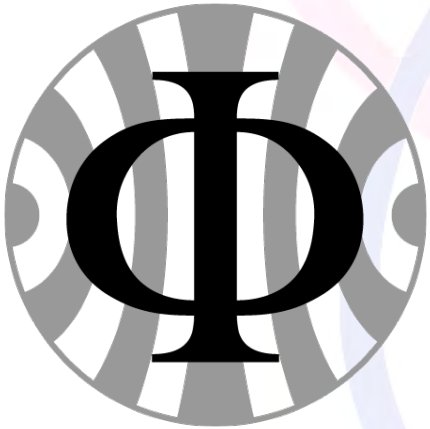
Track Reconstruction on GPUs for the Mu3e Experiment

Dorothea vom Bruch
for the Mu3e Collaboration

DPG Frühjahrstagung, T41: Detektoren und DAQ 1

March 10, 2015

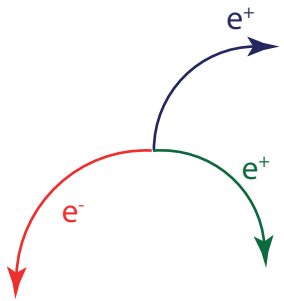


Physikalisches Institut Heidelberg

The Mu3e Experiment

Mu3e searches for the charged lepton-flavour violating decay $\mu^+ \to e^+e^+e^-$ with a sensitivity better than 10^{-16}

μ decays at rest $\to E_e < 53\,\mathrm{MeV}$



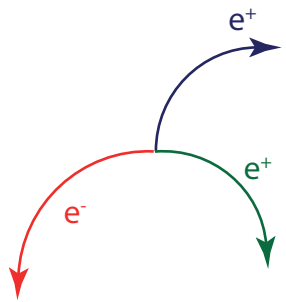
Signal

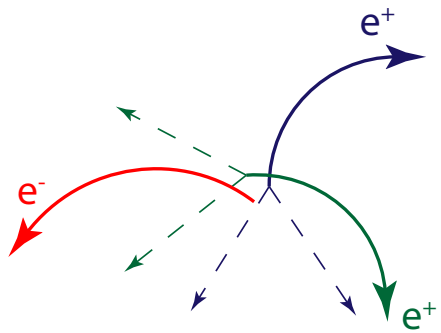
- Coincident in time
- Single vertex
- $\Sigma \vec{p_i} = 0$
- $E = m_\mu$

The Mu3e Experiment

Mu3e searches for the charged lepton-flavour violating decay $\mu^+ \to e^+e^+e^-$ with a sensitivity better than 10^{-16}

μ decays at rest $\to E_e < 53\,\mathrm{MeV}$





Signal

- Coincident in time
- Single vertex
- $\Sigma\vec{p_i} = 0$
- $E = m_\mu$

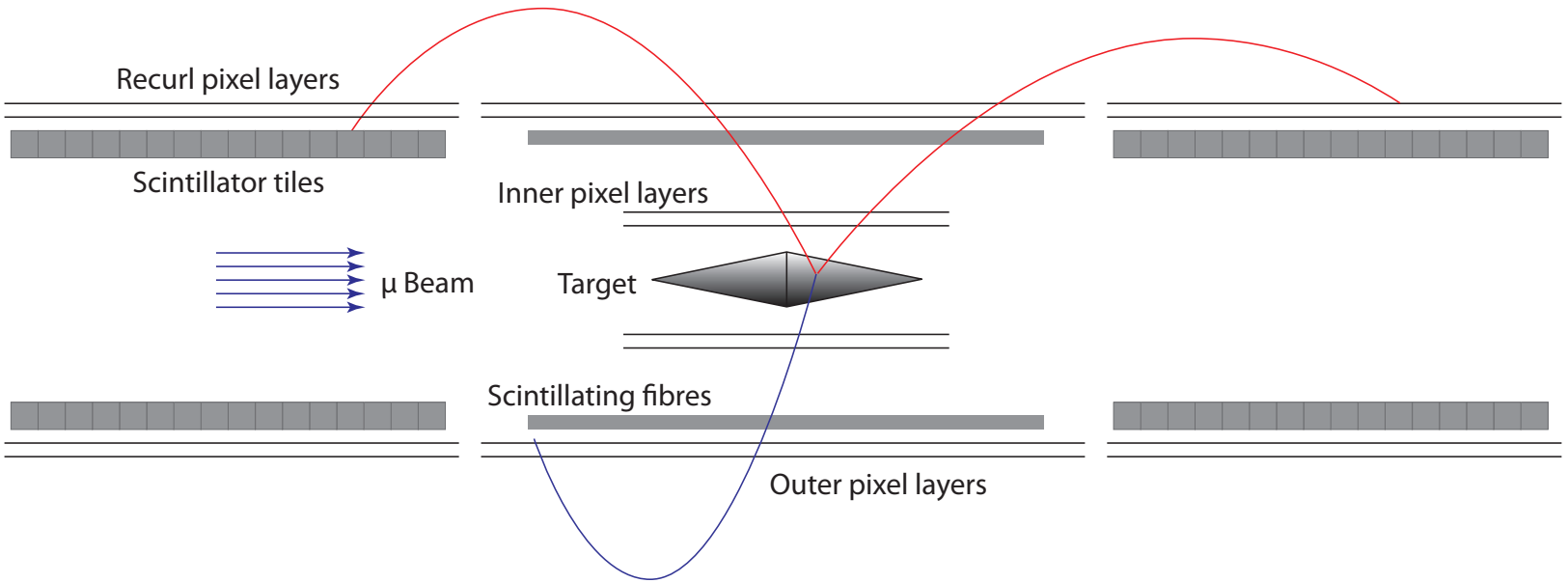
Random Combinations

- Not coincident in time
- No single vertex
- $\Sigma\vec{p_i} \neq 0$
- $E \neq m_\mu$

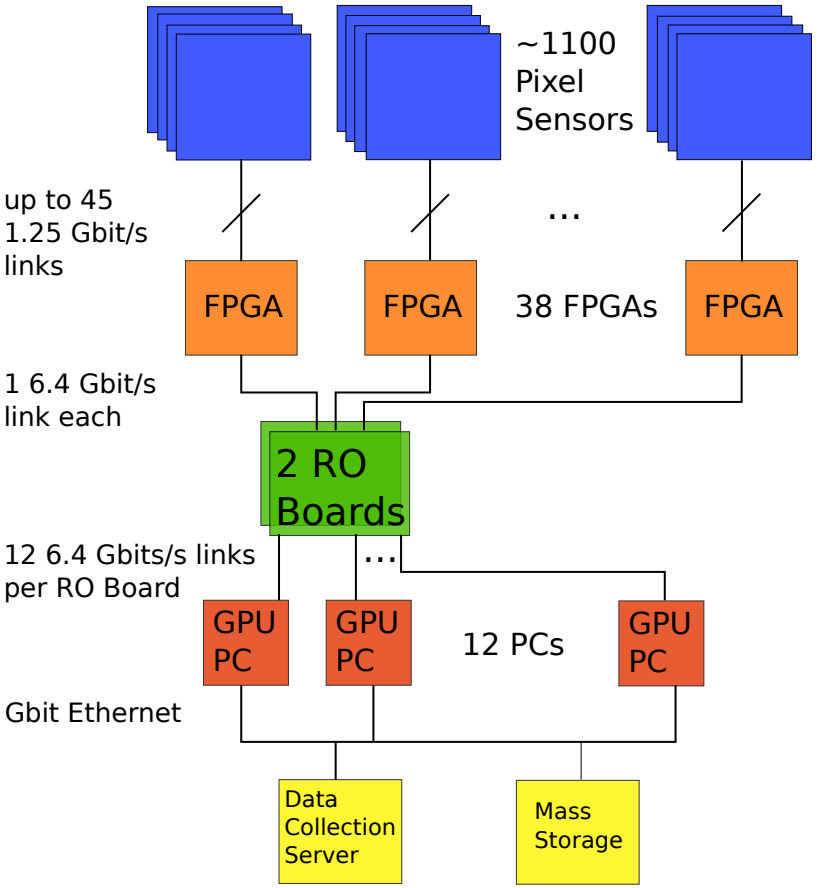
The Mu3e Detector

Requirements:

- Excellent momentum resolution: $< 0.5\,\mathrm{MeV/c}$
- Good timing resolution: 100 ps for tiles, 1 ns for fibres, $< 20\,\mathrm{ns}$ for pixels
- Good vertex resolution: $300\,\mu\mathrm{m}$
- High rates $\mathcal{O}(10^8 - 10^9 \mu/s)$



Readout Scheme



- Triggerless readout $\rightarrow$ 1 Tbit/s data rate
- Online data reduction required
- Track reconstruction and vertex fitting on GPUs
- $\Rightarrow$ Reduction factor of $\sim$ 1000

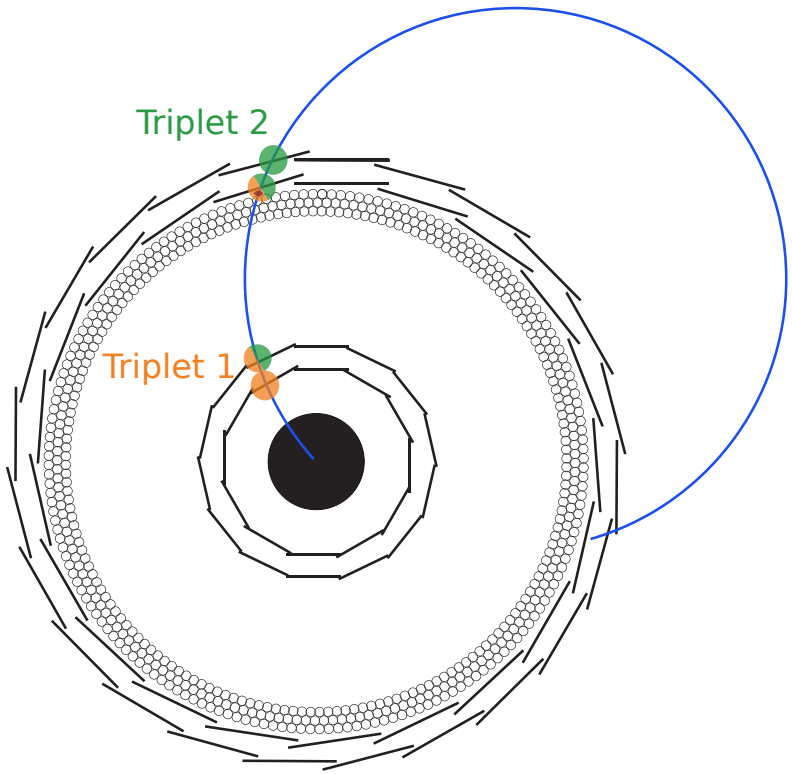
Multiple Scattering Fit

Low Momentum

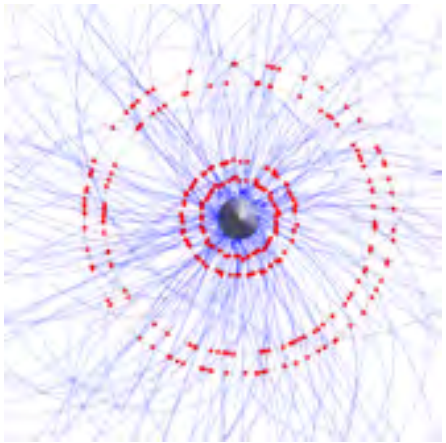
Electrons: 15 - 53 MeV
Resolution dominated by multiple Coulomb scattering

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

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



Fit on the GPU

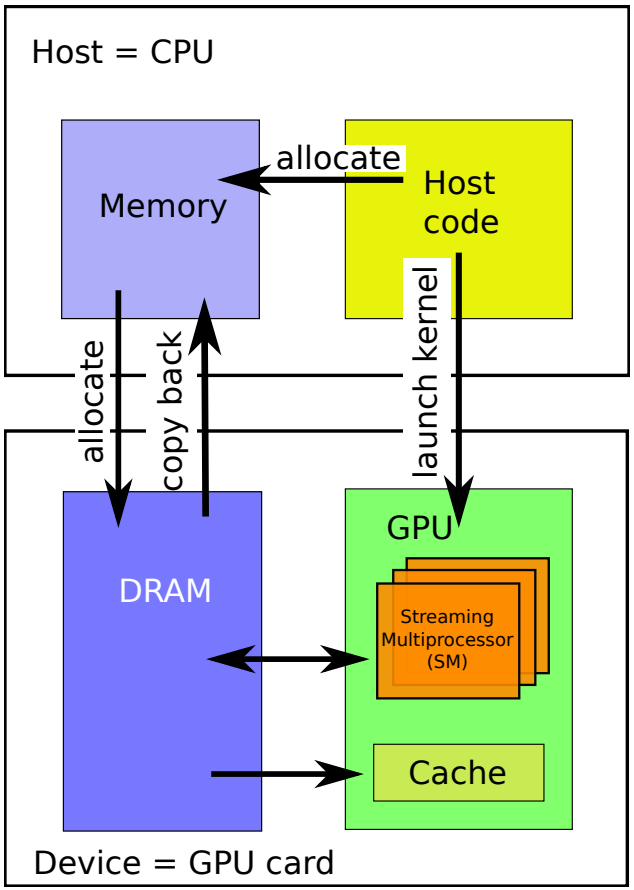


- Number of possible track candidates $\sim n^3$
- With n: # hits per layer
- On GPU: Loop over all possible combinations
  - Geometrical selection cuts
  - Triplet fit
  - Vertex fit

$\Rightarrow$ Reduction factor of $\sim$ 1000

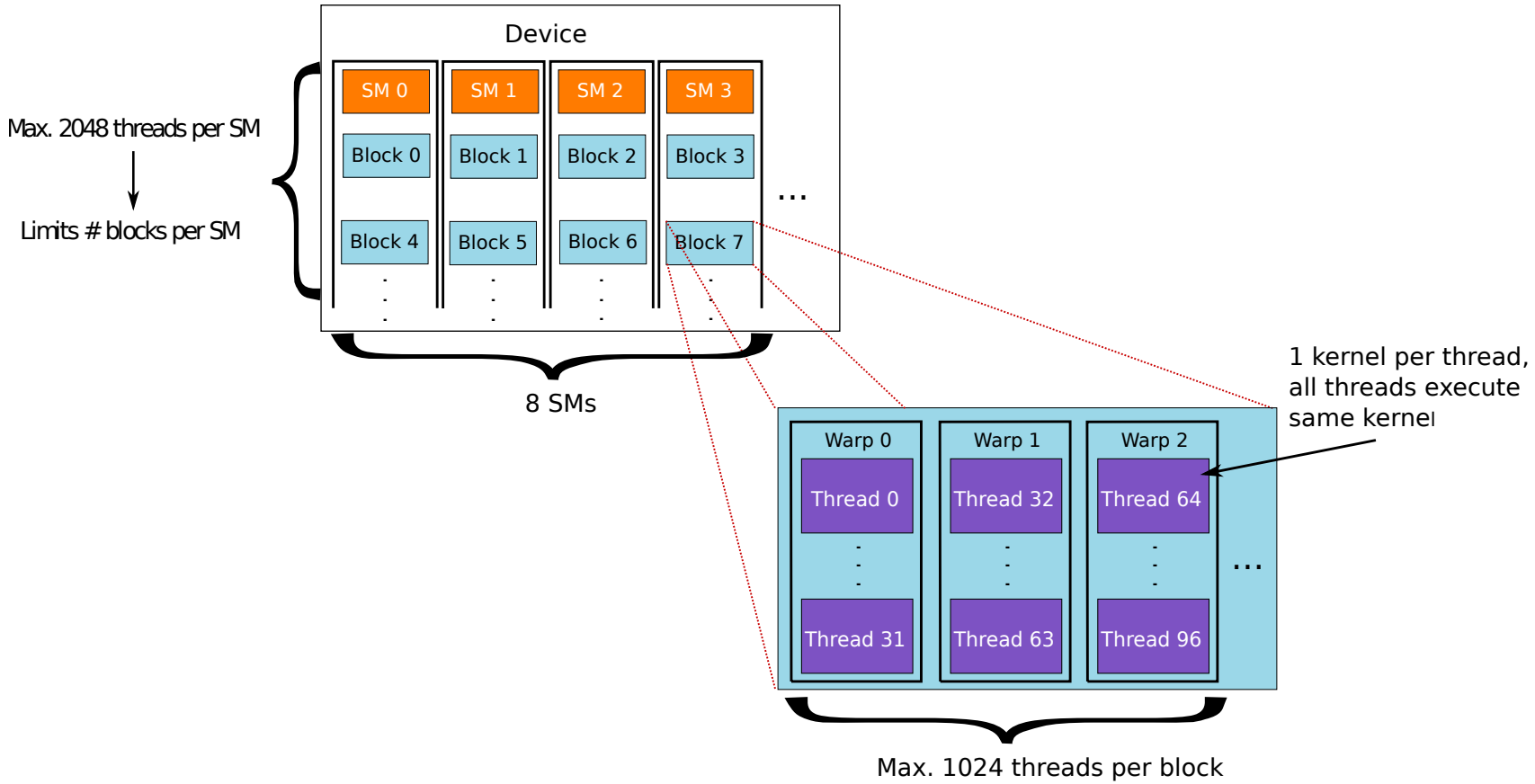
Image source: http://www.pcmag.com/article2/0,2817,2401953,00.asp

GPU Properties

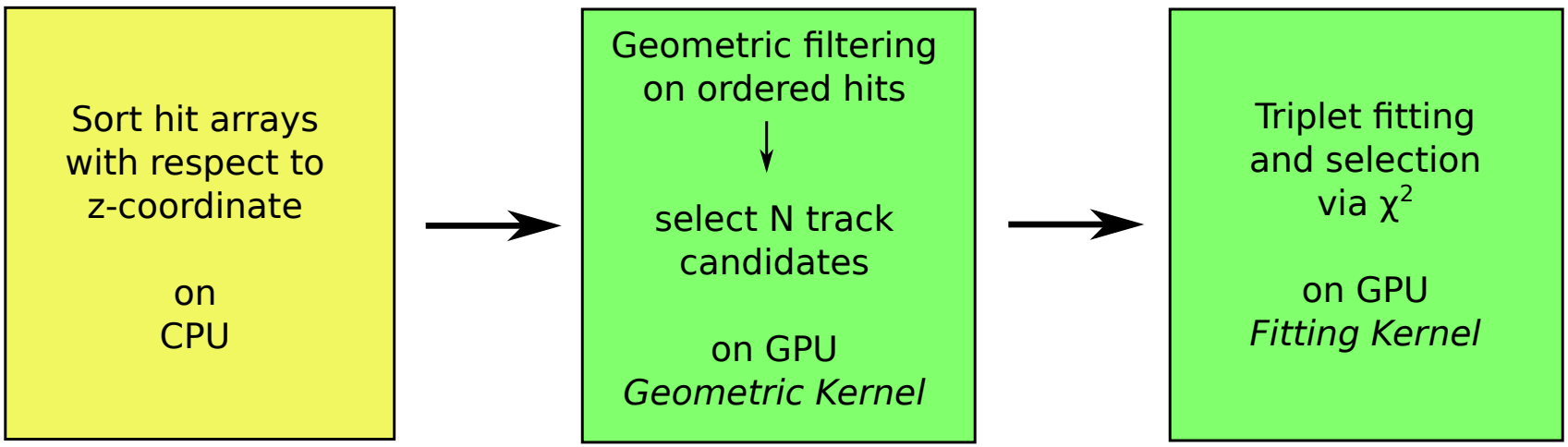


- Highly parallel structure
- Process large blocks of data
- Nvidia: API extension to C: CUDA (Compute Unified Device Architecture)

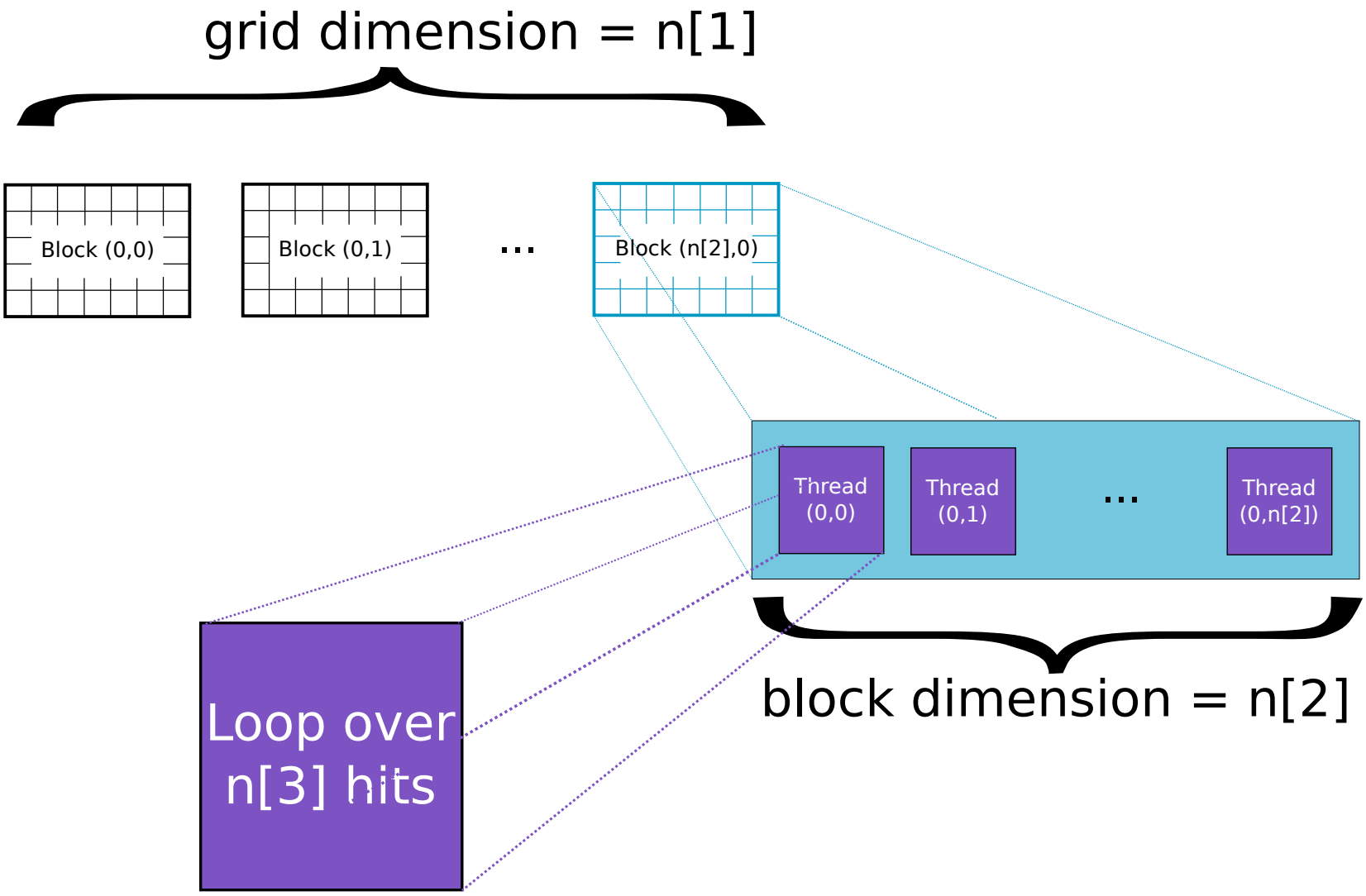
GPU Architecture (GTX 680 as example)



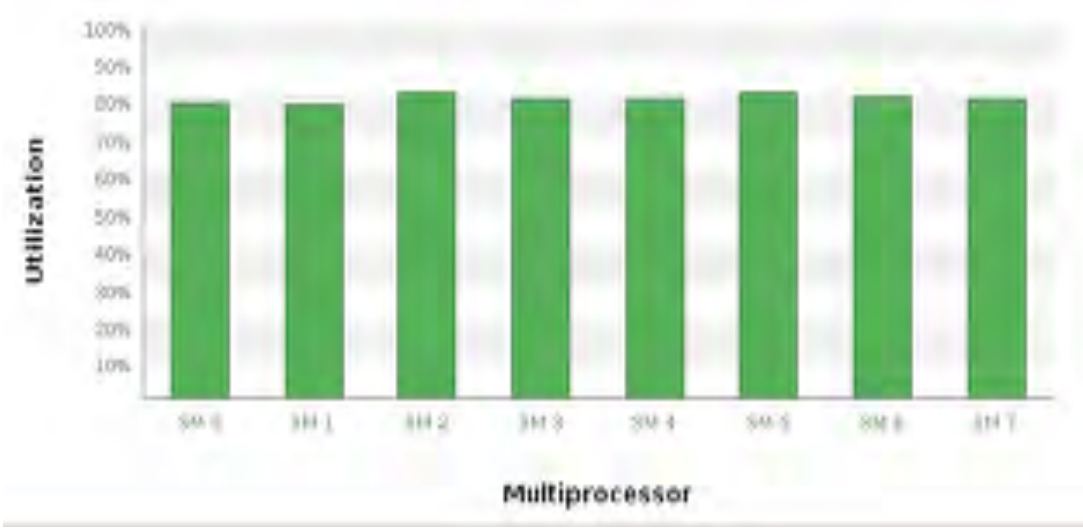
Main Steps

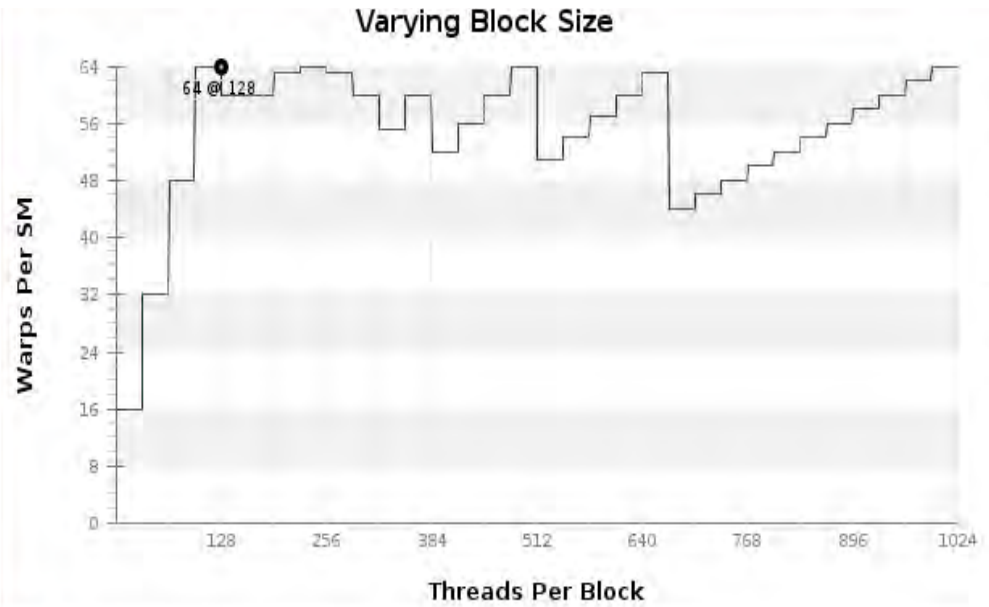

Sort hit arrays
with respect to
z-coordinate
on
CPU
Geometric filtering
on ordered hits
select N track
candidates
on GPU
Geometric Kernel
Triplet fitting
and selection
via χ^2
on GPU
Fitting Kernel

Grid for Geometric Kernel



Compute Utilization for Fitting Kernel





Performance

- Process 1.4×10^{10} triplets / s
- Most time spent on geometric kernel $\rightarrow$ outsource to FPGA
- Ratio of copying data from CPU to GPU to compute time: 40 %, will improve when selection cuts are applied on FPGA
- For 10^8 μ/s: 10^{12} hit combinations $\rightarrow$ 48 GPU computers
- For 10^9 μ/s: More improvements needed

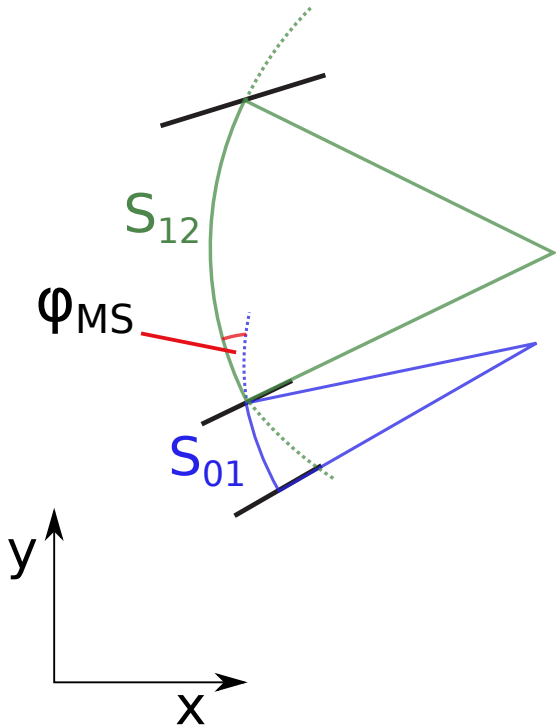
Outlook

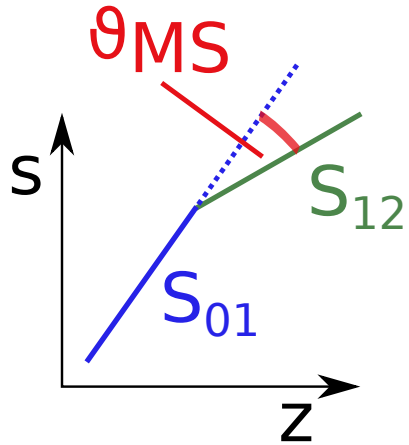
- Study data transmission from FPGA to GPU:
  - FPGA $\xrightarrow{PCIe, DMA}$ CPU $\xrightarrow{PCIe}$ GPU
  - FPGA $\xrightarrow{PCIe, DMA}$ GPU
- Outsource selection to FPGA
- Include vertex fit

Backup Slides

Multiple Scattering at middle hit of triplet

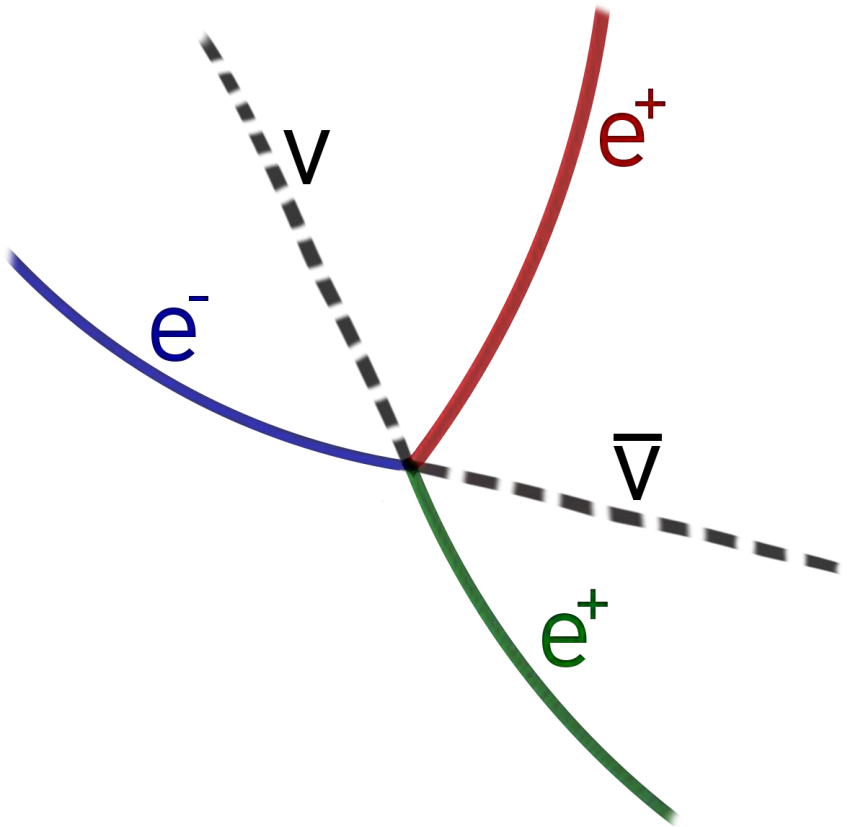
Multiple scattering at middle hit of triplet





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

Irreducible Background

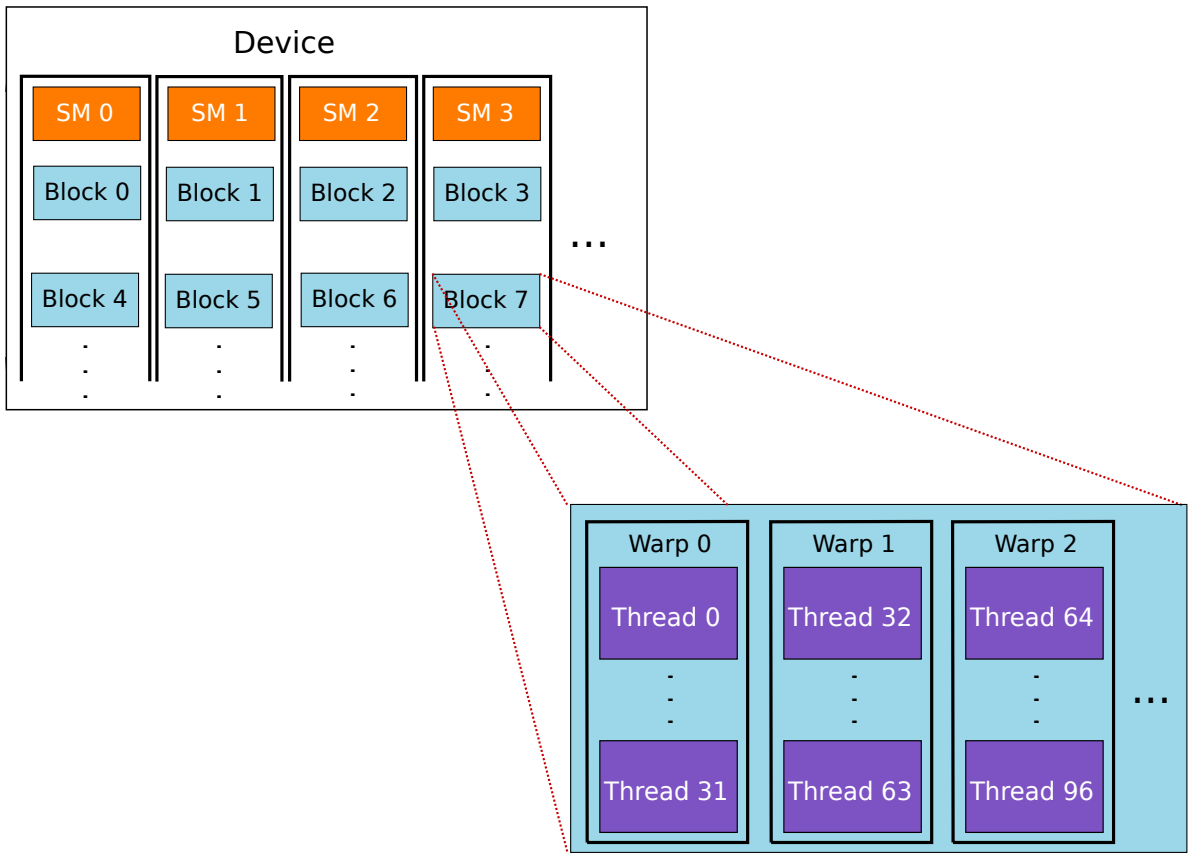


Phases of Mu3e

Phase 1	Phase 2
$\mathcal{O}10^8 \mu/s$	$\mathcal{O}10^9 \mu/s$
Central part	Central part
+	+
1 recurl station	2 recurls stations

Hardware Implementation: Warps

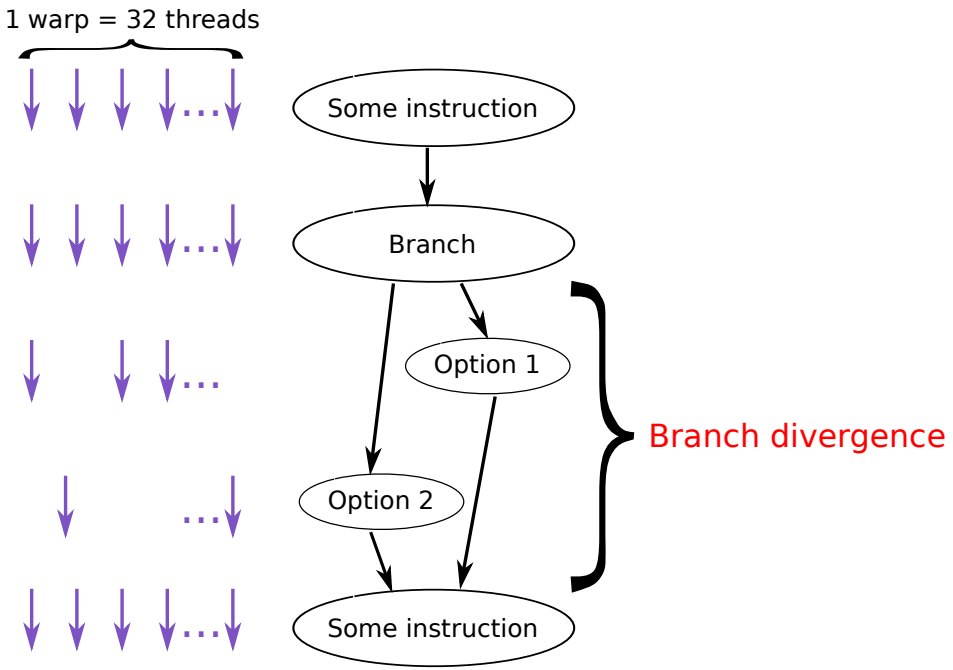
- After block is assigned to SM
  $\rightarrow$ Division into units called warps
- On GTX 680: 1 warp = 32 threads

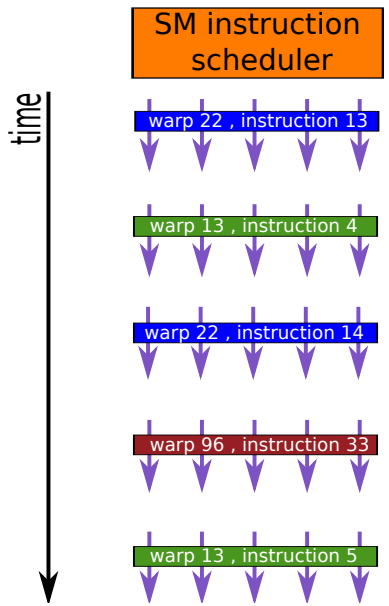


Warp Scheduling

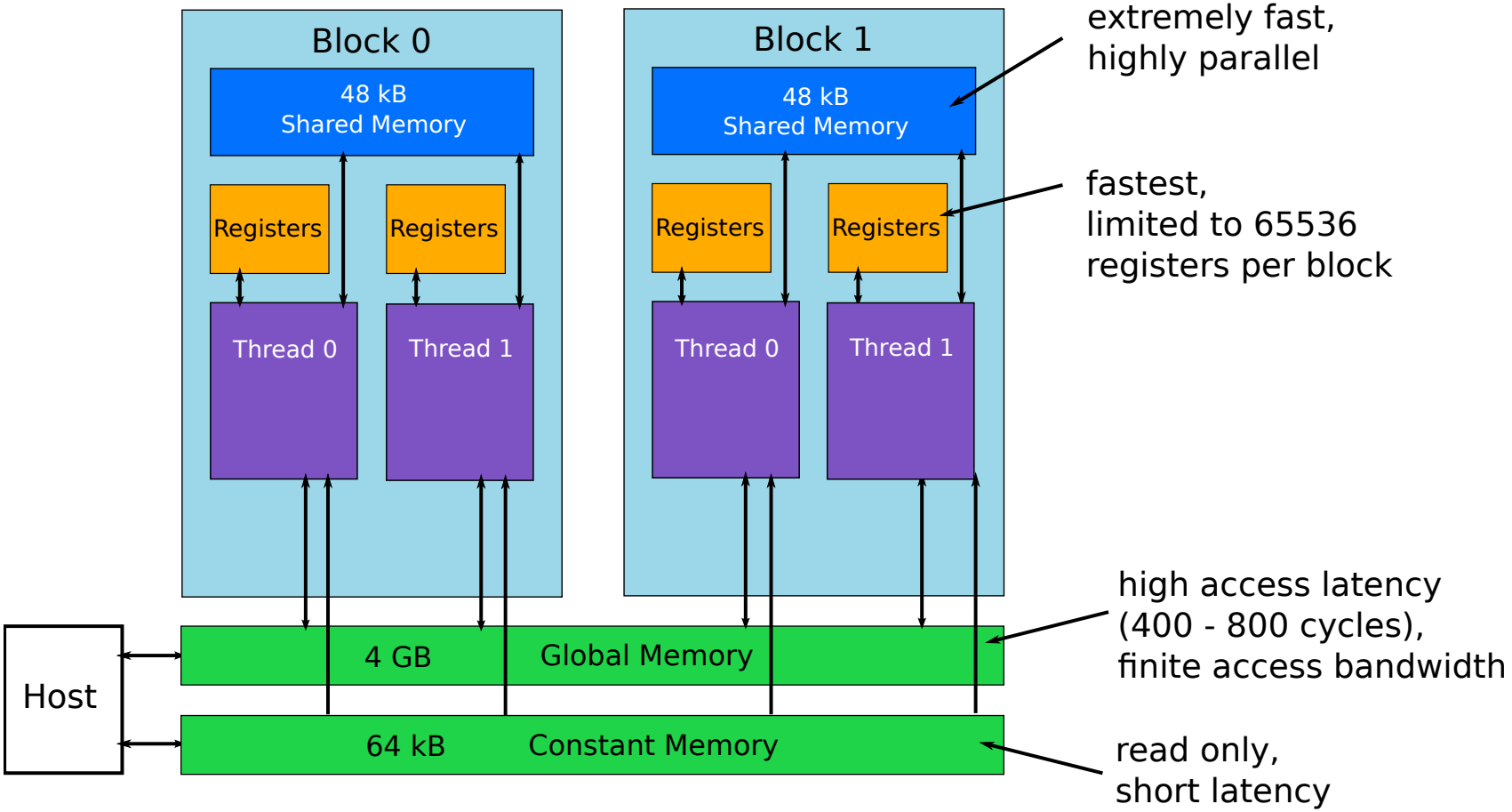
Warps execute

In SIMD fashion (Single Instruction, Multiple Data)
Not ordered

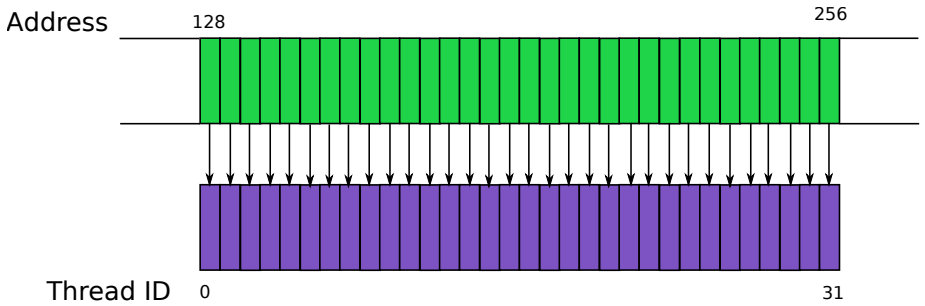




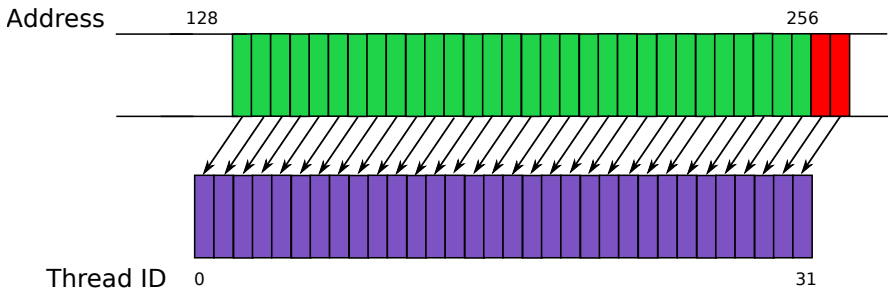
GPU Memory


Block 0
48 kB
Shared Memory
Registers
Registers
Thread 0
Thread 1
Block 1
48 kB
Shared Memory
Registers
Registers
Thread 0
Thread 1
extremely fast,
highly parallel
fastest,
limited to 65536
registers per block
Host
4 GB
Global Memory
64 kB
Constant Memory
high access latency
(400 - 800 cycles),
finite access bandwidth
read only,
short latency

Memory Access



Coalesced memory access



Non-coalesced memory access

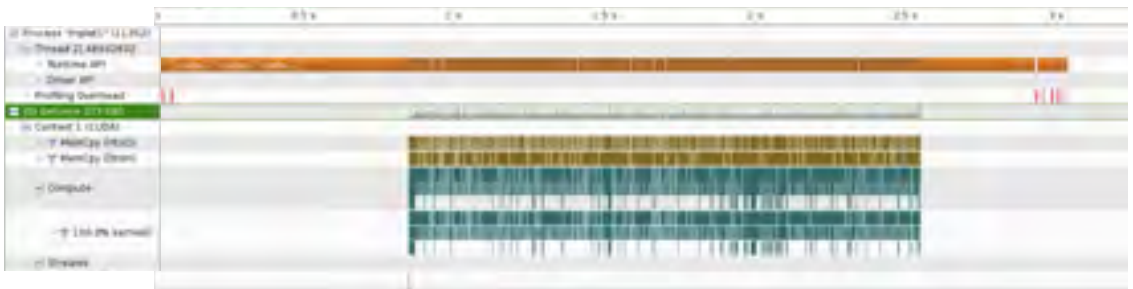
Warp Memory Access

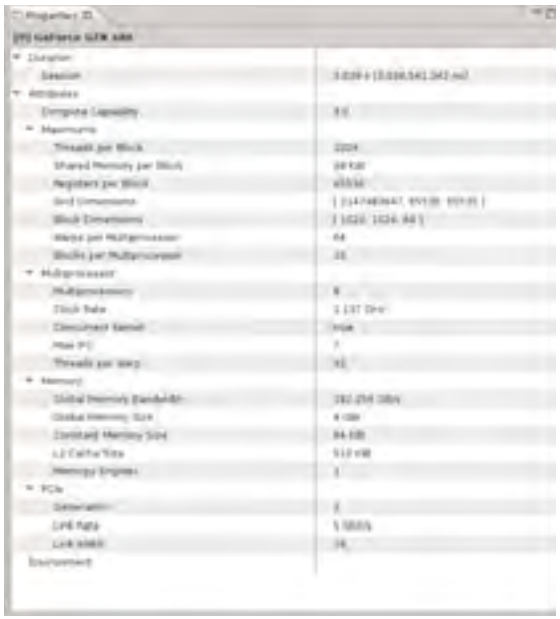
128 bytes in single transaction

Diagnostic Tools (One Kernel)

Nvidia provides several diagnostic tools:

- Profiler for terminal usage: Time spent by CPU and GPU
- Memory check: Memory allocation errors, misuse, ...
- Visual profiler: Diagnostics for performance of GPU code





Diagnostic Tools (One Kernel)

Kernel profile: Cuda code and machine code