Performance Evaluation of Sobel Edge Detection: CPU vs CUDA GPU

1. Introduction

This report presents an in-depth performance analysis of a Sobel edge detection implementation on CPU and CUDA-enabled GPU. The evaluation spans resolutions from 512×512 to 8K and uses runtime metrics, throughput measurements, speedup analysis, and roofline modeling to characterize compute and memory behavior.

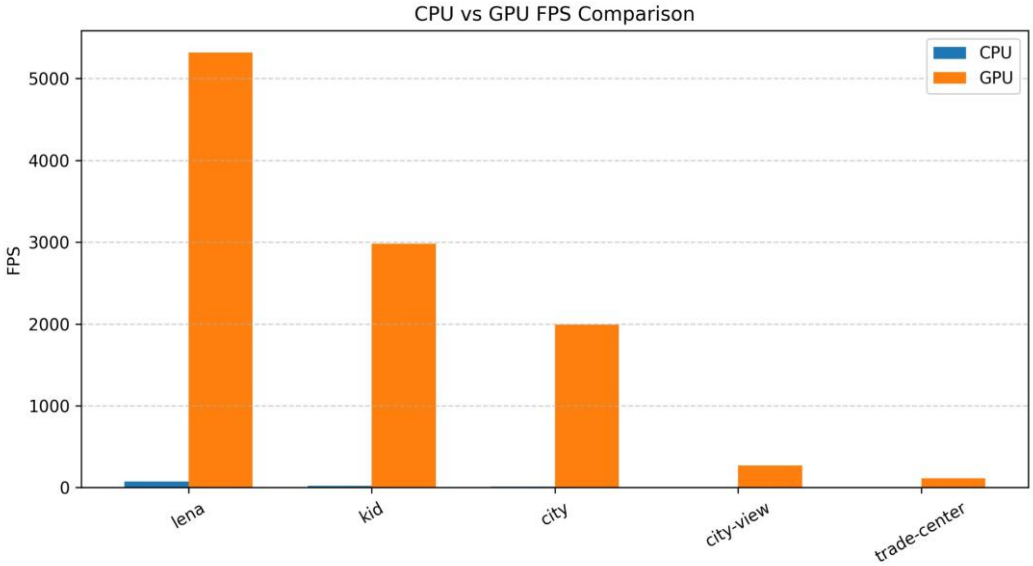
2. Experimental Setup

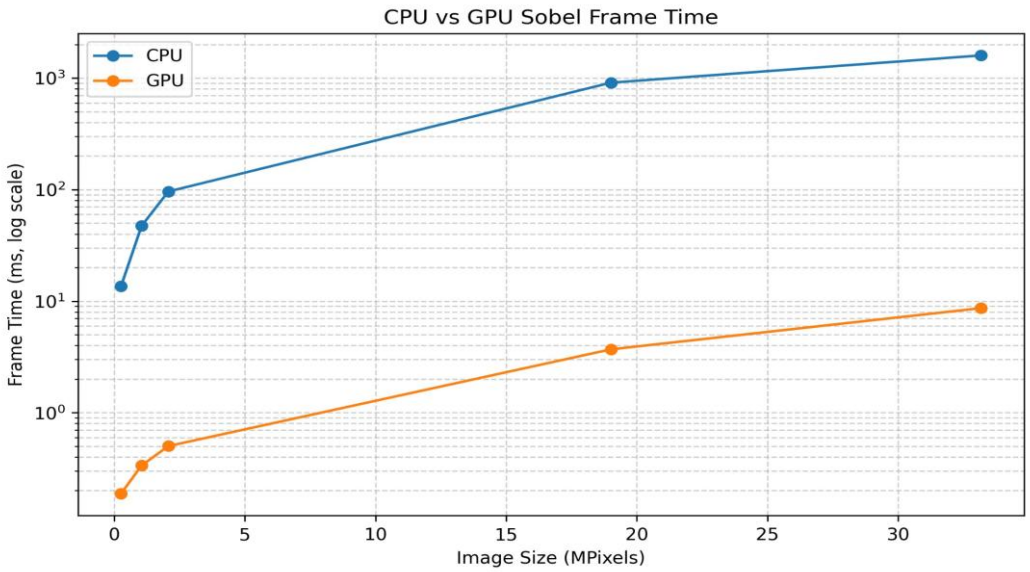
- CPU: Sequential baseline implementation
- GPU: CUDA kernel with 2D grid and thread blocks- Validation: Pixel-wise output match (CPU vs GPU) - Profiling Tools: NVIDIA Nsight Systems & Compute

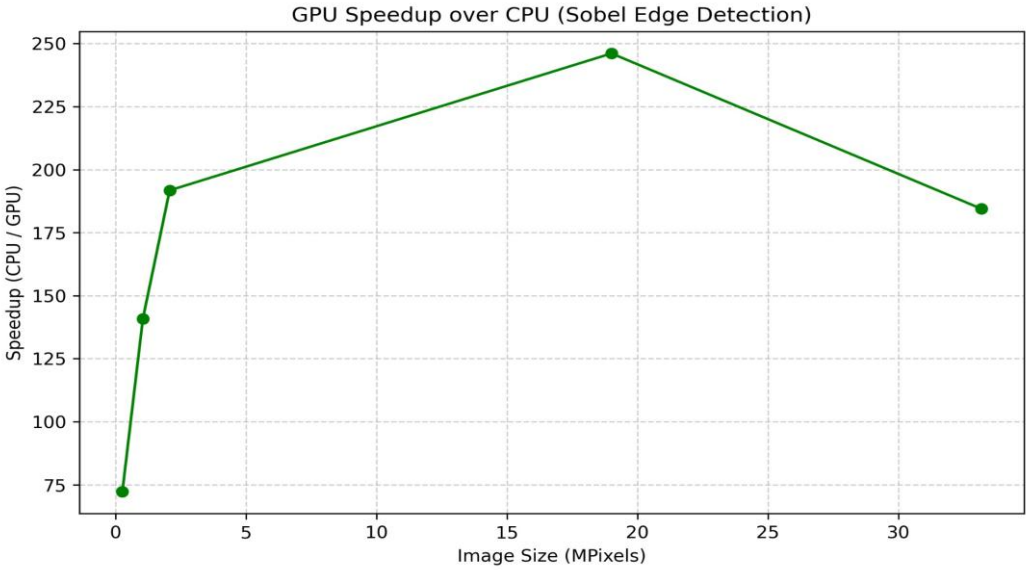
3. Image-wise Performance Summary

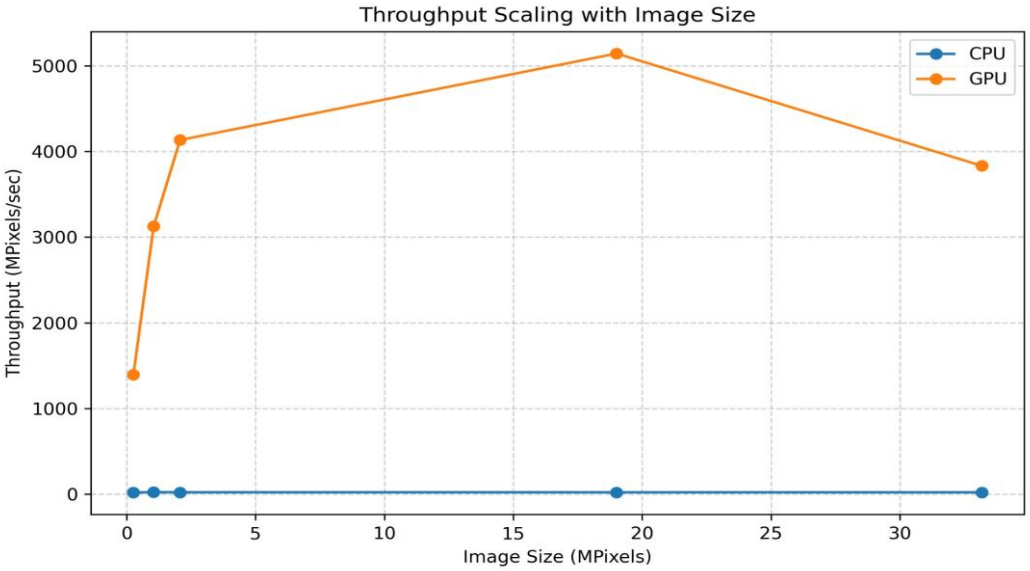
Image	**Resolution**	**MPixels**	**CPU Frame Time (ms)**	**GPU Frame Time** (ms)	**Speedup**
Lena	512×512	0.26	13.59	0.188	72×
Kid	1024×1024	1.05	47.27	0.336	141×
City	1920×1080	2.07	96.21	0.502	191×
City View	5824×3264	19.0	909.75	3.697	246×
Trade Center	7680×4320	33.2	1597.5	8.659	185x

4. Performance Scaling Graphs




CPU vs GPU Sobel Frame Time
CPU
GPU
Frame Time (ms, log scale)
10^3
10^2
10^1
10^0
0
5
10
15
20
25
30
Image Size (MPixels)


GPU Speedup over CPU (Sobel Edge Detection)
Speedup (CPU / GPU)
250
225
200
175
150
125
100
75
0
5
10
15
20
25
30
Image Size (MPixels)

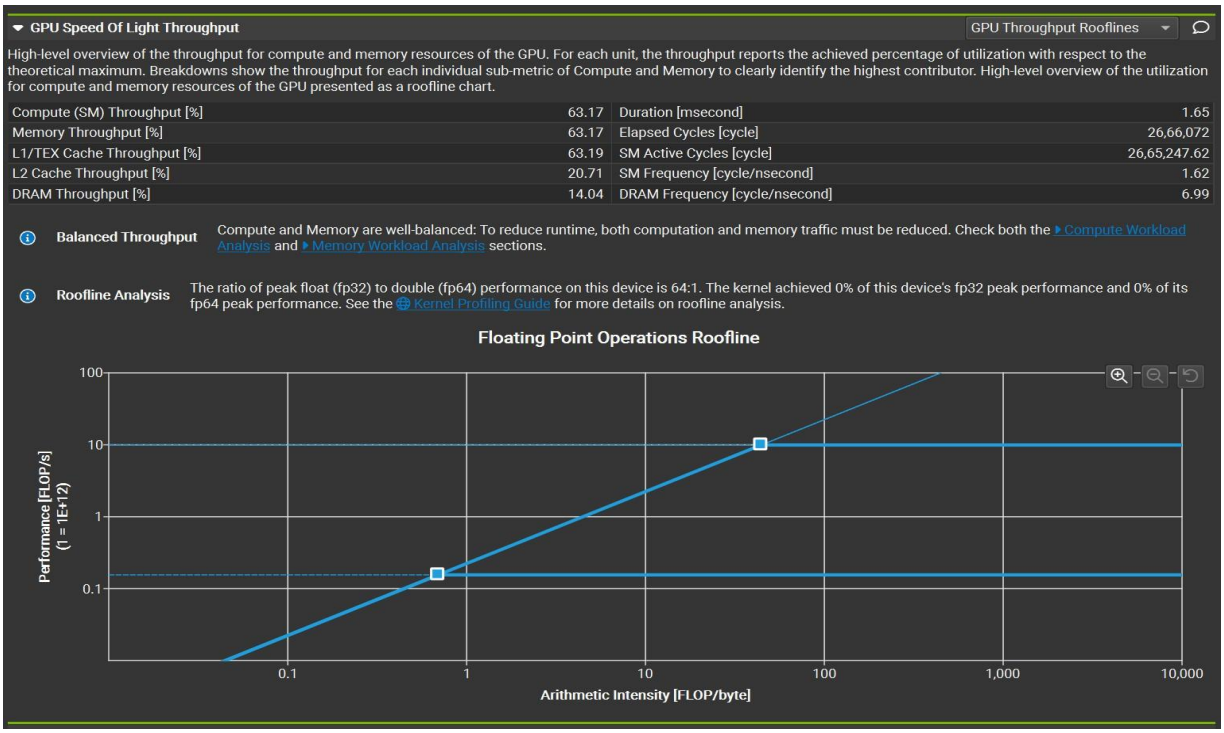

Throughput Scaling with Image Size
CPU
GPU
Throughput (MPixels/sec)
5000
4000
3000
2000
1000
0
0
5
10
15
20
25
30
Image Size (MPixels)

5. Key GPU Kernel Metrics (Nsight Compute) (for 8k image)

Metric	Value	Interpretation
Achieved Occupancy	91.46%	Excellent warp utilization
Kernel Runtime	1.65 ms	Pure compute time excluding overhead
Memory Throughput	31.4 GB/s	~63% of roofline bound

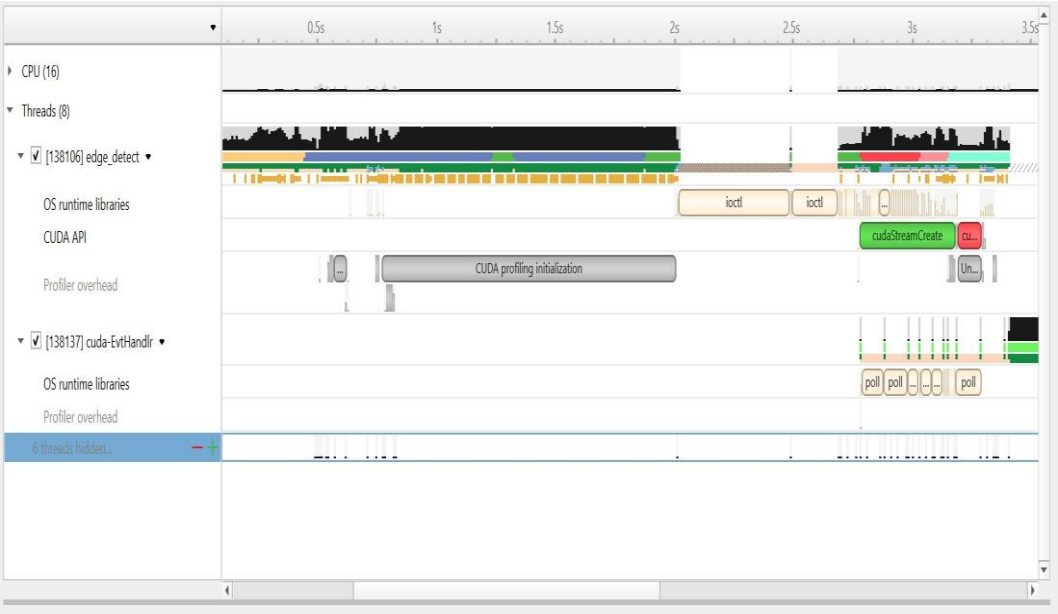
6. Roofline Analysis (for 8k Image)

The roofline model shows the Sobel kernel operating at approximately 63% of the attainable compute and memory throughput for its arithmetic intensity. The balanced slope and flat region intersection indicate that the kernel is neither purely memory-bound nor compute-bound, but well-balanced. Further optimizations would require reducing both arithmetic operations and memory traffic, such as shared memory tiling or fused kernels.



7. Nsight Systems Timeline Analysis

Timeline analysis confirms that kernel execution dominates GPU runtime, with minimal idle gaps and low synchronization overhead. Host-side setup costs are amortized across repeated runs, validating the benchmarking methodology.



8. Conclusion

This enhanced analysis demonstrates that the CUDA Sobel implementation achieves near-optimal occupancy, balanced roofline behavior, and up to 246× speedup over CPU for large images. The project exemplifies effective GPU utilization and serves as a strong foundation for further optimizations and research-grade extensions.