COLLECTIVE COMMUNICATION IN OPTICAL TRANSPOSED INTERCONNECTION SYSTEM MESH USING EXTENDED DOMINATING NODE APPROACH

By
Ruby Yahya Tahboub

Supervisor
Dr. Sami Serhan

Co-Supervisor
Dr. Basel Mahafzah

This Thesis was Submitted in Partial Fulfillment of the Requirements for the Master's Degree of Science in Computer Science

Faculty of Graduate Studies
The University of Jordan

May 2008

COMMITTEE DECISION

This thesis (Collective Communication in Optical Transposed Interconnection System Mesh Using Extending Dominating Node Approach) was successfully defended and approved on ---12/5/2008-------

Examination Committee   **Signature**
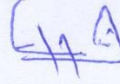
Dr. Sami Serhan, Chairman
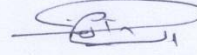Associate Professor of Compilers

Dr. Basel Mahafzah, Member
Assistant Professor of Parallel and Distributed Systems

Dr. Fawaz Zghoul, Member
Associate Professor of Software Engineering

Dr. Saleh Al-Sharaeh, Member
Associate Professor of Parallel Processing

Dr. Imad Eldeen Abu Rub, Member
Associate Professor of Parallel Processing

(Zarqa private university)

DEDICATION

To my family … To all of my friends … Thank You for always being there for me.

ACKNOWLEDGMENT

I would like to express my sincere thanks to Dr. Sami Serhan and Dr. Basel Mahafzah, my supervisors ... Thank You for your continuous support, efforts and dedication.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Term	Description
EDN	Extended Dominating Node
OTIS	Optical Transposed Interconnection System

APPENDICES

GLOSSARY OF TERMS

Term	Description
Barrier	A synchronization point in parallel algorithm where all participating processors must arrive to the barrier before any processor proceeds execution.
Collective Communication	A point where all the processors in the system, call the same routine, such as scatter, reduction, etc.
Dominating Node	A set of dominating nodes that have direct links to all other processors in a group so it can deliver a message to all other processors in a single step.
Extended Dominating Node (EDN)	A set of nodes that can deliver a message to all other processors in a group in a single message-passing step, where EDN do not necessarily have direct links to all other processors.
Matrix Transposition	Is the process of reflecting the data elements of a matrix by its main diagonal.
Optical Transposed Interconnection System (OTIS)	Is an interconnection system where processors are divided into groups, communication between processors within a group is realized by electronic interconnect links whereas communication between processors on different groups is realized by optical interconnect links.
Reduction	A global compute operation such as maximum, minimum, sum, etc. is performed on data items located in each processor in the system.
Routing	The process of selecting a path between two nodes in the network.
Scatter	One processor sends a different message for each processor.
Wormhole Switching	A switching method where the packet 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.

COLLECTIVE COMMUNICATION IN OPTICAL TRANSPOSED INTERCONNECTION MESH USING EXTENDED DOMINATING NODE APPROACH

By
Ruby Yahya Tahboub

Supervisor
Dr. Sami Serhan

Co-Supervisor
Dr. Basel Mahafzah

ABSTRACT

The Optical Transposed Interconnection System Mesh (OTIS-Mesh) is an optoelectronic interconnection network that employs electronic links to connect short distance processing elements and optical links to connect distant processing elements.

This thesis presents the design and implementation of a set of efficient collective communications' operations, which are scatter, reduction and barrier, for OTIS-Mesh by applying Extended Dominating Node (EDN) approach. Therefore, forming EDN OTIS-Mesh. Moreover, matrix transposition is presented as an application on data movement collective communication.

A performance evaluation study was conducted to evaluate and compare the performance of collective communications' operations, which are scatter, reduction and barrier, on the following interconnection systems: single port OTIS-Mesh, all port OTIS-Mesh and all port EDN OTIS-Mesh under the following performance metrics: number of steps, latency, execution time, total overhead, performance improvement and cost.

The experimental results show that the barrier operation in all port EDN OTIS-Mesh outperformed both single port and all port OTIS-Mesh under the following performance metrics: number of steps, latency, and performance improvement. For instance, the maximum latency of the barrier operation in all port EDN OTIS-Mesh outperformed all port OTIS-Mesh by 9.7 and 47.6 times for network sizes 256 and 1024 respectively.

The reduction operation in all port EDN OTIS-Mesh outperformed both single port and all port OTIS-Mesh under the following performance metrics: number of steps, execution time, performance improvement, total overhead and cost. For instance, the maximum execution time of the reduction operation in all port EDN OTIS-Mesh outperformed all port OTIS-Mesh by 12.63 and 29.47 times for network sizes 256 and 1024 respectively. On the other hand, the scatter operation in EDN OTIS-Mesh outperformed single port OTIS-Mesh only under the following performance metrics: number of steps, latency, performance improvement, and

showed relatively similar performance when it was compared with all port OTIS-Mesh under the pervious metrics.

1- Introduction

An optoelectronic parallel computing system called the Optical Transposed Interconnection Mesh (OTIS-Mesh) is a parallel computing system where electronic interconnect links are used to interconnect processing elements when the communication distance between processors is short, but in case of large communication distance optical links are used.

A graph theory approach called the Extended Dominating Node (EDN) defines a set of nodes (dominating nodes) that is capable to deliver a message to all other processors in a group within a single message-passing step.

Collective communication in parallel and distributed systems involves two or more processors calling the same operation. Collective communication operations are widely used in various applications such as graphs, sorting and image processing. Therefore, collective communication operations should be designed and implemented in an efficient way.

The problem statement of this thesis is presented in section 1.1. The motivations, objectives, contributions of this thesis are discussed in sections 1.2, 1.3, and 1.4, respectively. Finally, in section 1.5, the organization of the thesis is presented.

1.1 Problem Statement

This thesis presents the problem of designing and implementing a set of efficient (in terms of speed, cost, overhead, etc.) collective communication operations (such as scatter, reduction, etc) for an optoelectronic interconnection network called Optical Transposed Interconnection System Mesh (OTIS-Mesh) by applying the Extended Dominating Node (EDN) approach on the OTIS-Mesh. The thesis also presents the

matrix transposition as an application that on data movement collective communication.

1.2 Motivations

A typical parallel computing system consists of two or more interconnected elements within a single computer system. The choice of topology and the nature of interconnect links have a major role in determining cost, speed, power consumption, etc.

OTIS-Mesh is an optoelectronic interconnection system that divides the network into groups, communication within the group is realized by electronic interconnect links where communication between different groups is realized by optical links. Collective communication in parallel and distributed systems involves two or more processors calling the same operation. Collective communication operations such as scatter, reduction, barrier, etc. are widely used in solving problems such as sorting, graphs, image processing, etc., that is why collective communication operations must be designed and implemented in efficient way in order to maintain a good network performance.

1.3 Objectives

The objectives of this thesis are summarized as follows:

- Applying the EDN approach on each group in the OTIS-Mesh to form EDN OTIS-Mesh.

- Designing and implementing some collective communication operations such as scatter, reduction, barrier, etc. on the EDN OTIS-Mesh.

- Implementing the Matrix Transposition as an application on the EDN OTIS-Mesh.

- Conducting a comparative performance evaluation study between the collective communication operations on the OTIS Mesh and on the EDN OTIS-Mesh.

1.4 Contributions

The contributions of this thesis are summarized as follows:

- Implementing efficient collective communication operations on the OTIS-Mesh using the EDN approach.

- Increase the parallel send/receive, in EDN OTIS-Mesh collective communication operations.

- Exploit the wormhole switching to decrease the channel contention and latency in EDN OTIS-Mesh.

1.5 Thesis Organization

Chapter two presents background information about the main elements of the thesis; collective communication operations (scatter, reduction and barrier), the matrix transposition application, the concept of the Extended Dominating Node (EDN) approach from the graph theory, the Optical Transposed Interconnection System OTIS, some properties and routing of the OTIS-Mesh, in addition to literature review about some related work in this area. Chapter three presents the EDN OTIS-Mesh, collective communication operations in EDN OTIS-Mesh (scatter, reduction and barrier) in addition to the matrix transposition application. Chapter four presents the analytical model and the assumptions used to evaluate the following performance parameters:

number of steps, execution time, performance improvement, overhead and cost. Chapter five discusses some experimental environment; hardware and software in addition to the interconnection network simulator. Chapter six illustrates the experimental results and evaluation of number of steps, execution time, performance improvement, overhead and cost. Finally, chapter seven presents conclusions and future work.

2- Background and Related Work

This chapter discusses some relevant background and related work about the main elements of the problem. It presents the collective communication operations, the Extended Dominating Node (EDN) approach, and the Optical Transposed Interconnection System Mesh (OTIS-Mesh).

Section 2.1 presents the collective communication operations (scatter, reduction and barrier). The matrix transposition application is illustrated in section 2.2. The Extended Dominating Node (EDN) is illustrated in section 2.3 and section 2.4 illustrates the Optical Transposed Interconnection System Mesh (OTIS-Mesh) from properties, routing, basic operations, and applications.

2.1 Collective Communication

Collective communication in parallel and distributed systems involves two or more processors calling the same operation. Collective communication operations are widely used in solving problems that involve the exchange of huge amount of data among processors such as graphs and sorting applications.

In general, collective communications' operations are used for data movement such as scatter, global operation such as reduction, and process control as barrier. P.K. McKinley *et al.* (1997) discussed the design and use of collective communication operations such as (scatter, barrier, etc) in massively parallel computers that uses the wormhole routing as a switching method. H. Xu *et al.* (1992) implemented an efficient barrier synchronization operation for wormhole hypercube, and proposed a U-cube implementation for the hypercube systems that does not support broadcast or multicast in hardware. Moreover, W. Mao *et al.* (2007) implemented collective

communication operations such as scatter on store and forward all port torus, where the torus network was divided into subnetworks.

2.1.1 Scatter

In scatter operation, one processor sends a different message for each processor in the system. Figure 2.1 illustrates the scatter operation, where P_1 presents the root processor; P_1 sends a different message (m_1, m_2 and m_3) to each processor in the system (P_2, P_3 and P_4).
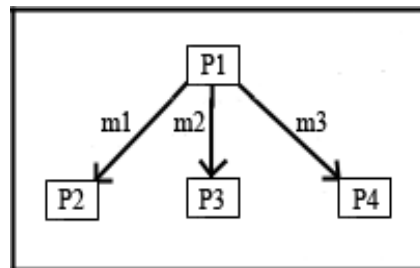
Figure 2.1: Scatter operation.

2.1.2 Reduction

In reduction operation, a global compute operation such as maximum, minimum, sum, etc. is performed on data items located in each processor in the system. Figure 2.2 illustrates the reduction operation, where P_1 presents the root processor; P_1 performs a global compute operation such as maximum, minimum, sum … etc (presented as * in Figure 2.2) on its data (m_1) and the received data (m_2, m_3 and m_4) from each processor in the system (P_2, P_3 and P_4).
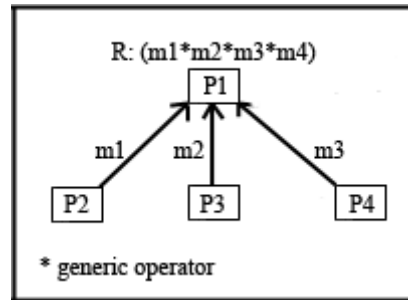
Figure 2.2: Reduction operation.

2.1.3 Barrier

Barrier is a synchronization point in parallel algorithm, where all participating processors must reach a certain point before any processor precedes execution. It consists of two phases: reduction and distribution as illustrated in Figure 2.3; in the reduction phase, one processor P_1 acts as a barrier processor. When a processor (such as P_2, P_3 and P_4) reaches the synchronization point it sends a message to the barrier processor. In the distribution phase, after all processors have reached the synchronization point, the barrier processor P_1 sends a message to all participating processors (P_2, P_3 and P_4) so they can continue execution.
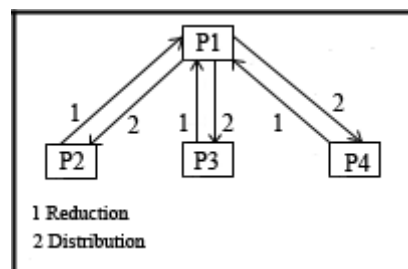
Figure 2.3: Barrier operation.

2.2 Application: Matrix Transposition

Matrix transposition is the process of reflecting the data elements of a matrix (distributed among the processors in the parallel system) by its main diagonal. Figure 2.4 part (a) illustrates a matrix distributed on 4×4 mesh, the data blocks (a_1, a_6, a_{11} and

a_{16}) are the diagonal (gray shaded squares), part (b) illustrates the transposed matrix, the diagonal elements are unchanged but the data blocks above the diagonal were transposed with the data blocks under the diagonal.
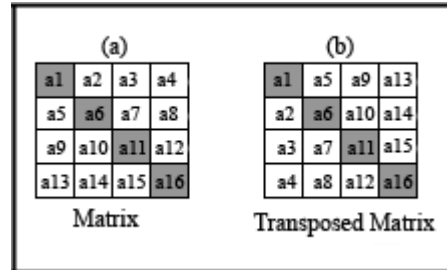
Figure 2.4: Matrix transposition.

2.3 Extended Dominating Node (EDN)

The dominating node approach introduced by Tsai *et al.* (1994) defines set of dominating nodes that have direct links to all other processors in a group so it can deliver a message to all other processors in a single step, as illustrated in Figure 2.5.
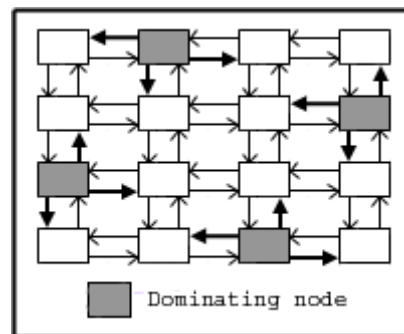
Figure 2.5: Dominating nodes in a 4× 4 2D mesh

The dominating node definition is by considering a set of nodes that can deliver a message to all other processors in a group in a single message passing step, as illustrated in Figure 2.6.
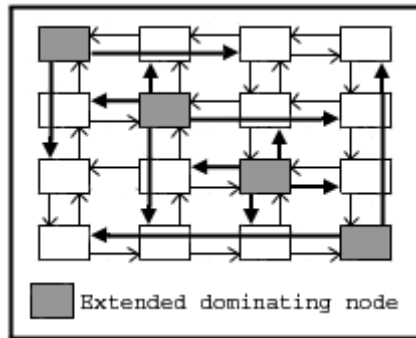
Figure 2.6: Extended dominating nodes in a 4× 4 2D mesh.

The definition of Extended Dominating Nodes (EDN) can be recursively used to form multiple levels of EDN processors, as illustrated in Figure 2.7, where gray nodes present level-2 EDN processors, dotted nodes present level-1 EDN processors, and white nodes present level-0 EDN processors.
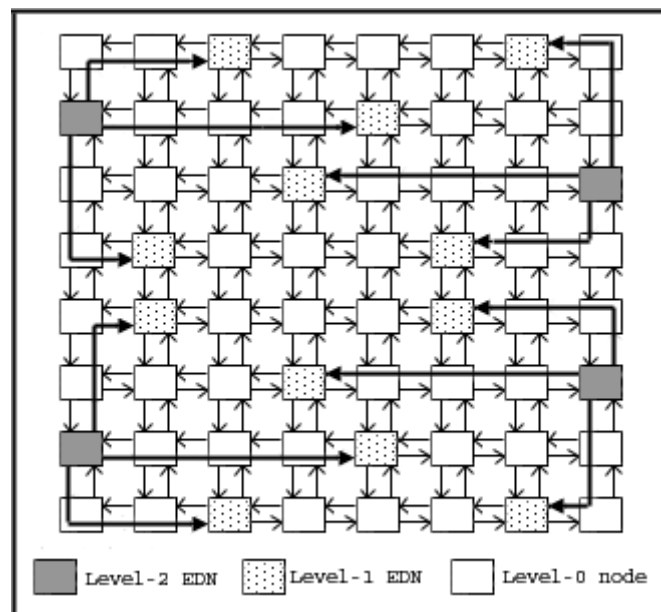
Figure 2.7: Extended dominating nodes in an 8×8 2D mesh.

The EDN approach used to design collective operations such as reduction by Tsai *et al.* (1994) and broadcast operations by Tsai *et al.* (1997) in all-port wormhole routed 2D mesh. Moreover, EDN was also applied on the problem of matrix transposition to minimize channel contention. The advantages of implementing collective communication operations using the EDN approach were confirmed over

the other approaches Xu *et al.* (1992) and Barnett (1993) by using simulation and analysis.

2.4 Optical Transposed Interconnection System (OTIS)

Optical Transposed Interconnection System (OTIS) that was introduced by Marsden *et al.* (1993) is an interconnection system that combines the advantages of electronic and optical interconnects. Processors in OTIS are divided into groups, communication between processors within a group is realized by electronic interconnect links whereas communication between processors on different groups is realized by optical interconnect links.

In general, an OTIS is divided into N groups, and each group consists of N processors. The processor within OTIS is modeled as a tuple (G, P) where G denotes group number and P is the processor number. The processor (G, P) is connected directly to its transposed processor (P, G) via optical link. The processors within a group can be interconnected as any of electronic topologies such as mesh, hypercube, star, mesh of trees, etc.

It has been verified by Krishnamoorthy *et al.* (1992) that the bandwidth in the OTIS network is maximized and the power consumption is minimized, when OTIS is divided into N groups and each group consists of N processors.

Mahafzah *et al.* (2008) developed a load balancing algorithm called Cluster Dimension Exchange Method (CDEM) for the OTIS-Hypercube and evaluated the performance of CDEM on OTIS-Hypercube and hypercube under the following metrics: execution time, load balancing accuracy, number of communication steps and speed. Experiment results showed that the OTIS-Hypercube outperformed the hypercube.

Najaf-Abadi and Sarbazi-azad (2004) implemented an adaptive routing algorithm for wormhole switched OTIS-hypercube and conducted an empirical performance evaluation study for adaptive wormhole routing in OTIS-hypercube under a number of conditions such as traffic loads, router delay, etc. Moreover, Najaf-Abadi and Sarbazi-azad (2005) evaluated the performance of OTIS-hypercube and OTIS-Mesh under different conditions such as cycle ratio, number of virtual channels, traffic, bisection width, etc.

2.4.1 OTIS-Mesh

The OTIS-Mesh consists of N groups having N processors in each group forming $\sqrt{N} \times \sqrt{N}$ mesh. Processors within the same group are connected as two-dimensional mesh. Figure 2.8 illustrates 4×4 OTIS-Mesh; it consists of 4 groups (named G_0, G_1, G_2, G_3), where each group consists of 4 processors interconnected as mesh. Each processor is interconnected with its transposed processor via an optical link (the bold arrows); for example, processor (0, 3) (that is processor 3 within group 0) shares an optical link with processor (3, 0) (that is processor 0 within group 3).
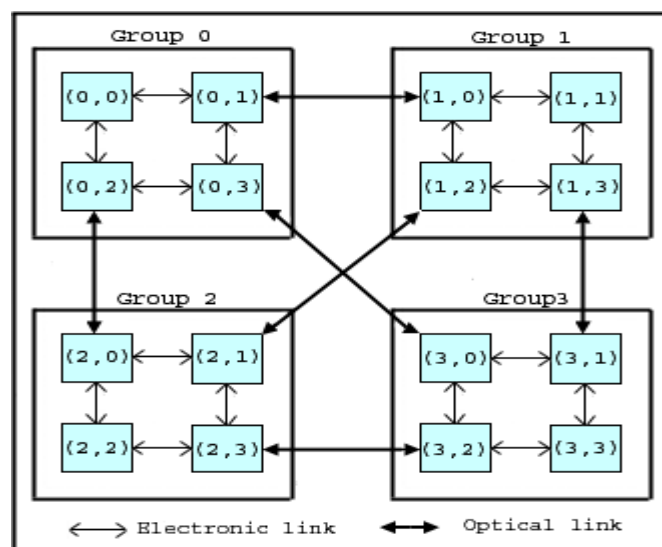
Figure 2.8: 4×4 OTIS-Mesh

2.4.2 OTIS-Mesh Properties

The topology of the interconnection network Culler (1998) is modeled as a graph, where processors present nodes and links present edges. The network topology determines properties, such as size, degree, diameter, etc; knowing these properties aids in choosing the appropriate topology to solve problems, such as sorting and matrix multiplication, in parallel systems. Many topological properties were investigated and derived on the OTIS-Mesh; Awwad *et al.* (2002) studied and defined the following topological properties (for N×N OTIS-Mesh):

a. Size is the number of nodes in the network equals to N^2.

b. Degree is the number of input/output channels of node (electronic and optic links):

$$\text{Degree} = \begin{cases} 4 & G = P \\ 5 & G \neq P \end{cases}$$

c. Diameter is the greatest distance between two nodes in the network.

Diameter = $4\sqrt{N}$-3

d. Number of links = (2*N)*N + (N^2-N)/2

2.4.3 OTIS-Mesh Routing

Routing Ni *et al.* (1991) is the process of selecting a path between two nodes in the network. Routing can be deterministic or adaptive (the message between two nodes can select one path among alternative paths). Awwad *et al.* (2002) implemented a deterministic routing algorithm for OTIS-Mesh that chooses the shortest path between two nodes. Najaf-Abadi *et al.* (2007) proposed an analytical performance model for fully adaptive wormhole routed mesh networks.

In this thesis, the routing in the OTIS-Mesh is deterministic. Intra group routing is selecting a path between two nodes within the same group, this is achieved by using the dimensional order routing where the message first routed in the X dimension then in the Y dimension. Whereas, Inter group routing is selecting a path between two nodes in different groups and it is achieved by the following steps:

1. The source node performs an intra group routing in order to transmit the message to the node that has an optical link with the group where the destination node exists.

2. An optical move is performed to transmit the message to the group where the destination node exists.

3. An intra group routing is performed to transmit the message to the destination node.

2.4.4 OTIS-Mesh Basic Operations

Many basic operations algorithms for OTIS-Mesh were developed. Wang *et al.*(1998) implemented many basic operations such as window broadcast operation where it starts with a data located in a submesh of one group, after that the data is transmitted to processors within group then the data is transmitted to the other groups in the OTIS-Mesh, data sum operation where it is performed on data items located in each processor in the OTIS-Mesh, data accumulation operation where each processor accumulates data from its neighboring processors, adjacent sum operation where it performs sum operation on the accumulated data, random access read operation where it reads data from one processor in the OTIS-Mesh, and random access write operation where it writes data on one processor in the OTIS-Mesh. The previous

operations are important in developing many applications on the OTIS-Mesh such as graph and matrix multiplication.

Rajaskekaran (1992) developed selection operation that finds the smallest number of a sequence of numbers distributed among the processor in the network, and Sorting operation that arranges a set of data distributed in ascending or descending order among the processors in the network. The previous operations are useful in numerical analysis applications.

Awwad *et al.* (2002) developed broadcast operation, where one processor transmits the same data message to each other processor in the OTIS-Mesh. The broadcast operation is one important operation for data movement in parallel and distributed systems.

2.4.5 OTIS-Mesh Applications

Many applications were developed on OTIS-Mesh such as:

- Matrix Multiplication algorithms Wang et al. (2001), several classes of multiplication algorithms were developed (such as multiplying two vectors, a vector and a matrix, and two matrices). In order to map the matrix into OTIS-Mesh the Group Row Major (GRM) mapping and Group Submatrix (GSM) mapping were used.

- Sorting algorithm Osterloh, (2000) presented k-k sorting algorithm (each processor in the OTIS-Mesh contain k elements), the algorithm is summarized by dividing the network into blocks, performing indexing and sorting blocks, after that the blocks are distributed in the network.

- Image processing algorithms: Wang (2000) developed algorithms that distribute a digital image on OTIS-Mesh processors such as histogramming algorithm which

calculates the histogram vector of a given image, and a histogram modification algorithm that changes the image histogram by using a mapping function, in addition to expanding and shrinking algorithms.

- Numerical analysis algorithms Jana (2006) developed polynomial interpolation and polynomial root finding, and row column mapping and group mapping were employed to arrange input data elements in the network. Moreover, Lagrange and Durand-Kerner methods were used to find the polynomial root.

3- Collective Communication in EDN OTIS-Mesh

This chapter presents the Extended Dominating Node Optical Transposed Interconnection System Mesh (EDN OTIS-Mesh) (Section 3.1), and illustrates how to perform the following collective communication operations: scatter, reduction and barrier on the EDN OTIS-Mesh in sections 3.2, 3.3, and 3.4, respectively. Also, section 3.5 presents the matrix transposition as an application on data movement collective communication.

3.1 EDN OTIS-Mesh

The OTIS-Mesh system consists of N groups and each group consists of N processing elements interconnected as mesh. Each processing element is presented as a tuple; Group number G and processor number P. Each processing element (G, P) is interconnected with its transposed processing element (P, G) with an optical link. The EDN approach is applied on each group in the OTIS-Mesh to form an EDN OTIS-Mesh. Figure 3.1 illustrates 16×16 EDN OTIS-Mesh, it consists of 16 groups (named G_0, G_1... G_{15}), each group consists of 16 processors (named 0, 1... 15) interconnected as mesh. Each processor is interconnected with its transposed processor via an optical link; for example, processor (0, 2) (that is Group 0, Processor 2) shares an optical link with processor (2, 0) (that is Group 2, Processor 0). 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 G_0 processors number 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, where level-0 nodes are the rest of the nodes.
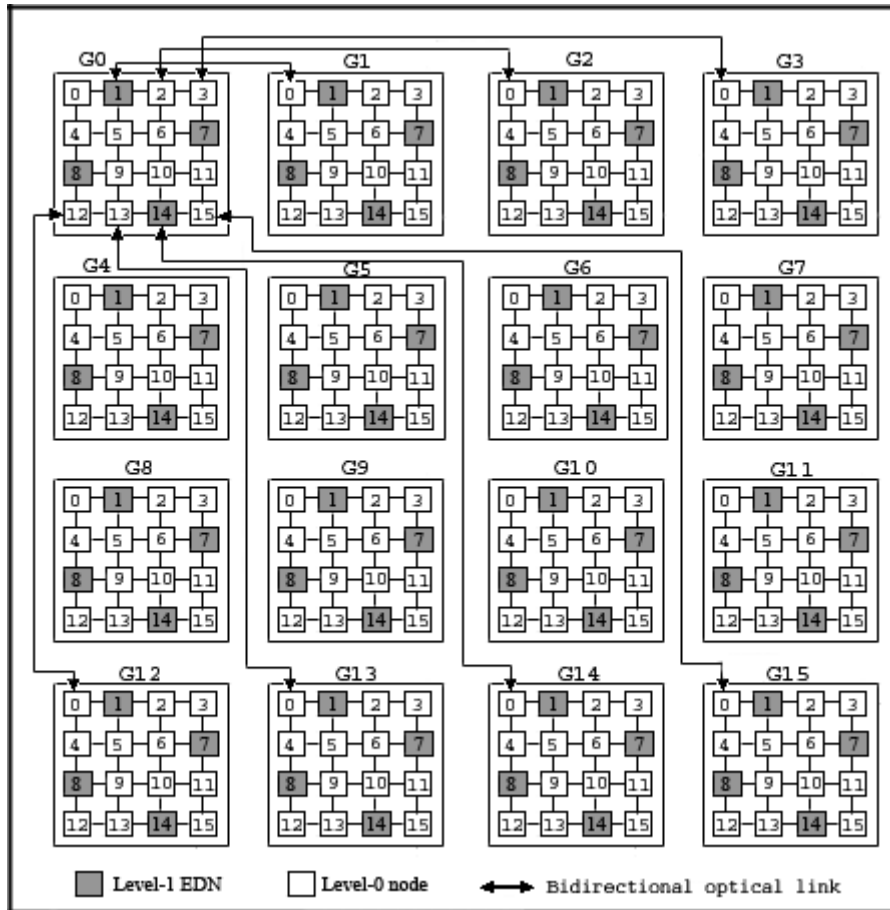
Figure 3.1: 16×16 EDN OTIS-Mesh.

3.2 Scatter

In scatter operation, one processor sends a different message for each processor in the EDN OTIS-Mesh. Scatter operation in the EDN OTIS-Mesh can be summarized by the following steps:

Step 1: the root processor assigns each group a set of messages and sends those messages to the processor that connects the control group with every other group. In order to accomplish that, the messages are first sent to the processors at the highest EDN level in the control group. Figure 4.2 illustrates the control group of 64×64 EDN OTIS-Mesh, in part (a) of the figure; the root processor (the square with a black circle) sends the messages to level-2 EDN processors (squares filled with gray color).

Next, part (b), the level-2 EDN processors send the messages to level-1 EDN processors (dot shaded squares). After that, in part (c), level-1 EDN processors send the received messages to level-0 nodes.
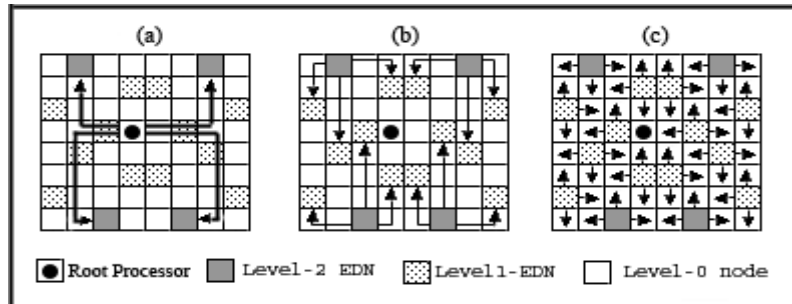
Figure 3.2: Step 1 - Scatter operation in 64×64 EDN OTIS-Mesh control group.

Step 2: processors at level-0 in the control group perform an optical move in order to send each group their messages. Figure 3.3 illustrates 64×64 EDN OTIS-Mesh, processor 0 of each group receives the messages through the optical link and sends those messages to level-2 EDN processors.
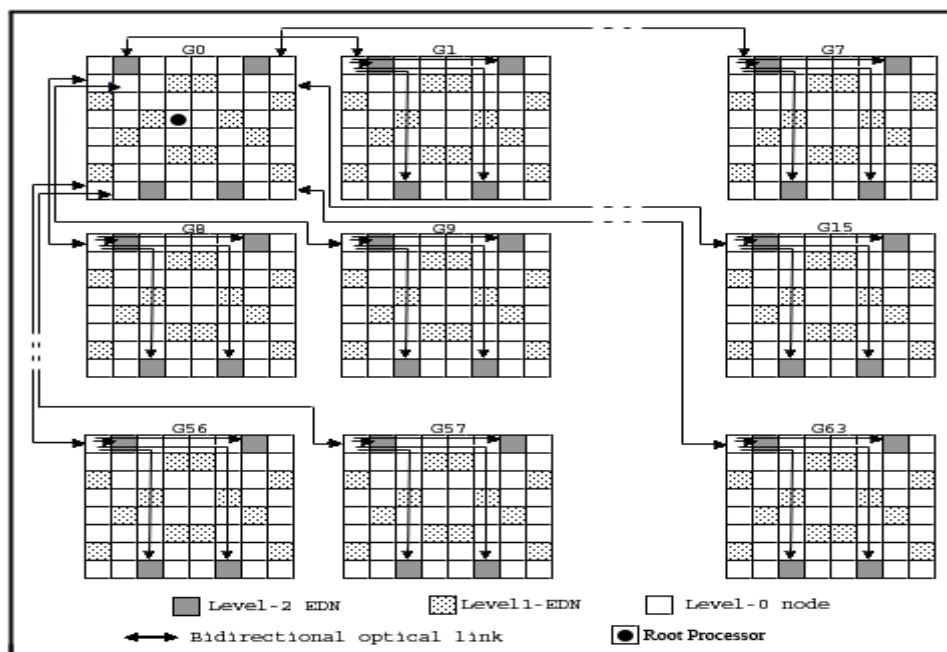
Figure 3.3: Step 2 - Scatter operation in 64×64 EDN OTIS-Mesh.

Step 3: within each group, the processors in the highest EDN level send the received messages to the next lower EDN processors. Figure 3.4 illustrates the scatter

operation in 64×64 EDN OTIS-Mesh where G_0 presents the control group, within each group (G_0 to G_{63}) level-2 EDN processors (squares filled with gray color) send the messages to level-1 EDN processors (dot shaded squares).
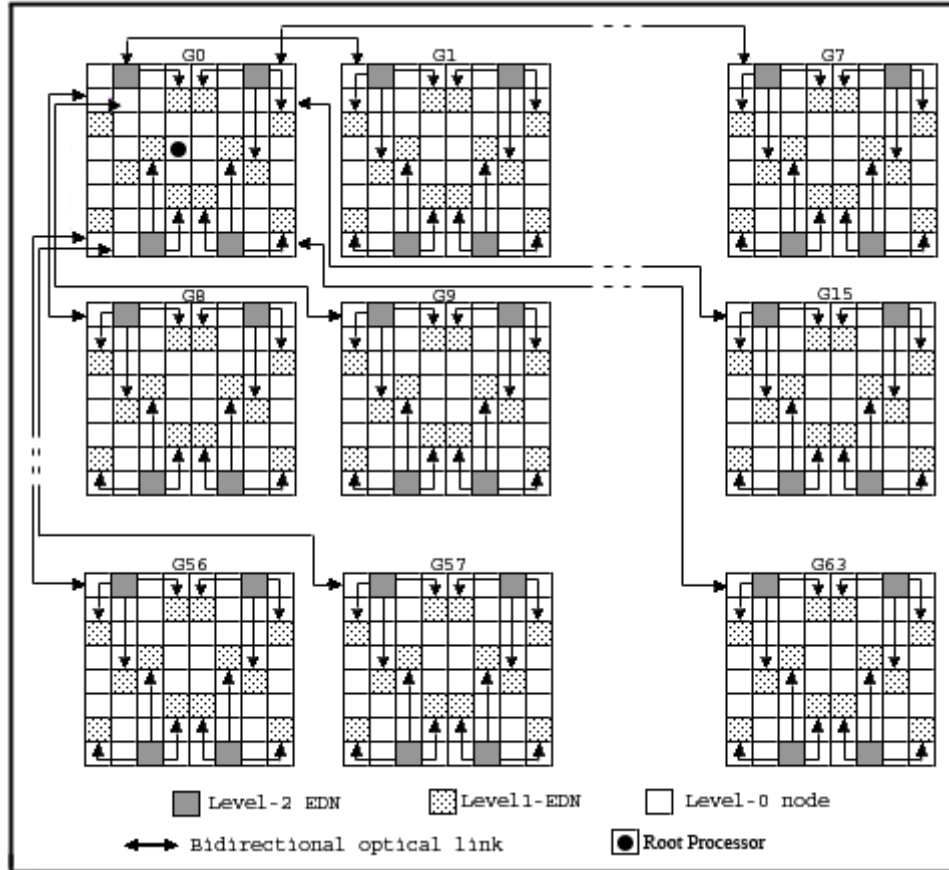
Figure 3.4: Step 3 - Scatter operation in 64×64 EDN OTIS-Mesh.

After that, as illustrated in Figure 3.5, the scatter operation in 64×64 EDN OTIS-Mesh continues as level-1 EDN processors (dot shaded squares) send the messages to level-0 nodes in parallel.
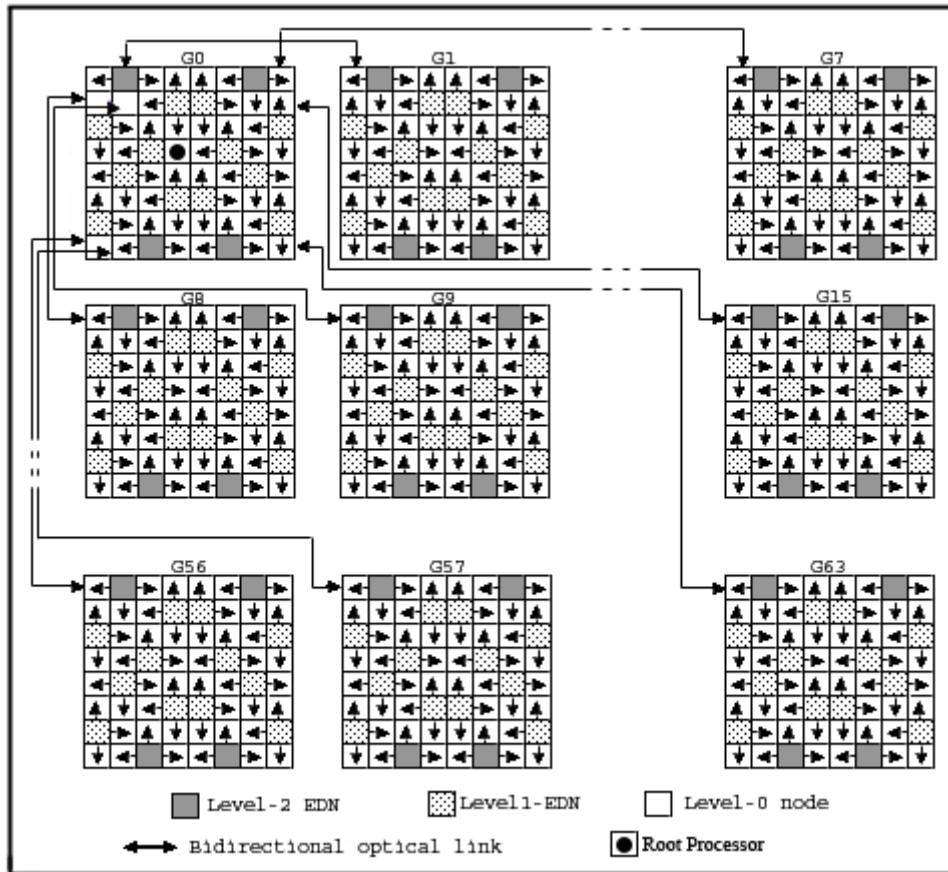
Figure 3.5 Step 3- Scatter operation in 64×64 EDN OTIS-Mesh.

3.3 Reduction

In reduction operation, a global operation such as maximum, minimum, sum, etc. is performed on data items located in each processor in the EDN-OTIS-Mesh. Reduction operation in the EDN OTIS-Mesh can be summarized in the following phases:

A. Reduction phase: every group within the EDN-OTIS-Mesh simultaneously starts to perform the compute operation as follows:

- Level-0 processors send their data to level-1 EDN processors where the compute operation is performed on the data items and the results are sent to the next EDN levels until the highest EDN level processors receive the results of computation. Figure 3.6 illustrates the reduction phase on 64×64 EDN OTIS-Mesh, where G_0

presents the control group. Within each group (G_0 to G_{63}) level-0 processors (white squares) send their data to level-1 EDN processors, where a global operation (such as maximum, minimum, sum, etc) is performed and the data result is then transmitted to level-2 EDN processors.
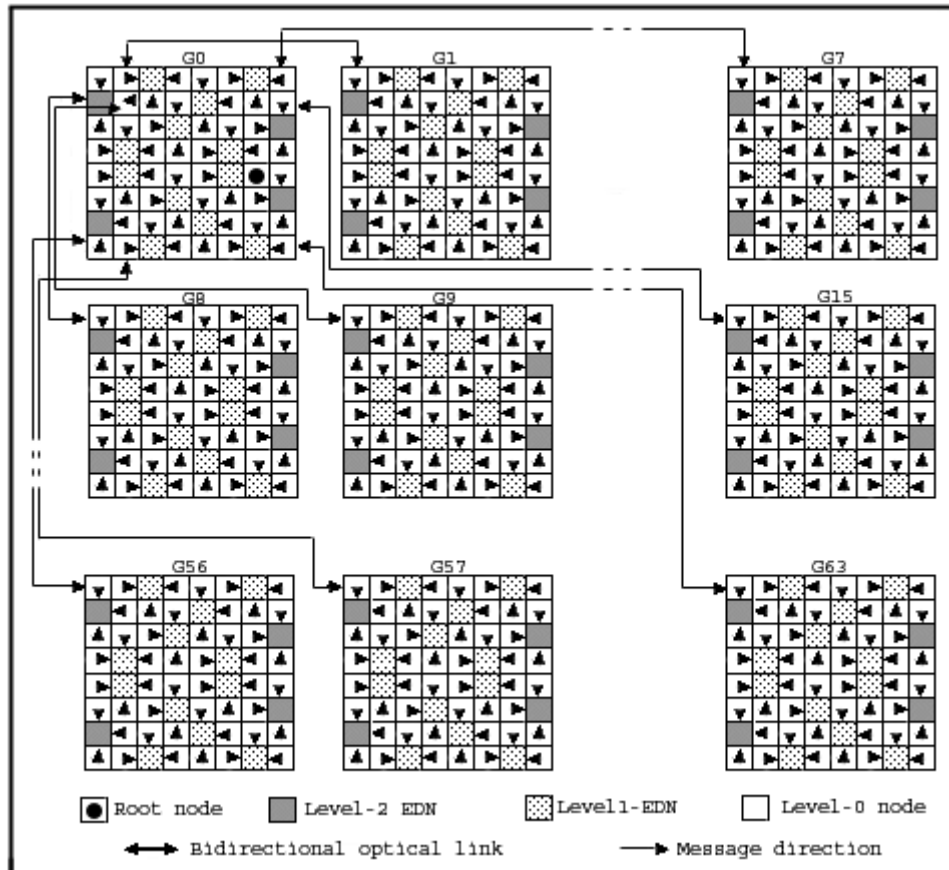
Figure 3.6: Reduction phase - reduction operation in 64×64 EDN OTIS-Mesh.

- The data results in the highest EDN level are then sent to the processor that has an optical link with the control group. The compute operation is performed on the results received from the highest EDN level processors and sent to the control group via the optical link. Figure 3.7 illustrates the reduction phase on 64×64 EDN OTIS-Mesh, where G_0 presents the control group. After level-2 EDN processors performed the global operation on the received data from level-1 EDN processors, the result is then transmitted to processor number 0 because this processor shares an optical link with the control group G_0.
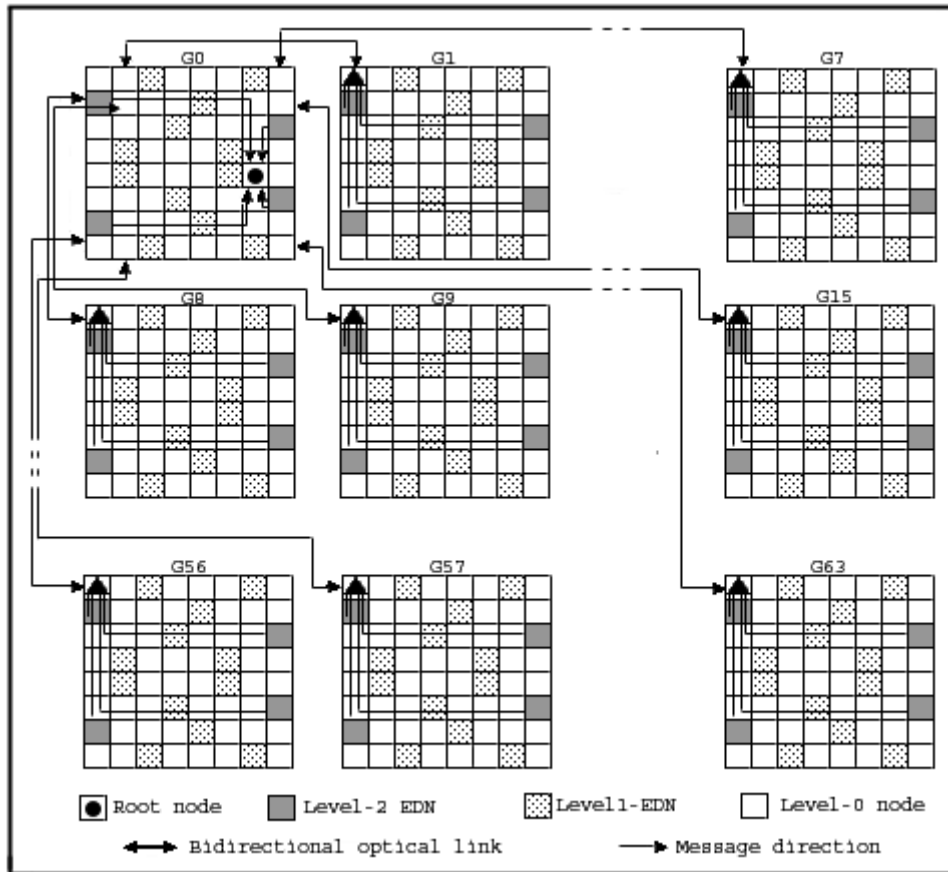
Figure 3.7: Reduction phase - Reduction operation in 64×64 EDN OTIS-Mesh .

B. Completion phase: the reduction phase is performed again in the control group; processors in the highest EDN level send their data to the root processor where the final computation is performed. Figure 3.8 illustrates the completion phase of the reduction operation in the 64×64 EDN OTIS-Mesh control group. Level-0 processors (white squares) send the result data (received by the optical link from the transposed group) to level-1 EDN processors (dot shaded squares) where the global computation is performed. Next, level-1 EDN processors send the result data to level-2 EDN processors (squares filled with gray color) where the global operation is performed on the received data. Finally, level-2 EDN processors send the data results to the root processor where the global computation is performed.
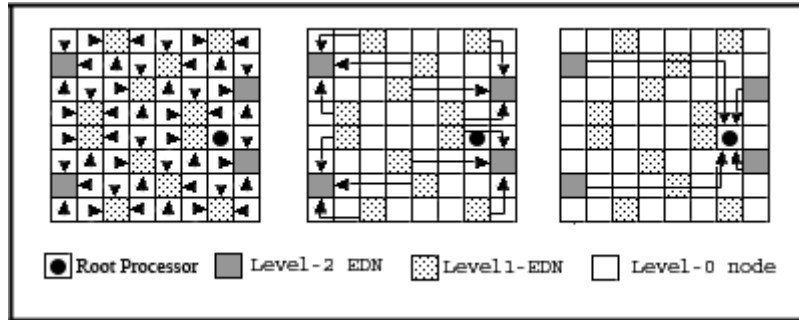
Figure 3.8: Completion phase - reduction operation in 64×64 EDN OTIS-Mesh.

3.4 Barrier

Barrier synchronization is a point in parallel algorithms, where all participating processors must reach before any processor proceeds execution. It consists of two phases: reduction and distribution. Moreover, barrier in EDN-OTIS-Mesh can be summarized in the following phases:

A. Reduction phase: the root processor acts as barrier. When each processor in the network reaches the synchronization point, it sends a message to the barrier processor. Reduction phase can be summarized in the following steps:

- Level-0 nodes send a synchronization message to level-1 EDN processors and block execution. Next, level-1 EDN processors send the collected synchronization message to the next EDN level and block execution until the nodes in the highest EDN level collect the entire synchronization message within the group. Figure 3.9 illustrates the reduction phase in 64×64 EDN OTIS-Mesh, where G_0 is the control group. Within each group (G_0 to G_{63}) level-0 processors send their synchronization messages to level-1 EDN processors. Next, level-1 EDN processors send the collected messages to level-2 EDN processors in order to be sent to the control group via processor number 0 as it shares an optical link with the control group.
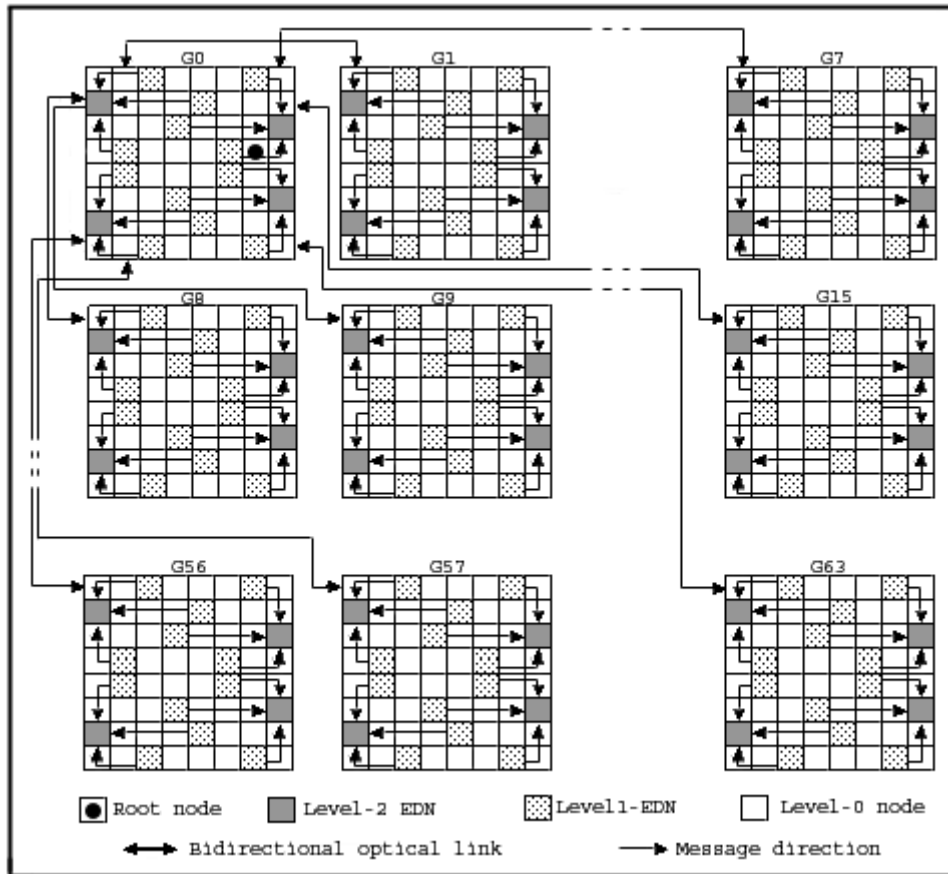
Figure 3.9: Reduction phase - barrier operation in 64×64 EDN OTIS-Mesh.

- The synchronization messages in the highest EDN level are then sent to the processor that has an optical link with the control group.

- The previous two steps are performed again in the control group. Processors that represent the highest EDN processors level send the collected synchronization messages to the root processor. Figure 3.10 illustrates the reduction phase in the control group. After the control group receives the messages via the optical link, level-0 nodes transmit the messages to level-1 EDN processors. Next level-1 EDN processors transmit the collected messages to level-2 EDN processors. Finally, level-2 EDN processors transmit the messages to the root processor (barrier).
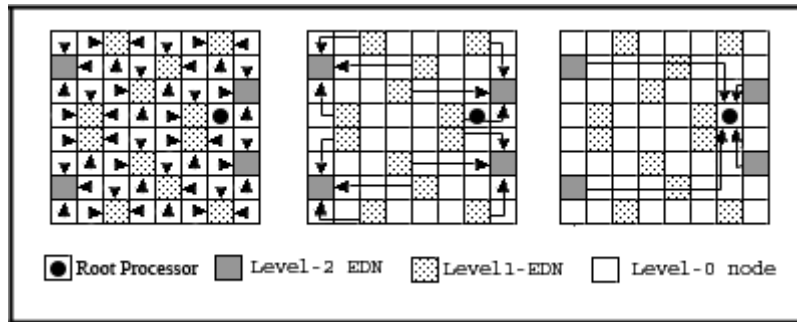
Figure 3.10: Reduction phase - barrier operation in 64×64 EDN OTIS-Mesh control group.

- The root processor counts the received messages and checks whether all the participating processors has sent a synchronization message.

B. Distribution phase: after all processors has reached the synchronization point and sent a message to the barrier processor, the root processor (barrier) sends a message to all participating processors so they can continue execution. Distribution phase can be summarized in the following steps:

- The root processor sends permission messages to the EDN processors in the highest level so they can continue their execution. Next, the EDN processors in the highest level send the messages to the next lower level EDN processors until level-0 processors receive the permission message. Figure 3.11 illustrates the distribution phase in the control group. The root processor sends level-2 EDN processors a permission message so they can continue execution. Next, level-2 EDN processors forward the permission message to level-1 EDN so they can continue execution. After that, level-1 EDN processors send the permission message to level-0 nodes.
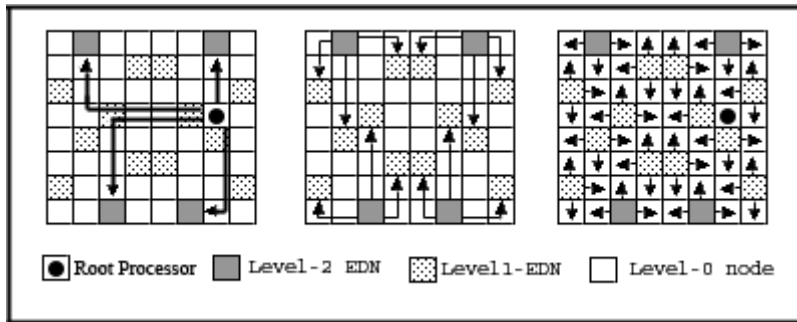
Figure 3.11: Distribution phase - barrier operation in 64×64 EDN OTIS-Mesh control group.

- Each processor has an optical link with another group sends a permission message to that group and the previous steps are performed again. Figure 3.12 illustrates distribution phase in 64×64 EDN OTIS-Mesh. Within each group of (G1 to G63), level-2 EDN processors (squares filled with gray color) send a permission message to level-1 processors (dot shaded squares). Next, level-1 EDN processors send the permission message to level-0 processors.
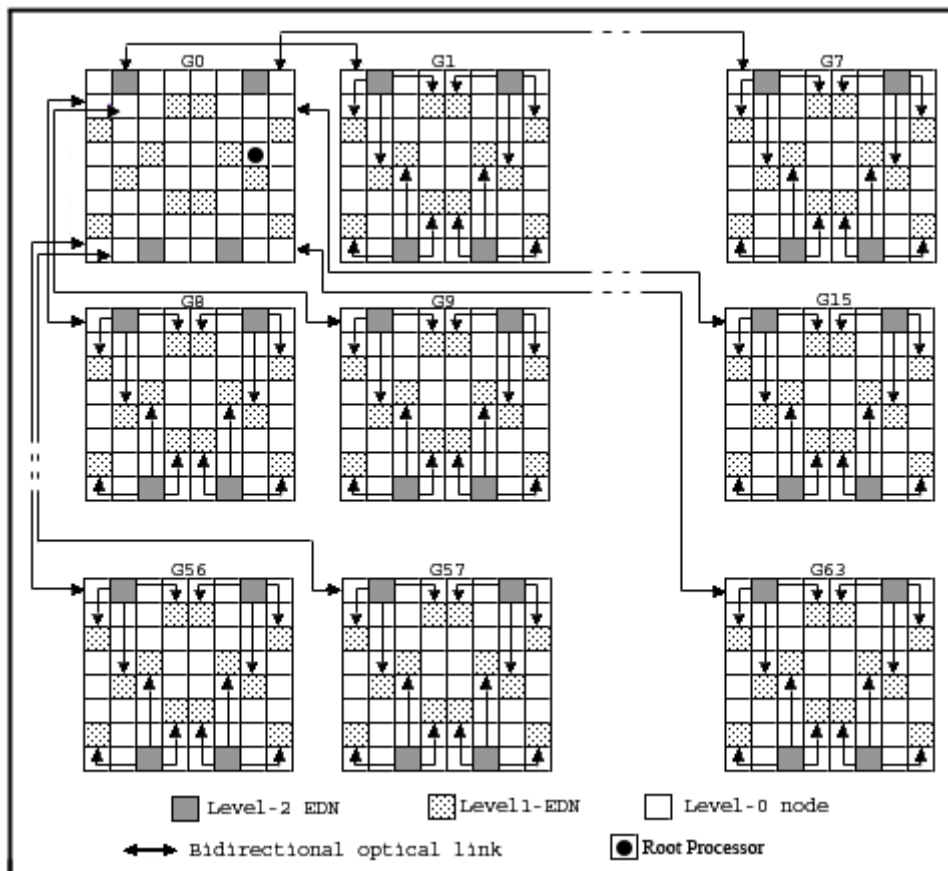
Figure 3.12: Distribution phase - barrier operation in 64×64 EDN OTIS-Mesh.

3.5 Application: Matrix Transposition in EDN OTIS-Mesh

Matrix transposition is the process of reflecting the data elements of a matrix (distributed among the processors in parallel system) by its main diagonal. Suppose that a square matrix $N_{n \times n}$ is distributed evenly on the processors of the EDN OTIS-Mesh. Figure 3.13 illustrates a square matrix distributed among 16×16 EDN OTIS-Mesh; each processor in the network holds a data block of the square matrix, the shaded squares present level-1 EDN processors and the white squares present level-0 processors.
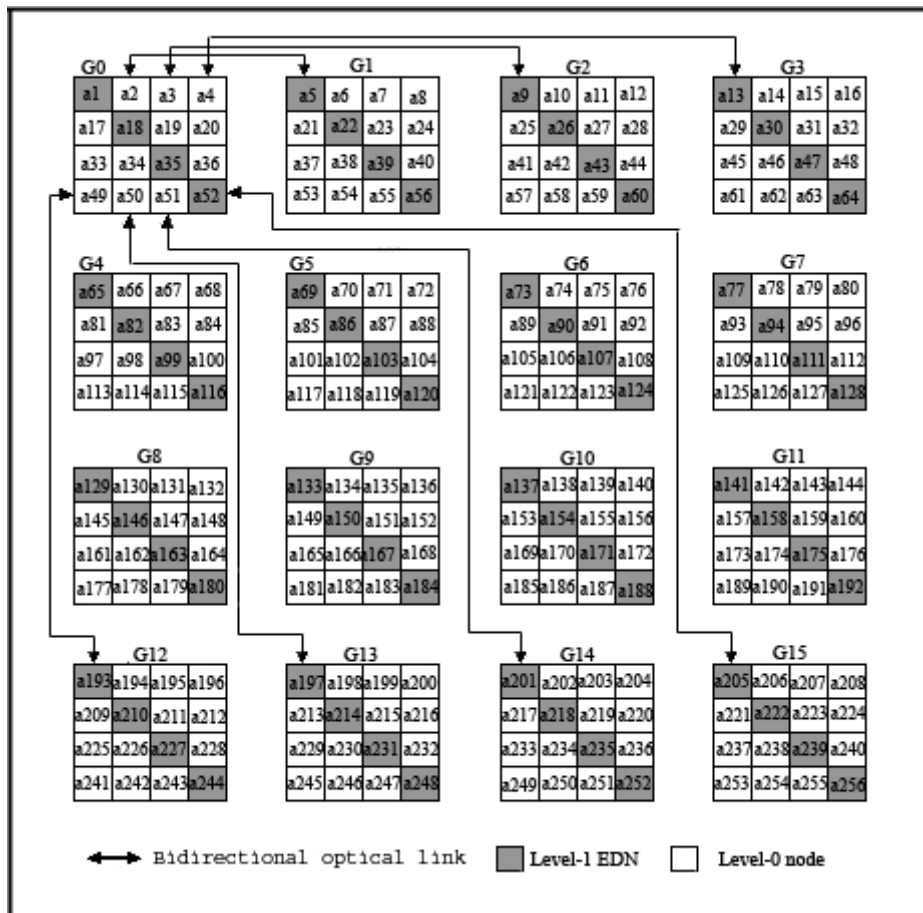
Figure 3.13: Square matrix $N_{n \times n}$ distributed on 16×16 EDN OTIS-Mesh.

Figure 3.14 illustrates a 16×16 EDN OTIS-Mesh; after the data blocks within each processor were transposed.
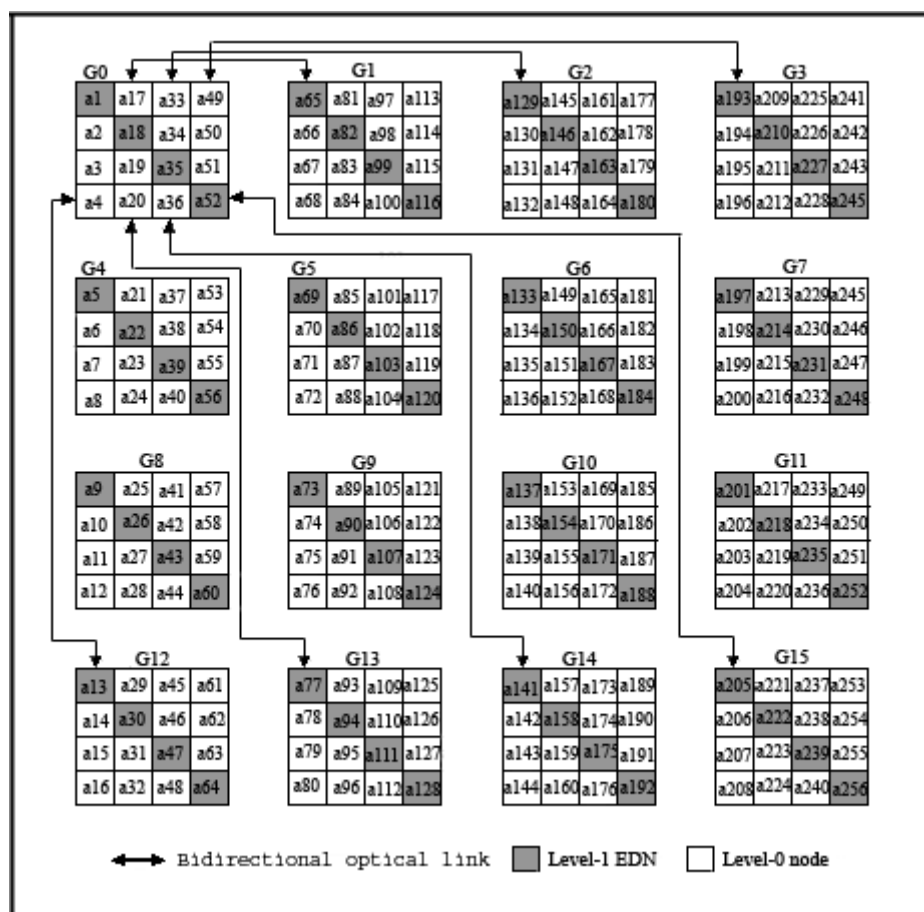
Figure 3.14: Transposed square matrix $N_{n \times n}$ distributed on 16×16 EDN OTIS-Mesh

Matrix transposition in EDN OTIS-Mesh can be summarized in the following steps:

Step 1: inside each group, processors within each 4×4 block send their data blocks to level-1 EDN processors. Figure 3.15 illustrates the first step of matrix transposition in 16×16 EDN OTIS-Mesh. Level-0 nodes (white squares) send their data blocks to the level-1 EDN processors (gray squares), where the message direction follows the black arrows within each group.
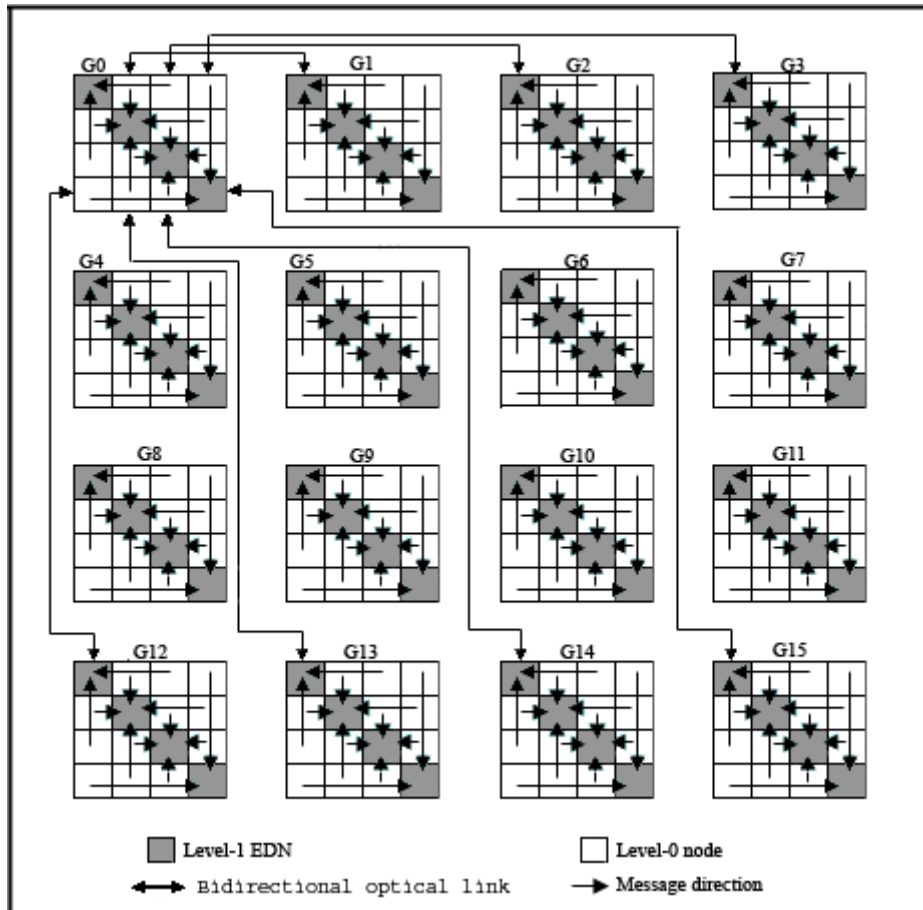
Figure 3.15: Matrix transposition - step 1.

Step 2: an optical move is performed to transmit the collected data in level-1 EDN of each group to another group located in the diagonal, assuming that the OTIS groups are arranged as two dimensional block as illustrated in Figure 3.16; the diagonal groups are G_0, G_5, G_{10} and G_{15}. Figure 3.18 illustrates 16×16 EDN OTIS-Mesh, where shaded squares present level-1 EDN processors, and solid shaded arrow presents the message direction, for example the EDN processors of G_3 and the EDN processors of G_{12} send their data blocks to the EDN processors in G_{15}.
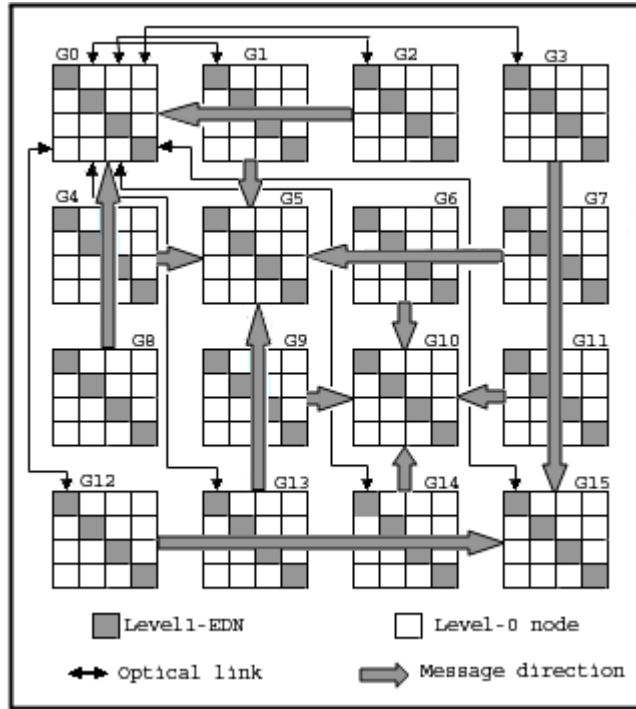
Figure 3.16: Matrix transposition - step 2.

Step 3: in this step, the opposite of step 2 process is performed; after the diagonal groups receive the data blocks, Level-1 EDN processors in each diagonal group perform an optical move to send the received data blocks to the appropriate groups. Figure 3.17 illustrates a 16×16 EDN OTIS-Mesh, where shaded squares present level-1 EDN processors and solid shaded arrows present the message direction for example, after G_{15} has received data blocks from groups G_3 and G_{12}, G_{15} sends the G_3 data block to G_{12} and sends G_{12} data block to G_3.
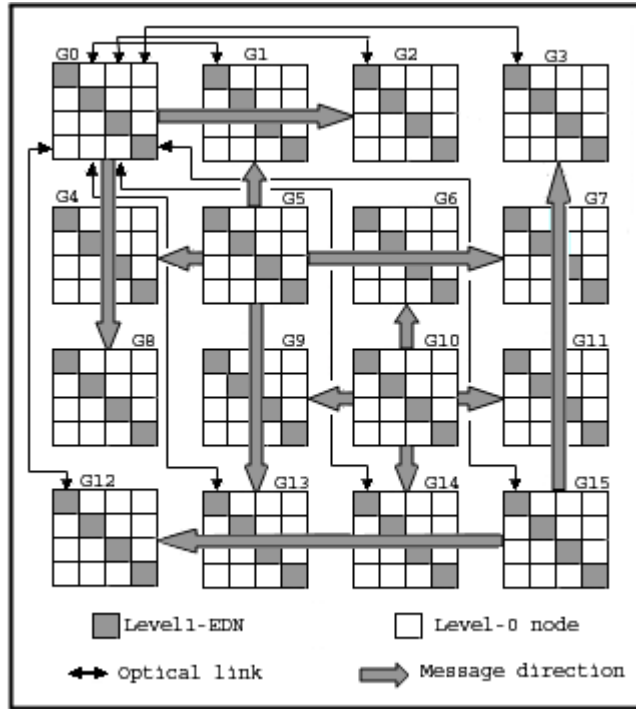
Figure 3.17: Matrix transposition - step 3.

Step 4: within each group, level-1 EDN processors send the received data blocks to the appropriate processors. Figure 3.18 illustrates the final step of matrix transposition in 16×16 EDN OTIS-Mesh, where level-1 EDN processors send the received data blocks to level-0 nodes and the message direction follows the black arrows within each group.
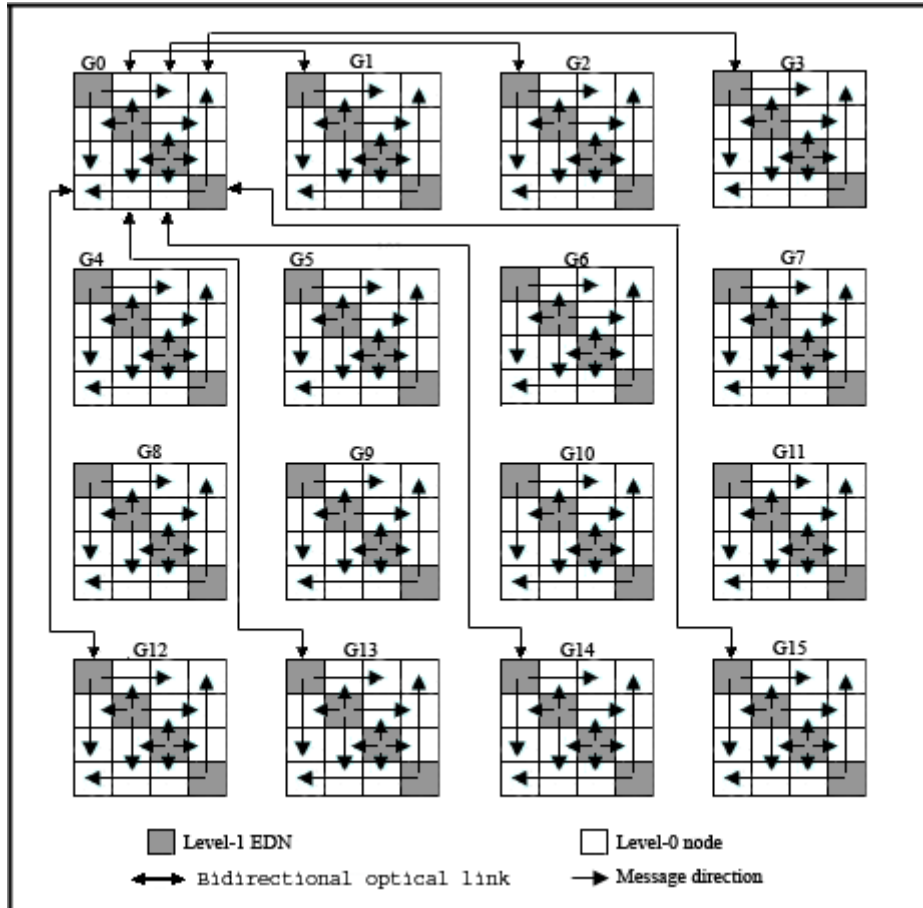
Figure 3.18: Matrix transposition - step 4.

4- Analytical Modeling

This chapter discusses the parameters that affect the performance of the collective communication's operations, presents the interconnection system assumptions and notations, and derives an analytical model to evaluate the performance metrics of interest; number of steps, latency, execution time, performance improvement, total overhead, and cost. The analytical model is used in building the interconnection network simulator and in conducting the performance evaluation experiments.

The analytical model assumptions are illustrated in section 4.1, the performance metric parameters: number of steps, latency, execution time, performance improvement, total overhead and cost are illustrated in sections 4.2, 4.3, 4.4, 4.5, 4.6 and 4.7, respectively.

4.1 Assumptions

The following are the assumptions used to evaluate the performance metrics:

1. Message length is equal to M.

2. Setup time t_{set} is the time needed to start an operation.

3. Router delay is characterized by the following parameters:

   a. Channel cycle time t_c is the time needed to transmit one flit from one router to the next.

   b. Switching delay t_s is the time needed to transmit a flit from the input channel to the required output channel in the crossbar.

   c. Routing time t_r is the time needed to route a message.

The overall router delay t_{period} (Equation 4.1) is the maximum of (t_c, t_s, t_r) since the three operations are overlapped.

$$\mathbf{t_{period}} = \text{Max } (t_c, t_s, t_r) \tag{4.1}$$

4. The number of virtual channels V is used per a physical channel.

5. Contention time t_{con} is the time which a flit spends in the router before it is transmitted to the next router due to channel contention.

6. Computation time t_{comp} is the time needed by the processor element to perform a computation.

7. Communication latency T_{lat} (Equation 4.2) is the time needed to transmit a message from the source node to the destination node. T_{lat} is affected by the number of hops D between the source and distention, message size M, contention time t_{con}, and router delay presented by t_c, t_s and t_r.

$$\mathbf{T_{lat}} = \quad t_{period} (D + M) + t_{con} \quad\quad\quad\quad (4.2)$$

8. The routing algorithm used is deterministic, the dimension order routing traverses the X dimension first then it traverses the Y dimension.

9. The number of nodes in the OTIS-Mesh group is equal to N (Equation 4.3).

$$N = 4^i \quad\quad\quad\quad\quad i = 2, 3\ldots n \quad\quad\quad\quad (4.3)$$

Where 4 is the number of EDN nodes in the smallest OTIS-Mesh group and i starts from 2 in order to have at least a group with N = 16 and one EDN level of 4 nodes.

10. The height of EDN tree (i.e. the number of EDN levels) within each EDN OTIS-Mesh group is equal to k (Equation4.4).

$$k = (\log_4 N) – 1 \quad\quad\quad\quad\quad\quad (4.4)$$

Where N is the number of nodes in the OTIS-Mesh group, $(\log_4 N)$ since 4 is the number of EDN nodes in the smallest OTIS-Mesh group.

4.2   Number of Steps

The number of steps to perform the collective communication operations is the sum of electronic message passing steps needed to complete the operation. In the all port OTIS-Mesh and the all port EDN OTIS-Mesh; the number of steps is affected by the root node location, so the minimum and maximum number of steps is calculated. The minimum number of steps (the best case) is obtained when the root node is located in 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 4.1 presents one group of 64×64 OTIS-Mesh group, the shaded processor is the root; it is located in the middle of the group thus presenting the best case.
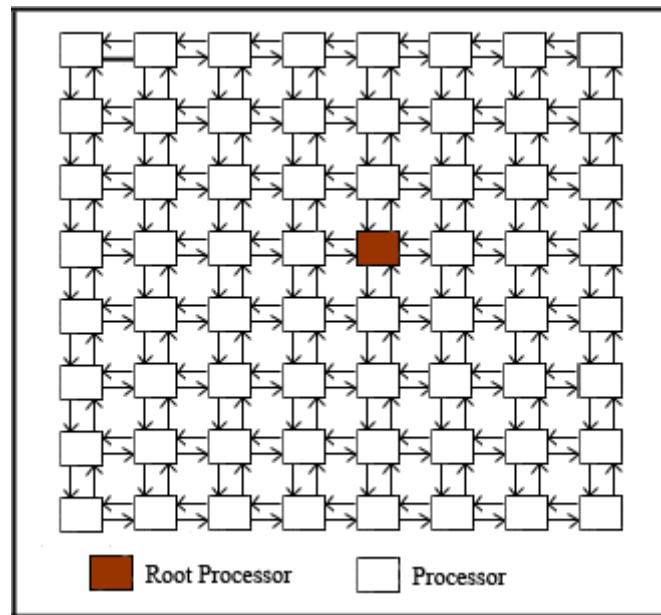
Figure 4.1: 64×64 OTIS-Mesh control group (best case).

The maximum number of steps (the worst case) is obtained when the root node is located in the endmost of the control group, which allows the root node to perform a limited number of simultaneous send/receive from all of its input/output channels. Figure 4.2 presents one group of 64×64 OTIS-Mesh group, the shaded processor is the root; it is located in the endmost of the group thus presenting the worst case.
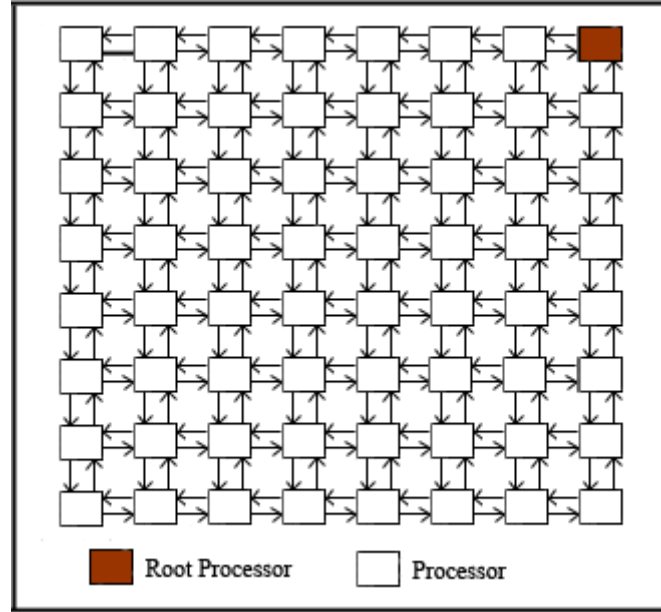
Figure 4.2: 64×64 OTIS-Mesh control group (worst case).

4.7.1 Scatter

The number of electronic message passing steps needed to perform scatter on single port OTIS-Mesh is given by Equation 4.5.

$$\text{Scatter\_Steps}_{\text{single\_port}} = 2*(N-1) \tag{4.5}$$

The number of required steps to perform scatter in the control group is equal to N-1 and multiplied by 2 because performing scatter on the other groups (in parallel) is N-1 steps.

The minimum number (best case) of electronic message passing steps needed to perform scatter in all port OTIS-Mesh is given by Equation 4.6.

$$\text{Scatter\_Min\_Steps}_{\text{all\_port}} = 2*\left(\frac{\sqrt{N}}{2}*\sqrt{N}\right) \tag{4.6}$$

The number of required steps to perform scatter in the control group is equal to $\left(\dfrac{\sqrt{N}}{2} * \sqrt{N}\right)$ and multiplied by 2 because performing scatter on the other groups (in parallel) is $\left(\dfrac{\sqrt{N}}{2} * \sqrt{N}\right)$ steps.

However, the maximum number (worst case) of electronic message passing steps needed to perform scatter in all port OTIS-Mesh is given by Equation 4.7.

$$\text{Scatter\_Max\_Steps}_{\text{all\_port}} = 2 * ((\sqrt{N} - 1) * \sqrt{N}) \tag{4.7}$$

The maximum number of required steps to perform Scatter in the control group is equal to $((\sqrt{N} - 1) * \sqrt{N})$ and multiplied by 2 because performing scatter on the other groups (in parallel) is $((\sqrt{N} - 1) * \sqrt{N})$ steps.

The minimum number (best case) of electronic message passing steps needed to perform scatter in all port EDN OTIS-Mesh is given by Equation 4.8.

$$\text{Scatter\_Min\_Steps}_{\text{all\_port\_EDN}} = 2 * \left(\dfrac{\sqrt{N}}{2} * \sqrt{N}\right) \tag{4.8}$$

The minimum number of required steps to perform scatter in the control group is equal to $\left(\dfrac{\sqrt{N}}{2} * \sqrt{N}\right)$ and multiplied by 2 because performing Scatter on the other groups (in parallel) is $\left(\dfrac{\sqrt{N}}{2} * \sqrt{N}\right)$ steps.

However, the maximum number (worst case) of electronic message passing steps needed to perform scatter in all port EDN OTIS-Mesh is given by Equation 4.9.

$$\text{Scatter\_Max\_Steps}_{\text{all\_port\_EDN}} = 2*(N-1) \qquad (4.9)$$

The maximum number of required steps to perform scatter in the control group is equal to (N-1) and multiplied by 2 because performing Scatter on the other groups (in parallel) is (N-1) steps.

4.7.2 Reduction

The number of electronic message passing steps needed to perform reduction on single port OTIS-Mesh is given by Equation 4.10.

$$\text{Reduction\_Steps}_{\text{single\_port}} = 2*(N-1) \qquad (4.10)$$

The number of required steps to perform the reduction phase is equal to N-1 and multiplied by 2 because the completion phase takes N-1 steps.

The minimum number (best case) of electronic message passing steps needed to perform reduction on all port OTIS-Mesh is given by Equation 4.11.

$$\text{Reduction\_Min\_Steps}_{\text{all\_port}} = 2*\left(\frac{\sqrt{N}}{2}*\sqrt{N}\right) \qquad (4.11)$$

The minimum number of required steps to perform the reduction phase is equal to $\left(\frac{\sqrt{N}}{2}*\sqrt{N}\right)$ and multiplied by 2 because the completion takes $\left(\frac{\sqrt{N}}{2}*\sqrt{N}\right)$ steps.

However, the maximum number (worst case) of electronic message passing steps needed to perform reduction on all port OTIS-Mesh is given by Equation 4.12.

$$\text{Reduction\_Max\_Steps}_{\text{all\_port}} = 2*((\sqrt{N}-1)*\sqrt{N}) \qquad (4.12)$$

The maximum number of required steps to perform the reduction phase is equal to ($\sqrt{N}$ -1) * $\sqrt{N}$ and multiplied by 2 because the completion phase takes ($\sqrt{N}$ -1) * $\sqrt{N}$ steps.

The minimum number (best case) of electronic message passing steps needed to perform reduction on all port EDN-OTIS-Mesh is given by Equation 4.13.

$$\text{Reduction\_Min\_Steps}_{\text{all\_port\_EDN}} = 2*(k+2) \qquad (4.13)$$

The minimum number of required steps to perform the Reduction phase is equal to k+2 and multiplied by 2 because the completion phase takes k+2 steps.

However, the maximum number (worst case) of electronic message passing steps needed to perform Reduction on all port EDN-OTIS-Mesh is given by Equation 4.14.

$$\text{Reduction\_Max\_Steps}_{\text{all\_port\_EDN}} = 2*(k+3) \qquad (4.14)$$

The maximum number of required steps to perform the Reduction phase is equal to k+3 and multiplied by 2 because the completion phase takes k+3 steps.

4.7.3 Barrier

The number of electronic message passing steps needed to perform barrier on single port OTIS-Mesh is given by Equation 4.15.

$$\text{Barrier\_Steps}_{\text{single\_port}} = 4*(N-1) \qquad (4.15)$$

The number of required steps to perform the reduction phase is equal to 2*(N-1) and the number of required steps to perform the distribution phase is takes 2*(N-1), so collectively the number of steps is equal to 4*(N-1).

The minimum number (best case) of electronic message passing steps needed to perform barrier on all port OTIS-Mesh is given by Equation 4.16.

$$\text{Barrier\_Min\_Steps}_{\text{all\_port}} = 4*\left(\frac{\sqrt{N}}{2} * \sqrt{N} \right) \qquad (4.16)$$

The minimum number of required steps to perform the reduction phase is equal to $2*\left(\frac{\sqrt{N}}{2} * \sqrt{N} \right)$ and the number of required steps to perform the distribution phase takes $2*\left(\frac{\sqrt{N}}{2} * \sqrt{N} \right)$, so collectively the number of steps is equal to $4*\left(\frac{\sqrt{N}}{2} * \sqrt{N} \right)$.

However, the maximum number (worst case) of electronic message passing steps needed to perform Barrier on all port OTIS-Mesh is given by Equation 4.17.

$$\text{Barrier\_Max\_Steps}_{\text{all\_port}} = 4*((\sqrt{N}-1)*\sqrt{N}) \qquad (4.17)$$

The maximum number of required steps to perform the reduction phase is equal to $2*((\sqrt{N}-1)*\sqrt{N})$ and the number of required steps to perform the distribution phase is takes $2*((\sqrt{N}-1)*\sqrt{N})$, so collectively the number of steps is equal to $4*((\sqrt{N}-1)*\sqrt{N})$.

The minimum number (best case) of electronic message passing steps needed to perform barrier on all port EDN OTIS-Mesh is given by Equation 4.18.

$$\text{Barrier\_Min\_Steps}_{\text{all\_port\_EDN}} = 4*(k+2) \qquad (4.18)$$

The minimum number of required steps to perform the reduction phase is equal to $2*(k+2)$ and the number of required steps to perform the distribution phase takes $2*(k+2)$, so collectively the number of steps is equal to $4*(k+2)$.

However, the maximum number (worst case) of electronic message passing steps needed to perform Barrier on all port EDN OTIS-Mesh is given by Equation 4.19.

$$\text{Barrier\_Max\_Steps}_{\text{all\_port\_EDN}} = 4*(k+3) \qquad (4.19)$$

The maximum number of required steps to perform the reduction phase is equal to $2*(k+3)$ and the number of required steps to perform the distribution phase takes $2*(k+3)$, so collectively the number of steps is equal to $4*(k+3)$.

4.3   Latency

The latency of the collective communication operations that have no computation, such as scatter, barrier, etc., is the time elapses from the beginning of communication until the moment the last processor finishes communication, and is given by the latency of the last node finishes communication (i.e. the node that has the maximum latency value), as shown in Equation 4.20.

$$\text{Latency} = t_{set} + \text{Max} (T_{lat}) \qquad (4.20)$$

The latency is equal to the sum of the start up time and the T_{lat} (where it is defined in Equation 4.2) of the last node that finishes communication.

4.4   Execution Time

The execution time for collective communication that contains computation, such as reduction, is the time elapses from the beginning of the parallel computation until the moment the last processing element finishes execution. Equation 4.21 presents the execution time.

$$\text{Execution Time} = t_{set} + t_{comp} + T_{lat} \qquad (4.21)$$

The execution time of the collective communication operation is obtained from summing the start up time, the computation time, and the communication latency time.

4.5 Performance Improvement

Performance improvement is the improvement gained from using the all port EDN-OTIS Mesh over single port OTIS or all port OTIS.

In reduction operation, the performance improvement is the ratio of the execution time of single port OTIS-Mesh or all port OTIS-Mesh over all port EDN OTIS-Mesh as shown in Equations 4.22 and 4.23, respectively.

$$\text{Improv\_over}_{Single\_Port} = \text{Exectution}_{SinglePort} / \text{Exectution}_{All\_Port\_EDN} \qquad (4.22)$$

The performance improvement of the all port EDN OTIS-Mesh over the single port OTIS-Mesh on the reduction operation is equal to the ratio of the execution time in the single port OTIS-Mesh over the execution time in all port OTIS-Mesh.

$$\text{Improv\_over}_{All\_Port} = \text{Exectution}_{All\_Port} / \text{Exectution}_{All\_Port\_EDN} \qquad (4.23)$$

The performance improvement of the all port EDN OTIS-Mesh over the all port OTIS-Mesh on the reduction operation is equal to the ratio of the execution time in the all port OTIS-Mesh over the execution time in all port OTIS-Mesh.

In scatter and barrier, the performance improvement is the ratio of the latency of single sort OTIS-Mesh or all port OTIS-Mesh over all port EDN OTIS-Mesh as shown in Equations 4.24 and 4.25, respectively.

$$\text{Improv\_over}_{Single\_Port} = \text{Latency}_{Single\_Port} / \text{Latency}_{All\_Port\_EDN} \qquad (4.24)$$

The performance improvement of the all port EDN OTIS-Mesh over the single port OTIS-Mesh on the operations scatter, and barrier, is equal to the ratio of the latency in the single port OTIS-Mesh over the latency in all port OTIS-Mesh.

$$\text{Improv\_over}_{\text{All\_Port}} = \text{Latency}_{\text{All\_Port}} / \text{Latency}_{\text{All\_Port\_EDN}} \qquad (4.25)$$

The performance improvement of the all port EDN OTIS-Mesh over the all port OTIS-Mesh on the operations scatter, and barrier, is equal to the ratio of the latency in the single port OTIS-Mesh over the latency in all port OTIS-Mesh.

4.6   Total Overhead

The total overhead is the passed time when processors are not performing computation operations due to contention and communication. The total overhead is evaluated for operations that have a computation part such as reduction. Equation 4.26 presents the total overhead.

$$\text{Total\_Parallel\_Overhead} = \text{Execution\_Time} - t_{\text{comp}} \qquad (4.26)$$

The total overhead evaluated by subtracting the total time processors spends in performing computation from the execution time.

4.7   Cost

Cost is defined as the sum of the time that each processing element spends solving the problem. Equation 4.27 presents cost.

$$\text{Cost} = \sum t_{\text{comp}} \qquad (4.27)$$

The cost of the collective communication operation that involves computation is obtained by summing the computation time.

5- Experimental Environment

This chapter presents the experimental environment of our work, discusses the specifications of the hardware and software used in implementing the interconnection network simulator and explains the simulator parts, classes, operations, parameters, etc. Prior implementing the simulator, a study was conducted to analyze the parameters that affect the simulation, such as port model, network switching method, parameters, etc., and accordingly a suitable development environment and tools were chosen.

The hardware experimental environment is illustrated in section 5.1, the software experimental environment is illustrated in section 5.2, and the interconnection network simulator is illustrated in section 5.3, where simulator parts, port model, interconnection network switching method, simulator classes, and simulator operations are illustrated in subsections 5.3.1 to 5.3.5.

5.4   Hardware Experimental Environment

The following are the hardware specifications that are used in the performance evaluation experiments:

- Processor: Intel(R) Core(TM) 2 Duo 1.5 GHz.

- Memory: 2038 MB.

- Cache memory: 2 MB L2.

- System type: 32-bit Operating System.

- Data bus speed: 667 MHz.

- Multithreaded architecture processor.

- Pipeline stages: 14 stages.

5.5 Software Experimental Environment

The following are the software specifications that are used in the performance
evaluation experiments:

- Platform: windows VISTA

- Programming language: C++

- Development tool: Microsoft Visual Studio 6.0

5.6 Interconnection Network Simulator

The OTIS-Mesh is a discrete event simulator that is developed in C++
programming language to capture the characteristics of the single port OTIS-Mesh, all
port OTIS-Mesh and all port EDN OTIS-Mesh in order to measure and compare the
performance metrics of the collective communication operations under the previous
systems.

5.6.1   Simulator Parts

The OTIS-Mesh simulator consists of the following parts: node, mesh, OTIS-
Mesh and EDN OTIS-Mesh.

- Node: depicts the main characteristics of the processing element; a local memory,
  four pointers present the input channels, four pointers present the output channels,
  as illustrated in Figure 5.1.
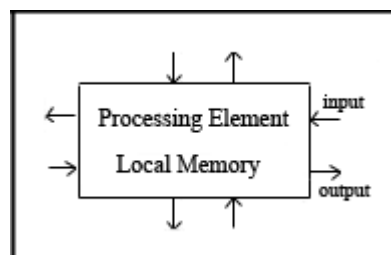
Figure 5.1: Processing element.

- Mesh: presents one OTIS-Mesh group, where there are N processing elements interconnected, as a mesh as illustrated in Figure 5.2.
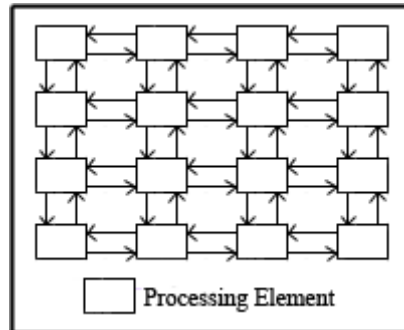
Figure 5.2: 4×4 2D Mesh.

- OTIS-Mesh: presents an OTIS-Mesh system that consists of N groups and each group consists of N processing elements interconnected as mesh. Each processing element is presented as a tuple; Group number G and Processor number P. Each processing element (G, P) is interconnected with its transposed processing element (P, G) with an optical link. Figure 5.3 illustrates 16×16 OTIS-Mesh, it consists of 16 groups (named G_0, G_1… G_{15}), each group consists of 16 processors (named 0,1,…, 15) interconnected as mesh. Each processor is interconnected with its transposed processor via an optical link; for example, processor (0, 2) (that is Group 0, Processor 2) shares an optical link with processor (2, 0) (that is group 2, processor 0).
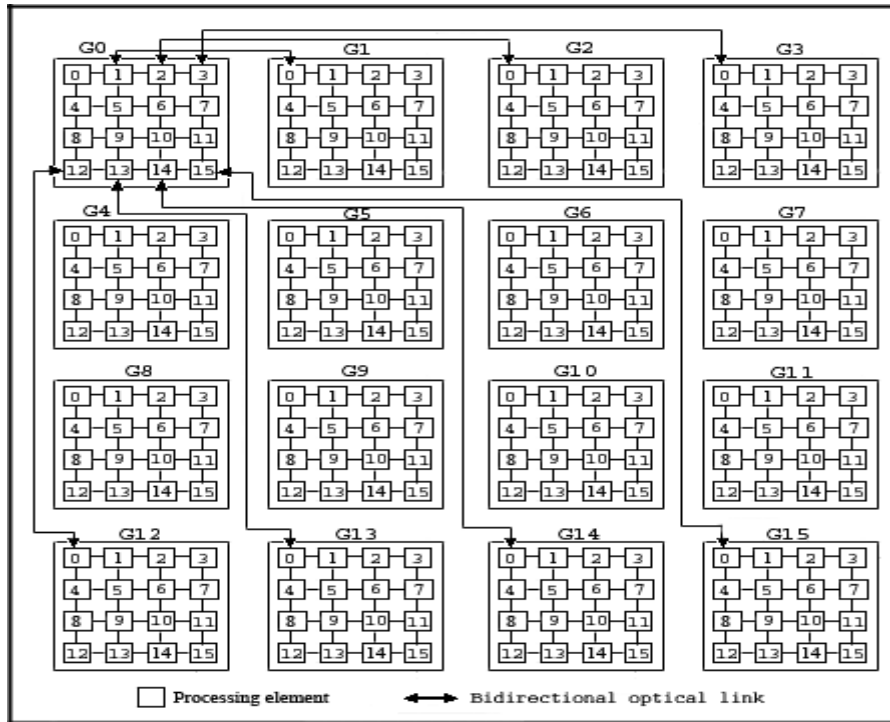
Figure 5.3: 16×6 OTIS-Mesh.

- EDN OTIS-Mesh: is an OTIS-Mesh system where a set of processing elements within each group is assigned as EDN processors. Figure 5.4 illustrates a 16×16 EDN OTIS-Mesh, it consists of 16 groups (named G_0, G_1… G_{15}), each group consists of 16 processors (named 0, 1… 15) interconnected as mesh. Each processor is interconnected with its transposed processor via an optical link; for example, processor (0, 2) shares an optical link with processor (2, 0). The shaded processors in each group present 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 G_0 processors number 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, where level-0 nodes are the rest of the nodes.
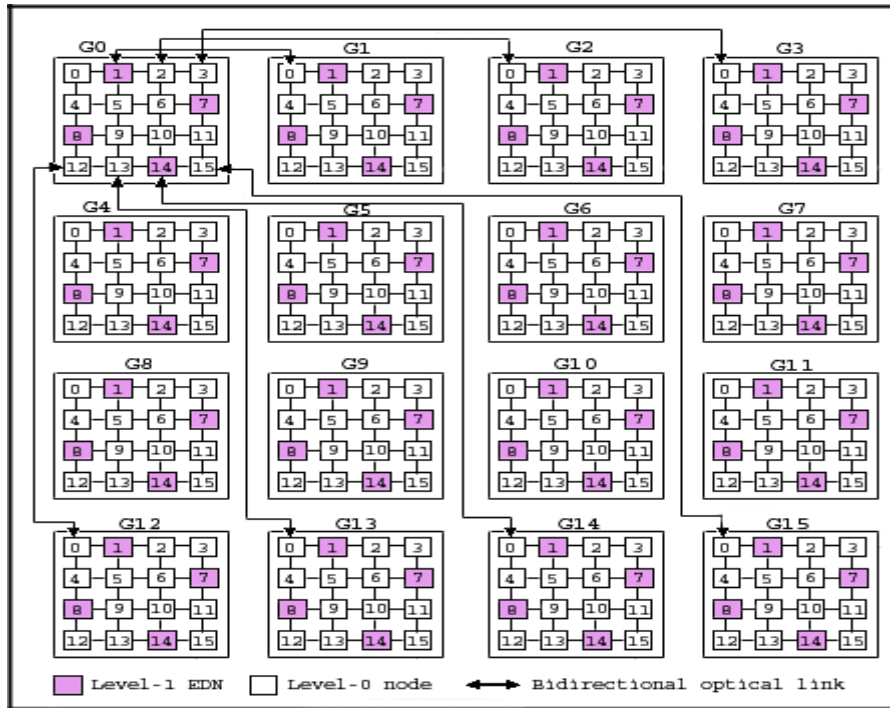
Figure 5.4: 16×16 EDN OTIS-Mesh.

5.6.2 Port Model

The node in the OTIS-Mesh simulator is either configured as a single port model; can perform one send/receive at a time, as illustrated in Figure 5.5 in bold lines, or it can be configured as an all port model; can perform simultaneous send/receive as many as channels it has, as illustrated in Figure 5.6 in bold lines.
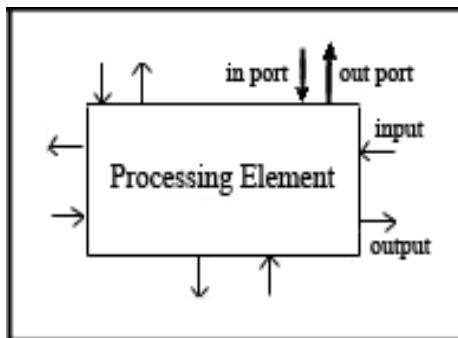
Figure 5.5: Single port node.
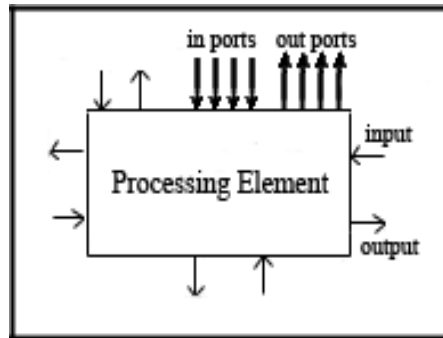
Figure 5.6: All port node.

5.6.3 Interconnection Network Switching Method

The network switching method used in the simulation is the wormhole switching introduced in Wilkinson (1996), where 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. 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.

5.6.4 Simulator Classes

The simulator consists of the following classes: node, mesh, OTIS_Mesh and EDN_OTIS_Mesh. Figure 5.7 illustrates the class diagram of the simulator. The node class is the center of the simulator as it presents one processing element. The mesh class is composed of Node objects and presents one OTIS group. The OTIS_Mesh class is composed of an array of mesh objects (a group of meshes). Finally the EDN_OTIS_Mesh class is composed of the OTIS_Mesh class with one additional parameter that makes the processor element as an EDN or not.
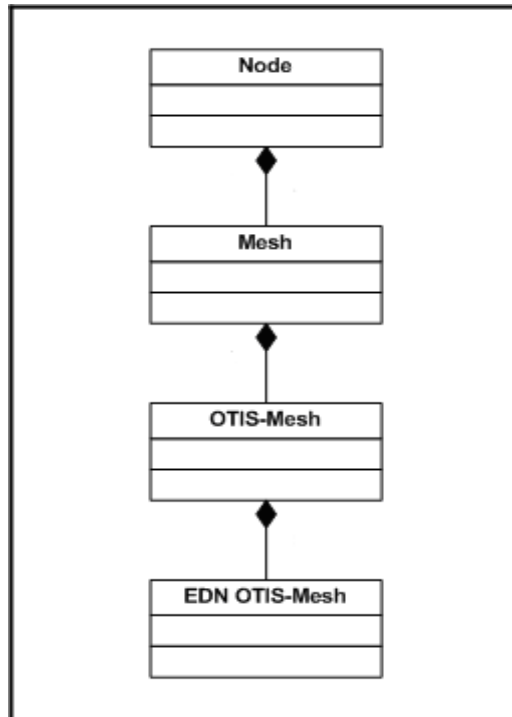
Figure 5.7: The simulator class diagram.

5.6.5 Simulator Operations

Table 5.1 describes the operations performed by the simulator. The input parameters of the operations are: network size, interconnection system (OTIS-Mesh or EDN OTIS-Mesh), message size, and port model.

Table 5.1: Operations performed by the simulator.

Operation	Description
Scatter_Latency	Calculates the maximum and minimum latency to perform scatter operation.
Reduction_Execution	Calculates the maximum and minimum execution time to perform reduction operation.
Reduction_Cost	Calculates the maximum and minimum cost to perform reduction operation.
Reduction_Overhead	Calculates the total overhead to perform reduction operation.
Barrier_Latency	Calculates the maximum and minimum latency to perform barrier operation.

6- Experimental Results and Evaluation

This chapter presents the results of the conducted experiments to evaluate and compare the performance of the collective communication operations (scatter, reduction and barrier) on the following interconnection systems: single port OTIS-Mesh, all port OTIS-Mesh, and all port EDN OTIS-Mesh under the following performance metrics: number of steps, latency, execution time, performance improvement, total overhead, and cost.

The results of the following performance metrics: number of steps, latency, execution time, performance improvement, total overhead, and cost are shown and discussed in sections 6.1, 6.2, 6.3, 6.4, 6.5 and 6.6 respectively.

6.1 Number of Steps

The number of electronic message passing steps to perform scatter, reduction, and barrier, on the interconnection systems, single port OTIS-Mesh, all port OTIS-Mesh, and all port EDN OTIS-Mesh, is evaluated and compared. Therefore, the following are observed:

- The minimum number of steps to perform scatter operation (Figure 6.1) in all port OTIS-Mesh and all port EDN OTIS-Mesh is the same, because the root node in the scatter operation sends a distinct message to each other node in the network which reduces the number of parallel send operation.

- The minimum number of steps to perform scatter operation (Figure 6.1) in single port OTIS-Mesh is as twice as to perform it in all port OTIS-Mesh and all port EDN OTIS-Mesh, because the single port OTIS-Mesh performs one send operation at a time.

- The maximum number of steps to perform scatter operation (Figure 6.2) in all port EDN OTIS-Mesh is slightly higher than to perform it in all port OTIS-Mesh, because in this case both of the networks perform limited number of parallel send operation and the EDN OTIS-Mesh needs a number of extra steps to maintain sending messages between the EDN levels.

- The maximum number of steps to perform scatter operation (Figure 6.2) in the single port OTIS-Mesh is slightly higher than to perform it in the all port OTIS-Mesh and all port EDN OTIS-Mesh, because the single port OTIS-Mesh performs one send operation at a time.

Figure 6.1: Scatter - minimum number of electronic steps.
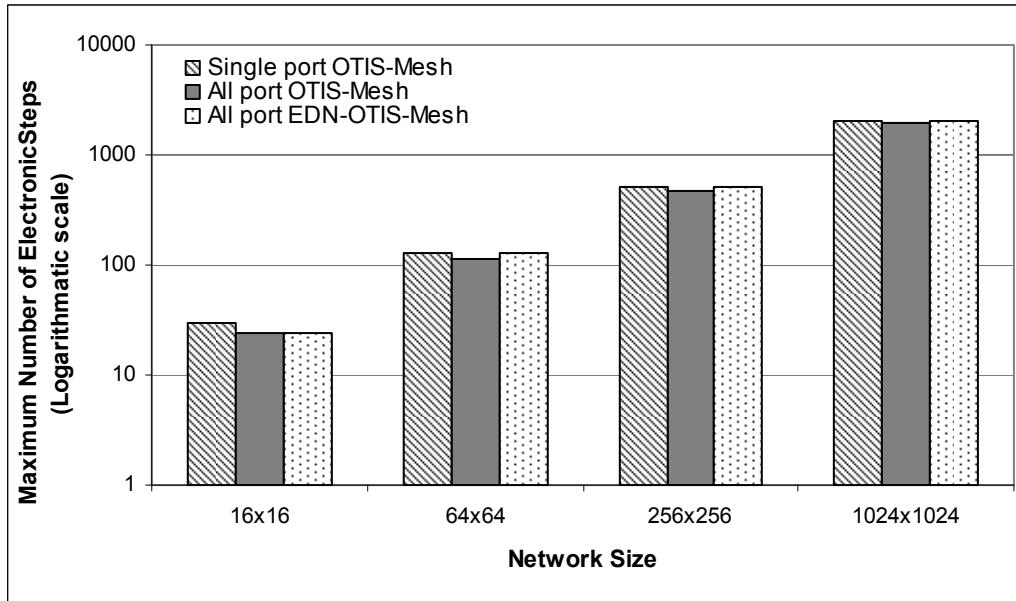
Figure 6.2: Scatter - maximum number of electronic steps.

- The number of steps to perform reduction operation (Figure 6.3 and Figure 6.4) in all port EDN OTIS-Mesh is significantly less than in single port OTIS Mesh and all port OTIS-Mesh, because the EDN processors in each group increase the number of parallel receive operation.

- The minimum number of steps to perform reduction operation (Figure 6.4) in single port OTIS-Mesh is as twice as to perform it in all port OTIS-Mesh, because the single port OTIS-Mesh performs one receive operation at a time.

- The maximum number of steps to perform reduction operation (Figure 6.4) in single port OTIS-Mesh is slightly higher than to perform it in all port OTIS-Mesh, because in this case the all port OTIS-Mesh performs a limited number of parallel receive operation.
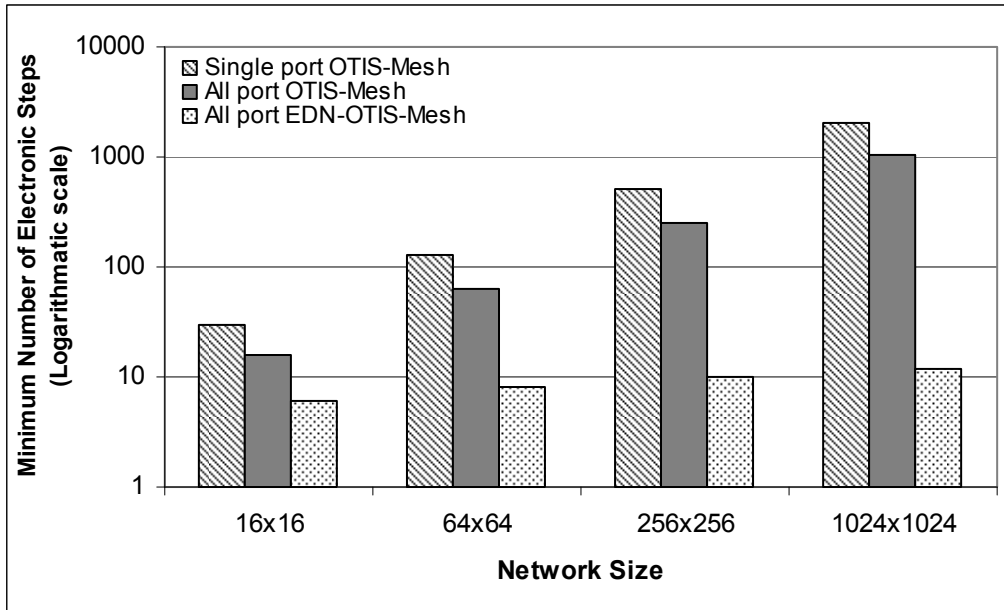
Figure 6.3: Reduction - minimum number of electronic steps.
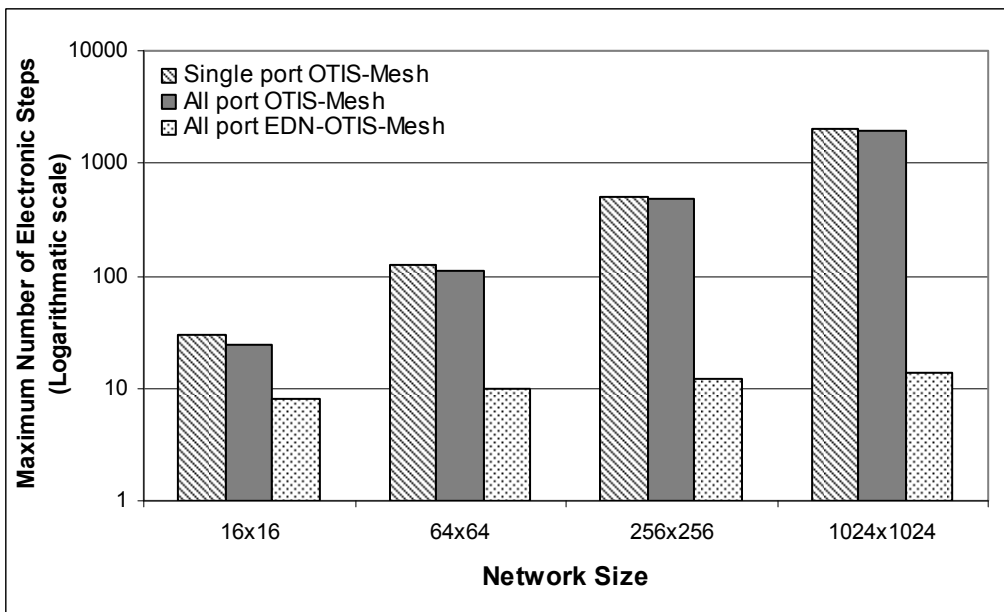
Figure 6.4: Reduction - maximum number of electronic steps.

- The number of steps to perform barrier operation (Figure 6.5 and Figure 6.6) in all

  port EDN OTIS-Mesh is significantly less than in single port OTIS Mesh and all

port OTIS-Mesh, because the EDN processors in each group increase the number of parallel send/receive operations.

- The minimum number of steps to perform barrier operation (Figure 6.5) in network sizes 16, 64 and 256 single port OTIS-Mesh is as twice as to perform it in all port OTIS-Mesh, because the single port OTIS-Mesh can perform one send/receive operations at a time.

- The maximum number of steps to perform barrier operation (Figure 6.6) in single port OTIS-Mesh is slightly higher than to perform it in all port OTIS-Mesh, because in this case the all port OTIS-Mesh performs a limited number of parallel send/receive operations.
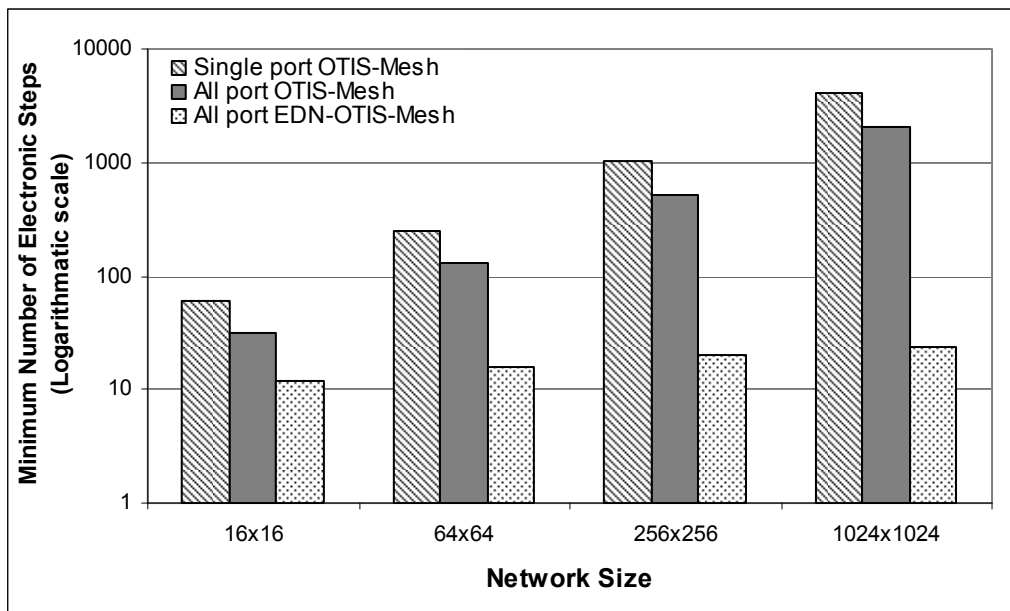
Figure 6.5: Barrier - minimum number of electronic steps.
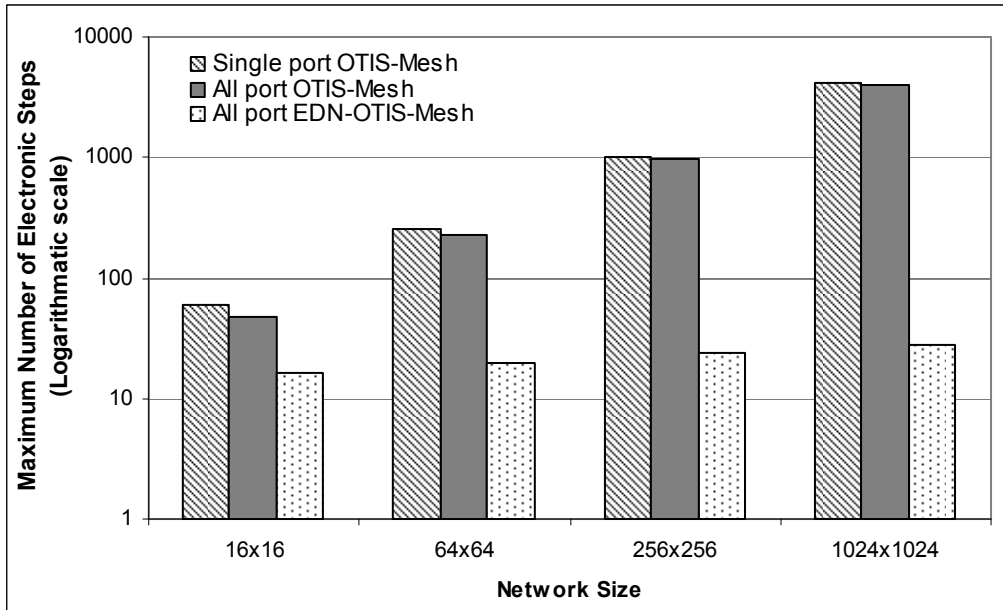
Figure 6.6: Barrier - maximum number of electronic steps.

6.2 Latency

The Latency of the collective communication operations that have no computation, such as scatter, and barrier, is evaluated on the interconnection systems (single port OTIS-Mesh, all port OTIS-Mesh, and all port EDN OTIS-Mesh) as best-case, when the root node is located in the middle of the OTIS-Mesh group, and as worst-case, when the root node is located in the endmost of the OTIS-Mesh group. Therefore, the following are observed:

- The latency to perform scatter operation (Figure 6.7 and Figure 6.8) in all port OTIS-Mesh is slightly higher than to perform it in all EDN OTIS-Mesh, because in the scatter operation the root node sends a distinct message to each other node in the network which reduces the number of parallel send operation. Moreover, the EDN OTIS-Mesh needs extra steps to maintain sending the message between the EDN's levels.

- The maximum latency to perform scatter operation (Figure 6.7 and Figure 6.8) in single port OTIS-Mesh is higher than perform it in all OTIS-Mesh, because the single port OTIS-Mesh performs one send operation while both of all port OTIS-Mesh and all port EDN OTIS-Mesh perform parallel send operation at a time.

- The maximum (worst-case) latency to perform scatter operation (Figure 6.8) for the network size 1024×1024 is almost the same in all systems, because in this case the root processor is located in the endmost of the OTIS-Mesh group so the number of parallel send operations is limited in both all port OTIS-Mesh and all port EDN OTIS-Mesh.
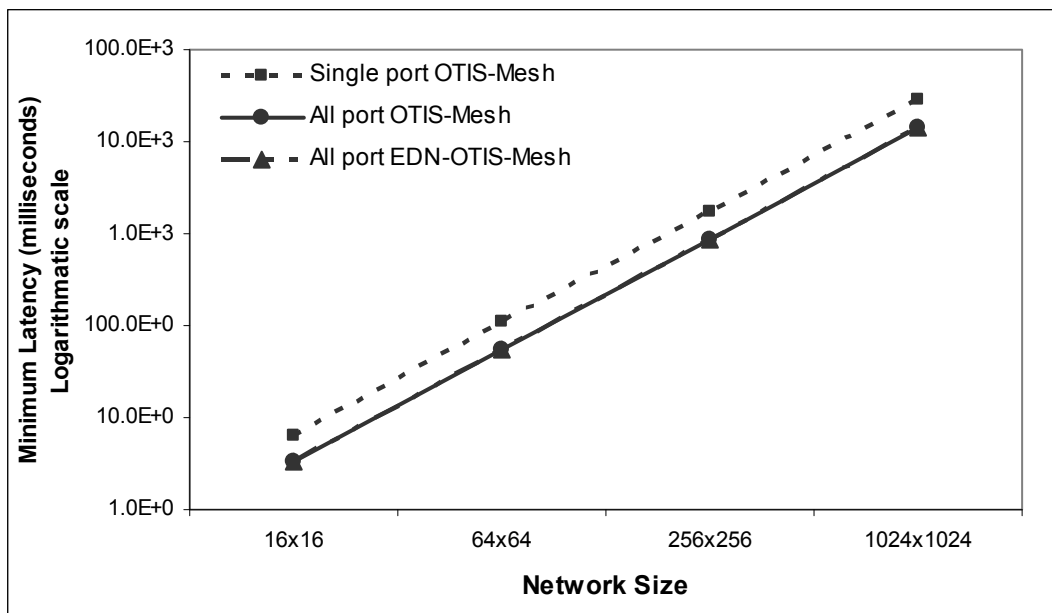
Figure 6.7: Scatter - minimum latency.
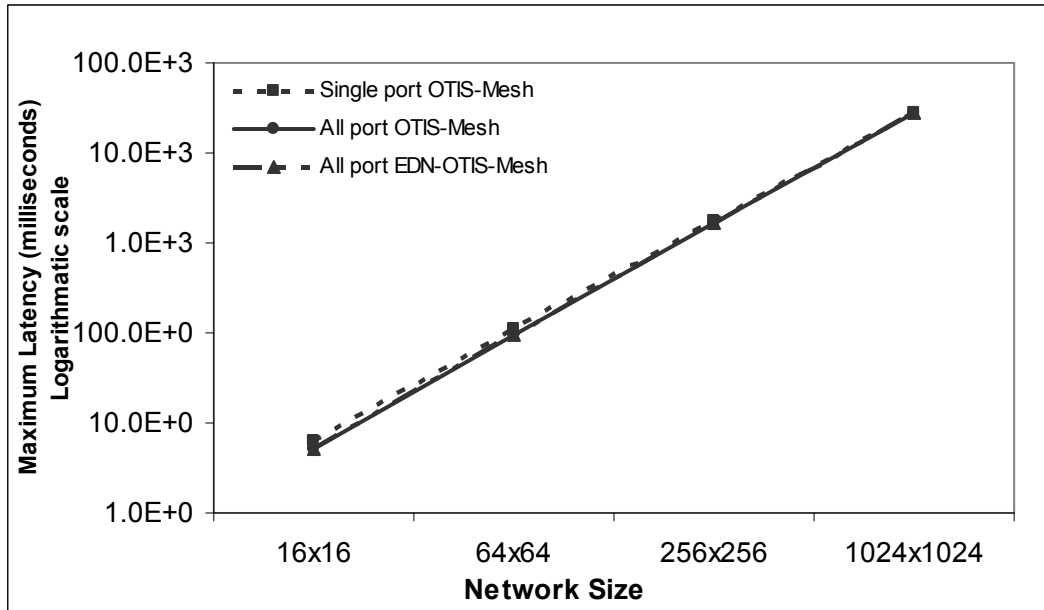
Figure 6.8: Scatter - maximum latency.

- The minimum latency to perform barrier operation (Figure 6.9) in single port OTIS-Mesh is higher than to perform it in all OTIS-Mesh and all port EDN OTIS-Mesh, because the single port OTIS-Mesh perform one send/receive operation at a time.
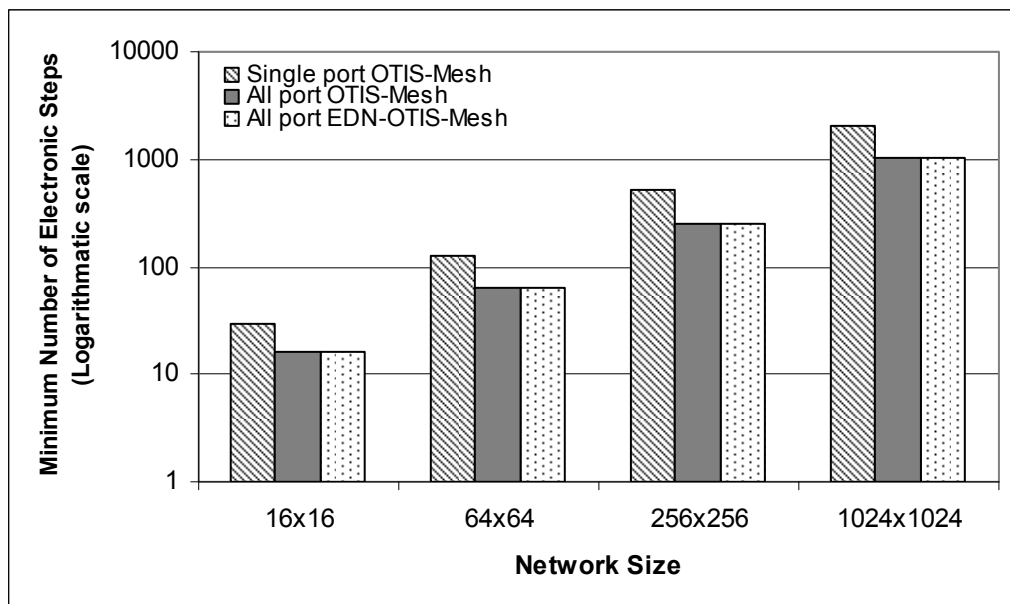
- The maximum (worst-case) latency to perform barrier operation (Figure 6.10) in single port OTIS-Mesh and all port OTIS-Mesh is higher than to perform it in all port EDN OTIS-Mesh, because the single port OTIS-Mesh performs one send/receive operation at a time and the all port OTIS-Mesh performs limited number send/receive operations. On the other hand, the EDN processors in the EDN OTIS-Mesh influences the numbers of parallel send/receive operation in the reduction and distribution phases of the barrier operation.

- The latency to perform barrier operation (Figure 6.9 and Figure 6.10) in all port OTIS-Mesh is higher than to perform it in all port EDN OTIS-Mesh, because the

EDN processors in the EDN OTIS-Mesh increases the parallel send/receive operations.
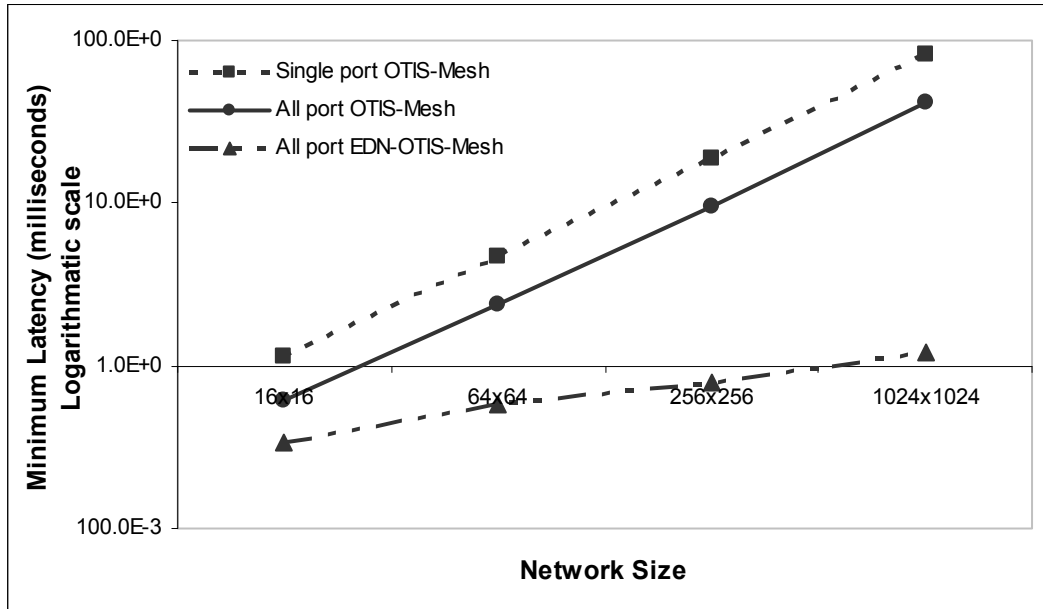
Figure 6.9: Barrier - minimum latency.
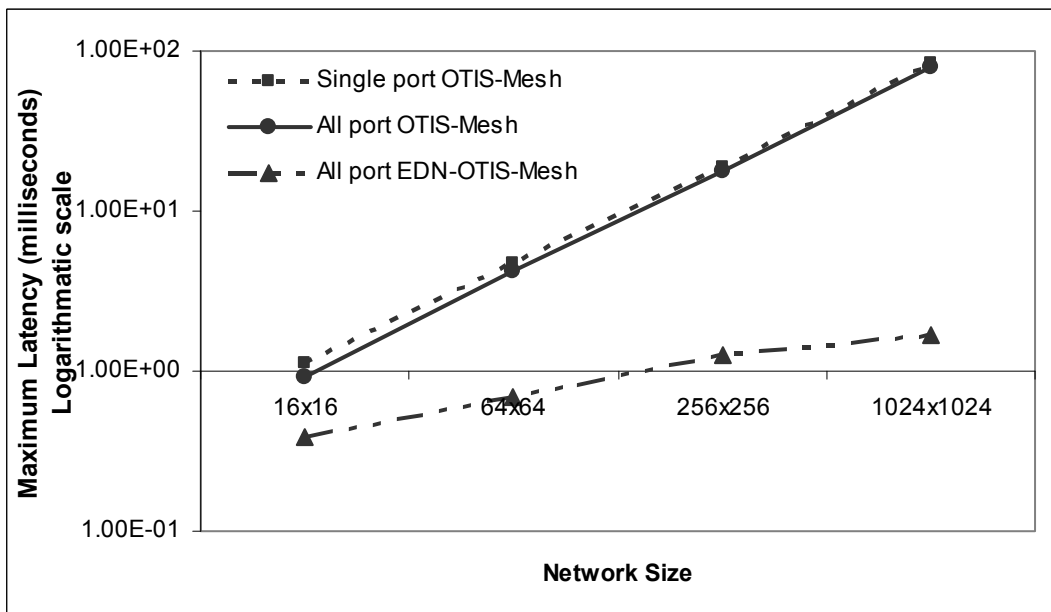
Figure 6.10: Barrier - maximum latency.

The latency of performing the following collective communication's operations scatter and barrier is illustrated in Figures 6.9 - 6.10. From these figures the latency increases as the network size increases. That is explained by the fact that large size networks such as 1024 are comprised of large number of processors compared to small size networks such as 16. As a result, the number of send/receive operations increases and consequently the latency to perform the collective communication's operations such as scatter, and barrier increases as well. For instance, the latency to perform barrier operation in 16×16 EDN OTIS-Mesh is equal to 0.43 millisecond while it takes 1.2 milliseconds to perform it on 1024×1024 EDN OTIS-Mesh.

6.3 Execution Time

The execution time for collective communication that contains computation, such as reduction, is evaluated in the following interconnection systems: single port OTIS-Mesh, all port OTIS-Mesh, and all port EDN OTIS-Mesh, for the best-case when the root node is located in the middle of the OTIS-Mesh group, and for the worst-case when the root node is located in the endmost of the OTIS-Mesh group. Therefore, the following are observed:

- The minimum execution time to perform reduction operation (Figure 6.11) in single port OTIS-Mesh is higher than to perform it in all OTIS-Mesh and all port EDN OTIS-Mesh, because the single port OTIS-Mesh can perform one receive operation at a time.

- The maximum execution time to perform reduction operation (Figure 6.12) for network size 256 and 1024 in single port OTIS-Mesh and all port OTIS-Mesh is almost the same because in this case the all port OTIS-Mesh performs a limited number of parallel receive operation at a time.

- The execution to perform reduction operation (Figure 6.11 and Figure 6.12) in all
  port OTIS-Mesh is higher than to perform it in all port EDN OTIS-Mesh, because
  the EDN processors in the EDN OTIS-Mesh increase the parallel receive
  operation.
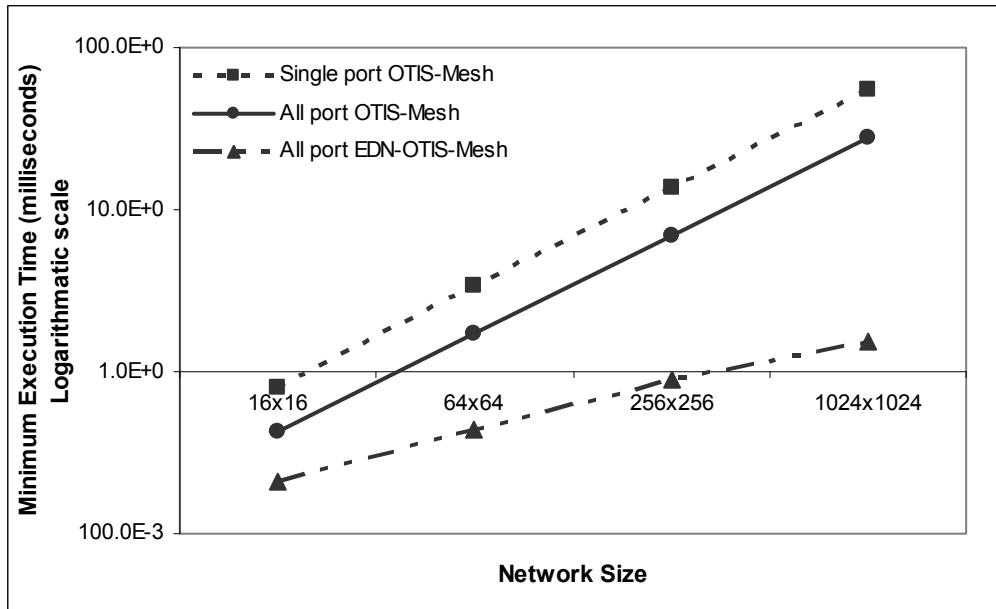
Figure 6.11: Reduction - minimum execution time.
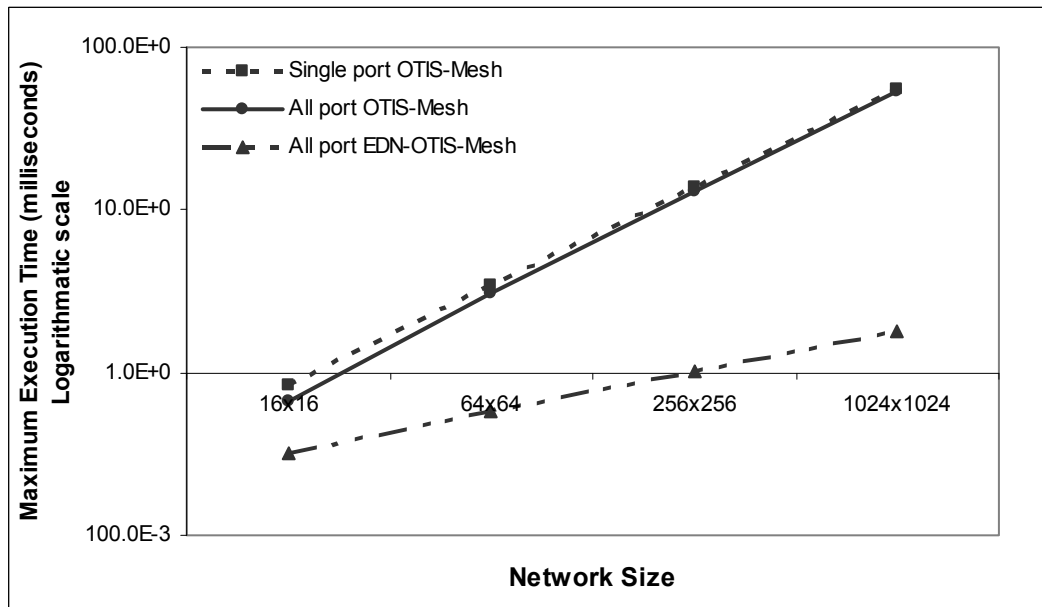
Figure 6.12: Reduction - maximum execution time.

The execution time of performing the reduction operation is illustrated in Figures 6.15 and 6.16 where the execution time increases as the network size increases. That is explained by the fact that large size networks such as 1024×1024 are comprised of large number of processors compared to small size networks such as 16×16. As a result, the number of receive operation increases and consequently the execution time to perform the reduction operation increases as well. For instance, the execution time to perform reduction operation in 16×16 EDN OTIS-Mesh is equal to 0.32 millisecond while it takes 1.81 milliseconds to perform it on 1024×1024 EDN OTIS-Mesh.

6.4 Performance Improvement

Performance improvement is the improvement gained from using all port EDN-OTIS Mesh to implement the collective communication's operations, such as scatter, reduction and barrier, over using single port OTIS-Mesh or all port OTIS-Mesh. Therefore the following are observed:

- The minimum performance improvement (best-case) experiment of the scatter operation (Figure 6.13) shows that all port EDN OTIS-Mesh slightly outperforms single port OTIS-Mesh by 1.97, 1.99, 2 and 2 times for network sizes 16, 64, 256 and 1024 respectively, because the single OTIS-Mesh performs one send operation at a time while the all port EDN OTIS-Mesh performs parallel send operation.

- The maximum performance improvement (worst-case), when the root node is located in the endmost of the OTIS-Mesh group, of the scatter operation (Figure 6.14) shows that the all port EDN OTIS-Mesh outperforms the single port OTIS-

Mesh by 1.27, 1.13, 1.06 and 1.03 for network sizes 16, 64, 256, and 1024 respectively. The performance improvement is slight because the all port EDN OTIS-Mesh performs a limited number of parallel send operations at a time.

- The minimum and maximum performance improvements' experiments of the scatter operation yield that all port EDN OTIS-Mesh does not outperform all port OTIS-Mesh because the scatter operation sends a distinct message to each node in the network, also sending the messages via the EDN hierarchy adds latency which affects the performance.

- The minimum performance improvement of performing scatter operation on all port EDN OTIS-Mesh over performing it on single port OTIS-Mesh and all port OTIS Mesh (Figure 6.13) are the same for network sizes 16, 64, 256, and 1024. This is because the ratio of latency on single port OTIS-Mesh over the latency on all port EDN OTIS-Mesh is the same on the previous network sizes. This is explained by the fact that the root processor in the scatter operation sends a distinct message to each other processor in the network so the all port EDN OTIS-Mesh performs limited numbers of parallel send operations.

- The maximum performance improvement of performing scatter operation on all port EDN OTIS-Mesh over performing it on single port OTIS-Mesh and all port OTIS Mesh (Figure 6.14) slightly decreases for different network sizes 1.27, 1.13, 1.06 and 1.03 times on network sizes 16, 64, 256 and 1024 respectively. This is explained by the location of the root processor at the endmost of the OTIS group so the number of parallel send operations in the EDN OTIS-Mesh is reduced

which leads to an increase in the latency value and consequently decreasing the performance improvement ratio.
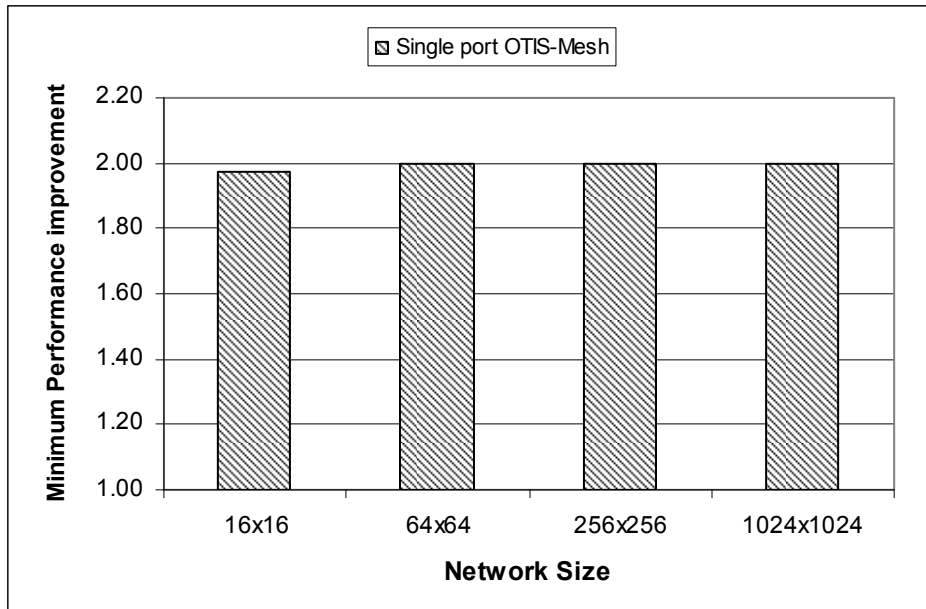
Figure 6.13: Scatter - minimum performance improvement.
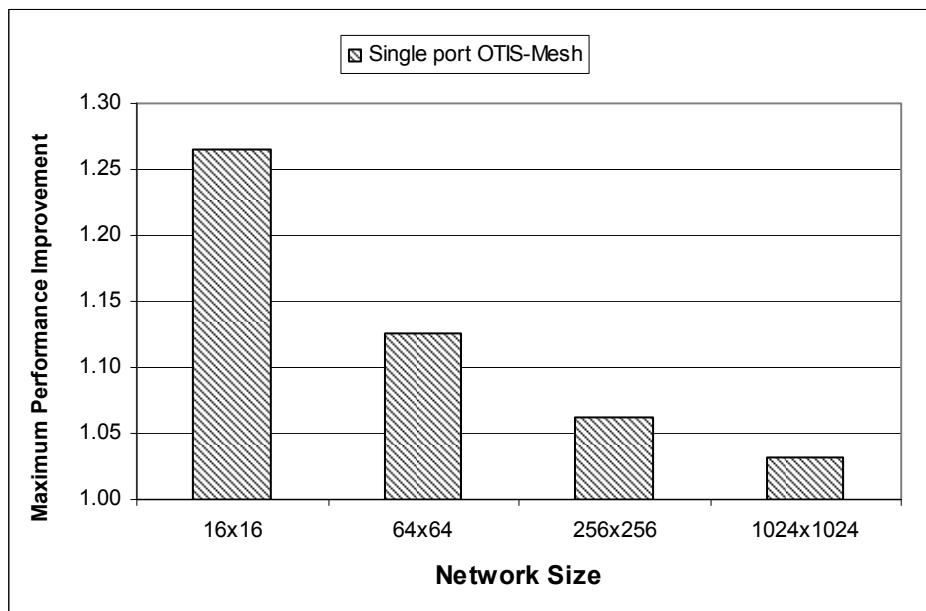
Figure 6.14: Scatter - maximum performance improvement.

- The minimum performance improvement (best-case) experiment of the reduction operation (Figure 6.15) shows that all port EDN OTIS-Mesh outperforms the

single port OTIS-Mesh by 3.8, 7.73, 15.42, and 33.67 times for network sizes 16, 64, 256 and 1024 respectively, and all port OTIS-Mesh by 2, 3.9, 7.73 and 17.85 times for previous networks, because the all port EDN OTIS-Mesh performs more parallel receive operation and global compute operation.

- The maximum performance improvement (worst-case) experiment of reduction operation (Figure 6.16) shows that all port EDN OTIS-Mesh outperforms single port OTIS-Mesh by 2.59, 5.86, 13.42 and 30.39 times for network sizes 16, 64, 256 and 1024 respectively, and all port OTIS-Mesh by 2.08, 5.21, 12.63 and 29.47 times for the previous network sizes, because the all port EDN OTIS-Mesh performs more parallel receive operation and global compute operation.

- The performance improvement of performing reduction operation on EDN OTIS-Mesh over performing it on single port OTIS-Mesh and all port OTIS-Mesh (Figures 6.15 and 6.16) increases when the network size increases. This is because the ratio of execution time on single port OTIS-Mesh over the execution time on all port EDN OTIS-Mesh and the ratio of all port OTIS-Mesh over all port EDN OTIS-Mesh increases as the network size increases. This is explained by the fact that the EDN OTIS-Mesh performs parallel receive operations so the execution time decreases and consequently the performance improvement ration increases.
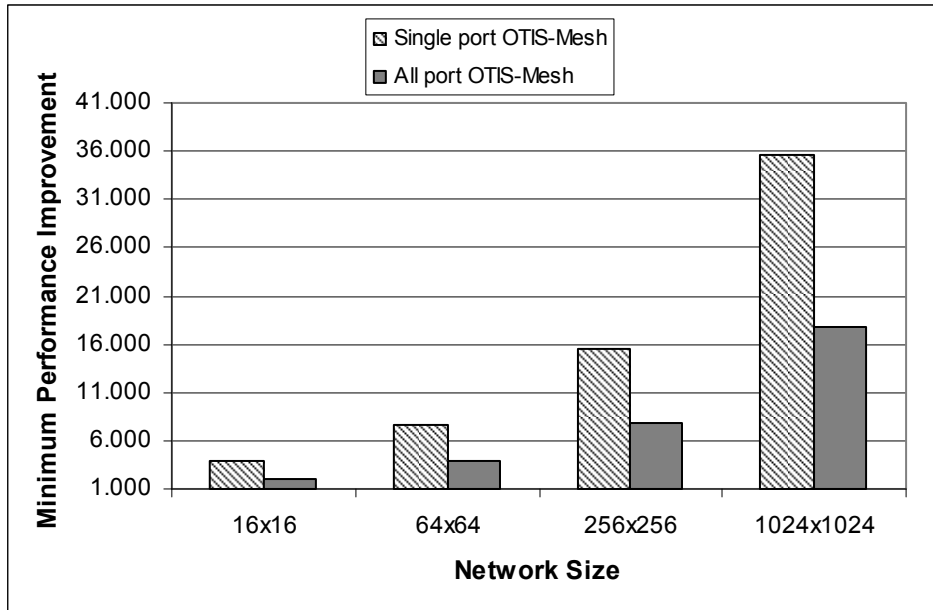
Figure 6.15: Reduction - minimum performance improvement.
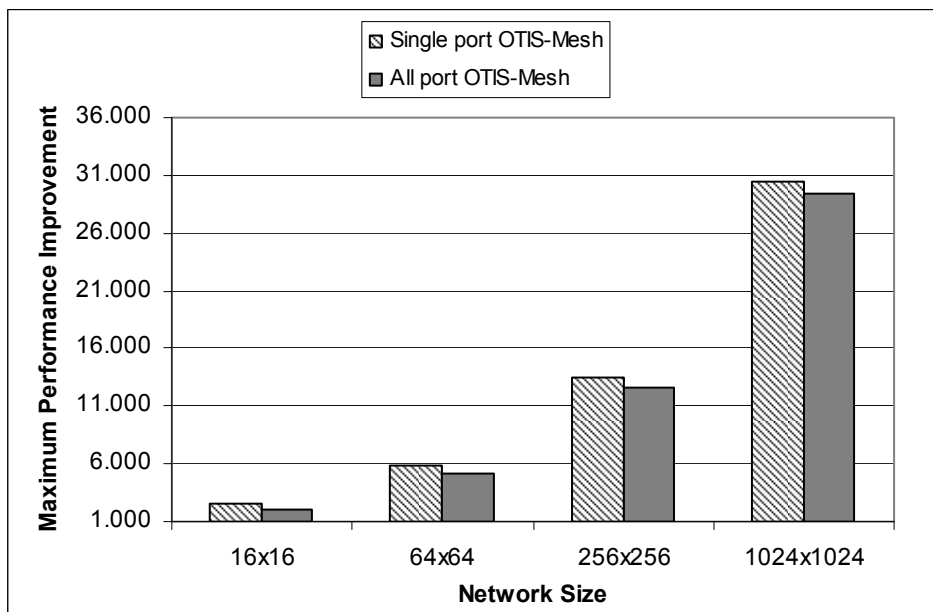
Figure 6.16: Reduction - maximum performance improvement.

- The minimum performance improvement (best-case) experiment of the barrier operation (Figure 6.17) shows that all port EDN OTIS-Mesh outperforms single port OTIS-Mesh by 3.3, 8.1, 24 and 69 times for network sizes 16, 64, 256, and 1024 respectively, and all port OTIS-Mesh by 1.8, 4.1, 12.1 and 34.5 times for the

previous network sizes, because the all port EDN OTIS-Mesh performs more parallel send/receive operations.

- The maximum performance improvement (worst-case) experiment of barrier operation (Figure 6.18) shows that all port EDN OTIS-Mesh outperformed single port OTIS-Mesh by 2.9, 6.8, 15.3, and 49.1 times for network sizes 16, 64, 256 and 1024 respectively, and all port OTIS-Mesh by 2.3, 6, 9.7, and 47.6 times on the previous network sizes, because the all port EDN OTIS-Mesh performs more parallel send/receive operation.

- The performance improvement of performing barrier operation on the EDN OTIS-Mesh over performing it on single port OTIS-Mesh and all port OTIS-Mesh (Figure 6.17 and Figure 6.18) increases when the network size increases. This is because the ratio of latency on single port OTIS-Mesh over the latency on all port EDN OTIS-Mesh and the ratio of all port OTIS-Mesh over all port EDN OTIS-Mesh increases as the network size increases. This is explained by the fact that the EDN OTIS-Mesh performs parallel send/receive operations so the latency decreases and consequently the performance improvement ratio increases.
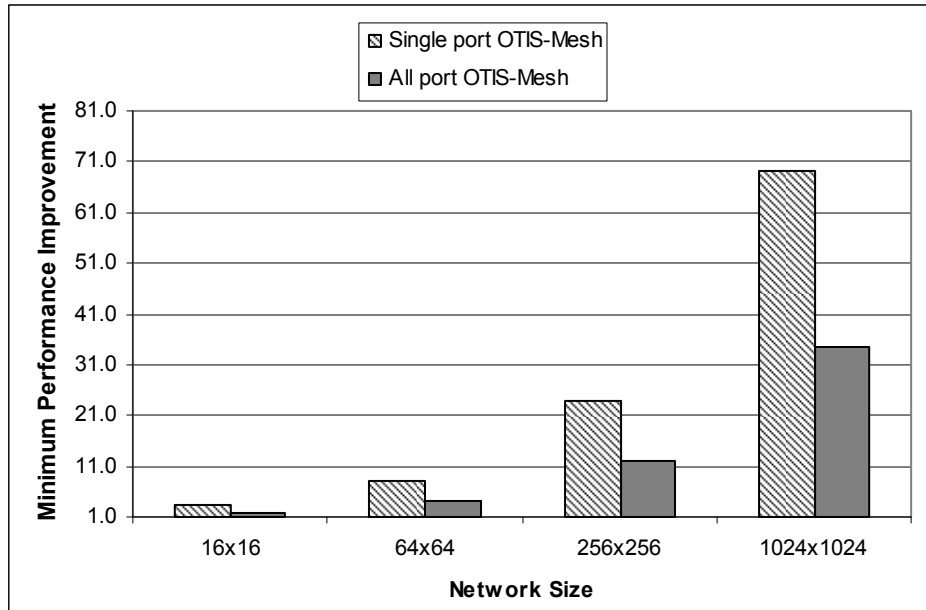
Figure 6.17: Barrier - minimum performance improvement.
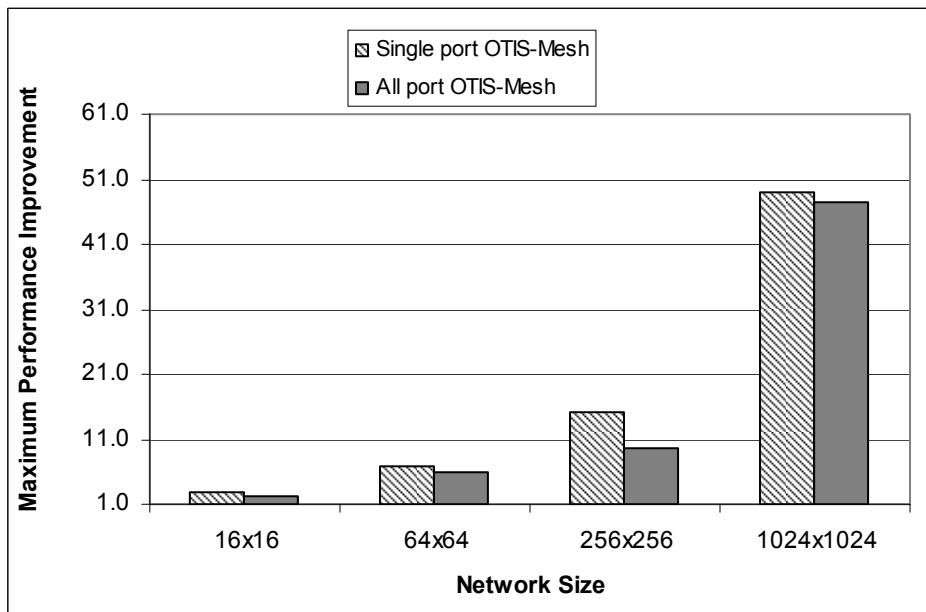
Figure 6.18: Barrier - maximum performance improvement.

6.5   Total Overhead

The total overhead is defined as the passed time when the processors are not performing computation operations due to contention and communication. The total overhead is evaluated for the reduction operation only because it has a computation

part, the previous operations scatter and barrier has no computation. The total overhead evaluated by subtracting the computation time from the execution time.

The total overhead of performing reduction on single port OTIS-Mesh, all port OTIS-Mesh, and EDN OTIS-Mesh is evaluated (Figures 6.19 and 6.20). The minimum total overhead is evaluated from the minimum execution time (best-case) while the maximum total overhead is evaluated from the maximum execution time (worst-case).

The total overhead of performing the reduction operation on EDN OTIS-Mesh increases as the network size increases (Figures 6.19 and 6.20). This is explained by the fact that large networks' size such as 1024×1024 are comprised of large number of processors compared to small networks' size such as 16×16. As a result, the overhead to perform the reduction operation increases as well. For instance, the maximum total overhead to perform reduction operation in 16×16 EDN OTIS-Mesh is equal to 0.30 milliseconds while it takes 1.78 milliseconds to perform it on 1024×1024 EDN OTIS-Mesh.
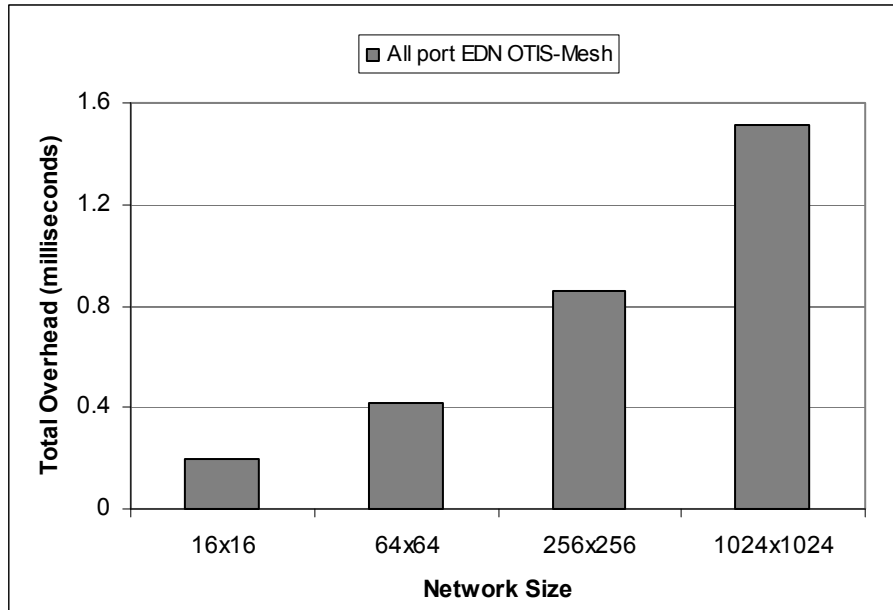
Figure 6.19: Reduction - minimum total overhead.
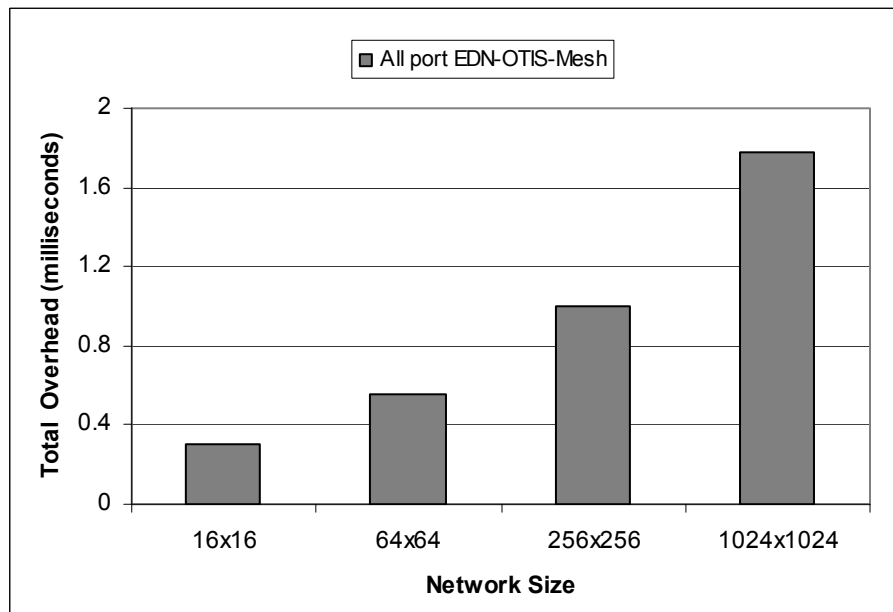
Figure 6.20: Reduction - maximum total overhead.

6.6   Cost

The cost is defined as the sum of the time that each processing element spends solving the problem (i.e. performing computation operation). The cost is evaluated for the reduction operation only because it has a computation part, while the previous operations scatter and barrier has no computation.

The cost of performing reduction on single port OTIS-Mesh, all port OTIS-Mesh, and EDN OTIS-Mesh is evaluated and compared. Therefore, the following are observed:

- The cost of performing the reduction operation (Figures 6.21 and 6.22) shows that all port EDN OTIS-Mesh has the least cost among single port OTIS and all port OTIS-Mesh, because it can perform more parallel send operation.

- The maximum cost of performing the reduction operation on single port OTIS-Mesh and all port OTIS-Mesh (Figure 6.22) is almost the same. This is explained by the fact that the root processor is located in the middle of the OTIS-Mesh group so the all port OTIS-Mesh performs a limited number of parallel receive operations.

- The cost of performing the reduction operation on all systems (single port OTIS-Mesh, all port OTIS-Mesh, and EDN OTIS-Mesh) increases as the network size increases (Figures 6.21 and 6.22). This is explained by the fact that large size networks such as 1024×1024 are comprised of large number of processors compared to small size networks such as 16×16. As a result, the number of global compute operation increases as well. For instance, the maximum cost to perform reduction operation in 16×16 EDN OTIS-Mesh is equal to 0.34 milliseconds while it takes 1398.1 milliseconds to perform it on 1024×1024 EDN OTIS-Mesh.
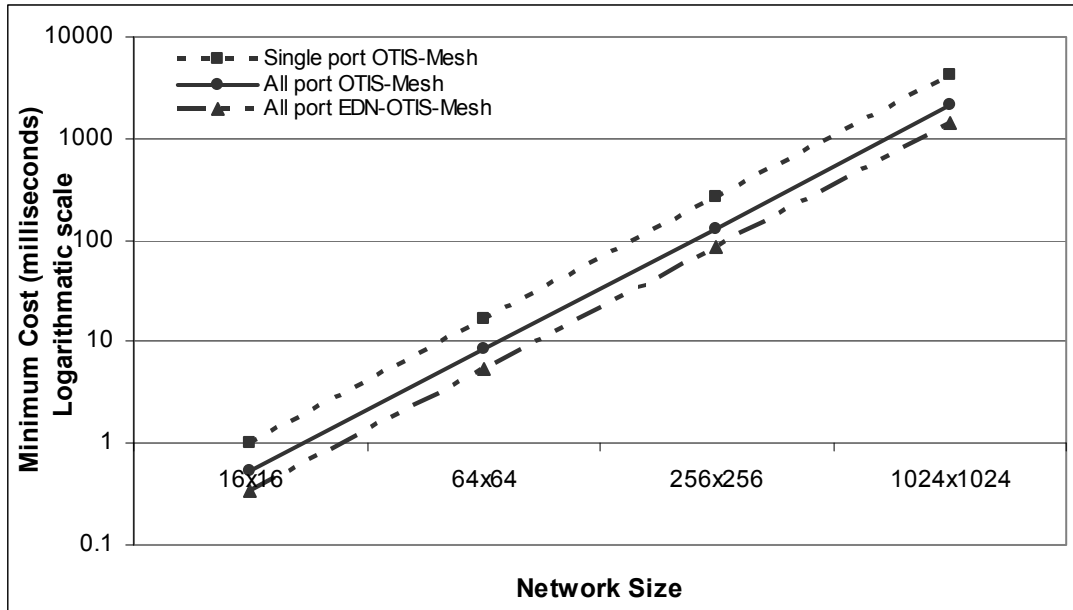
Figure 6.21: Reduction - minimum cost.
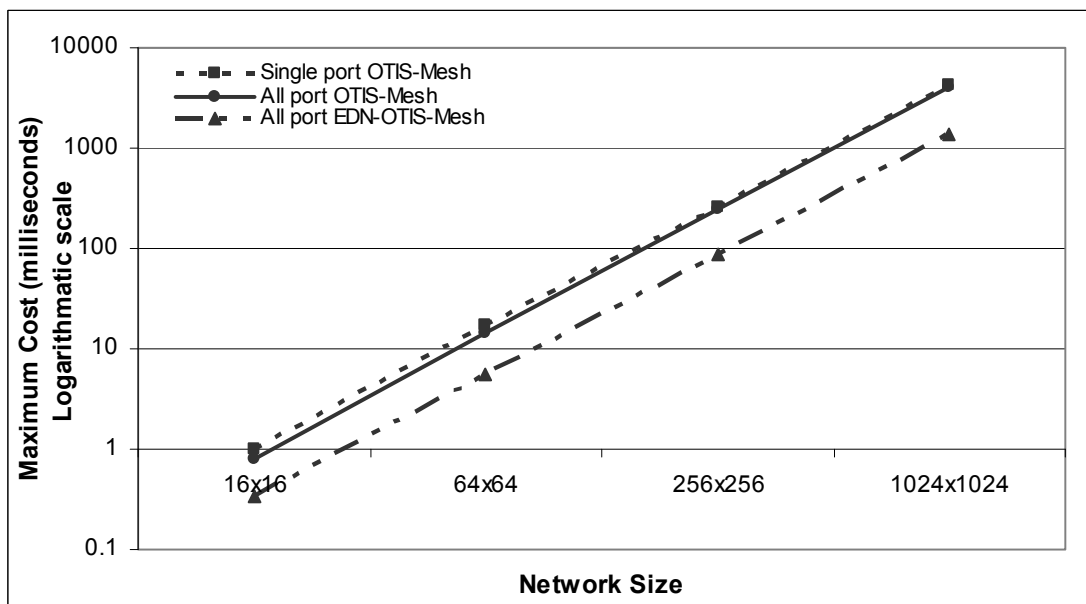
Figure 6.22: Reduction – maximum cost.

7- Conclusions and Future Work

This chapter provides a summary for the studied problem, where it explains briefly the methodology used in this thesis and discusses the results (Section 7.1), and draws future work suggestion in Section 7.2.

7.1 Conclusions

In this thesis, the problem of designing and implementing efficient collective communication operations, which are scatter, reduction, etc. in the Optical Transposed Interconnection System Mesh (OTIS-Mesh) was presented. The Extended Dominating Node (EDN) from the graph theory was used to implement efficient collective communication operation. The matrix transposition application was presented as an application on data movement in collective communication.

The EDN approach was applied on each group in the OTIS-Mesh to form the EDN OTIS-Mesh (an OTIS-Mesh that uses the EDN approach in collective communication).

The performance of some of the collective communication operations (scatter, reduction and barrier) was evaluated on the following interconnection systems: single port OTIS-Mesh, all port OTIS-Mesh and all port EDN OTIS-Mesh under the following performance metrics: number of steps, latency, execution time, performance improvement, total overhead, and cost.

The parameters that affect the performance of the collective communication's operations, the assumptions, and notations were illustrated. The analytical model to evaluate the performance metrics of interest (number of steps, latency, execution time, total overhead, performance improvement, and cost) was derived.

The analytical results showed that the number of steps to perform the operations (reduction and barrier) in all port EDN OTIS-Mesh is less than to perform them on both single port and all port OTIS-Mesh. For instance, the maximum number of steps to perform reduction operation in all port EDN OTIS Mesh equals to 6, 8, 10 and 12 for network sizes 16, 64, 256 and 1024 respectively, while it takes 24, 112, 480, 1948 steps to perform it on all port OTIS-Mesh and 30, 126, 510 and 2046 for single port OTIS-Mesh on the previous network sizes. The excellence of EDN OTIS-Mesh is justified by the fact that the all port EDN OTIS-Mesh performs parallel send/receive operations. On the other hand, the number of steps to perform the scatter operation on all port EDN OTIS-Mesh is the same as single port OTIS-Mesh. For instance, the maximum number of steps to perform scatter operation in all port EDN OTIS Mesh and single port OTIS-Mesh equals to 126, 510 and 2046 steps for network sizes 64, 256 and 1024 respectively. This is justified by the fact that in the scatter operation the root sends a distinct message to each node in the network, which minimizes the number of parallel send operations.

The experimental results showed that the operations (reduction and barrier) in the all port EDN OTIS-Mesh performed better than both of the single port and all port OTIS-Mesh in both of the best case (where the root node is located in the middle of the OTIS-Mesh group) and worst case (where the root node is located in the endmost of the OTIS-Mesh group) runs. For instance the minimum execution time of the reduction operation in the all port EDN OTIS-Mesh outperformed the single port OTIS-Mesh by 3.8, 7.73, 15.42 and 33.67 times for network sizes 16, 64, 256 and 1024 respectively. Also, it outperformed the all port OTIS-Mesh by 2, 3.9, 7.73 and 17.85 times for same previous network sizes. The maximum execution time of the

reduction operation outperformed the single port model by 2.59, 5.86, 13.42 and 30.39 times for network sizes 16, 64, 256 and 1024 respectively. It also outperformed the all port OTIS-Mesh by 2.08, 5.21, 12.63 and 29.47 times for same previous network sizes. The performance improvement is justified by the fact that the all port EDN OTIS-Mesh performs parallel send/receive operations. Moreover, the scatter operation on all port EDN OTIS-Mesh outperformed the single port OTIS-Mesh under the latency performance metric by 1.97, 1.99, 2 and 2 times for network sizes 16, 64, 256 and 1024 respectively in the best case and performed the same in the worst case. The all port EDN OTIS-Mesh performed the same as the all port OTIS-Mesh in the best case. On the other hand, the all port OTIS-Mesh slightly outperformed the EDN OTIS-Mesh in the worst case. This is justified by the fact that the scatter operation sends a distinct message to each node in the network and sending the messages via the EDN hierarchy adds latency that affects the performance.

The outcome of this study recommends the use of the EDN approach in implementing the following collective communication operations, reduction and barrier on OTIS-Mesh Interconnection Network.

7.2 Future Work

As a future work we suggest to use the Extending Dominating Node (EDN) approach to design and implement the broadcast and gather operations on the Optical Transposed Interconnection System Mesh (OTIS-Mesh).

8- References

Al-Ayyoub A., Awwad A., Day K. and Ould-Khaoua M., (2006) Generalized methods for algorithm development on optical systems, **Journal of Supercomputing**, vol. 38, no. 2.

Awwad A., Al-Ayyoub A. and Ould-Khaoua M., (2002) Efficient Routing Algorithm on the OTIS-Networks, **Proceedings of International Arab Conference.**

Barnett M., Littlefield R., Payn D., and van de Geijn R., (1993). Global Combine on mesh architectures with wormhole routing, **Proceedings of the 17ᵗʰ Parallel Processing Symposium,** pp 156-162.

Chien A., (1993). A Cost and Speed Model for K-ary n-cube Wormhole Routers, **Proceedings of Hot Interconnects.**

Culler D., Singh J. and Gupta A., (1998) **Parallel Computer Architecture: A Hardware/Software Approach,** (1ˢᵗ ed.), Morgan Kufmann.

Day K. and Al-Ayyoub A., (2002). Topological properties of OTIS-networks, **IEEE Transaction on Parallel and Distributed Systems** 13 359–366.

Fledman M., Esener S., Guest C., and Lee S., (1988). Comparison between Electronic and Free Space Optical Interconnects Based on Power and Speed Consideration, **Applied Optics**, vol.27, no.9.

Karonis N., De Supinski B., Foster I. and Gropp W., (2000) Exploiting Hierarchy in Parallel Computer Networks to Optimize Collective Operation Performance, **Proceedings of the 14th International Symposium on Parallel and Distributed Processing**.

Kouvalsos D., Assi A. and Ould-Khaoua .M, (2004). **Proceedings of 2ⁿᵈ International Working Conference on Performance Modeling and Heterogeneous Networks.**

Krishnamoorthy A., Marchand P., Kiamilev F., and Esener S., (1992). Grain-size Considerations for Optoelectronic Multistage Interconnection Networks, **Applied Optics**, vol.31, no. 26, p 5480-5507.

Mahafzah B. and Jaradat B., (2008). The load balancing problem in OTIS-Hypercube interconnection networks, The Journal of Supercomputing. Last access on April 10, 2008, from.
http://www.springerlink.com/content/vux017112073p8w7/

Mao W., Chen J., and Watson W., (2007). One-to-All Personalized Communication in Torus Networks, **Proceedings of the 25th conference on Proceedings of the 25th**

IASTED International Multi-Conference: parallel and distributed computing and networks, 291-296.

Marsden G. C., Marchand P. J., Harvey P., and Esener S. C., (1993). Optical transpose interconnection system architectures. **Optics Letters,** 18(13):1083–1085.

McKinley P., Tsai Y., and Robinson D., (1994) A survey of collective communication in wormhole-routed massively parallel computers, **Technical Report MSU-CPS-94-35, Communications Research Group, Michigan State University.**

Miller D. and Najjar W., (1997). Preliminary Evaluation of a Hybrid Deterministic/Adaptive Router, **Proceedings of Parallel Computing Routing and Communication workshop**.

Mondal D. and Jana P., (2004). Neighborhood property of OTIS-mesh optoelectronic computer, **Parallel Architectures, Algorithms and Networks, Proceedings. 7th International Symposium**.

Najaf-abadi H. H., (2004) A procedure for obtaining a behavioral description for the control logic of a non-linear pipeline, **Proceeding of the IEEE/ACM Asia & South Pacific Design Automation Conference (ASP-DAC)**.

Najaf-abadi H. H., Sarbazi-azad H., (2004) Comparative evaluation of adaptive and deterministic routing in the OTIS-hypercube, **Proceeding of Ninth Asia-Pacific Computer Systems Architecture Conference (ACSAC)**.

Najaf-abadi H. H., Sarbazi-azad H., (2005) An Empirical Comparison of OTIS-mesh and OTIS-hypercube Multicomputer Systems under Deterministic Routing, **19th IEEE International Parallel and Distributed Processing Symposium.**

Najaf-abadi H., Sarbazi-azad H. and P. Rajabzadeh, (2006). Performance Modeling of Fully Adaptive Wormhole Routing in 2-D Mesh-Connected Multiprocessors, **Proceedings of the 25th IEEE International Performance Computing and Communications Conference,** pp. 199- 206.

Najaf-abadi H. H., Sarbazi-azad H. and Rajabzadeh P., (2007) An accurate performance model of fully adaptive routing in wormhole-switched two-dimensional mesh multicomputers, **Microprocessors & Microsystems,** vol. 31, pp 445-455.

Osterloh A., (2000). Sorting on the OTIS-Mesh, **Proceedings of the 14th International Symposium on Parallel and Distributed Processing.**

Pjeˇsivac-Grbovi´c J., Angskun T., Bosilca G., Fagg G., Gabriel E. and Dongarra J., (2007) Performance analysis of MPI collective operations, **Cluster Computing**.

Rajasekaran S. and Sahni S., (1998) Randomized routing, selection, and sorting on the OTIS-mesh. **IEEE Transactions on Parallel and Distributed Systems**.

Raksapatchcharawong M. and Pinkston T., (1996). An Optical Interconnect Model for K-ary n-cube Wormhole Networks, **Parallel Processing Symposium.**

Sahni S., (2001) Models and Algorithms for Optical and Optoelectronic **Parallel Computers, Proceedings of the 15th International Parallel & Distributed Processing Symposium.**

Tanenbaum A., (1995). **Distributed Operating Systems**, (1st ed.). Printice Hall.

Tsai Y. and McKinley P., (1993). A Dominating Set Model for broadcast in all-port wormhole-routed 2D mesh networks. **Technical Report MSU_CPS_93_23, Michigan State University.**

Tsai Y. and McKinley P., (1994) A dominating set model for broadcast in all-port wormhole-routed 2D mesh networks, **Proceedings of the 8th international conference on Supercomputing.**

Tsai Y. and McKinley P., (1994) Broadcast in all-port wormhole-Routed 3D Mesh Networks Using Extended Dominating Sets, **Proceedings of International Conference on Parallel and Distributed Systems**.

Tsai Y. and McKinley P., (1994). An Extended Dominating Nodes to collective communication in Wormhole-Routed 2D meshes, **Proceedings of the Scalable High Performance Computing Conference.**

Tsai Y. and McKinley P., (1997). An Extended Dominating Node Approach to Broadcast and Global Combine in Multiport Wormhole-Routed Mesh Networks, **IEEE Transactions on Parallel and Distributed Systems**, Vol 8.

Wang C. and Sahni S., (2000). Image processing on the OTIS Mesh optoelectronic computer. **IEEE Transaction on Parallel and Distributed Systems**, 11(2):97–109.

Wang C. and Sahni S., (1998). Basic operations on the OTIS-mesh optoelectronic computer, **IEEE Transaction on Parallel and Distributed Systems,** 1226–1236.

Wang C. and Sahni S., (2001). Matrix multiplication on the OTIS-Mesh optoelectronic computer. **IEEE Transactions on Computers**, 50(7):635–646.

Wang C. and Sahni S., (2002) Computational Geometry on The OTIS-Mesh Optoelectronic Computer, **Proceedings of the International Conference on Parallel Processing**.

Wang C.and Sahni S., (1997) BPC permutations on the OTIS-mesh optoelectronic computer. **Proceedings on the Fourth International Conference on Massively Parallel Processing Using Optical Interconnections**, p130–135.

Wilkinson B. (1996). **Computer Architecture Design and performance**, (2nd ed.), Printice Hall.

Xu H., McKinely P., and Ni L., (1992). Efficient Implementation of Barrier Synchronization in Wormhole-Routed Hypercube Multicomputers, **Parallel and Distributed Computing** vol. 16, pp. 172-184.

Xu H., McKinley P. and Ni L., (1992) Efficient Implementation of Barrier Synchronization in Wormhole-Routed Massively Parallel Computers, **Proceedings 12th International Conference in Distributed Computing Systems**, p 118-125.

Yang J. and King C., (1998). Designing Tree-Based Barrier Synchronization on 2D Mesh Networks, IEEE Trans. Parallel and Distributed Systems, vol. 9, no. 6, pp. 526-534.

Zane F., Marchand P., Paturi R. and Esener S., (2000). Scalable network architectures using the optical transpose interconnection system, **Parallel and Distributed Computing,** pp 521–538.

Appendix A

Number of Electronic steps:

1. Scatter operation

Table 1: Scatter - Minimum Number of Electronic Steps

Size	Single port OTIS-Mesh	All port OTIS-Mesh	All port EDN-OTIS-Mesh
16x16	30	16	16
64x64	126	64	64
256x256	510	256	256
1024x1024	2046	1024	1024

Table 2: Scatter - Maximum Number of Electronic Steps

Size	Single port OTIS-Mesh	All port OTIS-Mesh	All port EDN-OTIS-Mesh
16x16	30	24	24
64x64	126	112	126
256x256	510	480	510
1024x1024	2046	1984	2046

2. Barrier operation

Table 3: Barrier - Minimum Number of Electronic Steps

Size	Single port OTIS-Mesh	All port OTIS-Mesh	All port EDN-OTIS-Mesh
16x16	60	32	12
64x64	252	128	16
256x256	1020	512	20
1024x1024	4092	2048	24

Table 4: Barrier - Maximum Number of Electronic Steps

Size	Single port OTIS-Mesh	All port OTIS-Mesh	All port EDN-OTIS-Mesh
16x16	60	48	16
64x64	252	224	20
256x256	1020	960	24
1024x1024	4092	3968	28

3. Reduction operation

Table 5: Reduction - Minimum Number of Electronic Steps

Size	Single port OTIS-Mesh	All port OTIS-Mesh	All port EDN-OTIS-Mesh
16x16	30	16	6
64x64	126	64	8
256x256	510	256	10
1024x1024	2046	1024	12

Table 6: Reduction - Maximum Number of Electronic Steps

Size	Single port OTIS-Mesh	All port OTIS-Mesh	All port EDN-OTIS-Mesh
16x16	30	24	8
64x64	126	112	10
256x256	510	480	12
1024x1024	2046	1984	14

Appendix B

Latency

1. Scatter operation

Table 7: Scatter - Minimum Latency (milliseconds)

Size	Single port OTIS-Mesh	All port OTIS-Mesh	All port EDN-OTIS-Mesh
16x16	6.46935	3.25635	3.2766
64x64	108.85335	54.44835	54.5631
256x256	1755.69135	877.86735	878.0631
1024x1024	28283.893	14141.96835	14142.35985

Table 8: Scatter - Maximum Latency (milliseconds)

Size	Single port OTIS-Mesh	All port OTIS-Mesh	All port EDN-OTIS-Mesh
16x16	6.46935	5.09235	5.1126
64x64	108.85335	84.68835	96.7371
256x256	1755.69135	1652.01135	1652.3011
1024x1024	28283.893	27425.96835	27426.3936

2. Barrier operation

Table 9: Barrier - Minimum Latency (milliseconds)

Size	Single port OTIS-Mesh	All port OTIS-Mesh	All port EDN-OTIS-Mesh
16x16	1.1341	0.61435	0.34235
64x64	4.6981	2.39635	0.5806
256x256	18.9541	9.52535	0.78985
1024x1024	82.93735	41.46535	1.2016

Table 10: Barrier - Maximum Latency (milliseconds)

Size	Single port OTIS-Mesh	All port OTIS-Mesh	All port EDN-OTIS-Mesh
16x16	1.1341	0.91135	0.3916
64x64	4.6981	4.17835	0.69535
256x256	18.9541	17.84435	1.24235
1024x1024	82.93735	80.35435	1.6876

Appendix C

Execution Time

1. **Reduction operation**

Table 11: Reduction - Minimum Execution Time (milliseconds)

Size	Single port OTIS-Mesh	All port OTIS-Mesh	All port EDN-OTIS-Mesh
16x16	0.8006	0.4226	0.2106
64x64	3.3926	1.7186	0.4386
256x256	13.7606	6.9026	0.8921
1024x1024	55.2326	27.6386	1.5481

Table 12: Reduction - Maximum Execution Time (milliseconds)

Size	Single port OTIS-Mesh	All port OTIS-Mesh	All port EDN-OTIS-Mesh
16x16	0.8276	0.6656	0.3186
64x64	3.4196	3.0416	0.5831
256x256	13.7876	12.9776	1.0271
1024x1024	55.2596	53.5856	1.8181

Appendix D

Performance Improvement

1. Scatter operation

Table 13: Scatter operation

Size	Single port OTIS-Mesh	All port OTIS-Mesh
16x16	1.97	0.994
64x64	1.99	0.998
256x256	2.00	1.000
1024x1024	2.00	1.000

Table 14: Scatter operation

Size	Single port OTIS-Mesh	All port OTIS-Mesh
16x16	1.27	1.00
64x64	1.13	0.88
256x256	1.06	1.00
1024x1024	1.03	1.00

2. Barrier operation

Table 15: Barrier operation

Size	Single port OTIS-Mesh	All port OTIS-Mesh
16x16	3.3	1.8
64x64	8.1	4.1
256x256	24.0	12.1
1024x1024	69.0	34.5

Table 16: Barrier - operation

Size	Single port OTIS-Mesh	All port OTIS-Mesh
16x16	2.9	2.3
64x64	6.8	6.0
256x256	15.3	9.7
1024x1024	49.1	47.6

3. Reduction operation

Table 17: Reduction - operation

Size	Single port OTIS-Mesh	All port OTIS-Mesh
16x16	3.802	2.007
64x64	7.735	3.918
256x256	15.425	7.737
1024x1024	35.678	17.853

Table 18: Reduction operation

Size	Single port OTIS-Mesh	All port OTIS-Mesh
16x16	2.598	2.089
64x64	5.865	5.216
256x256	13.424	12.635
1024x1024	30.394	29.473

Appendix E

Total Overhead

1. Reduction operation

Table 19: Reduction - Minimum Execution Time (milliseconds)

Size	All port EDN OTIS-Mesh
16x16	0.1946
64x64	0.4146
256x256	0.8601
1024x1024	1.5121

Table 20: Reduction - Maximum Execution Time (milliseconds)

Size	All port EDN-OTIS-Mesh
16x16	0.3026
64x64	0.5591
256x256	0.9951
1024x1024	1.7821

Appendix F

Cost

1. Reduction operation

Table 20: Minimum Cost (milliseconds)

Size	Single port OTIS-Mesh	All port OTIS-Mesh
16x16	1.02	0.544
64x64	16.9	8.32
256x256	262.14	131.584
1024x1024	4194.3	2099.2

Table 21: Maximum Cost (milliseconds)

Size	Single port OTIS-Mesh	All port OTIS-Mesh
16x16	1.02	0.816
64x64	16.9	14.56
256x256	262.14	246.72
1024x1024	4194.3	4066.92

الاتصال التجميعي في نظام الترابط المتنقل البصري الشبكي باستخدام نقطة الالتقاء الممتدة المسيطرة

اعداد
روبي يحيى طهبوب

المشرف
د.سامي سرحان

المشرف المشارك
د. باسل محافظة

ملخص

نظام الترابط المتنقل البصري الشبكي هو نظام ترابط شبكات الكتروضوئي يقوم بربط وحدات المعالجة التي تفصلها مسافات قصيرة باستخدام أسلاك توصيل الكترونية بينما يستخدم أسلاك توصيل بصرية لربط وحدات المعالجة التي تفصلها مسافات بعيدة.

تعرض هذه الرسالة تصميم و تنفيذ مجموعة من عمليات الاتصال التجميعي الفعّالة وهي: التبعثر، التخفيض، و الحاجز في نظام (OTIS-Mesh) من خلال تطبيق طريقة نقطة الالتقاء الممتدة المسيطرة(EDN) على نظام (OTIS-Mesh) لتكوين (EDN OTIS-Mesh). علاوة على ذلك، تعرض الرسالة مصفوفة التبديل كتطبيق على حركة البيانات في الاتصال التجميعي.

أجريت دراسة لتقييم ومقارنة لأداء عمليات الاتصال التجميعي التالية: التبعثر، التخفيض و الحاجز في أنظمة ترابط الشبكات التالية: (single port OTIS-Mesh)، (all port OTIS-Mesh)، (all port EDN OTIS-Mesh)، تحت مقاييس الأداء التالية: عدد الخطوات، مدة الانتظار، وقت التنفيذ، اجمالي العبء الزائد، مقدار تحسين الأداء و التكلفة.

أظهرت النتائج التجريبية أن عملية الحاجز في (all port EDN OTIS-Mesh) فاقت أداء كل من (single port OTIS-Mesh) و (all port OTIS-Mesh) بموجب مقاييس الأداء التالية: عدد الخطوات، مدة الانتظار، مقدار تحسين الأداء. على سبيل المثال، الحد الأقصى لمدة الانتظار لأداء عملية الحاجز في (all port EDN OTIS-Mesh) تفوقت على (all port OTIS-Mesh) بمقدار 9.7 و 47.3 من المرات لنظام ترابط الشبكات من حجم 256 و 1024 على التوالي.كما تفوقت عملية التخفيض في (-all port EDN OTIS Mesh) على كل من (single port OTIS-Mesh) و (all port OTIS-Mesh) في مقاييس الأداء التالية: عدد الخطوات، وقت التنفيذ،، اجمالي العبء الزائد ، مقدار تحسين الأداء و التكلفة. على سبيل المثال، الحد الأقصى لتنفيذ عملية التخفيض في (all port EDN OTIS-Mesh) فاقت (all port OTIS-Mesh) بمقدار 12.63 و 29.47 المرات لنظام ترابط الشبكات من حجم 256 و 1024 على التوالي. و من ناحية أخرى فان عملية التبعثر في (all port EDN OTIS-Mesh) تفوقت فقط على (single port OTIS-Mesh) بموجب مقاييس الأداء التالية: عدد الخطوات، مدة الانتظار، ، مقدار تحسين الأداء و أظهرت تقريباً أداءً مماثلاً مع (all port OTIS-Mesh) بموجب المقاييس السابقة.