CPSC 524 Assignment 3: Parallel Matrix Multiplication

Rami Pellumbi*

November 14, 2023

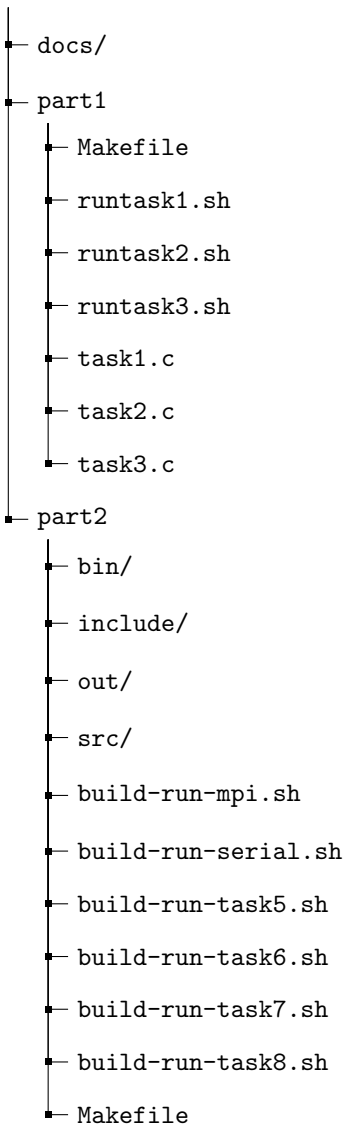
*M.S., Statistics & Data Science

1 Introduction

This assignment investigates the implementation and performance of an MPI- based program tasked with parallel computation of large double precision matrices. The program's objective is to effectively utilize multiple processors to perform the multiplication, optimizing the use of system resources and minimizing computation time. The implementation required careful consideration of data distribution and process synchronization, ensuring accuracy while maximizing computational efficiency. Performance analysis focuses on the scalability of the solution across various matrix sizes and the number of processors used. This report outlines the design considerations, describes the implementation strategy, and presents an evaluation of the program's computational performance.

2 Project Organization

The project is laid out as follows:



- **docs/**: This folder contains LaTeX files and other documentation materials that pertain to the report.
- **part1/**: Code for MPI basics and part 1 tasks.
- **part2/**: Code for parallel matrix multiplication.
  - **bin/**: The `bin` folder holds compiled objects and executable files, centralizing the output of the compilation process.
  - **include/**: Here, all the header files (`.h`) are stored.
  - **out/**: The `out` folder stores the outputs from each task. It also houses the csv file containing data generated by the programs.
  - **src/**: This directory houses the source files (`.c`) that make up the benchmarks.
  - **Shell Scripts**: The shell scripts are used to submit the job for the relevant task to slurm via `sbatch`.

3 Code Explanation, Compilation, and Execution

This section outlines the steps required to build and execute the code. The provided Bash scripts automate the entire process, making it straightforward to compile and run the code. All the below steps assume you are in the root of the project directory.

3.1 Automated Building and Execution

- **Part 1: Understanding MPI Basics** All part 1 related code is in the `part1/` directory. There are three scripts for part 1, one for each task The commands `sbatch runtask1.sh`, `sbatch runtask2.sh`, and `sbatch runtask3.sh` will compile and run the files `task1.c`, `task2.c`, and `task3.c`, respectively.
- **Part 2: Parallel Matrix Multiplication** All part 2 related code is in the `part2/` directory. There are multiple programs run in Part 2:
  - `serial.c` Simple, serial, matrix multiplication. Use `build-run-serial.sh` script to run.
  - `task5.c`: Parallel matrix multiplication of square matrices with blocking collectives. Assumes that the number of rows is evenly divisible by the number of processes.
  - `task6.c`: Parallel matrix multiplication of square matrices with non blocking collectives. Assumes that the number of rows is evenly divisible by the number of processes.
  - `task7.c`: Parallel matrix multiplication of square matrices with non blocking collectives, with a focus on interleaving communication and computation. Assumes that the number of rows is evenly divisible by the number of processes.
  - `taks8.c` Parallel matrix multiplication of square matrices with blocking collectives. Appropriately handles when the number of rows is not evenly divisible by the number of processes.

  To run all the MPI experiments, execute `sbatch build-run-mpi.sh` from the `part2` directory. Additionally, each task has its own individual run script: `build-run-taskX.c`, where `X` is one of $5, 6, 7$, or 8. The results will be stored in `part2/out/results.csv`.

3.2 Post-Build Objects and Executables

For part 2, upon successful compilation and linking, an `obj/` subdirectory will be generated within the directory. This directory will contain the compiled output files. Additionally, the executable files for running each part will be situated in the `bin/` subdirectory.

3.3 Output Files From `sbatch`

For part 1, the output files generated from running the code by submitting the relevant Bash script via `sbatch` will be in the `part1` directory. For part 2, the output files generated from running the code by submitting the relevant Bash script via `sbatch` will be stored in the `out` directory.

4 MPI Basics

4.1 Task 1: Basic Communication Between Processes

In task 1, we simply run the provided `task1.c` code. We observe that the print statements and appear in first-come-first-served order.

```
Node List:
r918u05n[01-02]
ntasks-per-node =  2

Run 1
Message printed by manager: Total elapsed time is 0.005460 seconds.
From process 2: I worked for 5 seconds after receiving the following message:
    Hello, from process 0.
From process 3: I worked for 10 seconds after receiving the following message:
    Hello, from process 0.
From process 1: I worked for 15 seconds after receiving the following message:
    Hello, from process 0.
real 0m15.668s
user 0m0.143s
sys 0m0.316s

Run 2
Message printed by manager: Total elapsed time is 0.006188 seconds.
From process 2: I worked for 5 seconds after receiving the following message:
    Hello, from process 0.
From process 3: I worked for 10 seconds after receiving the following message:
    Hello, from process 0.
From process 1: I worked for 15 seconds after receiving the following message:
    Hello, from process 0.
real 0m15.779s
user 0m0.146s
sys 0m0.318s

Run 3
Message printed by manager: Total elapsed time is 0.007760 seconds.
From process 2: I worked for 5 seconds after receiving the following message:
    Hello, from process 0.
From process 3: I worked for 10 seconds after receiving the following message:
    Hello, from process 0.
From process 1: I worked for 15 seconds after receiving the following message:
    Hello, from process 0.
real 0m15.614s
user 0m0.142s
sys 0m0.301s
```

4.2 Task 2: Handling Prints in the Manager

Here, we modify the provided `task1.c` remove prints from the non-manager processes and instead have the manager print out the messages in rank order. To handle this, each worker now sends its message back to the manager:

```
sprintf(message, "Hello manager, from process %d after working %d seconds.",
        rank, worktime);
/* Send message back to manager */
MPI_Send(message, strlen(message) + 1, MPI_CHAR, 0, type, MPI_COMM_WORLD);
```

which waits for them like so

```
/* Receive messages from the workers */
for (size_t i = 1; i < size; i++)
{
    MPI_Recv(message, 100, MPI_CHAR, i, type, MPI_COMM_WORLD, &status);
    sleep(3);
    printf("Message from process %d: %s\n", i, message);
}
```

The print out of this program correctly shows the messages printing out in rank order:

```
Node List:
r918u05n[01-02]
ntasks-per-node =  2

Run 1
Message from process 1: Hello manager, from process 1 after working 15 seconds.
Message from process 2: Hello manager, from process 2 after working 5 seconds.
Message from process 3: Hello manager, from process 3 after working 10 seconds.
Message printed by manager: Total elapsed time is 24.021217 seconds.
real 0m24.974s
user 0m15.186s
sys 0m0.231s

Run 2
Message from process 1: Hello manager, from process 1 after working 5 seconds.
Message from process 2: Hello manager, from process 2 after working 10 seconds.
Message from process 3: Hello manager, from process 3 after working 15 seconds.
Message printed by manager: Total elapsed time is 18.012530 seconds.
real 0m18.572s
user 0m9.184s
sys 0m0.281s

Run 3
Message from process 1: Hello manager, from process 1 after working 15 seconds.
Message from process 2: Hello manager, from process 2 after working 10 seconds.
Message from process 3: Hello manager, from process 3 after working 5 seconds.
Message printed by manager: Total elapsed time is 24.000228 seconds.
real 0m24.574s
user 0m15.176s
sys 0m0.219s
```

4.3 Task 3: Consistent Performance

In `task2.c`, we observe a different elapsed time based on the work done by each process. Since the receive is waiting on the ranks in rank order, there are scenarios where there is not an efficient overlap between simulated work and communication. To remedy this, the manager's receive is changed to any source and the messages are stored in a new array of messages instantiated by the manager. We use the `status.MPI_SOURCE`

from the status of the receive to appropriately place into the array. Concretely, the manager now does the following.

```
if (rank == 0)
{
    char **messages = (char **)malloc((size - 1) * sizeof(char *));
    for (int i = 0; i < size - 1; ++i)
    {
        messages[i] = (char *)malloc(100 * sizeof(char));
    }

    sparm = rwork(0, 0); // initialize the workers' work times

    /* Create the message using sprintf */
    sprintf(message, "Hello, from process %d.", rank);

    MPI_Barrier(MPI_COMM_WORLD);
    wct0 = MPI_Wtime();

    /* Send the message to all the workers, which is where the work happens */
    for (i = 1; i < size; i++)
    {
        MPI_Send(message, strlen(message) + 1, MPI_CHAR, i, type, MPI_COMM_WORLD);
        MPI_Send(&sparm, 1, MPI_INT, i, type, MPI_COMM_WORLD);
    }

    // Receive messages from workers as they complete
    for (int i = 1; i < size; i++)
    {
        MPI_Recv(message, 100, MPI_CHAR, MPI_ANY_SOURCE, type, MPI_COMM_WORLD, &status);
        sleep(3); // Simulate post processing
        strcpy(messages[status.MPI_SOURCE - 1], message);
    }

    /* print messages */
    for (size_t i = 1; i < size; i++)
    {
        printf("Message from process %d: %s\n", i, messages[i - 1]);
    }

    wct1 = MPI_Wtime();
    // timing(&wct1, &cput); // get the end time using the original timing routine
    total_time = wct1 - wct0;
    printf("Message printed by manager: Total elapsed time is %f seconds.\n", total_time);

    for (int i = 0; i < size - 1; i++)
    {
        free(messages[i]);
    }
    free(messages);
}
```

The printout from this execution observed consistent, optimal runtime:

```
Node List:
r918u05n[01-02]
ntasks-per-node =  2

Run 1
Message from process 1: Hello manager, from process 1 after working 15 seconds.
Message from process 2: Hello manager, from process 2 after working 5 seconds.
Message from process 3: Hello manager, from process 3 after working 10 seconds.
Message printed by manager: Total elapsed time is 18.001310 seconds.
real 0m18.730s
user 0m9.118s
sys 0m0.201s

Run 2
Message from process 1: Hello manager, from process 1 after working 5 seconds.
Message from process 2: Hello manager, from process 2 after working 15 seconds.
Message from process 3: Hello manager, from process 3 after working 10 seconds.
Message printed by manager: Total elapsed time is 18.000172 seconds.
real 0m18.589s
user 0m9.130s
sys 0m0.300s

Run 3
Message from process 1: Hello manager, from process 1 after working 10 seconds.
Message from process 2: Hello manager, from process 2 after working 5 seconds.
Message from process 3: Hello manager, from process 3 after working 15 seconds.
Message printed by manager: Total elapsed time is 18.000061 seconds.
real 0m18.584s
user 0m9.107s
sys 0m0.267s
```

It is evident that the post processing time is better overlapped with the communication by communicating in this way, thus leading to a faster runtime.

5 Parallel Matrix Multiplication

At a high level, we have input $N \times N$ matrices A and B, and an output $N \times N$ matrix C.. The input matrices, A and B, are divided into blocks of rows and columns, respectively. Each block row of matrix A is then multiplied with the corresponding block column of matrix B to compute a segment of the resultant matrix C. This block-wise computation enables the distribution of the workload among the available processors in the MPI environment. The distribution and collection of data blocks utilize MPI's collective communication operations, ensuring that each processor has the necessary data to perform its assigned calculations. The final step involves aggregating the computed blocks from all processors to form the complete resultant matrix. The implementation of the parallel matrix multiplication algorithm follows the ring-pass strategy, which is outlined as follows for the case where N is divisible by p (the number of processes):

1. The manager initializes memory of A, B, and C and initializes A and B with the data to be multiplied.
2. The manager partitions A and C into p blocks rows each. It partitions B into p block columns.
   - A and C are of dimension $(N/p) \times N$.
   - B is of dimension $N \times (N/p)$.
3. The manager permanently assigns to each MPI process one block row each of A and C, and it assigns each MPI process one block column of B as its initial assignment.

4. The computation proceeds iteratively until completed:
   (a) Each MPI process does all the computation it can with the data at hand.
   (b) MPI processes then pass their block columns of B to the next higher-ranked MPI process.
5. The manager uses MPI collective operations to assemble the full C matrix by collecting all block rows of C from the other MPI processes.

Task 5: Blocking Collectives

The first attempt at this algorithm was utilizing MPI's blocking collectives. Namely,

1. The manager process initializes matrices A and B with random double precision values and creates an empty matrix C to store the result.
2. `MPI_Scatter` is used to distribute the blockes of A, B, and C across all processes.
3. Each process computes a partial result by multiplying its block of A with the corresponding block of B and stores it in the appropriate sub-block of `blockC`.
4. The partial results are stored in the local block of matrix C, `blockC`.
5. Processes then pass their block of B to the next process using `MPI_Send` and `MPI_Recv`.
   - To avoid deadlock, even-ranked processes send before they receive and odd-ranked processes receive before they send.
   - Each process makes use of a secondary `blockB` buffer, named `tempB`, else a process may overwrite its initial segment of B before it is able to use it in a matrix multiplication. This allows one buffer to send while the other is being used in an operation.
6. After all computations complete, the processes gather their computed blocks into C using `MPI_Gather`.

The results of this approach are in Table 1 and Figure 1.

Table 1: Blocking Collectives - Average Performance (s)

p/N	1000	2000	4000	8000
1	0.320	3.315	40.153	328.265
2	0.213	1.828	21.412	168.646
4	0.138	0.852	10.187	86.116
8	0.099	0.517	4.048	43.899

Comparing MPI blocking collectives to serial performance as process count increases.

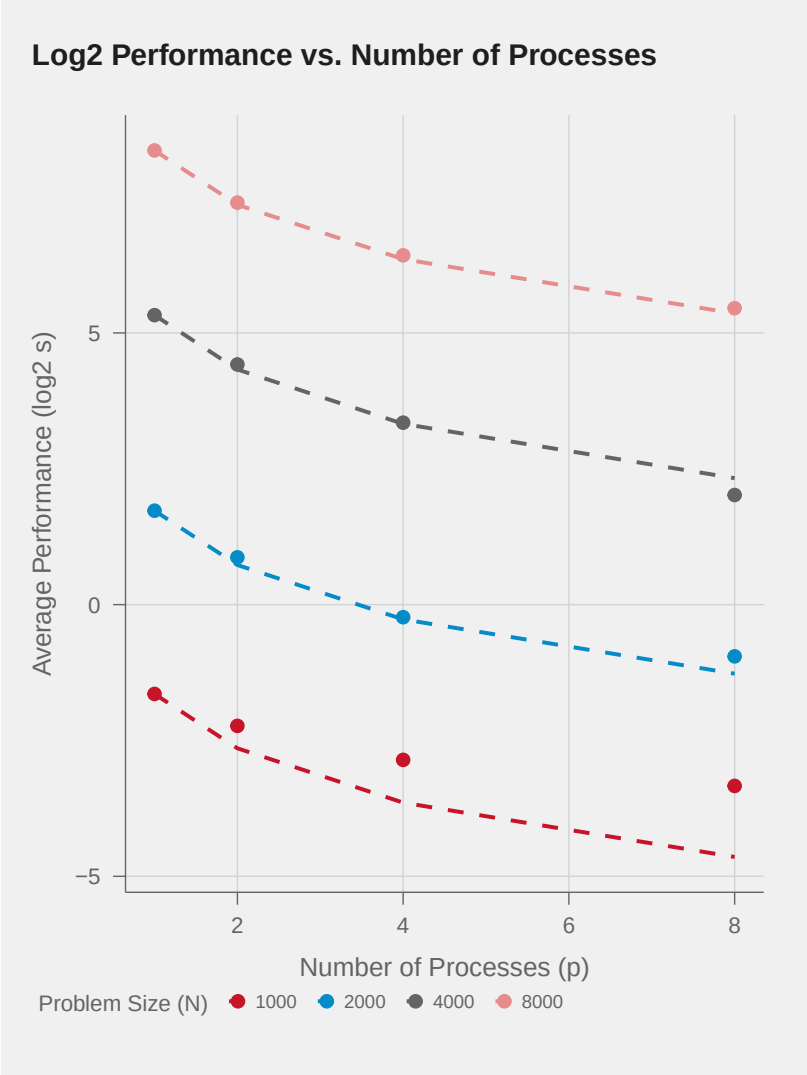


Figure 1: Blocking Communication: Log2 Performance vs. Number of Processes

The dashed line represents a doubling in performance (halving the runtime) as the number of processes doubles relative to the serial runtime. We see that for increasing N, we achieve closer to the desired throughput via scaling. For $N = 1000$, increasing the number of prcesses increases performance but at a smaller magnitude relative to larger N. It is likely that the overhead in communication here is to blame. As N increases, doubling the process count just about doubles performance. It seems as though the computational load is heavy enough that the communication overhead becomes a smaller fraction of the total runtime, allowing for better scalability. That is, the serial portion of a task (in this case, the communication overhead) becomes negligible as the problem size increases.

Task 6: Non Blocking Collectives

Next, we simply replace our blocking collectives with non blocking collectives. The algorithm and ring passing stayed entirely the same. What changes was:

- `MPI_Irecv` and `MPI_Isend` instead of `MPI_Recv` and `MPI_Send`, respectively.
- `MPI_Iscatter` and `MPI_Igather` instead of `MPI_Scatter` and `MPI_Gather`, respectively.
- Adding `MPI_Waitall` and `MPI_Wait` as necessary, e.g., to ensure the blocks have all been scattered.

The performance is seen in Table 2 and Figure 2.

Table 2: Non Blocking Collectives - Average Performance (s)

p/N	1000	2000	4000	8000
1	0.320	3.315	40.153	328.265
2	0.207	1.689	20.964	168.317
4	0.126	0.816	9.993	85.030
8	0.081	0.483	3.772	43.284

Comparing MPI non blocking collectives to serial performance as process count increases.

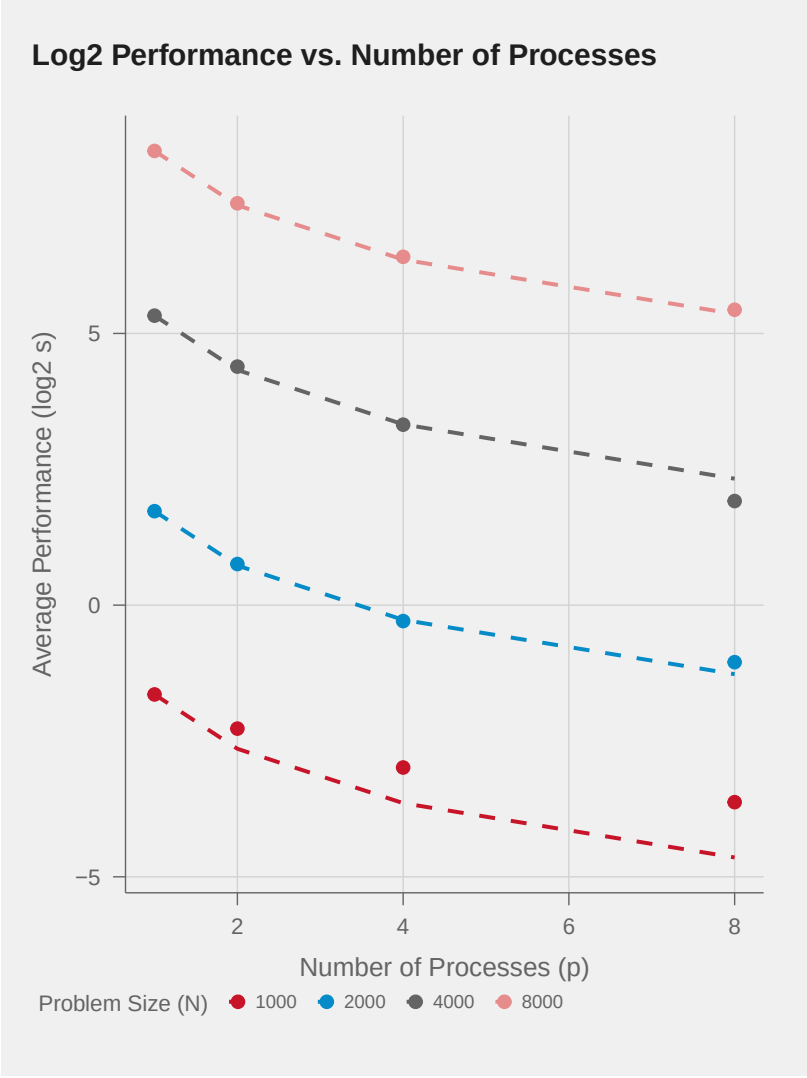


Figure 2: Non Blocking Communication: Log2 Performance vs. Number of Processes

The performance assessment here is precisely that of the blocking collectives. Since very little was changed in terms of computation and communication, we expect performance to be about similar, as observed.

Task 7: Overlapping Computation and Communication

Next, we aim to overlap communication and computation. The following items were considered:

1. Load Balancing: ideally, the work is evenly distributed among all processes. If some processes finish their work too early, they will be idle while waiting for the others to finish. This is not ideal.
2. Communication Minimization: We should communicate as little as needed for complete computation.
3. Overlap: since the `gemm` function for multiplication does not modify the `blockB` buffer, we can overlap the send + receive and computation.

The following is the updated ring-loop algorithm:

```
for (int step = 0; step < size; step++)
{
    MPI_Irecv(tempB, size_BLOCKxN, MPI_DOUBLE, prev_rank, 0, MPI_COMM_WORLD,
        &ring_pass_requests[0]);
    MPI_Isend(blockB, size_BLOCKxN, MPI_DOUBLE, next_rank, 0, MPI_COMM_WORLD,
        &ring_pass_requests[1]);

    // multiply permanant block of A on this process with current block of B on this process
    gemm(block_size, N, blockA, blockB, blockC, ((rank - step + size) % size) * block_size);

    MPI_Wait(&ring_pass_requests[0], MPI_STATUS_IGNORE);
    MPI_Wait(&ring_pass_requests[1], MPI_STATUS_IGNORE);

    double *swap = blockB;
    blockB = tempB;
    tempB = swap;
}
```

Table 3 and Figure 3 show the performance of these changes:

Table 3: Non Blocking Collectives - Average Performance (s)

np	1000	2000	4000	8000
1	0.320	3.315	40.153	328.265
2	0.206	1.647	21.035	168.214
4	0.135	0.816	9.970	85.009
8	0.076	0.481	4.034	43.118

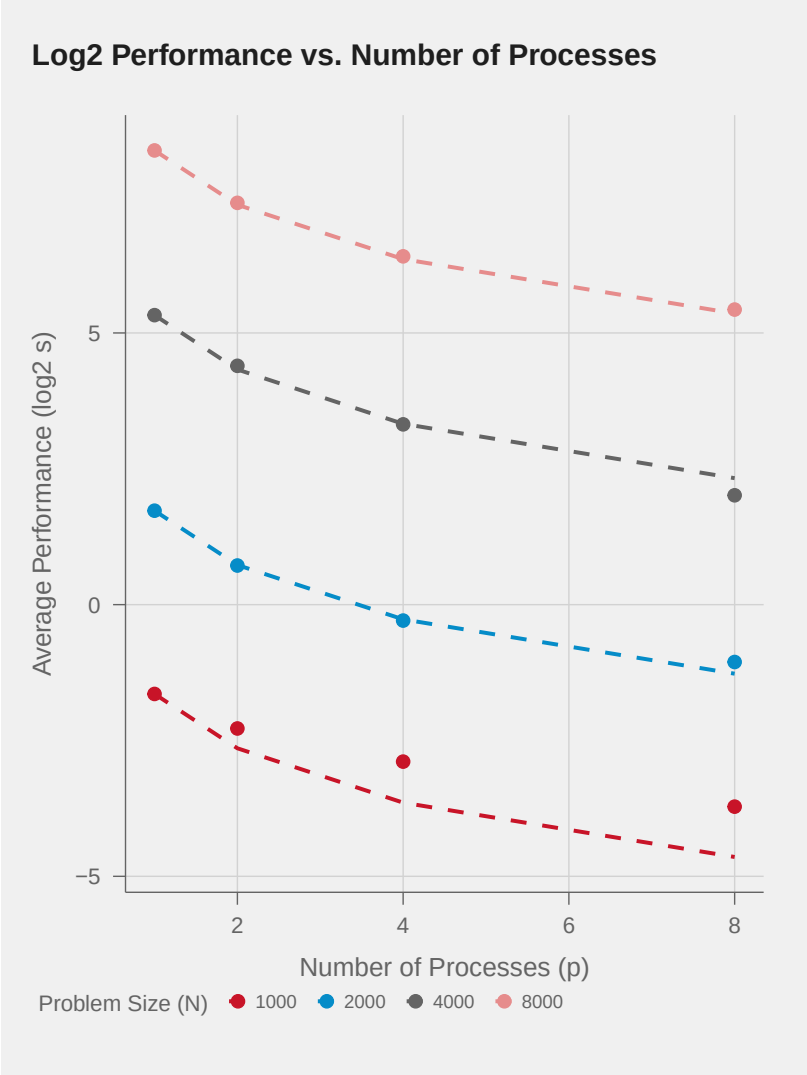


Figure 3: Non Blocking Communication Optimized: Log2 Performance vs. Number of Processes

We observe roughly the same performance as in task 6. Since average timings are reported and each has roughly the same variability, there is not enough of a difference to make a concrete statement about improvement.

Since each multiply takes the same time we are well load balanced and there is minimal idle time by any one process. Moreover, we do not overly communicate and we overlap the communication with computation as described. Since the communication is such a small amount of the total processing time, we do not expect much improvement in performance for this load balanced problem (since any optimizations would really come from the multiplication being optimized).

Task 8: Dealing with Non Uniform Buffer Sizes

Lastly, we modify the blocking collectives code to handle the case where N is not divisible by p. The following had to be considered (in continuation from task 5):

- N/p will have a remainder, i.e., $N\%p$. This remainder will be less than the total number of processes and so all ranks less than the remainder will get $\mathsf{floor}(N/p)+1$ rows (or columns) and the rest of the ranks get $\mathsf{floor}(N/p)$ rows (or columns).
  - Now, every `blockA`, `blockB`, or `blockC` has array size $(\texttt{floor}(N/p)+1)\times N$ or $\texttt{floor}(N/p)\times N$.
- For the initial assignments, `MPI_Scatterv` is used with the appropriately passed in displacements array to distribute the blocks.
  - Each process still gets a permanant allocation of `blockA` and `blockC` and starts with an initial assignment of `blockB`.
  - Each process allocates a size for `blockB` and `tempB` that is the largest possible size. In doing so, the differently sized buffers can be passed around without having to reallocate memory.
  - At each time step on a process, a multiply might be between:
    1. An $(\texttt{floor}(N/p)+1)\times N$ `blockA` and an $N\times(\texttt{floor}(N/p)+1)$ `blockB`.
    2. An $(\texttt{floor}(N/p)+1)\times N$ `blockA` and an $N\times\texttt{floor}(N/p)$ `blockB`.
    3. An $\texttt{floor}(N/p)\times N$ `blockA` and an $N\times(\texttt{floor}(N/p)+1)$ `blockB`.
    4. An $\texttt{floor}(N/p)\times N$ `blockA` and an $N\times\texttt{floor}(N/p)$ `blockB`.
- At step i in the ring swap, process p is working on the `blockB` initially allocated to rank `(rank - i + p) % p`. The size of each ranks initial allocation is kept track of to appropriately index the `blockB` array for multiplication (since it is allocated to the maximum possible size it may have data in it we do not want). The result of the multiply is placed in the correct sub-block of `blockC`.

The resulting code is an algorithm that produces the following performance and F-norm on an $N=7633$, $p=7$ program run on 4 nodes (the manager alone on one node and the other 3 nodes having two processes):[1]

Table 4: Non Uniform Size Allocation with Blocking Collectives

p	N	time	F-norm
7	7633	41.70652	1.535e-09

We see that the algorithm runs in a reasonably expected time and produces near 0 F-norm. Here is a sample of reported bindings

```
Task 8
[r918u05n01.grace.ycrc.yale.edu:1515798] MCW rank 0 bound to socket 0[core 0[hwt 0]]: [B][.]
[r918u05n03.grace.ycrc.yale.edu:636934] MCW rank 3 bound to socket 0[core 0[hwt 0]]: [B][.]
[r918u05n03.grace.ycrc.yale.edu:636934] MCW rank 4 bound to socket 1[core 1[hwt 0]]: [.][B]
[r918u05n02.grace.ycrc.yale.edu:380664] MCW rank 1 bound to socket 0[core 0[hwt 0]]: [B][.]
[r918u05n02.grace.ycrc.yale.edu:380664] MCW rank 2 bound to socket 1[core 1[hwt 0]]: [.][B]
[r918u05n04.grace.ycrc.yale.edu:1996065] MCW rank 5 bound to socket 0[core 0[hwt 0]]: [B][.]
[r918u05n04.grace.ycrc.yale.edu:1996065] MCW rank 6 bound to socket 1[core 1[hwt 0]]: [.][B]
```

6 Conclusion

In conclusion, we have demonstrated the effectiveness of OpenMPI for parallel matrix multiplication. Key insights include the significance of balancing computational load and communication overhead, particularly in larger matrix computations. The analysis revealed that while increasing the number of processors generally enhances performance, the impact varies with the problem size, indicating the necessity for a parallelization strategy that is based on the problems specific computational requirements.

[1] The distribution of the processes is done with a hostfile and was done via a back and forth consultation with ChatGPT. The resource allocation with the hostfile is NOT original work. The script is in `build-run-task8.sh`.