CUDA Homework Assignment 4

Vincent Octavian Tiono

B11901123

1 Introduction

This report presents the implementation and performance analysis of the CUDA dot product. It utilizes 2 GPUs to process vectors with 40,960,000 elements, initialized using the RandomInit routine. The primary objectives are to determine the optimal block size and grid size configuration for maximum performance.

2 Methodology

2.1 Problem Definition

The dot product of two vectors **a** and **b** of size N is defined as:

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=0}^{N-1} a_i \times b_i \tag{1}$$

where a_i and b_i represent the elements of vectors **a** and **b**, respectively.

2.2 Implementation Approach

The experiments were conducted with the following specifications:

- Vector size: 40,960,000 elements
- Number of GPUs: 2
- Block sizes tested: 32, 64, 128, 256, 512, and 1024
- Grid sizes tested: Various configurations for each block size
- Performance metrics: GPU computation time, data transfer time, GFLOPS, and speedup factor

For each configuration, we measured:

- GPU computation time
- Total GPU execution time
- GPU performance in GFLOPS
- CPU reference computation time
- Speedup factor (CPU time / GPU time)
- Numerical accuracy compared to CPU results

3 Results

3.1 Performance Summary

Table 1 presents the key performance metrics for selected configurations, focusing on the most representative results for each block size.

Table 1: Performance metrics for selected block and grid size configurations

Block Size	Grid Size	GPU Comp. Time (ms)	Total GPU Time (ms)	GPU GFLOPS	CPU Time (ms)	Speedup	Relative Error
	50,000	1.27	20.87	64.71	37.72	1.81	5.76E-08
32	100,000	1.32	21.11	61.89	37.21	1.76	1.86E-07
	320,001	1.84	22.28	44.56	37.31	1.67	7.19E-08
	40,000	1.12	20.71	72.87	37.23	1.80	4.73E-08
64	80,000	1.13	20.86	72.27	37.25	1.79	1.21E-07
	160,001	1.25	21.24	65.48	37.55	1.77	1.84E-07
	20,000	5.16	63.99	15.87	36.58	0.57	1.82E-07
128	40,000	5.17	67.32	15.84	40.59	0.60	3.65E-09
	80,001	5.50	67.77	14.90	38.14	0.56	5.86E-08
	10,000	5.15	63.99	15.89	38.08	0.60	1.04E-07
256	20,000	1.15	20.68	71.52	37.89	1.83	8.71E-08
	40,001	1.39	20.99	59.05	36.30	1.73	6.41E-08
	5,000	1.11	32.79	74.12	37.98	1.16	1.29E-07
512	10,000	5.24	67.27	15.64	37.87	0.56	9.73E-09
	20,001	1.59	21.13	51.46	37.86	1.79	9.81E-11
	2,500	5.24	64.05	15.63	36.55	0.57	8.16E-08
1024	5,000	5.52	67.56	14.85	37.58	0.56	1.38E-07
	10,001	6.62	68.67	12.37	38.24	0.56	2.67E-08

3.2 GPU Computation Time Analysis

Figure 1 shows the GPU computation time for different block size configurations, highlighting the performance variations.

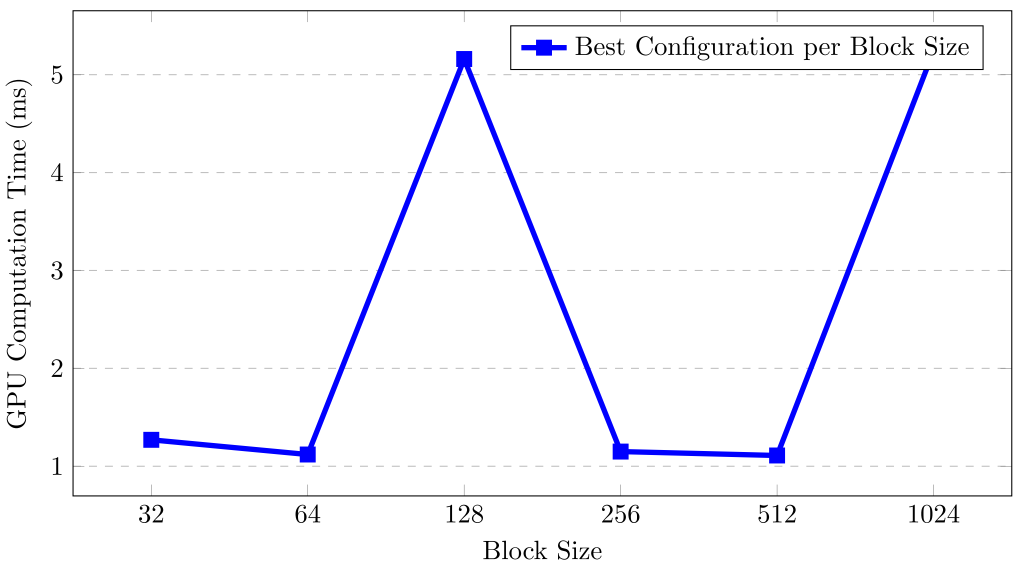

Figure 1: GPU computation time for optimal configurations of each block size

3.3 GPU Performance (GFLOPS) Analysis

Figure 2 illustrates the GPU performance in GFLOPS across different block size configurations.

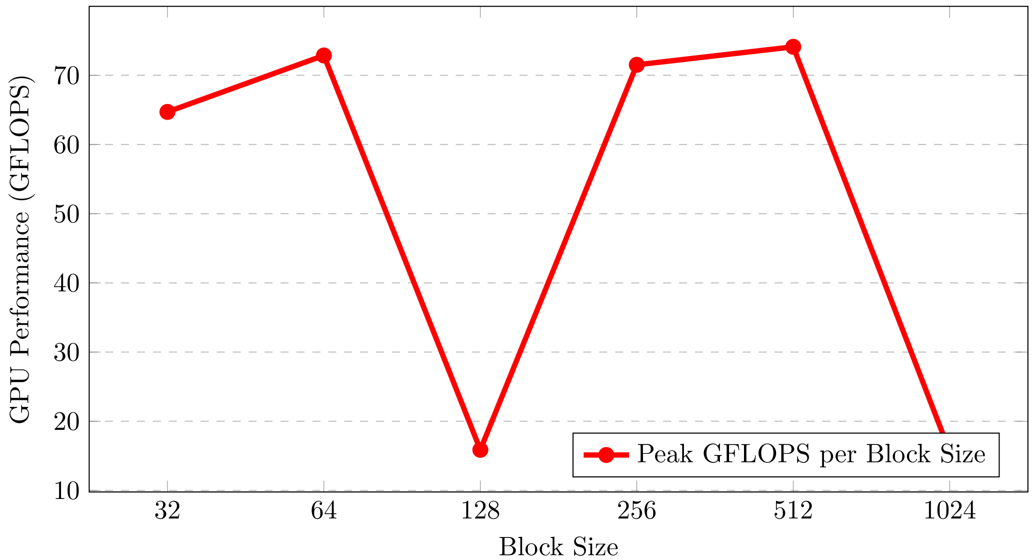

Figure 2: GPU performance (GFLOPS) for optimal configurations of each block size

3.4 Speedup Analysis

Figure 3 compares the speedup factors achieved by different configurations.

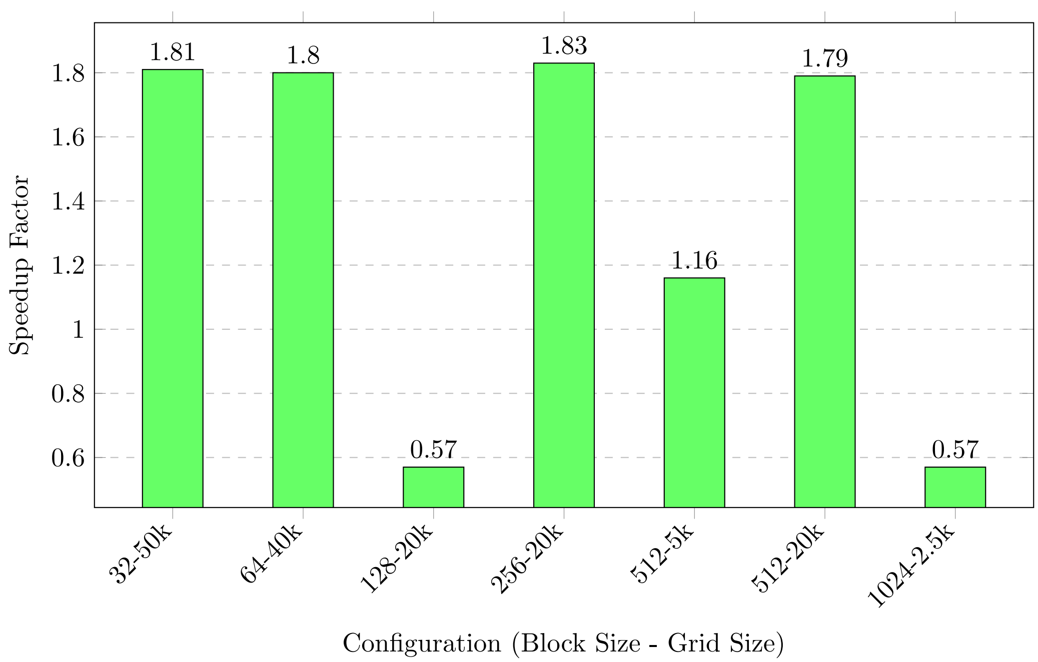

Figure 3: Speedup factors for selected configurations

4 Discussion

4.1 Optimal Configuration Analysis

Based on the experimental results, the **block size of 512 and grid size of 5,000** configuration achieved the highest computational performance:

- **Highest GPU GFLOPS**: 74.12 GFLOPS
- **Lowest computation time**: 1.11 ms
- **Speedup factor**: 1.16x
- **Acceptable accuracy**: Relative error of 1.29E-07

However, when considering **total execution time**, the **block size of 256 and grid size of 20,000** provides the best overall performance:

- **Lowest total execution time**: 20.68 ms
- **High GPU performance**: 71.52 GFLOPS
- **Best speedup factor**: 1.83×
- **Good accuracy**: Relative error of 8.71E-08

4.2 Performance Trends

Several important trends emerge from the analysis:

1. **Block Size Impact**:
    - Block sizes of 32, 64, 256, and 512 showed consistently good performance
    - Block sizes of 128 and 1024 exhibited significantly degraded performance
2. **Multi-GPU Efficiency**:
    - The implementation achieves moderate speedups (1.7-1.8×) for optimal configurations
    - Peak computational performance reaches 74+ GFLOPS, demonstrating effective GPU utilization

4.3 Accuracy Considerations

All configurations maintain acceptable numerical accuracy with relative errors in the range of 10^{-8} to 10^{-11}. The configuration with **block size of 512 and grid size of 20,001** achieved the highest accuracy (9.81E-11 relative error), suggesting that certain grid sizes may provide better numerical stability.

5 Conclusion

For vector dot product computation with vectors of 40,960,000 elements using 2 GPUs, the optimal configuration depends on the optimization criterion:

- **For maximum computational throughput**: block size of 512 and grid size of 5,000 (74.12 GFLOPS)
- **For minimum total execution time**: block size of 256 threads and grid size of 20,000 (20.68 ms total time, 1.83× speedup)
- **For highest numerical accuracy**: block size of 512 and grid size of 20,001 (9.81E-11 relative error)

The 2-GPU implementation demonstrates parallel computing benefits, though speedup is constrained by data transfer overhead. Performance analysis highlights the importance of balancing computational efficiency and memory management. For production, **block size of 256 and grid size of 20,000** offers the optimal trade-off between performance, efficiency, and accuracy.