2019

Computer interconnection networks with virtual cut-through routing

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 155 (2019) 449–455

Procedia Computer Science

www.elsevier.com/locate/procedia

The 14th International Conference on Future Networks and Communications (FNC)
August 19-21, 2019, Halifax, Canada

Computer interconnection networks with virtual cut-through routing

Lev B. Levitin[a], Yelena Rykalova[b]*

[a]*Boston University, 8 St.Mary's St.,Boston, MA 02215, USA*
[b]*UMass Lowell, 220 Pawtucket St, Lowell, MA 01854, USA*

Abstract

This paper considers a model of a toroidal computer interconnection network with the virtual cut-through routing. The interrelationships between network parameters, load and performance are analyzed. An exact analytical expression for the saturation point and expressions for the latency as a function of the message generation rate under the mean field theory approximation have been obtained. The theoretical results have been corroborated with the results of simulation experiments for various values of network parameters. The network behavior has been found not depending on the torus linear dimensions provided that they are at least twice as large as the message path length. The saturation point has been found to be inversely proportional to the message length in good agreement with the analytical results. A good agreement with Little's theorem has been found if the network remains in the steady state during the experiment.

Keywords: computer interconnection networks; network-on- chip; network torus topology; cut-through routing; latency; saturation in networks

1. Introduction

Many commercially available parallel computers use extensive communications between otherwise independent nodes. This communications network is implemented as a network of interconnected routers (each having its local processor) [1, 3, 4, 6-10]. Various routing techniques are used in interconnection networks [12, 15]. Store-and-forward approach is based on the assumption that an entire message must be received at any intermediate node before it can be forwarded to the next node. Obviously, for a long message, the total delivery time may turn out to be quite large. To the contrary, in wormhole routing, each message is divided into small packets – flits. The header flit contains information about source and destination, and is routed through the network according to this information and routing algorithm. Other flits of the message follow the header flit. When header flit of the message is blocked at an intermediate node because the requested link is occupied by another message, the flits are buffered at each node along the path up to the current

* Corresponding author. Tel.: +1-978-934-5288; fax: +1-978-934-3551.
E-mail address: Yelena_Rykalova@uml.edu

10.1016/j.procs.2019.08.062

node. This forms a long "worm" which remains in the network blocking other messages, thereby increasing their delivery time. Also, the problem of deadlocks emerges in this approach and should be dealt with [17, 18].

Virtual cut-through (VCT) routing algorithm is supposed to mitigate the drawbacks of both the above-mentioned techniques (e.g., [4, 6, 11-14, 15, 19-21]). Unlike the wormhole approach, in VCT routing, if the next node cannot accept the message, the current node must still be able to buffer the rest of the incoming message from previous nodes. Thus the VCT algorithm achieves a much higher throughput and avoids deadlocks at the expense of increased buffer capacity.

Several papers were devoted to the comparison between different routing techniques (e.g., [12, 15]). Analytical models of interconnection networks were considered in [2, 3, 20, 22]. Certain practical implementations were described in [4, 5, 16, 19, 21]. Network latency in VCT networks is defined as the average time from the moment a message is generated by the source processor to the moment when the last flit of the message enters the consumption channel of the destination processor. The network latency consists of propagation delay, router delay, and contention (blockage) delay.

In this paper, we study network latency and saturation using the VCT routing policy. At each network node, we introduce an (unlimited) storage buffers. The "unlimited" buffer model means in practice that the network throughput is limited by the link occupancy (utilization), rather than by the buffer capacity. So, after the header is blocked, the "worm" collapses ("condenses") into the storage buffer. This method prevents deadlocks and improves network performance (increases bandwidth, decreases latency, delays saturation).

2. Communication network model

The following assumptions have been made for the network model implementation:

- The storage buffers are unlimited and use FIFO (first-in, first-out) policy.
- The same clock is used for all network nodes.
- When message is generated, it takes one time unit for its header (if not blocked) to appear at the router internal input port.
- It takes two time units (a time unit can include one or more clock cycles) to move a header flit from a router input port to its output port.
- It takes one time unit to move a flit (except the header) from a router input port to its output port.
- It takes one time unit to move a flit from the router output port to the input port of the next node (to go through the link).

2.1. Network topology model

In many systems where the VCT routing may be used, the physical distance between communication nodes is small and thus unimportant. In such systems, real network topology can be abstracted without loss of generality as easily constructed lower-dimension meshes and tori. This paper deals with two-dimensional torus networks. The symmetry of toroidal networks leads to a more balanced utilization of communication links than "open" mesh topologies and improves scalability. Each node in such a network consists of a router and a local processor. Each router has four external input/output (I/O) channels, and one internal input/output channel to the local processor. All I/O channels are bidirectional so that two messages may travel simultaneously in the opposite directions between the nodes.

2.2. Routing

Each input/output port of the router contains three buffers: input buffer, output buffer and output storage buffer.

All input and output buffers can hold only one flit at a time. In our model, the storage buffer (which is an extension of the output buffer) is assumed to be "unlimited". This means that it can hold as many messages as needed at each moment of time. The local processor has the same one-flit sized input and output buffers. We also introduce an unlimited storage buffer, so the local processor can store a new incoming message in its output storage buffer until the required link is free. We assume that it takes two time units in the router for routing a header flit. Because all other flits of the message just follow the header, it takes one time unit in the router to send them to the correct output port.

We use the deadlock-free adaptive unicast VCT routing algorithm as described below.

- Every router has a (static) two-dimensional routing table relating minimum length path from each of the router's four output ports to each network node.
- The routing table is used to perform dynamic routing based on the following set of rules:
  - If the current node (node to which the message header has arrived) is the destination, the header is routed to the internal port connecting to the local processor.

- Else header of the message is sent from the input port to the output port which has the shortest path to the destination node.
- In general, more than one output port may have the minimum distance to destination, so the header is routed to the first available (free) port, where "first" refers to the port with the smallest number.
- If all ports with minimum distance to the destination are busy, the router sends the header to the storage buffer of that one from these ports whose number is the largest one.

- If more than one simultaneously arrived headers should be routed to the same output port based on the rules described above, the header of the message with the smallest identification number will be processed first, and the header(s) of message(s) with larger identification number(s) will remain in the output storage buffer.
- Flits follow the header. If header motion is blocked, the header is routed to the storage buffer, all flits follow the header and accumulate (condense) in the storage buffer.

2.3. Message generation

Assume that at each time unit, every node in the network can generate a message with probability λ independently of all other nodes. Destination nodes for generated messages are selected randomly among nodes having the specified distance l from the source node. Obviously, increasing λ increases the network load (the number of messages simultaneously traveling in the network), which, in turn, leads to the latency growth until network saturation is reached.

3. Theoretical Background

We consider three different network states: startup, steady state, and saturation.

When the network simulation starts up, initially there are no messages in the system. Then, new messages start appearing in the network. Even in the absence of other messages, certain time τ_{min} is required for a message to reach its destination. During this time more messages can be generated, so initially after the startup the number of messages in the network increases. When the network reaches its steady state, the average number of messages generated during time Δt equals the average number of messages delivered during the same time Δt, so that the number of messages in the system (in transit from source to destination) becomes approximately constant over time. For a meaningful evaluation of the latency, network must reach its steady state before data on the network behavior should be collected.

The network load increases with the number of messages present in the network and with the message length since longer messages occupy links for a longer time. As a result, for each message length, the network can accommodate only a limited range of message generation rates. If message generation rate is too high, the number of messages generated during time Δt exceeds the number of messages that can be delivered during this time, and the number of messages in the system is increasing with time with no bound. Thus, the latency tends to infinity, and network becomes dysfunctional. This is the state of saturation.

The theoretical analysis of the network performance with the VCT routing is a challenging problem. In [23], an expression for the saturation point (message generation rate at which network saturates) and approximate expressions for the latency (under the assumptions of the "mean field theory" have been obtained.

The critical value λ_{cr} is given by

$$\lambda_{cr} = \frac{4}{lm}. \tag{1}$$

Here m is the length of the message (in flits) and l is the distance (number of hops) from the source to the destination. (It is taken into account that the number of links in a bidirectional toroidal network with n nodes is $4n$.)

The approximate expression for the latency τ is

$$\tau = (l+1)\left(\frac{\rho}{1-\rho}+3\right)+m = (l+1)\left(\frac{\lambda lm}{4-\lambda lm}+3\right)+m, \tag{2}$$

where $\rho = \frac{\lambda lm}{4} = \frac{\lambda}{\lambda_{cr}}$ is the link utilization.

Expression (2) shows that transition to saturation is a second-order (continuous) phase transition with a critical exponent equal to 1, in agreement with the "mean field" theory.

Note that for small λ ($\lambda \ll \frac{4}{lm}$), the latency is a linear function of λ and depend linearly also on the length of the message m, while the dependence on the distance l has a small quadratic term. The plots of τ as function of ρ with m and l as parameters are given in Figure 1.

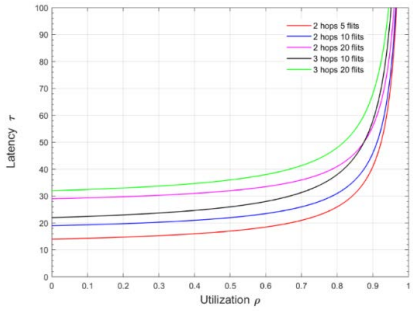

Fig. 1. Theoretical approximation of latency τ as a function of utilization ρ. $l = 2$, 3 hops, $m = 5$, 10, 20 flits.

4. Simulation

4.1. Simulation procedure

Numerical experiments were conducted for wrapped-around meshes with sizes ranging from 2×2 to 12×12 nodes. The relatively small network sizes have been chosen in order to reduce the simulation time. The distances from the source to destination and the message length varied for different experiments, but were assigned prior to the simulation and kept constant during simulation run.

The network performance is characterized by latency (average delivery time) as a function of the network load, and by its saturation load, which describes maximum network capacity. The latency is obtained by averaging delivery times for all messages generated during time T. The time T is selected according to the value of λ so that the total number of messages generated in the system during time T per each destination node will be about the same for all values of λ. Since the average number of messages generated per unit of time is proportional to λ the following empirical formula has been used to estimate T: $T \approx 10 \times 4l/\lambda$.

To ensure that the network is in its steady state during data collection period, the delivery time was recorded for messages generated within time interval $(t_{min}, t_{min} + T)$, where t_{min} was sufficiently large.

4.2. Simulation results

In each simulation run, we analyzed the number of messages in the network as a function of time to determine network state and make sure that it is computed using data collected in the steady state.

Message latency was analyzed as a function of network load (as parameterized by λ) using message lengths of 10 and 20 flits for mesh sizes ranging from 2×2 to 8×8 (network linear size $d = 2$, 4, 6, 8). To ensure steady state, $t_{min} = 50000$ was used.

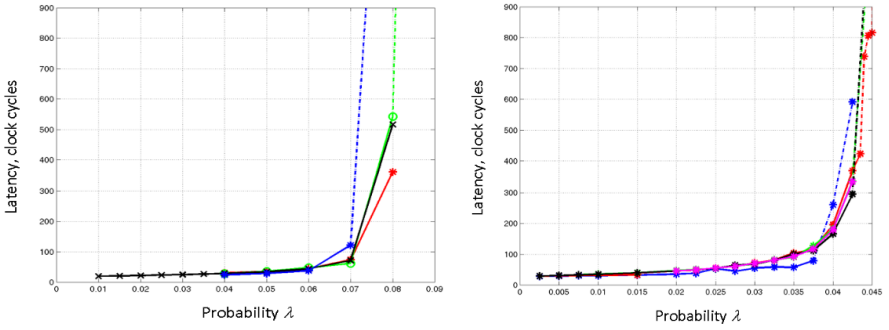

Fig. 2. (a) Latency for message length $m = 10$ flits. (b) Latency for message length $m = 20$ flits. Path length $l = 2$ hops. Solid line: steady state; dashed line: steady state could not be reached. Mesh size 2×2 (blue line), 4×4 (red line), 6x6 (green line), 8×8 (black line), and 12×12 (magenta line).

Results for $l = 2$ and message length of $m = 10$ and $m = 20$ flits are shown in Figures 2a and 2b, respectively. Since $d \geqslant 2l$ for $d = 6$, 8, 12 and $l = 2$, we expect similar network behavior while the results for $d = 2$ demonstrate early saturation.

Simulation results for $l = 3$ are shown in Figures 3a and 3b for message lengths $m = 10$ and 20, respectively. Again, similar network

behavior is observed, when $d \geqslant 2l$ condition is satisfied.

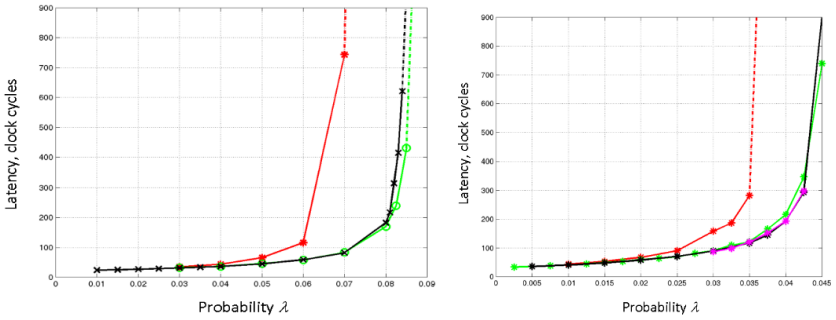

Fig. 3. (a) Latency for message length $m = 10$ flits. (b) Latency for message length $m = 20$ flits. Path length $l = 3$ hops. Solid line: steady state; dashed line: steady state could not be reached. Mesh size 4×4 (red line), 6×6 (green line), 8×8 (black line).

4.3. Dependence of the message latency and saturation on the message length

For the same message generation rate λ, longer messages result in higher network loads. Thus, we expect larger

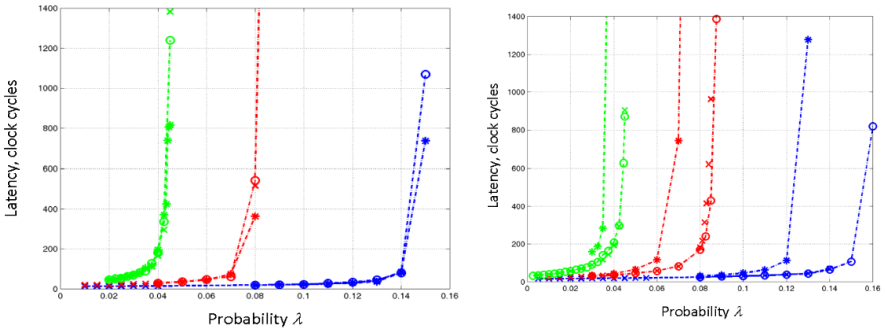

Fig. 4. Latency as function of λ various message lengths: $m = 5$ (blue lines), $m = 10$ (red lines), $m = 20$ (green lines), and for (a) $l = 2$ and (b) $l = 3$. Mesh sizes: 4×4 (∗), 6x6 (○), 8×8 (×).

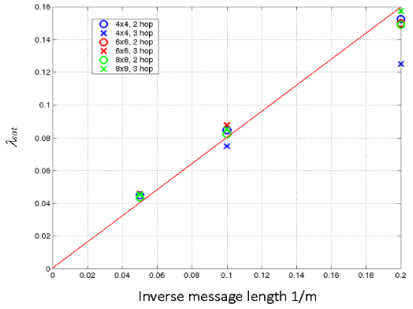

Fig. 5. Message generation probability at which network saturates (λ_{sat}) as a function of the inverse message length $1/m$. Red line $\lambda_{sat} = 0.8/m$ shown for reference.

latencies and earlier saturation for longer messages. Message latencies for the 2-hop path in 4×4, 6×6 and 8×8 meshes are shown in Figure 4(a) for message lengths $m = 5$ (blue), $m = 10$ (red), and $m = 20$ (green). Results are shown using star symbols (∗) for 4×4 mesh, circles (○) for 6×6 mesh, and crosses (×) for 8×8 mesh. Similar results for 3 hops paths are shown in Figure 4(b).

Obviously, the network load increases with the message length. According to expression (3) the message generation rate at which network saturates (λ_{sat}) is inversely proportional to the message length m. Our numerical experiments are in a good agreement with this

result (see Figure 5). It is seen that for all cases, when $d \geq 2l$ the dependence of λ_{sat} of the message length m can be closely approximated as $\lambda_{sat} = 0.8/m$ (solid red line in Figure 5).

4.4. Number of messages in the system as a function of message load

In the steady state, the relationship between the average number of messages in the system and the latency is given by Little's theorem [1] $N = \lambda n \tau$. Here N and τ are expected values of two random variables: number of messages in the network N_s, sampled over the total period of observation, and the sample delivery time τ_s. Therefore, the values of N_s and τ_s fluctuate with time and the relationship between N_s and τ_s satisfies Little's theorem only approximately.

We have measured values of N_s by averaging the number of messages in the network from t_{min} up to the end of simulation, as well as calculated the number of messages in the network using the observed values of τ_s. As shown in Figures 6(a) and 6(b), the directly measured values of N_s and those calculated by the use of Little's theorem are in a very good agreement, which supports the validity of the simulation experiments. However, note that when network state is closed to saturation, the calculated number of messages usually exceeds the measured value.

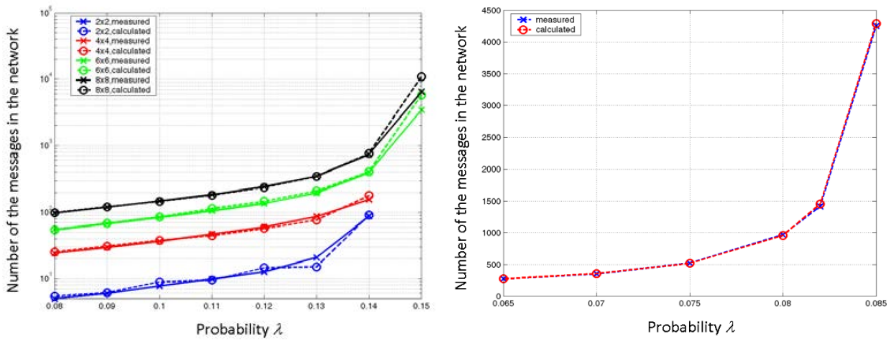

Fig. 6. The number of the messages in the network as function of λ. Solid line: measured during simulation; dashed line: calculated by the use of Little's theorem. (a) Message length $m = 5$ flits. Path length $l = 2$ hops; (b) Message length $m = 10$ flits. Path length $l = 3$ hops.

5. Conclusions

A model of a 2-dimensional toroidal interconnection network with virtual cut-through routing has been studied. An analytical expression for the saturation point and approximate expressions for the network latency for the ranges of small network loads and loads close to the critical value have been obtained.

- The critical value of the probability of message generation $l = l_{cr}$ is inversely proportional to the distance between the source and the destination l and the length of messages m: $\lambda_{cr} = \frac{2}{lm}$.
- For a given saturation rate λ the latency is a linear function of the message length m.
- The latency τ at the saturation point experiences a second-order (continues) phase transition with the critical exponent equal to 1.
- For small values of λ, the latency grows as a linear function of λ.

Simulation experiments have been performed in order to find out and analyze certain empirical relationships that can be used as a starting point for a deeper theoretical analysis and further research. In particular, the following results have been obtained.

- Network behavior (latency and saturation point) does not depend on the mesh size if the mesh is "large enough" compared to the path length. As an appropriate criterion, the mesh linear dimension should be at least twice as large as the message path length: $s \geq 2l$.
- For the same message generation rate, latency increases and saturation occurs earlier for longer messages. It appears that the saturation point λ_{cr} (message generation rate at which network saturates) is inversely proportional to the message length. If the condition $s \geq 2l$ is satisfied, numerical results are in a good agreement with a simple empirical relation $\lambda_{cr} = 0.8/m$ independently of the mesh size. (It seems to be consistent also with the theoretical expression (3)).
- If the network is in the steady state, the independently measured number of messages N_s and the average delivery time τ_s are in a good agreement with Little's theorem for their expected values $N = \lambda n \tau$.

References

[1] Kleinrock, L. (1975). Queueing Systems Volume I: Theory. New York: Wiley.

[2] Nikitin, N., & Cortadella, J. (2009). A performance analytical model for Network-on-Chip with constant service time routers. In Proceedings of the 2009 International Conference on Computer-Aided Design (ICCAD '09). ACM, New York, NY, USA,

571-578.DOI=10.1145/1687399.1687506

http://doi.acm.org/10.1145/1687399.168750

[3] Kiasari, A. E., Lu, Z., & Jantsch, A. (2013). An analytical latency model for networks-on chip. IEEE Trans. Very Large Scale (VLSI) Syst., 21(1), 113-123..

[4] Chen, D., et al., "The IBM Blue Gene/Q interconnection network and message unit," 2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC), Seatle, WA, 2011, pp. 1-10.

[5] Liu, N., & Carothers, C. D., "Modeling Billion-Node Torus Networks Using Massively Parallel Discrete-Event Simulation," 2011 IEEE Workshop on Principles of Advanced and Distributed Simulation, Nice, 2011, pp. 1-8. doi: 10.1109/PADS.2011.5936761

[6] Minkenberg, C. (2013) Interconnection Network Architectures for High-Performance Computing. Advanced Computer Networks – Guest Lecture, 21 May 2013, IBM. © 2013 IBM Corporation

[7] Rykalova, Y., Levitin, L. B., & Brower, R. 2010. Critical phenomena in discrete-time interconnection networks. Physica A: Statistical Mechanics and its Applications, 389(22), 5259-5278

[8] Levitin, L. B., Rykalova, Y. (2015). Analysis and Simulation of Computer Networks with Unlimited Buffers. In Al-Sakib Pathan, Muhammad Monowar, and Shafiullah Khan (Eds.) Simulation Technologies in Networking and Communications: Selecting the Best Tool for the Test. CRC Press 2015, 3-30

[9] Levitin, L.B., & Rykalova,Y. Latency and phase transitions in interconnection networks with unlimited buffers. International Journal of Modern Engineering (IJME), Vol. 17, No.1, 2016, 70-77.

[10] Rykalova, Y., & Levitin, L. 2017. Phase Transitions in Interconnection Networks with Finite Buffers. International Journal of Modern Engineering (IJME), Vol. 17, No.2, 2017, 33-40.

[11] Kermani, P., Kleinrock, L. 1979. Virtual Cut-Through: A New Computer Communication Switching Technique. Computer Networks, Volume 3, Issue 2, (Sept. 1979), pp. 267-286.

[12] Duato, J., Robles, A., Silla, F., & Beivide, R. 2001. A Comparison of Router Architectures for Virtual Cut-Through and Wormhole Switching in a NOW Environment. J. Parallel Distrib. Comput. 61, 2 (February 2001), 224-253. DOI=http://dx.doi.org/10.1006/jpdc.2000.1679

[13] Amandeep Kaushal, Sarbdeep Singh, Network on Chip Architecture and Routing Techniques: A survey. International Journal of Research in Engineering and Science (IJRES) ISSN (Online): 2320-9364, ISSN (Print): 2320-9356 www.ijres.org Volume 2 Issue 6, June 2014, pp.65-79.

[14] Domkondwar, P., & Chaudhari, D. (2012). Implementation of Five Port Router Architecture Using VHDL. International Journal Of Advanced Research In Computer Science And Electronics Engineering (IJARCSEE), 1(3), pp:17-20.

[15] Wang, P., Ma, S., Lu, H., & Wang, Z. (2014). A comprehensive comparison between virtual cut-through and wormhole routers for cache coherent Network on-Chips. IEICE Electronics Express. 11. 20140496-20140496. 10.1587/elex.11.20140496.

[16] Choudhary, S., & Qureshi, S. Performance Evaluation of Mesh-based NoCs: Implementation of a New Architecture and Routing Algorithm[J]. International Journal of Automation and Computing , vol. 9, no. 4, pp. 403-413, 2012.

[17] Levitin, L. B., Karpovsky, M. G., & Mustafa, M. (2009). Minimal Sets of turns for Breaking Cycles in Graphs Modeling Networks, IEEE Trans. Parallel and Distributed Systems, v. 21, 9, 2010, 1342-1353.

[18] Karpovsky, M. G., Levitin, L. B., & Mustafa, M. (2014) Optimal Turn Prohibition for Deadlock Prevention in Networks with Regular Topologies, IEEE Trans. On Control of Network Systems, v. 1, 1, 2014, 74-85.

[19] Sadawarte, Y. A., Gaikwad, M. A., & Patrikar, R. M. Implementation of Virtual Cut-Through Algorithm for Network on Chip Architecture. IJCA Proceedings on International Symposium on Devices MEMS, Intelligent Systems & Communication (ISDMISC) (1):5-8, 2011.

[20] Rexford, J., & Shin, K.G. (1996). Analytical Modeling of Routing Algorithms in Virtual Cut-Through Networks. U Michigan.

[21] Hag, A. A. Y., Hafizur Rahman, M.M., Nor, R. M., Sembok, T. M. T., Miura, Y., & Inoguchi, Y. (2015) Uniform Traffic Patterns using Virtual Cut-Through Flow Control on VMMN, Procedia Computer Science, Volume 59, 2015, Pages 400-409, ISSN 1877-0509, http://dx.doi.org/10.1016/j.procs.2015.07.553.

[22] Kodgire, S., & Shiurkar, U. (2015) An Analytical Router Model for Networks-On-Chip. IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume 5, Issue 4, Ver. II (Jul - Aug. 2015), 16-21 e-ISSN: 2319 – 4200, p-ISSN No.: 2319 – 4197. www.iosrjournals.org

[23] L.B. Levitin, Y Rykalova. 2018. Analysis and simulation of networks with virtual cut-through routing. Proc. of the 2018 IAJC International Conference, Oct. 11-14, 2018, Orlando, FL. ISBN 978-1-60643-379-9