MPI is Good, Control is Better: Checking Performance Guidelines of Collectives

Sascha Hunold **and Maximilian Hagn**

Research Group for Parallel Computing, Faculty of Informatics, TU Wien, Austria

MPI Collective Operations, such as `MPI_Bcast` and `MPI_Allreduce`, are important building blocks for many HPC applications. It is important to point out that only the semantics of each collective operation are defined in the MPI standard. Therefore, each MPI library may use different algorithms to implement MPI collectives. Additionally, some algorithms only work efficiently on a subset of communication problems, e.g., they were designed for communicating small messages, but they are rather inefficient for large messages.

Self-consistent Performance Guidelines can be formulated for each collective call [1]. These guidelines define naturally arising performance expectations for specialized collectives, e.g.,

$$\texttt{MPI\_Allreduce}(n) \preceq \texttt{MPI\_Reduce}(n) + \texttt{MPI\_Bcast}(n) \text{ or}$$
$$\texttt{MPI\_Scatter}(n) \preceq \texttt{MPI\_Bcast}(n).$$

A former guideline states that an Allreduce call with a payload of n data items should not be slower than the subsequent calls to Reduce and Bcast, which semantically perform the same operation. Another example is shown in Fig. 1, where the processes mimic the Scatter functionality by (1) first calling Bcast and then (2) copying the relevant pieces to form the correct result buffer.

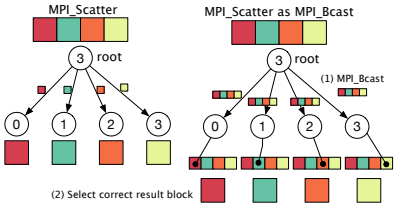

Fig. 1: MPI_Scatter as MPI_Bcast guideline.

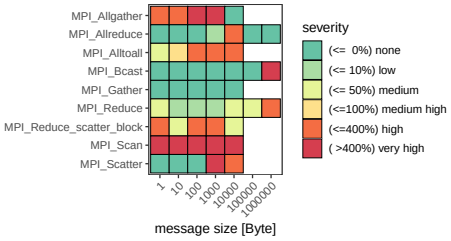

Fig. 2: `pgchecker` results for `16x128` processes on VSC-5, Open MPI 4.1.4. The severity denotes the slowdown of default MPI (in percent) compared to the best guideline implementation.

Goal and Approach: We examine the performance-consistency of modern MPI libraries with respect to defined performance guidelines [2] on current supercomputers, e.g., VSC-5 (Austria). We have implemented a tool, called `pgchecker`, which analyzes the performance consistency of MPI libraries. It is built on top of the ReproMPI benchmark [3] and implements several statistical tests (e.g., Mann-Whitney), which allows for an automatic and non-parametric testing of MPI libraries' performance-consistency.

Results: Fig. 2 shows `pgchecker` results for the VSC-5 and Open MPI, where the individual colors denote the severity of performance-guideline violations. Orange and red squares represent serious performance deficits. We observe several severe performance issues with Open MPI on VSC-5, e.g., with a payload of 1 MB, the default Bcast is more than 4 times slower than one of the mock-up implementations.

Acknowledgement: This work was partially supported by the Austrian Science Fund (FWF) project P33884-N.

References

[1] Jesper Larsson Träff, William D. Gropp, Rajeev Thakur: Self-Consistent MPI Performance Guidelines. *IEEE Trans. Parallel Distributed Syst.*, 21(5): 698–709, 2010.

[2] Sascha Hunold and Alexandra Carpen-Amarie: Autotuning MPI collectives using performance guidelines. In *Proceedings of the HPC Asia 2018*, pages 64–74. ACM, 2018.

[3] Sascha Hunold and Alexandra Carpen-Amarie: Reproducible MPI benchmarking is still not as easy as you think. *IEEE Trans. Parallel Distributed Syst.*, 27(12):3617–3630, 2016.