

Analysis Report

gpu4_column(int, int, int, double*, double*, double*)

Duration	90.723 ms (90,722,824 ns)
Grid Size	[257,9,1]
Block Size	[16,16,1]
Registers/Thread	94
Shared Memory/Block	0 B
Shared Memory Executed	0 B
Shared Memory Bank Size	4 B

[0] Tesla V100-PCIE-16GB

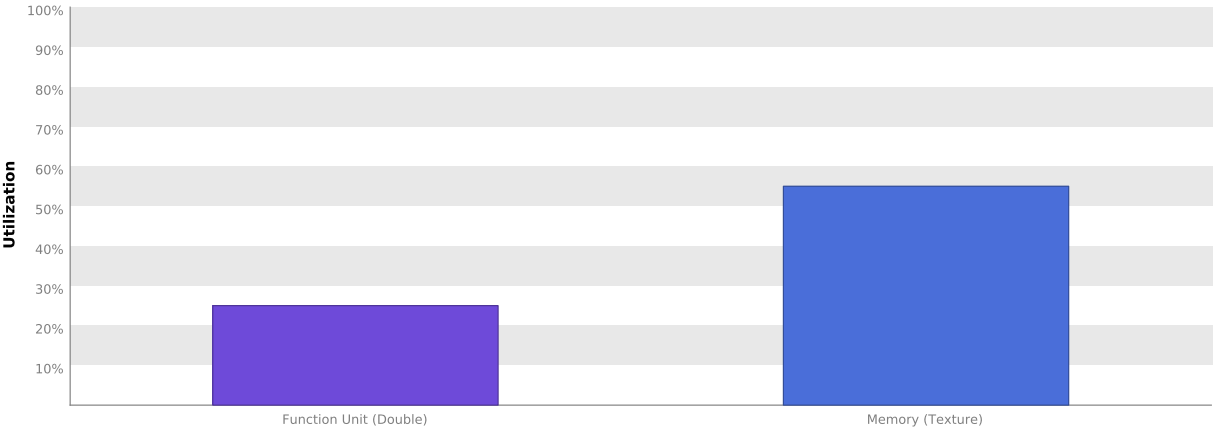
GPU UUID	GPU-297db011-cee1-e4b7-e4ef-f0bd9df9979a
Compute Capability	7.0
Max. Threads per Block	1024
Max. Threads per Multiprocessor	2048
Max. Shared Memory per Block	48 KiB
Max. Shared Memory per Multiprocessor	96 KiB
Max. Registers per Block	65536
Max. Registers per Multiprocessor	65536
Max. Grid Dimensions	[2147483647, 65535, 65535]
Max. Block Dimensions	[1024, 1024, 64]
Max. Warps per Multiprocessor	64
Max. Blocks per Multiprocessor	32
Half Precision FLOP/s	28.262 TeraFLOP/s
Single Precision FLOP/s	14.131 TeraFLOP/s
Double Precision FLOP/s	7.066 TeraFLOP/s
Number of Multiprocessors	80
Multiprocessor Clock Rate	1.38 GHz
Concurrent Kernel	true
Max IPC	4
Threads per Warp	32
Global Memory Bandwidth	898.048 GB/s
Global Memory Size	15.752 GiB
Constant Memory Size	64 KiB
L2 Cache Size	6 MiB
Memcpy Engines	7
PCIe Generation	3
PCIe Link Rate	8 Gbit/s
PCIe Link Width	16

1. Compute, Bandwidth, or Latency Bound

The first step in analyzing an individual kernel is to determine if the performance of the kernel is bounded by computation, memory bandwidth, or instruction/memory latency. The results below indicate that the performance of kernel "gpu4_column" is most likely limited by instruction and memory latency. You should first examine the information in the "Instruction And Memory Latency" section to determine how it is limiting performance.

1.1. Kernel Performance Is Bound By Instruction And Memory Latency

This kernel exhibits low compute throughput and memory bandwidth utilization relative to the peak performance of "Tesla V100-PCIE-16GB". These utilization levels indicate that the performance of the kernel is most likely limited by the latency of arithmetic or memory operations. Achieved compute throughput and/or memory bandwidth below 60% of peak typically indicates latency issues.



2. Instruction and Memory Latency

Instruction and memory latency limit the performance of a kernel when the GPU does not have enough work to keep busy. The results below indicate that the GPU does not have enough work because instruction execution is stalling excessively.

2.1. Kernel Profile - PC Sampling

The Kernel Profile - PC Sampling gives the number of samples for each source and assembly line with various stall reasons. The samples are collected at a period of 2048 [2¹¹] cycles. You can change the period under Settings->Analysis tab. The allowed values are from 5 to 31. Increasing the period would reduce the number of samples collected.

Using this information you can pinpoint portions of your kernel that are introducing latencies and the reason for the latency. Samples are taken in round robin order for all active warps at a fixed number of cycles regardless of whether the warp is issuing an instruction or not.

Instruction Issued - Warp was issued

Instruction Fetch - The next assembly instruction has not yet been fetched.

Execution Dependency - An input required by the instruction is not yet available. Execution dependency stalls can potentially be reduced by increasing instruction-level parallelism.

Memory Dependency - A load/store cannot be made because the required resources are not available or are fully utilized, or too many requests of a given type are outstanding. Data request stalls can potentially be reduced by optimizing memory alignment and access patterns.

Texture - The texture sub-system is fully utilized or has too many outstanding requests.

Synchronization - The warp is blocked at a __syncthreads() call.

Constant - A constant load is blocked due to a miss in the constants cache.

Pipe Busy - The compute resource(s) required by the instruction is not yet available.

Memory Throttle - Large number of pending memory operations prevent further forward progress. These can be reduced by combining several memory transactions into one.

Not Selected - Warp was ready to issue, but some other warp issued instead. You may be able to sacrifice occupancy without impacting latency hiding and doing so may help improve cache hit rates.

Other - The warp is blocked for an uncommon reason.

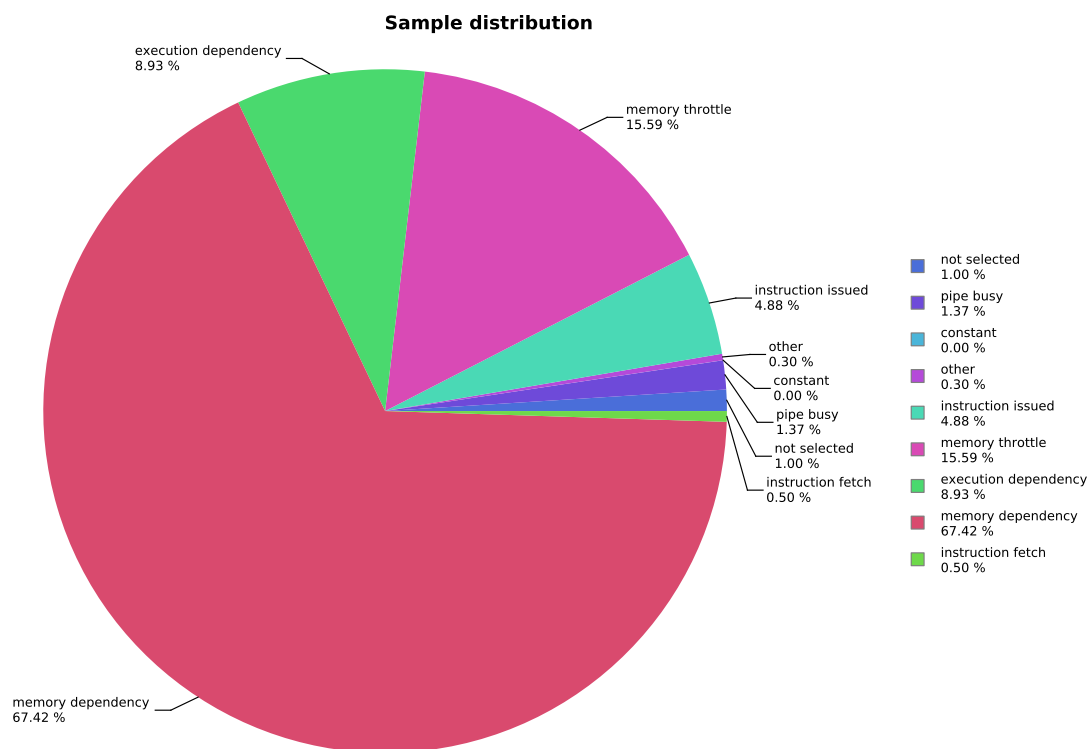
Sleeping -The warp is blocked, yielded or sleeping.

Examine portions of the kernel that have high number of samples to know where the maximum time was spent and observe the latency reasons for those samples to identify optimization opportunities.

Cuda Functions	Sample Count	% of Kernel Samples
gpu4_column(int, int, int, double*, double*, double*)	4307221	100.0

Source Files :

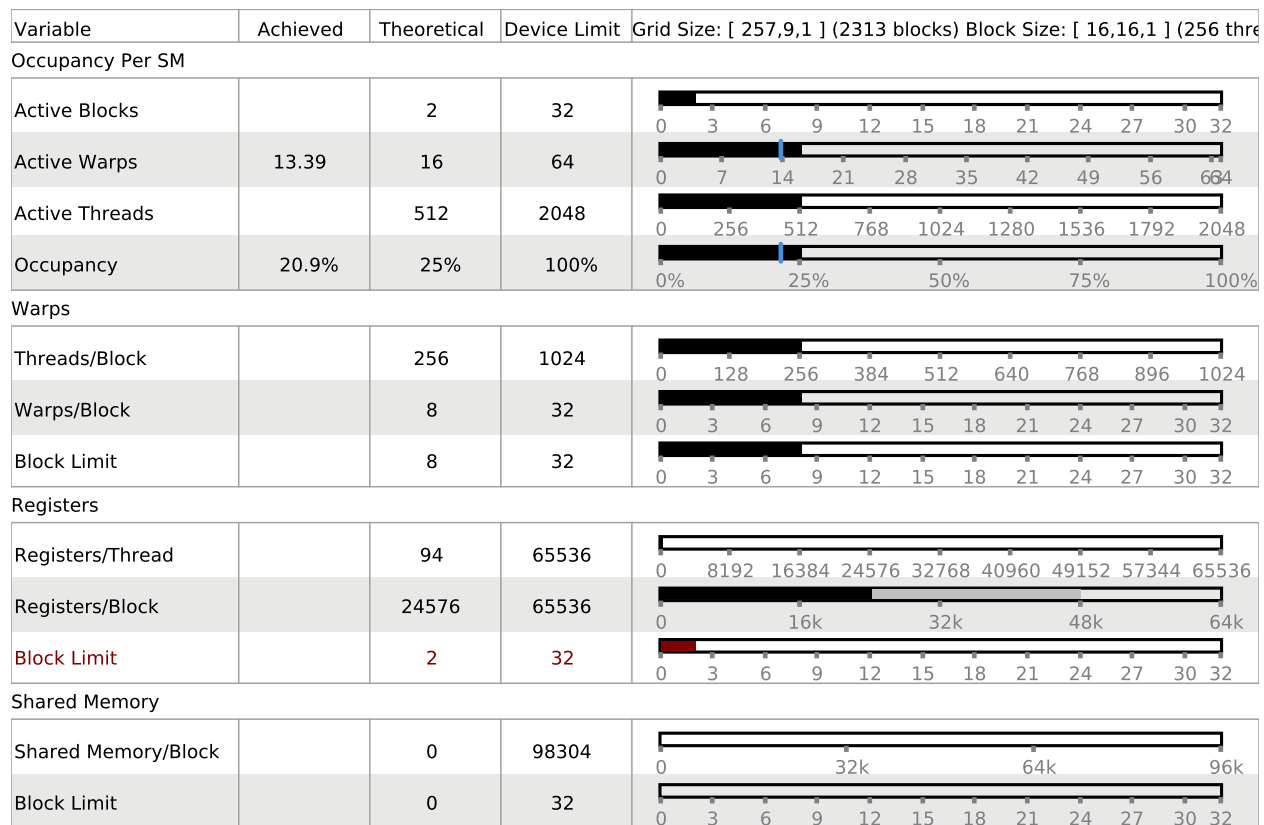
/zhome/a8/6/114633/hpc/week3/ass3/hpc-gpu/matmult/matmult_gpu4.cu



2.2. GPU Utilization Is Limited By Register Usage

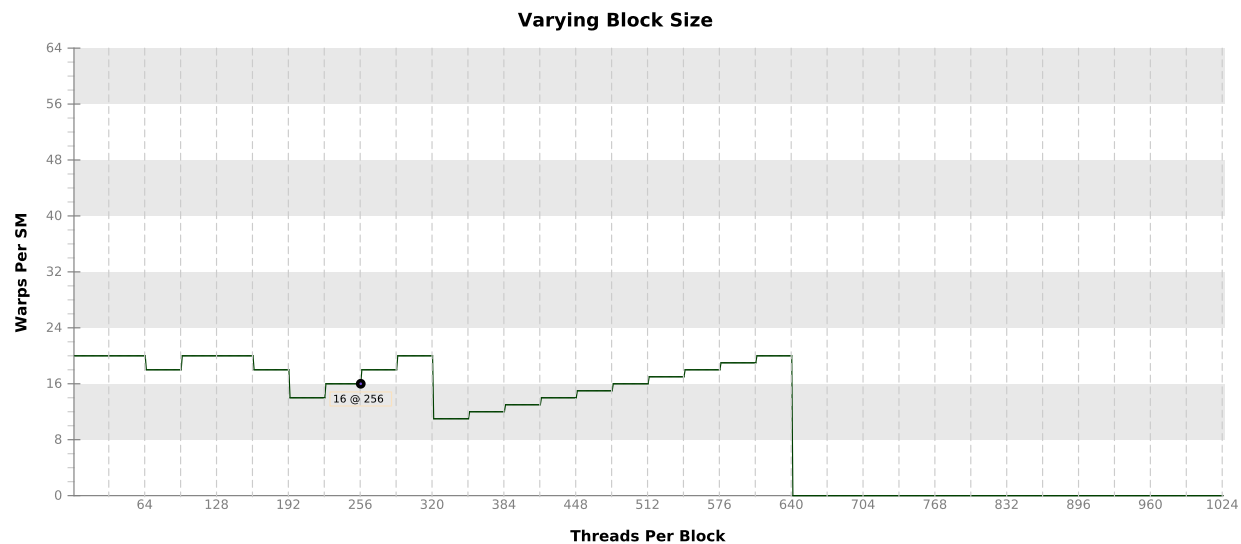
The kernel uses 94 registers for each thread (24064 registers for each block). This register usage is likely preventing the kernel from fully utilizing the GPU. Device "Tesla V100-PCIE-16GB" provides up to 65536 registers for each block. Because the kernel uses 24064 registers for each block each SM is limited to simultaneously executing 2 blocks (16 warps). Chart "Varying Register Count" below shows how changing register usage will change the number of blocks that can execute on each SM.

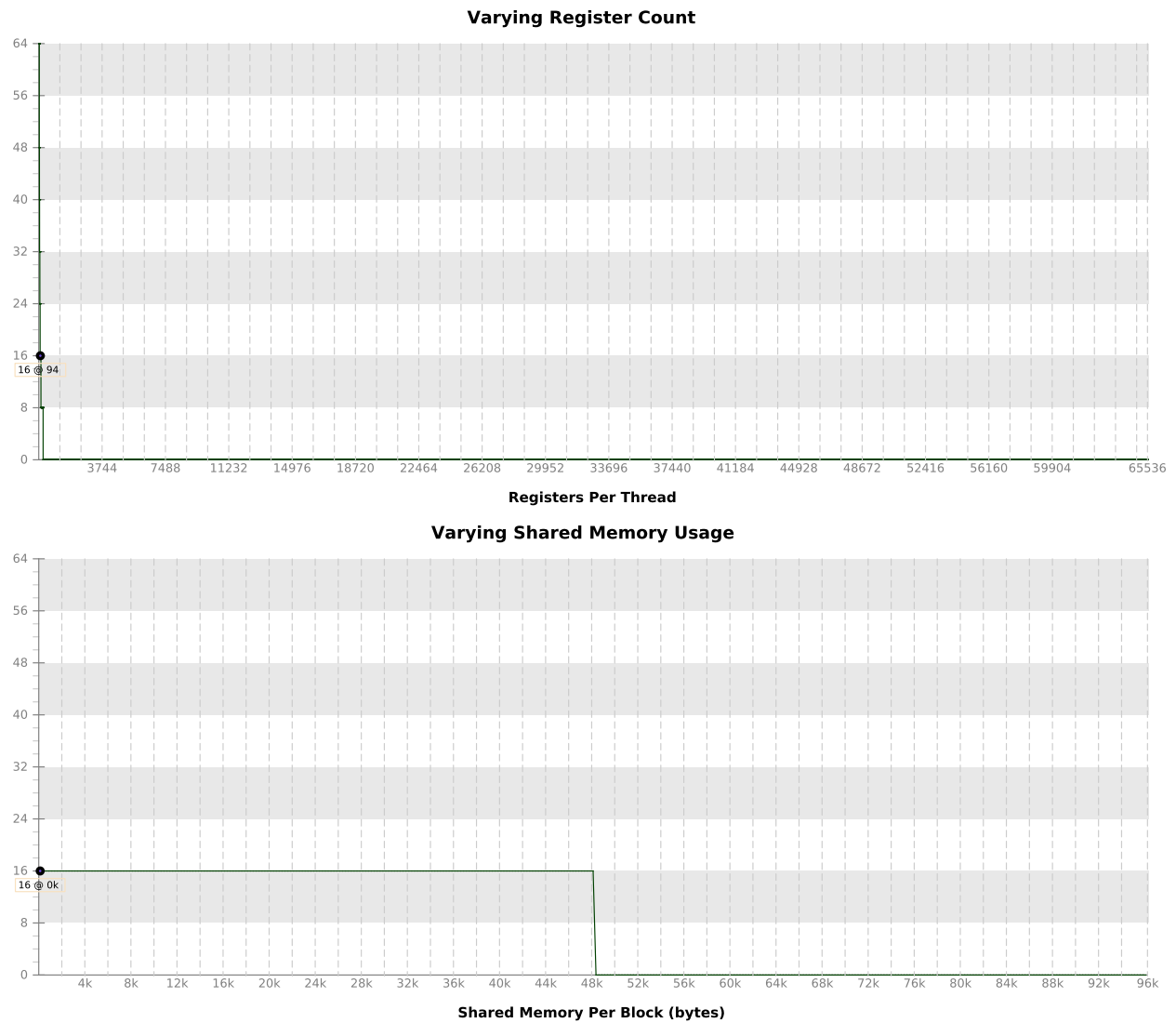
Optimization: Use the `-maxrregcount` flag or the `__launch_bounds__` qualifier to decrease the number of registers used by each thread. This will increase the number of blocks that can execute on each SM. On devices with Compute Capability 5.2 turning global cache off can increase the occupancy limited by register usage.



2.3. Occupancy Charts

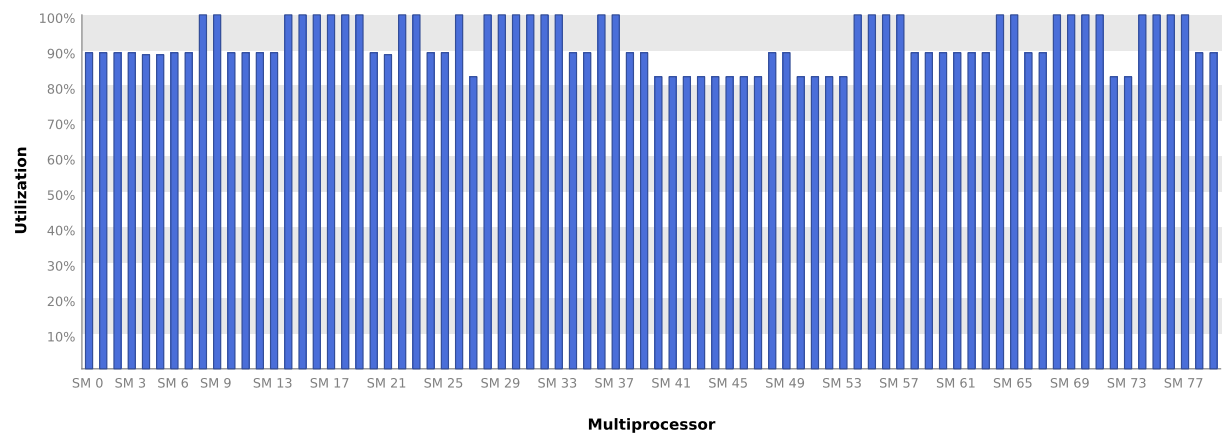
The following charts show how varying different components of the kernel will impact theoretical occupancy.





2.4. Multiprocessor Utilization

The kernel's blocks are distributed across the GPU's multiprocessors for execution. Depending on the number of blocks and the execution duration of each block some multiprocessors may be more highly utilized than others during execution of the kernel. The following chart shows the utilization of each multiprocessor during execution of the kernel.



3. Compute Resources

GPU compute resources limit the performance of a kernel when those resources are insufficient or poorly utilized.

3.1. Kernel Profile - Instruction Execution

The Kernel Profile - Instruction Execution shows the execution count, inactive threads, and predicated threads for each source and assembly line of the kernel. Using this information you can pinpoint portions of your kernel that are making inefficient use of compute resource due to divergence and predication.

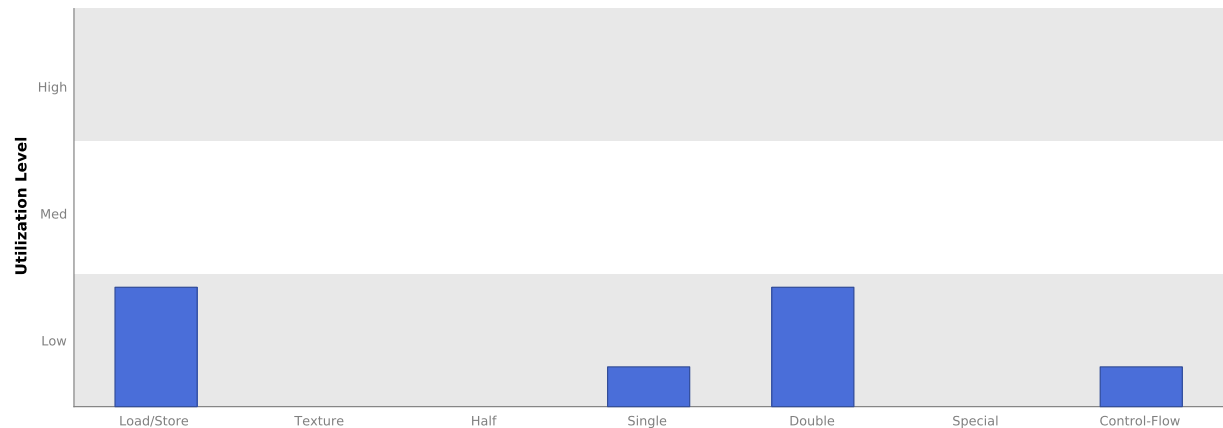
Examine portions of the kernel that have high execution counts and inactive or predicated threads to identify optimization opportunities.

Cuda Fuctions :
gpu4_column(int, int, int, double*, double*, double*)
Maximum instruction execution count in assembly: 67108864
Average instruction execution count in assembly: 15727142
Instructions executed for the kernel: 5221411160
Thread instructions executed for the kernel: 161798671104
Non-predicated thread instructions executed for the kernel: 161746675968
Warp non-predicated execution efficiency of the kernel: 96.8%
Warp execution efficiency of the kernel: 96.8%
Source files :
/zhome/a8/6/114633/hpc/week3/ass3/hpc-gpu/mattmult/matmult_gpu4.cu

3.2. Function Unit Utilization

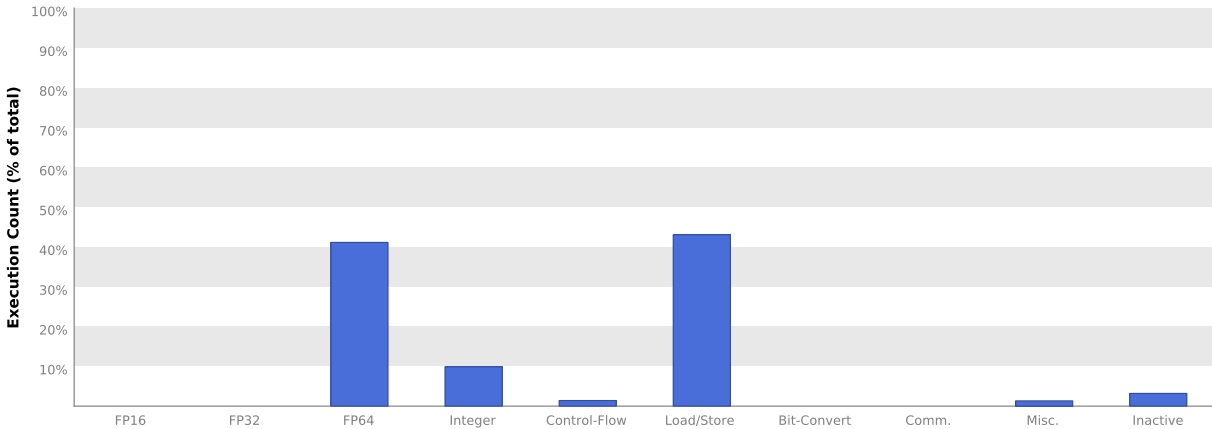
Different types of instructions are executed on different function units within each SM. Performance can be limited if a function unit is over-used by the instructions executed by the kernel. The following results show that the kernel's performance is not limited by overuse of any function unit.

- Load/Store - Load and store instructions for shared and constant memory.
- Texture - Load and store instructions for local, global, and texture memory.
- Half - Half-precision floating-point arithmetic instructions.
- Single - Single-precision integer and floating-point arithmetic instructions.
- Double - Double-precision floating-point arithmetic instructions.
- Special - Special arithmetic instructions such as sin, cos, popc, etc.
- Control-Flow - Direct and indirect branches, jumps, and calls.



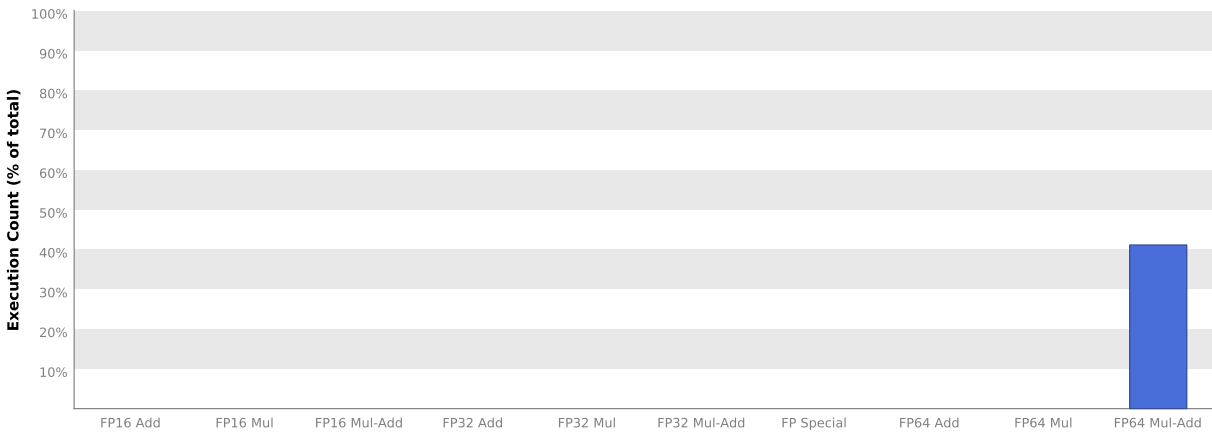
3.3. Instruction Execution Counts

The following chart shows the mix of instructions executed by the kernel. The instructions are grouped into classes and for each class the chart shows the percentage of thread execution cycles that were devoted to executing instructions in that class. The "Inactive" result shows the thread executions that did not execute any instruction because the thread was predicated or inactive due to divergence.



3.4. Floating-Point Operation Counts

The following chart shows the mix of floating-point operations executed by the kernel. The operations are grouped into classes and for each class the chart shows the percentage of thread execution cycles that were devoted to executing operations in that class. The results do not sum to 100% because non-floating-point operations executed by the kernel are not shown in this chart.



4. Memory Bandwidth

Memory bandwidth limits the performance of a kernel when one or more memories in the GPU cannot provide data at the rate requested by the kernel. The results below indicate that the kernel is limited by the bandwidth available to the unified cache that holds texture, global, and local data.

4.1. Global Memory Alignment and Access Pattern

Memory bandwidth is used most efficiently when each global memory load and store has proper alignment and access pattern. The analysis is per assembly instruction.

Optimization: Each entry below points to a global load or store within the kernel with an inefficient alignment or access pattern. For each load or store improve the alignment and access pattern of the memory access.

/zhome/a8/6/114633/hpc/week3/ass3/hpc-gpu/mattmult/matmult_gpu4.cu

Line 60	Global Load L2 Transactions/Access = 2, Ideal Transactions/Access = 2 [133169152 L2 transactions for 67108864 total executions]
Line 60	Global Load L2 Transactions/Access = 2, Ideal Transactions/Access = 2 [133169152 L2 transactions for 67108864 total executions]
Line 60	Global Load L2 Transactions/Access = 2, Ideal Transactions/Access = 2 [133169152 L2 transactions for 67108864 total executions]
Line 60	Global Load L2 Transactions/Access = 2, Ideal Transactions/Access = 2 [133169152 L2 transactions for 67108864 total executions]
Line 60	Global Load L2 Transactions/Access = 2, Ideal Transactions/Access = 2 [133169152 L2 transactions for 67108864 total executions]
Line 60	Global Load L2 Transactions/Access = 2, Ideal Transactions/Access = 2 [133169152 L2 transactions for 67108864 total executions]
Line 60	Global Load L2 Transactions/Access = 2, Ideal Transactions/Access = 2 [133169152 L2 transactions for 67108864 total executions]
Line 60	Global Load L2 Transactions/Access = 2, Ideal Transactions/Access = 2 [133169152 L2 transactions for 67108864 total executions]
Line 60	Global Load L2 Transactions/Access = 2, Ideal Transactions/Access = 2 [133169152 L2 transactions for 67108864 total executions]
Line 60	Global Load L2 Transactions/Access = 2, Ideal Transactions/Access = 2 [133169152 L2 transactions for 67108864 total executions]
Line 60	Global Load L2 Transactions/Access = 2, Ideal Transactions/Access = 2 [133169152 L2 transactions for 67108864 total executions]
Line 60	Global Load L2 Transactions/Access = 2, Ideal Transactions/Access = 2 [133169152 L2 transactions for 67108864 total executions]
Line 60	Global Load L2 Transactions/Access = 2, Ideal Transactions/Access = 2 [133169152 L2 transactions for 67108864 total executions]
Line 60	Global Load L2 Transactions/Access = 2, Ideal Transactions/Access = 2 [133169152 L2 transactions for 67108864 total executions]
Line 60	Global Load L2 Transactions/Access = 2, Ideal Transactions/Access = 2 [133169152 L2 transactions for 67108864 total executions]

/zhome/a8/6/114633/hpc/week3/ass3/hpc-gpu/matmult/matmult_gpu4.cu

[illegible]



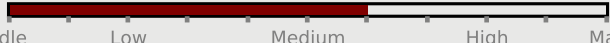



/zhome/a8/6/114633/hpc/week3/ass3/hpc-gpu/matmult/matmult_gpu4.cu

[illegible]

Line 66	Global Store L2 Transactions/Access = 31.8, Ideal Transactions/Access = 7.9 [520192 L2 transactions for 16384 total executions]
Line 66	Global Store L2 Transactions/Access = 31.8, Ideal Transactions/Access = 7.9 [520192 L2 transactions for 16384 total executions]
Line 66	Global Store L2 Transactions/Access = 31.8, Ideal Transactions/Access = 7.9 [520192 L2 transactions for 16384 total executions]
Line 66	Global Store L2 Transactions/Access = 31.8, Ideal Transactions/Access = 7.9 [520192 L2 transactions for 16384 total executions]
Line 66	Global Store L2 Transactions/Access = 31.8, Ideal Transactions/Access = 7.9 [520192 L2 transactions for 16384 total executions]
Line 66	Global Store L2 Transactions/Access = 31.8, Ideal Transactions/Access = 7.9 [520192 L2 transactions for 16384 total executions]
Line 66	Global Store L2 Transactions/Access = 31.8, Ideal Transactions/Access = 7.9 [520192 L2 transactions for 16384 total executions]
Line 76	Global Load L2 Transactions/Access = 1, Ideal Transactions/Access = 1 [8388608 L2 transactions for 8388608 total executions]
Line 76	Global Load L2 Transactions/Access = 16, Ideal Transactions/Access = 4 [16777216 L2 transactions for 1048576 total executions]
Line 76	Global Load L2 Transactions/Access = 1, Ideal Transactions/Access = 1 [8388608 L2 transactions for 8388608 total executions]
Line 76	Global Load L2 Transactions/Access = 1, Ideal Transactions/Access = 1 [8388608 L2 transactions for 8388608 total executions]
Line 76	Global Load L2 Transactions/Access = 1, Ideal Transactions/Access = 1 [8388608 L2 transactions for 8388608 total executions]
Line 82	Global Store L2 Transactions/Access = 16, Ideal Transactions/Access = 4 [32768 L2 transactions for 2048 total executions]
Line 82	Global Store L2 Transactions/Access = 16, Ideal Transactions/Access = 4 [32768 L2 transactions for 2048 total executions]
Line 82	Global Store L2 Transactions/Access = 16, Ideal Transactions/Access = 4 [32768 L2 transactions for 2048 total executions]
Line 82	Global Store L2 Transactions/Access = 16, Ideal Transactions/Access = 4 [32768 L2 transactions for 2048 total executions]

4.2. Memory Bandwidth And Utilization

The following table shows the memory bandwidth used by this kernel for the various types of memory on the device. The table also shows the utilization of each memory type relative to the maximum throughput supported by the memory.

Transactions	Bandwidth	Utilization	
Shared Memory			
Shared Loads	0	0 B/s	
Shared Stores	0	0 B/s	
Shared Total	0	0 B/s	
L2 Cache			
Reads	627885350	236.902 GB/s	
Writes	161453299	60.917 GB/s	
Total	789338649	297.819 GB/s	
Unified Cache			
Local Loads	134250496	50.653 GB/s	
Local Stores	142573568	53.793 GB/s	
Global Loads	5385438368	2,031.934 GB/s	
Global Stores	16777216	6.33 GB/s	
Texture Reads	4900920346	7,396.499 GB/s	
Unified Total	10579959994	9,539.21 GB/s	
Device Memory			
Reads	53950625	20.356 GB/s	
Writes	15875354	5.99 GB/s	
Total	69825979	26.345 GB/s	
System Memory			
[PCIe configuration: Gen3 x16, 8 Gbit/s]			
Reads	0	0 B/s	
Writes	5	1.886 kB/s	

4.3. Memory Statistics

The following chart shows a summary view of the memory hierarchy of the CUDA programming model. The green nodes in the diagram depict logical memory space whereas blue nodes depicts actual hardware unit on the chip. For the various caches the reported percentage number states the cache hit rate; that is the ratio of requests that could be served with data locally available to the cache over all requests made.

The links between the nodes in the diagram depict the data paths between the SMs to the memory spaces into the memory system. Different metrics are shown per data path. The data paths from the SMs to the memory spaces report the total number of memory instructions executed, it includes both read and write operations. The data path between memory spaces and "Unified Cache" or "Shared Memory" reports the total amount of memory requests made (read or write). All other data paths report the total amount of transferred memory in bytes.