Australian Journal of Basic and Applied Sciences, 6(6): 21-28, 2012
ISSN 1991-8178

An Analytical Model to Minimize EDP Using Virtual Channel Optimization in On-Chip Networks

[1]Mahdiar Ghadiry, [2]Mahdieh Nadi

[1]Young researcher club, Arak Branch, Islamic Azad University, Ashtian, Iran
[2]Department of Computer Engineering, Arak Branch Islamic Azad University, Arak, Iran

Abstract: One of the most important issues in on chip networks is reducing power consumption. The number of virtual channels has significant effect on power consumption and delay of the network. In this paper, an analytical approach will be presented to optimize the number of virtual channels in order to minimize EDP on NoCs. The proposed model is base on Orion power model. The new point in this model is this model is independent to simulation despite of Orion power model. To show the application of this model a network has been implemented using VHDL and the optimum number of virtual channels has been obtained using both VHDL model and analytical model. The results show that the values obtained from simulation and proposed model are close together. The proposed model can be used to identify optimum number of virtual channels fast and easily.

Key words: On Chip Networks, Power, Model, Performance, EDP

INTRODUCTION

As VLSI feature size shrinks, the density of transistors increases and it is possible to place many IPs in a single chip. The most important problem of the early on-chip systems is the IPs interconnection. It seems that scalability problem of current interconnection will be solved with proposing the NoCs. In this approach, switches are used to connect IPs instead of using shared buses. Some standard interfaces can be defined between network (collection of links and switching elements) and IPs. Therefore, the design of IPs will be independent of the network. On the other hand NoC improves bandwidth with the use of concurrent connections and decreases power consumption by removing the long interconnection wires (Dally, W.J., B. Towels, 2001).

Dally and Seitz (1987) have used the concept of *virtual channels* to develop deadlock free deterministic routing. A virtual channel has its own queue, but shares the bandwidth of the physical channel in a time-multiplexed fashion.

Power efficiency is one of the most important issues in early system design. For current process technologies, *dynamic power* is the primary power source consumed in CMOS circuits. The power is formulated as $P=Ef_{clk}$, and the energy $E=0.5\alpha CV_{DD}^2$, with clock frequency f_{clk}, switching activity α, total switch capacitance C, and supply voltage V_{DD}. Many analytical performance models for interconnection have been proposed so far, but for the case of power consumption more effort is required yet.

Wang, *et al*. (2002) have proposed a power and performance interconnection network simulator that is capable of providing detailed power characteristics, in addition to performance characteristics, to enable power-performance trade-off at the architectural level. They proposed an architectural-level parameterized power model as a part of that effort. Two routers have been modeled in (H. S. Wang, *et al*., 2003) using model proposed by Wang, *et al*. (2002). In (T.Ye, *et al*., 2004) the authors introduce a framework to estimate the power consumption on switch fabrics in network routers. They proposed different modeling methodologies for node switches, internal buffers and interconnection wires inside switch fabric architectures. A power model for the Nostrum NoC has been proposed in (S. Penolazzi, A. Jantsch, 2006). For this purpose, an empirical power model of links and switches has been formulated and validated by Synopsys Power Compiler. An architectural power modeling for interconnection networks proposed in (X. Chen, L-S. Peh, 2003). In (D. Rahmati, *et al*., 2006) WK-Recursive and Mesh topology are compared in the case of power and latency. They also proposed a novel approach in high-level power modeling based on latency for these topologies and showed that the power consumption of WK-Recursive topology is less than its equivalent mesh on a chip. In (P. Pande, *et al*., 2005) power and performance for various topologies in NoC have been studied. In that paper, some topologies such as BFT, Folded Torus and Mesh have been compared. They proposed a guideline for selection of best topology for a specific application in NoC. A High Level Power Analysis for On-Chip Networks proposed in (T.Ye, *et al*., 2002). Their analysis is based on link utilization as the unit of abstraction for network power, with contention among message flows modeled through propagation of overflow areas in link utilization functions. In (M. Nadi, *et al*., 2007; M. Nadi, *et al*., 2007; M. Nadi, *et al*., 2010). several routing algorithms modeled in VHDL and

Corresponding Author: Mahdiar Ghadiry, Young researcher club, Arak Branch, Islamic Azad University, Ashtian, Iran
E-mail: m.hoseuinghadiry@gmail.com

compared in case of power and performance using simulation. A new approach for communication and implementation of a NOC on FPGA has been reported in (S. A. Asghari, *et al*., 2009).

Note that this work is base on Orion model (H-S. Wang, *et al*., 2002). They have proposed model for most of capacitances of a router but they have used simulation to reach to switching activity values. As a part of our work, we have tried to calculate switching activities analytically in limited and averaged situation with some assumption, and then analyze the effect of virtual channel on EDP of NoC using the results as the other part. In addition, the optimum numbers of virtual channel for the network in various situations have been identified using both simulation and analytical model. Although this model have been provided for K-Ary N-cubes but it can be used for another topology using related performance model and changing a few parameters respect to performance model in the formulas.

Energy And Performance Measures:

We use two measures to calculate energy delay product; the energy and latency of a packet.

Energy:

When flits travel on the interconnection networks, both the inter-switch wires and the logic gates in the switches toggle and this will result in energy dissipation. Here, we are concerned with the dynamic energy dissipation caused by communication process in the network. The flits from the source nodes need to traverse multiple hops consisting of switches and wires to reach destination. Consequently, the energy dissipated by per flits per hop is given by equation 1.

$$E_{hop} = E_{switch} + E_{interconnect} \tag{1}$$

Where E_{switch} and $E_{interconnect}$ depend on the total capacitances and signal activities and each section of interconnect wire, respectively. They are determined as follows:

$$E_{router} = 0.5\alpha_{router}C_{router}V^2. \tag{2}$$

$$E_{interconnect} = 0.5\alpha_{interconnect}C_{interconnect}V^2. \tag{3}$$

α is a parameter between 0 and 1 and demonstrates the switching activity, C is the total switching capacitance and V is the operating voltage. The energy dissipated for transferring a packet with n flits over h hops can be calculated as stated in equation 4.

$$E_{packet} = n\sum_{j=1}^{h} E_{hop,j}. \tag{4}$$

Latency:

Message latency is defined as the time (in clock cycle) that elapses from the occurrence of a message header injection into the network at the source node and the occurrence of a tail flit reception at the destination node. We simply refer to this as *latency* from here on. In order to reach the destination node from some starting source node, flits must travel through a path consisting of set of switches and interconnects, called stages. Depending on the source/destination pair and the routing algorithm, each message may have a different latency. There is also some overhead in the source and destination that also contributes to overall latency. Therefore, for a given message *i*, the latency *Li* is:

$$Li = sender\ overhead + transport\ latency + receiver\ overhead \tag{5}$$

We use the average latency as a performance metric. Let *P* be the total number of message reaching their destination the average latency, L_{avg}, is then calculated accordingly as follows in equation 6.

$$L_{avg} = \frac{\sum_{1}^{P} L_i}{P} \tag{6}$$

EDP (Energy/Delay Product) The EDP obtained by production of the Latency and energy.

$$EDP_{avg} = L_{avg}.E_{avg} \quad (7)$$

In on chip networks, low EDP is desired, since it shows low latency and low energy, although these two parameters are in contrast, decreasing energy causes increasing in latency.

Energy Of A Packet Crossing A Wormhole Router:

Fig. 1 sketches the module representation of a wormhole router and its neighboring links. The source module injects a header flit into the write port of the input buffer module while E_{wrt} is dissipated.

When the flit emerges at the head of the FIFO buffer, it is checked via the read port of the buffer module, its route is read, and a request sent to the arbiter module for the desired output port. The arbiter performs the required arbitration so E_{arb} is dissipated.

Assuming the request is granted, the arbitration result is sent to the Config port of crossbar module. A grant signal also is sent to the grant port of input buffer and therefore the read port of buffer is activated and E_{read} is dissipated. The flit then traverses the crossbar module and dissipates E_{xb}. Finally, the flit leaves the router, enters link and traverses link and dissipates E_{link}

The total energy this header flit has consumed at this node and its outgoing link is as described in equation 8.

$$E_{flit} = E_{wrt} + E_{arb} + E_{read} + E_{xb} + E_{link} \quad (8)$$

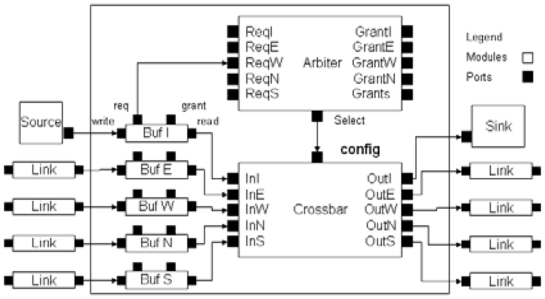


Fig. 1: A simple wormhole router modeled in Orion (H-S. Wang, *et al*., 2002).

Proposed Model To Calculate The Average Packet Energy:

In this section, we present the needed equations to calculate average packet energy when it crosses a router and the outgoing link. In this model we assume K-Ary n-Cubes topology, uniform traffic (each node can send packets to all other nodes with the same probability), random data in each packet such that total number of 1s is almost equal to 0s, and also Duato fully adaptive routing algorithm for routing algorithm. It should be noted that the energy for processing of routing algorithm is not covered and has been neglected. Although it is possible to implement the desired routing algorithm in a hardware description language such as VHDL and obtain the average energy using power simulators (e.g. Power compiler or XPower) and add it to the values derived from model (M. Nadi, et al., 2007).

In the following equations $E_x=0.5\alpha C_x V_{DD}^2$ (*) which x can be substituted with desired module and C_x is the total capacitance of that module calculated as is described in **(**H-S. Wang, et al., 2002). We refer to the equation as * in the rest of the paper. Note that for calculating this equation we should count each transition from 0 to 1 and 1 to 0 to obtain switching activity (α).

The average energy dissipated when a packet crosses a switch (router) contains header flit and non-header flit energy. In wormhole switching only header arbitrated and the other flits follow the header in the same route. Here we consider header size is one and the average packet size is L_p flits.

Let $\overline{E}_{packet\_hop}$ and $\overline{E}_{packet\_link}$ be the average energy of a packet which is dissipated due to the hop and link crossing respectively. Thus, when a packet goes on a hop, its energy is given as described in equations 9 and 10.

$$\overline{E}_{packet} = \overline{D}\,\overline{E}_{packet\_hop} + (\overline{D}-1)\overline{E}_{packet\_link} \quad (9)$$

$$\overline{E}_{packet\_hop} = (L_p - 1)\overline{E}_{body\_flit} + \overline{E}_{header\_flit} \quad (10)$$

$$\overline{E}_{header\_flit} = \overline{E}_{write} + \overline{E}_{arb} + \overline{E}_{read} + \overline{E}_{xbar\_header} \quad (11)$$

$$\overline{E}_{body\_flit} = \overline{E}_{write} + \overline{E}_{read} + \overline{E}_{xbar\_body} \quad (12)$$

$\overline{E}_{header\_flit}$ is one hop energy dissipation of a flit, which is described in equation 11.

In above equations $\overline{D}$ is the average distance of source to destination for a given packet. In K-Ary N-Cubes $\overline{D}$ is determined in equation 13.

$$\overline{D} = N\frac{k-1}{2} \quad (13)$$

Let W be the data width of link and equal to cross bar port bandwidth. Therefore the average number of bit flips on links is $W/2$. The link energy then is calculated according to equation 14.

$$\overline{E}_{packet\_link} = \frac{1}{2}\left(L_p C_{link\_unit}\frac{W}{2}\right)VDD^2 \quad (14)$$

let F be the flit size in bits, then the average read and write energy is calculated as followed in equations 15 and 16.

$$\overline{E}_{read} = E_{wl} + F\left(E_{br} + 2E_{chg} + E_{amp}\right) \quad (15)$$

$$\overline{E}_{write} = E_{wl} + \frac{F}{2}\left(E_{bw} + E_{cell}\right) \quad (16)$$

Which E_{amp} is sense amplifier energy and calculated from and the remaining energies can be calculated using * equation and (H-S. Wang, *et al*., 2002)

Matrix crossbar switch is used as switching element. Crossbar switch energy for a header traversing is summation of selected input and output lines and control of switches that connect input lines to output lines energy. The switch configurations remains fixed until the end of the packet transfer, therefore when non-header flits traverse the crossbar switch the control energy is omitted and we have:

$$\overline{E}_{xbar\_header} = \overline{E}_{xb\_in} + \overline{E}_{xb\_out} + \overline{E}_{xb\_ctr} \quad (17)$$

$$\overline{E}_{xbar\_body} = \overline{E}_{xb\_in} + \overline{E}_{xb\_out} \quad (18)$$

$$\overline{E}_{xb\_in} = 0.5VDD^2\left(C_{in\_sw}2NVW + C_a(T_{id}) + C_{Line\_unit}2NVWw_t\right) \quad (19)$$

$$\overline{E}_{xb\_out} = 0.5VDD^2\left(C_{out\_sw}2NVW + C_a(T_{od}) + C_{Line\_unit}2NVWh_t\right) \overline{E}_{xb\_ctr} = 0.5VDD^2\left(WC_{ctr\_sw} + \frac{C_{Line\_unit}2NVWw_t}{2}\right) \quad (21)$$

In above equations V is the number of virtual channels per physical channel, and W is the bandwidth of each link. C_{line_unit} is unit width capacitance of crossbar lines. h_t and w_t are vertical and horizontal line distances. T_{id} and T_{od} are the input and output drivers respectively, as shown in (H-S. Wang, *et al*., 2002).

Let E_{req} to be the header request signal energy to grant for an outgoing link, E_{pri} the energy to store grant priorities, E_{int} the energy dissipated in internal nodes, E_{clk} the flip-flop clocking energy and $E_{gnt,}$ to be the grant signal energy of the arbiter, then the arbitration energy is given by equation 22.

$$\overline{E}_{arb} = \overline{E}_{req} + \overline{E}_{pri} + \overline{E}_{\text{int}} + \overline{E}_{gnt} + \overline{E}_{clk}$$

With proper substitutions of parameters the average energy is calculated as described in equation 23:

$$\overline{E}_{arb} = \left(E_{req} + \frac{(2N-1)V-1}{2}E_{pri} + \frac{(2N-1)V((2N-1)V-1)}{2}E_{\text{int}}\right) + E_{gnt} + \overline{E}_{clk} \quad (23)$$

If we assume there is no U turn in packet path there are totally *(2N-1)V((2N-1)V-1)/2* flip-flop to store priority in arbiter which clocked due to one time clocking to arbiter. Therefore the average energy of one time clocking to arbiter is calculated as described in equation 24.

$$E_{clk} = \frac{1}{2}\left(\frac{(2N-1)V\left((2N-1)V-1\right)}{4}C_{FF\_clock}\right)VDD^2 \tag{24}$$

In equation 24 C_{FF_clock} is the flip-flop clock capacitance. The energy relates to more than one packet existing in that hop. The average energy each packet dissipates in clocking is derived from dividing total clocks energy to reach all packets to destination over total number of packets. Let $\overline{N}_{clk}$ be the average packet latency, -λ_g the packet generation rate per node per cycle, N_n and N_p are total number of nodes (K^N) and total number of packets generated respectively. N_n number of packets reach destination after $\overline{N}_{clk}$ clocks,and similarly $2N_n$ packets after $\overline{N}_{clk} + 2\lambda_g$ clocks. Finally N_p packets reach destination after number of clocks calculated in equation 25.

$$\left(\frac{N_p}{N_n}-1\right)\lambda_g + \overline{N}_{clk} \tag{25}$$

Thus total dissipated clock energy for all packets to reach the destination is given by equations 26.

$$E_{clk_{total}} = \left[\left(\frac{N_p}{N_n}-1\right)\lambda_g + \overline{N}_{clk}\right]E_{clk} \tag{26}$$

And the portion of a packet is given by equation 27.

$$\overline{E}_{clk} = \left[\left(\frac{N_p}{N_n}-1\right)\lambda_g + \overline{N}_{clk}\right]\frac{E_{clk}}{N_P} \tag{27}$$

The remaining energy can be calculated using * equation and (H-S. Wang, *et al.*, 2002) equations. Let B_i be the average blocking time in ith hop and W_{ej} the average blocking time in destination for ejection from the network, then the average packet latency calculated according to model presented (Ould-Khaoua, M., 1999) in as described in equation 28.

$$\overline{N}_{clk} = L_p + \overline{D} + \sum_{i=1}^{\overline{D}} B_i + W_{ej}$$

Note that B_i and W_{ej} are calculated using (Ould-Khaoua, M., 1999; N. Alzeidi, 2004).

Edp Model:

The mean network latency, S, consists of two parts: One is the delay due to the actual message transmission time and the other is due to blocking in the network. Given that a message makes, on average, $\overline{d}$ hops to reach its destination, S can be written as

$$S = M + \overline{d} + \sum_{i=1}^{\overline{d}} B_i + W_{ej} \tag{29}$$

Where M is the message length, B$_i$ is the mean blocking time seen by message at the *i*th hop channel ($1 \le i \le \overline{d}$), and W_{ej} is the mean waiting time at the ejection channel in the destination node.

Here we do not explain the details of using this model. Supporting information can be found in (Ould-Khaoua, M., 1999; N. Alzeidi, 2004).

EDP can be calculated by production of Dally delay and proposed model for average energy. This model can be used to obtain optimum number of virtual channels in order to design efficient NoCs.

Virtual Channel Optimization Using Proposed Model:

There is a trade-off between energy and delay. Increasing the number of virtual channels increases the energy consumption and usually decreases the delay. In low traffic load, time overhead of using more virtual channel causes to increase in delay. On the other hand, energy will be increased due to using more registers and larger modules such as multiplexers. Therefore, for each network there is an optimum number for virtual channels to minimize EDP. Below model can be used to determine the optimum number of virtual channels.

$$\frac{dEDP}{dV} = 0 \qquad (30)$$

Simulation Setup:

A cycle accurate simulator is implemented in VHDL. An 8×8 mesh is used as instance of K-Ary N-Cubes topology with 64 processing elements as the IPs. Totally 10000 packets each one with 32 flits are generated which each packet has a header flit and 31 body flits and all packets contain random data. Uniform traffic is assumed for destination addresses and Duato's fully adaptive routing algorithm is implemented to rout packets. In case of energy calculation we use Orion power model (H-S. Wang, *et al*., 2002). No power reduction codes is used and assumed no repeater is needed between two nodes.

Simulation Results:

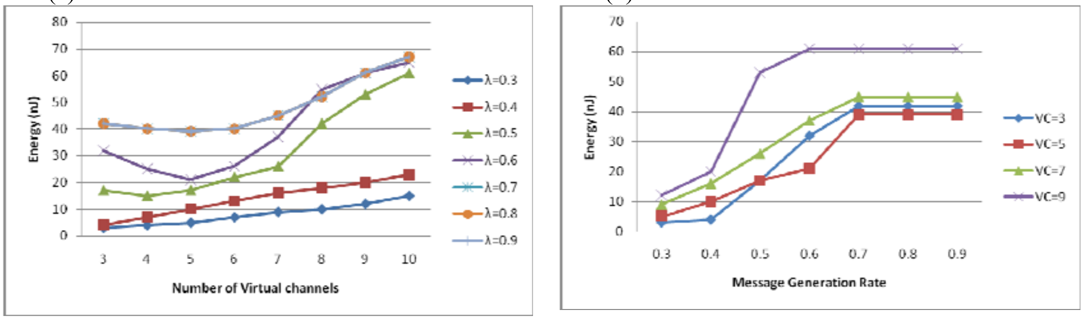
In this section above model is used to identify the best number of virtual channels. In order to verify the obtained values from modeling accurate VHDL simulator has been implemented. In these experiments supply voltage is $1.8V$ and technology is $0.18\mu m$. The number of virtual channels has been changed to see the effects on energy consumption and performance of the network. (Note that in remaining sections, we may use *message* and *packet* interchangeably). In this paper performance means delay.

Energy:

As can be seen in Fig. 2 (a) increasing the number of virtual channels in low traffic load, 0.3 increases the energy. In higher traffics, such as 0.5 there is a point that energy is the minimum value on that and after that point energy consumption increases as the number of virtual channels increase. In low traffic load, increasing in the number of virtual channels increases the line capacitances and clocking power while there are some free virtual channels on each clock that are unused. In 0.8 traffic load 5 virtual channels is optimum value to minimize the energy consumption. Fig. 2 (b) shows energy versus traffic load in various virtual channels. There are saturation points in energy. For example with 9 virtual channels after point 0.6 energy is almost constant. This figure shows that before load 0.5, 3 virtual channels is the optimum value but this value is 5 for loads higher than 0.5. Increasing number of virtual channels more than 5 has not benefit anymore.

Delay:

Fig. 2 (c) shows that increasing number of virtual channels in low traffic loads does not decrease the delay. In opposite it increases the delay due to more multiplexing time and larger line capacitances. However, in higher traffic loads increasing the number of virtual channels causes to decrease in latency. Fig. 2 (d) shows that in high loads, more virtual channels are necessary to decrease the delay. However in low loads high number of virtual channels cause to increase the delay. This is obvious from figure that increasing the message generation rate increase the delay due to blocking.


(a) (b)

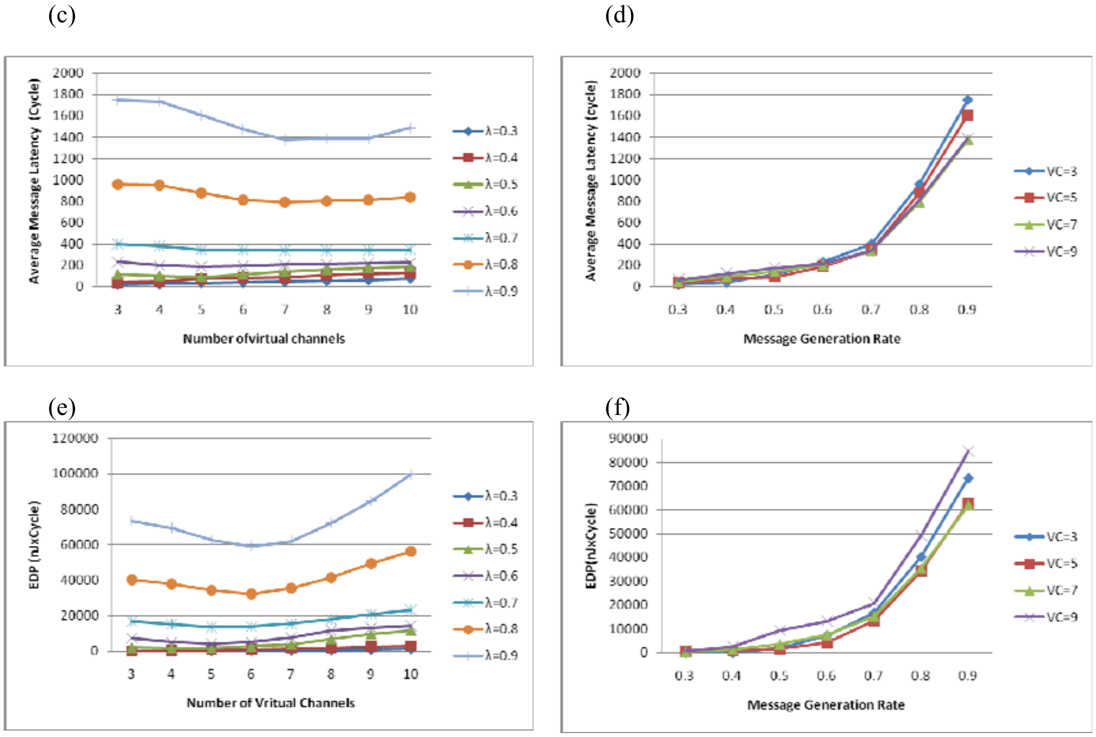


Fig. 2: (a) The effect of number of virtual channels on energy in various loads. (b) The effect of traffic variation on energy with different number of virtual channels. (c) The effect of number of virtual channels on energy in various loads. (d) The effect of traffic variation on energy with different number of virtual channels. (e) The effect of number of virtual channels on energy in various loads. (f) The effect of traffic variation on energy with different number of virtual channels

EDP:

Minimizing EDP is helpful to design networks for both high performance and low power applications. As can be seen in Fig. 2 (e) optimum number of virtual channels in point 0.9 in traffic load is 6. This value is shifted to 5 and 4 as the traffic load decreases. In low traffic load, EDP increases as number of virtual channel increases. the optimum number of virtual channels is 3 in low traffic load. Implementing network with low number of virtual channels decreases the latency when traffic load is high this is because of clocking energy consumption during blocking cycles. On the other hand high number of virtual channel cause to high delay and energy consumption in low traffic load. Therefore, for each network with certain properties certain number of virtual channel should be used to maintain both delay and energy low.

Fig. 2 (f) shows that as traffic load increases EDP increases too. The lowest EDP almost belongs to network with 5 number of virtual channels.

Table I, shows optimum number of virtual channels with aim to minimize the EDP of the NoC. As can be seen the results obtained from model and simulation are closed to each other. Therefore, this model can be used to identify optimum number of virtual channel with certain properties such as average load, number of IPs, technology parameters and all the parameters, which were discussed in Section IV.

Table 1: Optimum number of virtual channels to minimize the EDP.

Lambda	0.4	0.5	0.6	0.7	0.8	0.9
Simulation	3-4	3-4	4-5	5-6	5-6	6-7
Model	3.1	3.7	4.9	5.3	5.9	6.4

Conclusion:

In this paper, we proposed a model to calculate average packet energy in K-Ary n-Cubes. This model can be used to identify best number of virtual channels to minimize EDP on NoCs. Using this model time-consuming and complex simulations can be avoided which decrease the time of designing. In addition, the model shows the relation of the network parameters (e.g. number of virtual channels) with packet energy dissipation. The application of such relationship is to analyze the effect of each part on EDP in order to reduce the energy and delay. It was described that increasing the number of virtual channels improves the latency in high traffic loads, but increases the packet energy on the other hand. But for lower traffic loads there isn't any improvement in either latency or energy. It was shown that except in high traffic loads increasing the number of virtual channels has undesirable effects on EDP.

REFERENCES

Alzeidi, N., M. Ould-Khaoua, A. Khonsari, 2004. "A Queuing Model for Wormhole Routing with Finite Buffers", UKPEW.

Asghari, S.A., H. Pedram, M.P. Yaghini, 2009. "Designing and Implementation of a Network on Chip Router Based on Handshaking Communication Mechanism", World Applied Sciences J. 6(1): 88-93.

Chen, X., L-S. Peh, 2003. "Leakage Power Modeling and Optimization in Interconnection Networks", Symp. on Low power electronics and design.

Dally, W.J., B. Towels, 2001. "Rout Packets, Not Wires: On-Chip Interconnection Networks", DAC.

Dally, W.J., C.L. Seitz, 1987. "Deadlock-Free Message routing in Multi-computer Interconnection Networks", IEEE Trans. Computers, 36(5): 547-553.

Nadi, M., M. Hosein Ghadiry, M.T. Manzuri Shalmani, 2007. "Power and Performance Comparison of Routing Algorithms on NoC", ICEE.

Nadi, M., M. Hosein Ghadiry, M.T. Manzuri Shalmani, D. Rahmati, 2007. "Effect of Number of Faults on NoC Power and Performance", ICPADS.

Nadi, M., M.H. Ghadiry, M.K. Dermany, 2010. "The effect of number of virtual channel on NOC EDP", J. Appl. Math. and Informatics, (1): 539-551.

Ould-Khaoua, M., 1999. "A Performance Model for Duato's Fully Adaptive Routing Algorithm in k-Ary n-Cubes", IEEE Transaction on Computer Design, 48(12).

Pande, P., C. Grecu, M. Jones, A. Ivanov, R. Saleh, 2005. "Performance and Design Trade-Offs for Network-on-Chip Interconnect Architectures", IEEE Transaction on Computers, 54(8).

Penolazzi, S., A. Jantsch, 2006. "A High Level Power Model for the Nostrum NoC", EUROMICRO,

Rahmati, D., A. Kiasari, S. Hessabi, H. Sarbazi-Azad, 2006. "A Performance and Power Analysis of WK-Recursive and Mesh Networks for Network-on-Chips", ICCD.

Wang, H.S., L.S. Peh, S. Malik, 2003. "A Power Model for Routers: Modeling Alpha 21364", IEEE Micro,

Wang, H-S., X. Zhu, LS. Peh, S. Malik, 2002. "Orion: A Power Performance Simulator for Interconnection Networks", in Proc of Micro 35.

Ye, T., L. Benini, G.D. Micheli, 2002. "Analysis of Power Consumption on Switch Fabrics in Network Routers", DAC.