World Applied Sciences Journal 12 (5): 668-673, 2011
ISSN 1818-4952

Energy Reduction in Application Specific Network-on-Chips Using Performance-Power Exchange

[1]Mehdi Saeidmanesh and [2]Ehsan Ahvar

[1]Department of Computer Engineering, Islalmic Azad University, Aligudarz Branch, Iran
[2]Department of Communication, Information Technology and Payame Noor University, Iran

Abstract: Networks-on-Chips (NoCs) are required not only to provide ultra-low latency, but also to consume as little energy as possible. Because higher energy consumption leads to decreased mission duration, increased heat dissipation and decreased reliability. This paper exploit the well-known Dynamic Voltage Scaling (DVS) technique to reduce consumed energy in an application specific NoC without any performance overhead. This method includes two steps. In the first step, extra virtual channels are exploited in the switching element of an NoC node to increase the performance of the network. In the second step, the performance gain achieved in the first step is used to exchange by lower energy consumption using the DVS technique. A flit level VHDL-based simulator and Synopsys Power Compiler tool have been used to extract experimental results. The simulation results show that the method can efficiently reduce the energy consumption of an NoC node at least 20%.

Key words: Dynamic Voltage Scaling · on-Chip Network · Energy Consumption · Performance · Mesh

INTRODUCTION

Today's Systems on Chips (SoCs) consist of a large number of computing and storage cores that are interconnected by means of single or multiple layers of buses In order to cope with the large communication demands of such SoCs, a modular, scalable interconnect based on Networks on Chips (NoCs) is needed [2-4]. NoCs are mainly based on switch-based and packet-based communication.

Each NoC node employs a switching element and communication channels to transmit packets in the network. The switching element determines the way that packets visit intermediate nodes through their path to destinations. In order to increase the performance of the network each communication channel, namely physical channel, can be timely multiplexed between some virtual channels [5]. Each virtual channel consists of buffers associated to a physical channel. A virtual channel with a set of flit buffers associated to a physical channel shares the bandwidth of the physical channel with other virtual channels in a time multiplexed manner. As the number of virtual channels per each physical channel increases, the probability of a packet reaching a preoccupied channel decreases, resulting in lower average packet delivery time and better performance [7-9].

Although in NoCs parallel applications, like network or media processors, are characterized by independent data streams or by a small amount of inter-process communications [1]. However, many general-purpose parallel applications display a bulk-synchronous behavior: the processing nodes access the network according to a global, structured communication pattern. They can, for example, execute a personalized all-to-all information exchange, global synchronization, gather/scatter to/from one node. In contrast to the traditional direct networks in multi-computers, the NoCs have some specific features as follows:

- NoCs should consume as little energy as possible. Because higher energy consumption reduces mission duration (for battery-operated systems), increases temperature which may cause chip damage and decreases reliability [2, 5-7].
- Most NoCs are developed specifically for one application or as a platform for a small class of applications. Consequently, the designer has a good understanding of the traffic characteristics and can use this information to customize the NoC accordingly [10-11].

Corresponding Author: Mehdi Saeidmanesh, Department of Computer Engineering, Islalmic Azad University, Aligudarz Branch, Iran, E-mail: m.saeidmanesh@gmail.com.

One effective and widely used technique to reduce the energy consumption of a digital system such as an NoC is to scale down the supply voltage of the system [13-14]. This idea originates from the fact that the energy consumption is quadratically proportional to the supply voltage. Therefore, if the system can operate with a lower voltage level, its energy consumption will be reduced quadratically.

In this paper the voltage scaling technique has been exploited to reduce the energy consumption of an NoC. Voltage scaling allows devices to dynamically change their speed and voltage, increasing the energy efficiency of their operation [24]. In order to reduce the energy per operation in a system we can increase the delay, allowing an associated reduction in our current operating voltage [24]. Voltage scaling allows a device to dynamically change its voltage while in operation and thus tradeoff energy for delay. The key idea behind the proposed method is that at a predetermined traffic rate, which is a common assumption in application specific NoCs [7], it is possible to increase the performance by using extra virtual channels and then trade the performance gain for energy reduction objective. This can be done simply by reducing the supply voltage (DVS) of the NoC. It should be noted that, the extra virtual channels used to achieve performance gain impose some energy consumption overhead to the NoC. But due to quadratic proportionality of the energy consumption to the NoC supply voltage, the reduced energy is much more than the imposed energy. To prove the claim, a flit level VHDL-based NoC simulator is used and the detailed investigations of the proposed method have been done by the means of Synopsys Power Compiler tool.

The rest of the paper is organized as follows. Section 2 presents some preliminaries about NoCs. Section 3 discusses about energy and power consumption of an NoC. The proposed method has been introduced and evaluated in different traffic rate in terms of performance and energy consumption in section 4. Finally section 5 concludes the paper.

NoC Node Architecture: Three important items in an on-chip network are topology, switching method and routing algorithm. Fixed tile size mesh based topology is favored by many research groups because of its layout efficiency, good electrical properties and simplicity in addressing on-chip resources [3, 15-16]. Figure 1 shows a 4×4 mesh based NoC.

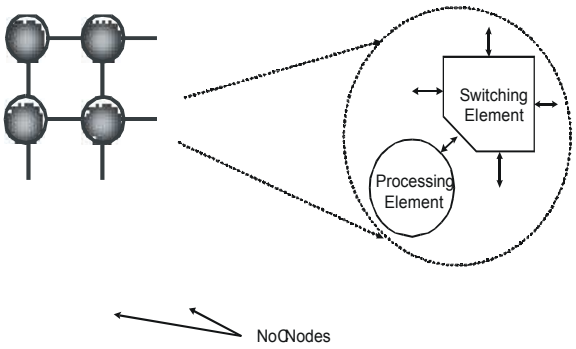

Fig. 1: A 4×4 mesh based NoC.

In wormhole switching, a packet is divided into a sequence of fixed-size units, called flits. The header flit (containing routing information) establishes a path through the network while the remaining data flits follow it in a pipelined fashion. Wormhole is a widely used switching method [16] due to its low buffering requirements and more importantly, because it makes average packet delivery time almost independent of the distance between source and destination nodes. The switching element of the node consists of 5 physical channels to communicate with its north, south, east and west neighbors and the local processor. Figure 2 shows an NoC node architecture which can be utilized in a 2 dimensional mesh based NoC.

When a header flit arrives at an input virtual channel of an NoC node from either the previous node or the local processor connected to the same node, the router decides the packet's destination direction (north, south, east, west or the local processor) and configures the crossbar switch by sending appropriate signals to the crossbar. Crossbar switch then connects the flit's incoming channel to the selected outgoing one. In the case of existence of a free virtual channel in the selected outgoing physical channel, the header flit will be transferred to the next node and the other flits in the packet follow it in a pipelined fashion, otherwise the header flit has to wait until a virtual channel of the selected outgoing physical channel becomes free.

In a high traffic rate situation in which the probability of having a free virtual channel is rather low, the header flit has to wait for a free virtual channel for a relatively long period of time. The use of more virtual channels per physical channel can improve network performance by increasing throughput and reducing packet blocking time [13, 17]. In the figure 1, each physical channel is timely shared between k virtual channels.

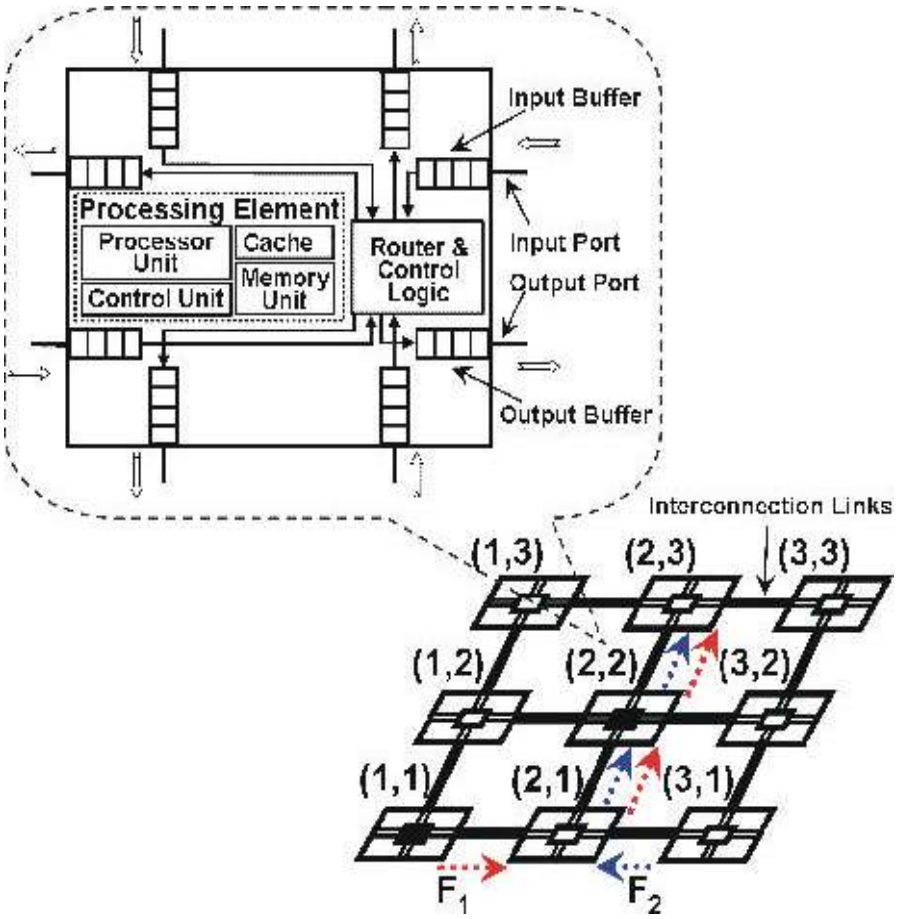

Fig. 2: A mesh based NoC node architecture consists of 4 bidirectional physical channels, each of them shared between k virtual channels

Energy and Power Consumption Analysis: Most of the consumed power in the switching element of an NoC node is due to switching logic. The average switching power consumed by the switching element in a cycle can be calculated by [17]:

$$P_{Ave} = N_s C_{Load} V_{dd}^2 F_{Clk} \quad (1)$$

Where N_s is the switching activity of the switching element, C_{Load} is the equivalent load capacitance of the switching element, V_{dd} is the supply voltage and F_{CIK} is the clock frequency.

Based on (1), total consumed power of the switching element of an NoC node can be reduced by scaling down each of the N_s, C_{Load}, V_{dd} or F_{CIK}.

N_s the switching activity, is a fixed value and is related to the traffic rate which in turn depends on the application and cannot be scaled to save power. C_{Load}, the average load capacitance of the switching element, is proportional to its transistor count, i.e., the more transistors are used in the switching element, the more load capacitance is produced. Circuit optimization techniques which try to reduce the transistor count of a circuit can be employed to decrease C_{Load}. But, circuit optimization techniques are hardly applicable in large designs such as the switching element of an NoC node.

Since most of power reduction techniques have direct effect on the performance [20, 25], it is recommended to use a metric which fairly considers both power consumption and delay at the same time. Power delay product, PDP, is a widely used parameter which simultaneously considers both power consumption and delay imposed to the system [19, 21, 22]. If a power reduction technique results into a lower PDP it can be inferred that the amount of power reduction of that technique is more than the amount of its imposed delay (performance overhead). In the other words, PDP discusses about energy (power multiplied by time) and lower PDP means lower energy consumption.

According to (1), power consumption of the switching element is linearly proportional to the operational frequency of the NoC. One can reduce the power consumption of an NoC by reducing its operational frequency. But frequency scaling has no improvement in PDP and consequently no energy saving is achieved.

Among the mentioned factors in (1), the supply voltage scaling, V_{dd} scaling, seems to be good solution especially in designs like NoCs in which lower energy consumption is more interested than lower power consumption. Although a system working with lower supply voltage has to work with lower operational frequency [22] which linearly increases the PDP

$$Delay \propto \frac{C_{Load}.V_{dd}}{(V_{dd}-V_{th})^{\alpha}}.$$

But it should be noted that V_{dd} scaling quadratically reduces the PDP and simultaneous scaling of V_{dd} and operational frequency reduce the PDP eventually.

Next section discusses about proposed method which is based on supply voltage scaling. This section reveals that in spite of the bad effect of supply voltage scaling on the operational frequency of the NoC, the proposed method improves the PDP of the switching element of the NoC and decreases consumed energy.

Proposed Method: As discussed in the previous section the voltage scaling is an efficient technique for energy saving. However, the voltage scaling can be done only if the system performance exceeds what is required by the used application. In the other word, only the extra performance gain with respect to what is needed by the application can be traded by the energy saving.

The proposed method is appropriate for application specific NoCs in which the network traffic rate can be estimated. In such NoCs, using extra virtual channels can increase the performance of the network for given specific traffic rate. Now by scaling down the supply voltage of the system, the network has still its initial performance while having lower energy consumption. To support this idea, a flit level VHDL-based simulator has been used and the network is simulated with different network traffic rates and different number of virtual channels. The simulator mimics the behavior of the XY routing algorithm. In each simulation experiment, a minimum of 20000 packets have been delivered and the average packet delivery time was calculated. Statistics gathering was inhibited for the first 1000 packets to avoid distortions due to startup transience. The average packet delivery time is defined as the average amount of time from the generation of a packet until the last data flit of that packet is delivered to its destination node. Packets are generated at each node according to a Poisson process with a fixed length of 32 flits and destination of each packet has been determined through a uniform random number generator. Figure 3 shows the average packet delivery time of a 4×4 mesh. It can be observed that increasing the number of virtual channels has increased the performance of the network which can be potentially traded to gain energy saving using voltage scaling. The bold line in Figure 3 represents the maximum acceptable packet delivery time in a typical application specific NoC (any other curve can be used as maximum acceptable latency). According to this figure, in a specific traffic rate, the number of virtual channels should be selected in a way that the average packet delivery time of the network does not exceed the maximum acceptable packet delivery time. In our proposed method another criteria is considered in order to choose the optimum number of virtual channels. This criterion is the amount of performance gain with respect to the maximum acceptable packet latency in a specific traffic rate, since more amount of performance gain gives the designers opportunity of scaling supply voltage more, hence having less energy consumption. For example in traffic rate of about 0.017 packet per cycle (Figure 3), the maximum acceptable packet latency is assumed to be about 310 cycles i.e., a packet generated by a source node should be delivered less than 310 cycles on average.

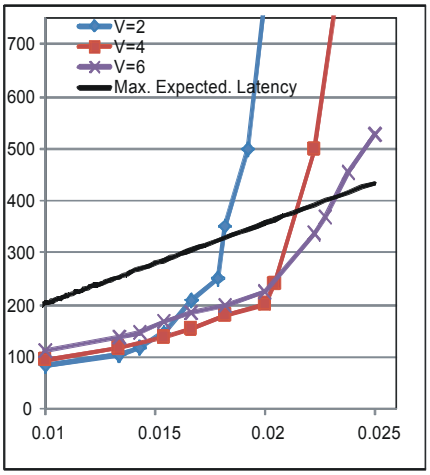

Fig. 3: Average packet delivery time of a 4×4 mesh NoC with 2, 4 and 6 virtual channels

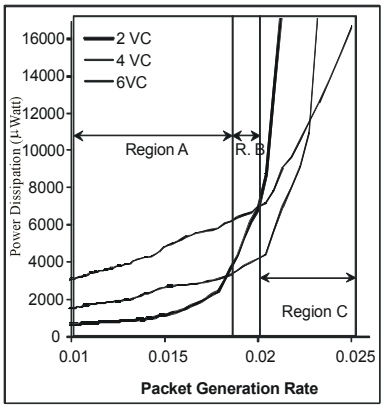

Fig. 4: Power consumption of a 4×4 mesh-based NoC node in different traffic rates with 2, 4 and six virtual channels per each physical channel

Now suppose that we use two virtual channels per each physical channel. In this configuration, the average packet delivery time of this NoC is about 250 cycles i.e., the supply voltage can be reduced by 25 percents (310/250). In the case of using four virtual channels per each physical channel, the average packet delivery time is about 200 cycles and the supply voltage can be reduced by about 35%.

To investigate the effect of using extra virtual channels in the proposed method, the Synopsys Power Compiler tool has been exploited and the energy consumption of the NoC with different number of virtual channels in presence of voltage scaling is measured and demonstrated in Figure 4. Power consumption simulations are done with the 160nm meter VLSI technology size. Library files of this technology are achieved from the predictive transistor model which is used by several other researchers.

Figure 4 has been classified in three regions i.e., region A, region B and region C. In each region, a specific number of virtual channels should be chosen with respect to the amount of energy consumption. For example in low traffic region i.e., region A, a switch with two virtual channels per each physical channel consumes less energy than a switch with four or six virtual channels. Therefore, in this region it is better to use two virtual channels. In medium traffic region i.e., region B, the use of a switch with two virtual channels make the NoC falling into saturation state (Figure 3) resulting in dramatically increase in energy consumption. In this region, the switch with four virtual channels is more efficient than a switch with six virtual channels from energy consumption criteria due to its more performance gain.

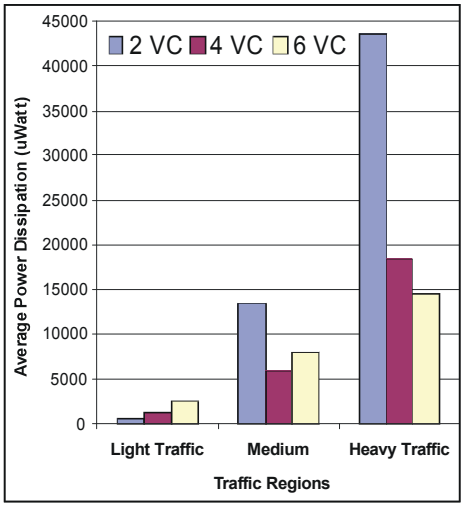

Fig. 5: Average PDP of each node in a 4×4 mesh-based NoC using voltage scaling in different traffic regions

Finally in high traffic region, region C, the switch with six virtual channels is the best candidate from both performance and energy consumption points of view.

Figure 5 shows the average PDP of the switching element of an NoC node in presence of voltage scaling. According to Figure 5, as traffic rate increases the use of extra virtual channels along with DVS technique can significantly reduce the PDP of the switching element. In this figure, the black un-marked line specifies the maximum acceptable latency in the network i.e., average message latency greater that this line cannot be accepted by the application.

CONCLUSIONS

An efficient power-performance exchange method is proposed in this paper for application specific NoCs. In this regard, performance improvement is achieved by the use of extra virtual channels in NoC switches i.e., the method relies on increasing the performance of an NoC in a given traffic rate by incorporating extra virtual channels. While, power reduction is achieved by the use of well-known voltage scaling technique. A wide range of simulations are done to evaluate the proposed method. The experimental results confirm that, the method can effectively reduce the energy consumption of the NoC while having no effect on performance.

REFERENCES

1. Guerrier, P. and A. Greiner, 2000. A generic architecture for on-chip packet-switched interconnections, In Proc. Design Automation and Test in Europe Conf. (DATE), pp: 250-256.
2. Benini, L. And G. De Micheli, 2002. Networks on chips: a new SoC paradigm, IEEE Computer, 35: 70-78.
3. Dally, W.J. and B. Towles, 2001. Route packets, not wires: on-chip interconnection networks, In Proc. Design Automatin Conf. (DAC), pp: 684-689.
4. Hemani, A., A. Jantsch, S. Kumar, A. Postula, J. Oberg, M. Millberg and D. Lindqvist, 2000. Network on a chip: an architecture for billion transistor era, In Proc. of the IEEE NorChip Conf., pp: 166-173.
5. Culler, D.E., J.P. Singh and A. Gupta, 1996. Parallel computer architecture, a hardware/software approach (second edition), Morgan Kaufmann.
6. Bertozzi, D., L. Benini and G. De Micheli, 2003. Energy-reliability trade-off for NoCs, in Networks on Chip, A. Jantsch and Hannu Tenhunen, eds., Dordrecht: Kluwer, pp: 107-129.
7. HU, J., 2005. Design Methodology for Application Specific Network-On-Chip, Ph.D Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University.
8. Jalabert, A., S. Murali, L. Benini and G. De Micheli, 2004. xPipesCompiler: a tool for instantiating application-speci_c NoCs, In Proc. Design Automation and Test in Europe Conf. (DATE).
9. Dall'Osso, M., G. Biccari, L. Giovannini, D. Bertozzi and L. Benini, 2003. xPipes: a latency insensitive parameterized network-on-chip architecture for multiprocessor socs, In Intl. Conf. on Computer Design, pp: 80-85.
10. Schmitz, M.T., B.M. Al-Hashimi and P. Eles, 2004. "System-Level Design Techniques for Energy-Efficient Embedded Systems", Kluwer Academic Publisher.
11. Burd, T.D., T.A. Pering, A.J. Stratakos and R.W. Brodersen, 2000. A dynamic voltage scaled microprocessor system, IEEE J. Solid- State Circuits, 35(11): 1571-1580.
12. Dally, W.J. and H. Aoki, 1993. Deadlock-free adaptive routing in multicomputer networks using virtual channels, IEEE Transaction on Parallel and Distributed System, 4(4): 466-475.
13. Dally, W.J. and B. Towles, 2001. Route packets, not wires: on-chip interconnection networks, In Proc. Design Automation Conf. (DAC), pp: 684-689.
14. Park, D., C. Nicopoulos, J. Kim, N. Vijaykrishnan and C.R. Das, 2006. Exploring Fault-Tolerant Network-on-Chip Architectures, International Conference on Dependable Systems and Networks (DSN), pp: 93.
15. Kumar, S., A. Jantsch, J.P. Soininen, M. Forsell, M. Millberg, J. Oberg, K. Tiensyrja and A. Hemani, 2002. A network on chip architecture and design methodology, IEEE Computer Society Annual Symposium on VLSI, pp: 105-112.
16. Duato, J., S. Yalamanchili and L. Ni, 2002. Interconnection Networks: An Engineering Approach, Morgan Kaufman Publication.
17. Dally, W.J. and J.W. Poulton, 1998. Digital Systems Engineering, CUP.
18. Park, D., C. Nicopoulos, J. Kim, N. Vijaykrishnan and C.R. Das, 2006. Exploring Fault-Tolerant Network-on-Chip Architectures, International Conference on Dependable Systems and Networks (DSN), pp: 93.
19. Dumitras, T., S. Kerner and R. Marculescu, 2003. Towards on-chip fault-tolerant communication, in Proc. of the Asia and South Pacific esign Automation Conference (ASP-DAC), pp: 225-232.
20. Kim, J., C. Nicopoulos and D. Park, 0000. A Gracefully Degrading and Energy-Efficient Modular Router Architecture for On-Chip Networks, 33rd International Symposium on Computer Architecture (ISCA'06), pp: 4-15.
21. Omana, M., D. Rossi and C. Metra, 2003. Novel transient fault hardened static latch, Proceedings International Test Conference, ITC pp: 886-892.
22. Rabaey J.M. and M. Pedram, 1996. Low Power Design Methodologies, Kluwer Academic Publishers.
23. Burd, T. and R.W. Brodersen, 1995. Energy ef?cient CMOS microprocessor design, Proc. 28th Hawaii International Conference on System Sci., 1: 288-297.
24. Asghari, S.A., H. Pedram, M. Khedemi and P.M. Yaghini, 2009. Designing and Implementation of a Network on Chip Router Based on Handshaking Communication Mechanis, in World Appl. Sci. J., 6(1): 88-93.