Paper title

General Failure Detection and Propagation in HPC and Distributed Systems

Statement

We present the design and implementation of an efficient runtime-level failure detection and propagation strategy targeting large-scale, dynamic systems that can detect both node and process failures. Multiple overlapping topologies are used to optimize the detection and propagation, minimizing the incurred overheads and guaranteeing the scalability of the entire framework.

This paper is an extended version of the conference proceedings [1] that considers not only the case of synthetic communication benchmarks, but also application benchmarks. We use the heavily communication-bound benchmark Graph500 which is an open specification effort to offer a standardized graph-based benchmark across large-scale distributed platforms which captures the behavior of common communication-bound graph algorithms. We conducted our experiments at a larger scale. At the same time, we expanded the study to consider the case of applications using parallel global address space (PGAS) programming models, and study and compare the overhead incurred by enabling failure detection in both MPI and OpenSHMEM applications, using RDaemon$^{\#}$ infrastructure as a shared backend.

Section 6, and figures 16, 17, 18, 19, 20 are additions that present the new application results with both OpenSHMEM and MPI on the Cori supercomputer. We have updated the abstract, introduction, conclusion and title of the paper to reflect the supplementary content.

References

[1] D. Zhong, A. Bouteiller, X. Luo, G. Bosilca, Runtime level failure detection and propagation in hpc systems, in: Proceedings of the 26th European MPI Users' Group Meeting, EuroMPI '19, ACM, New York, NY, USA, 2019, pp. 14:1–14:11. doi:10.1145/3343211.3343225.