	Result	Size	Time	Cycles	GPU	SM Frequency	Process	Attributes
Current	585 - distribution_elementwise_grid_stride_kernel	(120, 1, 1)x(256, 1, 1)	107.84 us	133,450	0 - NVIDIA GeForce RTX 3050 6GB Laptop GPU	1.24 Ghz	[109663] python3.13	

Summary | Details | Source | Context | Comments | Raw | Session

Compare | Tools | View | Export

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] (247.67 ms)	Runtime Improvement [ms] (106.41 ms)	Compute Throughput [%]	Memory Throughput [%]	# Registers [register/thread	Grid Size
0	7.35	distribution_elementwise_grid_stride_kernel	..distribution_elementwise_grid...	0.11	0.01	74.31	45.36	40	120, 1, …
1	7.35	distribution_elementwise_grid_stride_kernel	..distribution_elementwise_grid...	0.11	0.01	74.25	45.00	40	120, 1, …
2	7.35	distribution_elementwise_grid_stride_kernel	..distribution_elementwise_grid...	0.11	0.01	74.27	45.35	40	120, 1, …
3	41.08	_attention_fp8_e5m2_acc_fp32_tiled_kernel	_attention_fp8_e5m2_acc_fp32...	36.17	14.86	55.29	58.92	128	256, 4, …
4	40.74	_attention_fp8_e5m2_acc_fp32_tiled_kernel	_attention_fp8_e5m2_acc_fp32...	36.17	14.73	55.27	59.26	128	256, 4, …
5	7.34	distribution_elementwise_grid_stride_kernel	..distribution_elementwise_grid...	0.11	0.01	74.13	45.04	40	120, 1, …
6	7.34	distribution_elementwise_grid_stride_kernel	..distribution_elementwise_grid...	0.11	0.01	74.20	45.17	40	120, 1, …
7	7.35	distribution_elementwise_grid_stride_kernel	..distribution_elementwise_grid...	0.11	0.01	74.24	45.61	40	120, 1, …
8	62.53	_attention_fp8_e5m2_acc_fp32_kernel	_attention_fp8_e5m2_acc_fp32...	47.46	29.68	34.34	37.47	179	512, 1, …
9	62.60	_attention_fp8_e5m2_acc_fp32_kernel	_attention_fp8_e5m2_acc_fp32...	47.45	29.70	34.29	37.40	179	512, 1, …
10	6.98	distribution_elementwise_grid_stride_kernel	..distribution_elementwise_grid...	0.12	0.01	69.85	84.56	40	120, 1, …
11	6.87	distribution_elementwise_grid_stride_kernel	..distribution_elementwise_grid...	0.11	0.01	68.76	86.07	40	120, 1, …
12	6.94	distribution_elementwise_grid_stride_kernel	..distribution_elementwise_grid...	0.11	0.01	69.43	87.43	40	120, 1, …
13	1.26	ampere_sgemm_128x64_tn	ampere_sgemm_128x64_tn	17.48	0.22	75.00	63.46	122	64, 128, …
14	6.15	vectorized_elementwise_kernel	..vectorized_elementwise_kern...	3.41	0.21	3.10	93.85	28	65536, 1, …
15	7.90	cunn_SoftMaxForwardReg	..cunn_SoftMaxForwardReg..S...	3.47	0.27	42.56	92.10	32	8192, 1, …
16	51.37	Kernel2	..cutlass_80_simt_sgemm_128...	15.35	7.89	79.68	55.44	212	256, 1, …
17	1.26	ampere_sgemm_128x64_tn	ampere_sgemm_128x64_tn	17.47	0.22	75.00	63.47	122	64, 128, …
18	6.04	vectorized_elementwise_kernel	..vectorized_elementwise_kern...	3.41	0.21	3.11	93.96	28	65536, 1, …
19	7.91	cunn_SoftMaxForwardReg	..cunn_SoftMaxForwardReg..S...	3.48	0.27	42.42	92.09	32	8192, 1, …
20	52.54	Kernel2	..cutlass_80_simt_sgemm_128...	15.35	8.07	79.62	55.44	212	256, 1, …

The following performance optimization opportunities were discovered for this result. Follow the rule links to see more context on the Details page.
Note: *Speedup estimates provide upper bounds for the optimization potential of a kernel assuming its overall algorithmic structure is kept unchanged.*

FP32 Non-Fused Instructions
Est. Speedup: 7.35%

This kernel executes 1208192 fused and 669184 non-fused FP32 instructions. By converting pairs of non-fused instructions to their fused, higher-throughput equivalent, the achieved FP32 performance could be increased by up to 18% (relative to its current performance).

▸ Key Performance Indicators