Performance Evaluation of Wireless NoCs in Presence of Irregular Network Routing Strategies

Paul Wettin, Jacob Murray, Ryan Kim, Xinmin Yu, Partha Pratim Pande, Deukhyoun Heo
School of Electrical Engineering and Computer Science
Washington State University
Pullman, USA
{pwettin, jmurray, rkim, xyu, pande, dheo}@eecs.wsu.edu

Abstract—**The millimeter (mm)-wave small-world wireless NoC (mSWNoC) is an enabling interconnect architecture to design high performance and low power multicore chips. As the mSWNoC has an overall irregular topology, it is extremely important to design suitable deadlock-free routing mechanisms for it. In this paper we quantify the latency, energy dissipation, and thermal profiles of mSWNoC architectures by incorporating irregular network routing strategies. We demonstrate that the latency, energy dissipation, and thermal profile are affected by the adopted routing methodologies. In presence of the benchmarks considered, the variation in latency and energy dissipation is small. However, the network hotspot temperature can vary considerably depending on the exact routing strategy and the characteristics of the benchmark.**

Keywords—millimeter-wave wireless; network-on-chip; small-world; irregular; routing algorithms

I. Introduction

The wireless NoC (WiNoC) is capable of improving the latency, energy dissipation profile, and thermal profile of its traditional wireline counterparts [1]. The power law connectivity-based mm-wave small-world WiNoC (mSWNoC) principally is an irregular network topology [2]. Routing in irregular networks is more complex, because the routing methods need to be topology agnostic. Hence, it is necessary to investigate suitable routing mechanisms for mSWNoCs.

Routing algorithms for irregular network-based traditional large-scale parallel computers have been investigated in the past [3]. A well-known method for generating routing algorithms in irregular networks is the up/down strategy. A specific type of up/down routing algorithm that has been utilized for mSWNoCs is the Tree-based Routing Architecture for Irregular Networks (TRAIN) algorithm [1], [4]. A Minimum Spanning Tree (MST) of the network is created with a randomly selected node as the root, and data is routed along the MST. An allowed route never uses a link in the up direction along the tree after it has been in the down path once. Hence, channel dependency cycles are prohibited, and deadlock freedom is achieved [4]. However, a well-known weakness of this routing scheme is that it has a strong tendency to generate hotspots around the root of the tree structure. Another routing strategy that has been investigated for irregular networks is known as layered routing. The network resources are divided into layers and network deadlocks are avoided by preventing portions of traffic from using specific layers [5].

It can be conjectured that either up/down or layered routing with appropriate modifications necessitated by the use of wireless channels will suit the mSWNoC well. However, the design constraints of an mSWNoC are different than those in traditional parallel computing platforms. First, in mSWNoC, we should not only target higher performance; optimizing the NoC solely for performance may give rise to thermal imbalance. Also, in a traditional parallel computing system, the performance evaluation is carried out using synthetic traffic patterns, like uniform random, bit permutation, etc. [6]. The comparative performance evaluation between multiple routing algorithms is predominantly done in terms of achievable saturation throughput [5]. Conversely, NoCs in presence of general-purpose chip multiprocessor (CMP) benchmarks, such as SPLASH-2 and PARSEC, operate below network saturation and their traffic density varies significantly from one benchmark to the other. Hence, instead of looking at saturation throughput, we should quantify the associated network latency, energy dissipation, and thermal profiles in presence of frequently used non-synthetic benchmarks.

In this paper our aim is to undertake a detailed performance evaluation for the mSWNoC architecture by incorporating suitable irregular network routing strategies. We consider network latency, energy dissipation, and thermal profile as the relevant metrics in this performance evaluation. We demonstrate that depending on the specific benchmarks, the latency-energy-temperature trade-offs vary and the distribution of the on-chip traffic pattern has an important role to play.

II. Related Work

The limitations and design challenges associated with existing NoC architectures are elaborated in [6]. Conventional NoCs use multi-hop, packet switched communication. At each hop the data goes through a complex router/switch, which contributes to considerable power, throughput, and latency overheads. To improve performance, a methodology to automatically synthesize architectures with a few application specific long-range links inserted in a regular mesh was proposed in [7]. Subsequently, performance advantages of NoCs by insertion of long-range wireline links following principles of small-world graphs were elaborated in [8]. The concept of express virtual channels is introduced in [9]. Despite significant performance gains, in the above schemes the long-range links are designed with conventional wires. It is already shown that beyond a certain length, wireless links are more energy efficient than conventional metal wires. Hence, the performance improvements by using long-range wireless links will be more than that using wireline links [10].

This work was supported in part by the US National Science Foundation (NSF) grants CCF-0845504, CNS-1059289, and CCF-1162202, and Army Research Office grant W911NF-12-1-0373.

978-3-9815370-2-4/DATE14/©2014 EDAA

A comprehensive survey regarding various WiNoC architectures and their design principles are presented in [11]. WiNoC architectures can be divided into two sub categories, viz. mesh with wireless links inserted on top of it, and hierarchical architectures with long-range wireless shortcuts. Among the first category, notable examples include design of a WiNoC based on CMOS ultra wideband (UWB) [12], 2D concentrated mesh-based WCube architecture using sub-THz wireless links [13], and the inter-router wireless scalable express channel for NoC (iWISE) architecture [14]. Possibilities of creating novel architectures aided by the on-chip wireless communication have been explored in [10] and [15]. These two works proposed design of hierarchical and hybrid WiNoC architectures using long-range wireless shortcuts. The whole system is partitioned into multiple small clusters of neighboring switches called subnets. In the upper level of the network, the subnets are connected via wireline and wireless links. In both these designs, the subnets are connected in a regular structure like a mesh or a ring, in the second level of the hierarchy, long-range wireless shortcuts are placed on top of that. It is also shown that a WiNoC, where the network architecture is designed following the power-law based small-world connectivity [2], is more robust in presence of wireless link failures compared to the hierarchical counterpart [16].

In this work our aim is to quantify the performance of the mSWNoC architecture by incorporating irregular network routing algorithms in terms of latency, energy dissipation and thermal profiles.

III. Wireless NoC Architecture

Many naturally occurring complex networks, such as social networks, the Internet, the brain, as well as microbial colonies exhibit the small-world property [2], [17]. Small-world graphs are characterized by many short-distance links between neighboring nodes as well as a few relatively long-distance, direct shortcuts. Small-world graphs are particularly attractive for constructing scalable WiNoCs because the long-distance shortcuts can be realized using high-bandwidth, low-energy, wireless interconnects while the local links can be designed with traditional metal wires. In this work, we consider a small-world NoC architecture, where the long-range shortcuts are implemented through mm-wave wireless links operating in 10-100 GHz range. In the following sections we discuss the characteristics of this mSWNoC architecture and analyze its performance and temperature profiles with various irregular network routing strategies.

A. *mSWNoC Topology*

In the mSWNoC topology, each core is connected to a switch and the switches are interconnected using both wireline and wireless links. The topology of the mSWNoC is a small-world network where the links between switches are established following a power-law model [1], [2]. In this small-world network there are still several long wireline interconnects. As these are extremely costly in terms of power and delay, we use mm-wave wireless links to connect switches that are separated by a long distance. In [10], it is demonstrated that it is possible to create three non-overlapping channels with on-chip mm-wave wireless links. Using these three channels we overlay the wireline small-world connectivity with the wireless links such that a few switches get an additional wireless port. Each of these wireless ports will have wireless interfaces (WIs) tuned to one of the three different frequency channels. Each WI in the network is then assigned one of the three channels; more frequently communicating WIs are assigned the same channel to optimize the overall hop-count. One WI is replaced by a gateway WI that has all three channels assigned to it; this facilitates data exchange between the non-overlapping wireless channels. We have assumed an average number of connections from each switch to the other switches, $<k>$. The value of $<k>$ is chosen to be four so that the mSWNoC does not introduce any additional switch overhead with respect to a conventional mesh. Also an upper bound k_{max}, is imposed on the number of wireline links attached to a particular switch so that no switch becomes unrealistically large in the mSWNoC. This also reduces the skew in the distribution of links among the switches. Both $<k>$ and k_{max} do not include the local NoC switch port to the core.

B. *Communication and Channelization*

This section describes the overall communication mechanism, which includes routing and flow control, and the WI components for mSWNoC.

1) Routing and Flow Control

In the mSWNoC, data is transferred via a flit-based, wormhole routing [18]. Between a source-destination pair, the wireless links, through the WIs, are only chosen if the wireless path reduces the total path length compared to the wireline path. This can potentially give rise to hotspot situations in the WIs. Many messages will try to access the wireless shortcuts simultaneously, thus overloading the WIs, which would result in higher latency and energy dissipation. Token flow control [19] is used to alleviate overloading at the WIs. An arbitration mechanism is designed to grant access to the wireless medium to a particular WI, including the gateway WI, at a given instant to avoid interference and contention between the WIs that have the same frequency. To avoid centralized control and synchronization, the arbitration policy adopted is a wireless token passing protocol [10]. In this scheme, a single flit circulates as a token in each frequency channel. The particular WIs possessing a wireless token can broadcast flits into the wireless medium in their respective frequencies. The wireless token is forwarded to the next WI operating in the same frequency channel after all flits belonging to a message at a particular WI are transmitted. Packets are rerouted, through an alternate wireline path, if the WI buffers are full or, it does not have the token. To ensure deadlock-free routing in mSWNoC, we adopted three irregular network routing strategies and evaluated their performances.

The first routing strategy that we consider is an up/down tree-based routing architecture for the mSWNoC that utilizes a multiple tree roots (MROOTS)-based mechanism [5]. By allowing multiple routing trees to exist, where each tree routes on a dedicated virtual channel, traffic bottlenecks in the upper tree levels that is inherent in the TRAIN routing can be reduced. We adopt a traffic-weighted, minimized hop-count root-node placement policy. Selecting M tree roots will create M trees in the network, where the chosen M roots minimize the optimization metric μ as defined in (1) below.

$$\mu = \min_{\forall roots} \sum_{\forall i} \sum_{\forall j} h_{ij} f_{ij} \qquad (1)$$

Here, the minimum path distance in hops, h_{ij}, from switch i to switch j is determined following the up/down routing strategy [4], [5]. The frequency of traffic interaction between the switches is denoted by f_{ij}. As root selection only affects valid routing paths for deadlock freedom and does not alter the physical placement of links, any apriori knowledge of the frequency of traffic interaction aids in root selection. Incorporating f_{ij} helps minimize the routed path lengths for specific workloads on the mSWNoC architecture. Breadth-first trees were used during the tree creation process to balance the traffic distribution among the sub-trees, and to minimize bottlenecks in a particular tree. All wireless and wireline links that are not a part of the breadth-first tree are reintroduced as shortcuts. An allowed route never uses an up direction along the tree after it has been in the down path once. In addition, a packet traveling in the downward direction is not allowed to take a shortcut, even if that minimizes the distance to the destination. Hence, channel dependency cycles are prohibited, and deadlock freedom is achieved [4].

The second routing strategy is a layered shortest-path routing (LASH) algorithm as elaborated in [5]. LASH takes advantage of the multiple virtual channels in each port of the NoC switches in order to route messages along the shortest physical paths. In order to achieve freedom from deadlock, the network is divided into a set of virtual layers, which are created by dedicating the virtual channels from each switch port into these layers. The shortest physical path between each source-destination pair is then assigned to a layer such that the layer's channel dependency graph remains free from cycles. A channel dependency is created between two links in the source-destination path when a link from switch i to switch j and a link from switch j to switch k satisfies the following condition, $pathlength(i) < pathlength(j) < pathlength(k)$, where $pathlength(X)$ is the length of the minimal path between switch X and the original source switch. When a layer's channel dependency graph has no cycles, it is free from deadlocks as elaborated in [5]. The paths are also assigned in such a way that they are uniformly distributed among the layers so that they do not introduce an imbalance in distribution of paths. This ensures that in any particular layer, no switch has an unnecessarily high number of paths passing through it.

The third routing strategy is an adaptive layered shortest-path routing (ALASH) algorithm [5]. ALASH is built upon LASH but also allows each message to adaptively switch paths, allowing the message to choose its path at every intermediate switch. The decision to switch paths is based on current network conditions such as virtual channel availability and current communication density of the network. In order to increase the adaptability of the routing, multiple shortest paths between all source-destination pairs are found and then included into as many layers as possible while maintaining the original LASH deadlock freedom. It is possible to induce deadlock if a message is allowed to keep switching back and forth between two layers. Hence, a message is not allowed to revisit a layer that it has already traveled in to maintain deadlock freedom.

As mentioned earlier, in case the WI's buffers are full, or it does not have the token, packets attempting to access the WI need to be rerouted. As rerouting packets can potentially lead to deadlock, a rerouting strategy similar to Dynamic Quick Reconfiguration (DQR), as presented in [20], is used to ensure deadlock freedom. In this situation, the current WI becomes the new source for the packet, which then is forced to take a wireline only path to the final destination, still following the original routing restrictions that MROOTS, LASH, or ALASH impose.

2) Wireless Interface (WI)

The two principal WI components are the antenna and the transceiver, whose characteristics are outlined below.

The on-chip antenna for the mSWNoC has to provide the best power gain for the smallest area overhead. A metal zigzag antenna has been demonstrated to possess these characteristics [21]. This antenna also has negligible effect of rotation (relative angle between transmitting and receiving antennas) on received signal strength, making it most suitable for mm-wave NoC applications. Zigzag antenna characteristics depend on physical parameters like axial length, trace width, arm length, bend angle, etc. By varying these parameters, the antennas are designed to operate on different frequency channels [10]. Three different channels were obtained with 3dB bandwidths of 16 GHz and center frequencies of 31, 57.5, and 120 GHz respectively with a communication range of 20 mm. For optimum power efficiency, the quarter wave antennas use axial lengths of 0.73, 0.38, and 0.18 mm, respectively. The antenna design ensures that signals outside the communication bandwidth, for each channel, are sufficiently attenuated to avoid inter-channel interference.

The design of a low-power wideband wireless transceiver is the key to guarantee the desired performance of the mSWNoC. Therefore, at both the architecture and circuit levels of the transceiver, low-power design considerations need to be taken into account. At the architecture level, on-off-keying (OOK) modulation was chosen so as to simplify the circuit design. Non-coherent demodulation is used, therefore eliminating the power-hungry phase-lock loop (PLL) in the transceiver. Moreover, at the circuit level, body-enabled design techniques, including both forward body-bias (FBB) with DC voltages, as well as body-driven by AC signals, were implemented in several sub-blocks to further decrease power consumption.

The transceiver architecture is shown in Fig. 1. The receiver (RX) includes a wideband low-noise amplifier (LNA), an envelope detector for non-coherent demodulation, and a baseband amplifier. A local oscillator is not needed in the RX because non-coherent demodulation is used which results in a power reduction by more than 30% compared to the already proposed mSWNoC transceiver [10]. The transmitter (TX) has a simple direct up-conversion topology, consisting of a body-driven OOK modulator, a wideband power amplifier (PA), and a voltage-controlled oscillator (VCO).

IV. Experimental Results

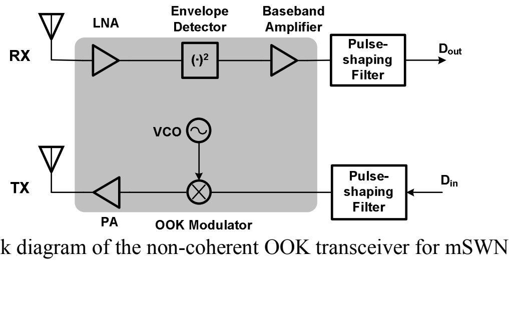

Fig. 1. Block diagram of the non-coherent OOK transceiver for mSWNoC.

In this section, we evaluate the performance and temperature profile of the mSWNoC and compare those with the conventional mesh-based NoC. We use GEM5 [22], a full system simulator, to obtain detailed processor- and network-level information. We consider a system of 64 alpha cores running Linux within the GEM5 platform for all experiments. Three SPLASH-2 benchmarks, FFT, RADIX, LU [23], and two PARSEC benchmarks, CANNEAL and BODYTRACK [24] are considered. These benchmarks vary in characteristics from computation intensive to communication intensive in nature and thus are of particular interest in this work. The behavior and problem size of the benchmarks are shown in Table 1. The width of all wired links is considered to be same as the flit width, which is 32 bits in this paper. Each packet consists of 64 flits. The NoC simulator uses switches synthesized from an RTL level design using TSMC 65-nm CMOS process in Synopsys™ Design Vision. All ports except those associated with the WIs have a buffer depth of two flits and each switch port has four virtual channels. Hence, four trees and four layers are created in MROOTS and LASH/ALASH, respectively. The ports associated with the WIs have an increased buffer depth of eight flits to avoid excessive latency penalties while waiting for the token. Increasing the buffer depth beyond this limit does not produce any further performance improvement for this particular packet size, but will give rise to additional area overhead [10]. Energy dissipation of the network switches, inclusive of the routing strategies, were obtained from the synthesized netlist by running Synopsys™ Prime Power, while the energy dissipated by wireline links was obtained through HSPICE simulations taking into consideration length of the wireline links. The processor-level statistics generated by the GEM5 simulations are incorporated into McPAT (Multicore Power, Area, and Timing) to determine the processor-level power values [25].

After obtaining the processor and network power values, these elements are arranged on a 20mm x 20mm die. The floorplans, along with the power values, are used in HotSpot [26] to obtain steady state thermal profiles. The processor power and the architecture-dependent network power values in presence of the specific benchmarks are fed to the HotSpot simulator to obtain the temperature profiles of each scenario.

A. Wireless Transciever Performance

The wireless transceiver circuitry was designed and laid out using TSMC 65-nm standard CMOS process and its characteristics were obtained through post-layout simulation. The overall power consumption of the transceiver is 31.1 mW, including 14.2 mW from the RX and 16.9 mW from the TX. With a data rate of 16 Gbps, this is equivalent to a bit energy of 1.95 pJ/bit. The total area overhead per wireless transceiver turns out to be 0.25 mm^2.

TABLE 1. CHARACTERISTICS OF BENCHMARKS UNDER CONSIDERATION.

Benchmark	**Busy %**	**Default Problem Size**	**Switch Traffic Interaction Rate (flits/cycle)**	
			Min	**Mean**
FFT	81.99	65,536 Data Points	4.9e-3	0.08
RADIX	84.98	262,144 Integers, 1024 RADIX	1.3e-3	0.04
LU	87.62	512x512 Matrix, 16x16 Blocks	1.8e-3	0.04
CANNEAL	56.74	200,000 Elements	0.03	0.35
BODYTRACK	32.11	2 Frames, 2000 Particles	0.09	0.22

B. Determination of the mSWNoC Topology

We determine the exact topology of the mSWNoC based on the principles of the small-world graph as discussed in section III.A. In [16] it is shown that the optimum values for k_{max} is 7 for a 64-core system size. We then augment the network by adding WIs. Also from [11], it is shown that WI placement is most energy efficient when the distance between them is at least 7 mm in the 65 nm technology node. The optimum number of WIs is 12 for a 64-core system size [1]. Increasing the number of WIs improves the connectivity of the network as they establish one-hop shortcuts. However, the wireless medium is shared among all the WIs and hence, as the number of WIs increases beyond a certain limit, performance starts to degrade due to the large token returning period [10].

C. mSWNoC Performance Evaluation

In this section, we present the latency, network-level energy dissipation and thermal characteristics of the mSWNoC by incorporating MROOTS-, LASH-, and ALASH-based routing strategies. We also compare the performance of the mSWNoC with respect to the traditional wireline mesh architecture. It is already demonstrated that using multiple-parallel metal wires as long-range shortcuts can match the achievable bandwidth of the mm-wave wireless links [10]. However, that wireline architecture dissipates significantly more energy than the corresponding WiNoC [10]. Hence, we do not consider the small-world architecture with wireline only long-range shortcuts in this performance evaluation.

1) Latency and Energy Characteristics

Figs. 2 and 3 show the latency and network energy respectively for the mSWNoC, using the three different routing strategies and considering the five benchmarks mentioned above. We also show the corresponding latency and energy profiles for the mesh. It can be observed from Fig. 2 that for all the benchmarks considered here, the latency of mSWNoC is lower than that of the mesh architecture. This is due to the small-world network-based interconnect infrastructure of the mSWNoC with direct long-range wireless links that enables a smaller average hop-count than that of mesh [16]. These three routing strategies are implemented on the same mSWNoC architecture. The difference in latency arises due to the difference in traffic distribution, from the benchmarks, and the relation of that traffic distribution with the specific routing mechanism adopted. However, it should be noted that the difference in latency among the routing algorithms is small due to the fact that the traffic injection load for all these benchmarks is low and the network operates much below saturation. As shown in Table 1, RADIX and LU have the

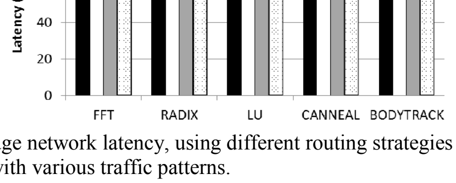

Fig. 2. Average network latency, using different routing strategies for mSWNoC, with various traffic patterns.

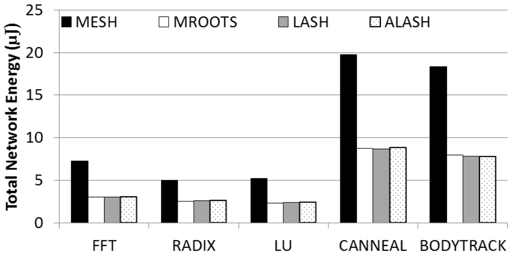

Fig. 3. Total network energy, using different routing strategies for mSWNoC, with various traffic patterns.

lowest overall inter-switch traffic interaction rates, as they have the lowest mean. When the injection rate is so low, the root nodes in MROOTS do not encounter enough traffic to become bottlenecks in the routing and hence MROOTS performs at its best. As a result, MROOTS has lower latency compared to LASH and ALASH. FFT and CANNEAL have higher injection rates than that of RADIX and LU, which can be seen by their mean inter-switch traffic interaction rates also. As the interaction rate increases, the root switches in MROOTS start to become bottlenecks and performance degradation becomes apparent. This is shown as LASH begins to improve in latency over MROOTS. BODYTRACK has the highest interaction rate among the benchmarks considered and it is near-uniform due to dynamic thread balancing [24]. In our experiments we also observed the same uniform trend by obtaining the f_{ij} matrix from the GEM5 simulation. CANNEAL has a higher mean interaction rate because it has a single skewed switch that has a very high interaction rate. When the interaction rate is higher, the network is stressed more and the aforementioned problems from MROOTS start. The weakness of MROOTS is that there is a strong tendency to generate traffic hotspots near the roots of the routing trees. The larger the hotspot, the longer messages get delayed in the network due to the root congestion. LASH and ALASH do not have the root congestion problem and hence, outperform MROOTS. At low injection loads, like that offered by these benchmarks, LASH achieves slightly less latency than ALASH as the latter has higher probability of choosing sub-optimal paths in each routing layer. The adaptiveness in ALASH helps to improve network performance for higher injection loads [5]. However, it does not help to reduce latency with respect to LASH when the injection load is low.

It can be observed from Fig. 3 that for each benchmark the network energy is much lower for the mSWNoC compared to the mesh architecture. The two main contributors of the energy dissipation are from the switches and the interconnect infrastructure. For mSWNoC, the overall switch energy decreases significantly compared to a mesh as a result of the better connectivity of the architecture. In this case, the hop-count decreases significantly, and hence, on the average, packets have to traverse through less number of switches and links. In addition, a significant amount of traffic traverses through energy efficient wireless channels; consequently allowing the interconnect energy dissipation to decrease. It can also be observed from Fig. 3 that the energy dissipation for the three different routing strategies follows the same trend as that of the latency. When messages are in the network longer (higher latency) they dissipate more energy. The difference in energy dissipation arising out of the logic circuits of each individual routing is very small and the overall energy dissipation is principally governed by the network characteristics.

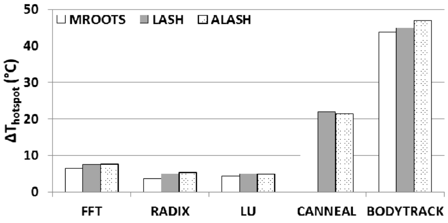

Fig. 4. Temperature decrease in switch hotspot compared to mesh with various traffic patterns.

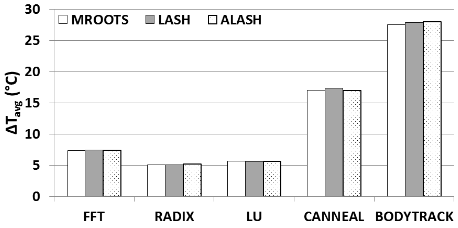

Fig. 5. Decrease in average switch temperature compared to mesh with various traffic patterns.

2) Thermal Characteristics

To quantify the thermal profile of the mSWNoC in presence of the three routing strategies, we consider the temperatures of the network switches. We consider the maximum and average switch temperature change between a mesh and mSWNoC, $\Delta T_{hotspot}$ and ΔT_{avg} respectively as the two relevant parameters. Figs. 4 and 5, show these parameters for the three routing strategies. It can be seen that the mSWNoC network architecture is inherently much cooler than the mesh counterpart. From Fig. 3, we can see that the difference in energy dissipation between mSWNoC and mesh is significant and hence, it is natural that mSWNoC switches are cooler. Fig. 4 helps to depict how well each routing strategy performs in distributing the power density, and hence heat, among the network switches. This is due to the fact that variations in $\Delta T_{hotspot}$ correspond to how well the routing mechanism balances the traffic within the network. The more interesting observation while analyzing the temperature profile lies in characterizing the differences among the routing strategies for the mSWNoC architecture. ALASH performs well in distributing the traffic among the network elements. Because of this, ALASH has the lowest maximum network temperature, which can be seen in Fig. 4, where ALASH has the highest $\Delta T_{hotspot}$. CANNEAL is an exception, with LASH having a higher $\Delta T_{hotspot}$. As mentioned above, CANNEAL has an extremely skewed traffic interaction pattern, allowing the thermal profile of LASH to match with that of ALASH. Fig. 6 displays the temperature distribution in the routing schemes of the BODYTRACK benchmark as an example. Here, it can be seen that the MROOTS routing strategy does a poor job on the well-balanced BODYTRACK benchmark as the temperature

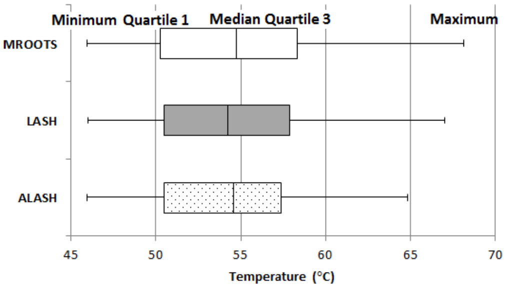

Fig. 6. Temperature distribution of switches using the BODYTRACK traffic pattern.

distribution is more spread than LASH and ALASH. As this benchmark has a near-uniform traffic distribution, the MROOTS routing strategy will form bottlenecks in the upper levels of its trees. In this case, the heat distribution will be spread further as the leaves among the trees see light traffic, while the near-root nodes see heavy traffic. The LASH routing strategy does not have nearly the same intensity of hotspots, as it follows routes specific to the source-destination pair. The bottlenecks in this situation involve the routing paths that overlap most; that is, the bottleneck will form in switches that are common among many of the shortest paths. ALASH attempts to avoid creating hotspots by having multiple shortest paths. By choosing the path that avoids local network hotspots, we can reduce the maximum network temperature quite well using the ALASH routing strategy. In the BODYTRACK example, LASH and ALASH reduce the hotspot switch temperature further compared to MROOTS by 1.16°C and 3.32°C, respectively. Additionally, ALASH reduces hotspot switch temperature compared to MROOTS by 1.18°C, 1.78°C, 0.65°C, and 2.22°C for FFT, RADIX, LU, and CANNEAL respectively. By observing Fig. 5, it can also be seen that the average temperature reduction in switches among the routing strategies is relatively unaffected. We can conclude that, reduction of the the maximum temperature using ALASH has not come at the cost of increasing the average network temperature due to the inherent rerouting efforts of this strategy. Overall, it can be seen that for the routing strategies implemented, we can obtain very similar latency and network energy profiles while reducing the hotspot switch temperature.

V. Conclusion

As we demand more from our computing systems, they will be limited by power, energy, and thermal constraints. Without new energy-efficient design paradigms, producing information and communication technologies (ICT) systems capable of meeting the computing, storage, and communication demands of emerging applications will be unlikely. Millimeter-wave wireless small-world NoC (mSWNoC) is an enabling technology to design energy efficient and high bandwidth multicore architectures with improved thermal profile over conventional wireline mesh-based counterparts. In this paper, we evaluated the latency, energy dissipation, and thermal profiles of mSWNoC in presence of three irregular routing methodologies, viz., MROOTS, LASH and ALASH. In presence of all these routing strategies mSWNoC provides lower latency and energy dissipation compared to a conventional wireline mesh. However, the difference in latency and energy dissipation profile among these routing strategies is small. The effects of these routing strategies are more pronounced when we analyze the thermal profile. ALASH reduces the temperature of the hotspot network switch for a well-distributed benchmark like BODYTRACK by an additional 3.32°C over MROOTS. When deciding which routing strategy is suitable for mSWNoC, it is clear that ALASH provides similar performance compared to MROOTS and LASH, while offering an improved temperature profile.

References

[1] P. Wettin, et al., "Energy-Efficient Multicore Chip Design Through Cross-Layer Approach." Proc. of DATE, 2013, pp. 725-730.

[2] T. Petermann and P. De Los Rios, "Spatial Small-World Networks: A Wiring Cost Perspective," arXiv: cond-mat/0501420v2.

[3] J. Flich, et al., "A Survey and Evaluation of Topology-Agnostic Deterministic Routing Algorithms," IEEE Trans. On Parallel and Distributed Systems, vol. 23, no. 3, 2012, pp. 405-425.

[4] H. Chi and C. Tang, "A Deadlock-Free Routing Scheme for Interconnection Networks with Irregular Topology," Proc. of ICPADS, 1997, pp. 88-95.

[5] O. Lysne, et al., "Layered Routing in Irregular Networks," IEEE Trans. on Parallel and Distributed Systems, vol. 17, no. 1, 2006, pp. 51-65.

[6] R. Marculescu, et al., "Out-standing Research Problems in NoC Design: System, Microarchitecture, and Circuit Perspectives," IEEE Trans. CAD of Integr. Circuits Syst., vol. 28, no. 1, 2009, pp. 3-21.

[7] U.Y. Ogras and R. Marculescu, "Application-Specific Network-on-Chip Architecture Customization via Long-Range Link Insertion," Proc. of ICCAD, 2005, pp.246-253.

[8] U.Y. Ogras and R. Marculescu, "It's a Small World After All: NoC Performance Optimization via Long-Range Link Inser-tion," IEEE Trans. Very Large Scale Integr. Syst., vol. 14, no. 7, 2006, pp. 693-706.

[9] A. Kumar, et al., "Toward Ideal On-Chip Communication Using Express Virtual Channels," IEEE Micro, vol. 28, no. 1, 2008, pp. 80-90.

[10] S. Deb, et al., "Design of an Energy Efficient CMOS Compatible NoC Architecture with Millimeter-Wave Wireless Interconnects, " IEEE Trans. Comput., 2012, pp. 2382-2396.

[11] S. Deb, et al., "Wireless NoC as Interconnection Backbone for Multicore Chips: Promises and Challenges," IEEE J. Emerg. Sel. Topic Circuits Syst., vol. 2, no. 2, 2012, pp. 228-239.

[12] D. Zhao and Y. Wang, "SD-MAC: Design and Synthesis of a Hardware-Efficient Collision-Free QoS-Aware MAC Protocol for Wireless Network-on-Chip," IEEE Trans. Comput., vol. 57, no. 9, 2008, pp. 1230-1245.

[13] S.B. Lee, et al., "A Scalable Micro Wireless Interconnect Structure for CMPs," Proc. of ACM MobiCom, 2009, pp. 20-25.

[14] D. DiTomaso, et al., "iWise: Inter-router Wireless Scalable Express Channels for Network-on-Chips (NoCs) Architectures," Proc. of HOTI, 2011, pp. 11-18.

[15] A. Ganguly, et al., "Scalable Hybrid Wireless Network-on-Chip Architectures for Multi-Core Systems," IEEE Trans. Comput., vol. 60, no. 10, 2011, pp.1485-1502.

[16] A. Ganguly, et al., "Complex Network Inspired Fault-Tolerant NoC Architectures with Wireless Links," Proc. of NOCS, 2011, pp. 169-176.

[17] D.J. Watts and S.H. Strogatz, "Collective Dynamics of 'Small-World' Networks," Nature, 393, 1998, pp. 440-442.

[18] P.P. Pande, et al., "Performance Evaluation and Design Trade-offs for Network-on-Chip Interconnect Architectures," IEEE Trans. Comput., vol. 54, no. 8, 2005, pp. 1025-1040.

[19] A. Kumar, L.-S. Peh, and N.K. Jha, "Token Flow Control," Proc. of MICRO, 2008, pp. 342-353.

[20] F.O. Sem-Jacobsen and O. Lysne, "Topology Agnostic Dynamic Quick Reconfiguration for Large-Scale Interconnection Networks," Proc. of CCGrid, 2012, pp. 228-235.

[21] B.A. Floyd, C.-M. Hung, and K.K. O, "Intra-Chip Wireless Interconnect for Clock Distribution Implemented with Integrated Antennas, Receivers, and Transmitters," IEEE J. Solid-State Circuits, vol. 37, no. 5, pp. 543-552.

[22] N. Binkert, et al., "The GEM5 Simulator," ACM SIGARCH Computer Architecture News, 39(2), 2011, pp. 1-7.

[23] S.C. Woo, et al., "The SPLASH-2 Programs: Characterization and Methodological Considerations," Proc. of ISCA, 1995, pp. 24-36.

[24] C. Bienia, "Benchmarking Modern Multiprocessors," Ph.D. dissertation, Dept. Computer Science, Princeton Univ., Princeton NJ, 2011.

[25] S. Li, et al., "McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures," Proc. of MICRO, 2009, pp. 469-480.

[26] K. Skadron, et al., "Temperature-Aware Microarchitecture," Proc. of ISCA, 2003, pp. 2-13.