Energy-efficient Adaptive Wireless NoCs Architecture

Dominic DiTomaso, Avinash Kodi, David Matolak, Savas Kaya, Soumyasanta Laha, and William Rayess
School of Electrical Engineering and Computer Science
Ohio University, Athens, OH 45701
E-mail: dd292006, kodi, matolak, kaya@ohio.edu

***Abstract*—With the increasing number of cores in chip multiprocessors, the design of an efficient communication fabric is essential to satisfy the bandwidth and energy requirements of multi-core systems. Scalable Network-on-Chip (NoC) designs are quickly becoming the standard communication framework to replace bus-based networks. However, the conventional metallic interconnects for inter-core communication consume excess energy and lower throughput which are major bottlenecks in NoC architectures. On-chip wireless interconnects can alleviate the power and bandwidth problems of traditional metallic NoCs. In this paper, we propose an adaptable wireless Network-on-Chip architecture (A-WiNoC) that uses adaptable and energy efficient wireless transceivers to improve network power and throughput by adapting channels according to traffic patterns. Our adaptable algorithm uses link utilization statistics to re-allocate wireless channels and a token sharing scheme to fully utilize the wireless bandwidth efficiently. We compare our proposed A-WiNoC to both wireless/electrical topologies with results showing a throughput improvement of 65%, a speedup between 1.4-2.6X on real benchmarks, and an energy savings of 25-35%.**

I. Introduction

The shrinking of silicon technology has given rise to chip multiprocessors (CMPs) that integrate hundreds to thousands of cores on a single chip. The traditional bus-based networks which connect these cores do not scale well due to high energy and latency bottlenecks. Additionally, with higher clock frequencies, the dissipated power rises and more clock cycles are required for data to traverse the bus. Network-on-Chip (NoC) designs are the response to the limitations of bus-based networks [1]. NoCs can provide high bandwidth communication for CMPs. However, the metal wires that connect cores in conventional NoC designs suffer from high energy costs and long propagation delays due to routers and intermediate hops, respectively. Additionally, the multi-hop communication of traditional NoC topologies such as a mesh or torus further increase power and latency [2]. Even though metal wires have limitations at long distances, they can still be highly effective and suitable for short distance communication. A 1 mm metal wire in 32 nm complementary metal-oxide-semiconductor (CMOS) technology consumes a low energy of 0.18 pJ/bit.

One potential solution is wireless interconnects that can alleviate the limitations of metal wires by providing low latency and energy efficiency [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. The work by Chiang et al. [4] used wireless interconnects operating at 2 pJ/bit to create long distance (30 cm) links between computing chassis. The RF-Interconnect [7] placed an radio frequency (RF) transmission line in a zig-zag pattern to transmit packets quickly across the chip at 1.2 pJ/bit. Ganguly et al. [3] created a hybrid network that organized cores into subnets in which communication within a subnet was wired and communication between subnets used 0.33 pJ/bit wireless links with a total 512 GHz bandwidth. The hybrid WCube design [5] used a wired mesh on one tier and a wireless hypercube network on the second tier with wireless transceivers operating at 4.5 pJ/bit with a 512 GHz bandwidth. Lastly, iWISE [6] was a hybrid network which distributed 1 pJ/bit wireless links to avoid additional hops to centralized hubs. These related works used long links to quickly propagate data across the chip at very low energies and designed hybrid networks to provide an additional 512 GHz wireless bandwidth without the area overhead and complexity of metal wires. However, each of these related NoCs used fixed wireless links that cannot adapt to dynamic traffic patterns during runtime. With the limited wireless spectrum, it is critical that wireless links be fully utilized by adapting them to traffic patterns.

Therefore, we propose to use energy efficient transceivers in our Adaptable Wireless NoC Architecture (A-WiNoC) to create a one-hop, low power design while improving network performance through adaptable transceivers. With the limited wireless spectrum, we use wired links for local communication while reserving wireless links for global communication. Wireless interconnects have been proven to be a low energy alternative to prior work; however, the inherent adaptability of wireless links have not been utilized before. Often, channel bandwidth can be under-utilized in real applications due to dynamic traffic patterns. We use adaptable wireless links to improve performance by adapting them to traffic demands, thereby, efficiently utilizing network resources. The major contributions of this work include:

(1) Adaptability: We use adaptability to give more bandwidth to hot spots caused by dynamic traffic patterns. Our adaptive algorithm reallocates transmission time slots to these high traffic spots to lower contention and improve performance.

(2) Energy Efficient Devices: We show that trends in various emerging fabrication technologies such as sub-50nm RF-CMOS and SiGe BiCMOS are moving towards wireless transceivers with the energies and data rates near the NoC

978-1-4673-6493-5/13/$31.00 ©2013 IEEE

requirements of ∼1 pJ/bit and ∼32 Gbps.

(3) Evaluation on Real Benchmarks: We evaluate our novel A-WiNoC architecture compared to other wired/wireless networks on the benchmarks PARSEC, SPLASH-2, and SPEC CPU2006 by collecting traces from SIMICS and GEMS [8]. We evaluate the network throughput and show an improvement of up to 65% as well as a speedup between 1.4X and 2.6X. Using the Synopsys Design Compiler, A-WiNoC was estimated to have an energy savings of 25% over a wireless network and up to 35% over electrical networks.

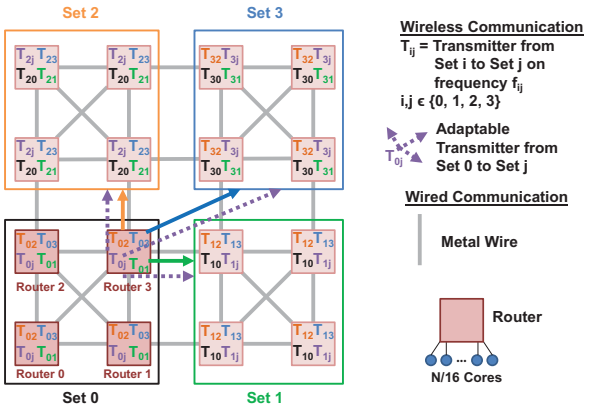
II. Adaptable Wireless NoC Architecture

A. NoC Design

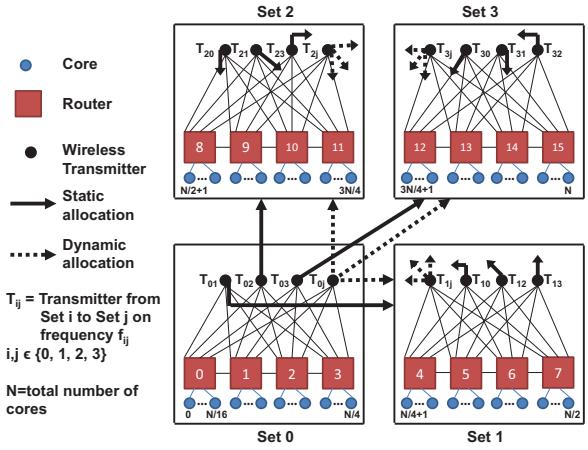
Architecture: The trends of wireless transceivers (Section III) show very low energies and high data rates making them ideal for our NoC architecture. The proposed architecture called A-WiNoC: Adaptable Wireless NoC Architecture is shown in Figure 1(a). The architecture has a total of N cores where N=64 in this paper. To minimize energy dissipation and reduce packet latency, we concentrate four cores by connecting to a single router [9] (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 0 has cores 1 to N/4, Set 1 has cores N/4+1 to N/2, Set 2 has cores N/2+1 to 3N/4, and Set 3 has cores 3N/4+1 to N (Also seen in the simplified Figure 1(b)). This creates four sets, each with four routers. Each router has four transmitters: T_{ij}, which indicates a transmitter from Set i to Set j. All the routers in each set share these four wireless transmitters. As explained in [6], the choice of four routers and four sets optimizes wireless channel sharing by giving a set an opportunity for every router to transmit 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 in a mesh topology. Wired links are, therefore, used for short distances because 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.

Deadlocks: Our network avoids deadlocks by routing packets to their destination in one hop. If a packet's source node is exactly one wired hop away from its destination node, then a wired link is used. Otherwise, if the source is farther than one wired link, then a single wireless hop 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 deadlocking can be avoided as there is no circular dependency for packet transmission.



(a) Detailed Architecture



(b) Wireless Communication

Fig. 1: Adaptable wireless architecture showing (a) router and transceiver organization and (b) the logical wireless communication between sets.

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 their own unique set of carrier frequencies. With a total available bandwidth of 512 GHz, each wireless channel has a data rate of 32 Gbps. There are 12 *static* wireless channels which are used to transmit packets at low energy (see Figure 1(b)). Static channels allow the network topology to be connected at all times. An additional, four *adaptable* wireless channels can be adapted based on traffic patterns to give additional bandwidth to certain portions of the chip. Four adaptable wireless channels are used so that each set has at least one adaptable channel. More than four adaptable channel can be used; however, this will unnecessarily add to the complexity of the network. The total 16 wireless channels are shared among multiple transceivers which 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 by reducing hot spots.

For wireless communication, each set has four transmitters. Three transmitters are used as static communication and one transmitter can be adapted to any set. For example, in Set 0 of Figure 1a, T_{01}, T_{02}, T_{03} are statically allocated from Set 0 to Set 1, Set 2, and Set 3, respectively. T_{0j} can be adapted to any Set 1-3. The transmitters are replicated at each router in the set to avoid additional hops to a centralized wireless hub. Transmitters T_{01}, T_{02}, T_{03}, and T_{0j} are replicated at routers 0-3 in Set 0.

Figure 1(b) shows a simplified version of A-WiNoC to illustrate the wireless communication. Each set has four shared transmitters. The notation T_{ij} is used to indicate a transmitter from Set *i* to Set *j*. For example, Set 0 in Figure 1(a) uses the four transmitters: T_{01}, T_{02}, T_{03}, and T_{0j}. For each transmitter, T_{ij}, a unique set of frequencies, f_{ij}, is allocated to avoid interference. Therefore, with four sets the total number of wireless channels is 16. Three transmitters are statically configured to the other three sets which are shown as solid arrows. This ensures that all the sets are always connected. One transmitter is adaptable, shown as a dotted arrow, and can transmit to any set depending on the traffic pattern. The thin black lines in Figure 1(b) show that each router has all four transmitters available for transmission. However, only one router can use a single transmitter at a time. For example, in Set 0, router 0 can use any of the four transmitters in Set 0, but not at the same time as routers 1-3.

Tokens: Since multiple routers in a set have transmitters tuned to the same wireless channel, time division multiplexing (TDM) is used to assign time slots to a router. Time slots indicate when a router can use a certain transmitter in order to avoid interference. Time slots are assigned by implementing a token sharing scheme. Tokens are passed between routers and represent the right to transmit on a certain wireless channel. When a router possesses a token, it is immediately given a time slot and starts transmitting data. If no data needs to be transmitted, it passes the token to the next router. Tokens were used because they can be quickly passed between routers so that routers do not wait long to transmit data. There are 16 tokens representing the 16 wireless channels. Since each set shares four wireless channels, only four tokens need to be passed between the routers within a set.

Figure 2 shows one example of communication for Set 0 and Set 1 across two cycles. For Set 0, the four tokens, 01, 02, 03, and 0j are passed between routers 0-3 where j indicates a adaptable token that can be used to send to any set 1-3. For this example, Router 3 has the token to transmit to Set 3. Router 3 will transmit to every router in Set 3. Each router will look into the packet header, compare the packet destination with its own address, and either accept or reject the packet. This is called single write multiple read (SWMR). Likewise for router 2, the packet will be transmitted to all routers in Set 2 and the correct destination will accept the packet. This approach will consume more power; however, it will reduce the number of hops for the packet. Router 0 in Figure 2 has heavy traffic going to Set 1. Therefore, it can use the token for its static transmitter as well as the token for its adaptable transmitter to double the data rate to Set 1. After each transmitter sends a packet in cycle 0, it will immediately pass on the token. Routers which need the token will capture it and transmit in cycle 1. In Figure 2, during cycle 1, the tokens 01 and 0j are captured and used. Tokens 02 and 03 are idle since no router currently requires transmission to Sets 2 and 3. For Set 1 in Figure 2, routers 4, 6, and 7 use their tokens to transmit to Sets 0, 2, and 3, respectively. During cycle 1, routers 5 and 6 capture tokens to transmit to Sets 0 and 2, respectively. Router 7 can capture tokens 13 and 1j again in the next cycle if no other routers require the tokens.

To scale A-WiNoC to a higher number of cores, such as 256 or 512, more cores per set can be added. As the maximum wireless spectrum is being used, 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. Four cores are concentrated to a single router as before; however, each router is directly connected to a wireless router. Flow control for all network sizes will be managed by credits which can be piggybacked onto packets.

B. Adaptability

We adjust the duty cycle, or the duration in which our wireless links transmit. When signal duty cycle increases for any link, if the signaling rate stays constant, throughput increases proportionally. The increased throughput afforded by a duty cycle increase comes at the expense of a proportional increase in dissipated power, and of course at the expense of throughput for other links that are time-multiplexed on the same frequency channel.

Unlike previous wireless NoC architectures, we take advantage of the inherent adaptability of wireless interconnects. Adaptability is used in our architecture to give more bandwidth to sets with the most traffic. 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. The global controller (GC) determines to which set an adaptable transmitter should allocate its resources. The local controller (LC) collects statistics on each link utilization and indicates to which set the adaptable transmitter should reconfigure. Each LC_i is attached to one of the four wireless transmitters. 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.

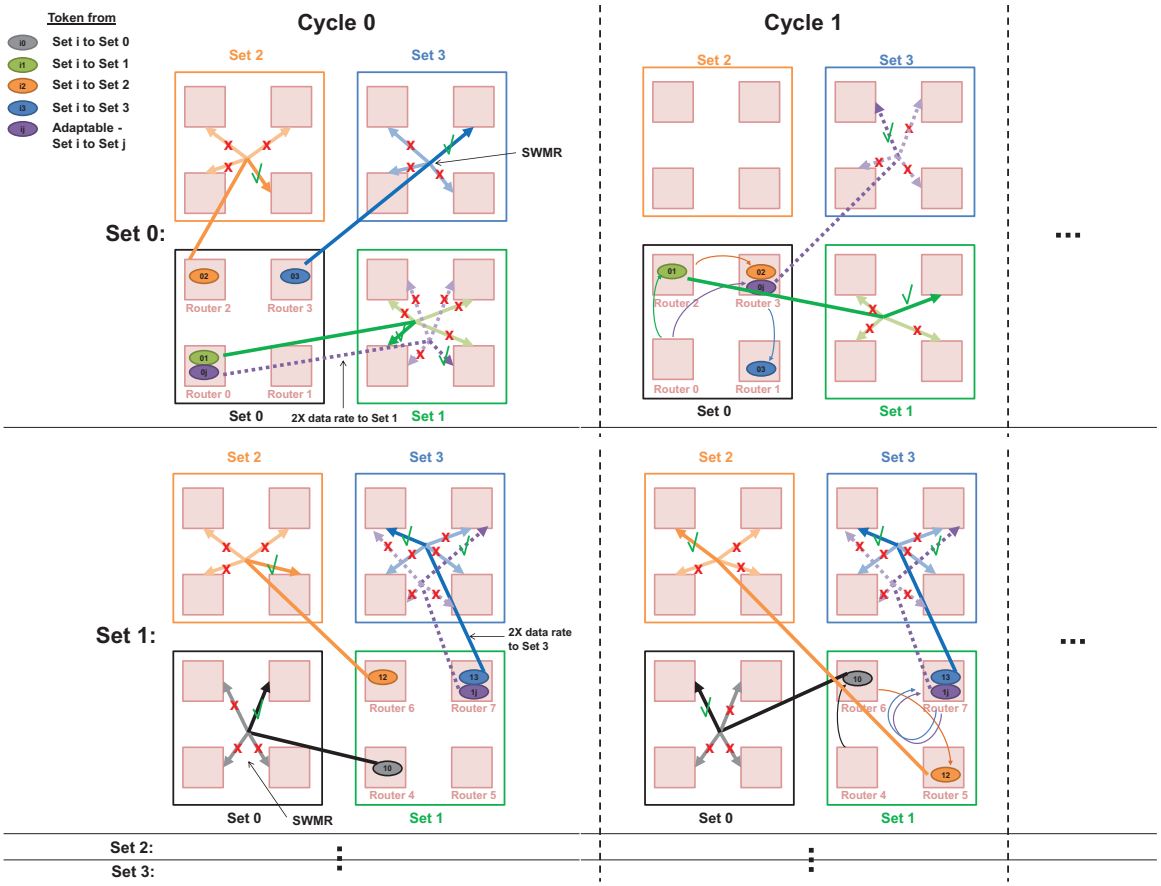


Fig. 2: Example of the token scheme for communication in Set 0 and Set 1 for two cycles.

The GC compares the data and determines which Set has the highest utilization. GC then communicates with LC_3, which is attached to the adaptable transmitter, to reconfigure to the set with the highest utilization.

The pseudo code for the adaptive algorithm is shown in Algorithm 1. GCs evaluate statistics and re-allocate resources for the current reconfiguration window, R_w, based on the previous R_w. After R_w, in Step 2, the GC will send a $Link_{Request}$ control packet to all LC_i, requesting utilization data. In Step 2a, each LC_i will update the field in the $Link_{Request}$ packet with the $Link_{util}$ information. The $Link_{util}$ information is the number of link traversals on the outgoing links and will be reset to zero to prepare for the next R_w. The $Link_{Request}$ packet is returned back to the GC.

In Step 3, GC receives $Link_{Request}$ packet containing the utilization information for all outgoing links for the previous R_w. In Step 3a, GC separates each $Link_{util}$ for each outgoing set: $Set0_{util}$, $Set1_{util}$, $Set2_{util}$, and $Set3_{util}$. The GC is able to separate utilization into different sets by knowing which set each transmitter sends to, including the adaptable transmitter. In Step 3b, GC finds the highest set utilization by using comparators and the utilization information from each set. The set with the maximum utilization should be the set to which the adaptable transmitter reconfigures.

In Step 4, GC sends a $Link_{Response}$ control packet to LC_3 which is attached to the adaptable transmitter. The $Link_{Response}$ packet requires two bits which will contain 00 if Set 0 has the highest utilization, 01 for Set 1, 10 for Set 2, and 11 for Set 3. In Step 4a, the adaptable transmitter , T_{ij}, reads the $Link_{Response}$ packet and reallocates time slots to the set with the maximum utilization. By reallocating time slots, T_{ij} will only accept packets destined for the reconfigured set during at least the time frame of R_w at which the algorithm is repeated.

Algorithm 1 Adaptive 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: $Set0_{util}$, $Set1_{util}$, $Set2_{util}$, and $Set3_{util}$,
Step 3b: GC finds max[$Set0_{util}$, $Set1_{util}$, $Set2_{util}$, $Set3_{util}$]
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

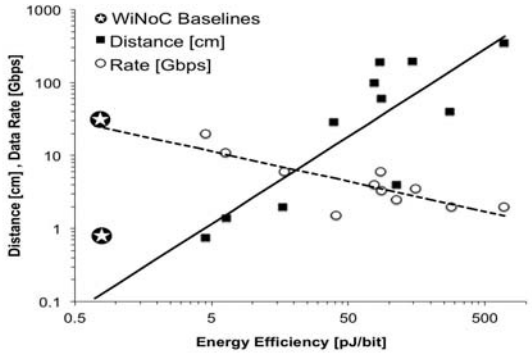


Fig. 3: Trends found in RF-CMOS transceivers designed for low-power and short-range links for WiNoC system requirements. Data adapted from [10], [11], [12].

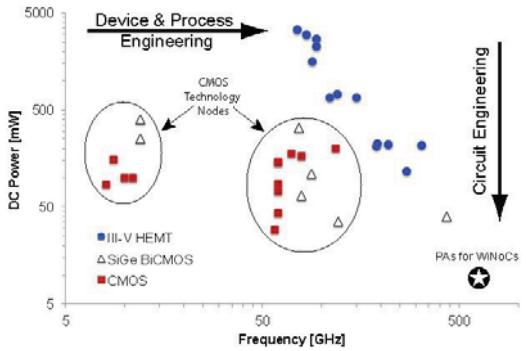


Fig. 4: Power amplifier trends in integrated transmitters implemented using compound (III-V) and silicon-based (SiGe HBT and CMOS) devices. Data collected from [17], [18], [19], [20]

III. Trends of Wireless Transceivers

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 (See Figure 3). In this summative plot, both data rate and link distance are plotted as a function of modulation energy efficiency, which must be lower than 1 pJ/bit for WiNoC systems to be able to compete with wired links. It appears that 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. While these objectives are not trivial to achieve, it is reassuring to note that they are within the reach of general trends in RF-CMOS, especially when the closest data points are considered for the link distance that use 65nm CMOS generation.

For low-power CMOS integration on silicon, the ongoing adaptation of sub-90nm RF-CMOS back-end solutions for vehicular radar at 77 GHz will be a critical starting point as it will provide a complete technology base with on-chip antennas as well as compact transceivers that can reach mass-markets [13]. Encouraged by recent demonstration of a 410 GHz oscillator based on 90nm CMOS devices [14] and empowered by ongoing device scaling, RF-CMOS circuitry will play a central role in the ultra low power integration up to 600GHz [15]. 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 [16]. Such hybrid SiGe BiCMOS solutions, already popular for high-throughput optical modulators operating around 30 Gbps, is 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 4 which shows measured DC power dissipation at state-of-the-art 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 LNA/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 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 3 show that CMOS circuits are moving towards target WiNoC data rates near 32 Gbps and energies near 1 pJ/bit.

IV. Antenna Considerations

The THz 'performance gap' is especially evident in antenna design [21]. On one side of the gap, the existing WLAN RF-CMOS radios use off-chip antennas because of their relatively low frequency of operation ($\leq$5.4GHz). On the other side, on-chip optical networks utilize infra-red free-space solutions with nanostructures used as efficient nanoantennas for resonant absorption. Also, the dominant applications in the 1 THz range have been medical and security imaging technologies that can do without on-chip integration. However, thanks to the vehicular anti-collision radar and multi-media driven indoor wireless network applications in the 60-90 GHz range, on-chip antenna solutions have recently gained momentum [13]. Thus, while much less developed compared to other pieces of the WiNoC puzzle, compact on-chip antenna solutions also appear to be within the reach of silicon mm-wave integration as will be discussed in more detail below.

The easiest case for the design will be when frequency is large enough to employ conventional antenna theory. Even in this case, analysis of mutual coupling and pattern deformation due to the WiNOC landscape may still be needed, which would require a rigorous 3D FDTD simulation that can only happen after actual digital floor design and wireless router placement. For low/moderate operating frequencies, additional

power must be transmitted to compensate for the reduced antenna efficiency when the antennas are "electrically small" ($l \ll \lambda$). For an example, a patch antenna of area 0.9 mm^2, mounted on a CMOS substrate and operating at 60 GHz, was analyzed and measured in [22] with gains ranging from $\sim$ 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 large aperture antennas or arrays due to size limitations. Moreover, high gain antennas are also needed for time-frequency resource reuse that rely on spatial isolation. Luckily, several novel solutions can be adapted for compact high gain antennas including special materials as in [21], where a micro-strip patch antenna design with gain $\sim$ 8 dB was obtained with approximately 70% radiation efficiency in the THz band, using a metamaterial substrate.

V. Performance Evaluation

In this section, we compare A-WiNoC to electrical NoC designs including mesh, Concentrated Mesh (CMesh), and Flattened Butterfly (FB) architectures and the wireless network WCube. A packet size of four 64 bit flits was used. For a fair comparison, the bisectional bandwidth for all networks was kept the same. Additional cycle delays were added for wired links longer than 5 mm. We assume a total wireless bandwidth of 512 GHz [5], [6]. Therefore, with the 16 channels in A-WiNoC, each wireless link is 32 Gbps.

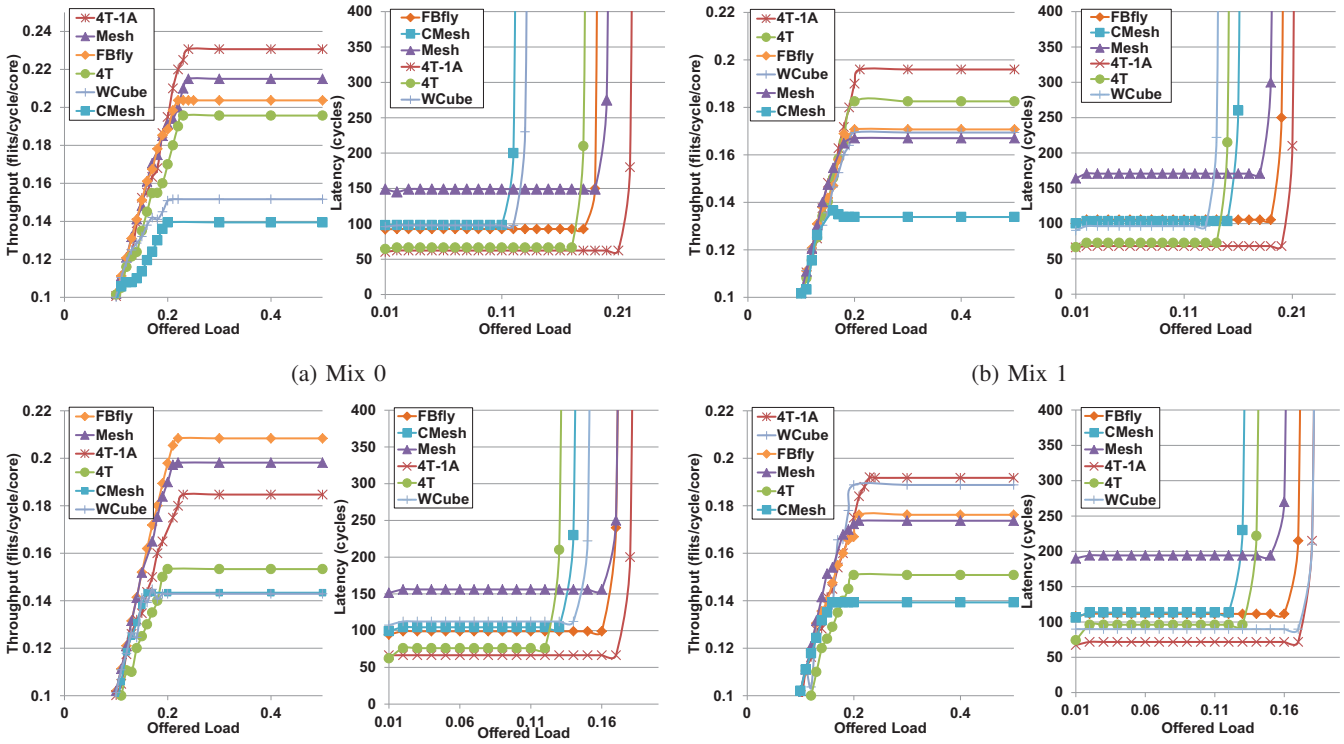
For open-loop measurement, we varied the network load from 0.1-0.9 of the network capacity. The simulator was warmed up under load without taking measurements until steady state was reached. Then a sample of injected packets were labeled during a measurement interval. The simulation was allowed to run until all the labeled packets reached their destinations. For closed-loop measurement, the full execution-driven simulator SIMICS from Wind River with the memory package GEMS was used to extract traffic traces from real applications. The Splash-2, PARSEC, and SPEC CPU200 workloads were used to evaluate the performance of 64-core networks. 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. 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.

A. Throughput and Latency

Figure 5 shows the throughput and latency for the 64 core networks for four different mixes of synthetic traffic. Mix 0 is a mix of non-uniform (NUR), matrix transpose (MT), and neighbor (NBR) traffic. Mix 1 is NUR, bit reversal (BR), and perfect shuffle (PS). Mix 2 is uniform (UN), butterfly (BFLY), and MT. Lastly, mix 3 is UN, BR, complement (COMP), and PS. For each mix, the traffic randomly switches between the different patterns every 500 cycles. 4T-1A is A-WiNoC as described earlier with 4 transmitters per set; 1 of which is adaptable (4T-1A). R=100 indicates that the reconfiguration window is 100 cycles. 4T serves as our non-adaptable baseline in which 4T is A-WiNoC with 4 transmitters per set; none of which are adaptable. For mix 0, 4T-1A shows an increase in throughput between 7% and 65%. For mix 1, 4T-1A 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, 4T-1A 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, 4T-1A still increases throughput by at least 29% over 4T, CMesh, and WCube due to the BFLY and MT patterns in the mix. For mix 3, 4T-1A shows a throughput higher than all other networks. As the traffic changes between four patterns, the reconfiguration algorithm adapts the network accordingly. The latency plots show the networks saturating at a similar point as the throughput plots. However, the low load latencies show that the wireless links of 4T and 4T-1A consistently have a lower latency than the other networks. Networks such as mesh have high hop counts and networks such as FBfly have long wired links which cause high latencies.

B. Speedup

Figure 6 shows the speedup on real applications for a miss status handling registers (MSHR) that allow 2 requests at a time per core. Simulations for a MSHR=4 and 8 were also evaluated, but not shown. A core sends a 1 flit request to another core which will send back a 4 flit response. For a MSHR of 2, 4T-1A 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 4T-1A and 4T 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 4T-1A improves the performance over 4T for the slightly less uniform PARSEC and SPEC CPU2006 benchmarks. As the MSHR increases from 2 to 8, the network load will be increasing. This results in 4T-1A improving its average speedup over 4T 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 4T-1A 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.



(a) Mix 0

(b) Mix 1

(c) Mix 2

(d) Mix 3

Fig. 5: Throughput and Latency for different mixes of traffic with traffic changing every 500 cycles.

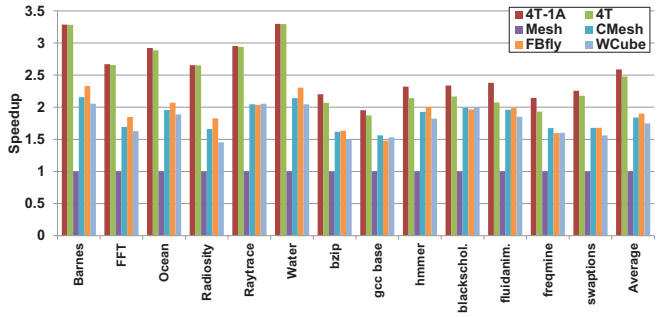


Fig. 6: Speedup on real applications.

C. Energy

Figure 7 shows the energy of each network for each of the traffic patterns. The energy consumption of a whole flit traversing a wireless link, a 5 mm wired link, a 5x5 crossbar and a buffer are shown in Table I. 4T-1A has an average energy savings of 35% over CMesh. The main reason for the savings is 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, 4T-1A has an average energy savings of approximately 25% over WCube. This savings is due to the higher ratio of wireless transmission compared 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 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 is 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.

D. Area

Table I shows the area estimates for a wireless link, a 5 mm wired link, a 5x5 crossbar used in mesh, and a buffer for a flit. For the wireless transceiver area, from our study of existing

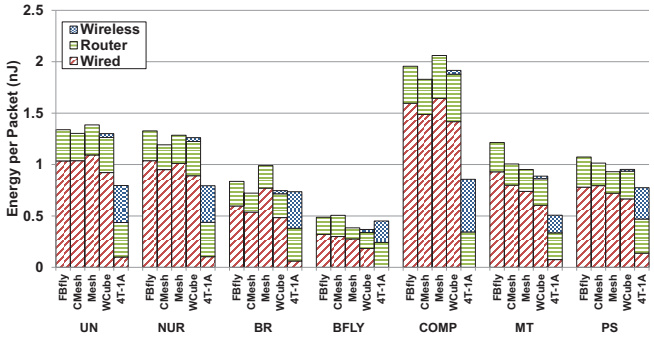


Fig. 7: Energy breakdown for different traffic patterns for A-WiNoC and other wireless/wired networks.

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

	Energy (pJ)	Area (mm^2)
Wireless Link	64	0.05-0.10
5 mm Wired Link	102	0.0394
5x5 Crossbar	7.5	0.0273
Packet Buffer	4.0	0.002949

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 be between 11x11 to 13x13 ports depending on its location in the topology. This area increase is the trade-off for the throughput, speedup, and energy benefits.

VI. 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 one-hop, hybrid architecture called A-WiNoC which uses adaptable wireless transceivers with low energies ($\sim$1 pJ/bit) and high data rates ($\sim$32 Gbps). We design a reconfiguration algorithm to adapt to traffic patterns and a token sharing scheme to fully utilize wireless bandwidth. 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.

Acknowledgements

We thank the anonymous reviewers for their excellent feedback. This work was partially supported by the National Science Foundation grants CCF-0915418, CCF-1054339 (CAREER) and ECCS-1129010.

References

[1] W. J. Dally and B. Towles, "Route packets, not wires: On-chip interconnection networks," in *Proceedings of Design Automation Conference (DAC)*, June 2001, pp. 684–689.

[2] J. Davis, R. Venkatesan, A. Kaloyeros, M. Beylansky, S. Souri, K. Banerjee, K. Saraswat, A. Rahman, R. Reif, and J. Meindl, "Interconnect limits on gigascale integration (gsi) in the 21st century," *Proceedings of the IEEE*, vol. 89, no. 3, pp. 305–324, Mar. 2001.

[3] A. Ganguly, K. Chang, S. Deb, P. Pande, B. Belzer, and C. Teuscher, "Scalable hybrid wireless network-on-chip architectures for multi-core systems," *IEEE Transactions on Computers*, vol. 60, pp. 1485–1502, August 2010.

[4] P. Y. Chiang, S. Woracheewan, C. Hu, L. Guo, R. Khanna, J. Nejedlo, and H. Lui, "Short-range, wireless interconnect within a computing chassis: Design challenges," *IEEE Design and Test of Computers*, vol. 27, no. 4, pp. 32–43, July 2010.

[5] S. B. Lee, S. W. Tam, I. Pefkianakis, S. Lu, M. F. Chang, C. Guo, G. Reinman, C. Peng, M. Naik, L. Zhang, and J. Cong, "A scalable micro wireless interconnect structure for CMPs," *Mobicom '09*, pp. 217–228, September 2009.

[6] D. DiTomaso, A. Kodi, S. Kaya, and D. Matolak, "iWISE: Inter-router wireless scalable express channels for network-on-chips (NoCs) architecture," *19th Annu. IEEE Symp. High-Performance Interconnects*, pp. 11–18, Aug. 2011.

[7] M. Chang, J. Cong, A. Kaplan, M. Naik, G. Reinman, E. Socher, and S. Tam, "CMP network-on-chip overlaid with multi-band RF-interconnect," *IEEE International Symposium on High Performance Computer Architecture*, pp. 191–202, February 2008.

[8] P. Magnusson, M. Christensson, J. Eskilson, D. Forsgren, G. Hallberg, J. Hogberg, F. Larsson, A. Moestedt, and B. Werner, "Simics: A full system simulation platform," *Computer*, vol. 35, no. 2, pp. 50–58, February 2002.

[9] J. Balfour and W. J. Dally, "Design tradeoffs for tiled cmp on-chip networks," in *Proceedings of the 20th ACM International Conference on Supercomputing (ICS)*, Cairns, Australia, June 28-30 2006, pp. 187–198.

[10] J. Gorisse, D. Morche, and J. Jantunen, "Wireless transceivers for gigabit-per-second communications," in *IEEE International NEWCAS*, June 2012, pp. 545–548.

[11] C. Wang, W.-H. Hu, and N. Bagherzadeh, "A wireless network-on-chip design for multicore platforms," in *19th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*, Feb. 2011, pp. 409–416.

[12] J. Lee, Y. Chen, and Y. Huang, "A low-power low-cost fully-integrated 60-ghz transceiver system with ook modulation and on-board antenna assembly," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 2, pp. 264–275, Feb. 2010.

[13] Y.-A. Li, M.-H. Hung, S.-J. Huang, and J. Lee, "A fully integrated 77ghz fmcw radar system in 65nm cmos," in *IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, Feb. 2010, pp. 216 –217.

[14] O. Momeni and E. Afshari, "High power terahertz and millimeter-wave oscillator design: A systematic approach," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 3, pp. 583 –597, Mar. 2011.

[15] U. Pfeiffer, E. Ojefors, A. Lisauskas, and H. Roskos, "Opportunities for silicon at mmwave and terahertz frequencies," in *Bipolar/BiCMOS Circuits and Technology Meeting*, Oct. 2008, pp. 149–156.

[16] H. Rucker, B. Heinemann, and A. Fox, "Half-terahertz sige bicmos technology," in *IEEE 12th Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems (SiRF)*, Jan. 2012, pp. 133–136.

[17] L. Samoska, "An overview of solid-state integrated circuit amplifiers in the submillimeter-wave and thz regime," *IEEE Transactions on Terahertz Science and Technology*, vol. 1, no. 1, pp. 9–24, Sept. 2011.

[18] N. Deferm and P. Reynaert, "A 120ghz 10gb/s phase-modulating transmitter in 65nm lp cmos," in *IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, Feb. 2011, pp. 290–292.

[19] S. Hu, L. Wang, Y. Z. Xiong, B. Zhang, and T. G. Lim, "A 434ghz sige bicmos transmitter with an on-chip siw slot antenna," in *IEEE Asian Solid State Circuits Conference (A-SSCC)*, Nov. 2011, pp. 269–272.

[20] R. Minami, K. Matsushita, H. Asada, K. Okada, and A. Matsuzawa, "A 60 ghz cmos power amplifier using varactor cross-coupling neutralization with adaptive bias," in *Asia-Pacific Microwave Conference Proceedings (APMC)*, Dec. 2011, pp. 789–792.

[21] G. Singh, "Design considerations for rectangular microstrip patch antenna on electromagnetic crystal substrate at terahertz frequency," *Elsevier Journal of Infrared Physics and Technology*, vol. 53, pp. 17–22, 2010.

[22] D. Titz, F. B. Abdeljelil, S. Jan, F. Ferrero, C. Luxey, P. Brachat, and G. Jacquemod, "Design and characterization of cmos on-chip antennas for 60 ghz communications," *Radioengineering Journal*, vol. 21, no. 1, pp. 324–331, April 2012.