Int. J. Communications, Network and System Sciences, **2011, 4, 351-355**
doi:10.4236/ijcns.2011.45040 Published Online May 2011 (http://www.SciRP.org/journal/ijcns)

New Approach of QoS Metric Modeling on Network on Chip

Salem Nasri[1,2]
[1]*CES Laboratory*, *ENIS*, *Tunisia*
[2]*College of Computer*, *Qassim University*, *Qassim, Kingdom of Saudi Arabia*
E-mail: *snasri@qu.edu.sa*

Received February 28, 2011; *revised March* 21, 2011; *accepted April* 12, 2011

Abstract

This paper presents a new NoC QoS metrics modeling shaped on mesh architecture. The new QoS model is based on the QoS parameters. The goal of this work is to quantify buffering requirements and packet switching techniques in the NoC nodes by analyzing some QoS metrics such as End-to-End delays (EEDs) and packet loss. This study is based on simulation approach of a 4 × 4 mesh NoC behavior under multimedia communication process. It proposes a study of NoC switching buffer size avoiding packet drop and minimizing EED. Mainly, we focus on percent flit losses due to buffer congestion for a network loading. This leads to identify the optimal buffer size for the switch design. The routing approach is based on the Wormhole Routing method.

Keywords: NoC, QoS, End-to-End Delay, Packet Loss, Throughput

1. Introduction

According to ITRS, in 2018, ICs will be able to integrate billions of transistors, with feature sizes around 18 *nm* and clock frequencies near to 10 GHz [1]. In this context, a network on chip (NoC) appears as an attractive solution to implement future high performance networks and more suitable QoS managements. A NoC is composed by IP cores and switches connected among them by communication channels [2]. End-to-end communication system is accomplished by the exchange of data among IP cores. Often, the structure of particular messages is not adequate for the communication purposes. This leads to the concept of packet switching. In the context of NoCs, Packets are composed by header, payload, and trailer. Packets are divided into small pieces called Flits [3,4]. It appears of importance, to meet the required performance in NoC hardware resources. It should be specified in an earlier step of the system design the main attention should be given to the choice of the physical buffer size in the node. The EED and packet loss are some of the critical QoS metrics. Some real-time and multimedia applications bound up these parameters and require specific hardware resources and particular management approaches in the NoC switch. The best case is to provide the shortest constant EED or at least with the minimum fluctuation [5,6].

The paper is organized as follows. Section 2 presents the network on chip internal architecture and network routing packets. Section 3 introduces the notion of QoS metric modeling based on the QoS parameters. Simulation results for the NoC architecture target are presented and discussed to bring out some physical requirements enabling QoS metric evaluation based on the QoS parameters for one class of application in the Section 4. We finish by the conclusions and perspectives.

2. NoC Architecture and Packet Routing

NoC topologies are defined by the switches connection structure. The studied NoC architecture assumes each switch has a set of bi-directional ports linked to its neighbor switches and to an IP core. It is built on 4 × 4 mesh topology as shown in **Figure 1**.

Each switch has routing control unit and five bi-directional ports: East, West, North, South, and Local. Each port has an input buffer for temporary information storage. The local port establishes a communication between the switch and its IP core. The other ports are connected to the neighbor switches. The routing control unit implements the logic arbitration and packet-switching algorithm. The main critical parameters driving the switch

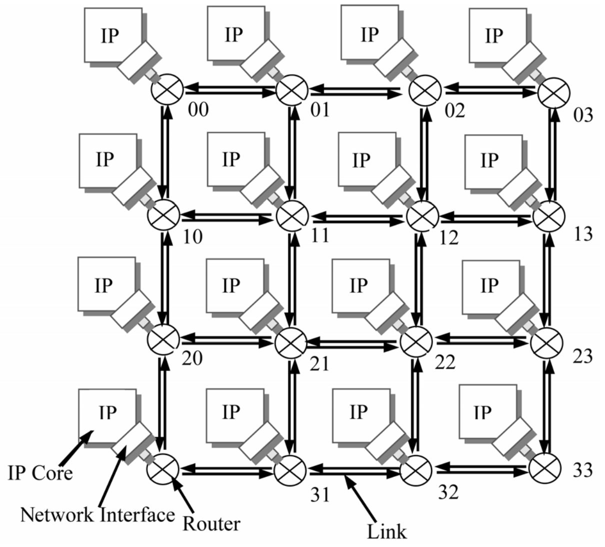

Figure 1. 4 × 4 mesh NoC structure.

performances are the memory access time (reading and writing) and the transition time through the switch (from input to output). In fact, it is important to minimize data bufferisation time because it reduces flits throughput, increases EED, causes jitter, and can lead to data loss if there is insufficient memory space to store all incoming data flows waiting to be transmitted. Theses communication parameters must be considered together with hardware system constraints related to circuit area and computing frequency optimization [7].

The main part of the switch is the flit scheduler. It is based on the Deficit Weighted Round Robin (DWRR) technique for the management of the data queuing. In this technique the switch defines many application classes and it associates a weight to each class. The switch bandwidth is then divided to input traffic classes according to their bandwidth requirement.

The scheduling approach managing the output switch buffer based on DWRR defines mainly two parameters:

- *The Counter* which specifies the total number of bytes that the queue is permitted to transmit at each time it is visited by the scheduler.
- A *quantum* of service proportional to the *weight* of the queue, it is expressed in bytes.

The *Counter* for a queue is incremented by the quantum value each time the queue is visited by the scheduler. In the DWRR the scheduler algorithm starts by determining the number of bytes at the head of the queue.

$$Counter = counter + quantum$$

Data in the queue is sent only if the size of the packet at the head of the queue is less than or equal to the variable *Counter*. The variable *Counter* is reduced by the number of bytes being sent and data is transmitted on the output port. The scheduler continues to send data from this queue until data in the queue is less than the value of *Counter* or the queue is empty. In this case the variable *Counter* will be set to zero. Then the scheduler moves on, to serve the next non-empty queue [8].

3. QoS Requirements in NoC Design

3.1. QoS Tentative Definition

The Quality of Service (QoS) refers to a broad collection of networking technologies and parameters. The goal of QoS is to provide guarantees on the ability of a network to deliver predictable performances. Elements of network performance within the scope of QoS often include availability (uptime), bandwidth (throughput), latency (delay), and error rate. QoS involves, also, prioritization of network traffic classes. It can be targeted at a network interface, toward a given server or router's performance. In terms of specific applications a network monitoring system must typically be deployed as part of QoS, to insure that networks are performing at the desired service level [9]. In packet-switched networks it refers to the probability of the network meeting a given traffic contract [10].

3.2. QoS Parameters

QoS is especially important for the new generation of Internet applications such as VoIP, video-on-demand and other consumer services. Some core networking technologies like Ethernet were not designed to support prioritized traffic or guaranteed performance levels, making much more difficult the QoS implementation solutions. In communication networks, such as Ethernet, throughput is the average rate of successful packets delivery over a communication channel. People are often concerned about measuring the maximum data throughput rate of a communications link or network access. A typical simple method of performing a measurement is to transfer a file F and measure the time T taken to do so.

EED concerns the time for a packet to reach its destination, because it gets held up in long queues, or takes a more indirect route to avoid congestion. Alternatively, it might follow a fast direct route. The delay is very unpredictable. Also the amount of time it takes a packet to move across a network connection defines the Latency. Latency and bandwidth are the two factors that determine a network connection speed. Latency and throughput are two fundamental measures of network performance. Moreover, sometimes the routers might fail to deliver (*drop*) some packets (packet loss) if they arrive when their buffers are already full. Some, none, or all of the packets might be dropped, depending on the state of the network, and it is impossible to determine what happened in advance. The receiving application must ask for this information to be retransmitted, possibly causing severe delays in the overall transmission [11].

3.3. QoS Modeling and Measurements

A traffic contract (SLA, Service Level Agreement) specifies the ability of a network or protocol to give guaranteed performance, throughput or latency bounds based on mutually agreed measures, usually by prioritizing traffic. A defined Quality of Service may be required for some types of network real time traffic or multimedia application [12-14]. We propose an approach of QoS-metric based on QoS-parameter prioritization factors α_i for one application-service using the relation:

$$Q(p_1, p_2, p_3, \cdots, p_m) = F(a_i, p_i), \quad i = 1, \cdots, m \tag{1}$$

We define k, α_i, p_i, and $Q(p_1, p_2, p_3, \cdots, p_m)$ such as:

1) $k \geq 1$: network efficiency coefficient (in our case we chose $k = 1.1$ for example).

2) α_i: parameter prioritization factor, with:

$$\sum_{i=1}^{m}(\alpha_i) = 1 \tag{2}$$

3) p_i: QoS performance parameter, p_i should be normalized p_{in}

$$p_{i\max} = \max\{p_i\},\ p_{i\min} = \min\{p_i\},$$

a) For increasing parameters when data rate increases

$$p_{in} = \left|\frac{p_i - p_{i\min}}{k * p_{i\max} - p_{i\min}}\right| \tag{3}$$

b) For decreasing parameters when data rate increases

$$p_{in} = \left|\frac{p_{i\max} - p_i}{k * p_{i\max} - p_{i\min}}\right| \tag{4}$$

4) Then the QoS expression can be defined by:

$$Q(p_i, p_2, \cdots, p_m) = \sum_{i=1}^{m}(\alpha_i p_{in}) \tag{5}$$

In this model we consider the packet loss parameter as p_1 and the EED parameter as p_2 for FIFO and DWRR scheduling techniques for 64 and 128 bytes of buffer size, α_1, α_2 are arbitrarily fixed referring to the Equation (2).

4. QoS Behavior Simulation of Target Architecture

We use for this study the available network simulator NS. This tool is becoming one of the most popular platforms for performances analysis in the network research community. Traffic is transferred over the network between two IPs connected respectively to the 00 switch (source) and the 33 switch destination. The packet size is 4 bytes (32 bits) based on 8 bits/flits (4 flits per packet). The maximum bandwidth link is fixed to 2GB/s. The purpose of the study is to give a QoS measurements approach according to the general network loading states and also according to the interconnected IPs throughput [15].

4.1. Packets Loss and Buffer Size

As being discussed, a reasonable buffer size may drive the NoC high performances. The main idea is to keep the minimum buffer size avoiding dropped packets and EED. We focus on percent flit losses due to buffer congestion for a network loading. The routing approach is based on the Wormhole Routing method. It reduces the store-and-forward delay at each switch, and requires less buffer size. **Figures 2** and **3** show the relationship between percent dropped packet and available switch buffer size. These figures show that the percentage of dropped packets increases with application rate and decreases with the Buffer Size increase. We compare the behavior of some parameters such as delay, dropped packet and buffer size for the same architecture based on the DWRR and FIFO scheduling. These figures testify the capability of the DWRR to provide best results. In

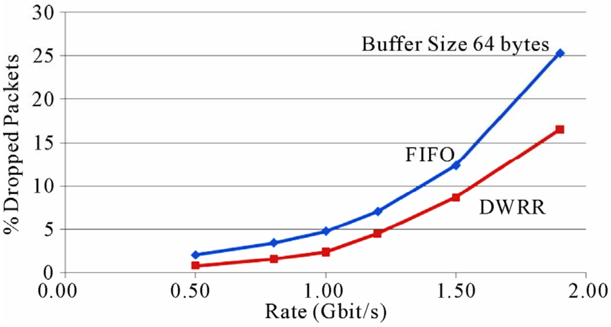

Figure 2. Dropped packets for 64 bytes buffer size.

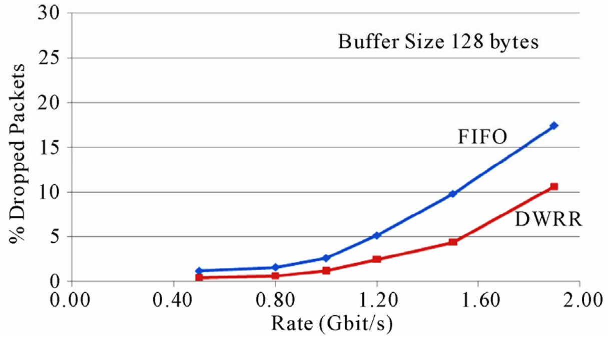

Figure 3. Dropped packets for 128 bytes buffer size.

fact the percentage of dropped packets is significantly less with DWRR compared to FIFO scheduling.

4.2. End to End Delay and Buffer Size

The EED is one of the most critical QoS metrics. Some real-time applications bound up this value and require specific hardware resources and particular management approaches in the NoC switch. The best case is to provide the shortest constant EED or at least with the minimum fluctuation. This can avoid synchronization between communication processes. **Figures 4** and **5** sum up the EED when the switching buffer is managed with DWRR and FIFO scheduling approach. The scheduler should improve service quality according to the flit requirement, using priority queuing technique.

Figures 4 and **5** show that the EED average when DWRR and FIFO scheduling technique are applied. It decreases significantly with a buffer size value.

4.3. QoS Measurements

Referring to the proposed model the following figures give the QoS measurements for 2 parameters: p_1: packet loss and p_2: EED, with prioritization factors: $\alpha_1 = \alpha_2 = 0.5$ and $\alpha_1 = 0.2$, $\alpha_2 = 0.8$.

Figures 6, **7** and **8** show the % QoS in relation with the Buffer Size, the rate, the scheduling techniques and prioritization factors. It appears that the % QoS increases with the rate. The prioritization factors have also an impact on the QoS values.

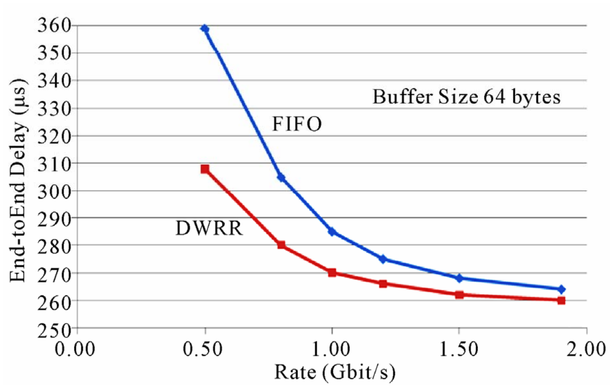

Figure 4. End to end delay according to the application rate with DWRR and FIFO scheduling (64 bytes buffer size).

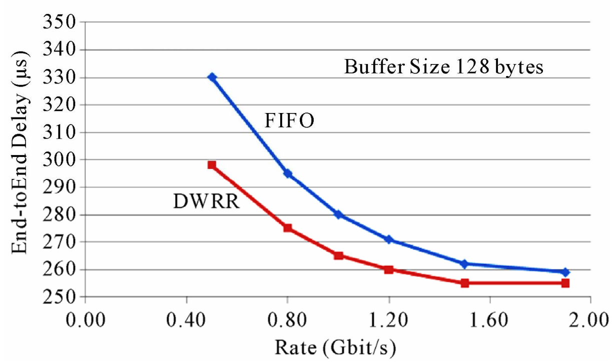

Figure 5. End-to-end delay according to the application rate with DWRR and FIFO scheduling. (128 bytes buffer size).

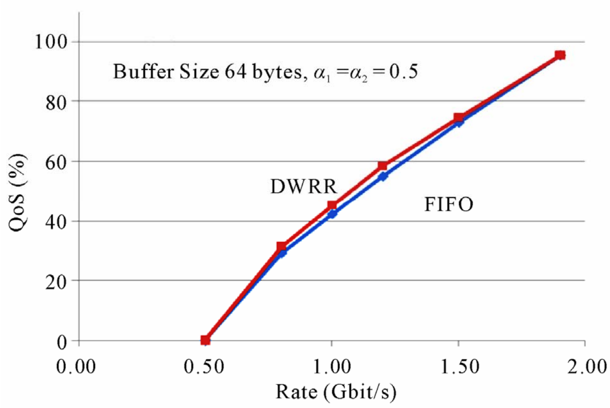

Figure 6. %QoS for 64bytes buffer size with prioritization factors $\alpha_1 = \alpha_2 = 0.5$.

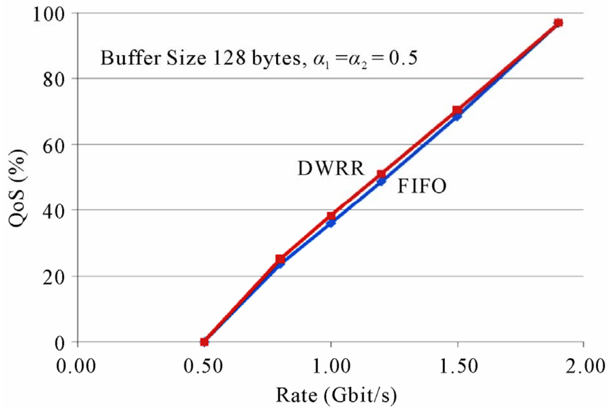

Figure 7. %QoS for 128 bytes buffer size with $\alpha_1 = \alpha_2 = 0.5$.

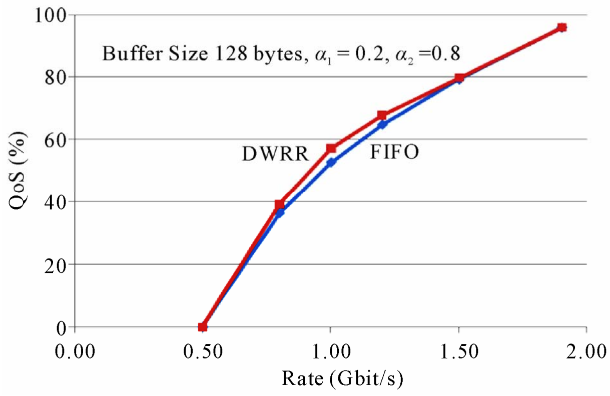

Figure 8. %QoS for 128 buffer size with $\alpha_1 = 0.2$, $\alpha_2 = 0.8$.

5. Conclusions and Perspectives

In this paper we have proposed a new QoS metric model for Network on Chip based on the QoS paremeters for one class of application. This model is a new approach of QoS metric leading to quantify and measure a QoS value in a network. We have focused our study on the switch buffering requirements. In fact, we have showed that the adequate buffer size in the switch drives to a better QoS values. During NoC communication processes, QoS metric can be affected by the switch buffering capacity and its management approach. We have shown that the DWRR is the best approach to manage packets scheduling in the NoC switch (**Figures 2**, **3**, **4** and **5**). We think that the QoS metric evaluation in the switch, and then in the network, can be ameliored by the adaptation of an appropriate approach for packets queuing in the switch buffer such as priority queuing (**Figures 6**, **7** and **8**).

For this purpose, we are now working on the specification of a new QoS model taking in consideration multiple applications with multiple QoS classes modeling. The idea is to meet both network required performances through the QoS metrics requirement, quantification and measurments.

6. References

[1] International Sematech, "International Technology Roadmap for Semiconductors," 2006. http://public.itrs.net

[2] T. Bjerregaard and S. Mahadevan, "A Survey 0f Research and Practice of Network on Chip," *ACM Computing Survey*, Vol. 38, March 2006, pp. 1-51. doi:10.1145/1132952.1132953

[3] J. Kim, D. Park, Ch. Nicopoulos, N. Vijaykrishnan and R. Chita. Das, "Design and Analysis of an NoC Architecture from Performance, Reliability and Energy Perspective," *ACM symposium on Architecture for networking and communications systems*, Princeton, NJ, USA, 2005, pp. 173-182, ISBN: 1-59593-082-5.

[4] S. Murali, M. Coenen, A. Radulescu, K. Goossens and G. De Micheli, "A Methodology for Mapping Multiple Use-Cases onto Networks on Chips," *Conference on design, automation and test in Europe,* Munich, Germany, 2006, pp. 118-123, ISBN:3-9810801-0-6.

[5] E. Rijpkema, K. Goossens and A. Radulescu, "Tradeoffs in the Design of a Router With Both Guaranteed And Best-Effort Services for Networks on Chip," *Conference on design, automation and test in Europe* (*DATE*'03), March 2003, pp. 350-355.

[6] C. Grecu, A. Ivanov, P. Pande, A. Jantsch, E. Salminen, U. Ogras and R. Marculescu, "Towards Open Network-on-Chip Benchmarks," *First International Symposium on Networks-on-Chip* (*NOCS'*07), IEEE Computer Society, May 2007, pp. 205-213.

[7] H. Gyu Lee, N. Chang, U. Y. Ogras and R. Marculescu, "On-Chip Communication Architecture Exploration: A Quantitative Evaluation of Point to Point, Bus, and Network-on-Chip Approaches," *ACM transaction on Design Automation of Electronic Systems*, Vol. 12, No. 3, August 2007, pp. 1-20.

[8] A. Helali and S. Nasri, "Network on Chip Switch Scheduling Approach for QoS and Hardware Resources Adaptation," *International Journal of Computer Sciences and Engineering Systems (IJCSES)*, Vol. 3, No. 1, 2009, pp. 29-35.

[9] T. Samak, E. Al-Shaer and H. Li, "QoS Policy Modeling and Conflict Analysis," *IEEE Workshop on policy for distributed systems and networks*, 2008, pp. 19-26.

[10] R. P. Liu, G. J. Sutton and I. B. Collings, "A New Queuing Model for QoS Analysis of IEEE 802.11 DCF with Finite Buffer and Load," *IEEE Transaction on Wireless Communications*, Vol. 9, No. 8, 2010, pp. 2664- 2675. doi:10.1109/TWC.2010.061010.091803

[11] V. X. Tran and H. Tsuji, "A Survey and Analysis on Semantics in QoS for Web Services," *International Conference on advanced information networking and applications*, October 2009, pp. 1-19.

[12] Y. Liu and H. He, "Grid Service Selection Using QoS Model," *Third international conference on semantics, knowledge and grid*, 2007, pp. 576-577.

[13] J. Luo, L. Jiang, and C. He, "Finite Queuing Model Analysis for Energy and QoS Tradeoff in Contention-Based Wireless Sensor Networks," *International conference on communication (ICC)*, IEEE Communications Society, 2007, pp. 1-6.

[14] K. Kunavut and T. Sanguankotchakorn, "Multi-Constrained Path (MCP) QoS Routing in OLSR Based on Multiple Additive QoS Metrics," *International symposium on communications and information technologies (ISCIT-IEEE)*, 2010, pp. 226-231.

[15] Z. Zhou, W. Xu, D. T. Pham, C. Ji, "QoS Modeling and Analysis for Manufacturing Networks: A Service Framework," 7*th IEEE International conference on industrial informatics* (*INDIN* 2009), 2009, pp. 825-230.