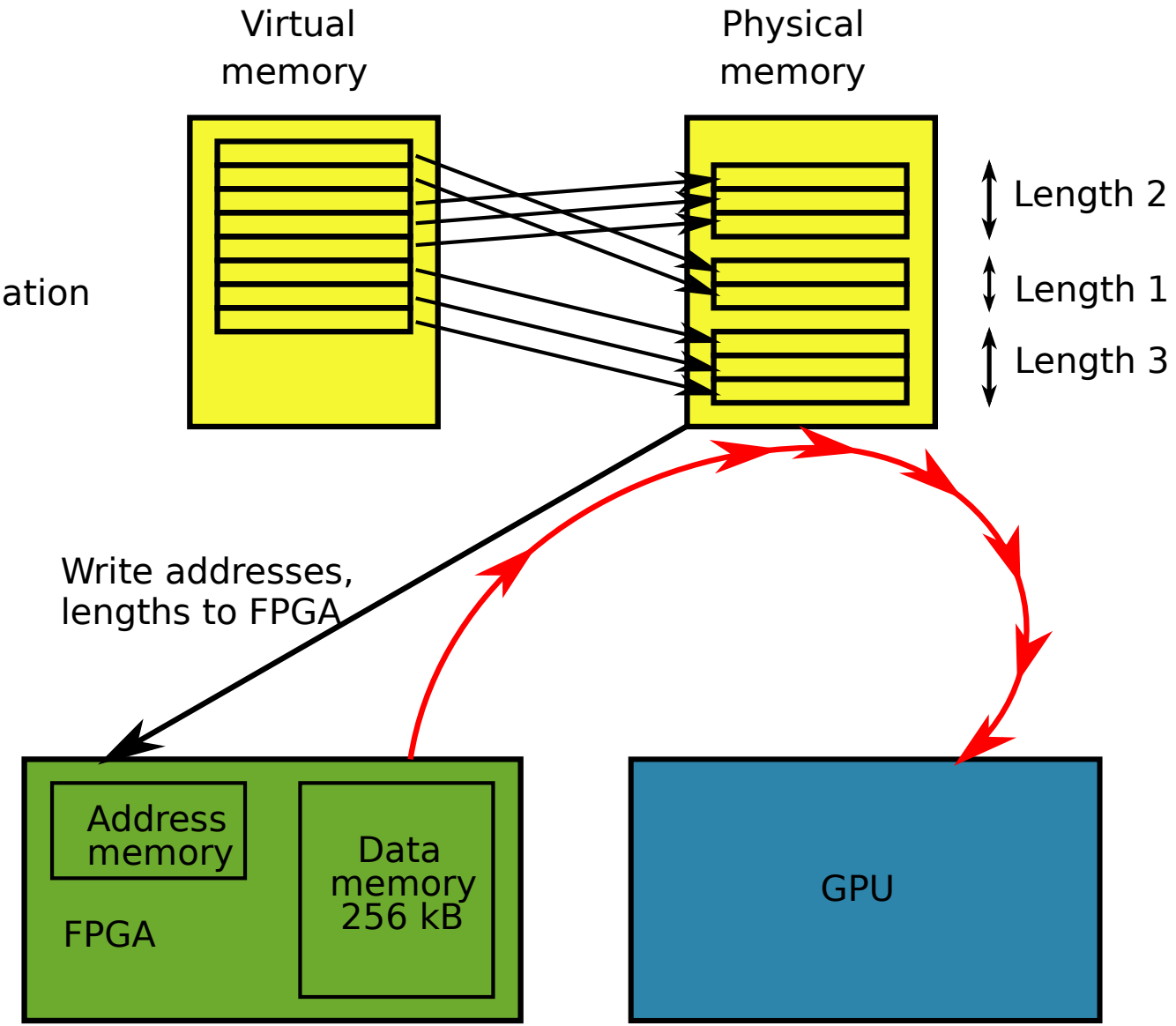


DMA: Implementation

- Stratix V / IV development board: DMA engine, PCIe interface
- Kernel module for communication with FPGA
  - Mapping of memory addresses
  - Read, write functions
  - Interrupt handling
- CUDA API: memory allocation of page-locked memory, usable for DMA from FPGA to RAM and from RAM to GPU memory
- Use DMA with scatter / gather mapping
  - Large (GB) memory buffers possible

DMA: Implementation

CUDA API:
memory allocation



Segmentation, Interrupt messages