Page: Details | Result: 0 - 130 - _Z16matmul_pi | Add Baseline | Apply Rules | Occupancy Calculator | Save as PDF

Result		Time	Cycles	Regs	GPU	SM Frequency	CC	Process
Current	130 - _Z16matmul_pipelined13CuDeviceArrayI7Float32Li2ELi1EES_IS0_Li2ELi1EES_IS0_Li2ELi1EES_IS0_Li2ELi1EE11ElementwiseI8identityES1_IS2_ES1_IS2…	2.33 msecond	2,482,301	103	0 - NVIDIA A800 80GB PCIe	1.06 cycle/nsecond	8.0	[1832454] julia

GPU Speed Of Light Throughput

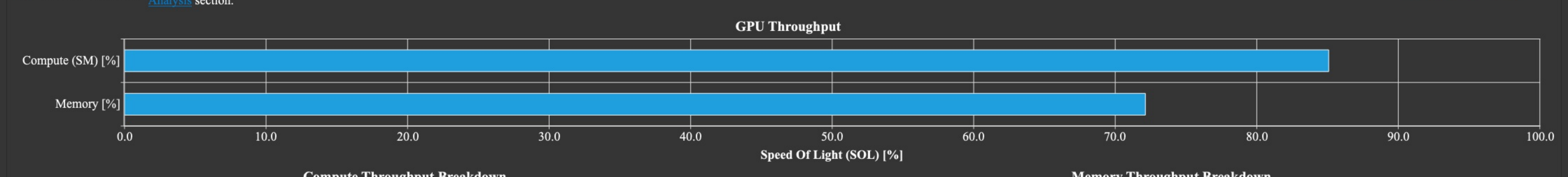
All

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 [%]	85.08	Duration [msecond]	2.33
Memory Throughput [%]	72.13	Elapsed Cycles [cycle]	2482301
L1/TEX Cache Throughput [%]	73.17	SM Active Cycles [cycle]	2444434.78
L2 Cache Throughput [%]	11.15	SM Frequency [cycle/nsecond]	1.06
DRAM Throughput [%]	4.69	DRAM Frequency [cycle/nsecond]	1.51

High Throughput The kernel is utilizing greater than 80.0% of the available compute or memory performance of the device. To further improve performance, work will likely need to be shifted from the most utilized to another unit. Start by analyzing workloads in the Compute Workload Analysis section.

GPU Throughput



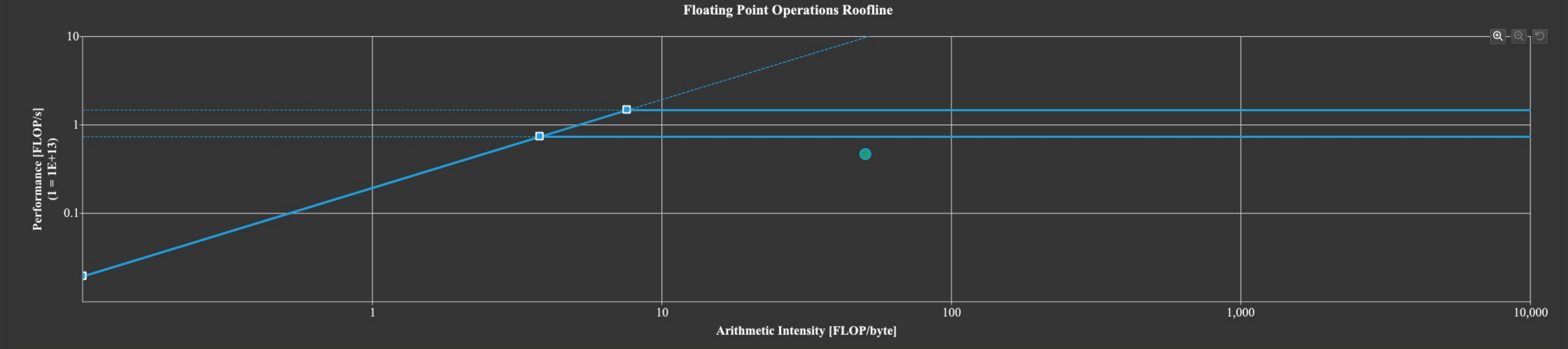
Compute Throughput Breakdown

SM: Issue Active [%]	85.08
SM: Inst Executed [%]	85.02
SM: Pipe Alu Cycles Active [%]	69.95
SM: Pipe Fma Cycles Active [%]	69.07
SM: Inst Executed Pipe Lsu [%]	49.19
SM: Mio2rf Writeback Active [%]	33.41
SM: Mio Inst Issued [%]	24.97
SM: Inst Executed Pipe Uniform [%]	4.34
SM: Mio Pq Write Cycles Active [%]	4.16
SM: Mio Pq Read Cycles Active [%]	4.16
SM: Inst Executed Pipe Adu [%]	1.51
SM: Inst Executed Pipe Cbu Pred On Any [%]	0.26
IDC: Request Cycles Active [%]	0
SM: Inst Executed Pipe Xu [%]	0
SM: Inst Executed Pipe Tex [%]	0
SM: Inst Executed Pipe Ipa [%]	0
SM: Inst Executed Pipe Fp16 [%]	0
SM: Pipe Fp64 Cycles Active [%]	0
SM: Pipe Shared Cycles Active [%]	0
SM: Pipe Tensor Cycles Active [%]	0

Memory Throughput Breakdown

L1: Data Pipe Lsu Wavefronts [%]	72.13
L1: Lsu Writeback Active [%]	66.81
L1: Lsuin Requests [%]	49.19
L2: Lts2xbar Cycles Active [%]	11.15
L1: Data Bank Reads [%]	9.92
L2: T Sectors [%]	9.48
L2: D Sectors [%]	7.95
L1: M Xbar2l1tex Read Sectors [%]	7.92
L2: T Tag Requests [%]	7.19
DRAM: Cycles Active [%]	4.69
L2: Xbar2lts Cycles Active [%]	4.54
L1: Data Bank Writes [%]	4.06
DRAM: Dram Sectors [%]	3.48
L1: M L1tex2xbar Req Cycles Active [%]	3.22
L2: D Sectors Fill Device [%]	3.08
L1: Texin Sm2tex Req Cycles Active [%]	0.09
L1: F Wavefronts [%]	0.00
L2: D Sectors Fill Sysmem [%]	0.00
L2: D Atomic Input Cycles Active [%]	0
L1: Tex Writeback Active [%]	0
L1: Data Pipe Tex Wavefronts [%]	0

Floating Point Operations Roofline



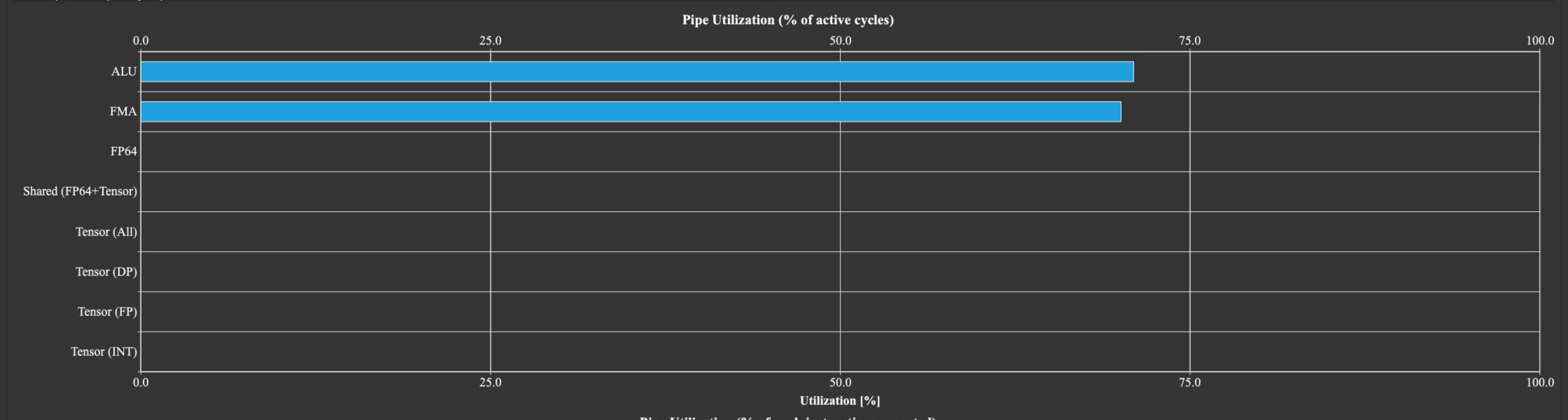
Compute Workload Analysis

All

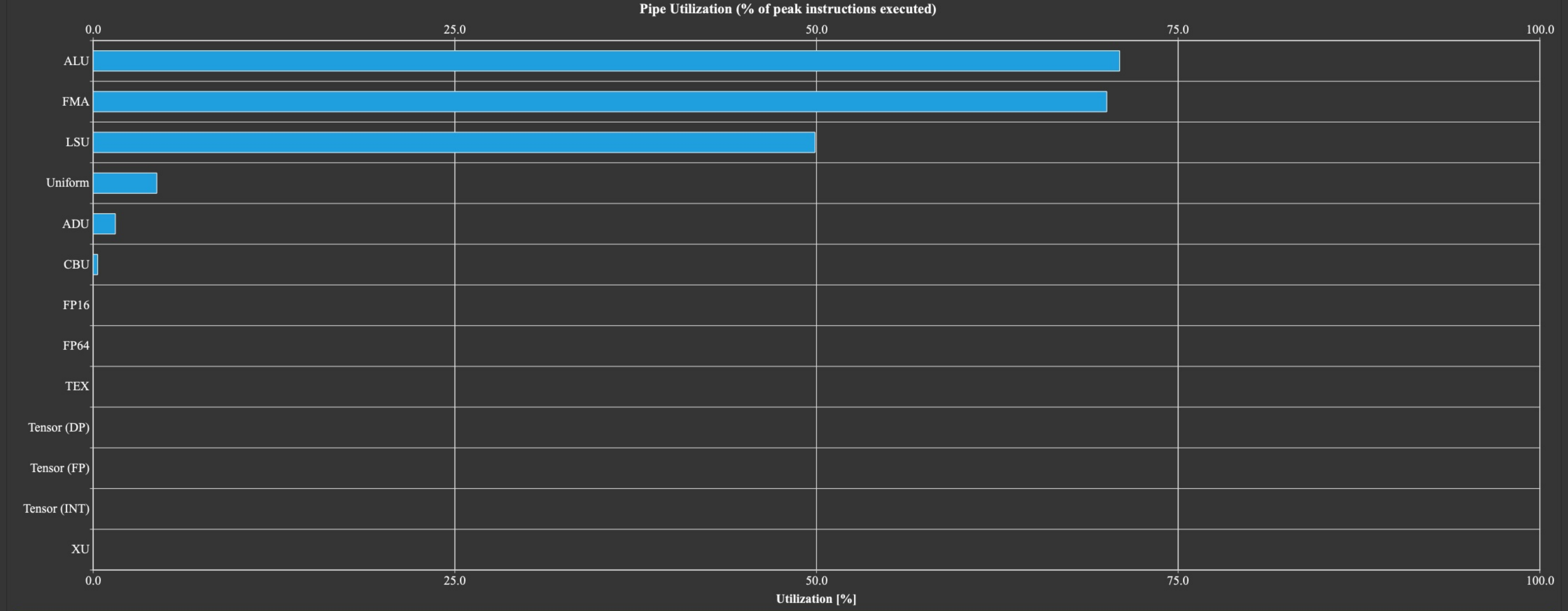
Detailed analysis of the compute resources of the streaming multiprocessors (SM), including the achieved instructions per clock (IPC) and the utilization of each available pipeline. Pipelines with very high utilization might limit the overall performance.

Executed Ipc Elapsed [inst/cycle]	3.40	SM Busy [%]	86.31
Executed Ipc Active [inst/cycle]	3.45	Issue Slots Busy [%]	86.31
Issued Ipc Active [inst/cycle]	3.45		

Pipe Utilization (% of active cycles)



Pipe Utilization (% of peak instructions executed)



Memory Workload Analysis

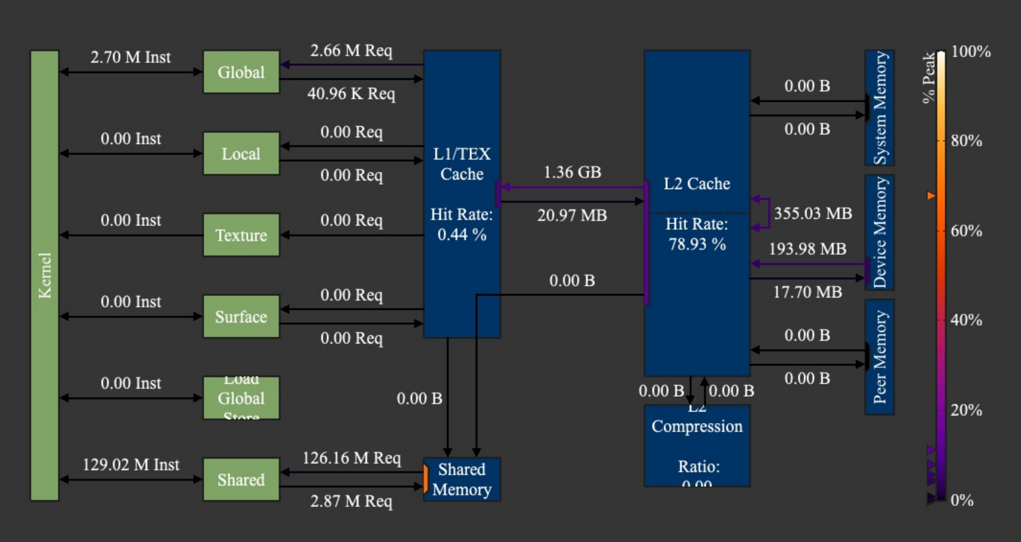
All

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/second]	90.79	Mem Busy [%]	72.13
L1/TEX Hit Rate [%]	0.44	Max Bandwidth [%]	66.81
L2 Hit Rate [%]	78.93	Mem Pipes Busy [%]	49.19
L2 Compression Success Rate [%]	0	L2 Compression Ratio	0

Memory Chart

Show As: Transfer Size



Shared Memory

	Instructions	Requests	Wavefronts	% Peak	Bank Conflicts
Shared Load	126156800	126156800	169419501	63.26	1150595
Shared Load Matrix	0	0			
Shared Store	2867200	2867200	12176696	4.55	1198257
Shared Store From Global Load	0	0	0	0	989
Shared Atomic	0	0	0	0	0
Other	-	-	260975	0.10	0
Total	129024000	129024000	181857172	67.90	2349841

L1/TEX Cache

	Instructions	Requests	Wavefronts	% Peak	Sectors	Sectors/Req	Hit Rate	Bytes	Sector Misses to L2	% Peak to L2	Returns to SM	% Peak to SM
Local Load	0	0	0	0	0	0	0	0			10654639	3.98
Global Load	2662400	2662400	4109317	1.53	42598400	16	0.45	1363148800	42403612	7.92		
Global Load To Shared Store (access)	0	0			0	0		0			-	-
Global Load To Shared Store (bypass)	0	0			0	0	-	0	0	0	-	-
Surface Load	0	0	0	0	0	0	0	0	0	0	0	0
Texture Load	0	0	0	0	0	0	0	0	0	0		
Global Store	40960	40960	68524	0.03	655360	16	0	20971520	655360	0.24	-	-
Local Store	0	0	0	0	0	0	0	0			-	-
Surface Store	0	0	0	0	0	0	0	0	0	0	-	-
Global Reduction	0	0	0	0	0	0	0	0	0	0	-	-
Surface Reduction	0	0	0	0	0	0	0	0	0	0	-	-
Global Atomic ALU	0	0	0	0	0	0	0	0	0	0	see above	see above
Global Atomic CAS	0											
Surface Atomic ALU	0	0	0	0	0	0	0	0	0	0	see above	see above
Surface Atomic CAS	0											
Loads	2662400	2662400	4109317	1.53	42598400	16	0.45	1363148800	42403612	7.92	10654639	3.98
Stores	40960	40960	68524	0.03	655360	16	0	20971520	655360	0.24	-	-
Atomics & Reductions	0	0	0	0	0	0	0	0	0	0	-	-
Total	2703360	2703360	4177841	1.56	43253760	16	0.45	1384120320	43058972	8.16	10654639	3.98

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	10601055	42404492	4.00	11.15	78.68	1356943744	581999615701.35	8906024	0	0
L1/TEX Store	163840	655360	4	0.23	100	20971520	8994784518.25	0	0	0
L1/TEX Atomic ALU	0	0	0	0	0	0	0	0	0	0
L1/TEX Atomic CAS	0	0	0	0	0	0	0	0	0	0
L1/TEX Reduction	0	0	0	0	0	0	0	0	0	0
L1/TEX Total	10765267	43060396	4.00	11.32	79.05	1377932672	591001866593.47	9079564	0	0
ECC Total	-	19	-	0.00	-	608	260774.09	19	-	-
L2 Fabric Total	2777303	11094690	3.99	5.83	77.06	355030080	152274087290.69	2555744	0	0
GPU Total	13548343	54059770	3.99	9.48	79.07	1729912640	741967746362.89	11085163	2	0

L2 Cache Eviction Policies

	First	Hit Rate	Last	Hit Rate	Normal	Hit Rate	Normal Demote	Hit Rate
L1/TEX Load	0	0	0	0	41814620	79.71	0	0
L1/TEX Store	0	0	0	0	655360	100	0	0
L1/TEX Atomic	0	0	0	0	0	0	-	-
L1/TEX Total	0	0	0	0	43156428	79.56	0	0
L2 Fabric Total	0	0	0	0	5851880	56.20	0	0
GPU Total	12662	99.84	0	0	49160281	76.04	0	0

Device Memory

	Sectors	% Peak	Bytes	Throughput
Load	6061732	4.30	193975424	83196980510.57
Store	553092	0.39	17698944	7591161130.94
Total	6614824	4.69	211674368	90788141641.50

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]	3.88	No Eligible [%]	13.68
Eligible Warps Per Scheduler [warp]	2.13	One or More Eligible [%]	86.32
Issued Warp Per Scheduler	0.86		

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.

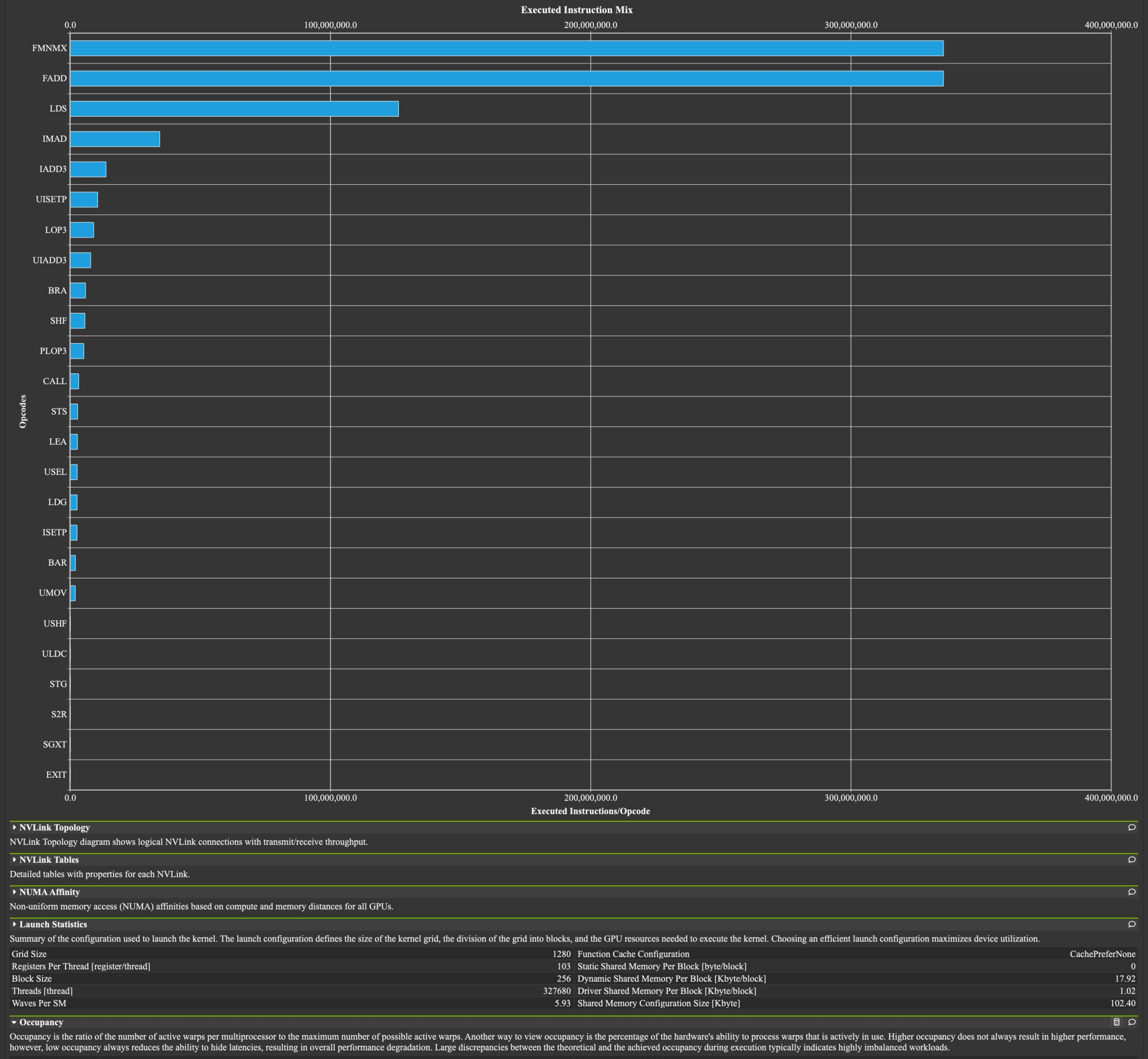
Warp Cycles Per Issued Instruction [cycle]	4.49	Avg. Active Threads Per Warp	32
Warp Cycles Per Executed Instruction [cycle]	4.50	Avg. Not Predicated Off Threads Per Warp	31.86

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]	910786560	Avg. Executed Instructions Per Scheduler [inst]	2108302.22
Issued Instructions [inst]	911477306	Avg. Issued Instructions Per Scheduler [inst]	2109901.17

Executed Instruction Mix



NVLink Topology

NVLink Topology diagram shows logical NVLink connections with transmit/receive throughput.

NVLink Tables

Detailed tables with properties for each NVLink.

NUMA Affinity

Non-uniform memory access (NUMA) affinities based on compute and memory distances for all GPUs.

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	1280	Function Cache Configuration	CachePreferNone
Registers Per Thread [register/thread]	103	Static Shared Memory Per Block [byte/block]	0
Block Size	256	Dynamic Shared Memory Per Block [Kbyte/block]	17.92
Threads [thread]	327680	Driver Shared Memory Per Block [Kbyte/block]	1.02
Waves Per SM	5.93	Shared Memory Configuration Size [Kbyte]	102.40

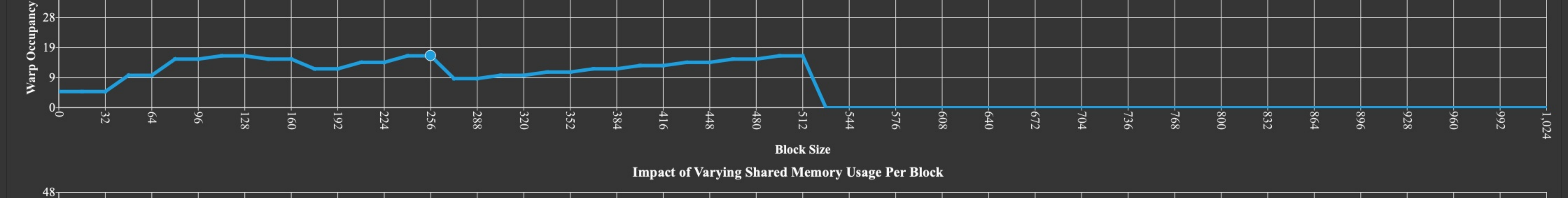
Occupancy

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.

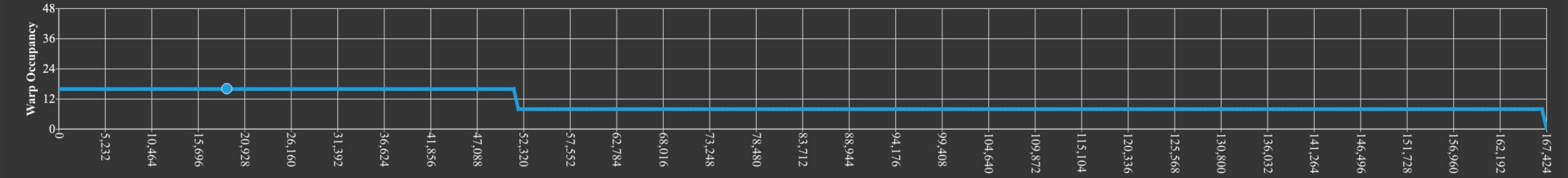
Theoretical Occupancy [%]	25	Block Limit Registers [block]	2
Theoretical Active Warps per SM [warp]	16	Block Limit Shared Mem [block]	5
Achieved Occupancy [%]	24.24	Block Limit Warps [block]	8
Achieved Active Warps Per SM [warp]	15.51	Block Limit SM [block]	32

Occupancy Limiters This kernel's theoretical occupancy (25.0%) is limited by the number of required registers. See the CUDA Best Practices Guide for more details on optimizing occupancy.

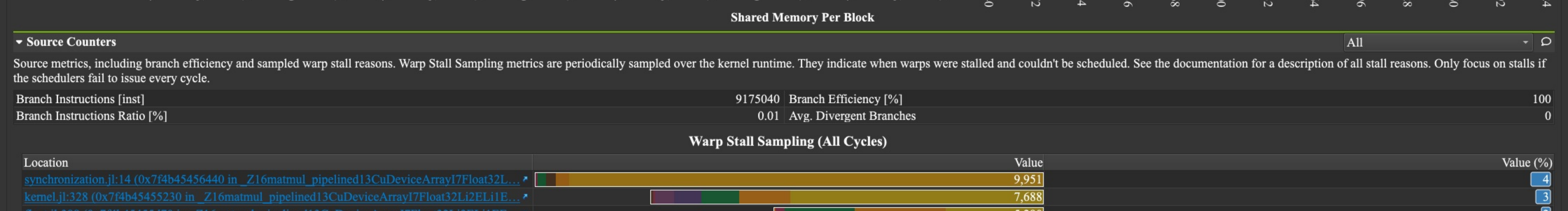
Impact of Varying Register Count Per Thread



Impact of Varying Block Size



Impact of Varying Shared Memory Usage Per Block



Source Counters

All

Source metrics, including branch efficiency and sampled warp stall reasons. Warp Stall Sampling 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]	9175040	Branch Efficiency [%]	100
Branch Instructions Ratio [%]	0.01	Avg. Divergent Branches	0

Warp Stall Sampling (All Cycles)

Location	Value	Value (%)
synchronization.jl:14 (0x7f4b45456440 in _Z16matmul_pipelined13CuDeviceArrayI7Float32L…	9,951	4
kernel.jl:328 (0x7f4b45455230 in _Z16matmul_pipelined13CuDeviceArrayI7Float32Li2ELi1E…	7,688	3
float.jl:398 (0x7f4b45455d70 in _Z16matmul_pipelined13CuDeviceArrayI7Float32Li2ELi1EE…	5,290	2
synchronization.jl:14 (0x7f4b454564b0 in _Z16matmul_pipelined13CuDeviceArrayI7Float32L…	4,413	2
float.jl:398 (0x7f4b45455810 in _Z16matmul_pipelined13CuDeviceArrayI7Float32Li2ELi1EE…	4,116	2

Most Instructions Executed

Location	Value	Value (%)
kernel.jl:328 (0x7f4b45456760 in _Z16matmul_pipelined13CuDeviceArrayI7Float32Li2ELi1E…	2,621,440	0
base.jl:44 (0x7f4b45456750 in _Z16matmul_pipelined13CuDeviceArrayI7Float32Li2ELi1EES…	2,621,440	0
base.jl:44 (0x7f4b45456740 in _Z16matmul_pipelined13CuDeviceArrayI7Float32Li2ELi1EES…	2,621,440	0
base.jl:44 (0x7f4b45456730 in _Z16matmul_pipelined13CuDeviceArrayI7Float32Li2ELi1EES…	2,621,440	0
base.jl:44 (0x7f4b45456720 in _Z16matmul_pipelined13CuDeviceArrayI7Float32Li2ELi1EES…	2,621,440	0

Follow the *rules outputs* to get guidance on how to navigate through the report and quickly discover performance bottlenecks in this kernel.
You could also disable individual sections to focus on selected performance aspects and make profiling faster.