

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

Jintao Peng<sup>1</sup>, Jie Liu<sup>2,\*</sup>, Min Xie<sup>3</sup>

Changsha, China

## Abstract

All-to-all communication is commonly used in parallel applications like FFT. In modern supercomputers, there are multiple cores, NUMAs and network endpoints. These features bring much parallelism. However, there is no method which makes use of the parallelism to improve the all-to-all communication. In this paper, we introduce an optimized multi-leader all-to-all library which explores the parallelism on network, CPU cores and overlaps the intra- and inter-node communication. The results show that, compared to MPI, our library achieves up to 20x speedup. 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 scalability 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) [1] and graph computing. However, each time we double the processes, the all-to-all communication workload is quadrupled. On modern 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, an effective way is node-aware all-to-all method. It replaces 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 reduces the number of messages by  $M^N$  times ( $M$  is number of processors in each node). 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 parallelism to optimize a node-aware all-to-all method:

- (1) Multiple network endpoints can simultaneously process multiple communication requests.

- (2) Processes in different NUMA can simultaneously access its local memory without contention.
- (3) Multiple processes can simultaneously gather/scatter data and compose communication requests.
- (4) Inter-node communication can be overlapped with intra-node communication.

As we know, no methods combine these parallelism together to improve a node-aware all-to-all collective communication.

In this paper, we proposed a multi-leader node-aware all-to-all method. It uses multiple leaders on different NUMA which open the different network endpoints to gather/scatter data, compose communication requests, and transpose local matrix. It explores the parallelism existing in modern multi-core processor with NUMA memory architecture and multi-port network. For intra-node gather/scatter, we proposed a shared-heap-based remote accessible memory which is similar to intra-node MPI RMA. Inter-node communication is based on Remote Direct Memory Access (RDMA) which provides high throughput and low latency. The results show that, compared to MPI.alltoall, our implementation achieves up to 20x speedup and 4x speedup on average.

## 2. Related Work

## 3. 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.

\*Corresponding author

<sup>1</sup>JintaoPengCS@gmail.com

<sup>2</sup>liujie@nudt.edu.cn

<sup>3</sup>xiemin@nudt.edu.cn

#### 4. Bibliography styles

60 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 BibT<sub>E</sub>X to generate your bibliography and include DOIs whenever available.

65 Here are two sample references: [2, 3].

#### References

#### References

- [1] C. Mehta, A. Karthi, V. Jetly, B. Chaudhury, Parallel fast multipole method accelerated fft on hpc clusters, *Parallel Computing* (2021) 102783.
- 70 [2] 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.
- [3] P. Dirac, The lorentz transformation and absolute time, *Physica* 19 (1–12) (1953) 888–896. doi:10.1016/S0031-8914(53)80099-6.