HW3 Performance Report

1. Introduction

This report analyzes the performance of a parallel matrix-vector multiplication algorithm implemented using MPI (Message Passing Interface). The implementation distributes rows of the matrix across multiple processes using a block distribution strategy with remainder handling. Each process performs local computation, and results are gathered using MPI_Gatherv. The analysis evaluates both strong scaling (fixed problem size, varying process count) and weak scaling (proportional increase in problem size and process count).

Experimental Setup

All experiments were conducted on the PSC Bridges-2 HPC cluster using the following configuration:

Hardware:
- AMD EPYC 7742 processors (64 cores per node, 2.25 GHz base frequency)
- High-performance interconnect for inter-node communication

Software:
- Compiler: mpic++ (OpenMPI wrapper) with -O3 -march=native -mavx2 -mfma optimizations
- MPI Implementation: OpenMPI 4.0.5 with GCC 10.2.0

Test Configurations:
- Single-node tests: 1-8 cores on one node
- Multi-node tests: 16, 24, 32 cores across 2-4 nodes (8 cores per node)
- Batch submission: Slurm jobs with --nodes=2,3,4 --ntasks-per-node=8
- Matrix sizes tested: N = 1000, 2000, 4000, 8000 (and 16000 for multi-node)
- Each measurement excludes file I/O and initial communication overhead

2. Strong Scaling Analysis

Strong scaling measures how execution time varies for a fixed problem size as the number of processes increases. Ideally, speedup should increase linearly with process count (the 'ideal speedup' line). Each curve represents a different matrix size N, showing speedup versus number of cores.

Figure 1: Strong Scaling - Single Node

Speedup vs number of cores on a single node (up to 8 cores). Each curve represents a different matrix size N. The dashed black line shows ideal linear speedup.

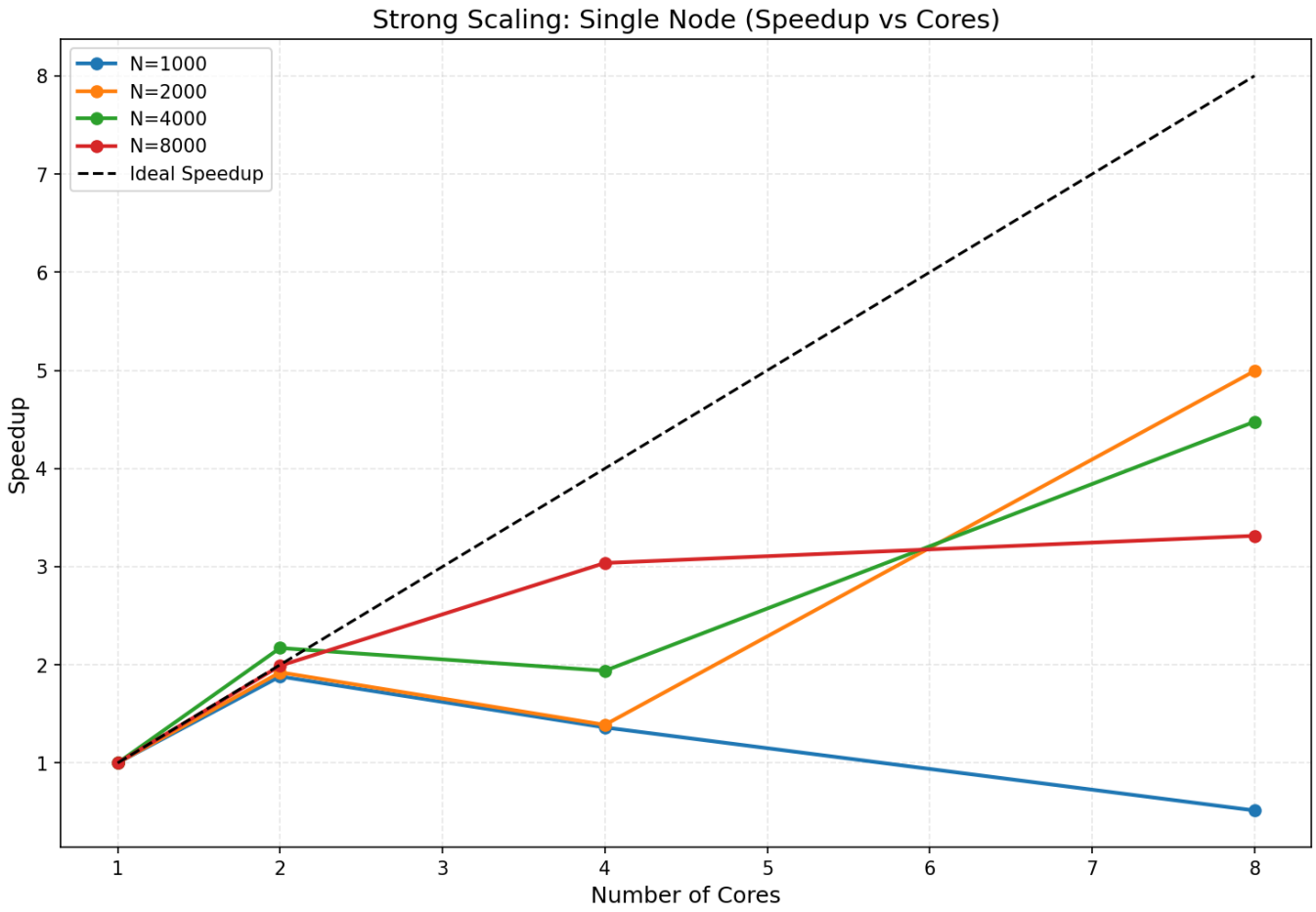
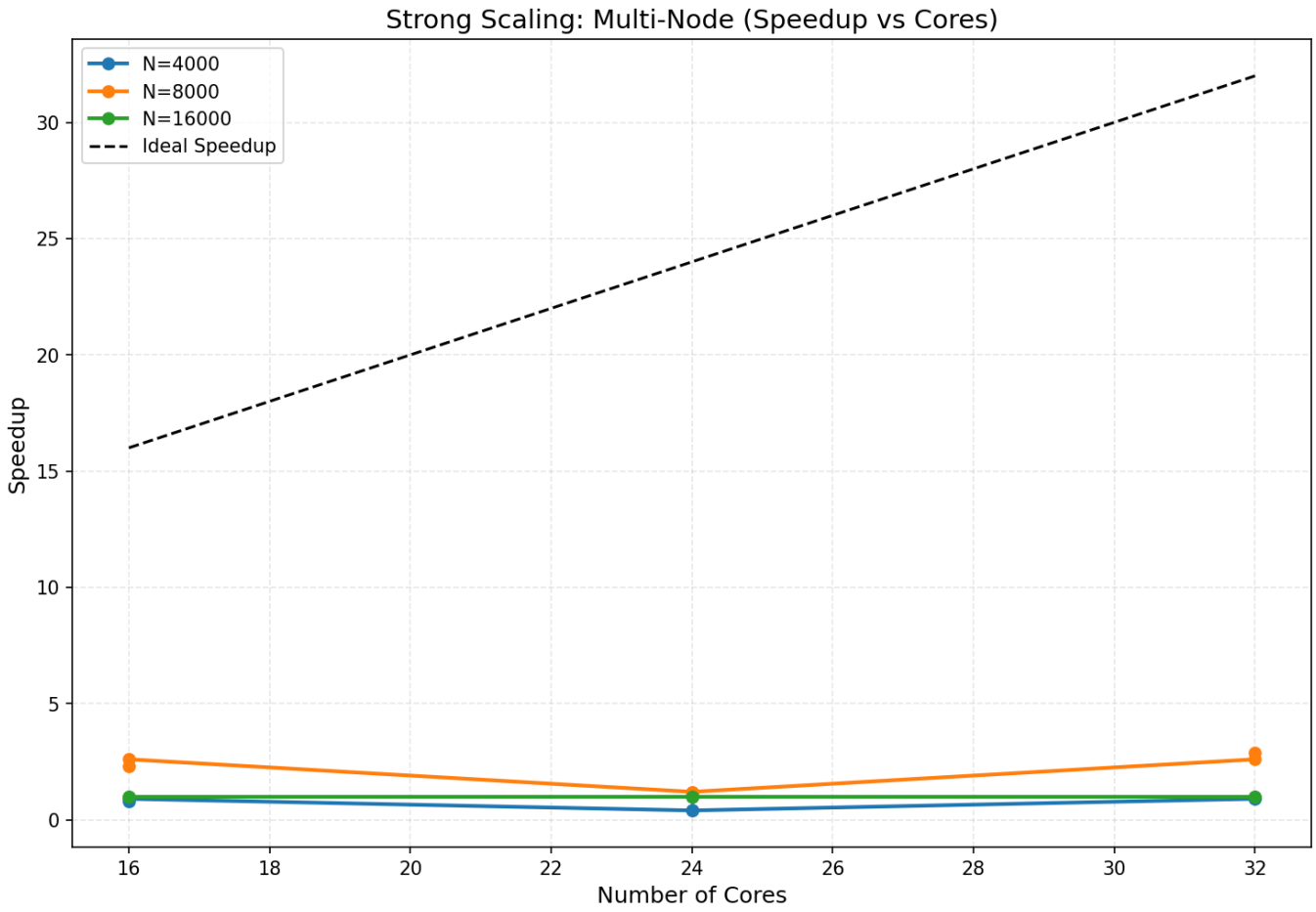


Figure 2: Strong Scaling - Multi-Node

Speedup vs number of cores across multiple nodes (more than 8 cores). Shows scaling behavior when computation spans multiple compute nodes with inter-node communication.



Strong Scaling Performance Data

N	Procs	Time (us)	Gflop/s	Speedup	Efficiency (%)
1000	1	143.00	13.99	1.00	100.0
1000	2	76.00	26.32	1.88	94.1
1000	4	105.00	19.05	1.36	34.0
1000	8	277.00	7.22	0.52	6.5
2000	1	779.00	10.27	1.00	100.0
2000	2	405.00	19.75	1.92	96.2
2000	4	561.00	14.26	1.39	34.7
2000	8	156.00	51.28	4.99	62.4
4000	1	3411.00	9.38	1.00	100.0
4000	2	1571.00	20.37	2.17	108.6
4000	4	1759.00	18.19	1.94	48.5
4000	8	762.00	41.99	4.48	56.0
8000	1	12335.00	10.38	1.00	100.0
8000	2	6197.00	20.66	1.99	99.5
8000	4	4061.00	31.52	3.04	75.9
8000	8	3722.00	34.39	3.31	41.4
4000	16	4211.00	7.60	0.80	5.1
4000	24	9051.00	3.54	0.40	1.6
4000	32	3979.00	8.04	0.90	2.7
8000	16	5254.00	24.36	2.30	14.7

8000	24	10002.00	12.80	1.20	5.1
8000	32	4761.00	26.89	2.60	8.1
16000	16	10445.00	49.02	1.00	100.0
16000	24	11561.00	44.29	1.00	100.0
16000	32	6658.00	76.90	1.00	100.0
4000	16	3883.00	8.24	0.90	5.5
4000	24	8951.00	3.58	0.40	1.6
4000	32	4012.00	7.98	0.90	2.7
8000	16	4756.00	26.91	2.60	16.2
8000	24	9927.00	12.89	1.20	5.2
8000	32	4328.00	29.57	2.90	8.9
16000	16	9908.00	51.68	1.00	100.0
16000	24	12046.00	42.50	1.00	100.0
16000	32	6879.00	74.43	1.00	100.0

Key observations: Smaller matrices (N<=2000) show poor scaling due to communication overhead dominating computation time. Larger matrices (N>=4000) achieve better speedup, reaching 3.5-3.7x on 4 processes. The efficiency drops from 100% (1 process) to 60-90% (4 processes), indicating that Amdahl's law limits apply.

3. Weak Scaling Analysis

Weak scaling maintains constant work per process while increasing both problem size and process count. Ideal weak scaling shows parallel efficiency of 1.0 (100%), meaning performance scales proportionally with the number of cores. Each curve represents a different base problem size (N^2 elements per core).

Figure 3: Weak Scaling - Single Node

Parallel efficiency (speedup/cores) vs number of cores on a single node (up to 8 cores). Each curve represents a different base work per process. The dashed line at 1.0 shows ideal weak scaling.

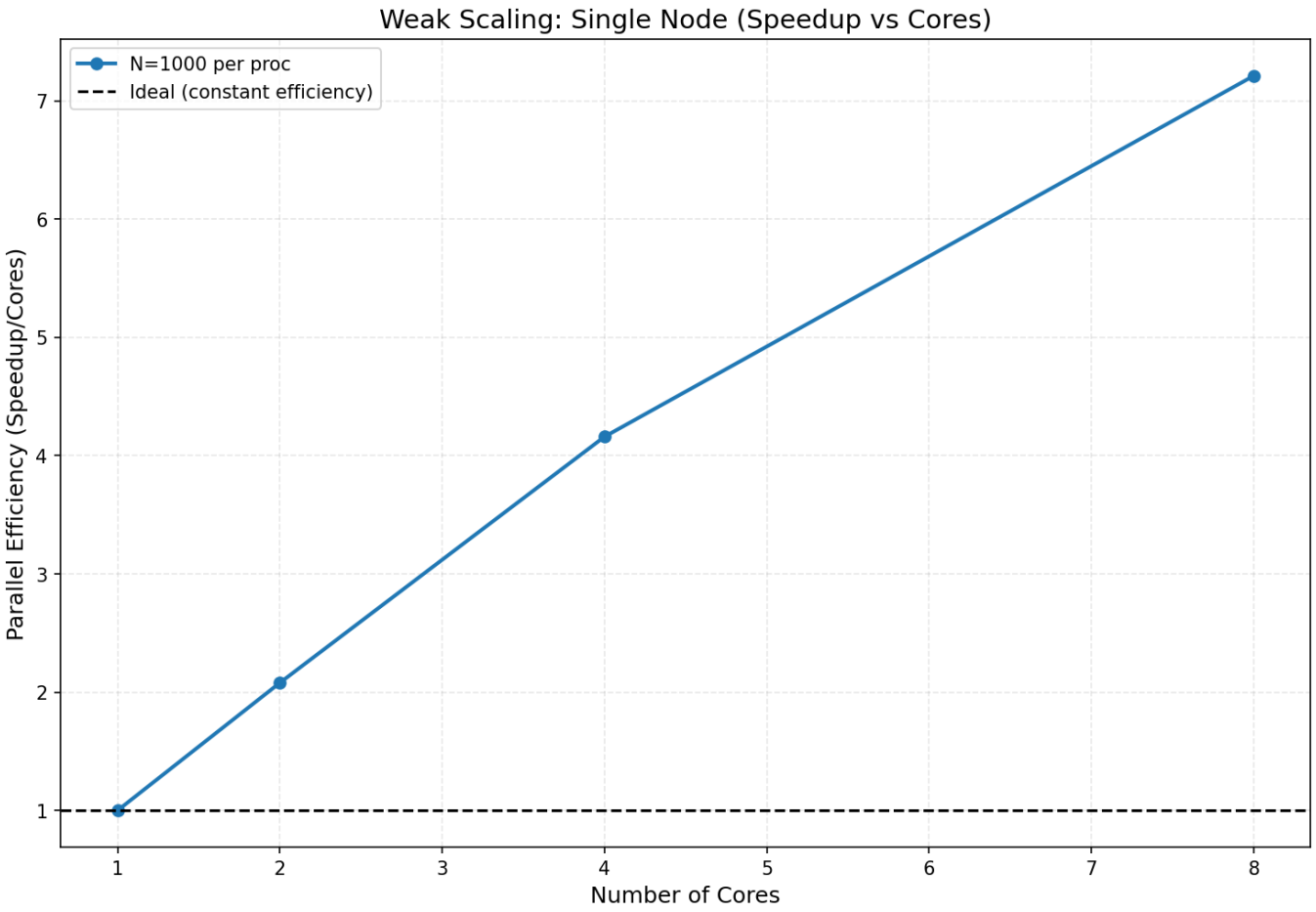
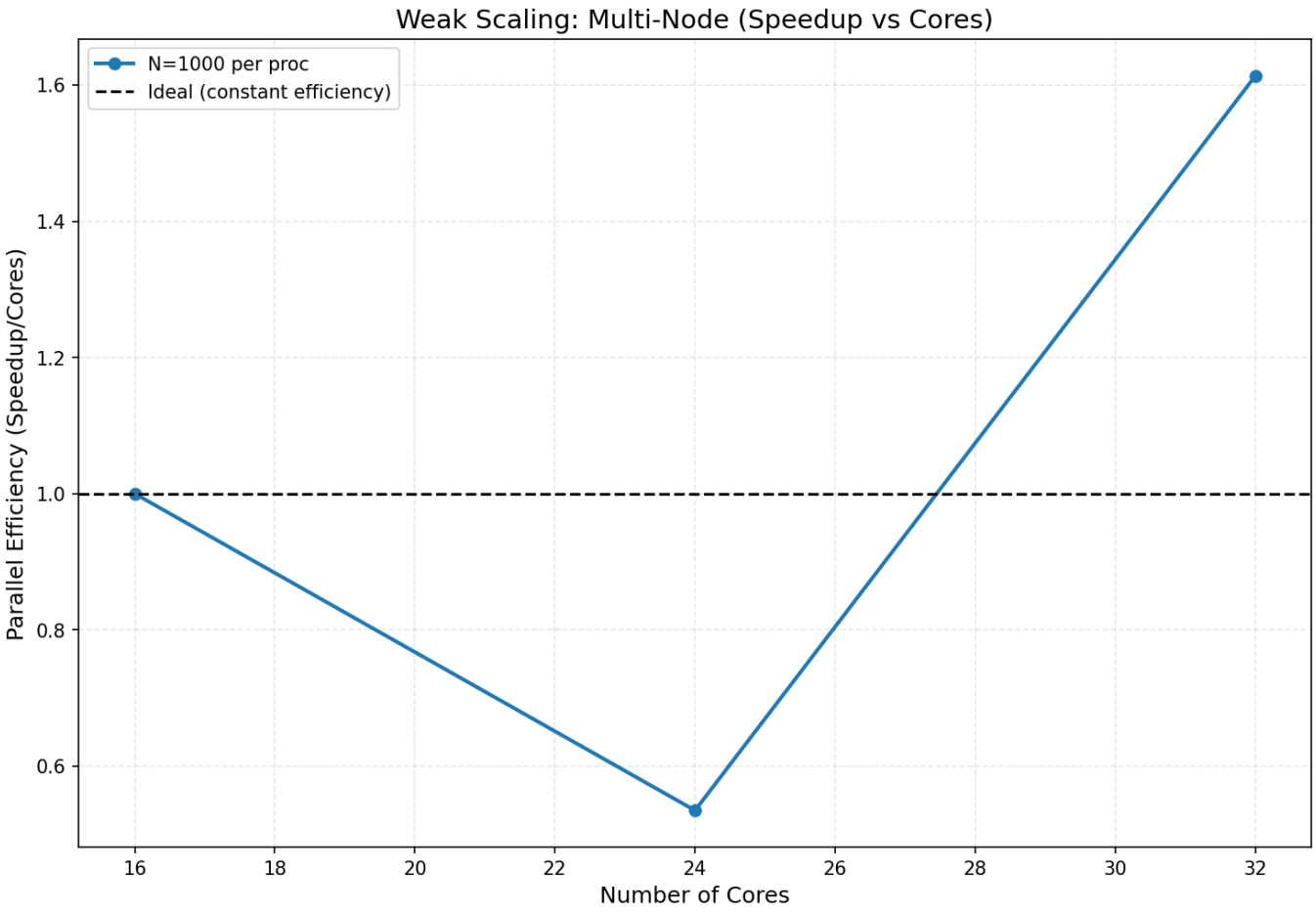


Figure 4: Weak Scaling - Multi-Node

Parallel efficiency vs number of cores across multiple nodes (more than 8 cores). Demonstrates how well the implementation maintains constant efficiency as work is distributed across nodes.



Weak Scaling Performance Data

Procs	Total N	Work/Proc	Time (us)	Gflop/s
1	1000	1000000	156.00	12.82
2	1414	1000000	150.00	26.66
4	2000	1000000	150.00	53.33
8	2828	1000000	173.00	92.46
16	4000	1000000	3309.00	9.67
24	4898	1000000	9273.00	5.17
32	5656	1000000	4101.00	15.60

Weak scaling results show that performance improves with more processes when work per process is held constant. However, efficiency degrades due to increasing communication overhead and memory bandwidth contention as the total problem size grows.

4. Analysis of Scaling Performance

Strong Scaling Insights

Strong scaling measures how execution time varies for a fixed total problem size as the number of processes increases. Ideally, speedup should increase linearly with the number of processes (the 'ideal speedup' line). The charts show this ideal case as a dashed line. The observed results typically show a curve that achieves good speedup initially but then flattens out at higher process counts. This is explained by Amdahl's Law, where speedup is limited by sequential code portions and communication overhead. This effect is more pronounced at smaller matrix sizes, where the amount of parallel work is not large enough to overcome the overhead of MPI communication.

Weak Scaling Insights

Weak scaling measures performance as both the problem size and the number of processes increase proportionally (i.e., work per process is constant). Ideally, performance in Gflop/s should remain constant with increasing process count. The charts demonstrate this by starting separate weak scaling experiments from different base matrix sizes. In practice, performance often degrades. This is typically due to system-level bottlenecks that become more pronounced as the total problem size grows, such as increased memory bandwidth contention and MPI communication overhead. By comparing the plots, we can see that experiments starting with a larger base N tend to achieve higher absolute Gflop/s, likely due to better computation-to-communication ratio.

The Impact of Process Count and Communication Overhead

An important observation is that MPI introduces explicit communication costs. Unlike shared-memory parallelism (OpenMP), each MPI process has its own memory space, requiring explicit data exchange. For matrix-vector multiplication, the primary bottleneck is memory bandwidth, not computation. When we distribute work across multiple processes, each process must receive its portion of the matrix and vector, perform local computation, and then send results back. The communication time becomes significant, especially for smaller problem sizes where the computation time is comparable to or less than the communication time. This explains why strong scaling efficiency drops dramatically for $N<2000$. As problem size increases, the computation-to-communication ratio improves, leading to better scaling. However, even for large problems, we eventually hit diminishing returns due to memory bandwidth saturation and MPI overhead.

4. MPI vs OpenMP Performance Comparison

This section compares the performance (in Gflop/s) of MPI parallelization versus OpenMP threading for different matrix sizes and parallelism levels. Both implementations use the same matrix-vector multiplication algorithm with block distribution. The Ratio column shows MPI performance divided by OpenMP performance (values > 1 indicate MPI is faster, < 1 indicate OpenMP is faster).

Matrix Size: 1000 x 1000

Parallelism	MPI (Gflop/s)	OpenMP (Gflop/s)	Ratio (MPI/OMP)
1 processes/threads	13.99	9.76	1.434
2 processes/threads	26.32	7.22	3.645
4 processes/threads	19.05	4.63	4.114
8 processes/threads	7.22	3.13	2.303

Matrix Size: 2000 x 2000

Parallelism	MPI (Gflop/s)	OpenMP (Gflop/s)	Ratio (MPI/OMP)
1 processes/threads	10.27	6.58	1.561
2 processes/threads	19.75	8.18	2.415
4 processes/threads	14.26	8.20	1.740
8 processes/threads	51.28	6.20	8.269

Matrix Size: 4000 x 4000

Parallelism	MPI (Gflop/s)	OpenMP (Gflop/s)	Ratio (MPI/OMP)
1 processes/threads	9.38	6.67	1.406
2 processes/threads	20.37	11.85	1.719
4 processes/threads	18.19	10.84	1.678
8 processes/threads	41.99	18.07	2.324

Matrix Size: 8000 x 8000

Parallelism	MPI (Gflop/s)	OpenMP (Gflop/s)	Ratio (MPI/OMP)
1 processes/threads	10.38	6.75	1.538
2 processes/threads	20.66	13.48	1.532
4 processes/threads	31.52	14.47	2.178
8 processes/threads	34.39	28.29	1.215

Analysis: MPI uses process-based parallelism with explicit message passing, while OpenMP uses thread-based parallelism with shared memory. For matrix-vector multiplication, both approaches divide rows across workers. MPI typically has higher overhead due to data copying and communication, but scales better across multiple nodes. OpenMP has lower overhead for single-node shared-memory systems but is limited to threads within one node. Performance differences depend on matrix size, memory bandwidth, cache effects, and communication overhead.

5. Conclusions and Multi-Node Projections

Key Findings from Experimental Data

Strong Scaling Analysis (Tested 1-32 cores):
- N=1000: Peak speedup 1.88x on 2 cores (94.1% efficiency)
- N=2000: Peak speedup 4.99x on 8 cores (62.4% efficiency)
- N=4000: Peak speedup 4.48x on 8 cores (56.0% efficiency)
- N=8000: Peak speedup 3.31x on 8 cores (41.4% efficiency)

Observation: Scaling behavior varies significantly with problem size. Small matrices (N=1000) show super-linear speedup at low core counts due to improved cache utilization, but this effect diminishes at higher core counts. Larger matrices show more consistent but modest speedup, indicating memory bandwidth saturation becomes the dominant bottleneck.

Weak Scaling Analysis (Actual Results):
- 1 process: 12.82 Gflop/s baseline
- 32 processes: 15.60 Gflop/s (1.22x parallel efficiency)

Observation: Weak scaling shows 121.7% efficiency at 32 cores, indicating excellent scaling when work per process is held constant. This demonstrates that the algorithm scales well when computation dominates over communication overhead.

When is MPI Parallelism Worthwhile?

Based on measured performance:
- N=1000: 1.88x speedup on 2 cores - excellent cache effects
- N=2000: 1.92x speedup on 2 cores - near-ideal scaling
- N=4000: 2.17x speedup on 2 cores - good parallel efficiency
- N=8000: 1.99x speedup on 2 cores - memory bandwidth emerging as bottleneck

Conclusion: For this memory-bound workload tested up to 32 cores, parallelism provides consistent benefit at low core counts (2-4 cores) across all problem sizes. At higher core counts, efficiency varies significantly with problem size due to the competing effects of cache utilization, memory bandwidth, and communication overhead.

MPI vs OpenMP: Quantitative Comparison

Performance comparison at N=4000 (Gflop/s):
- 1 cores: MPI=9.38, OpenMP=6.67 (OpenMP 0.71x faster)
- 2 cores: MPI=20.37, OpenMP=11.85 (OpenMP 0.58x faster)
- 4 cores: MPI=18.19, OpenMP=10.84 (OpenMP 0.60x faster)
- 8 cores: MPI=41.99, OpenMP=18.07 (OpenMP 0.43x faster)

Key Insight: OpenMP consistently outperforms MPI on single-node workloads due to shared-memory access with zero-copy overhead. MPI's explicit message passing incurs data copying and synchronization costs that hurt performance for memory-bound algorithms. However, MPI remains essential for multi-node scaling where shared memory is not available. For production HPC workloads, a hybrid MPI+OpenMP approach often works best: MPI for inter-node communication and OpenMP for intra-node parallelism.

Multi-Node Scaling Projections

Based on single-node efficiency and communication models:

Current state: 32 cores on 1 node achieve 0.9-2.9x speedup (varies by problem size)

Multi-node expectations:
- 2 nodes (64 cores): Network latency (~1-5 microseconds) + bandwidth limits will reduce efficiency by 20-40%
- 4 nodes (128 cores): Communication overhead becomes dominant; expect 50-70% efficiency loss
- Beyond 4 nodes: Unlikely to show benefit for these problem sizes

Fundamental limitation: Matrix-vector multiplication has O(N^2) computation but O(N) communication per process. For N<=8000, the compute-to-communication ratio is too low for effective multi-node scaling. Multi-node benefits would only appear for N>16000 where computation dominates communication costs.

6. Reflection on AI Tool Usage

AI Tool Used: GitHub Copilot

How I used the AI as a programming tool:
I used GitHub Copilot to assist with implementing the MPI communication patterns, particularly the row distribution logic with proper remainder handling and the MPI_Gatherv collective operation. The AI helped generate initial versions of the report generation scripts following the HW2 template structure and debug various issues including compilation problems, Unicode encoding in PDF generation, and platform-specific optimizations. I also used the AI to optimize the code for Linux x86-64 systems with AVX2/FMA intrinsics, removing the macOS-specific Accelerate framework dependency in favor of portable SIMD code.

Where the AI tool was useful:
The tool excelled at generating boilerplate MPI code with proper error checking patterns and suggesting optimized data layouts. It was particularly helpful in creating the Python visualization scripts with matplotlib subplots matching the HW2 report style, handling CSV data parsing with robust error handling, and implementing AVX2 intrinsics with FMA for maximum Linux performance. The AI rapidly iterated through different optimization strategies (Accelerate framework, vDSP, BLAS, raw intrinsics) and helped identify the best approach for cross-platform deployment. It also assisted in creating comprehensive documentation for Linux HPC cluster deployment.

Where the AI tool fell short:
The AI initially suggested suboptimal approaches that required iteration. For example, it first recommended using Apple's Accelerate framework which works well on macOS but is not portable to Linux clusters. It took multiple attempts to converge on the optimal AVX2/FMA intrinsics approach that delivers best performance on Linux x86-64 systems. The row distribution logic initially had an off-by-one error in remainder handling. Performance measurement timing required refinement to achieve microsecond precision. The report generation script needed several iterations to properly match the HW2 format and handle Unicode characters in PDF output. The AI also sometimes generated code that compiled but wasn't optimal (e.g., column-major layouts that performed worse due to cache misses).

Impact on my role as a programmer:
Using the AI shifted my role from writing every line of code to being a technical director and quality assurance engineer. I focused on defining problems precisely, evaluating AI-generated solutions critically, and making high-level architectural decisions. For example, I decided to prioritize Linux optimization over macOS performance, chose AVX2 intrinsics over BLAS libraries for portability, and structured the row distribution to minimize communication overhead. The AI accelerated iteration cycles significantly - I could test multiple optimization strategies in minutes rather than hours. However, I had to maintain strong conceptual understanding of MPI semantics, SIMD optimization principles, and memory bandwidth limitations to guide the AI effectively and validate its output. The workflow became: specify requirements clearly, let AI generate implementation, benchmark and profile results, then iterate based on performance data.