On the relation between network throughput and delay curves

Mehrnaz Moudi & Mohamed Othman

To cite this article: Mehrnaz Moudi & Mohamed Othman (2020) On the relation between network throughput and delay curves, Automatika, 61:3, 415-424, DOI: 10.1080/00051144.2020.1774731

To link to this article: https://doi.org/10.1080/00051144.2020.1774731

© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 04 Jun 2020.

Submit your article to this journal

Article views: 90

View related articles

View Crossmark data
On the relation between network throughput and delay curves

Mehrnaz Moudi\textsuperscript{a} and Mohamed Othman\textsuperscript{b}\textsuperscript{b}

\textsuperscript{a}Department of Computer and Electrical Engineering, University of Torbat Heydari, Razavi Khorasan Province, Iran; \textsuperscript{b}Department of Communication Technology and Network, Universiti Putra Malaysia, Serdang, Malaysia

ABSTRACT

The theoretical background of network throughput and delay has been widely used in the previous studies to efficiently study the behaviour of different interconnection networks. In this paper, we derive a new relationship between the network throughput and delay curves. Specifically, we prove that this relation on the average delay and throughput in x-Folded TM topology by considering the tangent line of each curve. Based on the achievements, we introduce the reflection relation between these two performance metrics, while considering the gradient of the metric curves. Moreover, we show the superiority of x-Folded TM topology, previously introduced with same authors, is obvious with this relation in interconnection networks. Consequently, the obtained results have been verified with simulation results to signify how to relate the performance metrics for various topologies in interconnection networks.

1. Introduction

With the advance of the technology, smart interconnect architectures integrate the huge number of nodes or processing elements in the forms of processors available on a network. For a considerable time, the number of processors simply increase to get a considerable improvement in the performance. By adding more processors to the network, their communication performance is degraded in sending and receiving packets. The evolutionary advances in interconnection networks are proper and improve the majority of network performance [1–5].

One of the most difficult challenges in these networks is how to design the processors connection without limiting the network performance. It makes an enthusiasm to solve the limitation and propose a novel topology with an efficient routing scheme to find the successful routes from each source to each destination without deadlock in interconnection networks. According to previous studies, we proposed a new topology in [6] for interconnection networks first. Then, the routing algorithm has been investigated in this topology [7, 8]. Through this study, the design of a novel topology is declared to solve the deadlock problem and had an effect on the performance of interconnection networks.

Now in this paper, the contribution is the investigation of the relations among the two performance metrics; throughput and delay, based on the notion of reversed process. Specifically, we prove that the conditions for low delay are sufficient to get a strong throughput in interconnection networks with respect to the saturation point. We also study the relation between throughput and delay using simulation under different traffic patterns for the applied topologies. To end with introducing the simulation environment and presenting the results, both performance metrics are compared in terms of various topologies.

The remainder of the paper is organized as follows. The theoretical background of the proposed contribution is described in Section 2. Section 3 proposes a relationship between throughput and delay according to introducing a new Theorem. In Section 4, the applied topologies and routing are introduced and key simulation parameters are defined as well. Numerical and simulation results are illustrated for various topologies and comparisons made between five different traffic patterns in Section 5. Conclusion and future work are summarized in Section 6.

2. Theoretical background

In general, many approaches have been used to improve the performance of the interconnection networks. To have good performance in the interconnection networks, low delay and high throughput must be achieved. The network throughput and delay are two performance metrics used to assess the simulation results in any studies on interconnection networks [9, 10]. Moreover, the tradeoff between these two metrics

CONTACT Mehznaz Moudi mmoudi@torbath.ac.ir

© 2020 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
or these metrics with other significant metrics is presented in various networks [11–13].

For simulation, Equations (1) and (2) represent the measurement of the performance metrics. Throughput defined when the first channel saturates in the network and delay is another key metric that represents the network latency. These metrics can be assessed analytically or produced by simulations. They measure in flits/node/cycle for each traffic pattern.

\[
D = \frac{1}{N_p} \times \sum_{i=1}^{N_p} D_i \tag{1}
\]

The elapsed time during, which a packet traverses from a source to destination is called the delay in the network. The average delay is the average value of the elapsed time from injection of the first flit of the message from the source to reception of the last flit of the message at the destination. The average delay \(D\) in Equation (1) through this simulation has been computed for the total number of packets \(N_p\).

\[
TP = \frac{\text{Total Received Flits}}{\text{Number of Nodes} \times \text{Total Cycles}} \tag{2}
\]

Throughput refers to the largest amount of traffic accepted for a particular traffic pattern through the network. It can also be defined as the amount of delivered information per unit of time by the network in Equation (2). Network throughput \(TP\) is determined by simulation and measured in flits per node and per clock cycle.

3. Throughput versus delay

For the communication onto the network topology, the routing algorithm offers different trade-offs for achievable performance in different traffic patterns. In switching and flow control, router efficiency and conflict effects are introduced to promote the network performance limitation. Figure 1(a) presents the delay curve with increasing the number of injected packets, while it has been introduced in [14] for first time.

According to Figure 1(a), the throughput curve has been plotted in Figure 1(b). With increasing the number of injected packets, the throughput is increased until reaching the saturation point. All interconnection network classifications have an effect on the throughput curve as well. There is the same saturation point for both delay and throughput curves. Since the throughput curve is the reflection of the delay curve, the exchange of them is presented in this section. First, if we assume:

- Function \(f(b) = a^b\) for several values of \(a\), \(f(b)\) is the value of delay and \(b\) is the value of the packet injection rate at the saturation point.
- Function \(g(b) = \log_a b\) for several values of \(a\), \(g(b)\) is the value of throughput and \(b\) is the value of the packet injection rate at the saturation point.
- Saturation point \(b = \log_a f(b)\) is the value of the saturation point based on the delay.
- Line \(y = x\) is the estimated tangent line to the saturation point for the delay and throughput curves.

**Theorem 3.1:** The throughput curve is a reflection of the delay curve based on the saturation point, while the gradient of the tangent line touches the delay and throughput curves at this point.

**Proof:** By substituting the saturation point.

\[
g(b) = \log_a b = \log_a \log_a f(b) \tag{3}
\]

Since \(b\) has been assumed as the saturation point, choosing \(a\), as the base of the functions \(f(b)\) and \(g(b)\) makes the relationship much simpler. It is calculable based on saturation point and delay or throughput

![Figure 1](image-url). Measured curves for delay and throughput: (a) delay and (b) throughput.
presented in the simulation results. Thus,

$$g^{(b)} = \log_a f(b). \quad (4)$$

Consequently, Theorem 3.1 is suited to finding the relation of delay and throughput based on the saturation point. The throughput curve is the reflection of the delay curve.

4. Simulation framework

The application of the related studies was simulated through two approaches: designing new network topology and developing fast router to minimize the deadlock as a severe problem in interconnection networks. When a network congested, deadlock is an anomalous state among the network resources without a careful design for a network. Deadlock is prone to occur while the deadlocked packets completely stop and another packet delivery is delayed. Deadlock problem plays a vital role in interconnection networks and it can be solved using a new design. Thus, the network topology and the employed routing within its influence on the network performance. The key element in the network performance is having an attractive topology with considering deadlock problem in terms of the delay and throughput.

4.1. Interconnection topologies

For low communication delay and having improved network performance, it needs an interconnection network with the low diameter and average distance topology. In previous studies [15–18], different interconnect topologies, such as k-ary n-cubes where n is the number of dimensions and k is the number of nodes in each dimension, have been utilized as the popular topologies. Among most popular topologies, Mesh and Torus topologies have achieved more importance by researcher due to their simplicity. The definitions of these topologies according to the coordinate of their nodes are as follows.

Definition 4.1: In a Mesh network, each node is represented by a two element coordinate \((x, y)\) and is a valid node if \(0 \leq x \leq (k - 1)\) and \(0 \leq y \leq (k - 1)\). The next nodes along the y-axis are \((x + 1, y)\) if \(x < (k - 1)\) and \((x - 1, y)\) if \(x > 0\). Along the x-axis, the next nodes are \((x, y + 1)\) if \(y < (k - 1)\) and \((x, y - 1)\) if \(y > 0\).

Definition 4.2: A Torus has \(2n\) added wraparound links, connecting the leftmost node to the rightmost node in the same row or connecting the uppermost node to the lowermost node in the same column. That is, two nodes \((x, 0)\) and \((x, k - 1)\) are connected by a wraparound link for \(0 \leq x \leq (k - 1)\), and two nodes \((0, y)\) and \((k - 1, y)\) are connected by a wraparound link for \(0 \leq y \leq (k - 1)\).

As an example for \(k = 8\) and \(n = 2\), Mesh and Torus topologies are shown in Figure 2(a,b) respectively.

Moreover, one of the recent topologies is the proposed solution by [19]. They proposed TM topology, which is a new and simple topology in interconnection networks. Its idea is removing one link in each row (column) of the Torus to split the cycles in each dimension of the Torus to form the TM topology.

Definition 4.3: Along x-axis in TM network, node \(a = (x, y)\) and node \(b = ((x + 1) \mod k, y)\) are the coordinates of the link \((a, b)\). Then, the links \((a, b)\) and \((b, a)\) are removed from the Torus if \((x + y + 1) \mod k = 0\). The links \((a, b)\) and \((b, a)\) along the y-axis are removed if \((x + y) \mod k = 0\), where \(a = (x, y)\) and \(b = (x, ((y - 1 + k) \mod k))\).

TM topology is a combination of Torus and Mesh topologies and includes some of the advantages of both the Mesh and Torus topologies to form its design. Figure 2(c) shows a TM topology when \(k = 8\) and \(n = 2\).

In [19], TM has been rearranged to form the nodes with their connections clearly. In rearranged TM topology (Figure 3 a), the node coordinates will be changed to new coordinates. Node \((x, y)\) is mapped to \((x, y)\) from the original topology (Figure 3) to the rearranged topology if \(x + y < k\) or else node \((x, y)\) is mapped to \((x, y - k)\). In other words, TM network can be divided into two subnetworks, UpTriangle and DownTriangle, which are shown in Figure 3(b). A node belongs to the UpTriangle network if its rearranged coordinate along dimension \(Y\) is no less than 0. In the DownTriangle, a node belongs to it if its rearranged coordinate along dimension \(Y\) is no more than 0.

By folding the rearranged TM topology based on the imaginary x-axis, which has been illustrated with dash line, an x-Folded TM topology is created. An x-Folded TM is derived from a TM topology by removing several links and sharing several nodes.

Definition 4.4: In x-Folded TM topology, node \((x, y)\) is a valid node if \(0 \leq x \leq (k - 1)\) and \(0 \leq y \leq (k - 1)\). Along y-axis, the nodes connecting to node \((x, y)\) are: \((x + 1, y)\) if \(x < (k - 1)\) and \((x - 1, y)\) if \(x > 0\). Along x-axis, nodes \((x, y + 1)\) if \(y < (k - 1)\) and \((x, y - 1)\) if \(y > 0\) are connected to this node. Then, node \((x, y)\) is removed from the x-Folded TM if \((x + y) \mod k = 0, (k - 3), \) where \(x > y\) and \((k - 3) \leq x \leq (k - 1)\) and \(1 \leq y \leq (k - 2)\). In addition, there is no link between two nodes \((x, y)\) and \((x + 1, y)\) if \(x = i\) and \(y = i + 1\), when \(k\) is even and \(i = \left(\frac{k}{2}\right) - 1\).

Eventually, x-Folded TM is illustrated in Figure 4. Considering the topology properties as basic requirements for each interconnection network, the properties of x-Folded TM topology proved the superiority of the
4.2. Routing algorithms

An Interconnection Network comprises a complex connection of switches and links that communicate among the cores. Efficient communication has a vital role to meet high network performance. A problem on the routing in each network topology appears when the received packets blocked in the network, which reduces communication efficiency and reliability of the network performance considerably. To solve this problem in applied interconnection topologies, the applied routing algorithms are deterministic and adaptive routing algorithms.

The simplest and reliable deterministic routing is dimension-order. Figure 5(a) displays the dimension order routing scheme in a Mesh topology when $k = 6$ and $n = 2$. This routing scheme determines packets direction in each dimension only for the routing between each source and destination pair. A route is successful in each dimension when the packet arrives at the proper coordinates and the distance is zero, then the route will proceed to the next dimension. The dimension-order algorithm is popular in interconnection networks because it has minimal hardware requirements and is exceedingly simple and fast; being as fast as
possible is important for the routing decision to reduce network delay [20, 21].

The second algorithm is adaptive routing algorithm, which by adding adaptive features to the routing can improve the deterministic routing algorithm. Figure 5(b) displays the adaptive routing scheme in a Mesh topology \( k = 6 \) and \( n = 2 \). An adaptive routing algorithm is more suitable for dynamic networks to choose the routing paths. The chosen path is used if the links are not busy when the packets arrive. A number of pairs of nodes transmit packets simultaneously using alternate paths without blocking. The algorithm improves the performance potential and its delay is near to deterministic algorithm, however, it increases the cost of preventing deadlock, which overweights the adaptive routing advantages [22, 23].

In this study, both of these algorithms are developed for the applied interconnection topologies to avoid congestion and improve the performance metrics.

### 4.3. Simulation model

In general, simulation is performed to evaluate the performance of a system. The simulation model in this study is carried out by Booksim2.0 simulator to evaluate the performance of the proposed topology along with other benchmark topologies. It is a cycle-accurate interconnection network simulator that was originally introduced with the Principles and Practices of Interconnection Networks book [24]. The current simulator supports a range of topologies, provides diverse routing algorithms, and includes many options for customizing a networks router microarchitecture. The simulator comprises a collection of routers and channels with topologies, which defines the interconnection between routers and channels. By using this simulator, the simulation is accompanied based on the parameters in Table 1.

In an interconnection network, the message length, number of VCs, and destination distribution have significant implications on the performance. Simulation studies are conducted for different topologies by varying the routing algorithms in this research. The simulation involves all topologies, when \( k = 8 \) and \( n = 2 \), under different traffic patterns.

### Table 1. Simulation parameters.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topology</td>
<td>Specify the network connectivity with size (k and n) and configuration for the selected topology.</td>
</tr>
<tr>
<td>Routing</td>
<td>Specify the routing algorithm for the selected topology.</td>
</tr>
<tr>
<td>Switching Strategy</td>
<td>Specify the switching strategy.</td>
</tr>
<tr>
<td>Flow Control</td>
<td>Specify number of virtual channel and buffer size.</td>
</tr>
<tr>
<td>Traffic Pattern</td>
<td>Traffic patterns used such as uniform, transpose.</td>
</tr>
</tbody>
</table>

In Figure 4, the x-Folded TM topology for an \((8 \times 8)\) interconnect network is shown.

In Figure 5, various routing schemes are displayed: (a) Dimension-Order and (b) Adaptive.
Table 2. Simulation parameters values.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topology</td>
<td>Mesh, Torus, TM, x-Folded TM</td>
</tr>
<tr>
<td>Routing</td>
<td>Deterministic and Adaptive Algorithms</td>
</tr>
<tr>
<td>Network Size</td>
<td>$k = 8, n = 2$</td>
</tr>
<tr>
<td>Switching Strategy</td>
<td>Wormhole Switching</td>
</tr>
<tr>
<td>Virtual Channel</td>
<td>2 VCs</td>
</tr>
<tr>
<td>Buffer Size</td>
<td>2</td>
</tr>
<tr>
<td>Simulation Time</td>
<td>10000 Cycles</td>
</tr>
<tr>
<td>Traffic Pattern</td>
<td>Uniform, Bit-Reversal, Transpose, Hotspot</td>
</tr>
</tbody>
</table>

A set of parameters for use in the simulations are presented in Table 2. All simulated use the packets with 8 flits size and flits are transmitted at 10,000 cycles. A packet consists of one or more flits, the shortest kind of packets has been selected for this simulation. During each cycle, a single flit is read from its input and written to its output. For fair performance comparisons, two virtual channels per link are simulated.

5. Performance evaluation

The performance evaluation is based on standard $(8 \times 8)$ different topologies with comparing their performance on deterministic and adaptive routing algorithms in this section. The comparison between topologies is made in terms of average delay and network throughput in the presence of four traffic patterns, while we presented the relationship between these two performance metrics in x-Folded TM topology as well.

5.1. Average delay

A comparative study of developing the dimension-order and adaptive routing algorithm on various topologies under different traffic patterns shows that x-Folded TM topology gives the higher performance in terms of the average delay in interconnection networks.

Figures 6 and 7 present the average delay for different topologies based on the dimension-order and adaptive routing algorithms under different traffic patterns. Obviously, the efficiency of the x-Folded TM topology performance is presented, while this improvement based on adaptive routing algorithm is more than dimension-order routing algorithm. We are sure that x-Folded TM topology can be accepted as the efficient topology under different traffic patterns to overcome the deadlock problem. It is worthy to mention that the saturation point of the x-Folded TM topology will be in more packet injection rate, however, this saturation will be sooner for this topology under hotspot traffic pattern in Figure 6(d) and Bit-reversal Traffic pattern in Figure 7(b).

Figure 6. Average delay performance by dimension-order routing under four different traffic patterns: (a) Uniform, (b) Bit-Reversal, (c) Transpose and (d) Hotspot.
Figure 7. Average delay performance by adaptive routing under four different traffic patterns: (a) Uniform, (b) Bit-Reversal, (c) Transpose and (d) Hotspot.

Figure 8. Network throughput performance by dimension-order routing under four different traffic patterns: (a) Uniform, (b) Bit-Reversal, (c) Transpose and (d) Hotspot.
5.2. Network throughput

Next, we compare the network throughput of these topologies using dimension-order and adaptive routing algorithms. The same set of conclusions hold with respect to the throughput performance of all topologies under four traffic patterns. Results are shown in Figures 8 and 9. Figure 8 shows the network throughput for different topologies by dimension-order routing algorithm. We can see from these figures that all topologies, such as Mesh, Torus, TM and x-Folded TM, have similar network throughput, however the injection rate at the saturation point is less in bit-reversal traffic pattern. Furthermore, it is clear that maximum throughput or x-Folded TM topology is less than the other topologies under hotspot traffic pattern.

Figure 9 shows the worse network throughput under bit-reversal traffic pattern for Torus and TM topologies. It could be noted from the figure that the saturation points significantly differ for x-Folded TM topology, where Mesh, Torus and TM topologies achieve lower maximum network throughput. It is proven the design of x-Folded TM topology improves the performance than other topologies.

5.3. Relation between delay and throughput

Since there is a reflection relation between the delay and throughput at the saturation point which was proved in Section 3, we now present this relation in x-Folded TM topology as one of the latest based on its saturation point. Tables 3 and 4 show the simulation results of x-Folded TM topology at the saturation point in terms of the average delay and network throughput under different traffic patterns with considering dimension-order and adaptive routing algorithms.

According to the simulation results in previous Sections (5.1 and 5.2), the average delay has a reflection relation with the network throughput based on the saturation point in these tables. The estimated tangent lines

Table 3. Delay vs throughput for x-Folded TM topology at saturation point based on dimension-order routing.

<table>
<thead>
<tr>
<th>Traffic patterns</th>
<th>Saturation point</th>
<th>Average delay</th>
<th>Network throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>0.0090</td>
<td>49.06</td>
<td>0.072</td>
</tr>
<tr>
<td>Bit-Reversal</td>
<td>0.0080</td>
<td>60.67</td>
<td>0.060</td>
</tr>
<tr>
<td>Transpose</td>
<td>0.0090</td>
<td>53.60</td>
<td>0.072</td>
</tr>
<tr>
<td>Hotspot</td>
<td>0.0005</td>
<td>98.60</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Table 4. Delay vs. throughput for x-Folded TM topology at saturation point based on adaptive routing.

<table>
<thead>
<tr>
<th>Traffic patterns</th>
<th>Saturation point</th>
<th>Average delay</th>
<th>Network throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>0.0100</td>
<td>58.38</td>
<td>0.082</td>
</tr>
<tr>
<td>Bit-Reversal</td>
<td>0.0060</td>
<td>96.58</td>
<td>0.045</td>
</tr>
<tr>
<td>Transpose</td>
<td>0.0090</td>
<td>54.93</td>
<td>0.080</td>
</tr>
<tr>
<td>Hotspot</td>
<td>0.0009</td>
<td>49.93</td>
<td>0.008</td>
</tr>
</tbody>
</table>

Figure 9. Network throughput performance by adaptive routing under four different traffic patterns: (a) Uniform, (b) Bit-Reversal, (c) Transpose and (d) Hotspot.
to the saturation point of delay and throughput curves are $y = x^4$ and $y = 4 \times x$, respectively. The gradient of this tangent line touches the saturation point in each curve. Thus the approximated value of average delay is calculable based on the value of network throughput considering the estimated tangent line to the saturation point. All the results presented in these tables prove the reflection relation between the average delay and network throughput on the same saturation point.

Based on this relation, the reduction in the average delay is in agreement with the increment in the network throughput obviously. Moreover, the improvement in the average delay and network throughput obtained with increasing the injection rate in x-Folded TM topology was presented in comparison with other topologies in the simulation results. Consequently, the superiority of x-Folded TM topology is that the saturation point happens in the larger injection rate, which proves the efficiency of x-Folded TM topology in solving deadlock problem over other topologies.

6. Conclusion

This study has been devoted to present the relation between two main performance metrics: Throughput and Delay. We began with an introduction and continued with proposing the metrics relation, which is the main contribution in this research. Then, the application of these metrics in interconnection networks has been presented. Our simulation environment for performance evaluation has been described as well. The applied performance metrics for evaluation and their relation have been mentioned and we contributed how to relate the performance metrics for various topologies in interconnection networks. The obtained results have been verified with simulation results and proved the relation between the performance metrics.

There are still new relations among the important performance metrics in interconnection networks. We believe that the relation between Throughput and Delay will cause significant impact on other performance metrics such as Packet Loss. Thus we will carry out related future work.

Acknowledgments

This work was supported by the Universiti Putra Malaysia under High Impact Putra Grant: UPM/700-2/1/GPB/2017/9557900.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Universiti Putra Malaysia under High Impact Putra Grant: UPM/700-2/1/GPB/2017/9557900.

ORCID

Mohamed Othman http://orcid.org/0000-0002-5124-5759

References

[16] Abd-El-Barr M, Al-Somani TF. Topological properties of hierarchical interconnection networks: a review


