IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)
Volume 5, Issue 4, Ver. II (Jul - Aug. 2015), PP 16-21
e-ISSN: 2319 – 4200, p-ISSN No. : 2319 – 4197
www.iosrjournals.org

An Analytical Router Model for Networks-On-Chip

Mrs. Shilpa Kodgire[1], Dr. Ulhas Shiurkar[2]
[1](Electronics and communication Engg. Department, MIT Aurangabad, India)
[2](Electronics and Telecommunication Engineering Department, DIEMS, Aurangabad, India

Abstract : *The proposed system is based on analytical model of Router and queueing theory. This approach not only provides aggregate performance metrics such as average latency and throughput, but also feedback about the network characteristics such as buffer utilization, average latency per router and per flow at a fine-level of granularity. The proposed analytical model can predict the average packet latency faster than an accurate simulation*

Keywords: *Network-on-Chip (NoC), Arrival Rate(λ), Latency*

I. Introduction

The design of router involves determining the flow control technique, number of virtual channels, buffer organization, switch design, pipelining strategy while adhering to target clock frequency and power budgets. All these issues require careful design since they have significant impact in terms of performance, power consumption and area. The main focus in designing a router is to minimize the latency through it, while meeting bandwidth requirement.

NoC communication architectures offer a scalable and modular solution for implementing complex systems which consist of a large number of heterogeneous components. The complexities of such systems, as well as the tight requirements in terms of power, performance, cost and time to market place a tremendous pressure on the design team. To solve this problem, application and platform models are usually developed separately [1]. After this application is mapped onto the target platform and the resulting system is evaluated to ensure its compliance with the design specifications. The application model comes with usually probabilistic terms and platform may come with some low level information from the designers, depending on the targeted level of accuracy. These models are used during the mapping step when the target application is mapped onto the target architecture. Performance analysis step is needed to determine whether the chosen application architecture combination satisfies the design constraints. The information provided during the performance analysis step is used to refine the communication architecture and communication topology and buffer space etc.

In this paper a performance queueing (PQ) model, is proposed and evaluated for NoCs. The PQ model, which is based on M/G/1 Queueing model, has been developed for deterministic routing and wormhole switching. The model is proposed for 2D Mesh topology and traffic pattern. The estimated performance metrics such as average latency and router blocking time can be conveniently used for optimization purposes to find appropriate design parameters, as well as obtaining quick performance estimates.

II. Related Work

Much of the previous analytical models have been formulated for a specific topology and traffic pattern[2], [3], [4],the authors utilized a queueing model and presented a performance model to overcome the problem of buffer allocation in NoC based systems, but the approach cannot handle the wormhole-switched networks. The authors in [5] addressed the allocation of link capacities in NoCs through an analytical latency model. Their proposed model, however, only works for networks with single flit buffers and also ignores the queueing delays and network contentions .A more accurate analytical router model has been proposed in [6]. This work assumes the packet arrivals to the network follow the Poisson distribution. As a result, such models lack the accuracy for use in applications with bursty traffic. The author in [7] proposes the PQ model for predicting the communication performance of wormhole-switching NOC platform. This proposed model is validated through simulation experiments. A worst case analysis of flow latency in the NoC based system was considered in [8]. In [9] a mathematical performance model for NoC based system was proposed to predict performance metrics in NoCs. However, the modeling approach was limited to k-ary n-cube networks with single flit buffers and dimension order routing algorithm.

2.1 Router Model

Heart of the proposed methodology is a new router model which is based on a set of FIFO buffers connected by a switch .The new proposed router model enables us to drive a closed–form expressions for the average number of packets at each input buffer in the router under Poisson traffic arrival assumption for the header flit.

Proposed Network performance approach provides three performance Metrics:

- Average Buffer utilization of each buffer.
- Maximum Network throughput.
- Average Packet Latency per flow.

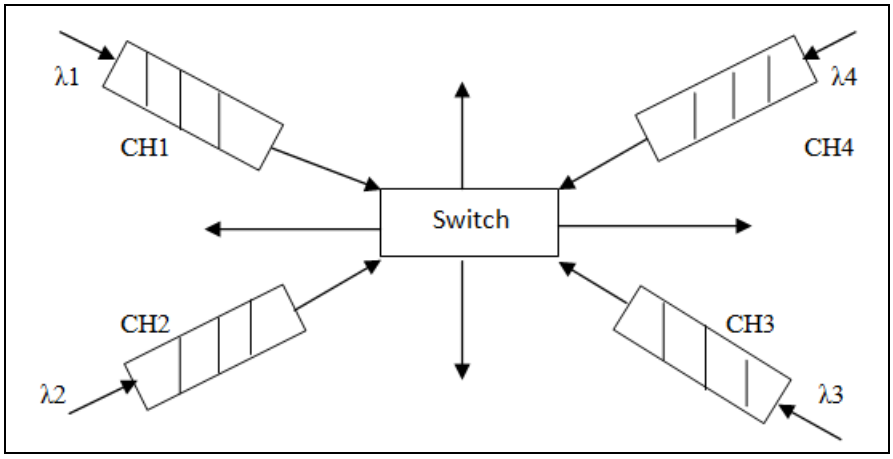

Fig. 2.1: Router Model as a collection of FIFO Buffer

Queueing theory is an appropriate and useful modeling tool for system analysis and performance evaluation in communication networks. It is the mathematical study of waiting lines or queues. In queueing theory a model is constructed so that queue lengths and waiting times can be predicted. The proposed model is M/G//1 queue.

a. M/G/1 Queue

It is a finite capacity, single server queue. Here consider finite capacity, single server queue with Poisson arrivals and generally distributed service times. The queue has (k-1) waiting positions where job can wait if they find the server busy on arrivals. This implies that the system states will range over 0, 1, -------k. customers arriving when the system is full i.e. system is in state k, is not allowed to enter the queue and has to leave without service. This queue therefore rejects customers who encounter blocking.
The queue of this type is finite capacity single server M/G/1/K Queue

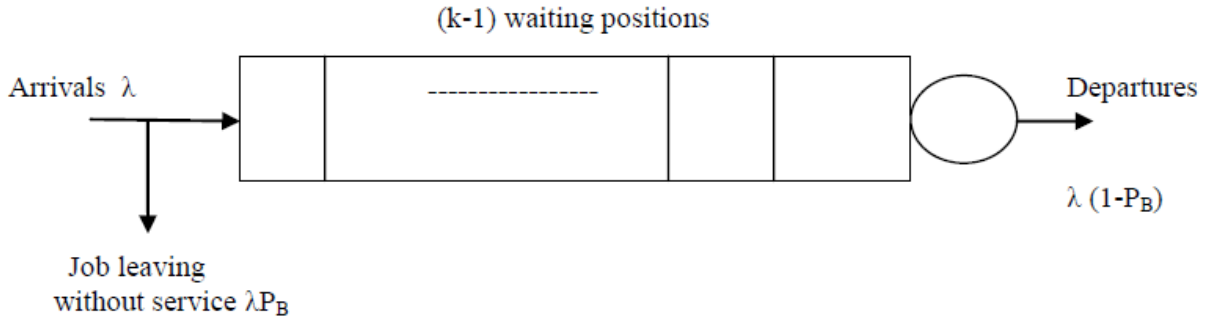

Fig. 2.2: M/G/1 Queue Model

Probability of blocking P_B = P {arrival finds queue full}
Only a fraction $(1-P_B)$ of the arrivals actually enters the system. Therefore effective arrival rate of job entering the queue is only $\lambda (1-P_B)$

b. Priority Queue

Usually not all jobs have the same urgency. Some jobs are supposed to be ready within a day or a week, while other jobs have a delivery date of 4 to 6 weeks from now. Further some customers are regular ones with contracts specifying short throughput times; others are occasional and receive a delivery date according to the present amount of work in process. This makes it quite natural to study a simplified production system in which arriving jobs belong to different job classes and these job classes have different throughput time requirements. To further simplify this we say that the job classes have different priorities. If we number the priority classes from 1 up to r, then class 1 is top priority, class 2 has the second highest priority, etc. Further the job classes may have different processing time characteristics. We denote the processing time of class i by Bi, with mean E(Bi) and mean residual processing time E(Ri) with E(Ri) = $E(B^2 i)/2E(Bi)$ Class i jobs arrive according to a Poisson process with rate _i. We will consider two variants of the priority rule. In the first one a job that has started cannot be interrupted; in the second one the processing of a job can be interrupted by newly arrived jobs of higher priority classes. If all higher priority jobs are served, the servicing of the job is resumed where it was preempted, i.e., no work is lost. The first type of priority is called non-preemptive; the second type is called preemptive-resume. If we think of the situation in which all production is done on one machine, non-preemptive priorities are far more natural, since an interrupt might lead to extra setup time or even to destruction of the product. If however the production capacity is mainly labour, then switching from one job to another might be fairly easy.

Performance Analysis

The following assumptions are made when developing the proposed performance model.

- The PQ model works for deterministic routing algorithms which may be minimal or non-minimal.
- The switching method is wormhole and messages are broken into packets.
- There is one finite FIFO queue per channel and channels are allocated per packets. It means that the channel is released when the whole packet has passed through the channel.
- Packets are consumed immediately by the destination node.

In order to characterize network performance, architecture and application models are essential.

a) Architectural Model

Fig. (a) shows a directed graph can represent the topology of an NoC architecture. Vertices and edges of the graph showing nodes and channels of the NoC respectively. The structure of a single node is depicted in fig (b).

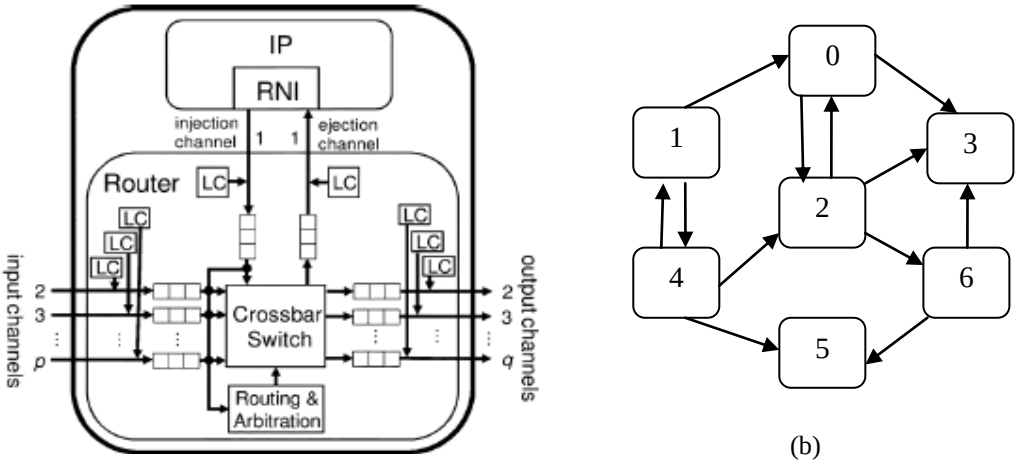

Fig. 3.1(a) Structure of a node in an NoC based System (b) Graph representation of a general NoC architecture

Every node contains an IP core and a router with p input channel and q output channels. Each IP core performs its own computational, storage or I/O processing functionality, and is equipped with a resource-network-interface (RNI). The RNI translates data between IP cores and routers by packing/unpacking data packets and also manages packets injection process. Packets are injected into the network on the injection channel (I/P PORT 1) and leave the network from the ejection channel(O/P port 1) generally each channel connects output port j of node N to input port 1 of node M .Therefore we denote this channel as j^{th} output channel of router N. Consider the general reference architecture for router in [10] and it comprises the following major components.

- Buffer-This is a finite FIFO buffers for storing packets in transit. In the model shown in (b),a buffer is associated with each input physical channel and each output physical channel.
- Link Controller (LC)- The flow of packets across the physical channel between adjacent router is implemented by the link controller. Link controllers on either side of a channel coordinate to transfer flits.
- Crossbar Switch- this component is responsible for connecting router input channels to router output channels
- Routing and Arbitration Unit- this component implements the routing algorithm, selects the output channel for an incoming packet, and accordingly sets the crossbar switch. Routing is only performed with the head flit of a packet .if two or more packets simultaneously request the same output channel. The incoming packets from injection channel have the highest priority in each priority group. A control mechanism prevents the network from being overloaded

B) Application Model

Applications are typically described as a set of concurrent tasks that have been already assigned and scheduled onto a set of selected IP cores. The mapping problem for NoCs is to decide how to topologically place the selected set of cores onto the PEs of the network, such that some metrics of interest are optimized**.**

Basic Assumptions and Notations

S - Random variable (rv) denoting the packet size (bits)
Xsd – packet transmission rate from node s to node d (packets/sec)
R – Residual packet waiting time
H_s – Router Service rate / time for the header flit
W – Network channel bandwidth (bits/sec)
B_{ij} – size of the input buffer at router i, channel j
T, T^2 - Random variables **T** denotes the packet service time
T & T^2 are the first and second order moment
C_{ij} , C – Contention probability between channels i and j
λ_j – mean arrival rate at input buffer of channel j
λ_{ij} – traffic arrival rate at router i channel j (packets/sec)
H_s is a function of router design and includes the time to traverse the router (t_R) and link (t_L).
Since the remaining flits are in a pipelined fashion, the service time of a packet, excluding the queuing delay is given by-

$$\mathbf{T} = H_s + \left[\frac{S}{W}\right] \qquad \text{……..(1)}$$

Assume that the arrival process of the header flits to the router inputs $\lambda_{ij\ follows}$ a Poisson process. Under this model, the arrival process for the body flits is not assumed to be Poisson. This assumption is actually quiet common in network performance analysis [12, 13], enables us to drive closed loop solution and show that model generalizes the classical results for single queue systems.

III. Analytical Model Of The Router

Modeling a single router as a set of first-come-first serve (FCFS) buffers connected by a crossbar switch Parameter of interest is the average number of packets at the input buffers, at each input channel 1, -----,p
$N = [N_1, N_2, -------, N_P]^T$
Poisson's arrival see time averages (PASTA),
The Equilibrium equation is

$$\lambda_j = \frac{N_j}{\tau_j} \qquad \text{……(2)}$$

τ_j is average time an incoming packet spends in queue

τ_j is composed of following components

1. Service time of the packets already waiting in the same buffer
2. The packets waiting in the other buffers of the same router and served before the incoming packet
3. The residual service time seen by an incoming packet (R)

$$\tau_j = TN_j + T\sum_{k=1,k\neq j}^{P} C_{jk}N_k + R \qquad \text{…….(3)}$$

C_{jk} is contention probability that channel j and k compete for the same output

Let C_j be the row vector

$C_j = [C_{j1}, C_{j2}, \text{-------}, C_{jp}]$ of the contention probabilities,

Where $C_{jj} = 1$.

$$\lambda_j = \frac{N_j}{TC_jN + R} \qquad \text{..... (4)}$$

$$\lambda_j TC_j N + \lambda_j R = N_j \qquad \text{..... (5)}$$

This equation describes the equilibrium condition of the buffer at the input channel j only.

$$A = \begin{bmatrix} \lambda_1 & 0 & --- & 0 \\ 0 & \lambda_2 & --- & 0 \\ - & - & --- & - \\ 0 & 0 & --- & \lambda_3 \end{bmatrix}_{PXP}, C = \begin{bmatrix} C_1 \\ C_2 \\ --- \\ C_P \end{bmatrix}_{PXP}, \overline{R} = R \begin{bmatrix} 1 \\ 1 \\ --- \\ 1 \end{bmatrix}_{PX1}$$

Then the equilibrium condition for the router can be designed as:

TACN + A = N

(I – TAC) N = A

Finally,

$N = (1 - TAC)^{-1} A$ ..... (6)

The router model described by equation (6) provides a closed form expression for the average number of packets at each buffer of the router, given the traffic arrival rates (A), the packet contention probabilities (C), router design specification (H_S, W) and packet size distribution(S).

The equation generalizes the single queue model we note that when denominator (1 – TAC) = 0, the packet population in the router grows to infinity. This corresponds to the case when the utilization is 1 for a system with a single queue.

If Contention probability = 1, single queue system and infinite buffers,

In this case equation (1) becomes

$$N = \frac{\lambda R}{1 - T\lambda} \qquad \text{..... (7)}$$

The Residual waiting time is $R = 1/2\lambda\overline{T^2}$

Where $\overline{T^2}$ is the second moment of the service time

$$N = \frac{\lambda^2 \overline{T^2}}{2(1 - T\lambda)} \qquad \text{.....(8)}$$

This is precisely the average number of packets in an M/G/1 (Single server system with general service time) queuing system

Overview of the Performance Analysis Methodology

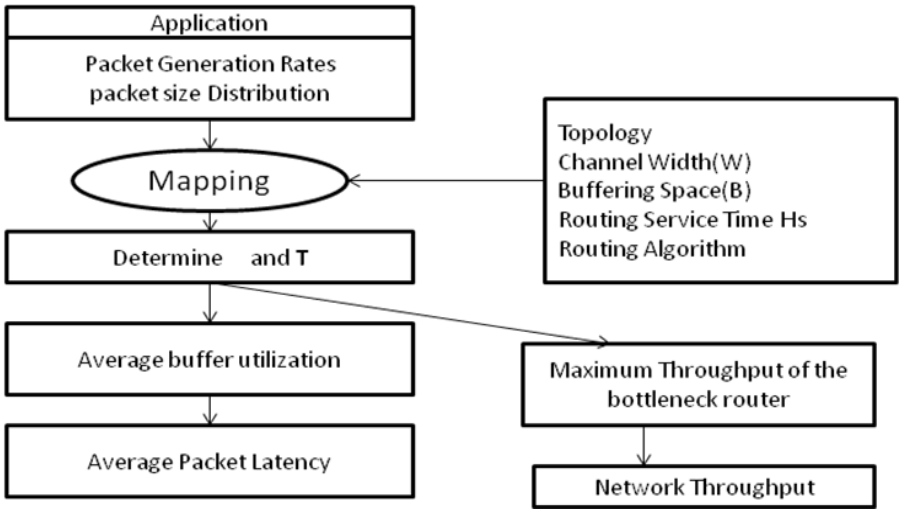

Fig. 3.2: Overview of the proposed Performance analysis Approach

IV. Conclusion

Presented a novel router model for NoC Performance analysis. This approach provides not only aggregate performance metrics such as average latency and Throughput, but also feedback about the network characteristics such as buffer utilization, average latency per router and per flow.

As a result the proposed approach can be used as a powerful **Design** and **Optimization** tool.

Reference

Journal Papers

[1]. R. Marculescu, U. Ogras, L. Peh, N. Jerger, and Y. Hoskote, "Outstanding Research Problems in NoC Design: System, Microarchitecture, and Circuit Perspectives," IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems, Vol. 28, No. 1, January 2009.

[2]. E. Kiasari, H.Sarbazi-Azad and M.quld Khaoua "An accurate mathematical Performance model of adaptive routing in star graph," Future generation computer System, Vol. 24 no.6, pp 461-471, 2008

[3]. J. Kim and C.R.Das, "Hypercube communication delay with wormhole routing," IEEE Transaction on Computer vol.43, no.7, pp.806-814, July 1994.

[4]. Z. Guz, I. Walter et.al. "Network delays and link capacities in application Specific wormhole NoCs," J. VLSI Design, vol 2007,2007 Article ID 90941

[5]. U.Y.Ogras,P.Bogdan and R. Marculescu, "An analytical approach for network-on-Chip performance analysis," IEEE Trans .Computer-Aided Design Integr.Circuits Syst.vol29,no.12 pp2001-2013

[6]. Abbas Eslami Kiasari,Zhonghai Lu,and Axel Jantsch, " An Analytical Latency Model for Network- on –Chip," IEEE Transaction on VLSI Systems,vol.21.No. 1,Jan 2013

[7]. F.Jafari, Z. Lu, A.Jantsch, and M.H.Yaghmace, "Buffer Optimization in networks on chip through flow regulation," IEEE Transaction Computer Aided Design Integr. Circuits Systems, vol 29, no.12 pp.1973-1986,Dec 2010

[8]. A.E.Kiasari,H.Saebazi-Azad and S.Hessabi, " Caspian A tunable performance model for multi-core systems," in Euro par 2008 parallel processing E.Luque, T. Margalef and D.Benitez Eds New Yark Springer Verlag 2008,pp100-109, Lecture notes in Computer Science.

[9]. M. Palesi , R.Holsmark, S.Kumar and V. Ctania, " Application Specific Routing Algorithms for network on c hip," IEEE Trans. Parallel Distributed System.vol.20 no.3 pp.316-330 March 2009

[10]. Draper J.Ghosh J, "A Comprehensive analytical model for wormhole routing in multicomputer systems," J.Parallel Distrib. Comput Vol 23, no.2 pp. 202-214,1994

[11]. Hu P.Kleinrock L., "An analytical model for wormhole routing with finite size input buffers," 15th International teletraffic congress, June 1997

Books

[1]. J.Duato, C. Yalamanchili and I. Ni, Interconnects Networks: An Engineering Approach, San Francisco, CA: Morgan Kaufmann, 2003

[2]. J. Kleinrock, Queueing Systems, New York Wiley, 1995 Vol.1

[3]. Heera Gupta Operation Research .

[4]. Umit Y. Ogras,Radu Marculescu, Modeling,Analysis and Optimization of Network-on-Chip Communication Architectures, Springer2013.