

# SMARTHEP

REAL-TIME ANALYSIS FOR  
SCIENCE AND INDUSTRY

## ESR12: Accelerated Anomaly Detection

Pratik Jawahar

Supervisors:

Caterina Doglioni, Jiri Masik, Alex Oh,  
Maurizio Pierini

### Overview:

- Qualification Task - Heterogenous Tracking
- ML based Data Compression - Baler
- Misc. Activities



SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

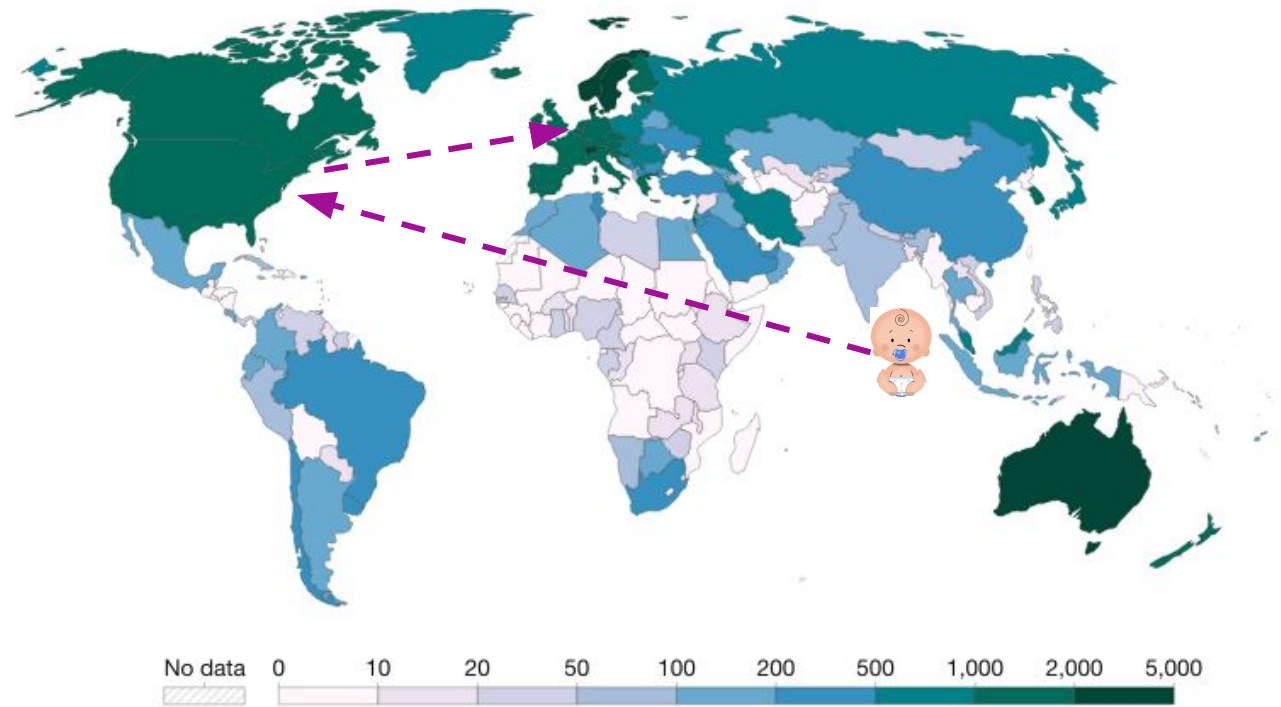
# About me:

- BS: Mechanical Engineering (2019)
- MS: Robotics Engineering (2022)
- PhD: Robo.. \*sike\* Particle Physics
- Experience:
  - Summer student (CERN, 2018)
    - Control algorithms for GEM detectors (CMS)
  - Technical student (CERN, 2021), DIANA HEP Fellowship (2020)
    - ML based anomaly detection for new physics searches
- Website: <https://www.pratikjawahar.com/>

Annual articles published in scientific and technical journals per million people, 2018

Our World in Data

Includes physics, biology, chemistry, mathematics, clinical medicine, biomedical research, engineering and technology, and earth and space sciences.



Data source: World Bank (2022); United Nations (2022)

[OurWorldInData.org/research-and-development](https://OurWorldInData.org/research-and-development) | CC BY

Note: Articles are counted by the country of the author's institution.

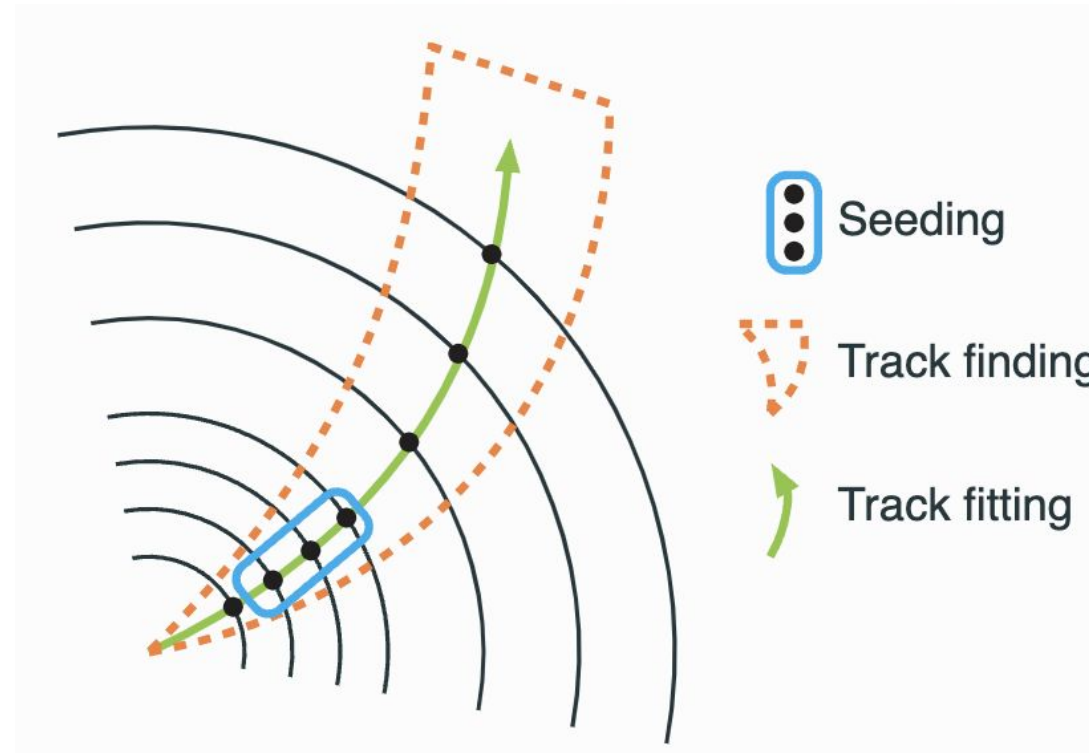


# Qualification Task: Heterogenous Track Reconstruction



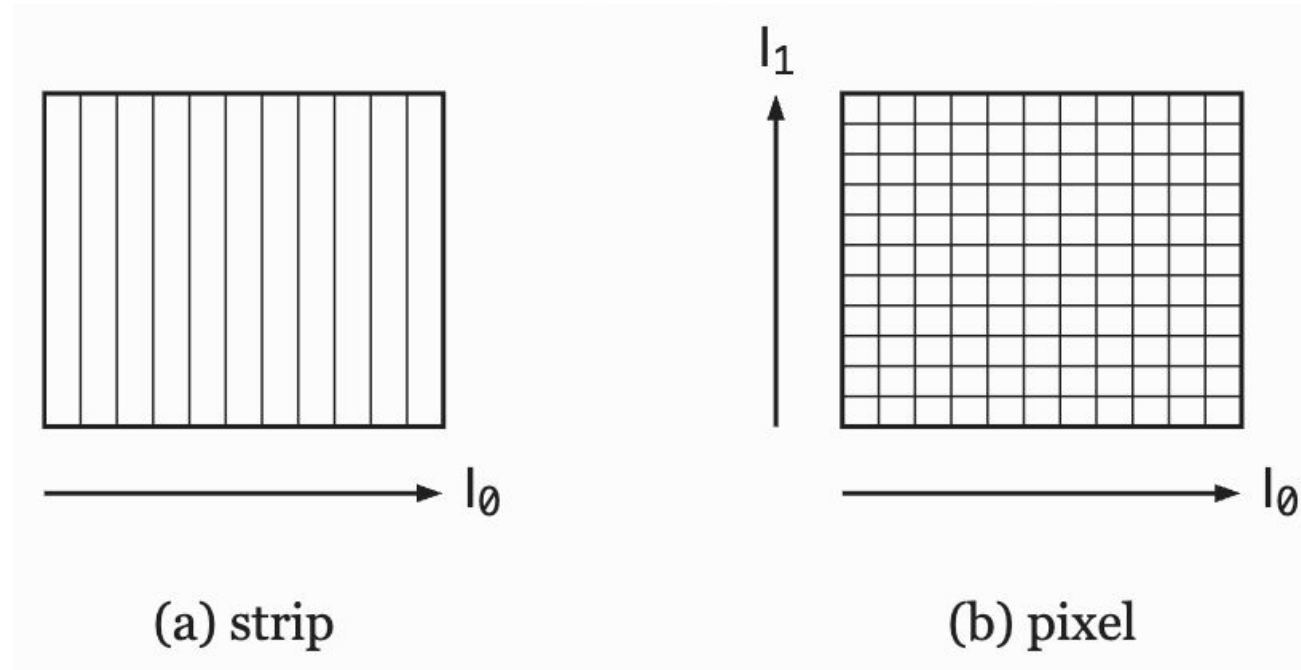
# Track Reconstruction:

- Reconstruct trajectories of charged particles as they pass through different layers of the detector, subject to a magnetic field



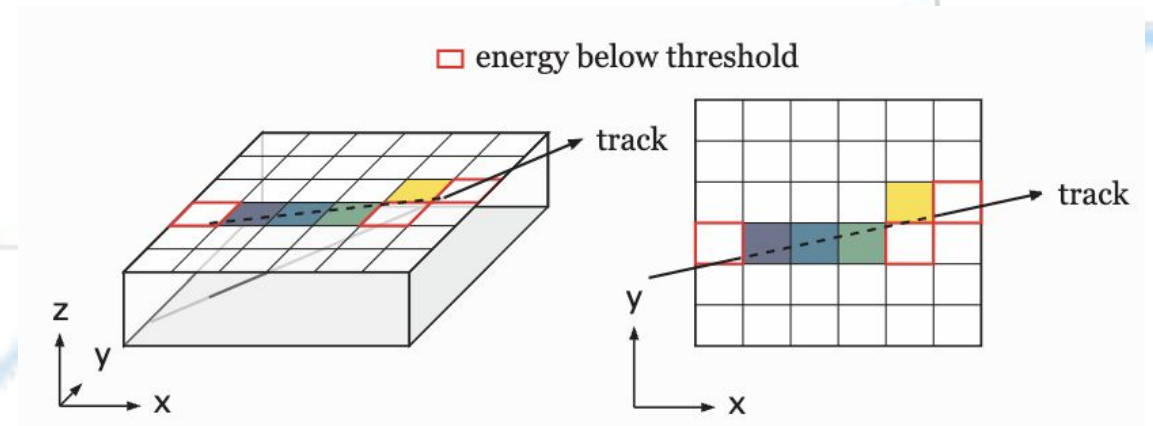
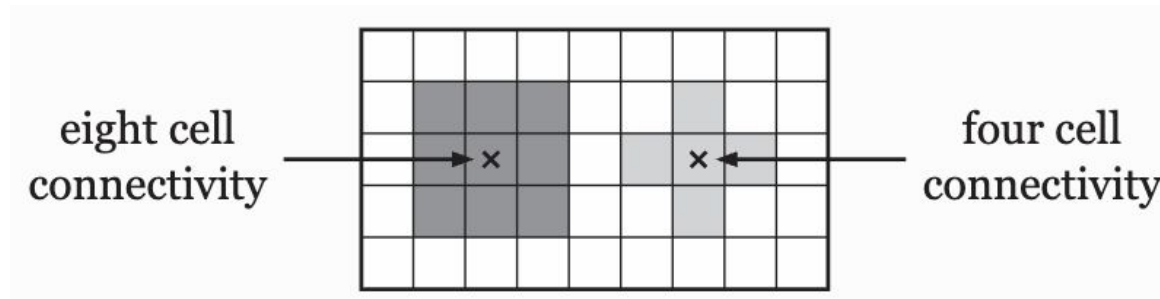
# Track Reconstruction:

- Step 1: Detect a charged particle when it passes through a layer
  - Ionization in semiconductors



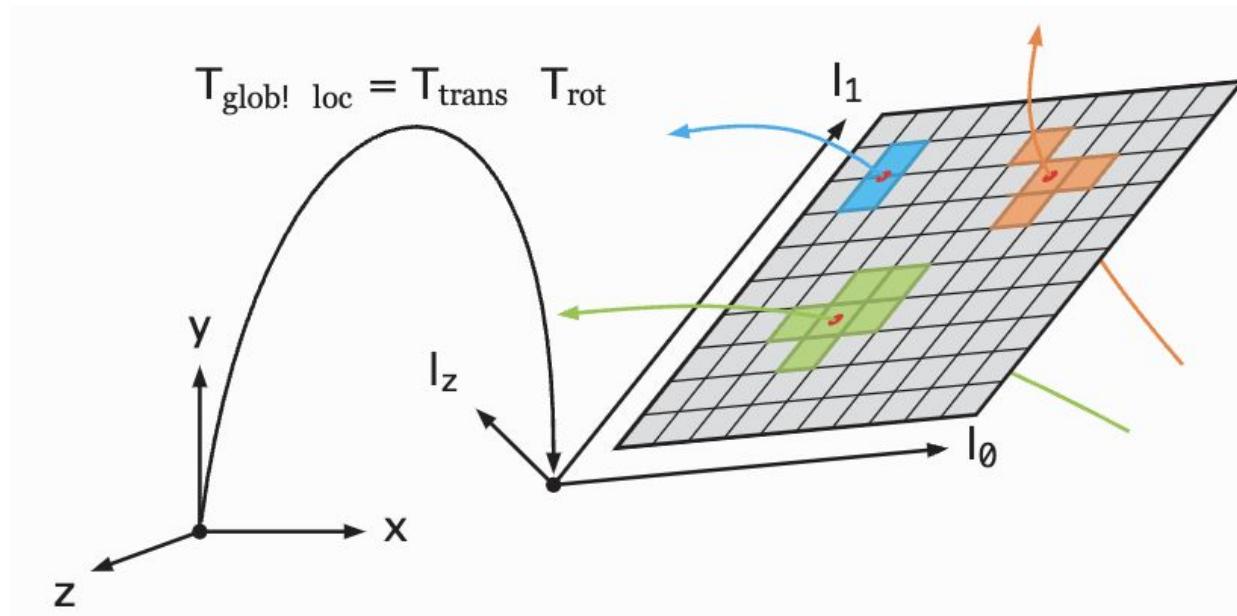
# Track Reconstruction:

- Step 2: Clusterization
  - Depends on how you treat readout, detector design etc.
  - Connected Components Algo (CCA) iterated on edges



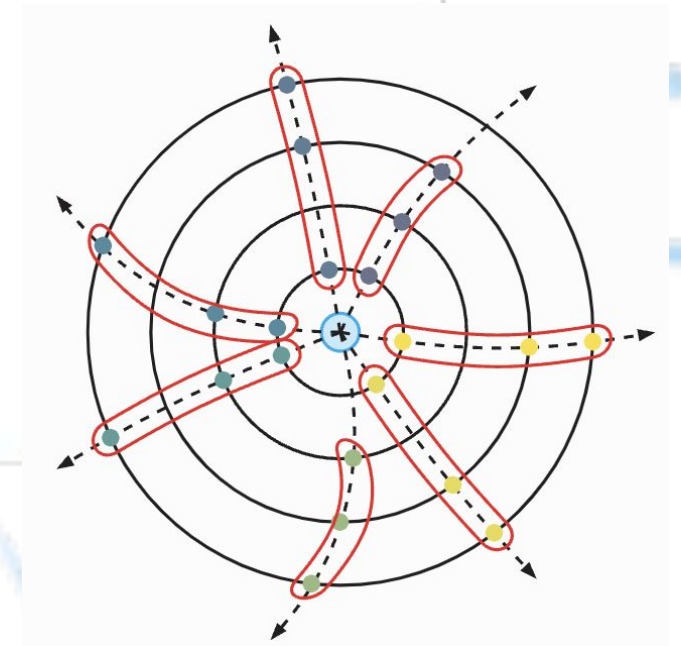
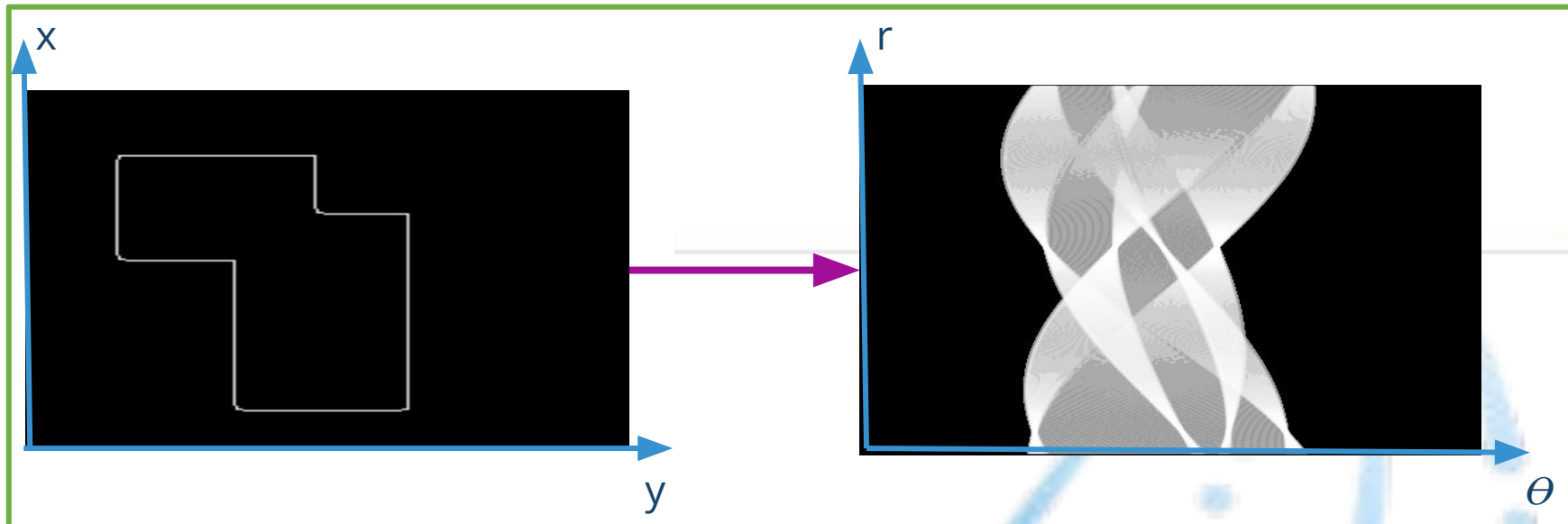
# Track Reconstruction:

- Step 3: Spacepoint Formation
  - Move from local sub-detector frame to global detector frame



# Track Reconstruction:

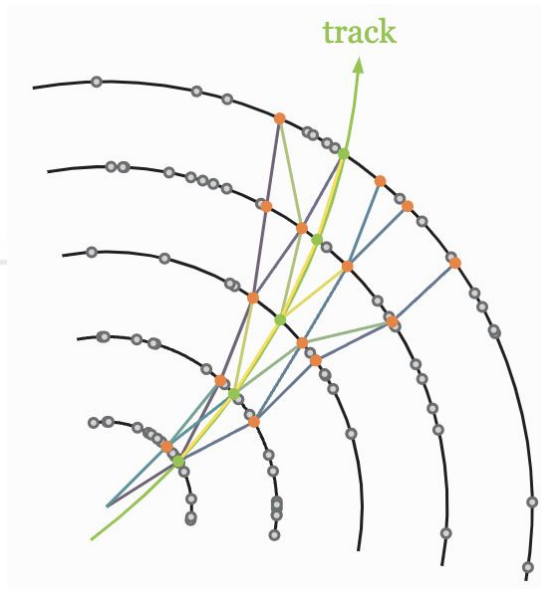
- Step 4: Seeding
  - Find triplets of spacepoints that could potentially belong to the same track
    - Conformal maps => Hough Transform





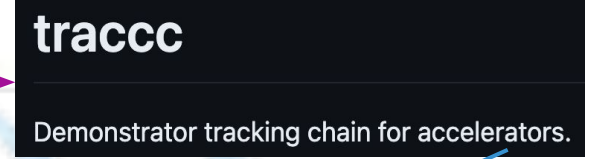
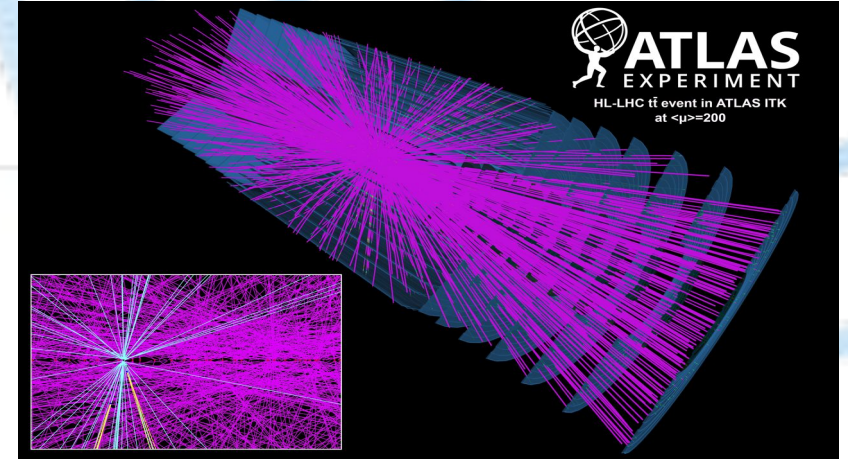
# Track Reconstruction:

- Step 5: Track finding and fitting
  - Start from a seed
    - Kalman formalism (eg. Combinatorial KF) => Smoothing (eg. global  $\chi^2$  fit, or walk back with Kalman fit) => Ambiguity resolution



# Problem(s)!

- No. of tracks per event for HL-LHC, expected to increase 2.5x
- Current R&D: Use GPUs for speedup via parallel computation
- **However**, sequential algorithms like CKF do much better on CPUs than GPUs!



Computing term for specific purpose architectures (eg. GPU, TPU, IPU etc.)

# Track Reconstruction:

- Heterogeneous Track Reconstruction:
  - Step 1: Profiling CPU and GPU code to identify speed-up in inference time
    - Ideally without drop in tracking efficiency
  - Step 2: Identify bottlenecks
    - Points where one architecture outperforms the other
  - Step 3: Calculate data-transfer latencies at bottlenecks
    - Data transfer latencies between host (CPU) and device (GPU) eat up speed-up

## Step 1: Ensure GPU computations do not decrease physics performance

### D2H Calculations - CPU vs D2H CPU

```
Running seeding algo with CUDA spacepoints moved to host
Number of D2H seeds: 28346 (host), 28346 (device)
Matching rate(s):
- 88.6898% at 0.01% uncertainty
- 98.9275% at 0.1% uncertainty
- 99.2592% at 1% uncertainty
- 99.2839% at 5% uncertainty
Running track param est with CUDA spacepoints and CUDA seeds moved to host
Number of D2H track parameters: 28346 (host), 28346 (device)
Matching rate(s):
- 99.2944% at 0.01% uncertainty
- 99.7989% at 0.1% uncertainty
- 99.8024% at 1% uncertainty
- 99.8024% at 5% uncertainty
```

### D2H Calculations - CPU vs CUDA

```
Number of seeds: 28346 (host), 28346 (device)
Matching rate(s):
- 87.6208% at 0.01% uncertainty
- 98.797% at 0.1% uncertainty
- 99.2345% at 1% uncertainty
- 99.2733% at 5% uncertainty
Number of track parameters: 28346 (host), 28346 (device)
Matching rate(s):
- 98.7617% at 0.01% uncertainty
- 99.7989% at 0.1% uncertainty
- 99.8024% at 1% uncertainty
- 99.8024% at 5% uncertainty
```

# CUDA Profiling - NSight Compute

Step 2: Profile code in as close-to-deployment conditions as possible

CUDA tools: NSight Compute, NSight Systems

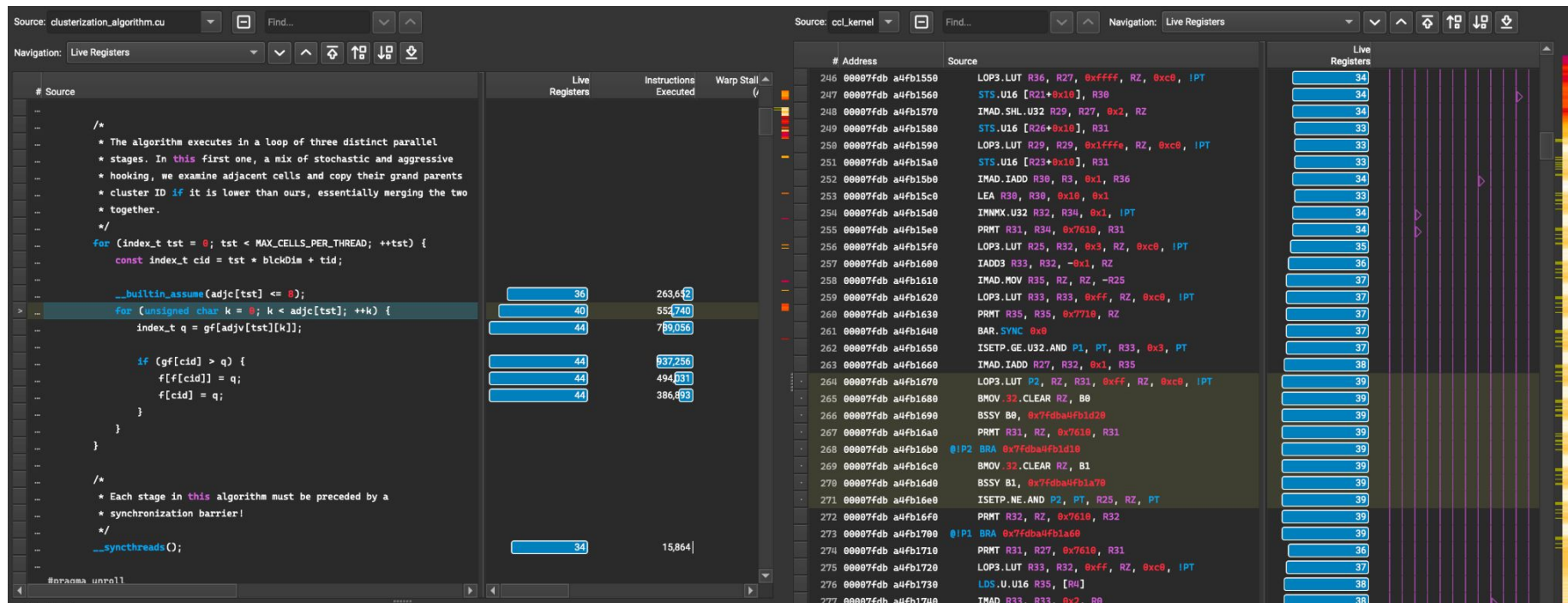
ID	Issues Detected	Function Name	Demangled Name	Process	Device Name	Grid Size	Block Size	Cycles [cycle]	Duration [msecond]	Compute Throughput	Memory Throughput [%]	# Registers [register/thread]
0	20	ccl_kernel	tracc::cuda::kernels::ccl_kernel(v...	[1138765] tracc...	Tesla T4	474, 1, 1	128, 1, 1	472435	0.81	43.75	43.72	49
1	14	form_spacepoints	tracc::cuda::kernels::form_space...	[1138765] tracc...	Tesla T4	137, 1, 1	1024, 1, 1	95281	0.16	5.95	36.43	36
2	11	count_grid_capacities	tracc::cuda::kernels::count_grid...	[1138765] tracc...	Tesla T4	545, 1, 1	256, 1, 1	18692	0.04	22.31	50.49	20
3	14	populate_grid	tracc::cuda::kernels::populate_gri...	[1138765] tracc...	Tesla T4	545, 1, 1	256, 1, 1	35039	0.06	11.99	54.14	22
4	15	fill_prefix_sum	tracc::cuda::kernels::fill_prefix_su...	[1138765] tracc...	Tesla T4	18, 1, 1	32, 1, 1	23989	0.04	7.45	26.34	28
5	10	count_doublets	tracc::cuda::kernels::count_doubl...	[1138765] tracc...	Tesla T4	1131, 1, 1	64, 1, 1	848561	1.45	71.08	16.92	42
6	13	find_doublets	tracc::cuda::kernels::find_doublet...	[1138765] tracc...	Tesla T4	479, 1, 1	64, 1, 1	1211907	2.07	33.78	35.93	38
7	9	count_triplets	tracc::cuda::kernels::count_triplet...	[1138765] tracc...	Tesla T4	24496, 1, 1	64, 1, 1	1957486	3.33	68.04	23.78	45
8	15	reduce_triplet_counts	tracc::cuda::kernels::reduce_tripl...	[1138765] tracc...	Tesla T4	479, 1, 1	64, 1, 1	19298	0.03	2.52	11.65	16
9	15	find_triplets	tracc::cuda::kernels::find_triplets(...	[1138765] tracc...	Tesla T4	720, 1, 1	64, 1, 1	294522	0.51	29.96	27.57	45
10	18	update_triplet_weights	tracc::cuda::kernels::update_tripl...	[1138765] tracc...	Tesla T4	930, 1, 1	64, 1, 1	18140	0.04	13.81	48.89	23
11	19	select_seeds	tracc::cuda::kernels::select_seed...	[1138765] tracc...	Tesla T4	479, 1, 1	64, 1, 1	102384	0.18	26.16	26.16	39
12	14	estimate_track_params	tracc::cuda::kernels::estimate_tra...	[1138765] tracc...	Tesla T4	443, 1, 1	64, 1, 1	78502	0.14	5.62	44.97	52



## Uncoalesced Global Accesses

This kernel has uncoalesced global accesses resulting in a total of 5097660 excessive sectors (68% of the total 7519840 sectors). Check the L2 Theoretical Sectors Global Excessive table for the primary source locations uncoalesced device memory accesses.

Name	Value	Info
derived__memory_l2_theoretical_sectors_global_excessive	5.09766e+06	memory_l2_theoretical_sectors_global > memory_l2_theoretical_sectors_global_ideal



Source: clusterization\_algorithm.cu

Navigation: Live Registers

```

...
/*
 * The algorithm executes in a loop of three distinct parallel
 * stages. In this first one, a mix of stochastic and aggressive
 * hooking, we examine adjacent cells and copy their grand parents
 * cluster ID if it is lower than ours, essentially merging the two
 * together.
 */
for (index_t tst = 0; tst < MAX_CELLS_PER_THREAD; ++tst) {
    const index_t cid = tst * blockDim.x + threadIdx.x;

    __builtin_assume(adjc[tst] <= 0);
    for (unsigned char k = 0; k < adjc[tst]; ++k) {
        index_t q = gf[adjv[tst][k]];

        if (gf[cid] > q) {
            f[cid] = q;
            f[cid] = q;
        }
    }

    /*
     * Each stage in this algorithm must be preceded by a
     * synchronization barrier!
     */
    __syncthreads();
}
...
#pragma unroll

```

# Source	Live Registers	Instructions Executed	Warp Stall
...	...	263,642	...
> ...	40	554,740	...
...	44	789,056	...
...	44	837,256	...
...	44	494,831	...
...	44	386,803	...
...	34	15,864	...

Source: ccl\_kernel

# Address	Source	Live Registers
246 00007fdb a4fb1550	LOPS.LUT R36, R27, 0xffff, RZ, 0xc0, IPT	34
247 00007fdb a4fb1560	STS.U16 [R21+0x10], R30	34
248 00007fdb a4fb1570	IMAD.SHL.U32 R29, R27, 0x2, RZ	34
249 00007fdb a4fb1580	STS.U16 [R26+0x10], R31	33
250 00007fdb a4fb1590	LOPS.LUT R29, R29, 0x1ffff, RZ, 0xc0, IPT	33
251 00007fdb a4fb15a0	STS.U16 [R23+0x10], R31	33
252 00007fdb a4fb15b0	IMAD.IADD R30, R3, 0x1, R36	34
253 00007fdb a4fb15c0	LEA R30, R30, 0x10, 0x1	33
254 00007fdb a4fb15d0	IMM.X.U32 R32, R34, 0x1, IPT	34
255 00007fdb a4fb15e0	PRMT R31, R34, 0x7618, R31	34
256 00007fdb a4fb15f0	LOPS.LUT R25, R32, 0x3, RZ, 0xc0, IPT	35
257 00007fdb a4fb1600	IADD3 R33, R32, -0x1, RZ	36
258 00007fdb a4fb1610	IMAD.MOV R35, RZ, RZ, -R25	37
259 00007fdb a4fb1620	LOPS.LUT R33, R33, 0xffff, RZ, 0xc0, IPT	37
260 00007fdb a4fb1630	PRMT R35, R35, 0x7718, RZ	37
261 00007fdb a4fb1640	BAR.SYNC 0x0	37
262 00007fdb a4fb1650	ISETP.GE.U32.AND P1, PT, R33, 0x3, PT	37
263 00007fdb a4fb1660	IMAD.IADD R27, R32, 0x1, R35	38
264 00007fdb a4fb1670	LOPS.LUT P2, RZ, R31, 0xffff, RZ, 0xc0, IPT	39
265 00007fdb a4fb1680	BMOV.32.CLEAR RZ, B0	39
266 00007fdb a4fb1690	BSSV B0, 0x7fdbba4fb1d28	39
267 00007fdb a4fb16a0	PRMT R31, RZ, 0x7618, R31	39
268 00007fdb a4fb16b0	IP2.BRA 0x7fdbba4fb1d18	39
269 00007fdb a4fb16c0	BMOV.32.CLEAR RZ, B1	39
270 00007fdb a4fb16d0	BSSV B1, 0x7fdbba4fb1a78	39
271 00007fdb a4fb16e0	ISETP.NE.AND P3, PT, R25, RZ, PT	39
272 00007fdb a4fb16f0	PRMT R32, RZ, 0x7618, R32	39
273 00007fdb a4fb1700	IP1.BRA 0x7fdbba4fb1a68	39
274 00007fdb a4fb1710	PRMT R31, R27, 0x7618, R31	36
275 00007fdb a4fb1720	LOPS.LUT R33, R32, 0xffff, RZ, 0xc0, IPT	37
276 00007fdb a4fb1730	LDS.U.U16 R35, [R4]	38
277 00007fdb a4fb1740	IMAD R33, R33, 0x2, B0	38



# CUDA Profiling

Data File	Detector Geometry	No. of Events
tml_full/ttbar_mu300	tml_detector/trackml-detector	1

## Old Version

## New Version - with FastSV

CPU		Duration [mu-sec]	GPU		Duration [mu-sec]
Parent process			Parent process		
Container Instantiation		8	Container Instantiation		6
File reading		1,720,207	File reading		NA
Clusterization		40,197	Clusterization		NA
Spacepoint Formation		6,078	Spacepoint Formation		NA
Clusterization + Spacepoints		46,275	Clusterization + Spacepoints		40,197
Seeding		1,075,583	Seeding		80,072
Track param est		27,909	Track param est		2,651

CPU		Duration [mu-sec]	GPU		Duration [mu-sec]
Parent process			Parent process		
Container Instantiation		8	Container Instantiation		6
File reading		1,288,893	File reading		NA
Clusterization		36,070	Clusterization		NA
Spacepoint Formation		3,131	Spacepoint Formation		NA
Clusterization + Spacepoints		39,201	Clusterization + Spacepoints		2,433
Seeding		792,729	Seeding		11,686
Track param est		9,316	Track param est		411



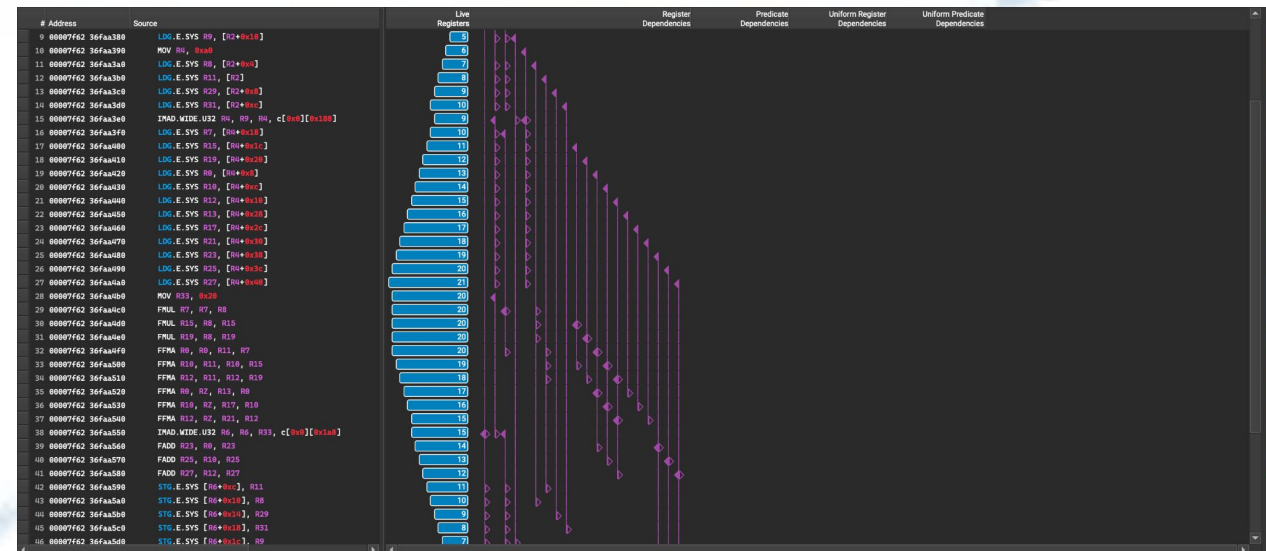
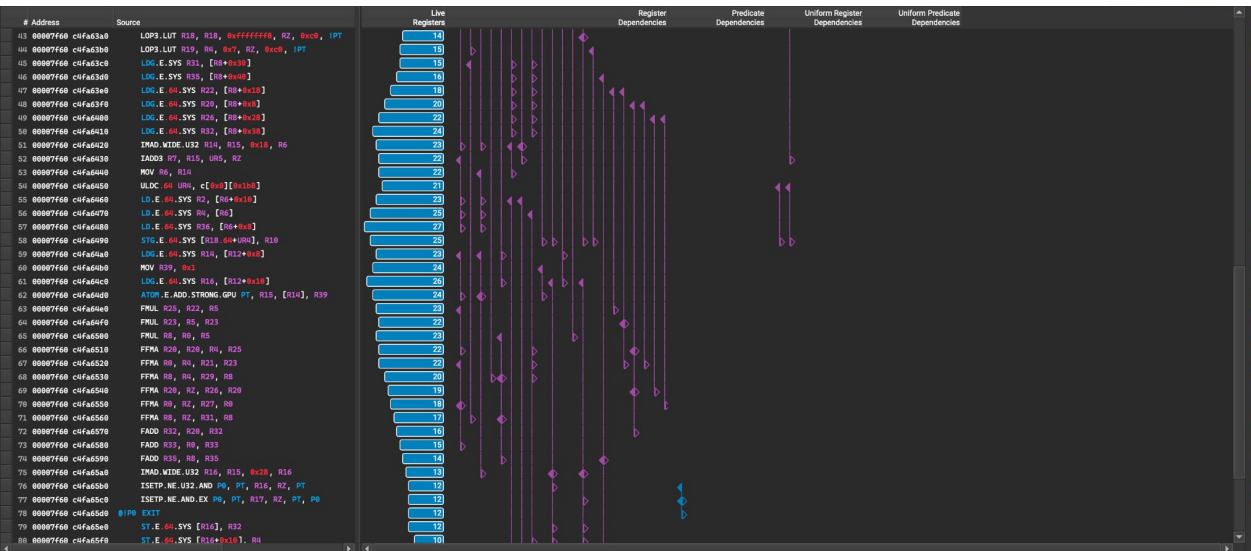


# CUDA Profiling - NSight Compute

Old Version

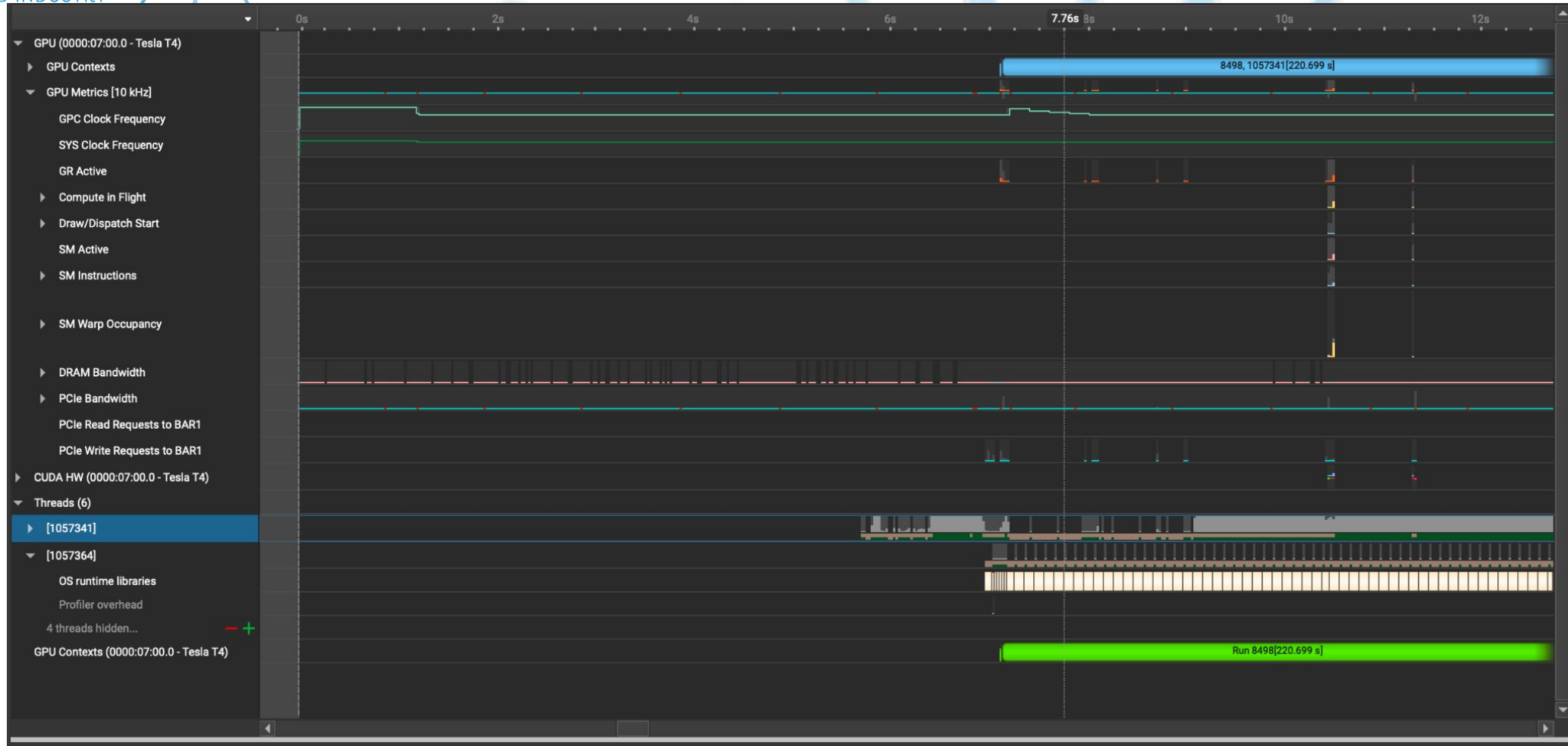
Data File	Detector Geometry	No. of Events
tml_full/ttbar_mu300	tml_detector/trackml-detector	1

New Version - with FastSV





# CUDA Profiling - NSight Systems



# CUDA Profiling

- POC feasibility example:
  - Clusterization, Spacepoint formation, Seeding are significantly faster on GPU
  - Considering Host-Device and Device-Host wall-time overheads,
    - there is still a speedup of  $\sim O(800 \text{ msec})$
- *\*Note\** This is only an example for one event

Data File	Detector Geometry	No. of Events
tml_full/ttbar_mu300	tml_detector/trackml-detector	1

CPU		GPU	
Parent process	Duration [mu-sec]	Parent process	Duration [mu-sec]
Container Instantiation	15	Container Instantiation	5
File reading	1,341,172	File reading	NA
Clusterization	36,070	Clusterization	NA
Spacepoint Formation	3,131	Spacepoint Formation	NA
Clusterization + Spacepoints	39,201	Clusterization + Spacepoints	2,433
Seeding	814,789	Seeding	11,686
Track param est	9,316	Track param est	411

# CUDA Profiling

- POC feasibility example:
  - Clusterization, Spacepoint formation, Seeding are significantly faster on GPU
  - Considering Host-Device and Device-Host wall-time overheads,
    - there is still a speedup of  $\sim O(800 \text{ msec})$
- If CKF on GPU is considerably slower, it may still be feasible to run the first part of the chain on Device and move data to Host for CKF
  - Potential solution for net speedup in the tracking chain
- \*Note\* This is only an example for one event

Data File	Detector Geometry	No. of Events
tml_full/ttbar_mu300	tml_detector/trackml-detector	1

CPU		GPU	
Parent process	Duration [mu-sec]	Parent process	Duration [mu-sec]
Host to Device [Cells]	<b>3035</b>	Device to Host [Cells]	NA
Host to Device [Spacepoints]	<b>2,703</b>	Device to Host [Spacepoints]	<b>1087</b>
Host to Device [Seeds]	<b>781</b>	Device to Host [Seeds]	<b>349</b>
Host to Device [Track params]	<b>1,655</b>	Device to Host [Track params]	<b>1,161</b>

# Next Steps

- Extend studies to larger number of events with more realistic data (eg. mu=60, 140, 200 etc)
- Use seeds generated by tracc in ACTS CKF function to understand Device-Host costs
- Repeat overhead studies wrt other parameters such as Memory, GPU Occupancy etc.

Data File	Detector Geometry	No. of Events
tml_full/ttbar_mu300	tml_detector/trackml-detector	1

CPU		GPU	
Parent process	Duration [mu-sec]	Parent process	Duration [mu-sec]
Host to Device [Cells]	<b>3035</b>	Device to Host [Cells]	NA
Host to Device [Spacepoints]	<b>2,703</b>	Device to Host [Spacepoints]	<b>1087</b>
Host to Device [Seeds]	<b>781</b>	Device to Host [Seeds]	<b>349</b>
Host to Device [Track params]	<b>1,655</b>	Device to Host [Track params]	<b>1,161</b>

# Baler: ML based Data Compression

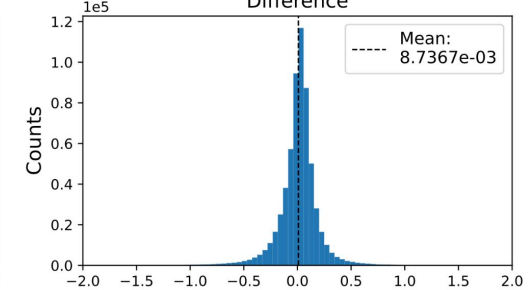
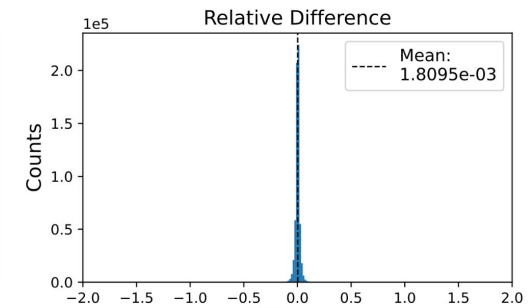
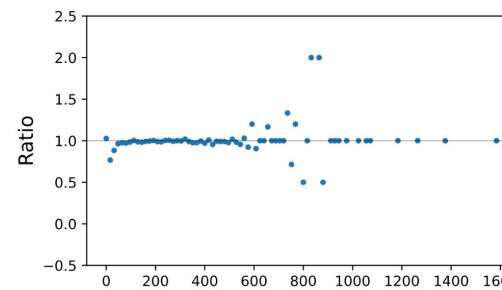
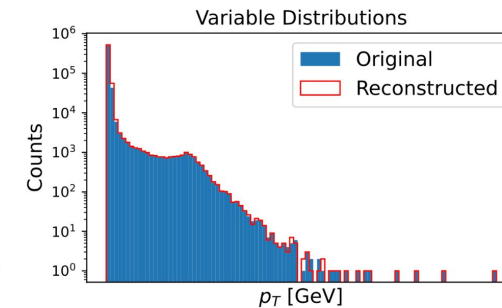
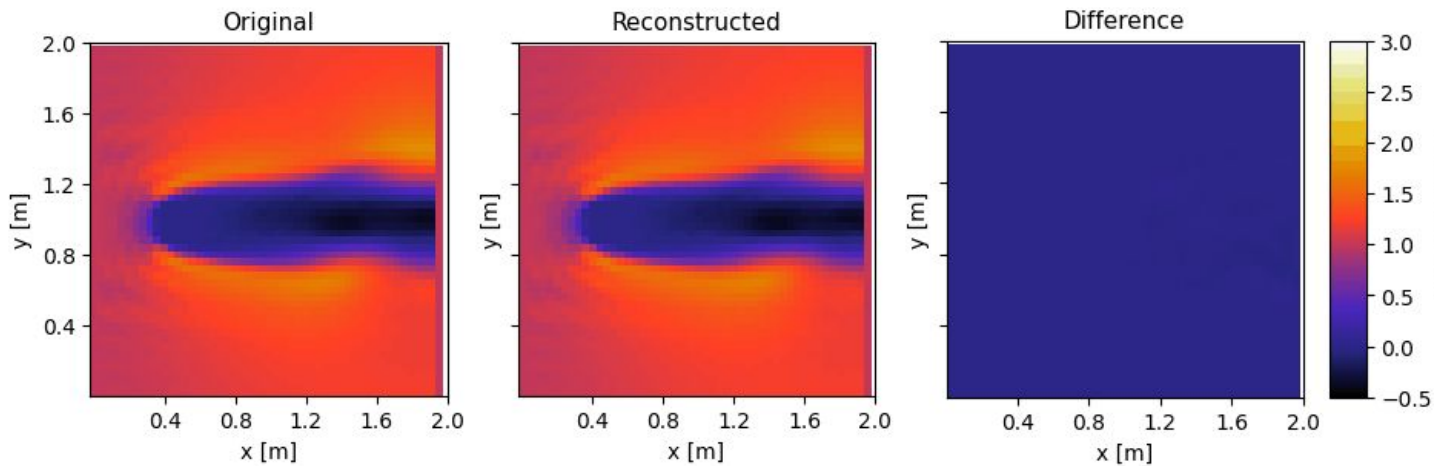
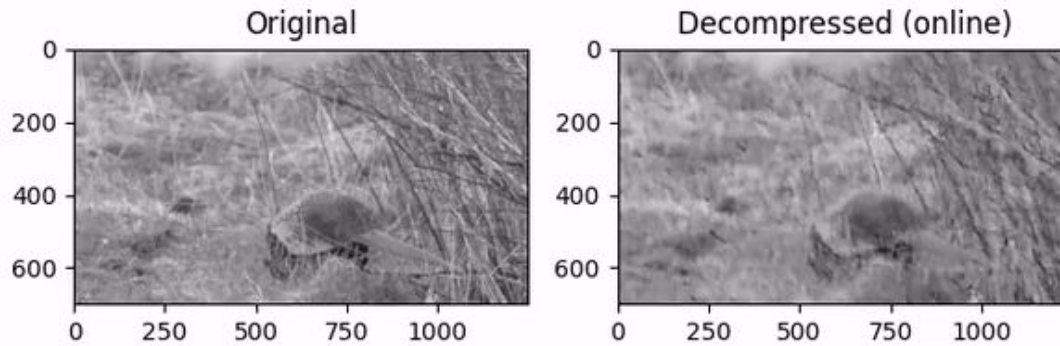


# Baler: ML based Data Compression



Paper: <https://arxiv.org/abs/2305.02283>

Code: <https://github.com/baler-collaboration/baler>



# Misc Activities

## Side Quests:

- Helped lift new wire bonder into the Lund Uni clean room
- Working on DDMTD corrections for mu-CTP trigger chip for temperature variance with Kalman Filters

## Hackathons:

- DL in HEP - MIT
- Baler (Lund)

## Workshops/Conferences:

- IOP APP&HEPP UK National - Talk
- FastML - Talk
- Hammers and Nails - Poster

## Schools:

- tCSC Split
- HASCO particle physics
- STFC UK theory summer school

## Teaching/Outreach:

- Jupyter Notebooks Intro - Lund
- Advanced C++ - Manchester

# Conclusions & Future Work

- Track reconstruction may not scale to HL-LHC
  - R&D required to speed up track reconstruction
    - without loss in tracking efficiency
    - GPUs could be a viable option
      - GPUs may not be able to speed up the full chain due to sequential algo (eg. CKF)
  - Heterogeneous computing solutions may be feasible
    - POC studies show:
      - First half of the chain on GPUs, followed by offloading to CPUs for CKF, could show order of 100msec speedup per event
- Work on novel anomaly detection solutions?
- Contribute to the HLS4ML project?
- Analysis: Semi-visible jet signatures for dark sector searches - ML based signal-background discrimination?
- TBD!





# Thank you!

## Questions?

People are hungry, just putting it out there :)

Feel free to chat later as well! :)

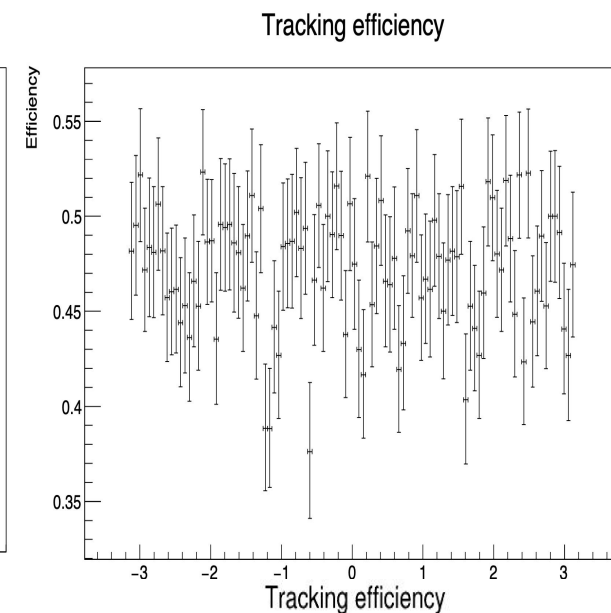
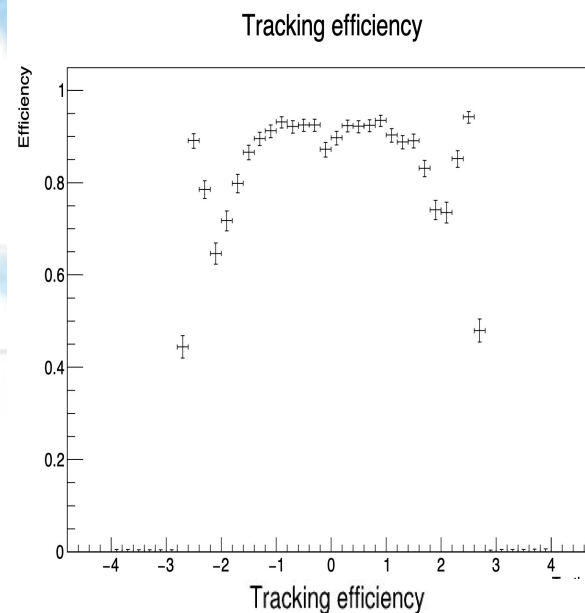
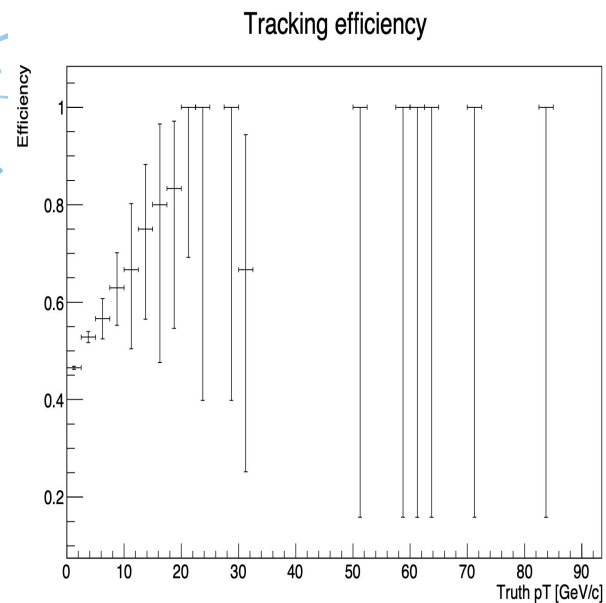


**BACKUP**

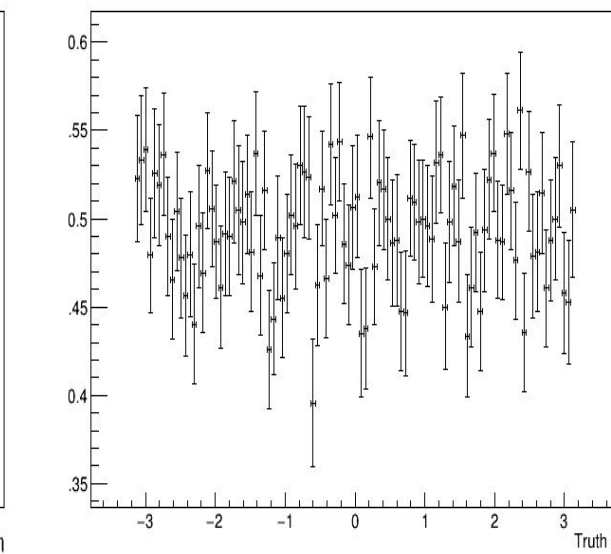
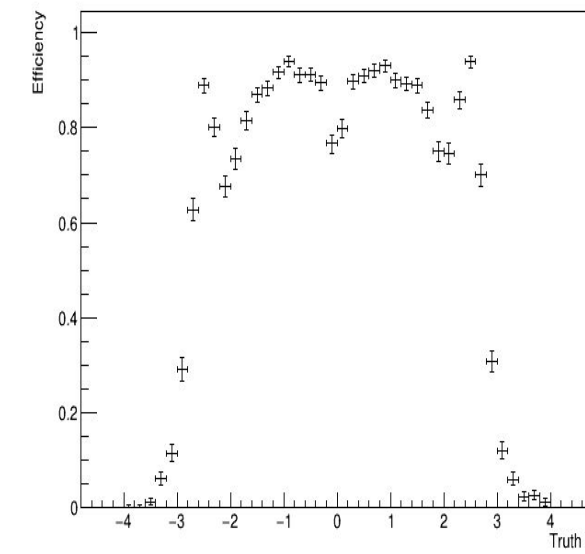
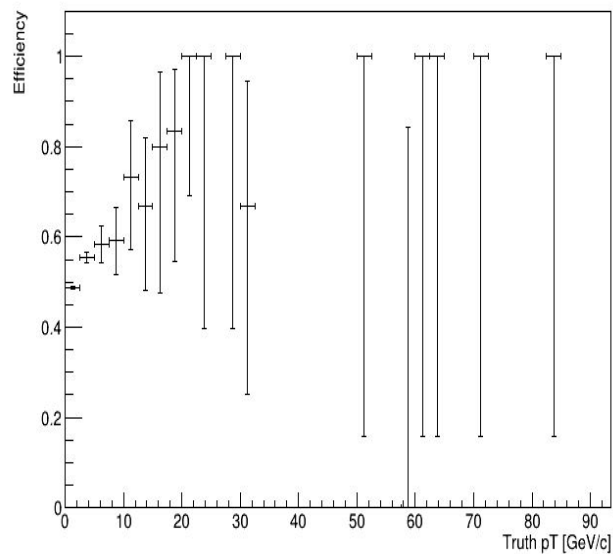


SMARTHEP is funded by the European Union's Horizon 2020 research and innovation programme, call H2020-MSCA-ITN-2020, under Grant Agreement n. 956086

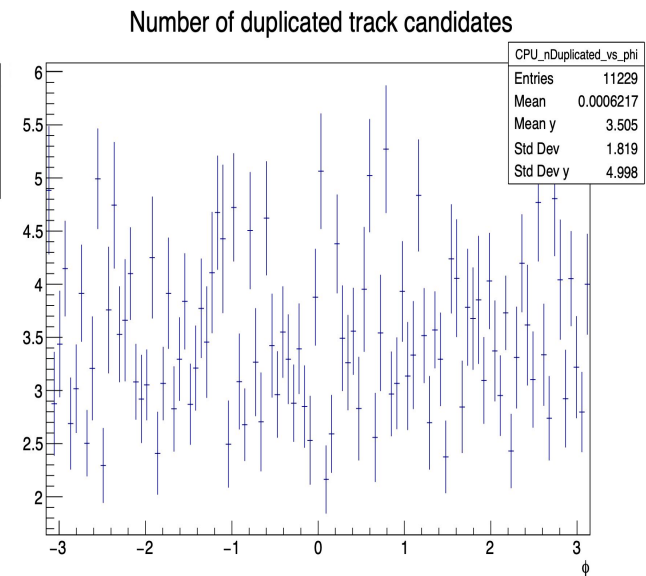
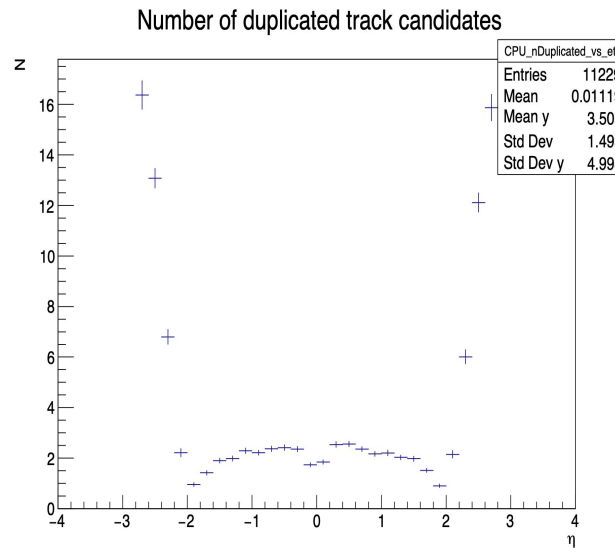
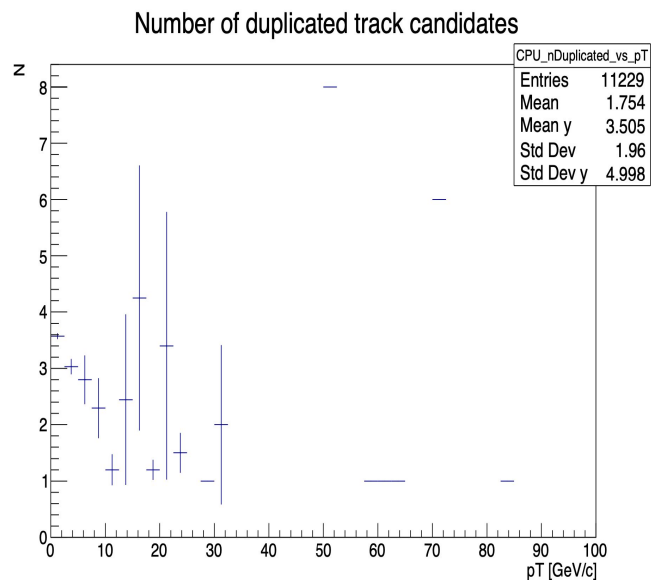
CPU



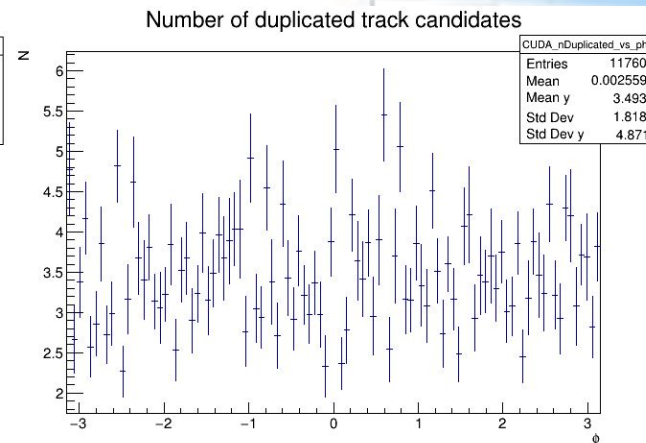
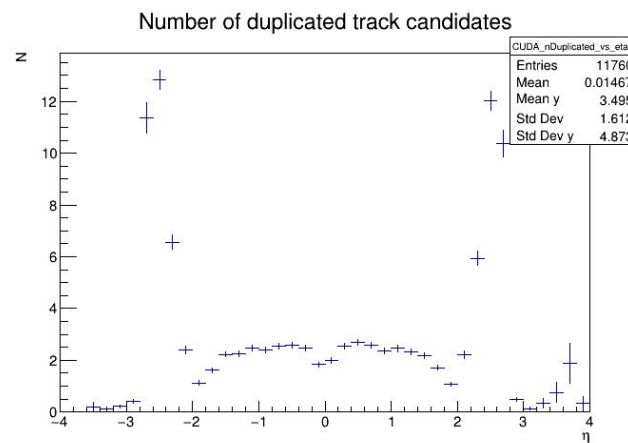
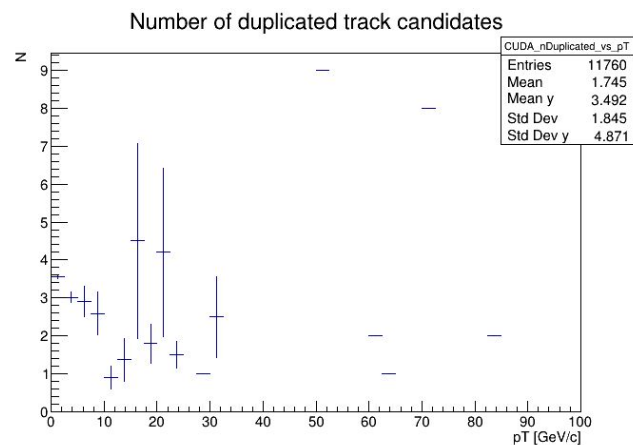
CUDA



## CPU



## CUDA



# CUDA Profiling

## Old Version

- Seeding - Synchronous kernel launches
  - fill\_prefix\_sum used for synchronization

## New Version - with FastSV

- Seeding - Asynchronous kernel launches
  - Increases parallelization

ID	Issues Detected	Function Name	Demangled Name
0	2	find_clusters	tracc::cuda::kernels::find_...
1	2	fill_prefix_sum	tracc::cuda::kernels::fill_p...
2	2	count_cluster_c...	tracc::cuda::kernels::coun...
3	2	connect_compo...	tracc::cuda::kernels::conn...
4	3	create_measure...	tracc::cuda::kernels::creat...
5	2	fill_prefix_sum	tracc::cuda::kernels::fill_p...
6	2	form_spacepoints	tracc::cuda::kernels::form...
7	2	fill_prefix_sum	tracc::cuda::kernels::fill_p...
8	2	count_grid_capa...	tracc::cuda::kernels::coun...
9	2	populate_grid	tracc::cuda::kernels::popu...
10	3	fill_prefix_sum	tracc::cuda::kernels::fill_p...
11	2	count_doublets	tracc::cuda::kernels::coun...
12	3	fill_prefix_sum	tracc::cuda::kernels::fill_p...
13	2	find_doublets	tracc::cuda::kernels::find_...
14	3	fill_prefix_sum	tracc::cuda::kernels::fill_p...
15	2	count_triplets	tracc::cuda::kernels::coun...
16	3	fill_prefix_sum	tracc::cuda::kernels::fill_p...
17	3	find_triplets	tracc::cuda::kernels::find_...
18	3	fill_prefix_sum	tracc::cuda::kernels::fill_p...
19	3	update_triplet_w...	tracc::cuda::kernels::upda...
20	2	select_seeds	tracc::cuda::kernels::sele...
21	2	estimate_track_...	tracc::cuda::kernels::esti...

ID	Issues Detected	Function Name	Demangled Name
0	3	ccl_kernel	tracc::cuda::kernels::ccl_kernel(v...
1	3	form_spacepoints	tracc::cuda::kernels::form_space...
2	2	count_grid_capacities	tracc::cuda::kernels::count_grid_...
3	2	populate_grid	tracc::cuda::kernels::populate_gri...
4	3	fill_prefix_sum	tracc::cuda::kernels::fill_prefix_su...
5	2	count_doublets	tracc::cuda::kernels::count_doubl...
6	2	find_doublets	tracc::cuda::kernels::find_doublet...
7	2	count_triplets	tracc::cuda::kernels::count_triplet...
8	2	reduce_triplet_counts	tracc::cuda::kernels::reduce_trip...
9	3	find_triplets	tracc::cuda::kernels::find_triplets(...
10	3	update_triplet_weights	tracc::cuda::kernels::update_trip...
11	3	select_seeds	tracc::cuda::kernels::select_seed...
12	2	estimate_track_params	tracc::cuda::kernels::estimate_tra...





# CUDA Profiling- NEW

Old Version  
count\_doublets()

Data File	Detector Geometry	No. of Events
tml_full/ttbar_mu300	tml_detector/trackml-detector	1

New Version - with FastSV  
count\_doublets()

