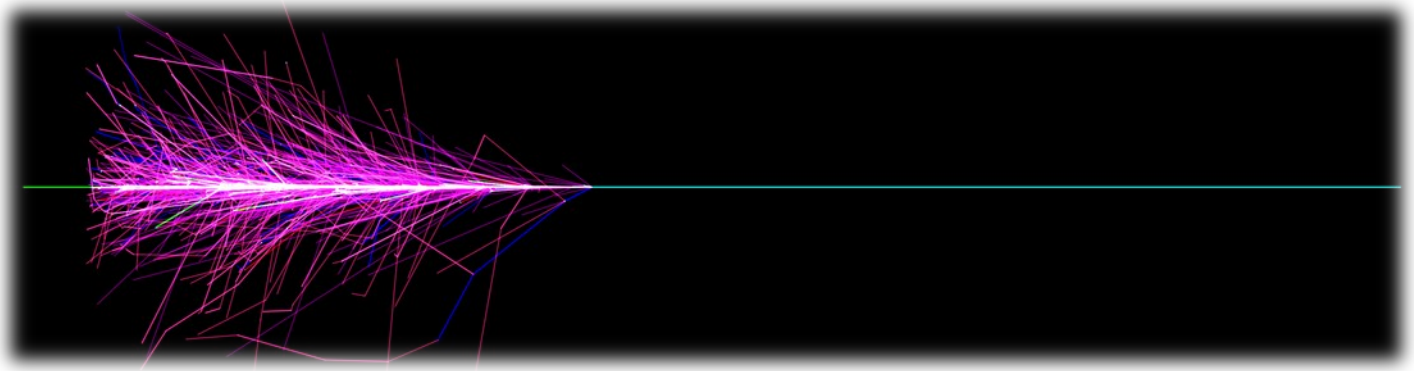
technische universität dortmund

SFB 876 Providing Information by Resource-Constrained Data Analysis

lehrstuhl physik e5



GPU based photon propagation for CORSIKA 8

Dominik Baack

Corsika Telcon

Online

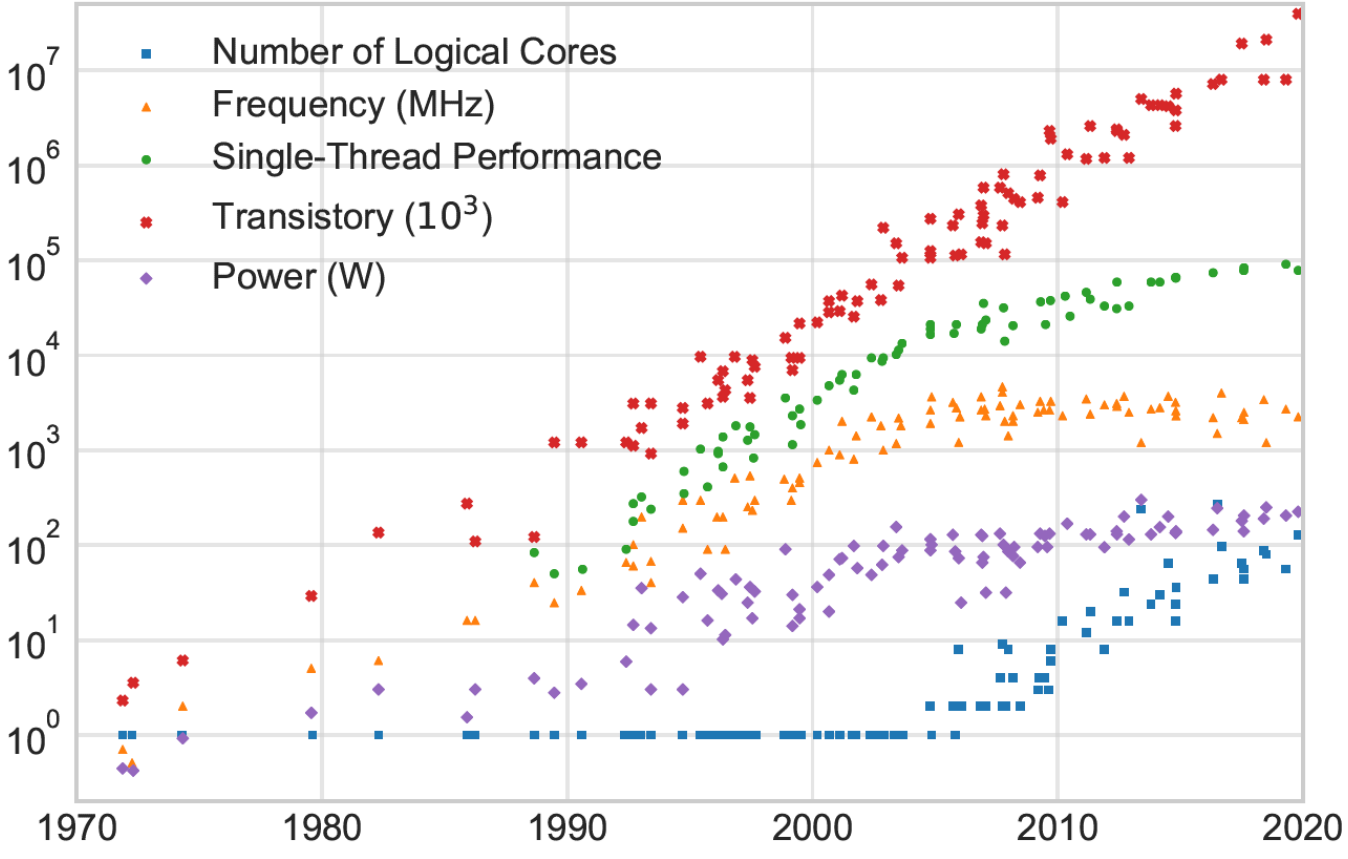
03. 12. 2020

Function	CPU Time: Total ▼	CPU Time: Self	Instructions Retired: Total
aamain	2232.101s	0.020s	100.0%
__libc_start_main	2232.101s	0s	100.0%
main	2232.101s	0s	100.0%
_start	2232.101s	0s	100.0%
box3	2208.159s	0.020s	99.0%
cerenk	2078.871s	533.937s	93.8%
egs4	2026.466s	0s	90.8%
em	2026.466s	0s	90.8%
shower	2026.466s	0.070s	90.8%
electr	2023.865s	34.329s	90.7%
distip	449.052s	203.941s	16.8%
__ieee754_acos_sse2	358.305s	318.770s	21.0%
tofip	336.495s	214.020s	15.5%
rmmard	250.622s		
__cos_avx	192.107s		
rhof	190.668s		
updatc	187.214s		
__GI___exp	180.922s		
update	180.016s		
__ieee754_exp_avx	166.768s		
mutrac	165.054s		
do_sincos_1	110.374s		
__sin_avx	104.359s		
telout_	55.212s		
do_sincos_1	47.284s		
__doasin	37.784s		
do_cos	32.290s		

Callees	CPU Time: Total ▼	CPU Time: Self
▼ cerenk	2078.871s	533.937s
▶ distip	449.052s	203.941s
▶ tofip	336.495s	214.020s
▶ rmmard	231.852s	231.438s
▶ rhof	190.258s	20.957s
▶ __cos_avx	179.945s	38.086s
▶ __sin_avx	95.118s	25.432s
▶ telout_	55.212s	54.892s
▶ __tan_avx	2.140s	1.960s
▶ thick	1.880s	0.290s
▶ __ieee754_log_av	1.280s	1.280s

lehrstuhl physik e5

Corsika 7



https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/

High parallel Cherenkov light

- Completely independent calculations for each photon
- “Simple” arithmetic no complex processes
- OpenCL allows platform independent vectorization & parallelization, first test shows good performance

  … but …
- CUDA – Nvidia proprietary but higher performance possible

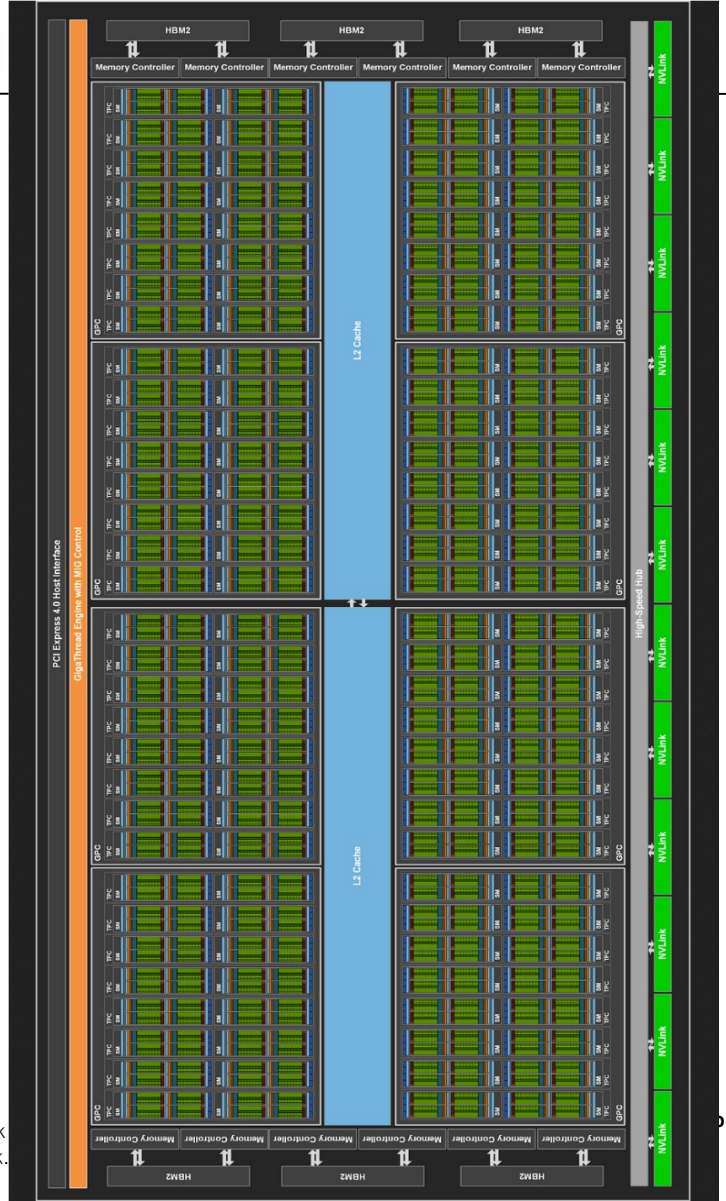


Figure 5. The GA100 streaming multiprocessor (SM).

Dominik
dominik.

on propaga

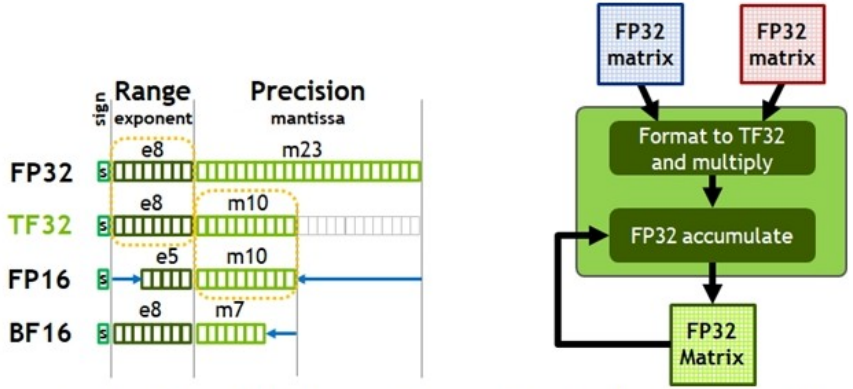


Figure 7. TensorFloat-32 (TF32) provides the range of FP32 with the precision of FP16

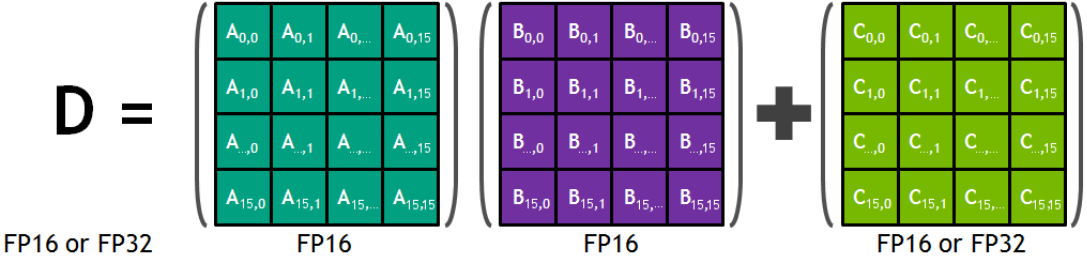


Figure 5: A warp performs D=A*B+C where A, B, C and D are 16×16 matrices. (Note the change in numbering from Figure 1: multiple Tensor Core operations are combined by the WMMA API to perform 16×16 matrix-multiply and accumulate operations.)

SYSTEM SPECIFICATIONS (PEAK PERFORMANCE)

	NVIDIA A100 for NVIDIA HGX™	NVIDIA A100 for PCIe
GPU Architecture	NVIDIA Ampere	
Double-Precision Performance	FP64: 9.7 TFLOPS FP64 Tensor Core: 19.5 TFLOPS	
Single-Precision Performance	FP32: 19.5 TFLOPS Tensor Float 32 (TF32): 156 TFLOPS | 312 TFLOPS*	
Half-Precision Performance	312 TFLOPS | 624 TFLOPS*	
Bfloat16	312 TFLOPS | 624 TFLOPS*	
Integer Performance	INT8: 624 TOPS | 1,248 TOPS* INT4: 1,248 TOPS | 2,496 TOPS*	
GPU Memory	40 GB HBM2	
Memory Bandwidth	1.6 TB/sec	
Error-Correcting Code	Yes	
Interconnect Interface	PCIe Gen4: 64 GB/sec Third generation NVIDIA® NVLink®: 600 GB/sec**	PCIe Gen4: 64 GB/sec Third generation NVIDIA® NVLink®: 600 GB/sec**
Form Factor	4/8 SXM GPUs in NVIDIA HGX™ A100	PCIe
Multi-Instance GPU (MIG)	Up to 7 GPU instances	
Max Power Consumption	400 W	250 W
Delivered Performance for Top Apps	100%	90%
Thermal Solution	Passive	
Compute APIs	CUDA®, DirectCompute, OpenCL™, OpenACC®	

* Structural sparsity enabled

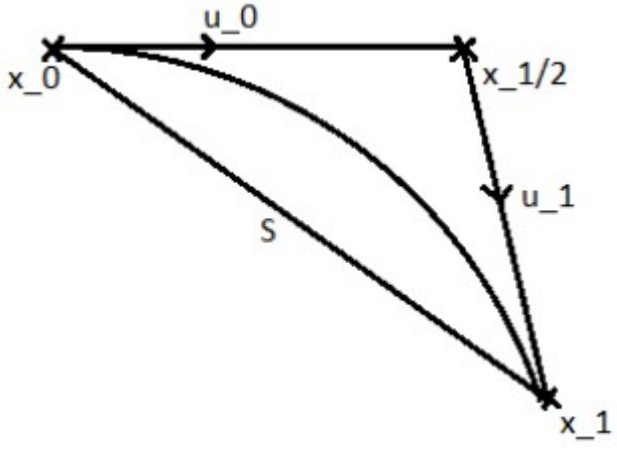
** SXM GPUs via HGX A100 server boards; PCIe GPUs via NVLink Bridge for up to 2 GPUs

https://developer.nvidia.com/blog/nvidia-ampere-architecture-in-depth/

Dominik Baack
dominik.baack@tu-dortmund.de

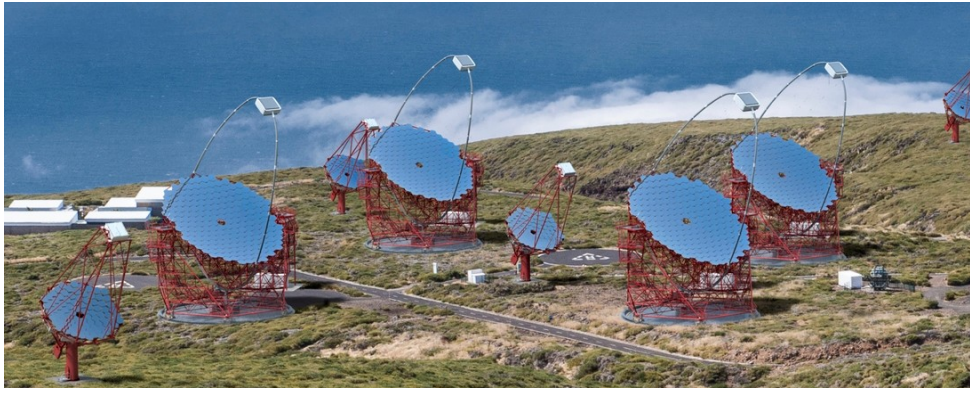
Prerequisite

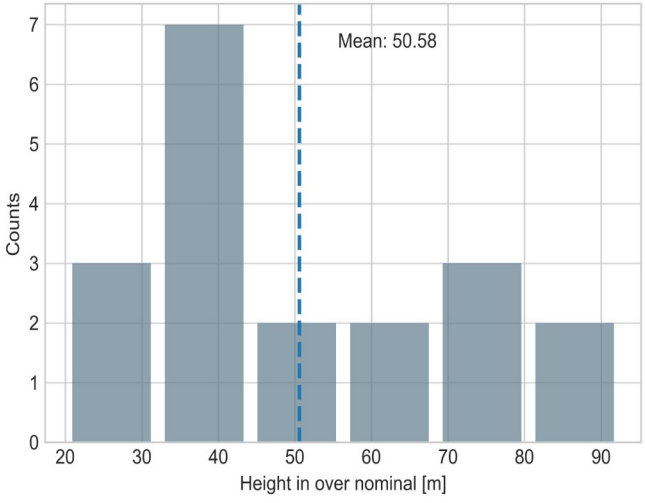
- Linear particle tracks
- >>1024 track segments accumulated to reach good performance

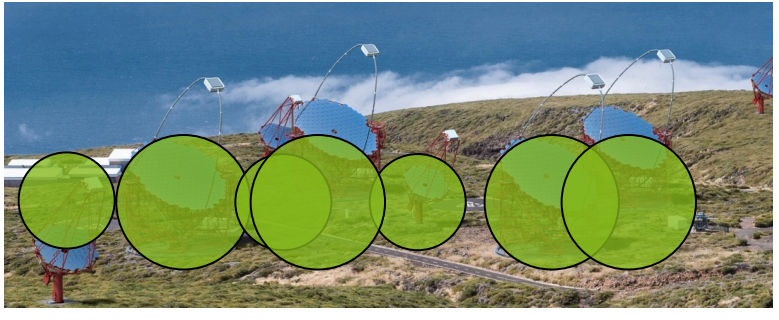


Andre Schmidt - KIT

Geometry

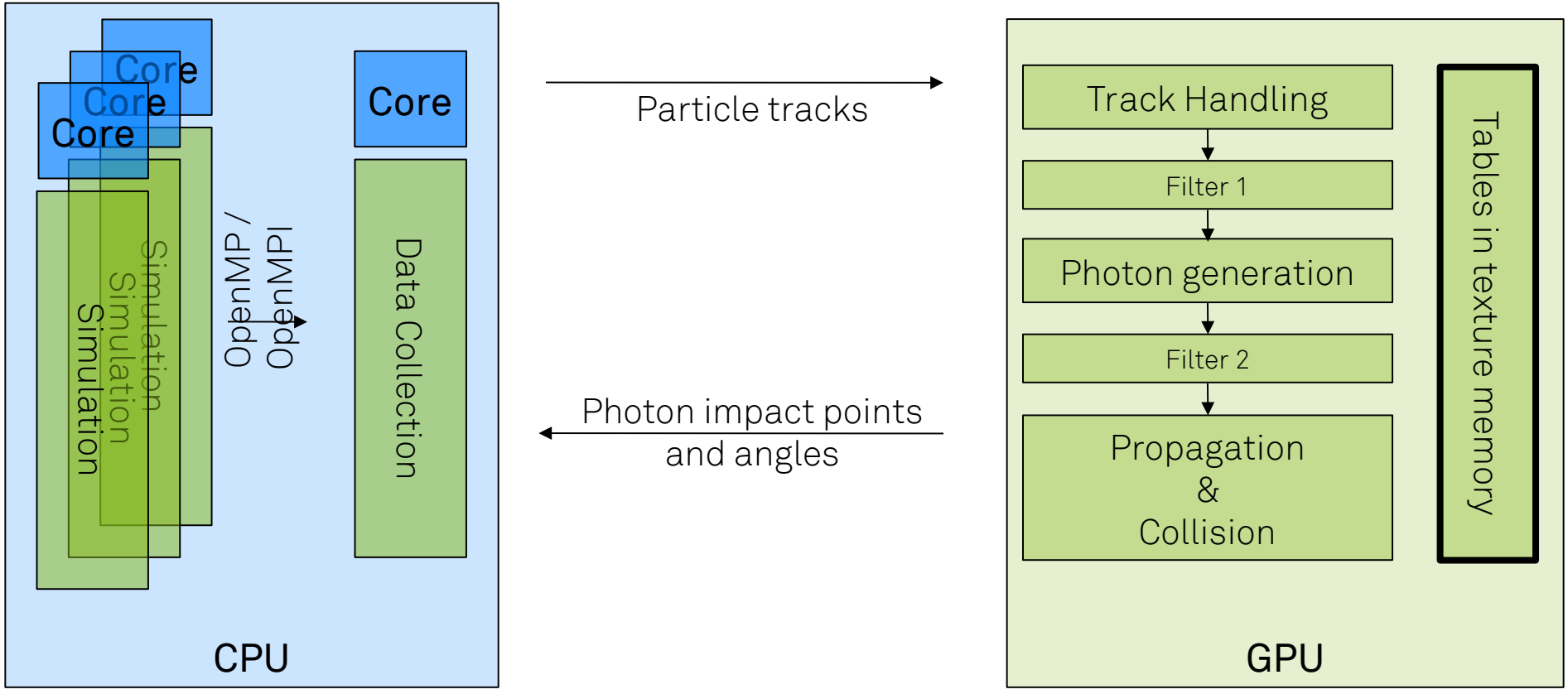




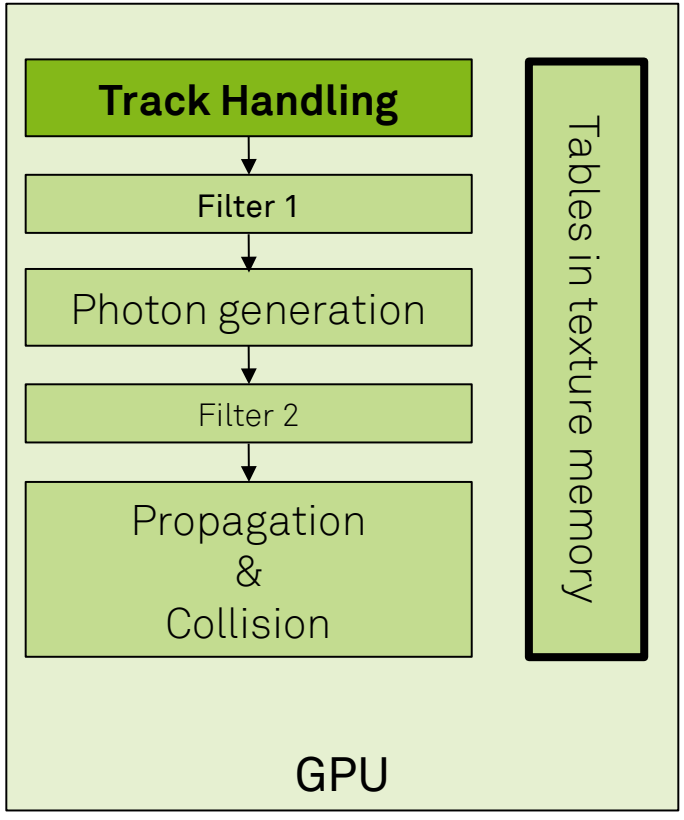


CTA Observatory - https://www.cta-observatory.org/about/how-cta-works/optimized-cta_orm_comp_webupdated_1800x590/

Computing Structure



Computing Structure

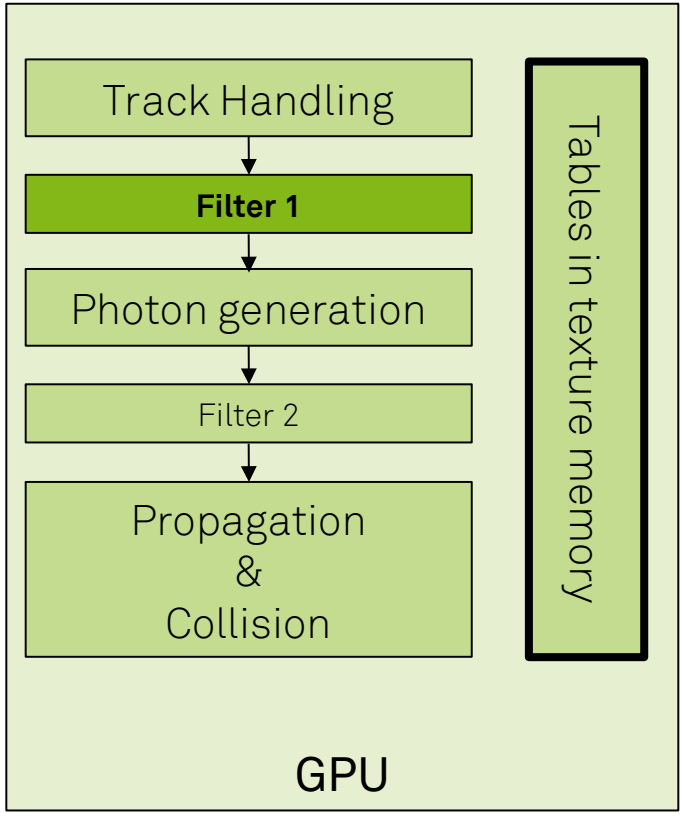


- All Tracks to GPU
  - More work to GPU
  - Less data transfer
- Load, convert and distribute tracks to warps
- Split tracks if necessary partial

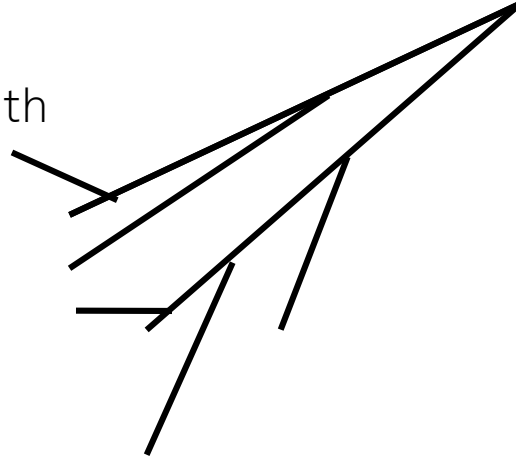
$$\beta_{start} < 1 < \beta_{end}$$
$$\beta_{end} < 1 < \beta_{start}$$

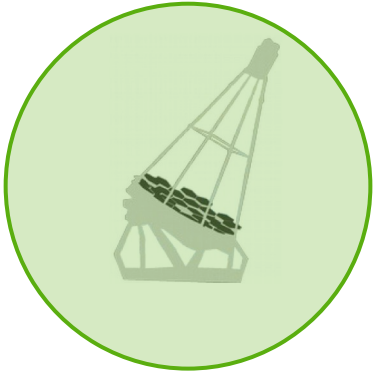
- Remove any tracks with $\beta < 1$

Computing Structure

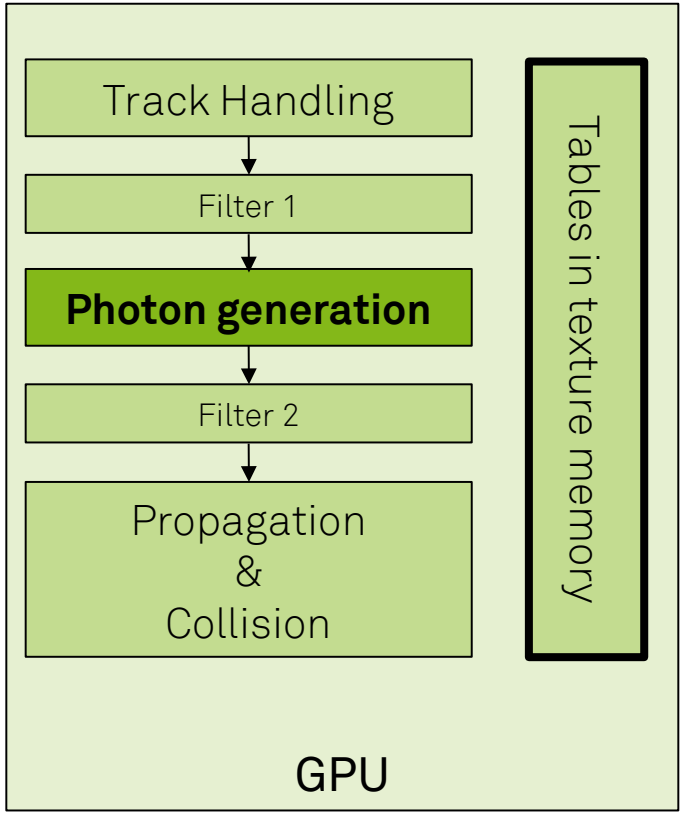


- Remove particle $\vec{v} \cdot \vec{p} \leqslant \text{cut}$
- Angular cut not possible with
  - diffuse Emission
  - Fluorescence





Computing Structure

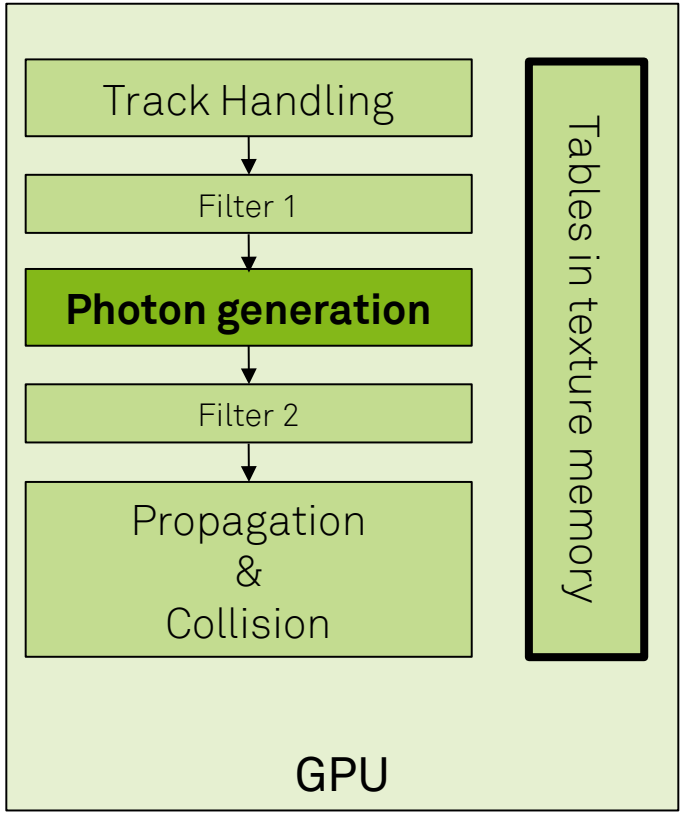


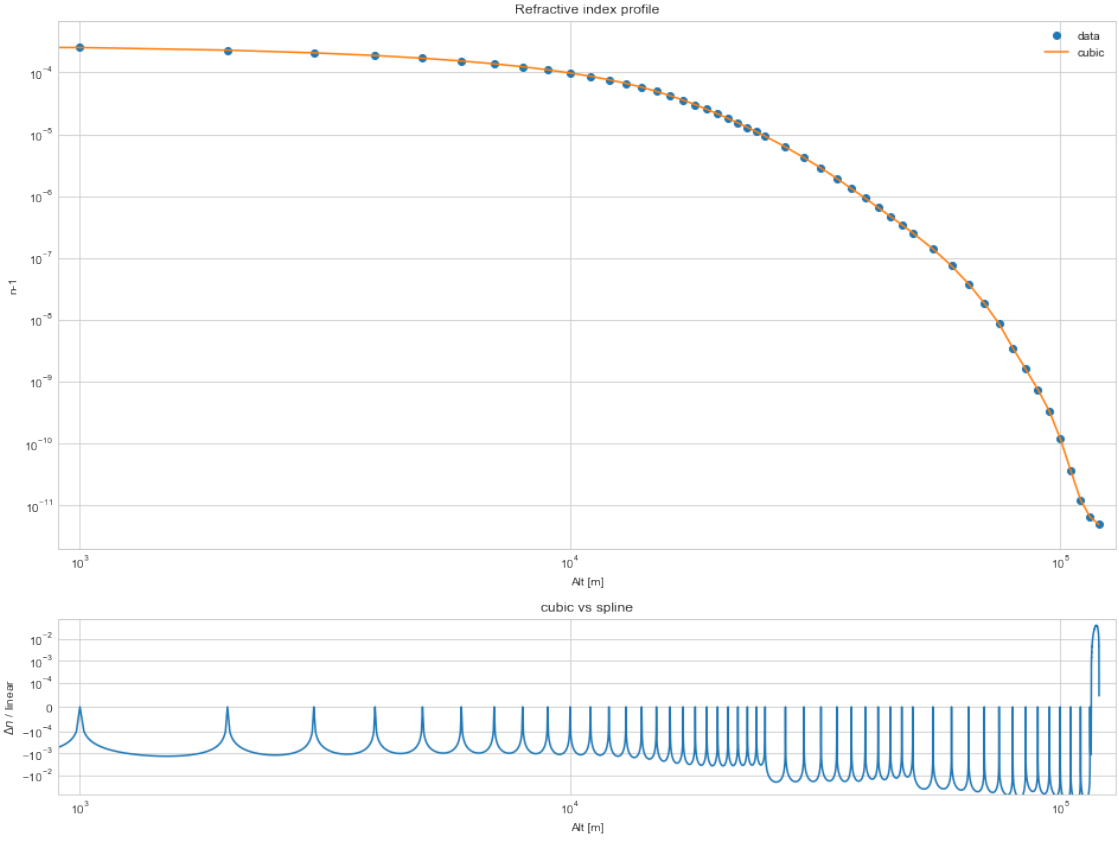
- Use presolved Frank-Tamm Formula

$$\frac{d^2N}{dxd\lambda} = \frac{2\pi\alpha z^2}{\lambda^2} \cdot sin^2(\theta_c)$$

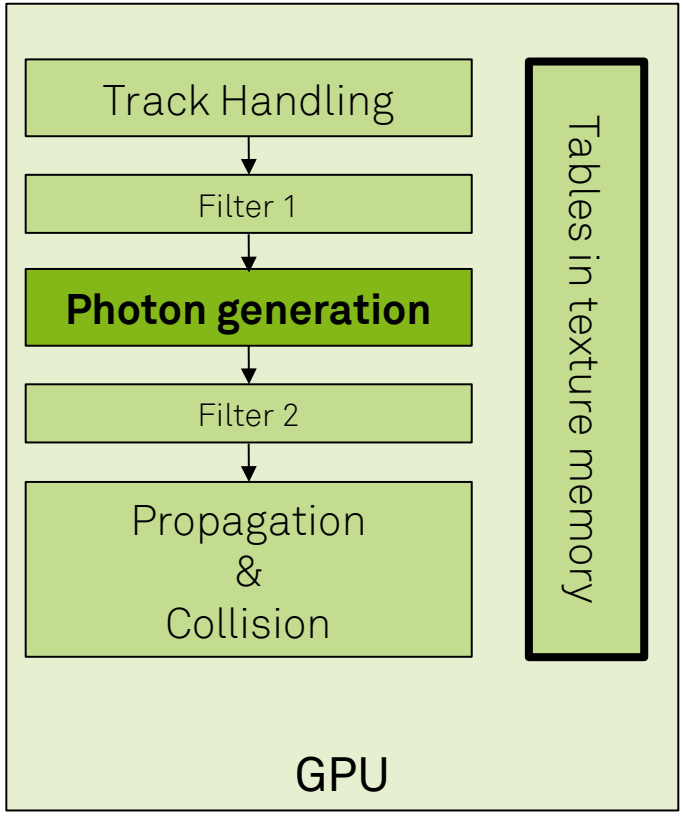
to calculate N and α

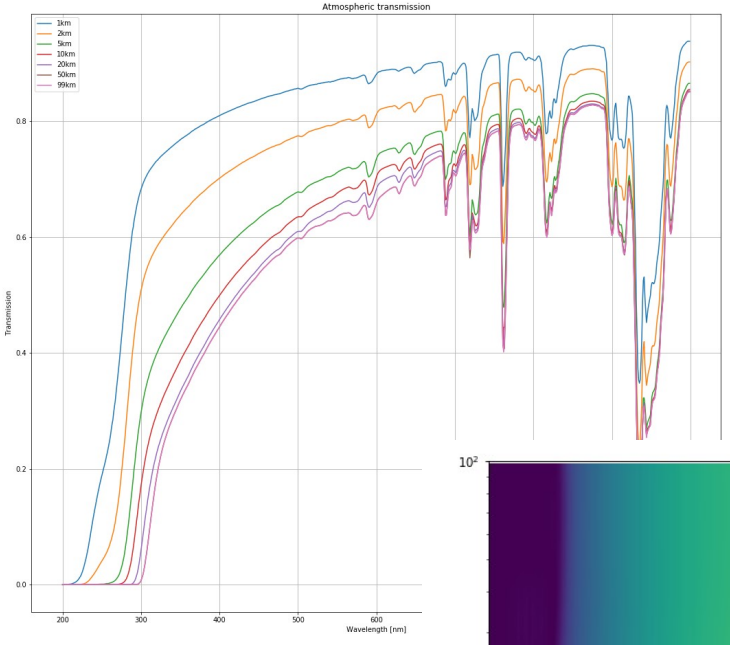
Computing Structure

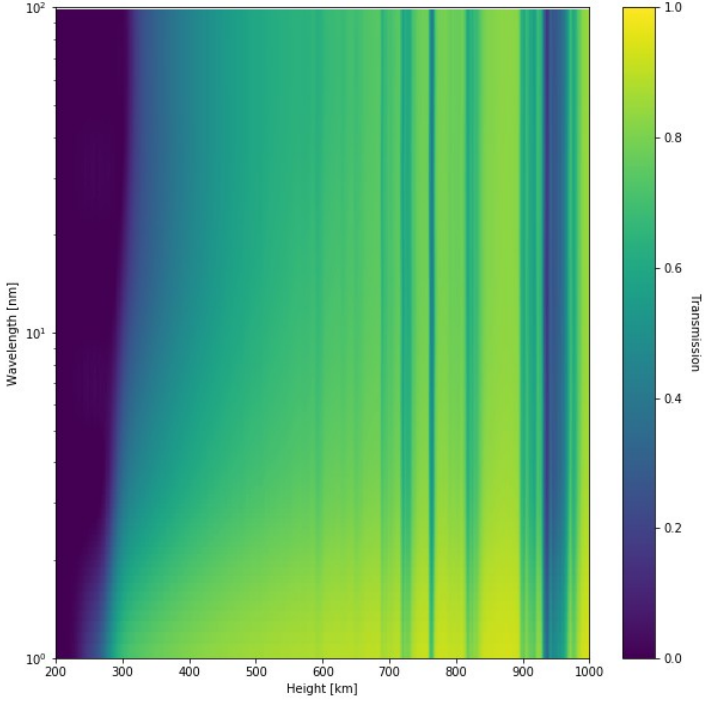




Computing Structure



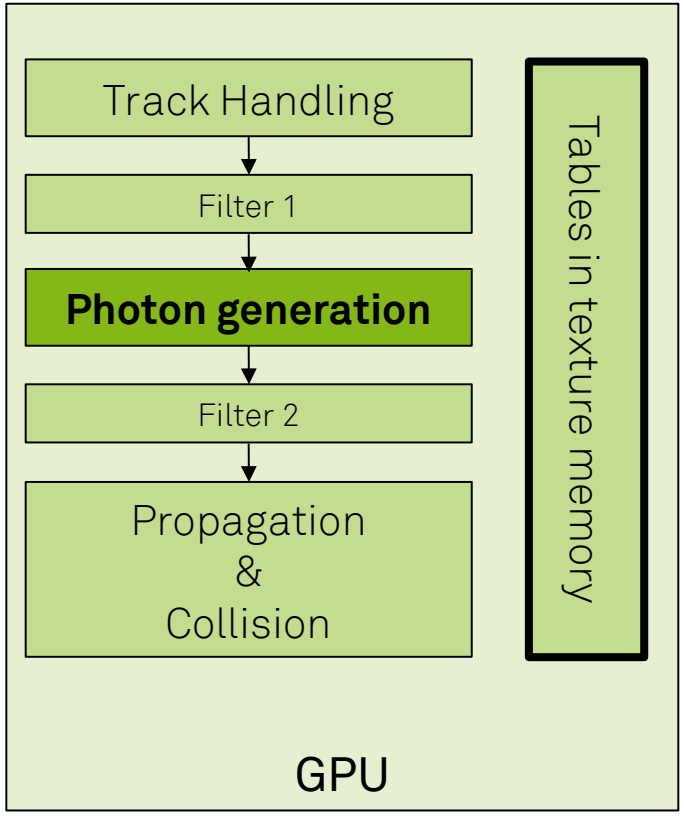


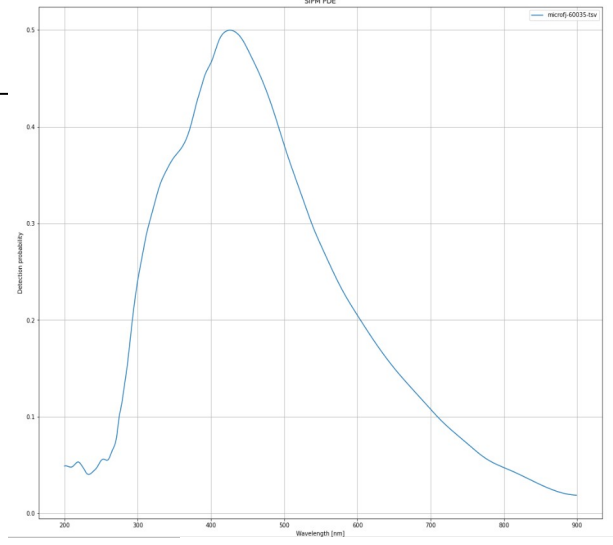


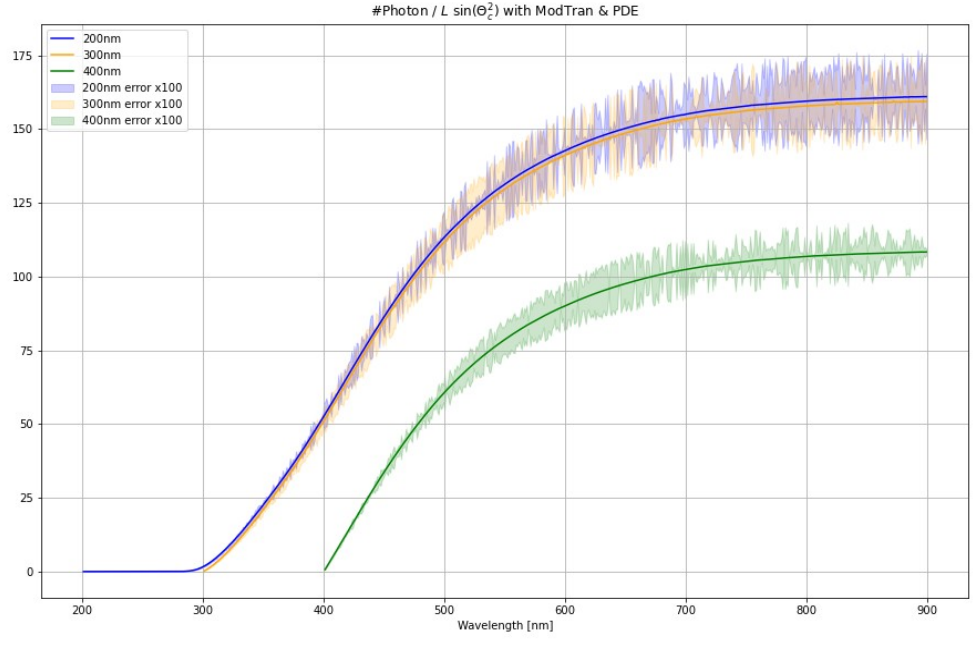
GPU based photon propagation for CORSIKA 8

Dominik Baack
dominik.baack@tu-dortmund.de

Computing Structure



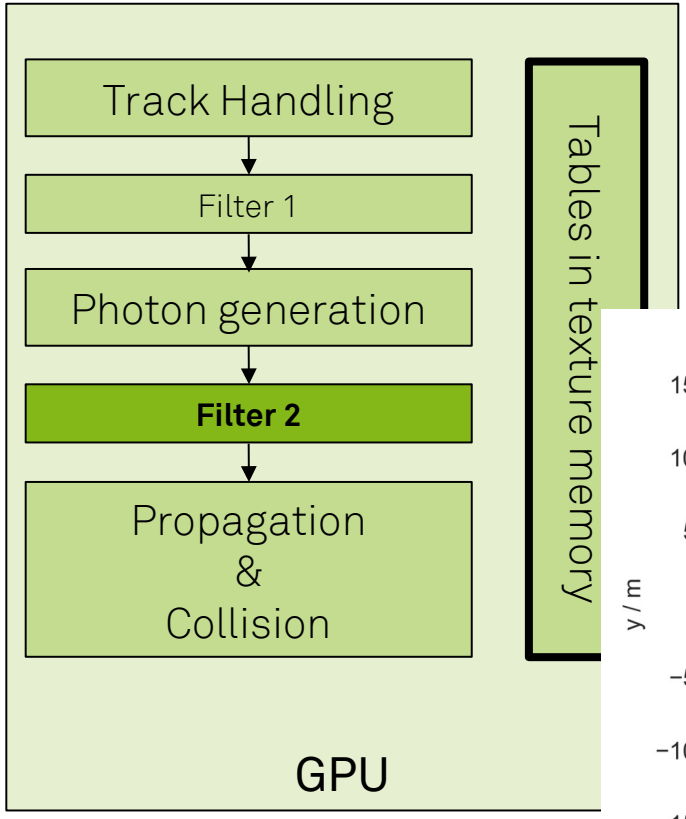


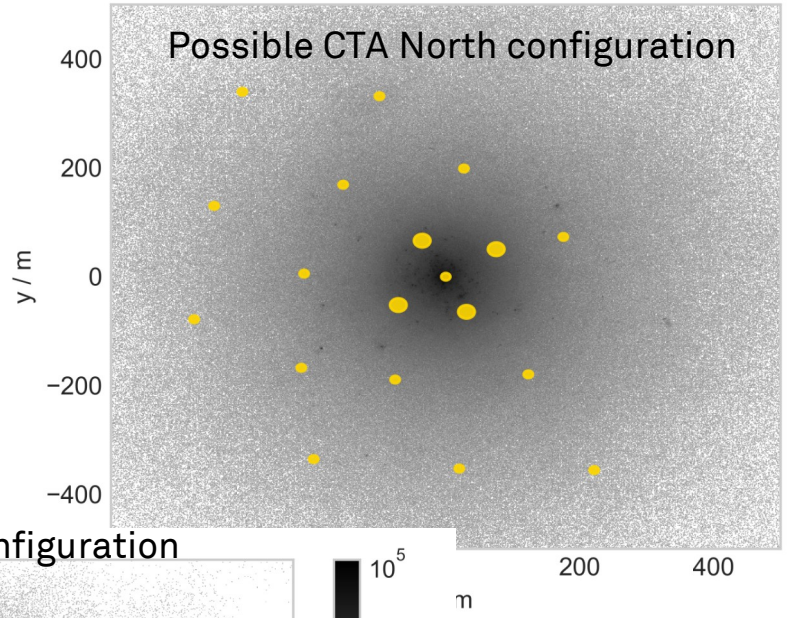


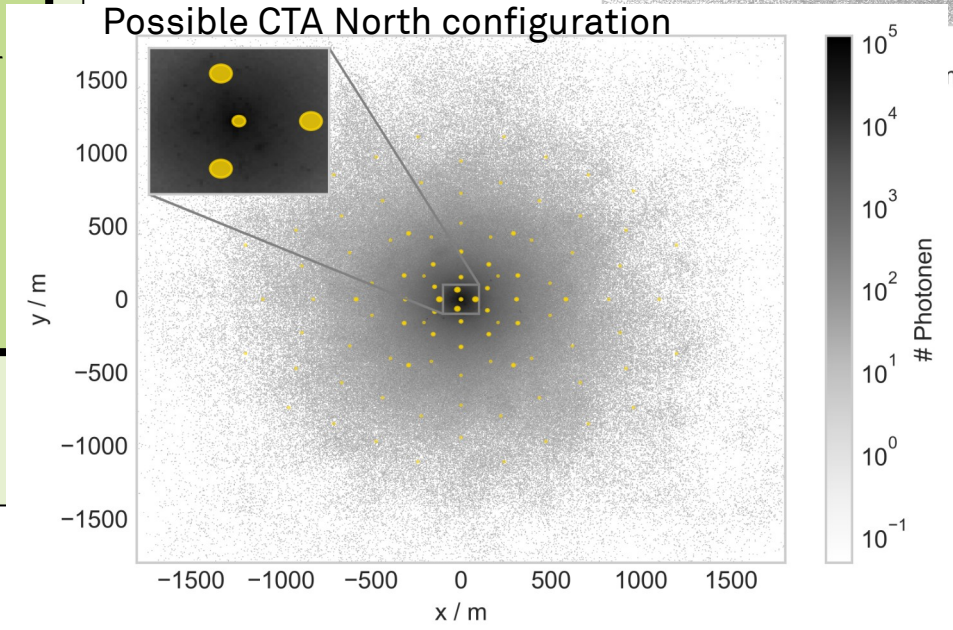
GPU based photon p

Dominik Baack
dominik.baack@tu-dortmund.de

Computing Structure

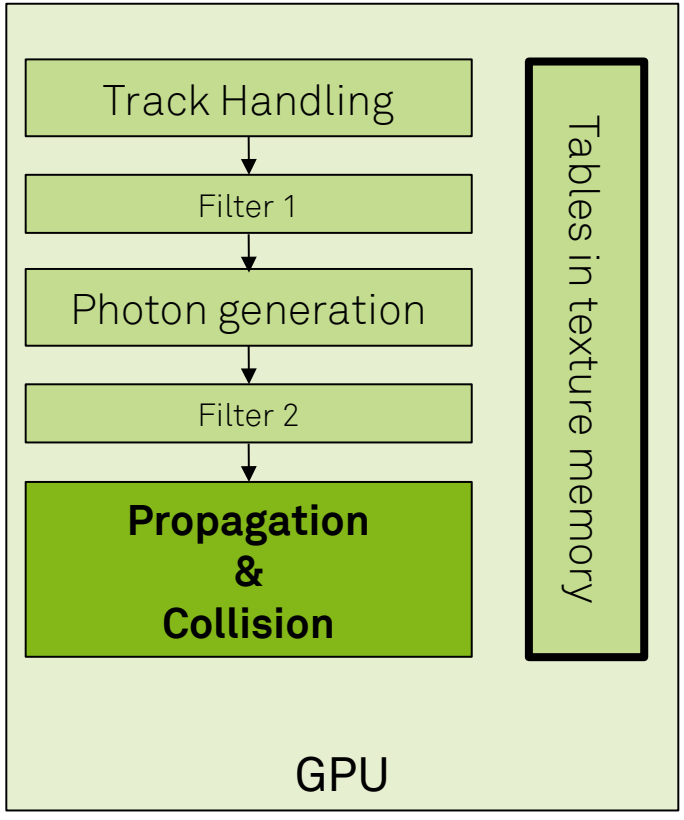






Dominik Baack
dominik.baack@tu-dortmund.de

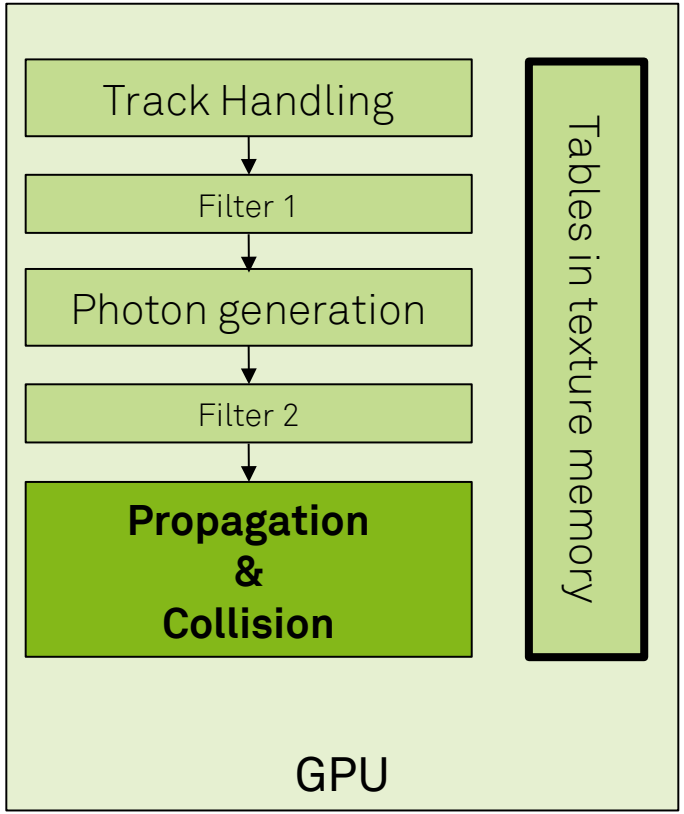
Computing Structure

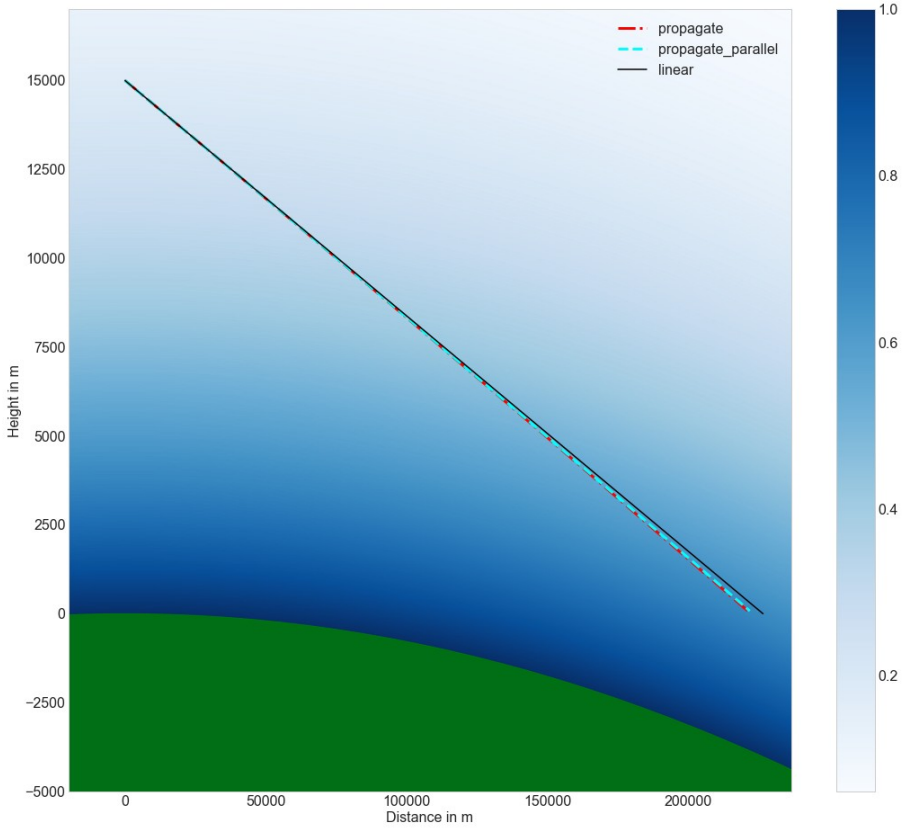


- Propagation linear in first order …

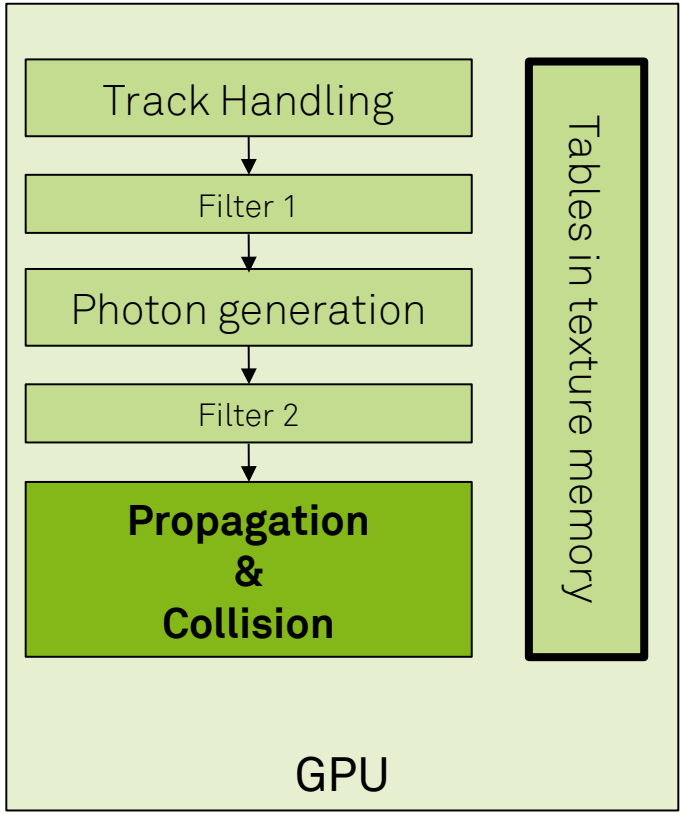
  … but corrections necessary for modern IACT's

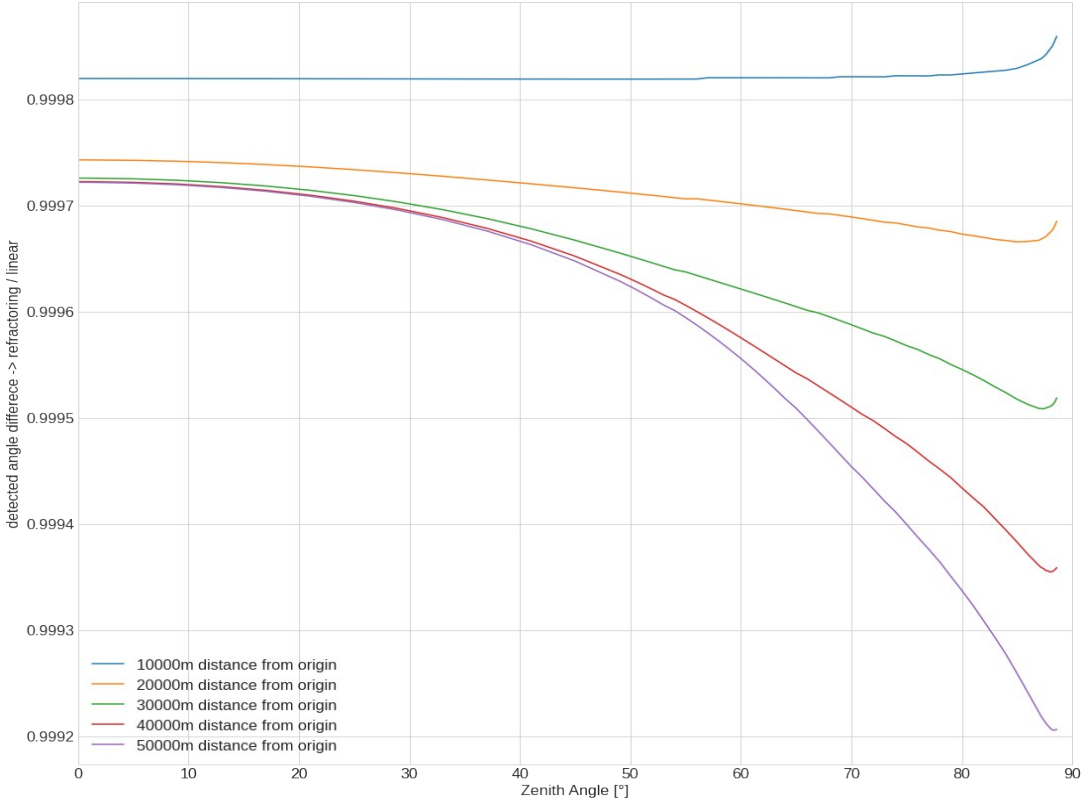
Computing Structure



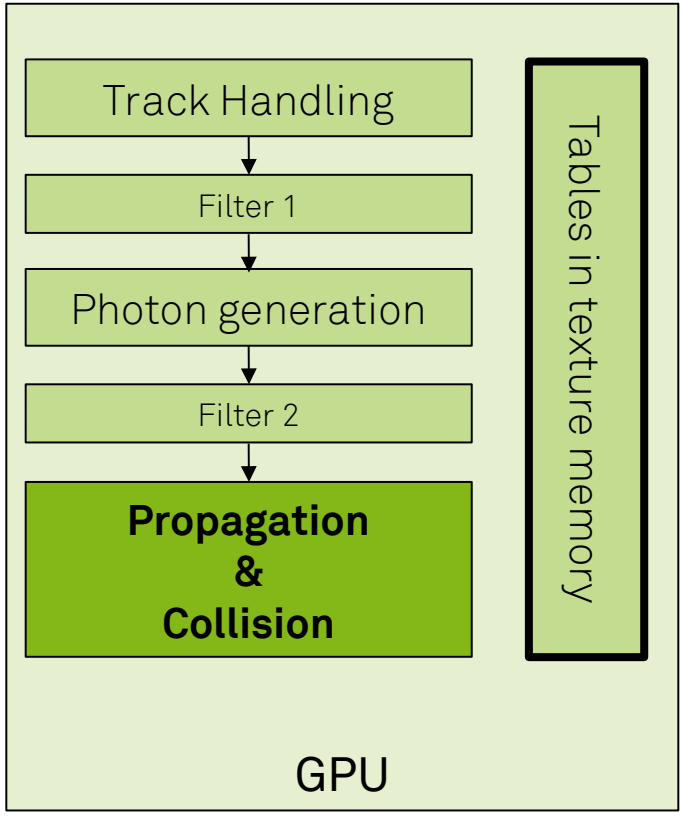


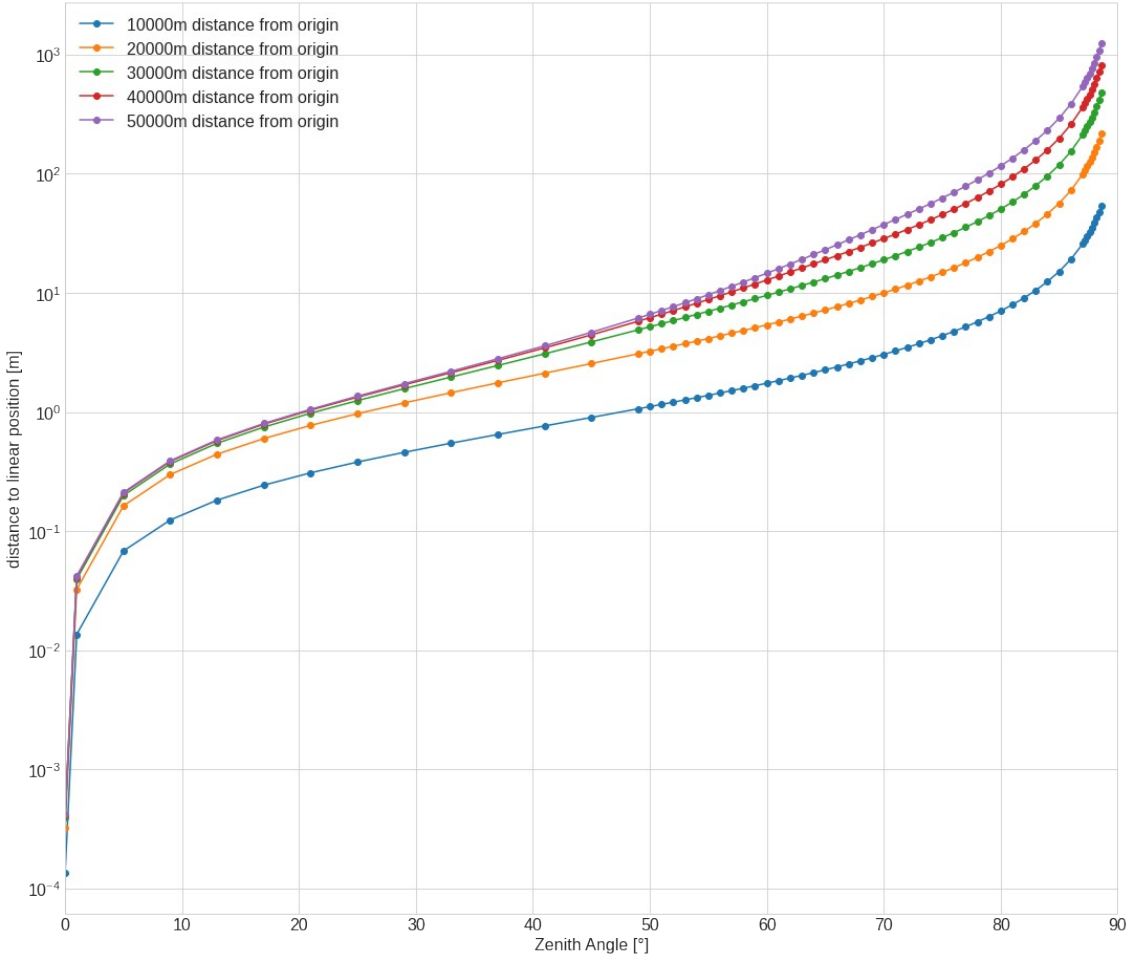
Computing Structure





Computing Structure

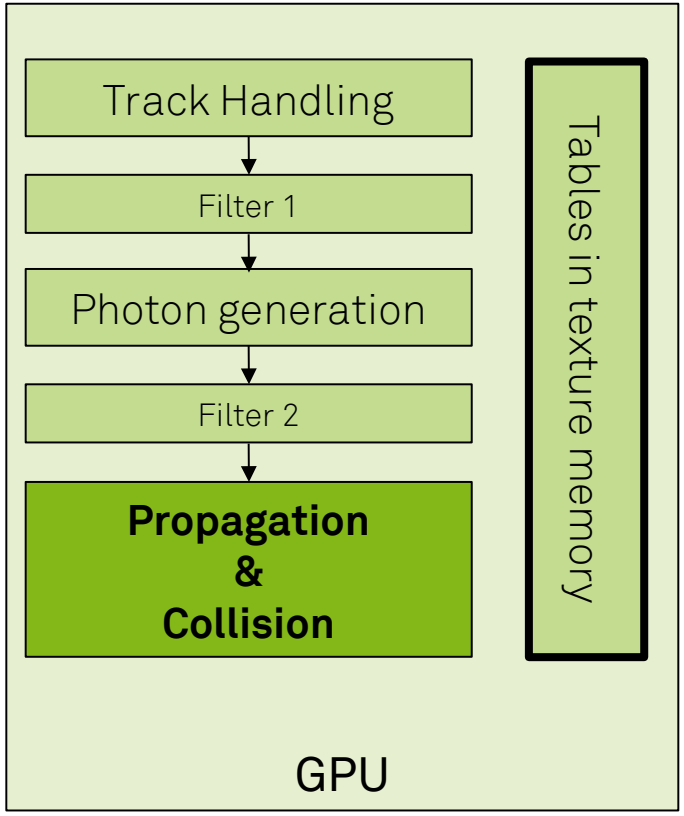




GPU based photon propagation for CORSIKA 8

Dominik Baack
dominik.baack@tu-dortmund.de

Computing Structure



- Propagation linear in first order
- Tabulated correction factors generated offline and applied on GPU

Conclusion & Outlook

- GPU based light propagation possible and likely fast than CPU based approach
- Corsika 8 is a open source and modular framework for cosmic ray simulation: https://gitlab.ikp.kit.edu/AirShowerPhysics/corsika
- In Vivo tests still necessary with very new EM-Model (Proposal) for Corsika8
- Improvement by sorting to reduce “starvation” of threads or more efficient thinning
- Dedicated test of a more efficient fluorescence propagation