Analysis Report

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

Duration	3.248 ms (3,248,345 ns)
Grid Size	[31,1,1]
Block Size	[256,1,1]
Registers/Thread	48
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

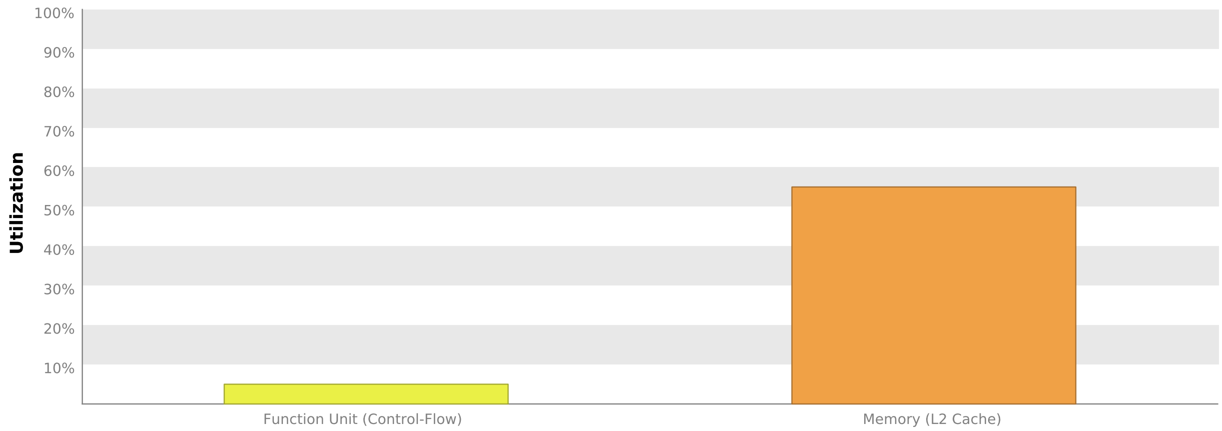
GPU UUID	GPU-0db32734-f94e-48a7-8b5d-4604317dc554
Compute Capability	5.2
Max. Threads per Block	1024
Max. Shared Memory per Block	48 KiB
Max. Registers per Block	65536
Max. Grid Dimensions	[2147483647, 65535, 65535]
Max. Block Dimensions	[1024, 1024, 64]
Max. Warps per Multiprocessor	64
Max. Blocks per 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 computeHOGlocalPred<fl..." 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.



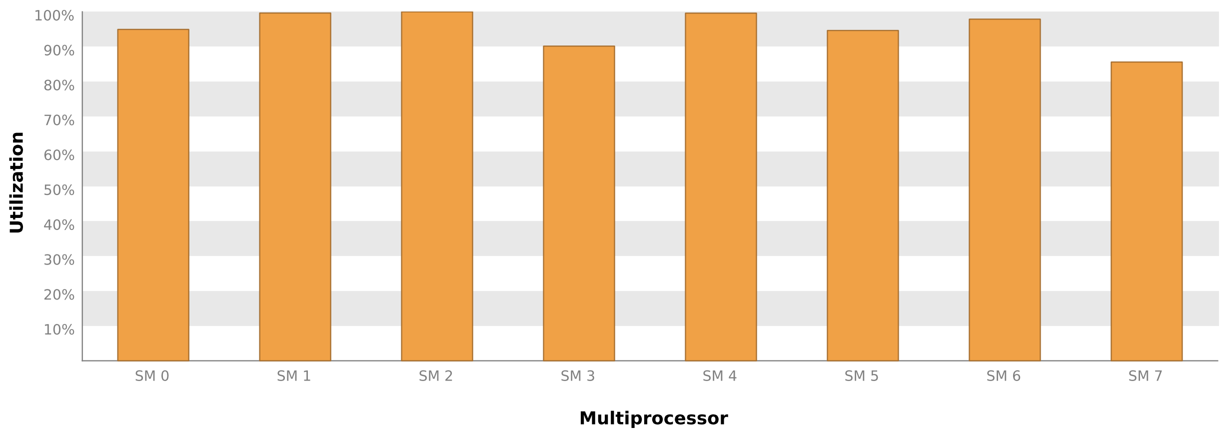
2. Instruction and Memory Latency

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 differences in the execution time of the kernel's blocks leads to poor load balancing across the SMs.

2.1. Achieved Occupancy Is Low

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 kernel's achieved occupancy of 11.8% is significantly lower than its theoretical occupancy of 62.5%. Most likely this indicates that there is an imbalance in how the kernel's blocks are executing on the SMs so that all SMs are not equally busy over the entire execution of the kernel. The following chart shows the utilization of each multiprocessor during execution of the kernel.

Optimization: Make sure that all blocks are doing roughly the same amount of work. It may also help to 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.

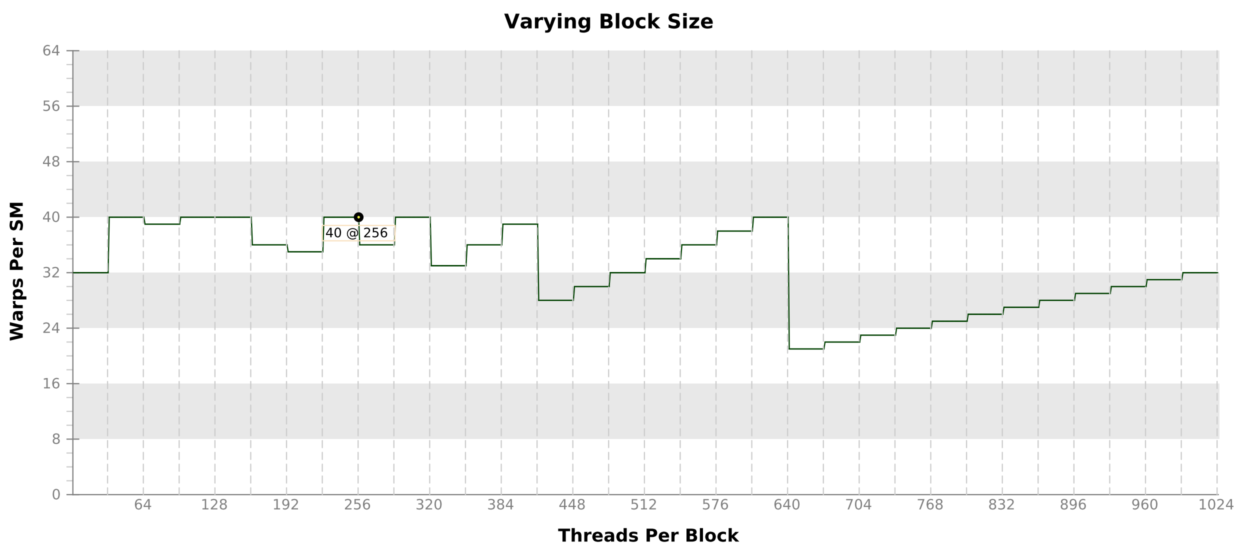
The kernel uses 48 registers for each thread (12288 registers for each block). This register usage is likely preventing the kernel from fully utilizing the GPU. Device "GeForce GTX 960" provides up to 65536 registers for each block. Because the kernel uses 12288 registers for each block each SM is limited to simultaneously executing 5 blocks (40 warps). Chart "Varying Register Count" below shows how changing register usage will change the number of blocks that can execute on each SM.

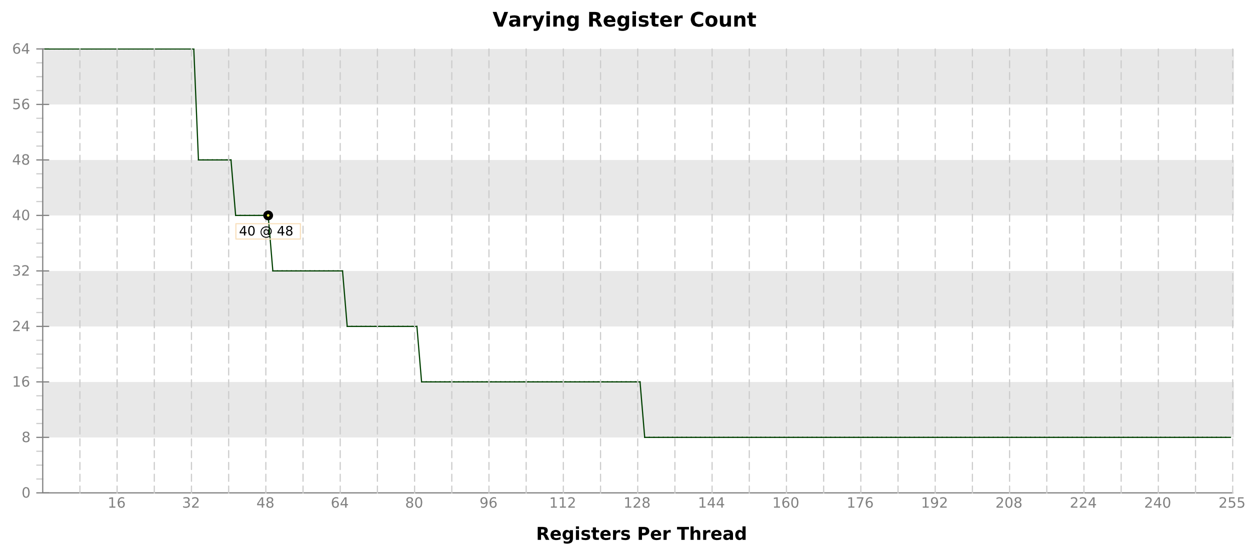
Optimization: Use the -maxrregcount flag or the __launch_bounds__ qualifier to decrease the number of registers used by each thread. This will increase the number of blocks that can execute on each SM. On devices with Compute Capability 5.2 turning global cache off can increase the occupancy limited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

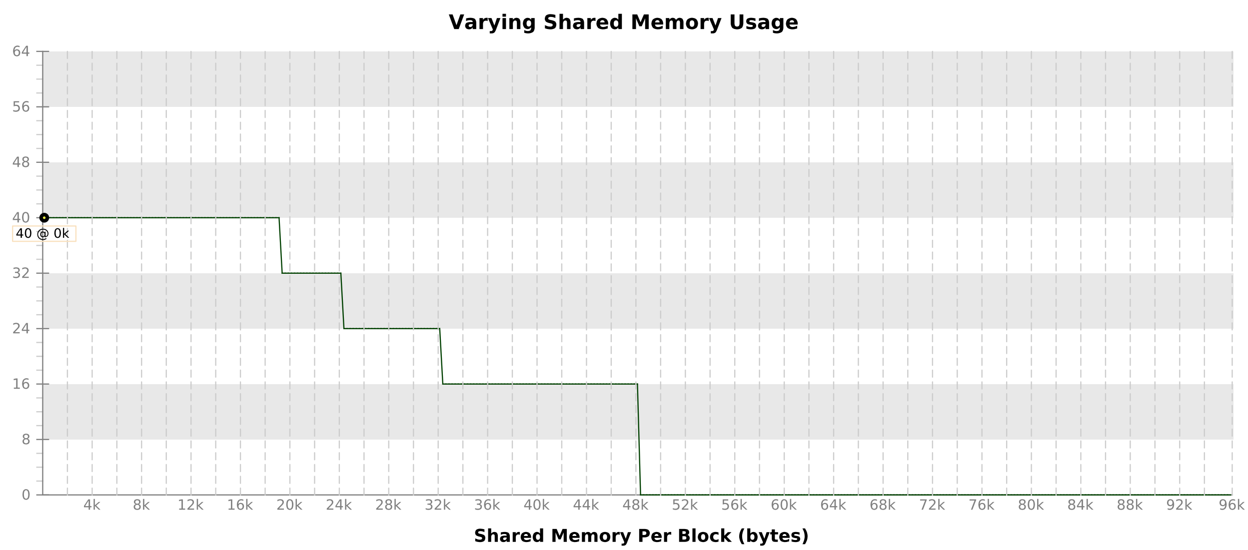
Variable	Achieved	Theoretical	Device Limit	Grid Size: [31,1,1] (31 blocks) Block Size: [256,1,1] (256 threads
Occupancy Per SM				
Active Blocks		5	32	
Active Warps	7.55	40	64	
Active Threads		1280	2048	
Occupancy	11.8%	62.5%	100%	
Warps				
Threads/Block		256	1024	
Warps/Block		8	32	
Block Limit		8	32	
Registers				
Registers/Thread		48	255	
Registers/Block		12288	65536	
Block Limit		5	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
64
56
48
40
32
24
16
8
0
40 @ 48
16
32
48
64
80
96
112
128
144
160
176
192
208
224
240
255
Registers Per Thread


Varying Shared Memory Usage
64
56
48
40
32
24
16
8
0
40 @ 0k
4k
8k
12k
16k
20k
24k
28k
32k
36k
40k
44k
48k
52k
56k
60k
64k
68k
72k
76k
80k
84k
88k
92k
96k
Shared Memory Per Block (bytes)

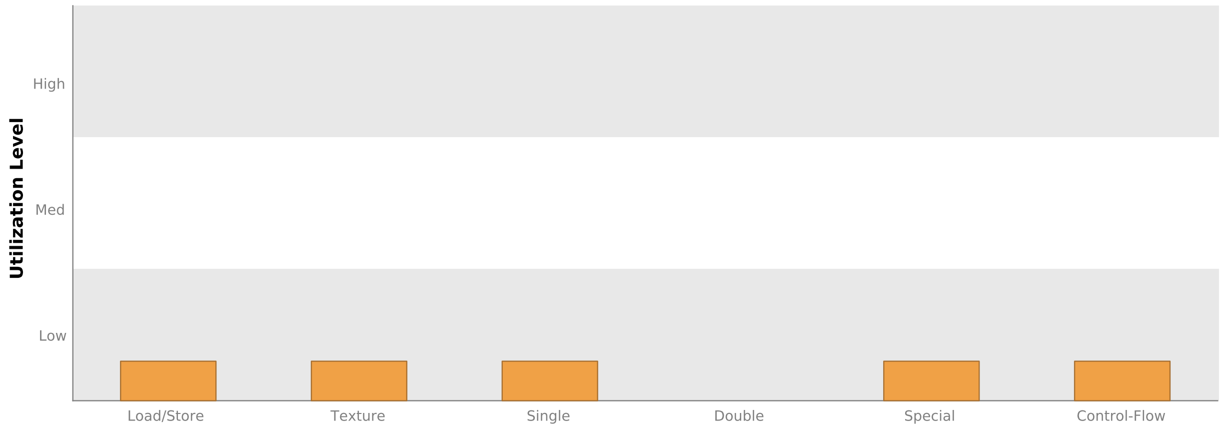
3. Compute Resources

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

3.1. Function Unit Utilization

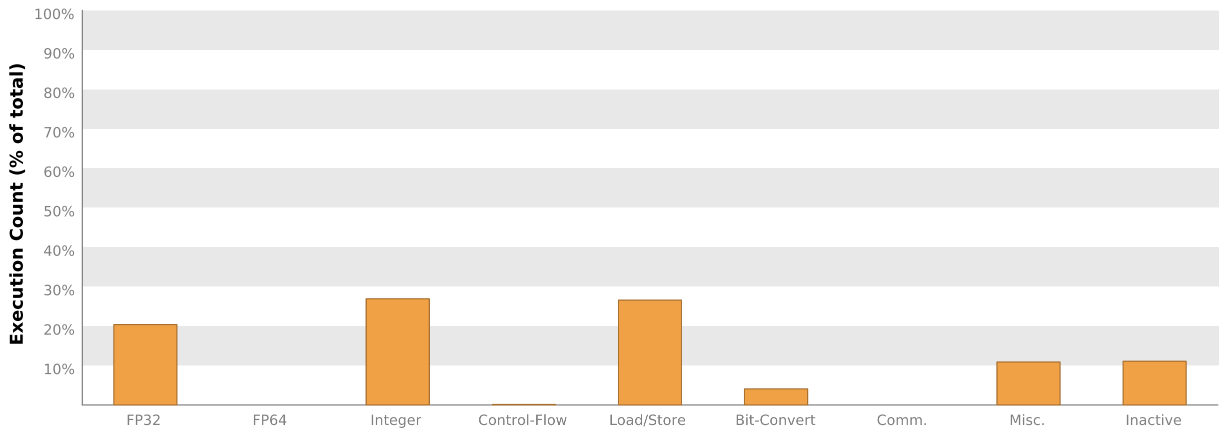
Different types of instructions are executed on different function units within each SM. Performance can be limited if a function unit is over-used by the instructions executed by the kernel. The following results show that the kernel's performance is not limited by overuse of any function unit.

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.



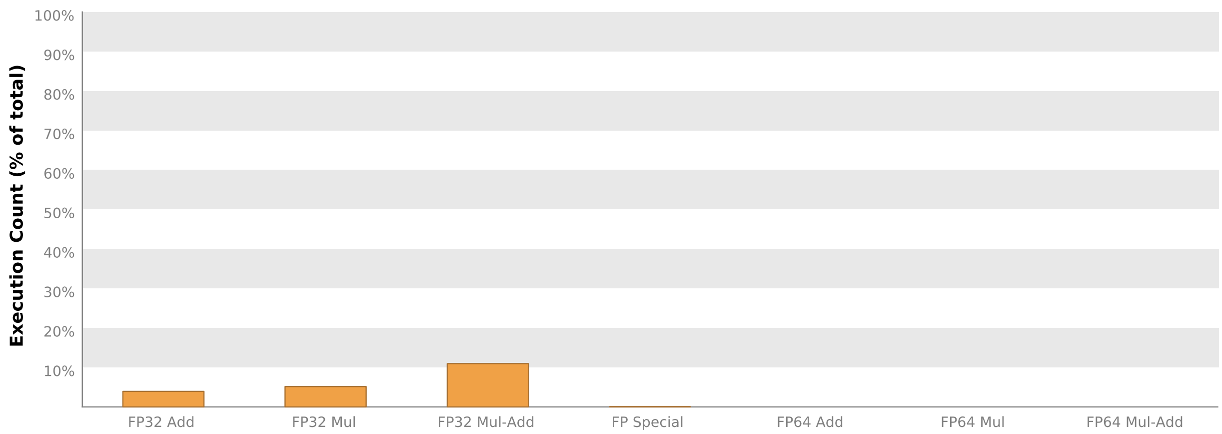
3.2. Instruction Execution Counts

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 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total

	executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 178	Global Load L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [124976 L2 transactions for 3920 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total

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

	executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total

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

	executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]
Line 188	Global Store L2 Transactions/Access = 31.9, Ideal Transactions/Access = 4 [7811 L2 transactions for 245 total executions]

4.2. High Local Memory Overhead

Local memory loads and stores account for 72% 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	9253841	93.091 GB/s	
Writes	6771310	68.117 GB/s	
Total	16025151	161.208 GB/s	Idle Low Medium High Max
Unified Cache			
Local Loads	6548266	65.874 GB/s	
Local Stores	6489981	65.287 GB/s	
Global Loads	4608880	41.769 GB/s	
Global Stores	281196	2.829 GB/s	
Texture Reads	2270548	22.841 GB/s	
Unified Total	20198871	198.6 GB/s	Idle Low Medium High Max
Device Memory			
Reads	3466288	34.87 GB/s	
Writes	2580857	25.963 GB/s	
Total	6047145	60.832 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	50.298 kB/s	Idle Low Medium High Max