Dynamic NoC Platform for Varied Application Needs

Sidhartha Sankar Rout[1], Hemanta Kumar Mondal[2], Rohan Juneja[1], Sri Harsha Gade[1], Sujay Deb[1]
[1]Indraprastha Institute of Information Technology Delhi, India
[2]University of Southern Brittany, Lab STICC, Lorient, France
E-mail: [1]{sidharthas, rohan14156, harshag, sdeb}@iiitd.ac.in, [2]hemanta.kumar-mondal@univ-ubs.fr

Abstract

Many-core processing platforms are gaining significant interest for a wide range of applications, viz., Internet of Things (IoT), consumer electronics, single-chip cloud computers, supercomputers, defense applications etc. Networks-on-Chip (NoCs) are accepted as the communication backbone for these many-core platforms. However, energy consumption in NoC components still remains considerably high. Specifically for large systems with many nodes in the network, a significant amount of energy is consumed by the communication infrastructure. The usage of the routers and resources associated with it are application dependent and for most applications performance requirements can be met without operating the whole communication infrastructure to its maximum limit. Dynamic reconfigurable system that can switch between both high performance and low power modes will be able to exploit the variable workload conditions provided by different applications. Among all the NoC components, Virtual Channels (VCs) are the most power hungry modules. This paper proposes a dynamic NoC platform (DNoC) that optimizes VC utilization for different applications using a smart router architecture. Power Management Controller (PMC) along with Utilization Computation Unit (UCU) controls and predicts the number of active VCs to achieve the required performance with minimum overhead. In our experiments the proposed solution provides 83.3% power benefit (best case scenario) with negligible throughput penalty compared to a baseline mesh router.

Keywords

Wide-range applications, network-on-chip, virtual channels, energy efficiency

1. Introduction

High performance computing demand is no more restricted to scientific applications only. Emergence of smart sectors like Internet of Things has propelled the computing demand significantly in multiple domains. Internet of Things, a revolutionary new paradigm refers to interconnection of physical devices and meaningful communication among them. With billions of physical devices interconnected to each other communicating continuously, huge amount of data is expected to be transferred, stored, analyzed and computed [1]. Data centers and servers involved are equipped with many-core processing units which analyze the data, perform arithmetic and logical operations on them, and take decisions based on the results for multiple applications. Since the applications are quite diverse, the demand on compute platform will vary significantly. As it is not feasible to have customized solutions for all different applications, a platform that can be easily modified to suit particular application demands will be highly desirable. A platform based solution to sustain and provide scalability to this trend is proposed in this paper.

Many-core processing systems require an efficient communication infrastructure to cater to the high throughput, while maintaining high performance. Network-on-Chip is accepted as the preferred interconnect solution owing to its scalability and the effectiveness to connect different cores in the many-core architectures [2]. To improve the performance of on-chip long distance communication, Wireless NoC (WNoC) has emerged as a radical solution [3, 4]. Fault tolerance and better load balancing properties of NoC make it a better communication infrastructure than other traditional interconnect networks. Despite of its various advantages, power consumption in NoC has been a major limiter [5]. Several research works have been carried out to deal with NoC power issue [6-10]. Virtual Channels claim a significant portion of the total NoC power [11] but are essential for high performance of the interconnection network. This necessitates an efficient VC management solution and therefore, we propose an optimal VC utilization method for NoC platform.

This paper deals with an efficient communication model for multi core systems suitable for varied application needs. Given the emergence of data science centric applications like IoT and big data exploration and the significant importance of virtual channels in NoC architectures, we thereby introduce an on-chip interconnection network model DNoC: a smart VC utilization scheme. This significantly improves the performance by providing optimum number of VC resources according to the network traffic pattern and thereby saving power.

The main contributions of this work are summarized as follows:

1. A dynamic NoC: DNoC that switches between a wide-range of applications with different requirements employing optimal VC utilization.
2. We used drowsy and power-gating schemes for VCs to improve NoC energy efficiency.
3. Dynamic load monitoring is achieved using the Utilization Computation Unit which provides the requisite information to the Power Management Controller for optimal operating conditions.

The remainder of this paper is organized into following sections. Section 2 describes related work and insight into our motivation to pursue this work. The proposed router architecture and power saving techniques have been

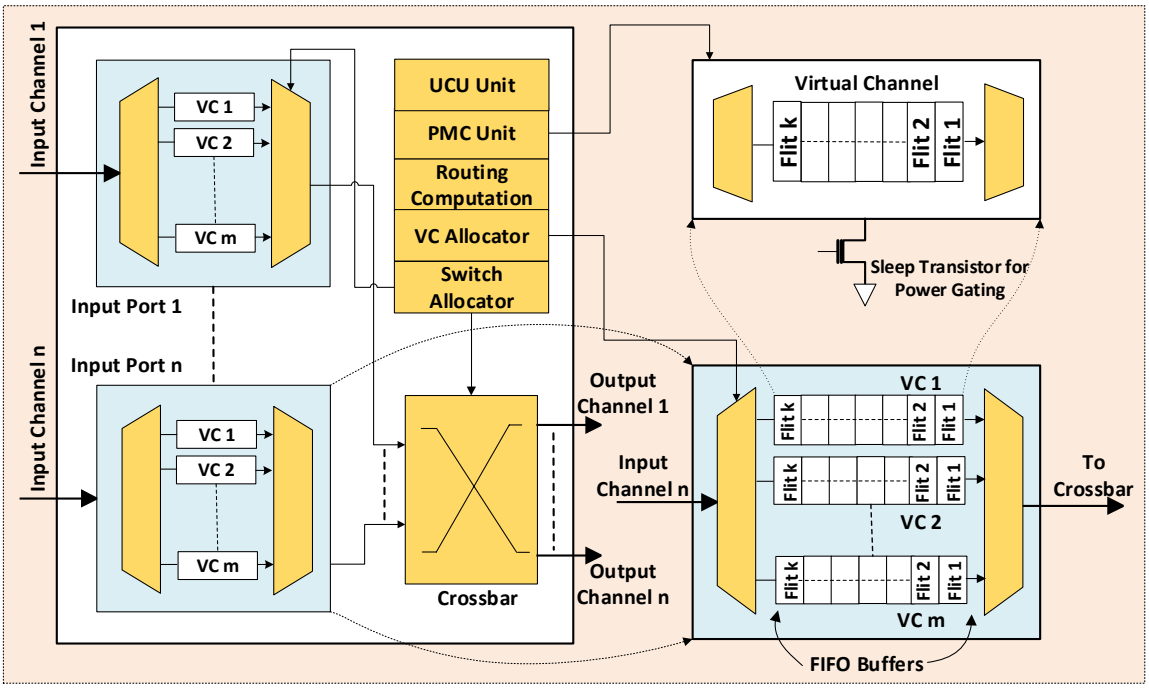

Figure 1: Block Diagram of Proposed DNoC Router

presented in section 3. Performance of DNoC router has been evaluated in section 4. This section also shows the power measurement and simulation results of the implemented system. Section 5 concludes this work.

2. Related work

Buffer/VC power consumption is a key issue in NoC routers which needs to be taken care. To deal with this issue, researchers have come up with bufferless (BLESS) routing on-chip network [11]. But this introduces extra latency and reduced bandwidth, which make it suitable only for low traffic network. It has been shown in research [12] that for high injection rate operations, NoC router architecture should be designed with more number of virtual channels whereas less number of VCs should be provided for low injection rates. Rate of injection varies with application and therefore in [13] a system-level algorithm has been presented which would allocate application-specific buffer space for NoC. But this is not a customized solution for simultaneous multiple applications.

In their research, people have exploited the delay requirements of traffic [14] and developed an active buffer sizing algorithm for low power NoC based on network calculus. But varied application needs in areas like IoT where multiple systems are live continuously with different traffic conditions actually demand a dynamic management of buffer resources of the communication infrastructure for power efficient solution. The idea of dynamically allocating VC resources based on the traffic conditions was presented by [15] which supports fully adaptive routing by dynamically and efficiently maintaining queue and network resources. An implementation in [16] deals with dynamic allocation of VC resources through a control logic, but it allows a VC to be consumed by a single packet only.

In this work we have proposed a runtime dynamic VC management mechanism DNoC which is a customizable power and performance efficient solution for varied traffic demands. We are using a UCU unit in the proposed router to estimate the router utilization based on the load conditions. A PMC unit is used to control the power saving schemes on the VCs of the router.

3. Proposed power-efficient NoC

This section gives the implementation details of the proposed router and the power saving schemes used. The physical channels at each input port of the router are provided with multiple virtual channels to maintain maximum efficiency during heavy workload condition. Figure 1 illustrates the block diagram of proposed DNoC router. Basic working of the proposed router is presented as follows. The incoming message flits through the input channel gets stored in the virtual channels before being routed to the desired output port. Routing Computation (RC) unit decides the next destination router for the packet and assigns the output port whereas VC allocator (VA) assigns an output virtual channel by arbitrating the requests from each input packet competing for the output channel. Switch Allocator (SA) assigns switch time slots to the input flits from different input ports bidding for the connection to the output ports and thereby controls the state of the crossbar switch. Crossbar unit provides the physical connection between the input and output channels of the router. A Utilization Computation Unit observes activity of the router and computes runtime utilization of the neighboring routers based on the traffic condition. This utilization information is provided to the Power Management Controller unit of the downstream router. PMC reads necessary information about

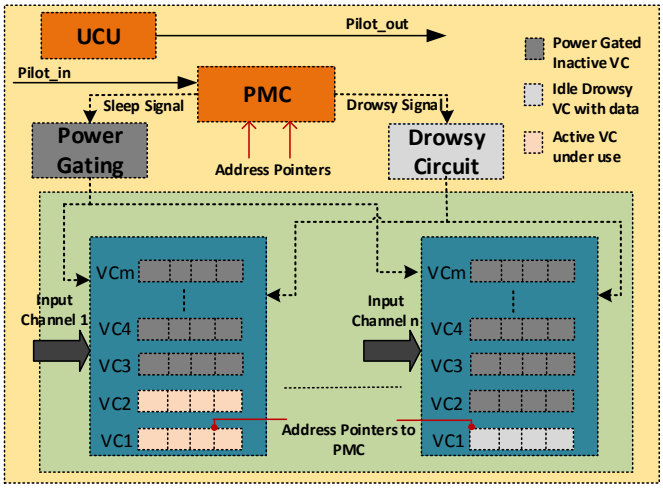

Figure 2: DNoC Low Power Management at Router level

the status of router utilization and generates different operating voltage levels based on that.

The core of DNoC is composed of two parts: (1) VC Utilization Estimation (2) Power Management Controller.

3.1. VC utilization estimation

The number of VCs per physical channel and VCs depth are two important parameters that interplay the buffer utilization, throughput, latency and energy consumption. Since buffer resources come at a premium, efficient management of it is desired. Fixed buffer structures will either be underutilized or will underperform at certain traffic conditions. DNoC architecture is proposed to best serve the varied workload demand. Different applications need different traffic condition and thereby claim different performance efficiency requirement. Routers in DNoC platform are provided with multiple VCs per channel to satisfy the maximum demanding application. But at the other extreme, tasks who claim less traffic are provided with few active VCs out of the multiple available ones. An efficient use of VC resources is done by calculating the runtime traffic demand for varied applications in terms of router utilization. Utilization Computation Unit in each router computes runtime utilization for all neighboring routers that are at one hop distance from it [17]. The estimation of router utilization is done periodically on an epoch-to-epoch basis. A fixed duration epoch is empirically set at 1K cycle for the set of benchmarks used in our simulation. At the start of an epoch, UCU in each router processes the header flits that are queued in its input virtual channels through the header decoder. Thus it determines the number of packets that will be routed to each output port of the router and thereby the utilization estimation of the routers connected to the output ports is done. Once UCU completes its operation, it transfers the estimated utilization information to the corresponding neighbor routers through the *Pilot_out* signal as shown in Figure 2.

3.2. Power management controller

With the drastic reduction of transistor feature size in the deep submicron era, static power dissipation has increased dramatically. At the same time gigahertz switching elevates the dynamic power dissipation. Virtual Channels being one of the major power consuming component among all the NoC resources [11], proper power management techniques need to be incorporated with the VCs to lower the overall power consumption. DNoC architecture supports multiple VCs per input port out of which variable number of channels can be active simultaneously depending on the application need. For our simulation, we have kept 8 VCs per channel with each virtual channel depth of 4 flits. In the proposed architecture, usage of VCs is categorized into three different types based on traffic load condition (i) VC is full and active when it is occupied with flits and an active packet transfer is going on, (ii) VC is full and idle when message flits are waiting in the VC due to congestion in the next destination router, and (iii) VC is empty when there are no packets on the input channel. According to the usage conditions, a VC can be in one of the three different states; Active, Drowsy and Power Gated (PG). When a VC is involved in packet transfer, the same will be in active state with a supply of normal operating voltage. Idle VCs with flits in it are provided with a lower voltage level which would ensure no data loss and thereby will be in drowsy state. Empty VCs are power gated by disconnecting the supply voltage from them. All the VC states are controlled by a Power Management Controller as shown in Figure 2. Based on the router utilization information, the PMC unit generates different voltage levels and drives the VCs to one of the three states.

PMC receives the estimated router utilization information for all of its input channels through the *Pilot_in* signals. Utilization information are provided by the

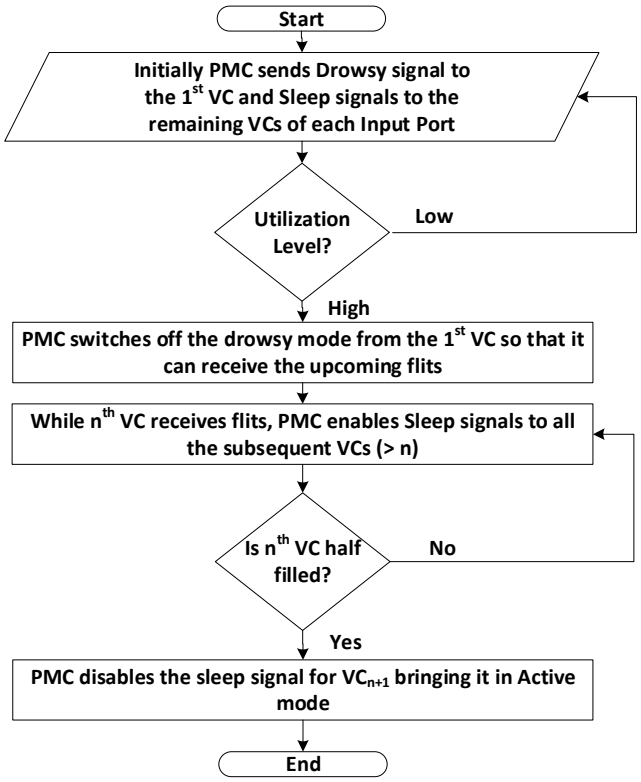

Figure 3: Work flow of the PMC unit in a Router

Pilot_out signals from the upstream neighboring routers. Initially PMC provides sleep signals to all the virtual channels and drives them to PG state to save the maximum possible power. At the same time to reduce the wake-up penalty and avoid any data loss, the controller sends drowsy signal to the first virtual channel of each port. This is shown for the input channel *n* in Figure 2. The utilization information on the *Pilot_in* signal instructs the PMC when to remove or activate the drowsy signal. PMC removes the drowsy signal from the first VC of a particular input port, when the router utilization information indicates that there are packets in the upstream router to be routed to the same input port. Once a particular port of a router starts getting utilized, the address pointers shown in Figure 2 indicate the exact location till which the VC is occupied. When the address pointer indicates the first buffer to be half/more filled, PMC removes the sleep signal from the second VC to save the wakeup time of the subsequent one. This enables the second VC to receive the upcoming flits as illustrated for the input channel 1 in the Figure 2. Work flow of the Power Management Controller in the proposed router is presented in Figure 3. Next two sub-sections explain the two power saving mechanisms used in DNoC design.

3.2.1. Power gated VCs

Whenever a VC is completely empty, it is unutilized and PMC sends the sleep signal to drive it to PG state. VCs in DNoC router are connected to sleep transistors which couple them to ground rail. When PMC sends a sleep signal, the corresponding sleep transistor gets into cut-off region and thereby disconnects the associated VC from the ground terminal. This saves a significant amount of leakage power. Depending on channel occupancy, the address pointer may invoke the controller to remove the sleep signal and bring the subsequent VC to active state.

3.2.2. Drowsy VCs

When a VC holds message flits, but no active communication is going on because of the congestion in subsequent routers, PMC sends a lower level supply voltage to the VC to reduce the power consumption. The retention voltage that ensures no loss of stored data in VCs is applied and the scheme is called as drowsy scheme. In our implementation, D-flip-flop (DFF) based buffer is used to design the VCs. From our experiment, we estimated the data retention voltage for DFF is 0.63V at 32nm technology node. Consequently, the circuit is operated under two level of voltages; 0.63 (drowsy state) and 1V (active state).

DNoC's architecture power gates the VCs while unutilized. Under low traffic condition, only a limited number of VCs are activated and utilized. This way the total energy dissipation in the VCs for the whole network reduces drastically. With the rise in traffic, more VCs are activated dynamically to handle the increased load. This proves DNoC as a real time manager of buffer resources which provides power efficient solution for variable load condition without having any impact on the performance of the system.

4. Simulation results

In this section, we discuss the performance of DNoC by injecting different synthetic traffic patterns to a 16x16 mesh system. Simulations are carried out for variable load conditions and the corresponding results are presented here. Power consumption of the router is measured for 32nm technology implementation. Simulation setup is presented in Table 1.

Table 1: Simulation setup

Topology	16x16 Mesh
Routing	XY
Flit Size	32 bits
Packet Size	2 flits
Clock Frequency	1 GHz
Workload	Synthetic

4.1. Simulation platform

A cycle accurate network simulator Noxim [18] is used and several synthetic traffic patterns are injected to evaluate the system performance. The evaluation is performed for *bitrevarsal*, *butterfly*, *random*, *shuffle*, *transpose1*, and *transpose2* traffic patterns. System Performance is observed for 1, 2, 4 and 8 VCs. DNoC is evaluated in terms of the throughput, and average latency. Leakage as well as dynamic power consumption are measured and compared for both the baseline mesh router and the proposed router using DSENT

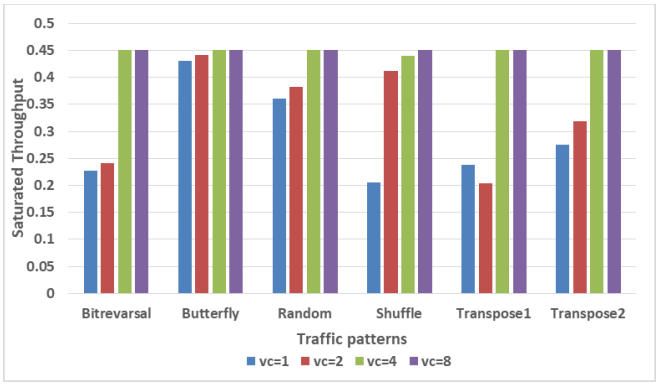

Figure 4: Saturated throughput vs traffic patterns for different number of VCs (16x16 system size)

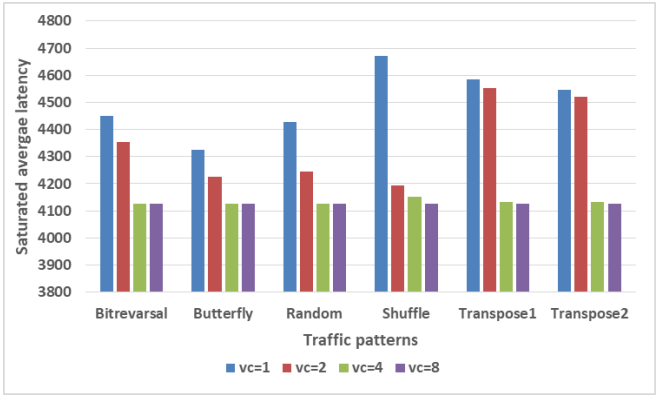

Figure 5: Saturated average latency vs traffic patterns for different number of VCs (16x16 system size)

tool [19]. Power overhead associated with power gating sleep transistors, UCU unit and PMC unit are also measured.

4.2. Throughput and latency evaluation for varied number of VCs

A 16x16 mesh based system is simulated with different synthetic benchmarks and the throughputs as well as latencies are observed for variable numbers of VCs per input port. The results for throughput and latencies with respect to different traffic patterns are shown in Figure 4 and Figure 5 respectively. Simulation result shows that for all the traffic patterns considered, 4 VCs provide equivalent throughput in comparison to 8 VCs. So, here in most of the cases only 4 VCs per router port performs with near equal efficiency as 8 VCs. The remaining unused VCs are power gated to save power. The result also presents that for the traffic pattern *shuffle*, 2 VCs and for *butterfly*, 1 VC serve the purpose with very less degradation in throughput and latency. This leaves us with more number of unused VCs which are again power gated to save considerable amount of power with a little impact in throughput and latency.

4.3. Power calculation for varied number of VCs

In the previous sub-section we find that for several applications we can achieve desired throughput with 4 VCs, sometime 2 VCs and even for few application 1 VC would do the task. So power measurements for the proposed DNoC router are done for 1, 2, 4 as well as 8VCs active along with the consideration of overhead introduced due to the extra hardware used. Similar power measurements are done for a baseline mesh router with 8 VCs per port for the comparison purpose.

For a baseline mesh router having 8 VCs per port shows channel power consumption of 24.6mW per port. In our proposed router the PMC unit adds 0.267mW whereas the UCU unit introduces 0.043mW of power overhead. Sleep transistors introduced in the design for power gating, consumes 0.712mW of extra power per port with 8 VCs. Table 2 shows the power consumption per port for both baseline mesh and proposed router with variable numbers of VCs active at a time.

Table 2: Power comparison when router port is active

Router Architecture	Power consumption for different number of Active VCs per router port in mW			
	1 VC	*2 VCs*	*4 VCs*	*8 VCs*
Baseline Mesh Router	24.6	24.6	24.6	24.6
Proposed Router	4.10	7.18	13.3	25.7

Form the observed data in Table 2 we realize that for applications not requiring all the VCs for packet communication can reduce power consumption drastically at individual router level. Calculation shows that with 4 VCs active, the proposed router saves upto 46% of total power per port in comparison to the baseline router. For the traffic conditions which need only 2 or 1 active VC, DNoC provides power saving upto 71% and 83.3% respectively. With all the VCs in active mode in few cases, the proposed router consumes 4.47% extra power than the baseline architecture due to the power consumed in the additional hardware integrated.

Table 2 holds the power consumption data when the router port is engaged in packet transfer. The proposed router can save even more power while the router port is inactive. In case of baseline router all the VCs dissipate leakage energy during inactive mode, a total of 24.6mW for a router port with 8 VCs. Whereas in this mode our proposed router would consume 3.17mW including the power overhead for extra hardware.

Table 3 illustrates the optimum number of VCs used in the proposed router for all the benchmarks considered along with power benefit, throughput and latency penalty in comparison to the baseline router.

Table 3: Power benefit, latency and throughput penalty for different workloads with the proposed router

Synthetic Workloads	Optimum number of VCs	Power Benefit in %	Throughput Penalty in %	Latency Penalty in %
Bitrevarsal	4	46	0	0.048
Butterfly	1	83.3	4.2	4.84
Random	4	46	0	0
Shuffle	2	71	8.33	1.63
Transpose1	4	46	0	0.18
Transpose2	4	46	0	0.18

Table 3 shows that the proposed router enables the use of optimum number of VCs required for an application. Though the routers have 8 VCs to handle high load demanding applications, the benchmarks we used are served by 4 or less number of VCs. *Butterfly* and *Shuffle* need 1 and 2 VCs respectively and thereby incurring a power benefit of 83.3% and 71% respectively. These power benefits come at a cost of negligible throughput and latency penalty.

5. Conclusion

In this work we have proposed DNoC, a dynamic reconfigurable platform with smart router architecture which can allocate optimum number of VCs to the input channels depending on the traffic densities. DNoC can switch between both high performance and low power modes by exploiting the variable workload conditions provided by different applications. The proposed architecture is suitable for varied application needs like IoTs and has been proved to be a power and performance efficient solution. Simulation shows a power benefit of 83.3% with only 4.2% of throughput penalty with DNoC in comparison to a baseline mesh router for a best case scenario.

6. References

[1] D. Miorandi, S. Sicari, F. De Pellegrini, and I. Chlamtac, “Internet of things: Vision, applications and research challenges,” Ad Hoc Netw., vol. 10, no. 7, pp. 1497–1516, 2012.

[2] L. Benini and G. D. Micheli, “Networks on Chips: A New SoC Paradigm,” in IEEE Computers, vol. 35, no. 1, pp. 70-78, 2002.

[3] P. Wettin, P. P. Pande, D. Heo, B. Belzer, S. Deb, and A. Ganguly, “Design space exploration for reliable mm-wave wireless NoC architectures,” Proc. 24th

IEEE Conf. Application-Specific Syst. Architectures and Processors, vol. 79, no. 82, pp. 5–7, Jun. 2013.

[4] A. Samaiyar, S. S. Ram, and S. Deb, "Millimeter-wave planar log periodic antenna for on-chip wireless interconnects," in Proc. 8th European Conf. Antennas Propag., pp. 1007–1009, Apr. 6-11, 2014.

[5] S. Borkar, "Networks for multi-core chips: a contrarian view," in: Special Session at ISLPED, 2007.

[6] H. K. Mondal, S. Deb, "An energy efficient wireless Network-on-Chip using power-gated transceivers," in Proc. 27th IEEE International System-on-Chip Conference (SOCC), vol., no., pp.243,248, 2-5 Sept. 2014.

[7] M. S. Shamim, N. Mansoor, A. Samaiyar, A. Ganguly, S. Deb, S. S. Ram, "Energy-efficient wireless network-on-chip architecture with log-periodic on-chip antennas." Proceedings of the 24th edition of the great lakes symposium on VLSI. ACM, 2014.

[8] H. K. Mondal, G. N. S. Harsha, and S. Deb, "An efficient hardware implementation of dvfs in multi-core system with wireless networkon-chip," in VLSI (ISVLSI), 2014 IEEE Computer Society Annual Symposium on, July 2014.

[9] H. K. Mondal, and S. Deb, "Energy Efficient On-chip Wireless Interconnects with Sleepy Transceivers," IEEE International Design and Test Symposium (IDT), pp.1-6, December 2013.

[10] H. K. Mondal, S. H. Gade, R. Kishore, S. Kaushik, and S. Deb, "Power efficient router architecture for wireless network-onchip," in Proc. 17th Int. Symp. Quality Electron. Des., pp. 227–233, Mar. 2016.

[11] T. Moscibroda, and O. Mutlu, "A case for bufferless routing in on-chip networks," ISCA-36, 2009.

[12] N. Banerjee, P. Vellank, K. S. Chatha, "A power and performance model for network-on-chip architectures," Proc. Des. Autom. Test Eur. Conf., pp. 1250-1255, 2004-Feb.

[13] H. Jingcao, R. Marculescu, "Application-specific buffer space allocation for networks-on-chip router design", Proceedings of the IEEE/ACM International Conference on Computer Aided Design (ICCAD), pp. 354-361, 2004.

[14] J. Wang, Y. Qia, J. Lu, B. Li, W. Dou, "An active buffer sizing algorithm for power-efficient NoC." Industrial Electronics and Applications (ICIEA), 2014 IEEE 9th Conference on. IEEE, 2014.

[15] Y. Choi and T. M. Pinkston, "Evaluation of queue designs for true fully adaptive routers", in Journal of Parallel and Distributed Computing, vol. 64(5), pp. 606-616, 2004.

[16] C. A. Nicopoulos, D. Park, J. Kim, N. Vijaykrishnan, M.S. Yousif and C. R. Das, "ViChaR: A Dynamic Virtual Channel Regulator for Network-on-Chip Routers", in Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture, pp, 333-346, 2006.

[17] H. K. Mondal, S. H. Gade, R. Kishore, and S. Deb, "Adaptive multi-voltage scaling in wireless NoC for high performance low power applications," in Proc. Des. Autom. Test Europe Conf. Exhib., pp. 1315–1320, Mar. 2016.

[18] V. Catania, A. Mineo, S. Monteleone, M. Palesi, D. Patti, "Cycle-Accurate Network on Chip Simulation with Noxim," ACM Transactions on Modeling and Computer Simulation. Volume 27 Issue 1, November 2016.

[19] C. Sun, C. O. Chen, G. Kurian, L. Wei, J. Miller, A. Agarwal, L. Peh, V. Stojanovic, "DSENT-a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling," Networks on Chip (NoCS), 2012 Sixth IEEE/ACM International Symposium on. IEEE, 2012.