Wavelength-routing interconnect "Optical Hub" for parallel computing systems

Yutaka Urino[†]
Tsukuba R&D Center
PETRA
Tsukuba, Ibaraki, Japan
y-urino@petra-jp.org

Kenji Mizutani
Tsukuba R&D Center
PETRA
Tsukuba, Ibaraki, Japan
k-mizutani@petra-jp.org

Tatsuya Usuki
Tsukuba R&D Center
PETRA
Tsukuba, Ibaraki, Japan
t-usuki@petra-jp.org

Shigeru Nakamura
Tsukuba R&D Center
PETRA
Tsukuba, Ibaraki, Japan
s-nakamura@petra-jp.org

ABSTRACT

To solve the inter-node bandwidth bottleneck in parallel computing systems, we propose a wavelength-routing inter-node interconnect "Optical Hub". The physical topology of Optical Hub is star network, which leads to advantages in term of its throughput, size, energy consumption and life-time cost. The logical topology is full-mesh network, which leads to advantages in term of its latency and reliability. We introduced multi-path routings, which expand the effective bandwidth with the full-mesh topology such as Optical Hub, by replacing conventional MPI functions with our wrapper functions. We simulated execution time of parallel benchmarks on the parallel computing system with Optical Hub using parallel computing simulator SimGrid. As a result, we have confirmed that the parallel computing system with Optical Hub can achieve higher performance and lower energy consumption than conventional ones. We also examined the scalability of Optical Hub and showed that recursive hierarchical configurations of Optical Hub can save cable count drastically in case of large number of nodes against Dragonfly networks.

CCS CONCEPTS

- Networks → Network components
- Networks → Network algorithms → Data path algorithm
- Networks → Network performance evaluation → Network simulations

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

HPCA2020, January 15–17, 2020, Fukuoka, Japan

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-1-4503-7236-7/20/01...\$15.00

<https://doi.org/10.1145/3368474.3368495>

KEYWORDS

parallel computing, inter-node interconnects, wavelength routing, full-mesh networks, MPI, collective communications, parallel benchmarks, simulations

1 INTRODUCTION

In recent years, while the amount of data to be processed has increased exponentially, the growth of CPU's computing power has slowed down, and parallelism has been main driver to fill the gap [1]. In distributed memory parallel computing systems, if communication performance among computing nodes is not enough, it may become a bottleneck of the overall system performance. To solve this problem, we have developed "Optical I/O Core", which is a high-bandwidth-density optical transceiver using silicon photonics and chip-scale packaging technologies, and a cluster system connecting between FPGA boards using it [2]. Optical I/O Core can contribute to eliminate the communication bottlenecks by its wide communication bandwidth when relatively large data are passed among relatively small number of nodes, that is, in case of network bandwidth bound. On the other hand, when smaller data are passed among more nodes or more frequently, that is, in case of network latency bound, Optical I/O Core may not sufficiently contribute to eliminate the communication bottleneck. For example, in a tree topology, the system scale and performance tend to be limited by the number of ports and the throughput of the hub switch. Also, in a torus topology, the performance tends to be bound by the multi-hop counts. In this paper, in order to solve the above problems caused by the topology and routing of inter-node interconnect, we have proposed Optical Hub which is an inter-node interconnect using wavelength routing. Multi-path routing algorithms, which expand the effective bandwidth with Optical Hub, have been introduced and have been implemented as MPI wrapper functions. The execution time, operation time and communication time of parallel computing benchmarks have been analyzed using the parallel computing simulator SimGrid. We have also examined the scalability of Optical Hub by taking recursive hierarchical configurations. As a result, we have confirmed that the parallel computing system with Optical Hub can achieve smaller size,

higher performance and lower energy consumption than conventional ones.

This paper is organized as follows. In chapter 2, we propose a wavelength-routing interconnect Optical Hub and describe its configuration, operation principle, characteristics, advantages and power consumption. In chapter 3, we introduce two routing methods “direct routing” and “multi-path routing”, which are suitable for Optical Hub and are properly chosen according to the message size. And the methods to realize the multi-path routings by replacing conventional MPI functions with our wrapper functions are described. In chapter 4, we simulate MPI communication time and execution time of parallel benchmarks on the parallel computing system with Optical Hub using parallel computing simulator SimGrid. Energy consumption of the parallel benchmarks are also simulated. In chapter 5, the scalability of Optical Hub is discussed. Previous related studies are introduced in chapter 6. Finally, this paper is summarized in chapter 7.

2 Proposal of Optical Hub

The proposed inter-node wavelength-routing interconnect Optical Hub will be described in this chapter.

2.1 Configuration and operation principle

The physical topology of the connection among eight computing nodes by Optical Hub is shown in Figure 1(a). It is a star topology with a wavelength router (WR) as a hub. Although the connection between each node and the WR is represented by a single line for simplification in the figure, it consists of two optical fibers for bi-directional communication. In this case, the WR has eight input and eight output ports. In a source node, carrier light, of which wavelength corresponds to a destination node, is modulated, then the modulated light signals for all destination nodes are wavelength division multiplexed (WDM), then the WDM light signals are transmitted to the WR via single optical fiber. A signal of a wavelength λ_m input from an input port $\#i$ of the WR is routed to an output port $\#j$ determined by the cyclic equation (1), where N_p is the port count of the WR [3]. The cyclic routing table of the 8-port WR derived from the equation (1) is shown in Figure 1(b).

$$j = \text{mod}(i + m, N_p), \\ 0 \leq i \leq N_p - 1, 1 \leq m \leq N_p - 1. \quad (1)$$

Figure 1: Physical topology of Optical Hub and routing table of the wavelength router.

The WDM light signals input to a destination (reception) node via a single fiber are demultiplexed, then the light signal of which wavelength corresponds to the source node is input to a receiver corresponding to the source node. Note that since the WR does not perform any electric signal processing, the latency hardly increases by passing through the WR. Therefore, the WR can be ignored in logical topology. The logical topology of Optical Hub is full mesh (complete graph in graph theory) as shown in Figure 2(a). In the figure, the single line between each pair of nodes represents a bi-directional link.

An example of the computing node configuration of Optical Hub is shown in Figure 2(b). Each computing node has an accelerator card, which consists of a memory, an application logic, a distributed switch and many optical transceivers (OTRx). The distributed switch, which consists of a direct memory access controller (DMAC), a crossbar switch and a control register, connects any memory addresses to any OTRx. We are now developing the accelerator card with an FPGA. The OTRx are also being developed with silicon photonics and chip-scale packaging technologies like Optical I/O Core. More detailed configuration and preliminary operational experiment results about inter-memory communications via distributed switches are reported in [4].

Figure 2: Logical topology of Optical Hub and configuration in the computing node.

2.2 Characteristics and advantages

Characteristics of Optical Hub are listed in Table 1 as a comparison with the conventional fat-tree packet switch networks. First, the characteristics on their physical topologies will be described. The feature difference between the two networks are mainly due to the difference of their hubs, i.e. the electric packet switch and the WR. First, regarding the power consumption of the hubs, the power consumption in the conventional electric switch is in principle proportional to its throughput (product of line rate and port count), while that of the WR is independent of its throughput. Next, regarding the port count, the upper limit of the port count of the electric switch is mainly determined by the number of electric connectors that can be mounted on the front panel and is up to 36 ports per standard 1 U (19×1.75 inches). On the other hand, the upper limit of the WR is mainly determined by the number of available wavelengths. Assuming that the symbol rate of the signal is 25 GBaud and the channel spacing is 50 GHz, about 80 wavelengths are available throughout the C-band, which has been used in optical fiber telecommunication industry [5]. Since the optical fibers can be arrayed at a pitch of 250 μm or less by using an MT connector, 160 fibers necessary for connection among 80 nodes can be accommodated in a 1 U front panel. In addition, since the electric switch depends on its line rate and modulation format, it is necessary to exchange the switch together when the network interface in the computing node is upgraded. On the other hand, since the WR does not include any electric signal processing, it does not depend on its line rate or modulation format. Therefore, the WR can continue to be used for several generations and is expected to be low life-time cost.

Next, the characteristics on their logical topologies will be described. The features of Optical Hub are mainly due to its logical full-mesh links. Conventional packet switch routing uses a time division manner, and while it has high flexibility, it requires complicated procedures to avoid packet collisions, and causes communication delay and power consumption. On the other hand, since all the nodes in Optical Hub are always directly connected via full-mesh links, there is no need to consider collisions between packets, therefore it is possible to communicate with simpler control processing, and it is expected that communication latency and power consumption can be reduced. Also, since each node of Optical Hub has many small-grained OTRxs and is connected to other nodes via many logical links, even if an OTRx fails, it is possible to keep connecting via other OTRxs and logical links. Therefore, the systems connected with Optical Hub is expected to have higher availability than conventional ones.

As a more specific example, Table 2 lists a specification comparison between Optical Hub and conventional fat-tree network in the case where 64 nodes are connected in full-bisection manner and the total network bandwidth per node is 1.6 Tbps. In the conventional fat-tree, it is assumed that a switch unit has a 400 Gbps / port $\times$ 36 ports, the switch network is composed with two layers (as shown in Figure 6), and the switch network is further

parallelized by four to achieve 1.6 Tbps per node. In this case, in the conventional type, half of interconnect resources are consumed for connections between switches, and signal conversions from optical to electric (OE) and electric to optical (EO) is required at every time when a signal passes through a switch in addition to at the source and destination nodes. On the other hand, in Optical Hub, the EO and OE conversions are required only at source and destination nodes, respectively. Therefore, Optical Hub has advantages in terms of switch size, number of cables, number of OTRxs, and power consumption.

Another option is a physical full-mesh network with parallel link using MPO breakout cables and optical shuffle wiring instead of WDM and wavelength routing. We have chosen WDM and wavelength routing from the following two viewpoints. The first is the bandwidth density on the front panel of the router board. With WDM, the bandwidth density can be multiplied by the number of wavelengths. For example, in the case of 64 nodes, while 8064 fibers, 384 24-core-MPO connectors and 8 U front panels are required for the parallel link, only 128 fibers, 6 MPOs (or 64 DLUs) and 1 U front panel are required for the WDM. The second is the energy efficiency of the optical amplifier, which is essential at least in the hierarchical configuration described in Chapter 5. While the parallel link requires individual amplification for each channel, all WDM channels can be amplified simultaneously by a single fiber amplifier efficiently, so the WDM enables much higher energy efficiency than the parallel link does. After all, although the parallel links are not so bad for a relatively small scale, we have chosen the WDM and wavelength routing for scalability.

Table 1: Characteristics comparison between Optical Hub and conventional fat-tree network.

Item		Fat-tree (Conventional)	Optical Hub
Physical topology		Tree	Star
Hub	Routing method	Electric switch	Wavelength router
	Power consumption	Proportional to throughput	Independent from throughput $\Rightarrow$ Low power
	Limiting factor of port count	Electric connector density on the front panel	Number of wavelengths
	In case of NW upgrade	Switch has to be upgraded together.	Router does not have to be upgraded.
Logical topology		Tree	Full mesh
Link	Connection	Time division (Packet)	Always connected
	Packet processing	Complicated	Simple $\Rightarrow$ Low latency and power
	Port count per node	1	$N - 1$
	Bandwidth granularity	Larger	Smaller
	Tolerance against TRx failures	Low	High

Table 2: Specification comparison between Optical Hub and conventional fat-tree network in case of full-bisection connection of 64 nodes with 1.6-Tbps/node bandwidth.

Item	Fat-tree (Conventional)	Optical Hub
Number of layers	2	1
Bandwidth per link	400 Gbps/link	25 Gbps/link
Switch / Router size	24 U	1 U
Cable count	1,024	128
Transceiver count (in 25Gbps equivalent)	16,384	4,032
Number of E-to-O or O-to-E conversions	4 ~ 8	2

2.3 Configuration of wavelength routing and estimation of communication power

The physical link configuration of Optical Hub is shown in Figure 3. A laser diode (LD) as a light source for generating light carriers with equal frequency spacing used for WDM is shared by all computing nodes in a rack, and the light split by the optical splitter is distributed to each computing node. Thereby, stable wavelength management can be achieved without being affected by temperature change in the computing node. The transmitter (Tx) and receiver (Rx) perform light modulation by optical modulator (Mod) and light reception by photo diode (PD) at 25 Gbps per channel, respectively. The WR is configured by many arrayed waveguide gratings (Mux / Dmx) in order to suppress crosstalk among the same wavelength signals input from different ports (referred to as coherent crosstalk). In order to secure a bit error rate of 10^{-12} or less and a signal to noise ratio of 23 dB or higher at the receiving end with an enough margin, optical fiber amplifiers (EDFAs) were placed at three locations. The first one from the left in the figure is used to compensate for optical loss due to optical splitter, the second one is used to compensate for loss due to optical transmitter, and the third one is used to secure reception power.

The breakdown of communication power consumption per bandwidth or communication power efficiency (in mW/Gbps) is also listed in Figure 3. The communication power efficiency of the transceiver is 5 mW/Gbps like Optical I/O Core [2]. The power efficiency of the distributed switch (crossbar switch and DMAC) and PCIe interface implied in FPGA was estimated by Intel's Early Power Estimator [6]. The power efficiencies of the optical amplifiers were estimated with EO conversion efficiency of 5% [7].

Figure 3: Physical link configuration and communication power efficiency breakdown of Optical Hub.

3 Routing in Optical Hub

3.1 Direct routing and multi-path routing

Consider the case of point-to-point (P2P) communication in a logical full-mesh network. The case of using only single link between source and destination nodes as a communication path (referred to as direct routing) is shown in Figure 4(a). In this case, as shown in the figure, since the only one link, of which bandwidth is relatively narrow, is used, and there are a lot of unused links, the direct routing may be inefficient in some cases. Therefore, the multi-path routing via relay nodes shown in Figure 4(b) is introduced. That is, data to be transmitted is divided into N in the source node, where N is the number of nodes, then $N-1$ of them are scattered to $N-1$ nodes other than the source node (1st-hop), and then the one remaining in the source node and $N-2$ in the relay nodes are gathered to the destination node (2nd-hop). By this method, the effective bandwidth is expanded by $N/2$ times with respect to direct routing, but the communication latency is doubled. Therefore, assuming that the message size is s , the bandwidth of the link is B , and the communication latency per link is L , the direct routing is used when $s < BLN/(N-2)$, i.e. in case of latency bound, while the multi-path routing is used when $s > BLN/(N-2)$, i.e. in case of bandwidth bound. Similarly, for other communication patterns (Bcast, Reduce, Allreduce, etc.), $N/2$ times bandwidth expansion is also realized by the similar multi-path routings.

Figure 4: Direct routing and multi-path routing for point-to-point communications.

3.2 Multi-path routing for collective communications

A pseudo code of a wrapper function to make the multi-path routing for a collective communication by calling multiple P2P MPI functions is shown in Figure 5. For example, the multi-path routing for MPI_Bcast can be performed by replacing MPI_Bcast functions in a source code of an existing parallel computing application with Optmpi_Bcast wrapper functions shown in

Figure 5. The reason why the multi-path routing via relay nodes can be realized with such a relatively simple code is that MPI functions for collective communications are called at all nodes within a communicator including the relay nodes and barrier synchronization are also involved at all nodes.

```
#define MPI_Bcast Optmpi_Bcast
int Optmpi_Bcast(..., root, ...)
{
    /* Scatter at 1st-hop */
    if (rank == root) {
        for (i = 1; i <= size - 1; i++) {
            MPI_Isend(..., (root + i) % size, ...);
        }
    } else if ((rank - root + size) % size <= size - 1) {
        MPI_Recv(..., root, ...);
    }
    /* Allgather at 2nd-hop */
    if (rank != root) {
        for (i = 1; i <= size - 1; i++) {
            if ((rank - i + size) % size != root) {
                MPI_Irecv(..., (rank - i + size) % size, ...);
            }
            if ((rank + i) % size != root) {
                MPI_Isend(..., (rank + i) % size, ...);
            }
        }
    }
    Return 0;
}
```

Figure 5: Pseudo code of multi-path routing for MPI_Bcast.

3.3 Multi-path routing for point-to-point communication

Unlike the above-mentioned collective communications, where all nodes call a unique MPI function with barrier synchronization, in case of P2P communication, a P2P MPI function such as MPI_Send or MPI_Recv is called at only source or destination nodes, respectively. Therefore, the relay nodes for multi-path routing have no way to know that the MPI function has been called at the source or destination node, and it is impossible to make multi-path routing via the relay nodes by a combination of MPI two-side communication functions such as MPI_Send and MPI_Recv. To solve this problem, MPI one-side communication functions such as MPI_Put and MPI_Get, which are supported by MPI-2 or later, are used to write sending data directly from the source node to the public memory in the relay nodes (Put), and to read the receiving data directly from the public memory of the relay node to the destination node (Get) with a passive target synchronization mode (MPI_Win_lock / unlock). Furthermore, the timing control procedure between the source and destination nodes is also incorporated so that the destination node does not start reading from the relay node before the source node finishes writing to the relay node.

4 Performance evaluation

4.1 Parallel computing simulator SimGrid

In order to predict the performance and advantages of parallel computing systems with Optical Hub, we simulated 3 types of parallel computing systems, which are that with Optical Hub and those with conventional networks (2D-torus and fat-tree), using a parallel computing simulator, SimGrid, which can execute existing parallel computing source code using MPI on a virtual parallel computing platform and simulate its execution time [8]. The number of the computing nodes and the computing power per node of the 3 systems were commonly set to 64 and 1 TFlops/node, respectively. The configuration of the 2-layer fat-tree network composed of 36-port packet switches without oversubscription is shown in Figure 6. Specifications of the 3 networks are listed in Table 3. The link bandwidth of 2D-torus and fat-tree is assumed to be 400 Gbps/link, which is expected to be widely penetrated in the early 2020s, while that of Optical Hub is 25 Gbps/link. The communication latency is tentatively made common at 100 ns/link. The routing algorithms for conventional 2D-torus and fat-tree depend on MVAPICH2 [9], and those for Optical Hub depend on the OptMPI wrapper functions as shown in Figure 5. Therefore, in the case of multi-path routing in Optical Hub, the overall communication latency is doubled to 200 ns per path.

Figure 6: Configuration of 2-layer fat-tree network composed of 36-port packet switches without oversubscription for 64 nodes.

Table 3: Specifications of inter-node networks for SimGrid simulations

Topology	2D-torus (8 × 8)	Fat-tree (2 layers)	Optical Hub (Full mesh)
Number of switches/routers	None	6	1
Port count per switch/router	None	36	64
Bandwidth per link (Gbps)	400	400	25
Latency per link (ns)	100	100	100
Routing algorithm	MVAPICH2	MVAPICH2	OptMPI (Wrapper)

4.2 MPI communication time

The communication time of MPI_Bcast in the 2D-torus, fat-tree and Optical Hub simulated by SimGrid is shown in Figure 7. The theoretical prediction values in Optical Hub with the direct routing, $s/B+2L$, and the multi-path routing, $2[s/(NB)+2L]$, are also plotted by a broken line and a dash-dot line, respectively. The communication time in Optical Hub was simulated with the direct routing or the multi-path routing depending on whether the message size is smaller than $BLN/(N-2)$. The cross point of the 2 routing methods is located around 512 Byte in this case. In both regions, the simulated values are in good agreement with theoretical ones. The communication time in the fat-tree is asymptotic to $2(\log_2 N)(s/B+L)$ when the message size is sufficiently large, so the routing algorithm by MVAPICH2 is estimated to be a binary tree. We confirmed that the multi-path routing in Optical Hub is about $N/2$ (32 in this case) times faster than the direct routing for a large-size message as we expected. Furthermore, it was confirmed that Optical Hub is about 10 times and 20 times faster than the conventional networks (2D-torus and fat-tree) in latency bound region and in bandwidth bound region, respectively.

Figure 7: Communication time of MPI_Bcast simulated by SimGrid.

4.3 Parallel computing benchmarks

Some parallel computing benchmarks, such as Integer Sort (IS), Fourier Transform (FT), Data Traffic (DT) in NAS Parallel Benchmarks (NPB) [10] and Graph500 [11], were simulated by using SimGrid. The simulation results for each maximum model size we simulated are summarized in Figure 8. The vertical axis represents the relative inverse number of execution time normalized by that of 2D-torus. The dominant communication patterns of each benchmark were Alltoallv for IS, Alltoall for FT, Send/Recv for DT and Allgather for Graph500. For benchmarks where collective communications are dominant, such as IS, FT, and Graph500, we confirmed that the system with Optical Hub is an order of magnitude faster than that with 2D-torus and about 5 times faster than that with fat-tree. On the other hand, for benchmark where P2P communications are dominant, such as DT,

Optical Hub has not shown a remarkable speed-up effect. The cause of this problem is estimate that the pre-processing and synchronization processing of the MPI one-side communications (Put / Get) to make P2P communication multi-path routing tend to be simulated taking longer time than real on SimGrid. It is expected that there will still be room for speeding up about twice as fast.

Figure 8: Relative performance of parallel computing benchmarks with 3 types of networks by SimGrid.

4.4 Evaluation of operation time and communication time

It is important to know the ratio of operation time and communication time in order to consider speeding up parallel computing systems. The execution time t of a job is divided into an operation time t_p and a communication time t_c , and it is assumed as equation (2).

$$t(P, B, L) = t_p + t_c, \quad t_p = \frac{C}{P}, \quad t_c = \frac{S}{B} + ML, \quad (2)$$

where t_c is the communication time excluding the overlapping part with the operation time, P is the computing power per node (in Flops), B is the network bandwidth per node (in Byte/s), C , S and M are the total operation amount per node (in Flop), the total communication data size per node (in Byte), and the total number of communications per node, respectively. Since SimGrid can set the platform parameters P , B and L to arbitrary values, the application parameters C , S , and M can be derived by simulating the execution time $t(P, B, L)$ with two value of platform parameters and using the following equations (3) to (5). Finally, the operation time t_p and the communication time t_c are obtained by the equation (2).

$$C = \frac{t(P_1, B, 0) - t(P_2, B, 0)}{1/P_1 - 1/P_2}, \quad (3)$$

$$S = \frac{t(P, B_1, 0) - t(P, B_2, 0)}{1/B_1 - 1/B_2}, \quad (4)$$

$$M = \frac{1}{L} [t(P, B, L) - t(P, B, 0)]. \quad (5)$$

4.5 Evaluation of energy consumption in parallel computing system

Next, we will compare the energy consumption per job of parallel benchmark on parallel computing systems using Optical Hub and conventional 2-layer fat-tree network.

First, communication power efficiency or power consumption per bandwidth in mW/Gbps of the 2 networks will be compared. The 2-layer fat-tree is assumed to be composed of 6 packet switches (InfiniBand EDR, 36 ports) as shown in Figure 6, the upper links (inter-switch links) are connected by OTRxs while the lower links (links to/from nodes) are connected by copper cable. The communication power efficiencies of the fat-tree and Optical Hub are listed in Table 4. The power efficiencies of the conventional OTRx, HCA (Host Channel Adapter), and the switch were calculated from specifications of Mellanox's MFA1A00, MCX456, and SB7800, respectively. In the 2-layer fat-tree network, the communication power efficiency of the overall switch was estimated to be 3 times of the value of each single unit. The communication power efficiency of Optical Hub, whose breakdown are referred from the table in Figure 3, is estimated to be about 1/3 of that of the conventional network.

Table 4: Comparison of communication power efficiency breakdown in mW/Gbps between fat-tree and Optical Hub.

Component	Fat-tree (Conventional)	Optical Hub
Switch ASIC	113.3	0.0
HCA	81.5	0.0
Optical transceiver	44.0	5.0
Light source and optical amplifier	0.0	5.3
FPGA	0.0	60.9
Total	238.8	71.2

Next, we will estimate and compare the energy consumption of parallel computing system with conventional 2-layer fat-tree network and Optical Hub. It is assumed that the energy consumption E per node of the conventional system while a certain job is executed is expressed by equation (6).

$$E = (P_C + P_N + P_O)t_P + (P_N + P_O)t_C, \quad (6)$$

where P_C is the CPU power consumption at maximum load, P_N is the communication power (including the power of inter-node switches etc.), P_O is the power other than the CPU and communication. The first term in the equation (6) is the energy consumption during operation and the second term is that during communication (the CPU is idle). To simplify, the CPU power

consumption during communication or idle is assumed to be zero. Similarly, it is assumed that the energy consumption E' per node of a system with Optical Hub is expressed by equation (7).

$$E' = (P_C + P'_N + P_O)t_P + (P'_N + P_O)t'_C. \quad (7)$$

If the ratio of each power consumption and time is defined as equation (8),

$$\alpha = \frac{P_N + P_O}{P_C + P_N + P_O}, \quad \beta = \frac{P'_N - P_N}{P_C + P_N + P_O},$$

$$b = \frac{t_C}{t_P}, \quad e = \frac{t_P + t'_C}{t_P + t_C}, \quad (8)$$

the reduction rate of energy consumption by using Optical Hub against the conventional system is expressed as equation (9).

$$1 - \frac{E'}{E} = \frac{[\alpha(1 - e) + \beta e](1 + b)}{1 + \alpha b}. \quad (9)$$

Among the variables on the right side of equation (9), α and β are mainly determined by the hardware configuration of the parallel computing system. These parameters of both parallel computing systems with conventional network and Optical Hubs are listed in Table 5. Power consumption values other than that of communication were referred from Lenovo Power Configurator [12]. The communication power efficiencies are referred from Table 4.

Table 5: Power parameters of parallel computing system.

Item	InfiniBand fat-tree (Conventional)	Optical Hub
Communication power efficiency (mW/Gbps)	239	71
Bandwidth per node (Gbps/node)	400	1600
Communication power (W/node)	96	114
CPU power at maximum load (W/node)	150	150
Other power (W/node)	100	100
α	0.57	0.59
β		-0.05

On the other hand, b and e in equation (9) also depend on their applications as well as the hardware configuration and can be estimated using SimGrid as described above. The reduction rate of energy consumption per parallel benchmark job by using Optical Hub against the conventional system ($1 - E'/E$) calculated by equation (9) is listed in Table 6. In case of NPB-IS, FT and Graph 500, in which collective communication is the dominant communication pattern, we confirmed that the reduction rates of energy consumption per job are about 80%. On the other hand, for benchmark where P2P communications are dominant, such as DT, Optical Hub has not shown a remarkable energy saving effect. As mentioned in section 4.3, the cause of this problem is estimate that

the pre-processing and synchronization processing for the MPI one-side communications to make P2P communication multi-path routing tend to be simulated taking longer time than real on SimGrid. It is expected that there will still be room for speeding up about twice as fast. If the execution time of DT is half of this simulation result, it is estimate that the reduction rate of energy consumption per job is about 36%.

Table 6: Reduction rate of energy consumption per parallel computing benchmark job by Optical Hub.

Benchmark	b	e	Reduction rate of energy consumption by Optical Hub
NPB-IS	1.7E+05	0.20	77.9%
NPB-FT	1.3E+05	0.19	79.7%
NPB-DT	1.0E+06	1.18	-28.8%
Graph500	4.4E+04	0.19	79.2%

5 Scalability

As mentioned in section 2.2, the port count of WR in Optical Hub is mainly limited by the number of available wavelengths, and the upper limit is about 80. So, the upper limit of node count is also about 80 with single-layer Optical Hub. A recursive hierarchical configuration of Optical Hub is effective to connect more nodes. For example, a physical topology of 2-layer Optical Hub with 5 5-port WRs and 4 wavelengths is shown in Figure 9. In this case, 4 nodes are connected each other via a 1st-layer WR, and make a group, or a super-node. Furthermore, 4 super-nodes are also connected via a 2nd-layer WR, and make a group again if necessary, and so on recursively. Table 7 shows a routing table of each WR in Figure 9. In the routing table, the node #0 represents the WR in the upper layer.

In the 2-layer configuration, a signal of a wavelength λ_m in a source node $\#i_1$ in a super-node $\#i_2$ is input to an input port $\#i_1$ of a 1st-layer WR in the super-node $\#i_2$ and then is routed to a destination node $\#j_1$ in a super-node $\#j_2$ determined by the equation (11), which means that a signal that satisfies the condition $k_1 \neq 0$ is routed within its super-node while that with $k_1 = 0$ is routed across super-nodes.

$$[j_2, j_1] = \begin{cases} [i_2, k_1] & \text{if } k_1 \neq 0 \\ [k_2, m] & \text{if } k_1 = 0, \end{cases}$$

where

$$\begin{aligned} k_l &= \text{mod}(i_l + m, N_p), \\ 0 \leq i_l \leq N_p - 1, \quad 1 \leq m \leq N_p - 1, \\ l &= 1, 2. \end{aligned} \quad (11)$$

For example, a node identification vector $[i_2, i_1]$ represents the node whose node identification within a super-node is i_1 and whose super-node node identification is i_2 . More generally, in a d -layer hierarchical configuration, a node identification is represented by a d -dimensional vector. The relation between

source node identification $[i_d, \dots, i_l, \dots, i_1]$ and destination node identification $[j_d, \dots, j_l, \dots, j_1]$ is expressed with equation (12).

$$j_l = \begin{cases} i_l & \text{if } \sum_{n=1}^{l-1} k_n \neq 0 \\ k_l & \text{else if } k_l \neq 0 \\ m & \text{else if } k_l = 0, \end{cases} \quad 1 \leq l \leq d. \quad (12)$$

In this type configuration, since one WR port is used for uplink and the others are for downlink, the oversubscription ratio is $N_p - 1 = W$, where W is the number of wavelengths. The number of nodes can be scaled up to $(N_p - 1)^d = W^d$. Since up to 80 wavelengths are available, the node count is up to 6,400 with 2 layers and 512,000 with 3 layers. Note again that since the WRs do not perform any electric signal processing, no matter how many times a signal passes through the WRs, the latency hardly increases except the flight time of light. Therefore, the WRs can be ignored in logical topology. Figure 10 shows the logical topology of the 2-layer Optical Hub, which is equal to that of a kind of Dragonfly network [13]. The single line between each pair of nodes represents a bi-directional link. Since any nodes are directly connected to a node in a particular super-node and the nodes within a super-node are directly connected each other in full-mesh manner, the network diameter is 3, each is for intra-source-super-node hop, inter-super-node one and intra-destination-super-node one. Therefore, in the worst case, the communication time in this 2-layer configuration is triple of that in the single layer. Table 8 lists cable count of Dragonfly and the 2-layer Optical Hub as functions of node count N . Since the cable count in Dragonfly is $N^{3/2}$, while that in the 2-layer Optical Hub is $2(N+N^{1/2})$, the 2-layer Optical Hub can save cables drastically in case of large number of nodes against Dragonfly network.

Figure 9: Physical topology of 2-layer Optical Hub with oversubscription. The network diameter is 3.

Table 7: Routing table of each wavelength router in Figure 9.

Wave-length	Output port to destination node #				
	0	1	2	3	4
Input port from source node #	0	λ_1	λ_2	λ_3	λ_4
	1	λ_4		λ_1	λ_2
	2	λ_3	λ_4		λ_1
	3	λ_2	λ_3	λ_4	
	4	λ_1	λ_2	λ_3	λ_4

Figure 10: Logical topology of 2-layer Optical Hub, whose physical topology and routing table are shown in Figure 9 and table 7, respectively.

Table 8: Cable count of Dragonfly and 2-layer Optical Hub as functions of node count.

Number of nodes	Number of cables	
	Dragonfly	Optical Hub
4	8	12
16	64	40
64	512	144
256	4,096	544
1,024	32,768	2,112
4,096	262,144	8,320

Another example of 2-layer Optical Hub with 8 8-port WRs and 7 wavelengths is shown in Figure 11. The topology is similar to that of 2-layer fat-tree. The routing table of the 1st and 2nd layer WRs are shown in Table 9(a) and (b), respectively. The output port numbering of the 2nd-layer WRs is shift by 1 port from that of the 1st-layer WRs.

In the 2-layer configuration, a signal of a wavelength λ_m in a source node $\#i_1$ in a super-node $\#i_2$ is input to an input port $\#2i_1$ of a 1st-layer WR in the super-node $\#i_2$ and then is routed to a destination node $\#j_1$ in a super-node $\#j_2$ determined by the equation (13), which means that a signal with an even wavelength identification number m is routed within its super-node while that with an odd number is routed across super-nodes.

$$\begin{aligned}
 & [j_2, j_1] \\
 &= \begin{cases} [i_2, \text{mod}(2i_1 + m, N_p)/2] & \text{if } m \text{ is even} \\ [\text{mod}(2i_2 + m + 1, N_p)/2, \text{mod}\{2(i_1 + m), N_p\}/2] & \text{otherwise,} \end{cases} \\
 & 0 \leq i_l \leq N_p - 1, \quad 1 \leq m \leq N_p - 1, \quad l = 1, 2. \quad (13)
 \end{aligned}$$

In this type configuration, since $N_p/2$ WR ports are used for both uplink and for downlink, respectively, the oversubscription ratio is equal to 1. The number of nodes can be scaled up to $(N_p/2)^2 = [(W+1)/2]^2$, i.e. about 1600. Figure 12 shows the logical links of the 2-layer Optical Hub. In order to avoid overcomplicating the figure, the inter-super-nodes links presented with dashed arrows are only shown those from source nodes in super-node #0. Since any nodes are directly connected to nodes in any super-nodes and nodes within a super-node are directly connected each other in full-mesh manner, the network diameter is 2, each is for intra-super-node hop within either source or destination super-node and

for inter-super-node one. Therefore, in the worst case, the communication time in this 2-layer configuration is double of that in the single layer.

Figure 11: Physical topology of 2-layer Optical Hub without oversubscription. The network diameter is 2.

Table 9: Routing table of 1st and 2nd layer wavelength routers in Figure 11.

(a) Routing table of 1st-layer WRs							
Wavelength	0	1	2	3	4	5	6
Input port from source node #	0	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
1	λ_7	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
2	λ_6	λ_7	λ_1	λ_2	λ_3	λ_4	λ_5
3	λ_5	λ_6	λ_7	λ_1	λ_2	λ_3	λ_4
4	λ_4	λ_5	λ_6	λ_7	λ_1	λ_2	λ_3
5	λ_3	λ_4	λ_5	λ_6	λ_7	λ_1	λ_2
6	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_1
7	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7

(b) Routing table of 2nd-layer WRs							
Wavelength	0	1	2	3	4	5	6
Input port from source node #	0	λ_7	λ_1	λ_2	λ_3	λ_4	λ_5
1	λ_6	λ_7	λ_1	λ_2	λ_3	λ_4	λ_5
2	λ_5	λ_6	λ_7	λ_1	λ_2	λ_3	λ_4
3	λ_4	λ_5	λ_6	λ_7	λ_1	λ_2	λ_3
4	λ_3	λ_4	λ_5	λ_6	λ_7	λ_1	λ_2
5	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_1
6	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
7	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7

Figure 12: Logical links of 2-layer Optical Hub, whose physical topology and routing table are shown in Figure 11 and table 9, respectively. The inter-super-nodes links presented with dashed arrows are only shown those from source nodes in super-node #0.

We can set any oversubscription ratios between $W-1$ to 1 by taking an intermediate configuration of 2-layer Optical Hub mentioned in the first and second examples.

6 Related works

Kudo et al. proposed a routing method that implements multiple paths on an L2 Ethernet network by using a VLAN, and an example is shown in the case where switches are connected by a

complete graph [14]. Unlike our proposal in which nodes are directly connected with logically full-mesh topology, there are restrictions on the number of ports and throughput in the switches. Fukunaga proposed a PC cluster with complete graph network using a multi-port Ethernet NIC [15]. Although the routing algorithms via relay nodes were similar to those of our approach, there were some limitations in terms of shortening the latency and scalability of the number of nodes due to using the existing multi-port Ethernet NIC.

Regarding logical full-mesh connection by wavelength routing using cyclic arrayed waveguide gratings (AWGs), research and development have been conducted mainly by NTT Group in the field of optical fiber telecommunication industry. The field trials had been reported in [16]. As studies of applying wavelength routing to inter-node interconnect for parallel computing systems, there is a group by S. J. B. Yoo et al. In the paper [17], by layering the wavelength-routing network, they showed that it was possible to scale over 100,000 nodes, and 2.5 times throughput was possible compared to a conventional fat-tree network with 3:1 oversubscription. However, there was no mention of speeding up and energy saving effect by multi-path routing within the range of the author's knowledge.

7 Summary

In order to eliminate the inter-node communication bottleneck in parallel computing systems, we proposed an inter-node interconnect "Optical Hub" using wavelength routing.

Optical Hub has following features. Up to 80 nodes can be directly connected to the central wavelength router via two fibers per node with a star physical topology. And the signal conversions between electric and optical ones are only once at the source and destination node each. Therefore, Optical Hub has advantages in terms of its size and power consumption in comparison with the prior art. Since the wavelength router does not include any electric signal processing, it does not depend on its line rate or modulation format. Therefore, the wavelength router can continue to be used for several generations and is expected to be low life-time cost. In addition, since all the nodes in Optical Hub are always directly connected via full-mesh links, there is no need to consider collisions between packets, therefore it is possible to communicate with simpler control processing, and it is expected that communication latency and power consumption can be reduced. Also, since each node of Optical Hub has many small-grained optical transceivers and is connected to other nodes by many logical links, even if one transceiver fails, it is possible to keep connecting via other transceivers and logical links. Therefore, the systems connected with Optical Hub is expected to have higher availability than conventional ones.

We introduced multi-path routing algorithms which expands effective bandwidth with Optical Hub and implemented them as

MPI wrapper functions. The execution time, operation time and communication time of parallel computing benchmarks with Optical Hub and conventional networks were simulated by using SimGrid. We confirmed it is possible to speed up the job execution and to save energy consumption per job by using Optical Hub compared to conventional networks.

We also examined the scalability of Optical Hub and showed that recursive hierarchical configurations of Optical Hub can save cable count drastically in case of large number of nodes against Dragonfly networks.

ACKNOWLEDGMENTS

This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO). We would like to thank Professor Koibuchi of National Institute of Informatics (NII) for useful guidance and discussions.

REFERENCES

- [1] S. Rumleya, M. Bahadori, R. Polster, S. D. Hammond, D. M. Calhoun, K. Wen, A. Rodrigues, and K. Bergman. 2017. Optical interconnects for extreme scale computing systems. *Parallel Computing*, 64, 65-80.
- [2] T. Nakamura, K. Yashiki, K. Mizutani, T. Nedachi, J. Fujikata, M. Tokushima, J. Ushida, M. Noguchi, D. Okamoto, Y. Suzuki, T. Shimizu, K. Tanaka, A. Ukita, Y. Ibusuki, M. Kurihara, K. Kinoshita, T. Horikawa, H. Yamaguchi, J. Tsuchida, Y. Hagiwara, and K. Kurata. 2119. Fingertip-Size Optical Module, "Optical I/O Core", and Its Application in FPGA. *IEICE Trans. Electron.*, E102-C(4), 333-339.
- [3] K. Okamoto, T. Hasegawa, O. Ishida, A. Himeno, and Y. Ohmori. 1997. 32×32 arrayed-waveguide grating multiplexer with uniform loss and cyclic frequency characteristics. *Electron. Lett.*, 33(22), 1865-1866.
- [4] K. Mizutani, H. Yamaguchi, and Y. Urino. 2019. Characteristics of inter-memory communications via distributed switches for inter-accelerator full mesh interconnect by Optical Hub. *IEICE Technical Report*, CPSY2019-32 (in Japanese).
- [5] E. Yamazaki, S. Yamanaka, Y. Kisaka, T. Nakagawa, K. Murata, E. Yoshida, T. Sakano, M. Tomizawa, Y. Miyamoto, S. Matsuoka, J. Matsui, A. Shibayama, T. Abe, Y. Nakamura, H. Noguchi, K. Fukuchi, H. Onaka, K. Fukumitsu, K. Komaki, O. Takeuchi, Y. Sakamoto, H. Nakashima, T. Mizuochi, K. Kubo, Y. Miyata, H. Nishimoto, S. Hirano, and K. Onohara. 2011. Fast optical channel recovery in field demonstration of 100-Gbit/s Ethernet over OTN using real-time DSP. *Optics Express*, 19(14), 13179-13184.
- [6] <https://www.intel.co.jp/content/www/jp/ja/programmable/support/support-resources/operation-and-testing/power/pow-powerplay.html>.
- [7] L. Lundberg, P. A. Andrekson, and M. Karlsson. 2017. Power Consumption Analysis of Hybrid EDFA / Raman Amplifiers in Long-Haul Transmission Systems. *J. Lightwave Technol.*, 35, 2132-2142.
- [8] <http://simgrid.gforge.inria.fr/>.
- [9] <http://mvapich.cse.ohio-state.edu/>.
- [10] <https://www.nasa.gov/publications/npb.html>.
- [11] <https://graph500.org/>.
- [12] <https://support.lenovo.com/jp/ja/solutions/lvno-pwrconf>.
- [13] J. Kim, W. J. Dally, S. Scott, and D. Abts. 2008. Technology-Driven, Highly-Scalable Dragonfly Topology. *IEEE International Symposium on Computer Architecture*. 77-88.
- [14] T. Kudo, M. Matsuda, H. Tezuka, Y. Kodama, O. Tatebe, and S. Sekiguchi. 2004. VLAN-based multi-path L2 Ethernet network for clusters. *IPSSJ Journal*, 45(SIG 6), 35-440 (in Japanese).
- [15] T. Fukunaga. 2014. Fully Connected PC Cluster with Indirect Routes. *IPSSJ Journal*, 55(1), 484-493 (in Japanese).
- [16] K. Noguchi, Y. Koike, H. Tanobe, K. Harada, and M. Matsuoka. 2004. Field Trial of Full-Mesh WDM Network (AWG-STAR) in Metropolitan/Local Area. *J. Lightwave Technology*, 22(2), 329-336.
- [17] Z. Cao, R. Proietti, and S. J. B. Yoo. 2015. Hi-LION: Hierarchical Large-scale Interconnection Optical Network with AWGRs [Invited]. *IEEE/OSA J. Optical Communications and Networking*, 7(1), A97-A104.