OpTree: An Efficient Algorithm for All-gather Operation in Optical Interconnect Systems

Fei Dai*, Yawen Chen, Zhiyi Huang, Haibo Zhang
Department of Computer Science, University of Otago, Dunedin, New Zealand
Email:daitr616@student.otago.ac.nz, {yawen.chen, zhiyi.huang, haibo.zhang}@otago.ac.nz

***Abstract*—All-gather collective communication is one of the most important communication primitives in parallel and distributed computation, which plays an essential role in many high performance computing (HPC) applications such as distributed Deep Learning (DL) with model and hybrid parallelisms. To solve the communication bottleneck of All-gather, optical interconnection network can provide unprecedented high bandwidth and reliability for data transfer among the distributed nodes. However, most traditional All-gather algorithms are designed for electrical interconnection, which cannot fit well for optical interconnect systems, resulting in poor performance. This paper proposes an efficient scheme, called OpTree, for All-gather operation on optical interconnect systems. OpTree derives an optimal m-ary tree corresponding to the optimal number of communication stages, which achieves the minimum communication time. We further analyze and compare the communication steps of OpTree with existing All-gather algorithms. Theoretical results exhibit that OpTree requires much less number of communication steps than existing All-gather algorithms on optical interconnect systems. Simulation results show that OpTree can reduce communication time by 72.21%, 94.30%, and 88.58% compared to three existing All-gather schemes WRHT, Ring, and NE, respectively.**

***Index Terms*—Optical interconnects, All-gather, communication, WDM**

I. INTRODUCTION

As the number of GPUs or other accelerators integrated into systems increases, efficient communication among these devices becomes crucial for running HPC applications. All-to-all communication methods, such as the message passing interface (MPI) All-gather operation, are widely used in many legacy scientific applications to perform Fast Fourier Transforms (FFTs) when data is distributed among multiple processes [1]. The All-gather operation is also gaining attention for performing model or hybrid parallelisms in the training of distributed Deep Neural Networks (DNNs) on GPU clusters [2]. Given the current software and hardware development for machine learning, optimizing All-gather communication in dense-GPU systems is of great importance, both in current systems and in next-generation systems.

Although many algorithms have been proposed to improve the performance of the All-gather operation in electrical interconnect systems, communication bottlenecks still exist in these systems [3]. For instance, in the training of distributed model-parallel DNNs, frequent communications with large data transfers between distributed nodes are required, which can lead to communication bottlenecks in electrical interconnect systems. However, with the advancement of silicon photonics, emerging optical networks can offer increased concurrent communication capacities through Wavelength Division Multiplexing (WDM), promising data transmission rates that are several orders of magnitude higher than current electrical networks. By using optical networks, communication bottlenecks can be avoided for many time-sensitive applications in data centers, such as scientific visualization, high-speed simulation of climate change, parallel and distributed training of deep learning, etc.

Most All-gather algorithms are designed for electrical interconnect systems using the packet-switching mechanism and have to share bandwidth in the electrical wires. In contrast, optical transmission is usually based on circuit switching, where each optical communication path monopolizes the bandwidth by occupying one wavelength for transmission. As a result, traditional All-gather algorithms do not work well for optical interconnect systems as they do not take into account the unique features of optical communication. While some research has been done on all-to-all broadcast in optical networks, the focus has primarily been on fault tolerance [4] or reducing the number of wavelengths used [5], rather than reducing communication time.

Inspired by the uniform routing and multi-stage model [5] in optical networks, we propose an efficient scheme, OpTree, for All-gather operation on optical ring interconnect systems, aiming at minimizing the communication time. We use the ring topology because it is more feasible to be deployed and existing prototypes such as TeraRack [3] have already implemented it. To the best of our knowledge, OpTree is the first scheme to optimize the communication time for All-gather operation in optical interconnect systems. Our main contributions are summarized as follows:

- We propose an efficient scheme OpTree based on m-ary tree structure for All-gather operation in optical interconnect networks. OpTree derives the optimal m-ary tree corresponding to the optimal number of communication stages and achieves the minimum communication time among all options of the m-ary tree.
- We analyze and compare the communication steps of OpTree with existing All-gather algorithms. Theoretical results show that OpTree requires a much smaller number

978-1-5386-7462-8/23/$31.00 ©2023 IEEE

of communication steps than existing All-gather algorithms on optical interconnect systems.

- We evaluate OpTree with extensive simulations. Compared to three existing All-gather algorithms WRHT [6], Ring [7], and NE [8], OpTree can reduce communication time by 56.36%, 92.76%, and 85.54% on average in optical interconnect systems with different number of nodes. OpTree can also reduce communication time by 88.06%, 95.84%, and 91.69%, respectively, in a 1024-node optical interconnect system using different wavelengths.

The rest of this paper is organized as follows. Section II presents the related works. Section III illustrates the optical interconnect architecture, motivation, the design of OpTree, the analysis of communication steps and communication time. Section IV evaluates OpTree under different scenarios. Section V provides this paper's limitations and further discussions. Finally, Section VI concludes the article.

II. Related Work

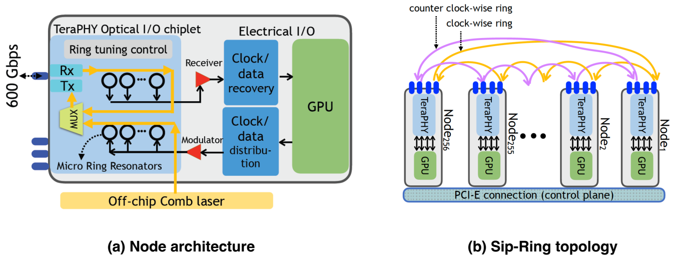
Since the standardization of MPI, All-gather operation in electrical interconnect systems has been extensively studied by researchers and engineers. Various All-gather algorithms have been proposed to optimize the performance in different situations. To reduce the bandwidth bottleneck of the electrical link, hierarchical algorithms [9]–[11], multi-leader approach [12], and multi-lane communication [13] were proposed. To minimize the number of links traversed during inter-node communication, a topology-aware collective algorithm [14] was proposed. A general optimization method that can be applied in different architectures was also proposed in [15]. However, they are not suitable for optical interconnect systems.

Collective communication in optical networks was first proposed in [16]. The authors discussed the bounds on the number of wavelengths and communication steps needed for broadcast and gossiping based on one-stage and multi-stage models. Many studies focused on specific topologies for one-stage collective communication to reduce the number of wavelengths, such as ring, torus, mesh, hypercubes, and trees [17]–[21]. In [22], the authors proposed a one-stage broadcasting model in WDM networks considering the tap-and-continue feature of optical nodes. Many follow-up investigations were carried out based on multi-stage models in different topologies. In [23], the authors studied the uniform all-to-all routing problem using the multi-stage model in a symmetric directed ring to reduce the number of wavelengths. Furthermore, many researches were conducted based on multi-stage model in different topologies. In [24], the authors addressed the collective communication problem in multi-stage WDM networks for some special topologies: lines, ring, 2-D square tori, and 3-D square tori. In [5], the authors proposed a general multi-stage routing and wavelength assignment to reduce the number of wavelengths in the optical network. Our work differs from these works as we aim to optimize the time for All-gather operation in optical interconnect systems.

III. The OpTree Scheme

A. Optical Interconnect Architecture

We assume that the OpTree system uses the microring resonator (MRR)-based optical switch known as TeraRack [25], although it is also compatible with other similar optical interconnect systems. As shown in Fig. 1 (a), a TeraRack node consists of components based on TeraPHY silicon photonics technology. Although other accelerators can be used in the TeraRack node, we assume that homogeneous GPUs are used as the computing devices. Each of the four optical interfaces on a TeraRack node has 64 micro-ring resonators that can select and forward any combination of 64 wavelengths. The laser source is a comb laser called SuperNova Light Supply [25]. On the transmission side (Tx), an off-chip comb laser generates light that is directed into the node via a fiber coupler and modulates the accelerator's transmission data at 40 Gbps per wavelength using an array of MRRs. On the receiving side (Rx), a second array of MRRs selects the wavelengths intended for the accelerator and passes through the remaining wavelengths. As shown in Fig. 1 (b), the nodes interconnect in a ring topology, with two rings for clockwise and counterclockwise communication. In the data plane, traffic is transmitted across four single-mode fiber rings, and the Routing and Wavelength Assignment (RWA) configuration is performed in the control plane, where the wavelengths can be dynamically allocated around the fiber rings. The details of the system parameters are shown in Section IV.



(a) Node architecture

(b) Sip-Ring topology

Fig. 1. Optical interconnection architecture: (a) TeraRack node, (b) Double ring topology [25].

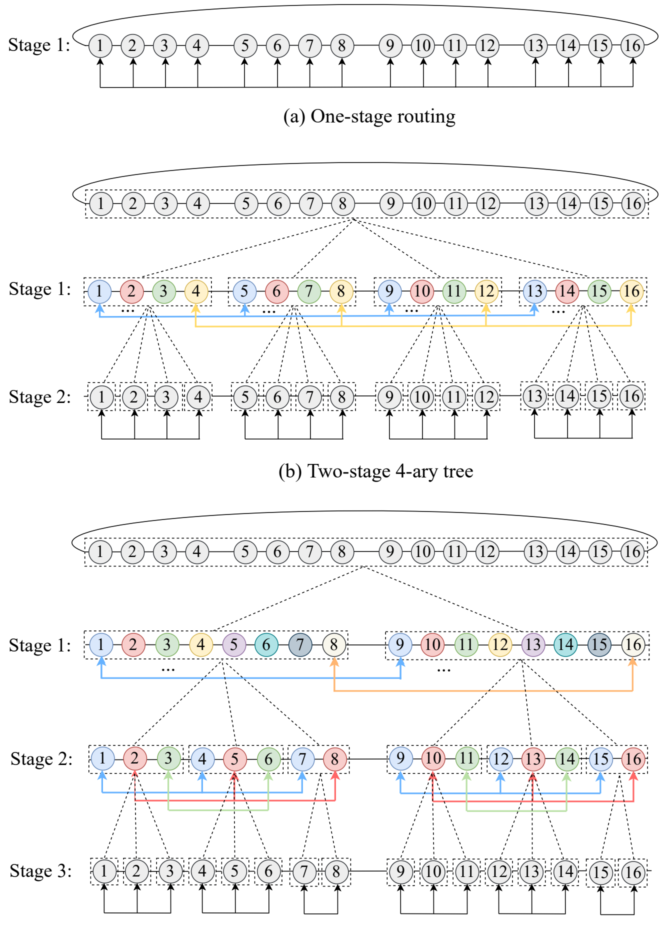
B. Preliminary

The role of the All-gather operation is to make every worker concurrently broadcast the data item until every worker receives the data item from each other. In an optical interconnect system, the All-gather operation corresponds to all-to-all broadcast routing using concurrent optical communications by multiple wavelengths. When the network size is large, the basic approach of one-stage model can lead to an unrealistically large number of wavelengths when realizing all-to-all routing, as can be seen in the following lemma [5].

Lemma 1. *In an N-node optical interconnect system, the minimum number of wavelengths needed for all-to-all routing under one-stage model is $\lfloor \frac{N^2}{4} \rfloor$ in line and $\lceil \frac{N^2}{8} \rceil$ in ring.*

Since the number of wavelengths in the optical interconnect system is limited, multi-stage (k-stage, $k \geq 2$) model can be used to reduce the number of wavelengths, which takes k communication stages to complete all-to-all optical communication [16]. For the k-stage model, the optical signals must be converted to electronic form $k-1$ times before the All-gather operation is completed. However, most designs for multi-stage model only focus on reducing the number of wavelengths but fail to address the performance of All-gather. Therefore, we propose an efficient scheme, OpTree, for All-gather operation on optical interconnect systems based on m-ary tree [26] structure. OpTree can optimize the communication time of All-gather. In graph theory, an m-ary tree is a tree structure in which each internal node has no more than m children. A binary tree is a special case where $m = 2$.

Lemma 2. *For an m-ary tree with height k, the upper bound for the maximum number of leaves is m^k.*



(a) One-stage routing

(b) Two-stage 4-ary tree

(c) Three-stage 3-ary tree

Fig. 2. (a) One-stage, (b) Two-Stage 4-ary tree, and (c) Three-stage 3-ary tree.

C. Motivation

We use a motivation example to illustrate the trade-off between the communication time and the number of stages for our design. We assume that the system has 16 nodes based on the architecture of Fig. 1, the number of available wavelengths is 2, and every node has an amount of data d to collect from all the other nodes. We compare the communication time for the following three schemes: (a) One-stage routing, (b) Two-stage 4-ary tree, and (c) Three-stage 3-ary tree.

One-stage routing: As shown in Fig. 2 (a), one stage routing conducts all-to-all broadcast among all nodes without any intermediate relay nodes. This requires $\lceil \frac{16^2}{8} \rceil = 32$ wavelengths according to Lemma 1. Since the number of available wavelengths is 2, it takes $\lceil \frac{32}{2} \rceil = 16$ communication steps (time slots) to finish the All-gather operation.

Two-stage 4-ary tree: As shown in Fig. 2 (b), the whole group of 16 network nodes is taken as a super root node of a full 4-ary tree, which is partitioned into 4 children (subgroup) in stage 1. Each child tree node contains 4 network nodes. During stage 1, the network nodes on the ith position in each sibling node perform all-to-all routing using one stage model along the ring, and the amount of data sent is d. So, the network nodes 1, 5, 9, and 13 (marked in blue) perform all-to-all routing as a subset, and 2, 6, 10 and 14 (marked in red) perform all-to-all routing as a subset, and so on. Since these four subsets share the optical links in the ring, the wavelength requirement for stage 1 is $4 \cdot \lceil \frac{4^2}{8} \rceil = 8$. This requires 4 communication steps (time slots) to complete. In stage 2, each subgroup (child node of the root node) is further partitioned into 4 children, with each child node having one network node (leaf of the 4-ary tree). During stage 2, sibling nodes conduct all-to-all routing using one stage model on the segment of the ring, and the amount of data sent is $4d$. So, network nodes 1, 2, 3, and 4 perform all-to-all routing as a subset, and 5, 6, 7 and 8 perform all-to-all routing as a subset, and so on. To achieve load balancing in each stage, each wavelength is loaded with a data item of size d. Since these four subsets do not share the optical links in the segment of the ring (line), the wavelength requirement of stage 2 is $4 \cdot \lfloor \frac{4^2}{4} \rfloor = 16$. This requires 8 communication steps (time slots) to finish. Therefore, the total number of communication steps by two-stage 4-ary tree is $\lceil \frac{8}{2} \rceil + \lceil \frac{16}{2} \rceil = 12$.

Three-stage 3-ary tree: As shown in Fig. 2 (c), the three-stage 3-ary tree scheme allows each node to contain at most 3 children with depth of $log_3 16 = 3$ (i.e., 3 stages). Accordingly, we can calculate the wavelength requirements for stages 1, 2, and 3 are $8 \times \lceil \frac{2^2}{8} \rceil = 8$, $2 \times 3 \times \lfloor \frac{3^2}{4} \rfloor = 12$, and $6 \times \lfloor \frac{3^2}{4} \rfloor = 12$ respectively. Therefore, the total number of communication steps to finish the All-gather operation by the three-stage 3-ary tree is $\lceil \frac{8}{2} \rceil + \lceil \frac{12}{2} \rceil + \lceil \frac{12}{2} \rceil = 16$.

We can see that 4-ary tree scheme has the least number of communication steps among these three schemes. One-stage model only uses one stage but requires more time slots than 4-ary tree scheme, whereas 3-ary tree scheme uses more stages than 4-ary tree scheme, but still requires more time slots than 4-ary tree scheme. So, there is a trade-off between communication time and the number of stages by m-ary tree to conduct All-gather operation on optical interconnect systems. The challenging problem is: for a given number of nodes and available wavelengths in the optical interconnect system, how to find the optimal m-ary tree with the optimal number of stages that minimizes communication steps and communication time among all the options of the m-ary tree.

D. Design of OpTree

To address the above challenge, we design an efficient scheme OpTree, represented by m-ary tree structure with the number of All-gather stages represented by the tree depth of k. We first present the routing algorithm and then analyze the communication steps and communication time by OpTree.

1) Routing of OpTree: Assume that the number of nodes in the optical interconnect system is N, and the number of available wavelengths is w. Therefore, the ith level of the m-ary tree corresponds to the ith stage of All-gather operation. As illustrated in Fig. 3, the k-level m-ary tree is constructed by recursively partitioning the group of network nodes into m sub-groups as child tree nodes. According to Lemma 2, each parent tree node has at most $m = N^{\frac{1}{k}}$ child tree nodes, with each child tree node containing $N^{\frac{k-1}{k}}$ network nodes at the ith level of m-ary tree. The depth of the m-ary tree is $k = log_m N$, which is also the number of stages for All-gather operation. The detailed working principle of OpTree is described below.

Stage 1: Initially, the whole group of N network nodes is regarded as the root tree node, which is divided into m child tree nodes denoted as $G_1^1, ..., G_m^1$ with each child node containing $\lceil \frac{N}{m} \rceil$ network nodes. During stage 1, the network node on the ith ($i \in [1, \lceil \frac{N}{m} \rceil]$) position in each of these m sibling nodes can form a subset to perform all-to-all routing by one-stage model. For example, the blue node (node 1) in each of $G_1^1, ..., G_m^1$ broadcasts its data item with size d to all the other blue nodes by one-stage model along the ring.

Stage j: Recursively, each node in stage $j-1$ containing $\lceil \frac{N}{m^{j-1}} \rceil$ network nodes is further partitioned into m child nodes denoted as $G_1^j, ..., G_m^j$ with each child node having $\lceil \frac{N}{m^j} \rceil$ network nodes, until each child node only contains one network node at the last stage k. During stage j, the network node in the ith ($i \in [1, \lceil \frac{N}{m^j} \rceil]$) position in each of its m sibling nodes can form a subset to perform all-to-all routing by one-stage model. For example, during stage 2, the blue node (node 1) in each of $G_1^2, ..., G_m^2$ broadcasts its data item to all the other blue nodes among sibling nodes. During stage k, each leaf node contains only one network node, so $G_1^k, ..., G_m^k$ exchange data with each other among sibling nodes by one-stage model. In stage j, the number of network nodes in each tree node is $\lceil \frac{N}{m^j} \rceil$, and each network node has $m^{j-1}d$ data to send by m^{j-1} wavelengths for load balance.

2) Analysis of Communication Steps: In order to achieve load balancing, OpTree algorithm specifies that each wavelength carries the same amount of data d during each communication step. Therefore, the communication time of All-gather by OpTree is mainly determined by the toal number of communication steps, which we calculate as follows.

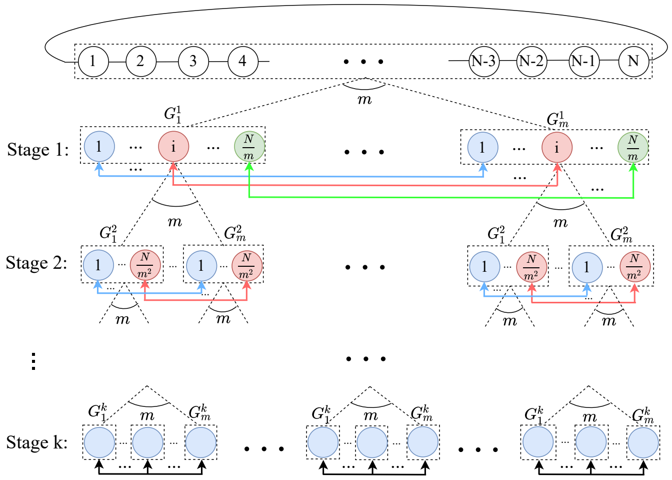


Fig. 3. The working principle in OpTree algorithm.

Theorem 1. *In an N-node optical ring interconnect system with w available wavelengths, the total number of communication steps needed to perform All-gather operation during k stages ($k \geq 2$) by OpTree is $\lceil \frac{(2k-1)N^{1+\frac{1}{k}}}{8w} \rceil$.*

Proof. In stage 1, since there are $\lceil \frac{N}{m} \rceil$ subsets sharing the optical links, the number of communication steps is $\lceil \frac{m^2}{8w} \rceil \times \lceil \frac{N}{m} \rceil$. For each of the subsequent $k-1$ stages, the required number of communication steps in stage j ($j \in [2, k]$) is $m^{j-1} \times \lceil \frac{N}{m^j} \rceil \times \lfloor \frac{m^2}{4w} \rfloor = \lfloor \frac{m^2}{4w} \rfloor \times \lceil \frac{N}{m} \rceil$. Let S represent the total number of communication steps for all k stages of OpTree, which can be calculated as follows.

$$S = \lceil \frac{(2k-1)N^{1+\frac{1}{k}}}{8w} \rceil. \tag{1}$$

Therefore, the theorem holds. □

By Theorem 1, the optimal m-ary tree corresponding to the optimal number of communication stages can be derived as follows, which can achieve the minimum communication time among all the m-ary tree options.

Theorem 2. *In an N-node optical ring interconnect system with w available wavelengths, the optimal number of communication steps by OpTree for finishing the All-gather operation is obtained at $k^* = \left[\frac{lnN+\sqrt{lnN(lnN-2)}}{2} \right]$.*

Proof. Assuming that S is a continuous function and k is a variable in Eq. (1), the number of communication steps is minimized when $\frac{\partial S}{\partial k} = 0$. It can be derived that

$$\frac{\partial S}{\partial k} = \frac{N^{1+\frac{1}{k}}}{4w} - \frac{N^{1+\frac{1}{k}} \cdot lnN \cdot (2k-1)}{8wk^2}.$$

Let $\frac{\partial S}{\partial k} = 0$. S is minimized when

$$k^* = \left[\frac{lnN + \sqrt{lnN(lnN - 2)}}{2} \right], \tag{2}$$

where $[\cdot]$ represents the integer rounding operation. □

Next, we compare the number of communication steps of OpTree with two existing All-gather algorithms on electrical networks and then with a state-of-the-art algorithm in the optical interconnect system. For traditional Ring All-gather [7], it takes $N - 1$ steps, with each step sending amount of data d for the All-gather operation. The Neighbor Exchange (NE) All-gather algorithm [8] requires $\frac{N}{2}$ steps, with each step transmitting data of size 2d, except for the first step. The WRHT algorithm is designed for All-reduce [6] and can be extended to the All-gather operation. It takes $1 + \left\lceil \frac{p(p^{\theta-1}-1)}{p-1} \right\rceil$ steps to collect all the data into one or a few nodes, and either $\lceil (\theta - 1)p^{\theta-1} \rceil$ or $\lceil \theta p^{\theta-1} \rceil$ steps to broadcast the data to any other nodes. The total number of communication steps for the WRHT algorithm is either $1 + \left\lceil \frac{N-p}{p-1} \right\rceil + (\theta - 1)p^{\theta-1}$ or $1 + \left\lceil \frac{N-p}{p-1} \right\rceil + \theta p^{\theta-1}$. We summarize the comparison of the number of communication steps of these algorithms for All-gather operation in Table I.

TABLE I
COMMUNICATION STEP COMPARISON FOR DIFFERENT ALL-GATHER ALGORITHMS IN OPTICAL INTERCONNECT SYSTEMS.

Algorithm	Communication steps	Number of steps
		N = 1024, w = 64
Ring	$N - 1$	1023
NE	$\frac{N}{2}$	512
WRHT	$\left\lceil \frac{N-p}{p-1} \right\rceil + \left\lceil \frac{(\theta-1)N}{p} \right\rceil + 1$	259
One-Stage	$\lceil \frac{N^2}{8w} \rceil$	128
OpTree	$\lceil \frac{(2k^*-1)N^{1+\frac{1}{k^*}}}{8w} \rceil$	70 (k^* = 7)

3) Communication Time of OpTree: Since we obtain the optimal number of communication steps, we can further derive the total communication time to finish the All-gather operation by OpTree, denoted as T_{comm}:

$$T_{comm} = (\frac{d}{B} + a)\lceil \frac{mlog_m N N}{4w} \rceil, \tag{3}$$

where a is the O/E/O conversion delay and the reconfiguration delay of the MRRs, B is the bandwidth per wavelength, d is the amount of transferred data to be received initially for each node, and S represents the total number of communication steps. As d, B, and a are all constant values, the optimal communication time using OpTree can be achieved when the number of communication steps is minimized, as indicated by Theorem 2, as shown below.

Theorem 3. *In an N-node optical ring interconnect system with w available wavelengths, the optimal communication time by OpTree for All-gather operation is* $(\frac{d}{B} + a)\lceil \frac{Nm^* log_{m^*} N}{4w} \rceil$.

Proof. It can be seen from Eq. (3) that the parameters of d, B, and a are all constant values. The optimal communication time using OpTree can be achieved when the number of communication steps for All-gather operation S is minimized. According to Theorem 2, the optimal communication time is $\frac{d\lceil \frac{(2k^*-1)N^{\frac{k^*+1}{k^*}}}{8w} \rceil}{B} + a\lceil \frac{(2k^*-1)N^{\frac{k^*+1}{k^*}}}{8w} \rceil$ for All-gather operation in an N-node optical ring interconnect system with w available wavelengths. □

IV. EVALUATION

A. Simulation setup

We used the same optical interconnect simulator as in [6] to test the performance of OpTree. The simulator, built in Python, uses mathematical modeling with DNN profiling traces as input to simulate optical communication time in optical interconnect systems. Initially, the simulator only implemented all-reduce algorithms, but we modified it to include the proposed OpTree and three All-gather algorithms. The simulation parameters for the optical interconnect system are as follows: the bidirectional ring topology used in TeraRack [25] is employed. The default number of wavelengths is 64, with a bandwidth of 40 Gbps per wavelength. The optical transmission packet size is 128 bytes, and the flit size is 32 bytes. The data type for the All-gather operation is set as float32, and each node in the optical interconnect system is equipped with one GPU. As per [3], the reconfiguration delay of MRRs is 25 μs, and according to [27], the conversion latency of O/E/O is one cycle/flit.

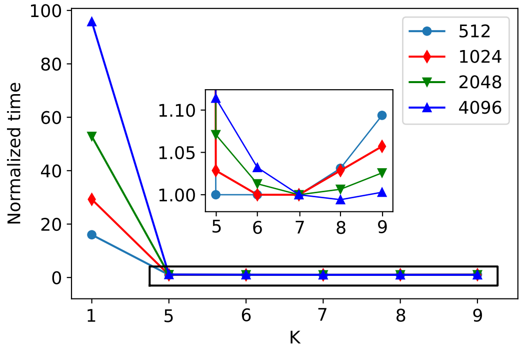


Fig. 4. Performance comparison of OpTree with different depths (denoted by k) for different number of nodes.

B. Optimal k-stage m-ary Tree

In this set of simulations, we verify that OpTree can find an optimal m-ary tree for All-gather operation with minimum communication time among different m-ary tree options under different numbers of nodes. In OpTree, the depth of the m-ary tree with N nodes is k, where $k = log_m N$. Therefore, we can use k to denote the different m-ary trees for different numbers of nodes. We vary the number of nodes from 512 to 4096 with the message size set to 4M. All results in Fig. 4 are normalized by dividing the optimal result of OpTree.

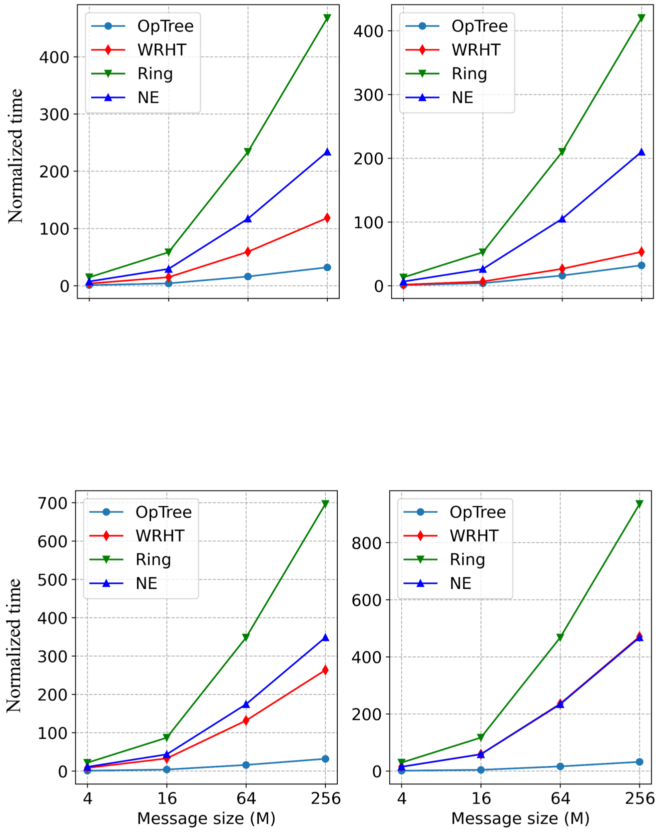
Fig. 4 compares the performance of m-ary trees with different depths (k) for the All-gather operation with 512, 1024, 2048, and 4096 nodes. A depth of $k = 1$ represents the one-stage model for the All-gather operation. The inset of the figure provides a closer look at the performance trends under different numbers of nodes. As the tree depth increases from 5 to 9, it can be seen that there is a trade-off between the depth of the m-ary trees and the number of nodes in the optical interconnect system. For 512 nodes, the performance remains stable at first and then increases dramatically, whereas for 1024 nodes, it decreases initially and remains the same before increasing rapidly. For 2048 and 4096 nodes, the performance decreases dramatically at first before reaching a minimum at depths 7 and 8, respectively, and then increas performance can be achieved at depth 1024, 2048, and 4096 nodes, respec in line with the theoretical findings i confirming the correctness of OpTree to the one-stage model in a ring for dif OpTree reduces communication time

C. Performance Comparison

In this set of simulations, we first c of OpTree with existing schemes incl NE under 1024 and 2048 nodes in system with message sizes ranging 64 wavelengths. Then, we further co by increasing the available waveleng 1024 nodes. All results in Fig. 5 and dividing the first result of OpTree.

Fig. 5 shows the performance com WRHT, Ring, and NE algorithms for A different numbers of nodes. From Fi see that the time of OpTree is the Ring is the highest among different of these algorithms increases slowly with the increasing of message size. As the number of nodes increases from 1024 to 2048, the total time of WRHT gradually reduces, getting close to OpTree. That is because, with the increasing number of nodes, the number of communication steps of OpTree grows faster than that of WRHT. Fig. 6 compares the performance of OpTree, WRHT, Ring, and NE algorithms for the All-gather operation using different wavelengths. As seen in Figs 6 (a) and (b), OpTree has the lowest time cost and Ring has the highest time cost under both 96 and 128 wavelengths. When the wavelength increases from 96 to 128, the time for OpTree slightly decreases while WRHT shows an increasing trend with the increasing message size. Specifically, the time of WRHT is smaller than Ring but slightly larger than NE when the wavelength is set to 128.

Compared to WRHT, Ring, and NE, OpTree can reduce communication time by 56.36%, 92.76%, and 85.54% on average in the optical interconnect system under different numbers of nodes, and can reduce communication time by 88.06%, 95.84%, and 91.69% on average in the 1024-node optical interconnect system under different wavelengths, respectively.



(a) 96 wavelengths

(b) 128 wavelengths

Fig. 6. Performance comparison of different All-gather algorithms in 1024-node optical interconnect system with different wavelengths

V. Limitations and Discussions

It is important to acknowledge the limitations of this work, as there are still areas for improvement. The simulations in this study use a mathematical modeling-based optical interconnect simulator to evaluate different algorithms for all-gather operations. A more thorough analysis of the communication time for the optical interconnect system would require a hardware-specific execution model using a real optical interconnect, which would provide a more accurate evaluation of its performance in real-world scenarios. However, this is outside the scope of this work.

Additionally, the proposed scheme is well-suited for the TeraRack-like architecture, which implements the FlowRing algorithm in the control plane for routing and wavelength assignment. However, it is also important to consider the traditional fat-tree architecture in data centers, as the proposed scheme can easily fit into this architecture with only differences in switch configurations.

VI. Conclusion

In this paper, we propose an efficient scheme called OpTree for All-gather operation in optical interconnect systems. OpTree is based on a m-ary tree structure and we derive the optimal tree corresponding to the minimum communication time among all possible OpTree options. We further analyze and compare the communication steps of OpTree with existing All-gather algorithms. Theoretical results demonstrate that OpTree requires much less number of communication steps than existing methods for All-gather operation in optical interconnect systems. The simulation results demonstrate the effectiveness of OpTree, which can reduce communication time by an average of 72.21%, 94.30%, and 88.58% compared to three existing citep algorithms in the simulated optical interconnect system. Future work can be done by extending to other interconnect topologies and heterogeneous computing device scenarios.

References

[1] K. S. Khorassani, C.-H. Chu, Q. G. Anthony, H. Subramoni, and D. K. Panda, "Adaptive and hierarchical large message all-to-all communication algorithms for large-scale dense gpu systems," in *2021 IEEE/ACM 21st International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*. IEEE, 2021, pp. 113–122.

[2] T. Ben-Nun and T. Hoefler, "Demystifying parallel and distributed deep learning: An in-depth concurrency analysis," *ACM Computing Surveys (CSUR)*, vol. 52, no. 4, pp. 1–43, 2019.

[3] M. Khani, M. Ghobadi, M. Alizadeh, Z. Zhu, M. Glick, K. Bergman, A. Vahdat, B. Klenk, and E. Ebrahimi, "Sip-ml: high-bandwidth optical network interconnects for machine learning training," in *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, 2021, pp. 657–675.

[4] Y. Zhu and J. P. Jue, "Reliable collective communications with weighted srlgs in optical networks," *IEEE/ACM transactions on Networking*, vol. 20, no. 3, pp. 851–863, 2011.

[5] W. Liang and X. Shen, "A general approach for all-to-all routing in multihop wdm optical networks," *IEEE/ACM transactions on networking*, vol. 14, no. 4, pp. 914–923, 2006.

[6] F. Dai, Y. Chen, Z. Huang, H. Zhang, and F. Zhang, "Wrht: Efficient all-reduce for distributed dnn training in optical interconnect system," *arXiv preprint arXiv:2207.10982*, 2022.

[7] J. Chen, L. Zhang, Y. Zhang, and W. Yuan, "Performance evaluation of allgather algorithms on terascale linux cluster with fast ethernet," in *Eighth International Conference on High-Performance Computing in Asia-Pacific Region (HPCASIA'05)*. IEEE, 2005, pp. 6–pp.

[8] J. Chen, Y. Zhang, L. Zhang, and W. Yuan, "Performance of a new allgather algorithm on terascale deepcomp 6800," in *Proceedings of the 2005 Joint DCABES and ICPACE Meeting*, 2005, pp. 73–76.

[9] R. Graham, M. G. Venkata, J. Ladd, P. Shamis, I. Rabinovitz, V. Filipov, and G. Shainer, "Cheetah: A framework for scalable hierarchical collective operations," in *2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*. IEEE, 2011, pp. 73–83.

[10] J. L. Träff, "Efficient allgather for regular smp-clusters," in *European Parallel Virtual Machine/Message Passing Interface Users' Group Meeting*. Springer, 2006, pp. 58–65.

[11] N. T. Karonis, B. R. De Supinski, I. Foster, W. Gropp, E. Lusk, and J. Bresnahan, "Exploiting hierarchy in parallel computer networks to optimize collective operation performance," in *Proceedings 14th international parallel and distributed processing symposium. IPDPS 2000*. IEEE, 2000, pp. 377–384.

[12] K. Kandalla, H. Subramoni, G. Santhanaraman, M. Koop, and D. K. Panda, "Designing multi-leader-based allgather algorithms for multi-core clusters," in *2009 IEEE International Symposium on Parallel & Distributed Processing*. IEEE, 2009, pp. 1–8.

[13] J. L. Träff and S. Hunold, "Decomposing mpi collectives for exploiting multi-lane communication," in *2020 IEEE International Conference on Cluster Computing (CLUSTER)*. IEEE, 2020, pp. 270–280.

[14] S. H. Mirsadeghi and A. Afsahi, "Topology-aware rank reordering for mpi collectives," in *2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE, 2016, pp. 1759–1768.

[15] A. Faraj, X. Yuan, and D. Lowenthal, "Star-mpi: self tuned adaptive routines for mpi collective operations," in *Proceedings of the 20th annual international conference on Supercomputing*, 2006, pp. 199–208.

[16] J.-C. Bermond, L. Gargano, S. Perennes, A. A. Rescigno, and U. Vaccaro, "Efficient collective communication in optical networks," in *International Colloquium on Automata, Languages, and Programming*. Springer, 1996, pp. 574–585.

[17] M. Sabrigiriraj and M. Meenakshi, "All-to-all broadcast in optical wdm networks under light-tree model," *Computer communications*, vol. 31, no. 10, pp. 2562–2565, 2008.

[18] B. Beauquier, "All-to-all communication for some wavelength-routed all-optical networks," *Networks: An International Journal*, vol. 33, no. 3, pp. 179–187, 1999.

[19] L. Narayanan, J. Opatrny, and D. Sotteau, "All-to-all optical routing in chordal rings of degree four," in *Proceedings of the Symposium on Discrete Algorithms*. Citeseer, 1999, pp. 695–703.

[20] S. Kumar and L. V. Kale, "Scaling all-to-all multicast on fat-tree networks," in *Proceedings. Tenth International Conference on Parallel and Distributed Systems, 2004. ICPADS 2004*. IEEE, 2004, pp. 205–214.

[21] X. Zhang and C. Qiao, "On scheduling all-to-all personalized connection and cost-effective designs in wdm rings," *IEEE/ACM Transactions On Networking*, vol. 7, no. 3, pp. 435–445, 1999.

[22] S. A. Pascu and A. A. El-Amawy, "On conflict-free all-to-all broadcast in one-hop optical networks of arbitrary topologies," *IEEE/ACM Transactions on Networking*, vol. 17, no. 5, pp. 1619–1630, 2009.

[23] J. Opatrny, "Uniform multi-hop all-to-all optical routings in rings," *Theoretical computer science*, vol. 297, no. 1-3, pp. 385–397, 2003.

[24] Q.-P. Gu and S. Peng, "Multihop all-to-all broadcast on wdm optical networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 14, no. 5, pp. 477–486, 2003.

[25] M. Khani, M. Ghobadi, M. Alizadeh, Z. Zhu, M. Glick, K. Bergman, A. Vahdat, B. Klenk, and E. Ebrahimi, "Terarack: A tbps rack for machine learning training," 2020.

[26] J. Tolentino, R. M. Marcelo, and M. A. C. Tolentino, "On twin edge colorings in m-ary trees," *Electronic Journal of Graph Theory and Applications (EJGTA)*, vol. 10, no. 1, pp. 131–149, 2022.

[27] F. Dai, Y. Chen, Z. Huang, H. Zhang, H. Zhang, and C. Xia, "Comparing the performance of multi-layer perceptron training on electrical and optical network-on-chips," *The Journal of Supercomputing*, pp. 1–22, 2022.