Journal of Zhejiang University SCIENCE A
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

A low-power high-throughput link splitting router for NoCs*

Mohsen SANEEI, Ali AFZALI-KUSHA, Zainalabedin NAVABI

(*Nanoelectronics Center of Excellence, School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran*)

E-mail: msaneei@ut.ac.ir; afzali@ut.ac.ir; navabi@cad.ece.ut.ac.ir

Received Dec. 12, 2007; revision accepted Apr. 24, 2008; CrossCheck deposited Nov. 16, 2008

Abstract: In this paper, we propose a technique for lowering the latency of the communication in a NoC (network on chip). The technique, which can support two qualities of service (QoS), i.e., the guaranteed throughput (GT) and best effort (BE), is based on splitting a wider link into narrower links to increase throughput and decrease latency in the NoC. In addition, to ease the synchronization and reduce the crosstalk, we use the 1-of-4 encoding for the smaller buses. The use of the encoding in the proposed NoC architecture considerably lowers the latency for both BE and GT packets. In addition, the bandwidth is increased while the power consumption of the links is reduced.

Key words: Low-power, Latency, Throughput, Network on chip (NoC), Delay-insensitive, Router
doi:10.1631/jzus.A0720117 **Document code:** A **CLC number:** TN402

INTRODUCTION

The integration of a SoC (system on chip) in near future will become very difficult due to the physical characteristics of nanoscale technologies. Global signal delays will span multiple clock cycles making some problems for the synchronization (Benini and de Micheli, 2002; Bertozzi and Benini, 2004). Signal integrity will also be compromised due to increased resistor-capacitor (RC) effects, inductance, and cross-coupling capacitances (Davis and Meindl, 2000). Power consumption, bandwidth, scalability and reusability are some other problems, most of which originate from long wire delays. A good solution to reduce these problems is the NoC (network on chip) where tiles of IP cores communicate with one another through a network-centric architecture realized using an interconnection network (Benini and de Micheli, 2002; Bertozzi and Benini, 2004). All the inter-core communication requests coming from different cores of the chip are performed using the interconnection network. Since the requests may have different latency requirements, different classes of packets may be defined. Packets which can tolerate high latency may be categorized in the best effort (BE) group while packets with a specific deadline may be grouped in the guaranteed throughput (GT) category.

In this paper, we propose an interconnection network architecture where wide links are split into two independent links. The links provide two separate paths for packets, enabling us to treat the GT and BE services differently using a simple priority mechanism for the links. Here, a higher priority is considered for the GT service and a lower priority for the BE service. In this scheme, not until resource requests for GT packets exist, will the transmission of BE packets take place. Along with the link splitting and the priority mechanism, the circuit switching scheme provides the GT service in the proposed architecture. In addition to splitting the link, we have used the 1-of-4 delay-insensitive encoding scheme to transmit 2 bits using 4 lines (Verhoeff, 1988). The scheme eases the synchronization (Nigussie *et al*., 2007), which could be a challenge for the communication between cores where the receiver and transmitter operate in different clock domains (Benini and de Micheli, 2002; Bertozzi and Benini, 2004). The use of the 1-of-4 encoding also leads to a lower crosstalk noise and less switching power in the link.

* Project supported by the Iranian National Science Foundation

The rest of the paper is organized as follows. Section 2 reviews some of the related works in the area of quality of service (QoS) and different services, while the delay of the 1-of-4 encoding scheme and its comparison with the parallel encoding technique are discussed in Section 3. We describe the implementation of the router architecture using the 1-of-4 encoding scheme in Section 4. The results of comparing the proposed scheme with two traditional schemes are presented in Section 5. Finally Section 6 concludes the paper.

RELATED WORKS

As mentioned in the previous section, one can categorize packets into BE and GT groups. An example of the former class is the packets used to transfer large data blocks, while an example of the latter could be the set of control signals exchanged between different cores (Santi *et al*., 2005). GT services are used for critical (e.g., real-time) communication, and BE services are used for noncritical communication (Rijpkema *et al*., 2003; Santi *et al*., 2005). As another example of diverse service, the requirements consider a system with a video processing core, which typically requires a lossless, in-order video stream with GT, but possibly allows corrupted samples. In this system, the cache update services, however, require uncorrupted, lossless and low-latency data transfer, but ordering and GT are less important (Rijpkema *et al*., 2003). To improve the QoS, the communication architecture should provide both BE and GT services. Some researchers have proposed architectures and related optimizations for on-chip interconnection networks with only the BE traffic class, but some other researchers have proposed architectures that support both BE and GT traffic classes. Nostrum (Millberg *et al*., 2004) and Æthereal (Vellanki *et al*., 2005) are two architectures that support both BE and GT traffic classes. Nostrum ensures the bandwidth for the GT traffic by reserving time slots called 'looped containers' for its transmission on inter-router links. If no GT traffic packet is injected into the network, the time slots are not utilized. Æthereal is a mesh-based NoC architecture that also supports the GT traffic by utilizing a centralized scheduler for the allocation of the link bandwidth. In (Vellanki *et al*., 2005), another NoC architecture that supports both BE and GT traffic classes is presented. It supports the GT traffic by reserving a certain number of virtual channels (buffers). In the synchronization area, globally asynchronous locally synchronous (GALS) and mesochronous clocking techniques are some of interesting schemes proposed in the literature (Saneei *et al*., 2006). The mesochronous scheme is typically used for the communication between two cores that have the same clock frequency but with two different phases. The problem with this communication protocol is metastability, which may occur if the sampling edge of the clock occurs when the input data are changing (Saneei *et al*., 2006).

DELAY-INSENSITIVE COMMUNICATION

A NoC system consists of many processing elements (blocks) which have different timing requirements and can operate at different clock frequencies. The communication between these blocks needs synchronization, which is error-prone. A viable solution for this is to use the GALS design approach where the communication between the (synchronous) processing blocks is performed asynchronously (Vellanki *et al*., 2005). Another possibility is to use a delay-insensitive encoding technique where the data are encoded using a one-hot scheme. As an example, one can consider the 1-of-4 encoding where 2 bits are transferred using 4 wires. The transition on each wire represents one out of 00, 01, 10, or 11. Several one-hot line groups can be combined to form a larger bus width. When the data are present on all line groups, a completion detection unit (normally implemented using a C-element) is used to generate an acknowledge signal on a common wire (Nigussie *et al*., 2007) and hence there is no need for matched delays (Santi *et al*., 2005). This makes the technique delay insensitive and the communication based on this technique robust. As a result, the communication has an average-case performance rather than the worst-case performance, which is the case in the communication based on timing constraints (Nigussie *et al*., 2007). Therefore, the circuitry will work properly independent of the wire and gate delays. In addition, the temperature, process, and supply voltage variations only affect the communication speed but

not the functionality. In this encoding scheme, the two neighboring wires do not change simultaneously and it eliminates any crosstalk delay and error. In addition, the power consumption is lower while the speed is higher. In this scheme, the number of transitions is independent of the actual data making the power usage predictable.

In this work, we use the 1-of-4 delay-insensitive encoding scheme in which the clock is injected into the data stream at the transmitter side and retrieved at the receiver side from all links for both GT and BE services. This solves the synchronization problem.

Delay model for a single wire

We can model an interconnect with a distributed RC model of a single wire and its driver. Using this model, the delay of the interconnect ($t_{p,\text{line}}$) and its pre-driver ($t_{p,\text{pre-driver}}$) can be modeled as

$$t_{p,\text{line}} = (0.693 r_o c_o + 0.693 R_w c_i + 0.377 R_w C_w) + 0.693 r_o (C_w + c_i)/h = A + B/h, \quad (1)$$

$$t_{p,\text{pre-driver}} = (N-1) t_{p0} (1 + \sqrt[N-1]{h}/\gamma), \quad (2)$$

where r_o, c_o, and c_i are the output resistance, the output capacitance, and the input capacitance of a minimum-size inverter, respectively; R_w and C_w are the resistance and capacitance of the interconnect line, respectively; h is the optimum size for the last stage of the driver; $\gamma=c_o/c_i$; t_{p0} is the delay of a minimum-size inverter without external load; N is the number of stages of the driver.

The total delay ($t_{p,\text{total}}$) is equal to the sum of the driver and line given by Eqs.(1) and (2). By equating the derivative of the total delay with respect to h to zero, one can obtain the optimal value of h as

$$h = \left(\gamma B / t_{p0}\right)^{(N-1)/N}, \quad (3)$$

which can be used to obtain the optimum (i.e., minimum) total delay ($t_{p,\text{optimum}}$) as

$$t_{p,\text{optimum}} = A + (N-1) t_{p0} + N \sqrt[N]{B (t_{p0}/\gamma)^{(N-1)}}. \quad (4)$$

To assess the accuracy of the above analytical model for obtaining the optimal delay, we compared the results of the model with those of HSPICE simulations for the interconnect lengths of 1~3 mm interconnects in 130, 90, 65, and 45 nm technologies, respectively (Im *et al*., 2005). The results for the optimum delay had errors less than 7% revealing a very good accuracy for the analytical model.

Delay model for parallel wires

Ignoring the inductance, an n-bit parallel bus with a single metal layer wire can be modeled as a distributed RC network with a coupling capacitance between the adjacent wires. The delay of the kth wire of the bus is obtained from (Kim *et al*., 2006)

$$T_l = \begin{cases} \tau_0[(1+\lambda)\Delta_1^2 - \lambda\Delta_1\Delta_2], & l = 1, \\ \tau_0[(1+2\lambda)\Delta_l^2 - \lambda\Delta_l(\Delta_{l-1} + \Delta_{l+1})], & 1 < l < n, \\ \tau_0[(1+\lambda)\Delta_n^2 - \lambda\Delta_n\Delta_{n-1}], & l = n, \end{cases} \quad (5)$$

where τ_0 is the delay of a crosstalk-free wire, λ is the ratio of the coupling capacitance (c_c) to the bulk capacitance (c_w), and Δ_k is the transition occurring on wire k [0 for the stable (or no) transition, +1 for the rising transition, and −1 for the falling transition].

Using Eq.(5), we can determine the clock period, T_c, for an n-bit parallel bus and a 1-of-4 encoded bus as (Kim *et al*., 2006)

$$T_c \geq \begin{cases} (1+4\lambda)t_0, & \text{for a parallel bus,} \\ (1+2\lambda)t_0, & \text{for a 1-of-4 encoded bus,} \end{cases} \quad (6)$$

where t_0 is the maximum value of T_k. For large values of λ, the clock frequency of the 1-of-4 encoded bus could be two times more than that of the parallel bus. Therefore, for the same pitch (space+width) and area for the parallel and 1-of-4 encoded buses, they can have the same bandwidth. To reduce λ and increase the throughput in the 1-of-4 encoded bus, we have doubled the pitch while keeping the same width for interconnects. This 1-of-4 encoded bus has the same area as an 8-bit parallel bus.

Table 1 shows the results of the analytical model for an 8-bit parallel and a 2-bit (4-line) 1-of-4 encoded buses for different technologies.

In the 8-bit parallel bus we need 8 line drivers with a higher capacitive load, while in the 1-of-4 bus we need only 4 line drivers with a lower capacitive load. Since the capacitance is reduced in the 1-of-4 encoding, the bandwidth and the power consumption

Table 1 Delay, energy per bit and bandwidth of parallel and 1-of-4 encoded bus

Bus type	Technology	Pitch (nm)	Width (nm)	Delay (ps)	Bandwidth (GB/s)	Energy per bit (fJ)	EDP* (ps.fJ)	EDP improvement (%)	Bandwidth improvement (%)	Energy improvement (%)
Parallel bus	130 nm	604	292	1038	7.7	54.0	56060			
	90 nm	428	205	642	12.5	58.5	37543			
	65 nm	304	145	731	10.9	59.7	43623			
	45 nm	216	103	1335	6.0	60.8	81245			
1-of-4 bus	130 nm	1208	292	218	9.2	27.4	5958	89.4	19.2	49.3
	90 nm	856	205	133	15.0	29.1	3872	89.7	20.6	50.3
	65 nm	608	145	150	13.3	29.0	4352	90.0	21.8	51.4
	45 nm	432	103	275	7.3	29.0	7968	90.2	21.3	52.4

* Energy-delay product

of the bus should be improved considerably. Based on the results given in Table 1, in the 1-of-4 encoded bus, the bandwidth increase is about 20%, while the power reduction is about 50%. These improvements are the results of the capacitance reduction and the crosstalk minimization in the 1-of-4 encoded bus.

PROPOSED ROUTER ARCHITECTURE

In this section, we describe the proposed low power router scheme, which is based on splitting a wider link into narrower links. The splitting is expected to improve the performance of the router. In this scheme, each router with 2*k*-bit links may be split into two independent *k*-bit links (Fig.1). The proposed router architecture, which is based on the 1-of-4 encoded bus (referred to as 'X2 scheme'), is shown in Fig.2a. The use of two independent links allows a better support for GT and BE services.

The router has five ports with two 1-of-4 encoded links in every port, one arbiter, and one crossbar switch. Every packet is split into some 32-bit flits. The first flit, which is the header flit, is decoded into binary and saved in a 32-bit register. Then, the 'router logic' determines a proper path of the header and sends a request to the arbiter, where a proper output link for the packet is assigned. If the output link is ready, the header will be sent to the output link after being encoded to the 1-of-4 code again. When the total path from the source to the destination is determined, the other flits will be transferred through a second path (dashed path in Fig.2b) without being decoded into binary. Note that the changing path from

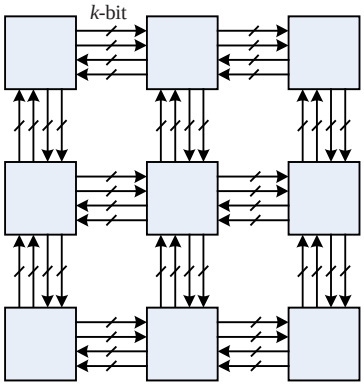

Fig.1 A 3×3 NoC with the link splitting scheme. Each link has *k* bits and the IP cores are not shown

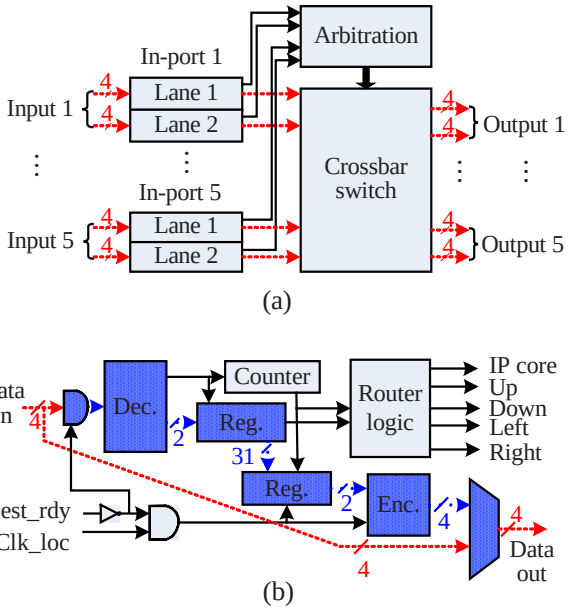

Fig.2 (a) Router architecture based on the 1-of-4 encoded bus (X2 scheme); (b) Logic of one lane

the initial path (dotted path) to the dashed path generates some additional transitions (glitches). To eliminate these transitions, after the header we send another flit which has only one '1' in the most significant bit (MSB). We call this flit the 'alignment flit'. All other flits are payloads which carry data.

There are no specific limitations on the payload length except for the latency. In our design, we can use any source-based routing algorithm in which the entire path is determined by the source. The path information travels through the path routers by the header. In our simulation, we have used the *XY*-routing algorithm. Fig.3 shows the format of the header where two first least significant bit (LSB) bits of the header have the global direction information (00: northeast; 01: northwest; 10: southeast; 11: southwest) and bits 2 to k+1 determine the *X* or *Y* local direction information (0: *X* direction; 1: *Y* direction). Here, k is the number of links between the source and the destination. Finally, the (k+2)th bit is 1 and all other header bits are 0. For example, when the source core is at the coordinates of (1, 2) and the destination core is at the coordinates of (6, 6), the header is '0000000000000000000110011010000'. An example of a routing path for this packet is shown in Fig.4. In every router, one bit of the local direction information bits is consumed and the content of the header flit (except for the two LSB bits) are shifted right one bit. Therefore, in the destination router, bits from 3 to 31 will be 0 while bit 2 will be 1. When this occurs, the received packet is routed to the IP core.

In the proposed NoC architecture, in every port we use two independent 1-of-4 encoded links for transmitting the data which could be equivalent of using one physical link but with two virtual channels. The advantage of the proposed approach is the elimination of the virtual channel (VC) buffers, which are very expensive in terms of the area and power, from the router. These two links can have the same or different routing and/or arbitration logics and each can transfer one GT or BE packet. The latency of the GT packets, which have more priority compared to the BE packets, can be reduced. In addition, as discussed in the previous section, the use of the 1-of-4 encoded bus can reduce the power consumption and increase the bandwidth of the links (see Table 1). Another advantage of the proposed architecture is the crosstalk error reduction. This has been achieved through tripling the spacing between the wires and the fact that two neighboring wires do not change simultaneously in the 1-of-4 encoded links.

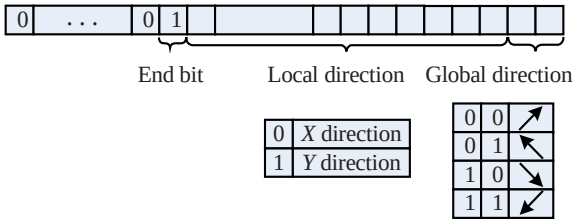

Fig.3 Format of header of packets in the proposed architecture

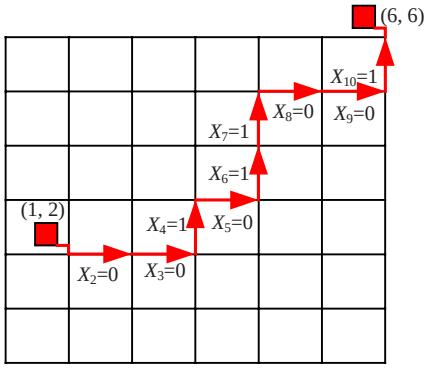

Fig.4 Routing path when the header is '0000000000 00000000000110011010000'

RESULTS AND DISCUSSION

To evaluate the proposed scheme, we have designed the following three different routers:

(1) X1 router: A 5-port router with one 1-of-4 link in every port;

(2) X2 router: A 5-port router with two independent 1-of-4 links in every port (the proposed scheme in this work);

(3) X1_4bit router: It is similar to the X2 router except that the two 1-of-4 links are dependent and are used jointly to transfer the same data. In other words, the X1_4bit router is similar to the X1 router but the former has wider links (every link has two 1-of-4 encoded buses).

We evaluate these routers in a 4×4 NoC with random destination address packets. In every source (packet generator), we consider a 32-kB buffer which is split into two parts—GT and BE. All packets in the sources will enter the corresponding part of the buffer. These packets transfer to the destination in order but

the GT packets have a higher priority. In the NoC with the X2 router, there are two buffers in every source (one buffer for every link) and the bandwidth of every source is twice of the bandwidth of the NoC with the X1 router. In the NoC with the X1_4bit router, the size of the buffer and the operating frequency are similar to those of the X2 router. The simulations have been repeated several times with different initial seeds in the destination, i.e., the random address generator. The results presented here are the average values of these simulation outputs.

To determine the operating frequency of the routers, we used the HSPICE simulations for a 130-nm technology. The results show that an input port with the 1-of-4 encoded link can operate at frequencies of more than 4 GHz, which yields a bandwidth of more than 1 GB/s for every link. Based on these results, we used 1, 2 and 2 GB/s for the X1, X2 and X1_4bit sources respectively when the packet injection rate into the network is 100%. The frequency of the X1 and X2 routers are assumed to be 250 MHz while the frequency of the X1_4bit router is assumed to be 500 MHz.

The three NoCs were designed using VHLD and simulated under uniform traffic patterns with different packet injection rates and GT to BE ratios. Fig.5 shows the results for 10% and 50% ratios of GT to BE packets when the packet injection rate varies from 10% to 50%. The results show that latencies of the GT packets in the X2 NoC are 27% (resp. 30%) to 47% (resp. 52%) lower than those in the X1 (resp. X1_4bit) NoC when the GT to BE ratio is 10%. These latency improvements are 27% (resp. 30%) to 56% (resp. 59%) when the GT to BE ratio is 50%. The latency improvement for the BE packets in the X2 NoC compared to those in the X1 (resp. X1_4bit) NoC is 32% (resp. 38%) to 94% (resp. 98%) when the GT to BE ratio is equal to 10%, and is 27% (resp. 36%) to 96% (resp. 98%) when the GT to BE ratio is equal to 50%. The bandwidths of the networks for a 50% packet injection rate are 15.5, 6.6 and 12.6 GB/s in the X2, X1 and X1_4bit NoCs, respectively. As the results suggest, the proposed scheme has improved the network bandwidth up to 135% and 23% compared to the X1 and X1_4bit schemes, respectively.

Fig.6 shows the bandwidth of the source vs. the packet injection rate for the three schemes. The bandwidth in the X1_4bit NoC is twice more than that

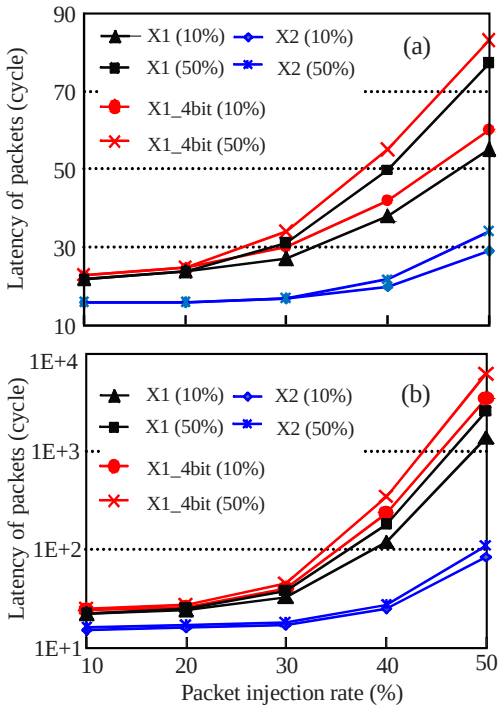

Fig.5 Latency of GT (a) and BE (b) packets for 10%~50% packet injection rate of three different routers when the rate of GT packets is 10% and 50%, respectively

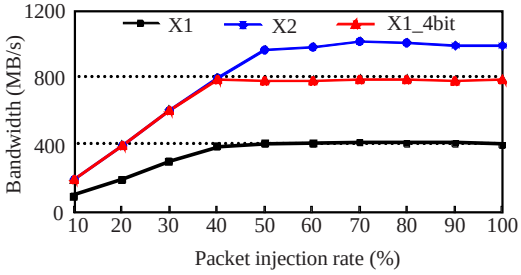

Fig.6 Bandwidth of every core (source) for packet injection rate 10% to 100%

in the X1 scheme. The reason is that the links of the X1_4bit router are two times wider than those of the X1 router. The bandwidth of the source in the X2 case is about 25% more than that in the X1_4bit router when the network saturates. This improvement originates from existence of the two independent links in the X2 router. This diminishes the effect of the congestion present when the injection rate increases.

The latencies of the GT and BE packets in the schemes as a function of the packet length are depicted in Fig.7, where the horizontal axis (packet length) has logarithmic scale with the base of 1.5. The

packet length changes from 5 to 432 flits. As the results show for both GT and BE packets, the X2 router has the lowest latency and the X1_4bit router has the highest latency. In all the schemes, the latency of the GT (resp. BE) packets increases (resp. decreases) when the packet length becomes larger.

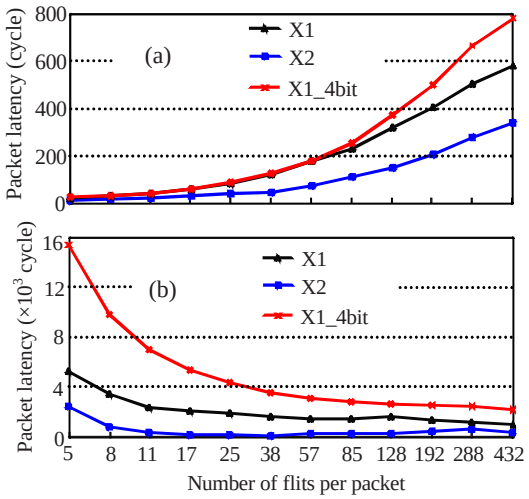

Fig.7 Latency of GT (a) and BE (b) packets of three different routers vs the packet length when the packet injection rate is 50% and the rate of GT packets is 30%

CONCLUSION

In this paper, we proposed an architecture for NoCs which had lower latency, crosstalk, and power for the link and router as well as higher bandwidth. The improvements were achieved by splitting a wider link into two thinner links and using a 1-of-4 encoding scheme which also resolved the synchronization problem. The results showed that if instead of transmitting data via a wide link, by using two thinner links (e.g., two independent 1-of-4 encoded links instead of a link with two dependent 1-of-4 encoded links), the latencies of the GT and BE packets can be reduced while the bandwidth of the network can be increased. Both VHDL and HSPICE simulations were used to assess the performance of the schemes. We also presented a model for delay of the link and driver with the optimum size and showed that the 1-of-4 encoding can improve the bandwidth by about 20% and power of links by about 50%. The results also showed that the new architecture could improve the QoS parameters (latency and bandwidth) of NoCs.

References

Benini, L., de Micheli, G., 2002. Network on chips: a new SoC paradigm. *IEEE Computer*, **35**(1):70-78. [doi:10.1109/2.976921]

Bertozzi, D., Benini, L., 2004. Xpipes: a network-on-chip architecture for gigascale systems-on-chip. *IEEE Circuts Syst. Mag.*, **4**(2):18-31. [doi:10.1109/MCAS.2004.1330747]

Davis, J., Meindl, D., 2000. Compact distributed RLC interconnect models—Part II: coupled line transient expressions and peak crosstalk in multilevel networks. *IEEE Trans. on Electr. Dev.*, **47**(11):2078-2087. [doi:10.1109/16.877169]

Im, S., Srivastava, N., Banerjee, K., Goodson, K.E., 2005. Scaling analysis of multilevel interconnect temperatures for high-performance ICs. *IEEE Trans. on Electr. Dev.*, **52**(12):2710-2719. [doi:10.1109/TED.2005.859612]

Kim, M., Kim, D., Sobelman, G.E., 2006. Network-on-Chip Link Analysis under Power and Performance Constraints. Proc. Design Automation and Test in Europe Conf. and Exhibition, Island of Kos, Greece, p.4163-4166. [doi:10.1109/ISCAS.2006.1693546]

Millberg, M., Nilsson, E., Thid, R., Jantsch, A., 2004. Guaranteed Bandwidth Using Looped Containers in Temporally Disjoint Networks within the Nostrum Network on Chip. Proc. Conf. on Design, Automation and Test in Europe, Paris, France, p.890-895. [doi:10.1109/DATE.2004.1269001]

Nigussie, E., Lehtonen, T., Tuuna, S., Plosila, J., Isoaho, J., 2007. High-performance long NoC link using delay-insensitive current-mode signaling. *VLSI Design*, p.1-13. [doi:10.1155/2007/46514]

Rijpkema, E., Goossens, K., Radulescu, A., Dielissen, J., Meerbergen, J.V., Wielage, P., Waterlander, E., 2003. Trade-offs in the design of a router with both guaranteed and best-effort services for networks on chip. *IEE Proc.-Comput. Digit. Tech.*, **150**(5):294-302. [doi:10.1049/ip-cdt:20030830]

Saneei, M., Afzali-Kusha, A., Navabi, Z., 2006. A Mesochronous Technique for Communication in Network on Chips. Proc. 18th Int. Conf. on Microelectronics, Saudi Arabia, p.32-35. [doi:10.1109/ICM.2006.373260]

Santi, S., Lin, B., Kocarev, L., Maggio, G.M., Rovatti, R., Setti, G., 2005. On the Impact of Traffic Statistics on Quality of Service for Networks on Chip. IEEE Int. Symp. on Circuits and Systems, Kobe, Japan, p.2349-2352. [doi:10.1109/ISCAS.2005.1465096]

Vellanki, P., Banerjee, N., Chatha, K.S., 2005. Quality-of-service and error control techniques for mesh-based network-on-chip architectures. *Integr., VLSI J.*, **38**(3):353-382. [doi:10.1016/j.vlsi.2004.07.009]

Verhoeff, T., 1988. Delay-insensitive codes—an overview. *Distrib. Comput.*, **3**(1):1-8. [doi:10.1007/BF01788562]