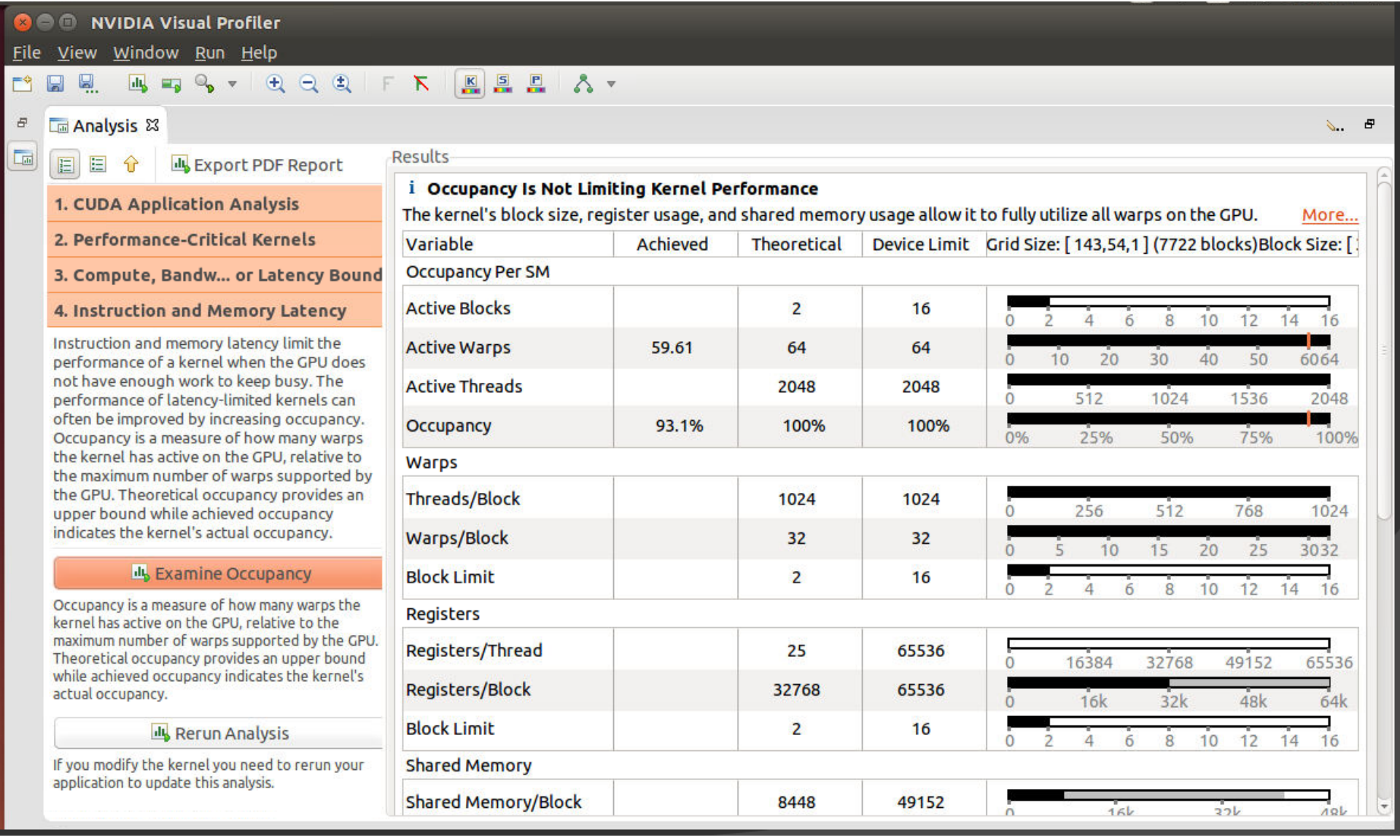
NVIDIA Visual Profiler
File View Window Run Help
Analysis
Export PDF Report
1. CUDA Application Analysis
2. Performance-Critical Kernels
3. Compute, Bandw... or Latency Bound
4. Instruction and Memory Latency
Instruction and memory latency limit the performance of a kernel when the GPU does not have 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.
Examine 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.
Rerun Analysis
If you modify the kernel you need to rerun your application to update this analysis.
Results
Occupancy Is Not Limiting Kernel Performance
The kernel's block size, register usage, and shared memory usage allow it to fully utilize all warps on the GPU.
More...
Variable
Achieved
Theoretical
Device Limit
Grid Size: [143,54,1] (7722 blocks)Block Size: [
Occupancy Per SM
Active Blocks
2
16
0 2 4 6 8 10 12 14 16
Active Warps
59.61
64
64
0 10 20 30 40 50 6064
Active Threads
2048
2048
0 512 1024 1536 2048
Occupancy
93.1%
100%
100%
0% 25% 50% 75% 100%
Warps
Threads/Block
1024
1024
0 256 512 768 1024
Warps/Block
32
32
0 5 10 15 20 25 3032
Block Limit
2
16
0 2 4 6 8 10 12 14 16
Registers
Registers/Thread
25
65536
0 16384 32768 49152 65536
Registers/Block
32768
65536
0 16k 32k 48k 64k
Block Limit
2
16
0 2 4 6 8 10 12 14 16
Shared Memory
Shared Memory/Block
8448
49152