A survey on energy-efficient methodologies and architectures of network-on-chip


Department of Electrical and Computer Engineering, North Dakota State University, Fargo, ND, USA

Abstract
Integration of large number of electronic components on a single chip has resulted in complete and complex systems on a single chip. The energy efficiency in the System-on-Chip (SoC) and its communication subset, the Network-on-Chip (NoC), is a key challenge, due to the fact that these systems are typically battery-powered. We present a survey that provides a broad picture of the state-of-the-art energy-efficient NoC architectures and techniques, such as the routing algorithms, buffered and bufferless router architectures, fault tolerance, switching techniques, voltage islands, and voltage-frequency scaling. The objective of the survey is to educate the readers with the latest design-improvements that are carried out in reducing the power consumption in the NoCs.

1. Introduction
An exponential demand for the smart and portable computing has resulted in a shift from the conventional board designs to the System-on-Chip (SoC) where different elements, such as the Central Processing Unit (CPU), memory, Graphics Processing Unit (GPU), modem, router, radios, and bridges, are all integrated on the same chip. Currently, the SoC is being used in the domain specific devices, such as the military, aerospace, microprocessors, medical science, physical science, and cellular technology. Typically, the SoC devices are battery-powered. Therefore, the energy efficiency is by far the single biggest challenge for the performance-scaling in these devices. The challenge of reducing the power consumption, while improving the SoC computing performance is a primary goal for researchers and design engineers.

The high density of the components in the SoC complicate the design and implementation of shared bus architecture. For the effective and fast communication among different components, the Network-on-Chip (NoC) is used [1]. Therefore, the NoC architecture is a subset/component of the SoC technology [2]. Replacing the conventional bus structures with the NoC provides significant improvements in the SoC communication architecture. The cores in the SoC are connected to a network through a network interface. The communication among the cores occurs by sending the packets on a path comprising of the switches and inter-switch links [3]. The NoC architecture allows higher bandwidth and reduces the latency, by integrating the synchronization and concurrency of the data flow among various modules [2]. The architectural components of

* Corresponding author.
E-mail addresses: assad.abbas@ndsu.edu (A. Abbas), mazhar.ali@ndsu.edu (M. Ali), ahmad.fayyaz@ndsu.edu (A. Fayyaz), ankan.ghosh@ndsu.edu (A. Ghosh), anshul.kalra@ndsu.edu (A. Kalra), samee.khan@ndsu.edu (S.U. Khan), ushahid.khan@ndsu.edu (M.U.S. Khan), thiago.demenezes@ndsu.edu (T. De Menezes), sayantica.pattanayak@ndsu.edu (S. Pattanayak), alarka.sanyal@ndsu.edu (A. Sanyal), saeeda.usman@ndsu.edu (S. Usman).

http://dx.doi.org/10.1016/j.compeleceng.2014.07.012
0045-7906/© 2014 Elsevier Ltd. All rights reserved.
the NoC include the routers, network interfaces, and communication links, whereas switches, buffers (for the input and output), schedulers, and Virtual Channel (VC) allocators are components of NoC routers [3].

Cost efficiency in terms of the energy consumption is a vital concern for the NoC. Consequently, the energy efficiency in the communication infrastructure of the NoC becomes a cornerstone to minimize the system-level power consumption. The increase in the number of transistors on the chip results in the decreased voltage supply [4]. The decrease in the energy supply elevates the need of careful power utilization. Around 40% of the total system power consumption is caused by the leakage power [4]. Ref. [5] claims that the NoC may consume up to 35% of the total SoC power. Likewise, Ref. [6] claims that a considerable amount of SoC power is consumed by interconnections. Therefore, the energy consumption is an important consideration while designing the efficient routing algorithms.

This survey provides a comprehensive overview of the state-of-the-art research conducted in the NoC domain. The major areas focussed in this survey are: (a) routing algorithms, (b) buffered and bufferless architecture, (c) fault tolerance, (d) switching techniques, (e) voltage islands, and (f) voltage-frequency scaling. The survey focuses on energy efficiency in each of the domains with a comprehensive analysis of the recent research techniques in the aforesaid domains. The NoC is much broader topic having various other sub domains, such as the deadlock avoidance and recovery, topologies, reconfiguration, and alternative interconnect technologies to the wireless interconnects. However, the survey does not encompass the above mentioned domains because they already have been discussed in detail in the various past studies.

The NoC performance is largely dependent on the routing algorithms that significantly impact the throughput and latency. Therefore, the optimization of routing algorithms for the NoC is a key concern in enhancing the NoC performance and to minimize the energy consumption [7]. The selection of the most optimal path to route the packets in the NoC minimizes the traffic flow on the network that in turn results in energy efficiency.

Likewise, the buffers also consume substantial portion of the router’s power in the NoC [8]. In the conventional designs, the buffers were assumed to be a necessary part of the routers within the NoC. However, recent techniques have shown that removing the buffers from the NoC reduces the power consumption. Conversely, the performance of the NoC is also affected. The hybrid techniques tend to find the delicate balance between the low power consumption and high performance, by reducing the number of required buffers in the NoC [1].

Faults in the communication links of the NoC can degrade the system performance. The escalated probability of the permanent and transient faults due to the swift aging effects and development/testing challenges brings the reliability issue at the forefront [9]. The deadlocks on the communication links can severely affect the working of the NoC. Similarly, the Processing Entity (PE), routers, and memory units can face faults that hinder the achievement of the overall system’s objectives. The fault tolerance techniques are applied to ensure the safe operation of the chip. The fault tolerance techniques require extra processing and/or redundancy in the hardware that in turn, increases the power consumption. Increased power consumption may lead to an increase in the temperature of the devices (on-chip) that intensifies the probability of the faults and diminishes the reliability [10]. Therefore, power-efficient fault tolerance techniques need to be developed and implemented in the NoC to: (a) minimize the system-level power consumption and (b) increase the reliability of the communication. The fault-tolerant schemes encompassing the flow control and routing, have been presented in this study.

The switch design in the NoC is an important factor that affects the energy efficiency and performance of the overall network. The latency and throughput of the network is also dictated by the micro-architecture of the switches. In the NoC paradigm, the interconnects have become a focus and the constraints and trade-offs have changed resulting in a more pronounced switch design in terms of the latency and power consumption [3].

The energy consumption of the SoCs can be reduced by utilizing the Dynamic Voltage and Frequency Scaling (DVFS) [11] and Voltage Islands (VIs) [12]. The Multi-Supply Voltage (MSV) allows for the partitioning of the circuit into multiple VIs, with each of the islands operating at a particular voltage level. The voltage islanding not only improves power efficiency but also reduces the chip area. The DVFS is widely used as an effective method to reduce the energy consumption of a SoC and the multi-core NoC by varying the voltage and frequency of the NoC in real-time.

Although the current literature contains the previous attempts pertaining to the various areas in the NoC, this survey differs significantly in terms of its extensiveness, focus on latest developments, and the primary goal of evaluating the energy efficiency in the various techniques and designs. For example, Ref. [13] discussed the NoC programming models and gave a tabulated analysis of the topologies and switching techniques only. However, the study did not focus on the energy efficiency. The routing algorithms for the NoCs have been discussed in [14], but again the study did not take into account the energy efficiency. Agarwal et al. [2] provided a comprehensive survey of the NoC encompassing diverse areas, such as the topologies, router architecture, routing protocols, switching techniques, flow control, VC, error correction, and detection. However, the review did not take into account the power consumption in the NoC. The authors in [15] provided a sufficient discussion on the existing approaches for the NoC and analyzed the networks in terms of the probabilistic and deterministic design approaches, alongside the designs inspired by statistical physics. Our survey in its entirety encompasses most of the recent research endeavors that have influenced the energy efficiency in the SoC/NoC within each of the areas of our focus. Moreover, we believe that the study will be beneficial for the researchers in finding the most recent research on energy-efficient designs and techniques for the NoC, at a single place.

The rest of this survey is organized as follows. In Section 2, a survey of the energy-efficient routing algorithms for the NoC is presented. Section 3 provides a discussion on the buffered and bufferless NoCs, with an emphasis on the energy conservation. Energy-efficient fault tolerance schemes are presented in Section 4, and Section 5 discusses the switch...
synchronization. Power management by using the VIs and voltage-frequency scaling is discussed in Section 6. Section 7 presents the discussion and Section 8 concludes the survey.

2. Energy-efficient routing algorithms for NoC

Optimized routing algorithms, lead to a better NoC performance and reduced energy consumption [7]. The routing algorithms for the NoCs are classified into the adaptive routing algorithms and oblivious routing algorithms [14]. Besides the energy efficiency, the routing algorithms also facilitate the improvements in the NoC performance, alongside avoiding the congestion and deadlocks. Moreover, various routing algorithms are also fault-tolerant.

The adaptive routing algorithms have the capability to route the traffic dynamically, depending on the traffic conditions in the network. However, implementing the adaptive routing algorithms that adjust the path at the run-time is not a trivial task. To fulfill the requirements for a deadlock-free packet transmission, the adaptive routing algorithms require intelligent design to perform the routing decisions on-the-fly. The adaptive routing algorithms are classified into the: (a) minimal and non-minimal path adaptive routing algorithm, (b) fully adaptive routing algorithm, and (c) Odd–Even routing [14]. In the minimal path routing, the packets follow the shortest paths; whereas, in the non-minimal path routing, the packet is led away from the destination [16]. The fully adaptive routing algorithms always utilize a path that is not congested. The Odd–Even adaptive turn model algorithms restrict the locations, where, certain turns can be taken to avoid a circular turn.

The oblivious routing algorithms are categorized into the deterministic routing algorithms, and Dimension Order Routing (DOR) algorithms [14]. The deterministic routing algorithms are a subset of the oblivious routing that always follow a pre-termined path between the source and the destination. The DOR determines the direction of the routed packets during every routing stage. The oblivious routing lacks in traffic balancing capabilities and also do not consider the congestion status of the available ports. The taxonomy of the various routing techniques is presented in Fig. 1.

2.1. Adaptive routing algorithms

A congestion-aware routing algorithm, called the Destination Based Adaptive Routing Algorithm (DBAR), is proposed by Ma et al. [17]. The DBAR overcomes the limitations of locally and globally adaptive routing schemes, and provides efficient, fully adaptive routing strategy to avoid the congestion, even beyond adjacent nodes. Moreover, the DBAR minimizes the interference among the applications by separating the regions dynamically. The task of separating regions is accomplished through a selection function that is integrated with the destination. The power overhead, in terms of the transmitted information about the congested routes, was measured as compared to the DOR routing algorithm. The energy-delay product for the DBAR was observed to be 20% lesser, in comparison to the DOR algorithm.

Ebrahimi et al. [7] proposed a non-minimal adaptive routing algorithm, called the LEAR that avoids congested paths from the source to the destination. The LEAR is based on the mad-y routing strategy that has the inability to use the appropriate turns for routing the packets over the paths with the least amount of congestion. Moreover, the LEAR overcomes the aforementioned limitation of the mad-y, by increasing the ability of existing VCs to handle the misrouted packets. The packets are routed on the non-minimal routes when the congested traffic conditions are observed on the neighboring routers. The LEAR algorithm was compared to the XY, DyXY, and mad-y routing algorithms, in terms of the power consumption. On an average, the power consumption in the LEAR was 10% lower.

Ebrahimi et al. [18] presented another approach to distribute the unicast and multicast traffic in 3D mesh based NoCs. An algorithm based on Hamiltonian path called Minimum and Adaptive Routing (MAR) algorithm is proposed that does not
require any virtual channel for both the unicast and multicast messages. The experimental results show that the Recursive Partitioning (RP) method equally distributes the multicast traffic and minimizes the network latency considerably. The power dissipation of the RP method under the multicasting model was observed 16% and 8% less than Two-Block Partitioning (TBP) and Vertical-Block Partitioning method (VBP) partitioning methods, respectively.

Ref. [19] also proposed an adaptive algorithm for unicast and multicast traffic. Without using any virtual channel the degree of adaptivity is maximized while ensuring the deadlock free routing operations. The technique minimizes the power dissipation significantly as compared to the Multi-Path and Column-Path multicast routing algorithms.

2.2. Oblivious routing algorithms

A deterministic routing algorithm that had been designed for the 2D Mesh [16] is called the XY routing algorithm. In the XY routing algorithm, the packets are first routed towards the X direction and afterwards in the Y direction to reach the destination. Although the algorithm is capable of preventing the deadlocks, it is not adaptive. The XY routing algorithm consumes more power for the torus based NoC, as compared to the mesh based NoC [16]. The basic working of the XY routing is presented in the Fig. 2.

Saad et al. [20] proposed an algorithm, called the “Power Efficient, Oblivious, Path-Diverse, and Minimal,” (POPM) Routing, which makes the routing decisions locally instead of establishing a fixed path. The packets are routed separately on a path having the minimum cost. The POPM distributes the packets at each of the intermediate nodes from the source and makes the next hop routing decisions that are based on the local information and the position of the destination node, in comparison to the current node. The algorithm is maximally diverse, due to the uniform probability of the path selection. The performance comparison with the DOR, North-Last, and PROM algorithms shows that the POPM algorithms not only outperforms in terms of the throughput and latency, but also consumes lesser power (by 10%).

Rahamti et al. [21] introduced a deterministic routing algorithm, called the “Torus Routing Algorithm for NoCs (TRANC),” which makes use of only one VC per physical channel. The algorithm is deadlock free and also significantly reduces the energy consumption for a mesh based NoC. The performance of the algorithm is no only equivalent to torus based NoC, using the XY routing but it also consumes 26% lesser power, as compared to the mesh based routing algorithm. Furthermore, the TRANC algorithm can be mapped easily on the partially adaptive and fully adaptive routing algorithm.

To overcome the issues of area and power overhead, a routing algorithm called the Region-Based Routing (RBR) is presented in [22]. The key feature of the RBR scheme is its capability to minimize the routing table entries that ultimately result in energy consumption while making routing decisions. Although the RBR is deterministic in nature, it can also be used with the dynamic routing strategies. The energy consumption of the RBR scheme was minimized by a factor of ten, when compared to the corresponding table based implementation.

Table 1 presents a brief analysis of the routing schemes in terms of the performance-improvement, energy efficiency, and congestion avoidance. Moreover, the heading “Time” represents whether the particular adaptive algorithm deals with the congestion in “Time”; whereas, “Local” and “Global” represent if the algorithm utilizes the information on the local or the global congestion, respectively. In Table 1 and the subsequent tables, the “✓” and “✗” symbols represent whether a particular characteristic/parameter is fulfilled or not, respectively; whereas, “-” represents that a particular characteristic/parameter is not discussed.

2.3. Routing and energy efficiency

Performing the routing tasks in the NoC, while improving the performance and energy efficiency, is imperative because of the ever-increasing network traffic. The study of the adaptive and deterministic routing algorithms for the NoCs, discussed in this survey reveals that energy consumption has direct association with the traffic on the network. The oblivious routing
algorithms, due to their pre-defined operational mechanisms, impose very low resource requirements and in terms of the energy consumption, their performance can also be predicted. The adaptive routing algorithms, on the other hand, are capable of routing the traffic dynamically that results in a change in the power state of the router components; thereby, reducing the energy consumption. Notably, the adaptive routing schemes dealing with the local traffic, contribute to significant reductions in the NoC power consumption. However, the energy consumptions of different routing algorithms differ because the implementations of the routing algorithms vary to different levels of granularity.

3. Buffered/bufferless NoC

The traditional NoC designs handled the contention among the flits (small pieces of network packets) by using the buffers (temporary memory areas) as queues in the router to stall the flits temporarily. The buffers were assumed to be necessary for each of the routers in the early NoC designs. The buffers help in improving the bandwidth efficiency, but are also the main culprits in consuming most of the energy in the NoCs. The buffers not only consume dynamic energy during the read/write process, but also consume static energy, when not in use[23]. Recently, techniques to design efficient buffers or to eliminate the use of buffers from the NoCs have remained a high trend in the research community. From the perspective of the energy efficiency, the bufferless techniques are suitable for the NoCs; whereas, the performance remains the goal for the buffered architectures. The hybrid techniques tend to find the delicate balance between the energy efficiency and the performance. We present the techniques recently published for the design of the buffered, bufferless, and hybrid NoCs.

3.1. Buffered NoC

Most of the work on the router/NoC is based on the idea of VC routing, proposed back in 1992. The idea behind the architecture is to use the buffers as a queue, for storing the incoming flits. If contention occurs among the flits, then the routing logic can select the heads of the queues in selecting the output and the loser of the contention will remain in the buffer. In case of a completely filled buffer, the router can put back pressure on the sending flit router to stop sending more flits. The VCs are used in the buffered router architectures, as multiple input queues. The proposed technique solves the problem of Head-of-line blocking, as even if the first flit in one queue is stalled/stopped, the flits in other queues can route to the other free ports and can utilize the outgoing links[24]. The technique was tested on 8 × 8 mesh topology.

Optimized versions of the bufferless architecture, BLESS (Bufferless)[23] and the buffered VC flow control architecture, have been compared in[25]. The authors have used two topologies 8 × 8 2D mesh and 2D FBFly during comparison. The SRAM buffers and empty buffer bypassing have been used to optimize the buffered architecture. The authors have shown that bufferless routing reduces the power consumption by only 1.5%, due to the buffer leakage power at very high loads. However, better performance (17% lower latency) and 17% lower power consumption, with 21% more throughput per unit power, have been observed at medium or high loads in the buffered networks. Similarly, the authors have also observed that the buffer power relates to the flit injection rate. At 7% flit injection rate, the buffered network is more energy-efficient. At the same time, the authors have shown that bypassing the empty buffer technique can help in reducing the energy consumption by a ratio of 8.5. As a result, the authors have concluded that by using the SRAM and bypassing the empty buffer techniques, the buffered network performs better; whereas, at very high loads, bufferless architectures consume marginally (5%) lesser energy.

Michelogiannakis et al.[26], named the Elastic Buffer (EB) router as the three-slot EBs, due to the use of three EBs within the network. The authors enhanced the design and presented two new designs for the routers of 8 × 8 2D mesh topology, the enhanced two-slot EB routers and the single-staged EB routers. The intermediate registers in the pipeline and the EBs at the output ports were replaced by the two-slot EBs to increase the throughput. The proposed design was further modified by merging the two stages into a single-stage EB to reduce the overhead. The first two-staged design reduced the area by 20% and the cycle time by 42%. The reduction in the stages was accompanied by a 29% reduction in the energy requirement per transferred bit, as compared to the two-staged EB routers. The results have also shown a 30% reduction in the occupied...
area as well. The authors have claimed that the three-slot EB routers [26], outperformed the enhanced two-staged and the single-staged EB routers, in terms of the energy consumption. The single-staged EB routers performed the best, in terms of latency. For the area requirements, the single-staged routers outperformed the others, when all of the three operated at the same frequencies. When operated at maximum frequencies, the enhanced two-staged EB routers covered a lesser area than the three-staged and the single-staged EB routers.

3.2. Bufferless NoC

The authors in [23], presented the idea of bufferless routing in the NoC architecture that is based on the notion of “hot-potato” routing in the network theory. The authors defined the bufferless routing as, “Bufferless routing is to route a packet (or a flit) to an output port, regardless of whether or not that output port results in the lowest distance to the destination of the packet.” The authors also proposed a simple, bufferless routing algorithm (BLESS) and had compared them with the buffered algorithms on 2D mesh topology. The results have shown that the bufferless routing results in an average energy-reduction of around 40%, without any significant impact on the performance. There was only a 3.2% decrease in the throughput. The covered area was also reduced by 60%. The BLESS has shown that the architecture of the NoC is different from the Internet or the Local Area Network (LAN), as the deflected packets eventually reach the destination and that is not the case with the Internet or the LAN. The evaluated latency for the BLESS, was also found to be too small that testified that the requirement of the buffers in the NoC router architecture is not necessary.

For the photonics NoC architectures, Kao et al. [8] presented a hybrid NoC router architecture design named the Buffered Clos Network (BLOCON). The BLOCON uses its own custom 64 × 64 topology. The BLOCON does not use the VC buffers and could thus reduce the power consumption and complexity. The BLOCON has been compared to the Buffered Photonic Clos Network (BPCN), Flattened Butterfly (FBfly), 2DMesh, and CMESHx2 networks, respectively. The results have indicated that the increase in the throughput for the BLOCON was about 128%, 116%, 43%, and 38%, as compared to the 2DMesh, CMESHx2, FBfly, and BPCN, respectively. The BLOCON has also consumed 62%, 61%, 60%, and 40%, lesser energy, as compared to the 2DMesh, CMESHx2, FBfly, and BPCN, respectively.

Bufferless routing has been criticized for the long critical path in port allocation, complexity of priority scheme for livelock freedom, the reassembly buffer requirement for packet fragmentation, and have been declared to be marginally energy-efficient, in case of very high loads in [25]. The bufferless router architecture, named the Cheap-Interconnect Partially Permuting Router (CHIPPER), had been proposed in [27] that could address the unaddressed problems, evaluated by the authors in [25] on 8 × 8 mesh topology. In the CHIPPER, the expensive port allocator and the crossbar have been eliminated that resulted in a reduced power per area-cost. The cost and the requirement of the reassembly buffers have been reduced by using the cache miss buffers, such as the MSHRs, for the purpose of reassembly. For the low to medium loads, the CHIPPER consumes 54.9% lesser network power, with 13% lesser average performance (multi-programmed) and 73.4% lower power with a 1.9% slow down (multi-threaded). The buffered area of CHIPPER was also reduced by 36.2%, as compared to the buffered counterpart. The CHIPPER was also compared to the other bufferless architectures, presented in [23]. The power consumption of the CHIPPER was found to be 8.8% lower. However, the performance of the CHIPPER was affected. The results indicated that the increase in the latency was about 13.6% and 9.6%, as compared to the buffered and bufferless architectures.

3.3. Hybrid approaches

The NoC architectures having complete buffer elimination used to be simple and power-efficient, but on the other hand, faced performance issues [28]. The buffered approaches improve the performance at the cost of the complexity and increased power consumption [28]. Recently, the research trend has shifted towards the hybrid designs to get more performance than the bufferless approach, but lesser energy consumption than the buffered approach.

A hybrid router architecture, named the Adaptive Flow Control (AFC), has been proposed in [29], which dynamically switches between the bufferless and the buffered router architecture of the NoCs. The authors used the terms “back pressured,” for the buffered routers and “back pressureless” flow control, for the bufferless routers. The bufferless routers do not put backpressure and should be termed as the “backpressureless” routers. The AFC is a threshold based, hybrid architecture for the NoC. The routers switch from the bufferless to the buffered mode, as soon as the link contention crosses the threshold, and vice versa. The proposed mechanism minimizes the energy consumption (and maximizes the performance) for the conditions with an unvarying (high or low) load. For the second mechanism, the gossip-induced mode, switches have been introduced to force the bufferless routers to switch to the buffered mode, when contention occurs at the neighboring buffered router. The proposed mechanism ensures correctness under varying load conditions. The third mechanism, the lazy VC allocation, exploits flit-by-flit routing in the AFC’s buffered mode, reduces the need for the buffers, and simplifies the VC allocation. The authors used 3 × 3 mesh topology during experimentations.

Lin et al. [28] presented a hybrid NoC router architecture design named the Buffered/Bufferless (BR/BLR) router. The authors proposed an idea that buffers should be removed from most of the routers. A reduction in the number of buffers improves the power efficiency; whereas, keeping a few necessary buffers intact, improves the performance. The question targeted in the research was, “the decision of how many routers should be deployed with the buffers, using a case study?” A seamless routing algorithm to address the question of how the bufferless routers will talk to the buffered routers was also presented in the paper. The authors used a 10 × 10 2D mesh topology for conducting experiments. The results show that the
BR/BLR reduces the power consumption and network latency by 40%, as compared to the purely buffered routers, whereas, the architecture reduces the network latency by 56% and the average deflection count by 87%.

A hybrid scheme proposed in [1] uses the EBs during the normal traffic load and switches to the VCs for buffering, if the traffic load crosses the pre-defined threshold. The proposed scheme, named the hybrid EB–VC, showed 21% and 12% more throughput per unit power than the VC routers and the EB routers respectively on 2D mesh and 2DFBFly topologies. However, 41% lesser throughput per unit area was noted by EB–VC in comparison to VC routers.

The bufferless routers are energy-efficient, but cause performance degradation. The bufferless routers either drop the packets in contention or forward through any open port. However, some of the studies claim that the energy consumption in the buffered routers is comparable to the bufferless routers, in low, average, and high traffic rates and the buffered routers consume slightly more energy (5%) in very high traffic rates [25]. Most of the recent researches have focused on the hybrid designs to target all of the types of traffic rates, increased throughput, and lower energy consumption. The trend in the hybrid buffered routers is towards utilizing the unused buffers in the NoCs with the routers, where, there is an extra need of the buffer and thus towards decreasing the overall demand of the buffers.

A comparison, encompassing all of the presented techniques, is presented in a form of taxonomy (Fig. 3) and in Table 2. The power consumption, latency, throughput, and area covered by the buffers, are taken as the comparison metrics in the literature and we have also used them for the comparison. Most of the energy-efficient designs are noted to cover a low buffered area on the NoC.

### 3.4. Buffered/bufferless routing architecture and energy efficiency

The energy consumption and performance of a router depend highly on the decision to use or not to use the buffers, in the router architecture of the NoC. Therefore, the researchers have focused on the bufferless versions of the NoC architectures. The works on bufferless router architectures, such as [8,23,27] reveal that these architectures have considerably reduced the power consumption of the NoC router architectures. However, the performances of the architectures have also degraded as a side-effect. Consequently, the trend in the research community has shifted towards the hybrid approach. In the hybrid approach, the researchers have combined the techniques of the bufferless architectures, along with those of the buffered ones. Moreover, in the hybrid techniques, the buffers are switched on or off, based on the traffic load [1]. Some of the hybrid techniques also focus on finding the optimum number of buffers to maintain the balance between the performance and the energy efficiency.

### 4. Fault tolerance

The number of faults in the NoC tends to increase with an increase in the number of on-chip components and reduced feature-size [30]. Permanent faults can occur during manufacturing or as a result of the wearing out of the components/links.
making the components/links unusable. The transient faults or soft errors are introduced during the transmission of the data and mostly affect the data in transit [10]. These faults can drastically decrease the performance and increase energy consumption of the chip. Any erroneous link can isolate the working components of the chip. A single link failure can cause the entire chip to discontinue communication or create a situation of deadlock [30]. Similarly, the faulty components can create isolated areas on the chip. The faults can arise in various components, such as the links, processing elements, memory units, routers, and may also be caused by the routing algorithms, elevating the reliability issues in the NoC [30]. A large number of fault-tolerant schemes have been developed to maintain the connectivity and correctness of the data, in the presence of permanent and/or transient faults. Fault-tolerant flow control schemes ensure the on-chip transmission by introducing link-level fault recovery between the routers. Likewise, fault-tolerant routing increases the performance of the chip by safeguarding the transmission of messages within the on-chip network and eliminating the deadlocks. In this section, our discussion is limited to the fault-tolerant flow control schemes and fault-tolerant routing strategies.

4.1. Fault-tolerant flow control schemes for the NoC

Yu and Ampada [9] used the benefits of hop-to-hop and end-to-end Error Control Coding (ECC) covering the changing noise conditions. The mutual cooperation between the network layer and data link layer reduces the energy consumption. Initially, the network layer conducts end-to-end ECC with a separate unit that collects the information about erroneous packets. On crossing the threshold level (in terms of number of faulty packets), hop-to-hop ECC at the data link layer is also enabled. Records of the flit (flow control digits) error detection are maintained that are used to decide when to switch back to the single-layer mode. The error detection algorithm switches to the dual-layer mode in the case of high noise conditions and to the single-layer mode in the case of low noise conditions. The results showed a 72% reduction in the energy consumption and a 64% improvement in the average latency.

Yu and Ampada [31] presented an error-management technique for both, the transient, as well as the permanent errors. The presented scheme works cooperatively at the data link and physical layers, and operates in two different modes. While mode one consists of the simple ECC, the second mode uses the composite ECC with packet re-organization. Mode one is used for the low noise conditions. For the high noise conditions and multi-bit errors, mode two is used. The unexploited surplus wires that are held in reserve for parity checking are utilized as spare wires. No added spare wires are required in the proposed scheme. If the number of faulty links surpasses the available spare links, then the split transmission and packet re-organization are performed. The authors also formulated an algorithm for the packet re-organization. Simulation results show up to a 70% reduction in the energy consumption on average. Nevertheless, the aforementioned scheme gives diminished throughput for the low transient error rates and small packet-size.

Vitkovskiy et al. [32] presented an error detection and correction technique that deals with both the transient, as well as the permanent errors. If any wire fails, then the aforesaid scheme does not label the whole link as faulty. Instead, the link is marked as a Partially-Faulty Link (PFL). The traffic on the remaining functional wires of the PFLs continues for upholding the network connectivity; thereby, allowing for graceful degradation of the link. With the increase in the number of faulty wires, the link is marked as a Fully-Faulty Link (FFL). Although capitalizing on a PFL requires multiple cycles for the flit transmission, it avoids the overhead of multiplexing, pipelining of supplementary routers, and misrouting. The shown scheme provides an increased throughput, but the energy overhead increases up to 26.26%, when used with the synthesized NoC routers.

Crosstalk and multiple errors are collectively handled in the Joint Crosstalk Avoidance and Triple Error Correction (JTEC) code, presented in [10]. During the encoding process, k-bits are encoded by using the Single Error Correction (SEC) Hamming code. Later, every encoded bit is duplicated and the parity bit is calculated over the encoded bits. The original bits, encoded bits (two copies, encoded and duplicated), and the parity bits are sent to the destination. At the decoding side, the parity bits for both of the hamming copies are calculated. If the parity bit of the first hamming code copy is non-zero, then the corresponding hamming copy can have one or two errors. The number of errors is determined by comparing the calculated parity bit with the received parity bit. If the calculated parity bit is non-zero, then the hamming copy can have three or no errors. Depending on the number of errors, one out of the two hamming code copies is chosen and the SEC decoding is applied for a corrected output. The duplication of the hamming code reduces the transitions of converse bits in the adjoining wires and reduces the crosstalk over the medium. The simulation results show energy-savings of up to 43.6%, as compared to the schemes, such as the Dual Rail, Duplicate Add Parity, Boundary Shift Code, and Modified Dual Rail. However, the aforementioned scheme has slight latency and area overhead.

Tables 3 and 4 present the analysis of the important characteristics of the presented techniques. Both of the tables implement the energy, latency, and area-costs of the discussed techniques, as the comparison parameters because a change in any of the three mentioned parameters affect the others. The conventional fault tolerance techniques in the communication systems engineering are hindered by area constraints for their implementation in the NoC. Therefore, these techniques have to be modified and tailored for the NoC design [10]. The fault tolerance techniques/algorithms must minimize the impact on the latency (for the packet transmission), so that the throughput of the systems is not degraded. The covered area and processing time (processing of the fault tolerance scheme that affects latency) directly affect the energy consumption of the system [33]. Therefore, Tables 3 and 4 compare the encountered latency, area overhead, and energy consumption, by different techniques.
4.2. Fault-tolerant routing

The “Neighbor Aware Turn Model-Based Fault Tolerant Routing for NoCs” (NARCO) utilizes the Odd–Even (OE) and Invert Odd–Even (IOE) models, for avoiding deadlock in routing [33]. In a 2D Mesh network, the OE turn model does not allow a packet at an even numbered column to take an east-north or an east-south turn. Similarly, a packet at an odd numbered column cannot take a north-west or a south-west turn. The IOE turn model works with the opposite directions. The abovementioned scheme uses replication, only if the fault rate is above a certain threshold level. The original packets are sent with the OE turns and replicated along the IOE turns. The results show a higher packet arrival rate for the NARCO. The NARCO has lower energy consumption for the lower fault rates and vice versa.

Patooghy and Miremadi simulated the DOR and duatos’ routing algorithms [34]. The simulation results under the various traffic conditions revealed that during the routing process, there are certain channels that are underutilized while some of the channels have a high traffic load. The two routing algorithms are said to be complementary, if both of the algorithms have disjoint underutilized and over utilized paths. The authors utilized the complementary algorithms to introduce the complement routing methodology for having reliability and fault tolerance in the NoCs. In the complement routing, one routing algorithm is used to route the original packets and the complementary routing algorithm forwards the redundant packets. The DOR and duatos’ routing algorithms are used as the base complementary algorithms in [34]. The scheme is compared with the flood routing and the results show that the complement routing scheme exhibits lesser power overhead, while maintaining the reliability of the flood routing schemes.

A redundancy approach to both, the routers, as well as the links for fault tolerance was adopted in [35]. The first row of the routers in a 2D Mesh, are declared as the spare routers. Every PE is connected to two adjacent routers, introducing redundancy in the links. If any router fails, then it is replaced by the adjacent upper router. The scheme addresses the problems of the isolated PE and faulty regions. Moreover, the authors claimed that the scheme is transparent to the routing algorithms, requiring no reconfiguration. Nonetheless, the dense use of spare hardware increases the energy demands of the NoC.

The authors in [36] presented the Reconfigurable, Adaptive and Fault-Tolerant (RAFT) routing algorithm for the NoCs that is not only low-cost, but also has the capability of avoiding the faulty paths in the mesh based architectures. Moreover, the proposed algorithm is congestion-aware that allows only two channels for fault tolerance and for adapting the routes dynamically. Furthermore, the proposed algorithm can accommodate the other adaptive routing approaches to distribute the traffic dynamically. The performance of RAFT was compared to the DyXY in terms of the hardware overhead and power consumption. The RAFT results in slight hardware and power overheads of 1% and 2.2%, respectively.

4.3. Energy efficiency and fault tolerance

The NoC is the primary adopted communication infrastructure for the SoCs. Due to the limited nature of the resources, such as the energy and area, the NoCs require specialized techniques for every aspect of their operation, including error handling and fault tolerance. The importance of energy consumption in the NoC is elevating as the technology scales down. For the NoC to work properly, the fault-tolerant schemes utilize extra processing, transmission, and/or redundancy in the hardware. Therefore, the design of fault-tolerant schemes affects the energy requirements of the NoC. Lesser energy consumption leads to decreased temperature of the fabric, resulting in increased reliability. Appropriate fault recovery at the link level decreases the necessity of the retransmission buffers; thereby, decreasing the energy consumption. Consequently, the schemes discussed in [9,31] showed a greater reduction in the energy usage. The schemes presented in [10] showed a lesser cut down in the energy consumption, as compared to the ones discussed in [9,31], because of the higher levels of redundancies. The partially correct links can be utilized up to their strength that reduces the redundancy in the hardware, resulting in a reduction of the power consumption. Smaller and simplified components and algorithms add to the energy efficiency. However, Tables 3 and 4 reveal that most of the fault tolerant schemes did not show a positive impact on the area of the chip. The information redundancy results in the use of extra hardware, thereby, increasing the size. Nevertheless, the schemes that do not add extra circuitry for fault tolerance [31] resulted in area reduction. Likewise, the schemes that are operating at different communication layers ([10]) are more complex. Therefore, they eventually increase in latency. In sum, the schemes involving time redundancy are more prone to latency overhead [37]. For coding schemes the overhead depends on the complexity and implementation of the technique [37].

<table>
<thead>
<tr>
<th>Work</th>
<th>Type of fault(s)</th>
<th>Latency-improvement</th>
<th>Energy efficiency</th>
<th>Area-reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>[9]</td>
<td>Permanent</td>
<td>✔</td>
<td>✔</td>
<td>✗</td>
</tr>
<tr>
<td>[31]</td>
<td>✗</td>
<td>✔</td>
<td>✔</td>
<td>✗</td>
</tr>
<tr>
<td>[32]</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>[10]</td>
<td>✔</td>
<td>✗</td>
<td>✗</td>
<td>✔</td>
</tr>
</tbody>
</table>

Please cite this article in press as: Abbas A et al. A survey on energy-efficient methodologies and architectures of network-on-chip. Comput Electr Eng (2014), http://dx.doi.org/10.1016/j.compeleceng.2014.07.012
5. Energy-efficient switching techniques in NoC

Switches are the building blocks of interconnections on a network. The design of switches highly affects the performance and energy efficiency of NoC. The latency and throughput of the network is also dictated by the micro-architecture of the switches. As the technology has moved to the NoC, the interconnects have become a focus with the change in constraints and trade-offs, consequently evolving a more pronounced switch design in terms of the latency and power consumption [3]. We present an overview of the different switch design techniques particular to the NoC, with the energy efficiency, latency, and area, being the primary metrics of focus.

5.1. Switching techniques

Dimitrakopoulos et al. [3] proposed a technique, named the “Merged Arbiter and Multiplexer (MARX),” to design efficient, high-radix, and reduced-latency switches. The MARX performs the functions of the arbiter and the multiplexer in parallel, and adapts to complex arbitration policies, along with the simple round-robin in an efficient manner. The transition from the simple to complex arbitration policies is achieved without any significant overhead in terms of the cycle time. Therefore, the MARX optimizes the functionality of the arbiter and multiplexer. The proposed technique is implemented in two ways: (a) the use of Fixed Priority Arbiters (FPAs) to build switch allocators that leads to a faster arbiter and multiplexer, giving delay-efficient solutions and (b) the use of dynamic arbitration policies in the MARX results in better area/energy efficiency, as compared to the separate arbiter and multiplexer.

The authors in [38] proposed a switch allocator that could operate with bufferless switches, eliminating the serial dependency of the output port allocation. In the proposed scheme, the routing algorithm determines the output port of a received flit, based on the destination address. The flits are assigned weights proportional to the time spent by the flit in the network. At each of the output ports, the flit with the maximum weight is selected for deflection. An input port having a flit, ready for deflection, generates a Deflection Request (DR) signal that is passed to another circuit called the load distributor. The load distributor deflects the input to an appropriate output port, in a way to maximize the utilization of the output ports. According to the authors, the delay of the proposed switch allocator is \( \log_2 n + n \) and is less than that of the previously-proposed schemes. Due to the lower delay, the proposed scheme is power-efficient.

El-Moursy et al. [39] discussed the asynchronous switching, with the aim to achieve low power consumption in the NoC. The “Globally Asynchronous Locally Synchronous (GALS)” method is used to minimize the global clocking network that in turn, minimizes the power consumption of the whole communication network. Additionally, the system reliability is increased by using GALS method and eliminating the global clock distribution. The power consumption of the asynchronous system is reduced by 70.8%, at the cost of increased (50% more) area consumption, as compared to the synchronous counterparts. Three network topologies, the Butterfly Fat-Tree (BFT), ClicHE, and Octagon, are used and results are compared. The ClicHE is the most efficient of the three with reduction in power dissipation by 75%.

Ref. [40] proposed a design that performs the function of an arbiter and a multiplexer in parallel, and simplifies the design of the high-radix and reduced-latency switches. The Dynamic Priority Arbiters (DPAs) and multiplexers are merged together in [40]. The RTL soft macros used in the design, increase the efficiency of the NoC switches and can adjust to both the simple as well as the more complex arbitration policies. The delay, area, and energy-efficient, implementations are achieved with the proposed scheme.

The operation of a switch can be affected in several ways by errors that occur in an arbiter’s logic. An on-line checking system was proposed for the arbiters of the switch allocator that detects and rectifies every transient and permanent error in the arbiter [41]. The on-line checker adds reliability to the operation of the switch. The proposed low-cost error protection scheme covers all of the fault scenarios, incurs minimum costs in terms of the delay and area overhead, achieves energy efficiency, and gives high-performance switches.

Nema et al. [42] proposed a NoC router using the wormhole switching. The wormhole switching is a flow control process in which the packets are disintegrated into small flits and the flits are then delivered using pipelining. The buffering requirements for the proposed scheme are reduced and the performance is enhanced. The average latency is minimized and the power consumption is reduced using the proposed approach.

Hybrid packet/circuit switching NoC with source synchronous operation, hybrid packet/circuit switched flow control and ultra-low-voltage optimization is fabricated in CMOS [43]. The proposed design achieves 93% circuit-switched latency and 55% improvement in energy efficiency. Table 5 summarizes the achievements of the works discussed in the current section.

---

Table 4
Comparison of the presented fault-tolerant routing schemes.

<table>
<thead>
<tr>
<th>Work Type of fault(s)</th>
<th>Latency-improvement</th>
<th>Energy efficiency</th>
<th>Area-reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Permanent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transient</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[35]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>[33]</td>
<td>✓</td>
<td>✓</td>
<td>–</td>
</tr>
<tr>
<td>[34]</td>
<td>✓</td>
<td>✓</td>
<td>–</td>
</tr>
<tr>
<td>[36]</td>
<td>✓</td>
<td>✓</td>
<td>–</td>
</tr>
</tbody>
</table>

---

Please cite this article in press as: Abbas A et al. A survey on energy-efficient methodologies and architectures of network-on-chip. Comput Electr Eng (2014), http://dx.doi.org/10.1016/j.compeleceng.2014.07.012
5.2. Switching and energy efficiency

The switch design highly affects the performance and energy efficiency of the whole network. With the advent of the NoC technology, switch design has become a focus in terms of the performance and energy efficiency. The switches, being the main building blocks of the interconnections on a network, highly affect the performance and energy efficiency of the whole system. The latency, power consumption, and throughput, of the network are also dictated by the micro-architecture of the switches. As the technology moved to the NoC environment, the interconnects became a focus as their constraints and trade-offs changed, making the switch design more pronounced in terms of the latency and power consumption [3]. We observe from the above discussion that by decreasing the size of the interconnects, the power consumption can also be reduced. Moreover, when the area of the switch is reduced, the delay is also reduced. Therefore, the energy efficiency is achieved by reducing the interconnect-size and the associated delay.

6. Power management

Increasing the number of cores on a single chip leads to the issue of power consumption [44]. Changing the workload conditions can cause bungling power consumptions, resulting in the temperature hot-spots [45]. The hot-spots can in turn degrade the performance and the lifetime of the chip [12]. Therefore, the power management is a key issue towards the performance and energy efficiency of the NoCs. While the Voltage Islands (Vis) rely on partitioning a circuit into islands that operate at specific voltage levels for lowering the power consumption [12], the DVFS, on the other hand, attempts to optimize the power consumption by lowering the operational voltage and clock-speed of a processor in real-time [11]. In Sections 6.1 and 6.2 the techniques pertaining to the aforementioned topics are reviewed.

6.1. Voltage Islands

The power dissipation in the NoCs can either be static or dynamic. The static power dissipation occurs due to the leakage current, whereas the dynamic power dissipation occurs during the switching activities. One of the techniques that can reduce the consumptions of the static and dynamic power (in the NoC) is the utilization of the Vis. The voltage islanding not only improves the power efficiency, but also the area overhead of the chip. The MSV allows for the partitioning of the whole circuit into different Vis [12]. Each of the islands operates at a specific voltage level. The techniques to reduce the static and dynamic power consumptions based on the Vis are presented below.

The authors in [46] proposed two algorithms: (a) the MCF and (b) the “Value-Oriented Branch-and-Bound” (VOBB), to solve the voltage assignment problem. During the floor-planning stage, the authors proposed a “MSV-driven floorplanning” to optimize the physical layout and power consumption of the circuits. The power consumption was reduced by 32.96%. To optimize the power consumption due to leakage, the authors in [45] concentrated on two areas, namely: (a) the voltage level assignment and (b) the Level-Shifters (LSs) assignment. The authors assigned the voltage levels and the LSs, during the floor-planning. The post-floor-planning stage considered the “White Space Redistribution” (WSR). Two frameworks were proposed in [45] to assign the voltage and the LSs, respectively: (a) the “convex cost network flow” algorithm and (b) the MCF algorithm. The LSs were inserted into the interconnects to reduce the leakage currents in the circuits. The design achieved an energy efficiency of 17%.

Kapadia et al. [47] introduced a different “core-to-tile” mapping, which consists of three variants: (a) the “incremental swapping”, (b) the “Probabilistic Incremental Swapping”, and (c) the “Incremental Swapping using Branch & Bound.” The proposed technique lowered the communication power by 32% and the total power dissipation by 13%.

The IP and hierarchical design are commonly used for the VI design. The authors in [44] presented a scheme that uses the “greedy heuristic” technique for assigning the voltage to the subsets of the IP blocks. After assigning the voltage to the subsets of the IP blocks, the floor-planning is performed. The technique reduces the power consumption by 9.4% and the area overhead by 8.7%. In [48] the complication of voltage assignment in the 3D SoCs, due to MSV is handled by using Integer Linear Programming (MILP) model for voltage Island generation. The power consumption is reduced by 16% by this method and also the performance is increased by 20%. Table 6 considers the energy, area, and latency overheads, as the comparison

Table 5
Summary of the energy-efficient switching techniques for the NoCs.

<table>
<thead>
<tr>
<th>Work</th>
<th>Switching technique(s)</th>
<th>Reduced latency</th>
<th>Reduced area</th>
<th>Energy-efficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>[3]</td>
<td>MARX</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[38]</td>
<td>Bufferless switch allocation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[39]</td>
<td>Asynchronous switching for low-power NoC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[40]</td>
<td>DPA and multiplexer</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[41]</td>
<td>Low-cost, fault-tolerant switch allocator</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[42]</td>
<td>Wormhole switching in NoC router</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[43]</td>
<td>Hybrid packet/circuit switching NoC</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Please cite this article in press as: Abbas A et al. A survey on energy-efficient methodologies and architectures of network-on-chip. Comput Electr Eng (2014), http://dx.doi.org/10.1016/j.compeleceng.2014.07.012
metrics. The following acronyms are used in Table 6: “EE” (Energy-Efficient), “RA” (Reduced Area), and “RHC” (Reduced Hop Count).

The VIs contribute significantly in reducing the power consumption and enhancing the performance of the NoCs. We presented different voltage islanding techniques for the NoCs. From the above discussion, it is evident that the VIs reduce the power consumption, area overhead, and hop count.

6.2. Voltage and frequency scaling in NoC

The DVFS plays a key role in the power management of the SoCs/NoCs. The DVFS is an efficient technique that saves the power by lowering the operational voltage and clock-speed of a processor in real-time [11]. The key concept used in the implementation of the DVFS is providing the circuit with only the sufficient clock frequency and voltage that it requires to operate. The implementation of the DVFS can be fragmented into two parts: (a) the adjustment of the required frequency and (b) the scaling of the applied voltage to a core on the basis of the workload being processed. In the modern CPU, the processor stalls when a data request is not entertained. The stall is an indication of the memory bound phase and is an area of focus pertaining to the energy conservation of the processors. The reduction in the power consumption of the electronic devices is desirable because it decreases the temperature of the devices and increases the time elapsed between the charges [4].

The DVFS is an operative technique used to improve the energy efficiency of the microprocessors. The DVFS achieves the energy efficiency by minimizing the applied voltage and clock-speed, when the workload decreases. When the applications are not sensitive to the frequency provided, for instance, in the memory bound state, the DVFS can be applied to reduce the power budget. As the voltage and frequency are assumed to be linearly dependent on each other, reducing the voltage reduces the frequency. The phenomenal property of balancing the trade-off between power consumption and performance has made the DVFS, a de facto choice in the multi-core designs.

David et al. [12] exploited the above mentioned benefits of the DVFS in the design of the Single-Chip Cloud Computer (SCC). The power management using the Voltage/Frequency Islands (VFI) was performed in real-time and the result assured that the aforementioned approach is almost two times better, as compared to the peer energy-saving techniques.

Wang et al. [49] proposed an adaptive power control algorithm for the implementation of the DVFS. Moreover, the temperature control and cache re-sizing enhance the efficacy of the proposed algorithm. Nevertheless, no power cap is implemented to confine the maximum power drawn.

For the single-threaded applications, a control knob, along with the DVFS is used to maximize the efficiency of the multi-core architecture. The thread packing specifies the quantity of threads running per core and uses the power cap mechanism to maximize the performance for the multi-threaded workloads with a varied number of cores. The thread reduction, coupled with the power cap mechanism, limits the maximum power that a particular core can use at any given instance of the time. Seo et al. [11] had also adopted the power capping mechanism using the DVFS to restrict the power consumption in the mobile devices and achieved a 10% saving in the energy consumption.

Jin et al. [4] accomplished 61.69% saving in the energy consumption, using the Hilbert Transform and the DVFS to implement the workload estimation model. A high accuracy and a negligible deadline miss ratio were attained. Nonetheless, the proposed design is only viable for the video encoding applications in the mobile devices.

Gorti and Somani [50] developed a thermal and performance management technique to improve chip lifetime and leverage application awareness. Chip heating is overcome by the intelligent selection of voltage-frequency (V, F) pair using the DVFS. The level of the (V, F) pair is chosen in accordance with the operating point (OP) of the core. Significant performance improvement and energy savings are obtained using the aforementioned scheme. Moreover, the reliability of the chip is increased by a factor of sixteen fold.

Table 7 exhibits a comparative analysis of the research works presented in this section. The comparison is based on the key metrics that are crucial for all of the DVFS techniques to achieve a high throughput and a smooth operation. The acronyms used in Table 7 are: “EI” (Energy Improvement), “VA” (Variation-Aware), “PDVFS” (Per-Core DVFS), “ML” (Multiple V/F Lines), “PG” (Performance Gain), “PC” (Power Cap), “AR” (Area-Reduction), and “LMT” (Latency-Minimization in V/F Transition).

6.3. Energy efficiency and power management

The hazard of the increased power consumption makes the energy efficiency a key metric to be considered in the design of the future-generation multi-core NoC architectures. As the CMOS devices are scaled down to the 45 nm technology and

Table 6
Summary of energy-efficient voltage islanding techniques in the NoCs.

<table>
<thead>
<tr>
<th>Work</th>
<th>Feature(s)</th>
<th>EE</th>
<th>RA</th>
<th>RHC</th>
</tr>
</thead>
<tbody>
<tr>
<td>[44]</td>
<td>IP blocks divided into sub-blocks</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[46]</td>
<td>MCF, VOBB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[45]</td>
<td>Voltage and level-shifter assignment driven floorplanning</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[47]</td>
<td>Incremental swapping, probabilistic incremental swapping, and incremental switching using branch and bound</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[48]</td>
<td>Mixed integer linear programming model (MILP)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Please cite this article in press as: Abbas A et al. A survey on energy-efficient methodologies and architectures of network-on-chip. Comput Electr Eng (2014), http://dx.doi.org/10.1016/j.compeleceng.2014.07.012
advancements are made towards the 8 nm technology node, shrinking the power budget has become a challenge. The core count per chip has increased to a significant level, but due to the power constraint, it becomes difficult to power all of the cores at the highest speed. One solution is to restrict the power consumption by setting the cores to the sleep mode, when they are idle for a specific amount of time using the DVFS. Nevertheless, in the active mode, to reduce the power and thermal budgets, alongside sustaining the improvements in the performances, the DVFS needs to be explored further to refrain the energy consumptions.

7. Discussion

The review of various techniques and architectures in the above sections testifies that the NoC significantly enhances the performance characteristics of the CMPs. As can be observed that energy consumption is directly associated with the network traffic. Effectively designing the routing algorithms not only improves the NoC performance but also improves the energy efficiency. Energy consumption and the router performance are also highly dependent on the decision of whether to use the buffers or not. Bufferless architectures though being able to minimize the energy consumption substantially come across with the performance degradation. Similarly, appropriate fault recovery at the link level alleviates the need of the retransmission buffers, consequently decreasing the energy consumption. However, fault tolerant schemes employing higher redundancy do not decrease energy consumption significantly. Moreover, the switch design also highly affects the performance of the system in general and the energy consumption in particular because the latency, power consumption, and throughput, of the network are also dictated by the micro-architecture of the switches. The voltage islanding and the DVFS equally prove effective to optimize the power consumption by partitioning circuits into islands and by lowering the operational voltage and clock speed of the processors. Although the voltage islanding and DVFS contribute significantly in reduced power consumption, they suffer in terms of area overhead, latency, flexibility, and computational time overhead.

In this survey, we reviewed the performance of approaches pertaining to different subdomains of the NoC from individual standpoints. Moreover, in the literature we seldom find the proposals that discuss the effects of combining the subdomains on the NoC performance improvement in general and energy efficiency in particular. Consequently, there is a significant need to investigate about how the energy efficiency can be improved by combining different subdomains of the NoC. For example, one possible direction of the future research on NoC energy efficiency may be to conduct a benchmark study with the focus on routing techniques and bufferless routers. The routing algorithms tend to improve the performance and reduce the congestion. The bufferless routers have reduced the power consumption by removing the buffers. Without a benchmark study it is hard to predict which scheme is better than the others. A bufferless router can also be more efficient than a fault tolerant scheme because a fault tolerant scheme requires extra calculating power that will cost more energy. At the same time, distance between different elements on the network is too small that will not incur any significant cost in the retransmission even if the packets are lost in case of bufferless routers in NoCs. All these assumptions require detailed studies to label any subdomain truly better than the others as such studies are missing till date. Topologies also significantly affect the power consumption of the NoC. The architectures in this survey largely are based on the topologies, such as ring, mesh (2-D and 3-D), tree, torus, and BFT. However, we make an important that most of the regular topologies, such as the 2D meshes, ring, and tree network are not much dependent on the characteristics of the underlying systems. Therefore, their performance in terms of energy consumption and power utilization is not consistent throughout. The bufferless routing that helped in decreasing the energy consumption of NoCs is mostly implemented on mesh topologies due to its simple design. However, any network topology that satisfies the following two constraints can be implemented with bufferless routing to decrease energy usage: (i) every router should have same number of input and output ports and (ii) each router should be reachable from every other router. Trees, Hypercubes, Mesh, and Torus satisfy the criteria, whereas, Butterfly networks do not satisfy the criteria [23].

8. Conclusions

Designing energy efficient methodologies for various NoC domains, such as the routing algorithms, buffered and bufferless router architectures, fault tolerance, switching techniques, VIs, and voltage-frequency scaling significantly affects the NoC performance. Besides the detailed discussion and comparisons of various approaches pertaining to the aforementioned NoC domains, we also highlighted several future research directions. We believe that the study will be beneficial for the researchers in finding the most recent research works carried out on the NoC in a consolidated form.

Table 7

Comparative analysis of various DVFS techniques.

<table>
<thead>
<tr>
<th>Work</th>
<th>EI</th>
<th>VA</th>
<th>PDVFS</th>
<th>ML</th>
<th>PG</th>
<th>PC</th>
<th>AR</th>
<th>LMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>[12]</td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
<td>✔</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>[49]</td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[4]</td>
<td>61.69%</td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[11]</td>
<td>10%</td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[50]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
References


Assad Abbas is a PhD candidate at the Department of Electrical and Computer Engineering, North Dakota State University, USA. His research interests include cloud computing, information systems, and information retrieval.

Mazhar Ali is a PhD student at the Department of Electrical and Computer Engineering, North Dakota State University, USA. His research interests include information security, formal verification, modeling, and cloud computing networks.

Ahmad Fayyaz is a PhD student at the North Dakota State University, Fargo, ND, USA. His research interests are mainly but not limited to digital system design, asynchronous circuit design using NULL Conventional Logic.

Ankan Ghosh is pursuing his MS degree from the North Dakota State University, Fargo, ND, USA. His research interests include the wired and wireless networks, delay tolerant networks, mobile, and telecommunication networks.

Anshul Kalra is a PhD candidate in the Department of Electrical and Computer Engineering, at the North Dakota State University, USA. His research interest include optical communication systems, fiber lasers, and optical coherence tomography.

Samee U. Khan is Assistant Professor of Electrical and Computer Engineering at the North Dakota State University, Fargo, ND, USA. His research interests include optimization, robustness, and security of: cloud, grid, cluster and big data computing, social networks, wired and wireless networks, power systems, smart grids, and optical networks. His work has appeared in over 225 publications.

Muhammad Usman Shahid Khan is a PhD candidate at the North Dakota State University, USA in the Department of Electrical and Computer Engineering. His areas of interest are recommendation systems, data mining, cognitive radio networks, and network security.

Thiago D.S. De Menezes is a PhD Student at the Department of Electrical and Computer Engineering, North Dakota State University, Fargo, USA. His research interests include fiber optics communications regarding advanced modulation formats and robustness to fiber effects such as PMD, photonic network on-chip, optical processing, and DSP.

Sayantica Pattanayak is pursuing her MS degree from the Department of Electrical and Computer Engineering, North Dakota State University, Fargo, USA. Her research interests include cryptography and embedded systems.

Alarka Sanyal is pursuing his MS degree from the North Dakota State University, Fargo, ND, USA. His research interests include the radio frequency integrated circuits, VLSI, and biomedical engineering.

Saeeda Usman is pursuing her PhD, from the North Dakota State University, Fargo, USA. Her research interests include energy efficiency systems, performance improvement in digital circuits, and robustness in cloud computing.