A Survey of Power-Aware Network-on-Chip Design Techniques

Emmanuel Ofori-Attah
Faculty of Art, Science and Technology
Univeristy of Northampton
Northamptonshire, UK
Emmanuel.Ofori-Attah@northampton.ac.uk

Michael Opoku Agyeman
Faculty of Art, Science and Technology
Univeristy of Northampton
Northamptonshire, UK
Michael.OpokuAgyeman@northampton.ac.uk

***Abstract*—The Network-on-Chip paradigm has been heralded as the solution to the communication limitation that System-on-Chip poses. As we usher into the billion-transistor era, Network-on-Chip which was once deemed as the solution is defecting due to its power demanding components. Several techniques have been proposed over the years to improve the performance of the Network-on-Chip, trading off power efficiency. However, low power design solution is one of the essential requirements of future Network-on-Chip based System-on-Chip applications. Power dissipation can be reduced by efficient routers, architecture saving techniques and communication links. This paper presents recent contributions and efficient saving techniques at the router, Network-on-Chip architecture and Communication link level.**

Keywords: Network-On-Chip; Power Consumption; Many-Core; Routers.

I. Introduction

Transistor size reduction in Complementary Metal-Oxide Semiconductor (CMOS) technology has resulted in the duplication of many-cores integration on the same die [1] [2]. As a result of this complexity, the conventional bus interconnect becomes inadequate for future System-on-Chip (SoC) designs as they are constrained from delivering an assured Quality of Service (QoS) and bound to scalability issues and bandwidth limitation [3]. To mitigate this problem, Network-on-Chip (NoC) is proposed to enable simultaneous communication in a high integrated Multi-Processing-Element (MPE) system [4]–[6]. At the same time, this multiplies the power consumption and contributes to the excessive thermal issues (reliability, lifetime of systems, hotspots, chip damage) currently limiting the number of Processing Elements (PE) that can powered-on in MPE systems [7] [8]. NoC contributes to 40% of the chip's total power and increases as the network is extended. As a result, the number of powered-on PE in an MPE system will be limited leading to a degrade in performance.

To overcome this challenge, extensive research has been done in reducing the power consumption of NoC's resources with more emphasis on leakage power due to its consumption of the majority of the on-chip power [9]–[13]. Depending on the workload, the leakage power of a network can vary between 30-90% of the total power and therefore, optimal solutions are required for the development of deeper scaling in technology [14] [15]. Nevertheless, the switching activities of NoC's components have also received some attention [16] [17]. To reduce dynamic power consumption, Dynamic Voltage Frequency Scaling (DVFS) schemes are generally employed.

DVFS vary the voltage frequencies of working components based on the network load. Such techniques predict the network load and supply the amount of voltage required for a successful operation without impacting the performance [18]. Phan et al. [18] proposed an DVFS algorithm which incorporates a controller to monitor the activity of a router through its ports. Based on the traffic load of the ports, the controller supplies the required amount of voltage and frequency. However, varying the frequency of some tasks may result in errors and missed deadlines. Subsequently, a widely popular technique employed to reduce the leakage power consumption in technology is Power-Gating (PG). PG is normally used to power-off NoC components when the network is under-utilized. However, PG can affect the performance of the network and therefore, power efficiency becomes an important factor in many-core design architectures.

This paper presents a study of recent contributions on low power techniques for NoC design. Particularly, to achieve this, we investigated the power saving technique into three main levels. The rest of the paper is organised as follows. Section II explains about the NoC router architecture. Section III discusses efficient techniques for power in NoC architectures. Section IV provides techniques for communication links and finally, Section V concludes the paper.

II. Router Architecture

Routers occupy majority of the NoC's power consumption and therefore are widely selected as the best candidate for mitigating leakage power consumption. A typical router architecture consists of a buffer, crossbar, input and output ports. Moreover, existing literature [16] reveals that 33% of dynamic power in routers are consumed by buffers. Notably, the input buffers are shown to consume 44% of router power and occupy 15% of area [19]. Since buffers occupy majority of the power and occupy large area, many architectural designs have been proposed to overcome this challenge. One such proposal is the bufferless router concept. Bufferless routers employ

 ISBN: 978-1-61208-641-5

algorithms to route packets without temporally storing it [20]–[22]. However, this develops a bottleneck when the network packet injection rates rise eventually leading to livelock and deadlock. Deadlock in bufferless routers occur when two packets arise at the same time and contest for the same port. To solve this problem, existing work propose algorithms which grants access to one packet while the other is redirected on another route leading to livelock thus increasing the power consumption. Therefore, it is not practical to design a network which completely dismisses buffers.

Subsequently, the crossbar switch also consumes a significant amount of power. Increase of PE in MPE systems increases crossbar sizes and therefore adds complexity, scalability issues and large area as well as power consumption as shown by the amount of the power Intel's teraflops processor and MIT RAW crossbars consume [23]–[25]. Therefore, to reduce the power consumption of the routers, existing work have utilized novel techniques to improve power efficiency, the crossbar size, arbiter, and buffer designs. For this purpose, this paper focuses on optimized power saving techniques at the buffer and crossbar level.

A. Router Architecture Techniques

The employment of buffered routers trade-off area and power consumption to prevent deadlock, livelock, and high throughput. Alternatively, Virtual Channels (VC) are employed in buffers to enable parallelism. VC allows multiple access to a physical channel simultaneously. However, they consume a significant portion of NoC routers. Unfortunately, completely dismissing buffers leads to a poor performance network. To defuse this situation, many architectural designs have been proposed. This section of paper presents techniques which can improve the power performance of buffers.

1) Reduction in the pipeline stages: The many stages (Buffer Write (BW), Route Computation (RC), Virtual Channel Allocation (VA), Switch Allocation (SA), and Switch Stage (ST)) that a packet must traverse through to reach its destination increases latency and power consumption. For this purpose, Noghondar et al. [26] proposed an arbitration method which reduces latency and power consumption as well as contention among flits. The proposed arbitration method assigns priority levels to each input port and the port with the highest priority is granted access to its required output port. To prevent latency of low priority ports, a counter is employed to keep track of the cycles that a flit has been delayed for. When the delay exceeds the threshold of the flit, the low priority port is granted access to the output port.

Likewise, Postman et al. highlights in [27] and propose the SWIFT NoC. The SWIFT NoC reduces power consumption by allowing flits to bypass the buffering stage in one (1) cycle; averting the use of read and write power.

Shenbagavalli et al. [28] on the other hand approach this issue by proposing a hybrid scheme which combines circuit and packet switching to allow flits to traverse through the network with only one (1) stage. Compared to virtual point to point connections, this hybrid scheme achieves a reduction of 6.8% in latency and 11.3% in power.

2) Power-gating Techniques: As previously stated before, PG disconnects idle resources from their voltage source to save more power for actual computation. The following authors employed PG to power-off VCs. Muhammad et al. [29] employed PG to activate and deactivate VCs based on the network's workload. In this architecture, VCs are divided into separate groups. Based on the workload, the groups are activated and deactivated. Similarly, Zhan et al. [17] employed PG to activate and deactivate grouped VCs. However, the power savings of the Zhan's architecture, is higher than Muhammad's because of the employment of Spin-Transfer Torque Magnetic Random-Access Memory (STT-RAM) which consumes less power than the conventional Static Random-Access Memory (SRAM). In addition to this, Zhan's VCs can either be powered-off or set into a drowsy state which decreases the wake-up time. Consequently, Nasirian et al. [30] employs PG to disable idle buffers by tracking its inactivity cycle state.

Unfortunately, to achieve a high-power savings, routers must be shut down for long periods of time because the continuous shut down of components incur non-negligible power overhead. For this purpose, existing work propose complex algorithms to route packets through different channels. Unfortunately, such algorithms require knowledge of the network to prevent deadlock and livelock [31]. Consequently, this adds extra complexity and functionality and increases the power consumption as well as possible latency issues. Additionally, the wake-up time for powered-off routers to be activated in time for incoming traffic introduces delay and causes performance degradation. Chen et al. [32] proposed a solution to transmit wake-up signal 3 hops ahead, ensuring intersecting routers which are powered-off are activated in time.

3) Substitutes to Input Buffers: Input buffers particularly consumes a staggering amount of power. To maintain performance whilst also mitigating the power consumption, many designs propose alternatives to input buffers. Kodi et al. [33] proposed an architecture which employs dual-function links and utilizes dynamic router allocation to assign flits to any free buffer. DiTomaso et al. [16] proposed QORE, an architecture which improves power consumption using power-efficient Multi-Function Channel buffers (MFC) and enhances the performance through reversible links. The use of MFC enables the channel buffers to be utilized instead of the routers in the buffers. Li et al. [34] deals with power consumption by replacing the conventional SRAM with Embedded Dynamic Random-Access Memory (eDRAM). Significantly, the buffer area was reduced by 52% and power by 43%.

B. Crossbar Switches

A crossbar switch is composed of individual switches arranged in a matrix form between several inputs and outputs. Crossbar size increases as the network gets larger. This amplifies the power consumption. Crossbar switches can be categorized into two groups, single stage and multi stage.

 ISBN: 978-1-61208-641-5

1) Crossbar size: To achieve low power consumption and small area, existing work focuses on splitting large crossbars into smaller crossbars. Kim et al. [35] proposed a router architecture composed of two crossbars. In the proposed router architecture, the employment of smaller crossbars reduced the size of the Virtual Channel Arbiter (VCA), Switch Arbiter units (SA) and shorter logic depth. Similarly, an optimized crossbar is proposed by the Park et al. in [36] which combines decomposition and segmentation to effectively reduce power consumption by 35%. The crossbar has been disassembled in two small crossbars to reduce area and power. However, in a large-scale network, there will an increase to latency and contention issues.

Emerging crossbars are being built based on multi stage crossbars such as the Clos and Benes network [37] [38] [39] because of their provision of low power and smaller area. Yikun et al. [40] conducted a study on Circuit design and concluded that the Clos network outperform their counterparts (Benes and Single stage Crossbars) in several ways. In the Clos network, there is a reduction in the number of logic unit used. The Benes network suffers from 65% delay in timing and less power in Clos consumed because of the size of crossbars.

Naik et al. [41] proposed a heterogeneous NoC embedded with circuit switched routers composed of buffered and bufferless routers and a 3-stage Clos network. In comparison to a crossbar switch of the same size, the results of this is a reduction of 26% in power consumption and 32% in area. However, circuit switched network causes additional latency when a transmission is established between a source and its destination.

2) Switching Algorithm: In theory, there are two different types of routers; circuit switching routers and packet switching routers. In packet switching routers, data is encoded into packets and routed individually through the network. Circuit switching routers on the hand establishes a connection between the source and destination and specifically allocate resources which will be used for transmission [42]. In Circuit switching routers, there is guaranteed throughput because all packets can be transmitted at the same time without delay in any router. However, there is an increase in latency. This is because during the transmission process, the resources allocated cannot be accessed.

For this purpose, CirKET switching mechanism has been proposed by the authors in [43] to effectively use the benefits of packet switching and circuit switching. In this architecture, messages are split into different groups; High priority and Low priority. High priority messages are transmitted using circuit switching and low priority messages are transmitted using packet switching. The employment of these two mechanisms allow power rails to be disconnected and power-gating to be used to disconnect parts of the router which are not used during a transmission.

III. Low Network Architecture

Novel NoC architectures have been proposed to reduce the average packet latency while increasing the throughput. However, this is usually at the expense of power consumption. To combat the challenges imposed by these power Hungry NoCs, various architectures have been proposed.

The exponential increase in the number of cores in multi-core over the last decade has resulted in the emergence of Three-Dimension (3D) NoC as the platform for On-Chip communication [44] [45]. 3D NoC allows multiple silicon layers to be stacked together to not only enhance the throughput and latency, but also to reduce power consumption [46] [47]. In 3D NoC, the lengthy wires are replaced with short wires through silicon vias (TSVs) to minimize the number of hops it takes for a packet to traverse through the network. Particularly, the increase in the number of links in 3D Integrated Circuits (IC) allows the transmission of more messages around the network [48].

Debora Matos et al. [49] proposed the 3D HiCIT, an architecture comprised of two hierarchical levels with a mesh topology at the top level. In comparison with the traditional 3D spin and 3D mesh topologies, the proposed architecture reduces the average latency to 50% and 54% respectively, with the 3D spin been the latter [49]. In addition to this, the architecture is comprised of a crossbar and low-cost routers. Compared to the 3D spin and 3D butterfly fat tree topologies, the proposed architecture uses less TSVs.

Stacking of multiple silicon layers in 3D IC reduces hop-count in comparison with the long interconnect wires in Two Dimension (2D) NoC. However, limitations such as power density caused by the chip size, the cost of TSV and its defects [50] [49] prevents 3D NoC from reaching its potential. For this purpose, the author in [51] recommends the use of monolithic 3D. One approach to reduce power consumption is to use fewer buffers at the router port [52]–[54]. Similarly, Fang et al. [55] proposed Reduce Router Counts and Increase Efficiency Service (RRCIES), an architecture based on a mesh topology. RRCIES allows multiple cores to be connected between one router. As a result, hop distance is reduced. The use of fewer routers constitutes to a reduction in power hungry components such as buffers, crossbar, switches, and virtual channels.

Another alternative is to employ PG. The study of vertical slit field effect transistors led to the proposal of a 3D Hybrid architecture in [56]. PG and clock gating are employed in this architecture to enable different level of buffers to be deactivated. The proposed architecture splits the input buffers into three (3) levels. Each input port is designed to access all three levels and permit any virtual channel destination to be chosen. In addition, the buffers from ports which are not being used are shared among busy ports.

IV. Communication Links

Although routers consume more power in NoC, the communication links can be optimized to accommodate this. According to [57] and [58], routers and communication channels contributes to most of the power consumption in NoC. Therefore, existing work have developed techniques to reduce the amount of power consumed by the links.

 ISBN: 978-1-61208-641-5

1) Voltage Scaling : The voltage swing in the communication links can be optimized to reduce the amount of power it consumes. However, this is at a cost of a rise in error bit rate. For this purpose, Mineo et al. [59] proposed to reduce power consumption by using a technique which permits two working levels in a link. A flag is attached to each communication to identify their priority. Low prioritised communications (Body and Tail flit) can be transmitted on a low-level voltage swing while the others (Head flit) can be sent using a normal level voltage.

2) Half-cycle Flits : The longer flits traverse through the links in NoC, the more power is consumed. Therefore, decreasing the number of cycles of it takes from a flit to transmit between routers would not only enhance the performance of the network but also save power. A. Psarras et al. [60] proposed a technique which allows flits to only use half a cycle to hop between routers. By allowing flits to spend less time in the links, less power is consumed compared to single cycle routers where one cycle is used to execute all operations in the router and one is used to hop between routers.

V. Summary of contribution

Table I presents a summary of the techniques. The table presents three categories (Techniques, Performance degradation and power saving). The Technique categories presents all the techniques which have been presented in the paper. The performance degradation category shows the impact that the applied technique has on the performance on the architecture. Lastly, the power saving technique shows the amount of power savings that the applied technique saves when applied to the architecture.

From the table, it can be concluded that there is a balance between performance and power when alternative buffers are used. Additionally, employing PG also impacts on the performance and power of the architecture.

VI. Conclusion

In this paper, several NoC power saving techniques have been critically evaluated. Particularly, the effect of buffered and bufferless routers on power consumption have been presented. Moreover, a summary of these techniques has been presented to compare their trade off. The combination of some of the architectures presented, if employed, can help improve the amount of power consumed by NoC resources which can either be removed or switched off. Whether it being the adjustment (crossbar size, buffers, virtual channels) of the components in the router architecture, modifying the architectures (resource management) and the amount of voltage used in the communication links. Also, we explored low power techniques used in emerging NoC Architectures; 3D NoC and WNOC. Based on our discussions, we can conclude power dissipation can be reduced in all areas of a network infrastructure. Our future work will be focused on accessing the power consumption of the memory subsystem. Particularly, on the Last Level Cache [61] [62].

References

[1] M. O. Agyeman, A. Ahmadinia, and A. Shahrabi, "Heterogeneous 3d network-on-chip architectures: Area and power aware design techniques," *Journal of Circuits, Systems, and Computers*, 2013.

[2] M. O. Agyeman, Q. Vien, A. Ahmadinia, A. Yakovlev, K. Tong, and T. S. T. Mak, "A resilient 2-d waveguide communication fabric for hybrid wired-wireless noc design," *IEEE Trans. Parallel Distrib. Syst.*, pp. 359–373, 2017.

[3] M. H. Neishaburi and Z. Zilic, "A fault tolerant hierarchical network on chip router architecture," in *IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems*, 2011, pp. 445–453.

[4] M. O. Agyeman and A. Ahmadinia, "Optimised application specific architecture generation and mapping approach for heterogeneous 3d networks-on-chip," in *IEEE 16th International Conference on Computational Science and Engineering*, 2013, pp. 794–801.

[5] ——, "An adaptive router architecture for heterogeneous 3d networks-on-chip," in *NORCHIP*, 2011, pp. 1–4.

[6] ——, "Optimising heterogeneous 3d networks-on-chip," in *Sixth International Symposium on Parallel Computing in Electrical Engineering (PARELEC 2011), 4-5 April 2011, Luton, United Kingdom*, 2011, pp. 25–30.

[7] E. Ofori-Attah, W. Bhebhe, and M. O. Agyeman, "Architectural techniques for improving the power consumption of noc-based cmps: A case study of cache and network layer," *Journal of Low Power Electronics and Applications*, 2017.

[8] M. O. Agyeman, Q. Vien, and T. S. T. Mak, "An analytical channel model for emerging wireless networks-on-chip," in *IEEE Intl Conference on Computational Science and Engineering, CSE 2016, and IEEE Intl Conference on Embedded and Ubiquitous Computing, EUC 2016, and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering, DCABES 2016, Paris, France, August 24-26, 2016*, 2016, pp. 9–15.

[9] J. Howard, S. Dighe, Y. Hoskote, S. Vangal, D. Finan, G. Ruhl, D. Jenkins, H. Wilson, N. Borkar, G. Schrom, F. Pailet, S. Jain, T. Jacob, S. Yada, S. Marella, P. Salihundam, V. Erraguntla, M. Konow, M. Riepen, G. Droege, J. Lindemann, M. Gries, T. Apel, K. Henriss, T. Lund-Larsen, S. Steibl, S. Borkar, V. De, R. V. D. Wijngaart, and T. Mattson, "A 48-core ia-32 message-passing processor with dvfs in 45nm cmos," in *IEEE International Solid-State Circuits Conference - (ISSCC)*, 2010.

[10] R. Parikh, R. Das, and V. Bertacco, "Power-aware nocs through routing and topology reconfiguration," in *51st ACM/EDAC/IEEE Design Automation Conference (DAC)*, 2014, pp. 1–6.

[11] E. Ofori-Attah and M. O. Agyeman, "A survey of low power noc design techniques," in *Proceedings of the 2Nd International Workshop on Advanced Interconnect Solutions and Technologies for Emerging Computing Systems*, ser. AISTECS '17. New York, NY, USA: ACM, 2017, pp. 22–27. [Online]. Available: http://doi.acm.org/10.1145/3073763.3073767

[12] M. O. Agyeman, J. X. Wan, Q. T. Vien, W. Zong, A. Yakovlev, K. Tong, and T. Mak, "On the design of reliable hybrid wired-wireless network-on-chip architectures," in *IEEE 9th International Symposium on Embedded Multicore/Many-core Systems-on-Chip*, 2015.

[13] E. Ofori-Attah and M. O. Agyeman, "A survey of recent contributions on low power noc architectures," in *Computing Conference*, 2017, pp. 1086–1090.

[14] R. Das, S. Narayanasamy, S. K. Satpathy, and R. G. Dreslinski, "Catnap: Energy proportional multiple network-on-chip," *SIGARCH Comput. Archit. News*, 2013.

[15] B. K. Daya, C. H. O. Chen, S. Subramanian, W. C. Kwon, S. Park, T. Krishna, J. Holt, A. P. Chandrakasan, and L. S. Peh, "Scorpio: A 36-core research chip demonstrating snoopy coherence on a scalable mesh noc with in-network ordering," in *ACM/IEEE 41st International Symposium on Computer Architecture (ISCA)*, 2014.

[16] D. DiTomaso, A. K. Kodi, A. Louri, and R. Bunescu, "Resilient and power-efficient multi-function channel buffers in network-on-chip architectures," *IEEE Transactions on Computers*, no. 12, pp. 3555–3568, 2015.

[17] J. Zhan, J. Ouyang, F. Ge, J. Zhao, and Y. Xie, "Hybrid drowsy sram and stt-ram buffer designs for dark-silicon-aware noc," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, no. 10, pp. 3041–3054, 2016.

 ISBN: 978-1-61208-641-5

TABLE I
LLC SRAM AND STT-RAM TECHNOLOGY COMPARISONS

Techniques	Performance Degradation	Power Saving
Power Gating	Medium	Medium
Alternative Buffers	Medium	High
Pipeline Stages	Low	Low
Voltage Scaling	Low	Low
Half-cycle Flits	Low	Low

[18] H. P. Phan, X. T. Tran, and T. Yoneda, "Power consumption estimation using vnoc2.0 simulator for a fuzzy-logic based low power network-on-chip," in *IEEE International Conference on IC Design and Technology (ICICDT)*, 2017.

[19] P. Kundu, "On-die interconnects for next generation cmps, in workshop on on- and off-chip interconnection networks for multicore systems," 2006.

[20] C. Feng, Z. Liao, Z. Lu, A. Jantsch, and Z. Zhao, "Performance analysis of on-chip bufferless router with multi-ejection ports," in *IEEE 11th International Conference on ASIC (ASICON)*, 2015, pp. 1–4.

[21] C. Fallin, C. Craik, and O. Mutlu, "Chipper: A low-complexity bufferless deflection router," in *IEEE 17th International Symposium on High Performance Computer Architecture*, 2011, pp. 144–155.

[22] B. K. Daya, L. S. Peh, and A. P. Chandrakasan, "Towards high-performance bufferless nocs with scepter," *IEEE Computer Architecture Letters*, pp. 62–65, 2016.

[23] K. Sewell, R. G. Dreslinski, T. Manville, S. Satpathy, N. Pinckney, G. Blake, M. Cieslak, R. Das, T. F. Wenisch, D. Sylvester, D. Blaauw, and T. Mudge, "Swizzle-switch networks for many-core systems," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, pp. 278–294, 2012.

[24] M. B. Taylor, J. Kim, J. Miller, D. Wentzlaff, F. Ghodrat, B. Greenwald, H. Hoffman, P. Johnson, J.-W. Lee, W. Lee, A. Ma, A. Saraf, M. Seneski, N. Shnidman, V. Strumpen, M. Frank, S. Amarasinghe, and A. Agarwal, "The raw microprocessor: a computational fabric for software circuits and general-purpose programs," *IEEE Micro*, pp. 25–35, 2002.

[25] S. R. Vangal, J. Howard, G. Ruhl, S. Dighe, H. Wilson, J. Tschanz, D. Finan, A. Singh, T. Jacob, S. Jain, V. Erraguntla, C. Roberts, Y. Hoskote, N. Borkar, and S. Borkar, "An 80-tile sub-100-w teraflops processor in 65-nm cmos," *IEEE Journal of Solid-State Circuits*, pp. 29–41, 2008.

[26] A. F. Noghondar, M. Reshadi, and N. Bagherzadeh, "Reducing bypass-based network-on-chip latency using priority mechanism," *IET Computers Digital Techniques*, 2018.

[27] J. Postman, T. Krishna, C. Edmonds, L. S. Peh, and P. Chiang, "Swift: A low-power network-on-chip implementing the token flow control router architecture with swing-reduced interconnects," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, pp. 1432–1446, 2013.

[28] S. Shenbagavalli and S. Karthikeyan, "An efficient low power noc router architecture design," in *Online International Conference on Green Engineering and Technologies (IC-GET)*, 2015, pp. 1–8.

[29] S. T. Muhammad, M. A. El-Moursy, A. A. El-Moursy, and A. M. Refaat, "Optimization for traffic-based virtual channel activation low-power noc," in *Energy Aware Computing Systems Applications (ICEAC), International Conference on*, 2015, pp. 1–4.

[30] N. Nasirian and M. Bayoumi, "Low-latency power-efficient adaptive router design for network-on-chip," in *28th IEEE International System-on-Chip Conference (SOCC)*, 2015, pp. 287–291.

[31] D. DiTomaso, A. Sikder, A. Kodi, and A. Louri, "Machine learning enabled power-aware network-on-chip design," in *Design, Automation Test in Europe Conference Exhibition (DATE),*, 2017.

[32] L. Chen, D. Zhu, M. Pedram, and T. M. Pinkston, "Power punch: Towards non-blocking power-gating of noc routers," in *IEEE 21st International Symposium on High Performance Computer Architecture (HPCA)*, 2015, pp. 378–389.

[33] A. K. Kodi, A. Sarathy, A. Louri, and J. Wang, "Adaptive inter-router links for low-power, area-efficient and reliable network-on-chip (noc) architectures," in *Asia and South Pacific Design Automation Conference*, 2009, pp. 1–6.

[34] C. Li and P. Ampadu, "A compact low-power edram-based noc buffer," in *Low Power Electronics and Design (ISLPED), IEEE/ACM International Symposium on*, 2015, pp. 116–121.

[35] J. Kim, C. Nicopoulos, and D. Park, "A gracefully degrading and energy-efficient modular router architecture for on-chip networks," in *33rd International Symposium on Computer Architecture (ISCA'06)*, 2006, pp. 4–15.

[36] D. Park, A. Vaidya, A. Kumar, and M. Azimi, "Mode-x: Microarchitecture of a layout-aware modular decoupled crossbar for on-chip interconnects," *IEEE Transactions on Computers*, no. 3, pp. 622–636, 2014.

[37] Y. Xia, M. Hamdi, and H. J. Chao, "A practical large-capacity three-stage buffered clos-network switch architecture," *IEEE Transactions on Parallel and Distributed Systems*, 2016.

[38] S. Yang, S. Xin, Z. Zhao, and B. Wu, "Minimizing packet delay via load balancing in clos switching networks for datacenters," in *International Conference on Networking and Network Applications (NaNA)*.

[39] J. Zhang and H. Gu, "A partially adaptive routing algorithm for benes network on chip," in *2nd IEEE International Conference on Computer Science and Information Technology*, 2009.

[40] Y. Jiang and M. Yang, "On circuit design of on-chip non-blocking interconnection networks," in *27th IEEE International System-on-Chip Conference (SOCC)*, 2014.

[41] A. Naik and T. K. Ramesh, "Efficient network on chip (noc) using heterogeneous circuit switched routers," in *International Conference on VLSI Systems, Architectures, Technology and Applications (VLSI-SATA)*, 2016, pp. 1–6.

[42] N. Chin-Ee and N. Soin, "A study on circuit switching merits in the design of network-on-chip," in *Computer and Communication Engineering (ICCCE), International Conference on*, 2010, pp. 1–5.

[43] M. FallahRad, A. Patooghy, H. Ziaeeziabari, and E. Taheri, "Cirket: A performance efficient hybrid switching mechanism for noc architectures," in *Euromicro Conference on Digital System Design (DSD)*, 2016, pp. 123–130.

[44] M. O. Agyeman and A. Ahmadinia, "An adaptive router architecture for heterogeneous 3d networks-on-chip," in *2011 NORCHIP, Lund, Sweden, November 14-15, 2011*, 2011, pp. 1–4.

[45] ——, "Power and area optimisation in heterogeneous 3d networks-on-chip architectures," *SIGARCH Computer Architecture News*, vol. 39, no. 4, pp. 106–107, 2011.

[46] C. Chao, "Traffic and thermal aware run time thermal management scheme for 3d noc systems," pp. 223 – 230, 2010.

[47] W. R. Davis, "Application exploration for 3d integrated circuits: Tcam, fifo and fft case studies," pp. 496 – 506, 2009.

[48] A. W. Yin, "Change function of 2d/3d network-on-chip," 2011.

[49] D. Matos, M. Prass, M. Kreutz, L. Carro, and A. Susin, "Performance evaluation of hierarchical noc topologies for stacked 3d ics," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2015, pp. 1961–1964.

[50] D. e. a. Velenis, "Impact of 3d design choices on manufacturing cost," 2009.

[51] D. K. Nayak, S. Banna, S. K. Samal, and S. K. Lim, "Power, performance, and cost comparisons of monolithic 3d ics and tsv-based 3d ics," in *SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), IEEE*, 2015.

[52] A. T. Tran and B. M. Baas, "Roshaq: High-performance on-chip router with shared queues," in *Computer Design (ICCD), IEEE 29th International Conference on*, 2011, pp. 232–238.

[53] K. Latif, A. M. Rahmani, L. Guang, T. Seceleanu, and H. Tenhunen, "Pvs-noc: Partial virtual channel sharing noc architecture," in *19th International Euromicro Conference on Parallel, Distributed and Network-Based Processing*, 2011, pp. 470–477.

[54] R. S. Ramanujam, V. Soteriou, B. Lin, and L. S. Peh, "Design of a high-throughput distributed shared-buffer noc router," in *Networks-on-*

 ISBN: 978-1-61208-641-5

Chip (NOCS), Fourth ACM/IEEE International Symposium on, 2010, pp. 69–78.

[55] J. Fang, J. Lu, and C. She, "Research on topology and policy for low power consumption of network-on-chip with multicore processors," in *International Conference on Computational Science and Computational Intelligence (CSCI)*, 2015, pp. 621–625.

[56] V. S. Nandakumar and M. Marek-Sadowska, "A low energy network-on-chip fabric for 3-d multi-core architectures," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, no. 2, pp. 266–277, 2012.

[57] J. Balfour and W. J. Dally., "Design tradeoffs for tiled cmp on-chip networks in ics,," 2006.

[58] S. M. Hassan and S. Yalamanchili, "Centralized buffer router: A low latency, low power router for high radix nocs," in *Networks on Chip (NoCS), Seventh IEEE/ACM International Symposium on*, 2013, pp. 1–8.

[59] A. Mineo, M. Palesi, G. Ascia, and V. Catania, "Runtime online links voltage scaling for low energy networks on chip," in *Digital System Design (DSD), Euromicro Conference on*, 2013, pp. 941–944.

[60] A. Psarras, J. Lee, P. Mattheakis, C. Nicopoulos, and G. Dimitrakopoulos, "A low-power network-on-chip architecture for tile-based chip multi-processors," in *International Great Lakes Symposium on VLSI (GLSVLSI)*, 2016, pp. 335–340.

[61] E. Ofori-Attah, X. Wang, and M. O. Agyeman, "A survey of low power design techniques for last level caches," in *Applied Reconfigurable Computing. Architectures, Tools, and Applications - 14th International Symposium, Santorini, Greece, May 2-4,, Proceedings*, 2018, pp. 217–228.

[62] S. Borkar, "How to stop interconnects from hindering the future of computing," in *Optical Interconnects Conference*, 2013.

 ISBN: 978-1-61208-641-5