

Analysis Report

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

Duration	846.511 μ s
Grid Size	[6,1,1]
Block Size	[256,1,1]
Registers/Thread	40
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

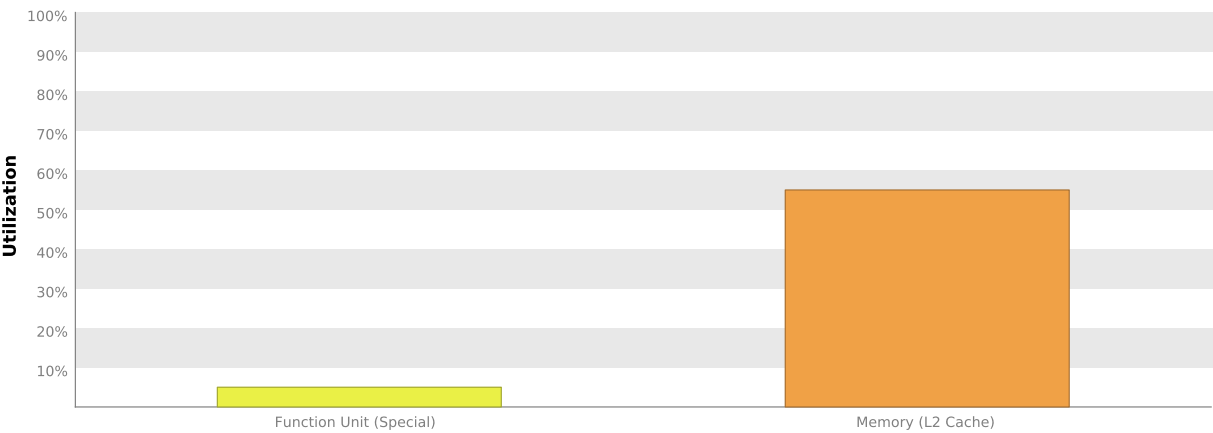
GPU UUID	GPU-0db32734-f94e-48a7-8b5d-4604317dc554
Compute Capability	5.2
Max. Threads per Block	1024
Max. Shared Memory per Block	48 KiB
Max. Registers per Block	65536
Max. Grid Dimensions	[2147483647, 65535, 65535]
Max. Block Dimensions	[1024, 1024, 64]
Max. Warps per Multiprocessor	64
Max. Blocks per 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 computeHOGdescriptor<f...>" 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 results below indicate that the GPU does not have enough work because the kernel does not execute enough blocks.

2.1. Grid Size Too Small To Hide Compute And Memory Latency

The kernel does not execute enough blocks to hide memory and operation latency. Typically the kernel grid size must be large enough to fill the GPU with multiple "waves" of blocks. Based on theoretical occupancy, device "GeForce GTX 960" can simultaneously execute 6 blocks on each of the 8 SMs, so the kernel may need to execute a multiple of 48 blocks to hide the compute and memory latency. If the kernel is executing concurrently with other kernels then fewer blocks will be required because the kernel is sharing the SMs with those kernels.

Optimization: Increase the number of blocks executed by the kernel.

2.2. GPU Utilization May Be Limited By Register Usage

Theoretical occupancy is less than 100% but is large enough that increasing occupancy may not improve performance. You can attempt the following optimization to increase the number of warps on each SM but it may not lead to increased performance.

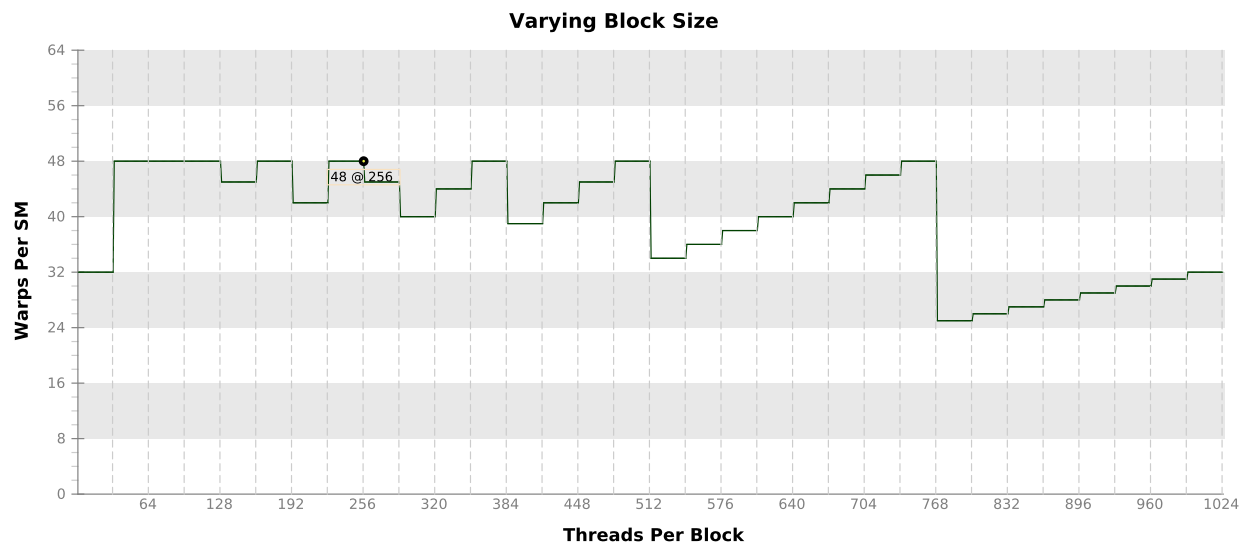
The kernel uses 40 registers for each thread (10240 registers for each block). This register usage is likely preventing the kernel from fully utilizing the GPU. Device "GeForce GTX 960" provides up to 65536 registers for each block. Because the kernel uses 10240 registers for each block each SM is limited to simultaneously executing 6 blocks (48 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.

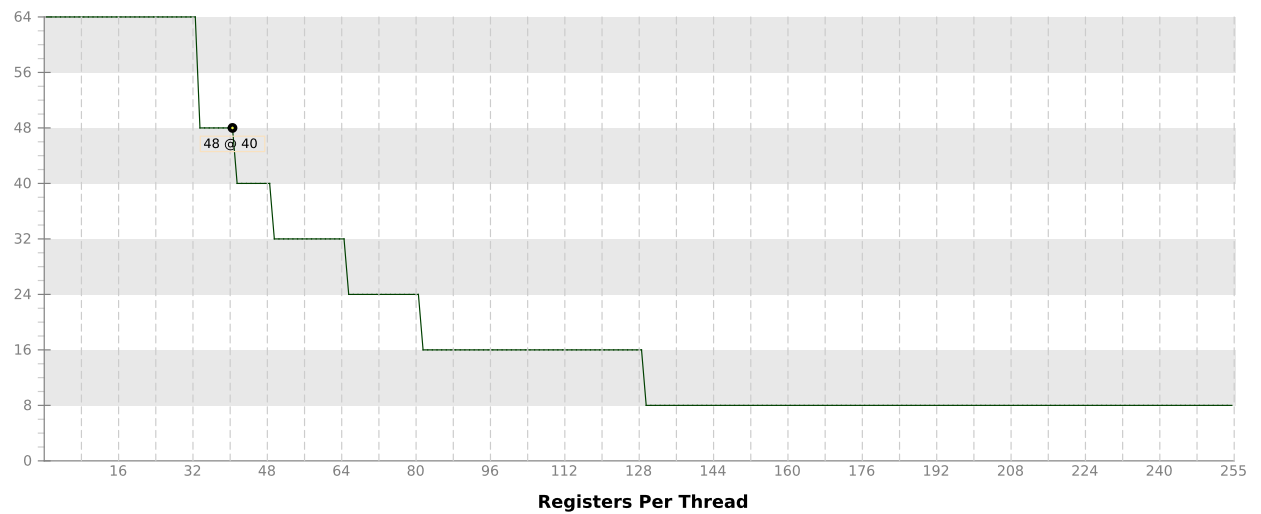
Variable	Achieved	Theoretical	Device Limit	Grid Size: [6,1,1] (6 blocks) Block Size: [256,1,1] (256 threads)
Occupancy Per SM				
Active Blocks		6	32	
Active Warps	7.55	48	64	
Active Threads		1536	2048	
Occupancy	11.8%	75%	100%	
Warps				
Threads/Block		256	1024	
Warps/Block		8	32	
Block Limit		8	32	
Registers				
Registers/Thread		40	255	
Registers/Block		10240	65536	
Block Limit		6	32	
Shared Memory				
Shared Memory/Block		0	98304	
Block Limit			32	

2.3. Occupancy Charts

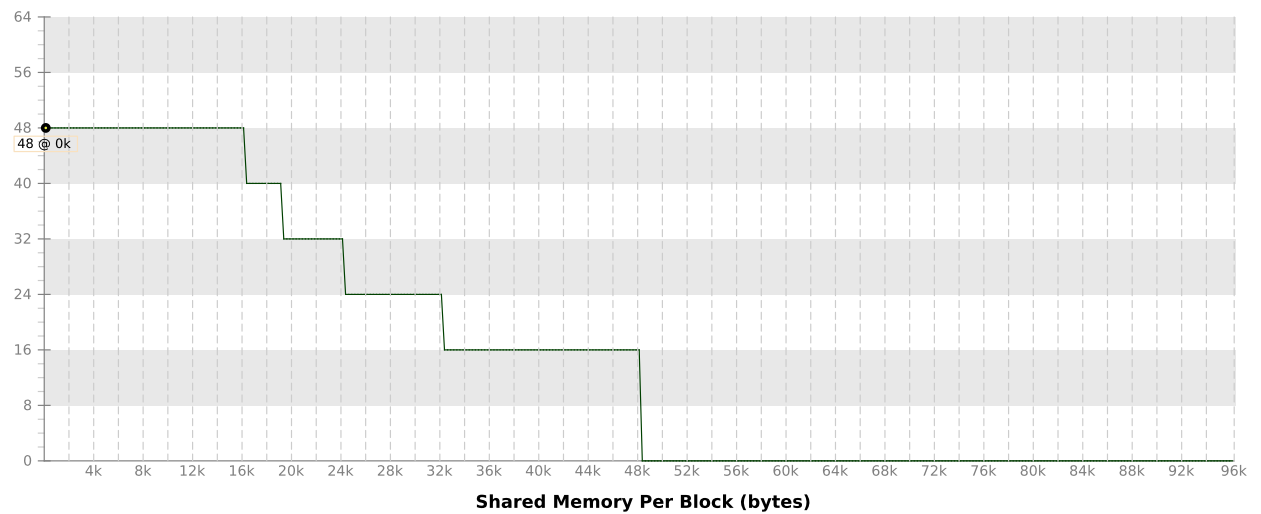
The following charts show how varying different components of the kernel will impact theoretical occupancy.



Varying Register Count



Varying Shared Memory Usage



3. Compute Resources

GPU compute resources limit the performance of a kernel when those resources are insufficient or poorly utilized. Compute resources are used most efficiently when all threads in a warp have the same branching and predication behavior. The results below indicate that a significant fraction of the available compute performance is being wasted because branch and predication behavior is differing for threads within a warp.

3.1. Divergent Branches

Compute resource are used most efficiently when all threads in a warp have the same branching behavior. When this does not occur the branch is said to be divergent. Divergent branches lower warp execution efficiency which leads to inefficient use of the GPU's compute resources.

Optimization: Each entry below points to a divergent branch within the kernel. For each branch reduce the amount of intra-warp divergence.

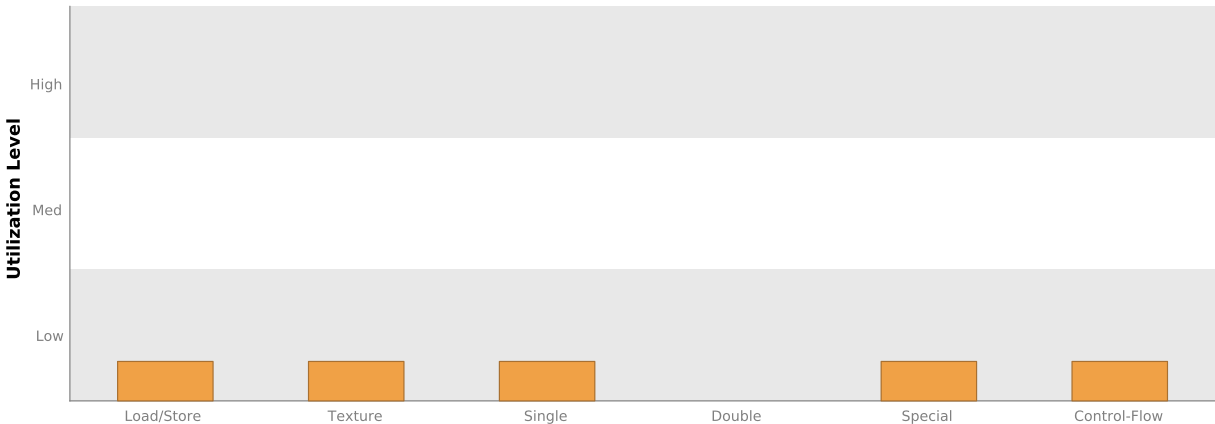
```
/home/adas/cuda-workspace/CudaVisionSysDeploy/Release/./src/init/./device/HOG/HOGdescriptor.cuh
```

Line 186	Divergence = 2.1% [1 divergent executions out of 48 total executions]
----------	---

3.2. Function Unit Utilization

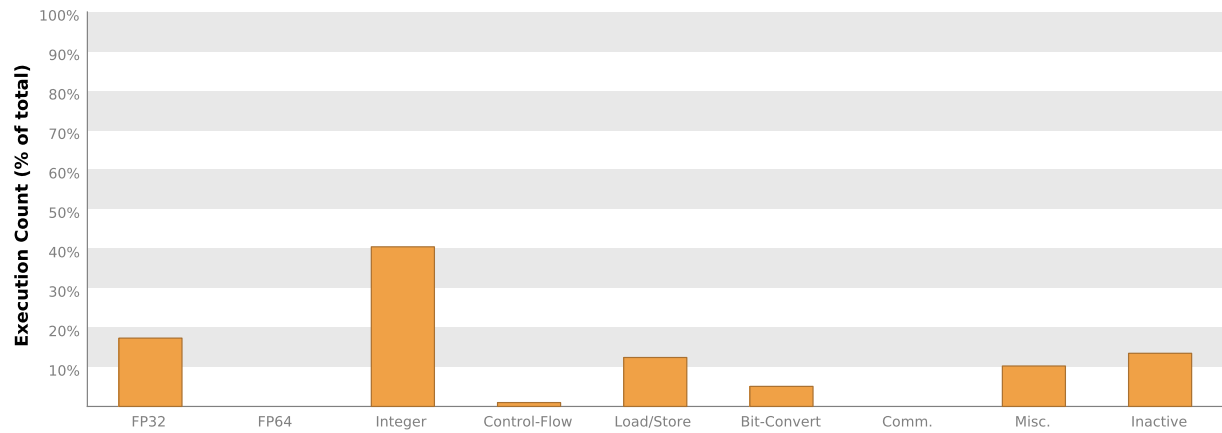
Different types of instructions are executed on different function units within each SM. Performance can be limited if a function unit is over-used by the instructions executed by the kernel. The following results show that the kernel's performance is not limited by overuse of any function unit.

- Load/Store - Load and store instructions for shared and constant memory.
- Texture - Load and store instructions for local, global, and texture memory.
- 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.



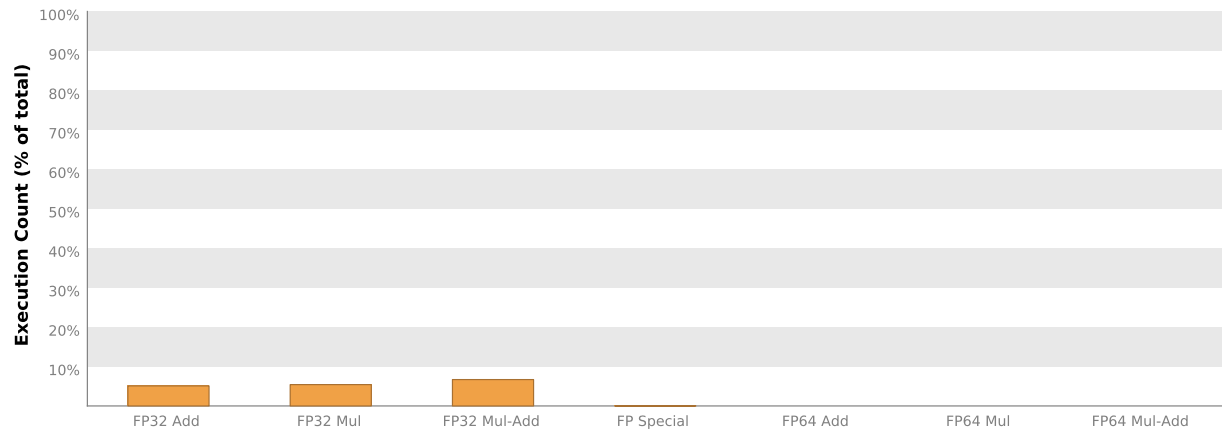
3.3. Instruction Execution Counts

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 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 189	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [371456 L2 transactions for 11776 total executions]
Line 189	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [371456 L2 transactions for 11776 total executions]

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

Line 48	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [208944 L2 transactions for 6624 total executions]
Line 48	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [208944 L2 transactions for 6624 total executions]
Line 49	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [208944 L2 transactions for 6624 total executions]
Line 49	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [208944 L2 transactions for 6624 total executions]
Line 57	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [208944 L2 transactions for 6624 total executions]
Line 57	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [208944 L2 transactions for 6624 total executions]
Line 58	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [208944 L2 transactions for 6624 total executions]
Line 58	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [208944 L2 transactions for 6624 total executions]
Line 64	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [208944 L2 transactions for 6624 total executions]
Line 64	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [208944 L2 transactions for 6624 total executions]
Line 65	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [208944 L2 transactions for 6624 total executions]
Line 65	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [208944 L2 transactions for 6624 total executions]
Line 72	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [208944 L2 transactions for 6624 total executions]
Line 72	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [208944 L2 transactions for 6624 total executions]


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

[illegible]

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

[illegible]

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

Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]

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

[illegible]






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

Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]

Line 61	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Load L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 total executions]
Line 61	Global Store L2 Transactions/Access = 31.5, Ideal Transactions/Access = 4 [1451 L2 transactions for 46 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	1	150.116 kB/s	
Shared Stores	1	150.116 kB/s	
Shared Total	2	300.232 kB/s	
L2 Cache			
Reads	2583003	96.938 GB/s	
Writes	1776030	66.653 GB/s	
Total	4359033	163.591 GB/s	
Unified Cache			
Local Loads	0	0 B/s	
Local Stores	0	0 B/s	
Global Loads	2617764	96.936 GB/s	
Global Stores	1776024	66.653 GB/s	
Texture Reads	369096	13.852 GB/s	
Unified Total	4762884	177.441 GB/s	
Device Memory			
Reads	46136	1.731 GB/s	
Writes	16073	603.207 MB/s	
Total	62209	2.335 GB/s	
System Memory			
[PCIe configuration: Gen2 x16, 5 Gbit/s]			
Reads	0	0 B/s	
Writes	5	187.646 kB/s	