Study on Network-on-Chip Saturation

Ming-hua TANG¹ and Jing LIN²,*

¹School of Internet Finance and Information Engineering, GuangDong University of Finance, China
²School of Information Management, Shanghai Lixin University of Accounting and Finance, China

*Corresponding author

Keywords: Network-on-Chip, Routing algorithm, Saturation.

Abstract. In Network-on-Chip (NoC), although the term “saturation” is always used while discussing about performance and power figures exhibited by a NoC, there is not, however, an unanimous definition of the saturation status. There are at least six representative usages about saturation, which shows that there is lack of systematical study on saturation in the literature. In this paper, we carry out detailed study on saturation. The current usages about saturation are analyzed. The simulation results show that NoC saturation status is quite complex. Consequently, saturation should not be used without careful consideration.

Introduction

In many NoC research, network saturation status is widely considered. The concrete number of researches related with saturation is too large to be counted. However, there is not a unanimous definition about saturation status. To the best of our knowledge, at least six representative usages about saturation in the literature could be found. These usages are respectively termed as twice usage [1, 2], three times usage [3], infinity usage [4], nonlinear usage [5], maximum usage [6] and undefined usage [7] in this paper.

We analyze the most frequently used saturation usages and find that they are misleading in several circumstances. Specifically, we have detected several shortcomings in their usage as follows.

In the twice usage and three times usage of saturation, that is, average latency reaches twice or three times of the zero-load latency, the shortcomings include: Firstly, it is difficult to accurately determine which load leads to saturation; Secondly, the network buffer occupancy level is very low when saturation occurs; Thirdly, there is little performance difference for various routings at the saturation point.

The shortcoming of the nonlinear usage of saturation is that it is difficult to determine saturation PIR since the latency will slowly increases to infinity under some routings and traffics.

The shortcomings of the infinity usage and the maximum usage of saturation are that the average packet latency is on the order of a few thousands of cycles when the throughput no longer grows linearly with traffic load or the throughput reaches its maximum sustainable throughput. It is meaningless to compare the performance of routings at such large latency.

Considering the complex of the saturation status, it is inappropriate to use saturation without definition.

We choose the following representative routing algorithms to carry out our study through a cyclic accuracy simulator Noxim [8]. The typical deterministic routing is XY routing, negative-first (NF), north-last (NL), and west-first (WF), Odd-Even (OE) turn model, Segment-Based Routing (SR), and Pressure-Model based routing (PM). In addition to routing algorithms, selection strategies, such as random, buffer-level, NoP, CAIS are also considered.

The saturation status is a widely used concept in NoC researches. The following usages could be found in the literature.

1) Saturation is defined when the average latency is twice of the zero-load latency. This usage is termed as twice usage in this paper.
2) When average latency reaches three times the average no-load latency the network is considered to be saturated. This usage is termed as three times usage in this paper.

3) Network enters into saturation status when the latency increases to infinity. This usage is termed as in this paper.

4) Saturation is defined as throughput no longer grows linearly with traffic load. This usage is termed as in this paper.

5) Network is called saturated when network throughput reaches its maximum sustainable throughput. This usage is termed as maximum usage in this paper.

6) In most cases, saturation is used without specifically defined. Here is a short list of those paper. This usage is termed as undefined usage in this paper.

**Network Buffer Utilization**

In the presence of destination packets and deterring packets then router's buffer utilization will not be high. However, passing packets tend to make router's buffer space be used up. Within a network, all the three types of packets exist. The network's buffer utilization has the following characteristics.

(1) The network's buffer space will never use up no matter how serious the congestion is.

As PIR increases, the network buffer utilization will firstly rise up and then keep at a stable level under any combination of routing and traffic. For example, the variations of network buffer utilization under XY routing and uniform traffic are depicted in Figure 1.

![Figure 1. Variations of network buffer utilization.](image)

The stable network buffer utilizations under various routings and traffics are summarized in Table 1. The stable network buffer utilizations range from 36% to 75%.

Under uniform and butterfly traffics, the standard deviations of the network buffer utilizations across different routings are 0.029 and 0.018, respectively. Under hs-c and shuffle traffics, the standard deviations of network buffer utilizations are 0.046 and 0.037. Under transpose1, transpose2, hs-tr and bitreversal traffics, the standard deviations of the network buffer utilizations are range from 0.081 to 0.087. Consequently, the routing algorithm does not have a relevant impact on network buffer utilization under uniform and butterfly traffics. Conversely, its impact under the remaining traffic scenarios is relevant. Such difference can be explained as follows. In uniform and butterfly traffics, the network traffic is quite uniform. The adaptive routing algorithms have little chance to make the network traffic uniform. Consequently, the network buffer utilization variations across different routings are small. While in transpose1, transpose2, hs-tr and bitreversal traffics, the network traffic is not quite non-uniform. The adaptive routing algorithms have more chance to improve the network traffic uniformity. As a result, the network buffer utilization variations across different routings are large.
Table 1. The stable network buffer utilizations under various routings and traffics, with random selection strategy.

|                | XY  | NF  | NL  | WF  | OE  | PM  | SR  |
|----------------|-----|-----|-----|-----|-----|-----|-----|
| uniform        | 0.6 | 0.63| 0.65| 0.59| 0.65| 0.67| 0.65|
| transpose1     | 0.39| 0.39| 0.51| 0.51| 0.58| 0.58| 0.59|
| transpose2     | 0.39| 0.65| 0.51| 0.51| 0.58| 0.58| 0.57|
| hs-c           | 0.5 | 0.6 | 0.6 | 0.6 | 0.63| 0.63| 0.63|
| hs-tr          | 0.51| 0.75| 0.61| 0.69| 0.71| 0.61| 0.61|
| bitreversal    | 0.43| 0.63| 0.6 | 0.63| 0.67| 0.66| 0.64|
| butterfly      | 0.36| 0.42| 0.39| 0.39| 0.39| 0.38| 0.4 |
| shuffle        | 0.53| 0.62| 0.6 | 0.59| 0.65| 0.6 | 0.61|

Finally, when NoP selection strategy is used, the stable network buffer utilizations have not been significantly affected, as shown in Table 2.

Let us now investigate on the low network buffer utilization results obtained in the above analysis. For this purpose, let us consider XY routing and transpose1 traffic. Under XY routing and transpose1 traffic, all packets from data flows 0-63, 1-55, 2-47, 3-39, 4-31, 5-23, 6-15 pass through router 7. At a first glance, router 7 will be mostly congested. However, congestion never occurs at router 7. Every router's stable buffer utilization is shown in Table 3. Router 7's buffer utilization is only 8.3%.

Table 2. The stable network buffer utilizations under various routings and traffics, with NoP selection strategy.

|                | XY  | NF  | NL  | WF  | OE  | PM  | SR  |
|----------------|-----|-----|-----|-----|-----|-----|-----|
| uniform        | 0.6 | 0.62| 0.65| 0.59| 0.65| 0.67| 0.65|
| transpose1     | 0.39| 0.39| 0.49| 0.49| 0.58| 0.57| 0.58|
| transpose2     | 0.39| 0.6 | 0.49| 0.49| 0.58| 0.57| 0.57|
| hs-c           | 0.5 | 0.59| 0.59| 0.59| 0.62| 0.62| 0.62|
| hs-tr          | 0.51| 0.72| 0.6 | 0.69| 0.7 | 0.6 | 0.6 |
| bitreversal    | 0.43| 0.62| 0.58| 0.61| 0.67| 0.66| 0.64|
| butterfly      | 0.36| 0.37| 0.36| 0.36| 0.36| 0.37| 0.4 |
| shuffle        | 0.53| 0.59| 0.59| 0.56| 0.64| 0.58| 0.59|

All the packets that try to reach router 7 should have to contend the east output port of router 6 with packets come from node 6. The contention makes only one flit could reach router 7 in two cycles, and arrived flits could be timely forwarded by router 7. Consequently, there is no congestion at router 7, and serious congestion takes place from router 0 to router 6.

Although there is serious congestion in a part of routers (such as from router 0 to router 6), the other part of routers have free buffer slots (such as router 7). As as result, although no more packets could be injected into network, the network buffer space is not fully occupied.

Table 3. The stable router buffer utilizations under XY routing and transpose1 traffic.

|   | 0   | 1   | 2   | 3   | 4   | 5   | 6   | 7   |
|---|-----|-----|-----|-----|-----|-----|-----|-----|
| 0 | 0.33| 0.56| 0.56| 0.56| 0.56| 0.56| 0.5 | 0.083|
| 1 | 0.25| 0.45| 0.45| 0.45| 0.45| 0.45| 0.05| 0.062|
| 2 | 0.31| 0.45| 0.45| 0.45| 0.45| 0.1 | 0.4  | 0.25 |
| 3 | 0.31| 0.45| 0.45| 0.45| 0.1 | 0.4 | 0.45 | 0.31 |
| 4 | 0.31| 0.45| 0.4 | 0.1 | 0.4 | 0.45| 0.45 | 0.31 |
| 5 | 0.25| 0.35| 0.1 | 0.45| 0.45| 0.45| 0.45 | 0.31 |
| 6 | 0.062| 0.05| 0.4 | 0.45| 0.45| 0.45| 0.45 | 0.31 |
| 7 | 0.083| 0.5 | 0.56| 0.56| 0.56| 0.56| 0.56 | 0.33 |
The network's buffer utilization will slowly rise up and finally maintain at a stable level, the speed depending on the routing and the traffic scenario.

In order to show the rising speed variations, a deterministic routing XY and an adaptive routing OE are considered. Figure 2 and Figure 3 are the simulation results of XY and OE routing, respectively.

Under hs-c, hs-tr, uniform and bitreversal traffics, the network buffer utilizations rise up quickly, in both XY and OE routings. Under transpose1, transpose2, butterfly and shuffle traffics, the network buffer utilizations rise up quite slowly. Such difference can be explained as follows. Let us take uniform and transpose1 traffics as example. Under uniform traffic node (i, j) can send packets to any other node through its four output channels. Consequently, node (i, j) can still inject packets into network even when three of its output channels are busy. Whereas under transpose1 traffic, node (i, j) only sends packets to node (N-1-j, N-1-i). When its output channel(s) leading to the destination node is busy, node (i, j) will stop injecting packet into network, thus decreasing network buffer utilization. As a result, the network buffer utilization rises up more quickly under uniform traffic than under transpose1 traffic.

Analyzing Saturation Usages

In this section, we analyze the most frequently used saturation usages based on the previous observations. In twice usage and three times usage, saturation is defined as the average packet latency equals twice or three times of the zero-load average packet latency. These two usages have several shortcomings. To show these shortcomings, we conduct simulations with hs-c traffic and various routings. Figure 4 depicts the latency and throughput variations vs. PIR. As it can be observed, the zero-load average packet latency is 11 cycles. Twice of zero-load latency is 22 cycles, and three times of zero-load latency is 33 cycles.

In the nonlinear usage, network is said to enter into saturation status when the latency increases to infinity. This usage has difficulties to accurately determine when network enters into saturation. Figure 5 shows the latency variations under transpose1 traffic and the considered routing algorithms.
routings. The average packet latency increases gradually. Consequently, it is difficult to determine at which point the latency increases to infinity.

Figure 6 depicts the latency variations for eight traffics under OE routing. Under hs-c, hs-tr, uniform and bitreversal traffics, average packet latencies quickly rise up to infinity. The saturation point could be identified. However, it is not so for the other four traffics. The saturation PIR is hard to be determined in such traffic scenarios traffics.

Figure 5. Latency variations under transpose1 traffic and various routings.

In the infinity usage and maximum usage, network is said to have saturated when throughput no longer grows linearly with traffic load or network throughput reaches its maximum sustainable throughput, respectively. Figure 7 plots the throughput variations under bitreversal traffic and various routings. As shown in Figure 2 and Figure 3 the network buffer utilization will rise up with increment of PIR, and finally reaches a stable level. Similarly, the throughput will rise up, although the speed and the final level are different as routing varies. The throughput does reach a sustainable level after a quite long and slow increment. Unfortunately, the average packet latencies are larger than 4000 cycles when the maximum throughput is reached.

Figure 7. Throughput variations under bitreversal traffic and various routings.

The average packet latency values at the maximum sustainable throughput under various routings and traffics are shown in Table 4. When gets the stable status, the average packet latency is several thousand cycles. It is meaningless to compare network performance at such large packet latency.
Table 4. The average packet latency at the maximum sustainable throughput under various routings and traffics.

|         | XY  | NF  | NL  | WF  | OE  | PM  | SR  |
|---------|-----|-----|-----|-----|-----|-----|-----|
| uniform | 2287| 4212| 4792| 4119| 3445| 3481| 4643|
| transpose1 | 3748| 3874| 3912| 3899| 3792| 3786| 3883|
| transpose2 | 3866| 3966| 3895| 3896| 3936| 3790| 4404|
| hs-c     | 4433| 1001| 1329| 1853| 1712| 1390| 1574|
| hs-tr    | 3129| 2343| 2232| 2052| 2757| 2224| 2027|
| bitreversal | 4185| 4672| 4843| 4177| 4862| 4689| 4295|
| butterfly | 2186| 2104| 2680| 2023| 2005| 2408| 2175|
| shuffle  | 5060| 4695| 5158| 4862| 4892| 4628| 5085|

Consequently, from these analysis, assessing the actual network status is a complex task. The undefined usage of saturation without definition is inappropriate.

Conclusions

In this paper, based on the analysis on NoC router buffer occupancy level, we study the most frequently used saturation usages. Each usage has some shortcomings. It suggests that it should be cautious to use saturation concepts. Future work will be devoted on defining a more appropriate definition of saturation status that can be easily and effectively used in simulation-based analysis.

Acknowledgement

This work is supported in part by the Natural Science Foundation of Guangdong under Grants No. 2017A030313391.

References

[1] E. J. Chang, H. K. Hsin, C. H. Chao, S. Y. Lin and A. Y. Wu, Regional ACO-Based Cascaded Adaptive Routing for Traffic Balancing in Mesh-Based Network-on-Chip Systems, IEEE Transactions on Computers, 64 (2015) 868-875.

[2] H. K. Hsin, E. J. Chang and A. Y. Wu, Spatial-Temporal Enhancement of ACO-based Selection Schemes for Adaptive Routing in Network-on-Chip Systems, IEEE Transactions on Parallel and Distributed Systems, 25 (2014) 1626-1637.

[3] C. Iordanou, V. Soteriou, K. Aisopos, Hermes: Architecting a Top-Performing Fault-Tolerant Routing Algorithm for Networks-on-Chips, Computer Design (ICCD), 2014 32nd IEEE International Conference on, (2014) 424-431.

[4] M. Lis, M. H. Cho, K. S. Shim and S. Devadas, Path-Diverse Inorder Routing, Green Circuits and Systems (ICGCS), 2010 International Conference on, (2010) 311-316.

[5] C. F. Wang, W. H. Hua, S. E. Lee and N. Bagherzadeh, Area and power-efficient innovative congestion-aware Network-on-Chip architecture. Journal of Systems Architecture, 57 (2011) 24-38.

[6] S. Foroutan, Y. Thonnart and F. Petrot, An Iterative Computational Technique for Performance Evaluation of Networks-on-Chip, Computers, IEEE Transactions on, 62 (2013) 1641-1655.

[7] R. S. Ramanujam, B. Lin, Randomized Throughput-Optimal Oblivious Routing for Torus Networks, Computers, IEEE Transactions on, 62 (2014) 561-574, 2014.

[8] V. Catania, A. Mineo, S. Monteleone, M. Palesi and D. Patti, Noxim: An Open, Extensible and Cycle-accurate Network on Chip Simulator. IEEE International Conference on Application-specific Systems, Architectures and Processors 2015. (2015) 27-29.