GLEX-Alltoall: Multi-leader All-to-all Communication on Multi-core Supercomputer

Jintao Peng[1], Jie Liu[2,*], Min Xie[3,*]

Changsha, China

Abstract

All-to-all communication is commonly used in parallel applications like FFT. In mordern supercomputers, there are multiple cores, NUMAs and network endpoints. These features bring much parallelism. However, there is no method which makes uses the parallelism to improve the all-to-all communication. In this paper, we introduce an optimized multi-leader all-to-all library which explore the parallelism on network, CPU cores and overlap the intra- and inter-node communication. The results show that, compared to MPI, our library achieves up to 20x speedup and 4x speedup on average. For application, our method achieves up to 1.75x speedup on peak performance for 16384 cores.

Keywords: Collective Communication, Multi-core processor, MPI all-to-all, RDMA, Shared Heap
2010 MSC: 00-01, 99-00

1. Introduction

Many parallel applications may suffer from global communication. Especially for communication-intensive applications, their time-to-solution and scalablily may be affected by global communication. Message Passing Interface (MPI) provides a set of commonly used collective communication. MPI_Alltoall is one of the collective communication where each process will send a different message to all processes. It is broadly used in some parallel applications like Fast Fourier Transform (FFT) and graph computing. However, each time we double the processes, the all-to-all communication workload is quadrupled. On mordern supercomputers, network throughput has a linear relationship with the number of nodes. This brings great challenges to large-scale all-to-all communications.

For multiple-core processes, a effective way is node-aware all-to-all method. It's replace a N nodes global all-to-all into N-1 times intra-node gather + inter-node all-to-all + N-1 times intra-node gather. This method is very effective for small messages. Because, compared to original method, a node-aware all-to-all reduce the number of messages by M^N times (M is number of processers in each nodes). The size of the message is increased by M^N times, which makes effective use of the network bandwidth. In the current supercomputer, a node has multiple CPU cores, NUMA and network endpoints. This architecture brings 4 kinds of parallism to optimize a node-aware all-to-all method.

(1) Group the authors per affiliation.
(2) Use footnotes to indicate the affiliations.

*Corresponding author
[1]JintaoPengCS@gmail.com
[2]liujie@nudt.edu.cn
[3]xiemin@nudt.edu.cn

2. Front matter

The author names and affiliations could be formatted in two ways:

(1) Group the authors per affiliation.
(2) Use footnotes to indicate the affiliations.

See the front matter of this document for examples. You are recommended to conform your choice to the journal you are submitting to.

3. Bibliography styles

There are various bibliography styles available. You can select the style of your choice in the preamble of this document. These styles are Elsevier styles based on standard styles like Harvard and Vancouver. Please use BibTEX to generate your bibliography and include DOIs whenever available.

Here are two sample references: [1, 2].

References

[1] R. Feynman, F. Vernon Jr., The theory of a general quantum system interacting with a linear dissipative system, Annals of Physics 24 (1963) 118–173. doi:10.1016/0003-4916(63)90068-X.

[2] P. Dirac, The lorentz transformation and absolute time, Physica 19 (1-12) (1953) 888–896. doi:10.1016/S0031-8914(53)80099-6.