A-WiNoC: Adaptive Wireless Network-on-Chip Architecture for Chip Multiprocessors

Dominic DiTomaso, Student Member, IEEE, Avinash Kodi, Senior Member, IEEE, David Matolak, Senior Member, IEEE, Savas Kaya, Senior Member, IEEE, Soumyasanta Laha, Student Member, IEEE, William Rayess, Student Member, IEEE,

Abstract—With the rise of chip multiprocessors, an energy-efficient communication fabric is required to satisfy the data rate requirements of future multi-core systems. The Network-on-Chip (NoC) paradigm is fast becoming the standard communication infrastructure to provide scalable inter-core communication. However, research has shown that metallic interconnects cause high latency and consume excess energy in NoC architectures. Emerging technologies such as on-chip wireless interconnects can alleviate the power and bandwidth problems of traditional metallic NoCs. In this paper, we propose A-WiNoC, a scalable, adaptable wireless Network-on-Chip architecture that uses energy efficient wireless transceivers and improves network throughput by dynamically re-assigning channels in response to bandwidth demands from different cores. To implement such adaptability in our network at run-time, we propose an adaptable algorithm that works in the background along with a token sharing scheme to fully utilize the wireless bandwidth efficiently. Since no wireless NoC design has been completely realized with current technology, we describe technology trends in designing energy-efficient wireless transceivers with emerging technologies. We compare our proposed A-WiNoC to both wireless and wired topologies at 64 cores, with results showing a 1.4-2.6X speedup on real applications and a 54% improvement in throughput for synthetic traffic. Using Synopsys Design Compiler, our results indicate that A-WiNoC saves 25-35% energy over other state-of-the-art networks. We show that A-WiNoC can scale to 256 cores with an energy improvement of 21% and a saturation throughput increase of approximately 37%.

Index Terms—Emerging technologies, Low-power design, On-Chip Interconnection Network, Wireless communication

1 INTRODUCTION

The scaling down of silicon technology has facilitated the phenomenal increase in the number of processing cores that can be integrated within a single chip (called Chip Multiprocessors (CMPs)). The Network-on-Chips (NoCs) design paradigm solves several of the problems of traditional bus-based networks, including limited bandwidth and scalability [1]. Regular NoCs topologies such as meshes and tori are implemented using metallic links that are energy efficient and provide high date rate links at short communication distances. However, as the links become longer, the global interconnects suffer from higher energy usage (extra hops) and longer propagation delays. The higher energy and longer latency will significantly degrade the overall network performance and reduce the throughput of future CMPs.

Wireless interconnects are a potential solution that can provide energy efficient communication while providing high bandwidth and low latency [2], [3], [4], [5], [6], [7]. The unique benefits of wireless interconnects include, (1) high energy efficiency for long, one-hop communication, (2) reduced complexity compared to systems with waveguides or wires, and (3) compatibility with complementary metal-oxide-semiconductor (CMOS) wireless technology designs. Wireless interconnects can be used to transmit data across the chip in one-hop with low energy. Previous on-chip wireless/RF technologies have shown estimated energies of 0.33 pJ/bit [2], 1 pJ/bit [8], [9], 1.2 pJ/bit [5], and 4.5 pJ/bit [4]. On the other hand, wired interconnects can have an energy of approximately 3.2 pJ/bit to transmit across chip. Additionally, wired interconnects often require multiple intermediate routers increasing latency as well as energy. Wireless transmission requires no waveguides or wires, which reduces the area overhead and complexity of the chip design. In addition, wireless technology is a familiar form of communication with existing applications in wireless networking, cell phones, etc. The existing research in the field of wireless communication will facilitate the design of on-chip wireless technology.

As wireless-NoCs (WiNoCs) is a relatively new field and no prior work has completely realized a NoC wireless transceiver, there are several critical challenges in the design of architecture, modeling the wireless channel and implementing the transceivers. At the architecture level, such short wireless links allow data to propagate across the chip in one clock cycle, essentially independent of distance. Ideally, all communication on the chip should be wireless to implement an energy-efficient as well as a high-throughput network. However, with limited wireless frequency spectrum, it becomes essential to maximize the wireless channel utilization while minimizing the use of wireless channels for all on-chip communication. Wireless
channels have different path losses and dispersion that varies with frequency and these impairments have a direct impact on the design of the transceiver. Lastly, the transceiver technology should meet stringent, yet sometimes incompatible energy/bitrate/distance requirements for WiNoCs at the desired frequency band to be competitive with electronics.

In this paper, we propose A-WiNoC, an adaptable wireless NoC architecture that improves energy-efficiency and performance by restricting wireless links to global communication (long distance) and wired links for local (or near-neighbor) communication. An adaptable wireless algorithm is implemented that dynamically allocates channel bandwidth on application demand, thereby maximizing the wireless channel utilization. We propose a 64 core A-WiNoC architecture as well as a scalable 256 core architecture. Moreover, we provide a discussion of unique technology trends that indicate the feasibility of transceivers implementation across different technologies (RF-CMOS, SiGe BiCMOS). The major contributions of this work are as follows:

(1) **Adaptability**: Adaptable wireless networks can maximize the use of the limited wireless bandwidth and improve the performance (throughput and latency) for diverse traffic patterns without user intervention.

(2) **Energy Efficient Devices**: We evaluate the trends of low energy wireless devices across various emerging fabrication technologies such as sub-50nm RF-CMOS and SiGe BiCMOS.

(3) **Evaluation on Real Traffic**: In addition to synthetic traffic, we evaluate A-WiNoC on the real traffic PARSEC [10], Splash-2 [11], and SPEC2006 [12] benchmark traces collected from SIMICS [13] and GEMS [14]. Our results show an improvement of up to 54% in throughput, a speedup between 1.4X and 2.6X and energy savings of 25-35% over electrical and other wireless networks. A-WiNoC is shown to be scalable with results at 256 cores showing an increase in throughput of 37% and improvement in energy of 21% on average.

This paper is organized as follows: In section 2, we discuss related wireless NoCs architecture; in section 3, the A-WiNoC architecture and adaptable algorithm is explained; in section 4, we discuss the wireless channel modeling; in section 5, wireless technology trends and the proposed wireless technology for A-WiNoC are discussed; in section 6, we compare the throughput and energy of A-WiNoC to other competitive networks and in section 7, we conclude the paper.

## 2 Related Work

Recent research has utilized the unique advantages of wireless/RF transceivers for on-chip communication. The work in [5] used a RF transmission line to propagate packets on a RF signal across the chip at nearly the speed of light. With a slight area tradeoff due to the RF transmission line as well as electrical wires, the design was able to increase the throughput of the network while using a low energy of 1.2 pJ/bit. The WCube design was proposed in [4] which used a 2-tier network with an electrical wired mesh and a wireless backbone. A centralized wireless hub was used to connect different areas of the chip in a hypercube topology. Fixed wireless links were used for long distance communication while wires were used for short range. The wireless transceivers operated in the 100-500 GHz frequency range and consumed 4.5 pJ/bit. The network improved latency while consuming little power.

Another hybrid network was proposed in [2] which used fixed centralized wireless transceivers operating at only 0.33 pJ/bit and considered the use of carbon nanotube antennas and on-chip optical modulators. This hybrid design organized cores into subnets in which communication within a subnet was wired and communication between subnets was wireless. Each subnet had a centralized wireless hub that packets needed to route to before using a wireless link. Additionally, wireless interconnects were used in [3] to create long wireless links between computing chassis. The links used an energy of 2 pJ/bit to transmit a maximum distance of 30 cm. The iWISE design in [8] used distributed wireless transceivers for shared long distance communication and wires for short distances. The distribution of wireless transceivers reduced the need for additional hops to a centralized hub. However, the wireless links in all of these designs were fixed and did not take advantage of the adaptable nature of wireless transceivers. The work in [9] uses fixed wireless as well as a limited number of adaptable wireless links on a 64 core architecture. Our work extends this work by: (i) proposing a scalable architecture and evaluating a 256 core network, (ii) performing a sensitivity study by varying the number of adaptable wireless links, and (iii) modeling the path gain of the wireless channel in terms of frequency.

## 3 A-WiNoC: Adaptable Wireless NoC Architecture

A-WiNoC is a scalable wired/wireless hybrid architecture with adaptable links. A wired/wireless hybrid is used to supplement the wireless bandwidth as well as provide more energy-efficient communication. Wired links help provide the required high bandwidth demands of CMPs as well as the desired energy-efficiency at short distances. Wireless links, on the other hand, can provide high energy efficiencies at long distances. Another unique advantage of wireless links is their adaptability. We use adaptability in A-WiNoC since this can improve channel utilization and no previous work has dynamically allocated wireless links during runtime. Lastly, we create a scalable design for future CMPs that will implement more cores with the same wireless bandwidth.

### 3.1 NoC Design

**Architecture**: As wireless technology projections (low energy, high bitrates) are promising for WiNoC, we now propose our architecture called A-WiNoC, an adaptable wireless NoC as shown in Figure 1(a). Adaptability of our
architecture will be discussed in the next subsection. The proposed architecture consists of N cores and each core is connected to at least one router. To minimize energy dissipation and reduce packet latency, we concentrate four cores by connecting to a single router [15] (for N=64, N/16 cores are concentrated). Routers are organized into sets in order to systematically distribute static and dynamic wireless links. Figure 1(a) shows the set organization. Each set has N/4 cores - Set k has cores kN/4 to (k+1)N/4-1, for k=0,1,2,3 (Also seen in the simplified Figure 1(b)). The architecture is divided into four sets, each with four routers. Routers 0-3 are in Set 0, routers 4-7 are in Set 1, routers 8-11 are in Set 2, and routers 12-15 are in Set 3 (Also seen in Figure 1(b)). Each router has four transmitters: T_{ij}, which indicates a transmitter from Set i to Set j. The next subsection on communication will explain that all the routers in each set share these four wireless transmitters. As explained in [8], the choice of four routers and four sets balances channel access and transceiver hardware by giving a set an opportunity for every router to use a transmitter to send to a different set. Additionally, since we have 16 wireless channels available, the choice of four total sets each with four transmitters was made to evenly distribute wireless bandwidth. Therefore, the four routers share four transmitters for wireless communication between sets.

Figure 1(a) also shows the wired/wireless connections between routers. These routers are placed on the chip in a grid-like fashion. Wired links connect the routers similar to a mesh topology except routers within a set are fully connected. Wired links are, therefore, used for short distances as short metal wires consume low energy and have lower propagation delays compared to long metal wires. Additionally, diagonal wired links are used to fully connect routers within a set. This reduces the total wireless spectrum requirement while still maintaining a single hop network. Routing is based on the distance from the packet’s source node to its destination node. If the distance is only one wired hop then a wired link is used. If the distance is greater than one wired hop then a wireless link is used in order to reduce packet latency and power. Therefore, a packet will always take at most one hop from source to destination (wired or wireless) and deadlock can be avoided as there is no circular dependency for packet transmission.

**Communication:** The proposed adaptable wireless NoC architecture uses statically and dynamically configured wireless channels for communication between routers. The architecture uses 16 wireless channels as there are 16 routers. Each wireless channel has its own unique carrier frequency and each channel is only used by one transceiver at a time so that all interference can be avoided at the transmitting and receiving end. Additionally, we use passive bandpass filters in each transmitter to suppress any adjacent channel interference. With a total available bandwidth of 512 GHz, each wireless channel has a bandwidth of 32 GHz, corresponding to a 32 Gbps data rate for our binary modulation. There are 12 static wireless channels (see Figure 1(b)) which are used to transmit packets at low energy. Static channels allow the network topology to be connected at all times. An additional, four adaptable wireless channels can be dynamically reconfigured based on traffic patterns to give additional bandwidth to certain portions of the chip. Note that the adaptable wireless channels are adaptable in which set they transmit to; not adaptable in frequency, so transceivers always send and receive on the same frequency. The total 16 wireless channels are shared among multiple transceivers; these are replicated at each router (see Figure 1(a)). However, to avoid interference, a time division multiplexing (TDM) scheme is used to ensure that multiple transceivers do not use the same wireless channel simultaneously. This virtually creates more wireless links from the 16 wireless channels without increasing the total wireless bandwidth. Therefore, multiple transceivers are distributed at each router to share wireless communication and improve network performance.

For wireless communication, each set has four transmitters. Three transmitters are used for static communication and one transmitter can be reconfigured to any set. For example, in Set 0 of Figure 1(a), transmitters T_{01}. T_{02}. T_{03} are statically allocated from Set 0 to Set 1, Set 2, and Set 3, respectively. Transmitter T_{0j} can be reconfigured to any Set 1-3. The transmitters are replicated at each
A certain transmitter in order to avoid interference. Time slots to a router. Time slots indicate when a router can use the same wireless channel, TDM is used to assign time slots if there is no congestion. In order to hide the latency of token passing, the token can be passed before transmission is complete. By the time the token is received at the next router, transmission will have completed. Finally, a router will only send data one time when it receives a token in order to avoid starvation.

**Deadlocks:** Our 64 core network avoids deadlocks by routing packets to their destination in one hop. As previously described, depending on the distance from source to destination either a single wired link or a single wireless link will be used. Therefore, a packet will always take at most one hop from source to destination (wired or wireless) and deadlocking can be avoided as there is no circular dependency for packet transmission.

### 3.2 A-WiNoC for 256 cores

The architecture described in the examples above is for 64 cores. To scale A-WiNoC to a higher number of cores, such as 256 or 512, more cores per set can be added. We assume that the maximum wireless spectrum is being used, hence the number of wireless channels will remain at 16. Therefore, the set organization and number of transmitters remains the same while the number of cores attached to the transmitters will increase. Wireless communication with tokens and the reconfiguration algorithm (explained in the next Section) is the exact same as the 64 core version.
For example, at 256 cores, there will be 64 cores in each set connected via a wired mesh. Four wireless transmitters will be shared by 16 cores via a direct wired connection as shown in the inset of Figure 3. Four cores are concentrated to a single router as before; however, each router is directly connected to a wireless router. The wireless routers use the same communication protocol as previously including reconfigurability.

The routing for A-WiNoC at 256 cores will send a packet to its destination using the shortest path (wired or wireless) measured in number of hops. The only exception is when the destination is in the same set as the source. In this case, the packet must use all wired communication, as shown in Figure 4(a) where source 1 (S1) and destination 1 (D1) are in the same set. The packet must use wires because there is no transceiver for wireless communication within a set due to limited wireless bandwidth; there is only wireless communication outside of a set. If the source and destination are in different sets, such as S2 and D2 in Figure 4(a), the packet can still take a wired path if it is shorter than the wireless path. Wireless communication will be used for long distance communication. For example, S2 and D2 in Figure 4(b) will use a three hop communication path instead of the four hop wired path. The packet will take one wired hop from source to the wireless router. The packet will then capture the wireless token and transmit using a wireless link. Finally, one more wired hop will be required to reach the destination. Each wireless communication path is exactly 3 hops. Therefore, the routing can be simplified to the following: If the source and destination are in the same set or the path from source to destination is less than three wired hops then use an all wired path; else, use a wireless link. The distance of the path can be easily calculated by using the x and y coordinates of the source and destination. Dimension ordered XY routing can be used for metal wire hops as well as escape VCs to avoid network and protocol deadlocks.

3.3 Reconfiguration

Unlike previous wireless NoC architectures, we take advantage of the inherent adaptability of wireless interconnects. Reconfiguration is used in our 64 core and 256 core architectures to give more bandwidth to sets with the most traffic. This will improve performance by decreasing packet latency and improving throughput. The A-WiNoC architecture reconfigures time slots to the adaptable transmitter. Time slots are defined as cycles in which a transmitter can send data and are allocated by the passing of tokens. Each static transmitter allocates all of their available time slots to their fixed sets. Whereas the adaptable transmitter can allocate time slots to different destination sets depending on the traffic pattern. This gives more time slots to packets with destinations in the busiest set, which will reduce contention and increase network throughput and decrease packet latency.

The global controller (GC) makes the decision to which set an adaptable transmitter should allocate its resources. The local controller (LC) collects statistics on each wireless link utilization and indicates to the adaptable transmitter that a reconfiguration is needed. Link utilization is used because it reacts better to changes in traffic than buffer
utilization [16]. Each $LC_i$ is attached to one of the four wireless transmitters as shown in Figure 2(b). Each $LC_i$ uses hardware counters to collect historical statistics. Each time a packet is sent, each $LC_i$ updates the counter, $Link_{util}$. At the end of the reconfiguration window, $R_w$, each $LC_i$ sends $Link_{util}$ to the GC. $R_w$ equals 100 cycles in this paper. In the sensitivity study we show results for different $R_w$. The size of this counter in bits is equal to $\log_2(R_w/num_{flits})$, where $num_{flits}$ is the number of flits in a packet. Figure 2(b) shows the communication between each GC and $LC_i$ for Set 0. Other sets use similar communication. The GC compares the data and determines which Set has the highest utilization. GC then communicates with $LC_i$ attached to the adaptable transmitter to reconfigure to the set with the highest utilization. The pseudo code for the reconfiguration algorithm is shown in Algorithm 1.

4 WIRELESS TRENDS AND TECHNOLOGY

4.1 Modeling the WiNoC Channel

The allocation of frequencies to wireless links will depend on the distance from the wireless transmitter to the receiver. An example of the channel attenuation effects versus frequency is shown in Figure 5. This figure plots “free-space” (vacuum) path gain vs. frequency from 150 to 500 GHz for two different link distances. The dashed line is for a link distance of 1 mm, and the dotted line for a distance of 1 cm. Conceptual signal spectra are also shown across the band, at their relative received power levels, assuming equal transmit powers at all frequencies. For either distance, the variation of attenuation across frequency, from minimum to maximum, is approximately 10.5 dB; this requires a transmit power level more than 10 times larger at 500 GHz than at 150 GHz. Similarly, there is a 20 dB difference at any given frequency between the attenuation at 1 mm and

Algorithm 1: Reconfiguration Algorithm

Step 1: Wait for reconfiguration window, $R_w$
Step 2: GC sends $Link_{Request}$ control packet to all $LC_i$
Step 2a: Each $LC_i$ computes the $Link_{util}$ for previous $R_w$ and updates the field in the $Link_{Request}$ packet and returns back to GC
Step 3: GC receives $Link_{Request}$ packet containing information for all outgoing links
Step 3a: GC separates each $Link_{util}$ for each outgoing set: $Set_{0util}$, $Set_{1util}$, $Set_{2util}$, and $Set_{3util}$
Step 3b: GC finds max[Set_{0util}, Set_{1util}, Set_{2util}, Set_{3util}]
Step 4: GC sends $Link_{Response}$ control packet to adaptable transmitter, $T_{ij}$. $Link_{response} \in 00, 01, 10, 11$, where 00 indicates maximum utilization is Set 0, 01 is Set 1, 10 is Set 2, and 11 is Set 3.
Step 4a: Transmitter $T_{ij}$ reallocates time slots to set with maximum utilization by only accepting packets for that outgoing set
Step 5: Go to step 1

that at 1 cm. This clearly means that the lowest possible frequency should be used for the largest link distances. Finally, results in Figure 5 assume that antenna gains do not vary with frequency; over such a large frequency band this is unlikely to be true, and at best gains might increase with frequency to compensate somewhat for the path loss difference.

4.2 Wireless Technology Trends

As wireless NoC (WiNoC) is an emerging technology, the most practical guideline to assess the viability of WiNoC technology is to refer to trends in important figures of merits measured for ultra-low power and short range CMOS transceivers in literature. Figure 6 shows both data rate and link distance plotted as a function of modulation energy efficiency. Each circle represents the data rates of a specific transceiver design and each square represents the maximum transmission distance of a transceiver design. The dotted line shows the trend of data rates and the solid line shows the trend of transmission distance. The stars show our target rate of 32 Gbps and our target distance of approximately 1 cm both at an energy of 1 pJ/bit. Since the closest data points use the 65 nm CMOS generation, both figures can be extrapolated with an acceptable certainty to meet the requirements for WiNoC systems, i.e. a typical link distance \(\leq 1\) cm and data rates \(\geq 30\) Gbps.

Encouraged by recent demonstration of a 410 GHz oscillator based on 90 nm CMOS devices [20] and empowered by ongoing device scaling, RF-CMOS circuitry will play a central role in the ultra low power integration up to 600 GHz [21]. For the acceptable noise and gain performance beyond 150 GHz, the use of SiGe BiCMOS technology, which integrates ultrafast SiGe heterojunction bipolar transistors (HBT) with sufficient gain performance, will be crucial in an otherwise purely CMOS architecture [22]. Such hybrid SiGe BiCMOS solutions, already popular for high-throughput optical modulators operating around 30 Gbps, are the most practical route to surmounting the impasse between ultra-low power performance and high frequency operation. To illustrate this trend, we refer to Figure
7 which shows measured DC power dissipation at state-of-the-art power amplifiers (PAs) based on high-performance III-V devices (high electron mobility transistors - HEMTs), SiGe HBTs and RF-CMOS technology, as a function of carrier/modulation frequency. SiGe HBTs are more suitable for WiNoCs due to their power levels and material engineering techniques on silicon bipolar transistors compared to high performance III-V HEMTs with poor integration potential. While CMOS devices do not yet match the frequency response needed for low-noise amplifier (LNA) and PA designs around 500GHz, the ongoing device scaling and process refinement appears to scale up the frequency response exactly at the right direction. Additionally, circuit engineering and better understanding of devices in a given technology generation can bring about significant reduction in power levels, thus making CMOS circuits a very strong contender for WiNoC implementation in the long term. The trend lines in Figure 6 show that CMOS circuits are moving towards target WiNoC data rates near 32 Gbps and energies near 1 pJ/bit. Furthermore, this trend line is in accordance with the energy and data rates found in related works which shown energies of 0.33 pJ/bit [2] and 4.5 pJ/bit [4] as well as data rates of 32 Gbps [2], [8].

4.3 Proposed Wireless Technology
The wireless transceiver technology in A-WiNoC must be energy-efficient and produce high data rates. Double-gate transistors are excellent high-performance devices that will endow mature RF-CMOS platforms with unique tunable capabilities via the additional gate used for dynamic threshold control and additional signal (de)modulation [27]. Therefore, we use DG-MOSFETs (FinFETs with two independent gates), that will be introduced to fabrication lines in 2013 by several leading manufacturers, as an excellent basis for a reconfigurable WiNoC technology that can reach the projected 150 GHz CMOS operation without the use of more power hungry SiGe HBT counterparts.

Due to their energy efficient and compact nature, simple on-off keying (OOK) transceivers are considered as the most suitable platform for building WiNoCs [17]. Based on the RF-CMOS trends in Figures 6 & 7 and best practices in OOK transceiver design, each transceiver will be built using 22 nm DG-CMOS devices and consume 32 mW (1 pJ/bit*32 Gbps), 6 mW of which will be used by the PA. Although the design of a fully developed transceiver architecture is beyond the scope of this work, we can exemplify the use of DG-CMOS in novel circuit engineering approaches to lower power consumption and provide reconfigurability in WiNoC applications via a PA design. Since PAs determine the amplitude of the transmitted signals and are often the dominant consumer of power and area within transceivers, such an example should be especially meaningful. In order to determine the appropriate signal levels and the required amplification levels, a link-budget analysis is presented in Figure 8. According to this figure, which considers losses in air and a 10 dB error margin, typical signal levels for a 30 Gbps link over a 1 cm distance are below -13 dBm.

With the signal levels determined from Figure 8 for a particular data rate and distance, we can decide the required gain for a WiNoC link. While such an allocation will be permanent for static links, it may be dynamically chosen in a reconfigurable one to save power. Figure 9 shows a practical PA design for carrying out such a dynamic
allocation using 32 nm DG-CMOS devices up to 100 GHz. Using the additional back gates in this novel breed of MOSFETs, it is possible to tune the gain typically by 5 to 10 dB [28]. Although the limitations of the current device model and the simulator prevents us from extending this design to 500 GHz at this time, the general transistor-scaling trends indicate that they can comfortably operate at this range when scaled to 15 nm level as foreseen by the ITRS roadmap (2011 edition). Most importantly the same approach can be used in other components such as the LNA in the receiver as well as oscillator, mixer and filter circuits to build a truly reconfigurable and compact WiNoC router that can adapt well to the changing link requirements.

4.4 Antenna Considerations

For large frequencies, the design of the antenna can employ conventional antenna theory. However, for low/moderate operating frequencies, additional power must be transmitted to compensate for the reduced antenna efficiency when the antennas are “electrically small” (l ≪ λ). For an example, a patch antenna of area 0.9 mm², mounted on a CMOS substrate and operating at 60 GHz, was analyzed and measured in [29] with gains ranging from approximately 7 dB to -9 dB. Use of such an antenna at both Tx and Rx would require from 14 to 18 dB larger transmit power than if an omnidirectional antenna of gain 0 dB were used. Thus increasing antenna gain (directivity) is a prime concern which cannot be tackled via traditional approaches such as use of large aperture antennas or arrays, due to size limitations. Luckily, several novel solutions can be adapted for compact high gain antennas including special materials as in [30], where a micro-strip patch antenna design with gain approximately 8 dB was obtained with approximately 70% radiation efficiency in the THz band. Additional solutions for antennas as well as inductors can be also pursued on non-CMOS platforms that can be can be flip-bonded to the main chip or built on top of the planarized passivation layers or via the bonding wires. Thus, despite the challenges, we assume that the emission and reception of signals up to 600 GHz via planar (metallic) elements in approximately 100 µm scale can be attainable, given the time scale expected for WiNoC deployment.

5 PERFORMANCE EVALUATION

In this section, we compare A-WiNoC to electrical NoC designs including mesh, Concentrated Mesh (CMesh) [15], and Flattened Butterfly (FB) [31] architectures and the wireless networks WCube [4] and iWISE [8]. A packet size of four 64 bit flits was used. The router uses a four stage pipeline with four VCs each four flits deep. CMesh has a concentration of four cores and the electrical networks use XY routing. For a fair comparison, the bisectional bandwidth for all networks was kept the same by adding cycle delays. Additional cycle delays were added for wired links longer than 5 mm. We assume a total wireless bandwidth of 512 GHz. A-WiNoC uses 16 wireless channels each 32 Gbps and each wired link is 64 bits wide with a network clock of 1 GHz. All results consider the token overhead including latency and energy.

For open-loop measurement, we varied the network load from 0.1-0.9 of the network capacity. The simulator was warmed up under load without taking measurements until steady state was reached. Then a sample of injected packets were labeled during a measurement interval. The simulation was allowed to run until all the labeled packets reached their destinations. All designs were tested with different synthetic traffic traces such as (1) Uniform Random (UN), where each node randomly selects its destinations with equal probability and (2) Permutation Patterns, where each node selects a fixed destination based on the permutations. We evaluated the performance on the following permutation patterns: Bit-Reversal (BR), Butterfly (BFLY), Matrix Transpose (MT), Complement (COMP) and Perfect Shuffle (PS). We also tested on two different loads, a non-uniform random (NUR) and workload completion traffic traces. In NUR, 25% of the traffic is directed to a certain destination node creating hot-spot traffic with the rest being uniform random traffic.

For closed-loop measurement, the full execution-driven simulator SIMICS from Wind River [13] with the memory package GEMS [14] was used to extract traffic traces from

![Fig. 9: A tunable gain PA design based on 32nm DG-CMOS devices, with a wide-band performance up to 100GHz.](image)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1/L2 coherence</td>
<td>MOESI</td>
</tr>
<tr>
<td>L2 cache size/assoc</td>
<td>4MB/16-way</td>
</tr>
<tr>
<td>L2 cache line size</td>
<td>64</td>
</tr>
<tr>
<td>L2 access latency (cycles)</td>
<td>4</td>
</tr>
<tr>
<td>L1 cache/assoc</td>
<td>64KB/4-way</td>
</tr>
<tr>
<td>L1 cache line size</td>
<td>64</td>
</tr>
<tr>
<td>L1 access latency (cycles)</td>
<td>2</td>
</tr>
<tr>
<td>Core Frequency (GHz)</td>
<td>5</td>
</tr>
<tr>
<td>Threads (core)</td>
<td>2</td>
</tr>
<tr>
<td>Issue policy</td>
<td>In-order</td>
</tr>
<tr>
<td>Memory Size (GB)</td>
<td>4</td>
</tr>
<tr>
<td>Memory Controllers</td>
<td>16</td>
</tr>
<tr>
<td>Memory Latency (cycle)</td>
<td>160</td>
</tr>
<tr>
<td>Directory latency (cycle)</td>
<td>80</td>
</tr>
</tbody>
</table>

TABLE 1: Cache and core parameters used for Splash-2, PARSEC, and SPEC2006 application suite simulation.
real applications. The Splash-2 [11], PARSEC [10], and SPEC CPU2006 [12] workloads were used to evaluate the performance of 64-core networks. Table 1 shows the parameters for the cache and core used for the Splash-2, PARSEC, and SPEC2006 benchmarks. We assume a 2 cycle delay to access the L1 cache, a 4 cycle delay for the L2 cache, and a 160 cycle delay to access main memory. For Splash-2 traffic, the assumed kernels and workloads are as follows: FFT (16K particles), LU (512 × 512 with a block size of 16 × 16), Radiosity (Largeroom), Raytrace (Teapot), Radix (1 Million integers), Ocean (258 × 258), FMM (16K particles) and Water (512 Molecules). We consider seven PARSEC applications with medium inputs (blacksholes, facesim, fluidanimate, freqmin, streamcluster, ferret, and swaptions) and three workloads from SPEC CPU2006 (bzip, gcc base, and hmmer). The energy and area results for the NoC components were estimated using the Synopsys Design Compiler with the 40 nm TSMC technology library. In the following sections, we will compare A-WiNoC to other networks by providing energy and area estimates along with speedup and throughput simulation results.

### 5.1 Throughput

Figure 10 shows the throughput for the 64 core networks for four different mixes of synthetic traffic. The different patterns in each traffic mix is shown in Table 2. The patterns were chosen in order to stress the network in a variety of ways. For example, mix 0 has MT and NBR patterns to represent a mix of both short and long distance traffic. NUR was included to create a hot spot of traffic in order to test the effectiveness of adaptability. For each mix, the traffic randomly switches between the different patterns every 500 cycles. The reconfiguration window of A-WiNoC is R=100 cycles. iWISE serves as our non-adaptable baseline.

For mix 0, A-WiNoC shows an increase in throughput between 7% and 65%. For mix 1, A-WiNoC shows an increase in throughput between 7%-46%. Both of these mixes use NUR traffic which creates a hot spot. The main reason for the increase in throughput is mainly due to the reconfiguration algorithm which gives more bandwidth to hot spots. For mix 2, A-WiNoC shows a decrease of 11% in throughput compared to FBfly and mesh. This is due to the more uniform mix of traffic patterns which is beneficial for the long links of FBfly and the non-concentrated mesh network. A uniform mix balances the load across all links, thereby having few under-utilized links. However, A-WiNoC still increases throughput by at least 29% over iWISE, CMesh, and WCube due to the BFLY and MT patterns in the mix. For mix 3, A-WiNoC shows a throughput higher all other networks. Mix 3 is the only mix with four traffic patterns. As the traffic changes between these four patterns, the reconfiguration algorithm adapts the network accordingly.

### 5.2 Speedup

Figure 11 shows the speedup on real applications for three different miss status handling registers (MSHR) that allow
2, 4, or 8 requests at a time per core. A core sends a 1 flit request to another core which will send back a 4 flit response for a mix of short and long traffic. The total execution time of mesh relative to the other networks for each application is the speedup. For a MSHR of 2, A-WiNoC has an average speedup of 2.59X over mesh as well as a 48% improvement over WCube. This is mainly because of the one-hop diameter of A-WiNoC which is possible due to our architecture utilizing long wireless links and our fair token scheme. The performance of A-WiNoC and iWISE are similar due to the overall uniform pattern and low traffic load of many of the benchmarks. The uniform nature of the Splash-2 benchmarks leave few links under-utilized. On the other hand, the adaptability of A-WiNoC improves the performance over iWISE for the slightly less uniform PARSEC and SPEC CPU2006 benchmarks. As the MSHR increases from 2 to 8, the network load will increase. This results in A-WiNoC improving its average speedup over iWISE from 4.4% (MSHR=2) to 8.5% (MSHR=4) to 11.1% (MSHR=8). Although the improvement of the reconfiguration is increasing with network load, the improvement of A-WiNoC relative to the other networks is decreasing. The speedup of A-WiNoC over mesh decreases from 2.59X (MSHR=2) to 2.17X (MSHR=4) to 1.4X (MSHR=8). This decrease in improvement may be due to the type of utilization used in the reconfiguration algorithm. Link utilization is used which is effective for low-medium loads, but less effective at higher loads [16].

5.3 Energy

Figure 12 shows the energy of each network when at saturation for the traffic patterns of uniform random (UN), non-uniform random (NUR), bit reversal (BR), butterfly (BFLY), complement (COMP), matrix transpose (MT), and perfect shuffle (PS). The energy is broken down into wired, wireless, and router energy. The energy consumption, including dynamic and static energy, of a whole flit traversing a wireless link, a 5 mm wired link, a baseline 5x5 crossbar and a buffer are shown in Table 3. The energy overhead for the reconfiguration controllers, GC and LC, are very small compared to the other router components.

A-WiNoC has an average energy savings of 35% over CMesh. The main reason for these savings are due to the use of the low energy wireless links. A-WiNoC shows a reduction in electrical wire energy dissipation for all traffic patterns. Furthermore, A-WiNoC has an average energy savings of approximately 25% over WCube. These savings are due to the higher ratio of wireless transmission compared to wired transmissions in A-WiNoC. By using a token sharing scheme, more wireless links can be used compared to the centralized wireless hubs of WCube. However, the many wireless links of A-WiNoC increases the router inputs and outputs, thereby increasing the crossbar.

Fig. 11: Speedup on real applications for a MSHR that allows 2, 4, or 8 requests.

Fig. 12: Energy breakdown for different traffic patterns for A-WiNoC and other wireless/wired networks.

TABLE 3: Power and Area estimates from Synopsys Design Compiler with the 40 nm TSMC library for a 64 bit flit.

<table>
<thead>
<tr>
<th>Component</th>
<th>Power (pJ)</th>
<th>Area (nm²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wireless Link</td>
<td>64</td>
<td>0.05-0.10</td>
</tr>
<tr>
<td>5 mm Wired Link</td>
<td>102</td>
<td>0.0394</td>
</tr>
<tr>
<td>Baseline Crossbar</td>
<td>7.5</td>
<td>0.0273</td>
</tr>
<tr>
<td>Packet Buffer</td>
<td>4.0</td>
<td>0.002949</td>
</tr>
<tr>
<td>GC</td>
<td>0.09627 μm²</td>
<td>1.41 μm²</td>
</tr>
<tr>
<td>LC</td>
<td>0.09664 μm²</td>
<td>1.42 μm²</td>
</tr>
</tbody>
</table>
size and energy. This causes A-WiNoC to have the largest router energy dissipation for most traffic patterns. However, the one-hop nature of A-WiNoC reduces the number of crossbar traversals. Overall, the slight increase in router energy can be compensated for by the large savings in link energy.

Across different traffic patterns, A-WiNoC improves energy over FBfly between 7% for BFLY traffic and 58% for MT. The differences across different traffic patterns are due to the total number of wired link traversals in each network. In traffic patterns such as MT and COMP, there is a high percentage of long distance traffic. With many packets traversing from one edge of the chip to the other, the energy dissipation due to wired links will be high in the electrical networks. However, in A-WiNoC the low energy wireless links can be utilized more and there will be a large energy savings. WCube is also a wireless network, but the centralized wireless hubs create more electrical hops as packets must route from the source to the wireless hub then from another wireless hub to the destination. In traffic patterns such as BFLY, there is less long distance traffic. This type of traffic causes the energy dissipation of the electrical networks to be lower and more competitive with A-WiNoC and WCube. A-WiNoC has energies similar to iWISE since the communication patterns are similar with the exception that iWISE has a wireless communication link to its own set.

Next, we examine the throughput/energy (TPE) cost metric. A network with a high throughput/energy indicates an efficient network. We compare A-WiNoC to various wired and wireless networks using the traffic patterns UN, NUR, BR, BFLY, COMP, MT, and PS. A-WiNoC has an average TPE of 38.7 Gbps/nJ which is 15% lower than WCube due to the low energy cost of A-WiNoC. The TPE of iWISE is 37.9 Gbps/nJ which is approximately 2% lower than A-WiNoC. These two networks perform similarly because the average energy of both networks are similar but the throughput of A-WiNoC is slightly higher. A-WiNoC has a higher TPE than the wired networks (29% over mesh, 46% over FBfly, 102% over CMesh) due to both a higher throughput and lower energy of A-WiNoC.

### 5.4 Area
Table 3 shows the area estimates for the wireless link, a 5 mm wired link, a 5x5 crossbar, and a buffer for a flit. For the wireless transceiver area, from our study of existing trends we estimate the transceiver area to be between 0.05 \(mm^2\) and 0.1 \(mm^2\). A-WiNoC will have a total network area increase of 1.7-2.2X over the mesh network and an increase between 1.8-2.4X over FBfly. This increase is due to the area of the wireless links and the increase in router size. A router in A-WiNoC will have a size between 11x11 to 13x13 depending on its location in the topology. Corner routers will be 11x11 due to fewer wired ports, other routers around the edge of the topology will be 12x12, and the routers in the center of the network will be 13x13. This area increase is the trade-off for the throughput, speedup, and energy benefits. The area overhead of the GC and LC are negligibly small compared to the other router components.

### 5.5 Sensitivity Study
In this section, we evaluate the effect of various changes to the A-WiNoC network. The first change is using a second adaptable transmitter. 4T-1A is A-WiNoC as described earlier with 4 transmitters per set; 1 of which is adaptable (4T-1A). 4T-2A is A-WiNoC with 4 wireless transmitters per set, 2 of which are adaptable. 4T-2A will increase the number of receivers required at each router, but will provide more adaptability and up to 3X data rate to one set. Another
disadvantage of using a second reconfigurable link is that a set may be disconnected. For example, if both adaptable transmitters in Set 0 get reconfigured to Set 1 and the two fixed transmitters send to Set 1 and Set 3 then Set 2 will become disconnected. To solve this, we allocate 50% of $R_w$ for transmission to the busiest set and the other 50% for transmission to the “disconnected” set.

Figure 13 shows the saturation throughput of different traffic mixes for 4T-2A compared to the baseline-iWISE and other electrical/wireless networks. The reconfiguration window for 4T-2A is 100 again. The traffic mixes are the same as before. However, the figure also shows results for the traffic changing every 100, 250, 500, 1000, 2000, or 4000 cycles. First, 4T-2A has an average higher throughput of 12% for mix 0, 1.5% for mix 1, 5.3% for mix 2, and 9.3% for mix 3 compared to 4T-1A. This is expected as the additional reconfigurable link adds more bandwidth to hot spots. The instances where 4T-1A outperforms 4T-2A may be due to the disconnected set that is caused by 2A. Additionally, differences may be due to the randomness of the mixes. During simulation 4T-1A may have had a more favorable traffic pattern for a longer period of time. Second, as the traffic period changes from 100 to 4000 cycles, the saturation throughput of 4T-2A seems to stay fairly similar with spikes for some traffic change periods. The volatile nature of the mixes in traffic may cause the throughput to saturate at varying loads. However, averaged over all traffic mixes, 4T-2A saturates at a load approximately 41% higher than WCube while 4T-1A saturates 34% higher than WCube.

The next study evaluates the effect of changing the reconfiguration window, R, of A-WiNoC. Figure 14 shows the speedup on real application for different R=100, 500, or 1000. Also included in the results is 3T which is A-WiNoC with 3 fixed wireless transmitters; one for each other set. Figure 14(a), and 14(b) show speedup relative to 3T for a MSHR allowing up to 2 and 4 at a time per core. A MSHR allowing 8 requests was also evaluated but the figure was omitted due to space constraints. On average, 4T-1A with R=100 has the highest speedup. R=100 performs the best compared to other R values because it is the smallest and can adapt quicker to the changes in traffic. The advantage of a higher R is that link utilization needs to be calculated less which can save some power. For the Splash-2 benchmarks, there is little difference between the different reconfiguration windows. This is due to the uniformity of the Splash-2 benchmarks. The PARSEC and SPEC CPU2006 benchmarks show a much higher speedup for R=100, 500, and 1000. As the MSHR increases from 2 to 8, the speedup of R=100 increases from 1.04X to 1.11X to 1.16X. This increase is due to an increasing network load that results from a larger MSHR. A higher network load means that the adaptable wireless link can be utilized more.

### 5.6 Scalability

A-WiNoC is scaled to a larger number of cores by maintaining the same wireless communication but adding more cores per set as explained in Section 3.2. To evaluate the effect of adding more cores to a set, we scale A-WiNoC to 256 cores by creating sets with 64 cores each. The saturation throughput for 256 core networks is shown in Figure 15 for four different mixes of synthetic traffic. Real application benchmarks were not evaluated due to the large size of the networks. It is assumed that the traffic changes every 500 cycles and the reconfiguration window of A-WiNoC is 100 cycles. A-WiNoC has a throughput approximately 33.4% higher than mesh on average. The wireless links of A-WiNoC and WCube allow packets to avoid additional hops, increasing throughput. Additionally, the adaptability of A-WiNoC increases the saturation throughput 37.2% over WCube on average and 7.9% over iWISE. In mix 0, WCube saturates at a throughput approximately 17% higher than A-WiNoC due to less wireless traffic in this mix. In most mixes, iWISE outperforms WCube due to the distributed wireless links. However, the lack of adaptability in iWISE causes a lower throughput compared to A-WiNoC. CMesh and FBfly have the lowest throughput due to the concentration of cores and long wired delays. Therefore, A-WiNoC is able to scale to a larger number of cores with minimal performance overhead by adding more cores to each set and maintaining the same wireless communication.

Figure 16 shows the normalized energy of an average packet for the wired and wireless networks when the number of cores scales to 256. The electrical networks mesh, CMesh, and FBfly consume a high energy due to the long electrical links and high router degree, similar to 64 cores. On average, A-WiNoC consumes 40% less energy than mesh. Energy-efficient wireless links contribute
to these power savings. Additionally, A-WiNoC has comparable energy values to the wireless networks WCube and iWISE, consuming 1.4% less energy than WCube and 3.8% more energy than iWISE. Since iWISE assumes more wireless bandwidth at 256 cores, the increase in wireless link causes more wireless link traversals, decreasing energy. Compared to 64 cores, the energy improvement may be less depending on traffic patterns due to the increase in wired link traversals. The limited wireless bandwidth demands wireless routers to become more centralized, increasing hop count. However, the energy savings of wireless links is still great enough to lower overall energy consumption.

6 CONCLUSIONS

The trends in wireless technologies have shown that on-chip wireless interconnects are a potential solution to alleviate the higher power and latency of metallic NoCs. We proposed a hybrid architecture called A-WiNoC which uses adaptable wireless transceivers with low energies (∼1 pJ/bit) and high data rates (∼32 Gbps). We design a reconfiguration algorithm to adapt to traffic patterns and a token sharing scheme to fully utilize wireless bandwidth. A 64 core and a 256 core A-WiNoC design are discussed which take advantage of the limited wireless bandwidth. Our determined frequency band is 50-500 GHz and we show path loss at various frequencies. Since a low energy, high data rate NoC wireless transceiver has not yet been realized in current technologies, we use trends in RF-CMOS devices and DG-CMOS technology to estimate parameters for our OOK wireless transceivers. Our results on real applications show a 1.4-2.6X speedup and our energy estimates from the Synopsys Design Compiler show an energy savings of 25-35% over wireless and electrical networks. Furthermore, our reconfiguration algorithm improves throughput by an additional 18%. The scalability results of A-WiNoC shows that throughput can be increased by 37% and energy can be improved by 21% at 256 cores.

ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation grants ECCS-1129010, ECCS-1342657, CCF-1054339, and CNS-1342984.

REFERENCES

Avinash Karanth Kodi received the Ph.D. and M.S. degrees in Electrical and Computer Engineering from the University of Arizona, Tucson in 2006 and 2003 respectively. He is currently an Associate Professor of Electrical Engineering and Computer Science at Ohio University, Athens. He is the recipient of the National Science Foundation (NSF) CAREER award in 2011. His research interests include computer architecture, optical interconnects, chip multiprocessors (CMPs) and network-on-chips (NoCs).

David Matolak received his B.S. degree from Pennsylvania State University, University Park, his M.S. degree from the University of Massachusetts, Amherst, MA, and the Ph.D. degree from the University of Virginia, Charlottesville, all in electrical engineering. He has worked for over 20 years on communication systems, with the Rural Electrification Administration, Washington, DC, the UMass LAMMDA Laboratory, Amherst, AT&T Bell Laboratories, North Andover, Massachusetts, the University of Virginia’s Communication Systems Laboratory, Lockheed Martin Tactical Communication Systems, Salt Lake City, Utah, the MITRE Corporation, McLean, Virginia, and Lockheed Martin Global Telecommunications, Reston, Virginia. From 1999 to August 2012 he was with the School of Electrical Engineering and Computer Science at Ohio University, and since August 2012 he has been with the Department of Electrical Engineering at the University of South Carolina.

Savas Kaya obtained his PhD in 1998 from Imperial College of Science, Technology and Medicine, London, for his work on strained Si quantum wells on vicinal substrates, following his MPhil in 1994 from the University of Cambridge. He was a post-doctoral researcher at the University of Glasgow between 1998-2001, carrying out research in transport, and fluctuation phenomena in decanano MOSFETs, and fluctuation phenomena in decanano MOSFETs. He is currently with the Russ College of Engineering at Ohio University, Athens. His other interests include transport theory, device modeling and process integration, nanofabrication, nanostructures and nanosensors.

Soumyasanta Laha obtained his MSc. in Embedded Digital Systems with distinction from the University of Sussex, UK in 2007. Since 2008, he is with the Russ College of Engineering, Ohio University pursuing a PhD in Electrical Engineering in the area of nanoscale energy efficient RF Transceivers. He also has more than three years of industrial work experience in India and the UK in Embedded Systems and Analog Electronics.

Dominic DiTomaso received his B.S. and M.S. degrees in Electrical Engineering and Computer Science from Ohio University, Athens in 2010 and 2012. He is currently pursuing his PhD degree in Electrical Engineering and Computer Science at Ohio University. His research interests include wireless interconnects, network-on-chips (NoCs) and computer architecture.

William Rayess received his B.E. in computer and communications engineering from Notre Dame University in Lebanon in 2008, a MCTP from Ohio University in 2009, and is currently pursuing his PhD in Electrical Engineering at the Russ College of Engineering, Ohio University.