1. CPU Performance

The CPU line increases steeply as the matrix size grows, demonstrating an exponential growth in execution time. At 2000x2000, the CPU task takes significantly longer (~41 seconds), while GPU-based methods remain below 0.1 seconds. The CPU method is computationally expensive and scales poorly with increasing matrix sizes.

2. GPU Standard Matrix Multiply

The GPU line shows vastly better performance compared to the CPU, with execution times consistently in the millisecond range. For smaller matrix sizes (e.g., 100x100), the GPU method is slower than cuBLAS and the Tiled GPU due to kernel launch overhead and memory transfers. As the matrix size increases, the GPU method performs significantly better, with execution times below 0.1 seconds even for 2000x2000 matrices.

3. Tiled GPU Matrix Multiply

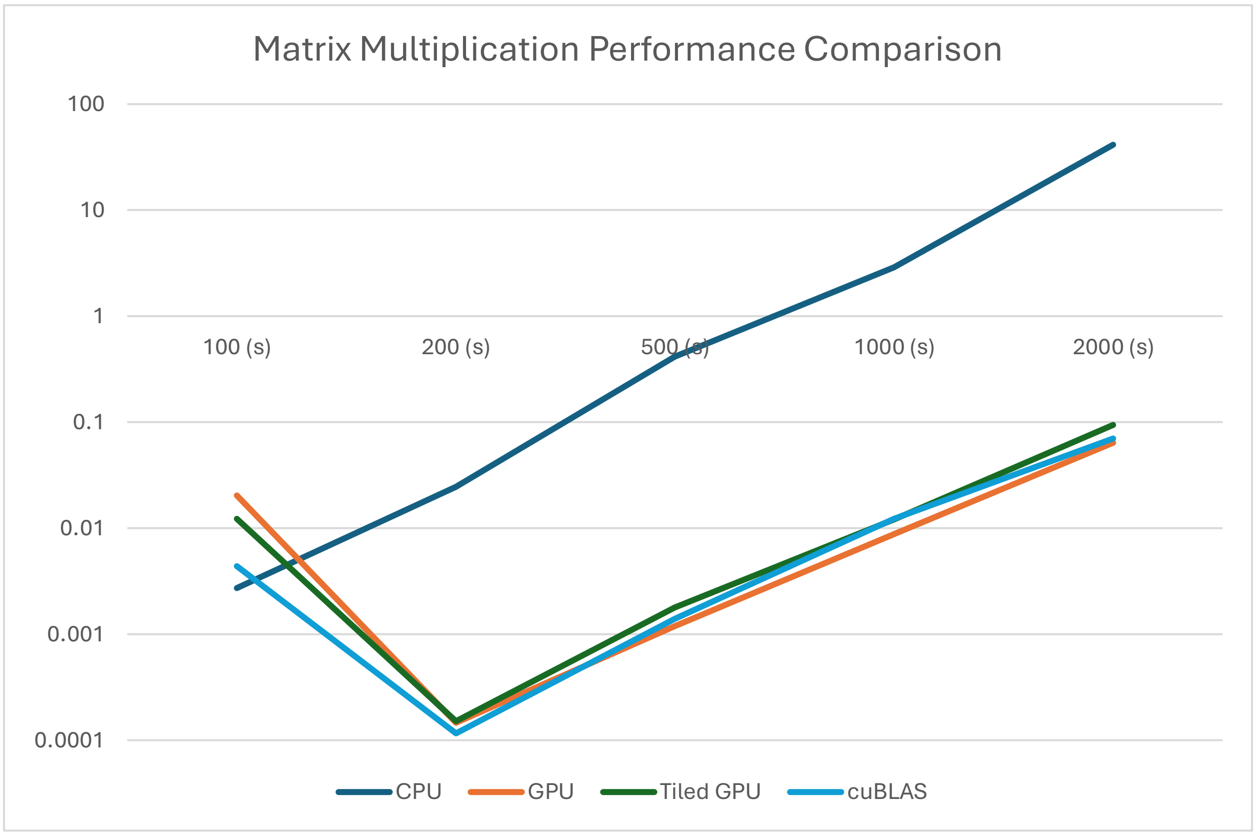
The Tiled GPU line shows consistent improvement over the standard GPU method, especially for larger matrices. At 500x500 and beyond, the tiled method's optimizations (e.g., using shared memory) result in slightly better performance compared to the standard GPU implementation. However, it is still slower than the cuBLAS method due to the latter's highly optimized library functions.

4. cuBLAS Matrix Multiply

The cuBLAS line consistently demonstrates the best performance across all matrix sizes. For 100x100 matrices, cuBLAS incurs minimal overhead, achieving the fastest times even for small sizes. At 2000x2000, cuBLAS remains competitive and performs just slightly slower than the tiled method for some sizes. This method is optimal for production-grade GPU matrix multiplication due to its efficiency.

5. Chart analysis

The chart uses a logarithmic scale on the y-axis, which helps visualize large variations in execution time across different tasks. The x-axis represents matrix sizes (100, 200, 500, 1000, 2000), while the four lines correspond to the different tasks: CPU, GPU, Tiled GPU, and cuBLAS.


Matrix Multiplication Performance Comparison
100
10
1
0.1
0.01
0.001
0.0001
100 (s)
200 (s)
500 (s)
1000 (s)
2000 (s)
CPU
GPU
Tiled GPU
cuBLAS