Networks-on-Chip in Emerging Interconnect Paradigms: Advantages and Challenges

Luca P. Carloni
Columbia University
luca@cs.columbia.edu

Partha Pande
Washington State University
pande@eecs.wsu.edu

Yuan Xie
Pennsylvania State University
yuanxie@cse.psu.edu

Abstract

Communication plays a crucial role in the design and performance of multi-core systems-on-chip (SoCs). Networks-on-chip (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 demand for energy-efficient and high-performance interconnect fabrics, which continues to grow with each new process generation. Multiple options have been envisioned as compelling alternatives to the existing planar metal/dielectric communication structures. In this paper we outline the opportunities and challenges associated with three emerging interconnect paradigms: three-dimensional (3-D) integration, nanophotonic communication, and wireless interconnects.

1 Introduction

The current trend in SoC design in the ultra deep sub-micron (UDSM) regime and beyond is to integrate a large number of functional and storage cores onto a single die [1]. The possibility of this enormous degree of integration gives rise to new challenges in designing the interconnection infrastructure for these complex SoCs. Extrapolating from the existing CMOS scaling trends, traditional on-chip interconnect systems have been projected to be limited in their ability to meet the performance needs of SoCs at the UDSM technology nodes and beyond. Wire delays that span large fractions of the chip do not scale as well as local wire delays with respect to gate delays [2] and global interconnect has an increasing impact on the performance of the overall SoC [3] as well as on the effectiveness of well-established design methodologies and CAD flows [4]. While copper and low-k dielectrics have been introduced to decrease the global interconnect delay, they only extend the lifetime of conventional interconnect systems by a few technology generations. According to the International Technology Roadmap for Semiconductors (ITRS), material innovation with traditional scaling will no longer satisfy the performance requirements in the long term and radically new interconnect paradigms are needed. The continued progress of interconnect performance will require approaches that introduce materials and structures beyond the conventional metal/dielectric system, and may require information carriers other than charge. Multiple options have been envisioned to provide alternatives to the metal/dielectric system. In particular, three emerging interconnect technologies are three-dimensional (3-D) integration, nanophotonic communication, and RF/wireless interconnects. Meanwhile, networks-on-chip (NoC) have been proposed as a promising solution to structure the design of the on-chip communications infrastructure and enable a high degree of integration in multi-core SoCs [5–7] In the following sections we discuss the opportunities and challenges associated with the possibility of building NoCs with these emerging interconnect technologies.

2 3D NoC

Three-dimensional integrated circuits (3D ICs) [8] offer an attractive solution for overcoming the barriers to interconnect scaling, thereby offering an opportunity to continue performance improvements using CMOS technology, with smaller form factor, higher integration density, and the support for the realization of mixed-technology chips. Among several 3D integration technologies [9], TSV (Through-Silicon-Via) approach is the most promising one and therefore is the focus of the majority of 3D integration R&D activities [8]. Even though both 3D integrated circuits and NoCs [10, 11] are proposed as alternatives for the interconnect scaling demands, the challenges of combining both approaches to design three-dimensional NOCs have not been addressed until recently [12–19]. This section gives a brief introduction on the exploration of possible architectural designs for three-dimensional NoC architectures, and discuss the tradeoffs among various design options.

Symmetric NoC Router Design. The natural and simplest extension to the baseline NoC router to facilitate a

978-1-4244-4143-3/09/$25.00 ©2009 IEEE

Crossbar Type	Area	Power (500 Mhz)
5×5 Crossbar	8523 μm^2	4.21 mW
6×6 Crossbar	11579 μm^2	5.06 mW
7×7 Crossbar	17289 μm^2	9.41 mW

Table 1. Area and power comparison of the crossbar switches in a 90nm technology.

3D layout is simply adding two additional physical ports to each router; one for Up and one for Down, along with the associated buffers, arbiters (VC arbiters and Switch Arbiters), and crossbar extension. We can extend a traditional NoC fabric to the third dimension by simply adding such routers at each layer (called a symmetric NoC, due to symmetry of routing in all directions). We call this architecture a 3D Symmetric NoC, since both intra- and inter-layer movement bear identical characteristics as hop-by-hop traversal. For example, moving from the bottom layer of a 4-layer chip to the top layer requires 3 network hops.

This architecture, while simple to implement, has two major inherent drawbacks: (1) It wastes the beneficial attribute of a negligible inter-wafer distance in 3D chips (for example, the thickness of a die could be as small as 10s of μm). Since traveling in the vertical dimension is multi-hop, it takes as much time as moving within each layer. Of course, the average number of hops between a source and a destination does decrease as a result of folding a 2D design into multiple stacked layers, but inter-layer and intra-layer hops are indistinguishable. Furthermore, each flit must undergo buffering and arbitration at every hop, adding to the overall delay in moving up/down the layers; (2)The addition of two extra ports necessitates a larger 7×7 crossbar. Crossbars scale upward very inefficiently, as illustrated in Table 1. This table includes the area and power budgets of all crossbar types investigated in this section, based on synthesized implementations in a 90nm technology. Clearly, a 7×7 crossbar incurs significant area and power overhead over all other architectures. Therefore, the 3D Symmetric NoC implementation is a somewhat naive extension to the baseline 2D network.

3D NoC-Bus Hybrid Router Design. There is an inherent asymmetry in the delays in a 3D architecture between the fast vertical interconnects and the horizontal interconnects that connect neighboring cores due to differences in wire lengths (a few tens of μm in the vertical direction as compared to a few thousands μm in the horizontal direction). Consequently, a symmetric NoC architecture with multi-hop communication in the vertical (inter-layer) dimension is not desirable.

Given the very small inter-layer distance, single-hop communication is, in fact, feasible. This technique revolves around the fact that vertical distance is negligible compared to intra-layer distances; the bus can provide single-hop

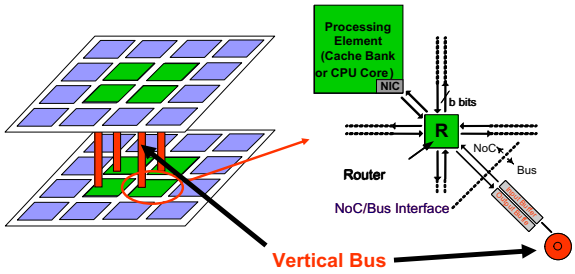

Figure 1. A hybrid 3D NoC/Bus architecture. The router has one additional input/output ports to connect with the vertical bus.

traversal between any two layers. This realization opens the door to a very popular shared-medium interconnect, the bus. The NoC router can be hybridized with a bus link in the vertical dimension to create a 3D NoC-Bus Hybrid structure, as shown in Fig. 1. This hybrid system provides both performance and area benefits. Instead of an unwieldy 7×7 crossbar, it requires a 6×6 crossbar (Fig. 1), since the bus adds a single additional port to the generic 2D 5×5 crossbar. The additional link forms the interface between the NoC domain and the bus (vertical) domain. The bus link has its own dedicated queue, which is controlled by a central arbiter. Flits from different layers wishing to move up/down should arbitrate for access to the shared medium.

Despite the benefits over the 3D Symmetric NoC router, the bus approach also suffers from a major drawback: it does not allow concurrent communication in the third dimension. Since the bus is a shared medium, it can only be used by a single flit at any given time. This severely increases contention and blocking probability under high network load. Therefore, while single-hop vertical communication does improve performance in terms of overall latency, inter-layer bandwidth suffers. More details on the 3D NoC-Bus hybrid architecture can be found in [12].

True 3D Router Design. Moving beyond the previous options, we can envision a true 3D crossbar implementation, which enables seamless integration of the vertical links in the overall router operation. Fig. 2 illustrates such a 3D crossbar layout. It should be noted at this point that the traditional definition of a crossbar - in the context of a 2D physical layout - is a switch in which each input is connected to each output through a single connection point. However, extending this definition to a physical 3D structure would imply a switch of enormous complexity and size (given the increased numbers of input- and output-port pairs associated with the various layers). Therefore, we chose a simpler structure which can accommodate the interconnection of an input to an output port through more than one connection points. While such a configuration can be viewed as a multi-stage switching network, we still call

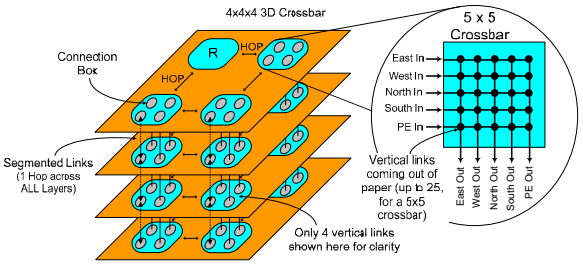

Figure 2. The true 3D router design.

this structure a crossbar for the sake of simplicity. The vertical links are now embedded in the crossbar and extend to all layers. This implies the use of a 5×5 crossbar, since no additional physical channels need to be dedicated for inter-layer communication.

As shown in Table 1, a 5×5 crossbar is significantly smaller and less power-hungry than the 6×6 crossbar of the 3D NoC-Bus Hybrid and the 7×7 crossbar of the 3D Symmetric NoC. Interconnection between the various links in a 3D crossbar would have to be provided by dedicated connection boxes at each layer. These connecting points can facilitate linkage between vertical and horizontal channels, allowing flexible flit traversal within the 3D crossbar.

The 2D crossbars of all layers are physically fused into one single three-dimensional crossbar. Multiple internal paths are present, and a traveling flit goes through a number of switching points and links between the input and output ports. Moreover, flits re-entering another layer do not go through an intermediate buffer; instead, they directly connect to the output port of the destination layer. For example, a flit can move from the western input port of layer 2 to the northern output port of layer 4 in a single hop.

However, despite this encouraging result, there is an opposite side to the coin which paints a rather bleak picture. Adding a large number of vertical links in a 3D crossbar to increase NoC connectivity results in increased path diversity. This translates into multiple possible paths between source and destination pairs. While this increased diversity may initially look like a positive attribute, it actually leads to a dramatic increase in the complexity of the central arbiter, which coordinates inter-layer communication in the 3D crossbar. The arbiter now needs to decide between a multitude of possible interconnections, and requires an excessive number of control signals to enable all these interconnections. A full crossbar with its overwhelming control and coordination complexity poses a stark contrast to this frugal and highly efficient design methodology. Moreover, the redundancy offered by the full connectivity is rarely utilized by real-world workloads, and is, in fact, design overkill [13].

Multi-layer 3D NoC Router Design. All the 3D router design options discussed earlier (symmetric 3D router, 3D NoC-Bus hybrid router, true 3D router, and 3D dimensionally-decomposed router) are based on the assumption that the processing element (PE) (which could be a processor core or a cache bank) itself is still a 2D design. For a fine-granularity design of 3D design, one can split a PE across multiple layers. For example, 3D cache design [20] and 3D functional units [21] have been proposed before. Consequently, a PE in the NoC architecture is possible to be implemented with such fine-granularity approach. Although such a multi-layer stacking of a PE is considered aggressive in the current technology, it could be possible as 3D technology matures with smaller TSV pitches.

With such a multi-layer stacking of processing elements in the NoC architecture, it is necessary to design a multi-layer 3D router that is designed to span across multiple layers of a 3D chip. Logically, such NoC architecture with multi-layer PEs and multi-layer routers is identical to the traditional 2D NoC case with the same number of nodes albeit the smaller area of each PE and router and the shorter distance between routers. Consequently, the design of a multi-layer router requires no additional functionality as compared to a 2D router and only requires distribution of the functionality across multiple layers.

The router modules can be classified into two categories - separable and non-separable, based on the ability to systematically split the module into smaller sub-modules across layers with the inter-layer wiring constraints and the need to balance area across layers [14]. Input buffers, crossbar, inter-router links are classified as separable modules, while arbitration logic and routing logic are classified as non-separable since they cannot be systematically broken into subsets. The saving in chip area can be used for enhancing the router capability, for example, adding express paths between non-adjacent PEs to reduce the average hop count, and help in boosting the performance and reducing the power. Furthermore, because a large portion of the communication traffic consists of short flits and frequent patterns, it is possible to dynamically shut down some layers of the multi-layer router to reduce the power consumption.

3D NoC Topology Design. All the router designs discussed in previous subsections are based on the mesh-based NoC topology. There exists various NoC topologies, such as concentrated mesh or flattened butterfly topology, all of which have advantages and disadvantages. By employing different topologies rather than the mesh topology, the router designs discussed above could also have different variants. For example, in 2D concentrated mesh topology, the router itself has a radix of 8 (i.e. an 8-port router, with four to local PEs and the others to four cardinal directions). With such topology, the 3D NoC-bus hybrid approach would result in a 9-port router design. Such high-radix router designs are power-hungry with degraded per-

formance, even though the hop count between PEs is reduced. Consequently, a topology-router co-design method for 3D NoC is desirable, so that the hop count between any two PEs and the radix of the 3D router design is as small as possible. Xu *et al.* [16] proposed a 3D NoC topology with low hop count (which is defined as low diameter) and low radix router design. The level 2D mesh is replaced with a network of long links connecting nodes that are at least m mesh-hops away, where m is a design parameter. In such a topology, long distance communications can leverage the long physical wire and vertical links to reach destination, achieving low total hop count, while the radix of the router is kept low. For application-specific NoC architecture, Yan *et al.* [15] also proposed a 3D-NoC synthesis algorithm that is based on a rip-up and reroute formulation for routing flows and a router merging procedure for network optimization to reduce the hop count.

Impact of 3D Technology on NoC Designs. Since TSV vias contend with active device area, they impose constraints on the number of such vias per unit area. Consequently, the NoC design should be performed holistically in conjunction with other system components such as the power supply and clock network that will contend for the same interconnect resources.

The 3D integration using TSV (through-silicon-via) can be classified into one of the two following categories; (1) *monolithic approach* and the (2) *stacking approach*. The first approach involves a sequential device process, where the frontend processing (to build the device layer) is repeated on a single wafer to build multiple active device layers before the backend processing builds interconnects among devices. The second approach (which could be wafer-to-wafer, die-to-wafer, or die-to-die stacking) processes each active device layer separately using conventional fabrication techniques. These multiple device layers are then assembled to build up 3D ICs using bonding technology. Dies can be bonded face-to-face (F2F) or face-to-back (F2B). The microbump in face-to-face wafer bonding does not go through a thick buried Si layer and can be fabricated with a higher pitch density. In stacking bonding, the dimension of the TSVs is not expected to scale at the same rate as feature size because alignment tolerance and thinned die/wafer height during bonding poses limitation on the scaling of the vias.

The TSV (or micropad) size, length, and the pitch density, as well as the bonding method (face-to-face or face-to-back bonding, SOI-based 3D or bulk CMOS-based 3D) can have a significant impact on the 3D NoC topology design. For example, relatively large size of TSVs can hinder partitioning a design at very fine granularity across multiple device layers, and make the true 3D router design less possible. On the other hand, the monolithic 3D integration provides more flexibility in the vertical 3D connection because the vertical 3D via can potentially scale down with feature size due to the use of local wires for connection. Availability of such technologies makes it possible to partition the design at a very fine granularity. Furthermore, face-to-face bonding or SOI-based 3D integration may have a smaller via pitch size and higher via density than face-to-back bonding or bulk CMOS based integration. Such influence of the 3D technology parameters on the NoC topology design should be thoroughly studied and suitable NoC topologies for different 3D technologies should be identified with respect to the performance, power, thermal, and reliability optimizations.

3 Nanophotonic Interconnection Networks

The combination of increasing requirements on on-chip and off-chip communication bandwidths and strict limitations on the maximum on-chip temperature and power budget imposed by packaging constraints will make the interconnect power consumption become perhaps the most critical problem for multi-core SoC design in the foreseeable future. How electronic NoCs can continue to satisfy future bandwidths and latency requirements within the chip power budget is an important open area of research [22].

Photonic networks-on-chip (NoC) have been proposed as a solution to reduce the impact of intra-chip and off-chip communication on the overall power budget [23, 24]. Thanks to the unique properties of optical communication, such as bit-rate transparency and low loss of optical waveguides, photonic NoCs are expected to reach levels of performance-per-watt scaling that cannot be matched by all-electronic interconnects. Recently various research projects have been started in both academia and industry to understand how to realize the potential of photonic NoCs. Shacham *et al.* have proposed a hybrid approach to photonic NoC [25,26], which is discussed in more detail below. Kırman *at al.* have studied the performance improvement obtained by using a CMOS-compatible photonic on-chip bus for future chip multi-processors (CMPs) [27]. Batten *et al.* have proposed power-constrained processor-memory network architectures for future many-core systems [28]. Vantrease *et al.* have presented a 3D many-core architecture that uses photonic communication for both inter-core communication and off-stack communication to memory or I/O devices [29]. Beausoleil *et al.* have made the case for a high-performance many-core computing system divided into multiple silicon compute clusters [30]. Krishnamoorthy *et al.* have provides a comprehensive overview on the opportunities of using photonic communication into a high-performance computing system at the chassis, chip-package and silicon micro-system levels [31].

Advances in Photonic NoC Components. The photonics opportunity is made possible now by recent breakthroughs in nanoscale silicon photonics and considerably improved photonic integration with commercial CMOS

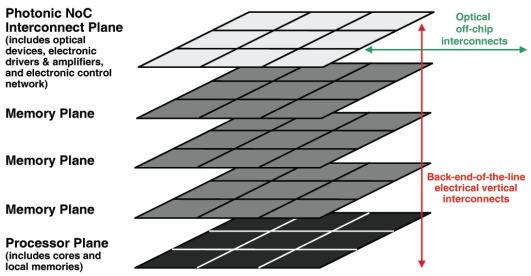

Figure 3. Possible organization of a SoC architecture combining 3DI and photonic NoC.

chip manufacturing [32]. The gains in power efficiencies for optical modulators and receivers are driven by the nanoscale device footprints and corresponding capacitances, as well as by the tight proximity of electronic drivers enabled by the monolithic CMOS platform integration. Advances in the fabrication of CMOS-compatible photonic devices have been reported in the literature for all the basic elements necessary to build photonic NoCs: links, modulators, switches, and receivers. In particular, thanks to design and fabrication improvements, sub-μm dimensional photonic links now achieve propagation losses around $1.7dB/cm$, and off-chip coupling losses of $0.5dB$/facet. Various NoCs functionalities can be implemented using the *micro-ring resonator*, a device of impressive versatility that can be used to build modulators, switches, and wavelength multiplexers. Micro-ring silicon modulators that are capable to reach modulation speeds in excess of $10Gbps$ with a dissipation of $85fJ/bit$ have been fabricated and their energy efficiency is expected to continue to improve in the future [33, 34]. Simple micro-ring switches have been demonstrated with throughput bandwidths of $250Gbps$ and are expected to scale to over $1Tbps$ [35, 36]. A four-port non-blocking photonic switch has been fabricated and characterized: it supports multi-wavelength routing through thermally tuned and stabilized micro-heaters [37]. Receivers based on SiGe or Ge photodetectors combined with high-gain CMOS amplifiers have been demonstrated to operate at $15Gbps$ with limited power dissipation [38].

Combining Photonic NoC and 3D Integration. In order to optimize the fabrication process it is reasonable to expect that the introduction of a photonic NoC in the design of future multi-core SoCs will take advantage of the progress in 3D Integration [39] and, particularly, the Through-Silicon-Via technology [40]. A possible organization of a future 3D SoC with a photonic NoC is illustrated in Fig. 3 and consists of three types of layers [26,41]: (a) a computational layer hosts multiple, possibly heterogeneous, processing cores together with their local memories and network interfaces; (b) one or more storage layers provide the bulk of on-chip memory; and (c) a communication layer host the optical components and opto-electronic devices that are combined to realize the photonic NoC and provide the main communication infrastructure to connect the cores among themselves and with off-chip memories and devices. This multi-layer organization will allow to separately optimize logic, memory, and Si photonics planes.

Hybrid Approach to Photonic NoC Design. Photonic NoCs can leverage two critical properties of the photonic medium:

- *bit-rate transparency:* Differently from electronic routers that must switch with every bit of the transmitted data, leading to a dynamic power dissipation that scales with the bit rate, photonic switches switch on and off once per message, and their energy dissipation is essentially independent from the bit rate. Hence, photonic NoCs can deliver very high bandwidth without incurring the levels of power dissipation that would be necessary for equivalent NoCs based on traditional electronic design.

- *low loss in optical waveguides:* At the chip and board scale the power that is dissipated on a photonic link is independent of the transmission distance. Energy dissipation remains essentially the same whether a message travels between two processing cores that are a few millimeters or a few centimeters apart. In fact, photonic NoCs based on low-loss off-chip interconnects would enable the seamless scaling of the optical communication infrastructure beyond the chip boundary to connect multiple multicore SoCs as well as DRAM memories.

While photonic technology offers unique advantages for energy-efficient high-bandwidth communication, its limitations in terms of computing and storage capabilities pose some critical challenges to the design of photonic NoCs. In particular, flit buffering and control-flit processing, two important functions of any packet-switched NoCs, are impractical to implement with optical devices. In order to address these challenges and leverage the best advantages provided by the electronic and photonic media respectively, Shacham *et al.* have proposed a *hybrid architecture* where a high-bandwidth circuit-switched photonic network is combined with a low-bandwidth packet-switched electronic network [25, 26]. While the electronic network carries small-size control (and data) packets, the photonic network transfers large-size data messages between pairs of cores. The NoC operates as follows: (1) a photonic circuit is reserved through the exchange of a path-setup packet over the electronic network between the source and the destination, followed by a short acknowledgment-pulse over the photonic network (path-setup process); (2) a large data transfer is completed on the high-bandwidth photonic circuit; and (3) at the end of the communication the photonic

circuit is released by the source through the transmission of a tear-down packet (path-teardown process).

A possible physical implementation of this photonic NoC and its performance are discussed in [42], which presented also the design of the 4-way ring-resonator photonic switch, whose prototype has been recently fabricated [37]. Alternative physical implementations, including two non-blocking topologies, have been proposed and analyzed by Petracca *et al.* [43], who presented also one of the first assessments of the expected benefits of using a photonic NoC for a real application, i.e. computing a very large Fast Fourier Transform. The simulation-based results confirm that on-chip photonic communication can support energy-efficient high-bandwidth data transfers among processing cores. Its potential can be realized especially by building photonic NoC that connect large multi-threaded cores or clusters of cores so that data aggregation can be performed to build and transfer large messages.

Challenges in Photonic NoC Design. Besides the need for continued advances in the design and the fabrication of the nanophotonic components, there are several issues that require further investigation. At the system-level more research is necessary to understand how to exploit most effectively the high-bandwidth and low-power connectivity offered by optical links to increase the performance of real applications. In particular, photonic NoCs could allow the seamless extension of this connectivity to off-chip devices such as other SoCs as well DRAM memories, thus potentially removing the off-chip communication bottleneck and enabling the rethinking of the overall system architecture. From a physical-layer viewpoint, photonic NoC designers face important challenges in looking for an optimal implementation of a given network architecture because, for instance, a certain layout could increase the aggregated insertion losses due to waveguide crossing [44]. Other important issues include the effect of scaling the number of wavelengths per link given a particular nonlinear power threshold within the optical waveguide and the effect of temperature variations on photonic components. In fact, from a computer-aided design (CAD) viewpoint there is a need for new design environments and tools that account for the many optical physical-layer aspects that have no electronic analogue.

4 Wireless NoC

A radical alternative to the existing metal/dielectric 2-D interconnect infrastructures is to use transmission of signals via RF/wireless interconnects. According to Chang et al. [45] intra-chip low latency communication can be achieved through RF interconnects, where transmission of data is guided through on-chip transmission lines [45]. Using Frequency Division Multiple Access (FDMA) with multiband frequency synthesizers and metal wires available

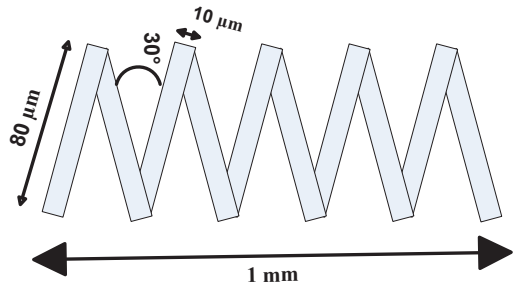

Figure 4. The mm-wave antenna configuration.

from current CMOS processes as transmission lines a high bandwidth RF interconnect can be created for on-chip data transport. In [46] a hybrid NoC architecture using such RF interconnects has been proposed. An underlying wireline mesh architecture overlaid with a high bandwidth FDMA based RF transmission is envisioned. The RF interconnect acts as an information "highway" enabling fast data transport across longer distances on the chip, thus reducing overall communication latency. In addition to reduced latency, the NoC power dissipation using overlaid RF shortcuts is demonstrated in [46]to be an order of magnitude less than that in traditional wireline NoCs. Miniaturized on-chip inductor based antennas and bank of high frequency precision oscillators and filters make such a hybrid NoC architecture viable. Unlike 3D and photonic NoCs, NoC with RF interconnects can be built using existing CMOS technology. But it requires laying of long on-chip transmission lines that serve as wave guides. To sustain high throughput, RF-interconnect based NoCs require multiple high frequency oscillators and high precision filters. In contrast to all these, the on-chip wireless communication network can be developed using existing and well-understood CMOS technology, and as elaborated later, it is capable of reducing the number of existing wired interconnects in the NoC. It also achieves on-chip effective speed of light signal propagation. By replacing multi-hop wireline links in a NoC through high-bandwidth single-hop long-range wireless channels the latency, power consumption and interconnect routing problems of a traditional NoC can be simultaneously addressed.

On-Chip Antenna. On-chip wireless interconnects were demonstrated first in [47] for distributing clock signals. Recently, design of a wireless NoC based on CMOS Ultra Wideband (UWB) technology was proposed [48]. The particular antennas used in [48] achieve a transmission range of 1 mm while being 2.98 mm in length. Consequently, for a NoC spreading typically over a die of 20 mm x 20 mm, this architecture essentially requires multi-hop communication through the on-chip wireless channels. Moreover, for 1 mm range of on-chip communication, a wireless link may not be more economical and efficient than metal wires. Having wireless nodes spread all over

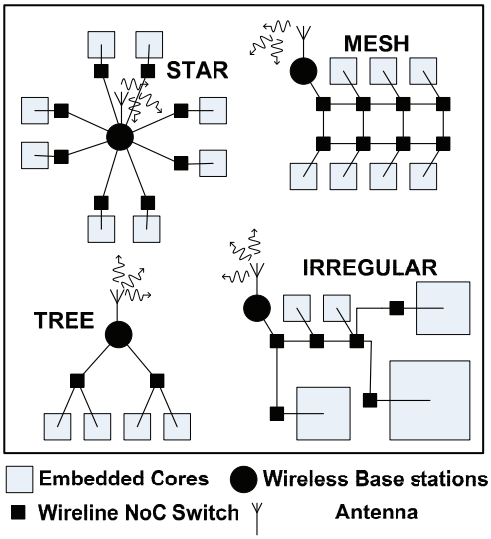

Figure 5. A hybrid wireless/wired NoC.

the die will introduce significant overhead due to antennas and associated transceiver circuits. This implementation achieves a peak bandwidth of 10 Gbps on a single channel for a system consisting of 16 embedded cores in the 0.18 um technology node. 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 [49]. They have primarily used metal zig-zag 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 [50]. It was shown that zig-zag monopole antennas of axial length 1-2 mm can achieve a communication range of about 10-15 mm. As explained in [49], antenna size can be reduced by using a monopole to utilize the virtual image below the ground plane to make it behave like a dipole with twice the length. A possible configuration of a zig-zag antenna is shown in Fig. 4.

Depending on antenna configuration and substrate characteristics, achievable frequency of the wireless channel can be in the range of 50-100 GHz. By varying the axial length, trace width, arm element length and bend angle the antenna bandwidth can be modified. Propagation of radio signal over intra-chip channels is mainly realized with surface waves guided on the air-wafer interface. A relatively long intra-chip communication range facilitates single-hop communication between widely separated blocks. This is essential to achieve the full benefit of on-chip wireless networks for multi-core systems by reducing long distance multi-hop wireline communication. Despite all these advantages, in the mm-wave range the antenna size (1-2 mm) is still a limitation. If the transmission frequencies can be increased to THz/optical range then the corresponding antenna sizes decrease, occupying much less chip real estate. One possibility is to use nanoscale antennas based on CNTs operating in the THz/optical frequency range [51–54]. Bundles of CNTs are predicted to enhance performance of antenna modules by up to 40dB in radiation efficiency and provide excellent directional properties in far-field patterns [55]. Moreover these antennas can achieve a bandwidth of around 500 GHz, whereas the antennas operating in the mm-wave range achieve a bandwidth of 10's of GHz. Thus antennas operating in the THz/optical frequency range can support much higher data rates. CNTs have several characteristics that make them suitable as on-chip antennas for optical frequencies. Given wavelengths of several hundreds of nanometers to several micrometers there is a need for virtually one-dimensional antenna structures for efficient transmission and reception. With diameters of a few nanometers and any length up to a few millimeters possible, CNTs are the perfect candidate. Such thin structures are almost impossible to achieve with traditional microfabrication techniques for metals. In CNTs, ballistic electron transport leads to quantum conductance, resulting in reduced resistive loss, which allows extremely high current densities, namely 4-5 orders of magnitude higher than copper. This enables high transmitted powers from nanotube antennas, crucial for long-range communications. By shining an external laser source on the CNT, radiation characteristics of multi-walled carbon nanotube (MWCNT) antennas are observed to be in excellent quantitative agreement with traditional radio antenna theory [52], although at much higher frequencies of hundreds of THz. The requirements of using external sources to excite the antennas can be eliminated if the electroluminescence phenomenon from a CNT is utilized to design linearly polarized dipole radiation sources [56]. It is demonstrated that, semiconducting CNTs emit infrared photons when an electron-hole pair recombines across the bandgap. Therefore, in addition to being the antenna, the CNT also acts as the signal generator, significantly reducing the transmitter circuit complexity. Consequently building an on-chip wireless interconnection network using optical frequencies for inter-core communications becomes feasible with much less overhead than the mm-wave antennas. But unlike the mm-wave antennas, CNTs face significant manufacturing challenges.

All these investigations regarding miniaturized antennas enable the design of novel wireless communication infrastructures for multi-core SoCs.

Network Architecture. The goal in on-chip communication system design is to transmit data with low latencies and high throughput using the least possible power and resources. Currently, the major challenges in wire-based traditional on-chip communication networks are the high latency and power consumption of their multi-hop links. By inserting single-hop long range wireless links in place of multi-hop wireline communication, overall system performance can be significantly improved. To do this the entire

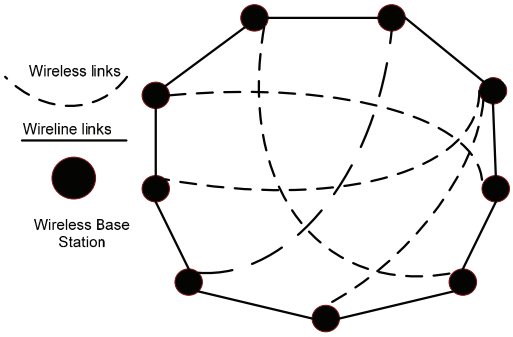

Figure 6. WiNoC with long-range wireless links.

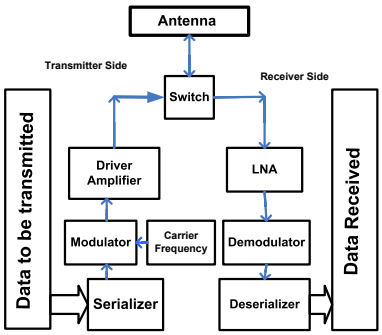

Figure 7. Components of mm-wave wireless transceiver.

network should be divided into multiple small clusters of neighboring cores called subnets. Wireless links will be introduced between the subnets, while intra-subnet communication will still be solely through wires. Each subnet is equipped with a wireless base station (WB), which transmits and receives data packets over the wireless channels. The advantage of this heterogeneous mode of data transport is that, as long as the antennas are placed within their communication range, only a single hop is required for inter-subnet communication, even if subnets are not adjacent. This also reduces the number of multi-hop wired links between distant cores. To reduce the antenna and wireless transceiver circuitry overhead, it is desirable to keep the number of WBs on a single chip as low as possible, without significantly compromising the performance benefit. This also helps reduce the load on the shared wireless medium. The optimal interconnect architecture and the number of subnets will vary depending on overall system size, mapping of the application on the entire chip, and the maximum number of WBs. Subnets will consist of relatively fewer cores, giving increased flexibility in designing their architectures. Instead of a single NoC spanning the entire system, as is traditional, there will be subnets with varying architectures for different parts of the chip. Fig. 5 shows a hybrid (wireless/wired) NoC architecture with heterogeneous subnets. Based on the research on small-world graphs [57], we can predict where to insert the wireless links in the network to improve performance. As shown in Fig. 6, we can view the wireless NoC (WiNoC) as a combination of clustered WBs with short-range wired links and a few long-range wireless links that produce shortcuts among the distant subnets. Allocating long range wireless links between distant subnets facilitates better utilization of limited wireless channels, as single-hop communication between distant subnets is the main contributing factor for overall performance gain. The existence of such links between the distant subnets basically helps in adopting the small-world phenomenon in the WiNoC. By incorporating the small-world network architecture through the wireless links, the latency and throughput profiles of the WiNoC are expected to be significantly improved.

Wireless Transceiver. The design of the wireless transceiver will depend on the specific frequency range of the wireless channels. For the WiNoC with mm-wave wireless links the necessary components of the wireless transceiver are shown in Fig. 7.

The design of the transceiver will be completely different when CNT antennas are used as the communicating frequency is in THz/optical range. The carrier has to be modulated by optical modulators. It is shown in [58] that high-speed silicon integrated Mach-Zehnder optical modulators are currently commercially available. Twenty four continuous wave laser sources of different frequencies and modulators operating at the rate of 10Gbps can be used. As noted in [59] this data rate is expected to increase many fold with technology scaling in future. Due to the very high frequency range, simple On-Off Keying (OOK) will be adopted.

The transceiver module will be much simplified if the electroluminescence phenomenon from a CNT is used to establish the communication channel. The transmitter will consist of a simple biasing circuit that, whenever the information bit to be transmitted is "1", applies a DC voltage across the CNT, which will accordingly generate the light signal, and at the same time act as the transmission antenna. The receiver will include a CNT antenna that generates a DC signal while the infrared signal is illuminating it, connected to a simple signal amplifier. In this case also depending on the bandgap of CNTs, multiple distinct frequency channels can be created. Moreover, it is found that, signals emitted by the nanotubes are polarized, with approximately a $\cos^2\theta$ dependence [56], where θ is the angle between the nanotube and the polarizer axis. This polarization dependence also exists in the case of absorption. This provides additional means of separating channels by polarization division multiplexing (PDM), using antennas positioned in various directions with respect to each other. The ultimate

		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	Direct point-to-point wireless links between smaller subnets
	Lower Power Dissipation	Shorter average path length	Negligible power dissipation in optical data transport	Multi-hop paths replaced by single hop links
Reliability		Vertical Via Failure	Temperature sensitivity of photonic components	Noisy wireless channel
Challenges		Heat dissipation due to higher power density, yield	Integration of on-chip photonic components	Low power mm-wave transceivers & Control over CNT growth

Table 2. Comparison of the three emerging interconnect paradigms.

effectiveness of this scheme will depend on the achievable communication range and signal power, which are yet to be fully characterized.

An important point to note here is that, the mm-wave metal antennas and the infrared CNT antennas are the enabling technologies to establish the on-chip wireless links. The characteristics of the on-chip wireless channels and hence the overall performance of the WiNoCs will depend on the specifics of these antennas, but the basic design principles of the WiNoC and its performance evaluation methodology remains the same. WiNoCs with mm-wave wireless links are more near-term solutions, which can be achieved with the help of existing CMOS technology. On-chip wireless communication links with CNT antennas will introduce much less overhead compared to the mm-wave ones, but it has to overcome several manufacturing challenges in the future.

5 Concluding Remarks

Three-dimensional integration, nanophotonic communication and on-chip wireless links are all promising alternative options to traditional planar metal/dielectric-based interconnects for building the communication infrastructure of future multi-core systems-on-chip. Each of these emerging interconnect paradigms, whose main features are summarized in Table 2, could offer remarkable advantages. 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.

Acknowledgments

The authors gratefully acknowledge the major contributions of their collaborators including B. Belzer, K. Chang, S. Deb, A. Ganguly, D. Heo, K. Bergman, A. Biberman, J. Chan, G. Hendry, J. Kash, B. Lee, M. Lipson, M. Petracca, A. Shacham, Y. Vlasov, C. Das, R. Das, J. Kim, C. Nicopoulos, D. Park, T. Richardson, and N.Vijaykrishnan.

This work is partially supported by the National Science Foundation (award numbers: 0702617 and 0811012) and the DARPA MTO office under grant ARL W911NF-08-1-0127.

References

[1] Y. Hoskote, S. Vangal, A. Singh, N. Borkar, and S. Borkar. A 5-GHz mesh interconnect for a teraflops processor. *IEEE Micro*, 27:51–61, September-October 2007.

[2] R. Ho, K. W. Mai, and M. A. Horowitz. The future of wires. *Proceedings of the IEEE*, 89(4):490–504, April 2001.

[3] J. A. Davis et al. Interconnect limits on gigascale integration (GSI) in the 21st century. *Proceedings of the IEEE*, 89(3):305–324, March 2001.

[4] L. P. Carloni and A. L. Sangiovanni-Vincentelli. Coping with latency in SoC design. *IEEE Micro*, 22(5):24–35, September/October 2002.

[5] A. Hemani et al. Network on chip: An architecture for billion transistor era. In *18th IEEE NorChip Conference*, November 2000.

[6] W. J. Dally and B. Towles. Route packets, not wires: On-chip interconnection networks. In *Proceedings of the Design Automation Conference*, pages 684–689, June 2001.

[7] L. Benini and G. De Micheli. Networks on chip: A new SoC paradigm. *IEEE Computer*, 49(2/3):70–71, January 2002.

[8] G. Philip, B. Christopher, and P. Ramm. *Handbook of 3D Integration*. Wiley-VCH, 2008.

[9] W. Davis et al. Demystifying 3D ICs: the pros and cons of going vertical. *IEEE Design and Test of Computers*, 22(6):498– 510, 2005.

[10] A. Jantsch and H. Tenhunen. *Networks on Chip*. Kluwer Academic Publishers, 2003.

[11] G. De Micheli and L. Benini. *Networks on Chips*. Morgan Kaufmann, San Francisco, CA, 2006.

[12] F. Li et al. Design and management of 3D chip multiprocessors using network-in-memory. In *Proceedings of International Symposium on Computer Architecture*, pages 130–141, Jun. 2006.

[13] J. Kim et al. A novel dimensionally-decomposed router for on-chip communication in 3D architectures. In *Proceedings of International Symposium on Computer Architecture*, Jun. 2007.

[14] D. Park et al. MIRA: A multi-layered on-chip interconnect router architecture. In *Proceedings of International Symposium on Computer Architecture*, pages 251–261, Jun. 2008.

[15] S. Yan and B. Lin. Design of application-specific 3D networks-on-chip architectures. In *Proceedings of International Conference of Computer Design*, pages 142–149, Oct. 2008.

[16] Y. Xu et al. A low-radix and low-diameter 3D interconnection network design. In *Proceedings of International Symoposium on High Performance Computer Architecture*, pages 30–41, Feb. 2009.

[17] I. Loi, F. Angiolini, and L. Benini. Developing mesochronous synchronizers to enable 3D NoCs. In *Proceedings of Design, Automation and Test in Europe Conference*, pages 1414–1419, Apr. 2008.

[18] I. Loi, S. Mitra, T. H. Lee, S. Fujita, and L. Benini. A low-overhead fault tolerance scheme for tsv-based 3D network on chip links. In *Proceedings of International Conference on Computer-Aided Design*, pages 598–602, Nov. 2008.

[19] V. F. Pavlidis and E. G. Friedman. 3-D topologies for networks-on-chip. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 15(10):1081–1090, 2007.

[20] Y. Tsai, F. Wang, Y. Xie, N. Vijaykrishnan, and M. Irwin. Design space exploration for three- dimensional cache. *IEEE Transactions on Very Large Scale Integration Systems*, 16(4):444 – 455, 2008.

[21] Y. Xie, G Loh, B Black, and K. Bernstein. Design space exploration for 3D architectures. *ACM Journal of Emerging Technologies in Compuing Systems*, (2):65–103, 2006.

[22] J. D. Owens et al. Research challenges for on-chip interconnection networks. *IEEE Micro*, 27(5):96–108, Sept.-Oct. 2007.

[23] A. Shacham, K. Bergman, and L. P. Carloni. Maximizing GFLOPS-per-Watt: High-bandwidth, low power photonic on-chip networks. In *Third Watson Conference on Interaction between Architecture, Circuits, and Compilers ($P = ac^2$)*, September 2006.

[24] N. Kırman et al. Leveraging optical technology in future bus-based chip multiprocessors. In *Intl. Symp. on Microarchitecture (MICRO)*, pages 492–503, December 2006.

[25] A. Shacham, K. Bergman, and L.P. Carloni. On the design of a photonic network-on-chip. In *Proceedings of the The First International Symposium on Networks-on-Chips (NOCS)*, May 2007.

[26] A. Shacham, K. Bergman, and L. P. Carloni. Photonic networks-on-chip for future generations of chip multi-processors. *IEEE Transactions on Computers*, 57(9):1246–1260, September 2008.

[27] N. Kırman et al. On-chip optical technology in future bus-based multicore designs. *IEEE Micro*, 27(1):56–66, 2007.

[28] C. Batten et al. Building manycore processor-to-dram networks with monolithic silicon photonics. In *16th IEEE Symp. on High Performance Interconnects*, pages 21–30, August 2008.

[29] D. Vantrease et al. Corona: System implications of emerging nanophotonic technology. In *35th Intl. Symp. on Computer Architecture*, pages 153–164, June 2008.

[30] Beausoleil et al. A nanophotonic interconnect for high-performance many-core computation. In *16th IEEE Symp. on High Performance Interconnects*, pages 182–189, August 2008.

[31] A.V. Krishnamoorthy et al. Optical interconnects for present and future high-performance computing systems. In *16th IEEE Symp. on High Performance Interconnects*, pages 175–177, August 2008.

[32] C. Gunn. CMOS photonics for high-speed interconnects. *IEEE Micro*, 26(2):58–66, March/April 2006.

[33] Q. Xu, S. Manipatruni, B. Schmidt, J. Shakya, and M. Lipson. 12.5 Gbit/s carrier-injection-based silicon microring silicon modulators. *Optics Express*, 15(2):430–436, 22 January 2007.

[34] M.R. Watts, D.C. Trotter, R.W. Young, and A. L. Lentine. Ultralow power silicon microdisk modulators and switches. In *2008 5th IEEE International Conference on Group IV Photonics*, pages 4–6, September 2008.

[35] A. Biberman, B.G. Lee, P. Dong, M. Lipson, and K. Bergman. 250 gb/s multi-wavelength operation of microring resonator-based broadband comb switch for silicon photonic networks-on-chip. In *34th European Conference on Optical Communication (ECOC)*, pages 1–2, September 2008.

[36] Y. Vlasov, W. M. J. Green, and X. Fengnian. High-throughput silicon nanophotonic wavelength-insensitive switch for on-chip optical networks. *Nature Photonics*, pages 2:242–246, March 2008.

[37] B.G. Lee, A. Biberman, K. Bergman, N. Sherwood-Droz, and M. Lipson. Multi-wavelength message routing in a non-blocking four-port bidirectional switch fabric for silicon photonic networks-on-chip. In *Optical Fiber Communications Conf. (OFC)*, March 2009.

[38] S.J. Koester et al. Ge-on-SOI-Detector/Si-CMOS-Amplifier receivers for high-performance optical-communication applications. *Journal of Lightwave Technology*, 25(1):46–57, January 2007.

[39] W. Haensch. Is 3D the next big thing in microprocessors? In *Intl. Solid State Circuits Conf.*, February 2007.

[40] K Bernstein et al. Interconnects in the third dimension: Design challenges for 3D ICs. In *Proceedings of the Design Automation Conference*, pages 562–567, 2007.

[41] J.A. Kash. Intrachip optical networks for a future supercomputer-on-a-chip. In *Photonics in Switching*, pages 55–56, August 2007.

[42] A. Shacham, B. G. Lee, A. Biberman, K. Bergman, and L. P. Carloni. Photonic noc for dma communications in chip multiprocessors. In *IEEE Symposium on High-Performance Interconnects*, August 2007.

[43] M. Petracca, B. G. Lee, K. Bergman, and L. P. Carloni. Design exploration of optical interconnection networks for chip multiprocessors. In *IEEE Symposium on High-Performance Interconnects*, pages 31–40, Stanford University, CA, August 2008.

[44] J. Chan, A. Biberman, B.G. Lee, and K. Bergman. Insertion loss analysis in a photonic interconnection network for on-chip and off-chip communications. In *Proc. of the 21st Ann. Mtg. of the IEEE Lasers & Electro-Optics Society (LEOS)*, November 2008, paper TuT3.

[45] M. F. Chang, E. Socher, S. W. Tam, J. Cong, and G. Reinman. RF interconnects for communications on-chip. In *International Symposium on Physical Design*, pages 78–83, April 2008.

[46] M. F. Chang et al. CMP network-on-chip overlaid with multi-band RF-Interconnect. In *IEEE International Symposium on High-Performance Computer Architecture*, pages 191–202, February 2008.

[47] B. A. Floyd, C. M. Hung, and K. K. O. Intra-chip wireless interconnect for clock distribution implemented with integrated antennas, receivers and transmitters. *IEEE Journal of Solid-State Circuits*, 37(5):543–552, May 2002.

[48] D. Zhao and Y. Wang. SD-MAC:Design and synthesis of a hradware-efficient collision-free QoS-aware MAC protocol for wireless network-on-chip. *IEEE Transactions on Computers*, 57(9):1230–1245, September 2008.

[49] J. Lin et al. Communication using antennas fabricated in silicon integrated circuits. *IEEE Journal of Solid-State Circuits*, 42(8):1678–1687, August 2007.

[50] Y. P. Zhang, Z. M. Chen, and M. Sun. 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*, 55(10):2900–2906, October 2007.

[51] G. Y. Slepyan, M. V. Shuba, S. A. Maksimenko, and A. Lakhtakia. Theory of optical scattering by achiral carbon nanotubes and their potential as optical nanoantennas. *Physical Review B (Condensed Matter and Materials Physics)*, 73:195416 (11), 2006.

[52] K. Kempa et al. Carbon naotubes as optical antennae. *Advanced Materials*, 19:421–426, 2007.

[53] M. Freitag et al. Hot carrier electroluminescence from a single carbon naotube. *Nano Letters*, 4(6):1063–1066, 2004.

[54] P. J. Burke, S. Li, and Z. Yu. Quantitative theory of nanowire and nanotube antenna performance. *IEEE Transactions on Nanotechnology*, 5(4):314–334, July 2006.

[55] Y. Huang, W. Y. Yin, and Q. H. Liu. Performance prediction of carbon nanotube bundle dipole antennas. *IEEE Transactions on Nanotechnology*, 7(3):331–337, May 2008.

[56] J. A. Misewich et al. Electrically induced optical emission from a carbon nanotube FET. *Science*, 300:783–786, 2003.

[57] 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*, 14(7):693–706, July 2006.

[58] W. M. J. Green, M. J. Rooks, L. Sekaric, and Y. A. Vlasov. Ultra-compact, low RF power, 10 Gb/s silicon Mach-Zehnder modulator. *Optics Express*, 15(25):17106–17113, December 2007.

[59] B.G. Lee et al. Ultrahigh-bandwidth silicon photonic nanowire waveguides for on-chip networks. *IEEE Photonics Technology Letters*, 20(6):398–400, March 2008.