

Dear Reviewer,

We would like to sincerely thank you for your thorough and constructive review of our manuscript ([An Effective Communication Topology for Performance Optimization: A Case Study of the Finite Volume WAve Modeling \(FVWAM\)](#)). Your insightful comments have been invaluable in improving the quality of our work. Please find below our detailed responses to each of the comments you raised.

Sincerely,

Renbo PANG, on behalf of the co-authors

Paper Review: "An Effective Communication Topology for Performance Optimization: A Case Study of the Finite Volume WAve Modeling (FVWAM)"

This paper presents an implementation of halo-exchanges in the FVWAM model using MPI's distributed graph topology and performance comparison over baseline implementation using point-to-point communication primitives.

The paper provides a detailed comparison from tests with 512 to 32,768 processes and shows that the speedup from the distributed graph topology with and without reordered processes ranged from 1.28 to 5.63.

Reply: Thank you very much for your valuable and insightful comments!

There is some important context missing from the article that can shed light on the significance of the performance improvements.

** What is the network interconnect and topology of the target system?*

Reply: The network topology used in the tests is a three-layer fat-tree topology, with the primary data exchange network connected by InfiniBand devices. The link bandwidth is 100 Gb/s.

** Were experiments repeated with different node allocations assuming there is a batch system scheduling resources?*

Reply: The computing nodes are allocated by the Slurm job scheduling system. In all tests, we specified only the number of CPU cores, while the allocation of nodes was determined by the Slurm. Each experiment was repeated twice, and the better result was selected. This means that the

Slurm may allocate different nodes depending on resource availability at the time of execution.

** Were the different experiment types (point-point, distributed, distributed with reordering) conducted using the same node allocation for consistency?*

Reply: For the different experiment types with the same number of processes, we did not specify the same node allocation. Node allocation was determined by the Slurm. Since the nodes are shared by multiple user-submitted jobs, specifying the same node list for different experiment types would result in longer wait times, especially for large-scale tests involving up to 32,768 cores.

** Were the experiments at each processor count (e.g., 512) conducted multiple times to rule out network variability and interference from other traffic on the network?*

Reply: We repeated the experiments at each processor count twice and selected the better performance result. No significant differences were observed between the repeated tests.

** Is there any performance variability across runs?*

Reply: Except potentially different computing nodes for different experiment types on the same process count and competition for network bandwidth among different jobs submitted by multiple users mentioned by you in previous comments, I/O operations are also a performance variability. Both the distributed graph communication topology and the point-to-point communication method use the blocking communication mode. As a result, communication time includes the waiting time for receiving data from other processes that have not yet sent data to the current process. Furthermore, I/O imbalances and competition in the global file system among different jobs can lead to varying wait times.

** Can the authors elaborate differences of their approach if any with using the MPI-3 neighbourhood collectives?*

Reply: MPI-3 offers two methods for neighborhood collectives: Cartesian topology and graph topology. The approach presented in this paper is same with the graph topology in MPI-3, which supports irregular grids and user-defined neighborhood communication. In contrast, the Cartesian topology only supports regular grids, and the number of sources and destinations is fixed at $2 \times \text{ndims}$ (ndims is the number of dimensions in the Cartesian grid).

The authors allude to the following factors as the primary contributors to the improvement:

> First, as the number of processes increases, the volume of exchanged data decreases, thereby reducing the speedup ratio achieved by the distributed graph communication topology.

It appears that the application is network bandwidth bound at low processor counts. It would be very enlightening to provide details of the communication volume and interconnect specifications to confirm if that's the case.

Reply: As suggested, we calculated the minimum, average, and maximum data volume received by each process, as shown in Figure 12(a) (attached in a separate file), for a range of process counts from 512 to 32,768. The data volume is computed using Formula 1(a). V_i represents the data volume for Process i , num_recv_j denotes the number of grids received from Process j , num_fre is the frequency of the wave spectrum (set to 35 in the test), num_dir is the number of directions in the wave spectrum (set to 36 in the test), len_data is the length of one single floating point element (4 bytes), and $steps$ is the number of iteration time steps (set to 60 in the test). To simplify the representation of V_i , the unit

of V_i is expressed in megabytes (MB), calculated by dividing by 1024×1024 .

$$V_i = \text{num_recv}_j * \text{num_fre} * \text{num_dir} * \text{len_data} * \text{steps} / (1024 * 1024) \quad (\text{Formula 1})$$

As the number of processes increases, the average data volume received by each process decreases. The data volume per process ranging from 512 to 32,768 processes is shown in Figure 12(b-h) (attached in a separate file). In these figures, the x-axis represents the receiving Process IDs, the y-axis represents the sending Process IDs, and the color indicates the volume of data received by each process from others. The process ordering is determined by the METIS tool. The results show that most process IDs exchanging data are neighbors, which explains why the performance improvement is less significant after enabling the reordered option in the distributed graph communication topology.

> Second, received data are continuously searched and inserted into wave action (N) at once in the distributed graph communication topology, which can improve cache hit rates.

The presumption about improved cache hit rates can be confirmed by obtaining hardware performance counter information. I'm skeptical that cache performance played such a big role. The improvement could better be explained by MPI library implementation ordering the communication operations optimally.

The biggest weakness of this work is the limited performance data from just one platform and MPI implementation. It would highly strengthen the work if the performance optimization can be demonstrated on multiple machines with different interconnects and topologies and recent versions of community standard libraries (OpenMPI, MPICH) or recent vendor implementations. It would make the case for neighbourhood collectives for earth system workloads stronger.

Reply: Thank you for your valuable recommendation to compare different communication methods across multiple MPI libraries. Due to the expiration of our rental contract for the high-performance computing system at the National Supercomputing Center of China in Jinan, we are currently unable to conduct additional experiments at the same scale (32,678 CPU cores) with different MPI implementations. However, we conducted smaller-scale tests in the West Pacific region using both Intel MPI Library and Open MPI Library on a different platform. The results indicate that the performance of the distributed graph topology is indeed

strongly dependent on the quality of the underlying MPI library implementation.

The software and hardware environment for the first set of tests is presented in Table 1.

Tab.1 Software and hardware environment

Name	Version
CPU	Intel(R) Xeon(R) E5-2680 v4 @ 2.40GHz (28 cores per node)
Memory	128GB
Hardware Architecture	X86_64
Network	Infiniband (100Gb/s)
Operating System	Red Hat Enterprise 7.6
Compiler	Ifort 17.0.3
Compilation Options	-O3
MPI	Intel(R) MPI Library 2017.3.191
NetCDF	NetCDF-Fortran 4.5.3

The cell resolution is 6-12 km, covering the region from 95° E to 145° E and 0° N to 40° N. The number of horizontal cells is 283,517, the count of the directional spectrum is 36, and the count of the frequency spectrum is 35. The time step of iterative computation in the test was 60 seconds, and the forecasting period was one hour. Each iteration involved a single neighboring communication for a 3D variable of wave action N . The total times of neighboring communication for N during the test was 60.

We performed a series of tests on the FVWAM using different numbers of

computing processes, ranging from 8 to 512 (28 processes per node), to evaluate and compare the efficiency of the point-to-point communication method versus the distributed graph communication topology with the Intel MPI Library, as shown in Figure 10. For intra-node communication with 8 and 16 processes, the performance of both communication methods was similar. However, for inter-node communication, the distributed graph communication topology significantly outperformed the point-to-point method.

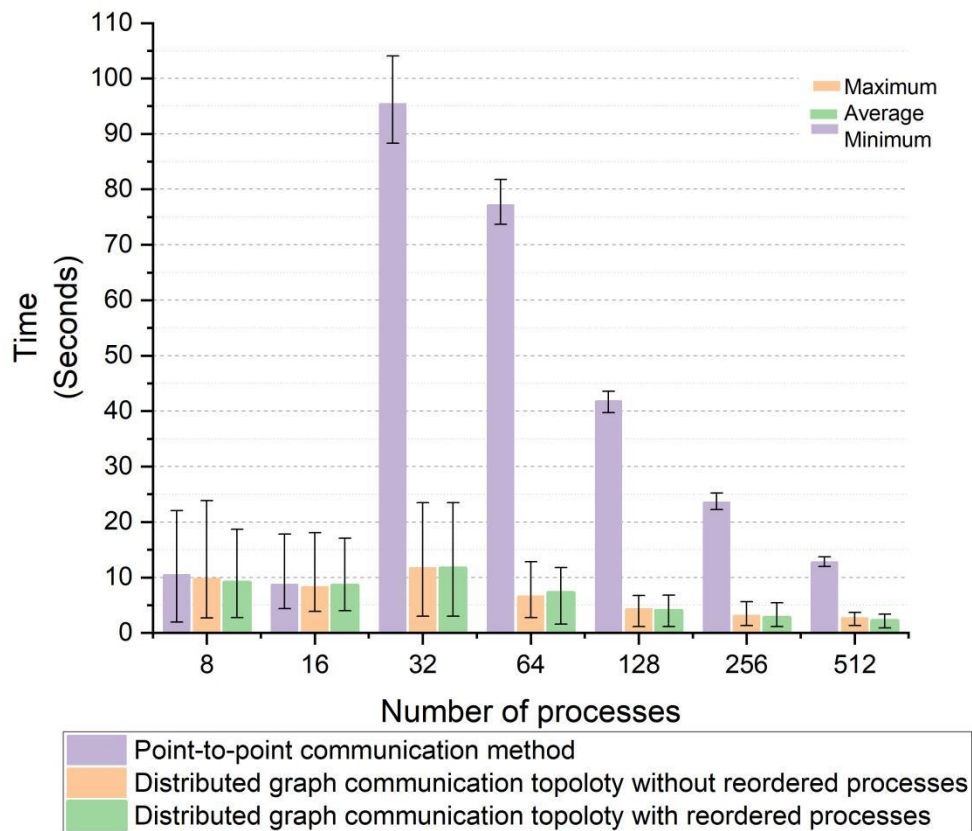


Figure 10. Time of neighborhood communication with the Intel MPI Library

The software and hardware environment for the second set of tests is presented in Table 2.

Tab.2 Software and hardware environment

Name	Version
CPU	Intel(R) Xeon(R) E5-2680 v4 @ 2.40GHz (28 cores per node)
Memory	128GB
Hardware Architecture	X86_64
Network	Infiniband (100Gb/s)
Operating System	Red Hat Enterprise 7.6
Compiler	GNU Fortran 10.2.0
Compilation Options	-O3
MPI	Open MPI 4.0.5
NetCDF	NetCDF-Fortran 4.5.3

The model configuration in this test is the same as the first test. The results of the FVWAM using different numbers of computing processes, ranging from 8 to 512 (28 processes per node), are shown in Figure 11 to evaluate and compare the efficiency of the point-to-point communication method versus the distributed graph communication topology with the Open MPI Library. The performance gap between the two methods was smaller, and there was no noticeable performance improvement in intra-node communication (with 8 or 16 processes) when using the Open MPI Library, compared to the Intel MPI Library.

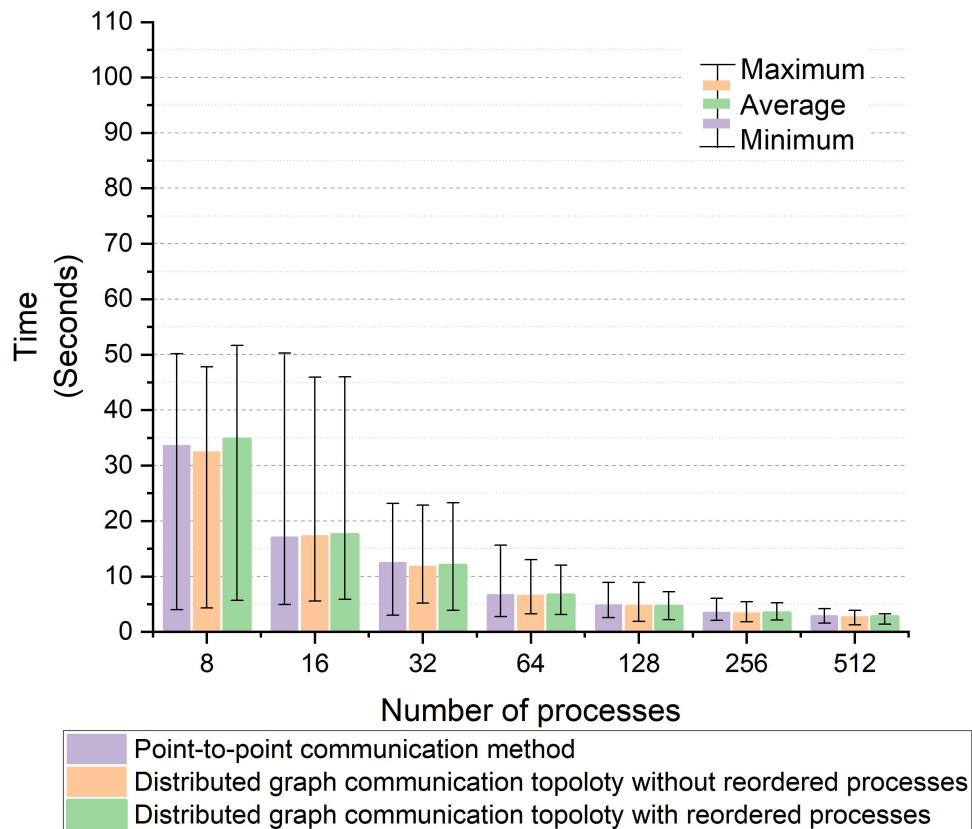


Figure 11. Time of neighborhood communication with the OpenMPI Library

There is work illustrating performance improvements from reordering MPI processes taking network topology into account. e.g., <https://dl.acm.org/doi/10.1145/2851553.2851575>

What is the current reordering strategy in case I missed? Did the authors consider any advanced reordering strategies?

Reply: Thank you very much for providing the reference! In both the

point-to-point communication method and the distributed graph communication topology without reordered processes, the process order is same. It is based on the output from the METIS partitioning tool, which is widely used in various models, including MPAS, WAVE WATCH III (WW3), and the Finite Volume Coastal Ocean Model (FVCOM). As shown in Figure 12(a-h), the results of the METIS tool are effective, as it places the majority of communicating processes as neighbors. The reordering strategy used in the distributed graph communication topology with reordered processes depends on the implementation details of the MPI library, which remains a black-box to users.

The paper refers to pre-posting receives using MPI_Irecv. However, they mention

> An alternative is to call the non-blocking communication interface MPI_Isend for sending data, but it is infrequently utilized due to the increased complexity that introduces to the sending operation.

It's not inherently that complex as a lot of applications use non-blocking sends effectively. I wonder what the performance impact would be if the authors used non-blocking operations.

Reply: In our tests, we use MPI_Isend to send data and MPI_Recv to receive data in the point-to-point communication method. The distributed graph communication topology, which is implemented using the MPI_Neighbor_alltoallv interface, also employs a blocking communication method. The statement "but it is infrequently utilized due to the increased complexity it introduces to the sending operation" is inaccurate, and it has been removed.

The section titled "Section 3.2 Point-to-point communication method" (Lines 199-214) has been revised and replaced with the following text and graph to better introduce the point-to-point communication method used in the test case.

The approach for implementing the point-to-point communication method for neighboring communication is illustrated in Figure 5. The process for determining ordered arrays of receiving grid IDs, receiving process IDs, sending grid IDs, and sending process IDs follows the same steps (1-4) as the distributed graph communication topology described in Figure 4.

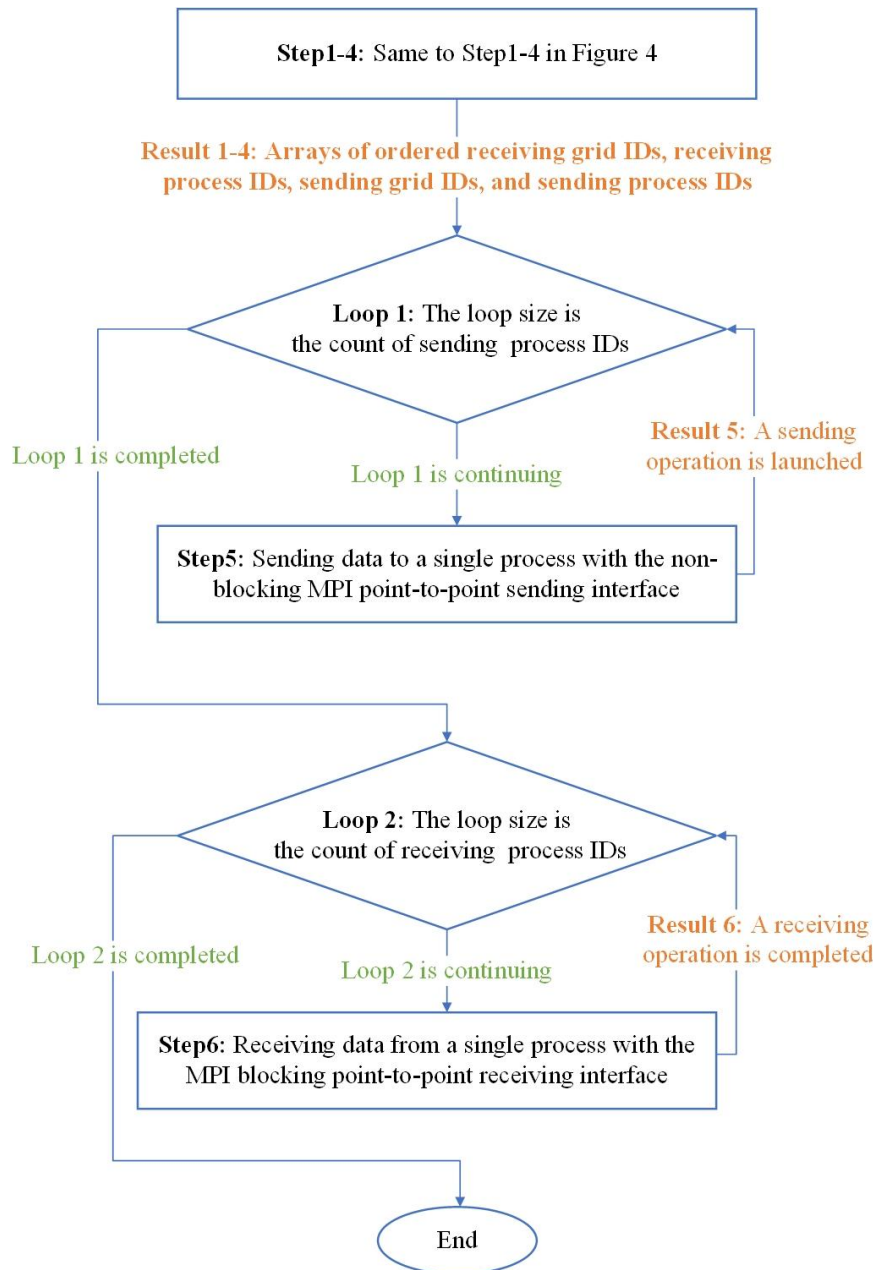


Figure 5. Implement data exchange with the point-to-point communication method

To prevent communication deadlocks, FVWAM initiates non-blocking sending operations before starting receiving operations in Step 5. Each process performs sending operations to transmit data to the corresponding receiving processes. These sending operations are executed by repeatedly

calling the MPI_Isend interface (*sendbuf*, *destination*, ...). The parameter *sendbuf* refers to the data buffer associated with the sending grid IDs from a single process, and the *destination* parameter corresponds to the receiving process ID. The number of calling the MPI_Isend interface depends on the number of sending process IDs for each process. Since MPI_Isend is non-blocking, it returns immediately without waiting for the completion of the send operation.

In Step 6, each process calls the MPI_Recv interface (*recvbuf*, *source*, ...) to receive data from the sending processes. The *recvbuf* parameter is used to store data corresponding to the receiving grid IDs, and the *source* parameter indicates the sending process ID. MPI_Recv is a blocking communication interface that only completes once the receiving operation has finished.

I was looking forward to the article to hear about novel techniques that could improve communication performance at scale. However, it was slightly disappointing to see that the benefit from the proposed optimization dramatically tapers off as we go from low (512) to high (32756) number of processes.

On heterogeneous GPU based supercomputers like the Frontier Exascale system, the number of nodes is relatively low (9,408) due to the fat node

architecture compared to CPU based supercomputers like Fugaku (158,976 nodes) out there. In this overall context, the benefit of a communication optimization is more relevant at scale when there are potentially hundreds of thousands of MPI endpoints at scale (e.g., 600k on Fugaku with 4 MPI ranks per node mapping optimally to the NUMA domains there).

Ref: https://docs.olcf.ornl.gov/systems/frontier_user_guide.html

<https://www.fujitsu.com/global/about/innovation/fugaku/specifications/>

Pg 15, lines 295-302:

To conclude, I understand the motivation of authors to improve their production simulations performance and the relative significance for their workload. Additional performance data would be highly informative and make this more generally applicable.

Reply: Thank you once again for your insightful and constructive comments! As mentioned in our previous response, we conducted additional tests, and the communication times for the three methods are presented in Figures 10 and 11.

Using the maximum and average communication times for the point-to-point communication method with the Intel MPI Library (Figure

10), we computed the speedup ratio for the distributed graph communication topology, as shown in Figure 13. We observed a significant performance gap between intra-node and inter-node communication for the point-to-point communication method. This resulted in similar speedup ratios for both communication methods in intra-node communication with 8 and 16 processes. However, starting from 32 processes, the speedup ratio increases as inter-node communication was introduced.

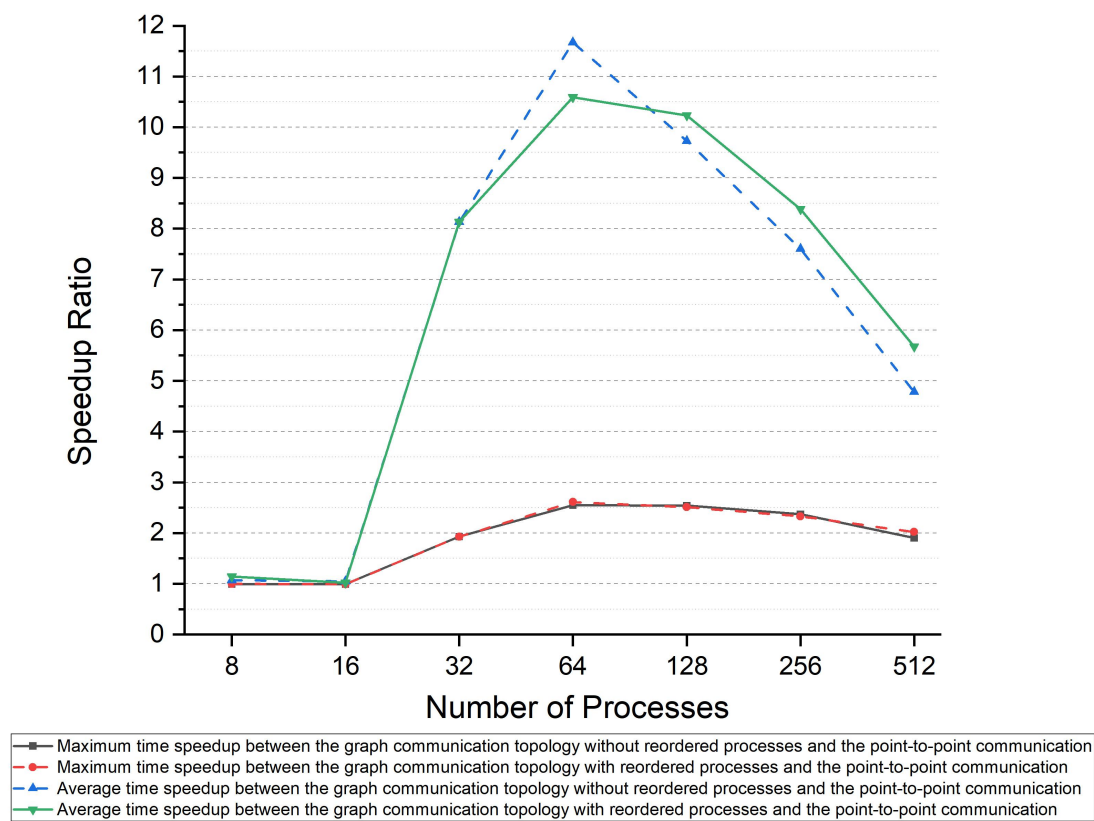


Figure 13. Speedup ratio of neighborhood communication with the Intel MPI Library

using the maximum and average communication times for the point-to-point method with the OpenMPI Library in Figure 11, we calculated the speedup ratio for the distributed graph communication topology, which is presented in Figure 14. The results show that the performance of both communication methods is comparable, indicating that the performance gap between the two methods depends on the MPI implementation library used.

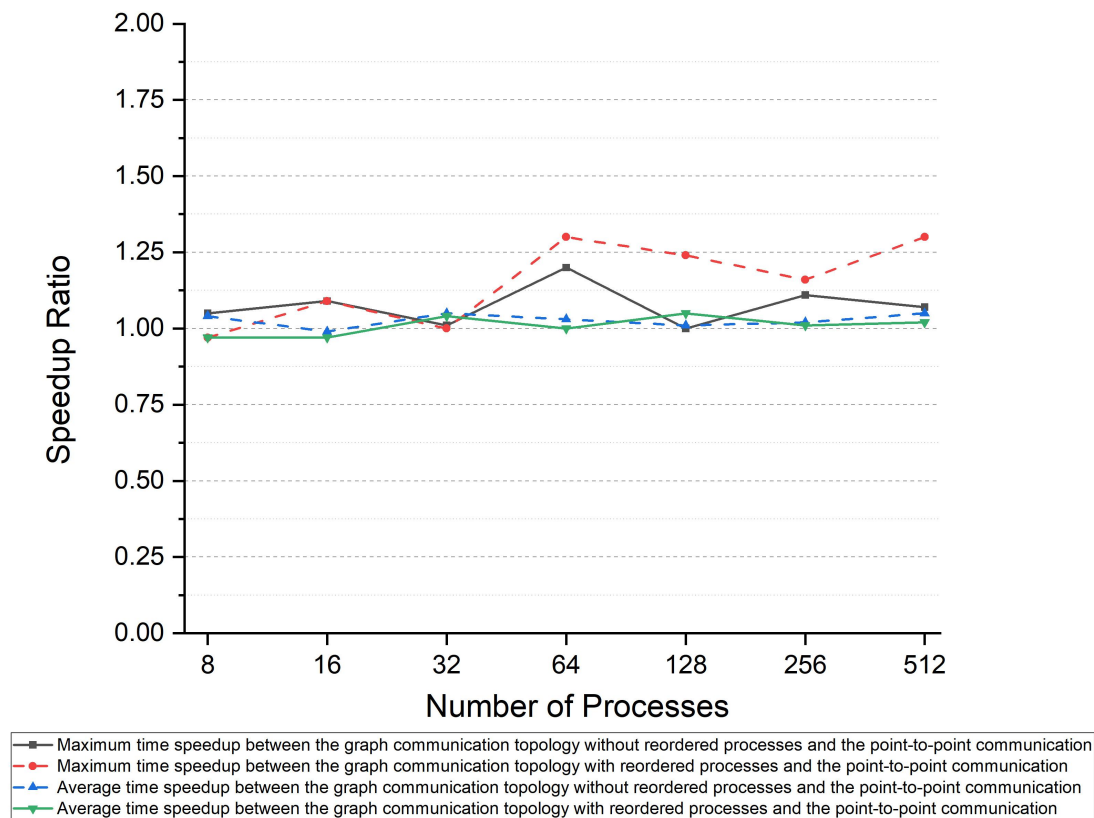


Figure 14. Speedup ratio of neighborhood communication with the OpenMPI Library

Minor comments:

Pg 2, line 29: There are better references than Sukhija et al., 2022 for the Frontier Exascale supercomputer. I suggest using one of the papers from the Supercomputing conference.

<https://dl.acm.org/doi/abs/10.1145/3581784.3607089>

> Sukhija, N., Bautista, E., Butz, D., and Whitney, C.: Towards anomaly detection for monitoring power consumption in HPC facilities, in: 380 Proceedings of the 14th International Conference on Management of Digital EcoSystems, pp. 1–8, 2022

Reply: As suggested, we have replaced the reference to (Sukhija et al., 2022) with (Atchley et al., 2023)

Atchley, S., Zimmer, C., Lange, J., Bernholdt, D., Melesse Vergara, V., Beck, T., Brim, M., Budiardja, R., Chandrasekaran, S., Eisenbach, M., et al.: Frontier: exploring exascale, in: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 1–16, 2023.

The performance sections in the paper are a bit verbose and redundant pointing to the information in the figures. It might be better to be succinct

in highlighting the results and elaborate further on the reasons behind the improvement.

Reply: We have revised this section to make it more concise, focusing on directly highlighting the key results rather than reiterating the details already presented in the figures. We have also expanded on the underlying reasons behind the observed improvements to provide a clearer understanding of the factors contributing to the performance gains. These revisions will be included in the revised manuscript submission.