Page: Details | Process: All | Launch: 7 - 799 - computeIntersections | Add Baseline | Apply Rules | Save as PDF

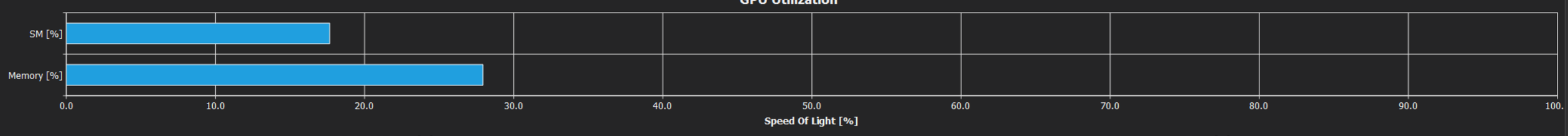
Current 799 - computeIntersections (119424, 1, 1) **Time:** 57.94 msecond **Cycles:** 51,679,618 **Regs:** 66 **GPU:** GeForce RTX 2070 with Max-Q Design **SM Frequency:** 891,972,680.49 cycle **CC:** 7.5 **Process:** [21456] cis565_path_tracer.exe

GPU Speed Of Light

High-level overview of the utilization for compute and memory resources of the GPU. For each unit, the Speed Of Light (SOL) reports the achieved percentage of utilization with respect to the theoretical maximum.

SOL SM [%]	17.66	Duration [msecond]	57.9
SOL Memory [%]	27.94	Elapsed Cycles [cycle]	51,679,61
SOL TEX [%]	25.85	SM Active Cycles [cycle]	50,887,83
SOL L2 [%]	26.27	SM Frequency [cycle]	891,972,680.4
SOL FB [%]	27.94	Memory Frequency [cycle]	1,360,668,956.9

GPU Utilization



SOL SM Breakdown

SOL SM: Pipe Alu Cycles Active [%]	17.66
SOL SM: Issue Active [%]	11.75
SOL SM: Inst Executed [%]	11.75
SOL SM: Inst Executed Pipe Lsu [%]	11.07
SOL SM: Mio Pq Read Cycles Active [%]	4.61
SOL SM: Mio Pq Write Cycles Active [%]	4.61
SOL SM: Mio Inst Issued [%]	3.97
SOL SM: Inst Executed Pipe Cbu Pred On Any [%]	3.17
SOL SM: Mio2rf Writeback Active [%]	1.65
SOL SM: Inst Executed Pipe Adu [%]	0.83
SOL SM: Pipe Fma Cycles Active [%]	0.32
SOL SM: Inst Executed Pipe Xu [%]	0.06
SOL IDC: Request Cycles Active [%]	0.00
SOL SM: Inst Executed Pipe Fp16 [%]	0
SOL SM: Inst Executed Pipe Ipa [%]	0
SOL SM: Inst Executed Pipe Tex [%]	0
SOL SM: Inst Executed Pipe Uniform [%]	0
SOL SM: Pipe Fp64 Cycles Active [%]	0
SOL SM: Pipe Shared Cycles Active [%]	0
SOL SM: Pipe Tensor Cycles Active [%]	0

SOL Memory Breakdown

SOL GPU: Dram Throughput [%]	27.94
SOL L2: T Sectors [%]	26.27
SOL L2: Xbar2lts Cycles Active [%]	17.34
SOL L2: Lts2xbar Cycles Active [%]	14.03
SOL L1: M L1tex2xbar Req Cycles Active [%]	12.93
SOL L1: Lsuin Requests [%]	11.07
SOL L2: D Sectors [%]	9.28
SOL L2: T Tag Requests [%]	7.23
SOL L1: Data Pipe Lsu Wavefronts [%]	6.25
SOL L1: M Xbar2l1tex Read Sectors [%]	4.21
SOL L2: D Sectors Fill Device [%]	3.82
SOL L1: Lsu Writeback Active [%]	2.88
SOL L1: Data Bank Reads [%]	1.89
SOL L1: Data Bank Writes [%]	1.52
SOL L1: F Wavefronts [%]	0.00
SOL L1: Texin Sm2tex Req Cycles Active [%]	0.00
SOL L1: Data Pipe Tex Wavefronts [%]	0
SOL L1: Tex Writeback Active [%]	0
SOL L2: D Atomic Input Cycles Active [%]	0
SOL L2: D Sectors Fill Sysmem [%]	0

Recommendations

Bottleneck High-level bottleneck detection — Apply

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.

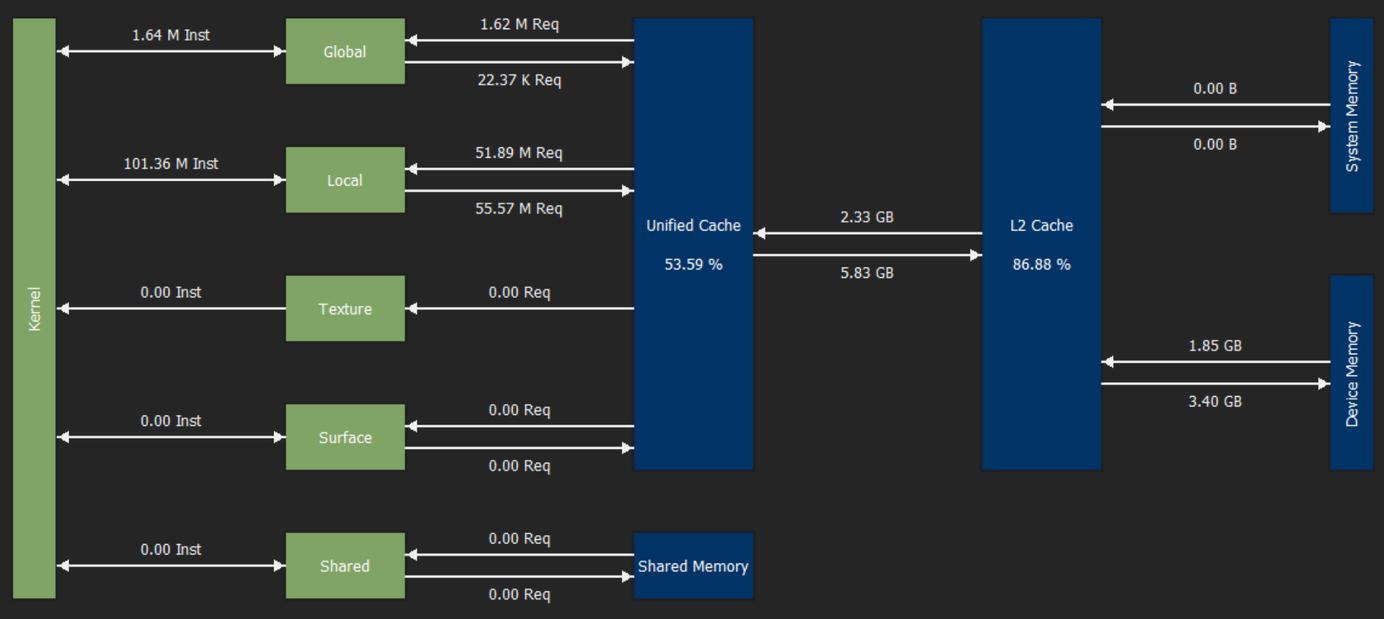
Executed Ipc Elapsed [cycle]	0.47	SM Busy [%]	17.9
Executed Ipc Active [cycle]	0.48	Issue Slots Busy [%]	11.9
Issued Ipc Active [cycle]	0.48	-	

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 [Gbyte]	97.32	Mem Busy [%]	26.2
L1 Hit Rate [%]	53.59	Max Bandwidth [%]	27.9
L2 Hit Rate [%]	86.88	Mem Pipes Busy [%]	11.0

Memory Chart



Shared Memory

	Instructions	Requests	% Peak	Bank Conflicts
Shared Load	0	0	0	0
Shared Store	0	0	0	0
Shared Atomic	0	-	-	-
Total	0	0	0	0

First-Level (Unified) Cache

	Instructions	SM->TEX Requests	% Peak	Hit Rate	TEX->L2 Requests	% Peak	L2->TEX Returns	% Peak	TEX->SM Returns	% Peak
Global Load Cached	1,615,523	1,615,523	0.09	85.56	-	-				
Global Load Uncached	-	-	-	-	-	-	78,261,327	4.27	53,646,690	2.93
Local Load Cached	49,162,618	51,886,621	2.83	56.96	-	-				
Local Load Uncached	-	-	-	-	-	-				
Surface Load	0	0	0	0	-	-	0	0	0	0
Texture Load	0	0	0	0	-	-	0	0		
Global Store	22,373	22,373	0.00	82.58	195,634,223	10.68	-	-	-	-
Local Store	52,196,677	55,573,657	3.03	49.93			-	-	-	-
Surface Store	0	0	0	0	0	0	-	-	-	-
Global Reduction	0	0	0	0	0	0	-	-	-	-
Surface Reduction	0	0	0	0	0	0	-	-	-	-
Global Atomic	0	0	0	0	0	0	0	0	see above	see above
Global Atomic Cas	0									
Surface Atomic	0	0	0	0	0	0	0	0	see above	see above
Surface Atomic Cas	0									
Loads	50,778,141	53,502,144	2.92	57.41	-	-	78,261,327	4.27	53,646,690	2.93
Stores	52,219,050	55,596,030	3.03	50.06	195,634,223	10.68	-	-	-	-
Total	102,997,191	109,098,174	5.96	53.62	195,634,223	10.68	78,261,327	4.27	53,646,690	2.93

Second-Level (L2) Cache

	TEX->L2 Requests	% Peak	L2->TEX Returns	% Peak	Total Bytes	Total Throughput
Global Load Cached	-	-			2,504,362,464	43,224,451,280.81
Global Load Uncached	-	-	78,261,327	4.27	-	-
Local Load Cached	-	-			-	-
Local Load Uncached	-	-			-	-
Surface Load	-	-	0	0	0	0
Texture Load	-	-	0	0	0	0
Global Store	195,634,223	10.68	-	-	6,260,295,136	108,050,582,133.90
Local Store			-	-		
Surface Store	0	0	-	-	0	0
Global Reduction	0	0	-	-	0	0
Surface Reduction	0	0	-	-	0	0
Global Atomic	0	0	0	0	0	0
Global Atomic Cas						
Surface Atomic	0	0	0	0	0	0
Surface Atomic Cas						
Loads	-	-	78,261,327	4.27	2,504,362,464	43,224,451,280.81
Stores	195,634,223	10.68	-	-	6,260,295,136	108,050,582,133.90
Total	195,634,223	10.68	78,261,327	4.27	8,764,657,600	151,275,033,414.71

Device Memory (FB)

	L2<->FB Sectors	% Peak	Bytes	Throughput
Load	62,221,700	9.87	1,991,094,400	34,365,617,647.38
Store	113,980,035	18.07	3,647,361,120	62,952,222,492.24
Total	176,201,735	27.94	5,638,455,520	97,317,840,139.62

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 [cycle]	6.33	Instructions Per Active Issue Slot [inst/issue]	
Eligible Warps Per Scheduler [cycle]	0.12	No Eligible [%]	88.0
Issued Warp Per Scheduler [cycle]	0.12	One or More Eligible [%]	11.9

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.

Warp Cycles Per Issued Instruction [warp/issue]	52.79	Avg. Active Threads Per Warp [thread/inst]	21.7
Warp Cycles Per Issue Active [warp/issue]	52.79	Avg. Not Predicated Off Threads Per Warp [thread/inst]	21.6
Warp Cycles Per Executed Instruction [warp/inst]	52.79	-	

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]	874,086,260	Avg. Executed Instructions Per Scheduler [inst]	6,070,043.4
Issued Instructions [inst]	874,142,842	Avg. Issued Instructions Per Scheduler [inst]	6,070,436.4

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	933	Registers Per Thread [register/thread]	6
Block Size	128	Static Shared Memory Per Block [byte/block]	
Threads [thread]	119,424	Dynamic Shared Memory Per Block [byte/block]	
Waves Per SM	3.70	Shared Memory Configuration Size [Kbyte]	32.7

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 [%]	87.50	Block Limit Registers [block]	
Theoretical Active Warps per SM [warp/cycle]	28	Block Limit Shared Mem [block]	1
Achieved Occupancy [%]	78.77	Block Limit Warps [block]	
Achieved Active Warps Per SM [cycle]	25.21	Block Limit SM [block]	1