Journal of
Low Power Electronics and Applications

MDPI

Article

The Advances, Challenges and Future Possibilities of Millimeter-Wave Chip-to-Chip Interconnections for Multi-Chip Systems

Amlan Ganguly [1,*], M. Meraj Ahmed [1], Rounak Singh Narde [2], Abhishek Vashist [1], Md Shahriar Shamim [3], Naseef Mansoor [1], Tanmay Shinde [2], Suryanarayanan Subramaniam [2], Sagar Saxena [1], Jayanti Venkataraman [2] and Mark Indovina [2]

1 Department of Computer Engineering, Rochester Institute of Technology, Rochester, NY 14623, USA; ma9205@rit.edu (M.M.A.); av8911@rit.edu (A.V.); nxm4026@rit.edu (N.M.); ss6010@rit.edu (Sa.S.)
2 Department of Electrical and Microelectronic Engineering, Rochester Institute of Technology, Rochester, NY 14623, USA; rn5949@rit.edu (R.S.N.); tvs6972@rit.edu (T.S.); ss4050@rit.edu (Su.S.); jnveee@rit.edu (J.V.); maieee@rit.edu (M.I.)
3 Intel Corp., Hillsboro, OR 97124, USA; shahriar.shamim@intel.com
* Correspondence: axgeec@rit.edu; Tel.: +1-585-475-4082

Received: 21 December 2017; Accepted: 26 February 2018; Published: 28 February 2018

Abstract: With aggressive scaling of device geometries, density of manufacturing faults is expected to increase. Therefore, yield of complex Multi-Processor Systems-on-Chips (MP-SoCs) will decrease due to higher probability of manufacturing defects especially, in dies with large area. Therefore, disintegration of large SoCs into smaller chips called *chiplets* will improve yield and cost of complex platform-based systems. This will also provide functional flexibility, modular scalability as well as the capability to integrate heterogeneous architectures and technologies in a single unit. However, with scaling of the number of chiplets in such a system, the shared resources in the system such as the interconnection fabric and memory modules will become performance bottlenecks. Additionally, the integration of heterogeneous chiplets operating at different frequencies and voltages can be challenging. State-of-the-art inter-chip communication requires power-hungry high-speed I/O circuits and data transfer over long wired traces on substrates. This increases energy consumption and latency while decreasing data bandwidth for chip-to-chip communication. In this paper, we explore the advances and the challenges of interconnecting a multi-chip system with millimeter-wave (mm-wave) wireless interconnects from a variety of perspectives spanning multiple aspects of the wireless interconnection design. Our discussion on the recent advances include aspects such as interconnection topology, physical layer, Medium Access Control (MAC) and routing protocols. We also present some potential paradigm-shifting applications as well as complementary technologies of wireless inter-chip communications.

Keywords: wireless interconnect; multi-chip system; heterogeneous system; in-package memory; IoT

1. Introduction

For decades, the silicon industry has exploited Moore's law to satisfy the exponential growth in the functionality required for high-performance computing nodes such as servers and embedded systems. However, in recent years continuous scaling down of transistors and scaling up of operating frequency resulted in a drastic increase in the power consumption of Integrated Circuits (ICs). Hence, multicore chips or Multi-Processor System-on-Chips (MP-SoCs) come into play to take advantage of parallel execution on the same die without the need to scale up frequency [1,2]. In modern processors designed using advanced process technologies, the number of individual functional cores has increased to the

order of hundreds and Network-on-Chip (NoC) has emerged as a scalable, modular interconnection architecture for such large multicore chips [3]. However, it is important to note that for advanced process nodes, different factors such as sub-wavelength lithography, line edge roughness, and random dopant fluctuation can cause a wide process variation, which can result in higher fault density [4] and hence, lower yield. Therefore, disintegration of large and complex multicore processors into smaller chips called, chiplets is used to alleviate the effect of higher fault densities in advanced technology nodes [5]. This is because the disintegration will decrease the area of individual chips and therefore, improve yield of the individual chips. Table 1 shows the increase in yield for implementing a system with increasing number of chips ranging from a single monolithic chip to 16 chips manufactured in a 300 mm round wafer. We consider a total of 64 cores in the system and a moderate fault density of $0.13/cm^2$ as reported by International Technology Roadmap for Semiconductors (ITRS) [6] for current process nodes. The yield has been calculated using the Murphy model [7] as shown in Equation (1):

$$y = \left[\frac{1 - e^{-AD}}{AD}\right]^2 \quad (1)$$

where Y is the yield, A is the die area, D is the defect density. Disintegrating a large chip with 64 cores into smaller chips or chiplets will result in steadily increasing yield. These smaller chips or chiplets are integrated onto a platform-based system and enable integration of chiplets from heterogeneous process technologies or functionalities. This in turn, offers both process and functional flexibility in the design while eliminating the design and manufacturing complexity of large SoCs.

Table 1. Yield improvement with chip size reduction.

Dies Per Package	Cores Per Die	Die Dimension (X × Y) (mm)	Good Die Per Wafer	Yield (%)
1	64	20 × 20	86	60.8
2	32	20 × 10	232	77.5
4	16	10 × 10	545	87.9
8	8	10 × 5	1186	93.7
16	4	5 × 5	2471	96.8

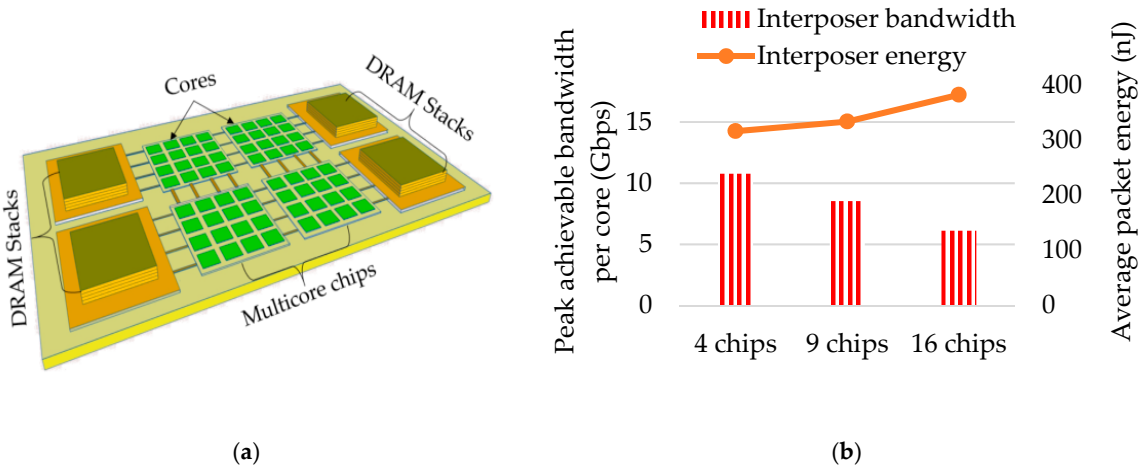
Multicore multi-chip computing modules with multiple processors or chiplets can be found in a wide range of platform-based designs from servers to embedded systems. An example of such multicore multi-chip module is the AMD EPYC series released in 2017 [8], which is a processor system designed for sophisticated servers. The EPYC Threadripper processor node is available as a 4-chip System-in-Package (SiP) with 8 cores in each chip, fabricated in 14 nm lithography technology. However, these new approaches require multiple chips to be interconnected efficiently to ensure desired performance with low cost. Moreover, the existing multi-chip systems can have processing chips such as multicore chips, CPUs, GPUs or a heterogeneous mix of such chips [9] (e.g., AMD's Fusion Accelerated Processor Units (APUs)) depending upon desired functionality. Applications running in a heterogeneous environment comprising of CPU and GPU chips require low latency and energy efficient memory access mechanism to ensure high performance. In such systems, different kinds of traffic interactions exist such as bandwidth-sensitive memory accesses and latency-sensitive control or cache-coherency messages. To ensure high performance for such multicore multi-chip system, a cache coherency aware interconnect is highly desirable to enable multicast/broadcast-based cache coherency/control message transmission without degrading system performance as even a small proportion of such messages can substantially impact the overall performance [10]. Moreover, due to scaling up of the number of individual chips in the multi-chip system, (homogeneous or heterogeneous) shared resources such as interconnect and memory subsystem are going to become a dominant bottleneck. As the number of multicore processors in the system increase, the average throughput per core in the system decreases. This is because, for a uniform random distribution

of hence it creates a bottleneck for the inter-chip interconnection fabric. The same phenomenon is observed for a decrease in localization of data packets within the same chip [11].

In traditional multi-chip platforms, inter-chip communication happens through C4 bumps or through Peripheral Component Interconnect (PCI) or PCI express (PCIe) which is a common local I/O bus standard. Recent trends according to the ITRS [6] predicts that the pitch of the wired I/O interconnects, solder bumps or pads in ICs is not scaling as fast as the gate lengths or pitch of on-chip interconnects. This implies a gap in density and performance of traditional I/O systems relative to on-chip interconnections. The wiring complexity of both on-chip and off-chip interconnects exacerbates the problem by posing design challenges, crosstalk, and signal integrity issues [12]. Moreover, typically intra-chip and inter-chip interconnections are designed separately to provide design flexibility and modular design approach. However, switching between protocols is necessary if the off-chip communication protocol is different from the on-chip one incurring overheads in inter-chip data transfer. Moreover, each of the heterogeneous components might operate at different voltages and frequencies and therefore will need voltage/frequency conversion while communicating with memory modules or other chips in the system. Therefore, we need an energy efficient, seamless, scalable interconnection network, which can address homogeneous, heterogeneous and memory integration challenges for different multi-chip environments/systems.

While metallic inter-chip interconnects are not scaling well, research in recent years have brought to light many emerging alternative interconnect solutions such as inter-chip photonics [13], vertically integrated monolithic 3D ICs [14] or silicon interposers [15,16] as solutions to the off-chip interconnection challenges. However, each interconnect technology has some challenges associated with them. The inability of the pitch scaling makes the adoption of photonic interconnects challenging for high complexity scalable multi-chip systems. Moreover, thermal tuning of on-chip electro-optic modulators can be power hungry and unreliable with thermal variations of the dies. In 3D integration of ICs, complex thermal management techniques are required to address the higher power dissipation densities in 3D ICs due to smaller footprints. These thermal management techniques vary from dynamic power management methods such as Dynamic Voltage Frequency Scaling (DVFS), temperature-aware task migration [17] or microfluidic cooling channels for better heat circulation [18]. Microfluidic channels, while very effective at cooling chips, need fluid intake pipes through the packaging making the whole system complex. Moreover, die thinning for fabrication of Through-Silicon-Vias (TSVs), particularly if co-existing with microfluidic channels, result in low yields. The use of silicon-interposers, which are themselves large dies with abundant wiring resources seem to emerge as a low-cost and high-yield alternative to photonic or monolithic 3D integration. However, the interposer-based integration methodology still engages metallic wires, which can incur high latency and power consumption over long distances affecting the scalability of the system [19]. To demonstrate the impact of increase in number of chips in a multi-chip system we have considered an example where, a total of 64 cores in the system are disintegrated into multiple multicore chips. The 64 cores are equally distributed among all the chips in the system resulting in three configurations with four 16-core chips, eight 8-core chips and sixteen 4-core chips. The cores within each multicore chip are interconnected with a 4×4 mesh and the inter-chip communication happens through a silicon interposer creating an extended mesh topology by connecting the switches on the boundaries of adjacent cores as shown in Figure 1a. Four DRAM stacks implementing a distributed memory subsystem is also considered. Each switch implements wormhole switching [20] with dimension order routing and is considered to have 4 Virtual Channels (VCs) per port and a buffer depth of 4 flits per VC. The wired link characteristics in each chip as well as in the interposer are modeled from global wire models with 0.2 pF/mm parasitic capacitance [21]. The switches are characterized from post-synthesis models using 65 nm technology node [22]. Cycle-accurate simulations for these systems, with a 20% memory access probability by all cores and an 80% uniform random traffic between all cores in all chips, are performed following the methodology outlined in detail in Section 5. All packets are broken down into flow control units (flits) and in our simulations each packet is considered to have 16 flits

where each flit is 64 bits wide. In Figure 1b we present the throughput per core and packet energy dissipation of an interposer-based multi-chip system with increasing number of multicore chips. It can be observed that the throughput per core decreases and packet energy increases with increase in the number of multicore chips. This indicates the need for an efficient inter-chip interconnection fabric. In addition, all these interconnect systems require physical links with wires or waveguides which require complex place and route algorithms for optimal designs. Moreover, photonic waveguides and rigid interposer platforms limit the range of motion between the individual ICs in the system. Wireless inter-chip interconnection systems will not have these limitations related to laying out waveguides or wired interconnects. Therefore, the wireless interconnect will enable system integration on flexible substrates for wearable electronic system.



(**a**) (**b**)

Figure 1. (**a**) Conceptual view of the interposer system (**b**) peak bandwidth and energy with respect to system size.

Research in recent years has demonstrated that on-chip and off-chip wireless interconnects can establish radio communications within as well as between multiple chips. On-chip antennas with multi GigaHertz bandwidths in millimeter-wave (mm-wave) bands, specifically, in the unlicensed 60 GHz band, are fabricated and demonstrated [23–26]. Using such on-chip antennas embedded in the chip [27] or waveguides [28] Wireless Network-on-Chip (WiNoC) architectures are shown to improve energy efficiency and bandwidth of on-chip data communication in multicore chips [29]. Recently, their feasibility in energy-efficient chip-to-chip data communication has also been investigated [30]. In addition to that, wireless interconnects do not require laying out physical interconnects eliminating additional place-and-route steps in the design process. Figure 2 shows a multi-chip system with multicore chips showing the propagation of the radiation for inter-chip wireless communication without the need for physical waveguides. The cores are interconnected using an intra-chip interconnection network and routers in the interconnection network are equipped with a wireless antenna and transceiver to enable the inter-chip wireless interconnects. In this paper, we will discuss and present the advances in design methodologies, challenges, and opportunities of such wireless interconnection architectures for of multi-chip systems. We present the state-of-the-art in design methodologies for topologies, physical layer, MAC, and routing protocols for wireless multi-chip systems. We will discuss challenges and opportunities in these aspects from a variety of perspectives, including design, electronic design automation (EDA), testing and qualitative benefits which are not easily measurable. Finally, we present some paradigm-shifting applications that might be enabled by high-speed wireless inter-chip communication and the implications of developments of complementary technologies such as Internet-of-Things (IoT), Industry 4.0, flexible electronic devices, and neuromorphic computing.

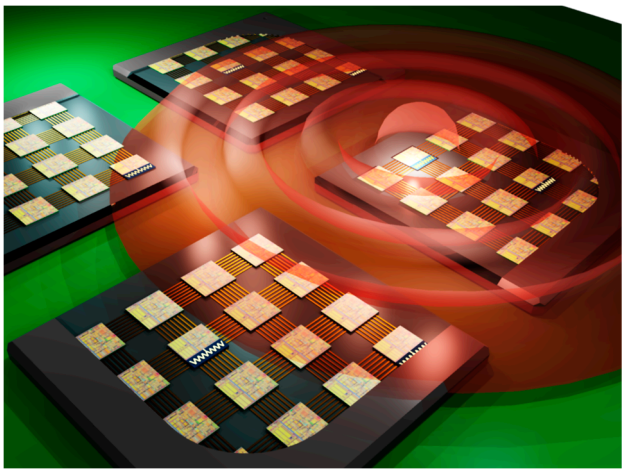


Figure 2. A conceptual view of a multichip system with inter-chip wireless communication.

2. Multi-Chip Interconnection Topology Design

Designing the wireless multi-chip interconnection topology essentially involves determining the intra-chip interconnection architecture for each chip as well as wireless interface (WI) placement for intra-chip or chip-to-chip communication. We first discuss the approaches for designing the intra-chip interconnections followed by the approaches for the inter-chip wireless communication.

2.1. The Intra-Chip Interconnection Design Approaches

Depending on the individual chips in the system, the intra-chip interconnection can be different and customized for the particular chips. For example, for small SoC chiplets with a few cores a bus-based interconnect mechanism might be sufficient. However, if each individual chip is large enough with several cores it might require a NoC for high performance and modular or scalable design. The intra-chip NoC should be carefully designed in a way that it avoids long multi-hop internal paths and therefore does not make the intra-chip network a bottleneck for the whole system. Therefore, it is desired to adopt an intra-chip topology that has small diameter with as well as high reliability. Keeping the small network diameter in mind, small-world architectures were proposed [31,32] and evaluated for intra-chip NoCs. Small world graphs are a type of complex network where both short distance as well as long distance links create a connected graph. It has both high local connectivity as well short diameters which scales well with size. While irregular topologies have better connectivity and diameter properties, unequal wire lengths in irregular topologies such as small world networks [32] make design, implementation, and verification very challenging. Therefore, to avoid complexity in design, verification and manufacturing of general-purpose multicore processors adopt regular intra-NoC architectures such as mesh, torus and folded torus [1,2].

As the performance of the multi-chip system also depends on the different traffic patterns, intra-NoC topology is chosen based on the support needed to handle such traffic. Therefore, for systems running memory intensive application a concentrated or hierarchical intra-chip topology is adopted to support multicast/broadcast-based cache coherent messages in the system. As a small amount of such broadcast traffic can significantly degrade system performance, adopting a concentrated, hierarchical or small-world topology for such traffic can improve the performance of the NoC compared to a regular mesh architecture [5,33]. While the intra-chip topology can have a regular, irregular, or concentrated architecture, different chiplets can implement different intra-chip architectures depending on the specific task or applications they are designed for. In a homogeneous system, each chip has the same intra-chip interconnection architectures as well as core architectures. Whereas in the heterogeneous system the intra-chip interconnection or core can have different architectures in different chips.

Various intra-chip NoC switch designs have been proposed to enable the variety of architectural needs of the intra-chip NoCs. These range from multi-cycle switch architectures [34] to low-latency single cycle ones [35]. Wide and ultra-wide switches enabling wide flit size for high throughputs have been investigated but are shown to be prohibitively power hungry for ultra-wide flit sizes of higher than 128 bits [36]. Heterogeneous, switch architectures altering between wide and low-latency designs are proposed for NoCs experiences heterogeneous traffic scenarios. Express virtual channels in switches have been proposed to bypass intermediate switches once path discovery has been accomplished with the help of routing have been shown to improve performance of NoCs [37]. A NoC router with power-gated WIs is proposed in [38] to reduce the power dissipation of the wireless transceivers when they are not used. However, as shown in [11] the contribution of the intra-chip interconnection for large multi-chip systems is marginal as throughput and latency of various architectures converge when the number of chips in the system increases. This is because when the system becomes large with many chips along with an increase in inter-chip traffic, the chip-to-chip interconnect becomes the main bottleneck instead of the intra-chip NoC. This makes the choice of the inter-chip communication architecture paramount in determining the overall performance and energy efficiency. Hence, in the subsequent section we will focus on the inter-chip wireless topology design.

2.2. The Inter-Chip Wireless Communication Topology

To alleviate the limitations of traditional inter-chip interconnects, WIs need to be deployed in the multicore chips to realize the inter-chip wireless interconnects. Deployment of the WIs in a wireless multicore multi-chip system is performed by answering two major questions namely, what are the appropriate locations of the WIs and what is an appropriate number of WIs for optimum performance? In the context of intra-chip WiNoCs, several approaches have been taken to deploy WIs optimally. Two-step nested optimization approaches were adopted in [39] where, in the first step, the location of the WIs were optimized for a varying number of WIs by using heuristics such as Simulated Annealing to optimize the average hop-count between cores in the NoC. In the second step, the throughput of each configuration corresponding to the specific number of WIs with optimal location was compared to choose the best performing configuration. Alternatively, optimizing Minimum Average Distance (MAD) of the WiNoCs to achieve optimal WI deployment was proposed in [40]. A Voltage-frequency Island (VFI)-aware wireless NoC design is proposed in [41]. In [42] a framework for using Machine Learning to deploy WIs in the intra-chip NoC is envisioned.

The key idea is to deploy the WIs in the chips in the multi-chip system in such a way that we achieve wireless deployment density that avoids long multi-hop paths between the internal cores and the WIs. For this purpose, we define the *wireless density* as the number of cores within each multicore chip that are serviced by a single WI. However, the exact WI deployment density is a function of various factors. One of the major factors is the choice of the physical layer and associated Medium Access Control (MAC) layer. For example, if non-interfering channels are possible by MAC mechanisms such as Frequency Division Multiple Access (FDMA) on separate non-interfering channels then it may be suitable to deploy a WI density to fully utilize all the channels possible using the specific physical layer technology. However, if a single frequency channel is shared among the WIs through MACs such as Time Division Multiple Access (TDMA) or Code Division Multiple Access (CDMA) then the WI density needs to be optimized. The optimization metric can be a metric of interest such as throughput, latency, or energy consumption. This optimization is necessary as a higher WI density will improve the connectivity of the fabric, which is expected to improve the performance. However, it will reduce the access time or share of the bandwidth for each WI due to more WIs sharing the channel. Further discussion on the dependence of the performance on the MAC and physical layers can be found later in Section 4.2. In case of mm-wave wireless inter-chip interconnects, the shared mm-wave physical layer can use a form of TDMA to obtain access to the shared wireless channel. Therefore, the WI density needs to be optimized such that each WI has a fair share of the channel access while providing a low average diameter or average hop-count between communicating cores. Figure 3,

which shows the peak achievable bandwidth and average packet energy as the WI density increases in a homogeneous 16-chip multi-chip system with each chip having 16 cores connected in a regular mesh topology. The intra-chip packet transfers occur through the intra-chip mesh wireline links whereas, the inter-chip data transfer utilizes the inter-chip mm-wave wireless links. The simulations for the optimization are performed considering both the intra-chip data transfer latency from the cores to the WIs as well as the latency between the WIs for the inter-chip transfer. The details of obtaining these models are described when discussing the simulation methodology in Section 5. The characteristics of the wired links are modeled from global wire characteristics [21] whereas the mm-wave inter-chip are adopted from [11] in 65 nm technology node. From Figure 3 it is observed that the multi-chip system has higher peak achievable bandwidth and low energy for a WI deployment density of 1/16 WIs per core. As each chip has 16 cores, this is the minimum density possible while still ensuring all chips are interconnecting using the wireless links. For higher WI density the system performance degrades due to larger starvation time. While Figure 3 gives an example of the optimization, the optimum WI density in each chip depends on the number of cores in each chip as well as the intra-chip topology adopted for each chip.

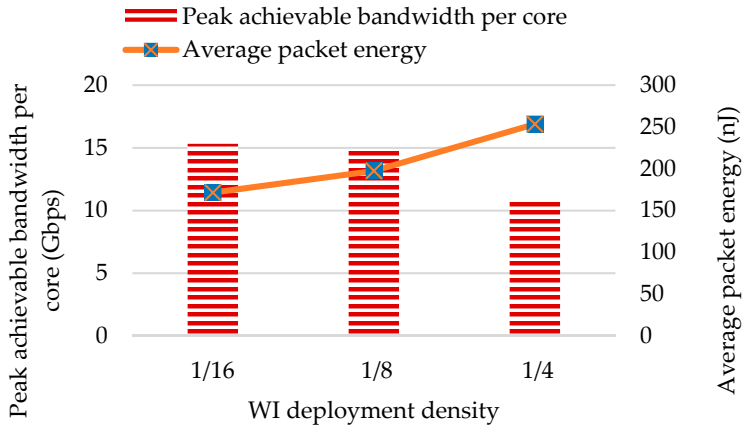


Figure 3. Multichip system performance vs. WI deployment density.

Optimization through system-level simulation as discussed above requires a priori knowledge of the expected traffic patterns of the multi-chip system. However, these systems are designed for embedded platforms or blade servers and are, therefore expected to cater to a wide variety of applications ranging from multimedia to parallel tasks characterized by benchmark suites like Parsec [43] or SPLASH-II [44]. Therefore, while a priori knowledge of precise traffic interactions between cores in multiple chips can result in better optimization, such knowledge may not always be available at design time, especially for general purpose platforms. On-chip traffic modeling has been done in [45,46]. However, in the absence of such knowledge about traffic in a multi-chip system, a uniform random traffic pattern can be assumed at design time, which is generic and captures the characteristics of both local as well long-distance traffic. This approach has been adopted in the design of intra-chip WiNoCs [31]. Alternatively, traffic-independent metrics such as average hop-count or zero-load latency, which is the average latency of a packet in the absence of any other packet in the interconnection network, can be adopted. Average zero-load latency is a more precise metric compared to average hop-count, but it does not capture delay due to traffic congestion in absence of traffic in the network. Therefore, the optimization method for the wireless interconnection fabric needs to be adopted depending on inter-dependent factors such as whether knowledge of traffic pattern is available at design-time, the adopted wireless MAC protocol, and the chosen optimization metric.

While the above method is suitable for a shared medium wireless interconnect topology, directional antennas with specific radiation patterns might govern the exact location of transmitting and receiving antennas to maximizing antenna transmission gains [47]. The challenging aspect of

that design is a rigorous analysis of wireless transmission and channel characteristics at relevant frequencies to determine optimal locations.

3. Physical Layer Design

The physical layer of the proposed wireless communication architecture consists of the mm-wave communication channel, the antennas, and the transceiver. In this section, we attempt to present the advancements in understanding and design along the various aspects of the physical layer for inter-chip communication in multi-chip systems.

3.1. Channel Characteristics

The medium of communication between chips is envisioned to be partially through the air between the chips and through a stack of silicon/silicon dioxide or packaging and encasement of the individual chips. Due to the wide variety of materials used in the fabrication of chips even more diversified by integration of heterogeneous chiplets in a multi-chip system specific channel models need to be developed to fine-tune consequent antenna and transceiver design. This introduces challenges in having universal solutions that will fit all environments. The distance between the chips will be the dominant factor affecting the path loss. For normal out-door and in-door communications, there are many channel models such as log-normal or Saleh-Valenzuela (SV) model [48]. One important requirement of channel model is that it should not contain any effects of antennas or circuits. Since, the on-chip antennas once fabricated are fixed in a dielectric, the usual techniques for removing antenna effects from channel models such as, compensation with the free-space characteristics of the antenna, cannot be adopted. This makes removing antenna effects on the channel model for both inter and intra-chip communication challenging, and we get an antenna-dependent channel model.

In modern processing platforms, chip encapsulant (epoxy mold), substrate (organic, ceramic, PTFE), underfill (epoxy composites), and other materials are used to securely pack the silicon die in an end-user package. Therefore, it is required to characterize their complex permittivity i.e., dielectric constant and loss tangent in the mm-wave spectrum as in [49]. Furthermore, reflections from the metallic components in package and encapsulations such as heat-sinks, heat-spreaders, can cause multipath propagation which may cause delay spreads and therefore, create inter-symbol interference (ISI). The delay-spread and ISI in these environments also need to be captured with small-scale channel model for transceiver designs. All these factors make the channel in a multi-chip environment quite hostile and extreme for wireless communication. Simple path-loss-based channel models for chip-to-chip wireless communication in mm-wave frequencies are presented in [11,50]. A small-signal channel model for the 3–10 GHz band for inter-chip wireless model was developed in [51]. A time-domain channel model at 15–24 GHz is presented in [52]. However, specific frequencies and structures make it difficult to generalize channel models to be used in other frequencies and scenarios.

In [53] the idea of using quilt packaging to enable direct through-air wireless communication between antennas is envisioned. Quilt packaging leaves patterns on the die not covered by the encasement materials, enabling direct radiation into the air, potentially reducing path loss. In [54] it is noted that the transmission from on-chip antennas mostly remain trapped as surface waves at the boundary of the dielectric and the silicon substrate of the dies. This reduces loss due to wasted radiation in space considerably favoring the communication across the dies. Etching waveguides in heat sinks or metal chucks sitting on top of the dies channeling radiation between transceivers can provide communication channels between chips. However, waveguides require layout and are not truly wireless in principle although they may improve power efficiency due to a reduction in path loss.

3.2. High-Bandwidth Antennas

The most important criteria for the antenna is that it should be embedded on a die while supporting high bandwidths. Therefore, it needs a small footprint, which in turn dictates the choice of

the carrier frequency as the antenna size is proportional to the wavelength of the carrier. Mm-wave antennas will have dimensions of the order of a millimeter or less making their footprint acceptable for on-chip implementations. Moreover, traditional off-chip interconnects such as substrate traces or PCIe provide multi-gigabit bandwidths. In order to provide an energy-efficient alternative to the traditional inter-chip interconnects the wireless interconnects must also provide competitive data rates. Therefore, to enable multi-GHz bandwidths and a small antenna footprint, the minimum carrier frequencies need to be in the mm-wave range higher than few tens of GHz.

Wireless interconnects using on-chip antennas tuned to mm-wave bands offer a mature interconnect technology as the antennas can be fabricated using the Back End of Line (BEOL) steps of CMOS fabrication processes. Many on-chip antennas such as meander dipole, zigzag dipole, slot, loop, inverted F, bow-tie, and Yagi have been designed and investigated [55]. To enable the inter-chip wireless interconnection, multiple cores in the chip needs to be equipped with antennas as discussed in Section 2.2. Therefore, the antennas should be compact enough to fit within the footprint of the core, which is typically a few square millimeters in current and future technology nodes. Relatively smaller antennas such as zigzag monopole antenna have been designed [56]. Zigzag monopole antenna provides a compact size and can be easily fabricated. The mm-wave zig-zag on-chip antennas and their co-planar waveguide feeds were designed to resonate in the 60 GHz frequency band [11]. Figure 4 shows this antenna with its major dimensions. This antenna has a footprint of about 0.385 mm^2. A co-planar feed structure is chosen for the antenna as it has low losses compared to other feed structures such as microstrips. A simple model of the chips was assumed without any heat sinks or complex structures that are typically found in such environments. All the antennas were tuned to 60 GHz with less than −25 dB of return loss and the worst transmission coefficient between the pairs was found to be around −35 dB [11].

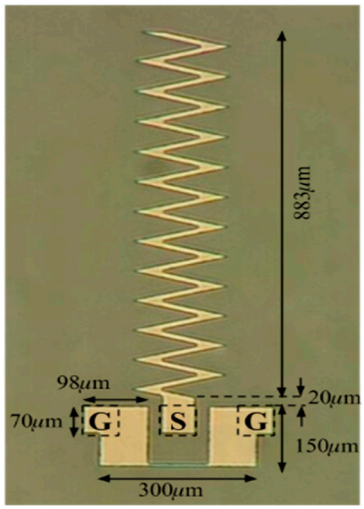


Figure 4. Fabricated metal zigzag antenna (microscopic picture) [56].

In addition to the above antenna designs, novel techniques can be used to enhance various aspects of the antenna characteristics. Compact antennas are generally not very directional. In order to improve directional gains of on-chip antennas fixed beam log-periodic antennas are also investigated [57]. On the other hand, phased antenna arrays provide the advantage of high directive gain with beam steering although, with additional feeding circuitry. However, given the wavelengths of a few millimeters and die sizes of few millimeters across, it is challenging to design antennas arrays to enable beam-steering for intra- and inter-chip wireless communication. Leaky wave antennas (LWA) are potentially viable alternatives to phased antenna arrays [58]. The bandwidth of on-chip antennas can be increased by using a combination of multiple antennas operating at different center frequencies. This will effectively create multiple frequency bands. Meta-surfaces such as an Artificial Magnetic

Conductor (AMC), which are periodic structures of elements can be designed in conjunction with the antennas to enhance the transmission of the on-chip antennas as well [59–61]. These structures can be used to reflect the antenna transmission outside the die reducing energy penetration into the substrate. However, the constraints on footprints in the on-chip environment need to be considered while designing meta-surfaces.

It may be noted that the transmission characteristics and radiation patterns of the antennas need to be estimated in presence of realistic features such as heat sinks and other structures. However, due to the variability of the nature of these platforms, it is challenging to formulate a generalized model suitable for all systems and configurations. Due to inability to isolate the antenna characteristics from that of the channel due to the antenna being embedded on-die, the wireless channel and antenna need to be viewed as a unified communication link, necessitating joint optimizations.

3.3. Low-Power Transceivers

Inter-chip wireless links require extremely low energy consumption to be competitive with high-speed serial I/O and other emerging technologies such as photonic interconnects. The power consumption in data transmission over wireless medium occurs at the transmitters and receivers. The choice of the transceiver is dependent on the choice of the physical layer and modulation technique. In the mm-wave physical layer designs, most research on wireless interconnects adopt simple modulation techniques such non-coherent On-Off-Keying (OOK) or Amplitude Shift Keying (ASK). Non-coherent modulations eliminate power-hungry, high-frequency carrier recovery circuits such as Phase Locked Loops (PLLs) resulting in low power consumption on the transceivers. A few on-chip mm-wave transceivers have been proposed for wireless interconnects in intra-chip NoCs [27,30,62]. Mm-wave transceivers have the advantage of being CMOS process compatible while providing high enough data rates for chip-to-chip interconnections. In [63], a 60 GHz Binary Phase Shift Keying (BPSK) transceiver is proposed. However, being a coherent modulation scheme, it requires carrier recovery circuitry resulting in relatively high-power consumption. Higher mm-wave carrier frequencies can potentially provide higher data rates. A 20 Gbps OOK transceiver operating at 260 GHz with a bit energy consumption of 58.65 pJ/bit is designed in 65 nm technology [64]. Due to the low voltage headroom in advanced technology nodes, the power output of the transmitter would be limited, therefore, most mm-wave transceivers discussed here are designed in 90 or 65 nm nodes. However, Fully Depleted Silicon on Insulator (FD-SOI) processes in advanced nodes such as 45 nm or 28 nm can provide higher power-performance trade-offs due to faster switching transistors [65]. In the next section, we present a non-coherent OOK transceiver designed at 45 nm for inter-chip wireless communication in the 60 GHz band.

3.3.1. 60 GHz Transmitter

Figure 5a shows the block diagram of an OOK transmitter consisting of a Voltage Controlled Oscillator (VCO), modulator and Power Amplifier (PA). In this section, we briefly discuss the design approaches for these components of the 60 GHz OOK transmitter designed using 45 nm technology.

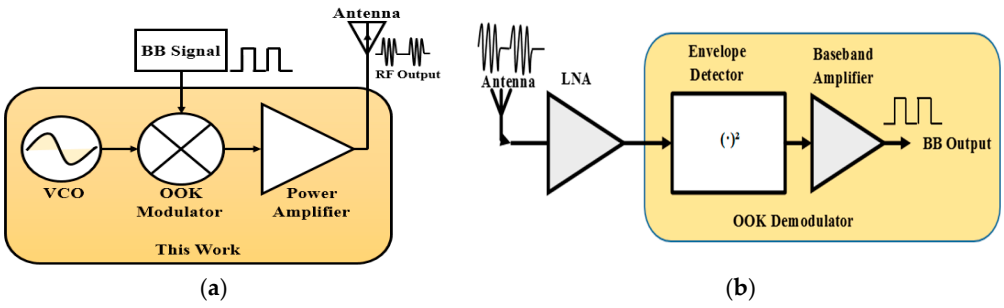
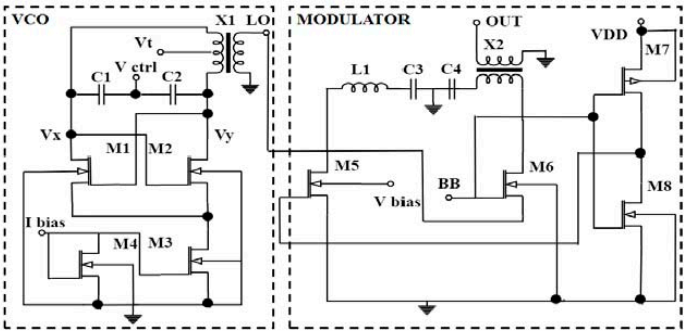


Figure 5. Block diagram of the OOK (**a**) Transmitter; (**b**) receiver.

Voltage Controlled Oscillator

The VCO is responsible for generating the 60 GHz RF carrier wave for the OOK modulator. The maximum single-ended voltage swing with minimum phase noise can be achieved by an NMOS cross-coupled VCO as it has comparatively less phase noise compared to the other oscillators such as ring or LC oscillator [66]. In the cross-coupled NMOS oscillator shown in Figure 6, the positive feedback is given via the M1, M2 transistors and the LC tank circuit. The startup time is a critical part of the oscillator design. To avoid the startup time from limiting the data rate we do not use the VCO as a modulator. Therefore, the output of the VCO, the Local Oscillator (LO) signal, is coupled to a separate OOK modulator input via a transformer, X1 with a 1:1 turns ratio to ensure symmetrical swing at the input of the modulator. We adopt transformer matching technique as it reduces overall power consumption by eliminating buffers or coupling capacitors.



X1, X2	M1n, M2n	M3n	M4n	M5n, M6n
1:1, 1:1.5	6 um/65 nm	60 um/65 nm	1 um/65 nm	2.25 um/45 nm
M7p	**M8n**	**C1, C2**	**C3, C4**	**L1**
480 nm/45 nm	450 nm/45 nm	10 fF	20.1 fF	350 pH

Figure 6. Schematic and component values of the proposed OOK modulator with NMOS cross-coupled VCO.

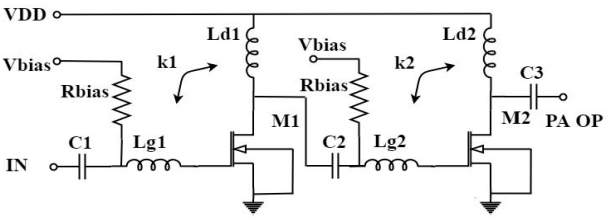
OOK Modulator

The proposed modulator is shown in Figure 6. The NMOS M6 works essentially as a pass transistor. The Base Band signal (BB) is fed into the gate terminal of M6. The BB signal when HIGH turns on the pass transistor allowing the carrier to pass to the drain of M6 from its source. Similar to the transformer X1 coupling the VCO to the modulator, another 1:1.5 transformer, X2 is designed to obtain single ended output from the modulator and for impedance matching of the modulator output with the input of the PA. NMOS M5 is used to drain the residual charge on the transformer when BB is LOW.

Two Stage PA

In the transmitter, power amplifier is one of the power-hungry blocks that affects the overall efficiency of the transmitter. In our design, a two-stage Common Source (CS) topology is used as it provides high frequency response and larger voltage swing compared to cascode designs. The two-stage PA proposed in this work is shown in Figure 7. It uses a CS topology with drain-to-gate transformer-feedback neutralization technique, which creates an additional signal path that neutralizes the current flow through Cgd [67]. The transformers are formed using inductors at the gate and drain terminals of M1 (Lg1 and Ld1) and M2 (Lg2 and Ld2) as shown in Figure 7 with coupling factors, k1 and k2 respectively. The values of the coupling factors are chosen such that they minimize the reverse transformation $Y_{12} = \frac{I_{gs}}{V_{ds}}\Big|V_{in} = 0$ of the two stages of the amplifier. The inductor Ld1 and Ld2 at the drain of the transistors M1 and M2 resonates with the drain capacitance. This increases the overall gain and bandwidth at the center frequency. Transistor M1 and M2 are biased using a voltage

divider-based biasing circuit. The output of the power amplifier is matched with 50Ohm by using L match impedance matching technique.



M1n, M2n	Ld1	Ld2	Lg1
10 um/90 nm	350 pH	420 pH	600 pH
Lg2	**C1**	**C2, C3**	**k1, k2**
350 pH	1 pF	14 fF	0.088, 0.128

Figure 7. Schematic and component values of the proposed power amplifier.

Characteristics of the 60 GHz Transmitter

Here, we present the key characteristics of the transmitter as described above. Figure 8a shows the frequency response of the VCO. From Figure 8a, it can be observed that the VCO is centered at 60 GHz. We simulated the on-off steady state gain of OOK modulator using Cadence Virtuoso analog design environment (ADE). At 60 GHz the modulator has an "on" and "off" state gain of −3.6 dB and −81.6 dB respectively as shown in Figure 8b. Therefore, the on-off steady state gain is above 78 dB for the frequency ranges from 45 GHz to 75 GHz. Figure 9 shows the S_{11}, S_{22}, and S_{21} for the proposed PA. The S_{11} is the ratio of reflected power to the incident power at the input port and is known as reflection coefficient or return loss. For the proposed PA design the value of reflection coefficient, S_{11} is found to be −30.15 dB at 60 GHz. Therefore, we conclude that the input is matched well with the 50Ohm antenna impedance and hence we find negligible reflections at the input at 60 GHz. Similarly, S_{22} is the ratio of reflected power to the incident power on the output port and it is found to be −15.42 dB at 60 GHz. S_{21} is the ratio of output power to the input power and represents the power gain for a well-matched PA. From Figure 9, S_{21} is found to be 14.48 dB.

The waveform shown in Figure 10 shows the amplified OOK modulated signal at the output of PA with a pseudo-random sequence of logic 0 s and 1 s at baseband signal input. As demonstrated 9 bits (000111100) are modulated between 1.06 ns and 1.59 ns indicating a data rate of nearly 17 Gbps. The output waveform indicates a 450 mVpp which translates to −3 dBm output power and is achieved with a total DC power consumption of 3.9 mW. This implies a bit energy efficiency of 0.23 pJ/bit at 17 Gbps on the transmitter side.

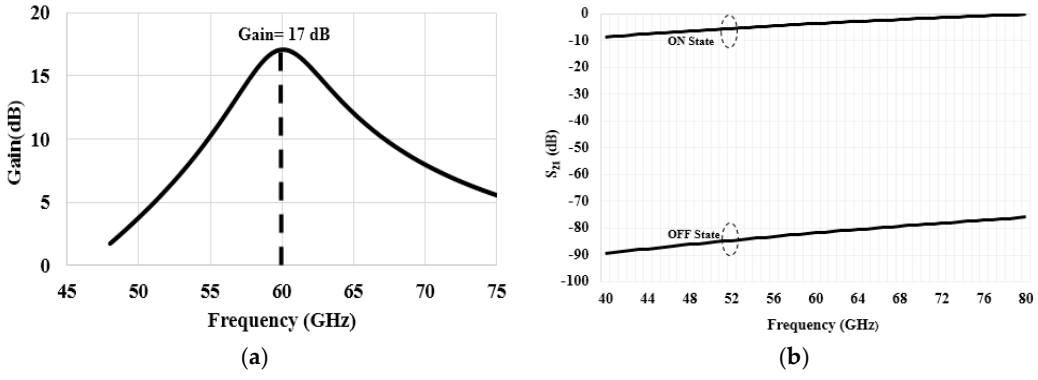


Figure 8. Transmitter characterization (**a**) VCO oscillation frequency (**b**) ON-OFF state gain of modulator.

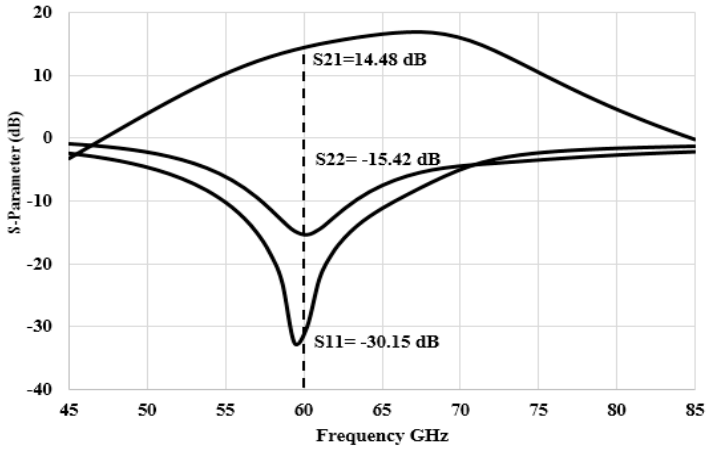


Figure 9. S parameter of the PA.

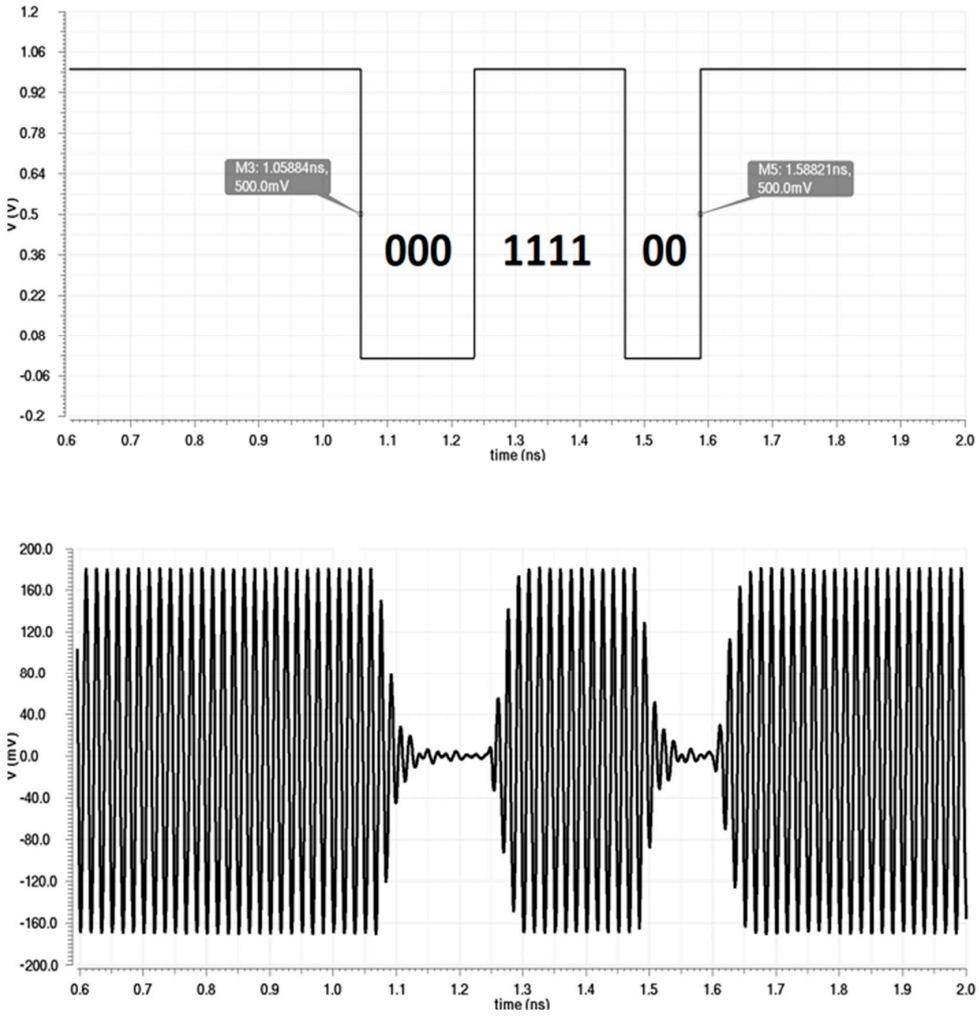


Figure 10. BB signal (**Top**) and transmitter output (**Bottom**) for a pseudo-random sequence.

3.3.2. 60 GHz Receiver

Here, we discuss a 60 GHz receiver proposed in [68] that can demodulate the signal received over inter-chip wireless interconnect transmitted from the above the transmitter. The Figure 5b shows

the components of the receiver that consists of a high gain Low Noise Amplifier (LNA) and an OOK demodulator to demodulate the amplified received signal. The LNA consists of a two-stage CS configuration as both the Common Gate (CG) and Resistive feedback-based topologies suffer from Noise Figure degradation due to the occurrence of noisy resistances in the signal path. Moreover, cascode structures, which are used commonly in low-frequency design for their high gains, are not suitable for the high frequency application. This is because the parasitic capacitances in the cascode transistors become dominant at higher frequencies, which reduces the inter stage impedance and hence, overall gain. The output of the LNA is then matched to the input of the demodulator. Since this information is modulated with a high frequency carrier signal, the OOK demodulator will exhibit a Low Pass Filter characteristic, removing the carrier wave and recovering the baseband digital signal. The proposed OOK demodulator consists of a source degenerated Envelope Detector (ED) at the input stage and a two-stage Baseband (BB) amplifier at the final output. The received waveform in Figure 11 shows that the receiver is capable of achieving a data rate of 17 Gbps with 12 mVpp at the input. The total DC power consumption of the OOK receiver is 6.1 mW, which results in a bit energy-efficiency of 0.36 pJ/bit [68] at a data rate of 17 Gbps. The LNA of the receiver has a Noise Figure of 2.8 dB making the noise floor −68.7 dB at 300 K for a bandwidth of 17 GHz Therefore, together with the transmitter the transceiver consume a total of 0.59 pJ/bit operating at 17 Gbps.

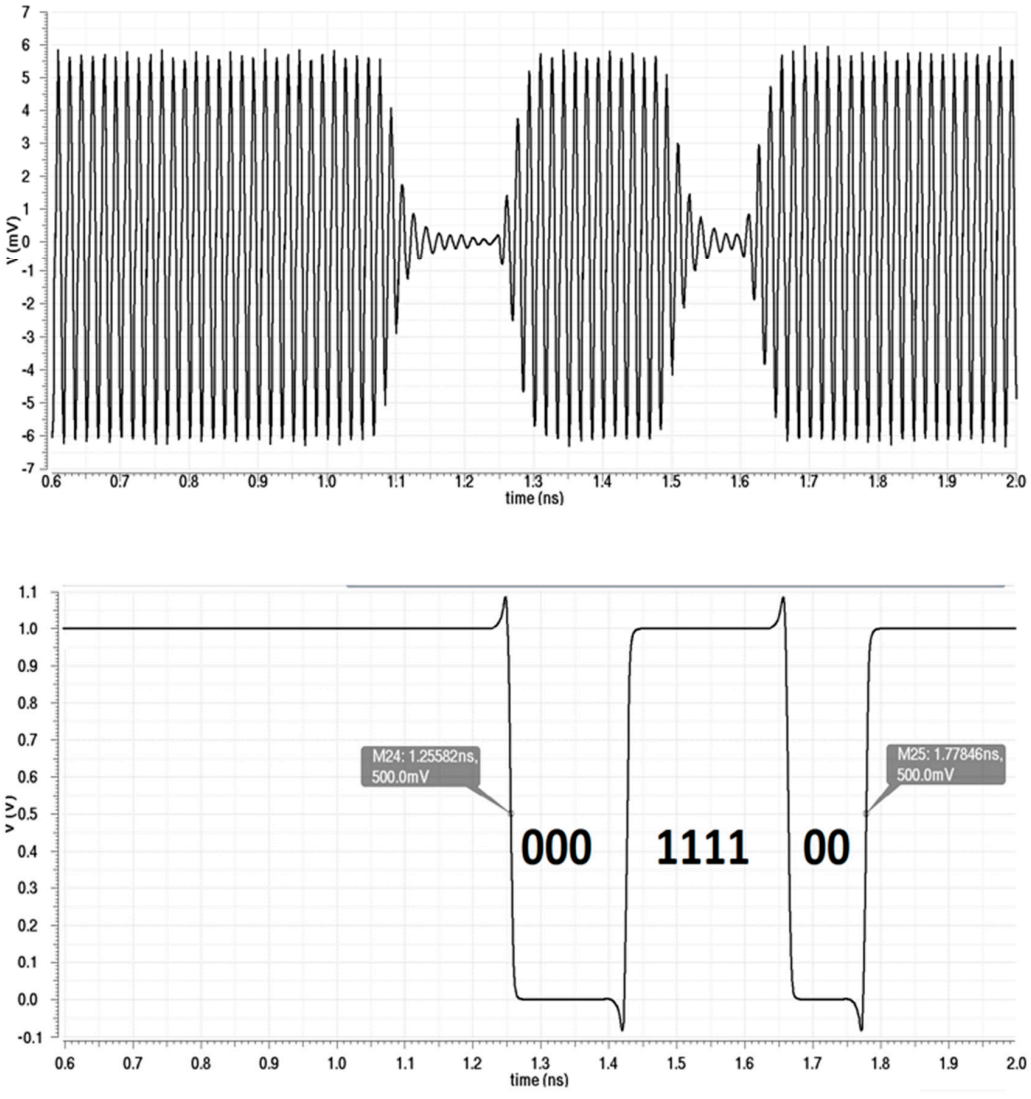


Figure 11. LNA input of the receiver (**Top**) and BB receiver output (**Bottom**) for the same sequence.

3.4. Link Budget Analysis

In this section, we estimate the Bit Error Rate (BER) that can be sustained by using the TX described here in typical application environments. We evaluate the BER for a range of path loss from 25 to 50 dB corresponding to typical intra and inter-chip communication distances [11]. The received signal power is given by:

$$P_R = P_T - PL \tag{2}$$

Here, P_R and P_T are the received and transmitted power respectively and PL is the path loss. At the receiving side we assume a non-coherent OOK receiver as these are most power efficient [68]. Considering Additive White Gaussian Noise (AWGN) at the receiver, the overall Noise Floor, N_{Floor} of such a receiver is given by,

$$N_{Floor} = 10\log(kT) + 10\log(BW) + NF. \tag{3}$$

where, k is the Boltzmann constant, T is the absolute temperature, *BW* is the bandwidth and NF is the Noise Figure of the receiver. We have considered an NF of 2.8 dB for the LNA designed in [68]. Therefore, from (2) the N_{Floor} of the receiver is -67.8 dBm at 300 K for a BW of 16 GHz. However, in addition to thermal AWGN, Inter-symbol Interference (ISI) can impact the reliability of the mm-wave link. Multipath channels existing in such environments as well as bandwidth limitation of the transceiver, the antennas and the channel, cause ISI. In the absence of thorough multipath channel models at 60 GHz in multi-chip environments as discussed in Section 3.1, we only consider bandwidth limited ISI in our link budget analysis. From Figure 10, the transmitter output is 450 mVpp and 67 mVpp when it is transmitting a '1' and a '0' respectively when measured at the center of the pulses. Therefore, the on-off ratio of overall OOK transmitter, R_{OOK} is 16.54 dB. Therefore, the value of bandwidth-limited ISI noise, N_{ISI} is dependent on path loss and is given by,

$$N_{ISI} = P_T - R_{OOK} - PL. \tag{4}$$

The received Signal to Interference and Noise Ratio (SINR) is given by,

$$SINR = P_R - N_{Floor} - N_{ISI} \tag{5}$$

For non-coherent OOK demodulation, the BER is given by [69],

$$BER_{OOK} = \frac{1}{2}\exp\left(-\frac{1}{2}SINR\right) + \frac{1}{4}\text{erfc}\left(\sqrt{\frac{1}{2}SINR}\right) \tag{6}$$

where erfc(.) is the complimentary error function. Figure 12 shows the BER achieved with this TX for a wide range of path loss values with and without ISI (for comparison). Consideration of multipath channel models may have a significant effect on this link budget analysis.

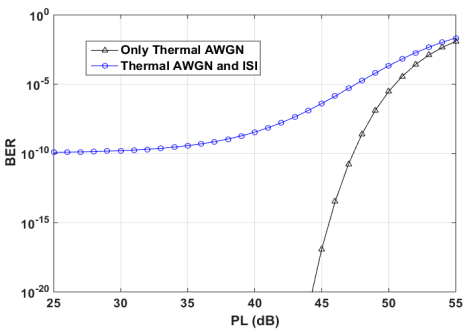


Figure 12. BER variation with path loss: with and without ISI.

3.5. Impact of Technology Scaling on Inter-Chip Wireless Interconnects

As device geometries continue to scale, the fabrication of two novel transistor structures are attaining maturity. These are the FinFET and the Silicon-on-Insulator (SOI) transistor. Both these transistors attempt to reduce the channel cross-section and increase the gate electric field ultimately increasing the switching speed. However, the FinFET is emerging as the device of choice for fast digital processes whereas, the Fully Depleted Silicon-on-Insulator (FDSOI) processes are emerging as the device of choice for high-speed analog applications [70]. Therefore, as the technology shrinks with 45 nm SOI, 28 nm SOI and beyond, the power consumption of the mm-wave transceivers is expected to keep reducing and provide higher data rate efficiency per milliwatt of power. In Tables 2 and 3, we present the trend in bit energy efficiency of recently designed mm-wave OOK transmitters and envelope detectors (for OOK receivers) respectively, having comparable data rates at 60 GHz with technology node scaling. The trends point towards a promising reduction of energy consumption per bit with technology scaling reaching sub-1pJ/bit at the 45 nm technology node.

Table 2. Performance Comparison of mm-wave OOK Transmitter across technology nodes.

References	This Paper	[67] [1]	[71] [1]
Technology node	45-nm CMOS	65-nm CMOS	90-nm CMOS
Carrier Frequency	60 GHz	60 GHz	60 GHz
Modulation	OOK	OOK	OOK
Maximum Data Rate	16 Gb/s	16 Gb/s	10.7 Gb/s
Power Consumption	3.9 mW	19 mW	31 mW
Supply Voltage	1 V	1 V	0.8/1.8 V
RF output P1dB	−3 dBm	1.5 dBm	5.1 dBm
Bit-Energy Efficiency	0.24 pJ/bit	1.2 pJ/bit	2.9 pJ/bit
Active-Foot Print	N/A	0.077 mm^2	0.15 mm^2
FoM	2.055	1.19	1.12

[1] Post-layout or post-fabrication results.

Table 3. Summary and comparison of the ED across technology nodes.

References	[68]	[69] [1]	[71] [1]
Technology	45 nm	65 nm	90 nm
Conversion Gain	16 dB	12.9 dB	16 dB
Carrier Frequency	60 GHz	60 GHz	60 GHz
Required input Power	−12 dBm	−16 dBm	−8 dBm
Max Data Rate	17 Gb/s	18.7 Gb/s	10.7 Gb/s
Power Consumption	1.3 mW	4.6 mW	19.2 mW
Supply Voltage	1 V	1 V	1.2 V
Responsivity	3.15 V/mW	6.1 V/mW	5.89 V/mW
Bit-Energy Efficiency	0.08 pJ/bit	0.25 pJ/bit	0.56 pJ/bit

While the power efficiency per Gbps of data rate can be expected to increase with the progress of SOI technology nodes, the performance of wireline interconnects, I/O are not predicted to scale at the same rate [72]. Non-scalable pin size or pitch of solder bumps and balls imposes limitations on the bandwidth density of inter-chip I/O and will prevent power consumption per Gbps of wireline communication from scaling significantly. In [73] Global Foundries projects that the percentage of die area occupied by only analog I/O components would be between 40–60% in the 7 nm node. Therefore, it is expected that as technology scales and SOI processes develop the power-performance gap between wireless inter-chip interconnects and traditional wireline interconnects will continue to widen, pushing the advantage towards the wireless interconnections even more.

4. Wireless Communication Protocol

The bandwidth for both on-chip as well as off-chip communication is limited by the transceiver performance and on-chip antenna design. However, adopting efficient wireless medium access mechanism, suitable flow control and routing protocols can significantly improve the bandwidth utilization and hence, performance. In this section, we discuss various flow control, routing, and MAC mechanisms for contention free, reliable, and energy-efficient wireless communication. The goal of the MAC should also be a seamless integration of the intra-chip and inter-chip wireless data transfer in order to reduce overheads of protocol transfer across different MACs.

4.1. Flow Control and Deadlock Free Routing

Routing protocols adopted for intra or inter-chip communication in a multicore multi-chip environment depends on the application and its performance goal. In most NoC-based systems, wormhole switching is adopted for wireline links in the multi-chip system where data packets are broken down into flow control units or flits [20]. Wormhole switching is known to reduce the buffering requirements which makes the switches consume low power and occupy lower area. We advocate the use of similar routing and switching protocols in both intra and inter-chip data communication in order to reduce overheads of protocol conversion and make the communication backbone seamless.

An adaptive routing protocol might be preferred for application requiring reliability but not low latency. While shortest-path routing should be used for application requiring low latency communication. Another important concern in routing for such environments is deadlock. Many conventional NoC routing algorithms leverage turn-model routing to achieve deadlock freedom. Dimension-order routing (DOR) or X-Y routing is an example of a turn-model routing algorithm for regular NoC architectures like mesh. Tree-based NoCs adopt Least Common Ancestor (LCA) as a deadlock-avoidance routing methodology. DOR and LCA algorithms have low computational complexity but are dependent on specific topologies and are not easily generalizable across multiple kinds of topologies. For irregular NoC architectures such as small-world topologies, deadlock-avoidance routing such as South-Last [74] and Layered Shortest Path (LASH) [75] have been proposed. Turn-model-based routing algorithms restrict certain turns to avoid deadlocks and therefore, may eliminate shortest paths in some cases. Layered routing algorithms such as LASH can be effective in deadlock avoidance for any arbitrary network topology. However, VC utilization maybe reduced if separate VCs are reserved for packets in specific virtual layers. Intra-chip WiNoCs with irregular small-world architectures have adopted LASH-based routing algorithms [76]. Dijkstra's algorithm, which extracts minimum spanning tree between any pair of node in a graph provides shortest path as well as deadlock free routing at the same time [77]. This because Dijkstra's algorithm extracts and routes packets over a Minimum Spanning Tree (MST) from any regular or irregular topology. This ensures the paths are shortest possible and the MST being a tree is inherently free from cyclic dependencies. Only information about next-hop from the MST for every possible final destination computed at design time is necessary at each switch, typically in the form of a Look Up Table (LUT). This eliminates the need for non-scalable global routing information and expensive route computation for every packet. However, the MST-based routing requires relatively higher area overhead due to the necessity of the LUT with next-hop information.

For multi-chip communication, intra and inter-chip communication should be seamless, and broadcast or multicast of latency sensitive messages should be given a preference. Broadcast or multicast traffic is commonly observed in multicore chips which can be generated by cache coherency and system management control or status update signals. Even a small number of broadcast/multicast messages can congest the whole interconnection as multiple copies of the same message are generated to ensure delivery to all destination cores [10,78]. Wireless interconnect has inherent broadcast capability to handle such broadcast traffic due to the shared wireless channel. Here we discuss one such possible broadcast routing in brief. Therefore, a forwarding-table-based routing algorithm over pre-computed shortest paths determined by Dijkstra's algorithm can be used to

transfer broadcast messages over the multi-chip communication fabric. To accommodate one-to-many multicast/broadcast messages encountered in these systems a tree-based routing policy can be chosen. For multicast packets, the header maintains a list of all destinations of the packet. If the routing paths for the possible destinations diverge at an intermediate switch, duplicate packets are generated and routed towards the divergent branches of the routing tree. The list of destinations is also split between the packets to prevent subsequent repeated splitting for same destinations.

Similar to the multicast messages an explicit list of final destinations can be maintained in the broadcast packet header. This will be split when the packet is split towards divergent destinations according to the adopted tree-based routing policy. This approach will require a large amount of destination information in the header. However, an alternative broadcast messaging protocol may be adopted where an explicit list of all destinations is not required, reducing the control overhead. In this case, the broadcast messages are not forwarded from a particular switch if it has already been received earlier. However, this condition of terminating further downstream flow of broadcast messages will cause a higher volume of packets compared to the earlier case.

The same tree-based routing strategy used for multicast/broadcast messages, can be used for unicast messages as well, as it routes messages along shortest paths. In Figure 13 we show the increase in system bandwidth and reduction in packet energy when wireless interconnects are used for a system with 4 chips, each having 16 cores, with increasing percentage of broadcast messages. Each chip is considered to be equipped with one mm-wave WI. The low-latency transmission of broadcast messages over the wireless links compared to the multi-hop wired paths results in the improvement of the bandwidth. The presence of long-distance wireless links over a shared wireless channel will ensure fast delivery of broadcast messages over the multi-chip system thereby dissipating less energy due to the faster data transfer for both broadcast and unicast messages.

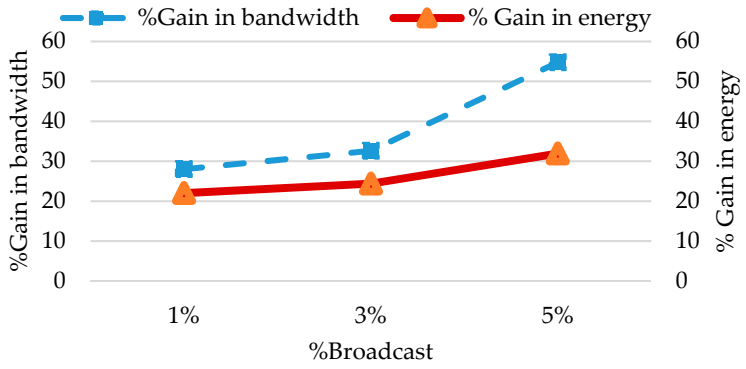


Figure 13. Performance evaluation under broadcast traffic.

4.2. Wireless Medium Access Control Layer

The MAC layer acts as an interface between the physical layer and the interconnection routers to enable chip-to-chip wireless communication. The MAC layer, along with the antenna and transceiver circuitry, enables the NoC routers to use the wireless channel. A MAC mechanism in this layer ensures the collision free communication among the switches equipped with wireless interfaces (WIs). However, adoption of complex MAC mechanisms from macro scale wireless networks is not feasible for multi-chip platforms due to the strict area, power, and buffer overhead constraints. Hence, design of simple, low-overhead, and fair MAC is one of the most challenging design tasks for inter-chip wireless interconnections.

To utilize the full potential of the novel wireless interconnect technology, the MAC should be distributed, simple, scalable, and efficient. Following these requirements, many MAC mechanisms has been designed for wireless interconnect fabrics. In [11] a distributed, asynchronous TDMA approach is proposed as it eliminates the need to synchronize transceivers deployed in different chips that have

no feasible way of sharing a synchronous clock. If the choice of physical layer enables the use of multiple frequency channels, Frequency Division Multiple Access (FDMA) can be used with graphene or Carbon Nanotube (CNT) antennas [31]. Code Division Multiple Access (CDMA) was proposed for inter-chip wireless MAC in [32]. To satisfy the performance and energy requirements of the future multi-chip systems, we advocate the following capabilities of the MAC along with the simplicity, fairness and scalability.

4.2.1. Smart Channel Access

In addition to basic requirements of providing channel access, we envision the MAC for wireless multi-chip systems can benefit from some smart features. The smart channel access needs to enable the WIs to gain channel access and acquire transmission opportunities or bandwidth based on their traffic demands. This will reduce the bandwidth underutilization by the WIs while transmitting data over the wireless channel which in turn, will improve the performance of chip-to-chip data transfer. Such capability is required because the demand of the WIs to access the wireless channel varies both temporally and spatially [79]. Traffic patterns are observed to be varying in both temporal and spatial sense in conventional intra-chip NoCs as well [45,80]. This is because traffic interaction between cores in a many-core chip is often non-uniform and time-varying in nature. The variation is expected to be even more in a multi-chip system with heterogeneous chips involving multicore CPUs, memory modules and GPU elements. A uniform channel access will result in waste of transmission opportunities, which will limit the performance gain achievable utilizing the novel wireless technology. Transmission opportunities can be in the form of token possession period, number of frequency/wavelength channels or code channels depending on the MAC mechanism. In [81–83] heterogeneous intra-chip NoC architectures are proposed. In [79] the authors proposed a simple history-based prediction mechanism to estimate the traffic demand of the WIs in the context of a single chip wireless NoC. Then, by allowing the WIs to transmit based on the predicted traffic demand, authors have shown 23% improvement in energy efficiency for application-based traffic patterns over the wired counterpart. Hence, we believe exploration of smart access mechanisms enabled MAC is going to be a driving force to meet the stringent bandwidth and energy requirements of the future multi-chip systems.

4.2.2. Robustness in MAC

With the shrinking feature size of the CMOS technology augmented with on-chip wireless transmission, the issues of failure in data transmission and single event upsets (SEU) are anticipated to be significant. Furthermore, the high frequency wireless transceivers are extremely vulnerable to noise. Such high risks of random failures and vulnerabilities can result in a failure to the MAC mechanism. This can either degrade the benefits of the novel interconnect technology or corrupt wireless data due to interference. Hence, to ensure reliable and interference-free wireless communication, the MAC mechanism should be robust against such failures. To mitigate such failures, authors in [84] proposed a Token Management Unit (TMU) that can recover from the failures in a token passing MAC mechanism. The TMU has intelligence about when the token should arrive to which WI, and in case the token is not possessed by the specific WI within a certain time, the TMU regenerates the token correctly while avoiding duplication. Such MAC mechanism also can restore the performance benefit of the wireless interconnect with minimal computation and power overheads.

5. CAD Tools and Simulation Environment

In this section, we will discuss the requirements and progress towards creating simulation environments for inter-chip wireless interconnections. Simulations are important to enable proof-of-concept, initial performance analysis and early design decisions. Infrastructure development at mm-wave frequencies is a very expensive enterprise making simulations even more important. However, simulation of wireless interconnection requires amalgamation of multiple simulation tools. At the physical layer, finite element methods need to be adopted to estimate channel characteristics

or design on-chip antennas. ANSYS High Frequency Structural Simulator (HFSS) [85] is among the most popular simulator used for simulation of channel characteristics and transmission characteristics of antennas. The transmitter and receiver circuits require Electronic Design and Automation (EDA) tools such as Cadence for design and simulations. RF design kits are required to model the mm-wave circuit components. Routers and interconnection network switches can be designed also using similar tools using RTL-level designs and extracting post-synthesis performance parameters using either ASIC design flows such as Synopsys or FPGA flow such as Xilinx.

Using the characteristics of the antenna, channel, transceiver circuits and routers, system-level simulators need to be used to capture the performance of the wireless inter-chip interconnection system. System-level simulators such as Gem-5 [86] or Multi2Sim [87] can be used to simulate multicore parallel systems even with heterogeneous processors and memory subsystems integrated with a pre-defined interconnection network. In case the data transport mechanism and switching protocol in the multi-chip system is same as that of an intra-chip NoC to create a seamless interconnection a NoC simulator can be extended to evaluate such multi-chip systems. A NoC simulator models the data flow by considering various simulation parameters and settings required for a specific system architecture. Although several NoC simulator exist [88–90], they have limited ability to model and evaluate wireless interconnection fabrics. A wireless NoC (WiNoC) simulator is available in the public domain [88]. In addition to implementing data flow and estimating various network performance metrics such as throughput, packet latency and packet energy WiNoC simulators have special features to simulate a system with specific applications. In [11] a cycle accurate WiNoC simulator has been used that characterizes the multi-chip architecture and models the progress of the flits over the switches and links per cycle accounting for those flits that reach the destination as well as those that are stalled. The WI is modeled as a port connected to the network switches where they are deployed. In [88] the author presented a configurable, extendible NoC simulator developed in system C that can analyze the performance and power figures of the both conventional wired NoC as well as emerging WiNoC architectures. The simulator not only supports synthetic traffic but also supports running application specific benchmarks under proper mapping. The simulator also supports basic routing algorithm and provide the flexibility to incorporate new algorithm from users. Figure 14 captures the interaction between simulators at various levels of abstraction.

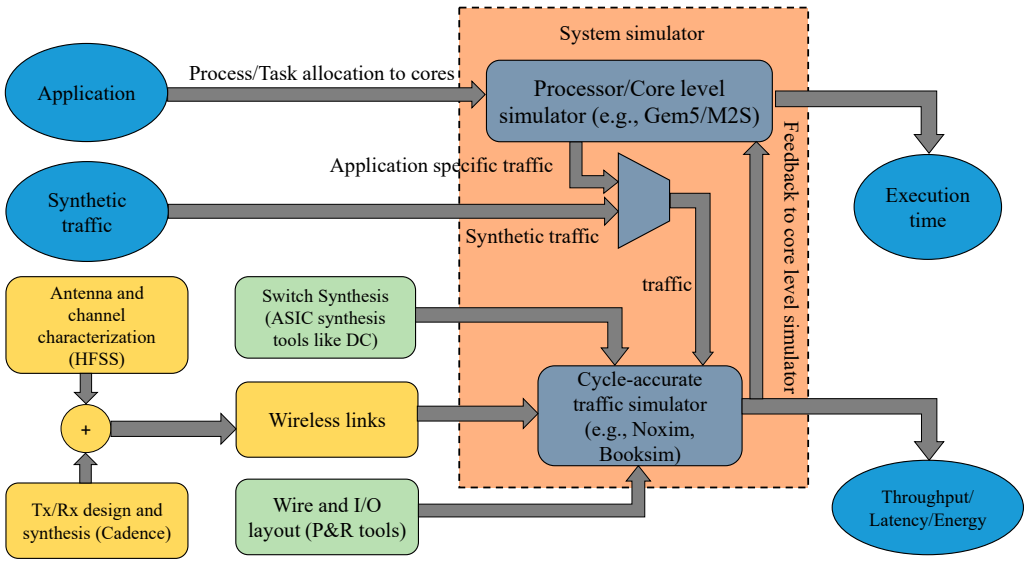


Figure 14. WiNoC simulation flow.

Using the above described simulation flow a wireless inter-chip interconnection fabric is evaluated for application specific traffic. We consider a wireless multi-chip configuration with 4 multi-core

processor chips and 4 DRAM memory stacks. The cores are considered to be out-of-order cores with 32 KB of L1 and 512 KB of L2 cache running a Directory-based MOESI cache coherency protocol. The memory stacks are considered to be 3D stacks with 4 layers on top of a logic layer responsible for memory addressing. The logic layer in each stack is equipped with a WI which can communicate with the WIs in the multicore chips. A single WI is deployed at one of the central cores of each multicore chip achieving the optimal WI deployment density of 1/16WI/core. Applications from the PARSEC [43] and SPLASH2 [44] benchmark suites are used to evaluate the systems. In order to map these traffic patterns to the multi-chip environment, we consider multiple threads of the same application kernel running on the multi-chip system where each processing chip executes a single thread, and the DRAM stacks are shared among threads. The average packet latency and average packet energy of the wireless architecture with respect to the interposer-based wire-line counterpart for different application-specific traffic patterns is shown in Figures 15 and 16 respectively. The latency best represents the performance in these cases as the interconnection network is not saturated in the steady-state. The reduction in average packet latency and average packet energy for the wireless multi-chip system varies between applications due to the variation in off-chip traffic patterns from different memory access patterns. However, for all application-specific traffic patterns considered here, the performance of the wireless multi-chip system is better than the interposer-based wireline configuration. The average reduction in packet latency and packet energy for the wireless multi-chip system is 54% and 45% compared to the interposer-based system, respectively. This is due to the energy efficient single-hop wireless links connecting processing chips and memory stacks.

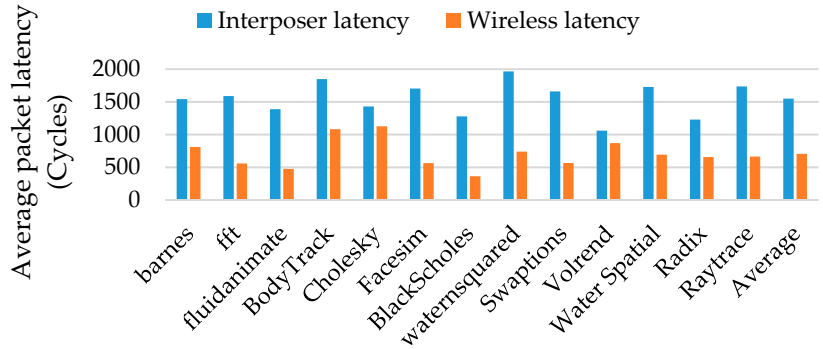


Figure 15. Performance evaluation (latency) under application specific traffic.

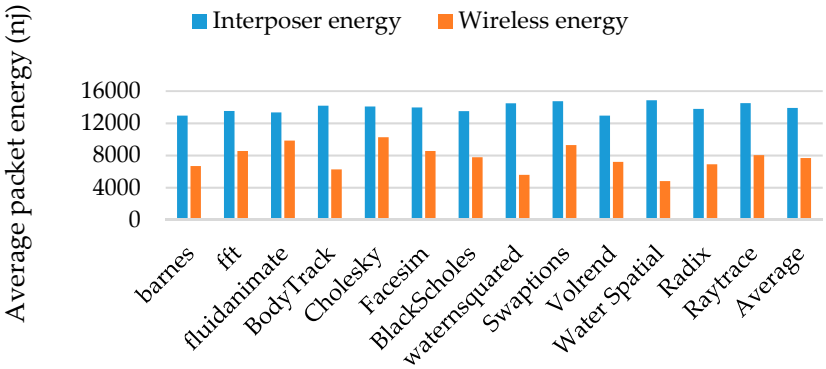


Figure 16. Performance evaluation (energy) under application specific traffic.

6. Possibilities of Paradigm Shifts in Computer Design and Methodologies

Interconnects are the main bottleneck to increasing performance of computing systems. Therefore, a solution to the interconnect problem has the capability to usher in innovative computer system design paradigms.

6.1. Non Von Neumann Architectures

For decades the dominant design paradigm for computing systems is the Von Neumann architecture that can be represented as Figure 17. The memory organization stretches from register files via multi-level caches to the main memory. Whenever there is a need to access memory that is farther away from the processor, longer physical distances degrade performance due to interconnect bottleneck. However, if the wireless interconnection system is adopted the data rate in the interconnect is independent of physical distance. Therefore, distant memories come virtually close to the processor. Therefore, programs and executions can be agnostic of the spatial locality of the data and/or instructions, ultimately resulting in eliminating the need for multi-level caches. An alternative to traditional Von Neuman architecture is envisioned in [91] where, in-memory computation is advocated to mitigate the cost of moving data as it is the most expensive atomic task. Such a system might need broadcast of instructions to various parts of the memory that contains the data. Shared medium wireless interconnections transferring and broadcasting instructions from instruction memory to data memory can enable in-memory computation more efficiently as well.

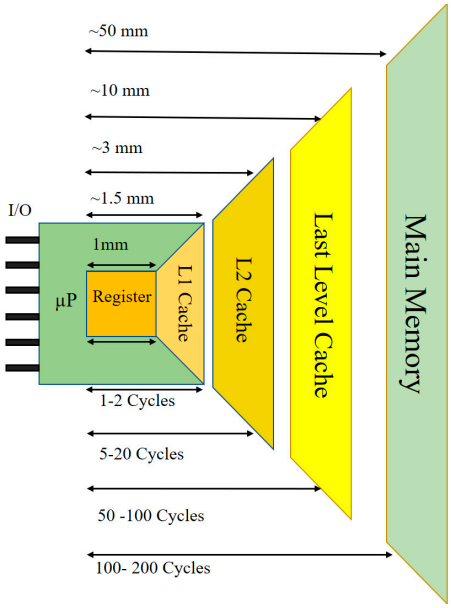


Figure 17. Comparative distance and access latency for hierarchical memory architecture.

6.2. Wafer Testing Made-Easy

Manufacturing testing is one of the most expensive and time-consuming steps in IC and system integration. After manufacturing, the wafers first go through the wafer testing to check functional defects by applying the special test vectors. Wafer-level testing presents huge challenges to high fault-coverage due to severely restricted controllability and observability, especially for complex Systems-on-Chips (SoCs) or multicore processors. After wafer testing, pre-packaged chips are cut and tested, packaged, and retested to undergo packaging test to check if the package induced defects. Finally, the packaged dies that pass the packaging test are then assembled onto a substrate such as a PCB board to make the final product, and the product level testing is done to check the interconnections and the assembly process. As a result, it is very important to catch defects at the early stages of the life cycle of a chip to save cost and ensure quality. This wafer-testing is an important but a very challenging step due to lack of controllability and observability of an uncut die. Hence, traditionally wafer-level testing is restricted to basic tests using probes. The probe needles come in physical contact with the wafer and can cause damage to the wafer due to the stress applied by the probes and it can also affect the packaging, assembly, and long-term reliability. Further probe needles require frequent

cleaning as they can accumulate debris on them and that causes contamination of needles which can lead to increase in contact resistance. The idea of reusing the NoC infrastructure as the Test Access Mechanism (TAM) for multicore SoCs to eliminate additional hardware overhead is proposed in [92]. Contact-less wafer testing has been proposed using capacitive coupling-based interconnects [93]. However, capacitive or inductive coupling requires close proximity of the test-bed to that of the chip and due to the footprint of on-die capacitors or inductors are not capable of providing sufficiently low-latency TAMs to deliver high volume test data for achieving high fault coverage. On-chip wireless transceivers can receive test data from an external equipment such as an Automatic Test Equipment (ATE) which will also be equipped with a wireless transceiver to send and receive test data and response respectively. This wireless TAM architecture is shown in Figure 18. Depending on the speed of the chosen physical layer and MAC efficiency, test data can be delivered at high rates to uncut dies in the wafer without requiring physical contact. The controllability and observability can be improved to the desired degree by deploying the wireless transceivers to receive and transmit test data at the required granularity or density.

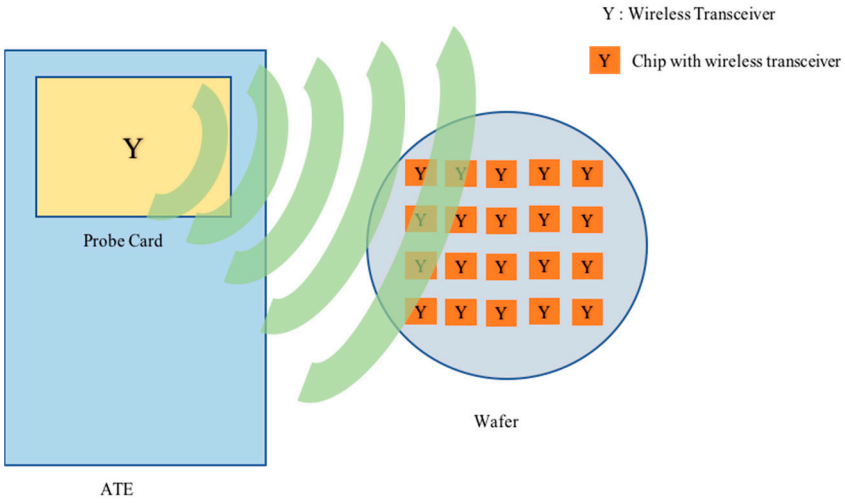


Figure 18. A conceptual view of wireless wafer testing.

The same wireless transceivers can also be used as part of the TAM [94] for testing of cut and packaged dies as well. On-chip wireless interconnects have been envisioned and investigated for several years now. This on-chip wireless interconnect can be used to communicate with an ATE equipped with a transceiver operating at the same frequency and uses a MAC integrated with the other wireless nodes in the chip. Therefore, inter-chip wireless interconnect can be used for testing cut and packaged dies as well, to provide a low-latency and energy-efficient TAM.

6.3. Enabling Voltage/Frequency Boundary Crossing

As the wireless interconnection virtually increases the proximity of the distant chips where the different chips can be functionally heterogeneous from different designers or vendors, they will most likely operate at different voltages and frequencies. Additionally, due to ubiquitous mechanisms for power management such as Digital Voltage/Frequency Scaling (DVFS) even the same chip in the system or the same island on a chip can operate at different voltage-frequency operating points. Typically, crossing voltage-frequency boundaries require clock domain synchronization or voltage conversion circuits resulting in overheads. In case of wireless transceivers, the serializer/deserializer (SERDES) buffers that convert the parallel data into serial form for transmission and vice versa for reception can operate in the voltage and frequency of the island they are in, thereby eliminating the need for separate synchronizers or voltage conversions. This enables simplifications of the design, EDA tools and verification or testing process as well. The designers only need to ensure that the

SERDES buffers can operate reliably within the operating voltage-frequency range of their islands or chips. The analog parts of the transceivers however, need to be powered by the voltages for which they are designed and characterized which can be different from those of the islands. This can be achieved in modern SoCs without overheads due to the presence of multiple voltage rails for standard power management techniques. If the analog transceivers are operated at different voltages their characteristics can change significantly affecting the reliability and link budget of the wireless interconnects. In absence of such power rails, the transceivers need to be characterized with the available power supply to evaluate their operating limits of speed and reliability. Lower voltages may force a reduction in transceiver data rate, but this may be acceptable in a voltage domain with lower supply.

7. Qualitative Benefits of Wireless Interconnects: Flexibility, Quick Design and Time to Market

Besides the quantitative benefits discussed so far, there are some qualitative benefits of adopting wireless interconnections for chip-to-chip communications. Primary among the qualitative benefits is the fact that there is no need for the layout of physical interconnects, neither metallic nor optical fibers. This implies that the designs will be more flexible and modular. Therefore, systems can be designed in a plug-and-play manner where the interconnect routing and placement will not impose additional constraints on the system design. Eliminating interconnect place-and-route procedures for large and complex systems can result in significant reduction in design, test, and verification times. Chiplets equipped with wireless transceivers can be just integrated on a physical platform without the need for physical interconnect layout, thus simplifying the design and integration process. This will reduce time-to-market for multi-chip systems, potentially reducing cost.

Due to the absence of physical layouts of interconnects the integration is possible between chips or components on a flexible substrate such as a polymer [95]. Bio-medical devices, smart watches and augmented vision devices are a few examples where wireless interconnects can eliminate the constraints of rigid platforms such as FR4 boards and interposers. This is suitable for wearable electronic gadgets where ICs on a flexible substrate can have relative motion between them. However, evaluation of the communication links in these specific systems are needed to enable their adoption in this case.

8. Comparison with Other Emerging Interconnect Technologies

In the last few years, several alternative interconnect technologies have been envisioned for novel intra-chip and inter-chip communications. 3D integration using TSVs is a methodology for stacking multiple active dies to reduce the planar distance between them. Dense TSVs over short vertical distances can provide high data rates between multiple dies for low power consumptions. However, 3D integration with TSVs require precise alignment of dies to avoid structural damage to the TSVs. Moreover, due to the reduced footprint of 3D ICs, the power dissipation density and temperature of 3D ICs can be very high requiring sophisticated cooling mechanisms such as microfluidic channels [18]. As an alternative to 3D integration, 2.5D integration has emerged more recently. In 2.5D integration, a substrate board or a silicon interposer is used to integrate the various chips in the system, which may include 3D stacks of memory subsystems [96]. The interposer itself is a recent technology where a relatively large bare silicon die with metal routing layers provides interconnection between the chips attached to it [15]. This reduces the stringent demands on the fabrication technologies of the 3D processes such as die alignments. However, the system performance is limited by the bandwidth density and on-die global wire characteristics of the interposer.

Chip-to-chip photonic interconnects have emerged as another enabling technology for high bandwidth interconnections [13]. In [97] the authors report the design of a large-scale microprocessor with 70 million transistors that are capable of communicating using light using fabrication CMOS manufacturing processes without any special steps for the silicon-photonic devices. A chip-to-chip monolithically integrated photonic interconnects is proposed in [98]. A unified inter and intra-chip

optical interconnection networks are discussed in [99]. In the photonic multi-chip system, the inter-chip communication happens through high bandwidth photonic interfaces. However, photonic interconnect poses many new challenges such as huge static power consumption, particularly due to the on-chip laser sources. The laser source used for optical signal generation typically has very low efficiency and are difficult to fabricate using SOI technology [100]. The micro-ring resonators used in the optical interconnect architectures needs thermal tuning as its wavelength selectivity varies with temperature. Moreover, the photonic signal cannot be buffered directly and therefore requires electrical conversion and hence introduce electro-optic conversion delay.

RF-Interconnects with microstrip waveguides were proposed to provide data freeways in intra-chip NoCs as well as chip-to-chip communications [28,101]. The Zenneck surface wave interconnect (SWI) is an emerging interconnect which is essentially an inhomogeneous 2D electromagnetic wave (EM) supported by a surface. The surface is a designed waveguide that traps the EM in two dimensional media [54]. As a result, the electrical-field decay rate in the SWI from the source horizontally along the boundary is around $(1/\sqrt{d})$ where d is the distance from the source [54]. This low power dissipation allows the SWI to offer relatively linear Joules per bit scaling compared to the higher order scaling of regular global buffered wire interconnects. However, though the surface wave interconnect provides energy efficient multicast/broadcast message transmission the surface itself needs to be designed carefully to match the desired impedance. Therefore, designing such surface with precise dimension and material is very challenging. Moreover, making the surface wave incident at the required Brewster angle for maximum transmission efficiency requires additional transducer and hence increase the complexity.

THz band communication using Graphene or Carbon Nanotube (CNT)-based antennas [102–105] has drawn much attention for intra and inter-chip communication as it provides low energy and high bandwidth. Surface Plasmon Polariton resonance in the specific Graphene-based structures can emit photons in the THz bands as a result of electronic pulse excitation [106]. Single walled metallic CNT structures have been shown to operate as dipole antennas when excited by a light source such as a laser source [103]. By controlling the length of the CNT elements, they can be tuned to specific frequencies making it suitable to employ Frequency Division Multiple Access (FDMA) as the medium access mechanism. This can create multiple non-overlapping THz channels greatly improving the interconnect bandwidth. Wireless NoCs utilizing CNTs for intra-chip communication have been proposed in [31]. Arrays of THz antennas based on Graphene have been proposed in [104]. A survey of THz based on Graphene has been presented in [105]. Graphene THz antennas have been used for broadcast enabled wireless NoCs [10]. While THz wavelengths are known to suffer exponential path loss due to molecular absorption for specific frequencies, the inter-chip distances being below 1 m will have negligible absorption [104,107]. Moreover, the multi-chip system can be packaged inside a controlled environment to reduce or eliminate specific molecules to mitigate the effect of molecular absorption if necessary.

Table 4 presents comparisons between various alternative inter-chip interconnection technologies in terms of qualitative advantage, disadvantage as well as energy/bit for a single point-to-point link and physical bandwidth per channel. The energy and bandwidth numbers vary depending upon design and implementation technologies. However, the numbers in Table 4 represent specific designs and implementations which are used for system-level comparative evaluation. Figure 19 shows the average packet energy and average bandwidth per core at network saturation for various interconnection technologies discussed here considering 65 nm technology node characteristics. We have considered a system with 4 multicore chips each with 16 cores in them and 4 memory stacks with 4 layers in each. In each multicore chip, the 16 cores interconnected with a mesh-based wireline NoC. For all the interconnect technologies, the memory stacks are connected to the I/O modules of the processing chips through 4 channels each consisting of 128 bit (assuming μ-bump pitch of 50 μm and 10 mm die edge) [96] wide channel operating at 1 GHz. Hence, this wide I/O provides a total bandwidth of 128 Gbps per channel with its neighboring processing chip with an

energy consumption of 6.5 pJ/bit [96]. The 4 channels of the HBM memory are connected with this wide I/O to the adjacent 4 cores along the boundary of the adjacent chips. For the conventional I/O-based system we have considered an architecture where a middle core along an edge is connected to its counterpart on the nearest edge of the adjacent chip. The high speed serial I/O channels are adopted from 65 nm designs and are shown to have a bandwidth of 15 Gbps with an energy consumption of 5 pJ/bit [108]. As analog I/O circuitry do not scale well with technology node these parameters are representative of current technology trends [73]. For the mm-wave wireless multi-chip system we have considered the inter-chip communication to utilize the mm-wave transceivers that are located at one of the central cores of each chip achieving a WI density of 1/16 WI/core. This design principle is adopted from the discussion on topology selection in Section 2.2. In the photonic multi-chip system, the inter-chip communication happens through high bandwidth photonic interfaces. To make all the systems comparable, we consider one photonic interface at each chip at the same location as the WIs. We connect these interface switches through a single waveguide. For our experiment, we consider one waveguide with 8-way Wavelength Division Multiplexing (WDM) channels to each enable concurrent Single Write Multiple Read (SWMR) communication between the 4 photonic interfaces using two WDM channels per link. As a representative of the THz band communication, we consider CNT-based nano-antennas. We adopt FDMA to enable concurrent communication among wireless nodes envisioning the use of CNT antennas tuned to different frequency bands [31]. To model the THz wireless multi-chip system for comparative evaluations, we consider the same system where the mm-wave transceivers are replaced at the central cores with CNT-based transceivers. A separate FDMA channel is reserved for each pair of communicating chips via their central cores. These systems are simulated using the cycle-accurate simulation methodology discussed in Section 5. Uniform random traffic is used with 20% packets generated from all cores being memory accesses. The memory accesses are randomly distributed among the memory stacks to emulate a shared memory system. From Figure 19 it can be seen that all the emerging technologies improve the performance compared to the conventional I/O-based system. Furthermore, mm-wave wireless technology outperforms the interposer-based system as it provides direct single-hop paths between internal cores of the chips.

Table 4. Advantage and disadvantages of various integration and interconnection technology.

Interconnect/ Integration Technology	Advantage	Disadvantage	Energy/Bit (pJ/Bit) at 65 nm Tech	Physical Bandwidth Per Channel (Gbps)
2D (I/O)	Matured technology and reliable communication.	Large latency and higher energy dissipation.	5.0 [108]	15 [108]
2.5D	Reduced cost and improved yield.	Number of device integrated is limited by interposer size. Challenges of metallic interconnects persist.	~0.2 pJ/bit/mm [21]	1–3 [21]
3D	Low latency and high bandwidth communication.	Higher heat dissipation density and less reliable TSV-based fabrication.	0.125 [109]	80 [109]
RF-I	Multi-band RF interconnect.	Requires physical waveguide design.	1 pJ/bit [101] [1]	4.0 [101]
Photonic	High bandwidth and low energy consumption.	Thermal variation on the chip potentially resulting in lack of tuning of electro-optic devices. Fabrication of Integrated laser sources for carrier generation in SOI processes.	0.43 [13]	10 [13]
Wireless (mm-wave)	CMOS compatible, single hop, energy efficient communication.	Limited physical bandwidth. Requiring architecture design to maximize bandwidth utilization.	2.075 [67]	16 [67]
Wireless (CNT—THz band)	High bandwidth and ultra-low energy consumption.	Fabrication and CMOS integration of CNT with desired properties is challenging.	0.48 [31]	10 [31]

[1] Value is in 45 nm technology node.

The mm-wave wireless interconnection outperforms the wired systems. However, the photonic and CNT-based systems perform better than the mm-wave technology due to higher channel capacity of optical and THz bands. The main limitation of the mm-wave wireless interconnection is the relatively less channel capacity of mm-wave bands compared to photonic waveguides using WDM. However,

the THz wireless inter-chip interconnections with CNT antennas can provide comparable performance to that of silicon photonics and has the advantage of not requiring physical layout of waveguides.

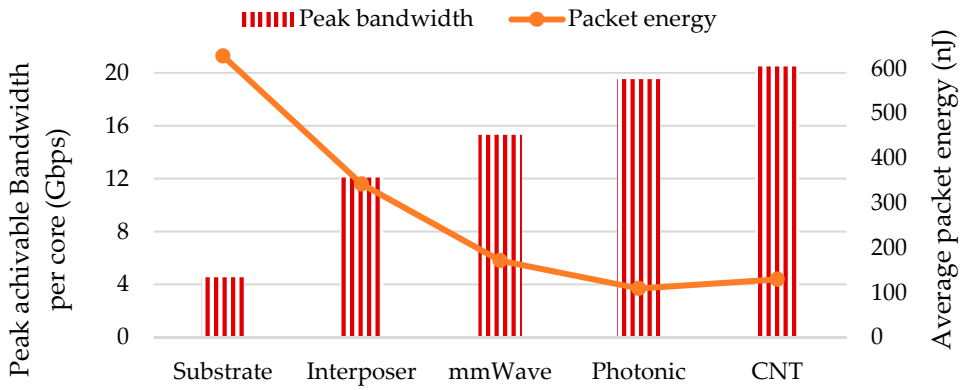


Figure 19. Performance Comparison for the 4C4M system using emerging technologies.

9. The Scope of Inter-Chip Wireless Interconnect with Advancements in Orthogonal Technologies Such as IoT, Industry 4.0, Wearable Electronics and Neuromorphic Computing

The world of IC design and communication is evolving at a rapid pace and several new directions seem to be emerging clearly, namely, IoT, Industry 4.0 and Neuromorphic Computing.

9.1. Fusion of Inter-Chip Wireless Interconnects with IoT

With IoT, fifty billion devices are predicted to be connected to the Internet by 2020 [110]. Chip-to-chip wireless interconnects can provide a seamless interconnection infrastructure to directly integrate chips into the IoT. One way of achieving this is to integrate the MAC of the inter-chip wireless transceivers with that existing in the IoT. The emerging IEEE802.11ad, IEEE802.11ay standards and compatible transceivers, modems, and MAC units, if integrated on the chip can connect chips directly to the IoT eliminating multi-level communication hierarchies thereby reducing significant overheads. Thus, using wireless inter-chip interconnects, it is possible to interconnect chips to IoT directly and efficiently. However, a challenge is expected to stem from the fact that the IEEE802.11 standards are developed for home WiFi environments target sparse network usage. Typically, they employ carrier sense approaches which are not scalable to many concurrent links. On the other hand, inter-chip interconnections require dense interconnects and high bandwidth. Therefore, novel MACs and standards need to be developed to enable this fusion of inter-chip wireless communication and IoT.

Wearable electronic devices such as smart wearable displays, watches, health monitoring equipment and vision devices are forming an integral part of modern IoT as all such devices will be equipped with IoT connectivity in the near future. These devices require integration of multiple ICs on a flexible substrate [111]. The design of active circuits and antennas on a flexible substrate require novel innovations in materials, circuit design as well as antenna design [112–114]. Due to constant bending of the substrate platform physical metallic wires could easily suffer from wear and tear. Therefore, inter-chip wireless interconnects can be used for power-efficient multi-chip integration without requiring physical waveguide-based interconnections.

9.2. Applicability of Inter-Chip Interconnects in Industry 4.0

Modern industrial revolution is seeing an increasing push towards automation, use of Artificial Intelligence (AI) and novel methodologies to combine manufacturing and material movement. Industry 4.0 will require automated and connected industrial elements such as material handling agents, robots, and sensing apparatus [113]. Robotic arms and material handling agents such as fork lifts have parts that need to communicate across moving joints and those with relative motion. Video sensors to aid in

localization of robots require capturing high-speed video feeds and transmitting them via high-speed cables to compute nodes. Cables hinder range of motion of robotic elements. Using short-range high bandwidth wireless interconnects video feeds from sensor cameras can be sent across moving parts to the computational elements inside chips directly. Modern and future 3D printing devices have robotic arms with moving parts similar to material handling agents. Wireless inter-chip connectivity between sensors, actuators and microcontrollers in 3D printers can also provide a low-power alternative to cables while increasing the range of motion between the parts with relative motion.

The future of material handling is going to include small-scale distributed warehouses in urban centers and material handling agents that are small and are able to move only light packages such as drones [114]. The weight and power budget on such agents will be more stringent than on larger agents such as forklifts. High-speed communication cables for communication between sensors and controller or compute nodes are generally metallic and can be heavy depending upon the required length. These can be eliminated by using wireless short-range communication. Therefore, short-range wireless interconnections have the potential of reducing the power requirement of small-scale distributed material handling agents such as drones or Unmanned Aerial Vehicles (UAV). In addition to providing a high speed communication channel, using the unlicensed 60GHz mm-wave frequency, indoor localization of autonomous industrial agents such as robots have been investigated. Mm-wave frequencies enable high-precision localization in the order of a few millimeters, which is impossible to achieve using lower frequency radar technologies [115].

9.3. Inter-Chip Wireless Communication in Neuromorphic Computing

Neuromorphic Computing enabling deep learning is emerging as the computation backbone for many applications [116]. Deep learning neural networks require several layers of artificial neurons where the number of interconnections between neurons scale exponentially with increase in the number of layers [117]. Typically, the neurons have a degree of connectivity in cortical networks in the order of 10,000. Therefore, traditional interconnection methodologies with physical interconnects will be incapable of realizing this kind of networks. Moreover, the communication between neurons in different layers need to be reconfigurable where a neuron communicates with different neurons at different times. Therefore, wireless interconnections render themselves most suitable for creating the communication backbone for such deep neural networks with reconfigurable interconnections. While the IBM TrueNorth [118] is among the most notable examples of neuromorphic chip, true computation power will require the integration of multiple neuromorphic chips to provide the scalability. In that case, the chip-to-chip interconnection network will become a bottleneck. This is because each individual chip will be neuromorphic, providing the performance per watt similar to the human brain while the communication between the chips will use traditional interconnect mechanisms. Therefore, novel inter-chip interconnections utilizing reconfigurable wireless links will provide the flexibility and scalability in performance needed for multi-chip neuromorphic computing platforms. Thorough investigation of the specific deep neural network will aid in determining the required data rates and power envelopes, thereby guiding the design choices for the wireless interconnection architecture including topology, physical layer, and MAC.

Thus, IoT, Industry 4.0 and Neuromorphic Computing present new opportunities for chip-to-chip wireless interconnections. However, these new paradigms may introduce their own challenges.

10. Challenges of Inter-Chip Wireless Interconnects

Beyond the opportunities discussed so far, there are several challenges of realizing an inter-chip wireless interconnection network. One such challenge is the security and privacy of data exchange over the wireless inter-chip interconnection, which needs to be ensured. While hardware security and trust are gaining importance among researchers [119,120] the inter-chip wireless introduces new challenges regarding security and privacy in chip-to-chip communication. Wireless communication between the chips in the multi-chip system can suffer from two imminent security threats namely,

eavesdropping and malicious data transfer. Eavesdropping can be done by a receiver tuned to the same frequency channel if the receiver is powerful enough to adequately amplify the signal. On the other hand, a malicious transmitter tuned to the same frequency can send unauthenticated malicious data flooding the interconnection ultimately causing Denial-of-Service (DoS). DoS-resilient wireless NoC architecture has been proposed [119]. In [120] authors present a mechanism for secure communication in wireless intra-chip NoCs. These techniques need to be evaluated for their application in multichip wireless interconnection networks. However, in multi-chip systems, the chances of such security threats can be mitigated by isolating the wireless interconnection system inside a metal package or case. However, in the absence of such possibility, the system needs to be protected against eavesdropping and unauthenticated messaging. For this, cryptographic and authentication solutions need to be investigated. Security and privacy measures existing in the larger system such as the IoT can be adopted for the on-chip wireless transceiver to equip the inter-chip wireless communication with a seamless security and privacy platform.

Another challenge stems from the fact that wireless bandwidth is a limited resource. In any wireless band, the total spectral range is specific and limited. Unlike wireless interconnects, the throughput of wireline links cannot be increased by tracing parallel wires to increase the bus-width. Such a technique is not possible in a shared medium interconnect system such as wireless interconnections. However, the data rate over wireless channels can be increased by adopting higher order modulation schemes. If each symbol represents multiple bits, the data rate over the wireless channel can be increased several times. However, the transceiver design for higher order modulation schemes will be more complicated and will, therefore, consume higher power compared to simple OOK transceivers. Carrier recovery circuits to implement coherent modulation schemes such as PLLs at high mm-wave frequencies are extremely power hungry [121]. Depending on the power-performance targets of the system appropriate modulation schemes need to be chosen. A design approach could be to have several modulators and demodulators in each transceivers to be selectively chosen depending on the data rate requirements of the system or of specific connections depending on applications or environments.

11. Conclusions

With ever increasing scale, complexity, and functionality in future high-performance computing nodes with multiple chips, the communication infrastructure is considered to be the dominant performance determinant. In this paper, we have discussed the advances and challenges of inter-chip wireless interconnects as communication backbone for multi-chip systems. Chip-to-chip wireless interconnect has the potential to introduce an energy efficient, high bandwidth unified communication architecture for homogeneous, heterogeneous and memory intensive multi-chip systems. Moreover, in future such communication architectures can introduce paradigm shifts in chip-memory communication and chip testing because of its inherent single hop and broadcast capability. In a broad sense, the multi-chip wireless interconnect architecture can be considered as the foundation of thriving future technologies such as the IoT where complex computing platforms maybe seamlessly integrated to both take advantage of and provide support the IoT-based systems. Other disciplines such as Neuromorphic Computing and industry 4.0 can also benefit from low-power high-speed wireless interconnects.

Acknowledgments: This work was supported in part by the US National Science Foundation (NSF) CAREER grant CNS-1553264.

Author Contributions: A.G., M.M.A. and M.S.S. are responsible for the vision and overall composition of the article as well the contributions towards discussions on architectures. R.S.N. and J.V. are responsible for the sections on wireless antennas. A.V. is responsible for the work on wafer testing. N.M. is responsible for the sections on wireless M.A.C. for multi-chip systems. Sa.S.'s contribution is on the THz communication technologies. T.S., Su.S. and M.I. contributed to the design of the mm-wave wireless transceiver circuits.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Vangal, S.R.; Howard, J.; Ruhl, G.; Dighe, S.; Wilson, H.; Tschanz, J.; Finan, D.; Singh, A.; Jacob, T.; Jain, S.; et al. An 80-tile sub-100-w teraflops processor in 65-nm cmos. *IEEE J. Solid-State Circuits* **2008**, *43*, 29–41. [CrossRef]
2. Howard, J.; Dighe, S.; Vangal, S.R.; Ruhl, G.; Borkar, N.; Jain, S.; Erraguntla, V.; Konow, M.; Riepen, M.; Gries, M.; et al. A 48-core IA-32 processor in 45 nm CMOS using on-die message-passing and DVFS for performance and power scaling. *IEEE J. Solid-State Circuits* **2011**, *46*, 173–183. [CrossRef]
3. Benini, L.; de Micheli, G. Networks on chips: A new SoC paradigm. *Computer* **2002**, *35*, 70–78. [CrossRef]
4. Bhunia, S.; Mukhopadhyay, S.; Roy, K. Process Variations and Process-Tolerant Design. In Proceedings of the 20th International Conference on VLSI Design, Bangalore, India, 6–10 January 2007; pp. 699–704.
5. Kannan, A.; Jerger, N.E.; Loh, G.H. Exploiting Interposer Technologies to Disintegrate and Reintegrate Multicore Processors. *IEEE Micro* **2016**, *36*, 84–93. [CrossRef]
6. ITRS Reports. Available online: http://www.itrs2.net/itrs-reports.html (accessed on 30 January 2018).
7. Murphy, B.T. Cost-size optima of monolithic integrated circuits. *Proc. IEEE* **1964**, *52*, 1537–1545. [CrossRef]
8. Threadripper Delidding Uncovers Epyc in Disguise—AMD 32 Core Desktop CPU Maybe in the Works. Available online: https://wccftech.com/amd-threadripper-delidding-epyc-32-core (accessed on 30 January 2018).
9. AMD Fusion APU Era Begins. Available online: http://www.amd.com/en-us/press-releases/Pages/amd-fusion-apu-era-2011jan04.aspx (accessed on 17 December 2017).
10. Abadal, S.; Sheinman, B.; Katz, O.; Markish, O.; Elad, D.; Fournier, Y.; Roca, D.; Hanzich, M.; Houzeaux, G.; Nemirovsky, M.; et al. Broadcast-Enabled Massive Multicore Architectures: A Wireless RF Approach. *IEEE Micro* **2015**, *35*, 52–61. [CrossRef]
11. Shamim, M.S.; Mansoor, N.; Narde, R.S.; Kothandapani, V.; Ganguly, A.; Venkataraman, J. A Wireless Interconnection Framework for Seam-less Inter and Intra-chip Communication in Multichip Systems. *IEEE Trans. Comput.* **2017**, *66*, 389–402. [CrossRef]
12. Ho, R.; Mai, K.W.; Horowitz, M.A. The future of wires. *Proc. IEEE* **2001**, *89*, 490–504. [CrossRef]
13. Wu, X.; Ye, Y.; Zhang, W.; Liu, W.; Nikdast, M.; Wang, X.; Xu, J. Union: A Unified Inter/Intrachip Optical Network for Chip Multiprocessors. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2014**, *22*, 1082–1095.
14. Topol, A.W.; La Tulipe, D.C.; Shi, L.; Frank, D.J.; Bernstein, K.; Steen, S.E.; Kumar, A.; Singco, G.U.; Young, A.M.; Guarini, K.W.; et al. Three-dimensional integrated circuits. *IBM J. Res. Dev.* **2006**, *50*, 491–506. [CrossRef]
15. Loh, G.H.; Jerger, N.E.; Kannan, A.; Eckert, Y. Interconnect-memory challenges for multichip, silicon interposer systems. In Proceedings of the International Symposium on Memory Systems, Washington, DC, USA, 5–8 October 2015; pp. 3–10.
16. Jerger, N.E.; Kannan, A.; Li, Z.; Loh, G.H. NoC Architectures for Silicon Interposer Systems: Why Pay for more Wires when you Can Get them (from your interposer) for Free? In Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), Cambridge, UK, 13–17 December 2014; pp. 458–470.
17. Annamalai, A.; Kumar, R.; Vijayakumar, A.; Kundu, S. A system-level solution for managing spatial temperature gradients in thinned 3D ICs. In Proceedings of the 14th International Symposium on Quality Electronic Design (ISQED), Santa Clara, CA, USA, 4–6 March 2013; pp. 88–95.
18. Sridhar, A.; Vincenzi, A.; Atienza, D.; Brunschwiler, T. 3D-ICE: A Compact Thermal Model for Early-Stage Design of Liquid-Cooled ICs. *IEEE Trans. Comput.* **2014**, *63*, 2576–2589. [CrossRef]
19. Shamim, M.S.; Ahmed, M.M.; Mansoor, N.; Ganguly, A. Energy-Efficient Wireless Interconnection Framework for Multichip Systems with In-Package Memory Stacks. Available online: https://arxiv.org/abs/1709.07529 (accessed on 27 February 2018).
20. Duato, J.; Yalamanchili, S.; Ni, L. *Interconnection Networks*, 1st ed.; Morgan Kaufmann: Burlington, MA, USA, 2002.
21. Kapur, P.; Chandra, G.; McVittie, J.P.; Saraswat, K.C. Technology and reliability constrained future copper interconnects. II. Performance implications. *IEEE Trans. Electron Devices* **2002**, *49*, 598–604. [CrossRef]
22. Circuits Multi-Projects (CMP). Available online: http://cmp.imag.fr/ (accessed on 9 February 2018).

23. Lin, J.J.; Wu, H.T.; Su, Y.; Gao, L.; Sugavanam, A.; Brewer, J.E. Communication using antennas fabricated in silicon integrated circuits. *IEEE J. Solid-State Circuits* **2007**, *42*, 1678–1687.
24. Yang, W.; Ma, K.; Yeo, K.S.; Lim, W.M. A 60GHz on-chip antenna in standard CMOS silicon Technology. In Proceedings of the IEEE Asia Pacific Conference on Circuits and Systems, Kaohsiung, Taiwan, 2–5 December 2012; pp. 252–255.
25. Yue, L. A 60-GHz Millimeter-Wave CMOS Integrated on-Chip Antenna and Bandpass Filter. *IEEE Trans. Electron Devices* **2011**, *58*, 1837–1845.
26. Gutierrez, F.; Agarwal, S.; Parrish, K.; Rappaport, T.S. On-chip integrated antenna structures in CMOS for 60 GHz WPAN systems. *IEEE J. Sel. Areas Commun.* **2009**, *27*, 1367–1378. [CrossRef]
27. Chang, K.; Deb, S.; Ganguly, A.; Yu, X.; Sah, S.P.; Pande, P.P.; Belzer, B.; Heo, D. Performance evaluation and design trade-offs for wireless network-on-chip architectures. *ACM J. Emerg. Technol. Comput. Syst.* **2012**, *8*, 1–25. [CrossRef]
28. Agyeman, M.O.; Vien, Q.T.; Ahmadinia, A.; Yakovlev, A.; Tong, K.F.; Mak, T. A resilient 2-d waveguide communication fabric for hybrid wired-wireless noc design. *IEEE Trans. Parallel Distrib. Syst.* **2017**, *28*, 359–373.
29. Deb, S.; Ganguly, A.; Pande, P.P.; Belzer, B.; Heo, D. Wireless NoC as interconnection backbone for multi-core chips: Promises and challenges. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2012**, *2*, 228–239. [CrossRef]
30. Laha, S.; Kaya, S.; Matolak, D.W.; Rayess, W.; DiTomaso, D.; Kodi, A. A New Frontier in Ultralow Power Wireless Links: Network-on-Chip and Chip-to-Chip Interconnects. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2015**, *34*, 186–198. [CrossRef]
31. Ganguly, A.; Chang, K.; Deb, S.; Pande, P.P.; Belzer, B.; Teuscher, C. Scalable hybrid wireless network-on-chip architectures for multicore systems. *IEEE Trans. Comput.* **2011**, *60*, 1485–1502. [CrossRef]
32. Vijayakumaran, V.; Yuvaraj, M.P.; Mansoor, N.; Nerurkar, N.; Ganguly, A.; Kwasinski, A. CDMA enabled wireless network-on-chip. *ACM J. Emerg. Technol. Comput. Syst.* **2014**, *10*, 1–20. [CrossRef]
33. Ahmed, M.M.; Shamim, M.S.; Mansoor, N.; Mamum, S.A.; Ganguly, A. Increasing Interposer Utilization: A Scalable, Energy Efficient and High Bandwidth Multicore-Multichip Integration Solution. In Proceedings of the IEEE International Green and Sustainable Computing Conference (IGSC), Orlando, FL, USA, 23–25 October 2017.
34. Pande, P.P.; Grecu, C.; Jones, M.; Ivanov, A.; Saleh, R. Performance evaluation and design trade-offs for network-on- chip interconnect architectures. *IEEE Trans. Comput.* **2005**, *54*, 1025–1040. [CrossRef]
35. Kumar, A.; Kundu, P.; Singh, A.P.; Pe, L.S.; Jha, N.K. A 4.6Tbits/s 3.6GHz single-cycle NoC router with a novel switch allocator in 65nm CMOS. In Proceedings of the International Conference on Computer Design, Lake Tahoe, CA, USA, 7–10 October 2007; pp. 63–70.
36. Lee, J.; Nicopoulos, C.; Park, S.J.; Swaminathan, M.; Kim, J. Do we need wide flits in networks-on-chip? In Proceedings of the IEEE Computer Society Annual Symposium on VLSI (ISVLSI), Natal, Brazil, 5–7 August 2013; pp. 2–7.
37. Kumar, A.; Peh, L.S.; Kundu, P.; Jha, N.K. Toward Ideal On-Chip Communication Using Express Virtual Channels. *IEEE Micro* **2008**, *28*, 80–90. [CrossRef]
38. Mondal, H.K.; Kaushik, S.; Gade, S.H.; Deb, S. Energy-Efficient Transceiver for Wireless NoC. In Proceedings of the 30th International Conference on VLSI Design, Hyderabad, India, 7–11 January 2017; pp. 87–92.
39. Deb, S.; Sah, S.P.; Cosic, M.; Chang, K.; Yu, X.; Heo, D.; Ganguly, A.; Belzer, B.; Pande, P.P. Design of an Energy-Efficient CMOS-Compatible NoC Architecture with Millimeter-Wave Wireless Interconnects. *IEEE Trans. Comput.* **2012**, *62*, 2382–2396. [CrossRef]
40. Yuan, M.; Fu, W.; Chen, T.; Wu, M. An exploration on quantity and layout of wireless nodes for hybrid wireless network-on-chip. In Proceedings of the High Performance Computing and Communications, Paris, France, 20–22 August 2014; pp. 100–107.
41. Kim, R.G.; Choi, W.; Liu, G.; Mohandesi, E.; Pande, P.P.; Marculescu, D.; Marculescu, R. Wireless NoC for VFI-Enabled Multicore Chip Design: Performance Evaluation and Design Trade-Offs. *IEEE Trans. Comput.* **2016**, *65*, 1323–1336. [CrossRef]
42. Kim, R.G.; Doppa, J.R.; Pande, P.P.; Marculescu, D.; Marculescu, R. Machine Learning and Manycore Systems Design: A Serendipitous Symbiosis. Available online: https://arxiv.org/abs/1712.00076 (accessed on 27 February 2018).

43. Biennia, C.; Kumar, S.; Singh, J.P.; Li, K. The PARSEC benchmark suite: Characterization and architectural implica-tions. In Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques, Toronto, ON, Canada, 25–29 October 2008; pp. 72–81.
44. Woo, S.C.; Ohara, M.; Torrie, E.; Singh, J.P.; Gupta, A. The SPLASH-2 Programs: Characterization and Methodological Considerations. In Proceedings of the 22nd International Symposium on Computer Architecture, S. Margherita Ligure, Italy, 22–24 June 1995; pp. 24–36.
45. Soteriou, V.; Wang, H.; Peh, L. A Statistical Traffic Model for On-Chip Interconnection Networks. In Proceedings of the 14th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, Monterey, CA, USA, 11–14 September 2006; pp. 104–116.
46. Bogdan, P.; Kas, M.; Marculescu, R.; Mutlu, O. QuaLe: A Quantum-Leap Inspired Model for Non-stationary Analysis of NoC Traffic in Chip Multi-processors. In Proceedings of the Fourth ACM/IEEE International Symposium on Networks-on-Chip, Grenoble, France, 3–6 May 2010; pp. 241–248.
47. Mondal, H.K.; Gade, S.H.; Shamim, M.S.; Deb, S.; Ganguly, A. Interference-Aware Wireless Network-on-Chip Architecture Using Directional Antennas. *IEEE Trans. Multi-Scale Comput. Syst.* **2017**, *3*, 193–205. [CrossRef]
48. Saleh, A.A.; Valenzuela, R. A Statistical Model for Indoor Multipath Propagation. *IEEE J. Sel. Areas Commun.* **1987**, *5*, 128–137. [CrossRef]
49. Toda, A.P.; Flaviis, F.D. 60-GHz Substrate Materials Characterization Using the Covered Transmission-Line Method. *IEEE Trans. Microw. Theory Tech.* **2015**, *63*, 1063–1075. [CrossRef]
50. Sikder, M.A.I.; Kodi, A.; Rayess, W.; DiTomaso, D.; Matolak, D.; Kaya, S. Exploring Wireless Technology for Off-Chip Memory Access. In Proceedings of the IEEE 24th Annual Symposium High-Performance Interconnects, Santa Clara, CA, USA, 24–26 August 2016; pp. 92–99.
51. Chen, Z.M.; Zhang, Y.P. Inter-Chip Wireless Communication Channel: Measurement, Characterization, and Modeling. *IEEE Trans. Antennas Propag.* **2007**, *55*, 978–986. [CrossRef]
52. Wu, H.T.; Lin, J.; Kenneth, K.O. Inter-chip wireless communication. In Proceedings of the 7th European Conference on Antennas and Propagation, Gothenburg, Sweden, 8–12 April 2013; pp. 3647–3649.
53. Bernstein, G.H.; Liu, Q.; Sun, Z.; Fay, P. Quilt packaging: A new paradigm for interchip communication. In Proceedings of the Electronic Packaging Technology Conference, Singapore, 7–9 December 2005.
54. Karkar, A.J.; Turner, J.E.; Tong, K.; Ra'ed, A.D.; Mak, T.; Yakovlev, A.; Xia, F. Hybrid wire-surface wave interconnects for next-generation networks-on-chip. *IET Comput. Digit. Tech.* **2013**, *7*, 294–303. [CrossRef]
55. Cheema, H.M.; Shamim, A. The last barrier: On-chip antennas. *IEEE Microw. Mag.* **2013**, *14*, 79–91. [CrossRef]
56. Narde, R.S.; Venkataraman, J.; Ganguly, A. Feasibility study of Transmission between Wireless Interconnects in Multichip Multicore systems. In Proceedings of the IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting, San Diego, CA, USA, 9–14 July 2017; pp. 1821–1822.
57. Shamim, M.S.; Mansoor, N.; Samaiyar, A.; Ganguly, A.; Deb, S.; Ram, S.S. Energy-efficient wireless network-on-chip architecture with log-periodic on-chip antennas. In Proceedings of the 24th Edition of the Great Lakes Symposium on VLSI, Houston, TX, USA, 21–23 May 2014; pp. 85–86.
58. Chang, L.; Zhang, Z.; Li, Y.; Wang, S.; Feng, Z. Air-Filled Long Slot Leaky-Wave Antenna Based on Folded Half-Mode Waveguide Using Silicon Bulk Micromachining Technology for Millimeter-Wave Band. *IEEE Trans. Antennas Propag.* **2017**, *65*, 3409–3418. [CrossRef]
59. Bao, X.Y.; Guo, Y.X.; Xiong, Y.Z. 60-GHz AMC-Based Circularly Polarized On-Chip Antenna Using Standard 0.18 µm CMOS Technology. *IEEE Trans. Antennas Propag.* **2012**, *60*, 2234–2241. [CrossRef]
60. Nafe, M.; Syed, A.; Shamim, A. Gain Enhanced On-Chip Folded Dipole Antenna Utilizing Artificial Magnetic Conductor at 94 GHz. *IEEE Antennas Wirel. Propag. Lett.* **2017**, *16*, 2844–2847. [CrossRef]
61. Barakat, A.; Allam, A.; Pokharel, R.K.; Elsadek, H.; El-Sayed, M.; Yoshida, K. Performance optimization of a 60 GHz Antenna-on-Chip over an Artificial Magnetic Conductor. In Proceedings of the Japan-Egypt Conference on Electronics, Communications and Computers (JEC-ECC), Alexandria, Egypt, 6–9 March 2012; pp. 118–121.
62. Mineo, A.; Rusli, M.S.; Palesi, M.; Ascia, G.; Catania, V.; Marsono, M.N. A closed loop transmitting power selfcalibration scheme for energy efficient WiNoc architectures. In Proceedings of the Design, Automation & Test in Europe Conference & Exhibition, Grenoble, France, 9–13 March 2015; pp. 513–518.

63. Lee, C.; Yao, T.; Mangan, A.; Yau, K.; Copeland, M.A.; Voinigescu, S.P. SiGe BiCMOS 65-GHz BPSK transmitter and 30 to 122 GHz LC-varactor VCOs with up to 21% tuning range. In Proceedings of the Compound Semiconductor Integrated Circuit Symposium, Monterey, CA, USA, 24–27 October 2004; pp. 179–182.
64. Park, J.D.; Kang, S.; Thyagarajan, S.V.; Alon, E.; Niknejad, A.M. A 260 GHz fully integrated CMOS transceiver for wireless chip-to-chip communication. In Proceedings of the IEEE Symposium on VLSI Circuits (VLSIC), Honolulu, HI, USA, 13–15 June 2012; pp. 48–49.
65. Jan, C.H.; Agostinelli, M.; Deshpande, H.; El-Tanani, M.A.; Hafez, W.; Jalan, U.; Janbay, L.; Kang, M.; Lakdawala, H.; Lin, J. RF CMOS technology scaling in High-k/metal gate era for RF SoC (system-on-chip) applications. In Proceedings of the International Electron Devices Meeting, San Francisco, CA, USA, 6–8 December 2010.
66. Dellsperger, T. *Design of a 5 GHz VCO in CMOS*; Swiss Federal Institute of Technology Zurich: Zurich, Switzerland, 2002.
67. Yu, X.; Sah, S.P.; Rashtian, H.; Mirabbasi, S.; Pande, P.P.; Heo, D. A 1.2-pJ/bit 16-Gb/s 60-GHz OOK Transmitter in 65-nm CMOS for Wireless Network-On-Chip. *IEEE Trans. Microw. Theory Tech.* **2014**, *62*, 2357–2369. [CrossRef]
68. Subramaniam, S.; Shinde, T.; Deshmukh, P.; Shamim, M.S.; Indovina, M.; Ganguly, A. A 0.36pJ/bit, 17Gbps OOK Receiver in 45-nm CMOS for Inter and Intra-Chip Wireless Interconnect. In Proceedings of the IEEE 10th International Workshop on Network on Chip Architectures, Cambridge, MA, USA, 14–15 October 2017.
69. Yu, X.; Rashtian, H.; Mirabbasi, S.; Pande, P.P.; Heo, D. An 18.7-Gb/s 60-GHz OOK Demodulator in 65-nm CMOS for Wireless Network-on-Chip. *IEEE Trans. Circuits Syst.* **2015**, *62*, 799–806.
70. FD- SOI. Available online: http://www.st.com/content/st_com/en/about/innovation---technology/FD-SOI.html (accessed on 17 December 2017).
71. Byeon, C.W.; Yoon, C.H.; Park, C.S. A 67-mW 10.7-Gb/s 60-GHz OOK CMOS transceiver for short-range wireless communications. *IEEE Trans. Microw. Theory Tech.* **2013**, *61*, 3391–3401. [CrossRef]
72. Chiang, P.; Dally, W.J.; Lee, M.J.E.; Senthinathan, R.; Oh, Y.; Horowitz, M. 20 Gb/s 0.13 µm CMOS serial link transmitter using an LC-PLL to directly drive the output multiplexer. In Proceedings of the IEEE Symposium on VLSI Circuits, Honolulu, HI, USA, 18–20 June 2004; pp. 272–275.
73. Wei, A.; Singh, J.; Bouche, G.; Zaleski, M.; Augur, R.; Senapati, B.; Stephens, J.; Lin, I.; Rashed, M.; Yuan, L.; et al. Challenges of analog and I/O scaling in 10nm SoC technology and beyond. In Proceedings of the IEEE International Electron Devices Meeting, San Francisco, CA, USA, 15–17 December 2014. [CrossRef]
74. Ogras, U.Y.; Marculescu, R. It's a small world after all: NoC performance optimization via long-range link insertion. *IEEE Trans. Very Large Scale Integr. Syst.* **2006**, *14*, 693–706. [CrossRef]
75. Skeie, T.; Lysne, O.; Theiss, I. Layered shortest path (LASH) routing in irregular system area networks. In Proceedings of the International Parallel and Distributed Processing Symposium, Lauderdale, FL, USA, 19 April 2002; p. 8.
76. Wettin, P.; Kim, R.; Murray, J.; Yu, X.; Pande, P.P.; Ganguly, A.; Heoamlan, D. Design Space Exploration for Wireless NoCs Incorporating Irregular Network Routing. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2014**, *33*, 1732–1745. [CrossRef]
77. Cormen, T.H.; Leiserson, C.E.; Rivest, R.L.; Cohn, C.S. *Introduction to Algorithms*, 3rd ed.; MIT Press: Cambridge, MA, USA, 2009.
78. Abadal, S.; Mestres, A.; Nemirovsky, M.; Lee, H.; González, A.; Alarcón, E.; Cabellos-Aparicio, A. Scalability of Broadcast Performance in Wireless Network-on-Chip. *IEEE Trans. Parallel Distrib. Syst.* **2016**, *27*, 3631–3645. [CrossRef]
79. Mansoor, N.; Shamim, S.; Ganguly, A. A demand-aware predictive dynamic bandwidth allocation mechanism for wireless network-on-chip. In Proceedings of the 18th System Level Interconnect Prediction Workshop, Austin, TX, USA, 4 June 2016; pp. 1–8.
80. Mishra, A.K.; Vijaykrishnan, N.; Das, C.R. A case for heterogeneous on-chip interconnects for CMPs. In Proceedings of the 38th Annual International Symposium on Computer Architecture, San Jose, CA, USA, 4–8 June 2011; pp. 389–399.
81. Harsha, G.S.; Deb, S. Hywin: Hybrid wireless NOC with sandboxed sub-networks for cpu/gpu architectures. *IEEE Trans. Comput.* **2017**, *66*, 1145–1158.

82. Zhan, J.; Kayıran, O.; Loh, G.H.; Das, C.R.; Xie, Y. OSCAR: Orchestrating STT-RAM cache traffic for heterogeneous CPU-GPU architectures. In Proceedings of the 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), Taipei, Taiwan, 15–19 October 2016; pp. 1–13.
83. Ziabari, A.K.; Abellán, J.L.; Ma, Y.; Joshi, A.; Kaeli, D. Asymmetric NoC Architectures for GPU Systems. In Proceedings of the 9th International Symposium on Networks-on-Chip, Vancouver, BC, Canada, 28–30 September 2015; p. 25.
84. Mansoor, N.; Iruthayaraj, P.J.S.; Ganguly, A. Design methodology for a robust and energy-efficient millimeter-wave wireless network-on-chip. *IEEE Trans. Multi-Scale Comput. Syst.* **2015**, *1*, 33–45. [CrossRef]
85. ANSYS HFSS. Available online: http://www.ansys.com/Products/Electronics/ANSYS-HFSS (accessed on 31 January 2018).
86. Binkert, N.; Beckmann, B.; Black, G.; Reinhardt, S.K.; Saidi, A.; Basu, A.; Hestness, J.; Hower, D.R.; Krishna, T.; Sardashti, S.; et al. The gem5 simulator. *ACM SIGARCH Comput. Arch. News* **2011**, *39*, 1–7. [CrossRef]
87. Ubal, R.; Jang, B.; Mistry, P.; Schaa, D.; Kaeli, D. Multi2Sim: A simulation framework for CPU-GPU computing. In Proceedings of the 21st International Conference on Parallel Architectures and Compilation Techniques, Minneapolis, MN, USA, 19–23 September 2012; pp. 335–344.
88. Catania, V.; Mineo, A.; Monteleone, S.; Palesi, M.; Patti, D. Noxim: An open, extensible and cycle-accurate network on chip simulator. In Proceedings of the IEEE 26th International Conference on Application-specific Systems, Architectures and Processors (ASAP), Toronto, ON, Canada, 27–29 July 2015; pp. 162–163.
89. Lis, M.; Shim, K.S.; Cho, M.H.; Ren, P.; Khan, O.; Devadas, S. DARSIM: A parallel cycle-level NoC simulator. In Proceedings of the Sixth Annual Workshop on Modeling, Benchmarking and Simulation, Saint Malo, France, 17 June 2010.
90. Jiang, N.; Balfour, J.; Becker, D.U.; Towles, B.; Dally, W.J.; Michelogiannakis, G.; Kim, J. A detailed and flexible cycle-accurate Network-on-Chip simulator. In Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), Austin, TX, USA, 21–23 April 2013; pp. 86–96.
91. Pawlowski, J.T. Hybrid memory cube (HMC). In Proceedings of the IEEE Hot Chips 23 Symposium (HCS), Stanford, CA, USA, 17–19 August 2011; pp. 1–24.
92. Vermeulen, B.; Dielissen, J.; Goossens, K.; Ciordas, C. Bringing communication networks on a chip: Test and verification implications. *IEEE Commun. Mag.* **2003**, *41*, 74–81. [CrossRef]
93. Kim, G.S.; Takamiya, M.; Sakurai, T. A capacitive coupling interface with high sensitivity for wireless wafer testing. In Proceedings of the IEEE International Conference on 3D System Integration, San Francisco, CA, USA, 28–30 September 2009; pp. 1–5.
94. Zhao, D.; Upadhyaya, S.; Margala, M. Design of a wireless test control network with radio-on-chip technology for nanometer system-on-a-chip. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2006**, *25*, 1411–1418. [CrossRef]
95. Liquid Crystalline Polymer Circuit Material Double-Clad Laminates. Available online: https://www.rogerscorp.com/documents/730/acm/ULTRALAM-3000-LCP-laminate-data-sheet-ULTRALAM-3850.aspx (accessed on 30 January 2018).
96. Kim, J.; Kim, Y. HBM: Memory Solution for Bandwidth-Hungry Processors. In Proceedings of the IEEE Hot Chips 26 Symposium (HCS), Cupertino, CA, USA, 10–12 August 2014; pp. 1–24.
97. Sun, C.; Wade, M.T.; Lee, Y.; Orcutt, J.S.; Alloatti, L.; Georgas, M.S.; Waterman, A.S.; Shainline, J.M.; Avizienis, R.R.; Lin, S.; et al. Single-chip microprocessor that communicates directly using light. *Nature* **2015**, *528*, 534. [CrossRef] [PubMed]
98. Sun, C.; Georgas, M.; Orcutt, J.; Moss, B.; Chen, Y.H.; Shainline, J.; Wade, M.; Mehta, K.; Nammari, K.; Timurdogan, E.; et al. A monolithically-integrated chip-to-chip optical link in bulk CMOS. *IEEE J. Solid-State Circuits* **2015**, *50*, 828–844. [CrossRef]
99. Yang, P.; Nakamura, S.; Yashiki, K.; Wang, Z.; Duong, L.H.; Wang, Z.; Chen, X.; Nakamura, Y.; Xu, J. Inter/intra-chip optical interconnection network: Opportunities, challenges, and implementations. In Proceedings of the Tenth IEEE/ACM International Symposium on Networks-on-Chip (NOCS), Nara, Japan, 31 August–2 September 2016; pp. 1–8.
100. Schoeniger, D.; Henker, R.; Ellinger, F. An 850-nm common-cathode VCSEL driver with tunable energy efficiency for 45 Gbit/s data transmission without equalization. In Proceedings of the IEEE Asia Pacific Microwave Conference (APMC), Kuala Lumpar, Malaysia, 13–16 November 2017; pp. 1103–1106.

101. Li, Y.; Cho, W.H.; Du, Y.; Du, J.; Huang, P.T.; Lee, S.J.; Chang, M.C. Carrier synchronisation for multiband RF interconnect (MRFI) to facilitate chip-to-chip wireline communication. *Electron. Lett.* **2016**, *52*, 535–537. [CrossRef]
102. Abadal, S.; Alarcón, E.; Cabellos-Aparicio, A.; Lemme, M.; Nemirovsky, M. Graphene-enabled wireless communication for massive multicore architectures. *IEEE Commun. Mag.* **2013**, *51*, 137–143. [CrossRef]
103. Kempa, K.; Rybczynski, J.; Huang, Z.; Gregorczyk, K.; Vidan, A.; Kimball, B.; Carlson, J.; Benham, G.; Wang, Y.; Herczynski, A.; et al. Carbon Nanotubes as Optical Antennae. *Adv. Mater.* **2007**, *19*, 421–426. [CrossRef]
104. Dragoman, M.; Muller, A.A.; Dragoman, D.; Coccetti, F.; Plana, R. Terahertz antenna based on graphene. *J. Appl. Phys.* **2010**, *107*, 104313. [CrossRef]
105. Diego, C.S.; Gomez-Diaz, J.S. Graphene-Based Antennas for Terahertz Systems: A Review. Available online: http://arxiv.org/abs/1704.00371 (accessed on 27 February 2018).
106. Jornet, J.M.; Akyildiz, I.F. Graphene-based plasmonic nano-transceiver for terahertz band communication. In Proceedings of the 8th European Conference on Antennas and Propagation (EuCAP), The Hague, The Netherlands, 6–11 April 2014; pp. 492–496.
107. Llatser, I.; Mestres, A.; Abadal, S.; Alarcón, E.; Lee, H.; Cabellos-aparicio, A. Time and Frequency Domain Analysis of Molecular Absorption in Short-range Terahertz Communications. *IEEE Antennas Wirel. Propag. Lett.* **2015**, *14*, 350–353. [CrossRef]
108. Balamurugan, G.; Kennedy, J.; Banerjee, G.; Jaussi, J.E.; Mansuri, M.; O'Mahony, F.; Casper, B.; Mooney, R. A scalable 5–15 Gbps, 14–75 mW lowpower I/O transceiver in 65 nm CMOS. *IEEE J. Solid-State Circuits* **2008**, *43*, 1010–1019. [CrossRef]
109. Ouyang, J.; Xie, J.; Poremba, M.; Xie, Y. Evaluation of using inductive/capacitive-coupling vertical interconnects in 3D network-on-chip. In Proceedings of the International Conference on Computer-Aided Design, San Jose, CA, USA, 7–11 November 2010; pp. 477–482.
110. State of the Market: Internet of Things 2017. Available online: https://www.verizon.com/about/sites/default/files/Verizon-2017-State-of-the-Market-IoT-Report.pdf (accessed on 1 February 2018).
111. Hsu, Y.P.; Liu, Z.; Hella, M.M. An ultra low-power front-end IC for wearable health monitoring system. In Proceedings of the 38th Annual International Conference of the Engineering in Medicine and Biology Society (EMBC), Orlando, FL, USA, 16–20 August 2016; pp. 1906–1909.
112. Abbosh, A.; Al-Rizzo, H.; Abushamleh, S.; Bihnam, A.; Khaleel, H.R. Flexible CPW-IFA antenna for wearable electronic devices. In Proceedings of the Antennas and Propagation Society International Symposium (APSURSI), Memphis, TN, USA, 6–11 July 2014; pp. 1720–1721.
113. ABOUT ROADMA 2.0. Available online: http://www.mhlroadmap.org/roadmap-2-0 (accessed on 1 February 2018).
114. Park, S.; Zhang, L.; Chakraborty, S. Design space exploration of drone infrastructure for large-scale delivery services. In Proceedings of the 35th International Conference on Computer-Aided Design, Austin, TX, USA, 7–10 November 2016; pp. 1–7.
115. Redant, T.; Ayhan, T.; De Clercq, N.; Verhelst, M.; Reynaert, P.; Dehaene, W. 20.1 A 40nm CMOS receiver for 60GHz discrete-carrier indoor localization achieving mm-precision at 4m range. In Proceedings of the IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), San Francisco, CA, USA, 9–13 February 2014; pp. 342–343.
116. Merolla, P.A.; Arthur, J.V.; Alvarez-Icaza, R.; Cassidy, A.S.; Sawada, J.; Akopyan, F.; Jackson, B.L.; Imam, N.; Guo, C.; Nakamura, Y.; et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* **2014**, *345*, 668–673. [CrossRef] [PubMed]
117. Calimera, A.; Macii, E.; Poncino, M. The Human Brain Project and neuromorphic computing. *Funct. Neurol.* **2013**, *28*, 191–196. [PubMed]
118. Cassidy, A.S.; Alvarez-Icaza, R.; Akopyan, F.; Sawada, J.; Arthur, J.V.; Merolla, P.A.; Datta, P.; Tallada, M.G.; Taba, B.; Andreopoulos, A.; et al. Real-Time Scalable Cortical Computing at 46 Giga-Synaptic OPS/Watt with. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, New Orleans, LA, USA, 16–21 November 2014; pp. 27–38.
119. Ahmed, M.Y.; Vidapalapati, A. A Denial-of-Service Resilient Wireless NoC Architecture. In Proceedings of the Great Lakes Symposium on VLSI, Salt Lake City, UT, USA, 2012; pp. 259–262.

120. Fernando, P.G.; Abellán, J.L. Secure communications in wireless network-on-chips. In Proceedings of the 2nd International Workshop on Advanced Interconnect Solutions and Technologies for Emerging Computing Systems, Stockholm, Sweden, 25 January 2017; pp. 27–32.
121. Szortyka, V.; Shi, Q.; Raczkowski, K.; Parvais, B.; Kuijk, M.; Wambacq, P. 21.4 A 42mW 230fs-jitter sub-sampling 60 GHz PLL in 40 nm CMOS. In Proceedings of the IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), San Francisco, CA, USA, 9–13 February 2014; pp. 366–367.