Enhancing Performance of Network-on-Chip Architectures with Millimeter-Wave Wireless Interconnects

Sujay Deb, Amlan Ganguly, Kevin Chang, Partha Pande, Benjamin Belzer, Deuk Heo
School of Electrical Engineering and Computer Science, Washington State University, USA
{sdeb,ganguly,jchang,pande,belzer,dheo}@eecs.wsu.edu

Abstract

In a traditional Network-on-Chip (NoC), latency and power dissipation increase with system size due to its inherent multi-hop communications. The performance of NoC communication fabrics can be significantly enhanced by introducing long-range, low power, high bandwidth direct links between far apart cores. In this paper a design methodology for a scalable hierarchical NoC with on-chip millimeter (mm)-wave wireless links is proposed. The proposed wireless NoC offers significantly higher throughput and lower energy dissipation compared to its conventional multi-hop wired counterpart. It is also demonstrated that the proposed hierarchical NoC with long range wireless links shows significant performance gains in presence of various application-specific traffic and multicast scenarios.

1. Introduction

The Network-on-Chip (NoC) has emerged as a revolutionary methodology to integrate numerous blocks in a single chip [1]. 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 [2]. 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 to use on-chip millimeter (mm)-wave wireless links designed in traditional CMOS technology as long-range communication channels between far apart cores in a NoC. 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 [3]. 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. Simulations demonstrate that the proposed mWNoC outperforms its more traditional wireline counterparts in terms of sustainable data rate and energy dissipation.

2. Related Work

Conventional NoCs use multi-hop packet switched communication. By using virtual express lanes to connect far apart cores in the network, it is possible to avoid the router overhead at intermediate nodes, and thereby greatly improve NoC performance in terms of power, latency and throughput [4, 5]. Performance of NoCs has also been improved by inserting long range wired links following principles of small world graphs [6].

The design principles of a photonic NoC are elaborated in various recent publications [7, 8]. It is estimated that a photonic NoC will dissipate an order of magnitude less power than an electronic planar NoC. Another alternative is NoCs with multi-band RF interconnects [9]. 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 [10]. It involves multi-hop communication through the on-chip short-range wireless channels. However the performance of silicon integrated on-chip antennas for intra- and inter-chip communication with longer range have been already demonstrated by the authors of [11]. In [12], the feasibility of designing miniature antennas and simple transceivers that operate at frequencies from 100-500 GHz for on-chip wireless communication has been demonstrated.

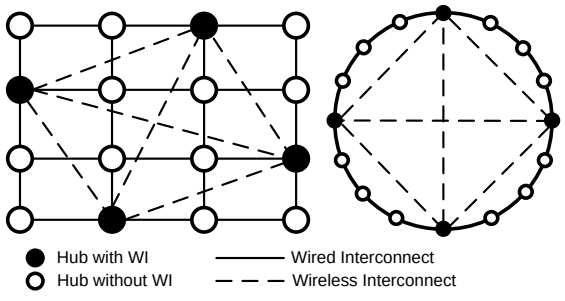


Figure 1. Network topology of hubs connected by a small-world graph with wired links and Wireless Interfaces (WIs)

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

3. MM-Wave Wireless NoC (mWNoC) Architecture

A traditional wired NoC consists of embedded cores that communicate via switches and wired links. Between a pair of source and destination cores, communication is generally multi-hop, resulting in high energy dissipation and latency. With increasing system size the problem of higher energy dissipation and latency becomes more profound. To alleviate this problem we propose a scalable hierarchical NoC architecture with wireless interfaces strategically placed for optimized performance. In the following subsections we discuss the topology of the proposed hierarchical architecture and the adopted performance optimization methodology.

3.1 Proposed Architecture

The problem of high latency in data transfer with increasing system size in a conventional NoC can be addressed by incorporating features of small-world networks. A small-world topology can be constructed from a locally connected network by re-wiring connections from any given node randomly to any other node, which creates short cuts between distant nodes in the network [2]. Use of conventional metal wires to establish direct links between such distant cores will incur significant penalty in terms of latency as well as energy dissipation. On the other hand, establishment of one-hop communication between the distant cores using on-chip wireless interconnects will result in significant performance gain due to low power and higher bandwidth over long distances. In our proposed design, we use mm-wave on-chip wireless links as shortcuts. The whole NoC is divided into smaller networks called *subnets* consisting of a relatively smaller number of cores interconnected through conventional wireline NoC architectures. Each subnet has a centrally located hub, which is connected to all cores of the subnet. The hubs are then interconnected using both wired and wireless links following a small world topology in the upper level of the hierarchy. In Fig. 1 mesh and ring topologies as examples for interconnection of the hubs are shown. A few chosen hubs are equipped with a *Wireless Interface* (WI) according to the methodology outlined in section 3.2. Thus the whole NoC is a hierarchical architecture with subnets at the lower level and a small-world topology of hubs at the upper level.

A hub is similar to a conventional NoC switch, with multiple ports connected to each core within its subnet via wired links. If a hub has a WI then that acts as an additional port connected to the wireless medium for wireless data transfer. Each hub has a routing block that routes flits to the next hub either through a wired or a wireless link depending upon the adopted routing strategy outlined in section 4.3. It should be noted that, instead of a ring or mesh, the hubs can be connected following any other interconnection architecture. The only restriction is that the wireless links need to be utilized as long-range links to fully leverage their low-power and high bandwidth. The design and components of the WI are discussed in section 4.

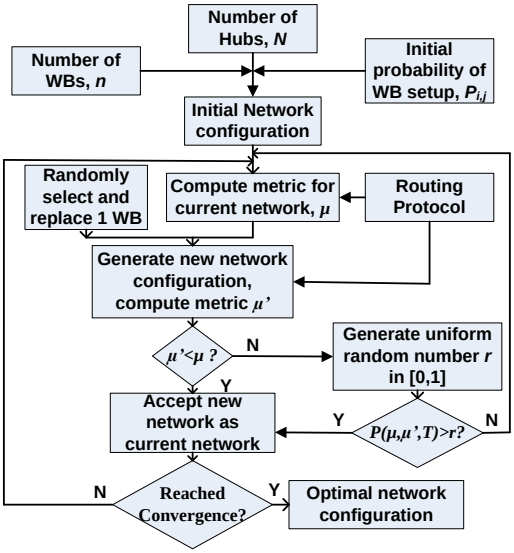


Figure 2. Flow diagram for the simulated annealing based optimization of mWNoC architectures

3.2 Placement of WIs

The WI placement is crucial for optimum performance gain as it establishes high-speed, low-energy interconnects on the network. We adopt a Simulated Annealing (SA) [13] based optimization technique for placement of the WIs to get maximum benefits of using this novel technology. Initially, the WIs are placed randomly with each hub having equal probability of having 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 optimization step is performed using SA. Since the deployment of WIs is only on the hubs, the optimization is performed solely on the 2nd level network of hubs. SA offers a simple, well established and scalable approach for the optimized placement of WIs as opposed to a brute force search. If there are N hubs in the network and n WIs to distribute, the size of the search space S is given by

$$|S| = \binom{N}{n} . \quad (1)$$

Thus, with increasing N, it becomes increasingly difficult to find the best solution by exhaustive search. To perform SA, a metric μ has been established, which is closely related to the connectivity of the network. The metric μ is the average distance, measured in number of hops, between all source and destination hubs. To compute μ the shortest distances between all pairs of hubs are computed following the

routing strategy outlined in the next section. The distances are then weighted with a normalized frequency of communication between the particular pair of hubs. The optimization metric, μ can be computed as

$$\mu = \sum_{i,j} h_{ij} f_{ij} , \quad (2)$$

where h_{ij} is the distance (in hops) between the i^{th} source and j^{th} destination. The frequency f_{ij} of communication between the i^{th} source and j^{th} destination is the apriori probability of traffic interactions between the subnets determined by particular traffic patterns depending upon the application mapped onto the NoC. In this case, equal importance is attached to both inter-hub distance and frequency of communication. The algorithm used to optimize the network is shown in Fig. 2. In this work we adopt the Cauchy annealing schedule where the temperature profile varies inversely with the number of iterations as

$$T = \frac{T_0}{k} , \quad (3)$$

where T is the temperature profile, T_0 is the initial temperature and k is the current annealing step. The convergence criterion 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. 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 zig-zag antenna [11] 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 [3]. The zig-zag antenna is designed using the top layer of the metal with 10μm trace width, 60μm arm length and 30° bend angle. The axial length depends on the operating frequency of the antenna which is determined in section 5.1. The details of the antenna structure are shown in Fig. 3.

4.2 Wireless Transceiver Circuit

The mm-wave transceiver is a part of the WIs in the hubs. The transmitter accepts data flits to be transmitted and after serialization and modulation, transmits the data stream through the antenna. The receiver picks up the signal using

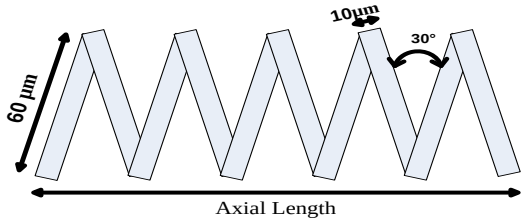


Figure 3. Zig-zag antenna configuration details

the on-chip antenna, and then amplifies, demodulates and de-serializes before passing the received data to the appropriate module. The block diagram of the wireless transceiver is shown in Fig. 4.

In this work, on-off-keying (OOK) modulation has been used, and the OOK modulator and demodulator design is adopted from [18]. The Low Noise Amplifier (LNA) in our work uses an inductively degenerated common-source amplifier coupled with an input network. The input network is a multi-section reactive network that results in the overall input reactance causing resonance over a wider bandwidth. The LNA consists of wideband input impedance matching networks, a two-stage cascode amplifier with shunt-peaked load, and an output buffer. The cascode topology with shunt-peaking is used to achieve better isolation and higher conversion gain. The Serializer De-Serializer (SERDES) is implemented with an oscillator block and MUX. The MUX sequentially selects one of the inputs per clock cycle. The oscillator provides the clock to the MUX. The amplifier in the transmitter consists of single-ended, common-source stages biased in the class AB region for linearity and efficiency. A low insertion loss, high isolation wideband single pole double throw switch (SPDT) is used to switch between the transmitting and receiving modes in the transceiver.

4.3 Adopted Routing Strategy

In this proposed hierarchical NoC data is transferred via flit-based wormhole routing [1]. Intra-subnet data routing is done according to the topology of the subnets. For example, if the subnet architecture is a mesh then the routing is dimension order or e-cube routing. Inter-subnet data routing however requires the flits to use the small-

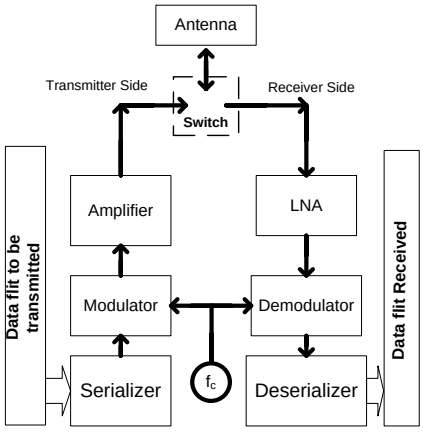


Figure 4. The Wireless Transceiver

world 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 between them. Therefore, 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 [19]. 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. However, only if the destination address matches the address of the receiving hub then the flit is accepted for further routing either to an adjacent hub or to a core in the subnet of that hub. The token is released to the next hub with a WI after all flits belonging to a single packet at that hub are transmitted. These flits could be either already buffered or could have arrived when other flits of the same packet were being transmitted from the WI. Hence, the time for which the token is held at any particular hub depends not only on the depth of the FIFO buffers but also the arrival rate of flits at the hubs. Inter-subnet data can be routed either through wired links connecting the hubs or through a wireless shortcut via the hubs with WIs. 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 broadcast on arrival of a token. 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 wireless link is chosen if it reduces the path length compared to the wired path.

5. Experimental Results

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

5.1 Wireless Channel Characteristics

The metal zig-zag antennas described earlier are used to establish the on-chip wireless communication channels. The characteristics of the antennas are simulated using the ADS momentum tool [21]. High resistivity silicon substrate (ρ=5kΩ-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. 5. As shown in Fig. 5, we are able to obtain a 3 dB bandwidth of 16 GHz with a center frequency of 62.5 GHz. For optimum power efficiency, the quarter wave antenna needs an axial length of 0.33 mm in the silicon substrate [14]. For a sufficiently low bit error rate (BER) of 10^{-21}, OOK modulation requires a SNR of 20dB. The receiver noise floor N_{floor} for this center frequency and bandwidth was computed to be $-59\,dBm$. (Power in dBm is equal to $10\log_{10}(\text{power in milliWatts})$.) The receiver sensitivity P_r and the required transmitter power are given by

$$P_r = N_{floor} + SNR, \quad (4)$$

$$P_t = P_r - G_a \ . \quad (5)$$

Using the values of SNR and N_{floor}, P_r is calculated to be $-39dBm$. From antenna simulations we measure the worst-case antenna pair gain G_a [11] to be $-37.52dB$ in the frequency band of interest. So, the worst-case required transmitter power P_t is calculated using (5) to be $-1.48dBm$ or 711 μW. This level of transmitted power can be easily generated on-chip.

At 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 [15]. It has been found that short metal lines running over antennas have negligible impact on antenna gain and phase, and the dummy fills needed for chemical mechanical polishing (CMP) also have negligible effect. Hence other metal wires in the vicinity of these antennas do not significantly affect their performance. 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 [16] 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 [15]. It is shown that immunity from digital switching noise can be improved by selecting the antenna signal frequency to be much greater than the circuit frequency. Using antennas with a differential or balanced feed structure can significantly reduce coupling of switching noise, which is mostly common-mode in nature [17]. Hence, the proposed zig-zag antenna operating at 62.5GHz can be designed to be immune to digital switching noise and is suitable for the mWNoC wireless communication links.

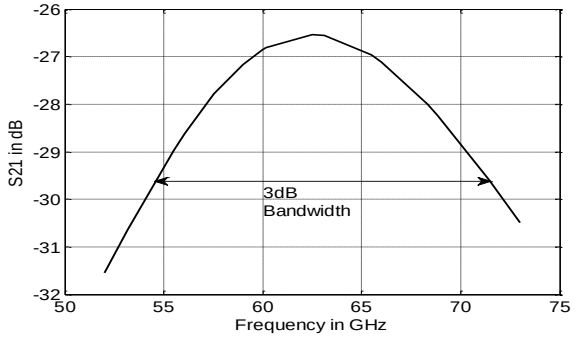


Figure 5. Simulated forward transmission gain (S21) for zig-zag antenna

The wireless transceiver was designed using the CMP [20] 65 nm CMOS process and its performance obtained through Cadence simulations. The designed transceiver can sustain a data rate of 16 Gbps with a power dissipation of 90 mW.

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 fact that the relative angle between transmitter and receiver antenna does not affect the received signal strength for zig-zag antennas significantly, enable any pair of wireless transceivers to be chosen on the chip. Thus the combination of token passing protocol and zig-zag antenna provides a great deal of flexibility in mWNoC design.

5.2 Characteristics of Token Passing Network

The WIs introduce hardware overhead, and hence we would like to limit the number of WIs on the chip without significantly affecting the overall performance. In Fig. 6 we present the variation of time taken by the token to return to a particular WI, as a function of the number of WIs in the NoC. We assume round-robin token circulation among WIs. The token is considered to be a single flit transmitted from the WI which currently holds it to the next one. The smaller this period of token return to a particular WI is, the better the network performance will be since the delay in acquiring the wireless medium will be minimized. We also show the reduction in hop-count due to introduction of WIs for a system with 32 hubs in the upper level of the network connected as a mesh. The hop-count decreases with the number of WIs due to higher connectivity. Since these are two opposing trends, a tradeoff needs to be established. From Fig. 6, it is clear that for a 32-subnet system 10 WIs give the best trade-off, which is validated from our network-level simulations as well. Similarly, for 8 and 16 subnet systems the corresponding numbers of WIs are 4 and 6 respectively. Since the circulation of the token is only between the WIs, it is independent of the wireline interconnection of the hubs. Hence, when the upper level network is a ring, simulations show the same trend.

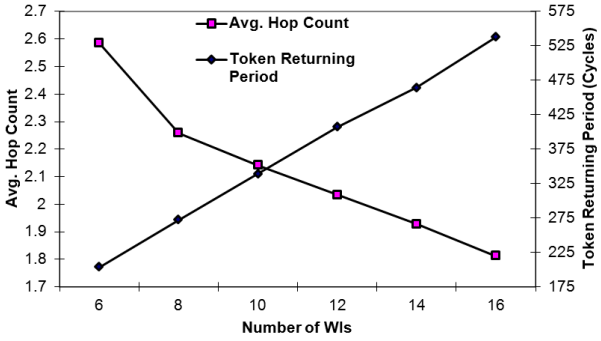
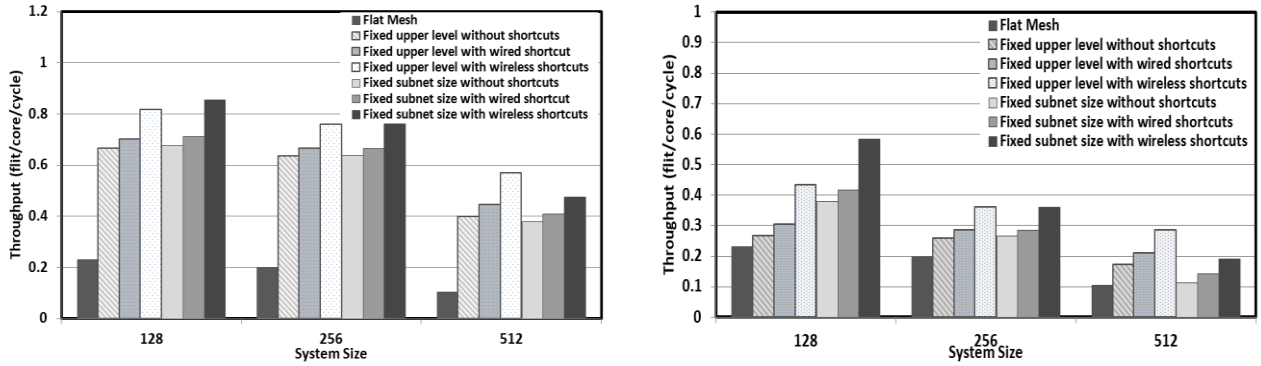


Figure 6. Avg. Hop Count vs. Token Returning Period with different number of WI for a 32 Subnet System

5.3 Performance Evaluation

In this section we analyze the characteristics of the proposed mWNoC architectures and study trends in their performance as the system size scales up. For our experiments we consider three different system sizes, namely 128, 256, and 512 cores. We observe results of scaling up the system size by following two strategies. In one, we increase the number of subnets by keeping the size of each subnet fixed. In the other, we increase the number of cores in a subnet while the total number of subnets is fixed. Hence, in one scenario, we fix the number of cores per subnet at 16 and vary the number of subnets between 8, 16, and 32. In the other case, we keep the number of subnets fixed at 16 and vary the size of the subnets from 8 to 32 cores.

Establishment of WIs using simulated annealing depends only on the number of hubs on the 2nd level of the network. The cores within each subnet are considered to be connected in a mesh topology. Each core also has a direct path to the central hub. The upper level of the network is considered to be connected using two different topologies namely, ring and mesh. For both these hub topologies, the appropriate number of WIs is optimally deployed among the hubs following the SA methodology. The subnet switch



(a) **(b)**

Figure 7. Saturation throughput with varying system size (a) Mesh-Mesh Architecture (b) Ring-Mesh Architecture

architecture is adopted from [1] and it has three functional stages, namely, input arbitration, routing/switch traversal, and output arbitration. 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. Similar to the intra-subnet communication, we have adopted wormhole routing in the wireless channels too. The hubs have similar architectures as the NoC switches in the subnets. Hence, each port of the hub has same input and output arbiters, and an equal number of virtual channels with same buffer depths. The ports associated with the WIs have an increased buffer depth of 8 flits to avoid excessive packet dropping while waiting for the token. Increasing the buffer depth beyond this number does not produce any further performance improvement for this particular packet size. The wireless ports of the WIs are assumed to be equipped with antennas and wireless transceivers. A self-similar traffic injection process is assumed.

The network architectures developed earlier are simulated using a cycle accurate simulator, which models the progress of data flits accurately per clock cycle accounting for flits that reach their destinations as well as those that are dropped.

The subnet switches and the digital components of the hubs are synthesized using 65 nm standard cell library from CMP [20] at a clock frequency of 2.5GHz. The delays in flit traversals along all the wired interconnects that are introduced to enable the proposed wireless NoC architecture were considered while quantifying the performance. These include the intra-subnet core to hub wired links and the inter-hub links in the upper level of the network. The delays through the switches and inter-switch wires of the subnets and the hubs are taken in account too.

Fig. 7 shows the saturation throughput of the proposed mWNoC for the 3 different system sizes considered under a uniform random spatial traffic distribution. Saturation throughput is defined as the average number of flits successfully received per embedded core per clock cycle when the NoC is in saturation phase. For comparison we also present the saturation throughputs of three alternative architectures of same size: (a) a flat mesh (b) the same hierarchical architecture as the mWNoC, but without any long-range links and (c) the same hierarchical architecture as the mWNoC, but with shortcuts implemented using normal metal wires instead of the wireless links. Figs. 7 (a) and (b) show the throughput for Mesh-Mesh and Ring-Mesh architectures where the basic upper level topology is a mesh and a ring respectively. The subnets are mesh-based in both the cases. It can be observed that the mWNoC outperforms all the other three alternatives for the various system sizes under consideration. Flat mesh architecture performs worst among all the cases due to its highest average hop count. The hierarchical architecture improves the performance by reducing hop count, but the best performance is obtained from hierarchical architecture with shortcuts due to small world nature of the network. Wireless shortcuts outperform wired shortcuts since the later provide multi-hop path between distant cores, whereas wireless links provide a single hop path, making mWNoC the best architecture.

It is also observed that, the maximum achievable throughput in mWNoCs degrades with increasing system size for both Mesh-Mesh and Ring-Mesh cases, but Mesh-Mesh always outperforms Ring-Mesh architecture. Moreover, by scaling up the subnet size or the upper level of the network, the trend in throughput change is similar for both Mesh-Mesh and Ring-Mesh architectures. By increasing the subnet size, we are increasing congestion in the wired subnets and load on the hubs whereas by increasing the upper level of network (i.e., the number of subnets and WIs) more traffic is routed through the wireless shortcuts. Since the wireless medium is shared among many hubs, the overall throughput deteriorates as upper level size increases. From these simulation results, an appropriate configuration for a particular system size can be selected depending on performance requirements.

5.4 Energy Dissipation

To quantify the energy dissipation characteristics of the proposed mWNoC architecture we estimate the packet energy dissipation, E_{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

Table I: Packet Energy /Bandwidth for different system sizes

System size	Subnet size	No. Of subnets	Flat mesh (nJ/Tbps)	Ring-mesh architecture (nJ/Tbps)			Mesh-mesh architecture (nJ/Tbps)		
				Without shortcut	With wired shortcuts	With wireless shortcuts	Without shortcuts	With wired shortcuts	With wireless shortcuts
128	16	8	560.04	161.02	115.11	30.25	75.11	58.32	17.87
	8	16		266.67	185.77	43.34	83.25	59.58	19.25
256	16	16	721.68	231.65	148.13	27.98	78.46	57.65	11.36
512	32	16	1171.87	330.83	155.85	29.36	124.69	57.64	10.21
	16	32		533.58	277.23	53.88	150.75	86.75	10.89

$$E_{pkt} = \frac{N_{intrasubnet} E_{subnet,hop} h_{subnet} + N_{intersubnet} E_{s-w} h_{s-w}}{N_{intrasubnet} + N_{intersubnet}} \quad (6)$$

where $N_{intrasubnet}$ and $N_{intersubnet}$ are the total number of packets routed within the subnet and between the subnets respectively. $E_{subnet,hop}$ is the energy dissipated by a packet traversing a single hop on the wired subnet including a wired link and switch, and E_{s-w} is the energy dissipated by a packet traversing a single hop on the 2nd level of the mWNoC network, which has the small-world property. h_{subnet} and h_{s-w} are the average number of hops per packet in the subnet and the small-world network respectively. $E_{subnet,hop}$ consists of the energy dissipation of the NoC switches and links of the subnets. E_{s-w} consists of the energy dissipated in the hubs as well the inter-hub links. The energy dissipation of the NoC switches and hubs are obtained through synthesis using Synopsys tools with 65nm standard cell libraries from CMP [20]. The energy dissipated by the wireless transceiver is calculated through Cadence simulations. The energy dissipation of all the wired links are obtained from actual layout in Cadence assuming a 20mm x 20mm die area.

As all the architectures produce a different peak bandwidth, the values of the packet energy per bandwidth for the 3 system sizes considered in this work are presented in Table I. For this comparison a uniform random spatial traffic distribution is considered. Once again all the three possible alternatives mentioned in the previous sub-section were considered for the performance benchmarking. It can be seen that the energy dissipation of the hierarchical wired NoCs with or without wireline shortcuts is significantly less than that of the flat mesh architecture. This is because a hierarchical network reduces the average hop count and hence latency between cores. Thus packets get routed faster and hence occupy resources for shorter time and dissipate less energy in the process. From Table I it is also evident that the mWNoC significantly outperforms the other two possible wired hierarchical architectures. This is because the energy dissipation in a wireless transmission is much less than along a long wired interconnect. Overall the mWNoC is capable of reducing the packet energy dissipation per bandwidth by at least an order of magnitude compared to the flat wireline architecture. For higher system size, the packet energy dissipation per bandwidth of the mWNoC is reduced by multiple orders of magnitude.

5.5 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 real application based traffic distributions. In the following analysis, the system size considered is 128 (with 16 hubs and each subnet consisting of 8 cores) with 6 WIs.

We considered 2 types of synthetic traffic to evaluate the performance of the proposed mWNoC architecture.

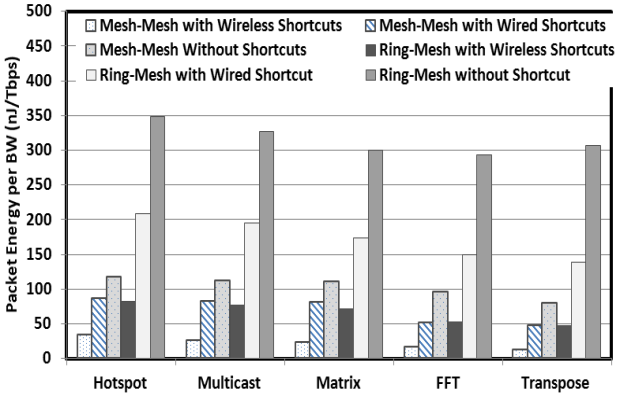


Figure 8. Packet Energy per BW with varying traffic for a 128 Ring-Mesh and Mesh-Mesh Architecture

First, a *transpose* traffic pattern [6] was considered where a certain number of hubs are considered to communicate more frequently with each other. We have considered 3 such pairs and 50% of packets generated from one of these hubs were targeted towards the other in the pair. The other synthetic traffic pattern considered was the *hotspot* [6], where each hub communicates with a certain number of hubs more frequently than with the others. We have considered three such hotspot locations to which all other hubs send 50% of the packets that originate from them. To represent a real application a 256-point Fast Fourier Transform (FFT) was considered on the same 128 core mWNoC subdivided into 16 subnets. Each core was considered to perform a 2-point radix 2 FFT computation. The traffic pattern generated in performing multiplication of two 128x128 matrices is also used to evaluate the performance of the mWNoC.

The mWNoC's performance is also evaluated in a multicasting scenario. Multicasting is delivery of information to a group of destinations simultaneously using the most efficient strategy and creating copies only when the paths to the multiple destinations split. Though traditional NoC supports many concurrent transactions, it does not directly support multicast. There exists a variety of SoC applications that require multicast, e.g., passing global states, managing and configuring the network, implementing cache coherency protocols, etc. Multicasting can be implemented efficiently in the proposed mWNoC by employing the wireless links. Efficient use of the WIs in broadcast mode results in significant performance improvements in multicast scenarios.

Fig. 8 presents packet energy per bandwidth for different traffic patterns. Results for both Ring-Mesh and Mesh-Mesh architectures are shown along with improvement achieved in presence of wireless shortcuts. From the results it is evident that for all the traffic patterns considered here highest performance gain is achieved in case of Mesh-Mesh architecture with wireless shortcuts.

6. Overheads

In this section we have quantified the area overhead due to the wireless deployment in the hierarchical network to form the mWNoC. The antenna used is a 0.33 mm long zig-zag antenna. The area of the transceiver circuits required per WI is the total area required for the OOK modulator/demodulator, the Serializer-deserializer and the LNAs. The total area overhead per wireless transceiver turns out to be 1.12 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 0.40mm^2. Therefore, total circuit area overhead per hub with a WI is 1.52 mm^2.

7. Conclusions and Future Work

In this paper we have shown that by adopting long range, high bandwidth, low power mm-wave wireless links, significant performance improvements can be achieved in a NoC. Small world network architecture along with the proposed communication infrastructure provides scalability and flexibility of design. The proposed architecture shows considerable performance gain in both uniform and non-uniform (application specific traffic) traffic scenarios.

As part of this on-going investigation, we intend to explore the possibility of multi-channel mm-wave links between far apart cores. We also intend to establish a detailed performance benchmark for the proposed mWNoC with respect to other emerging NoC architectures, like 3D and photonic NoCs and NoCs with THz wireless links.

8. Acknowledgement

This work was supported in part by the US National Sci-ence Foundation (NSF) CAREER grant (CCF-0845504).

9. References

[1] P. P. Pande, et al., "Performance Evaluation and Design Trade-offs for Network-on-chip Interconnect Architectures", IEEE Transactions on Computers, Vol. 54, No. 8, August 2005, pp. 1025-1040.

[2] D. J. Watts and S. H. Strogatz. "Collective dynamics of 'small-world' networks." Nature 393, 1998, pp. 440–442.

[3] Y. P. Zhang et al., "Propagation Mechanisms of Radio Waves Over Intra-Chip Channels with Integrated Antennas: Frequency-Domain Measurements and Time-Domain Analysis", IEEE Transactions on Antennas and Propagation, Vol. 55, No. 10, October 2007, pp. 2900-2906.

[4] A. Kumar et al., "Toward Ideal On-Chip Communication Using Express Virtual Channels", IEEE Micro, Vol. 28, Issue 1, January-February 2008, pp. 80-90

[5] T. Krishna et al., "NoC with Near-Ideal Express Virtual Channels Using Global-Line Communication", Proceedings of IEEE Symposium on High Performance Interconnects, HOTI, 26-28 August, 2008, pp. 11-20.

[6] U. Y. Ogras and R. Marculescu, ""It's a Small World After All": NoC Performance Optimization Via Long-Range Link Insertion", IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol. 14, No. 7, July 2006, pp. 693-706.

[7] A. Shacham et al., "Photonic Network-on-Chip for Future Generations of Chip Multi-Processors", IEEE Transactions on Computers, Vol. 57, no. 9, 2008, pp. 1246-1260.

[8] D. Vantrease et al., "Corona: System Implications of Emerging Nanophotonic Technology", Proceedings of IEEE International Symposium on Computer Architecture (ISCA), 21-25 June, 2008, pp. 153-164.

[9] M. F. Chang et al., "CMP Network-on-Chip Overlaid With Multi-Band RF-Interconnect", Proceedings of IEEE International Symposium on High-Performance Computer Architecture (HPCA), 16-20 February, 2008, pp. 191-202.

[10] 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 Transactions on Computers, vol. 57, no. 9, September 2008, pp. 1230-1245.

[11] J. Lin et al., "Communication Using Antennas Fabricated in Silicon Integrated Circuits", IEEE Journal of solid-state circuits, vol. 42, no. 8, August 2007, pp. 1678-1687.

[12] S. B. Lee et al., "A Scalable Micro Wireless Interconnect Structure for CMPs", Proceedings of ACM Annual International Conference on Mobile Computing and Networking (MobiCom), September, 2009, pp. 20-25.

[13] S. Kirkpatrick et al., "Optimization by Simulated Annealing". Science. New Series 220 (45978): 671-680.

[14] K. Kim and K. K. O, "Characteristics of Integrated Dipole Antennas on Bulk, SOI, and SOS Substrates for Wireless Communication," Proceedings of the International Interconnect Conference, pp 21-23, San Francisco, CA, June 1998.

[15] E. Seok and K. K. O, "Design Rules for Improving Predictability of On-Chip Antenna Characteristics in the Presence of Other Metal Structures", Proceedings of IEEE International Interconnect Technology Conference, 6-8 June 2005, pp. 120-122.

[16] J. Branch et al., "Wireless Communication in a Flip-Chip Package using Integrated Antennas on Silicon Substrates," IEEE Electron Device Letters, vol. 26, no. 2, Feb. 2005, pp 115-117

[17] J. Mehta, and K. K. O, "Switching Noise of Integrated Circuits (IC's) Picked up by a Planar Dipole Antenna Mounted Near the IC's," IEEE Transactions on Electro-Magnetic Compatibility, vol. 44, no. 5, May 2002, pp. 282-290.

[18] J. Lee et al., "A low-power fully integrated 60GHz transceiver system with OOK modulation and on-board Antenna assembly," Proceedings of IEEE Solid-State Circuits Conference, ISSCC 2009, pp.316-317,317a.

[19] William Stallings, Data and Computer Communications, Prentice Hall 2007.

[20] Circuits Multi-Projects. http://cmp.imag.fr

[21] Agilent EDA Design & Simulation Software: http://agilent.com