RTOIN: A New Scalable Optical Interconnection Network

Shaoying Wang  
School of Computer, Xidian University, Xi'an China 710071  
tsoft@163.com

Huaxi Gu  
State key lab of ISN Xidian University, Xi'an China 710071  
hxgu@xidian.edu.cn

Changshan Wang  
School of Computer, Xidian University, Xi'an China 710071  
chshwang@xidian.edu.cn

Abstract
An optical interconnection architecture is proposed, which is based on optical ring and torus. The proposed optical interconnection network combines advantages of both ring and torus, e.g., simple node interface, constant node degree, better support for local communication and remote communication, wide bisection and good scalability etc. This architecture could be used to connect thousands of processing elements in massively parallel computer systems while maintaining a low latency and high bandwidth. The comparison with other popular networks is made, and the result shows that RTOIN (Ring based Torus Optical Interconnection Network) has pretty good node degree, network diameter, the number of links, bisection width and average message distance. Also, the result shows its excellent scalability, so it will probably be widely used in the area of parallel computing in future.

Categories and Subject Descriptors
C.2.1 [Computer-Communication networks]: Network Architecture and Design - Network topology; Network communications, Packet-switching networks

General Terms
Design, Performance, Theory

Keywords
Parallel Computing; Optical interconnection network; Ring; Torus; Scalability

1. INTRODUCTION
With the ever-increasing demand of computer processing capacity, thousands of processors are interconnected by interconnection networks in parallel processing systems[1]. In massively parallel processing system, the performance of interconnection network, to a large extent, determines the system performance and cost. Optical interconnection is becoming more and more popular in parallel processing systems recent years because of its many attractive features such as high speed, wide bandwidth, low latency etc. As a very attractive property of optics, WDM (wavelength division multiplex) can be used to reduce the number of physical channels. So, optical interconnection is very suitable to build massively parallel processing systems. Besides the delay and connectivity, we also need to consider its modularity and scalability.

Several optical interconnection network architectures have been proposed worldwide up to now [2-9]. However, some of them are short of scalability and modularity, and hence limit their development and application in the area of parallel computing. For example: the most popular network in parallel computers is the binary n cube, i.e. hypercube network. Its main advantage is its small diameter. A binary n cube has 2^n nodes, and the diameter is n. This characteristic is fit for the transmission of information in the network. However, there exist some practical limitations. For example, the fact that the node degree grows with its dimension size is difficult for implementation. This is the most serious drawback and is considered the main limiting factor for the use of it in large systems[10]. The mesh network has similar limitation. Although its simple connection and small number of links make it easy to implement, its shortcoming is its large diameter. In order to keep a relatively small diameter, the number of processing elements should not be too large, this ultimately limits the scalability of mesh. We can see that the limitation of many of these networks for parallel computer restricts the applications on parallel computers, and ultimately limit the development of network-based computer software.

This paper presents the ring based torus optical interconnection network, which overcomes the limitation shown above. By using WDM, the interconnection network configuration flexibility is enhanced. With the combination of ring and torus, the problem of scalability is well solved. Also, the use of light as a transmission medium, the electric communication bottleneck between nodes is solved, hence the overall system performance can be greatly improved.

The rest of this paper is organized as follows: In section 2, the architecture of ring based torus optical interconnection network is described; In section 3, interconnection properties are discussed; The comparison with other popular networks is focused in section four; In the next section, routing scheme for RTOIN is depicted; In the last section of this paper, a conclusion is drawn.

2. RING BASED TORUS OPTICAL INTERCONNECTION NETWORK (RTOIN)

2.1 Ring based Torus
The Ring is the basic block of RTOIN and Torus structure is its base architecture. Its structure is shown in Fig.1. The proposed RTOIN network consists of PEs and SEs. PE is the abbreviation of “Processing Element” and SE is that of “Switching Element”. PEs are used to process messages and then send them to certain target PEs or SEs, while the function of SEs is only to send messages to the target SEs.

If there are n PEs in each ring, the size of torus is \( l \times m \), then
the total number of PEs in the network is $n \times l \times m$ and the total number of SEs is $l \times m$. Such a network can be denoted as $(n,l,m)$-RTOIN.

Each circle on the torus with slash lines inside is a ring. For each ring, the circle with nothing inside is a PE, the circle with dots inside is a SE. The horizontal line on each SE is the link of the horizontal direction on torus, and vertical line is the link of the vertical direction on torus.

Each Ring is made up of two unidirectional rings, i.e., the inner ring used for the local communication and the outer ring used for remote communication. This structure provides some degree of path diversity of the RTOIN network. The PE can communicate with the PE located in remote ring by the SE if necessary. We can construct the ring based torus network by connecting all the SEs with torus topology.

The number of wavelengths available has reached as many as a hundred, so if $n = 16$, $l = 16$, $m = 16$ (i.e. A torus whose size is $16^3$, and there are 16 PEs on each ring), then the total number of PEs in the interconnection network is $n \times l \times m = 16 \times 16 \times 16 = 4096$; If $n = 32$, $l = 32$, $m = 32$, then the total number of PEs in the interconnection network is $n \times l \times m = 32 \times 32 \times 32 = 32768$. As another sample, the total number of PEs will reach as many as 262144 if $n = 64$, $l = 64$, $m = 64$.

With the development of optical component, the number of wavelengths available will increase as time goes on. This provides great advantage for the scalability of the proposed network. Therefore, this network is able to meet the demands for upgrading the computer performance in the next few years.

### 2.2 Optical Wavelength Assignment for RTOIN

Assigning a unique wavelength to all processing elements would be an ideal solution since it would make packet routing become an easy task, but on one hand, the number of processing elements in parallel processing systems has reached to as many as thousands at present and will become even more in the future, on the other hand, the number of wavelengths available is far away from that degree, so this simple ideal scheme becomes unpractical. Fortunately, WDM technology can decrease the number of wavelengths because a wavelength assigned to one PE/SE can be assigned to another PE/SE again. In order to keep the number of wavelengths to be assigned as few as possible, it is necessary to take WDM into account.

For the remote communication, a PE on the local ring communicates with another PE on a remote ring. In order to do this, it is necessary to assign a wavelength to each of the rings. That’s because all the PEs on a ring can receive the wavelength assigned to the ring in which the PEs located. In RTOIN, since each WDM physical link carries several distinct wavelength channels, which allows up to several simultaneous data to be transferred. For the torus of $l > m$, $m$ different wavelengths for each row and $l$ different wavelengths for each column are needed, so there are $l$ physical links with $m$ wavelength channels a link in the horizontal direction and there are $m$ physical links with $l$ wavelength channels a link in the vertical direction. Because of WDM, the total number of wavelengths assigned to remote communication should be the bigger one between $l$ and $m$, which can be denoted by the symbol $N_w^{\text{max}}$ as follows:

$$N_w^{\text{max}} = \max\{|R(i)|, 1 \leq i \leq l \times m\}$$  \hspace{1cm} (1)

For example, for the ring on which there are $n = 15$ PEs, the wavelength assignment scheme can be seen in Fig. 2.

For each of the rings, a unique wavelength is assigned to each of the PE on it, so the number of wavelengths assigned to a ring equals to the number of PEs on that ring. Since the wavelengths assigned to one ring can be assigned to another ring again, all the rings can use the same wavelength assignment scheme. Since the number of PEs on each ring may be different, the number of wavelengths assigned to the rings may differ. For identity, the number of wavelengths assigned to each ring is set to be the maximum value of the numbers of PEs in all the rings, which can be denoted by the symbol $N_w^{\text{max}}$ as follows:

$$N_w^{\text{max}} = \max\{|R(i)|, 1 \leq i \leq l \times m\}$$  \hspace{1cm} (1)

For instance: for a $l \times m = 6 \times 5$ torus, the total number of wavelengths assigned to the torus is $N_w^{\text{max}} = \max\{6, 5\} = 6$. An illustration of wavelength assignment scheme for the example is shown in Fig. 2.
So, the total number of wavelengths needed for RTOIN will be the bigger one between the maximum number of wavelengths assigned to local rings and the number of wavelengths assigned to torus, which can be denoted by the symbol $N^m_{\text{total}}$ as follows:

$$N^m_{\text{total}} = \max\{ N^m_{\text{max}}, N^t_{\text{torus}} \} \quad (3)$$

We can see from above that, by using WDM, the number of wavelengths reduces greatly and thus simplify the implementation because the number of physical links will decrease sharply consequently. In this way, WDM enhances the scalability of the proposed network in a sense. In addition, either in basic ring or in the torus structure, WDM makes that different wavelength channels are used when sending messages, in this way, congestion is avoided.

### 3. INTERCONNECTION PROPERTIES

In the following analysis, we assume the $(n,l,m)$-RTOIN is used, in which $n$ is the number of PEs on a ring, $l$ and $m$ represent the number of rows and the number of columns of the torus respectively, $N$ is the total number of PEs on the whole RTOIN network.

#### 3.1 Node Degree

The node degree in a network is defined as the number of links at each node in the network [7]. In RTOIN network, there are two input physical links and two output links for each PE on a ring, so the node degree of each PE is:

$$D_{\text{PE}} = 4 \quad (4)$$

Since the SEs are connected by the torus structure, and there are two input physical links and two output physical links connected with the ring, so the node degree of each SE is:

$$D_{\text{SE}} = 4 + 4 = 8 \quad (5)$$

#### 3.2 Network Diameter

The diameter of a network is defined as the maximum distance between any pair of processors. The distance between a pair of processors is the smallest number of links that have to be traversed from one processor to the other [14]. In the traditional sense, one hop refers to the distance from one node to the other node, but in the RTOIN, because of the usage of WDM, one node can communicate with any of the other nodes directly in the same ring, so there is only one hop within the ring. Similarly, there are two hops at worst that the source node and the target node are not in the same row or in the same column. In this case, the number of hops will be summed to 4. The maximum distance between any PE on RTOIN is 4, i.e., the diameter of RTOIN is:

$$K_{\text{RTOIN}} = 4 \quad (6)$$

#### 3.3 Bisection Width

The bisection width of a network is defined as the minimum number of links that have to be removed to partition the network into two equal halves [16]. Since there are $m$ different wavelengths in the horizontal direction, so the bisection of RTOIN is:

$$B_{\text{RTOIN}} = 2 \times l \times m \quad (7)$$

#### 3.4 Average Message Distance

The average distance is calculated as the sum of the distance of a particular node to nodes within the same basic module and to nodes in the other basic modules divided by $N-1$ [15], which can be denoted as:

$$\hat{\ell} = \frac{1}{N-1} \sum_{i=1}^{N} \frac{d_i}{N_i} \quad (8)$$

where $N_i$ represents the number of PEs at a distance $i$ from the reference PE, $N$ is the total number of PEs in the network, and $K$ is the diameter of the network.

In order to calculate the average message distance of RTOIN clearly, we assume the size of RTOIN is $n \times l \times m$, where $n$ is the number of PEs in each ring, $l$ and $m$ are the number of rows and number of columns of the torus respectively.

For a certain reference PE, the PEs at distance 1 all are on the same ring, so the number of them is:

$$N_1 = n - 1 \quad (9)$$

As for the PEs at a distance 3 or 4, they are located in the rings either in the same row or in the same column with the ring in which the referenced PE located, so the number of PEs at distance 3 is:

$$N_3 = n \times (l - 1) \times (m - 1) \quad (10)$$

where $N_i$ represents the number of PEs at distance $i$.

The rest of PEs are neither in the same row nor in the same column with the ring in which the referenced PE located, the distance between any of them and the referenced PE is 4, so the number of PEs at distance 4 becomes:

$$N_4 = n \times (l - 1) \times (m - 1) \quad (11)$$

Substituting equations (9), (10) and (11) into equation (8) produces:

$$\hat{\ell} = \frac{1}{N-1} \left[ (l \times (n-1)) + 3 \times n \times (l - 1) \times (m - 2) + 4 \times n \times (l - 1) \times (m - 1) \right] \quad (12)$$

### 4. COMPARISON WITH OTHER POPULAR NETWORKS

In this section, we make comparison with other popular networks. These topologies include a traditional Crossbar network (CB), the Binary Hypercube (BHC), the Cube Connected Cycles (CCC) [10], Torus, the Spanning Bus Hypercube (SBH) [11] and the Spanning Multi-channel Linked Hypercube (SMLH) [12]. Each of these networks will be compared with respect to node degree, diameter, number of links, bisection bandwidth and average message distance. These characteristics will be shown in tables 1 and 2. Some of the results in the table are drawn from [7].

Fig. 3 shows a comparison of the node degree of various networks with respect to system size (number of PEs). It can be seen that for networks containing any number of processors, the two torus networks provide a node degree of 4 for torus($w$, $d$=2) and node degree of 6 for torus($w$, $d$=3) configuration respectively. The proposed RTOIN would, require a node degree of 4 for the same size system.

For medium size of networks, the node degree of RTOIN is better than most of the networks, and for very large size of networks (1000 PEs or more), RTOIN maintains the minimum node degree than all the networks except the CCC(d).

Fig. 4 shows a comparison of the diameter of various networks.
with respect to the system size. Since each node of the CB network is directly connected to every other node, so the diameter of the CB network equals to 1. Clearly, it is the best.

For medium size or large size of networks, RTOIN keeps a very good value of diameter 4, it is better than most of the networks.

Torus\((w,d=2)\) and torus\((w,d=3)\) may also keep a low diameter if WDM is fully used in optical interconnection networks. If so, a diameter of 2 for torus\((w,d=2)\) and 3 for torus\((w,d=3)\) can be achieved, but its scalability is largely restricted by the number of wavelengths. For the number of wavelengths 32, the size of a torus\((w,d=2)\) can only build a system whose size is \(32 \times 32 = 1024\), while a RTOIN\((n=32, l=m=32)\) can reach a size of \(32 \times 32 \times 32 = 32768\). The same case is applicable to the torus\((w,d=3)\).

The number of links (along with the node degree of the network) is a nice measurement of the total cost to implement the network, because each link will be converted into some kind of wire(s), waveguide(s), optical fiber(s), or at least some set of optical components (lenses, gratings, etc.) ultimately. So the overall cost of the system is proportional to the number of links [7].

Fig. 5 shows a comparison of the number of network links with respect to the size of the system. As we can see, the RTOIN network is comparable to most of the networks in the case of medium size and in the case of large size, RTOIN network is better than most of the networks. The RTOIN network configurations show a pretty good scalability in the number of links for very large-scale systems.

The bisection width of a network should scale linearly or near linearly with the number of PEs for a scalable network, otherwise the interconnection network will become a bottleneck as the number of PEs increases [7].

Fig. 6 shows a comparison of the bisection width of various network architectures with respect to the number of PEs in the system. Clearly, the CB provides the best bisection width because the number of links between PEs in a CB increases as fast as \(O(N^2)\) with respect to the number of PEs. We can also see that the RTOIN configurations are much better than some of the less scalable networks such as torus\((w,d=2)\) and torus\((w,d=3)\), and is better than SBH[8] and SMLH[9]. For example, for a RTOIN in which \(n = 8, l = 8, m = 8\), the bisection is \(B_{RTO\text{IN}} = 2 \times l \times m = 2 \times 8 \times 8 = 128\), while for the same size of other network \((N = 8 \times 8 \times 8)\).
Table 1 Comparison of Size, Node Degree, Diameter, Number of Links and Bisection Bandwidth with Several Popular Networks

<table>
<thead>
<tr>
<th>Network</th>
<th>Size</th>
<th>Node Degree</th>
<th>Diameter</th>
<th>Number of Links</th>
<th>Bisection Bandwidth(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CB(N)</td>
<td>N</td>
<td>N - 1</td>
<td>1</td>
<td>(\frac{(N^2-N)}{2})</td>
<td>(\frac{N^2}{4})</td>
</tr>
<tr>
<td>BHC(d)</td>
<td>(2^d)</td>
<td>(\log_2N)</td>
<td>(\log_2N)</td>
<td>(\frac{N}{2}\log_2N)</td>
<td>(\frac{N}{2})</td>
</tr>
<tr>
<td>CCC(d)</td>
<td>(d2^d)</td>
<td>3</td>
<td>(\frac{5d-2}{2})</td>
<td>(\frac{3}{2}N)</td>
<td>(\frac{N}{2n})</td>
</tr>
<tr>
<td>Torus(w, d)</td>
<td>(w^d)</td>
<td>2(\log_wN)</td>
<td>(\frac{w}{2}\log_wN)</td>
<td>(N\log_wN)</td>
<td>(2w^{\log_wN-1})</td>
</tr>
<tr>
<td>SBH(w, d)</td>
<td>(w^d)</td>
<td>(\log_wN)</td>
<td>(\log_wN)</td>
<td>(\frac{N}{w}\log_wN)</td>
<td>(2w^{\log_wN-1})</td>
</tr>
<tr>
<td>SMLH(w, d)</td>
<td>(w^22^d)</td>
<td>2 + (\log_w\frac{N}{w})</td>
<td>2 + (\log_w\frac{N}{w})</td>
<td>(\frac{N}{2}(\frac{4}{N}+\log_w\frac{N}{w}))</td>
<td>(N)</td>
</tr>
<tr>
<td>RTOIN(n, l, m)</td>
<td>(n \times l \times m)</td>
<td>4</td>
<td>4</td>
<td>(\frac{2 \times N}{n}(n+2))</td>
<td>(2 \times l \times m)</td>
</tr>
</tbody>
</table>

\(n = \) number of PEs per cluster, \(d = \) dimensionality, \(w = \) number of PEs per bus/ring/multi-channel link, and \(N = \) total number of PEs

Fig. 7 shows a comparison of the average message distance with respect to the number of PEs in the system. We can see that the CB provides the best possible average message distance of 1. The RTOIN network configurations display a good average message distance for medium to large-scale configurations, which is not as good as the average message distance of the SMLH networks, but is much better than most of the remaining networks.

When \(N\) becomes large enough, the average message distance can still maintain a pretty good value. For example, for a RTOIN in which \(n = 32, l = 32, m = 32\), the average message distance \(7 = \frac{32 \times 4 \times 32 \times 32 \times 32 \times 32 - 1}{32767} \approx 3.94 < 4\).

### 5. ROUTING SCHEME FOR RTOIN

The performance of the routing algorithm has much influence on the performance of the interconnection networks, so it is necessary to design a good routing algorithm.

In RTOIN, each ring is composed of two sub rings, i.e. the inner ring and the outer ring, the former one is used for local communication and the latter one is used for remote communication.

#### 5.1 Local Communication

The communication between any two of the PEs in the same ring is called local communication. Since each of the PE in the same ring is assigned a unique wavelength, any of them can communicate directly.

As an example in Fig. 2, we suppose the PEs from 1 to 15 are assigned the wavelengths from 1 to 15 respectively. If PE 1 is about to send message to PE 3, then PE 1 should first tune the wavelength to \(\lambda_3\), and then send the message by transmitter in a certain order. In this way, PE 3 can receive the message come from PE 1.

#### 5.2 Remote Communication

The communication between the local PE and any of the PEs which are in another ring is called remote communication. Apparently, because of the existence of multi-paths in RTOIN...
network, the path diversity is enhanced.

First of all, the source PE should send the message to the SE that is in the same ring with the source PE, the SE then turns the wavelength to what assigned to the ring in which target PE located and send the message out. According to the position of the target ring, the message is routed in horizontal direction or vertical direction or both directions.

If the ring in which the source PE located and the ring in which the target PE located are neither in the same row or in the same column, the source PE send the message to the SE who is in the same ring with the source PE, and then the SE send the message to the ring in which the target PE located in horizontal direction or vertical direction.

If the ring in which the source PE located and the ring in which the target PE located are neither in the same row nor in the same column, then it will be necessary for the certain SE to change the direction of the message.

For example, suppose a PE in the ring which is located in the fourth row and first column in the torus is going to communicate with a PE in the ring which is located in the third row and fifth column in the torus, the message will first be sent to the fourth row and the fifth column and then be sent to the third row and the fifth column. (Or send to the third row and the first column first and then to the target ring.) The routing path is shown in Fig. 2 with bold line. The circle filled with slashed line is the source PE and the circle filled with light shadow is the ring in which the target PE located.

The routing scheme of RTOIN network can be described by pseudocode as follows:

```c
/* Source PE and target PE are in the same ring*/
Turn the wavelength of the source PE to that of the target PE;
Send message out to the inner ring;
Else
/* Source and target PE are not in the same ring*/
Turn the wavelength of the source PE to that of the source ring;
Send message out to the outer ring;
If ( SSource.row = SEtarget.row )
/* The source and the target ring are in the same row of the torus*/
Turn the wavelength of the source SE to that of the target SE;
Send message out in horizontal direction;
Else
If ( SSource.column = SEtarget.column )
/* In the same column of the torus */
Turn the wavelength of the source SE to that of the target SE;
Send message out in vertical direction;
Else
/* Neither in same row nor in same column */
Turn the wavelength of the source SE to that of the ring in the same row with the source SE and in the same column with the target SE;
Send message out in horizontal direction;
The SE receives the coming message then turns the wavelength to that of the target ring;
Send message out in vertical direction

/* Now, the message is in the target ring*/
The target PE receives the message from the SE which is in the same ring with the target PE;
```

6. CONCLUSION

In this paper, a ring based torus optical interconnection network architecture is proposed. For this architecture, efficient routing is realized both in local communication and remote communication by reusing wavelengths. RTOIN has a very good scalability, the size of RTOIN can be increased by increasing the number of rows or the number of columns of the torus without changing the configuration of the basic rings. If there are \( n = 32 \) PEs in each ring and the size of torus is \( l \times m = 32 \times 32 \), then the total number of PEs in the system could reach as many as \( N = n \times l \times m = 32768 \).

At the time of \( n = 64, l = 64, m = 64 \), the size of the system will become \( N = n \times l \times m = 262144 \). So, this may meet the demand of architecture for high performance computing in the next few years. A comparison with other popular networks is made, and the result shows that RTOIN has pretty good properties especially the scalability. Also, the character of path diversity is analyzed. It shows that the architecture is highly fit to be implemented in optics. The physical components required are tunable transmitters, fixed tuned receivers, EDFAs, star couplers, ATOFs and passive couplers. With these components and some existing optical hardware, simple optical implementation of the proposed network is possible which can be used to construct large or very large scale high performance computing network systems.

Our future work is just to use the current available components to build our proposed RTOIN network, and try our best to use inexpensive components so as to reduce the total cost of the system. Also, more efficient components will be studied.

7. ACKNOWLEDGEMENT

This research was supported by the Zhongxing Telecommunication Equipment Corporation (ZTE) Research Fund under Grant No. ZXJS200609120159.

8. REFERENCES


