MILLIMETER-WAVE WIRELESS NETWORK-ON-CHIP: A CMOS COMPATIBLE INTERCONNECTION INFRASTRUCTURE FOR FUTURE MANY-CORE PROCESSORS

By

SUJAY DEB

A dissertation submitted in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

WASHINGTON STATE UNIVERSITY
School of Electrical Engineering and Computer Science

MAY 2012

To the Faculty of Washington State University:

The members of the committee appointed to examine the dissertation of SUJAY DEB find it satisfactory and recommend that it be accepted.

_____

Partha Pratim Pande, Ph.D., Chair

_____

Deukhyoun Heo, Ph.D.

_____

Benjamin Belzer, Ph.D.

ACKNOWLEDGEMENT

I would like to take this opportunity to express my gratefulness to my advisor Dr. Partha Pratim Pande for having guided me through the curriculum so well. His active involvement in my research and incessant inspiration has made this work possible. I also thank him for having allowed me freedom of thought and choice of research direction.

Special thanks go to Dr. Deukhyoun Heo for having helped me with his expertise in analog and RF circuits providing a strong buttress to my work. I would like to thank Dr. Benjamin Belzer for helpful insights about communication and coding theory essential to my work. I thank Dr. Amlan Ganguly for his inputs regarding future many-core interconnect architectures. Dr. Christof Teuscher deserves a special note of thanks for his help in the domain of complex networks. My work was partially supported by the US National Science Foundation (NSF) CAREER grant CCF-0845504 and NSF grant CCF-0635390.

I would also like to thank my colleagues Mr. Kevin Chang, Dr. Souradip Sarkar and Mr. Turbo Majumder for their frequent help and brainstorming which always helped me to strengthen the foundations of my conceptual understanding of the problems.

My parents, Mr. Sadhan Chandra Deb and Mrs. Rita Deb have always been extremely inspiring. Through their experience and caring they have made it possible for me to pursue research at a school of higher learning. Without their support none of this work would have been possible. Last but most importantly I thank my elder brother Mr. Suman Deb for his encouragement and unflinching faith in me that made my research experience even more rewarding.

MILLIMETER-WAVE WIRELESS NETWORK-ON-CHIP: A CMOS COMPATIBLE INTERCONNECTION INFRASTRUCTURE FOR FUTURE MANY-CORE PROCESSORS

Abstract

by Sujay Deb, Ph.D.
Washington State University
May 2012

Chair: Partha Pratim Pande

Multi-core platforms are emerging trends in the design of Systems-on-Chip (SoCs). Interconnect fabrics for these multi-core SoCs play a crucial role in achieving the target performance. The Network-on-Chip (NoC) paradigm has been proposed as a promising solution for designing the interconnect fabric of multi-core SoCs. But the performance requirements of NoC infrastructures in future technology nodes cannot be met by relying only on material innovation with traditional scaling. The continuing demand for low power and high speed interconnects with technology scaling necessitates looking beyond the conventional planar metal/dielectric-based interconnect infrastructures. Among different possible alternatives, the on-chip wireless communication network is envisioned as a revolutionary methodology, capable of bringing significant performance gains for multi-core SoCs. Millimeter-wave Wireless NoCs (mWNoCs) can be designed by using miniaturized on-chip antennas as an enabling technology. On-chip CMOS compatible millimeter-wave wireless

links provide high bandwidth and low power communication channels over long distances. Hence they can be used to create short cuts between distant cores on the chip to provide fast and efficient traffic freeways. Such long-range wireless links facilitate design of NoC architectures inspired by the natural complex networks like *Small-World* graphs. Such topologies inherently have low average inter-core distances and scale very well with increase in size. In this work, design methodologies and technology requirements for scalable mWNoC architectures are presented and their performance is evaluated. It is demonstrated that mWNoCs outperform their wired counterparts in terms of network throughput and latency, and that energy dissipation improves by orders of magnitude under various experimental and real-life scenarios.

TABLE OF CONTENTS

List of Tables

List of Figures

Dedication

This dissertation is dedicated to my parents and Dada

for whom this was possible

Chapter 1

INTRODUCTION

1.1 Multi-core System-on-Chip Design Challenges

Constraints of power, heat and reliability have forced the computer industry to shift focus from single processor core to incorporating multiple processor cores on a single chip. Current commercial Systems-on-Chips (SoCs) designs integrate an increasingly large number of pre-designed cores and their number is predicted to increase significantly in the near future. For example, molecular-scale computing promises single or even multiple orders-of-magnitude improvements in device densities. These state-of-the-art commercial SoC designs with a large number of intellectual property (IP) blocks, commonly known as cores, on a single die [1] [2] are made feasible by the continuing progress and integration levels in silicon technologies. This number of cores on a single die, which is currently between ten and hundred [3] depending on the application, is likely to go up manifold in the near future. This massive level of integration makes modern multi-core chips all pervasive in domains ranging from weather forecasting, astronomical data analysis and biological applications, to consumer electronics and smart phones. The growing complexity of integration as well as aggressive technology scaling introduces multiple challenges for the design of such big multi-core SoCs. An important feature of such Multi-Processor SoCs (MP-SoC) is the interconnection fabric, which must allow seamless integration of numerous cores performing various functionalities.

One of the major problems associated with future SoC designs arises from non-scalable global wire delays [4]. Global wires carry signals across a chip, but these wires typically do not scale in length with technology scaling. Though gate delays scale down with technology, global wire delays typically increase quadratically or, at best, linearly by inserting repeaters. Even after

repeater insertion, the delay may exceed the limit of one clock cycle or even multiple clock cycles. As a result the design efforts necessary towards meeting the timing requirements for computation and communication is widening. In ultra-deep submicron processes, eighty percent or more of the delay of critical paths is due to interconnects [5]. With supply voltage scaling down as ever and global wires becoming thinner the delay in transmission of signals over these wires will seriously affect the overall achievable performance of the system. Long wires with lengths of the order of the dimensions of the die can have delays well over multiple clock cycles. Even with architectural and design innovations interconnects still remain a critical bottleneck and it highlights the challenges for future chip designers associated with traditional scaling of conventional interconnects and material innovation. That is why a great need arises to explore radical alternative technologies for sustainable improvements in power dissipation and performance on future generation of multi-core chips.

1.2 The Network-on-Chip Paradigm

The network on chip (NoC) paradigm is an enabling solution to this problem of many core integration and has captured the attention of the academia and the industry [6]. Consequently, MP-SoCs have embarked a paradigm shift from computation-centric to communication-centric system design as the number of cores in a chip increases. NoC introduces a higher level of communication abstractions and it has emerged as the most preferred communication platform that enables partitioning of the design effort into minimally interdependent sub modules. The common characteristic of these NoC architectures is that the processor/storage cores communicate with each other through switches and links as shown in figure 1.1. Communication between constituent cores in a NoC takes place through packet switching. Generally wormhole switching is adopted for NoC's, which breaks down a packet into fixed length flow control units

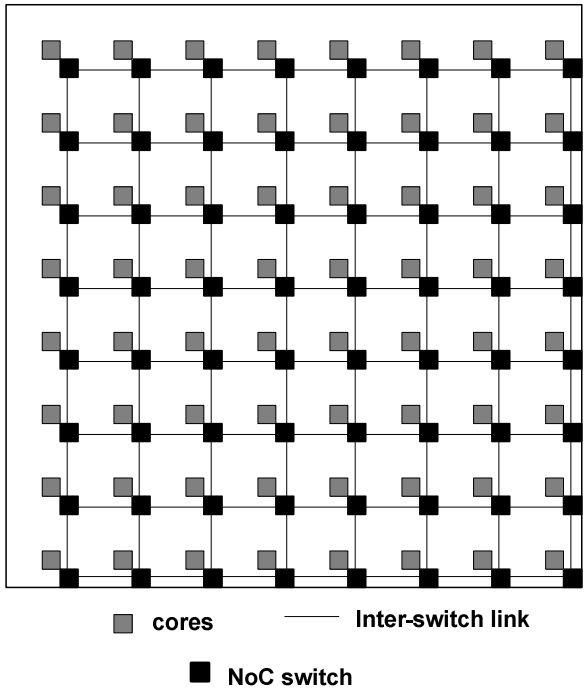


Figure 1.1. A regular tile based Mesh NoC

or *flits.* The first flit or the *header* contains routing information that helps to establish a path from the source to destination, which is subsequently followed by all the other *payload* flits. Advances in NoC research along several dimensions spanning architectural explorations, development of routing protocols, and reliability in communication have made it the choice for the communication backbone of complex and Multi-Processor Systems-on-Chip (MP-SoCs).

1.3 Limitations of Conventional NoCs

Despite their several advantages, an important performance limitation in traditional NoCs arises from planar metal interconnect-based multi-hop communications, wherein the data transfer between two distant blocks causes high latency and power consumption. To alleviate this problem, insertion of long-range links in a standard mesh NoC using conventional metal wires has been proposed [7]. Another effort to improve the performance of multi-hop NoC was undertaken by introducing ultra-low-latency and low power express channels between largely

separated communicating nodes [8][9]. But these express channels are also basically metal wires, though they are significantly more power and delay efficient compared to their conventional counterparts. According to the International Technology Roadmap for Semiconductors (ITRS) [5] for the longer term, improvements in metal wire characteristics will no longer satisfy performance requirements and new interconnect paradigms are needed. Different approaches have been explored already, such as 3D and photonic NoCs and NoC architectures with multi-band RF interconnect [10][11][12]. Though all these emerging methodologies are capable of improving the power and latency characteristics of the traditional NoC, they need further and more extensive investigation to determine their suitability for replacing and/or augmenting existing metal/dielectric-based planar multi-hop NoC architectures. Consequently, it is important to explore further alternative strategies to address the limitations of planar metal interconnect-based multi-hop NoCs.

1.4 A CMOS Compatible mm-Wave Wireless NoC

In this work, we propose an innovative and novel approach, which addresses simultaneously the latency, power consumption and interconnect routing problems: replacing multi-hop wired paths in a NoC by high-bandwidth single-hop long-range CMOS compatible wireless links.

The limitations of conventional NoCs can be addressed by drawing inspiration from the interconnection mechanism of natural complex networks. Modern complex network theory [13]provides us with a powerful method to analyze network topologies and their properties. Between a regular, locally interconnected mesh network and a completely random Erdõs-Rényi topology, there are other classes of graphs [13], such as small-world and scale-free graphs. 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. Also, such networks have

a high clustering parameter which is an index of the connectivity of the topology. The average shortest 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 [14][15]. Most complex networks, such as social networks, the Internet, as well as certain parts of the brain exhibit the small-world property. It has been shown that such "shortcuts" in NoCs can significantly improve the performance compared to locally interconnected mesh-like networks [7][15] with fewer resources than a fully connected system. This feature of small-world graphs makes them particularly interesting for efficient communication in modern multi-core chips with increasing levels of integration. This is because by using on-chip transceivers it is possible to establish long range, low power wireless links across the chip to create shortcuts which enable the small-world based topologies.

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. Two principal components of the wireless interface (WI) are the antenna and the transceiver. 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 [16]. The on-chip antenna has to provide the best power gain for the smallest area overhead. A metal zigzag antenna has been demonstrated to possess these characteristics and suits this application. Coupled with significant advances in mm-wave transceiver design this opens up new opportunities for detailed investigations of mm-wave wireless NoCs (mWNoCs). The design principles of mWNoC architectures using mm-wave wireless antennas are presented in this work.

The performance benefits of these mWNoCs due to the utilization of high-speed wireless links in a small-world based topology are evaluated through cycle accurate simulations. On-chip wireless links enable one-hop data transfers between distant nodes and hence reduce the hop counts in inter-core communication. In addition to reducing interconnect delay, eliminating multi-hop long distance wired communication reduces energy dissipation as well. In future the number of cores in a SoC is expected to increase manifold. Consequently it is imperative to have a scalable communication infrastructure without affecting system performance significantly. This work proposes a mWNoC architecture and evaluates its performance with respect to conventional wired NoCs. It is demonstrated that by utilizing the wireless medium efficiently, it is possible to minimize the effects of scaling up the system size on the performance of the mWNoCs. It is possible to create various configurations for the mWNoC depending on the number of available wireless interfaces (WIs) and their placement in the network. The various mWNoC architectures considered in this work are shown to dissipate significantly less energy and to achieve notable improvements in throughput and latency compared to traditional wired NoCs. The inherent broadcasting capability of mWNoC is also exploited to demonstrate its performance advantage. We demonstrate that mm-wave wireless interconnect based NoCs can be a viable CMOS compatible solution for future many core chips, which are capable of solving the performance limitations of traditional multi-hop wire line counterparts.

1.5 Multi-channel mWNoC

The proposed mWNoC shares a single wireless channel between all the WIs and hence the performance gain is limited. With technology scaling design of antennas and transceivers operating in different non-overlapping frequency bands in the range of hundred GHz is feasible and it is a very promising alternative for long-range, one-hop intra-chip communications.

According to the ITRS, the cut-off frequency and unity maximum available power gain frequency targets in 16 nm CMOS technology are 600 GHz and 1 THz respectively. With such scaling the required antenna and circuit areas for on-chip wireless interfaces will scale down. This allows easy on-chip integration. By varying the axial length, trace width, arm element length and bend angle the antenna operating frequency and bandwidth can be varied. Using such antenna elements and associated transceivers multiple simultaneously operating long distance wireless links can be deployed on-chip to create shortcuts between distant source and destination pairs. Multiple wireless shortcuts operating simultaneously provide an energy efficient solution for design of many-core communication infrastructures. This approach will support the current device-scaling strategy and will fit in the existing process of making more and more powerful chips without diverting/needing new technological innovation/compromise to incorporate interconnection. In this work we show that mWNoC framework can accommodate multiple simultaneously operating wireless channels resulting in significant improvement of overall performance. We also evaluate the performance of mWNoC with respect to two other small-world NoC architectures with emerging interconnects. In one of these architectures the long-range links are implemented with recently proposed RF interconnects (RFNoC) [17]. The other architecture is a hierarchical and small-world wireless NoC designed with carbon nano tube (CNT) enabled THz wireless links (THzNoC) [18]. We present simulation results to evaluate the performance of mWNoC with other emerging interconnect based NoCs, viz., THzNoC and RFNoC in both uniform and non-uniform traffic scenarios. We demonstrate the advantages and the limitations of each architecture and establish the relevant design trade-offs. The area overheads associated with these novel NoC architectures are also quantified and it is shown that performance benefits clearly outweigh the overheads. We also present various challenges and

emerging solutions regarding the design of efficient wireless NoC architectures.

1.6 Contributions

The principal contribution of this thesis can be summarized as below:

- **Architecture space exploration to enhance the performance of NoCs with wireless links**
  - Design of hybrid mm-wave wireless NoC (mWNoC) with hierarchical Small-World topologies with wireless shortcuts.
  - Design of efficient communication and routing protocols for mWNoC.
  - Optimized deployment of wireless transceivers with respect to varying traffic patterns.
  - Analysis and minimization of associated overheads for wireless link deployment.
- **Comparative analysis of multi-channel mWNoC with emerging interconnect technologies**
  - A comparative study of achievable performance advantages of NoC architectures with two emerging interconnects namely, CNT antenna based THz wireless interconnects and RF-Interconnect based NoCs.
  - Comparison of alternatives and establishment of benchmarks with various parameters like system size and traffic patterns.

1.7 Thesis Organization

The thesis is organized in five chapters. The first chapter introduces the complexity of the problem and the possible means of addressing those issues. Literature survey is presented in the second chapter. The third chapter presents the main design methodologies and performance of the proposed hybrid wireless NoC architectures. In this chapter it is demonstrated that the mWNoCs outperform the wireline counterparts in network performance. The fourth chapter

presents how multiple non-overlapping wireless channels can further make mWNoC more competitive with respect to other emerging and technologically more challenging alternatives. We also present a comparative analysis of mWNoC with other available alternative wireless NoCs and highlight the promises and challenges associated with each of them. Finally the last chapter summarizes the important conclusions and points to the direction of future research.

Chapter 2

Related Work

2.1 Background

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. The limitations and design challenges associated with existing NoC architectures are elaborated in [19]. This paper highlights interactions among various open research problems of the NoC paradigm. The concept of express virtual channels is introduced in [8] to improve performance of conventional NoCs. 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]. NoCs have been shown to perform better by insertion of long range wired links following principles of small-world graphs [7]. Despite significant performance gains, the above schemes still require laying out long wires across the chip and hence performance improvements beyond a certain limit cannot be achieved [20].

The performance improvements due to NoC architectural advantages will be significantly enhanced if 3D integration is adopted as the basic fabrication methodology. The amalgamation of two emerging paradigms, namely NoCs in a 3D IC environment, allows for the creation of new structures that enable significant performance enhancements over traditional solutions [10], [21][22]. Despite these benefits, 3D architectures pose new technology challenges such as thinning of the wafers, inter-device layer alignment, bonding, and interlayer contact patterning [23]. Additionally, the heat dissipation in 3D structures is a serious concern due to increased

power density [23][24] on a smaller footprint. There have been some efforts to achieve near speed-of-light communications through on-chip wires [25][26]. Though these techniques achieve very low delay in data exchange along long wires, they suffer from significant power and area overheads from the signal conditioning circuitry. Moreover the speed of communication is actually about a factor of one-half the speed of light in silicon dioxide. By contrast, on-chip data links at the true velocity of light can be designed using recent advances in silicon photonics [27][28]. The design principles of photonic NoCs are elaborated in various recent publications [27][28][29][30]. 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. Although the optical interconnect option has many advantages, some aspects of this new paradigm need more extensive investigation. The speed of light in the transmitting medium, losses in the optical waveguides, and the signal noise due to coupling between waveguides are other important issues that need more careful investigation. Moreover, Photonic NoCs demonstrated in [28]and [29] still require an underlying electrical network to establish the path through the photonic links due to lack of optical storage elements. However, in [31] a completely photonic CLOS network is shown to achieve significant performance benefits over the wireline counterparts. In [30] CORONA, an amalgamation of 3D architecture and photonic NoC, is presented and demonstrated to deliver high network bandwidths for various real-application based traffic models. Another alternative is NoCs with multi-band RF interconnects [17]. Various implementation issues of this approach are discussed in [12]. 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 [32]. The antennas used in [32] achieve a 1 mm transmission range, and are 2.98 mm in length. For 1 mm range, wireless links are less efficient than metal wires [20]. In [33] 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. 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 [34]. 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 [35]. Depending on antenna configuration and substrate characteristics, achievable wireless channel frequencies can be in the range of 50-100 GHz. 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 [36]. 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 [37] 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 [36]. Using antennas with a differential or balanced feed structure can significantly reduce coupling of switching noise, which is mostly common-mode in nature [38]. In [39], 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 [40] 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 [18]. Though this particular NoC is shown to improve the performance of traditional wireline NoC 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 hierarchical small-world network with CMOS compatible mm-wave wireless links for multi-core chips.

Chapter 3

A CMOS Compatible Millimeter-Wave Wireless NoC (mWNoC)

A generic wired NoC provides interconnection among embedded cores via switches and wired links. Communication between a pair 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 mm-wave wireless interfaces strategically placed for optimum performance. In this chapter we discuss the topology of the proposed hierarchical architecture and the adopted performance optimization methodology and demonstrate that by efficient utilization of the wireless medium, the proposed mWNoC outperforms the corresponding conventional wireline counterpart in terms of bandwidth and energy dissipation.

3.1 Design Methodologies

In the following subsections, the design methodologies essential for design of a mWNoC are discussed.

3.1.1. Interconnection Topology

Modern complex network theory provides powerful methods to analyze network topologies and their properties [41]. 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 [14]. 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 re-wiring selected node connections randomly to any other node, which creates short cuts in the network [15]. These random long-range links between nodes can also be established following probability distributions depending on the distance separating the nodes [42]. It has been shown that such *shortcuts* in NoCs can significantly improve the performance compared to locally interconnected mesh-like networks [44][15] 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 traditional 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 introducing a few long range mm-wave wireless links between distant hubs according to the placement scheme outlined in section 3.1.2. 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 Simulated Annealing (SA) [43] based optimization algorithm. The SA approach allows network design to be scalable with an increase in 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. 3.1 shows a representative interconnection topology of a hierarchical 256-core network with 16 hubs and 6 wireless interfaces. In Fig. 3.1 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 StarRing architectures, which consist of a ring with a central hub. The hubs are interconnected via both wireless and wired links while the subnets are wired only. The hubs with wireless links are equipped with

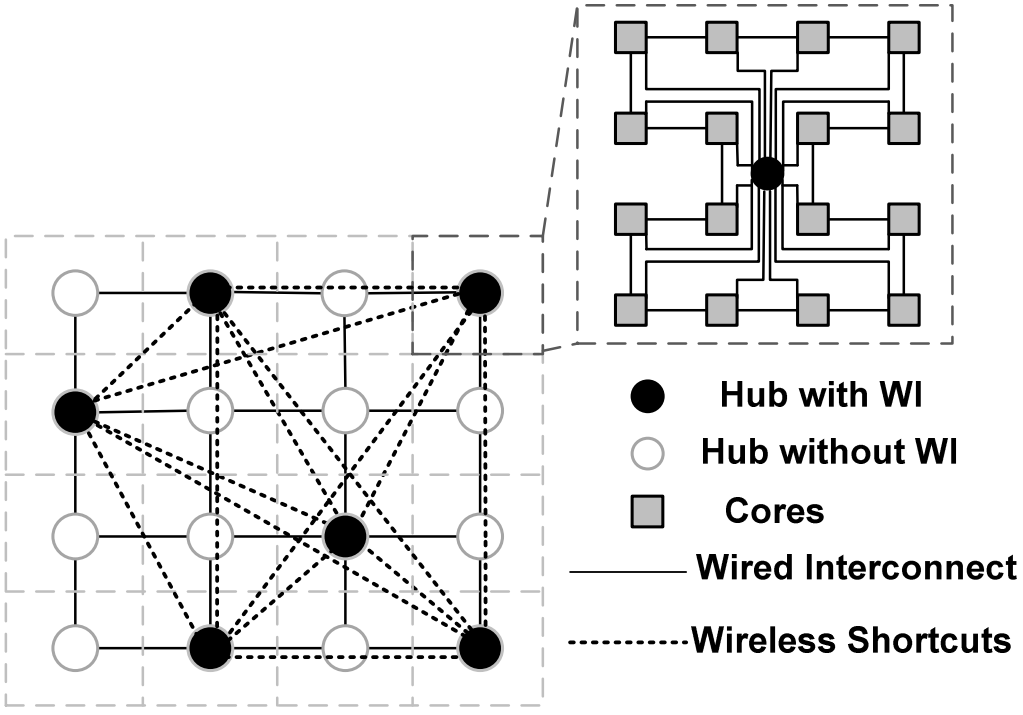


Figure 3.1. A hierarchical 256-core network where hubs are connected by a small-world graph.

wireless interfaces (WIs) that transmit and receive data packets over the wireless channels. For inter-subnet data exchange, a packet first travels to its respective hub and reaches the hub of the destination subnet via the small-world network, where it is then routed to the final destination core.

There can be various subnet architectures, like mesh, star, ring, etc. Similarly, the basic architecture of the 2nd level of the hierarchy may vary. As an example the hubs may be connected in a mesh architecture with a few long-range wireless links spread across them creating a small-world network in the 2nd level of the hierarchy. As case studies, in this work we consider two types of subnet architectures, viz. mesh and star-ring (a ring architecture with a central hub connecting to every core). Corresponding to each subnet architectures, we consider two upper level small-world configurations, mesh and ring, with long-range wireless shortcuts distributed among the hubs. Thus, the following four hierarchical mm-wave NoC architectures are considered: Ring-StarRing, Ring-Mesh, Mesh-StarRing, and Mesh-Mesh. As an example, in the Ring-StarRing architecture, the first term (Ring) denotes the upper level architecture and the second term (StarRing) indicates that the subnet is a star-ring topology. The same nomenclature applies to the rest of the hierarchical architectures in this paper. 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. However, since the architecture of the two levels can be different causing their traffic characteristics to differ from each other, the exact hierarchical division can be obtained by performing system level simulations as discussed in subsection 3.3.2.

3.1.2 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 about 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 have bearing on the performance as well as the cost of NoC. The first metric that indicates the approximate network performance is the *average shortest path, μ* between all pairs of hubs. 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. 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 [44]. The distances are then weighted with the normalized frequencies of communication between hub pairs. The metric, μ can be calculated as

$$\mu = \sum h_{i,j} * f_{i,j} / [(N^2 - N) * F], \tag{3.1}$$

$$h_{i,j} = p * d_{i,j\_with\_shortcut} + (1-p) * d_{i,j\_without\_shortcut} \tag{3.2}$$

where h_{ij} 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

$$F = \sum f_{i,j} . \tag{3.3}$$

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

$$p = 1/n \,. \tag{3.4}$$

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

$$Cost(\#of\,WI) = A + P \,, \tag{3.5}$$

where, A and P are normalized area and power 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 determine P. The two metrics, *average shortest path* μ and *cost* are thus the two objectives to be optimized. Many methods exist for evaluating multi-objective optimization problems [45]. We describe the *aggregate objective function* (*AOF*), which combines both of the metrics, as follows:

$$AOF = a * \mu + (1 - a) * Cost \,, \tag{3.6}$$

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 outline 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 [18] 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

$$|S| = \binom{N}{n} . \tag{3.7}$$

Thus, with increasing N, it becomes increasingly difficult 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

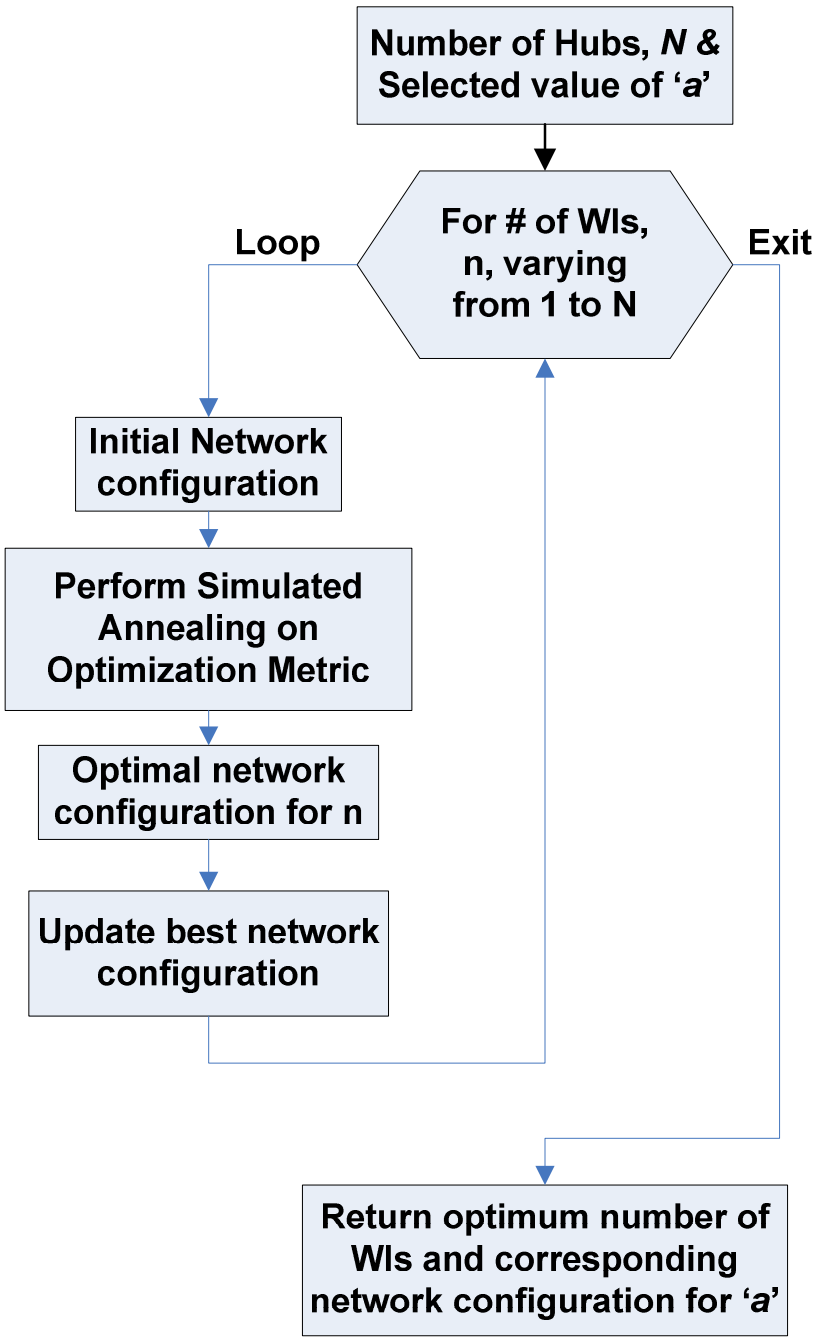


Figure 3.2. Flow diagram for the simulated annealing based optimization of mWNoC architectures.

worse solution. In this work we have used *Cauchy annealing schedule*, where the *temperature* profile varies inversely with the number of iterations [43]. 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. Fig. 3.2 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) [46] and co-evolutionary algorithms [47]. Although EAs are generally believed to give better results, SA reaches comparably good solutions much faster [48]. We have used SA in this work as an example.

3.2 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.

3.2.1 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 [49] has been demonstrated to possess these

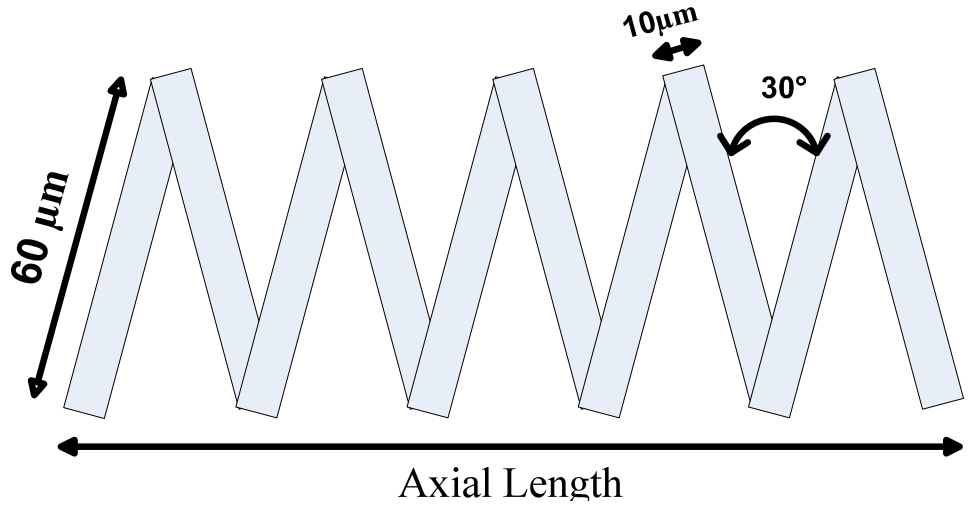


Figure 3.3. Zigzag antenna structure details.

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 [35]. The zigzag antenna is designed with 10μm 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 3.3.4. The details of the antenna structure are shown in Fig. 3.3.

3.2.2 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. 3.4, the

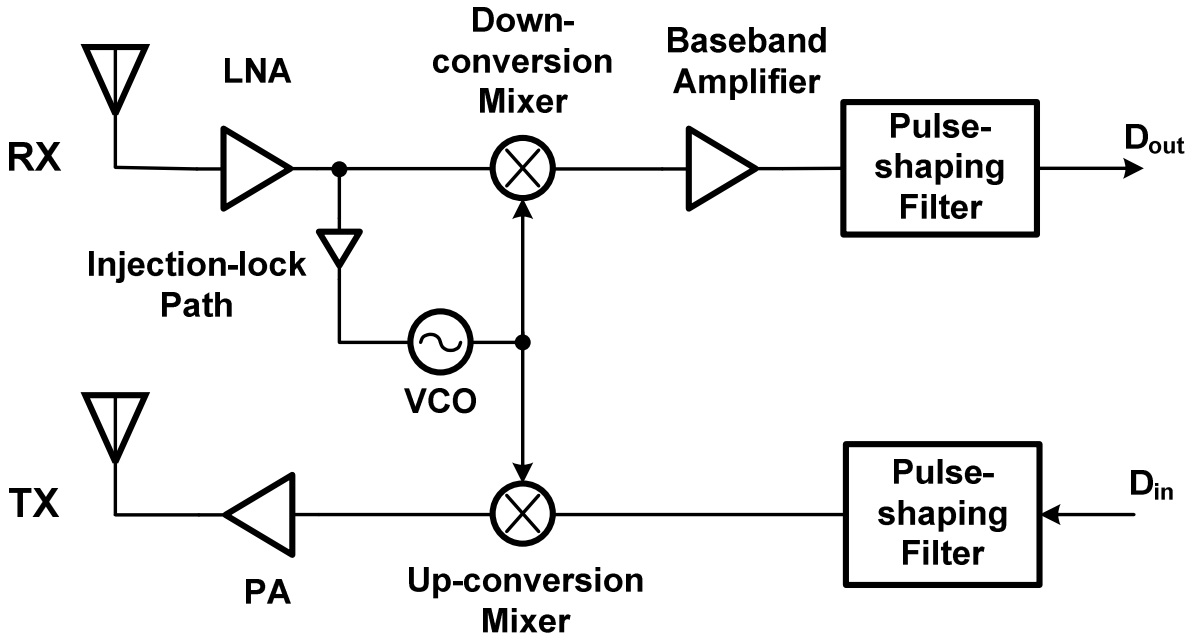


Figure 3.4. OOK transceiver block diagram.

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 [50], including both forward body-bias (FBB) with DC voltages, as well as body-driven by AC signals [51], are implemented to further decrease power consumption. Detailed design descriptions of the transceiver are presented in [20][52].

3.2.3 Adopted Data Transmission Strategy

In the proposed hierarchical NoC, data is transferred via flit-based wormhole routing [53]. Intra-subnet data routing is done according to the topology of the subnets. In this work, we consider two subnet topologies (i.e., mesh and star-ring). In a mesh subnet, the data routing follows a deadlock-free dimension order (e-cube) routing. 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 [54], 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 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

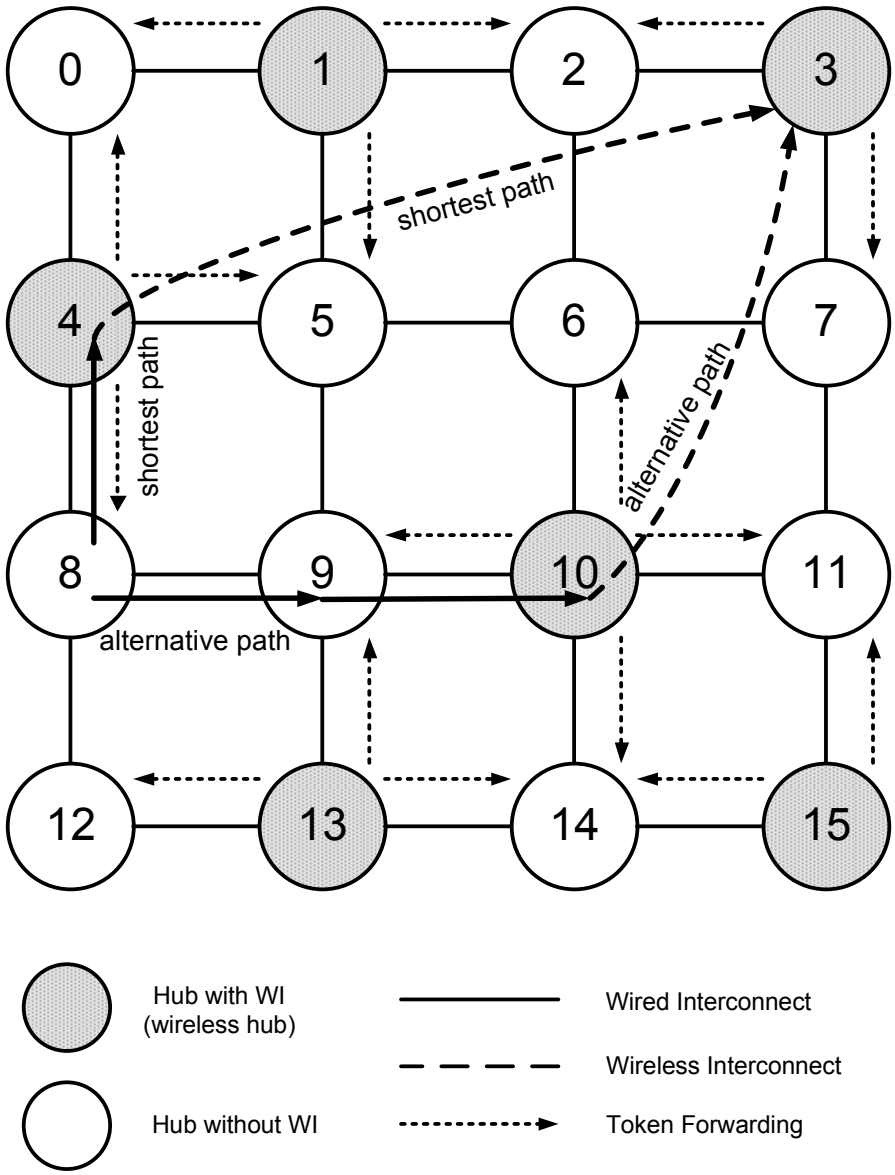


Figure 3.5. An example of token flow control based distributed routing.

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 [55] 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. Figure 3.5 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 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

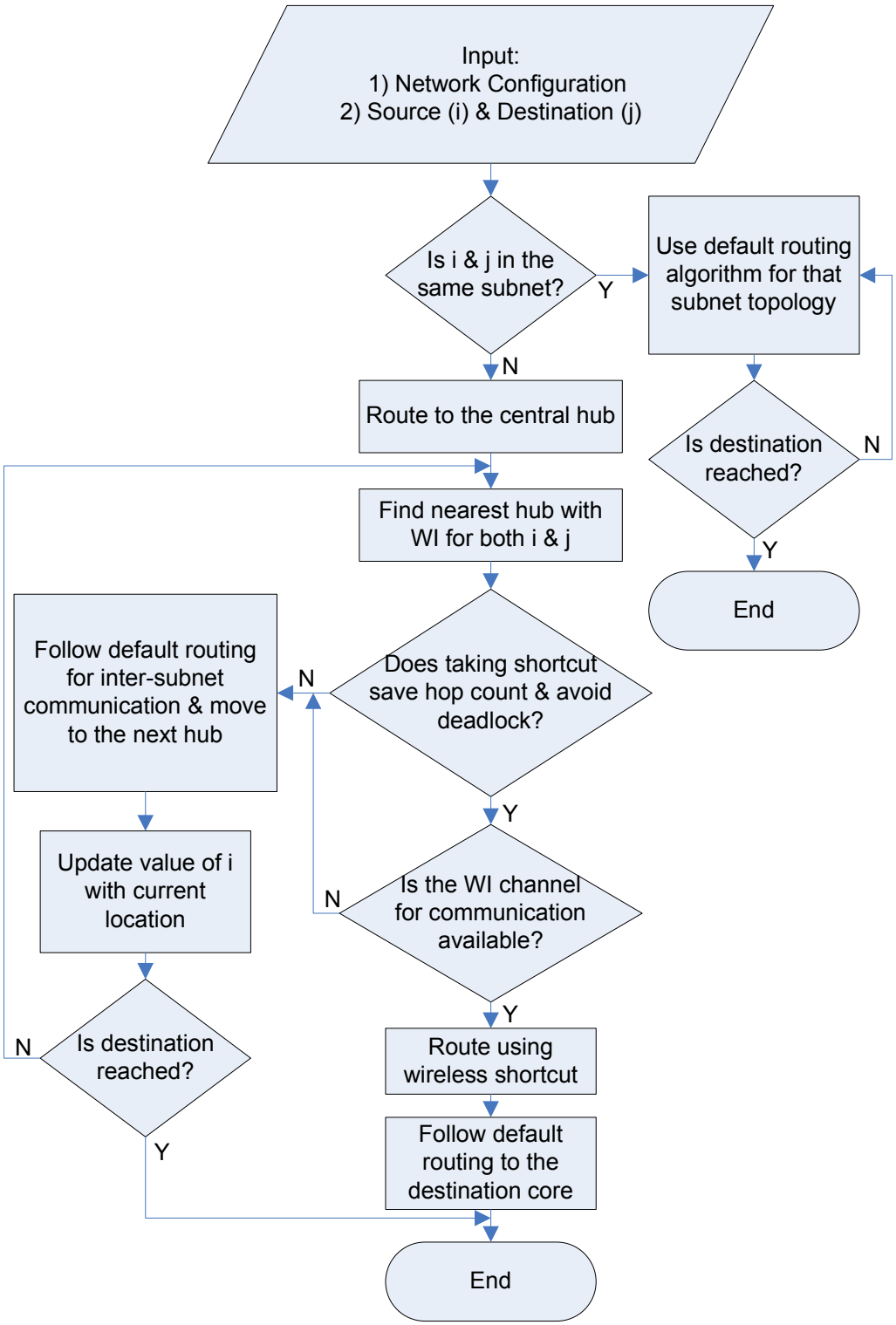


Figure 3.6. An algorithmic representation of the adopted data routing strategy.

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 [56]. 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. 3.6 shows the flow chart of the adopted data routing strategy.

According to [57], the mWNoC is deadlock free if both the subnets and the 2^{nd} level of the network are deadlock free and the boundary nodes are safe nodes. As explained above, the subnets and the 2^{nd} 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.

3.3 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.

3.3.1 Simulation Setup

An overview of the performance evaluation setup for the mWNoC is shown in Fig. 3.7. To obtain the gain and bandwidth of the antennas we use the ADS momentum tool [58]. The mm-wave wideband wireless transceiver is designed and simulated using Cadence tools with TSMC [59] 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 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 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 configuration generated from the SA 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

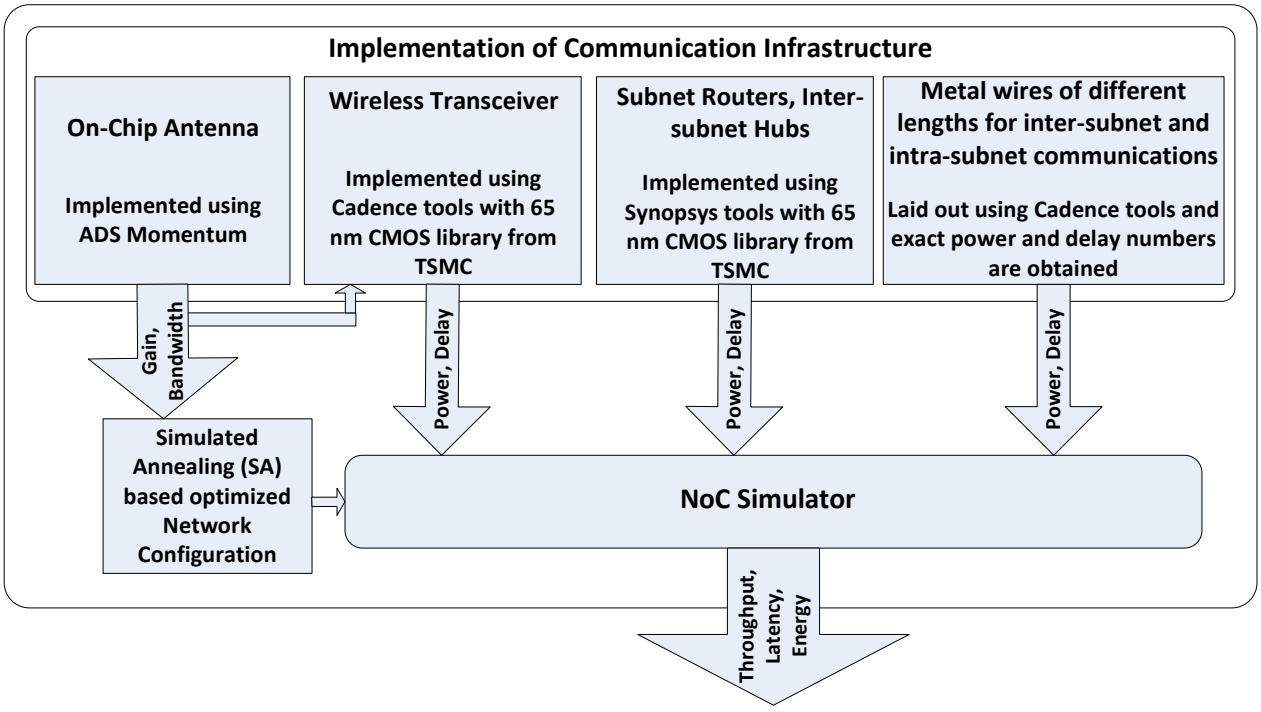


Figure 3.7. Performance evaluation setup for mWNoC.

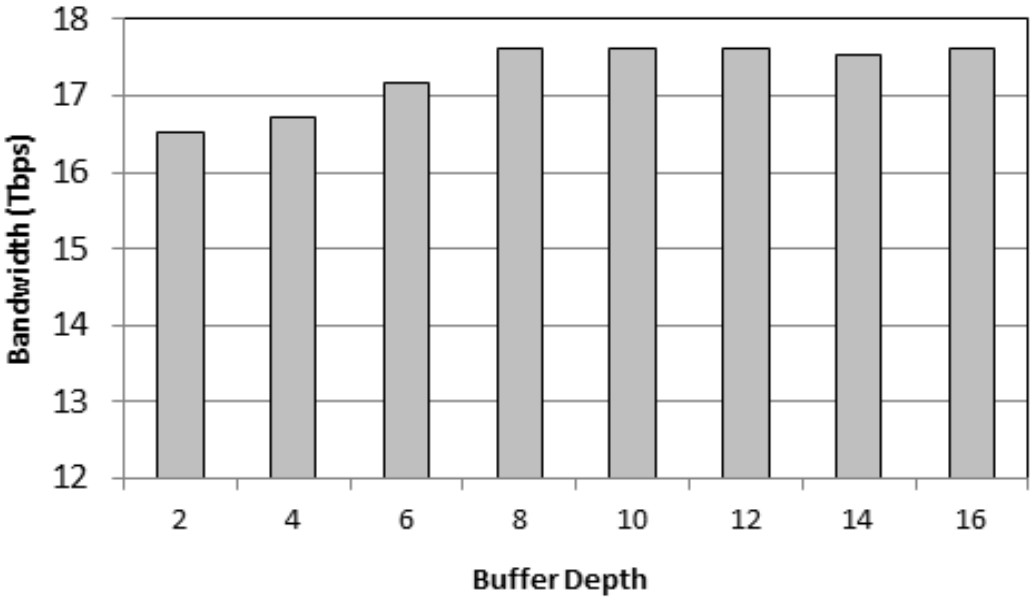
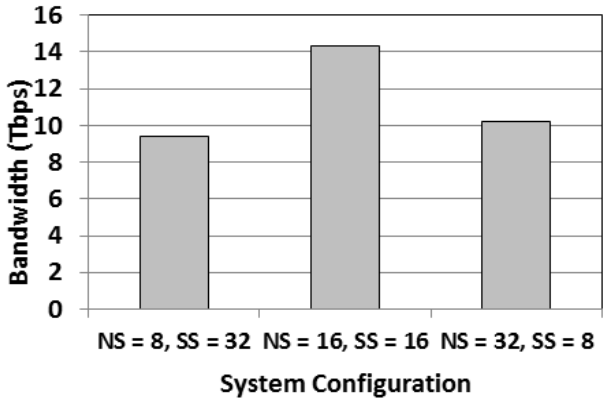


Figure 3.8. Performance variation with change in buffer depth for the ports associated with WIs for a 256-core Mesh-StarRing system.

cores, and the die area is kept fixed at 20 mm x 20 mm for all system sizes. The NoC switch architecture is adopted from [60]. 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. 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. This is shown in Figure 3.8 for a 256-core Mesh-StarRing system divided into 16 subnets. 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. 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



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

Figure 3.9. Achievable bandwidth of a 256-core Mesh-StarRing NoC for various hierarchical configurations.

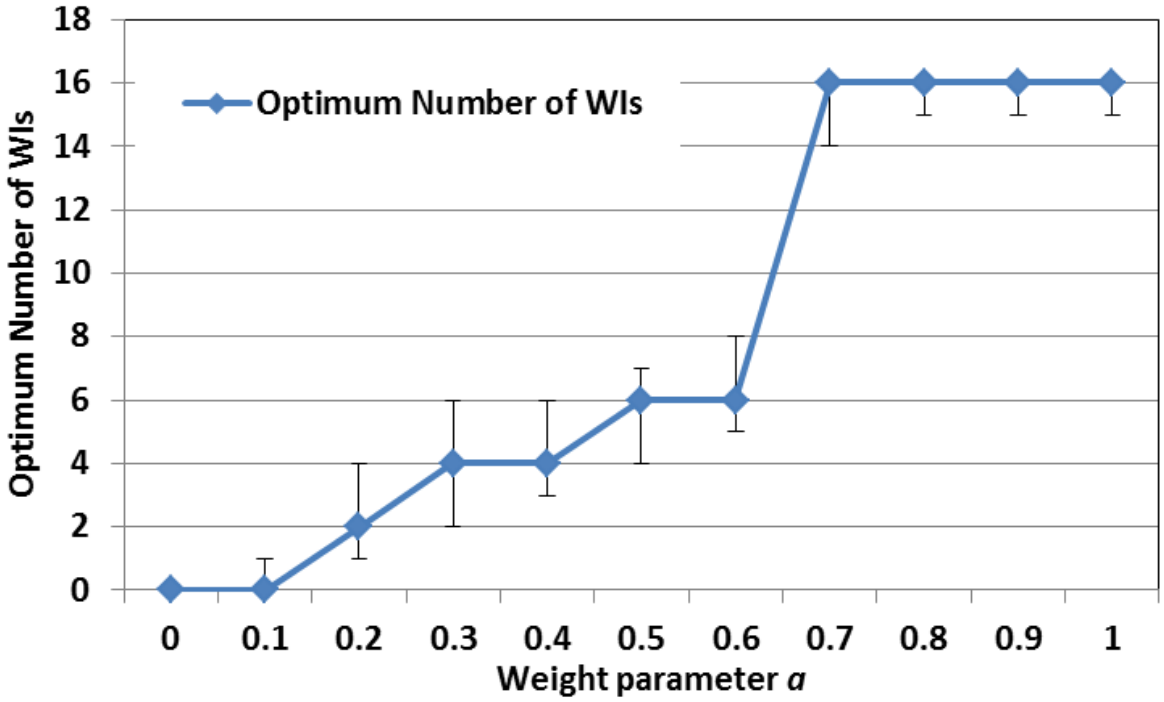
intra-subnet core-to-hub and the inter-hub wired links in the upper level of the network. The delays through the switches and inter-switch wires of the subnets and also the delays through the hubs are taken into account.

3.3.2 Optimum Hierarchical Division

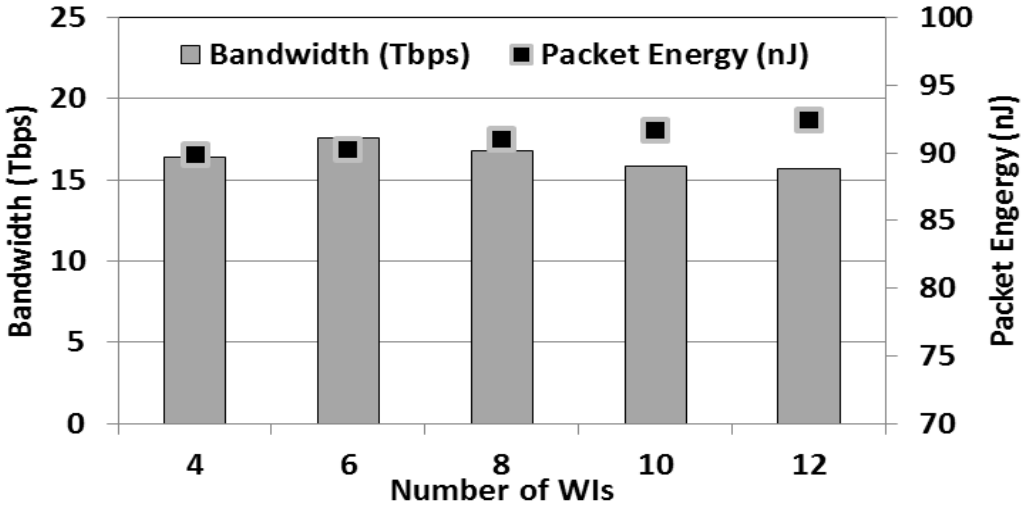
To determine the optimum division of the proposed hierarchical architecture in terms of achievable bandwidth, we evaluate the performance of the mWNoC by dividing the whole system in various alternative ways. This analysis is performed without any shortcuts to highlight the effect on performance resulting from different ways of doing the hierarchical division. Fig. 3.9 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.

3.3.3 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



(a)



(b)

Figure 3.10. Results obtained from (a) Cost function analysis and (b) Network simulation.

wireless medium, only the 2nd level of the network is considered. We performed a network quality analysis and optimum number of WIs (*n*) obtained for different values of *a* for a 16 hub system connected in a mesh with one wireless channel is shown in Fig. 3.10 (a). The weight parameter *a* determines how the cost versus the performance is weighted for the optimization fitness function. This was explained in section 3.1.2. From this result it can be observed that for a moderate weight value, *a* (varying from 0.30 to 0.65) 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 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 Fig. 3.10 (b). The token is considered to be a single flit transmitted from the WI which currently holds it to the next one. From Fig. 3.10 (b), it is seen that for a 256-core Mesh-StarRing 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

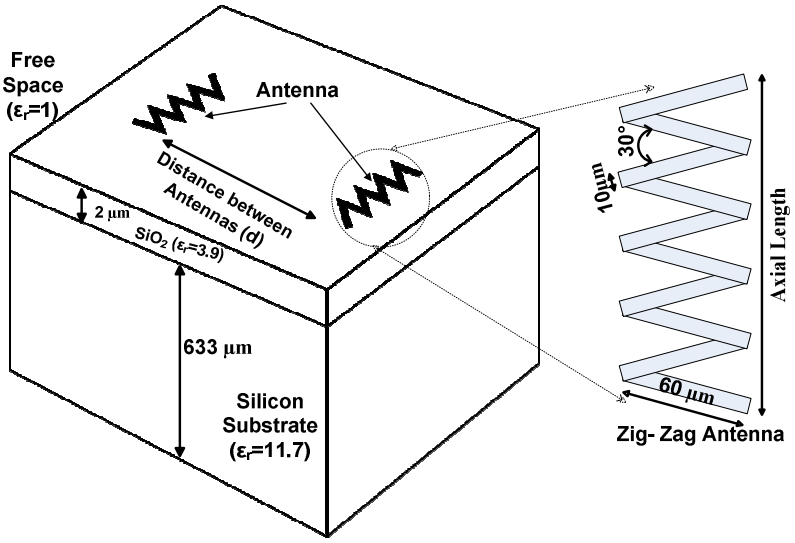


Figure 3.11. Zigzag antenna simulation set-up.

with 4 and 10 WIs respectively.

3.3.4 Wireless Channel Characteristics

The metal zigzag antennas described earlier are used to establish the on-chip wireless communication channels. High resistivity silicon substrate (ρ=5kΩ-cm) is used for the

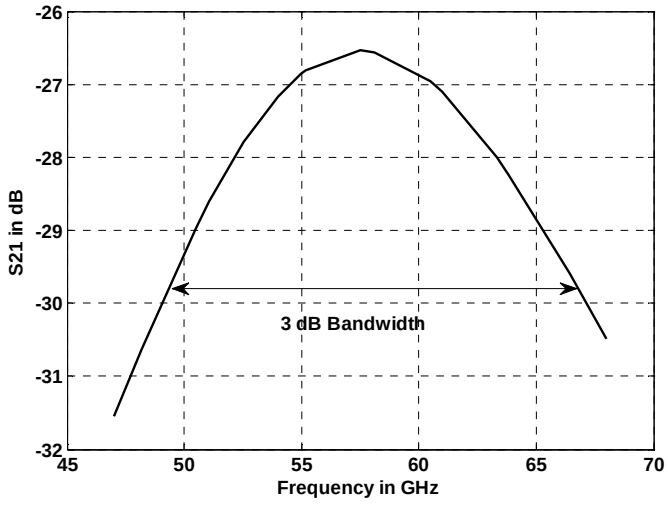


Figure 3.12. Antenna transmission gain (S21) response.

simulation. To represent a typical inter-subnet communication range the transmitter and receiver were separated by 20 mm. The details of the antenna simulation setup are shown in Fig. 3.11. The forward transmission gain (S21) of the antenna obtained from the simulation is shown in Fig. 3.12. As shown in Fig. 3.12, 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.

A system-level simulation using Simulink is performed in order to accurately define the circuit design specifications. From the system-level simulation the design targets are finalized and the wireless transceiver circuitry is designed and simulated using TSMC 65-nm standard CMOS process.

The achieved aggregate power consumption of the entire transceiver is 36.7 mW, 16% lower than the previous design without using body-enabled techniques [61]. It is able to support a data rate of at least 16-Gbps, and a BER $< 10^{-15}$ using an OOK modulation scheme [62] 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 omni-directionality 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.

3.3.5 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. Figures 3.13 and 3.14 show the achievable bandwidth of the proposed mWNoC under uniform random spatial traffic distribution for the three system sizes of 128, 256, and 512 cores divided into 8, 16 and 32 subnets respectively. Figure 3.13 considers two specific architectures (i.e., Mesh-Mesh and Mesh-StarRing), where the upper levels are mesh-based topologies, and the subnets are Mesh and StarRing respectively. Figure 3.14 represents the characteristics of Ring-Mesh and Ring-StarRing architectures. From our experiment we find that the Mesh-StarRing architecture always outperforms Mesh-Mesh architecture as it has better connectivity in the subnets. The same trend is also observed in the architectures with ring-based upper levels, where systems with StarRing subnets always achieve higher bandwidth than those with Mesh subnets. In terms of upper level topologies, systems with a mesh-based upper level always perform better than those with ring-based upper level due to a

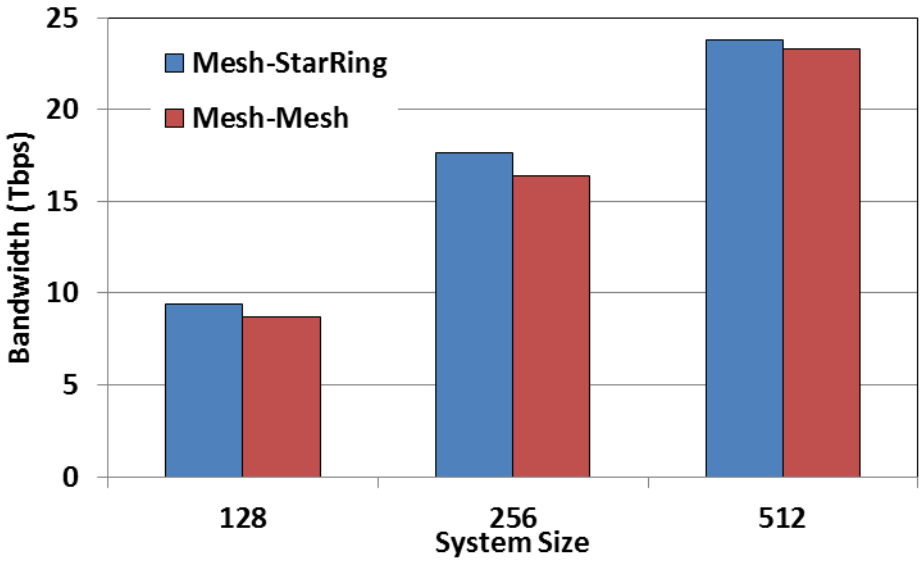


Figure 3.13. Achievable bandwidth with scaling for Mesh-StarRing and Mesh-Mesh architectures.

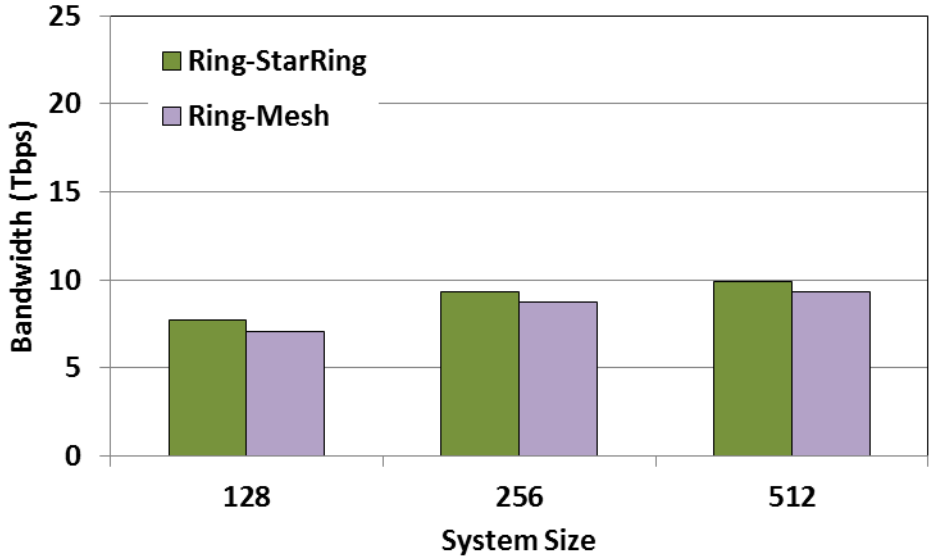


Figure 3.14. Achievable bandwidth with scaling for Ring-StarRing and Ring-Mesh architectures.

more efficient upper level network. These advantages of the mesh-based upper level network come at the cost of extra wiring overhead. For a 256-core system, a mesh-based upper level network has 24 wired links whereas a ring-based counterpart only has 16 wired links, or 33% less links. Since Mesh-StarRing architecture performs best among all the other alternatives considered, this architecture is considered for the subsequent analysis of mWNoC in this work.

Fig. 3.15 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) [12], (iv) hierarchical architecture as mWNoC, but long range links implemented with CNT antenna based THz wireless interconnects (THzNoC) [18] 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

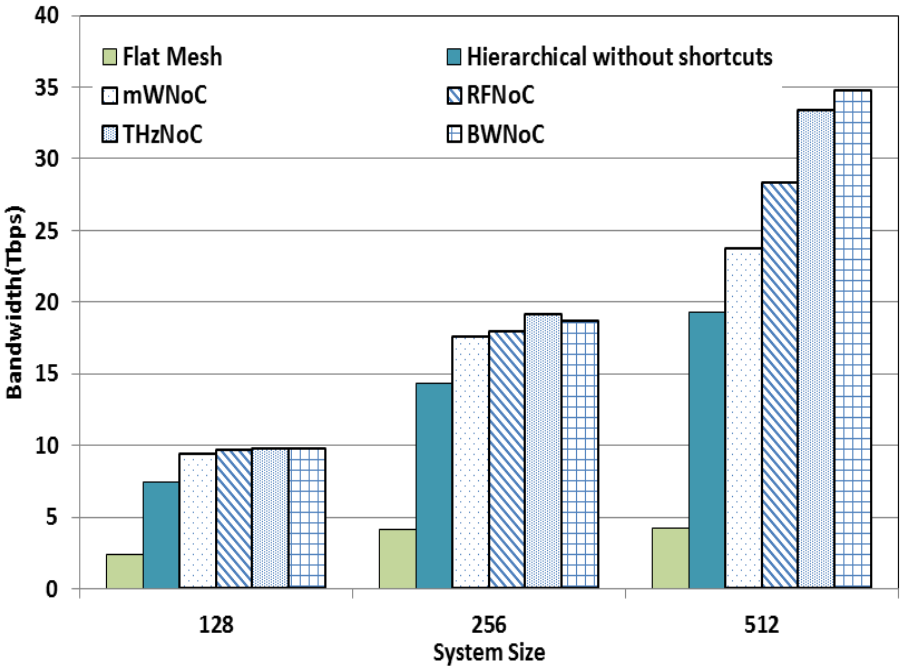


Figure 3.15. Achievable bandwidth with scaling for different NoC architectures.

interconnects proposed in [32] 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 [18].

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 [17], 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 [63]. These shortcuts are placed optimally using the same SA based optimization procedure.

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 SA-based optimization as used for the placement of WIs (section 3.1.2). Each wired shortcut is considered to be 32-bit wide. The wires are designed for minimum delay with an optimum number of uniformly placed and sized repeaters.

The flat mesh architecture performs worst among all the alternatives due to its highest 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. BWNoC, RFNoC and THzNoC 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 3.3.6. 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 the 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 [12]. 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 work.

3.3.6 Energy Dissipation

To quantify the energy dissipation characteristics of the proposed mWNoC architecture, we determine 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

$$E_{pkt} = \frac{N_{\text{int}\,rasubnet} E_{subnet,hop} h_{subnet} + N_{\text{int}\,ersubnet} E_{s-w} h_{s-w}}{N_{\text{int}\,rasubnet} + N_{\text{int}\,ersubnet}} \tag{3.8}$$

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 a switch, and $E_{s\text{-}w}$ 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. The average number of hops per packet in the subnet and the upper level small-world network are denoted by h_{subnet} and $h_{s\text{-}w}$ respectively. Fig. 3.16 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 [12] and [18] respectively. The flat mesh architecture dissipates highest packet energy among all the NoCs considered. A

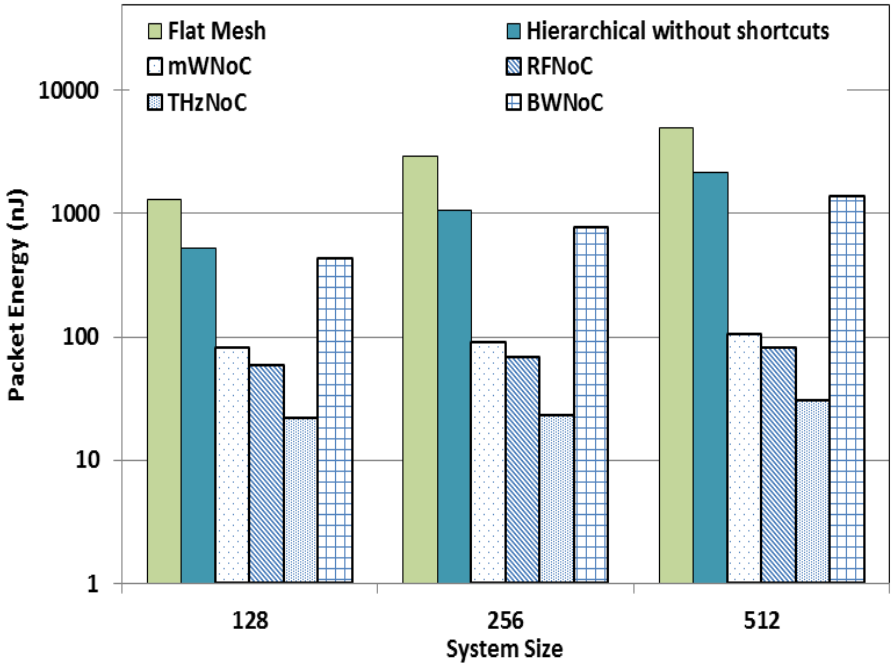
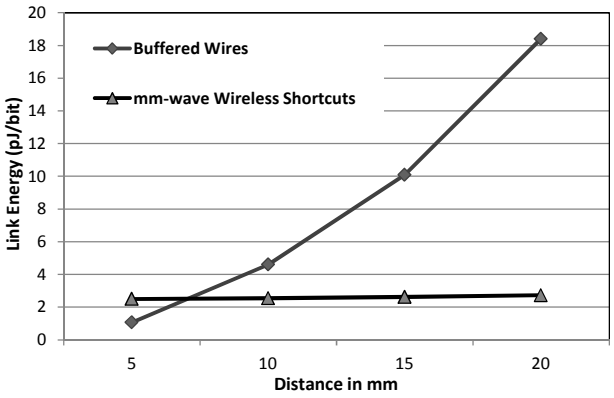


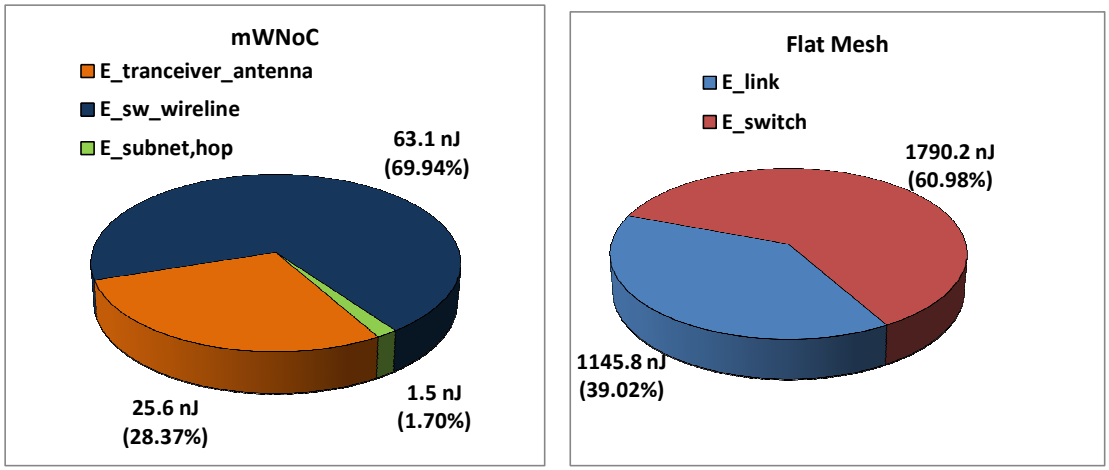
Figure 3.16. Packet energy for different NoC architectures.

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. 3.17 (a) we show the variation of per bit energy dissipation with distance for a wired and a mm-wave wireless link. Fig. 3.17 (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 of the hierarchical NoC with metal 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. 3.16, 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



(a)



(b)

Figure 3.17. (a) The variation of per bit energy dissipation with distance for a wired and a wireless link and (b) Components of packet energy dissipation for mWNoC and flat mesh.

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 3.2.3 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 [61] for the same system size.

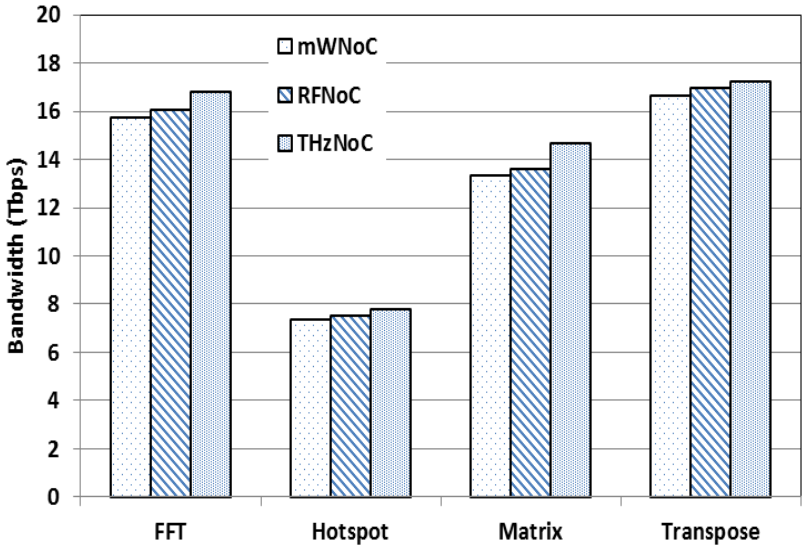


Figure 3.18. Achievable Bandwidth with different traffic scenarios.

3.3.7 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 2^{nd} 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.

Fig. 3.18 shows the achievable bandwidth for the different NoC architectures in non-uniform traffic scenarios. From the results it is evident that for all the traffic patterns considered here the Mesh-StarRing architectures with efficient shortcuts performs very close to each other. The difference between achievable bandwidth of the NoCs considered here is small and follows the same trend for non-uniform traffic scenarios as we have seen for uniform traffic. Due to the presence of higher number of simultaneously operating shortcuts in THzNoC, it performs best in all the different traffic scenarios considered here. The performance of THzNoC is closely followed by RFNoC and mWNoC for all the traffic scenarios considered in this work.

3.3.8 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.

Figure 3.19 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

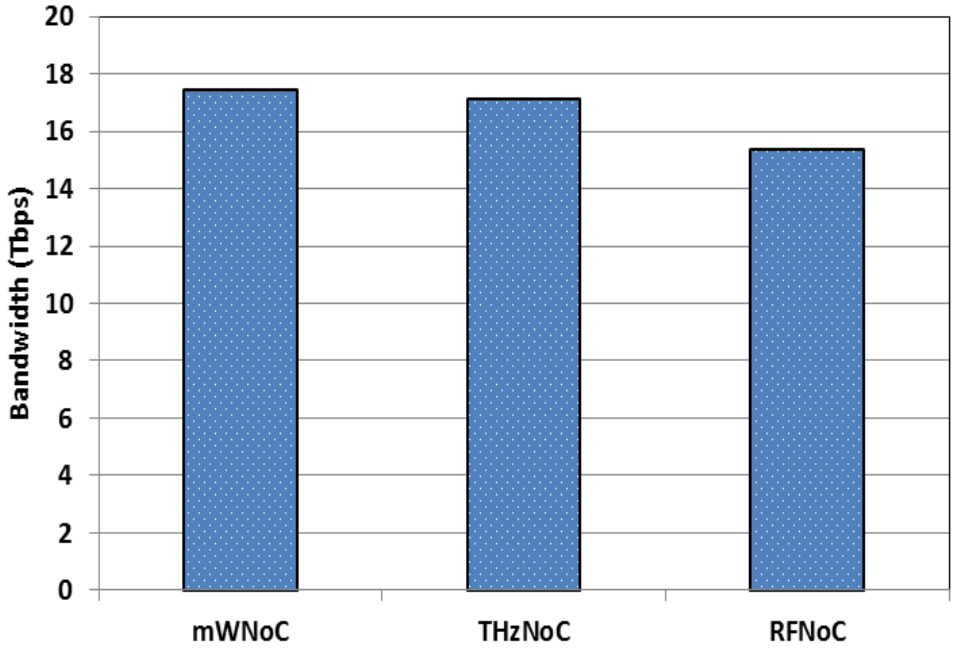


Figure 3.19. Achievable Bandwidth for different NoCs with broadcast traffic.

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.

3.4 Conclusions

In this chapter, we have proposed the design of a small-world NoC architecture with mm-wave wireless interconnects used as long-range links. Design of associated data routing mechanism and optimum placement of wireless hubs are highlighted. The architectural innovations proposed in this work are made possible by the use of low power and high speed wireless links capable of communicating directly between distant parts of the chip in a single hop. 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. The gains in network performance metrics are in part due to the architecture and the rest is due to the adopted high bandwidth, energy efficient wireless links. 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. 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 work.

Chapter 4

Multi-Channel mWNoC

The performance of mWNoCs can be significantly improved by optimally placing and using multiple simultaneously operating wireless shortcuts. The works of [39] and [40] have already discussed the feasibility and advantages of multiple non-overlapping wireless channels in NoC environment. We extend our single channel mWNoC design discussed in chapter 3 by incorporating multiple simultaneously operating non-overlapping channels. This chapter evaluates the performance of a small-world NoC with multiple non-overlapping mm-wave wireless channels as long-range links. Multiple wireless shortcuts operating simultaneously further enhance the performance and provide an energy efficient solution for design of communication infrastructures for multi-core chips. We demonstrate that the proposed multi-channel mWNoC outperforms its more traditional non-hierarchical wire line counterparts in terms of sustainable data rate and energy dissipation. It also performs better than other emerging NoC architectures like, NoC with RF-I and 3D NoC. Moreover, the performance gap with wireless NoC designed with THz wireless links (THzNoC) becomes smaller in case of mWNoC with multiple non-overlapping channels. The area overheads associated with these novel NoC architectures are also quantified and it is shown that performance benefits clearly outweigh the overheads.

4.1 mWNoC Architecture with Multiple Non-Overlapping Wireless Channels

We use the same hierarchical NoC architecture as developed in previous chapter with a limited number of wireless shortcuts operating in different frequencies strategically placed for

optimum performance. The small-world topology is incorporated in the hierarchical NoC by introducing long-range, high bandwidth, low power and simultaneously operating wireless links between distant cores. The hubs connected through wireless links require wireless interfaces (WIs) and here we also use a SA-based algorithm to optimally place the WIs so as to establish optimal overall network topology under given resource constraints, i.e., a limited number of WIs. The number of WIs sharing the same frequency channel is kept equal for different frequency bands along with one gateway hub (which can operate in all frequency channels). Multiple non-overlapping wireless channels are distributed between all the WIs and the WIs sharing the same channel form a cluster. Figure 4.1 shows a representative hierarchical 256-core network where the subnets have a Ring-Star (a ring with a central hub connecting to every core) topology and it has 16 hubs and 7 WIs. The hubs are connected in mesh architecture with overlaid long-range wireless shortcuts on the 2nd level of the hierarchy. In this chapter Mesh-RingStar architecture is used as an example since it is shown in the previous chapter to provide the best performance-overhead tradeoff among several possible mWNoC architectures.

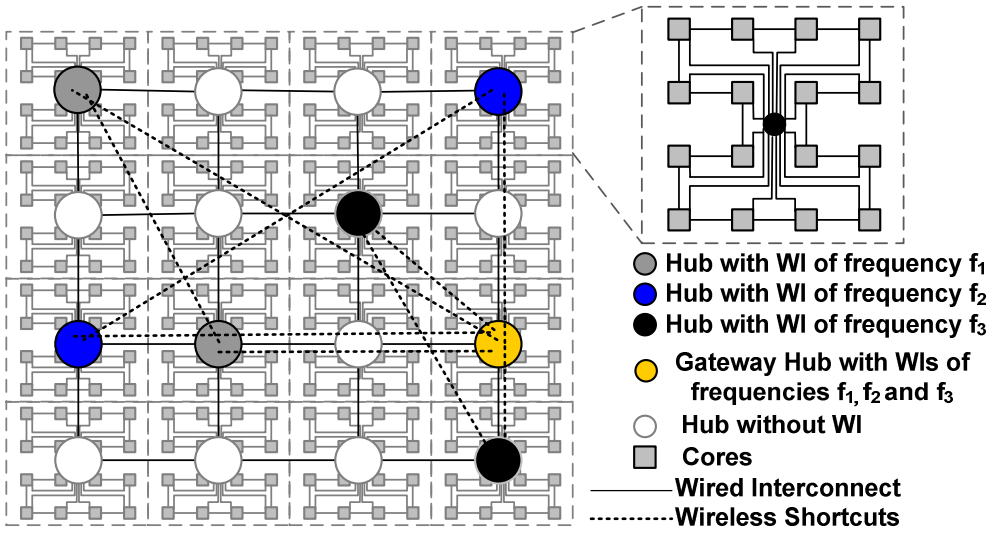


Figure 4.1. A hierarchical 256-core network with multiple simultaneously operating wireless shortcuts.

4.1.1 Optimum Placement of WIs of different frequency Bands

The SA based optimization method used for WI placement in the previous chapter is modified to assign multiple simultaneously operating frequency channels. These non-overlapping wireless channels are distributed among *n* WIs and WIs sharing the same channel form a cluster. Each frequency band has same number of WIs sharing the channel and one gateway hub (which can operate in all frequency channels) is used for inter-cluster communication. The optimization metric μ introduced in the previous chapter is used for SA. The optimization metric, μ can be computed as

$$\mu = \sum_{i,j} h_{ij} f_{ij} , \tag{4.1}$$

$$h_{ij} = p * d_{i,j\_with\_shortcut} + (1-p) * d_{i,j\_without\_shortcut} \tag{4.2}$$

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 normalized apriori probability of traffic interactions between subnets determined by particular traffic patterns depending upon the application mapped onto the NoC. Since multiple non-overlapping channels work simultaneously in this case, the token which grants access to the wireless medium circulates inside each cluster. Hence, 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 in a cluster (n_c) sharing the same frequency channel. With the assumption that all the WIs are equally likely to have access to wireless channel in a cluster, p can be computed as

$$p = 1/n_c . \tag{4.3}$$

The distance ($d_{i,j}$) between source and destination varies depending on whether or not

wireless shortcuts are used while routing. In this chapter as well, equal importance is attached to inter-hub distance and frequency of communication.

4.2 Communication Scheme

This section describes the WI components and the adopted routing strategy for multi-channel mWNoC.

4.2.1 Wireless Interfaces for non-overlapping frequency bands

The non-overlapping wireless interfaces are established using zigzag antennas and OOK modulation based transceivers. The metal zigzag antenna characteristics depend on physical parameters like axial length, trace width, arm length, bend angle, etc. By varying these parameters antennas are designed to operate on different non-overlapping frequency channels in this work. The transceivers for all the frequency bands use non-coherent OOK modulation for its relative simplicity and low-power consumption as discussed in the previous chapter.

4.2.2 Adopted Routing Strategy

In this proposed hierarchical NoC, intra-subnet data routing depends on the Ring-Star subnet topology. In the subnet, if the destination core is within two hops on the ring from the source, then the flit is routed along the ring; otherwise the flit goes through the central hub to its destination. To avoid deadlock within the subnet, we adopt the virtual channel management scheme from Red Rover algorithm [54], 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 requires flits to use the upper level network. By using the wireless shortcuts between the hubs with WIs of the same frequency channel, flits can be transferred in a

single hop between them. If the source and destination WIs are tuned to different frequencies, flits are first routed to a gateway hub via wireless links and are then transmitted using the destination WI's frequency channel. If the source hub has no 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 has no WI, then the nearest WI hub receives the data and routes it to the destination through wired links. Between a source and destination hub pair without WIs, the routing path with 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 [55] and distributed routing are used to alleviate this problem. 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]. Figure 4.2 shows a particular communication snapshot of a mesh-based upper level network where hub 12 wants to communicate with hub 3. In this example, the WIs at the following hubs are tuned to the same frequency: hub 0, 9; hub 3 & 8, and hub 6 & 15. Hub 11 is a gateway hub which can operate in all frequency channels and used for bridging from one frequency channel to another. First at source 12, the nearest WI (8 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 8 is checked and this path is taken only if the token is available. If this is not the case, the message at hub 12 follows dimension order routing towards the destination and arrives at hub 13. At hub 13, again the shortest path using WIs is searched and if the token from hub 9 allows the usage of wireless shortcuts, then the message is routed through hub 9. Since hub 9 and the destination hub 3 are tuned to different

frequencies, a direct communication between these two hubs is not possible. Here we utilize the gateway to bridge hub 9 and hub 3. The gateway is equipped to communicate with all the clusters, so it receives message from hub 9 and then retransmits the message to hub 3. 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 along with token flow control prevents deadlocks and effectively improves performance by distributing traffic though alternative paths.

The wireless hubs are grouped into clusters, each tuned to a particular frequency. As the wireless hubs in a particular cluster use the same frequency and can send or receive data from

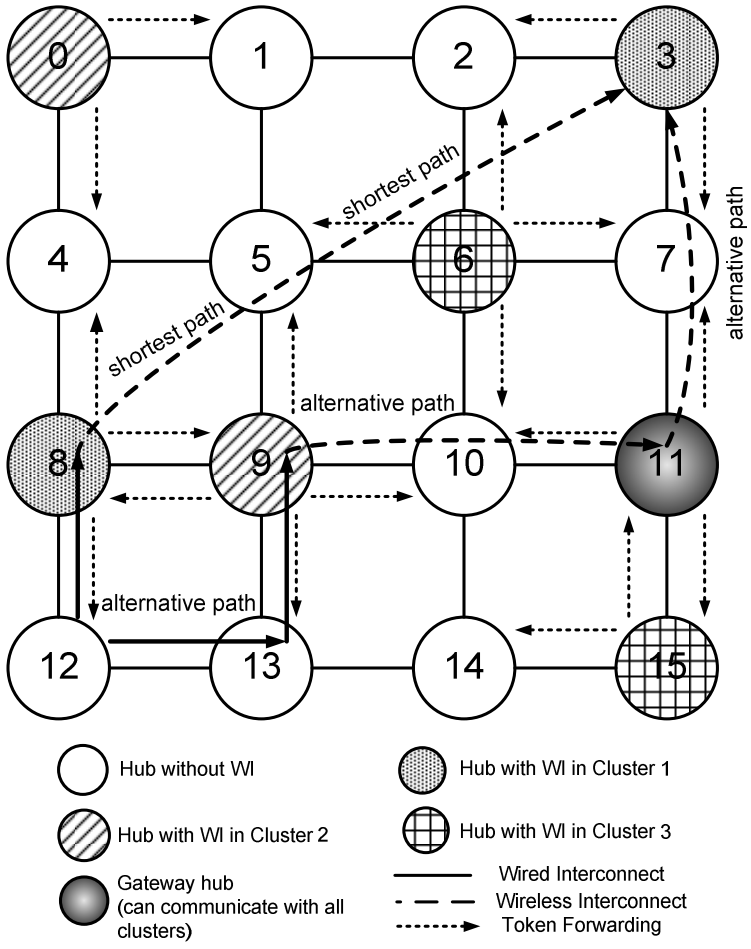


Figure 4.2. An example of token flow control based distributed routing.

any other wireless hub in that cluster, an arbitration mechanism must be designed to grant access to the wireless medium to a particular hub at a given instant to avoid interference and contention. To avoid centralized control and synchronization, the arbitration policy adopted is a wireless token passing protocol. (Note that the use of the word token in this case differs from the usage in the above mentioned token flow control.) In this scheme a dedicated token circulates in each cluster. The particular WIs possessing the wireless tokens can broadcast flits into the wireless medium in their respective clusters. The wireless token is forwarded to the next wireless hub in the same cluster after all flits belonging to a packet at the current hub are transmitted.

4.3 Experimental Results

This section characterizes multi-channel mWNoC performance through simulation and analysis in presence of various traffic patterns. Characteristics of the on-chip wireless communication channel and selection of the optimum number of WIs for different system sizes are presented, followed by detailed network simulations with various system sizes. We also present performance benchmarking with respect to other emerging NoC architectures, viz., NoC with RF interconnects (RF-I), 3D NoC and THzNoC.

4.3.1 Wireless Channel Characteristics

The metal zigzag antennas described earlier are used to establish the on-chip wireless links. Antenna characteristics 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 between WIs, the transmitter and receiver were separated by 20 mm. The antenna's forward transmission gains (S21) obtained via simulations are shown in Figure 4.3. We are able to obtain three different channels with 3 dB bandwidths of 16 GHz and center

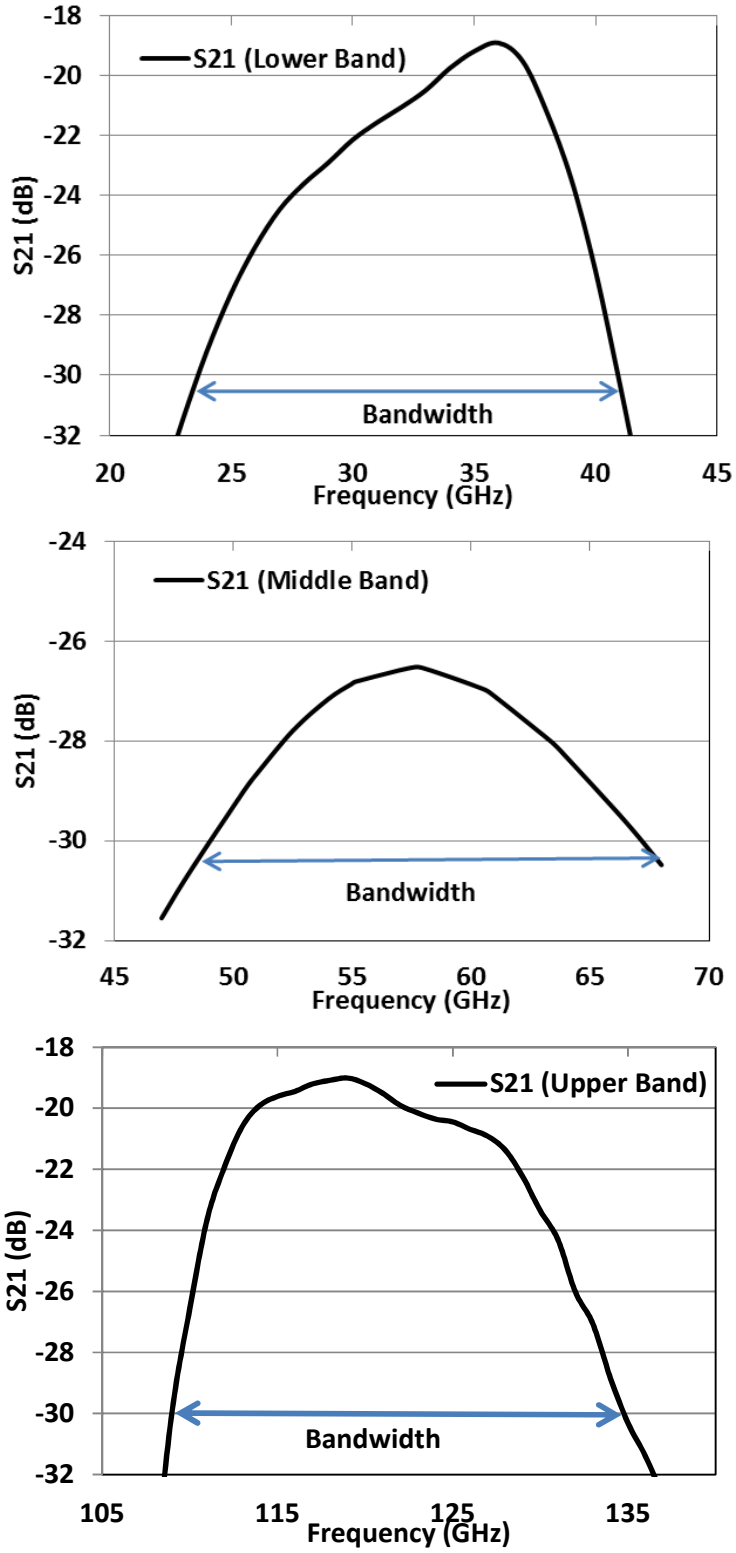


Figure 4.3. Antenna transmission gain (S21) for three non-overlapping channels.

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. The OOK transceiver can sustain a data rate of 16 Gbps with a power consumption of 43.6 mW [52].

4.3.2 Optimum Number of WIs

To reduce hardware overhead, we aim to limit the number of WIs on the chip without significantly compromising the overall performance. We assume round-robin token circulation among WIs. The token is considered to be a single flit transmitted from the WI currently holding it to the next one. The smaller the token return time to a particular WI is, the better the network performance is since wireless medium acquisition delay is minimized. On the other hand, hop-count decreases with more WIs due to higher connectivity as a result of introduction of additional WIs in the network's upper level. Since these are two opposing trends, a tradeoff needs to be established. Hence, we study achievable network bandwidth and packet energy as a function of the number of WIs. The upper level of the network is considered a mesh with three simultaneously operating wireless shortcuts and the subnet architecture is Ring-Star as shown in Figure 4.1. The WI clusters are equal in size and a single WI with transceivers of all frequencies acts as gateway between different clusters. Figure 4.4 shows that for a 512-core Mesh-RingStar system (32 subnets with 16 cores per subnet) bandwidth increases with number of WIs until reaching a maximum at 13 WIs (3 clusters of 4 WIs each and a gateway) and then it decreases. 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 512-core mWNoC as 13. Similarly, for 8 and 16

subnet systems optimum performance is achieved with 5 and 7 WIs respectively. More than one frequency channel can be used by each WI for a system with only 5 WIs.

4.3.3 Performance Evaluation

In this section we analyze mWNoC characteristics and study performance trends as the system size scales up. We consider three different system sizes, namely 128, 256, and 512 cores divided into 8, 16 and 32 subnets respectively. The die area is kept fixed at 20 mm x 20 mm for all system sizes. The NoC switch architecture is adopted from [60]. The hubs and NoC switches in the subnets have 4 virtual channels per port and have a buffer depth of 2 flits. Each packet consists of 64 flits. The WI ports have an increased buffer depth of 8 flits, which ensures that all messages trying to access wireless links are efficiently handled without compromising

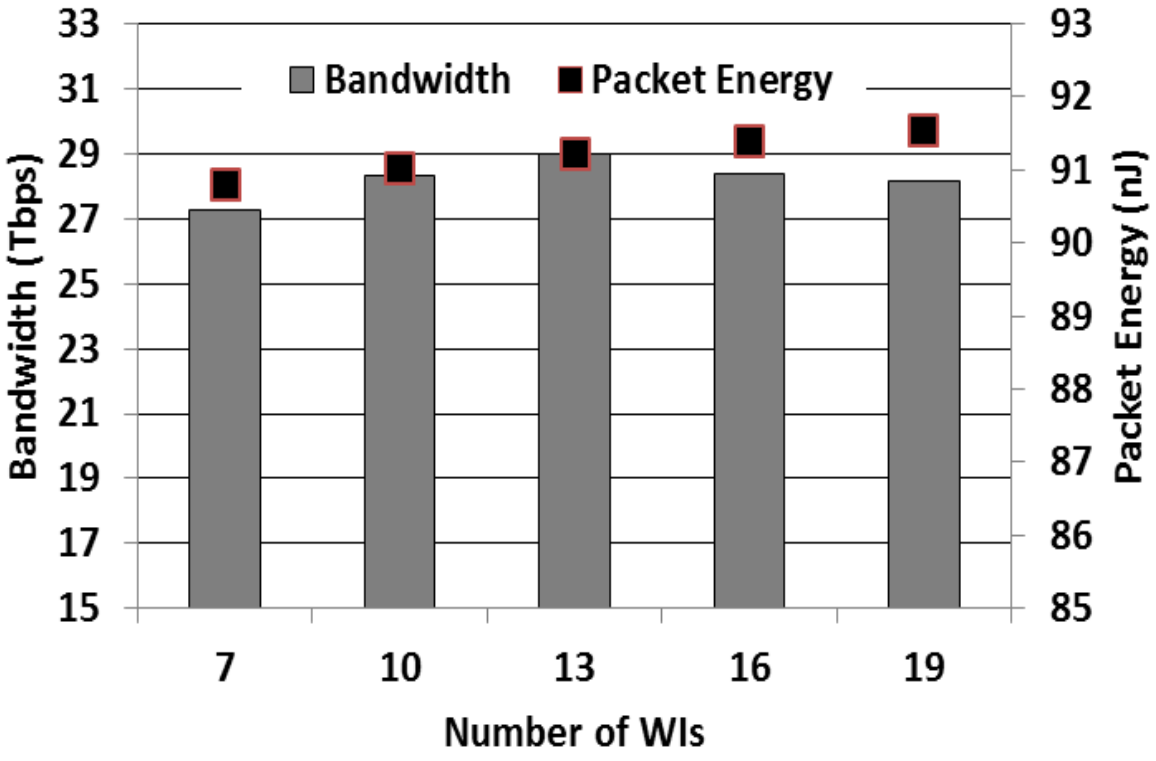


Figure 4.4. Performance variation with different number of WIs for a 512-core system with 32 subnets.

performance as shown in the previous chpter. A self-similar traffic injection process is assumed.

The Mesh-RingStar architecture introduced earlier is simulated using a cycle accurate simulator. The subnet switches and the hub digital components are synthesized using 65 nm standard cell library from TSMC at a clock frequency of 2.5 GHz. The delays in flit traversals along the wired interconnects of the hybrid NoC architecture are considered when quantifying the performance.

These include the intra-subnet core-to-hub wired links and the inter-hub links in the network's upper level. The delays through the switches and inter-switch wires of the subnets and the hubs are taken into account as well.

Figure 4.5 shows achievable bandwidth of the proposed mWNoC for three different system sizes considered under a uniform random spatial traffic distribution. We considered mWNoC with one, two and three simultaneously operating wireless channels. 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) hierarchical architecture as the mWNoC, but with shortcuts implemented using buffered metal wires instead of wireless links (BWNoC). The number of wired shortcuts is kept equal to the number of WIs for different system sizes and they are optimally placed using the same SA-based optimization used for the placement of WIs. Each wired shortcut is considered to be 32 bits, which is equal to the width of a flit considered here. The wires are designed with an optimum number of uniformly placed and sized repeaters. The mWNoC with three simultaneously operating channels outperforms all the other alternatives for the three system sizes, except for the system with buffered wired shortcuts. The flat mesh architecture performs the worst due to its high average hop count. The hierarchical architecture improves the performance by reducing hop

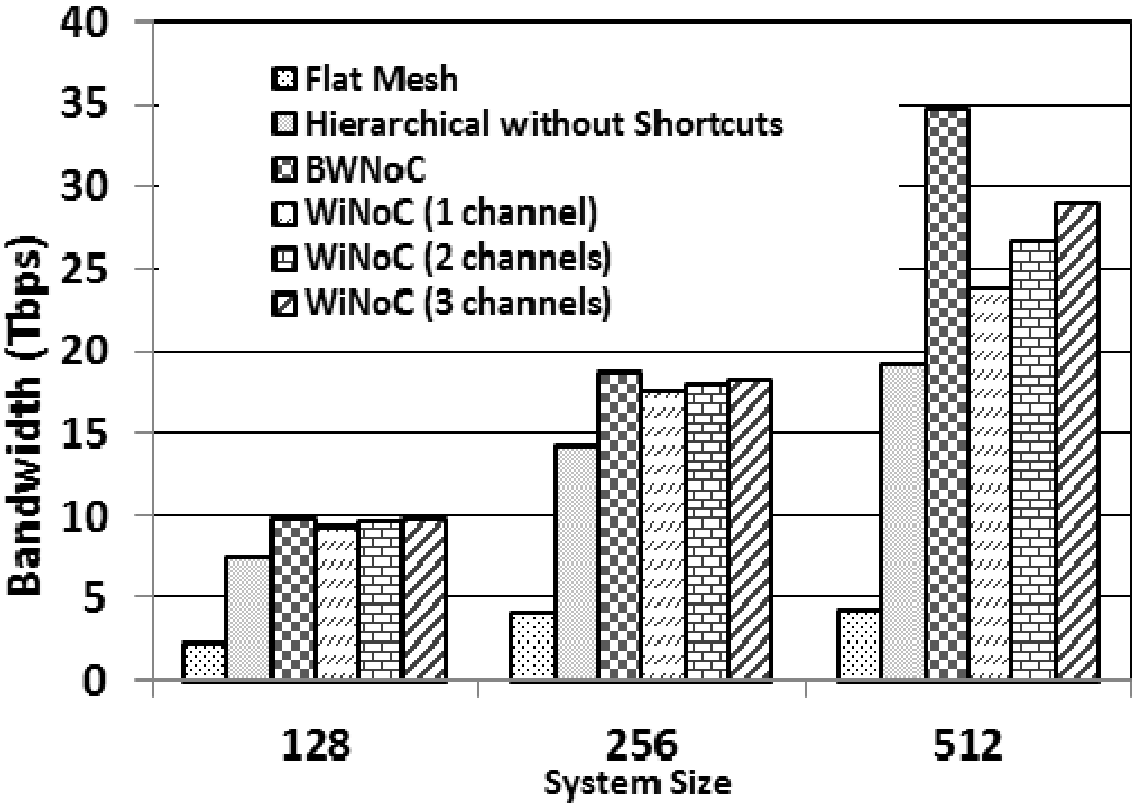


Figure 4.5. Achievable bandwidth for different system sizes.

count, but the best performance is obtained from the hierarchical architecture with multiple 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. It can be observed that the bandwidth of the mWNoC with three non-overlapping channels improves compared to the initially proposed mWNoC with single wireless channel [64] for all the system sizes considered. Specifically, for higher system size the performance gain is more.

4.3.4 Energy Dissipation

We determine the mWNoC's 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. The energy dissipations of the switches and hubs are obtained through synthesis using Synopsys

tools with 65 nm standard cell libraries from TSMC. The energy dissipated by the wireless transceiver is determined through Cadence simulations. The energy dissipation of the wired links is obtained from the Cadence layout, assuming a 20 mm x 20 mm die area.

Figure 4.6 shows the packet energy dissipation of the considered architectures for uniform random traffic. 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 the latency between the cores. Packets get routed faster and hence occupy resources for less time and dissipate less energy in the process. The mWNoC further improves performance by employing multiple energy efficient long range

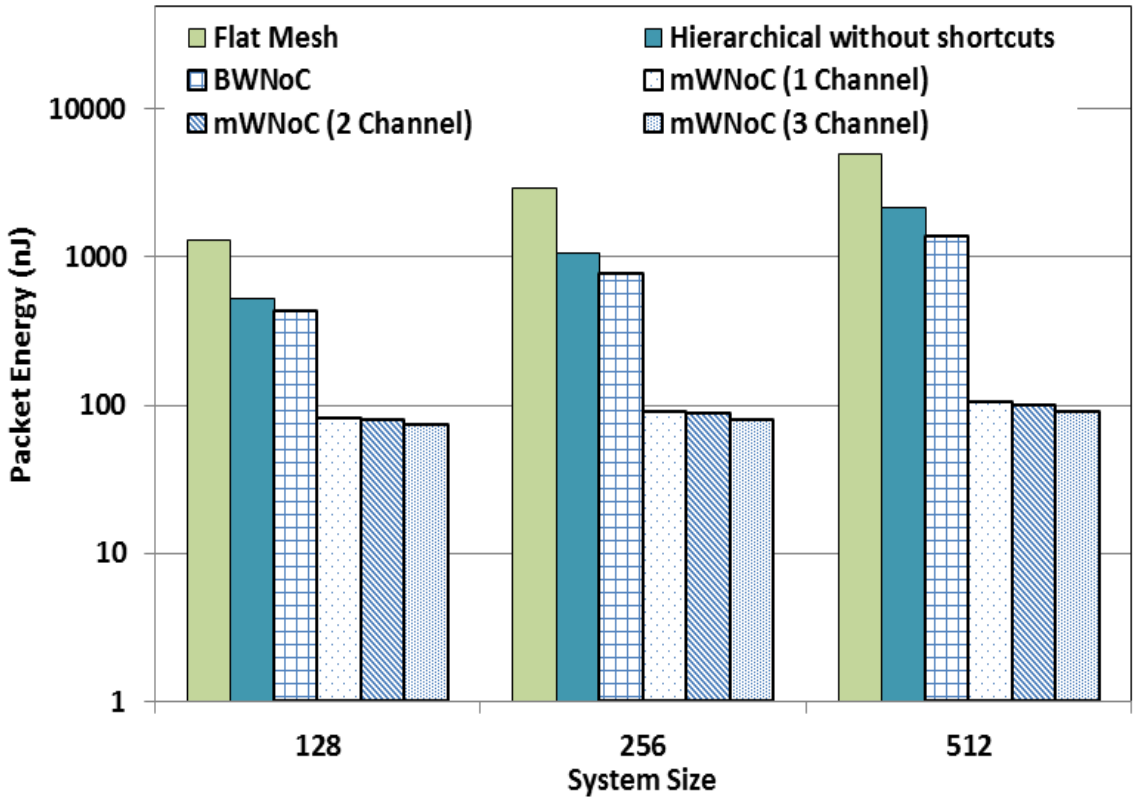


Figure 4.6. Packet energy for different NoC architectures.

shortcuts in the hierarchical network. As shown in the previous chapter, wireless shortcuts are always energy efficient whenever the link length is 7 mm or more. In our implementation, the minimum and maximum distances between the WIs communicating using the wireless channel are 7.07 mm and 18 mm respectively. Therefore, in this design, using the wireless channel is always more energy efficient. The mWNoC with multiple non-overlapping channels has reduced packet energy for all system sizes compared to the single channel mWNoC of [64]. Also, the mWNoC significantly reduces energy dissipation compared to the other two possible wired hierarchical architectures and can reduce the packet energy dissipation by at least an order of magnitude compared to the flat mesh.

4.3.4 Comparative Evaluation of Multi-Channel mWNoC

In this section we first perform a comparative analysis between the mWNoC and two other

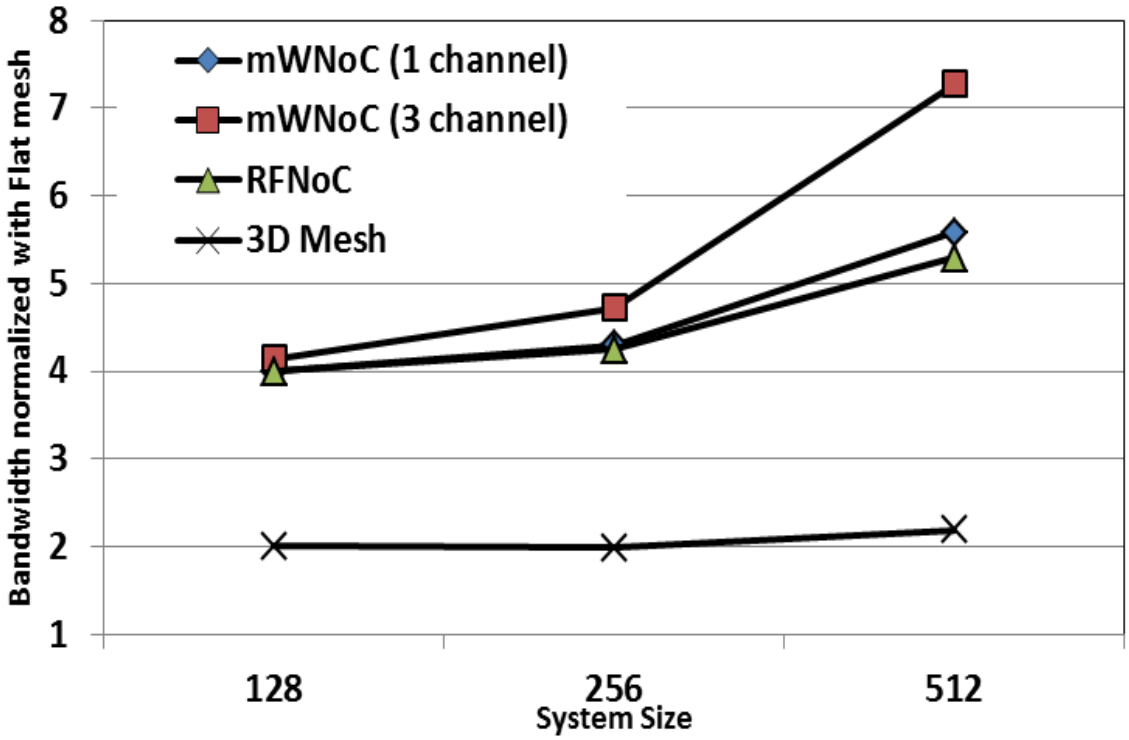


Figure 4.7. Comparative performance evaluation for different emerging NoCs.

emerging NoCs. The on-chip RF transmission line (RF-I) proposed in [17] is a new interconnect technology that can improve NoC performance. Hence, we designed a small-world NoC (RFNoC) by replacing the wireless communication links of the mWNoC by RF-Is, maintaining the same hierarchical topology. Like the wireless links, these RF links can be used as long-range shortcuts. These shortcuts are optimally placed using the same SA-based optimization used for placing WIs in the mWNoC. As mentioned in [17], in 65 nm technology it is possible to have 8 different frequency channels, each operating with a data rate of 6 Gbps and used for long-range inter-subnet communications. We also considered a 3D mesh-based NoC with four layers as in [21]. Due to the high energy dissipation hierarchical NoC without shortcuts and BWNoC are not considered for this analysis.

For this comparative evaluation, we first present the normalized bandwidth with respect to flat mesh for different system sizes in Figure 4.7 for uniform random spatial traffic distribution. From this result it is evident that performance benefits are more prominent for bigger systems and is highest for 512-core. Consequently, a 512-core system is considered for subsequent analysis.

We consider both uniform and non-uniform traffic patterns in this evaluation. For non-uniform traffic patterns we use synthetic and application-based traffic distributions. We considered two types of synthetic traffic. First, a transpose traffic pattern [7] is considered where cores in a certain number of subnet pairs are considered to communicate more frequently with each other. We consider three such pairs and 50% of packets originating from one of these subnets 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 subnets more frequently than with the others. We consider three such hotspot subnets to which all other cores

send 50% of the packets that originate from them. Transpose and hotspot traffics are mapped in 3D NoC by selecting sets of adjacent cores to form groups (equivalent to subnets of mWNoC). In the transpose traffic three of these groups communicate with each other and also we consider three hotspot groups. We consider two application-based traffics. A 1024-point FFT is considered where each core performs a 2-point radix 2 FFT computation. Multiplication of two 512x512 matrices is used to generate another application-based traffic pattern.

Figure 4.8 shows the achievable overall network bandwidth for the different NoC architectures in uniform and non-uniform traffic scenarios for a 512-core system. It can be observed that in case of non-uniform traffic, due to the skewed communication pattern, certain interconnects on the path are overloaded and become bottle-neck affecting the overall

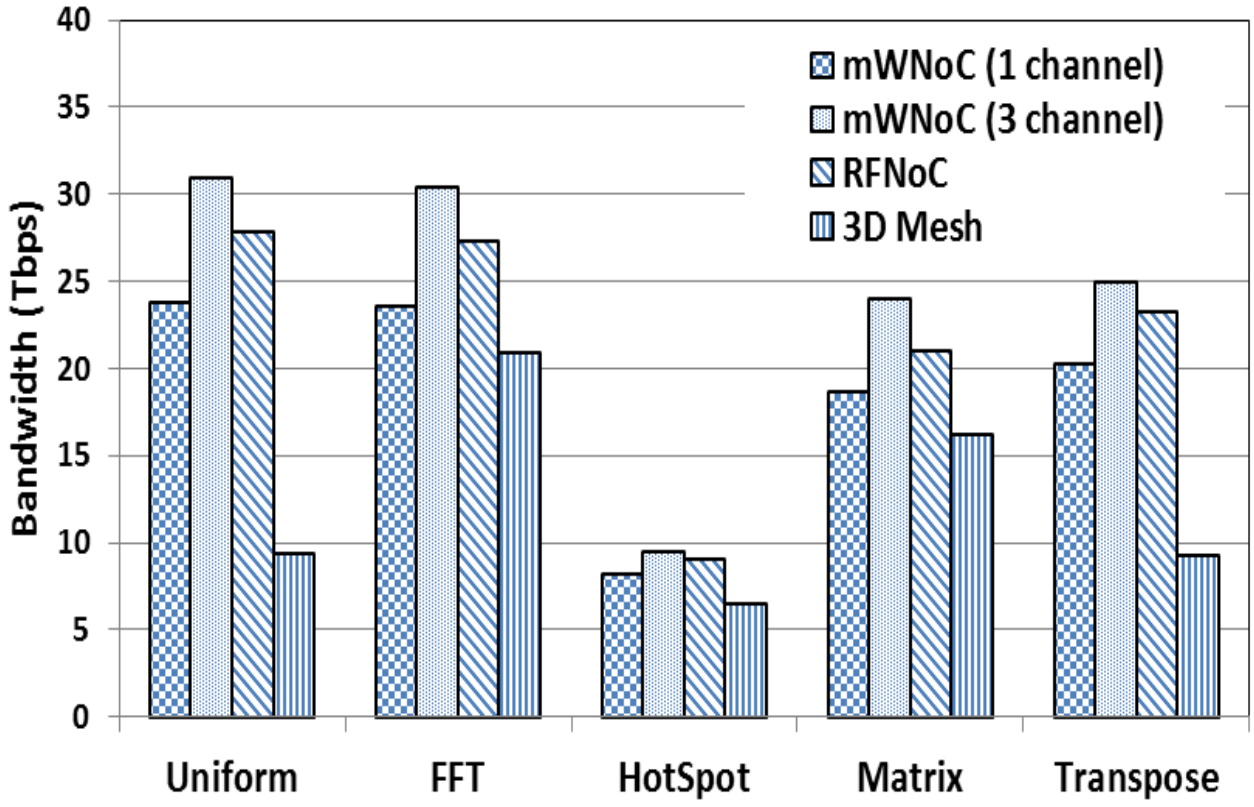


Figure 4.8. Achievable bandwidth for different traffic patterns.

performance of the NoC. This is most prominent in case of hotspot traffic. The 3D mesh based NoC suffers from the fact that number of layers is limited to four, which results in poor performance for 512-core system size. Though mWNoC and RFNoC have the same hierarchical architecture, mWNoC performs better. In RFNoC once the 8 shortcuts are placed they are fixed for that traffic pattern. But, in mWNoC, 13 WIs are placed depending on the traffic and wireless communication channel can be established between any of those pairs. Moreover, the total long-range link area overhead and the layout challenges of the 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 [17]. This is significantly higher than the combined length of all the antennas used in the mWNoC, which is 6.45 mm for the 512-core system.

We also compared multi-channel mWNoC with another emerging wireless NoC, namely THzNoC introduced in the previous chapter. The achievable bandwidths for 256-core and 512-core system mWNoCs with three simultaneously operating wireless channels and THzNoC is

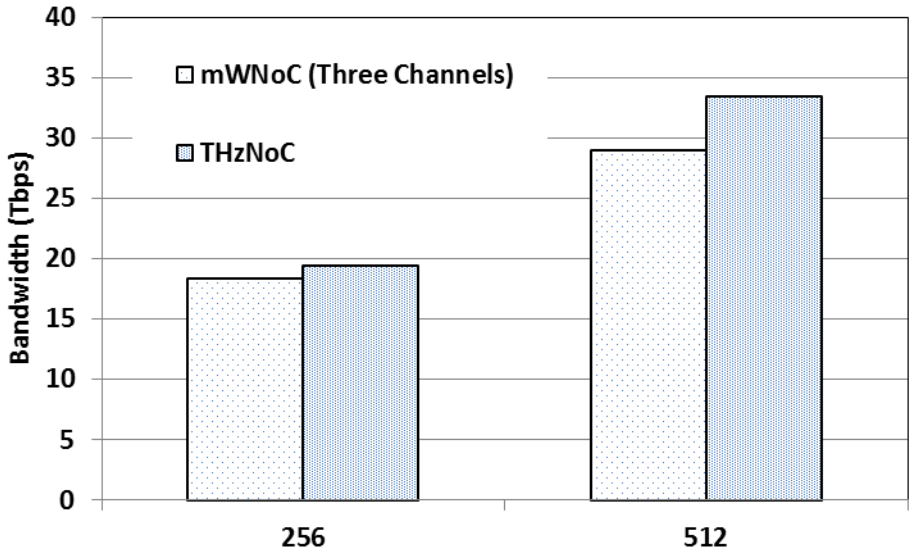


Figure 4.9. Achievable Bandwidth for NoCs with different interconnects.

shown in Figure 4.9 for uniform random traffic. 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 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 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.

4.4 Area Overhead

In this section we quantify the area overhead due to the wireless deployment in the mWNoC.

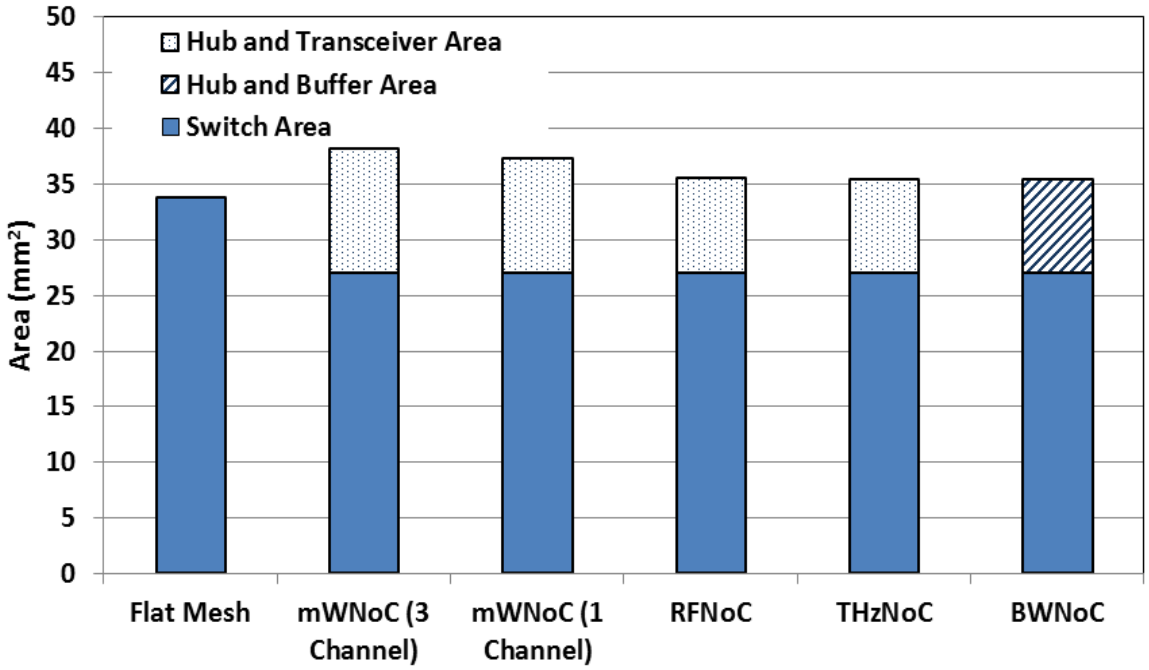


Figure 4.10. Silicon area overhead for three different NoCs of size 256 core.

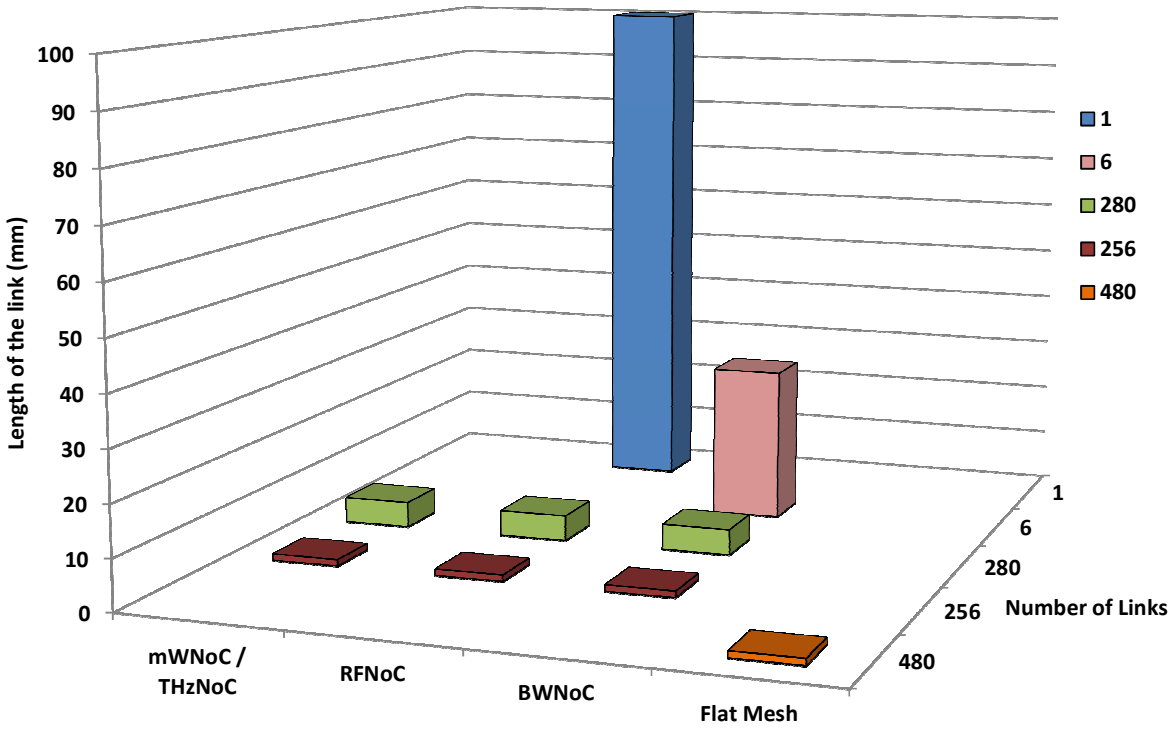


Figure 4.11. Total wiring requirements of various lengths for a 20 mm x 20 mm die for three different NoCs of 256 core system size.

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 [17] and [18] respectively.

Total silicon area overheads for flat mesh, mWNoC, RFNoC, THzNoC and BWNoC for a

256-core system are shown in Figure 4.10. 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. Figure 4.11 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.

4.5 Conclusions

This chapter demonstrates that with multiple simultaneously operating mm-wave wireless channels, the performance of mWNoC can be improved significantly. By incorporating a hierarchical small-world topology and multiple non-overlapping wireless channels, the proposed mm-wave NoC architecture gives considerable performance gains in presence of various traffic scenarios without incurring significant area overhead. Moreover, with multiple non-overlapping channels mWNoC performs better than RFNoC and the performance gap with THzNoC becomes smaller.

Chapter 5

Conclusions and Future Work

This chapter concludes the work undertaken in this thesis by summarizing the salient contributions. It also points towards various promising future directions emanating from this research endeavor.

5.1 Conclusions

Massive levels of integration is making modern Multi-core chips all-pervasive in several domains ranging from scientific applications like weather forecasting, astronomical data analysis, bioinformatics applications to even consumer electronics. Design of many-core integrated systems beyond the current CMOS era will present unprecedented advantages and challenges, the former being related to very high device densities and the latter to soaring power dissipation issues. Communication plays a crucial role in the design and performance of many-core SoCs. According to the International Technology Roadmap for Semiconductors (ITRS) in 2007, *the contribution of interconnects to chip power dissipation is expected to increase from 51% in the 0.13µm technology generation to up to 80% in the next five year period*. This clearly indicates the challenges facing future chip designers associated with traditional scaling of conventional metal interconnects and material innovation. NoCs have been proposed as a promising solution to simplify and optimize SoC design. However, it is expected that improving traditional communication technologies and interconnect organizations will not be sufficient to satisfy the increasing demand for energy-efficient and high performance interconnect fabrics. Among several emerging possibilities, wireless NoC is a promising option. In this work we have presented various design possibilities and challenges for CMOS compatible mm-wave wireless NoC (mWNoC) architectures as communication backbones for many-core chips. It is shown in

this work that NoC architectures inspired from Complex Network Theory using CMOS compatible wireless interconnects can achieve significant performance benefits compared to traditional wireline NoCs under both synthetic and real application based network traffic. It is also shown that the wireless interconnect enabled mWNoC can perform better than NoC with other alternative interconnect technologies.

The mWNoC paradigm is still in its initial stages. It needs extensive investigations to make it a viable alternative to existing interconnect infrastructures. Also, it should be noted that on-chip

TABLE 5.1

COMPARISON OF THE THREE EMERGING INTERCONNECT PARADIGMS

		3D Integration	**Optical Interconnects**	**Wireless Links**
Design Requirements		Multiple layers with active devices	Silicon photonic components	On-chip metal or CNT-based antennas
Performance Gains	Bandwidth Advantage	Higher connectivity & less hop count	High speed optical devices and links	Single hop high bandwidth wireless links
	Lower Power Dissipation	Shorter average path length	Negligible power dissipation in optical data transport	Multi-hop paths replaced by single hop links
Reliability Issues		Vertical Via Failure	Temperature sensitivity of photonic components	Noisy wireless channels
Challenges		Heat dissipation due to higher power density, yield	Integration of on-chip photonic components	Low power mm-wave transceivers & Control over CNT growth

wireless communication is not the only emerging paradigm alternative to traditional planar metal/dielectric-based interconnects. Three-dimensional integration and nanophotonic interconnects are other alternatives. All these new technologies have been predicted to be capable of enabling many-core designs which improve the speed and power dissipation in data transfer significantly. However, these alternative interconnect paradigms are in their formative stage and need to overcome significant challenges pertaining to integration and reliability. The main feature of all these emerging interconnect paradigms are summarized in Table I. All these emerging paradigms can offer remarkable performance advantage in many-core SoCs. However, in order to harvest their potential more research is necessary to address various challenges in multiple areas including system architecture, circuit design, device fabrication and CAD tool development. Moreover, the achievable performance benefits of mWNoCs need to be benchmarked and relevant design trade-offs need to be established with respect to these other alternative emerging paradigms.

5.2 Future Directions

The research performed for this thesis work can be carried forward in several far reaching directions as discussed below:

5.2.1 Thermal Modeling of mWNoC

Temperature-aware design has become a necessity as the power density is rising with the increasing amount of devices that can fit in the same area of a chip. To ensure proper functioning of the mWNoC in different operating conditions an accurate thermal model suitable for architectural studies is very important. There is a tremendous need for design techniques to help control and reduce heat dissipation, especially runtime techniques that can regulate operating temperature when the package's thermal capacity is exceeded. Runtime response provides safe

cooling and prevents thermal emergencies by changing the processor's behavior rather than relying on costly thermal packaging. Evaluating such techniques, however, requires a thermal model that is practical for architectural studies and we plan to develop such a thermal model to enhance the performance and functionality of mWNoC.

5.2.2 Resilient mWNoC with ECC schemes

The ITRS [5] has predicted signal integrity to be a major challenge in current and future technology generations. Transient errors are becoming increasingly important due to increase in crosstalk, ground bounce and timing violations. These transient events are made more and more probable due to several reasons. With increased device density, the layout dimensions are shrinking and hence the charge used for storing the information bits in memory as well as logic, is reducing in magnitude [65]. Shrinking storage charges also make the chips vulnerable to events like alpha particle hits. Increasing gate counts force designers to lower the supply voltages to keep power dissipation reasonable and thus reduce noise margins. Highly packed wires increases coupling between adjacent wires and opposing transitions induce crosstalk generated faults on these lines. Faster switching rates cause ground bounce and timing violations which manifest as transient errors. There are several ways to address signal integrity issues in an on chip environment like minimization of radiation exposure, careful layout, use of new materials and error control coding schemes. Moreover, the performance of the wireless links in the mWNoC depends on the transceiver circuit and the antennas which are vulnerable to manufacturing defect or anomalies. Error control coding (ECC) enables us to address the transient sources of errors at a higher level of abstraction in the system design phase rather than at a post design, layout phase. For an on chip environment we need simple coding schemes that will not impose a limiting overhead due to the encoding and decoding complexity. Different

error rates due to distinct events in the wireless and wireline links require different ECC schemes. We intend to evaluate the error rates of both type of links in the mWNoC and design appropriate ECC schemes to enable corrective intelligence in the mWNoC fabric.

5.2.3 Complex Network based mWNoC architectures

It is imperative that emerging interconnect paradigms replace or at least augment traditional metal interconnects for on-chip communication in future many-core chips. This will enable many-core chips to deliver the power-performance demands in the extremely demanding target application areas. To achieve this paradigm shift in design of interconnect infrastructures for massive many-core chips, fault-tolerance in inherently unreliable technology must be addressed with radical and effective techniques. A complex network theory based interconnection architecture is a step in this direction. Theoretical studies in complex networks show that certain types of network connectivity are inherently more resilient to faults and failures [66]. Adopting novel architectures inspired by complex network theory in conjunction with the emerging interconnection technologies will enable design of high-performance, robust multi-core chips.

Each of the emerging interconnect technologies, vis., 3D integration, photonic NoCs or wireless/RF NoCs pose significant challenges related to their reliable integration. Vertical Through-Silicon-Via (TSV) is an enabling technology for 3D integration of chips. However, misalignment of layers during 3D stacking can result in TSV failure impairing the performance benefits of 3D NoCs. High power densities in 3D chips lead to thermal issues which may aggravate metal via failure rates. The reliable integration of silicon nanophotonic devices and waveguides to make Photonic NoCs a reality is a major challenge and is hence a subject of on-going research. NoCs with RF interconnects need laying long on-chip transmission lines across the chip along with a bank of precision, high frequency filters and oscillators. Design of such

high-precision analog components is non-trivial. Wireless interconnects with either mm-wave metal antennas or Carbon Nanotube (CNT) based on-chip antennas may encounter significant failure rates pertaining to issues of integration and transceiver design respectively. NoCs using these emerging interconnects demand high performance from inherently unreliable technology. With technology scaling due to shrinking device geometries in the future this issue can be predicted to even increase in importance. Hence, traditional fault-tolerance techniques like adaptive routing strategies [67] and error control coding (ECC) [68] will not be sufficient to address these issues specially with technology scaling.

Such challenges in reliability and integration demand radically different approaches to make these emerging interconnect paradigms viable for large-scale adoption. Natural complex networks often demonstrate surprising robustness against high degree of malfunctions, viz., microbes are known to persist and reproduce even in the presence of harsh external interferences. Large networks having a connectivity structure known as Scale-Free graphs are characterized by a few highly connected nodes and many peripheral nodes with very few connections. Under these conditions random faults result mostly in failure of those nodes which have very few connections as they occur in large majority. However failure of these nodes only marginally affects the entire network due to their relatively few connections. On the other hand these networks are very vulnerable to preferential failures of the important nodes which are highly connected. In contrast, a connectivity pattern known as Small-World graphs are characterized by near equal connectivity of all nodes. These networks are consequently similar in performance to random as well as preferential failures. Hence, depending on the failure patterns of the emerging interconnect technologies network architectures inspired from complex network theory can be designed which may provide inherent reliability against such faults.

The goal of this research would be to explore reliable NoC architectures using emerging interconnect paradigms. Architectures inspired from scale-free or small-world graphs will have different performance characteristics in presence of various failure patterns. Development of failure models for these emerging interconnects will be undertaken. Using these failure models an extensive study of the impact of interconnect failures on the performance of interconnect infrastructures for multi-core chips will be undertaken. From this systematic exploration inherently fault resilient multi-core architectures can be identified, which will have negligible or marginal effects due to interconnect faults. Furthermore, novel NoC architectures depending upon application-specific workloads with high levels of error resilience will be developed based on complex network theory. These novel fault-tolerant NoCs will be compared for performance with more traditional fault-tolerant techniques based on adaptive routing and ECCs to establish the relative advantages of the proposed complex network-based approach. As an extension, conventional fault-tolerance techniques will be implemented in this environment which will further enhance the performance of the emerging NoCs in the face of inherent failures related to technology.

References

[1] P. Magarshack and P.G. Paulin, “System-on-Chip beyond the Nanometer Wall,” Proceedings of Design Automation Conference (DAC 03), ACM Press, 2003, pp. 419-424.

[2] L. Benini and G. De Micheli, “Networks on Chips: A New SoC Paradigm,” IEEE Computer, Jan. 2002, pp. 70-78.

[3] Tilera Corporation, www.tilera.com

[4] R. Ho, K. W. Mai, M.A. Horowitz, “The Future of Wires”, Proceedings of the IEEE, Vol. 89 Issue: 4, April 2001 pp. 490–504.

[5] ITRS 2007, http://www.itrs.net/Links/2007ITRS/Home2007.htm

[6] W. J. Dally, B. Towles, “Route Packets, Not Wires: On-chip Interconnection Networks”, Proceedings of Design and Automation Conference (DAC 01), ACM Press, 2001, pp. 684-689.

[7] 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.

[8] 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

[9] 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.

[10] V. F. Pavlidis and E. G. Friedman, “3-D Topologies for Networks-on-Chip,” IEEE Transactions on Very Large Scale Integration (VLSI), Vol. 15, Issue 10, October 2007, pp. 1081-1090.

[11] 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.

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

[13] R. Albert and A.-L. Barabasi. “Statistical mechanics of complex networks,” Reviews of Modern Physics, 74:47–97, January 2002.

[14] M. Buchanan. “Nexus: Small Worlds and the Groundbreaking Theory of Networks.” Norton, W. W. & Company, Inc, 2003.

[15] C. Teuscher, “Nature-Inspired Interconnects for Self-Assembled Large-Scale Network-on-Chip Designs,” Chaos, 17(2):026106, 2007.

[16] B. Razavi, “Design of millimeter-wave CMOS radios: a tutorial”, IEEE Trans. Circuits Syst. I, vol. 56, no. 1, Nov. 2009, pp. 4-16.

[17] M. F. Chang et al., “RF Interconnects for Communications On-Chip”, Proceedings of International Symposium on Physical Design, 13-16 April 2008, pp. 78-83.

[18] A. Ganguly et al., “Scalable Hybrid Wireless Network-on-Chip Architectures for Multi-Core Systems”, IEEE Transactions on Computers (TC), vol. 60, issue 10, 2011, pp. 1485-1502.

[19] R. Marculescu et al., “Outstanding Research Problems in NoC Design: System, Microarchitecture, and Circuit Perspectives”, IEEE Transaction on Computer-Aided Design of Integrated Circuits and Systems, vol. 28, no. 1, January 2009, pp. 3-21.

[20] K. Chang et al., “Performance Evaluation and Design Trade-Offs for Wireless Network-on-Chip Architectures”, accepted for publication in ACM Journal on Emerging Technologies

in Computing Systems (JETC).

[21] B. Feero and P. P. Pande, "Networks-on-Chip in a Three-Dimensional Environment: A Performance Evaluation", IEEE Transactions on Computers, Vol. 58, No. 1, January 2009, pp. 32-45.

[22] D. Park et al., “MIRA: A Multi-layered On-Chip Interconnect Router Architecture”, IEEE International Symposium on Computer Architecture, ISCA, 21-25 June 2008, pp. 251-261.

[23] W. R. Davis et al., “Demystifying 3D ICs: The pros and cons of going vertical.” IEEE Design and Test of Computers, Vol. 22, Issue 6, November-December. 2005, pp. 498-510

[24] A. W. Topol et al., “Three-dimensional integrated circuits,” IBM Journal of Research & Development. Vol. 50 No. 4/5 July/September 2006.

[25] A. P. Jose et al., “Pulsed Current-Mode Signaling for Nearly Speed-of-Light Intrachip Communication”, IEEE Journal of Solid-State Circuits, Vol. 41, No. 4, April 2006, pp. 772-780.

[26] R. T. Chang et al., “Near Speed-of-Light Signaling Over On-Chip Electrical Interconnects”, IEEE Journal of Solid-State Circuits, Vol. 38, No. 5, May 2003, pp. 834-838.

[27] I. O’Connor et al., “Systematic Simulation-Based Predictive Synthesis of Integrated Optical Interconnect”, IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol. 15, No. 8, August 2007, pp. 927-940.

[28] M. Briere et al., “System Level Assessment of an Optical NoC in an MPSoC Platform”, Proceedings of IEEE Design, Automation & Test in Europe Conference & Exhibition, DATE, 16-20 April, 2007, pp-1084-1089.

[29] 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.

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

[31] 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), May 2009, pp. 124-133.

[32] 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.

[33] D. Zhao et al., "Design of multi-channel wireless NoC to improve on-chip communication capacity", Proceedings of the fifth ACM/IEEE International Symposium on Networks-on-Chip 2011, pp. 177-184.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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. 217-228.

[40] D. DiTomaso et al., "iWise: Inter-router wireless Scalable Express Channels for Network-on-Chips (NoCs) Architecture", Proceedings of Annual Symposium of High Performance Interconnects (HOTI), 2011, pp. 11-18.

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

[42] T. Petermann and P. De Los Rios, "Physical realizability of small-world networks," Physical Review E, 73:026114, 2006.

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

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

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

[46] A. E. Eiben, and J. E. Smith, Introduction to Evolutionary Computing. Springer-Verlag, Berlin, Heidelberg, 2003.

[47] M. Sipper, Evolution of Parallel Cellular Machines: The Cellular Programming Approach. Springer-Verlag, Heidelberg, 1997.

[48] T. Jansen, and I. Wegener, “A comparison of simulated annealing with a simple evolutionary algorithm on pseudo-boolean functions of unitation”, Theoretical Computer Science, 386, 2007, pp. 73-93.

[49] 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, May 2002, pp. 543-552.

[50] 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.

[51] 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.

[52] 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.

[53] J. Duato et al., “Interconnection Networks – An Engineering Approach”, Morgan Kaufmann, 2002.

[54] J. Draper, et al., “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, 1997,1184-93.

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

[56] W. Stallings, “Data and Computer Communications”, Prentice Hall 2007.

[57] R. Holsmark, et al., “HiRA: A Methodology for Deadlock Free Routing in Hierarchical Networks on Chip”, In Proceedings of International Symposium on Network-on-Chip, 2009.

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

[59] Taiwan Semiconductor Manufacturing Company process technology: www.tsmc.com

[60] 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.

[61] 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.

[62] X. Yu et al., “Performance evaluation and receiver front-end design for on-chip millimeter-wave wireless interconnect”, in Proc. IEEE IGCC 2010.

[63] B.G. Lee et al., “Ultrahigh-Bandwidth Silicon Photonic Nanowire Waveguides for On-Chip Networks,” IEEE Photonics Technology Letters, vol. 20, no. 6, Mar. 2008, pp. 398-400.

[64] S. Deb et al., “Design of an Efficient NoC Architecture using Millimeter-Wave Wireless Links”, Proceedings of IEEE International Symposium on Quality Electronic Design (ISQED), 19th-21st March 2012.

[65] E. Dupont, M. Nicolaidis, P. Rohr, ‘’Embedded Robustness IPs for Transient-Error-Free ICs”, IEEE Design and Test of Computers, Volume 19, Issue 3, May-June 2002 pp: 54 – 68.

[66] R. Albert, H. Jeong and A. Barabási, “Error and Attack Tolerance of Complex Networks”, Nature, Vol. 406, July 2000, pp. 378-382.

[67] H. Zhu, P. P. Pande, C. Grecu, "Performance Evaluation of Adaptive Routing Algorithms for achieving Fault Tolerance in NoC Fabrics," Proceedings of 18th IEEE International Conference on Application-specific Systems, Architectures and Processors, ASAP 2007, July 9th - 11th, 2007.

[68] A. Ganguly, P. Pande and B. Belzer, “Crosstalk-Aware Channel Coding Schemes for Energy Efficient and Reliable NoC Interconnects”, IEEE Transactions on VLSI (VLSI) Vol. 17, No.11, November 2009, pp. 1626-1639.

Appendix A

Publications

Following is a list of publications published in reputed journals and conferences during the course of this research.

Book Chapters:

1. Partha Pratim Pande, Amlan Ganguly, **Sujay Deb** and Kevin Chang, Energy-Efficient Network-on-Chip Architectures for Multicore Systems, Handbook of Energy-Aware and Green Computing, Ishfaq Ahmad and Sanjay Ranka (Editors), Publisher: Chapman and Hall/CRC Press Taylor and Francis Group LLC.

Journals:

1. **Sujay Deb**, Kevin Chang, Amlan Ganguly, Xinmin Yu, Partha Pande, Deuk Heo, Benjamin Belzer, "Design of an Energy Efficient CMOS Compatible NoC Architecture With millimeter-wave Wireless Interconnects", IEEE Transactions on Computers (**TC**), *under review*.
2. **Sujay Deb**, Amlan Ganguly, Partha Pande, Benjamin Belzer, Deukhyoun Heo, "Wireless NoC as interconnection backbone for multicore chips: Promises and Challenges", IEEE Journal on Emerging and Selected Topics in Circuits and Systems (**JETCAS**), *under review*.
3. Kevin Chang, **Sujay Deb**, Amlan Ganguly, Xinmin Yu, Suman Prasad Sah, Partha Pande, Benjamin Belzer, Deukhyoun Heo, "Performance Evaluation and Design Trade-Offs for

Wireless Network-on-Chip Architectures", ACM Journal on Emerging Technologies in Computing Systems (**JETC**), *in press*.

4. Amlan Ganguly, Kevin Chang, **Sujay Deb**, Partha Pande, Benjamin Belzer, Christof Teuscher, "Scalable Hybrid Wireless Network-on-Chip Architectures for Multi-Core Systems", IEEE Transactions on Computers (**TC**), vol. 60, issue 10, 2011, pp. 1485-1502.

Conferences:

1. **Sujay Deb,** Kevin Chang, Xinmin Yu, Amlan Ganguly, Partha Pande, Deuk Heo and Benjamin Belzer, "CMOS Compatible Many-Core NoC Architectures with Multi-Channel Millimeter-Wave Wireless Links", *Submitted (Name of the conference omitted due to the requirement of blind review).*

2. **Sujay Deb**, Kevin Chang, Amlan Ganguly, Xinmin Yu, Christof Teuscher, Partha Pande, Deuk Heo and Benjamin Belzer, "Design of an Efficient NoC Architecture using Millimeter-Wave Wireless Links", Proceedings of IEEE International Symposium on Quality Electronic Design **(ISQED)**, 19th-21st March 2012.

3. Xinmin Yu, Suman Sah, **Sujay Deb**, Partha Pande, Benjamin Belzer, and Deukhyoun Heo, "A Wideband Body-Enabled Millimeter-Wave Transceiver for Wireless Network-on-Chip", Proceedings of IEEE International Midwest Symposium on Circuits and Systems (**MWSCAS**), 7th-10th August 2011, pp. 1-4.

4. **Sujay Deb**, Kevin Chang, Amlan Ganguly and Partha Pande, "Comparative Performance Evaluation of wireless and Optical NoC Architectures", Proceedings of IEEE International SOC Conference (**SOCC**), 27th-29th September 2010, pp. 487-492.

5. **Sujay Deb**, Amlan Ganguly, Kevin Chang, Partha Pande, Benjamin Belzer and Deuk Heo,

"Enhancing Performance of Network-on-Chip Architectures with Millimeter-Wave Wireless Interconnects", Proceedings of IEEE International Conference on Application-specific Systems, Architectures and Processors (**ASAP**), 7th – 9th July, 2010, pp. 73-80.