DESIGN AND PERFORMANCE ANALYSIS OF ROUTER IN NETWORK-ON-CHIP USING QUEUING THEORY

Immanuvel.C[1] B.Jaiganesh[2] S.A.Sivasankari[3]

[1,2,3]Saveetha University
Email: immanuvelc@gmail.com

Abstract

In response to meet higher routing complexity in System - On - Chip (SoC), Network - On - Chip (NoC) has been proposed as a flexible and scalable solution. Router design is one of the most important factors that significantly impacts NoC system performance. Therefore, acquiring an accurate estimation of the router performance is an important parameter in Networks-on-Chip. In this paper, optimum router model is designed for NoC and explains how it can be used to study the effect of changing the queue size, routing algorithm and the number of ports. Queuing Analysis is used to obtain an analytical model for an NoC based router for mesh topology with reduced overall delay and power consumption.

Keywords: Network- on- Chip (NoC), System - on- Chip (SoC), source(S), Destination (D).

I. INTRODUCTION

The growing complexity of system-on-chip (SoC) designs aims both industrial and academic researches to find proper solution to solve the communication problem between on-chip processing elements (PEs). With SoC designs that have hundreds of PEs, implementing on-chip communication using shared buses is no longer a practical solution. To address this problem, Networks – on – chip (NoC) is proposed to provide a solution that achieves an efficient communication infrastructure between the PEs.

Optimizing the power consumption of NoC – based designs has become more critical with the use of high speed, complex ICs in mobile and portable applications [1]. Power constraints are among the major bottlenecks that limit the functionality and performance of complex NoC- based designs [2]. Therefore, several methodologies have been proposed to address the power dissipation problem from system level approach [3, 4].

In system level approach, router design, mapping of PEs and first-in-first-out (FIFO) buffer resizing are introduced to address the power dissipation problem:

(1) Optimum router is designed in terms of Routing Algorithm (XY,YX,XY-YX), no of ports and depth of the buffer [5, 8], (2) Mapping of PEs depends on best matching between PEs physical placement and their average communication traffic pattern [6, 7], (3) FIFO resizing focuses the optimum buffer size that achieves the lowest power consumption [9]. To address such problems, the contributions of this paper are twofold:

First, the proposed methodology gives the total available buffering space, the arrival rates between different communicating IP pairs and other relevant architectural parameters (e.g., routing algorithm, no of ports and queue size).

Second, a novel analytical model is proposed which can be used to quickly analyze a given buffer size configuration and detect potential performance bottlenecks in the router channels. This is done by solving a set of nonlinear equations derived from detailed queuing models. This analytical model lies at the very heart of the algorithm for allocating buffer resources. The main advantage in using this analytical approach as opposed to straightforward simulation is the ability to quickly analyze the overall system's performance.

This paper is organized as follows. Section 2 discusses the NoC architecture. Section 3 explains the router problems. Section 4 shows solving the router buffer allocation problem. Section 5 explains the power analysis in NoC based systems. Conclusion is in Section 6.

II. NoC ARCHITECTURE

To analyze the power consumption of routers using NoC architecture we use queuing theory. Fig.1 shows an m * m output queuing router. In output queuing routers, packets arrive at the input of the router asynchronously. Then, the packet header of each

incoming packet, which contains the destination address, is determined by the routing module. As shown in fig 1, there are m queues for each output buffer serving as FIFO buffers. Finally the output buffer uses a round robin scheduling algorithm to serve backlogged queues one after another at each output port.

To analyze the power consumption of the shown NoC router, a set of synthesized routers (with different number of ports, routing algorithm and depth of the buffer) are modelled in verilog hardware description language (HDL).To calculate the power consumption of various ports of the routers various experiments are performed.

III. ROUTER PROBLEMS

The problem to be solved is to find the buffer depth assignment for each input channel, across all the on-chip routers, such that the communication performance is maximized. If the performance is measured in terms of average packet latency, then maximizing the performance means, in fact, minimizing the end-to-end packet latency.

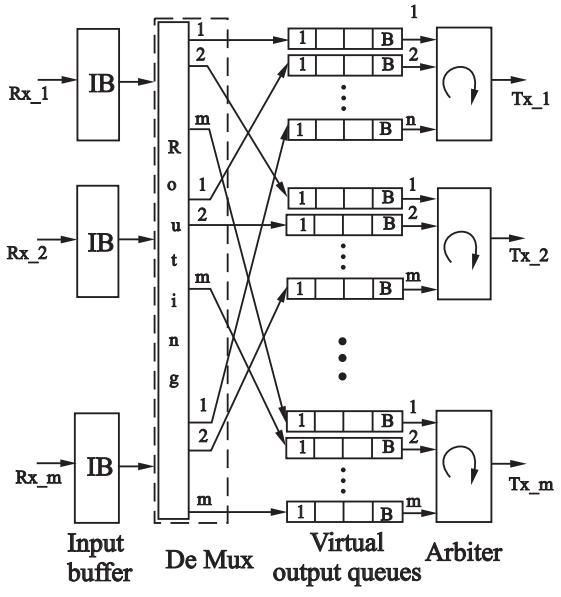

Fig. 1 $m \times m$ NoC - based queuting router

A. System characterization

The system under consideration is composed of $m \times m$ output queuing router interconnected by a 3D mesh network as shown in fig 2.

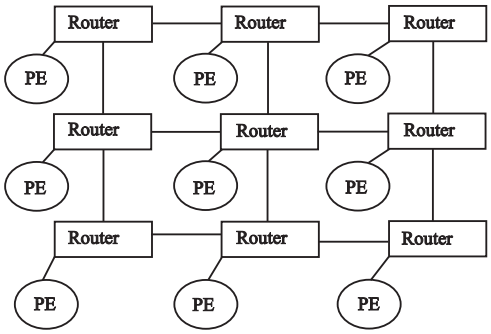

Fig. 2. Regular 3×3 Mesh topology

Deterministic routing is shown in fig 2(a), 2(b) and 2(c) (e.g. XY, YX, XY – YX Routing) are used to direct the packets across the network instead of adaptive routing because of the resource limitations, as well as the out-of-order packet delivery problem associated with the adaptive routing. For simplicity, assume that all the packets in the network have a fixed size. When a new packet is received, the address decoder processes the incoming packet and sends the destination address to the channel controller;

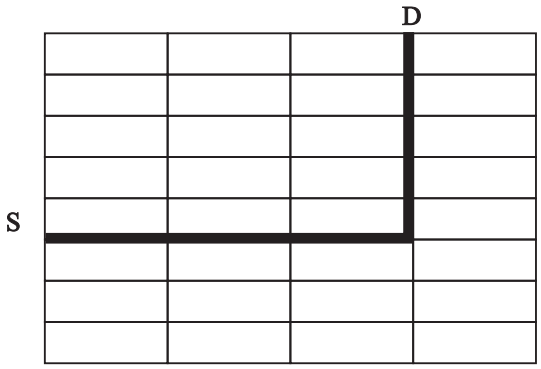

Fig. 2(a) X Y Routing

This controller determines which output channel the packet should be delivered to. Once the router has made the decision on which direction the data should be routed to, the channel controller sends the connection request to the Crossbar Arbiter in order to set up a path to the corresponding output channel. Once a packet arrives

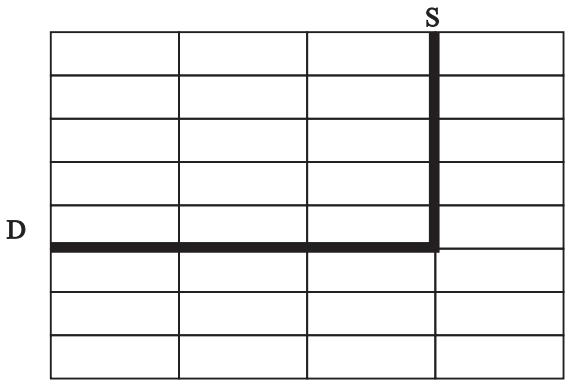

Fig. 2(b) Y X Routing

at an input buffer, it will be served on a first-come-first-served (FCFS) manner. Without loss of generality, we assume that the buffer size is measured in multiples of packet size. More specifically, let size of a packet be SP bytes; then the size of any input buffer must be $m \times SP$, where m is a positive integer. The Crossbar Arbiter maintains the status of the current crossbar connection and determines whether or not to grant connection permission to the channel controller. When there are multiple input channel controllers' requests for the same available output channel, the Crossbar Arbiter also uses the FCFS policy to decide which input channel gets the access, such that the starvation at a particular channel can be avoided.

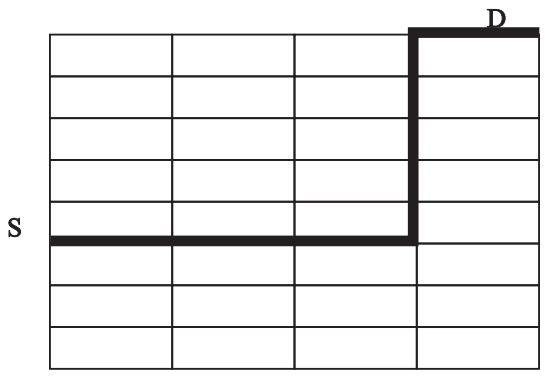

Fig. 2(c) X Y-Y X Routing

B. Problem formulation

For convenience, the basic parameters are given in Table1.1.With these notations; the problem of router buffer allocation for performance maximization under total buffering space constraints can be formulated. The average time, waiting time, depth of the buffer and queue lengths are calculated based on the parameters.

Table 1.1 Parameter Notations

Parameters	Description
Sp	Size of the packet
B	Total buffering space budget
L	Average packet latency
R	Routing function
$R_{x, y}$	The Router located at the (x, y)
$C_{x, y, dir}$	The dir direction input channel in router $R_{x, y}$
$PE_{x, y}$	The PE located at tile (x, y)
$1_{x, y}$	The buffer xize of channel $C_{x, y, dir}$
$a_{x, y}$	Packet injection rate of $PE_{x, y}$
$\mu_{x, y, dir}$	Service rate for packet at channel $C_{x, y, dir}$
$\lambda_{x, y, dir}$	Packet arrival rate at channel $C_{x, y, dir}$
$\rho_{x, y, dir}$	Utilization factor of channel $C_{x, y, dir}$
bx, y, dir	The probability of the buffer at $C_{x, y, dir}$ being full

IV. SOLVING THE ROUTER BUFFER ALLOCATION PROBLEM

In this section, a novel buffer allocation algorithm is proposed which starts from the minimum buffer size configuration (where each input channel has a buffer size of only one packet) and iteratively increases the buffer size of the bottleneck channels until the specified value of the buffer budget is reached. The proposed analytical model can be used to quickly analyze the current buffer size configuration and detect the performance bottlenecks in the router channels; this is done by solving a series of nonlinear equations derived from queuing models. The basic idea is that, given the system configuration (which includes the traffic pattern, the routing delay and the size of each FIFO in the current solution), the algorithm detects the FIFO which has the highest probability to be in the *"*full state*"*. The channel which owns this particular FIFO becomes the real performance bottleneck in the current configuration and thus its size should be increased. To solve this problem analytically, we resort to the theory of finite

queuing networks. The basic element in the model is a M/M/1/K finite queue (the first two "M" mean that the customer arrival time and server's service time follow exponential distributions,"1" tells that the queue has one server to provide the service, and finally "K" represents the capacity of the queue). In this case, the channel *Cx, y, dir* is modelled as a finite queue of length *lx, y, dir*, with the arrival rate λ, *x, y, dir*, served by one server with service rate μ *x, y, dir*. Both inter-arrival and service times are independent and identically distributed, following exponential distributions. With this model, further develop the queuing model of a router as shown in Fig. 3. The five bubbles in the left hand side represent the five channels of *Rx, y* (that is, the router placed at tile (*x, y*)), with *N, E, W, S* and *L* representing the directions of north, east, west, south and local, respectively. On the right hand side, the upper four bubbles represent the four corresponding input channels in router $R_{x, y}$'s neighbouring routers and the bottom bubble represents the output channel to $R_{x, y}$'s local PE ($PE_{x, y}$). These five bubbles on the right side give all the queues that the packets in $R_{x, y}$ can possibly go to during the next time step and thus directly affect the calculation of the parameters related to $R_{x, y}$. Now let us consider, for instance, the north input channel at router $R_{x, y}$. This channel is represented as $C_{x, y, N}$ in Fig. 4. Assuming the network is not overloaded (that is, λ *x, y, dir* < μ *x, y, dir*), then the arrival rate of $C_{x, y, N}$ can be calculated using the following.

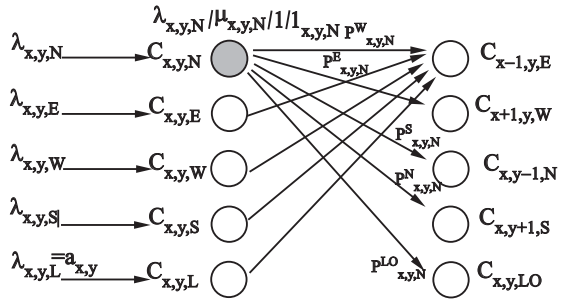

Fig. 3 Queuing model of Router

$$\lambda_{x, y, N} = \sum_{j, k} \sum_{j', K'} a_{j, k} \times d_{j, k}{}' K' \times R(j, k, j', K', x, y, N)$$

The routing function $R(j, k, j', K', x, y, N)$ equals 1 if the packet from $PE_{j, k}$ to $PE_{j', K'}$ uses the channel $C_{x, y, N}$; it equals 0 otherwise. A deterministic routing algorithm, thus the function of $RT(j, k, j', K', x, y, N)$ can be predetermined. Now, the only unknown parameter for $C_{x, y, N}$ is μ *x, y, N*. Once the value of μ *x, y, N* is determined, the probability of $C_{x, y, N}$ to be in *"*full state*"* can be calculated straightforward using the finite M/M/1/K queuing model with the following equations.

$$b_{x, y, N} = \frac{1 - \rho_{x, y, N}}{1 - \rho_{x, y, N}^{1_{x, y, N}+1}} \times \rho_{x, y, N}^{1_{x, y, N}}$$

The average waiting time for entering the FIFO of $C_{x+1, y, W}$ can be approximated as.

$$W_{x+1, y, W} = \frac{1}{\frac{1}{b_{x+1, y, W}} - \lambda_{x+1, y, W}}$$

By performing similar derivations for all input channels, we can finally build a series of equations which describe the system's behaviour. When given other parameters (*i.e.*, routing function $R(j, k, j', K', x, y, dir)$ and $l_{x, y, dir}$), these equations can be solved together by a nonlinear equation solver to

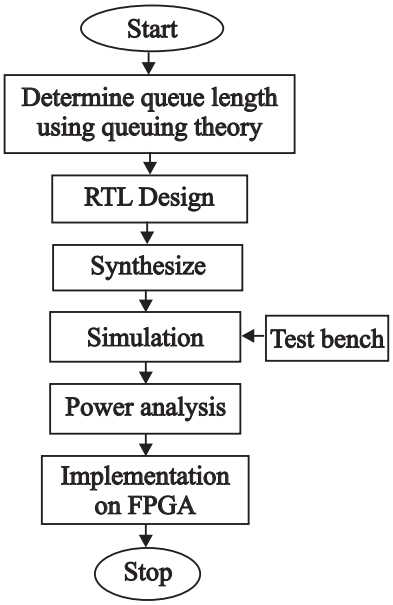

Fig. 4 Power measurement of NoC Routers

determine the important parameters related to the system performance (such as $\mu_{x, y, dir}$ and $b_{x, y, dir}$).

V. POWER ANALYSIS IN NoC – BASED SYSTEMS

To analyze the power consumption of the above NoC router, a set of synthesized routers (with different number of ports and queue sizes) are modelled in Verilog HDL.Based on the test bench waveform the simulation results are obtained. After this power is calculated and the results are implemented on FPGA. Acquiring the minimum value of total power depends on many design factors such as traffic distribution, routing algorithm number of ports per routers and depth of the buffer.

VI. EXPERIMENTAL RESULTS

Fig 5 and fig 6 shows the simulation results and explains the working of the memory and transmitter unit When the. R/W count is equal to zero, it indicates that memory is free and the data can be sent .The transmitter recieves the signals and processes them using round robin approach. The multiplexed signals are then sent to the output when the R/W count is equal to four, it indicates that memory is full and the data cannot be sent

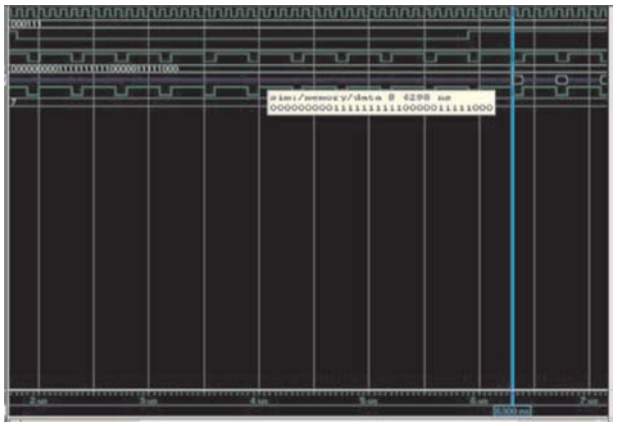

Fig. 5 Memory unit

VII. CONCLUSION

The proposed methodology is used to minimize the power consumption by varying depth of the buffer according to the traffic pattern. This increases the overall performance of the system.

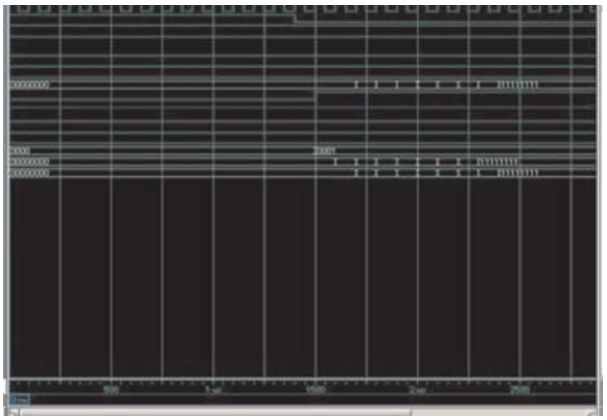

Fig. 6 Transmitter unit

REFERENCES

[1] Haytham Elmiligi et.al, 2009 power optimization for application – specific networks – on – chips: A Topology based approach, Microprocessors and Microsystems 33 343–355.

[2} Meyer. B.H., Pieper J.J., Paul J.M., Nelson J.E. , Pieper S.M., Rowe A.G., 2005. Power-performance simulation and design strategies for single-chip heterogeneous multiprocessors, IEEE Transactions on Computers 54 (6) 684–697.

[3] Banerjee. N., Vellanki P, Chatham K.S., 2004, A power and performance model for network-on-chip architectures, in: Proceedings of the Design, Automation and Test in Europe (DATE'04), Paris, France, pp. 21250–21256.

[4] Simunic. T., Boyd S., Glynn P., 2004 Managing power consumption in networks on chips, IEEE Transactions on Very Large Scale Integration (VLSI) Systems 12 (1) 96–107.

[5] Bhat S., 2005 Energy models for networks-on-chip components, Master's thesis, Technische University Eindhoven, Eindhoven, Netherlands, December.

[6] Benini. L., Siegel P., Micheli G.D., 1994, Saving power by synthesizing gated clocks for sequential circuits, IEEE Design and Test 11 (4) 32–41.

[7] Ogras. U.Y., Marculescu R., 2007, Analytical router modelling for networks-on-chip performance analysis, in: Proceedings of Design, Automation and Test in Europe Conference (DATE'07), Nice, France, pp. 1096–1101.

[8] Chandra. V., Xu A., Schmit H., 2004, A low power approach to system level pipelined interconnect design, in: Proceedings of the 2004 international workshop on System level interconnect prediction (SLIP'04), Paris, France, pp. 45–52.