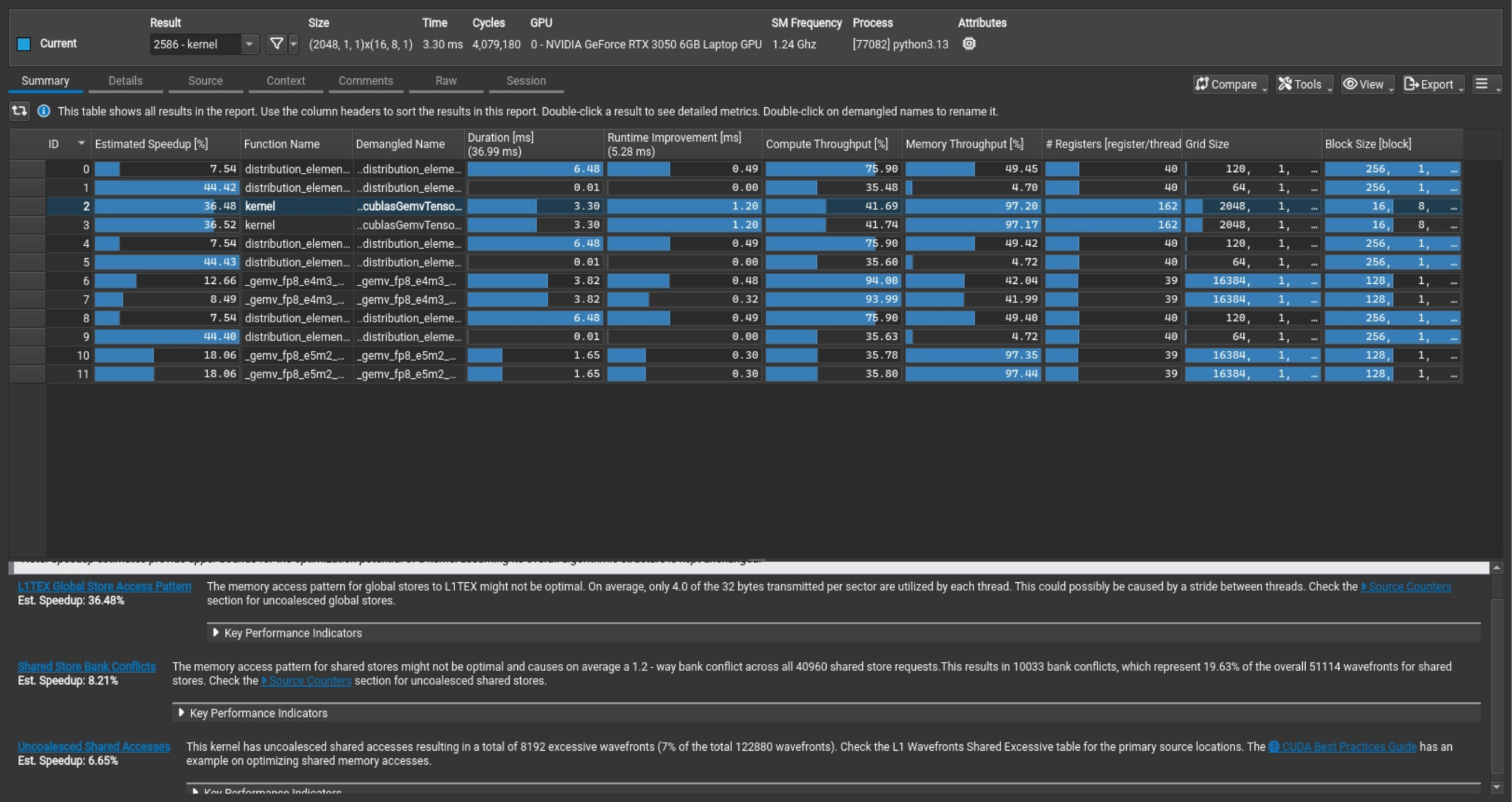

	Result	Size	Time	Cycles	GPU	SM Frequency	Process	Attributes
Current	2586 - kernel	(2048, 1, 1)x(16, 8, 1)	3.30 ms	4,079,180	0 - NVIDIA GeForce RTX 3050 6GB Laptop GPU	1.24 Ghz	[77082] python3.13	

This table shows all results in the report. Use the column headers to sort the results in this report. Double-click a result to see detailed metrics. Double-click on demangled names to rename it.

ID	Estimated Speedup [%]	Function Name	Demangled Name	Duration [ms] (36.99 ms)	Runtime Improvement [ms] (5.28 ms)	Compute Throughput [%]	Memory Throughput [%]	# Registers [register/thread	Grid Size	Block Size [block]
0	7.54	distribution_elemen…	..distribution_eleme…	6.48	0.49	75.90	49.45	40	120, 1, …	256, 1, …
1	44.42	distribution_elemen…	..distribution_eleme…	0.01	0.00	35.48	4.70	40	64, 1, …	256, 1, …
2	36.48	kernel	..cublasGemvTenso…	3.30	1.20	41.69	97.20	162	2048, 1, …	16, 8, …
3	36.52	kernel	..cublasGemvTenso…	3.30	1.20	41.74	97.17	162	2048, 1, …	16, 8, …
4	7.54	distribution_elemen…	..distribution_eleme…	6.48	0.49	75.90	49.42	40	120, 1, …	256, 1, …
5	44.43	distribution_elemen…	..distribution_eleme…	0.01	0.00	35.60	4.72	40	64, 1, …	256, 1, …
6	12.66	_gemv_fp8_e4m3_…	_gemv_fp8_e4m3_…	3.82	0.48	94.00	42.04	39	16384, 1, …	128, 1, …
7	8.49	_gemv_fp8_e4m3_…	_gemv_fp8_e4m3_…	3.82	0.32	93.99	41.99	39	16384, 1, …	128, 1, …
8	7.54	distribution_elemen…	..distribution_eleme…	6.48	0.49	75.90	49.40	40	120, 1, …	256, 1, …
9	44.40	distribution_elemen…	..distribution_eleme…	0.01	0.00	35.63	4.72	40	64, 1, …	256, 1, …
10	18.06	_gemv_fp8_e5m2_…	_gemv_fp8_e5m2_…	1.65	0.30	35.78	97.35	39	16384, 1, …	128, 1, …
11	18.06	_gemv_fp8_e5m2_…	_gemv_fp8_e5m2_…	1.65	0.30	35.80	97.44	39	16384, 1, …	128, 1, …

L1TEX Global Store Access Pattern
Est. Speedup: 36.48%

The memory access pattern for global stores to L1TEX might not be optimal. On average, only 4.0 of the 32 bytes transmitted per sector are utilized by each thread. This could possibly be caused by a stride between threads. Check the ▸ Source Counters section for uncoalesced global stores.

▸ Key Performance Indicators

Shared Store Bank Conflicts
Est. Speedup: 8.21%

The memory access pattern for shared stores might not be optimal and causes on average a 1.2 - way bank conflict across all 40960 shared store requests.This results in 10033 bank conflicts, which represent 19.63% of the overall 51114 wavefronts for shared stores. Check the ▸ Source Counters section for uncoalesced shared stores.

▸ Key Performance Indicators

Uncoalesced Shared Accesses
Est. Speedup: 6.65%

This kernel has uncoalesced shared accesses resulting in a total of 8192 excessive wavefronts (7% of the total 122880 wavefronts). Check the L1 Wavefronts Shared Excessive table for the primary source locations. The CUDA Best Practices Guide has an example on optimizing shared memory accesses.

▸ Key Performance Indicators