Bulletin of Networking, Computing, Systems, and Software – www.bncss.org, ISSN 2186-5140
Volume 13, Number 1, page 45, January 2024

Multi-Port All-to-All Communications in Two-Level Fully Connected Network Topology

Kien Trung Pham
National Institute of Informatics
SOKENDAI Tokyo, Japan
ptkien@nii.ac.jp

Truong Thao Nguyen
National Institute of Advanced Industrial Science and Technology (AIST)
Tokyo, Japan
nguyen.truong@aist.go.jp

Michihiro Koibuchi
National Institute of Informatics
Tokyo, Japan
koibuchi@nii.ac.jp

I. Introduction

Network topology and communication algorithms are two components that directly affect the performance of applications on supercomputers. In this work, we use two-level fully connected (2lvfc) as the scalable network topology that works for the next generation of supercomputers. 2lvfc is a regular 2-dimension topology. 2lvfc is isomorphic with 2D HyperX or diameter-two Dragonfly networks. We also propose three algorithms to execute all-to-all collective operation on 2lvfc.

II. Dimensional Order Algorithm

This is the simplest algorithm based on the nature of 2lvfc. The topology has two dimensions. We will do the all-to-all collective operation in two steps. First, we will do the all-to-all operation inside one dimension. The data for the second step is also prepared in the first step. Then, we exchange the data in the remaining dimension to finish the algorithm.

The Dimension Order (DO) algorithm has low latency because it finishes in two steps but it makes link utilization low. This is because DO algorithm sequentially performs two different sets of links. The sequential nature makes DO have low link utilization.

III. MULTITREE algorithm

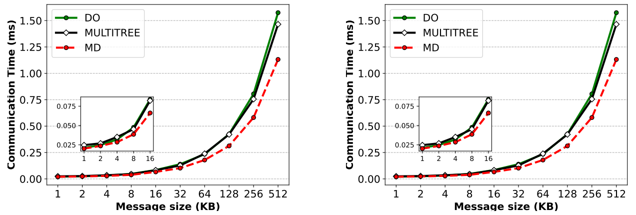
MULTITREE algorithm [1] is originally designed for all-reduce collective operation. It takes a topology-agnostic approach. It can also be applied for all-to-all operation. The strength of MULTITREE algorithm is that it provides the schedule for high network resource utilization. MULTITREE algorithm builds N spanning trees for a N-node topology. The schedule of MULTITREE makes it fully utilizes the resource of 2lvfc. However, it suffers high communication steps which are not good for small size messages.

IV. Multi-Dimension Algorithms

Multi-Dimension (MD) algorithm is developed to reduce the high communication step of MULTITREE algorithm while 100% uses the network resource of 2lvfc. MD algorithm is based on the following idea. DO algorithm uses two different sets of links sequentially. We can divide the data into two parts and send it in parallel in two reverse directions. In this way, we can reduce the high communication step of MULTITREE algorithm while solving the low link utilization of DO algorithm.

V. Result

The performance of the three algorithms is estimated in discrete event simulator — SimGrid 3.28. The start-up latency for each communication is set to 1μs. Bandwidth is set to 100 Gbps as current top supercomputers. The simulator is run on a host system with SMP Debian 3.16.64-2 operating system on Intel(R) Xeon(R) CPU E5-2667 v4 and 794 GB RAM.



(a) 8×8 2lvfc small data size (b) 8×8 2lvfc big data size

Fig. 1: Communication time of DO, MULTITREE and MD algorithm on 8×8 2lvfc.

Overall, MD algorithm is better than DO and MULTITREE algorithm because it has small communication step and efficiently uses the resources. In small message sizes, the communication time of MD is approximate to DO algorithm because they have the same communication step. In big message sizes, MD outperforms MULTITREE and DO algorithm.

VI. Conclusion

This study exploited efficient all-to-all collective algorithms using parallel non-blocking point-to-point communications in two-level fully connected network topologies, which can be considered as a basic block of cutting-edge high-radix network topologies. We firstly proposed an optimization of MULTITREE algorithm to two-level fully connected topologies. We secondly proposed their custom collective algorithms.

References

[1] J. Huang, P. Majumder, S. Kim, A. Muzahid, K. H. Yum, and E. J. Kim, "Communication algorithm-architecture co-design for distributed deep learning," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2021, pp. 181–194.