Analysis Of Factors Affecting Thresholds Of Cache-Miss Packet Transmission In A NOC Based Many-Core Architecture

Maddula N V Sesha Saiteja, K Sai Sumanth Reddy, K Pavan Kumar, D Radha

Abstract: The need for the high computational power has led to the birth of the many-core architecture and its accompanying interconnect system Network on Chip (NOC). Communication between the routers of the cores is increased due to cache misses in the core. The congestion in the network has led to an increase in the latency and decrease in the throughput even with many cores for computation. The injection rate in the network is what determines the capacity of network load that it can withstand. The paper discusses the Threshold injection rate (TIR) which serves as a parameter to determine the upper bound of the network. And also, various parameters that affect the Threshold injection rate are discussed. The parameters considered are the number of virtual channels (VC), traffic pattern, size of the network, packet size, flit size etc.

Index Terms: Latency, Throughput, Threshold, injection rate, Network on Chip, Manycore architecture,Virtual Channels,Performance

1 INTRODUCTION

The technology of many-core architecture is becoming vital with the development of other technologies which includes large computational tasks and large data analysis like in machine learning, big data and so on. Thereby the need for the high computation power [1] has been rising so rapidly. The Network on Chip is a core concept of many core architecture as it serves as an interconnect mechanism. A good network or system is one that produces high throughput and less latency and several works have been carried out in achieving the same [2]. But the congestion in the traffic of the network has led to a decrease in the throughput with the effect of higher latency [1]. Every core or router is used in carrying out one or the other task and leading to longer latency for the new tasks. The core generates cache miss and sent data request packets to the core having the missed data. The request packets from source are injected into the network of cores to the destination. If the request and reply packets are injected more into the network, the congestion increases which in turn increase the latency of the packets reaching its destination. If the rate which the packets are sent is under the control, congestion can be avoided. Injection rate is the rate at which the packets are being issued or injected into the network. The study has shown that lower injection rates are good for lower latency but the throughput is less and on increasing the injection rate. The performance and finally it reaches the saturation state thereby no improvement in performance and thereafter the network may collapse due to congestion. This stage or limit is known as the saturation injection rate [3] or Threshold injection rate (TIR). The TIR serves as the upper bound of the network and determines how best it can perform in the worst-case scenario of a network similar to the Big O Notation for that of the algorithms.

There are various synthetic traffic patterns which models the destination for the sources. Traffic pattern determines the math and logic behind choosing a source and destination node for the communication. When a packet is to be sent from a core(source), traffic pattern determines where the destination core is. In this paper, we use some of the synthetic traffic patterns [3] for our analysis such as uniform, bitcomp, bitrev, transpose, tornado.

i) Uniform: For a given source core, randomly any core can be selected as all cores have equal probability calculated using uniform probability distribution function.
ii) Transpose: For a source core at x^{th} row and y^{th} column destination core would be the one at y^{th} row and x^{th} column.
iii) Bitrev: For a given source core id, the destination it would be reverse of the bits of that are present in the source id.
iv) Bitcomp: For a given source id, the destination id would be complement of the bits that are present in the source id
v) Tornado: For a given source core, destination core would be (k-1)/2 number of steps to the right and (k-1)/2 number of steps above source core (where k is the radix of the network i.e., for 4x4 network k=4, for 8x8 network k=8)[3].

These are some of the patterns that can occur in real time traffic in Network on Chip of many-core architecture. Apart from these, size of the network for any topology is also a key element for the efficient NoC communication system. It plays an important role in performance parameters such as area, power consumption, latency, throughput, and scalability and reusability of the NoC design. Thus it is very much important to know the threshold of the packets that can be injected into the network as it saves the system from being congested or from a lower system performance and also serves as a measure for estimating the capacity of network at an abstract level. The Threshold of injection rate is dependent on various factors which include:

i) Number of virtual channels (VC's)
ii) Traffic pattern
iii) Packet size
iv) Size of the network

2 LITERATURE SURVEY

As of today there are several experiments and studies carried out in the area of observing the performance of a many core system with respect to its injection rate which gives an approximate value of the load that the system can bear and

- *Maddula N V Sesha Saiteja,Sumanth and Pavan are B.Tech students and D.Radha is Assitant Professor(Email:d_radha@blr.amrita.edu)*
- *Department of Computer Science Engineering*
- *Amrita School of Engineering,Bangalore*
- *Amrita Vishwa Vidyapeetam India*

what are the factors that affects its value across various routing algorithms that include oblivious routing, adaptive routing so as to decrease the latency and improve the overall performance of the system[1] and other parameters like power etc. The main interest laid only in determining or comparing performance over the various algorithms. But the amount of work carried to improve the threshold injection or load of the network is very limited. It is observed that the capacity of the system with increase in the number of nodes/cores and determining the threshold injection rate so as to increase the load of the system and what are the necessary parameters that need to be considered for improving overall capacity of the system for better performance. The growth of Artificial Intelligence and Big Data where data has become fuel for the latest technologies has opened lot of scope and necessity for the Network on chip and greater efforts are being required to invest in the other fields [2] rather than one single way to approach for better performance. One must always remember that it helps designers in computing latency and throughput values as well as in identifying traffic congestion regions providing scope for betterment. Several external and internal mechanisms have been carried out and another dimension for measuring performance could be identification of hot spots and critical paths created by packets with higher latency values allows the designer to optimize the NoC [3]. Out of the many proposed ways to improve the performance of the NoC Systems by reducing the congestion, a method by multiplexing using virtual channels (VCs) can be used as they help in the decrease of latency and increase in network throughput. VCs provide multiple buffers for each channel and thereby increasing resources for each packet. The insertion of VCs also enables to implement policies for allocating the physical channel bandwidth, which enables to support quality of service (QoS) in applications. There are various methods proposed for selecting the location of VCs there by maximizing performance with little effort and cost [4]. In addition to improving performance with use of VC's it is suggested that instead of having a single network with the complex allocation logic necessary to support VC flow control on large channels, it is possible to use simpler flow control mechanisms. Partitioning the channel widths across multiple independent and parallel networks leading to multi-plane NoCs, can be designed to have smaller power dissipation and area occupation by leveraging the fact that they consist of many simpler networks operating independently [5]. Thus, it is evident from the above discussions, the improvement performance metrics like latency and throughput is mainly concentrated only from the few dimensions like use of better routing algorithms or introduction of VC's into network maximizing resources. The other related work includes determining efficient number or placement policies of Virtual Channels. But experimenting with only parameters has got limitation in improving performance so we should also look at other parameters which are stated above like size of network, traffic pattern, packet size etc.

3 PROPOSED EXPERIMENT

As stated earlier there are various other approaches or parameters need to look into in achieving the high performance of the system and also increase the capacity of system i.e. the amount of load and traffic it can handle during its operation and this experiment and analysis gives a room for optimizing other variables for better performance rather than the existing approaches for optimizing performance which could be coming up with best routing algorithm or coming up with the best organization of the nodes or the topology. So as in the process of this traditional approach this paper provides a good insight for other ways to tackle the problem of optimizing the performance in bringing down the latency (which is time take to arrive at the required results should be relative low for better performance) and thereby achieving the results at a faster rate and this paper also shows the dependence and independence of the parameters and also motivate and decide the parameters of architecture in efficient manner at a low cost. In the following sections, we deeply explore each of the above factors which include the number of virtual channels, size of the network/architecture (number of cores), traffic pattern, packet size along with that of experimental results and also discuss their contribution towards the increase or to determine their role in the contribution for threshold injection rate. For the analysis, the experiments are carried out on Booksim2.0 Simulator over 4x4 and 8x8 mesh architectures across various packet sizes ranging from 1 to 6 and also across different synthetic traffic to patterns namely uniform, transpose, bitrev, tornado, bitcomp etc. over various packet sizes(1,2,3,4,5,6) and the necessary parameters are chosen and then each setup for the architecture with all the necessary parameters like traffic pattern,number of channels is fixed and then it is granted certain CPU cycles to produce the result and if it could not produce the results in the given amount of time then that threshold rate is taken to be the thresold injection rate of the architecture and the threshold injection rate is incremented slowly little at a little so as to determine or reach to the threshold injection rate of the system.This paper also lists out the results obtained by plotting the threshold rate vs the change in the parameters which could be the change in the traffic pattern or the change in the number of virtual channels or change in the number of cores i.e 8x8 network or 4x4 mesh network.The analysis and the inference of each parameter that is chosen for carrying out the experiment is explained in detail in the further sections.

4 EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Effect of virtual channels

Fig. 1 shows the range of the threshold injection rates across various packet sizes and that of various traffic patterns and their effect with the increase of the number of virtual channels. From the above figure, it is evident that there is a significant progress in the threshold injection rate (TIR) of the network. It clearly shows the nature of increase is independent of the traffic pattern but the value depends on the type of traffic**.** The graph is shown for normal traffic. The increase is too high for the lower traffic or fewer packet sizes but it decreases gradually for higher ranges. It finally converges to a flat injection rate for any number of virtual channels beyond which the system hangs up. These thresholds can be used for various analyses about many-core transmission. It can be a check to avoid congestion and for delay injection of packets in the real time scenario. The relation between Threshold Injection rate and packet size is as follows. Threshold Injection Rate α 1/ Number of Virtual Channels. Not only the packet size with that of VCS effect [4][5] is noticed but also one must notice the effect of the threshold injection rate with that of traffic pattern along with VCS. One must note that the behavior of traffic pattern on TIR is independent of the packet size as they show a similar nature or shape of the curve is the same

across all the packet sizes in the same order. It also shows that the regular and synthetic traffic patterns [6] follow the same trend of increasing TIR with the increase in the number of virtual channels. This increase in the value of TIR is obeyed across all the traffic patterns and its relation is studied in further sections below. A good performance and better interconnect mechanism or network is achieved only when it can deliver higher throughput and able to handle large loads at a time. The virtual channels are very useful in obtaining low latency as they provide additional paths along with that of true or hardwired physical channels [4][5]. They provide additional buffers which increase resource allocation for each packet.[6]

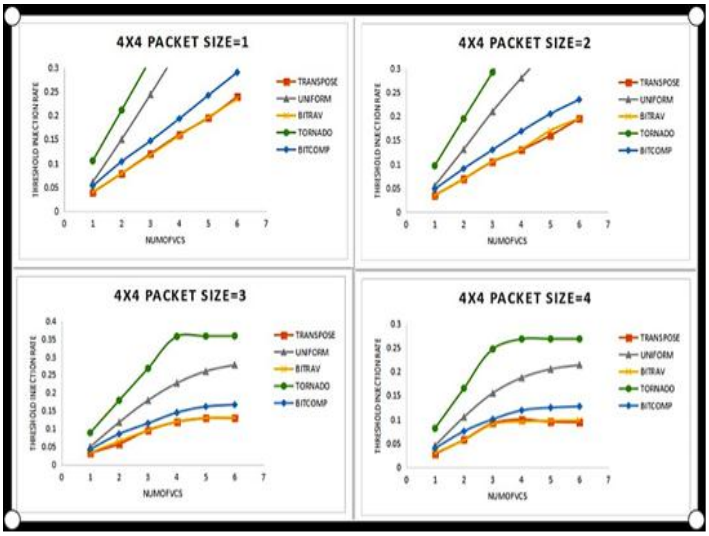


Figure 1. *Threshold of injection rates for 4x4 mesh topology*

Several studies and experiments have revealed the fact that increase in the virtual channels have always led to better performance and this increase is significant to make out. There is a gradual increase in the performance with that of the injection rate and remains unaltered after the threshold point known as the threshold injection rate. The increase in the threshold injection rate also follows a similar trend which can be drawn from the experimental results. Thus, it is shown that the number of VCs plays not only an important role in increasing the performance of the network [4][5] and also it helps to increase the value of the Threshold injection rate which helps in determining the maximum stable load that network can handle at a time. The number of VCs is the main factor for improving the TIR over the other factors and in the further sections that follow we explore other factors in detail.

4.2 Effect of Traffic Pattern

The figure (refer Fig.2) clearly shows the values of the threshold injection rate across various packet sizes and that of various traffic patterns and their effect with the increase of the number of virtual channels. Below figure shows the results obtained for different traffic pattern.

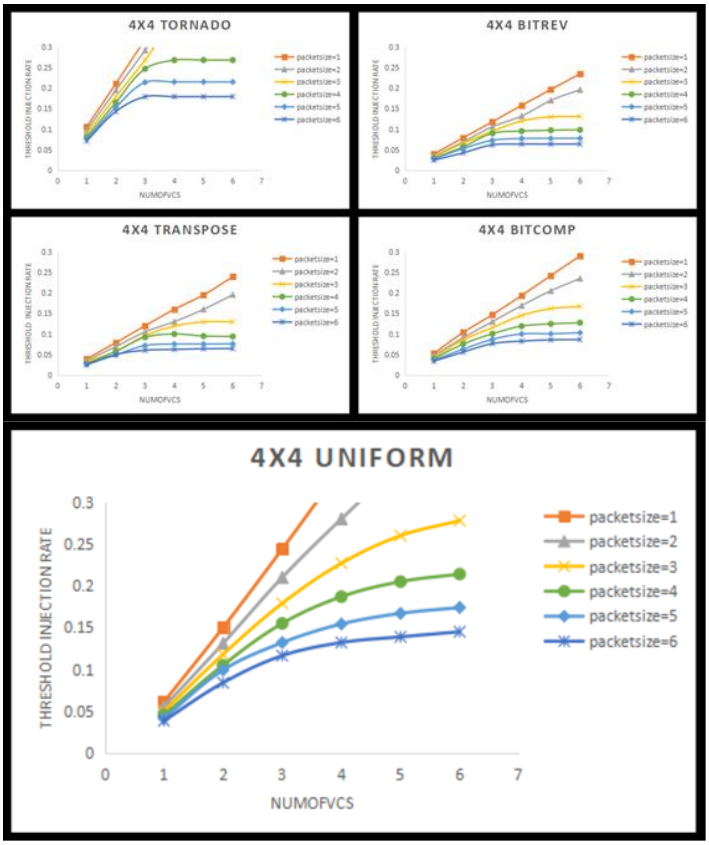
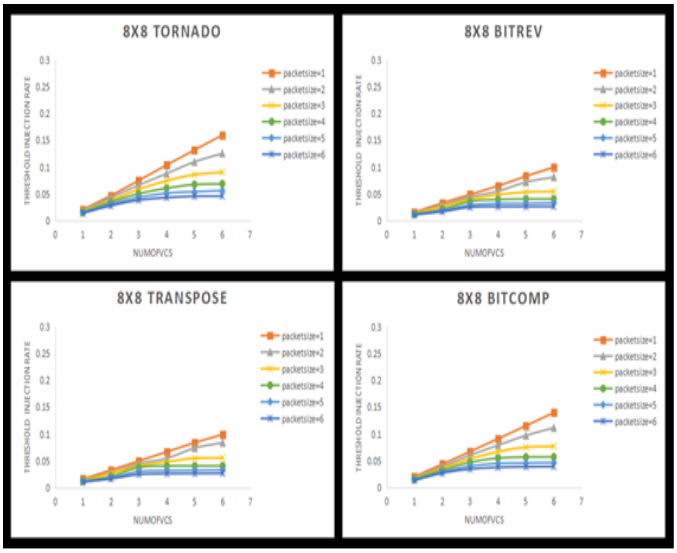


Figure 2. *Threshold of injection rates for different traffic patterns in a 4x4 mesh network*

Threshold injection rate (TIR) for different synthetic traffic patterns can be analyzed from these results. It is clear that for all traffic patterns as the number of VCs increases injection rate increases and finally reaches a threshold value. It is found that uniform and tornado traffic pattern have the highest TIR while others have low threshold value. For any traffic pattern threshold injection rate is higher for smaller packet size whereas it is lower for greater packet size i.e., as the packet size increases threshold injection rate decreases. So, this factor slightly affects the threshold injection rate [4]. In the following sections, we discuss the other parameters affecting the threshold injection rate.

4.3 Effect of packet size

As we are dealing with packet switched network it is important to consider the size of the packet. Packet size plays a vital role in effecting the TIR [7][8]. Usually, TIR decreases by increasing packet size. The decrease in the packet size also follows a trend which can be drawn from the below experimental results.



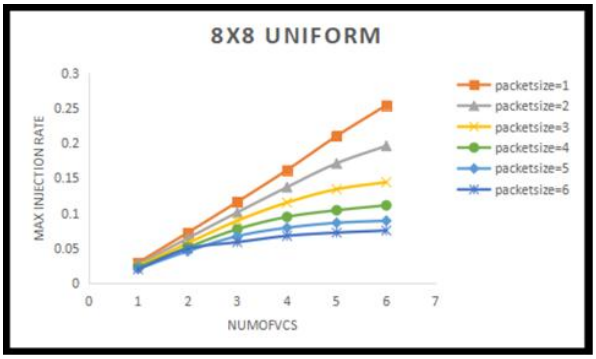


Figure 3. *Threshold of injection rates for 8x8 mesh topology*

Fig.3 shows the values of the threshold injection rate across various packet sizes and that of various traffic patterns and their effect with the increase in the number of virtual channels. From the Fig.3, it is clear that there is a significant decrease in the threshold injection rate (TIR) of the network as the packet size increases [7]. This trend is the same for all the traffic patterns. As number of VCs increases, there will be more difference in the injection rate, but if the packet size exceeds 3 there will be not much difference in injection rate as packet size increases because of increase in the congestion.

4.4 Effect of size of the network

Fig.4 shows the values of the threshold injection rate across various packet sizes and that of various traffic patterns and their effect with the increase in the number of virtual channels in different sizes of networks.

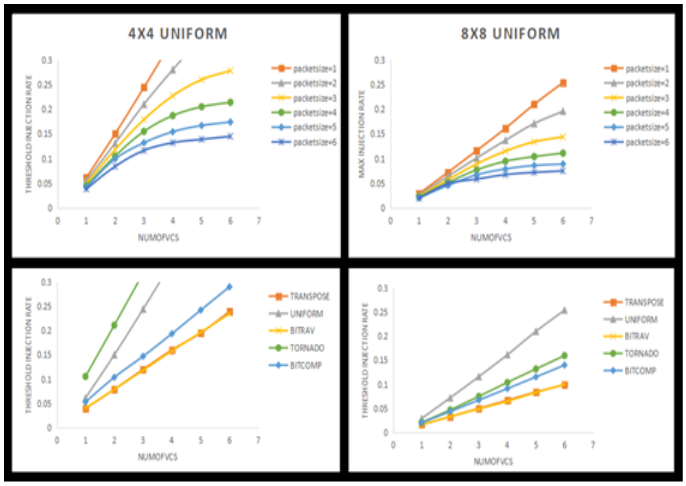


Figure 4. *Threshold of injection rates for 4x4 and 8x8 mesh topology*

From the Fig. 4 it is clear that there is a significant decrease in the threshold injection rate (TIR) of the network as the size of the network topology changes. There is a significant change that is almost half of the previous threshold values. This trend continues irrespective of network pattern [9]. But there is an exceptional decrement of tornado pattern is observed for the packet size=1.

5. Conclusion

In this paper, we discussed the factors effecting Threshold Injection Rate (TIR), which helps in reducing latency and increasing throughput of many-core architecture. From the above analysis for higher congestion to be handled, we need to have a higher number of virtual channels. Also, it is found that uniform traffic pattern can handle high workload irrespective of the size of the network [9]. Packet size is inversely proportional to the threshold injection rate which means that increasing the size of packet decreases congestion. And finally, it is observed that as the size of network doubles, TIR decreases by 50% approximately which indicates that congestion can be reduced by increasing network size. In this way Threshold Injection Rate can be helpful to increase throughput with low power consumption[10] and reduce latency in many-core architecture and this also helps to combine with other models and principles to improve injection rate like throttling[11] which is also known as delay injection as implementing it every time may reduce performance. Instead of calculating the current congestion status of the network every time in other aspects, maximum injection rate of every core can be restricted with the threshold injection rate for that scenario. This parameter helps to determine when to use throttling instead of always using it so that it can improve the performance of network in terms of latency, throughput and complexity of algorithms for congestion detection etc.

References

[1]. Haghi, Mostafa & Rani, Asha. (2014). Evaluation of Effect of Packet Injection Rate and Routing Algorithm on Network-on-Chip Performance. International Journal of Innovative Research in Science, Engineering and Technology. 3. 9589-9597.

[2]. Maddula N V Sesha Sai Teja, K Sai Sumanth Reddy, D Radha, Minal Moharir, “Multicore Architecture and

Network on Chip: Applications and Challenges", presented at ICIC 2018, Amrita School of Engineering, Bengaluru. India.

[3]. Tedesco, Leonel & Mello, Aline & Garibotti, Diego & Calazans, Ney & Moraes, Fernando. (2005). Traffic Generation and Performance Evaluation for Mesh-based NoCs. 184-189. 10.1109/SBCCI.2005.4286854.

[4]. Mello, Aline & Tedesco, Leonel & Calazans, Ney & Moraes, Fernando. (2005). Virtual channels in networks on chip: implementation and evaluation on hermes NoC.. 178-183. 10.1145/1081081.1081128.

[5]. Jin Yoon, Young & Concer, Nicola & Petracca, Michele & P. Carloni, Luca. (2013). Virtual Channels and Multiple Physical Networks: Two Alternatives to Improve NoC Performance. Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on. 32. 1906-1919. 10.1109/TCAD.2013.2276399.

[6]. Vijaya Bhaskar, Adusumilli & Gopalakrishnan Venkatesh, Tiruchirai. (2015). A study of the effect of virtual channels on the performance of Network-on-Chip. 255-260. 10.1109/SCORED.2015.7449335.

[7]. Haghi, Mostafa & Asha Rani, M & Javadi, Elham. (2014). EVALUATION OF EFFECT OF BUFFER SIZE ON NOC PERFORMANCE. International Journal of Electronics, Communication & Instrumentation Engineering Research and Development (IJECIERD). 4. 143-150.

[8]. Bindammas, Ahmed & Soudani, Adel & Al-Dhelaan, Abdullah. (2015). The efficiency of buffer and buffer-less data-flow control schemes for congestion avoidance in Networks on Chip. Journal of King Saud University - Computer and Information Sciences. 28. 10.1016/j.jksuci.2015.11.002.

[9]. Nishin Jude C Abraham, D Radha, "Detection and Analysis of Congestion of Nodes in Many-core Processor",presented at First International Conference on Sustainable Technologies for Computational Intelligence (ICTSCI-2019), Amity University Rajasthan, Jaipur, India.

[10]. M. Vinodhini, N.S Murty, "Reliable Low Power NoC Interconnect", Microprocessors and Microsystems, vol. 57, pp.15-22, 2018.

[11].]N S, Aswathy & Reshma Raj, R.S. & Jose, John & Josna, V.R.. (2017). Implementation and Analysis of Adaptive Packet Throttling in Mesh NoCs. Procedia Computer Science. 115. 626-634. 10.1016/j.procs.2017.09.149.