Raymond Zhu

1. From the experimental results, it can be seen that the latency values for the cache read (65.1131 ns) and main memory read (65.4872 ns) are quite close. The difference between the cache write latency (76.8761 ns) and the main memory write latency (78.8386 ns) is a bit bigger.
2. Comparing the bandwidth of the main memory under different data access granularities, it can be seen that the larger granularity leads to higher bandwidth. Comparing the bandwidth of the main memory under different read vs. write intensity ratios, it can be seen that the read-only takes the most bandwidth, then the write-only, then the 70:30 ratio, then finally the 50:50 ratio takes the least bandwidth.
3. As the thread count increases, the latency decreases, indicating better throughput with parallel execution. This follows the queuing theory predictions, where higher parallelism results in lower wait times, thus reducing latency.
4. As the size of the matrix increases, the execution time increases as well, reflecting the impact of higher cache miss ratios for larger matrices. Larger datasets lead to more cache misses, resulting in higher execution times.
5. As the strides got longer, the execution time decreased, indicating fewer TLB misses. Smaller strides cause frequent TLB misses, leading to significantly higher execution times as the system needs to access page tables more often.