Design of an Energy Efficient CMOS Compatible NoC Architecture with Millimeter-Wave Wireless Interconnects

Sujay Deb, Student Member, IEEE, Kevin Chang, Student Member, IEEE, Xinmin Yu, Student Member, IEEE, Suman Prasad Sah, Student Member, IEEE, Miralem Cosic, Amlan Ganguly, Member, IEEE, Partha Pratim Pande, Senior Member, IEEE, Benjamin Belzer, Member IEEE and Deukhyoun Heo, Member IEEE

Abstract— The Network-on-Chip (NoC) is an enabling technology to integrate large numbers of embedded cores on a single die. The existing methods of implementing a NoC with planar metal interconnects are deficient due to high latency and significant power consumption arising out of multi-hop links used in data exchange. To address these problems, we propose design of a hierarchical small-world wireless NoC architecture where the multi-hop wire interconnects are replaced with high-bandwidth and single-hop long-range wireless shortcuts operating in the millimeter (mm)-wave frequency range. The proposed mm-wave wireless NoC (mWNoC) outperforms the corresponding conventional wireline counterpart in terms of achievable bandwidth and is significantly more energy efficient. The performance improvement is achieved through efficient data routing and optimum placement of wireless hubs. Multiple wireless shortcuts operating simultaneously further enhance the performance, and provide an energy efficient solution for design of communication infrastructures for multi-core chips.


1 INTRODUCTION

THE Network-on-Chip (NoC) has emerged as a revolutionary methodology to integrate numerous blocks in a single chip. An important performance limitation in traditional NoCs arises from planar metal interconnect-based multi-hop communications, wherein the data transfer between two far apart cores causes high latency and power consumption. This limitation of conventional NoCs can be addressed by drawing inspiration from the interconnection mechanism of natural complex networks. Many networks, such as networks of neurons in the brain, the Internet, and social networks share the small-world (SW) property [1]. Compared to a purely locally and regularly interconnected network (such as a mesh interconnect), small-world networks have a very short average distance between any pair of nodes. This makes them particularly interesting for efficient communication in modern multi-core chips with increasing levels of integration. This small-world topology can be incorporated in NoCs by introducing long-range, high bandwidth and low power links between far apart cores. In this work, we propose a hybrid NoC architecture that uses on-chip millimeter (mm)-wave wireless links designed in traditional CMOS technology as long-range communication channels between widely separated cores, along with wired interconnects connecting adjacent cores. Recent investigations have established characteristics of the silicon integrated on-chip antenna operating in the mm-wave range of a few tens to one hundred GHz and it is now a viable technology [2]. Coupled with significant advances in mm-wave transceiver design this opens up new opportunities for detailed investigations of mm-wave wireless NoCs (mWNoCs). In this paper, we propose a design methodology and establish associated trade-offs for hierarchical NoCs with mm-wave wireless links in presence of various traffic patterns. We demonstrate that the proposed mWNoC outperforms its more traditional wireline counterparts in terms of sustainable data rate and energy dissipation. We also evaluate the performance of mWNoC with respect to two other small-world NoC architectures with emerging interconnect technologies. In one of these architectures the long-range links are implemented with recently proposed RF interconnects (RFNoC) [3]. The other architecture is a hierarchical and small-world wireless NoC designed with carbon nano tube (CNT) enabled THz wireless links (THzNoC) [4]. We demonstrate the advantages and the limitations of each architecture and establish the relevant design trade-offs.

The main contributions of this paper are as follows:
1. We demonstrate that mm-wave wireless interconnect based NoCs can be a viable CMOS compatible solution for future many core chips, and are capable of solving the performance limitations of traditional multi-hop wireline counter-
parts. The mWNoC can achieve performance comparable to that of other emerging and more technologically challenging on-chip RF/wireless interconnect solutions. It is also shown that mWNoC can accommodate multiple simultaneously operating wireless channels resulting in further improvement of overall performance.

2. We present simulation results to evaluate the performance of mWNoC against other emerging interconnect based NoCs, viz., THzNoC and RFNoC, in both uniform and non-uniform traffic scenarios. The inherent broadcasting capability of mWNoC is also exploited to demonstrate its performance advantage. We also demonstrate that the hierarchical and small-world based NoC architectures with emerging interconnects like wireless and RF provide significantly better performance than either conventional mesh-based NoCs or hierarchical architectures without any shortcuts. The area overheads associated with these novel NoC architectures are also quantified and it is shown that performance benefits clearly outweigh the overheads.

2 RELATED WORK

The limitations and design challenges associated with existing NoC architectures are elaborated in [5]. Conventional NoCs use multi-hop packet switched communication. At each hop the data packet goes through a complex router/switch, which contributes considerable power, throughput and latency overhead. To improve performance, a methodology to automatically synthesize an architecture with a few application specific long-range links inserted in a regular mesh was proposed in [6]. Subsequently, performance advantages of NoCs by insertion of long range wired links following principles of small-world graphs was elaborated in [7]. The concept of express virtual channels is introduced in [8]. By using virtual express lanes to connect distant cores in the network, it is possible to avoid the router overhead at intermediate nodes, and thereby improve NoC performance in terms of power, latency and throughput. Performance is further improved by incorporating ultra low-latency, multi-drop on-chip global lines (G-lines) for flow control signals [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 the conventional metal wires. Hence, the performance improvements by using long-range wireless links will be more than that using wired links [10].

The design principles of photonic NoCs are elaborated in various recent publications [11][12][13]. The components of a complete photonic NoC, e.g., dense waveguides, switches, optical modulators and detectors, are now viable for integration on a single silicon chip. It is estimated that a photonic NoC will dissipate an order of magnitude less power than an electronic NoC. Another alternative is NoCs with multi-band RF interconnects [3]. Various implementation issues of this approach are discussed in [14]. In this particular NoC, instead of depending on the charging/discharging of wires for sending data, electromagnetic (EM) waves are guided along on-chip transmission lines created by multiple layers of metal and dielectric stack. As the EM waves travel at the effective speed of light, low latency and high bandwidth communication can be achieved. This type of NoC is also predicted to dissipate an order of magnitude less power than the traditional planar NoC, with significantly reduced latency as well.

Recently, the design of a wireless NoC based on CMOS ultra wideband (UWB) technology was proposed in [15] and [16]. In [17] the authors propose multi-channel wireless NoC using UWB transceivers. As ultra-short pulses can be used with the UWB technology, the authors propose time-hopping multiple access to improve the performance of the NoC. In this scheme a transmitting RF node uses pseudorandom timing of its pulses within the UWB signal interval, which is unique for each receiver. This enables concurrent multiple channels between multiple transceiver pairs.

The performance of silicon integrated on-chip antennas for intra- and inter-chip communication have been already demonstrated by the authors of [18]. They have primarily used metal zigzag antennas operating in the range of tens of GHz. The propagation mechanisms of radio waves over intra-chip channels with integrated antennas were also investigated [19]. Depending on antenna configuration and substrate characteristics, achievable wireless channel frequencies can be in the range of 50-100 GHz. At these mm-wave frequencies the effect of metal interference structures such as power grids, local clock trees and data lines on on-chip antenna characteristics like gain and phase are investigated in [20]. The demonstration of intra-chip wireless interconnection in a 407-pin flip-chip package with a ball grid array (BGA) mounted on a PC board [21] has addressed the concerns related to influence of packaging on antenna characteristics. Design rules for increasing the predictability of on-chip antenna characteristics have been proposed in [20]. Using antennas with a differential or balanced feed structure can significantly reduce coupling of switching noise, which is mostly common-mode in nature [22]. In [23], the feasibility of designing miniature antennas and simple transceivers that operate in the sub-THz frequency range for on-chip wireless communication has been demonstrated. In [24] a combination of Time and Frequency Division Multiplexing is used to transfer data over inter-router wireless express channels. However, the issues of inter-channel interference due to multiple adjacent frequency channels remain unresolved in this work. Design of a small-world wireless NoC operating in the THz frequency range using carbon nanotube (CNT) antennas is elaborated in [4]. Though this particular NoC is shown to exceed the performance of traditional wireline NoCs by orders of magnitude, integration and reliability of CNT devices need more investigation.

This work aims to circumvent the performance limitations of traditional multi-hop NoCs by introducing a hi-
erarchical small-world network with CMOS compatible mm-wave wireless links for multi-core chips.

3 MM-WAVE WIRELESS NOC (mWNoC) ARCHITECTURE

A generic wired NoC provides interconnection among embedded cores via switches and wired links. Communication between pairs of source and destination cores is generally via multi-hop links, resulting in high energy dissipation and latency. With increasing system size the average hop count increases and consequently the problem of higher energy dissipation and latency becomes more profound. To alleviate this problem, we propose a hierarchical NoC architecture with wireless interfaces strategically placed for optimum performance. In the following subsections we discuss the topology of the proposed hierarchical architecture and the adopted performance optimization methodology.

3.1 mWNoC Architecture

Modern complex network theory provides powerful methods to analyze network topologies and their properties [25]. Networks with the small-world property have a very short average path length, which is commonly measured as the number of hops between any pair of nodes. The average path length of small-world graphs is bounded by a polynomial in $\log(N)$, where $N$ is the number of nodes, which makes them particularly interesting for efficient communication with minimal resources [26]. This feature of small-world graphs makes them attractive for constructing mWNoCs. A small-world topology can be constructed from a locally connected network by rewiring selected node connections randomly to other nodes, which creates short cuts in the network [27]. These random long-range links between nodes can also be established following probability distributions depending on the distance separating the nodes [28]. It has been shown that such shortcuts in NoCs can significantly improve the performance compared to locally interconnected mesh-like networks [7][27] with fewer resources than a fully connected system.

Our goal here is to use the small-world approach to build a highly efficient NoC based on both wired and wireless links. This topology can be incorporated in NoCs by introducing long-range, high bandwidth and low power wireless links between distant cores. This will enable the design of hierarchical NoC architectures, where closely spaced cores will communicate through tradition-al metal wires, but long distance communications will be predominantly achieved through high performance wireless links. Thus, for our purpose, we first divide the whole system into multiple small clusters of neighboring cores called subnets. As subnets are smaller networks, intra-subnet communication will have a shorter average path length than a single NoC spanning the whole system. These subnets have switches and links as in a standard NoC. The cores are connected to a centrally located hub through wired links, and the hubs from all subnets are connected in a second level network forming a hierarchical structure. This is achieved by interconnecting adjacent hubs with wireline links, and by introducing a few long range mm-wave wireless links between distant hubs according to the placement scheme outlined in section 3.2. The hubs connected through wireless links require wireless interfaces (WIs).

Reducing long-distance multi-hop wired communication is essential in order to achieve the full benefit of on-chip wireless networks for multi-core systems. The number of WIs and their placement are optimized for performance using a Simulated Annealing (SA) [29] based optimization algorithm. The SA approach allows network design to be scalable with increasing system size. The key to our approach is establishing optimal overall network topology under given resource constraints, i.e., a limited number of WIs. Fig. 1 shows a representative interconnection topology with 16 hubs and 6 wireless interfaces. In this paper, as an example, the hubs are considered to be connected in a mesh. Instead of the mesh the hubs can be connected in any other possible interconnect topology depending on the exact performance requirement. The subnets considered here have the StarRing architecture, which consists of a ring with a central hub. We have already shown that this architecture provides the best performance-overhead tradeoff point for mWNoC architectures [10]. The size and number of subnets should be chosen such that neither the subnets nor the upper level of the hierarchy become too large. If either level of the hierarchy becomes too large then it causes a performance bottleneck by limiting the data throughput in that level. Since the architecture of the two levels can be different causing their traffic characteristics to differ, the exact hierarchical division can be obtained by performing system level simulations as discussed in subsection 5.2.

3.2 Optimization of mWNoC Architecture

In this section we present the method used for determining the optimum mWNoC architecture. At first we define the optimization metric, and then we discuss the SA based optimization procedure for obtaining optimum number of WIs and their suitable placement.

i) Optimization metric

In order to determine the optimal number of WIs for a given network, we define two metrics, which account for the performance as well as the cost of the NoC. The first metric, which measures approximate network performance, is the average shortest path, $\mu$ between all pairs of hubs. Let $N$ be the number of hubs of the network. Let $d$ be an $N \times N$ matrix, where element $d_{u}$ is the distance
For each number of WIs, the minimum value of the number of WIs ($n$) to connect the shortest path between hub $i$ and hub $j$ measured in hops. A single hop in this work is defined as the path length between a source and destination pair that can be traversed in one clock cycle. The matrix $d$ is populated using Dijkstra's shortest path algorithm [30]. The distances are then weighted with the normalized frequencies of communication between hub pairs. The metric, $\mu$ can be calculated as

$$\mu = \sum h_{i,j} \cdot f_{i,j} \cdot [(N^2 - N) \cdot F],$$

where $h_{i,j}$ is the distance (in hops) between the $i^{th}$ source and $j^{th}$ destination. The frequency $f_{i,j}$ of communication between the $i^{th}$ source and $j^{th}$ destination is the apriori frequency of the traffic interactions between the subnets determined by a particular traffic pattern that depends on the application mapped onto the NoC. $F$ is then calculated as

$$F = \sum f_{i,j}.$$  

The probability of getting access to the wireless channel for communication between any source-destination pair is designated by $p$ which is inversely proportional to the number of WIs ($n$) sharing the same frequency channel. With the assumption that all the WIs are equally likely to have access to the wireless channel, $p$ can be computed as

$$p = 1/n.$$  

Here, equal importance is attached to inter-hub distance and frequency of communication.

The second metric needed to complete the quantification of a network's quality is the cost function

$$Cost(\# of WI) = A + P + L,$$

where, $A$, $P$, and $L$ are normalized area, power and wireless channel access delay overheads respectively arising from the WIs. $A$ is determined by dividing the total wireless hub area by the area of the communication infrastructure. The power dissipated by all WIs is divided by the total power consumed by the communication infrastructure to determine $P$. $L$ is determined by dividing the token returning period (described in section 4.3) by the average packet latency. The two metrics, average shortest path $\mu$ and cost are thus the two objectives to be optimized. Many methods exist for evaluating multi-objective optimization problems [31]. We describe the aggregate objective function (AOF), which combines both of the metrics, as follows:

$$AOF = a \cdot \mu + (1-a) \cdot Cost,$$

where, $a$ specifies the importance of the two metrics, i.e., $a = 0$ results in an analysis entirely dependent on cost, $a = 1$ results in an analysis entirely dependent on the network connectivity, while $a = 0.5$ makes for a balance between the two metrics. The choice of $a$ is a design decision and depends on the design requirements. For a chosen value of $a$, optimum number of WI ($n$) is selected that results in minimum value of $AOF$.

The AOF defined above is then used in the optimization step outlined in the next subsection to determine the optimal NoC architecture.

ii) Placement of WIs

This process takes $N$ and $a$ as inputs and for all possible number of WIs perform SA based placement optimization. WI placement is crucial for optimum performance gain as it establishes high-speed, low-energy interconnects on the network. It is shown in [4] that for placement of wireless links in a NoC, the SA algorithm converges to the optimal configuration much faster than the exhaustive search technique. Hence, we adopt a SA based optimization technique for placement of the WIs to get maximum benefits of using the wireless shortcuts. Initially, the WIs are placed randomly with each hub having equal probability of getting a WI. The only constraint observed while deploying the WIs to the hubs is that a single hub could have a maximum of one WI.

Once the network is initialized randomly, an SA based optimization step is performed. Since the deployment of WIs is only on the hubs, the optimization is performed solely on the second level network of hubs. If there are $N$ hubs in the network and $n$ WIs to distribute, the size of the search space $S$ is given by

$$|S| = \binom{N}{n}.$$  

Thus, with increasing $N$, it becomes more computationally expensive to find the best solution by exhaustive search. SA is performed on the optimization metric $AOF$ defined by (6). In each SA iteration, a new network is created by randomly reassigning a WI in the current network. The metric for the new network is calculated and compared to the current network’s metric. The new network is chosen as the current optimal solution if its metric is lower. However, even if the metric is higher we choose the new network probabilistically. This reduces the probability of getting stuck in a local optimum, which could happen if the SA process were to never choose a worse
solution. In this work we have used Cauchy annealing schedule [29] and it is preferred over the normal distribution because of its flatter tails, making it easier to escape from local minima. The convergence criterion chosen here is that the metric at the end of the current iteration differs by less than 0.1% from the metric of the previous iteration [4]. Fig. 2 shows the steps used to optimize the network.

An important point to note here is that similar results can also be obtained using other optimization techniques, like evolutionary algorithms (EAs) [32] and co-evolutionary algorithms [33]. Although EAs are generally believed to give better results, SA reaches comparably good solutions much faster [34]. We have used SA in this work as an example.

4 OVERALL COMMUNICATION SCHEME

In this section we describe the various components of the WIs and the adopted data routing strategy. As mentioned in the previous section, the WIs are optimally placed in some of the hubs to provide them with the capability to communicate using the wireless channel. The two principal components of the WI are the antenna and the transceiver. Characteristics of these two components are outlined below.

4.1 On-Chip Antennas

The on-chip antenna for the proposed mWNoC has to provide the best power gain for the smallest area overhead. A metal zigzag antenna [35] has been demonstrated to possess these characteristics. This antenna also has negligible effect of rotation (relative angle between transmitting and receiving antennas) on received signal strength, making it most suitable for mWNoC application [19]. The zigzag antenna is designed with 10 μm trace width, 60 μm arm length and 30° bend angle. The axial length depends on the operating frequency of the antenna which is determined in subsection 5.4. The details of the antenna structure are shown in Fig. 3.

4.2 Wireless Transceiver Circuit

To ensure the high throughput and energy efficiency of the mWNoC, the transceiver circuitry has to provide a very wide bandwidth as well as low power consumption. In designing the on-chip mm-wave wireless transceiver, the low power design considerations are taken into account at the architecture level. Non-coherent on-off keying (OOK) is chosen as the modulation method, as it allows relatively simple and low-power circuit implementation. As illustrated in Fig. 4, the transmitter (TX) circuitry consists of an up-conversion mixer and a power amplifier (PA). On the receiver (RX) side, direct-conversion topology is adopted, consisting of a low noise amplifier (LNA), a down-conversion mixer and a baseband amplifier. An injection-lock voltage-controlled oscillator (VCO) is reused for TX and RX. With both direct-conversion and injection-lock technology, a power-hungry phase-lock loop (PLL) is eliminated. Moreover, at the circuit level, body-enabled design techniques [36], including both forward body-bias (FBB) with DC voltages, as well as body-driven by AC signals [37], are implemented to further decrease power consumption. High isolation to other circuits is guaranteed by using triple-well CMOS with deep N-well, which is now common in most of the scaled CMOS processes. Detailed design descriptions of the transceiver are presented in our previous works [10][38].

4.3 Adopted Data Transmission Strategy

In the proposed hierarchical NoC, data is transferred via flit-based wormhole routing [39]. Intra-subnet data routing is done according to the topology of the subnets. For StarRing subnet topology if the destination core is within two hops on the ring from the source then the data is routed along the ring. If the destination core is more than two hops away then the data routing takes place via the central hub. To avoid deadlock within the subnet, we follow the virtual channel management scheme adopted from the Red Rover algorithm [40], in which the ring is divided into two equal sets of contiguous nodes. Messages originated from each group of nodes use a particular set of dedicated virtual channels regardless of destination. Furthermore, messages injected on a particular virtual channel are reused for TX and RX. With both direct-conversion and injection-lock technology, a power-hungry phase-lock loop (PLL) is eliminated. Moreover, at the circuit level, body-enabled design techniques [36], including both forward body-bias (FBB) with DC voltages, as well as body-driven by AC signals [37], are implemented to further decrease power consumption. High isolation to other circuits is guaranteed by using triple-well CMOS with deep N-well, which is now common in most of the scaled CMOS processes. Detailed design descriptions of the transceiver are presented in our previous works [10][38].

4.3 Adopted Data Transmission Strategy

In this section we describe the various components of the WIs and the adopted data routing strategy. As mentioned in the previous section, the WIs are optimally placed in some of the hubs to provide them with the capability to communicate using the wireless channel. The two principal components of the WI are the antenna and the transceiver. Characteristics of these two components are outlined below.

4.1 On-Chip Antennas

The on-chip antenna for the proposed mWNoC has to provide the best power gain for the smallest area overhead. A metal zigzag antenna [35] has been demonstrated to possess these characteristics. This antenna also has negligible effect of rotation (relative angle between transmitting and receiving antennas) on received signal strength, making it most suitable for mWNoC application [19]. The zigzag antenna is designed with 10 μm trace width, 60 μm arm length and 30° bend angle. The axial length depends on the operating frequency of the antenna which is determined in subsection 5.4. The details of the antenna structure are shown in Fig. 3.

4.2 Wireless Transceiver Circuit

To ensure the high throughput and energy efficiency of the mWNoC, the transceiver circuitry has to provide a very wide bandwidth as well as low power consumption. In designing the on-chip mm-wave wireless transceiver, the low power design considerations are taken into account at the architecture level. Non-coherent on-off keying (OOK) is chosen as the modulation method, as it allows relatively simple and low-power circuit implementation. As illustrated in Fig. 4, the transmitter (TX) circuitry consists of an up-conversion mixer and a power amplifier (PA). On the receiver (RX) side, direct-conversion topology is adopted, consisting of a low noise amplifier (LNA), a down-conversion mixer and a baseband amplifier. An injection-lock voltage-controlled oscillator (VCO) is reused for TX and RX. With both direct-conversion and injection-lock technology, a power-hungry phase-lock loop (PLL) is eliminated. Moreover, at the circuit level, body-enabled design techniques [36], including both forward body-bias (FBB) with DC voltages, as well as body-driven by AC signals [37], are implemented to further decrease power consumption. High isolation to other circuits is guaranteed by using triple-well CMOS with deep N-well, which is now common in most of the scaled CMOS processes. Detailed design descriptions of the transceiver are presented in our previous works [10][38].

Fig. 3. Zigzag antenna structure details.

Fig. 4. OOK transceiver block diagram.

Fig. 5. An algorithmic representation of the adopted data routing strategy.
channel will continue their traversals on that channel until reaching destinations. Since a message is confined to a particular channel for its entire traversal and each of these channels contains no cycles, the scheme is deadlock free.

Inter-subnet data routing however requires the flits to use the upper level network consisting of the wired and wireless links. By using the wireless shortcuts between the hubs with the WIs, flits can be transferred in a single hop. If the source hub does not have a WI, the flits are routed to the nearest hub with a WI via the wired links and are transmitted through the wireless channel. Likewise, if the destination hub does not have a WI then the hub nearest to it with a WI receives the data and routes it to the destination through wired links. Between a pair of source and destination hubs without WIs, the routing path involving a wireless link is chosen if it reduces the total path length compared to the wired path. This can potentially give rise to a hotspot situation in the WIs because many messages try to access wireless shortcuts simultaneously, thus overloading the WIs and resulting in higher latency. Token flow control [41] and distributed routing are used to alleviate this problem. Tokens are used to communicate the status of the input buffers of a particular WI to other nearby hubs, which need to use that WI for accessing wireless shortcuts. Every WI input port has a token and the token is turned on if the availability of the port’s buffer is greater than a fixed threshold and turned off otherwise. The routing adopted here is a combination of dimension order routing for the hubs without WIs and South-East routing algorithm for the hubs with wireless shortcuts. This routing algorithm is proved to be deadlock free in [7]. If the WIs that the message encounters along the way are not available, the message follows dimension order routing and keeps looking for the shortest path using WIs at every hub until the destination hub is reached. Consequently, the distributed routing and token flow control prevents deadlocks and effectively improves performance by distributing traffic though alternative paths. All the wireless hubs are tuned to the same channel and can send or receive data from any other wireless hub on the chip. Under these conditions an arbitration mechanism needs to be designed in order to grant access to the wireless medium to a particular hub at a given instant to avoid interference and contention.

To avoid the need for a centralized control and synchronization mechanism, the arbitration policy adopted is a token passing protocol [42]. It should be noted that the use of the word token in this case differs from the usage in the above mentioned token flow control. According to this scheme, the particular WI possessing the token can broadcast flits into the wireless medium. All other hubs will receive the flit as their antennas are tuned to the same frequency band. When the destination address matches the address of the receiving hub then the flit is accepted for further routing. It is routed either to a core in the subnet of that hub or to an adjacent hub. The token is released to the next hub with a WI after all flits belonging to a single packet at the current token-holding hub are transmitted. Fig. 5 shows the flow chart of the adopted data routing strategy.

According to [43], the mWNoC is deadlock free if both the subnets and the 2nd level of the network are deadlock free and the boundary nodes are safe nodes. As explained above, the subnets and the 2nd level of small-world network are deadlock free. Moreover, in this work the boundary nodes are the hubs, which allow inter-subnet communication. The hubs are safe nodes as there is no path from an internal output link to an internal input link.

5 EXPERIMENTAL RESULTS

In this section we discuss the experimental results that demonstrate performance of the proposed mWNoC. First we present the justification behind using SA based optimization followed by the characteristics of the on-chip wireless communication channel. Then we present detailed network level simulations with various system sizes and traffic patterns.

We evaluate the performance of SA for WI placement by comparing the number of iterations required to achieve optimal network configuration through SA and exhaustive search. Moreover, we present the justification behind using SA based optimization for the system configurations considered in this paper. Fig. 6 shows the number of iterations required to arrive at the optimal solution with SA and exhaustive search. The numbers of iterations with SA were measured as an average of 10 trials. Clearly the SA algorithm converges to the optimal configuration much faster than the exhaustive search technique. This advantage will further increase for larger system sizes.

5.1 Simulation Setup

An overview of the performance evaluation setup for the mWNoC is shown in Fig. 7. To obtain the gain and bandwidth of the antennas we use the ADS momentum tool [44]. For our experiments, we consider three different system sizes, namely 128, 256, and 512 cores, and the die area is kept fixed at 20 mm x 20 mm for all system sizes. The mm-wave wideband wireless transceiver is designed and simulated using Cadence tools with TSMC [45] 65-nm standard CMOS process to obtain its power and delay characteristics. The subnet switches and the digital components of the hubs are synthesized using Synopsis tools with 65-nm standard cell library from TSMC at a clock frequency of 2.5 GHz. Energy dissipation of all the wired links is obtained from Cadence layout assuming a 20 mm x 20 mm die area. All the power and delay numbers of various components along with the optimum network.

![Fig. 6. Number of iterations required to reach optimal solution by the SA and exhaustive search methods.](image-url)
configuration generated from the SA are then fed into the network simulator to obtain overall mWNoC performance. As an example, table 1 shows all the performance parameters for different components that the NoC simulator uses for deriving network performance for a 256-core system divided into 16 subnets.

The NoC switch architecture is adopted from [46]. The hubs and the NoC switches in the subnets all have 4 virtual channels per port and have a buffer depth of 2 flits. Each packet consists of 64 flits. The ports associated with the WIs have an increased buffer depth of 8 flits, which ensures that all the messages trying to access wireless links are efficiently handled without compromising performance [10]. Increasing the buffer depth beyond this tradeoff point does not produce any further performance improvement for this particular packet size, but will give rise to additional area overhead. The wireless ports of the WIs are assumed to be equipped with antennas and wireless transceivers. A self-similar traffic injection process is assumed.

We consider a Mesh-StarRing as the overall interconnection architecture for the mWNoC. The upper level of the hierarchy is a mesh with overlaid mm-wave wireless shortcuts and the subnets have StarRing architectures. The Mesh-StarRing network architecture is simulated using a cycle accurate simulator. The delays in flit traversals along all the wired interconnects that enable the proposed hybrid NoC architecture are considered while quantifying the performance. These delays include the intra-subnet core-to-hub and the inter-hub wired links in the upper level of the network. The delays through the switches and inter-switch wires of the subnets and also the delays through the hubs are taken into account.

### 5.2 Optimum Hierarchical Division

To determine the optimum division of the proposed hierarchical architecture in terms of achievable bandwidth,

<table>
<thead>
<tr>
<th>Components</th>
<th>Performance Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wireless Transceiver</td>
<td>36.7 mW supporting a data rate of 16 Gbps</td>
</tr>
<tr>
<td>Subnet Routers</td>
<td>7.7451 pJ with three pipeline stages operating at a clock freq. of 2.5 GHz</td>
</tr>
<tr>
<td>Inter-subnet Hubs</td>
<td>51.634 pJ with a clock freq. of 2.5 GHz</td>
</tr>
<tr>
<td>Inter-subnet and intra-subnet metal wires</td>
<td>1.06 pJ/bit for 5 mm</td>
</tr>
</tbody>
</table>

Fig. 7. Performance evaluation setup for mWNoC.

we evaluate the performance of the mWNoC by dividing the whole system in various alternative ways. This analysis is performed without any shortcuts to highlight the effect on performance resulting from different ways of doing the hierarchical division. Fig. 8 shows the achievable bandwidth for a 256-core Mesh-StarRing divided into different numbers of subnets. As can be seen from the plot, the division of the whole system into 16 subnets with 16 cores in each performs the best. Similarly, the suitable hierarchical division that achieves the best performance is determined for the other system sizes. For system sizes of 128 and 512, the optimum number of subnets turns out to be 8 and 32 respectively.

### 5.3 Optimum Number of WIs

The WIs introduce hardware overhead, and hence we aim to limit the number of WIs without significantly compromising the overall performance. As this is related to the utilization of the wireless medium, only the 2nd level of the network is considered. We undertook the network optimization analysis following the methodology elaborated in section 3. The optimum number of WIs \((n)\) obtained for different values of \(a\) for a 16 hub system with one wireless channel is shown in Fig. 9 (a). The weight parameter \(a\) determines how the cost versus the performance is weighted for the optimization fitness function. From this result it can be observed that for a moderate weight value, \(a\) (varying from 0.30 to 0.7) the optimum number of WIs varies from 4 to 12. The error bars represent the overall variation of the optimum number of WIs for different execution of the optimization process. As expected at the weight boundary values, the cost function
optimization ends with either zero or the maximum number of WIs. Thus, this analysis gives us a narrower window of possible optimum number of WIs for a particular system size. To exactly determine the optimum number of WIs, we carried out system-level simulations within this narrower window with the wireless token passing mechanism and the results are shown in Fig. 9 (b). The token is considered to be a single flit transmitted from the WI, which currently holds it to the next one. From Fig. 9 (b), it is seen that for a 256-core mWNoC (16-subnets with 16 cores in each subnet) bandwidth increases with number of WIs until reaching a maximum at 6 WIs and then it decreases. This is because although a higher number of WIs improves connectivity by reducing hop-count of the network, the shared wireless medium is distributed among the WIs, and as the number of WIs increases beyond a certain point, performance degrades due to the large token returning period. Moreover, as the number of WIs increases, the overall energy dissipation from the WIs becomes higher, and it causes the packet energy to increase as well. Considering all these factors, we determine the optimum number of WIs for 256-core mWNoC as 6. Similarly, for system sizes of 128 and 512 (consisting of 8 and 32 subnets respectively) the optimum performance is achieved with 4 and 10 WIs respectively.

5.4 Wireless Channel Characteristics

The metal zigzag antennas described earlier are used to establish the on-chip wireless communication channels. High resistivity silicon substrate ($\rho=5k\Omega\cdot cm$) is used for the simulation. To represent a typical inter-subnet communication range the transmitter and receiver were separated by 20 mm. The forward transmission gain ($S21$) of the antenna obtained from the simulation is shown in Fig. 10. As shown in Fig. 10, we are able to obtain a 3 dB bandwidth of 16 GHz with a center frequency of 57.5 GHz. For optimum power efficiency, the quarter wave antenna needs an axial length of 0.38 mm in the silicon substrate.

Since a flat channel and antenna frequency response over the entire signal bandwidth is not practical, a system-level simulation is performed in order to more accurately define the circuit design specifications.

The simulations were carried out in Simulink, with the block diagram given in Fig. 11. We assume that the gain and noise figure (NF) of the LNA are both fixed at 10 dB, and the down-conversion mixer has a gain of 0 dB and a NF of 20 dB. Thus, the overall NF is:

$$NF_{total} = 10\log\left(1 + \frac{F_{mixer}}{G_{LNA}}\right) = 13 \text{ dB},$$  \hspace{0.5cm} (8)

where $F_{LNA}$ and $F_{mixer}$ are the noise factors of the LNA and mixer, respectively, and $G_{LNA}$ is the gain of the LNA, all in linear scale. We assume that the temperature is at 323 K (50 °C). As for the transmitter, assume that the PA output power varies from –10 to –4 dBm. Therefore, the SNR at the demodulator can be calculated as:

$$SNR = P_t - \text{Path Loss} - \text{Noise Floor},$$  \hspace{0.5cm} (9)

$\text{Noise Floor} = 10\log(kT) + 10\log(BW) + NF_{total},$  \hspace{0.5cm} (10)

in which $P_t$ is the TX power, $k$ is the Boltzmann constant, $T$ is the absolute temperature, and BW is the bandwidth, which is set to be 32 GHz, considering the worst case noise forOOK modulation with 16-Gbps data rate. From Fig. 10, it is seen that the path-loss is 26.5 dB at the center frequency. Accordingly from (9), the SNR at the demodulator ranges from 19 to 25 dB. The BER performances were simulated within this range of SNR. Using these results, the required SNR was then extrapolated for the targeted BER of 10$^{-15}$, which is the BER of traditional wired links [23]. The resulting BER vs transmitted power curve is shown in Fig. 12. It can be seen that a power amplifier (PA) transmit power of at least 2.5 dBm (equivalent to an SNR of 31.5 dB) is needed.

In [47], simulations with root-raised cosine (RRC) pulse shaping filters in both TX and RX gave a 5 dB performance gain, which in principle would allow reduction of the PA transmit power by 5 dB. However, circuit implementation of a low-power and sufficiently wideband RRC matched filter in the receiver working in the 60 GHz range remains a challenging problem for the designed mWNoC. Hence, the present paper employs rectangular pulse shaping.

According to the simulation results, it is now possible to redistribute the SNR budget to RF building blocks in the system. Specifically, for a low power consumption of the TX, we decreased $P_{t,min}$ to -0.5 dBm, so that the corresponding $NF_{total}$ needs to be 10 dB at the receiver, in order to maintain the target BER. Therefore, a gain of greater than 13 dB and an NF less than 7 dB are required at the LNA. The mixer would still need to achieve less than 20 dB NF.
With the design target set by the system-level simulation, the wireless transceiver circuitry was designed and simulated using TSMC 65-nm standard CMOS process.

The overall conversion gain and double-sideband (DSB) NF of the receiver at 27 °C with typical-typical (TT) process corner are illustrated in Fig. 13. As can be seen, the conversion gain is 20 dB at the center frequency, and rises up to 20.5 dB at the peak. Fig. 13(a) also shows that the overall 3-dB bandwidth of the receiver front-end is 18 GHz. From Fig. 13(b), it can be seen that the overall NF stays below 6 dB.

Fig. 13 also demonstrates the process and temperature variation of the receiver performance. In the worst case of 85 °C with TT process corner, both the conversion gain and the NF show around 2 dB of degradation, and the 3-dB bandwidth shrinks to 16.5 GHz. Moreover, at 27 °C with slow-slow (SS) process corner, approximately 1.5 dB of gain degradation is observed, and the NF increases by 1 dB. However, the 3-dB bandwidth remains 18 GHz. Therefore, even with process and temperature variations, the achieved bandwidth and NF of the receiver are still better than the design targets of 16 GHz and 10 dB, respectively.

The conversion gain of the transmitter is illustrated in Fig. 14. The transmitter has a peak gain of 15 dB, and a 3-dB bandwidth of 18.1 GHz. Furthermore, circuit simulation also shows that the output 1-dB gain compression point ($P_{1db}$) of the transmitter is 0 dBm.

The achieved aggregate power consumption of the entire transceiver is 36.7 mW, 16% lower than the previous design without using body-enabled techniques [48]. It is able to support a data rate of at least 16-Gbps, and a BER $< 10^{-15}$ using an OOK modulation scheme [47] for a communication range of 20 mm.

All the transceivers work in the same frequency range, making the overall design modular and scalable. Area overhead is minimized since only one antenna per transceiver is needed. As mentioned earlier, a token passing protocol is used to select which transceiver will use the wireless channel at any particular time, thereby removing the possibility of channel contention. The omnidirectionality of the zigzag antennas allows essentially equal antenna gains for all pairs of wireless transceivers on the chip. Thus the combination of token passing protocol and zigzag antenna provides a great deal of flexibility in mWNoC design.

### 5.5 Achievable Bandwidth with Uniform Traffic

In this section we analyze the characteristics of the proposed mWNoC and study trends in its performance as the system size scales up. Fig. 15 shows the bandwidth of the proposed mWNoC for the three different system sizes considered under a uniform random spatial traffic distribution. For comparison, we also present the bandwidth of five alternative architectures of the same size: (i) flat mesh, (ii) the same hierarchical architecture as the mWNoC, but without any long-range links (iii) hierarchical architecture as the mWNoC, but long range links implemented with RF interconnects (RFNoC) [14], (iv) hierarchical architecture as mWNoC, but long range links implemented with CNT antenna based THz wireless interconnects (THzNoC) [4] and (v) hierarchical architecture as the mWNoC, but with shortcuts implemented using buffered metal wires (BWNoC) instead of the wireless links. Due to the short range of communication, UWB based on-chip wireless interconnects proposed in [15] is not considered as another alternative to establish the shortcuts in the hierarchical small-world networks. We have also shown in our previous studies that UWB NoC dissipates significantly more energy compared to the THzNoC [4].

We designed a small-world RFNoC by replacing the wireless communication channel of the mWNoC by the RF-I, maintaining the same hierarchical topology. As mentioned in [3], in 65nm technology it is possible to have 8 different frequency channels each operating with a data rate of 6 Gbps. Like the wireless channel, these RF links can be used as long-range shortcuts in the hierarchical NoC architecture. These shortcuts are optimally placed using the same SA based optimization as used for placing the WIs in the mWNoC.

We also designed THzNoC using nanoscale antennas based on CNTs operating in the THz/optical frequency range as long range wireless shortcuts in mWNoC architecture. There can be 24 different wireless shortcuts each operating at 10 Gbps data rate [49]. These shortcuts are placed optimally using the same optimization method.

In case of BWNoC, the numbers of wired shortcuts are kept equal to the number of WIs for different system sizes and they are also optimally placed using the same opti-
mWNoC, THzNoC and BWNoC perform better than mWNOC because multiple shortcuts can work simultaneously in them, whereas in mWNOC (where the wireless channel is a shared medium) only one pair can communicate at a particular instant of time. But, BWNoC suffers from significant energy dissipation overhead, which is quantified in section 5.6. Though THzNoC shows better performance than mWNOC, it is not a CMOS compatible solution and the integration and reliability of CNT devices need more investigation. Similarly, the total long-range link area overhead and the layout challenges of RFNoC are more significant compared to mWNOC. For example, for a 20 mm x 20 mm die, an RF interconnect of approximately 100 mm length has to be allocated for RFNoC following the layout of [14]. This is significantly higher than the combined length of all the antennas used in the mWNOC, which is 3.8 mm for the highest system size (512 cores) considered in this paper.

5.6 Energy Dissipation for Uniform Traffic

To quantify the energy dissipation characteristics of the proposed mWNOC architecture, we determine the packet energy dissipation, \( E_{\text{pkt}} \). The packet energy is the energy dissipated on average by a packet from its injection at the source to delivery at the destination. This is calculated as

\[
E_{\text{pkt}} = \frac{N_{\text{intrasubnet}}E_{\text{subnet,hop}}h_{\text{subnet}} + N_{\text{intersubnet}}E_{\text{sw}}+h_{\text{sw}}}{N_{\text{intrasubnet}} + N_{\text{intersubnet}}} \tag{11}
\]

where \( N_{\text{intrasubnet}} \) and \( N_{\text{intersubnet}} \) are the total number of packets routed within the subnet and between the subnets respectively, \( E_{\text{subnet,hop}} \) is the energy dissipated by a packet traversing a single hop on the wired subnet including a wired link and a switch, and \( E_{\text{sw}} \) is the energy dissipated by a packet traversing a single hop on the 2nd level of the mWNOC network, which has the smallest-world property. The average number of hops per packet in the subnet and the upper level small-world network are denoted by \( h_{\text{subnet}} \) and \( h_{\text{sw}} \) respectively. Fig. 16 shows the packet energy dissipation for the considered architectures under uniform random traffic. The energy dissipation for RF-I and CNT based interconnect is obtained from [14] and [4] respectively. The flat mesh architecture dissipates highest packet energy among all the NoC architectures considered. A hierarchical network reduces the average hop count, and hence the latency between the cores compared to a flat mesh. Packets get routed faster and hence occupy resources for less time and dissipate significantly less energy compared to flat mesh in the process.

In Fig. 17 (a) we show the variation of per bit energy dissipation with distance for a single wired and a mm-wave wireless link. Fig. 17 (b) highlights the contributions of the different components of the packet energy dissipation for 256-core mWNOC and flat mesh architecture. The contributions of the antenna and the transceiver, which constitute the wireless link energy, are shown separately from the wireline links of the upper level small-world network. The largest contribution to packet energy in mWNOC is from the wireless and wireline link traversals combined in the upper level small-world network. This is because on an average a large portion of the packets travels through the upper level of the mWNOC to reach other subnets. However as this level has very small average path length due to its small-world nature and due to the low power wireless channels the absolute value of this energy dissipation is small. It can be observed that the energy dissipation in wireless and RF-I transmission is much less compared to long metal wire interconnects. From Fig. 16, it can be observed that a 512-core hierarchical NoC with buffered wire shortcuts (BWNoC) is significantly more compared to the other NoC architectures (mWNOC, RFNoC and THzNoC). This is because the energy dissipation in wireless and RF-I transmission is much less compared to long metal wire interconnects. From Fig. 16, it can be observed that a 512-core hierarchical NoC with buffered wire shortcuts burns 12.79 times more energy yet achieves only 1.46 times more bandwidth compared to mWNOC. All three small-world NoC architectures with emerging interconnect technologies, viz., mWNOC, RFNoC and THzNoC dissipate significantly less packet energy than the other three alternatives. The THzNoC has the lowest packet energy dissipation and the difference in packet
energy values between RFNoC and mWNoC is small. But RFNoC and THzNoC have their implementation challenges compared to mWNoC as mentioned earlier.

Due to the high energy dissipation, flat mesh, hierarchical NoC without shortcuts and hierarchical NoC with multiple metal wire shortcuts are not considered for the subsequent analysis.

The Mesh-StarRing architecture along with the routing mechanism elaborated in section 4.3 results in 14.6% bandwidth improvement and 48% savings in packet energy for a 256-core system with 6 WIs in comparison with the previously proposed NoC architecture with mm-wave wireless links [48] for the same system size.

5.7 Performance Evaluation with Non-uniform Traffic

In order to evaluate the performance of the proposed NoC architecture with non-uniform traffic patterns we considered both synthetic and application based traffic distributions. In the following analysis, the system size considered is 256 (with 16 subnets and 16 cores per subnet) with 6 WIs as a representative case.

We considered two types of synthetic traffic to evaluate the performance of the proposed mWNoC architecture. First, a transpose traffic pattern [7] is considered where a certain number of cores are considered to communicate more frequently with each other. We considered three such pairs and 50% of packets originated from one of these cores are targeted towards the other in the pair. The other synthetic traffic pattern considered is hotspot traffic [7], where each core communicates with a certain number of cores more frequently than with the others. We considered three such hotspot locations to which all other cores send 50% of the packets that originate from them. In both of these situations, the communicating cores are considered to be in different subnets so that the 2nd level of the network is used in the data exchange. As an application-based traffic, a 512-point Fast Fourier Transform (FFT) is considered and each core is assigned to perform a 2-point radix 2 FFT computation. The traffic pattern generated in performing multiplication of two 256 x 256 matrices was also used to evaluate the performance of the mWNoC.

5.8 Performance Evaluation with Broadcast Traffic

Though traditional NoC supports many concurrent transactions, they do not directly support broadcast. There exists a variety of SoC applications that require broadcast, e.g., passing global states, managing and configuring the network, implementing cache coherency protocols, etc. Due to the broadcasting capability of the mm-wave wireless channels, mWNoC is capable of incorporating broadcasting efficiently. Broadcasting can be implemented in the proposed mWNoC by employing the wireless links in broadcast mode.

Fig. 19 shows the achievable bandwidth of mWNoC, RFNoC and THzNoC in presence of broadcast traffic for a 256-core system. The number of broadcast source and destinations are kept identical for all the NoCs under consideration. The results show that mWNoC performs better than RFNoC and THzNoC. Due to the inherent broadcasting capability of mWNoC, all the WIs can receive the broadcast at 16 Gbps data rate. This gives mWNoC higher overall bandwidth than RFNoC’s RF-I based point to point shortcuts (6Gbps each) and THzNoC’s point to point wireless shortcuts (10 Gbps each). Since all the WIs can receive the broadcast traffic at higher bandwidth, the overall performance of mWNoC is better in case of broadcast traffic scenario.

5.9 A mWNoC with Multiple Simultaneously Operating Channels

The Performance of mWNoCs can be significantly im-

Fig. 18. Achievable Bandwidth with different traffic scenarios

Fig. 19. Achievable Bandwidth for different NoCs with broadcast traffic.

Fig. 20. Antenna transmission gain (S21) for three non-overlapping channels.
proved by optimally placing and using multiple simultaneously operating wireless shortcuts. The works of [23] and [24] have already discussed the feasibility and advantages of multichannel wireless interconnects in the NoC environment. We extended our single channel mWNoC design to mWNoC with three simultaneously operating channels. The antenna's forward transmission gains (S21) obtained via simulations are shown in Fig. 20. We are able to obtain three different non-overlapping channels with 3 dB bandwidths of 16 GHz and center frequencies of 31, 57.5 and 120 GHz respectively. For optimum power efficiency, the quarter wave antennas use axial lengths of 0.73, 0.38 and 0.18 mm respectively in the silicon substrate. The antenna design ensures that signals outside the communication bandwidth for each channel are sufficiently attenuated to avoid inter-channel interference. The wireless transceiver circuitry is designed and simulated using TSMC 65-nm CMOS process.

Multiple non-overlapping wireless channels are distributed among the hubs and the WIs sharing the same channel form a cluster. Since each channel is shared between relatively smaller number of WIs, the optimum number of WIs increases from a single channel case. The technique for finding out optimum number of WIs and SA based optimization algorithm discussed in section 3.2 are used for optimally distributing the WIs belonging to three different clusters. The WI clusters are equal in size and a single WI with transceivers of all frequencies acts as gateway between different clusters for intercluster wireless communication. For a 256-core (16 subnets with 16 cores per subnet) and 512-core (32 subnets with 16 cores per subnet) mWNoC, with three non-overlapping wireless channels, the optimum number of WIs is found to be 7 (3 clusters of 2 WIs each and a gateway) and 13 (3 clusters of 4 WIs each and a gateway) respectively.

The achievable bandwidths for 256-core and 512-core system mWNoCs with three simultaneously operating wireless channels, RFNoC and THzNoC is shown in Fig. 21. Since the performance improvements are more prominent in larger systems, results for 128-core systems are not shown here. It can be observed that the performance difference decreases considerably among these NoCs as the number of simultaneously operating wireless channels increases for the mWNoC. The mWNoC (Three Channels) performs better than RFNoC and the performance difference between THzNoC and mWNoC (Three Channels) is smaller than that with mWNoC with single channel. The fact that mWNoC can establish communication channel between any WI pair unlike RFNoC and THzNoC where fixed point to point communication links are assigned, also helps in achieving improved performance. Therefore, it can be concluded that mm-wave based long range interconnects can be a viable and efficient alternative interconnect in future many core NoCs. The technological challenges of making mWNoC practically feasible are significantly lower than the THzNoC with CNT based wireless links.

6 AREA OVERHEAD

In this section we quantify the area overhead due to the wireless deployment in the mWNoC. The antenna used is a 0.38 mm long and 58 μm wide zigzag antenna. The area of the transceiver circuits required per WI is the total area required for the OOK modulator/demodulator, LNA, PA and VCO. The total area overhead per wireless transceiver turns out to be 0.3 mm$^2$ for the selected frequency range. The digital part for each WI, which is very similar to a traditional wireline NoC switch, has an area overhead of 0.40 mm$^2$. Therefore, the total area overhead per hub with a WI (inclusive of transceiver and antenna) is determined to be 0.72 mm$^2$. Since the number of WIs is kept limited, the overall silicon area overhead is dominated by the wireline NoC switches. For example, in case of a 256 core mWNoC, 6 wireless transceivers consume only 4.8 % of total silicon area overhead. The transceiver area overhead for RFNoC and THzNoC is obtained from [14] and [4] respectively.

Total silicon area overheads for flat mesh, mWNoC, RFNoC, THzNoC and BWNoC for a 256-core system are shown in Fig. 22. The required silicon areas are dominated by the NoC intra-subnet switches. The area overheads of the hubs along with the required transceivers (mWNoC, RFNoC, THzNoC) and buffers (BWNoC) are shown separately. The transceiver area overhead for mWNoC is marginally higher than RFNoC, THzNoC and BWNoC. Though the overall silicon area for mWNoC, RFNoC, THzNoC and BWNoC are higher than flat mesh, the performance benefit of these hierarchical NoCs with shortcuts clearly outweighs the associated overhead. Fig. 23 shows the total wiring requirements of various lengths for a 20 mm x 20 mm die for a 256-core system of Mesh-
StarRing configuration considered in this work. The wiring requirements for a flat mesh architecture are shown for comparison. The hierarchical architecture has no inter-subnet direct core to core links as inter-subnet communication occurs through the hubs; this eliminates a number of wireline links along the subnet boundaries which are present in the flat mesh topology. RFNoC and BWNoC require extra long range links for inter-subnet communication whereas for mWNoC and THzNoC these communications are predominantly carried out by wireless links.

7 CONCLUSIONS AND FUTURE WORK

In this paper we have proposed the design of a small-world NoC architecture with mm-wave wireless interconnects used as long-range links. Design of associated broadband and low power mm-wave transceivers is highlighted. The mm-wave wireless NoC (mWNoC) outperforms its more traditional wireline counterpart in terms of achievable bandwidth and energy dissipation in the presence of various synthetic and application specific traffic patterns. Performance of the proposed mWNoC is evaluated with respect to other small-world architectures where the long-range links are implemented with RF interconnects (RF-I) and CNT antenna based wireless links. Though RFNoC and THzNoC perform better compared to the mWNoC, the difference in performance is small. Moreover, with multiple non-overlapping channels mWNoC performs better than RFNoC and the performance gap with THzNoC becomes smaller. The THzNoC faces manufacturing and integration challenges; by contrast the mWNoC is CMOS compatible and does not require any new technology. Therefore, it can be concluded that mWNoC achieves the best performance-energy-area-technological challenge tradeoff among all of the emerging interconnects compared in this paper.

As part of this on-going investigation, we intend to establish a detailed performance benchmark for the proposed mWNoC with respect to other emerging NoC architectures, such as 3D and photonic NoCs.

ACKNOWLEDGMENT

This work was supported in part by the US National Science Foundation (NSF) CAREER grant (CCF-0845504) and CRI grant (CNS-1059289) and NSF CAREER Grant (ECCS-0845849).

REFERENCES
