Design of an Efficient NoC Architecture using Millimeter-Wave Wireless Links

Sujay Deb[1], Kevin Chang[1], Amlan Ganguly[2], Xinmin Yu[1], Christof Teuscher[3], Partha Pande[1], Deukhyoun Heo[1], Benjamin Belzer[1]
[1]Washington State University, Pullman, WA USA
[2]Rochester Institute of Technology, Rochester, NY USA
[3]Portland State University, Portland, OR USA
[1]E-mail: {sdeb, jchang, xyu, pande, dheo, belzer}@eecs.wsu.edu
[2]E-mail: amlan.ganguly@rit.edu
[3]E-mail: teuscher@pdx.edu

Abstract

The Network-on-Chip (NoC) is an enabling technology to integrate large numbers of embedded cores on a single die. Traditional multi-core designs based on the NoC paradigm suffer from high latency and power dissipation due to the inherent multi-hop nature of communication. The performance of NoC fabrics can be significantly enhanced by introducing long-range, low power, and high-bandwidth single-hop links between far apart cores. In this paper we present a design methodology and performance evaluation for a hierarchical small-world NoC with on-chip millimeter (mm)-wave wireless channels as long-range communication links. The proposed wireless NoC offers significantly better performance in terms of achievable bandwidth and energy dissipation compared to its conventional multi-hop wired counterpart in both uniform and non-uniform traffic scenarios. The performance improvement is achieved through efficient data routing, an optimum placement of the wireless hubs, and an energy-efficient transceiver design.

Keywords

Networks on Chip, Wireless, Small-world

1. Introduction

The continuing progress and integration levels in silicon technologies indicate that there will be a manifold increase in the number of cores on a single die over the next few years. Interconnection fabrics for these multi-core *Systems-on-Chip* (SoCs) play a crucial role to sustain the expected growth of computing performance. NoCs have emerged as the communication platform to enable a high degree of integration in multi-core SoCs. However, traditional NoCs suffer from the limitations arising from a multi-hop communication over conventional planar metal interconnects, resulting in high latency and power consumption. With a further increase in the number of cores, this problem will be significantly aggravated. This challenge can be addressed by drawing inspiration from the small-world property possessed by many natural complex networks. Such networks have a low average shortest path length and a high clustering cofficient. In small-world networks with distance-dependent link distributions, nodes in close proximity are connected via direct links with fewer long range shortcuts connecting largely separated nodes. Performance of NoCs has been shown to improve by the insertion of long-range wired links following principles of small-world graphs [1]. However, that approach doesn't scale up well because of the multi-hop wired links that are necessary for longer distances. Recent investigations in silicon integrated circuits have uncovered the possibility of implementing on-chip millimeter (mm)-wave wireless interconnects working in the range of 10s to 100 GHz. This opens up the possibility of desiging small-world NoC architectures with high-speed, long-range, energy efficient wireless interconnects as shortcuts. In this paper we apply mm-wave wireless communication channels as long-range links in a hierarchical small-world NoC and analyze various relevant design tradeoffs. The biggest advantage of using mm-wave on-chip wireless links are their CMOS-compatibility. But these wireless links have associated antenna and transceiver area and power overheads. Thus, to achieve the best performance without undue overhead, the wireless resources need to be placed and used optimally. To accomplish that goal, we implemented a hybrid and hierarchical network where nearby cores communicate through traditional metal wires, but long distance communications are predominantly achieved through high performance single-hop wireless links. We show that network performance can be significantly improved by using this hybrid approach and by choosing both an optimal number and an optimal location of the wireless hubs. A comprehensive approach and design flow towards the design of an efficient multi-core chip with long-range wireless links are presented in this paper. We also introduce a cost-function to determine the optimum number of wireless interfaces for different system sizes, which is subsequently verified through network simulation. The optimal placement of the wireless interfaces is done through an *Evolutionary Algorithm* (EA). This optimization also takes into account the target application or traffic for which the chip is being designed. Furthermore, we use a distributed flow-control-based routing algorithm that ensures an optimum utilization of the wireless links. We demonstrate that the proposed mm-wave wireless NoC (mWNoC) outperforms its more traditional non-hierarchical wireline counterparts in terms of bandwidth and energy dissipation in both uniform and non-uniform traffic scenarios.

2. Related Work

NoCs have emerged as suitable communication platforms for future multi-core processors. Conventional NoCs use multi-hop packet switched communication and to improve performance, the concept of express virtual channels is introduced in [2]. It is shown that 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. NoCs have been shown to perform better by inserting long-range wired links following principles of small-world graphs [1].

The design principles of Photonic NoC are elaborated in various recent publications [3][4]. It is estimated that a photonic NoC will dissipate significantly less power than its electronic counterpart.

NoC with multi-band RF interconnects [5] is another promising alternative. In this particular NoC, 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.

The design of a wireless NoC based on *CMOS Ultra Wideband* (UWB) technology was proposed in [6]. In [7], the feasibility of designing on-chip wireless communication network with miniature antennas and simple transceivers that operate at the sub-THz range of 100-500 GHz has been demonstrated. If the transmission frequencies can be increased to THz/optical range then the corresponding antenna sizes decrease, occupying much less chip real estate. One possibility is to use nanoscale antennas based on *Carbon Nanotubes* (CNTs) operating in the THz/optical frequency range [8]. Consequently, building an on-chip wireless interconnection network using THz frequencies for inter-core communications becomes feasible. The design of a small-world wireless NoC operating in the THz frequency range using CNT antennas is elaborated in [9]. Though this particular NoC is shown to improve the performance of traditional wireline NoC by orders of magnitude, the integration and reliability of CNT devices need more investigation. A preliminary design of a NoC with CMOS-compatible mm-wave wireless links was proposed in [10].

This work introduces a comprehensive and detailed design methodology of a hierarchical and small-world mWNoC. It proposes to improve the mWNoC performance by incorporating an efficient topology, an EA-based optimization for wireless interface placement, flow-control-based distributed routing, and the use of more energy-efficient body-enabled wireless transceivers.

3. Architectural Overview of mWNoC

We consider a hierarchical NoC architecture with strategic placement of a limited number of wireless interfaces (WIs) for optimum performance. Our goal is to build a highly efficient small-world NoC based on both wired and wireless links. We first divide the whole system into multiple small clusters of neighboring cores and call these smaller networks *subnets*. These subnets have NoC switches and links as in a standard NoC. The cores are connected to a centrally located hub through direct links and the hubs from all subnets are connected in a 2nd level network forming a hierarchical structure. The hubs connected through wireless links require WIs. Due to a limited number of WIs, as discussed in later subsections, neighboring hubs are connected by traditional wired links and a few of the WIs are distributed between hubs separated by relatively long physical distances. As will be described in section 3.2, we use an evolutionary algorithm to optimally place the WIs. The key to our approach is establishing an optimal overall network topology under given resource constraints, i.e., with a limited number of WIs. Figure 1 shows a representative hierarchical 256-core network where the subnets are connected in a StarRing (ring with a central hub) topology and the network has 16 hubs and 6 WIs. The hubs are connected in a mesh architecture with overlaid long-range wireless shortcuts on the 2nd level of the hierarchy. Instead of the Mesh-StarRing configuration used in this example, any other possible interconnection topology can be considered.

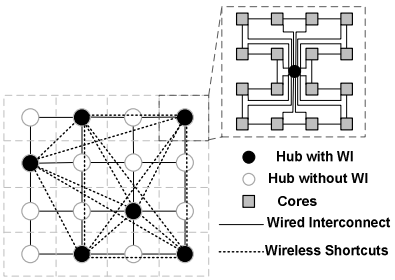


Figure 1. A Hierarchical 256 core network where hubs are connected by a small-world graph

3.1. Optimizing the Number of WIs

In order to determine the optimal number of WIs for a given network, we first define a metric that allows approximating a network's performance. Let n be the number of hubs of the network. Let d be an $n \times n$ matrix where $d_{i,j}$ is the distance (shortest path) between hub i and hub j measured in hops. The matrix d is populated using Dijkstra's shortest path algorithm [11]. The *average shortest path* (*ASP*) is then calculated by summing the length of all paths and dividing by the number of paths.

$$ASP = \sum d_{i,j} / (n^2 - n) \ . \quad (1)$$

In addition, to account for non-uniform traffic patterns, the average shortest path metric is then extended so that each path length is weighted by the probability of that path being traveled by a packet. The new metric is denoted as *ASPf*

$$ASPf = \sum d_{i,j} * f_{i,j} / [(n^2 - n) * F] \ . \quad (2)$$

Here, the frequency $f_{i,j}$ of communication between the i^{th} source and j^{th} destination is the apriori probability 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

$$F = \sum f_{i,j} \ . \quad (3)$$

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

$$Cost(\#\ of\ WI) = A + P + L \ , \quad (4)$$

where, A, P and L are normalized area, power and latency overheads respectively arising from the WIs. A is determined by dividing the total WI area by the chip area. The power dissipated by all WIs is divided by the total power consumed by the communication infrastructure to

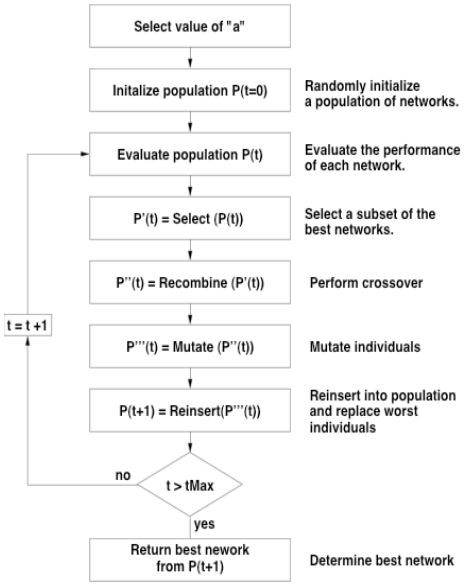


Figure 2. Flowchart illustrating the evolutionary algorithm we used for the network optimization.

determine *P*. *L* is obtained by dividing the message latency incurred in the wireless channel by the average message latency of the whole network. The two metrics, average shortest path and cost, are thus the two objectives to be optimized. Many methods exist for evaluating multi-objective optimization problems [12]. We describe the *aggregate objective function* (*AOF*), which combines both of the metrics, as following:

$$AOF = a * ASPf + (1-a) * Cost, \qquad (5)$$

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 shortest path (i.e., the network performance), while a = 0.5 makes for a balance between the two metrics. The choice of a is a designer's decision and depends on the design requirements.

3.2. Placement of WIs

An *Evolutionary Algorithm* (EA) [13] is a nature-inspired stochastic search algorithm that works with a population of candidate solutions. EAs are very powerful for complex problems with large and unknown search spaces. We use a standard evolutionary algorithm that is depicted in the flowchart of Figure 2. Each network is encoded in the form of genome, which represents a single solution in this population-based optimization process. Starting from a random initial solution of such solutions, genetic operators, such as crossover and mutation, are applied to the encoded solution (i.e., the genome) with the goal to find the optimum solution. The crossover (or recombination) operator combines two parent individuals and results in new individual, which will then be subjected to mutation. The mutation operator randomly changes an individual in order to introduce variation in the population and to explore the design space. For determining the optimal WI placement, we have thus only encoded the WIs location on the genome. Each gene is represented by an integer value, which specifies the ID of the location within the grid. We have implemented the regular genetic operators, namely single-point crossover and simple mutation. For our purpose, we have employed a standard canonical evolutionary algorithm with tournament selection [13]. ParadisEO [14], an object-oriented framework for evolutionary computation is used as a platform to implement the EA, which then interfaces with our C++ *Complex Network Framework* (*CNF*).

4. Communication Scheme

In this section we describe the physical layer design and adopted data routing strategy for the mWNoC.

4.1. Physical Layer

The WIs are responsible for establishing the wireless communication channel between the hubs. The two principal components of the WI are the antenna and the transceiver, whose characteristics are outlined below.

4.1.1 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 [15] 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. The zig-zag antenna is designed 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 5.2. The details of the antenna structure are shown in Figure 3.

4.1.2 Wireless Transceiver

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 with a design adopted from [16]. The detailed description of the transceiver circuit is beyond the scope of this paper. Non-coherent on-off keying (OOK) is chosen as the modulation method, as it allows relatively simpler and low-power circuit implementation. As illustrated in Figure 3, 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 circuit level, body-enabled design techniques [17], including both forward body-bias (FBB) with DC voltages, as well as body-driven by AC signals

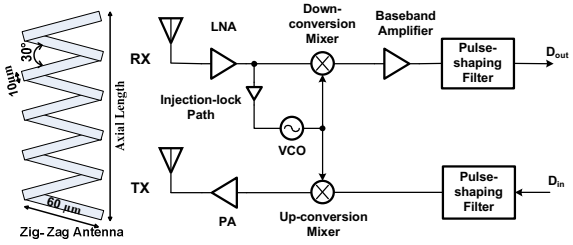


Figure 3. Zig-zag antenna and OOK transceiver block diagram.

[18], are implemented to further decrease power consumptions.

4.2 Adopted Routing

In our proposed hierarchical NoC, data is transferred via flit-based wormhole routing. For intra-subnet data routing in the StarRing subnet, 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 is more than two hops away, then the data goes through the central hub. To avoid deadlock, we adopt the virtual channel management scheme from the Red Rover algorithm [19], in which the ring is divided into two equal sets of contiguous nodes. Messages originating from each group of nodes use dedicated virtual channels. This scheme breaks cyclic dependencies and prevents deadlock.

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 between them. 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 then 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 the wireless medium is chosen if it reduces the total path length compared to the wired path. This can potentially give rise to a hotspot situation in all the WIs because many messages try to access wireless shortcuts simultaneously, thus overloading the WIs and resulting in higher latency. A token flow control [20] along with a distributed routing strategy is adopted to alleviate this problem. Tokens are used to communicate the status of the input buffers of a particular WI to the other nearby hubs, which need to use that WI for accessing wireless shortcuts. Every input port of a WI has a token and the token is turned on if the availability of the buffer at that particular port 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 WIs. This routing algorithm is proved to be deadlock free in [1]. Figure 4 shows a particular communication snapshot of a mesh-based upper level network where hub 8 wants to communicate with hub 3. First at source 8, the nearest WI (4 in this case) is identified. Then the routing algorithm checks whether taking this WI reduces the total hop count. If so, the token for the south input port of hub 4 is checked and this path is taken only if the token is available. If this is not the case, the message at hub 8 follows dimension order routing towards the destination and arrives at hub 9. At hub 9, again the shortest path using WIs is searched and if the token from hub 10 allows the usage of wireless shortcuts, then the message is routed through hub 10. Otherwise, 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 along with

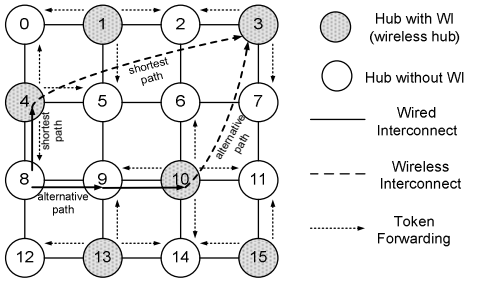


Figure 4. Token flow control based distributed routing.

token flow control prevents deadlocks and effectively improves performance by distributing traffic though alternative paths. It is also livelock free since it generates a minimal path towards the destination, as the adopted routing here ensures that the wireless shortcuts are only followed if that reduces the hop count between source and destination. As a result, this routing always tries to find the shortest path and never allows routing away from the destination.

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 wireless token passing protocol. 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 wireless 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, the flit is accepted for further routing either to a core in the subnet of that hub or to an adjacent hub. The wireless token is released to the next hub with a WI after all flits belonging to a packet at the current wireless token-holding hub are transmitted.

5. Performance Evaluation

In this section we characterize the performance of the proposed mWNoC through rigorous simulation and analysis in presence of various traffic patterns. First, we present the simulation setup followed by how the optimum hierarchical division and number of WIs is selected for different system sizes and characteristics of the on-chip wireless communication channel. Then we present the detailed network level simulations with various system sizes and traffic patterns.

5.1 Simulation Setup

An overview of the performance evaluation setup for a mWNoC is shown in Figure 5. To obtain the gain and bandwidth of the antennas we use the ADS momentum tool [21]. Bandwidth and gain of the antennas are necessary for establishing the required design specifications for the transceivers. The mm-wave wideband wireless transceiver is designed and simulated using Cadence tools with TSMC 65-nm standard CMOS process to obtain its power and delay characteristics. The subnet switches and the digital

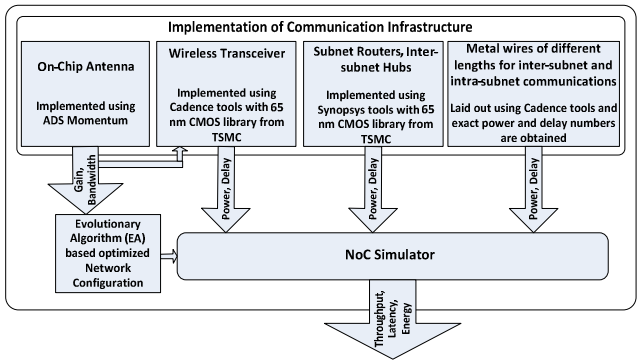


Figure 5. Performance evaluation setup for mWNoC.

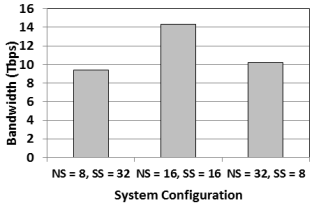
components of the hubs are synthesized using Synopsys tools with 65-nm standard cell library from TSMC at a clock frequency of 2.5 GHz. Energy dissipation of all the wired links are obtained from Cadence layout assuming a 20mm x 20mm die area. All the power and delay numbers of various components along with the optimum network configuration generated from the EA are then fed into the network simulator to obtain overall mWNoC performance.

For our experiments, we consider three different system sizes, namely 128, 256, and 512 cores, and the die area is kept fixed at 20mm x 20mm for all system sizes. The NoC switch architecture is adopted from [22]. 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 is efficiently handled without compromising performance. 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.

The network architectures developed are 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 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 the hubs are taken into account as well.

5.2 Optimum Hierarchical Division

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



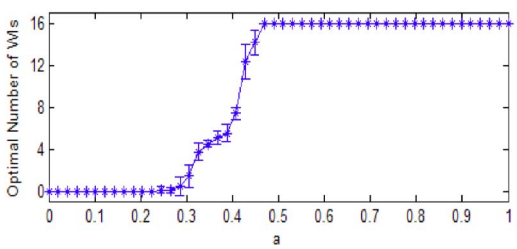
** NS = number of subnets, SS = subnet size

Figure 6. Achievable bandwidth of a 256-core Mesh-StarRing NoC for various hierarchical configurations.

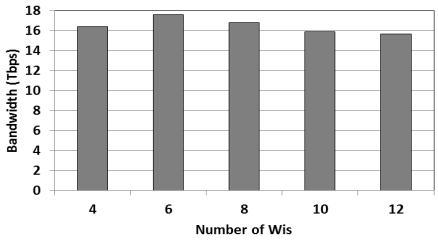
we evaluate the performance by dividing the whole system in various alternative ways. This analysis is performed without any shortcuts to highlights the effect on performance resulting from different ways of doing the hierarchical division. Figure 6 shows the achievable bandwidth for a 256-core Mesh-StarRing 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 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 WI

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 performed a network quality analysis and the result obtained for a 16 hub system is shown in Figure 7 (a). The weightage parameter *a* determines how the cost versus the performance is weighted for the optimization fitness function. This was explained in Section 3.1. From this result it can be observed that for a moderate weightage value, *a* (varying from 0.35 to 0.45), the optimum number of WIs varies from 4 to 12. As expected at the weightage boundary values, the cost function optimization ends with either no 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 verify the results obtained from the network quality analysis and exactly determine the optimum number of WIs, we carried out system-level simulations with the wireless token passing mechanism and the results are shown in Figure 7 (b). The token is considered to be a single flit transmitted from the WI, which currently holds it to the next one. From Figure 7 (b), it can be seen that for a 256-core mWNoC (16-subnets with 16 cores in each subnet)



(a)



(b)

Figure 7. Results obtained from (a) cost function analysis and (b) network simulation.

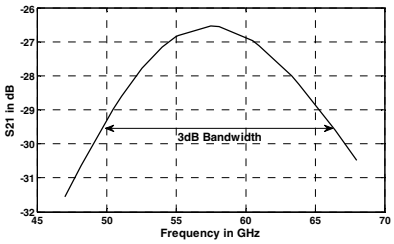


Figure 8. Antenna S21 response.

bandwidth increases with the number of WIs until reaching a maximum at 6 WIs, after which it decreases. This is because although a higher number of WIs improves connectivity by reducing the hop-count of the network, the shared wireless medium is distributed among the WIs, and as the number of WIs increase beyond a certain point, performance degrades due to the high token returning period. Similarly, for system sizes of 128 and 512 (consisting of 8 and 32 subnets respectively), the maximum bandwidth is achieved with 4 and 10 WIs respectively.

5.4 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. High resistivity silicon substrate (ρ=5kΩ-cm) is used for the simulation. To represent the worst case inter-subnet communication range considering the placement of the WIs, the transmitter and receiver are separated by 20 mm. The forward transmission gain (S21) of the antenna obtained from the simulation is shown in Figure 8. As can be seen, 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. The wireless transceiver circuitry is designed and simulated using TSMC 65-nm CMOS process. The transceiver can sustain a data rate of 16 Gbps with a power consumption of 36.7 mW with OOK modulation [16].

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. Figure 9 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 three alternative architectures of the same size: (i) a flat mesh (ii) the same hierarchical architecture as the mWNoC, but without any long-range links, and (iii) the same hierarchical architecture as the mWNoC, but with shortcuts implemented using buffered metal wires instead of the wireless links. The numbers of wired shortcuts are kept equal to the number of WIs for different system sizes and they are optimally placed using the same EA-based optimization as used for the placement of WIs (Section 3.2). Each wired shortcut is considered to be 32-bit wide. The wires are designed with an optimum number of uniformly placed and sized repeaters. It can be observed that the mWNoC outperforms all the other alternatives for the three

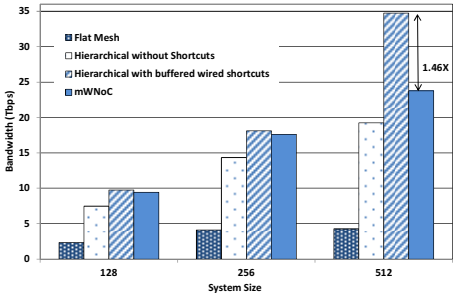


Figure 9. Achievable bandwidth with scaling for different architectures.

system sizes under consideration, with the exception of the system with buffered wire shortcuts. The flat mesh architecture performs the worst due to its high average hop count. The hierarchical architecture improves the performance by reducing hop count, but the best performance is obtained from the hierarchical architecture with shortcuts due to the small-world nature of the network. The hierarchical NoC with buffered wires as shortcuts results in a higher bandwidth as multiple parallel wires can operate together. But, it suffers from significant energy dissipation overhead, which is quantified in section 5.6.

5.6 Energy Dissipation

To quantify the energy dissipation characteristics of the proposed mWNoC architecture, we determine the packet energy dissipation. The packet energy is the energy dissipated on average by a packet from its injection at the source to delivery at the destination.

Figure 10 shows the packet energy dissipation for all the architectures under uniform random traffic. It can be observed that the energy dissipation of the hierarchical wired NoCs with or without wireline shortcuts is significantly less compared to the flat mesh architecture. This is because a hierarchical network reduces the average hop count, and hence the latency between the cores. Packets get routed faster and hence occupy resources for less time and dissipate less energy in the process. From Figure 10 it is also evident that the mWNoC significantly outperforms the other two possible wired hierarchical architectures. It can be ovserved that a 512-core hierarchical NoC with buffered wire shorcuts burns 12.79 times more energy where as achieves only 1.46 times more bandwidth compared to mWNoC. Therefore, it can be concluded that mWNoC achieves best performance-energy tradeoff among all the NoC architectures compared in this paper. Overall, the mWNoC is capable of reducing the packet energy dissipation at least an order of magnitude compared to the

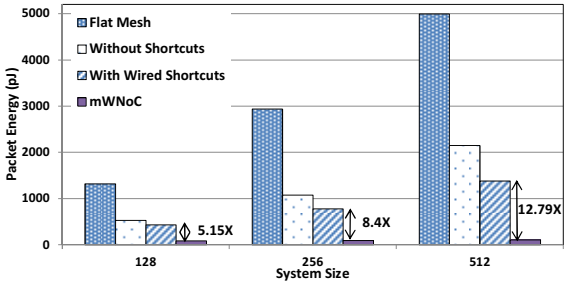


Figure 10. Packet Energy for different architectures.

flat wireline architecture.

The new Mesh-StarRing architecture along with the routing mechanism elaborated in section 4.2 results in 14.6% bandwidth improvement and 48% saving in packet energy for a 256 core system with 6 WIs in comparison with respect to the previously proposed NoC architecture with mm-wave wireless links [10] for same system size.

5.7 Comparative Study with RF-I

The on-chip RF transmission line (RF-I) proposed in [5] is another possible revolutionary interconnect technology that can improve the performance of the NoC. Like the wireless channel, these RF links can be used as long-range shortcuts in the hierarchical NoC architecture. For this comparative analysis, we consider a 128-core system as an example. The mWNoC has 8 subnets and each subnet has 16 cores. We design a small-world NoC (RFNoC) using RF-I links as shortcuts, maintaining the same hierarchical topology. As mentioned in [5], in 65nm technology it is possible to have 8 different frequency channels, each operating with a data rate of 6 Gbps and used for long-range communications. These shortcuts are optimally placed using the same EA-based optimization as used for placing the WIs in the mWNoC.

In order to evaluate the performance of the mWNoC and RFNoC architectures, we considered both uniform and non-uniform traffic patterns. For non-uniform traffic patterns both synthetic and application-based traffics are used.

We considered two types of synthetic traffic to evaluate the performance of the proposed mWNoC architecture. First, a transpose traffic pattern [1] 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 [1], 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 example of an application-based traffic, a 256-point *Fast Fourier Transform* (FFT) is considered and each core is assigned to perform a 2-point radix 2 FFT computation.

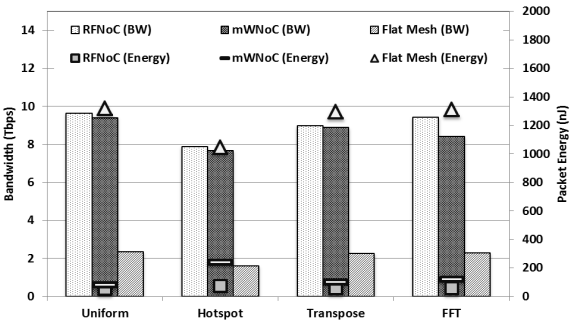


Figure 11. System performance variation with different traffic scenarios.

Figure 11 shows the achievable overall network bandwidth and packet energy for the different NoC architectures in uniform and non-uniform traffic scenarios. The energy dissipation for RFI is obtained from [5]. Compared to the flat mesh architecture both mWNoC and RFNoC provides orders of magnitude better performance. It can be observed that though mWNoC and RFNoC have the same hierarchical architecture, the latter performs better because with RFNoC, multiple shortcuts can work simultaneously whereas in mWNoC (where the wireless channel is a shared medium), only one pair can communicate at a particular instant of time. However, the total long-range link area overhead and the layout challenges of the RFNoC are more significant compared to mWNoC. For example, for a 20mm x 20mm die, an RF interconnect of approximately 100 mm length has to be allocated for RFNoC following the layout of [5]. 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.

6. Area Overheads

In this section, we first quantify the area overhead due to the wireless deployment in mWNoCs. The antenna used is a 0.38 mm long and 58 μm wide zig-zag antenna. The area of the transceiver circuits required per WI is 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.

Figure 12 shows the total wiring requirements of various lengths for a 256 core system of different NoC configurations considered in this work for a 20mm x 20mm die. For comparison, the wiring requirements for flat mesh architecture are shown as well. We observe that the two hierarchical small-world architectures, viz., mWNoC and RFNoC require extra wiring overhead compared to a flat mesh. However, we have seen that those extra wires significantly improve the performance compared to the conventional NoCs.

7. Conclusions and Future Work

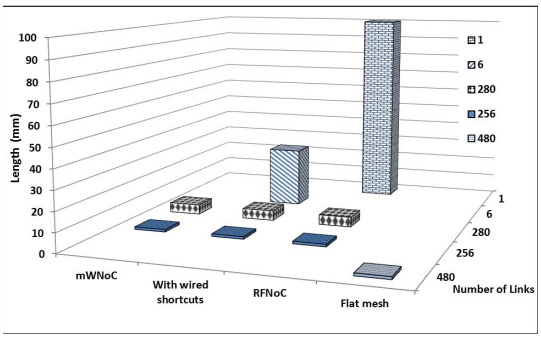


Figure 12. Total wiring requirements of various lengths for a 20mm x 20mm die for a 256-core system.

In this paper we have shown that by an optimal utilization of long-range, high bandwidth, low power mm-wave wireless channels, significant performance improvements can be achieved in a NoC. By incorporating a small-world topology, the proposed mm-wave NoC architecture shows considerable performance gain in the presence of various traffic scenarios.

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

8. 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).

9. References

[1] 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, 2006, pp. 693-706.

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

[3] 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.

[4] A. Joshi, et al., "Silicon-Photonic Clos Network for Global On-Chip Communication", Proceedings of the 3rd International Symposium on Networks-on-Chip (NOCS-3), 2009, pp. 124-133.

[5] 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), 2008, pp. 191-202.

[6] 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, 2008, pp. 1230-1245.

[7] 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), 2009, pp. 20-25.

[8] K. Kempa, et al., "Carbon Nanotubes as Optical Antennae", Advanced Materials, vol. 19, 2007, pp. 421-426.

[9] A. Ganguly, et al., "Scalable Hybrid Wireless Network-on-Chip Architectures for Multi-Core Systems", IEEE Transactions on Computers (TC), vol. 60, issue 10, 2010, pp. 1485-1502.

[10] S. Deb, et al., "Enhancing Performance of Network-on-Chip Architectures with Millimeter-Wave Wireless Interconnects", Proceedings of IEEE International Conference on ASAP, 2010, pp. 73-80.

[11] E. W. Dijkstra, "A note on two problems in connexion with graphs", Numerische Mathematik 1, 1959, pp. 269-271.

[12] S. Deb, Multi-Objective Optimization using Evolutionary Algorithms, Wiley, Chichester, UK, 2001.

[13] T. Bäck and H.P. Schwefel, "An Overview of Evolutionary Algorithms for Parameter Optimization". Evolutionary Computation, vol. 1, no 1, 1993, pp.1-23.

[14] S. Cahon, et al. "ParadisEO: A Framework for the Reusable Design of Parallel and Distributed Metaheuristics", Journal of Heuristics, vol. 10, no 3, 2004, pp. 357-380.

[15] B. A. Floyd, et al., "Intra-Chip Wireless Interconnect for Clock Distribution Implemented With Integrated Antennas, Receivers, and Transmitters", IEEE Journal of Solid-State Circuits, vol. 37, no. 5, 2002, pp. 543-552.

[16] X. Yu, et al., "A Wideband Body-Enabled Millimeter-Wave Transceiver for Wireless Network-on-Chip", Proceedings of the 54th IEEE Midwest Symposium on Circuits and Systems 2011, pp. 1-4.

[17] M. J. Deen and O. Marinov "Effect of forward and reverse substrate biasing on low-frequency noise in silicon PMOSFETs", IEEE Transactions on Electron Devices, 49, 3, 2002, pp. 409-413.

[18] G. Kathiresan and C. Toumazou, "A low voltage bulk driven down-conversion mixer core". In Proceeding of the IEEE International Symposium on Circuit and Systems, 2, 1999, pp. 598-601.

[19] J. Draper and F. Petrini, "Routing in Bidirectional k-ary n-cube switch the Red Rover Algorithm". In Proceedings of the International conference on Parallel and Distributed Processing Techniques and Applications, 1997, pp .1184-93.

[20] A. Kumar, et al. "Token flow control", Proceedings of the 41st IEEE/ACM International Symposium on Microarchitecture, MICRO-41, 2008, pp. 342-353.

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

[22] 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.