Cluster Comput (2010) 13: 87–110
DOI 10.1007/s10586-009-0117-8

Performance evaluation of broadcast and global combine operations in all-port wormhole-routed OTIS-Mesh interconnection networks

Basel A. Mahafzah · Ruby Y. Tahboub · Omar Y. Tahboub

Received: 13 April 2009 / Accepted: 27 December 2009 / Published online: 28 January 2010

Abstract OTIS (Optical Transpose Interconnection System) optoelectronic architecture is an attractive high-speed interconnection network. As a continuation for the research work performed on OTIS, this paper investigates broadcast and global combine communication operations on the promising all-port wormhole-routed OTIS-Mesh using the Extended Dominating Node (EDN) approach, referred to as EDN-OTIS-Mesh. The performance of broadcast and global combine operations is evaluated, both analytically and by simulation, in terms of the number of communication steps, latency, and latency improvement. A comparative study is conducted among three interconnection networks' architectures: the single-port wormhole-routed OTIS-Mesh, all-port wormhole-routed OTIS-Mesh, and all-port wormhole-routed EDN-OTIS-Mesh. The obtained analytical and simulation results show that the broadcast and global combine operations on all-port EDN-OTIS-Mesh significantly outperform the single-port and all-port OTIS-Mesh.

B.A. Mahafzah (✉)
Department of Computer Science, King Abdullah II School for Information Technology, The University of Jordan, Amman 11942, Jordan
e-mail: b.mahafzah@ju.edu.jo

R.Y. Tahboub
Department of Computer Science, School of Computer & Information Technology, Jordan University of Science & Technology, Irbid 22110, Jordan
e-mail: rytahboub6@just.edu.jo

O.Y. Tahboub
Department of Computer Science, Kent State University, Kent 44240, USA
e-mail: otahboub@cs.kent.edu

Keywords Broadcast · Global combine · Extended Dominating Node · OTIS Mesh · Interconnection networks · Performance evaluation

1 Introduction

Under the technological advancement and the rising need for huge computational power, a great attention has been directed towards parallel and distributed processing systems. A distributed processing system is mainly composed of a plurality of Processing Elements (PEs), interconnected through communication links over an interconnection network [1]. PEs represent the computational aspect of the interconnection network architecture. The communication part is achieved through communication links, which are categorized based on their communication nature into electronic and optical. Electronic links are generally used among PEs interconnected along short distances, while optical links are preferred under long distances.

The distributed computing system's performance is greatly affected by the underlying interconnection network. The network's topology and the nature of the communication links have a major effect on the system performance (in terms of its speed, overhead, etc.). Therefore, the choice of the interconnection network topology is a critical issue that must be taken into consideration when designing distributed systems and developing distributed applications [2]. Furthermore, designing and implementing efficient communications operations in interconnection networks and message-passing systems plays great role in the performance of distributed applications [2–4].

An emerging communication improvement is based on the use of both electronic and optical links in one system.

Such a system is referred to as optoelectronic. An attractive optoelectronic architecture that has gained a considerable attention in the recent years is the OTIS (Optical Transpose Interconnection System) architecture [5]. In an OTIS system, processors are organized into groups, where in each group; processors are connected electronically along short distances forming a basis network, such as a mesh, hypercube, ring etc., while the longer interconnection among groups is achieved optically. OTIS-Mesh, under study, is an instance of OTIS, in which each of the constituting groups is a mesh network.

In parallel and distributed systems, broadcast and global combine are two vital communication operations involving two or more processors that perform the same operation. The significance of such operations lies in their wide use in various applications, including graphs, sorting and image processing. Thus, efficient implementation of the broadcast and global combine communication operations forms crucial design and yet performance challenges. One graph theoretic approach to broadcast and global combine is called the Extended Dominating Set (EDS) [6, 7]. This approach defines a set of nodes, referred to as Extended Dominating Nodes (EDNs), which are capable of delivering a message to all other processors in a group within a single message-passing step.

In this paper, efficient broadcast and global combine communication operations in all-port wormhole-routed OTIS-Mesh interconnection networks is evaluated, both analytically and by simulation. The broadcast and global combine algorithms are based on Tsai and McKinley's approach (EDN approach) [6–8], where these algorithms have been modified and embedded in order to be applied and implemented on OTIS-Mesh. More specifically, two research goals are achieved in this paper. First, an extensive performance evaluation study is conducted to both analyze and compare the performance of broadcast and global combine operations on single-port wormhole-routed OTIS-Mesh, all-port wormhole-routed OTIS-Mesh, and all-port wormhole-routed EDN-OTIS-Mesh. However, the difference between all-port wormhole-routed OTIS-Mesh and all-port wormhole-routed EDN-OTIS-Mesh is that the later uses broadcast and global combine operations based on the EDN approach. Our performance metrics include the number of communication steps, latency, and latency improvement. Second, the EDN approach is applied in designing and implementing efficient broadcast and global combine communication operations on all-port wormhole-routed OTIS-Mesh interconnection networks, which is referred to as EDN-OTIS-Mesh.

The remainder of this paper is organized as follows: Sect. 2 provides background and related work on broadcast and global combine communication operations, EDN approach, OTIS systems, and more specifically, OTIS-Mesh. The broadcast and global combine communication operations on all-port wormhole-routed OTIS-Mesh using the EDN approach are presented in Sect. 3. This is followed by an analytical evaluation of the broadcast and global combine algorithms in Sect. 4. Our analytical results are validated in Sects. 5 and 6 through a comprehensive simulation work, which involves a comparative study among single-port, all-port wormhole-routed OTIS-Mesh, and all-port wormhole-routed EDN-OTIS-Mesh interconnection networks. Finally, Sect. 7 concludes the paper and suggests some future work.

2 Background and related work

This section presents some related background on broadcast and global combine communication operations, EDN approach, OTIS interconnection networks, and some basic operations in OTIS-Networks.

2.1 Broadcast and global combine operations

In a broadcast communication operation [9], one processor sends the same message for each processor in the interconnection network. Figure 1 illustrates the broadcast communication operation, where P_0 presents the root processor, which sends the same message M to all other processors (P_1, P_2, P_3, and P_4) in the interconnection network.

In a global combine communication operation [9], one processor receives a different message from each processor in the interconnection network. Figure 2 illustrates the global combine communication operation, where P_0 presents the root processor; P_0 receives a different message (M_1, M_2, M_3, and M_4) from each of the processors (P_1, P_2, P_3 and P_4) in the interconnection network.

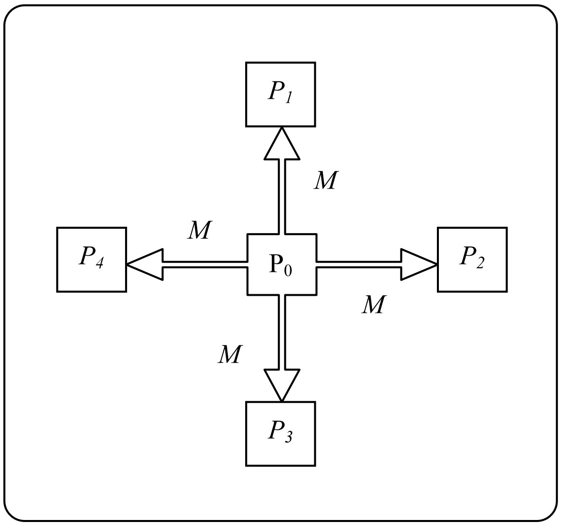


Fig. 1 Broadcast communication operation

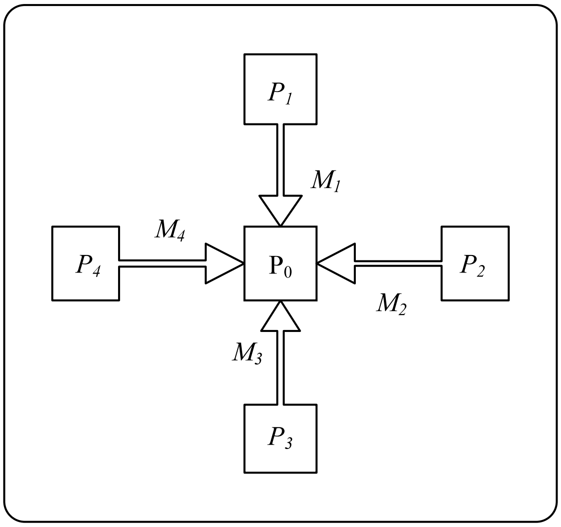


Fig. 2 Global combine communication operation

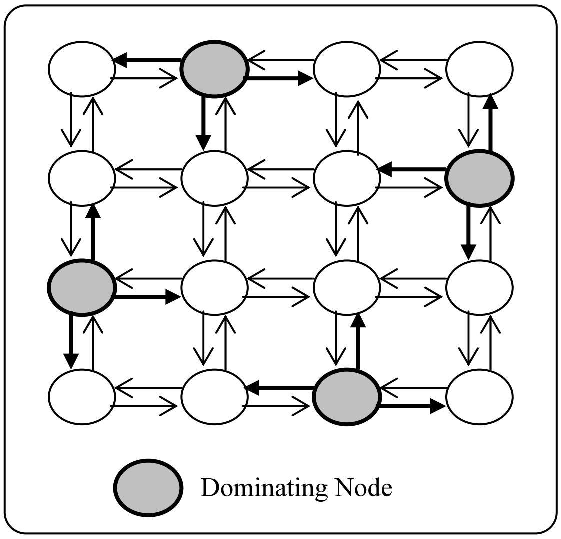


Fig. 3 Dominating nodes in a 4 × 4 2D-Mesh

Broadcast and global combine communication operations play a great role in developing message-passing programs and solving many parallel problems [7, 10–14] with their ability to rapidly distribute or collect large amounts of data on different interconnection networks, such as OTIS-Mesh, OTIS-Hypercube, two-dimensional mesh etc.

Hartmann et al. [10] presented an adaptive extension library for the Message Passing Interface (MPI) library, which is capable of improving the performance of specific collective communication operations. The extended library improved the collective communication operations' performance by decomposing these operations into a number of MPI communication steps. Experimental results revealed that the extension library for MPI significantly improves the performance of collective communication operations. Matsuda et al. [12] modified some collective operations, such as broadcast and all reduce algorithms, in MPI to effectively utilize fast wide-area inter-cluster networks and to control the number of nodes, which can transfer data concurrently through wide-area networks to avoid congestion. Furthermore, Pjesivac-Grbovic et al. [13] analyzed and improved some collective operations, such as broadcast, in MPI for high performance computing.

Dvorak [15] investigated the overhead effect of collective communication such as broadcast and global combine in single-port wormhole-routed ring-based and 2D-Mesh interconnection networks, where real collective communication algorithms were used to evaluate both of upper bound and lower bound communication steps.

Chen et al. [16] investigated a multi-node broadcasting algorithm in all-port torus using the deterministic dimension order routing. In order to perform efficient broadcasting, aggregation followed by distribution techniques were applied. The messages are first aggregated into positions. Then, the constructed sub-networks distribute the messages, achieving maximized parallelism.

Barnett et al. [17] presented broadcast algorithms for meshes that are non-power-of-two. Also, conflict-free minimum-spanning tree, pipelined, and a proposed scatter-collect approach broadcast algorithms were presented and evaluated in [17].

2.2 Extended dominating node approach

The dominating set approach is applied in many areas such as graph theory, unit disc graphs, and wireless sensor networks [18]. However, the dominating node approach [8] defines a set of dominating nodes that have direct links to all other processors in a group. Such a feature grants a processor the capability of delivering a message to all other processors in a single step, as illustrated in Fig. 3.

By considering the scenario illustrated in Fig. 4, the extended dominating nodes still preserve the single step message delivery (broadcasting) by also delivering to non-adjacent processors. Furthermore, the definition of Extended Dominating Nodes (EDNs) can be recursively applied to form multiple levels of EDN processors, as illustrated in Fig. 5. As shown in this figure, gray nodes present level-2 EDN processors, dotted nodes present level-1 EDN processors, and white nodes present level-0 EDN processors. The process of message broadcasting is performed in two steps as follows: EDN level-2 transmits the message to EDN level-1, which then transmits the message to EDN level-0.

The EDN approach was applied in the design of collective communication operations, such as reduction, broadcasting, and global combine operations in all-port wormhole-routed 2D-Mesh [6, 7]. EDN was also applied to solve the matrix transposition problem in order to minimize channel contention [6]. Moreover, the advantages of EDN approach-based collective communication operations over other approaches were confirmed in [6] through simulation and analysis.

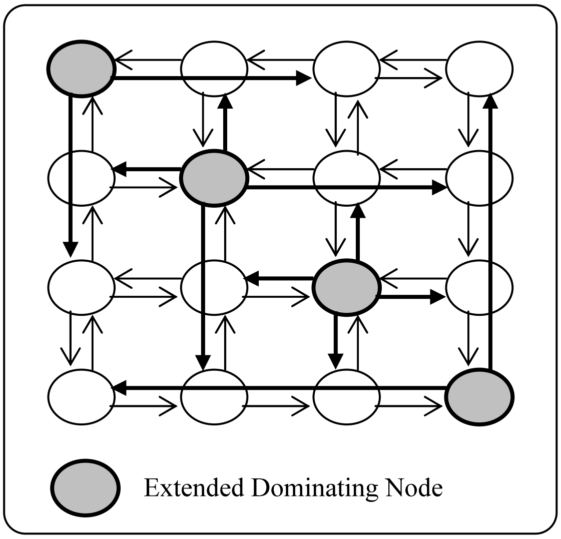


Fig. 4 Extended dominating nodes in a 4 × 4 2D-Mesh

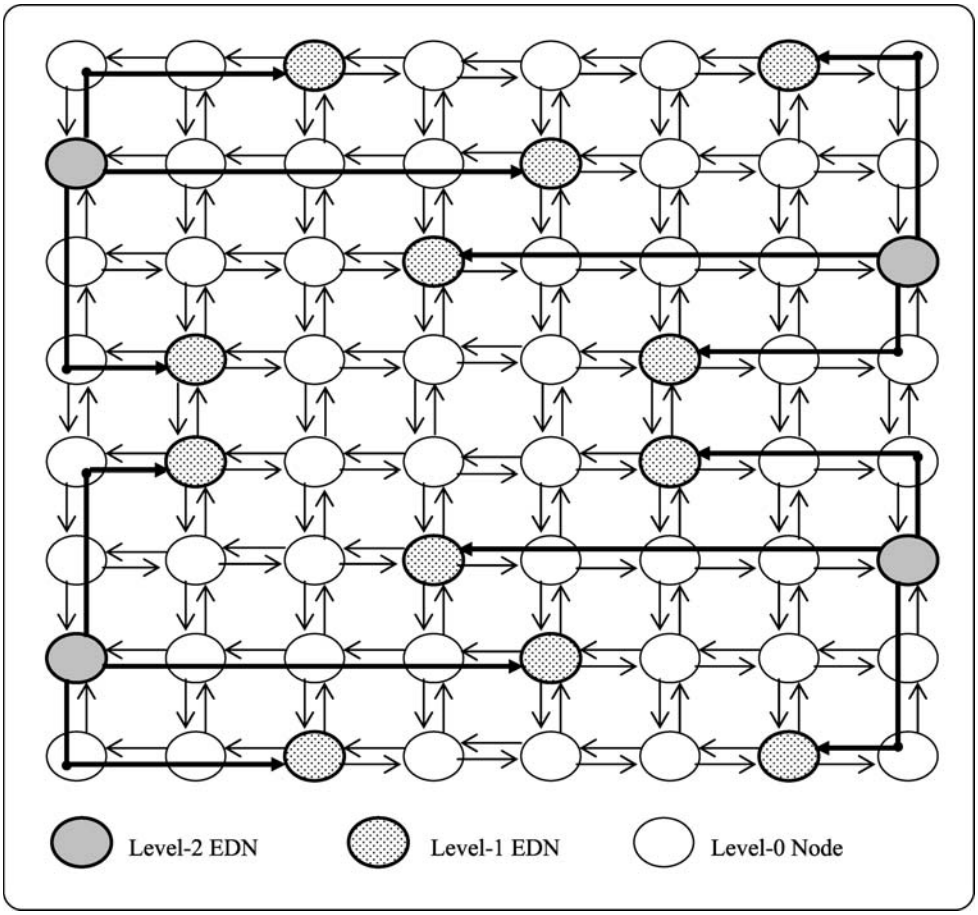
2.3 OTIS interconnection networks

The Optical Transpose Interconnection System (OTIS) was introduced by Marsden et al. [5] as an interconnection system that combines the advantages of electronic and optical interconnection links. Processors in OTIS are divided into groups, where intra-group (short-distance) communication is realized by electronic links and the longer inter-group communication is achieved through optical links.

In general, an OTIS is divided into N groups, each of which consists of N processors. A processor within an OTIS group is modeled as a tuple (G, P), where G denotes the group's number and P is the processor's number within the group. Processor (G, P) is connected directly to its transpose processor (P, G) via optical interconnection. Intra-group processors, on the other hand, are connected electronically forming a common interconnection network's topology, such as mesh, hypercube, star, mesh of trees etc.

OTIS-Mesh, under study, consists of N groups with N processors in each group organized as a two-dimensional $\sqrt{N} \times \sqrt{N}$ mesh. Figure 6 illustrates a 4 × 4 OTIS-Mesh, which consists of 4 groups (named Group 0, Group 1, Group 2 and Group 3), where each group consists of 4 processors interconnected as a mesh. Each processor is in-

Fig. 5 Extended dominating nodes in an 8 × 8 2D-Mesh



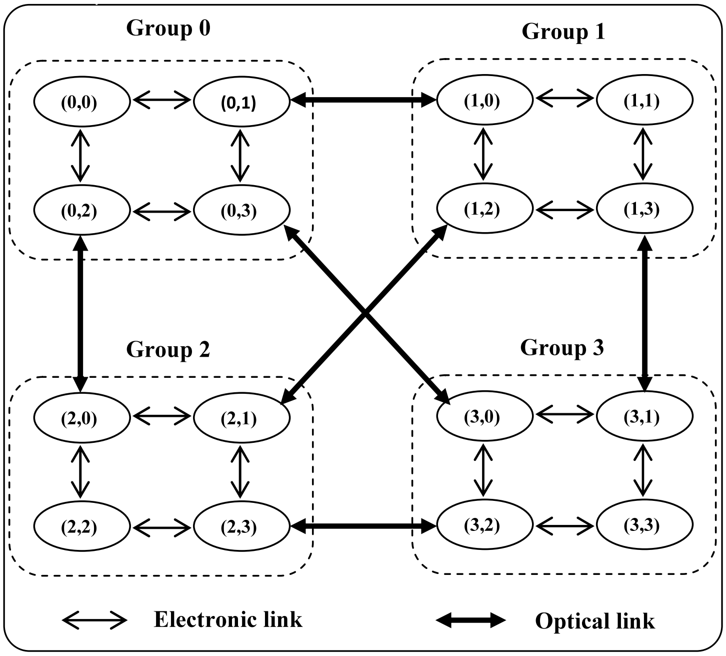


Fig. 6 4 × 4 OTIS-Mesh

terconnected with its transpose processor via an optical link (the bold arrows); for example, processor $(0, 3)$ shares an optical link with processor $(3, 0)$.

Several attractive features can be achieved by the OTIS architecture, especially when OTIS is divided into N groups, with N processors in each group. For example, it was verified by Krishnamoorthy et al. [19] that the OTIS bandwidth is maximized and the power consumption is minimized when the number of groups is equal to the number of processors within each group. Several research efforts have achieved significant performance optimization in OTIS through load-balancing and adaptive routing algorithms [11, 20–23]. For example, a load balancing algorithm called Clusters Dimension Exchange Method (CDEM) [11] was developed for OTIS-Hypercube and was further evaluated under the execution time, load balancing accuracy, number of communication steps, and speed metrics. The experimental work strongly indicates the outperforming results achieved by OTIS-Hypercube over the hypercube interconnection network. An adaptive routing algorithm was implemented in [24] for wormhole switched OTIS-Hypercube. The proposed algorithm was examined through an empirical performance evaluation study under a number of conditions, such as traffic load, router delay etc. Moreover, the performance of deterministic routing in OTIS-Hypercube and OTIS-Mesh were evaluated in [20] under different conditions such as the number of virtual channels, traffic, bisection width, cycle ratio etc.

2.4 Basic operations in OTIS-Networks

Several basic operations and topological properties were developed in OTIS-Networks [9, 25–28], including optimal routing, broadcast, data sum, size, degree, and diameter. Day [27] presented one-to-one routing and optimal broadcasting algorithms on OTIS k-ary n-cube interconnection networks, besides the embedding of OTIS k-ary $(n-1)$-cubes, cycles, meshes, cubes, and spanning trees. Topological properties were presented on the OTIS k-ary n-cube interconnection networks; such as size, degree, short distance, and diameter. Day and Al-Ayyoub [25] presented minimal one-to-one routing and optimal broadcasting algorithms in OTIS-Networks, where the optimal routing guarantees finding the minimum distance path and the broadcast algorithm employs a spanning tree. Also, some topological properties of OTIS-Networks were presented; including size, degree, shortest distance, and diameter. Moreover, Wei and Xiao [26] developed basic communication operations such as: broadcast, prefix sum and data sum on the Swapped Network; an optoelectronic interconnection network that resembles OTIS with as number of processors as in OTIS.

Wang and Sahni [28] presented some basic operations on OTIS-Mesh in particular; such as window broadcast, data sum, data accumulation, adjacent sum, and random accesses read and write. The window broadcast operation in OTIS-Mesh starts with a data located in a sub-mesh of one group, which is then transmitted to processors within a group, and finally to other groups. The data sum operation is performed on data items located in each processor in OTIS-Mesh. In data accumulation operation, each processor accumulates data from its neighboring processors. The adjacent sum operation performs sum operation on the accumulated data. The random access read operation reads data from one processor in OTIS-Mesh, while the random access write operation writes data onto another processor in OTIS-Mesh.

Table 1 Algorithm for broadcast in single- and all-port wormhole-routed OTIS-Mesh

Step 1: The root processor sends *m* broadcasted messages to each processor in the OTIS-Mesh group.
Step 2: Each processor in the root group performs an optical move to transmit the received message to the transposed group.
Step 3: When the processor in the transposed group receives the broadcasted message via its optical link, Step 1 is repeated.

Table 2 Algorithm for global combine in single- and all-port wormhole-routed OTIS-Mesh

Step 1: Each processor within each group in the OTIS-Mesh, sends the message (message need to be collected) to the processor that shares an optical link with the root group (the group where the root processor exists).
Step 2: An optical move is performed to send the collected messages within each group to the root group.
Step 3: Each processor in the root group sends the received messages to the root processor.

Table 3 Algorithm for broadcast in EDN-OTIS-Mesh

Step 1: The root processor performs broadcast on the root group as follows:
  a. The root processor sends the message to the processors located at the highest EDN level in the control group.
  b. Processors in the highest EDN level send the message to the processors located in the lowest EDN levels until the message is received by EDN level-1.
  c. Processors in EDN level-1 send the message to level-0 processors.
Step 2: Each processor in the root group performs an optical move to transmit the received message to the transposed group.
Step 3: When the processor in the transposed group receives the broadcasted message via its optical link, Step 1 is repeated.

The previous operations are important in developing many applications on OTIS-Mesh, such as graph-based and matrix multiplication.

McKinley, Tsai, and Robinson [9] presented broadcast and global combine algorithms in single- and all-port wormhole-routed massively parallel computers. In our paper, these algorithms (Tables 1 and 2) have been applied on single- and all-port wormhole-routed OTIS-Mesh, in order to be compared analytically and by simulation, with the broadcast and global combine algorithms presented in Sect. 3 (Tables 3 and 4) in all-port wormhole-routed OTIS-Mesh using EDN approach, referred to EDN-OTIS-Mesh.

Besides the above basic operations on OTIS-Networks, a number of other operations were developed [29], such as the selection operation that finds the smallest number in a sequence of numbers distributed among processors in the network, and the sorting operation that arranges a set of data distributed in ascending or descending order among the network's processors. Both operations have vital use in numerical analysis applications.

Table 4 Algorithm for global combine in EDN-OTIS-Mesh

Step 1: Each group perform global combine as follows:
  a. Within each group in the EDN-OTIS-Mesh, level-1 EDN processors collect messages from level-0 processors.
  b. Level-1 EDN processors send the collected messages to Level-2 EDN processors until the collected messages are received by the highest EDN level.
Step 2: An optical move is performed to send the collected messages within each group to the root group.
Step 3: When the transposed processors in the root group receive the collected messages Step 1 is repeated in the root group.

3 Extended dominating node approach for broadcast and global combine operations on OTIS-Mesh

This section presents efficient broadcast and global combine algorithms (Tables 3 and 4) on all-port wormhole-routed Optical Transpose Interconnection System Mesh (OTIS-Mesh) using Extended Dominating Node (EDN) approach [6–8] referred to EDN-OTIS-Mesh. However, the broadcast and global combine in the EDN-OTIS-Mesh only can be applied on the all-port model because the EDN approach can be applied only on that model [6–8]. The proposed broadcast and global combine communication operations are presented in Sects. 3.1 and 3.2, respectively.

OTIS-Mesh is organized as an interconnection of N groups, each of N Processing Elements (PEs) interconnected as a mesh, where each PE is presented as a two-element tuple (G, P); where G is the group number and P is the processor number within G. For each PE (G, P), the corresponding transpose PE is denoted by (P, G). In OTIS-Mesh, each PE is connected to its corresponding transpose via an optical link. Furthermore, on each group of OTIS-Mesh, the EDN approach is applied forming an EDN-OTIS-Mesh.

A 16×16 EDN-OTIS-Mesh is shown in Fig. 7; it consists of 16 groups $(G0, G1, \dots, G15)$, each of which contains 16 processors $(0, 1, \dots, 15)$ interconnected in the form of mesh. As shown by the figure, each processor is interconnected with its transpose processor via an optical link; for example, processor $(0, 2)$ (processor 2 in group 0) shares an optical link with processor $(2, 0)$ (processor 0 in group 2). The shaded processors in each group present the EDN processors, where a set of nodes can deliver a message to all other processors in a group in a single message-passing step. For example, within group $G0$, processors 1, 7, 8 and 14 are called level-1 EDN processors, where they can deliver a message to level-0 nodes in a single message-passing step, while level-0 nodes form the rest of the nodes.

3.1 Broadcast operation

In broadcast operation, one processor sends the same message to each processor in the EDN- OTIS-Mesh. The broad-

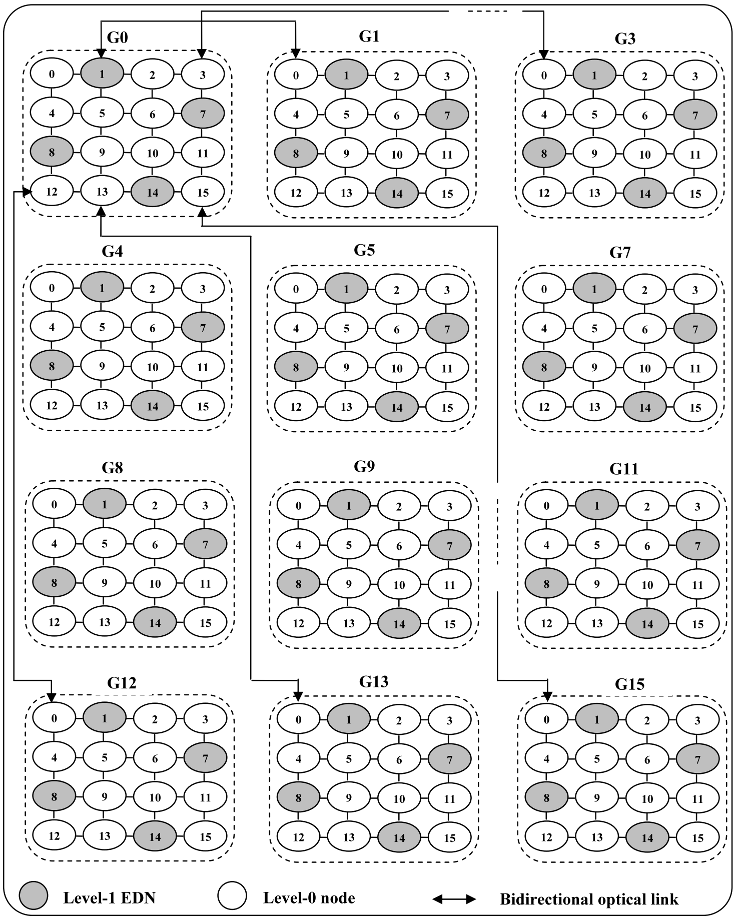


Fig. 7 16 × 16 EDN-OTIS-Mesh

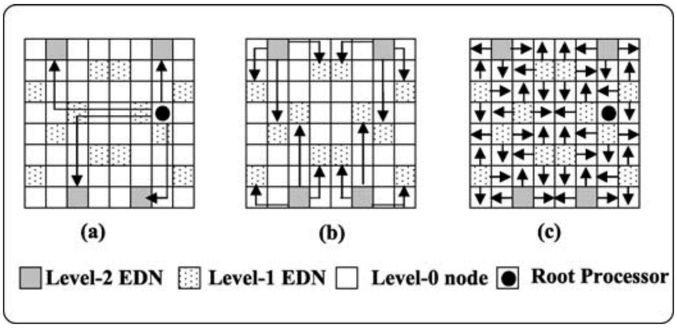


Fig. 8 Step 1—broadcast operation in 64 × 64 EDN-OTIS-Mesh control group

cast operation in EDN-OTIS-Mesh can be summarized by the following steps:

Step 1: the root processor sends the message to the processors located at the highest EDN level in the control group. Figure 8 illustrates the control group ($G0$ as shown in Fig. 9) of a 64 × 64 EDN-OTIS-Mesh, in part (a) of Fig. 8; the root processor (the square with a black circle inside) sends the message to level-2 EDN processors (gray-colored squares). Next, as shown in part (b), the level-2 EDN processors send the messages to level-1 EDN processors (dot-shaded squares). Then, as part (c) shows, level-1 EDN processors send the received message to level-0 nodes (white squares).

Step 2: processors at level-0 in the control group ($G0$ as shown in Fig. 9) perform an optical movement in order to send the message to each group in the EDN-OTIS-Mesh. Figure 9 illustrates a 64 × 64 EDN-OTIS-Mesh, in which processor 0 of each group receives the message through the optical link and sends it to level-2 EDN processors.

Step 3: within each group, the processors in the highest EDN level send the received message to the next lower EDN processors. Figure 10 illustrates the broadcast operation in a 64 × 64 EDN-OTIS-Mesh,

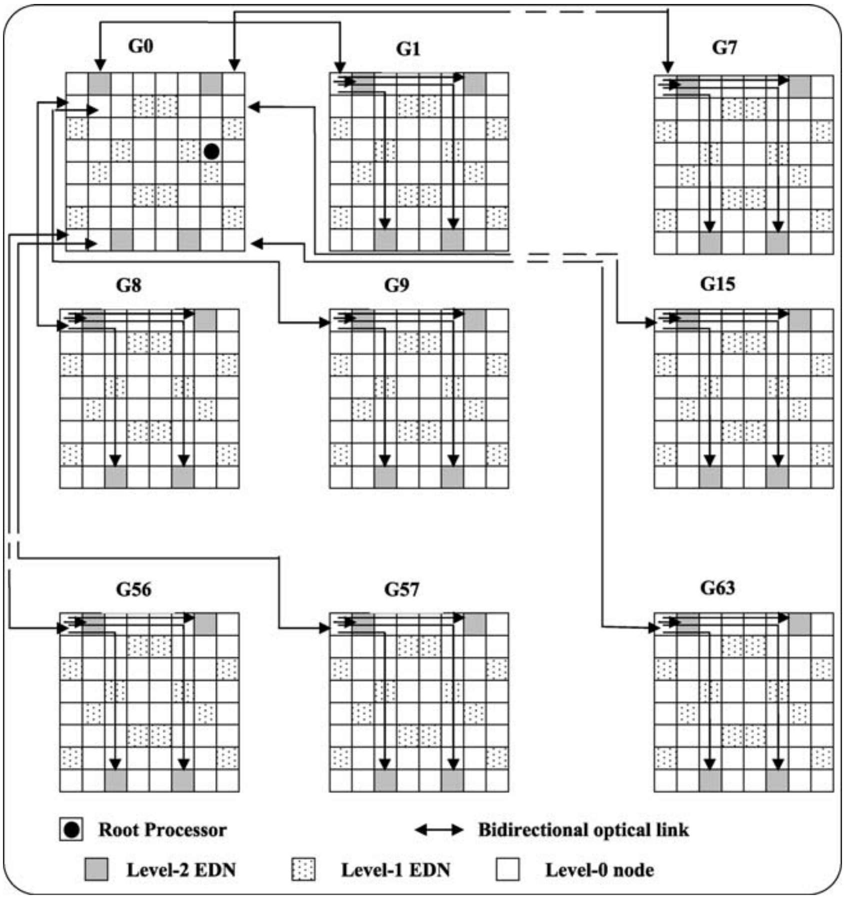


Fig. 9 Step 2—broadcast operation in 64 × 64 EDN-OTIS-Mesh

where within each group ($G1$ to $G63$) level-2 EDN processors (gray-filled squares) send the message to level-1 EDN processors (dot-shaded squares).

Step 4: as illustrated in Fig. 11, the broadcast operation in a 64 × 64 EDN-OTIS-Mesh is finalized by having level-1 EDN processors (dot-shaded squares) sending messages to level-0 nodes in parallel.

As a summary, Table 3 presents the above steps as an algorithm for broadcast communication operation in all-port wormhole-routed EDN-OTIS-Mesh.

3.2 Global combine operation

In a global combine operation, one processor (root processor) receives a different message from each processor in the EDN-OTIS-Mesh. The global combine operation in EDN-OTIS-Mesh can be summarized in the following phases:

Global combine phase: every group within the EDN-OTIS-Mesh starts to collect messages as follows:

- Within each group in the EDN-OTIS-Mesh, level-1 EDN processors collect messages from level-0 processors and send those messages to the next EDN level, until the messages reach the highest EDN level. Figure 12 illustrates the global combine phase in one group of a 64 × 64 EDN-OTIS-Mesh. As shown by part (a), level-0 nodes (white squares) send their messages to level-1 EDN processors (dot-shaded squares). In part (b), level-1 EDN processors from the previous 64 × 64 EDN-OTIS-Mesh group send their messages to level-2 EDN processors (gray-filled squares).
- The processors located at the highest EDN level send the collected messages to the processor that shares an optical link with the control group (the group where the root processor is located).

Completion phase: each processor in the control group receives a group of messages via its optical links from its transpose processors. The global combine phase will be performed again in the control group ($G0$ as shown in Fig. 13) until the root processor receives all the messages. Figure 13 illustrates the completion phase in a 64 × 64 EDN-OTIS-Mesh. The 64 × 64 EDN-OTIS-Mesh consists of 64 groups $(0, 1, \dots, 63)$ and each group consists of 64 processors; the white squares in the figure represent level-0 nodes, the dot-shaded squares are level-1 EDN processors and the gray-filled squares stand for level-2 EDN processors. The control

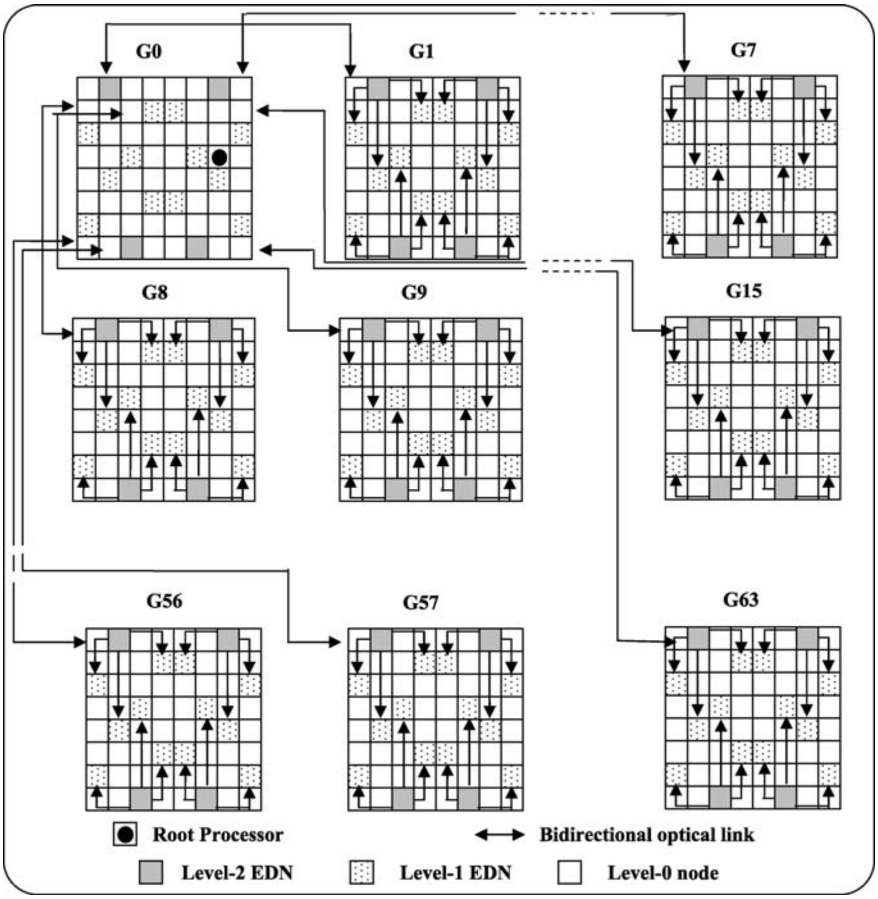


Fig. 10 Step 3—broadcast operation in 64 × 64 EDN-OTIS-Mesh

group is $G0$, in which the root processor is represented by the square with the dark circle inside. After performing the global combine operation in the control group, level-2 EDN processors send their collected messages to the root processor.

As a summary, Table 4 presents the above phases as an algorithm for global combine communication operation in all-port wormhole-routed EDN-OTIS-Mesh.

4 Analytical evaluation

This section provides an analytical evaluation of the broadcast and global combine communication operations in the following interconnection networks: single-port wormhole-routed OTIS-Mesh, all-port wormhole-routed OTIS-Mesh, and all-port wormhole-routed EDN-OTIS-Mesh. The analysis is performed in terms of the following performance metrics: number of communication steps, latency, and latency improvement. This analytical evaluation will further serve as a foundation for implementing the interconnection network's simulator developed for performance evaluation purposes.

4.1 Number of communication steps

The total number of communication steps is the sum of optical and electronic communication steps required to perform the broadcast or global combine operations. The required number of optical message-passing steps between the groups of OTIS-Mesh is one for the three interconnection networks architectures. For this reason, we consider the number of communication steps metric is the sum of the electronic message-passing steps needed to perform the broadcast operation. This also applies to the global combine operation. In all-port wormhole-routed OTIS-Mesh and all-port wormhole-routed EDN-OTIS-Mesh, this metric is affected by the root node's location. Hence, we calculate the minimum (best-case) and maximum (worst-case) number of communication steps. The minimum number of communication steps is achieved when the root node is located at the middle of the control group (the group that contains the root processor), which enables the root node to perform simultaneous send/receive from all of its input/output channels. Figure 14 presents one group of a 64 × 64 OTIS-Mesh, where the shaded processor is the root, whose location is at the middle of the group; thus presenting the best-case. On the other hand, the maximum number of communication steps occurs when the root node is located at the endmost of the

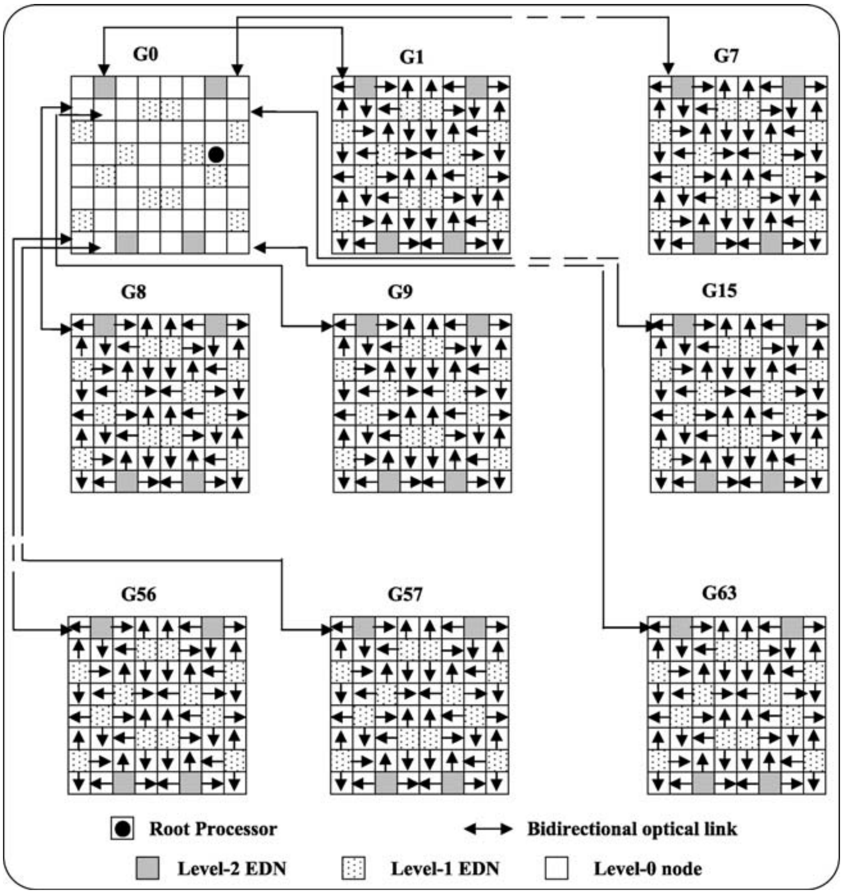


Fig. 11 Step 4—broadcast operation in 64 × 64 EDN-OTIS-Mesh

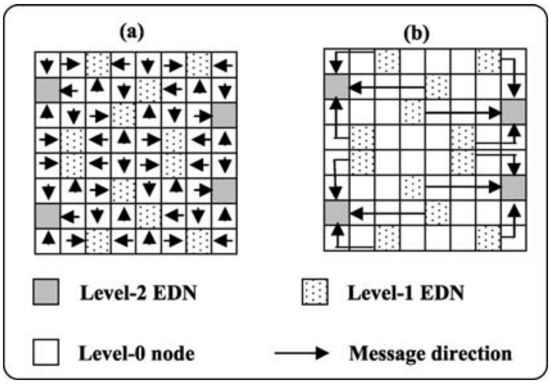


Fig. 12 Global combine phase in one group of a 64 × 64 EDN-OTIS-Mesh

control group. This leads to a limited number of simultaneous send/receive by the root node from all of its input/output channels. This scenario is shown in Fig. 15, representing one group of the 64 × 64 OTIS-Mesh, where the shaded processor is the root, which is located at the endmost of the group, thus presenting the worst-case.

In order to analytically evaluate the single-port OTIS-Mesh, all-port OTIS-Mesh, and all-port EDN-OTIS-Mesh in terms of the number of electronic message-passing steps, the following assumptions hold: the *routing algorithm* used is deterministic, and the dimension order routing traverses the X dimension first then the Y dimension in OTIS-Mesh, where wormhole-routed is used. Moreover, the *number of nodes* in an OTIS-Mesh group is N, as shown in (1), where $i = 2, 3, \ldots, n$ and 4 is the number of EDN nodes in the smallest OTIS-Mesh group. However, i starts from 2 in order to have at least a group with $N = 16$ and one EDN level of 4 nodes.

$$N = 4^i \tag{1}$$

The *height of EDN tree* (i.e. the number of EDN levels) within each EDN-OTIS-Mesh group is equal to L, as shown in (2), where N is the number of nodes in the OTIS-Mesh group and 4 is the number of EDN nodes in the smallest OTIS-Mesh group.

$$L = (\log_4 N) - 1 \tag{2}$$

Next, Theorems 1 to 5 show the number of electronic message-passing steps to perform a broadcast or a global combine operation in single-port wormhole-routed OTIS-Mesh, the minimum and maximum number of the electronic

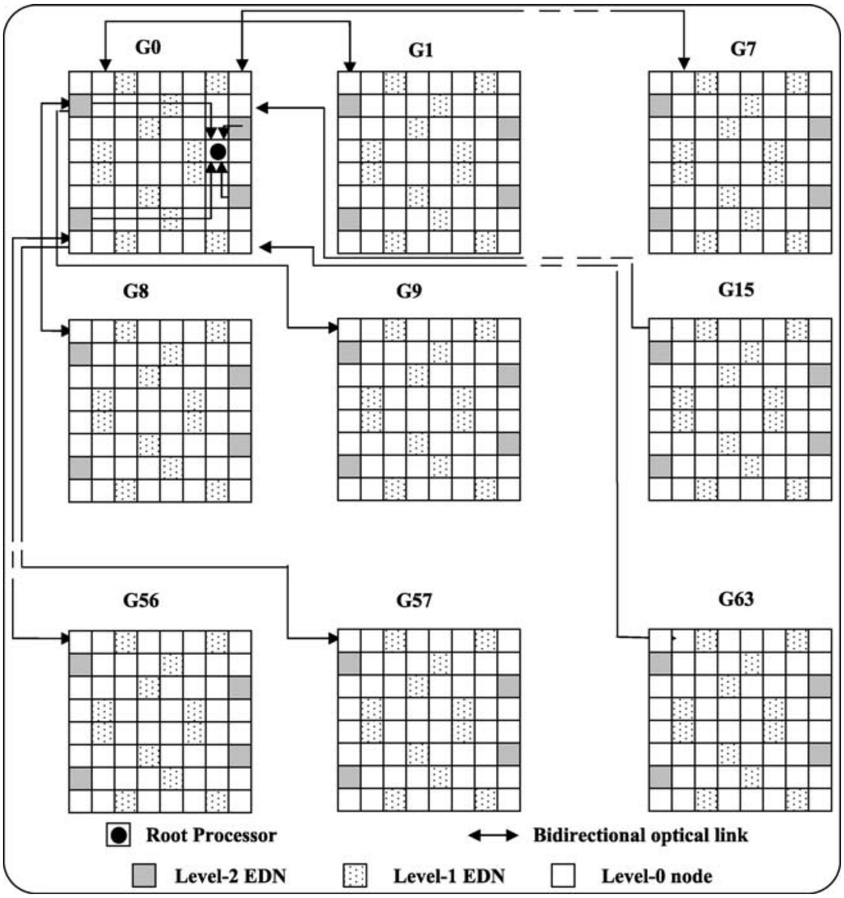


Fig. 13 Completion phase in 64×64 EDN-OTIS-Mesh

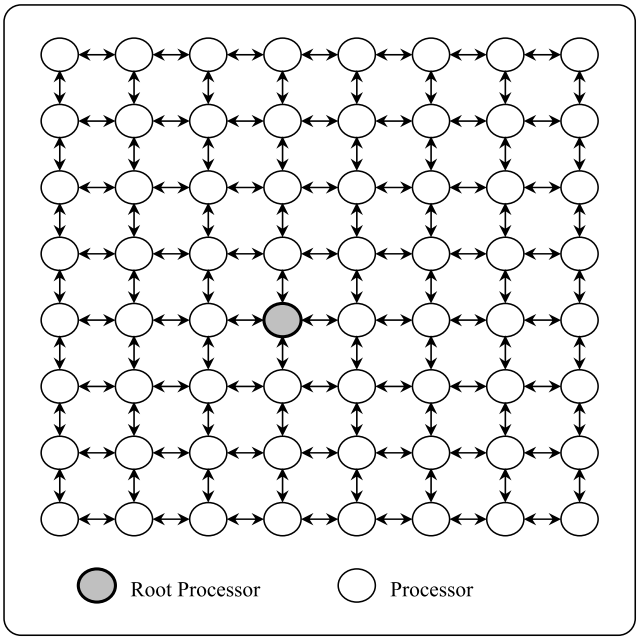


Fig. 14 64×64 OTIS-Mesh control group (best-case)

message-passing steps required to perform a broadcast or a global combine operation in both the all-port wormhole-routed OTIS-Mesh and the all-port wormhole-routed EDN-OTIS-Mesh, respectively.

Theorem 1 *It takes* $2 \times (N - 1)$ *electronic message-passing steps to perform a broadcast or a global combine operation in a single-port wormhole-routed OTIS-Mesh.*

Proof The number of communication steps required to perform a broadcast operation (Table 1) in the control group of the single-port wormhole-routed OTIS-Mesh is equal to $(N - 1)$ because the root processor is a single-port and is able to send one message at a time to each processor in the control group, which contains N processors. In addition, parallel broadcasting on other groups of the single-port OTIS-Mesh requires $(N - 1)$ steps, where each group contains N processors. Therefore, the required number of electronic message-passing steps to perform a broadcast operation in the single-port wormhole-routed OTIS-Mesh is equal to $2 \times (N - 1)$. Regarding the global combine operation (Table 2), the number of required communication steps to perform the global combine phase in a single-port OTIS-Mesh is $N - 1$ because the interconnection network is a single-port, where the processor responsible for the global combine phase is capable to receive one message at a time from each of the N-processors groups. The completion phase of the

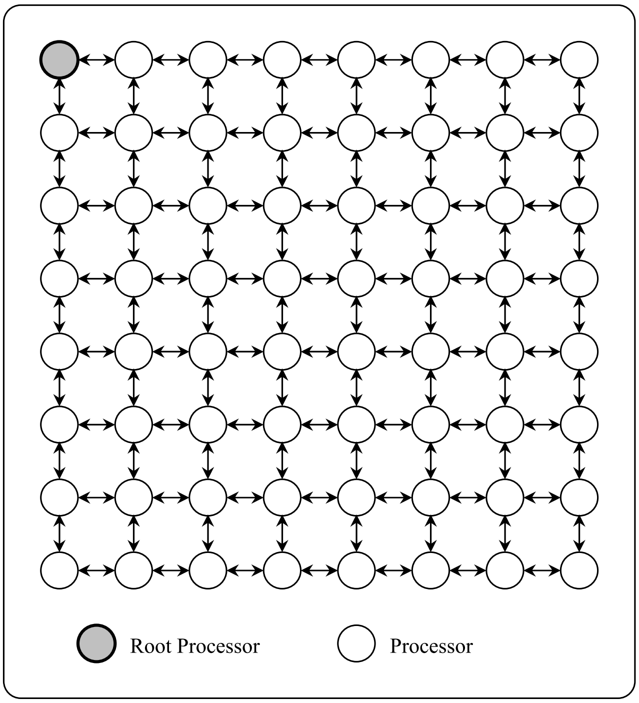


Fig. 15 64×64 OTIS-Mesh control group (worst-case)

global combine operation in the control group of the single-port OTIS-Mesh also takes $(N-1)$ steps, as the root processor in the control group will receive one message from each processor at a time. Therefore, the total number of electronic message-passing steps to perform a global combine operation in the single-port wormhole-routed OTIS-Mesh is $2 \times (N-1)$. □

Theorem 2 *The minimum number of the electronic message-passing steps required to perform a broadcast or a global combine operation in all-port wormhole-routed OTIS-Mesh is* $2 \times ((\sqrt{N}/2) \times \sqrt{N}))$.

Proof The minimum number of the required communication steps to perform a broadcast operation (Table 1) in the control group of the all-port wormhole-routed OTIS-Mesh is equal to $(\sqrt{N}/2) \times \sqrt{N}$, where the best-case scenario is obtained when the root processor is located at the middle of the control group (Fig. 14). Since the interconnection network is all-port, the root processor has the ability to send messages in parallel based on the number of its communication channels. The number of messages approximately reaches $(\sqrt{N}/2) \times \sqrt{N}$ since the control group's N-processors are organized in the form of a mesh, consisting of $\sqrt{N}$ rows, each of $\sqrt{N}$ processors, where the root processor is located at the middle of the group. In addition, the routing algorithm is deterministic (i.e. the message between two nodes always selects the same previously determined path), so half of the messages will be sent through one port and the rest of the messages will be sent through the other ports. Parallel broadcasting on other groups also takes $(\sqrt{N}/2) \times \sqrt{N}$ steps. Therefore, the minimum number of electronic message-passing steps to perform broadcasting in all-port wormhole-routed OTIS-Mesh is equal to $2 \times ((\sqrt{N}/2) \times \sqrt{N})$. With regard to the global combine operation (Table 2), both of its phases consume a number of communication steps. The global combine phase, at first, requires $(\sqrt{N}/2) \times \sqrt{N}$ steps because the interconnection network is all-port, in which the processor conducting the global combine phase is capable to receive, in parallel, a number of messages equal to the number of its communication channels, and since a deterministic routing algorithm is used, half of the messages which is equal to $(\sqrt{N}/2) \times \sqrt{N}$ will be received through one port and the rest of the messages will be received through the other ports. The completion phase of the global combine operation in the control group of the all-port OTIS-Mesh also takes $(\sqrt{N}/2) \times \sqrt{N}$ steps, as the root processor in the control group will receive one message from each processor in the group and half of the messages $(\sqrt{N}/2) \times \sqrt{N}$ will be received by one port and the rest of the messages will be received by the other ports. Therefore, the minimum number of communication steps to perform global combine operation in all-port wormhole-routed OTIS-Mesh is $2 \times ((\sqrt{N}/2) \times \sqrt{N})$. □

Theorem 3 *The maximum number of the electronic message-passing steps required to perform a broadcast or a global combine operation in all-port wormhole-routed OTIS-Mesh is* $2 \times ((\sqrt{N}-1) \times \sqrt{N})$.

Proof The maximum number of the required message-passing steps to perform a broadcast operation (Table 1) in the control group of the all-port wormhole-routed OTIS-Mesh is equal to $(\sqrt{N}-1) \times \sqrt{N}$, where the worst-case scenario is obtained when the root processor is located at the endmost of the control group (Fig. 15). Since the interconnection network is all-port, the root processor is capable to send messages in parallel according to the number of its communication channels. However, the location of the root processor limits the number of parallel send operations, and the used routing algorithm is deterministic, so $(\sqrt{N}-1) \times \sqrt{N}$ messages will be sent through one port since the control group's N-processors are organized in a mesh structure, consisting of $\sqrt{N}$ rows, each containing $\sqrt{N}$ processors, where the root processor is located at the endmost of the group. In addition, parallel broadcasting on other groups also takes $(\sqrt{N}-1) \times \sqrt{N}$ steps. Thus, the maximum number of electronic message-passing steps to perform a broadcast operation in all-port wormhole-routed OTIS-Mesh is equal to $2 \times ((\sqrt{N}-1) \times \sqrt{N})$. With respect to the global combine operation (Table 2) in the all-port wormhole-routed OTIS-Mesh, the maximum number of communication steps taken to perform the global combine phase is equal to $(\sqrt{N}-1)\times\sqrt{N}$ because the interconnection network is all-port, where the processor that performs the global combine phase is capable to receive, in parallel, a number of messages equal to the number of communication channels it has. Since the location of the root processor limits the number of parallel receive messages and the routing algorithm is deterministic, $(\sqrt{N}-1)\times\sqrt{N}$ messages will be received by one port and the rest of the messages will be received through the rest of ports. As well as, the completion phase of the global combine operation in the all-port OTIS-Mesh takes $(\sqrt{N}-1)\times\sqrt{N}$ steps. Therefore, the maximum number of electronic message-passing steps for global combine in all-port wormhole-routed OTIS-Mesh is equal to $2\times((\sqrt{N}-1)\times\sqrt{N})$. □

Theorem 4 *The minimum number of electronic message-passing steps to perform a broadcast or a global combine operation in all-port wormhole-routed EDN-OTIS-Mesh is* $2\times(L+2)$.

Proof The minimum number of the required electronic message-passing steps to perform a broadcast operation (Table 3) in the control group of the all-port wormhole-routed EDN-OTIS-Mesh is equal to $L+2$. Since the interconnection network is all-port architecture, and the best case scenario is obtained when the root processor is located at the middle of the control group (Fig. 14), the root processor sends parallel messages relevant to the number of its communication channels. It takes L steps, which is the number of EDN levels in this group passed during sending the message from the highest EDN level to the lowest EDN level, and it takes up to 2 steps to complete the broadcast operation in this group, which is sending the message from the root processor to the highest EDN level. In addition, performing broadcasting on other groups, in parallel, also takes $L+2$ steps. Therefore, the minimum number of electronic message-passing steps to perform a broadcast operation in all-port wormhole-routed EDN-OTIS-Mesh is equal to $2\times(L+2)$. As for the global combine operation (Table 4), the minimum number of communication steps taken to perform the global combine phase in all-port wormhole-routed EDN-OTIS-Mesh is equal to $L+2$. This refers to the fact that the interconnection network is all-port, in which the processor that is responsible for the global combine phase receives a number of messages in parallel that is related to the number of its communication channels. So, it requires L steps, which is the number of EDN levels in the group to receive messages from the lowest EDN levels up to the highest EDN levels, and it takes up to 2 steps to complete the global combine phase, which is sending the message from the highest EDN level to the processor that performs the global combine phase in all groups. The completion phase adds $L+2$ more steps in order to receive the messages from level-0 nodes to the highest EDN level, recall that it takes L steps to receive the messages from the lowest EDN levels to the highest EDN levels and up to 2 steps for the completion phase. This results in a minimum number of electronic message-passing steps to perform global combine operation in all-port wormhole-routed EDN-OTIS-Mesh equal to $2\times(L+2)$. □

Theorem 5 *The maximum number of electronic message-passing steps to perform a broadcast or a global combine operation in all-port wormhole-routed EDN-OTIS-Mesh is* $2\times(L+3)$.

Proof At most, $L+3$ electronic message-passing steps are required to perform broadcasting (Table 3) in the control group of all-port wormhole-routed EDN-OTIS-Mesh. The interconnection network's architecture is all-port, and the worst-case scenario is achieved when the root processor is positioned at the endmost of the control group (Fig. 15). Thus, the number of parallel messages sent by the root processor is relevant to the number of the root's communication channels. In order to send the message from the highest EDN level to the lowest EDN level, L steps are required, which is the number of EDN levels in this group, and up to 3 steps are necessary to complete the broadcast operation in this group, which is sending the message from the root processor to the highest EDN level. Furthermore, $L+3$ more steps are required for parallel broadcasting on other groups of all-port EDN-OTIS-Mesh. As a result, the maximum number of electronic message-passing steps to perform broadcasting in all-port wormhole-routed EDN-OTIS-Mesh is equal to $2\times(L+3)$. As to the global combine operation (Table 4) in the all-port wormhole-routed EDN-OTIS-Mesh, the maximum number of required electronic message-passing steps to perform the global combine phase is $L+3$. This can be clarified by referring to the interconnection network's all-port architecture, where the processor that performs the global combine phase can receive a number of parallel messages related to the number of its own communication channels. Based on this, L steps, which is the number of EDN levels in the group, are required to receive messages from the lowest EDN levels up to the highest EDN levels and it takes up to 3 steps to complete the global combine phase, which is sending the message from the highest EDN level to the processor that performs the global combine phase in all groups. By the completion phase of the global combine operation in the control group, $L+3$ additional steps are necessary to receive the messages from level-0 nodes to the highest EDN level, recall that it takes L steps to receive the messages from the lowest EDN levels to the highest EDN levels and up to 3 steps for the completion phase. Therefore, the maximum number of electronic message-passing steps to perform a global combine operation in all-port wormhole-routed EDN-OTIS-Mesh is equal to $2 \times (L + 3)$. □

4.2 Latency

In order to analytically define the latency of the single-port wormhole-routed OTIS-Mesh, all-port wormhole-routed OTIS-Mesh, and all-port wormhole-routed EDN-OTIS-Mesh, the following assumptions are taken into account: *message length* is M_{len}, setup time T_{set} is the time needed to start an operation, *router delay* is characterized by the following three parameters: *channel cycle time* T_{cc} which is the time needed to transmit one flit from one router to the next, *switching delay* T_{sw} which is the time needed to transmit a flit from the input channel to the required output channel in the crossbar, and *routing time* T_r which is the time needed to route a message. So, the overall *router delay* T_{rd} (3) is the maximum of (T_{cc}, T_{sw}, T_r) since the three operations are overlapped.

$$T_{rd} = \mathrm{Max}(T_{cc}, T_{sw}, T_r) \tag{3}$$

Moreover, *contention time* T_{cont} is the time spent by a flit in the router before it is transmitted to the next router due to channel contention, *communication latency* $T_{latency}$ (4) is the required time to transmit a message from the source node to the destination node, $T_{latency}$ is affected by the number of hops H between the source and destination, message size M_{len}, contention time T_{cont}, and router delay T_{rd}.

$$T_{latency} = T_{rd} \times (H + M_{len}) + T_{cont} \tag{4}$$

Based on the above, (5) gives the *Latency* to perform broadcast or global combine operations in single-port wormhole-routed OTIS-Mesh, all-port wormhole-routed OTIS-Mesh, or all-port wormhole-routed EDN-OTIS-Mesh.

$$Latency = T_{set} + T_{latency} \tag{5}$$

The *Latency* of either, a broadcast or a global combine operation, which has no computation phase, is the time elapses from the beginning of the communication until the moment the last processor finishes communication, which is equal to the sum of the setup time T_{set} and the communication latency $T_{latency}$ for the last processor finishing communication.

4.3 Latency improvement

The latency improvement (i.e., reduction in latency) ($Late.Imp._{over\_single}$) achieved by broadcast or global combine communication operations on all-port wormhole-routed EDN-OTIS-Mesh to single-port wormhole-routed OTIS-Mesh is presented in (6), where the latency of performing broadcast or global combine in single-port wormhole-routed OTIS-Mesh is $Latency_{single\_port}$ and the latency of performing the previous operations on all-port wormhole-routed EDN-OTIS-Mesh is $Latency_{all\_port\_EDN}$.

$$Late.Imp._{over\_single} = Latency_{single\_port}/Latency_{all\_port\_EDN} \tag{6}$$

Moreover, the latency improvement ($Late.Imp._{over\_all\_port}$) achieved by broadcast or global combine communication operations on all-port wormhole-routed EDN-OTIS-Mesh to all-port wormhole-routed OTIS-Mesh is presented in (7), where the latency of performing broadcast or global combine in all-port wormhole-routed OTIS-Mesh is $Latency_{all\_port}$ and the latency of performing the previous operations on all-port wormhole-routed EDN-OTIS-Mesh is $Latency_{all\_port\_EDN}$.

$$Late.Imp._{over\_all\_port} = Latency_{all\_port}/Latency_{all\_port\_EDN} \tag{7}$$

5 Simulation environment

This section presents the technical specifications of the simulation environment, under which, the simulation runs were conducted. These specifications include both the hardware and software modules used in the implementation of the interconnection network's simulator. Prior to implementing the simulator, an analytical study was conducted in order to quantify the simulation's parameters. These parameters include: port model, network switching method, and so forth. According to these parameters, suitable development environment and tools were chosen.

The performance evaluation of the simulation runs were conducted on a Dual-Core Intel Processor (CPU 1.5 GHz), with 14 pipeline stages and a multithreaded architecture, 2 MB L2 Cache per CPU, and 2 GB RAM. The simulation was developed using C++ programming language, within the Microsoft Visual Studio 6.0 programming environment. The simulation runs were performed under Windows Vista operating system.

Under the hardware and software specifications introduced above, the special-purpose simulator named OTIS-Mesh was designed and further implemented. OTIS-Mesh is a discrete event simulator, which mainly captures the characteristics of three interconnection networks: single-port wormhole-routed OTIS-Mesh, all-port wormhole-routed OTIS-Mesh, and all-port wormhole-routed EDN-OTIS-Mesh. Such capability supports both the measurements and the comparisons of both broadcast and global combine communication operations performance metrics under the previous three interconnection networks.

The OTIS-Mesh simulator consists of the following modules: node (i.e. processor or Processing Element (PE)),

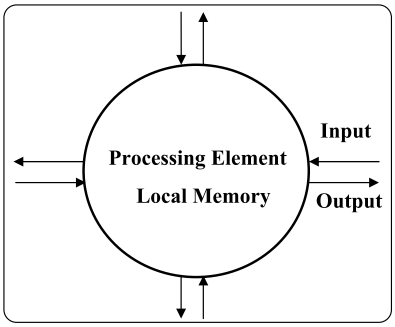


Fig. 16 Processing element

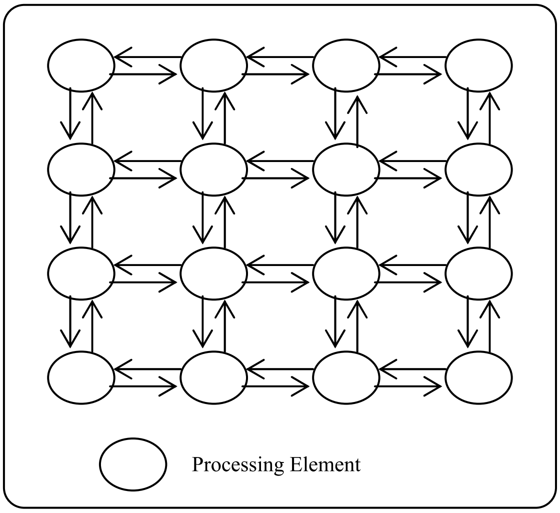


Fig. 17 4 × 42D Mesh

mesh, OTIS-Mesh, and EDN-OTIS-Mesh. The main characteristics of the PE module are depicted from the node hardware specifications mentioned previously. As illustrated in Fig. 16, these characteristics include: local memory, input channels, and output channels. The mesh module presents the interconnected PEs within a single OTIS-Mesh group, in which N PEs are interconnected, as illustrated in Fig. 17. The OTIS-Mesh module simulates the overall OTIS-Mesh interconnection network consisting of N groups, each of which contains N PEs interconnected in a mesh structure. Figure 18 illustrates a 16 × 16 OTIS-Mesh, which consists of 16 groups ($G0, G1, \ldots, G15$), where each group consists of 16 mesh-interconnected processors $(0, 1, \ldots, 15)$. The groups are interconnected based on the topological specifications of OTIS-Mesh. For instance, processor $(0, 2)$ (processor 2 of group 0) shares an optical link with its corresponding transpose processor $(2, 0)$ (processor 0 of group 2). Finally, the EDN-OTIS-Mesh module is an OTIS-Mesh interconnection network that uses EDN approach for communication operations purposes. Figure 19 presents the 16 × 16 EDN-OTIS-Mesh, which it consists of 16 groups ($G0, G1, \ldots, G15$), where each group consists of 16 processors $(0, 1, \ldots, 15)$ organized into a 4 × 4 2D-Mesh. This network topology instance preserves the topological characteristics of EDN-OTIS-Mesh. The shaded processors in each group represent the EDN processors (a set of nodes that can deliver a message to all other processors in a group in a single message-passing step). For example, within group $G0$, processors number 1, 7, 8, and 14 are called level-1 EDN processors where they can deliver a message to level-0 processors (nodes) in a single message-passing step, where level-0 processors are the rest of the processors.

The node in the OTIS-Mesh simulator is configured as a single-port model, which performs a single send/receive operation at a time as illustrated in Fig. 20. Alternatively, the node can also be configured as all-port model capable of performing simultaneous send/receive operations over multiple channels as illustrated in Fig. 21.

The network switching method used by the simulator is the wormhole switching introduced in [30]. Using this method, the message is partitioned into small fragments called flits. The header is the first flit of the packet that contains routing information and followed by data flits, which are transmitted in a pipelined manner. Wormhole switching eliminates the need of having large packet buffers at each intermediate node. Moreover, wormhole switching is convenient in the systems that use XY routing in which messages are routed first in X-dimension then in the Y-dimension.

6 Simulation results and performance evaluation

This section presents and provides a detailed discussion on the obtained simulation results. The performance of the broadcast and global combine communication operations is evaluated on the following three interconnection networks: single-port wormhole-routed OTIS-Mesh, all-port wormhole-routed OTIS-Mesh, and all-port wormhole-routed EDN-OTIS-Mesh, under the following performance metrics: number of communication steps, latency, and latency improvement.

6.1 Number of communication steps

The number of electronic message-passing steps is evaluated and compared under the single-port wormhole-routed OTIS-Mesh, all-port wormhole-routed OTIS-Mesh, and all-port wormhole-routed EDN-OTIS-Mesh interconnection networks. On the other hand, the number of optical communication steps required for broadcasting and global combine is the same on all the above three interconnection networks, which is only one optical step.

The number of electronic communication steps required to perform a broadcast or a global combine operation in all-port EDN-OTIS-Mesh is significantly less than that in

Fig. 18 16 × 16 OTIS-Mesh

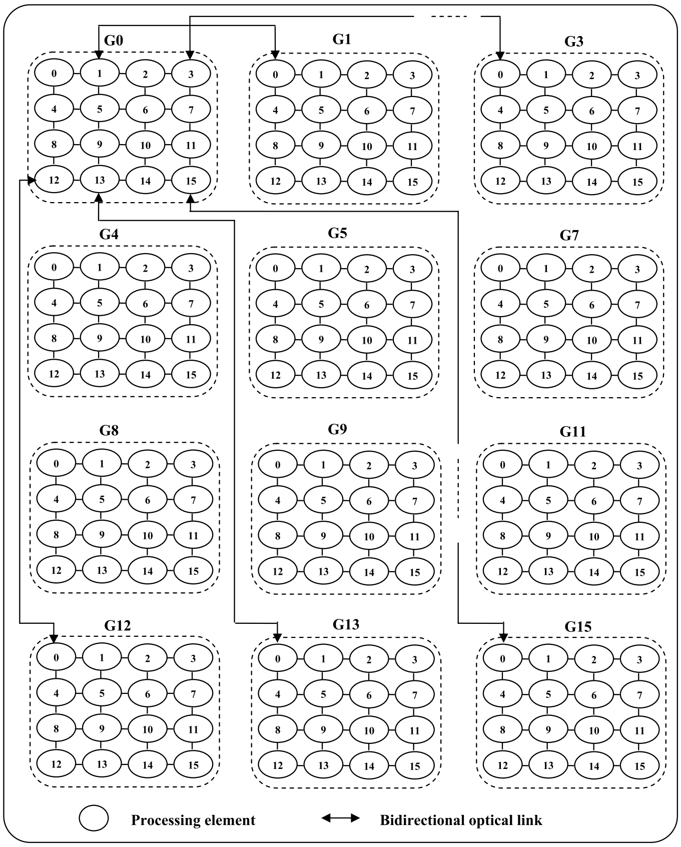
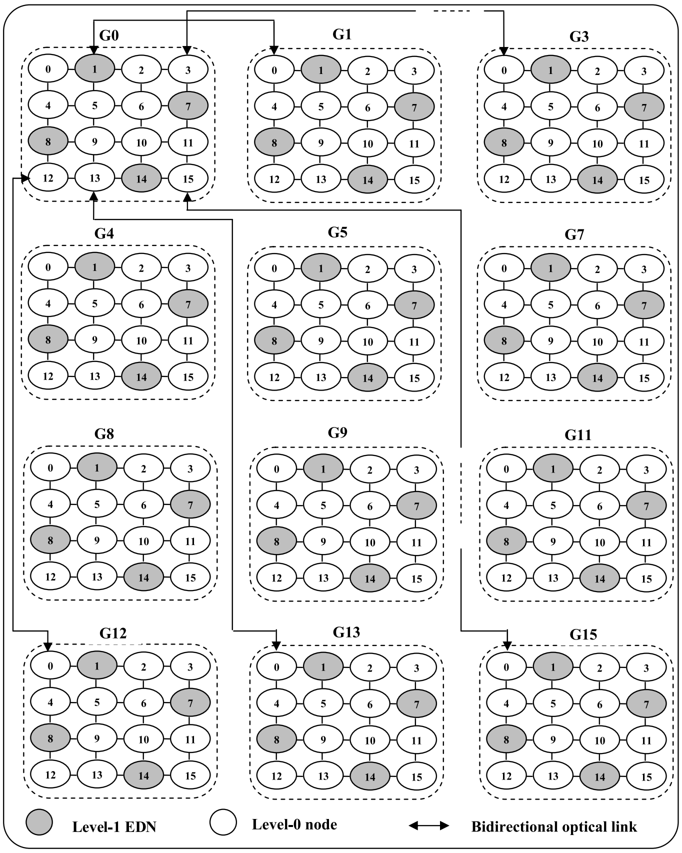


Fig. 19 16 × 16 EDN-OTIS-Mesh



single-port OTIS-Mesh and all-port OTIS-Mesh due to the fact that the EDN processors in each group increase the number of parallel send or receive operations. As shown in Fig. 22, the best-case, which is represented by the minimum number of electronic communication steps required to perform a broadcast or a global combine operation in all-port EDN-OTIS-Mesh is 6, 8, 10, and 12 for networks of sizes 16 × 16, 64 × 64, 256 × 256, and 1024 × 1024, respectively, whereas, in all-port OTIS-Mesh, the minimum number of electronic communication steps is 16, 64, 256, 1024, and in single-port OTIS-Mesh, it is 30, 126, 510, and 2046 for the same previous networks sizes, respectively. The worst-case, on the other hand, is shown in Fig. 23 which gives the maximum number of electronic communication steps required to perform the broadcast or global combine operations in all-port EDN-OTIS-Mesh, which is 8, 10, 12, and 14 for networks of sizes 16 × 16, 64 × 64, 256 × 256, and 1024 × 1024, respectively. For the same networks sizes, the maximum number of communication steps is 24, 112, 480, and 1984 in all-port OTIS-Mesh, and in single-port OTIS-Mesh, it is 30, 126, 510, and 2046. It is clear from Figs. 22 and 23 that the minimum and maximum number of electronic communication steps for performing a broadcast or a global combine operation in single-port OTIS-Mesh is the same for networks sizes 16 × 16, 64 × 64, 256 × 256, and 1024 × 1024 since it is single-port. Thus, performing one send or receive operation at a time. A look at Fig. 22 reveals that the minimum number of electronic communication steps for a broadcast or a global combine operation in all-port OTIS-Mesh is about as twice as lower than that to perform it in single-port OTIS-Mesh, because the single-port OTIS-Mesh performs one send or receive operation at a time. On the other hand, Fig. 23 shows that the maximum number of electronic communication steps to perform broadcast or global combine in all-port OTIS-Mesh is slightly lower than that to perform it in single-port OTIS-Mesh, because the all-port OTIS-Mesh performs a limited number of parallel send or receive operations. Furthermore, it can be noticed from Figs. 22 and 23 that as the network size increases, the number of electronic communication steps increases rapidly to perform broadcast or global combine in single-port and all-port OTIS-Mesh because the number of parallel send or receive operations in-

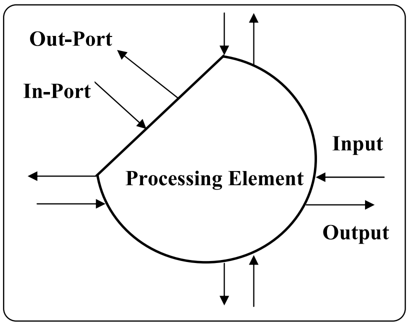


Fig. 20 Single-port node

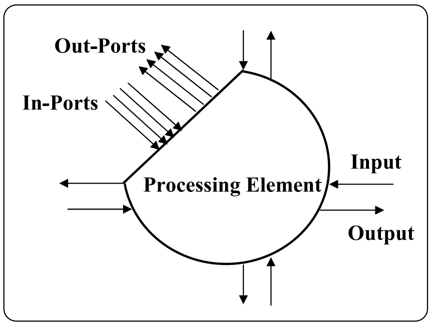
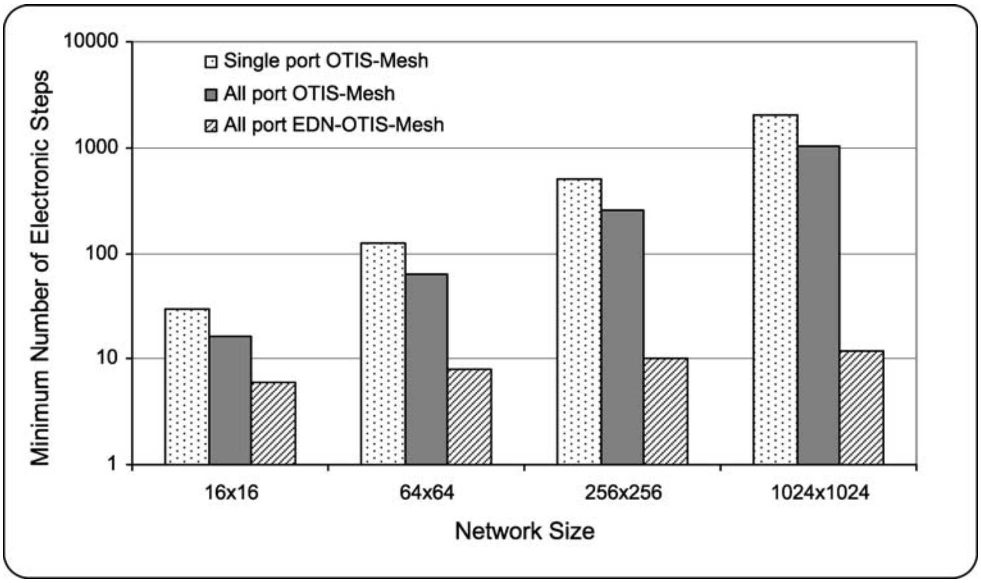


Fig. 21 All-port node

Fig. 22 Minimum number of electronic communication steps to perform broadcast operation or global combine operation



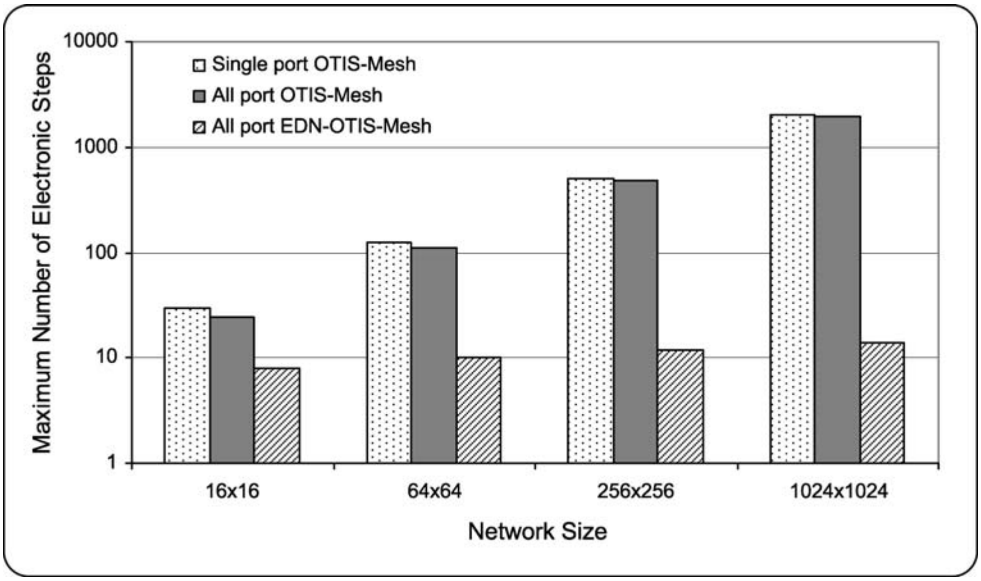


Fig. 23 Maximum number of electronic communication steps to perform broadcast operation or global combine operation

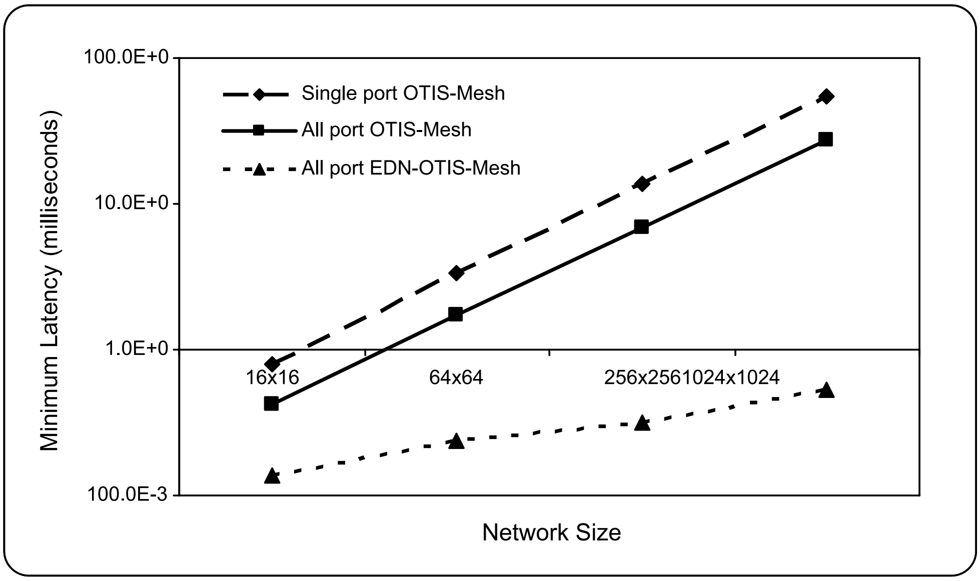


Fig. 24 Minimum latency to perform broadcast operation

creases, whereas the number of electronic communication steps in all-port EDN-OTIS-Mesh increases very slightly.

6.2 Latency

The latency is a measurement of the elapsed time since the beginning of communication until the last processor finishes communication. The latency, of both broadcast and global combine operations, is evaluated on the single-port OTIS-Mesh, all-port OTIS-Mesh, and all-port EDN-OTIS-Mesh interconnection networks, in both; the best-case and worst-case scenarios.

The minimum (best-case) latency to perform broadcast in single-port OTIS-Mesh is about as twice as higher than to perform it in all-port OTIS-Mesh and higher, significantly, than all-port EDN-OTIS-Mesh, as shown in Fig. 24, because the single-port OTIS-Mesh performs one send operation at a time. Moreover, the maximum (worst-case) latency to perform broadcast operation in all-port EDN-OTIS-Mesh is lower, significantly, than that to perform it in single-port and all-port OTIS-Mesh, as shown in Fig. 25, because the single-port OTIS-Mesh performs one send operation at a time and the all-port OTIS-Mesh performs a limited number of send operations. It can also be noticed from Figs. 24 and 25 that the latency of broadcasting in all-port EDN-OTIS-Mesh is lower, significantly, than the resulting latency to perform it in all-port OTIS-Mesh due to the existence of the EDN processors in the EDN-OTIS-Mesh, which increase the parallel send operation, yielding less latency. Furthermore, Figs. 24 and 25 make it clear that as the network size

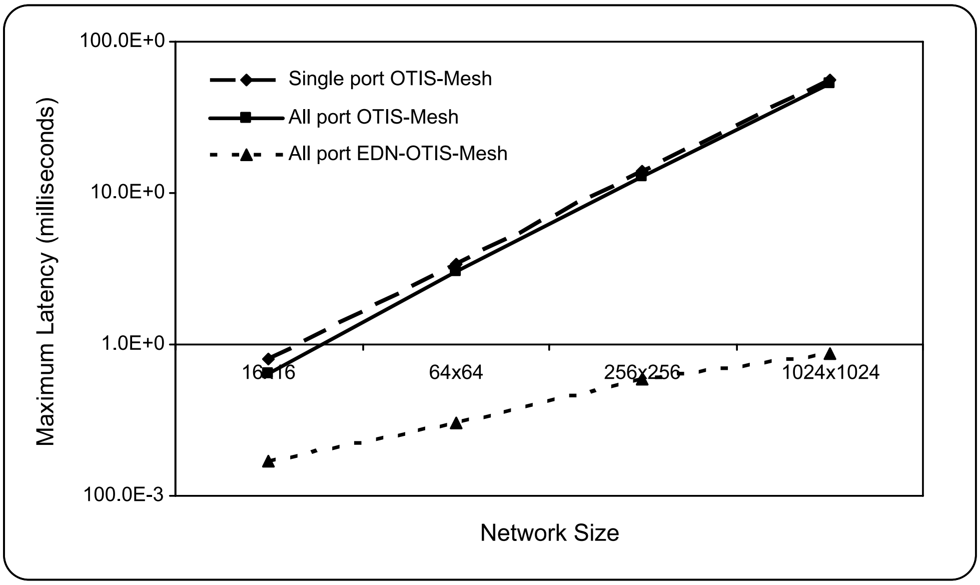


Fig. 25 Maximum latency to perform broadcast operation

Table 5 Minimum latency (milliseconds) to perform global-combine operation

Network Size	Single-Port OTIS-Mesh	All-Port OTIS-Mesh	All-Port EDN-OTIS-Mesh
16 × 16	6.5E+0	3.2E+0	2.4E+0
64 × 64	108.8E+0	54.4E+0	40.8E+0
256 × 256	1.8E+3	881.3E+0	661.0E+0
1024 × 1024	28.3E+3	14.1E+3	10.6E+3

Table 6 Maximum latency (milliseconds) to perform global-combine operation

Network Size	Single-Port OTIS-Mesh	All-Port OTIS-Mesh	All-Port EDN-OTIS-Mesh
16 × 16	6.9E+0	5.5E+0	4.1E+0
64 × 64	110.6E+0	98.3E+0	73.7E+0
256 × 256	1.8E+3	1.7E+3	1.2E+3
1024 × 1024	28.3E+3	27.5E+3	20.6E+3

increases, the latency increases rapidly to perform broadcast in single-port and all-port OTIS-Mesh because the number of parallel send operations increases, whereas the latency in all-port EDN-OTIS-Mesh increases very slightly.

The minimum and maximum latency to perform a global combine operation in single-port OTIS-Mesh is higher than that to perform it in all-port OTIS-Mesh and all-port EDN-OTIS-Mesh for different networks sizes, as shown in Tables 5 and 6, respectively. This refers to the fact that the single-port OTIS-Mesh performs one receive operation at a time, while the all-port EDN-OTIS-Mesh can perform parallel receive operations at a time. Moreover, it is shown in Tables 5 and 6 that the minimum and maximum latency to perform global combine in all-port EDN-OTIS-Mesh is lower than those required to perform it in all-port OTIS-Mesh because every group in the EDN-OTIS-Mesh collects the received messages in EDNs before sending them to the control group, which increases the number of parallel receive operations. An examination look at Tables 5 and 6 reveals that as the network size increases, the latency increases rapidly to perform global combine in single-port and all-port OTIS-Mesh, and all-port EDN-OTIS-Mesh due to the increasing number of the parallel receive operations.

6.3 Latency improvement

The latency improvement metric (i.e., reduction in latency) is defined as the improvement achieved by the broadcast and global combine communication operations using all-port EDN-OTIS-Mesh over single-port OTIS-Mesh or over all-port OTIS-Mesh. In particular, the latency improvement

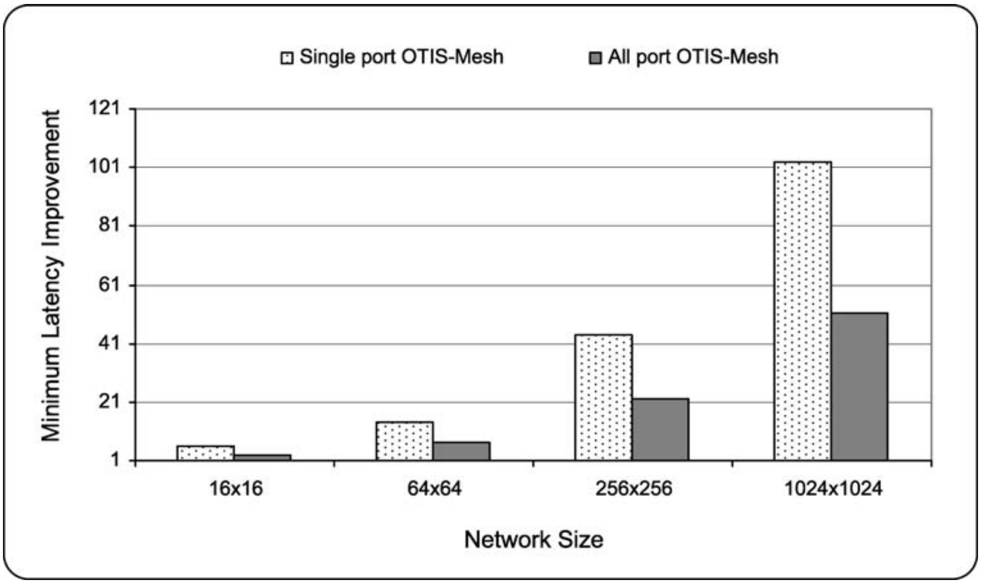


Fig. 26 Minimum latency improvement of broadcast operation on all-port EDN-OTIS-Mesh

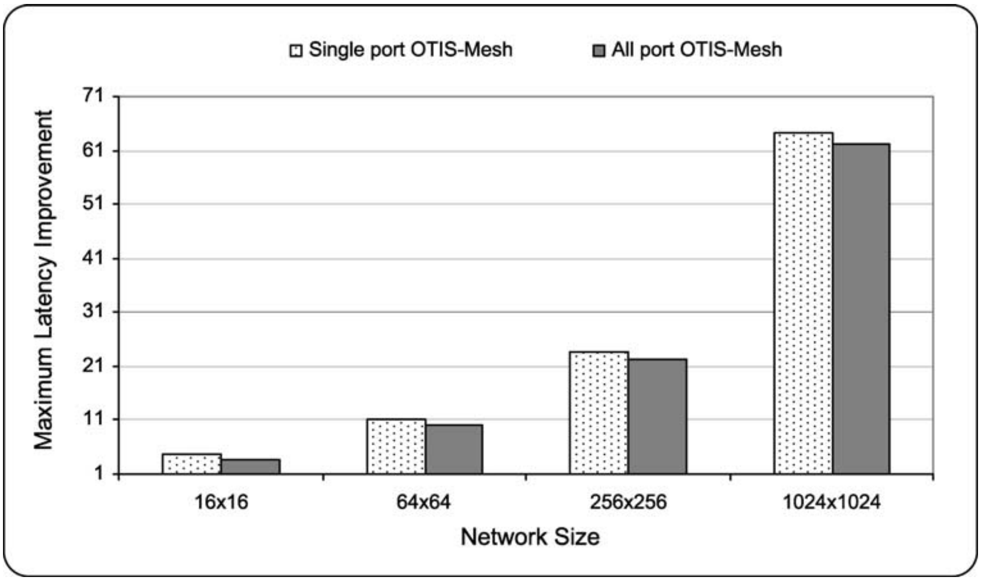


Fig. 27 Maximum latency improvement of broadcast operation on all-port EDN-OTIS-Mesh

is the ratio of the latency of single-port OTIS-Mesh or all-port OTIS-Mesh over all-port EDN-OTIS-Mesh.

The minimum (best-case) and maximum (worst-case) latency improvement simulation results of the broadcast operation show that the all-port EDN-OTIS-Mesh outperforms both the single-port OTIS-Mesh and all-port OTIS-Mesh interconnection networks significantly, as shown in Figs. 26 and 27. For example, the minimum latency improvement simulation results of the broadcast operation show that all-port EDN-OTIS-Mesh outperforms single-port OTIS-Mesh by 5.8, 14.18, 43.91, and 103.02 times for network sizes 16 × 16, 64 × 64, 256 × 256 and 1024 × 1024, respectively, as shown in Fig. 26. It also outperforms the all-port OTIS-Mesh by 3.05, 7.18, 22.02, and 51.44 times for the same previous network sizes, as shown in Fig. 26, because the all-port EDN-OTIS-Mesh performs parallel send operations, where the root processor sends the message only to the highest EDN level and the EDN processors send the message to the lowest EDN levels then to level-0 processors, so the number of parallel send operations is increased by the EDNs. Also, the maximum latency improvement simulation results of broadcast operation show that all-port EDN-OTIS-Mesh outperforms both the single-port OTIS-Mesh and all-port OTIS-Mesh by almost the same values, as shown in Fig. 27, since all-port OTIS-Mesh performs a limited number of parallel send operations, where the root processor sends the message to all processors in the group, so the number of parallel send is limited by the number of ports that the root processor has.

Fig. 28 Minimum latency improvement of global combine operation on all-port EDN-OTIS-Mesh

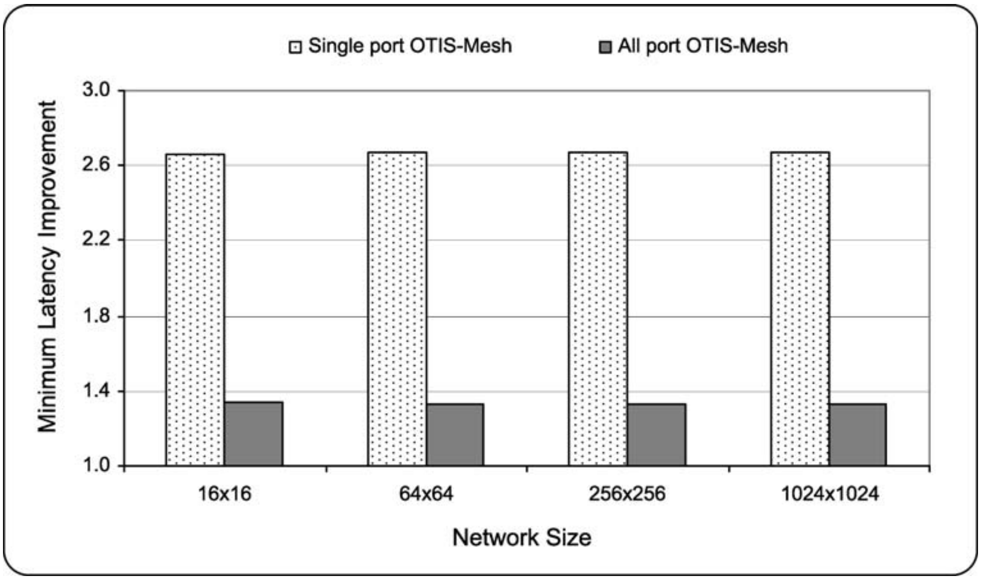
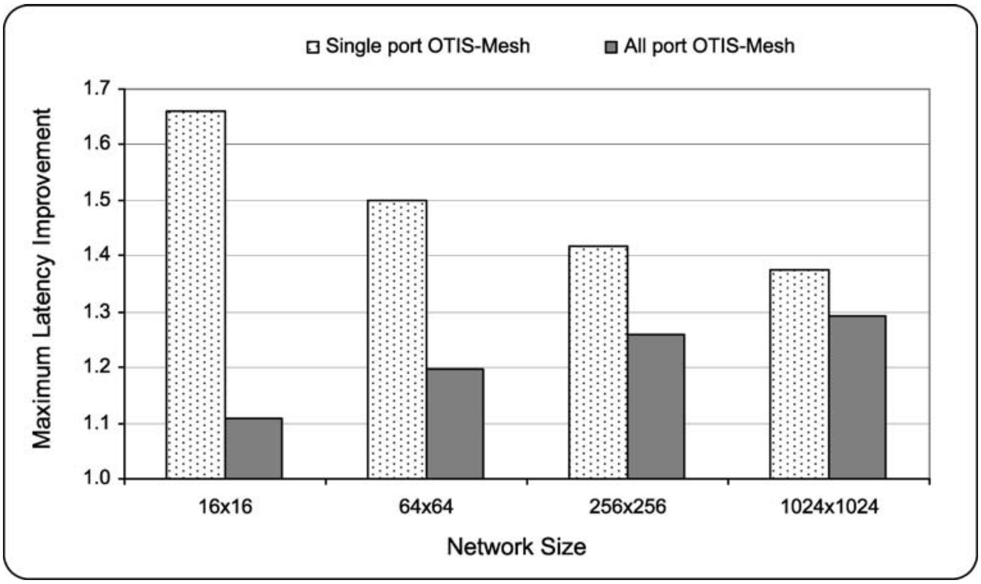


Fig. 29 Maximum latency improvement of global combine operation on all-port EDN-OTIS-Mesh



The minimum latency improvement (best-case) simulation results of the global combine operation show that all-port EDN-OTIS-Mesh outperforms single-port OTIS-Mesh and all-port OTIS-Mesh by 2.66 and 1.33 times respectively for network sizes 16 × 16, 64 × 64, 256 × 256 and 1024 × 1024, as shown in Fig. 28, since all-port EDN-OTIS-Mesh performs parallel receive operation. Also, the maximum latency improvement (worst-case) simulation results of global combine operation show that all-port EDN-OTIS-Mesh outperforms single-port OTIS-Mesh by 1.66, 1.5, 1.42 and 1.37 times for network sizes 16 × 16, 64 × 64, 256 × 256 and 1024 × 1024 respectively, and it outperforms the all-port OTIS-Mesh only by 1.11, 1.2, 1.26 and 1.29 times for the same previous network sizes, as shown in Fig. 29, because the all-port EDN-OTIS-Mesh performs a limited number of parallel receive operations in the worst-case.

The minimum latency improvement to perform global combine operation on all-port EDN-OTIS-Mesh over single-port OTIS-Mesh and over all-port OTIS-Mesh is the same on the networks of sizes 16 × 16, 64 × 64, 256 × 256 and 1024 × 1024, as shown in Fig. 28. This is because the ratio of latency on single-port OTIS-Mesh or all-port OTIS-Mesh over the latency on all-port EDN-OTIS-Mesh is the same on the previous networks sizes. Moreover, the EDNs on all-port EDN-OTIS-Mesh group the received messages from non-EDN nodes before the messages are forwarded to the destination; this grouping mechanism eliminates the unnecessary headers from the collected messages, which reduces the overall message size and consequently reduces the latency of the global combine operation. This amount of reduction in the latency is proportional on the previous network sizes.

The maximum latency improvement to perform a global combine operation on all-port EDN-OTIS-Mesh over single-port OTIS-Mesh slightly decreases by 1.66, 1.5, 1.42, and 1.37 on network sizes 16 × 16, 64 × 64, 256 × 256 and 1024 × 1024, respectively, as shown in Fig. 29. This is because the ratio of latency on single-port OTIS-Mesh to that of all-port EDN-OTIS-Mesh decreases on the previous networks. This slight decrease is explained by the location of the root processor. Since the root processor is located at the endmost of the OTIS-Mesh group, then the EDN-OTIS-Mesh performs fewer parallel receive operations and thus increasing the latency of the operation and consequently decreasing the improvement ratio. However, the maximum latency improvement to perform global combine on all-port EDN-OTIS-Mesh over all-port OTIS-Mesh slightly increases by 1.11, 1.2, 1.26, and 1.29 times on network sizes 16 × 16, 64 × 64, 256 × 256 and 1024 × 1024, respectively, as shown in Fig. 29. This is clarified by the location of the root processor at the endmost of the OTIS group, which causes the number of parallel receives in OTIS-Mesh to decrease, leading to an increase in the latency value. However, the EDNs in all-port OTIS-Mesh perform grouping of the collected messages, where the EDNs on all-port EDN-OTIS-Mesh group the received messages from non-EDN nodes before the messages are forwarded to the destination. Therefore, the latency in EDN-OTIS-Mesh decreases, when compared to all-port OTIS-Mesh, and consequently increasing the improvement ratio.

7 Conclusions and future work

This paper presented and evaluated broadcast and global combine communication operations on the Optical Transposed Interconnection System Mesh (OTIS-Mesh). The Extended Dominating Node (EDN) of the graph theory field along with the OTIS-Mesh were used as the underlying interconnection network architecture, where EDN approach was applied on each group of OTIS-Mesh to form the EDN-OTIS-Mesh in order to implement an efficient approach for broadcast and global combine communication operations.

The performance of broadcast and global combine communication operations was evaluated analytically and by simulation on the following interconnection networks: single-port wormhole-routed OTIS-Mesh, all-port wormhole-routed OTIS-Mesh, and all-port wormhole-routed EDN-OTIS-Mesh under the following performance metrics: number of communication steps, latency, and latency improvement.

The analytical results revealed that the number of communication steps to perform the broadcast and global combine operations in all-port EDN-OTIS-Mesh is less than that required to perform them on both; single-port and all-port OTIS-Mesh, significantly. The excellence of EDN-OTIS-Mesh is justified by the fact that the EDNs in all-port EDN-OTIS-Mesh increase the number of parallel send or receive operations. For example, in broadcast operation the message is sent by the EDN processors, not by a single processor as OTIS-Mesh without EDN; consequently, the number of communication steps is reduced.

As a complementary work, simulation runs were conducted, whose results indicated that both operations; broadcast and global combine, in the all-port EDN-OTIS-Mesh performed better, in terms of latency and latency improvement than both the single-port and all-port OTIS-Mesh in both the best-case and worst-case runs. The reason behind this is that the EDNs in all-port EDN-OTIS-Mesh reduces the latency as the number of EDNs increases the parallel send or receive operations, so the time to perform the operation is optimized. For example, in the global combine operation, messages' grouping is done by the EDN processors, not by a single processor. Consequently, the operation's latency is reduced. Moreover, the maximum (worst-case) latency improvement of the broadcast operation outperformed the single-port model.

As a future work, the EDN approach can be used to design and implement the broadcast, global combine, and other collective communications operations, such as scatter, reduction, barrier, etc. on other OTIS interconnection networks' instances, such as OTIS-Hypercube.

References

1. Duato, J., Yalamanchili, C., Ni, L.: Interconnection Networks: An Engineering Approach. IEEE Computer Society Press, Los Alamitos (1997)
2. Trobec, R., Brostnik, U., Janezic, D.: Communication performance of d-meshes in molecular dynamics simulation. J. Math. Chem. **45**(2), 503–512 (2009)
3. Park, S.-Y., Hariri, S.: A high performance message-passing system for network of workstations. J. Supercomput. **11**(2), 159–180 (1997)
4. Park, S.-Y., Hariri, S.: ACS: An adaptive communication system for heterogeneous wide-area ATM clusters. Clust. Comput. **2**(3), 229–246 (1999)
5. Marsden, G., Marchand, P., Harvey, P., Esener, S.: Optical transpose interconnection system architectures. Opt. Lett. **18**(13), 1083–1085 (1993)
6. Tsai, Y., McKinley, P.: An extended dominating nodes to collective communication in wormhole-routed 2D meshes. In: Proceedings of the IEEE Scalable High Performance Computing Conference, pp. 199–206, TN (1994)
7. Tsai, Y., McKinley, P.: An extended dominating node approach to broadcast and global combine in multiport wormhole-routed mesh networks. IEEE Trans. Parallel Distrib. Syst. **8**(1), 41–58 (1997)

8. Tsai, Y., McKinley, P.: A dominating set model for broadcast in all-port wormhole-routed 2D mesh networks. In: Proceedings of the Eighth ACM International Conference on Supercomputing, pp. 126–135, England (1994)
9. McKinley, P., Tsai, Y., Robinson, D.: Collective communication in wormhole-routed massively parallel computers. Computer **28**(12), 39–50 (1995)
10. Hartmann, O., Kühnemann, M., Rauber, T., Rünger, G.: An adaptive extension library for improving collective communication operations. Concurr. Comput.: Pract. Exp. **20**(10), 1173–1194 (2008)
11. Mahafzah, B., Jaradat, B.: The load balancing problem in OTIS-Hypercube interconnection networks. J. Supercomput. **46**(3), 276–297 (2008)
12. Matsuda, M., Kudoh, T., Kodama, Y., Takano, R., Ishikawa, Y.: The design and implementation of MPI collective operations for clusters in long-and-fast networks. Clust. Comput. **11**(1), 45–55 (2008)
13. Pjesivac-Grbovic, J., Angskun, T., Bosilca, G., Fagg, G., Gabriel, E., Dongarra, J.: Performance analysis of MPI collective operations. Clust. Comput. **10**(2), 127–143 (2007)
14. Kenyon, C., Schabanel, N.: The data broadcast problem with non-uniform transmission times. Algorithmica **35**(2), 146–175 (2008)
15. Dvorak, V.: Communication performance of mesh- and ring-based NoCs. In: Seventh International Conference on Networking (ICN 2008), pp. 156–161 (2008)
16. Chen, Y.-S., Chiang, C.-Y., Chen, C.-Y.: Multi-node broadcasting in all-ported 3-D wormhole-routed torus using an aggregation-then-distribution strategy. J. Syst. Archit. **50**(9), 575–589 (2004)
17. Barnett, M., Payne, D., Van de Geijn, R., Watts, J.: Broadcasting on meshes with wormhole routing. J. Parallel Distrib. Comput. **35**(2), 111–122 (1996)
18. Shang, W., Yao, F., Wan, P., Hu, X.: On minimum m-connected k-dominating set problem in unit disc graphs. J. Comb. Optim. **16**(2), 99–106 (2008)
19. Krishnamoorthy, A., Marchand, P., Kiamilev, F., Esener, S.: Grain-size considerations for optoelectronic multistage interconnection networks. Appl. Opt. **31**(26), 5480–5507 (1992)
20. Najaf-abadi, H., Sarbazi-azad, H.: An empirical comparison of OTIS-mesh and OTIS-hypercube multicomputer systems under deterministic routing. In: Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium (IPDPS'05), Workshop 14, vol. 15 (2005)
21. Zhao, C., Xiao, W., Qin, Y.: Hybrid diffusion schemes for load balancing on OTIS-Networks. In: Proceedings of the 7th International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP), China, 2007. Lecture Notes in Computer Science, vol. 4494, pp. 421–432. Springer, Berlin (2007)
22. Qin, Y., Xiao, W., Zhao, C.: GDED-X schemes for load balancing on heterogeneous OTIS-networks. In: Proceedings of the 7th International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP), China, 2007. Lecture Notes in Computer Science, vol. 4494, pp. 482–492. Springer, Berlin (2007)
23. Zhao, C., Xiao, W., Parhami, B.: Load-balancing on swapped or OTIS networks. J. Parallel Distrib. Comput. **69**(4), 389–399 (2009)
24. Najaf-abadi, H., Sarbazi-azad, H.: Comparative evaluation of adaptive and deterministic routing in the OTIS-hypercube. In: Proceeding of Ninth Asia-Pacific Computer Systems Architecture Conference (ACSAC 2004), Beijing, China. Lecture Notes in Computer Science, vol. 3189, pp. 349–362. Springer, Berlin (2004)
25. Day, K., Al-Ayyoub, A.: Topological properties of OTIS-networks. IEEE Trans. Parallel Distrib. Syst. **13**(4), 359–366 (2002)
26. Wei, W., Xiao, W.: Algorithms of basic communication operation on the biswapped network. In: Proceedings of the 8th International Conference on Computational Science (ICCS 2008), Part I, Krakow, Poland, 2008. Lecture Notes in Computer Science, vol. 5101, pp. 347–354. Springer, Berlin (2008)
27. Day, K.: Optical transpose k-ary n-cube networks. J. Syst. Archit. **50**(11), 697–705 (2004)
28. Wang, C., Sahni, S.: Basic operations on the OTIS-mesh optoelectronic computer. IEEE Trans. Parallel Distrib. Syst. **9**(12), 1226–1236 (1998)
29. Rajasekaran, S., Sahni, S.: Randomized routing, selection, and sorting on the OTIS-mesh. IEEE Trans. Parallel Distrib. Syst. **9**(9), 833–840 (1998)
30. Wilkinson, B.: Computer Architecture Design and Performance, 2nd edn. Prentice Hall, New York (1996)

Basel A. Mahafzah is an Assistant Professor of Computer Science at the University of Jordan, Jordan. He received his B.Sc. degree in Computer Science in 1991 from Mu'tah University, Jordan. He also earned from the University of Alabama in Huntsville, USA, a B.S.E. degree in Computer Engineering. Furthermore, he obtained his M.S. degree in Computer Science and Ph.D. degree in Computer Engineering from the same University, in 1994 and 1999, respectively. During his graduate studies he obtained a fellowship from Jordan University of Science and Technology. After he obtained his Ph.D. and before joining the University of Jordan, he joined the Department of Computer Science at Jordan University of Science and Technology, where Dr. Mahafzah held several positions; Assistant Dean, Vice Dean, and Chief Information Officer at King Abdullah University Hospital. His research interests include Performance Evaluation, Parallel and Distributed Computing, Interconnection Networks, Artificial Intelligence, Data Mining, Software Testing, and e-Learning. He received more than one million US dollars in research and projects grants. Moreover, Dr. Mahafzah supervised graduate students and developed several graduate and undergraduate programs in various fields of Information Technology. His experience in teaching extends to ten years.

Ruby Y. Tahboub is a full time instructor at the Computer Science Department of Jordan University of Science and Technology, Jordan. She received her B.Sc. degree in Computer Science from Jordan University of Science and Technology, Jordan in 2005, and M.S. degree in Computer Science from University of Jordan, Jordan in 2008. Her research interests include Parallel and Distributed Computing, Business Intelligence, Database Systems and Data Mining.

Omar Y. Tahboub is a Teaching Assistant at the Computer Science Department at Kent State University, USA. He received his B.Sc. and M.S. degrees in Computer Science from Jordan University of Science and Technology and University of Jordan, Jordan, in 2002 and 2004, respectively. He is currently perusing his Ph.D. degree in Computer Science at Kent State University, USA. His research interests include High Speed Communication Networking, Design of Next Generation Space Communication Architectures, and Parallel and Distributed Computing.