Page: Details | Launch: 0 - 13810 - compute_conditional_join | Add Baseline | Apply Rules | Save as PDF

	Report	Launch	Time	Cycles	Regs	GPU	SM Frequency	CC	Process
Current	report1	13810 - compute_conditional_join_output_size (7813, 1, 1)x(128, 1, 1)	2.22 second	3,131,778,702	94	Quadro RTX 8000	1.40 cycle/nsecond	7.5	[10023] JOIN_BENCH
Baseline 1	report2	13939 - compute_nested_loop_join_output_size (7813, 1, 1)x(128, 1, 1)	1.32 second	1,857,837,894	40	Quadro RTX 8000	1.41 cycle/nsecond	7.5	[10103] JOIN_BENCH

GPU Speed Of Light Throughput

High-level overview of the throughput for compute and memory resources of the GPU. For each unit, the throughput reports the achieved percentage of utilization with respect to the theoretical maximum. Breakdowns show the throughput for each individual sub-metric of Compute and Memory to clearly identify the highest contributor. High-level overview of the utilization for compute and memory resources of the GPU presented as a roofline chart.

Compute (SM) Throughput [%]	52.90 (-6.10%)	Duration [second]	2.22 (+68.87%)
Memory Throughput [%]	52.90 (-6.10%)	Elapsed Cycles [cycle]	3131778702 (+68.57%)
L1/TEX Cache Throughput [%]	61.25 (-37.87%)	SM Active Cycles [cycle]	3095049041.18 (+68.58%)
L2 Cache Throughput [%]	0.06 (-48.15%)	SM Frequency [cycle/nsecond]	1.40 (-0.15%)
DRAM Throughput [%]	0.00 (-39.30%)	DRAM Frequency [cycle/nsecond]	6.49 (-0.18%)

Latency Issue This kernel exhibits low compute throughput and memory bandwidth utilization relative to the peak performance of this device. Achieved compute throughput and/or memory bandwidth below 60.0% of peak typically indicate latency issues. Look at Scheduler Statistics and Warp State Statistics for potential reasons.

Roofline Analysis The ratio of peak float (fp32) to double (fp64) performance on this device is 32:1. The kernel achieved 0% of this device's fp32 peak performance and 0% of its fp64 peak performance.

Compute Workload Analysis

Detailed analysis of the compute resources of the streaming multiprocessors (SM), including the achieved instructions per clock (IPC) and the utilization of each available pipeline. Pipelines with very high utilization might limit the overall performance.

Executed Ipc Elapsed [inst/cycle]	2.02 (-7.53%)	SM Busy [%]	50.84 (-8.50%)
Executed Ipc Active [inst/cycle]	2.03 (-7.51%)	Issue Slots Busy [%]	50.84 (-7.51%)
Issued Ipc Active [inst/cycle]	2.03 (-7.51%)		

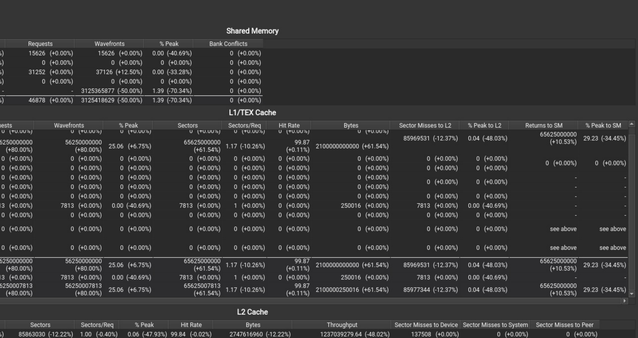
Balanced No pipeline is over-utilized.

Memory Workload Analysis

Detailed analysis of the memory resources of the GPU. Memory can become a limiting factor for the overall kernel performance when fully utilizing the involved hardware units (Mem Busy), exhausting the available communication bandwidth between those units (Max Bandwidth), or by reaching the maximum throughput of issuing memory instructions (Mem Pipes Busy). Detailed chart of the memory units. Detailed tables with data for each memory unit.

Memory Throughput [Mbyte/second]	2.63 (-39.41%)	Mem Busy [%]	30.62 (-37.87%)
L1/TEX Hit Rate [%]	99.87 (+0.11%)	Max Bandwidth [%]	52.90 (-6.10%)
L2 Hit Rate [%]	99.79 (-0.03%)	Mem Pipes Busy [%]	52.90 (-6.10%)

Memory Chart



Shared Memory

	Instructions	Requests	Wavefronts	% Peak	Bank Conflicts
Shared Load	15626 (+0.00%)	15626 (+0.00%)	15626 (+0.00%)	0.00 (-40.69%)	0 (+0.00%)
Shared Load Matrix	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)
Shared Store	31252 (+0.00%)	31252 (+0.00%)	37126 (+12.50%)	0.00 (-33.28%)	0 (+0.00%)
Shared Atomic	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)
Other	-	-	3125365877 (-50.00%)	1.39 (-70.34%)	0 (+0.00%)
Total	46878 (+0.00%)	46878 (+0.00%)	3125418629 (-50.00%)	1.39 (-70.34%)	0 (+0.00%)

L1/TEX Cache

	Requests	Sectors	Sectors/Req	% Peak	Hit Rate	Bytes	Sector Misses to L2	% Peak to L2	Returns to SM	% Peak to SM
Local Load	(+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	65625000000 (+0.53%)	29.23 (-34.45%)
Global Load	50000000 (+80.00%)	56250000000 (+80.00%)	56250000000 (+80.00%)	25.06 (-6.75%)	65625000000 (+61.54%)	1.17 (-10.26%)	99.87 (+0.11%)	2100000000000 (+61.54%)	85969531 (-12.37%)	0.04 (-48.03%)
Surface Load	(+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)
Texture Load	(+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)
Global Store	(+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	-	-
Local Store	(+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	-	-
Surface Store	(+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	-	-
Global Reduction	(+0.00%)	7813 (+0.00%)	7813 (+0.00%)	0.00 (-40.69%)	7813 (+0.00%)	1 (+0.00%)	0 (+0.00%)	250016 (+0.00%)	7813 (+0.00%)	0.00 (-40.69%)
Surface Reduction	(+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)
Global Atomic ALU	(+0.00%)									
Global Atomic CAS	(+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	see above
Surface Atomic ALU	(+0.00%)									
Surface Atomic CAS	(+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	see above
Loads	50000000 (+80.00%)	56250000000 (+80.00%)	56250000000 (+80.00%)	25.06 (-6.75%)	65625000000 (+61.54%)	1.17 (-10.26%)	99.87 (+0.11%)	2100000000000 (+61.54%)	85969531 (-12.37%)	0.04 (-48.03%)
Stores	(+0.00%)	7813 (+0.00%)	7813 (+0.00%)	0.00 (-40.69%)	7813 (+0.00%)	1 (+0.00%)	0 (+0.00%)	250016 (+0.00%)	7813 (+0.00%)	0.00 (-40.69%)
Total	50007813 (+80.00%)	56250007813 (+80.00%)	56250007813 (+80.00%)	25.06 (-6.75%)	65625007813 (+61.54%)	1.17 (-10.26%)	99.87 (+0.11%)	2100000250016 (+61.54%)	85977344 (-12.37%)	0.04 (-48.03%)	65625000000 (+10.53%)	29.23 (-34.45%)

L2 Cache

	Requests	Sectors	Sectors/Req	% Peak	Hit Rate	Bytes	Throughput	Sector Misses to Device	Sector Misses to System	Sector Misses to Peer
L1/TEX Load	85673473 (-11.87%)	85863030 (-12.22%)	1.00 (-0.40%)	0.06 (-47.93%)	99.84 (-0.02%)	2747616960 (-12.22%)	1237039279.64 (-48.02%)	137508 (+0.00%)	0 (+0.00%)	0 (+0.00%)
L1/TEX Store	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)
L1/TEX Atomic ALU	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)
L1/TEX Atomic CAS	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)	0 (+0.00%)
L1/TEX Reduction	7813 (+0.00%)	7813 (+0.00%)	1 (+0.00%)	0.00 (-40.68%)	99.97 (+0.00%)	250016 (+0.00%)	112562.86 (-40.78%)	2 (+0.00%)	0 (+0.00%)	0 (+0.00%)
L1/TEX Total	85671463 (-11.93%)	85866217 (-12.69%)	1.00 (-0.87%)	0.06 (-48.21%)	99.84 (-0.02%)	2747718944 (-12.69%)	1237087195.14 (-48.30%)	137510 (+0.00%)	0 (+0.00%)	0 (+0.00%)
GPU Total	85726656 (-11.89%)	86164474 (-12.60%)	1.01 (-0.81%)	0.06 (-48.15%)	99.79 (-0.03%)	2737263168 (-12.60%)	1241382220.58 (-48.25%)	189067 (+0.04%)	0 (+0.00%)	0 (+0.00%)

Device Memory

	Sectors	% Peak	Bytes	Throughput
Load	180895 (+2.12%)	0.00 (-39.42%)	5788640 (+2.12%)	2606176.61 (-39.53%)
Store	1560 (+33.56%)	0.00 (-20.77%)	49920 (+33.56%)	22475.11 (-20.91%)
Total	182455 (+2.32%)	0.00 (-39.30%)	5838560 (+2.32%)	2628651.72 (-39.41%)

Scheduler Statistics

Summary of the activity of the schedulers issuing instructions. Each scheduler maintains a pool of warps that it can issue instructions for. The upper bound of warps in the pool (Theoretical Warps) is limited by the launch configuration. On every cycle each scheduler checks the state of the allocated warps in the pool (Active Warps). Active warps that are not stalled (Eligible Warps) are ready to issue their next instruction. From the set of eligible warps the scheduler selects a single warp from which to issue one or more instructions (Issued Warp). On cycles with no eligible warps, the issue slot is skipped and no instruction is issued. Having many skipped issue slots indicates poor latency hiding.

Active Warps Per Scheduler [warp]	4.93 (-37.05%)	No Eligible [%]	49.16 (+9.17%)
Eligible Warps Per Scheduler [warp]	0.71 (-51.28%)	One or More Eligible [%]	50.84 (-7.51%)
Issued Warp Per Scheduler	0.51 (-7.51%)		

Issue Slot Utilization Every scheduler is capable of issuing one instruction per cycle, but for this kernel each scheduler only issues an instruction every 2.0 cycles. This might leave hardware resources underutilized and may lead to less optimal performance. Out of the maximum of 8 warps per scheduler, this kernel allocates an average of 4.93 active warps per scheduler, but only an average of 0.71 warps were eligible per cycle. Eligible warps are the subset of active warps that are ready to issue their next instruction. Every cycle with no eligible warp results in no instruction being issued and the issue slot remains unused. To increase the number of eligible warps, avoid possible load imbalances due to highly different execution durations per warp. Reducing stalls indicated on the Warp State Statistics and Source Counters sections can help, too.

Issue Slot Utilization The 5.00 theoretical warps per scheduler this kernel can issue according to its occupancy are below the hardware maximum of 8. Use the Occupancy section to identify what limits this kernel's theoretical occupancy.

Warp State Statistics

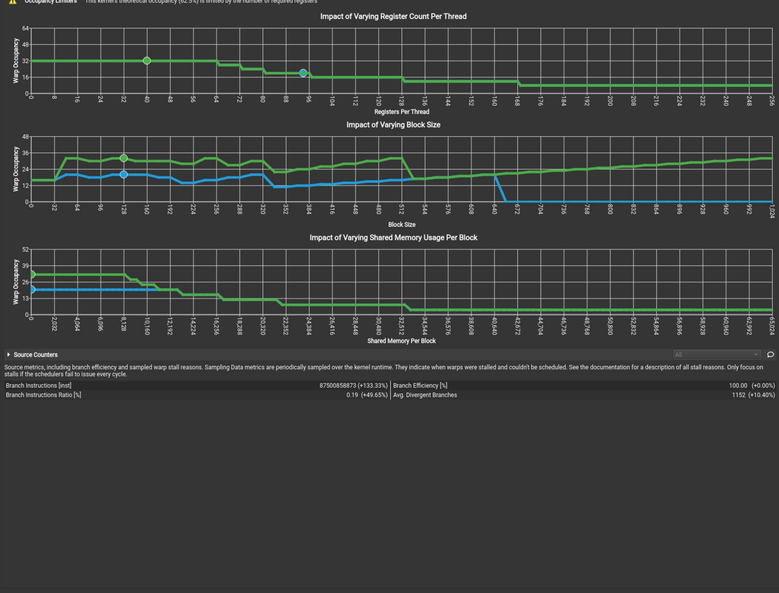
Analysis of the states in which all warps spent cycles during the kernel execution. The warp states describe a warp's readiness or inability to issue its next instruction. The warp cycles per instruction define the latency between two consecutive instructions. The higher the value, the more warp parallelism is required to hide this latency. For each warp state, the chart shows the average number of cycles spent in that state per issued instruction. Stalls are not always impacting the overall performance nor are they completely avoidable. Only focus on stall reasons if the schedulers fail to issue every cycle. When executing a kernel with mixed library and user code, these metrics show the combined values.

Warp Cycles Per Issued Instruction [cycle]	9.69 (-31.93%)	Avg. Active Threads Per Warp	32.00 (+0.00%)
Warp Cycles Per Executed Instruction [cycle]	9.69 (-31.93%)	Avg. Not Predicated Off Threads Per Warp	28.47 (+16.53%)

wait On average, each warp of this kernel spends 3.8 cycles being stalled on a fixed latency execution dependency. This represents about 39.3% of the total average of 9.7 cycles between issuing two instructions. Typically, this stall reason should be very low and only shows up as a top contributor in already highly optimized kernels. If possible, try to further increase the number of active warps to hide the corresponding instruction latencies.

Warp Stall Check the Source Counters section for the top stall locations in your source based on sampling data.

Warp State (All Cycles)



Instruction Statistics

Statistics of the executed low-level assembly instructions (SASS). The instruction mix provides insight into the types and frequency of the executed instructions. A narrow mix of instruction types implies a dependency on few instruction pipelines, while others remain unused. Using multiple pipelines allows hiding latencies and enables parallel execution. Note that 'Instructions/Opcode' and 'Executed Instructions' are measured differently and can diverge if cycles are spent in system calls.

Executed Instructions [inst]	4531314302051 (+55.91%)	Avg. Executed Instructions Per Scheduler [inst]	1573373022.05 (+55.91%)
Issued Instructions [inst]	4531314442295 (+55.91%)	Avg. Issued Instructions Per Scheduler [inst]	1573373070.47 (+55.91%)

NVLink Topology

NVLink Topology

NVLink Tables

Detailed tables with properties for each NVLink.

Launch Statistics

Summary of the configuration used to launch the kernel. The launch configuration defines the size of the kernel grid, the division of the grid into blocks, and the GPU resources needed to execute the kernel. Choosing an efficient launch configuration maximizes device utilization.

Grid Size	7813 (+0.00%)	Registers Per Thread [register/thread]	94 (+135.00%)
Block Size	128 (+0.00%)	Static Shared Memory Per Block [byte/block]	32 (+0.00%)
Threads [thread]	1000064 (+0.00%)	Dynamic Shared Memory Per Block [byte/block]	0 (+0.00%)
Waves Per SM	21.70 (+60.00%)	Driver Shared Memory Per Block [byte/block]	0 (+0.00%)
Function Cache Configuration	cudaFuncCachePreferNone (cudaFuncCachePreferNone)	Shared Memory Configuration Size [Kbyte]	32.77 (+0.00%)

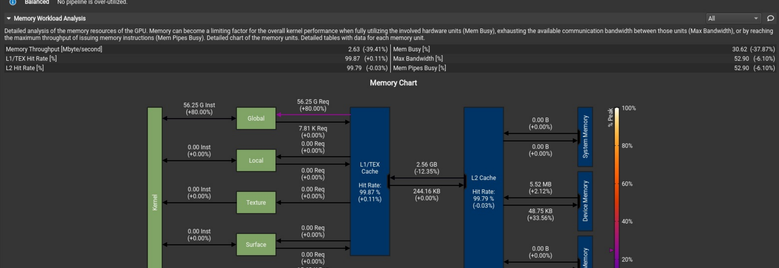
Occupancy

Occupancy is the ratio of the number of active warps per multiprocessor to the maximum number of possible active warps. Another way to view occupancy is the percentage of the hardware's ability to process warps that is actively in use. Higher occupancy does not always result in higher performance, however, low occupancy always reduces the ability to hide latencies, resulting in overall performance degradation. Large discrepancies between the theoretical and the achieved occupancy during execution typically indicates highly imbalanced workloads.

Theoretical Occupancy [%]	62.50 (-37.50%)	Block Limit Registers [block]	5 (-58.33%)
Theoretical Active Warps per SM [warp]	20 (-37.50%)	Block Limit Shared Mem [block]	256 (+0.00%)
Achieved Occupancy [%]	61.59 (-37.05%)	Block Limit Warps [block]	8 (+0.00%)
Achieved Active Warps Per SM [warp]	19.71 (-37.05%)	Block Limit SM [block]	16 (+0.00%)

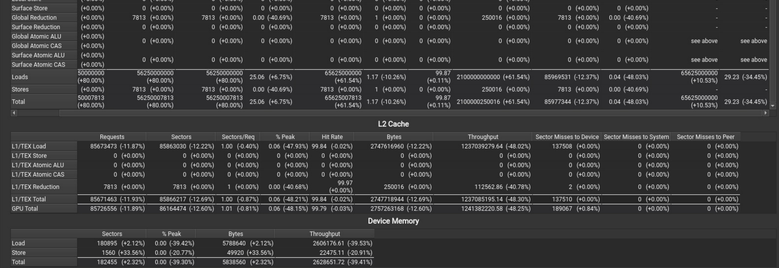
Occupancy Limiters This kernel's theoretical occupancy (62.5%) is limited by the number of required registers

Impact of Varying Register Count Per Thread



Impact of Varying Block Size

Impact of Varying Shared Memory Usage Per Block



Source Counters

Source metrics, including branch efficiency and sampled warp stall reasons. Sampling Data metrics are periodically sampled over the kernel runtime. They indicate when warps were stalled and couldn't be scheduled. See the documentation for a description of all stall reasons. Only focus on stalls if the schedulers fail to issue every cycle.

Branch Instructions [inst]	8750085873 (+133.33%)	Branch Efficiency [%]	100.00 (+0.00%)
Branch Instructions Ratio [%]	0.19 (+49.65%)	Avg. Divergent Branches	1152 (+10.40%)