# Analysis Report

## void computeHOGlocal<float, float, float, int=8, int=8, int=16, int=16, int=64>(float*, float*, float*, float*, int, int, int, int)

| | |
|---|---|
| Duration | 6.197 ms (6,197,140 ns) |
| Grid Size | [ 70,1,1 ] |
| Block Size | [ 256,1,1 ] |
| Registers/Thread | 72 |
| Shared  Memory/Block | 0 B |
| Shared Memory Requested | 96 KiB |
| Shared Memory Executed | 96 KiB |
| Shared Memory Bank Size | 4 B |

| [0] GeForce GTX 960 | |
|---|---|
| GPU UUID | GPU-0db32734-f94e-48a7-8b5d-4604317dc554 |
| Compute Capability | 5.2 |
| Max. Threads per Block | 1024 |
| Max. Shared Memory per Block | 48 KiB |
| Max. Registers per Block | 65536 |
| Max. Grid Dimensions | [ 2147483647, 65535, 65535 ] |
| Max. Block Dimensions | [ 1024, 1024, 64 ] |
| Max. Warps per Multiprocessor | 64 |
| Max. Blocks per Multiprocessor | 32 |
| Single Precision FLOP/s | 2.644 TeraFLOP/s |
| Double Precision FLOP/s | 82.624 GigaFLOP/s |
| Number of Multiprocessors | 8 |
| Multiprocessor Clock Rate | 1.291 GHz |
| Concurrent Kernel | true |
| Max IPC | 6 |
| Threads per Warp | 32 |
| Global Memory Bandwidth | 112.16 GB/s |
| Global Memory Size | 4 GiB |
| Constant Memory Size | 64 KiB |
| L2 Cache Size | 1 MiB |
| Memcpy Engines | 2 |
| PCIe Generation | 2 |
| PCIe Link Rate | 5 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 "void computeHOGlocal<float,..." 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 "GeForce GTX 960". 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.
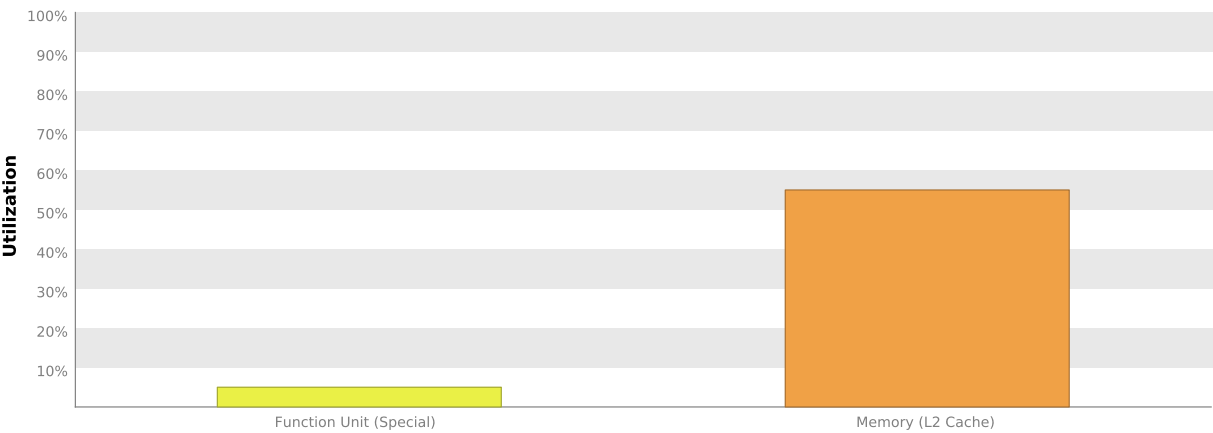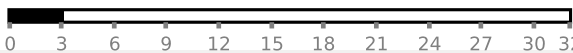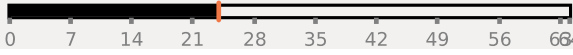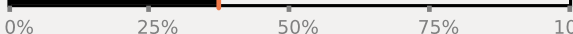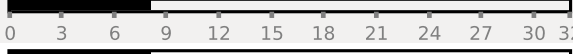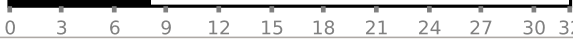
# 2. Instruction and Memory Latency

Instruction and memory latency limit the performance of a kernel when the GPU does not have enough work to keep busy. The performance of latency-limited kernels can often be improved by increasing occupancy. Occupancy is a measure of how many warps the kernel has active on the GPU, relative to the maximum number of warps supported by the GPU. Theoretical occupancy provides an upper bound while achieved occupancy indicates the kernel's actual occupancy. The results below indicate that occupancy can be improved by reducing the number of registers used by the kernel.

## 2.1. GPU Utilization Is Limited By Register Usage

The kernel uses 72 registers for each thread (18432 registers for each block). This register usage is likely preventing the kernel from fully utilizing the GPU. Device "GeForce GTX 960" provides up to 65536 registers for each block. Because the kernel uses 18432 registers for each block each SM is limited to simultaneously executing 3 blocks (24 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.*

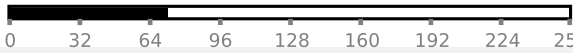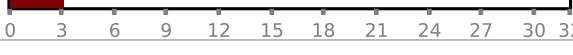| Variable | Achieved | Theoretical | Device Limit | Grid Size: [ 70,1,1 ] (70 blocks) Block Size: [ 256,1,1 ] (256 threads |
|---|---|---|---|---|
| **Occupancy Per SM** | | | | |
| Active Blocks | | 3 | 32 | |
| Active Warps | 23.54 | 24 | 64 | |
| Active Threads | | 768 | 2048 | |
| Occupancy | 36.8% | 37.5% | 100% | |
| **Warps** | | | | |
| Threads/Block | | 256 | 1024 | |
| Warps/Block | | 8 | 32 | |
| Block Limit | | 8 | 32 | |
| **Registers** | | | | |
| Registers/Thread | | 72 | 255 | |
| Registers/Block | | 18432 | 65536 | |
| Block Limit | | 3 | 32 | |
| **Shared Memory** | | | | |
| Shared Memory/Block | | 0 | 98304 | |
| Block Limit | | | 32 | |

## 2.2. Occupancy Charts

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

## Varying Block Size

**Warps Per SM** vs **Threads Per Block**

24 @ 256

## Varying Register Count

**Warps Per SM** vs **Registers Per Thread**

24 @ 72

## Varying Shared Memory Usage



24 @ 0k

Shared Memory Per Block (bytes)
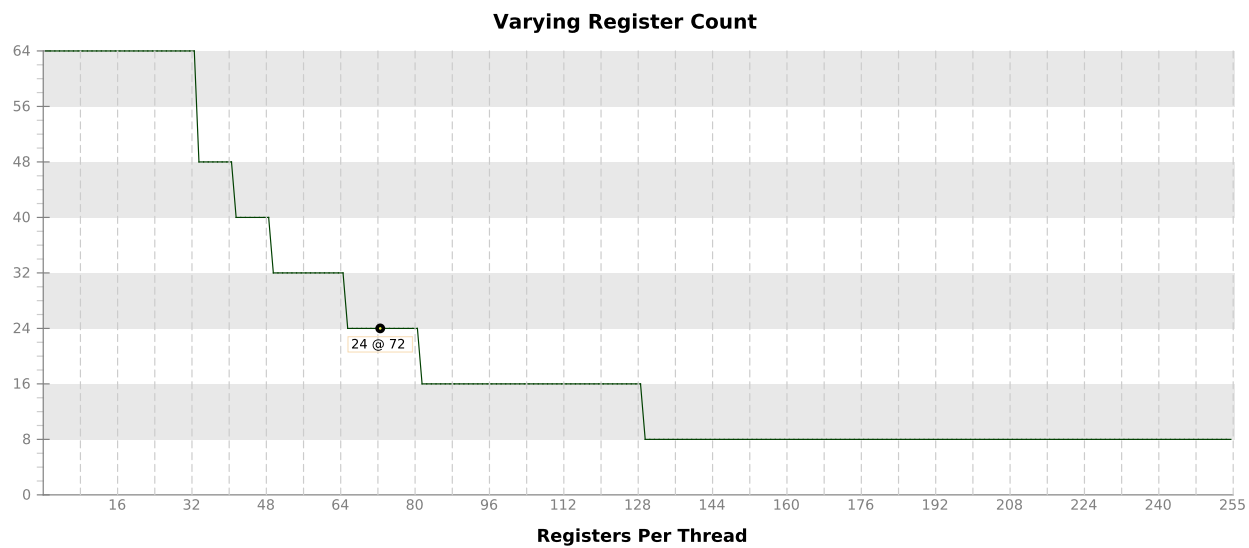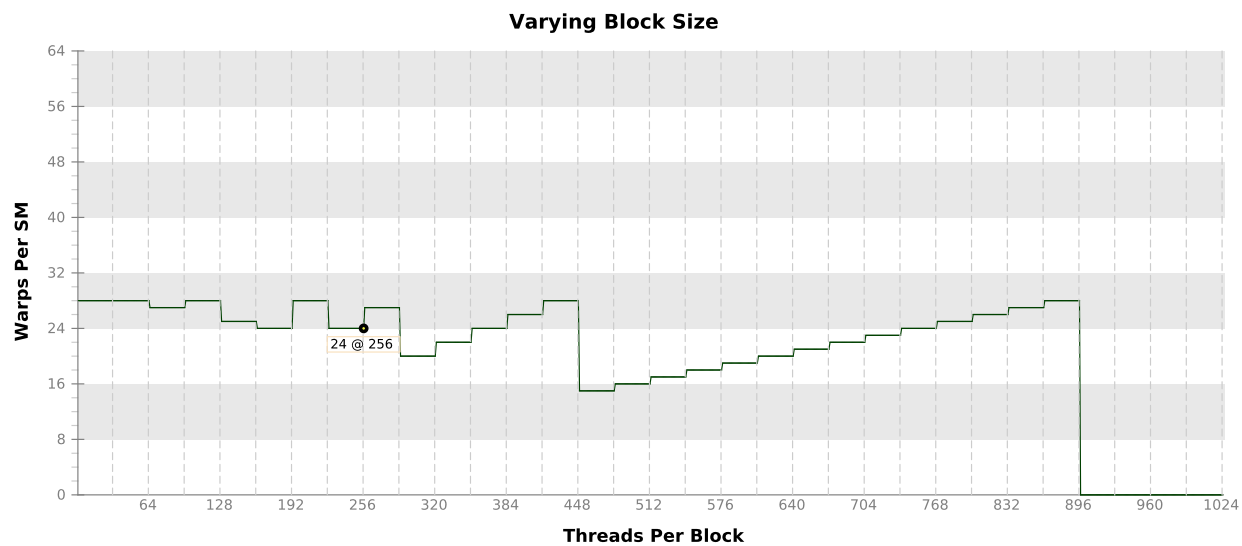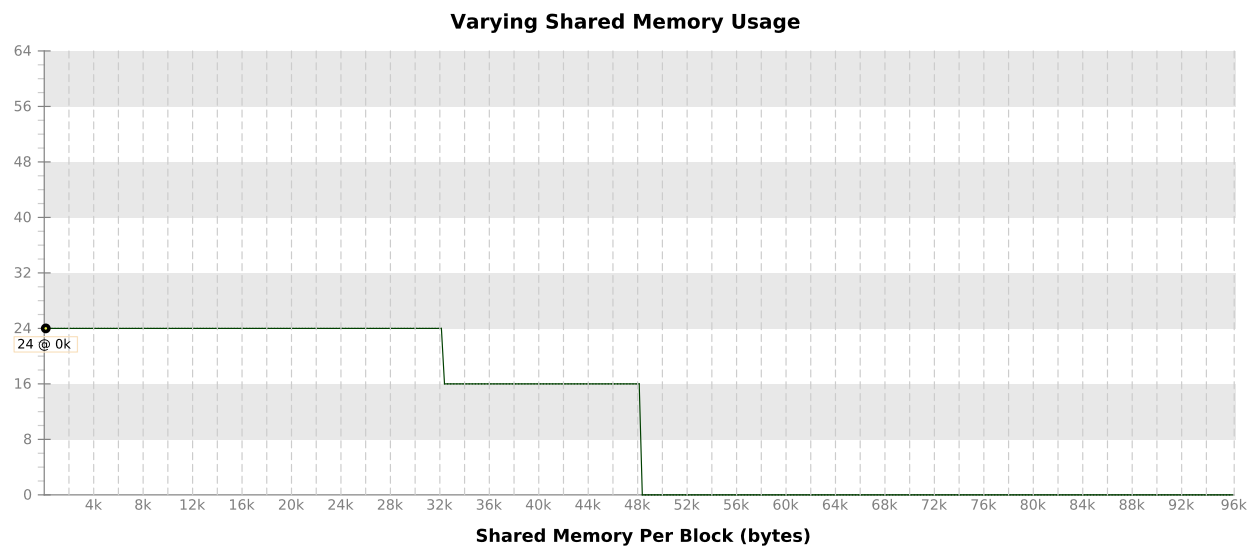
# 3. Compute Resources

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

## 3.1. 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.
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.
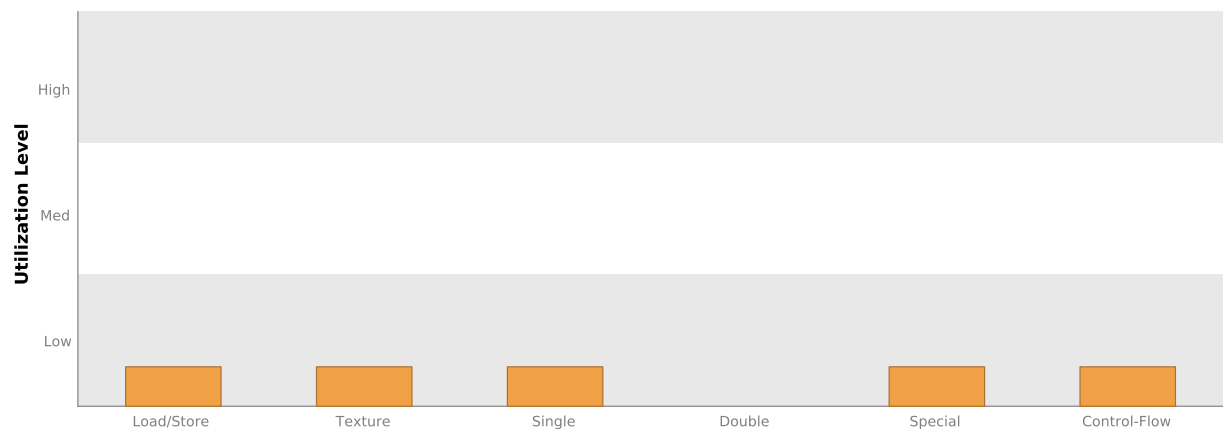


## 3.2. Instruction Execution Counts

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.
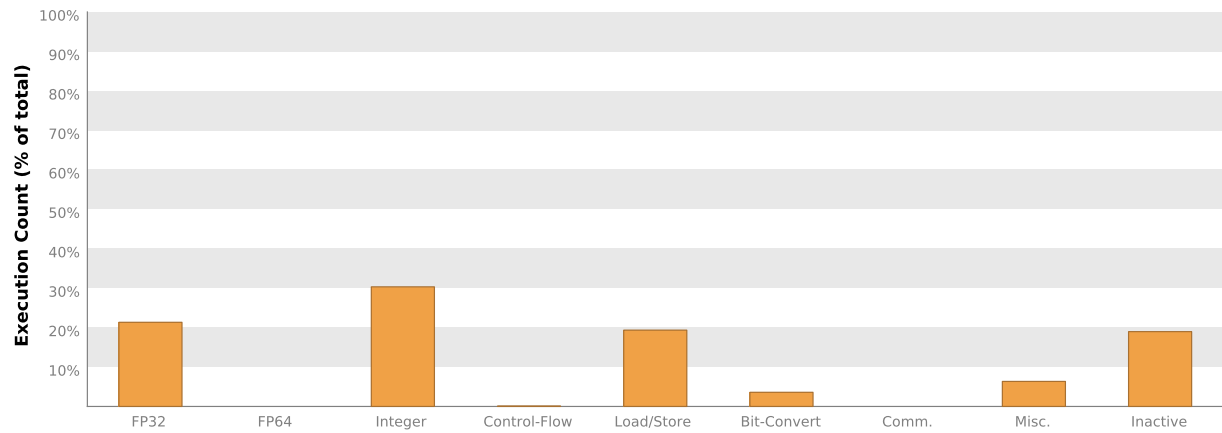
### 3.3. Floating-Point Operation Counts

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.
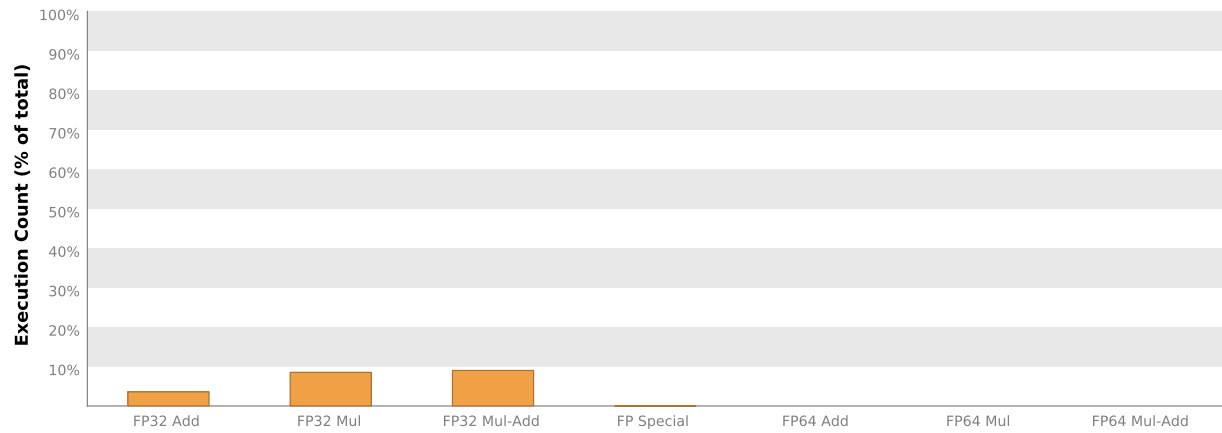
# 4. Memory Bandwidth

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 L2 cache.

## 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.

*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.*

/home/adas/cuda-workspace/CudaVisionSysDeploy/Release/../src/init/../device/HOG/HOGdescriptor.cuh

| | |
|---|---|
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total |

| | executions ] |
|---|---|
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 221 | Global Load L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 286000 L2 transactions for 8944 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total |

| | executions ] |
|---|---|
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total |

| | executions ] |
|---|---|
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |
| Line 230 | Global Store L2 Transactions/Access = 32, Ideal Transactions/Access = 4 [ 17875 L2 transactions for 559 total executions ] |

## 4.2. High Local Memory Overhead

Local memory loads and stores account for 74% of total memory traffic. High local memory traffic typically indicates excessive register spilling.

*Optimization: Use the -maxrregcount flag or the __launch_bounds__ qualifier to increase the number of registers available to nvcc when compiling the kernel.*

## 4.3. 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 | Idle　Low　Medium　High　Max |
| **L2 Cache** | | | |
| Reads | 21840991 | 118.14 GB/s | |
| Writes | 16565572 | 89.605 GB/s | |
| Total | 38406563 | 207.746 GB/s | Idle　Low　Medium　High　Max |
| **Unified Cache** | | | |
| Local Loads | 15980700 | 86.441 GB/s | |
| Local Stores | 15921980 | 86.124 GB/s | |
| Global Loads | 9724160 | 50.278 GB/s | |
| Global Stores | 643500 | 3.481 GB/s | |
| Texture Reads | 4371660 | 23.647 GB/s | |
| Unified Total | 46642000 | 249.971 GB/s | Idle　Low　Medium　High　Max |
| **Device Memory** | | | |
| Reads | 4389654 | 23.744 GB/s | |
| Writes | 5097209 | 27.571 GB/s | |
| Total | 9486863 | 51.316 GB/s | Idle　Low　Medium　High　Max |
| **System Memory** | | | |
| [ PCIe configuration: Gen2 x16, 5 Gbit/s ] | | | |
| Reads | 0 | 0 B/s | Idle　Low　Medium　High　Max |
| Writes | 5 | 27.045 kB/s | Idle　Low　Medium　High　Max |