Exploring Wireless Technology for Off-Chip Memory Access

Md Ashif I Sikder, Dominic DiTomaso, Avinash Kodi, Savas Kaya
School of Electrical Engineering and Computer Science
Ohio Univeristy
Athens, OH 45701
Email: (ms047914, dd292006, kodi, kaya)@ohio.edu

William Rayess and David Matolak
Department of Electrical Engineering
University of South Carolina
Columbia, SC 29208
Email: (rayess, matolak)@cec.sc.edu

***Abstract*—The trend of shifting from multi-core to many-core processors is exceeding the data-carrying capacity of the traditional on-chip communication fabric. While the importance of the on-chip communication paradigm cannot be denied, the off-chip memory access latency is fast becoming an important challenge. As more memory intensive applications are developed, off-chip memory access will limit the performance of chip multi-core processors (CMPs). However, with the shrinkage of transistor dimension, the energy consumption and the latency of the traditional metallic interconnects are increasing due to smaller wire widths, longer wire lengths, and complex multi-hop routing requirements. In contrast, emerging wireless technology requires lower energy with single-hop communication, albeit with limited bandwidth (at a 60 GHz center frequency). In this paper, we have proposed several hybrid-wireless architectures to access off-chip memory by exploiting frequency division multiplexing (FDM), time division multiplexing (TDM), and space division multiplexing (SDM) techniques. We explore the design-space of building hybrid-wireless interconnects by considering conservative and aggressive wireless bandwidths and directionality. Our hybrid-wireless architectures require a maximum of two hops and show 10.91% reduction in execution time compared to a baseline metallic architecture. In addition, the proposed hybrid-wireless architectures show on an average 62.07% and 32.52% energy per byte improvement over traditional metallic interconnects for conservative and aggressive off-chip metallic link energy-efficiency respectively. Nevertheless, the proposed hybrid-wireless architectures incur an area overhead due to the higher transceiver area requirement.**

I. Introduction

Technology scaling is primarily driving the growth in the number of processing cores on a single chip and thus, increasing the design complexity of the on-chip communication fabric called Network-on-Chip (NoC). However, due to the pin bandwidth limitation, the number of memory controllers used to access the off-chip dynamic random access memory (DRAM) is not proportionally increasing with the number of cores [1]. In a traditional mesh-based NoC architecture, the memory controllers are connected to the corner routers due to pin restrictions. Therefore, as core count increases, packets would require more hops to access off-chip memory which would increase the latency and energy consumption. For example, with private L1 and shared L2 caches, the on-chip communication delay which comprises of request packet from L1 to L2 and L2 to DRAM, and corresponding response packet from DRAM to L2 and L2 to L1 can be as high as 400 cycles [2]. Moreover, the metallic link connecting the memory controller to DRAM cannot be traversed in a single processor cycle due to fundamental electrical signaling limitations [3] which incurs additional latency for off-chip memory accesses. Thus, the overall performance of CMPs is limited due to the increased on-chip and off-chip latencies to access DRAM.

The problem of higher off-chip memory access latency can be addressed in two potential ways: (1) by reducing the processing core to memory controller (request message) latency and memory controller to processing core (response message) latency, and/or (2) reducing the link traversal latency that connects the memory controller to the DRAM. For the first solution, as connecting all the cores directly to the memory controllers using metallic links is not convenient, positioning the memory controllers carefully on the chip would dramatically reduce the latency [1]. However, this would partially solve the problem because the processing cores further away from the memory controller will still see significant latency. In addition, the transmission latency and energy requirement for a metallic link increases with the length [3]. For a standard DDR3 [4] or DDR4 with parallel termination [5], the trace length is approximately 2 inch with a transmission latency that exceeds one clock cycle. Moreover, the energy/bit requirement for the off-chip metallic link is about 5-10 pJ which is $\sim$0.2 pJ for an on-chip metallic link calculated using Dsent [6]. Furthermore, according to the International Technology Roadmap for Semiconductors (ITRS), the development of traditional metallic interconnects would not be sufficient to support the growing number of chip multi-core processors (CMPs) as metallic interconnects do not scale due to the increased energy and multi-hop requirements. Alternately, one could explore emerging technologies such as wireless to overcome the off-chip memory access latency and energy constraints.

Wireless technology offers several advantages over the metallic technology such as (1) distance independent one-hop communication, (2) lower energy requirement compared to a long metallic link, (3) multicasting and broadcasting with omnidirectionality, and (4) absence of any physical channels. However, on-chip wireless technology has limited bandwidth at 60 GHz center frequency and is not energy efficient at shorter distances. Hence, metallic interconnects are used for short distance communications whereas wireless

interconnects are used for long distance communications using frequency division multiplexing (FDM), time division multiplexing (TDM), and space division multiplexing (SDM) to overcome the bandwidth limitation [7], [8].

In this paper, we propose to use wireless technology for off-chip memory access to improve latency (execution time) and energy-efficiency. We explore the design-space of building hybrid-wireless interconnects by considering conservative and aggressive wireless bandwidths and directionality. We restrict the use of wireless technology to communicate between the on-chip routers and the memory controllers and between the memory controllers and the DRAM. We use a combination of FDM, TDM, and SDM to overcome the wireless bandwidth limitation. Our results indicate that, compared to the baseline metallic architecture, the proposed hybrid-wireless architectures can reduce the execution time by 10.91% with an average energy/byte reduction of 62.07% and 32.52% for conservative and aggressive off-chip metallic link transmission respectively. The major contributions of this paper are as follows:

1) **Wireless technology to access off-chip memory**: Traditionally, metallic technology is used to communicate with the DRAM. In this paper, we propose to use wireless technology to connect the memory controller and DRAM. We explore the advantages and disadvantages of using wireless technology for this purpose. In addition, we use wireless technology on the chip to reduce the router-to-memory controller latency. Therefore, we use wireless technology to send a message from a router to a memory controller, from a memory controller to the DRAM, from the DRAM to a memory controller, and from a memory controller to a router.
2) **Design-space exploration of hybrid-wireless architectures**: We propose several hybrid-wireless architectures to access off-chip memory. The proposed architectures consider different wireless bandwidth and directionality with different combination of FDM, TDM, and SDM. Moreover, we explore the impact of utilizing wireless technology at the on-chip level, at the off-chip level, and for both on-chip and off-chip levels. We compare the performance of the architectures to find the optimum use of wireless technology.
3) **Wireless technological feasibility study**: The primary concerns of a wireless design are the cross-talk between wireless channels and the limited wireless bandwidth at 60 Ghz center frequency. Therefore, we propose a helical antenna model that can provide the necessary bandwidth while resolving the interference issue. We perform a full-wave simulation of the proposed antenna using HFSS and show the channel characterization.

II. Related Work

Limited pin bandwidth and link contention to access off-chip memory are increasing the latency in CMPs. As the number of cores is increasing, wiring memory controller to each core is not practical. However, changing the location of the memory controllers on the chip has shown both latency and bandwidth improvement [1], but this cannot alleviate all off-chip memory access contentions. Sharifi et al. [2] proposed to reduce the off-chip memory access latency by prioritizing the messages that were delayed and/or destined for less utilized memory banks. This prioritization is achieved by giving priority to the flits in the virtual channel and switch arbitration stages and/or pipelining router stages. In [9], the authors proposed inter and intra-chip hybrid-wireless communication for multiple multicore chips using code division multiple access (CDMA) and mm-wave antenna. Inter-chip communications use wireless technology whereas intra-chip communications employ metallic or wireless technology.

Paper [1] placed memory controllers intelligently to reduce the off-chip memory access latency but did not propose any new communication network. In contrast, we propose novel network designs for off-chip memory access. Sharifi et al. [2] revealed the components responsible for off-chip memory access latency but focused only on the arbitration and scheduling techniques to reduce the latency. However, our goal is to reduce the latencies showed in [2] by employing wireless technology for off-chip memory access. Shamim et al. [9] focused on reducing multi-chip communication latency using CDMA technique whereas we are interested in off-chip memory access latency and energy reduction using FDM, TDM, SDM, and directionality. To the best of our knowledge, none of the other scholarly articles have employed wireless technology to reduce off-chip memory access latency and energy requirement.

TABLE I
Naming Convention of the Architectures

General Name Format*: (On-chip)-(Off-chip)-(Antenna Type)-(Bandwidth)	
”M”	stands for Metallic link
”W”	stands for Wireless link
”D”	stands for Directional Antenna
”O”	stands for Omnidirectional Antenna
”A”	stands for Aggressive assumption for wireless BW (512 Gbps)
”C”	stands for Conservative assumption for wireless BW (128 Gbps)

*”Antenna Type” and ”Bandwidth (BW)” stand only for wireless networks

III. Proposed Architectures

In this paper, all the proposed and baseline architectures are 16-core tile-based architecture where each tile consists of a processing core, two caches, and a router. The first level cache (L1) is private to the core and the last level cache (L2) is distributed among the cores. Each router is connected to the caches via input and output ports, neighboring routers, and memory controllers (if applicable). The memory controllers are considered as a switch that can arbitrate between multiple requests [1]. In addition, we have assumed distributed off-chip memory where each memory module is serviced by a specific memory controller. We use metallic links to communicate between L1 and L2 for all the architectures. However, the proposed hybrid-wireless architectures use wireless links to communicate between L2 and memory controller (we term this communication as *L2MC*) and/or to communicate between

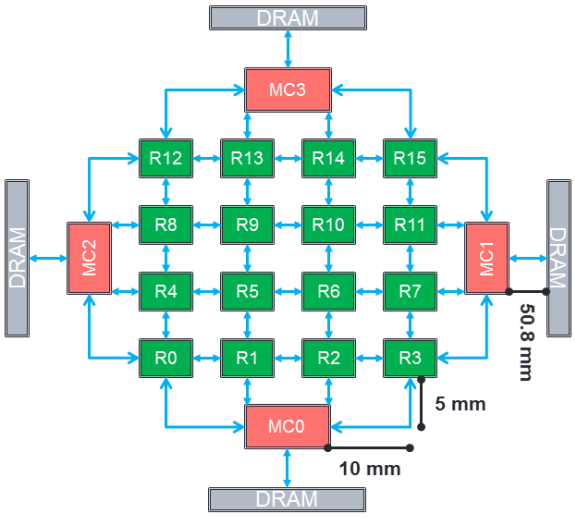


Fig. 1. Baseline architecture (M-M-X-X) with both L2 to memory controller (L2MC) and memory controller to DRAM (MCM) communications employing metallic interconnects.

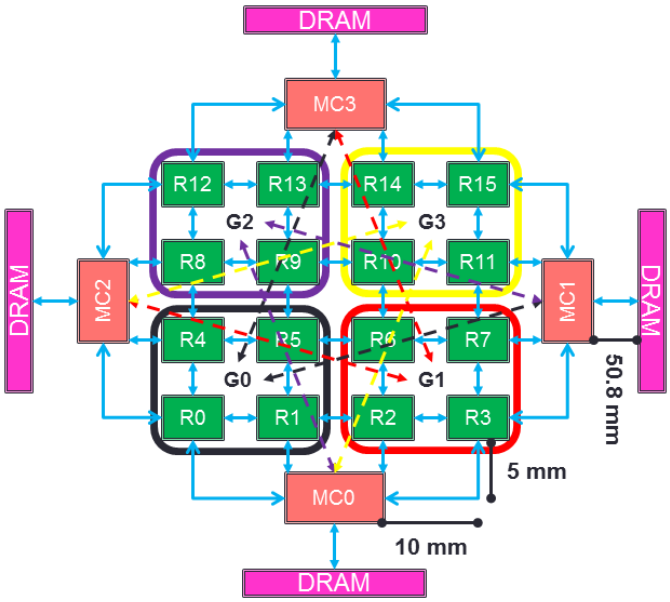


Fig. 2. General structure of proposed hybrid-wireless architectures (W-M-X-X). Wireless interconnects are used for L2 to memory controller (L2MC) and metallic interconnects are used for memory controller to DRAM (MCM) communications.

memory controller and DRAM (we term this communication as *MCM*). The bandwidth of the on-chip router-to-router and router-to-memory controller metallic links are assumed to be 64 bits whereas the off-chip metallic link bandwidth is 128 bits. The bandwidth of a wireless link is calculated by dividing the total wireless bandwidth with the number of wireless links.

The naming convention of the architectures used in this paper are given in Table I. For example, consider the architecture M-M-X-X. The first and second letter "M" suggests that both the L2MC and MCM communication use metallic links. Because the metallic interconnects are not constrained in terms of bandwidth and cannot be categorized as different types, the last two parts are written as "X" (Don't care). The name W-M-O-A indicates that the architecture uses wireless interconnects for L2MC communication, and metallic interconnects for MCM communication. "O" and "A" state that the antenna used for wireless network is omnidirectional in nature and the overall bandwidth is 512 Gbps respectively (shown in Table I). In the following sub-sections, we will discuss the baseline and the proposed architectures with proposed routing mechanism in detail.

A. Metallic Interconnect (M-M-X-X)

The structure of M-M-X-X is shown in Figure 1. It is used as the baseline architecture to compare the performance of the proposed architectures. We consider the router-to-router distance as 5 mm, the shortest router-to-memory controller distance as 5 mm [1], the longest router-to-memory controller distance as 10 mm, and the trace length as 50.8 mm (2 inch) for DDR3 or parallel terminated DDR4 technology [4], [5]. We have placed the memory controllers at the edge of the chip to provide maximum connectivity between the memory controllers and the routers using metallic links. The tradeoff is lower link and router contention with longer links that require more energy and area.

B. Hybrid Wireless Interconnect (W/M-W/M-X-X)

On top of the baseline architecture, M-M-X-X, hybrid wireless architectures are built by inserting wireless links for L2MC and/or MCM communications. Therefore, on-chip wireless links are used to transfer messages to and from memory controllers, and off-chip wireless links replace the metallic links that connect the memory controller to the DRAM. *Wireless bandwidth is determined by the technology and the antenna used, and is not the same for all the proposed architectures*. Different types of hybrid wireless architectures proposed in this paper are discussed next.

1) On-chip Hybrid-Wireless Interconnect (W-M-X-X): The routers of W-M-X-X use wireless technology to send request messages and receive response messages from the distant memory controllers that are more than two hops apart. However, metallic links are used for router-to-router and nearby (require a maximum of two hops) router-to-memory controller communications. One such general architecture is shown in Figure 2. The on-chip routers are divided into four groups where each group contains four routers. Each group is assigned a unique frequency channel to transmit messages to the distant memory controllers while metallic links are used for nearby memory controllers. Similarly, each memory controller is assigned a unique frequency channel to transmit data to the distant router-groups while it uses metallic links for nearby router-groups. We have considered two types of antennas- omnidirectional and directional and two wireless bandwidth assumptions- conservative and aggressive. This provides us with four different hybrid-wireless architecture designs. However, we have not considered W-M-O-C in this work as wireless bandwidth of 512 Gbps for omnidirectional

type antenna is well established [10], [8]. Other architectures considered are described below in more detail:

- **W-M-O-A:** As shown in Figure 2, the routers of a group share the frequency channel assigned to that group for sending messages to the memory controllers. For example, group G0 is assigned a frequency channel to send message to the memory controllers MC1 and MC3. The routers (R0, R1, R4, R5) of G0 share the frequency channel using a token to maintain signal integrity. As omnidirectional antenna is used for wireless communication, both MC1 and MC3 can receive the data at the same time and then discard if the message is not destined for it. Similarly, memory controller MC1 uses a frequency channel to send data to the groups G0 and G2 (R8, R9, R12, R13). Therefore, each router of a group contains one transmitter to send data to the distant memory controllers and two receivers to receive data from the distant memory controllers. In the same fashion, each memory controller contains a transmitter to send data to the distant router groups and two receivers to receive data from the distant router groups. In summary, W-M-O-A requires eight wireless channels with each having a bandwidth of 64 Gbps (=512 Gbps/8).
- **W-M-D-A:** The basic architecture of W-M-D-A is similar to W-M-O-A architecture. It requires the same number of wireless channels and the same wireless bandwidth per channel as W-M-O-A. However, two antennas are required to send data because the antenna used for wireless communication is of directional type in W-M-D-A. For example, router R0 of group G0 contains two transmitters- one for sending data to memory controller MC1 and the other to MC3. When router R0 has the token to transmit, it uses one of the two transmitters depending on the destination memory controller. Similarly, the memory controller say MC1 uses two transmitters to send data to the routers of groups G0 and G2. Both the transmitters of a router or a memory controller are tuned at the same frequency and do not operate at the same time. Although twice the number of transmitters is required in W-M-D-A as compared to W-M-O-A, the number of receivers required is the same.
- **W-M-D-C:** The structure of W-M-D-C is the same as W-M-D-A architecture except for the wireless bandwidth used. The wireless link bandwidth of W-M-D-C is 1/4 times the wireless link bandwidth of W-M-D-A. That is, the wireless link bandwidth of each wireless channel in W-M-D-C is 16 Gbps (=128 Gbps/8). Hence, the latency for W-M-D-C would be higher than W-M-D-A.

2) Off-chip Hybrid-Wireless Interconnect (M-W-X-X): M-W-X-X has the same on-chip architecture as of M-M-X-X. However, M-W-X-X employs wireless links to communicate with the off-chip memory instead of metallic link shown in Figure 1. For this purpose, each memory controller contains a transmitter and a receiver tuned at the frequency of the corresponding DRAMs transmitter. Hence the DRAM needs to facilitate a transmitter, and a receiver tuned at the frequency of the corresponding memory controllers transmitter. Therefore, M-W-X-X requires eight unidirectional wireless channels for MCM communication- each with a bandwidth of 64 Gbps (=512 Gbps/8). In this paper, we have only shown the off-chip hybrid wireless interconnect- M-W-O-A which uses omnidirectional antenna with a total bandwidth of 512 Gbps because of the space constraints and the performance similarity compared to other architectures.

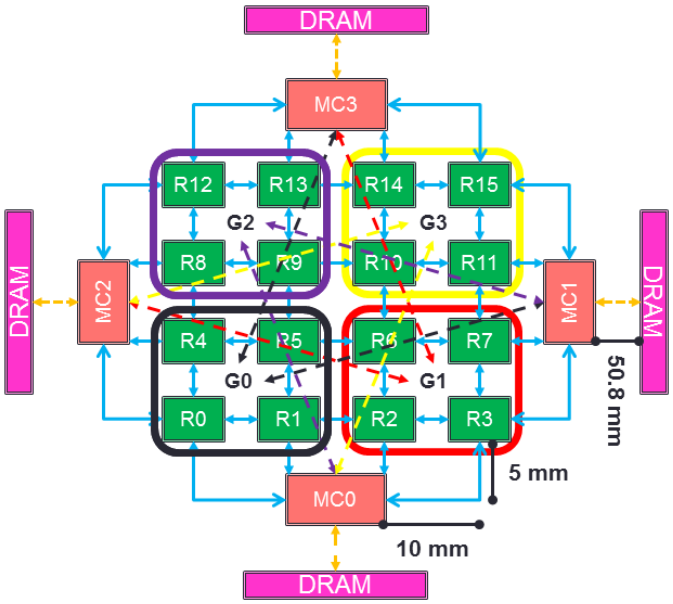


Fig. 3. General structure of the proposed on-chip and off-chip hybrid-wireless architectures (W-W-X-X). Wireless interconnects are used for both L2 to memory controller (L2MC) and memory controller to DRAM (MCM) communications.

3) On-chip and Off-chip Hybrid-Wireless Interconnect (W-W-X-X): This architecture combines the hybrid-wireless architectures- W-M-X-X and M-W-X-X. One such architecture is shown in Figure 3. As can be seen, W-W-X-X uses wireless technology for both L2MC and MCM communications to access off-chip memory. Hence, W-W-X-X requires 16 wireless channels (8 for L2MC and 8 for MCM communication). Since both the L2MC and MCM communications use wireless technology, we use SDM technique to overcome the frequency bandwidth limitation. The end result is that we can provide an on-chip wireless bandwidth of 16 Gbps (conservative) or 64 Gbps (aggressive) with an off-chip wireless bandwidth of 32 Gbps (conservative) or 128 Gbps (aggressive). All the architectures described above are summarized in Table II.

C. Communication Protocol: Metallic and Hybrid Wireless Interconnect

In this paper, we assume that each core requests necessary instruction and data from its private L1 cache. If there is a L1 miss, then a request message is sent through the router it is connected to the L2 cache containing the necessary data. On a L2 miss, a request message is sent to the memory controller that is servicing the memory module with the latest data. After

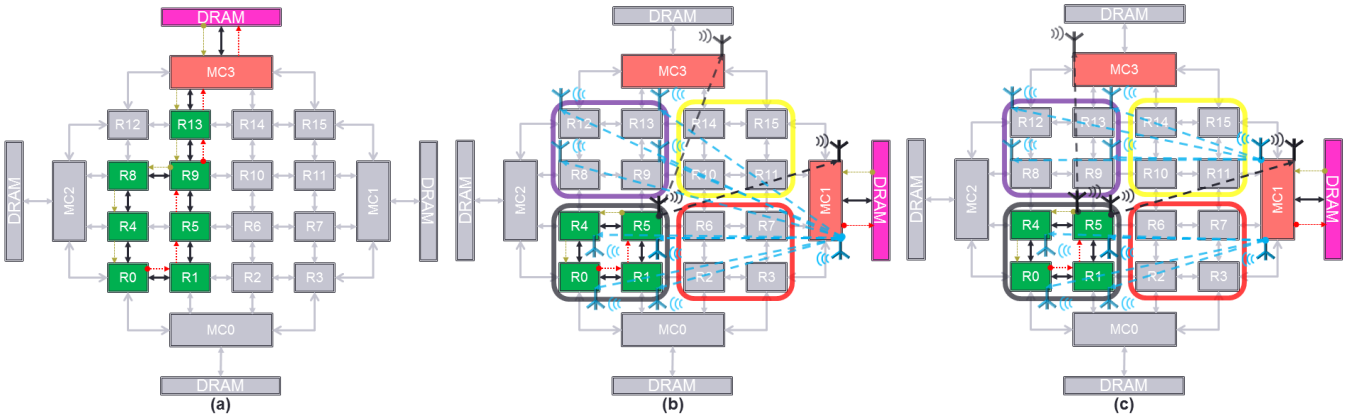


Fig. 4. Communication mechanism of the baseline and the proposed architectures. (a) On-chip metallic and off-chip metallic or wireless interconnects. (b) On-chip wireless interconnects with omnidirectional antenna and off-chip metallic interconnects. (c) On-chip wireless interconnects with directional antenna and off-chip metallic or wireless interconnects.

TABLE II
SUMMARY OF ARCHITECTURES

M-M-X-X	Metallic on-chip interconnects, and metallic off-chip interconnects (link BW 128 Gbps)
W-M-O-A	Hybrid wireless on-chip interconnects with omnidirectional antenna, metallic off-chip interconnects (link BW 128 Gbps), and total on-chip wireless bandwidth is 512 Gbps
W-M-D-C	Hybrid wireless on-chip interconnects with directional antenna, metallic off-chip interconnects (link BW 128 Gbps), and total on-chip wireless bandwidth is 128 Gbps
W-M-D-A	Hybrid wireless on-chip interconnects with directional antenna, metallic off-chip interconnects (link BW 128 Gbps), and total on-chip wireless bandwidth is 512 Gbps
M-W-O-A	Metallic on-chip interconnects, off-chip wireless interconnects (link BW 64 Gbps) with omnidirectional antenna, and total off-chip wireless bandwidth is 512 Gbps
W-W-D-C	Hybrid wireless on-chip interconnects with directional antenna, total on-chip wireless bandwidth is 128 Gbps, off-chip wireless interconnects (link BW 32 Gbps) with directional antenna, and total off-chip bandwidth is 128 Gbps employing SDM
W-W-D-A	Hybrid wireless on-chip interconnects with directional antenna, total on-chip wireless bandwidth is 512 Gbps, off-chip wireless interconnects (link BW 128 Gbps) with directional antenna, and total off-chip bandwidth is 512 Gbps employing SDM

performing a read operation, DRAM sends the data to the memory controller that requested the data. Since memories are inclusive, a response message carrying the data is sent to the requesting routers L2 and then this router sends the data to the source routers L1 cache. This is the basic communication protocol we followed. Following are the architecture specific communication mechanisms.

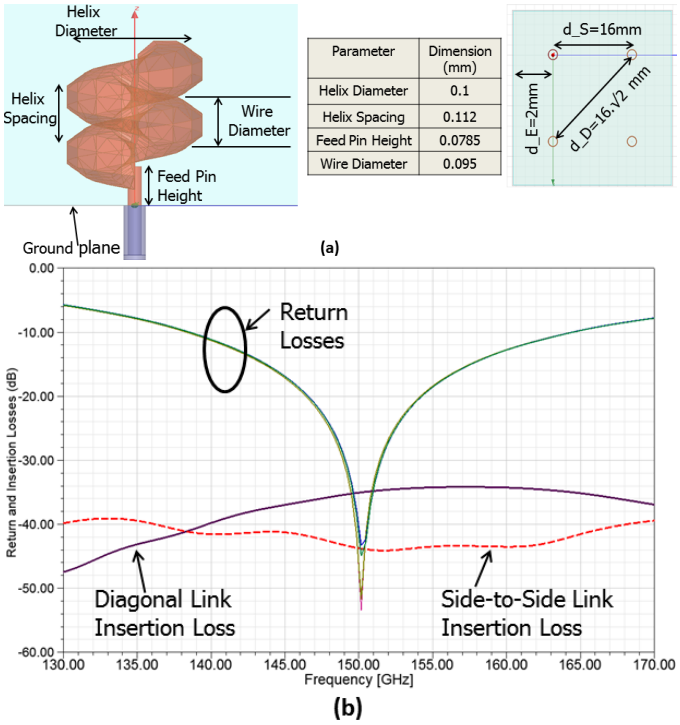
- **On-Chip Metallic with Off-Chip Metallic/Wireless Interconnects:** Figure 4(a) shows the communication mechanism for an architecture where the L2MC communication uses metallic links and MCM communication uses either wireless or metallic links. For example, if there is a miss at the L1 cache connected to router R0 and this address space is serviced by the L2 cache connected to router R9, then the L1 cache needs to send a request message through R0 to the L2 cache via R9. The request message follows the DOR protocol to reach R9 from R0 [11]. If the L2 cache has the updated data, a response message is sent to R0. However, if there is a L2 miss, then router R9 sends a new request message to the memory controller servicing that address space. Consider, the memory controller MC3 is servicing the address space of the L2 cache connected to router R9. Hence, R9 sends a message requesting updated data to MC3 and the message utilizes the DOR protocol to reach MC3. MC3 sends the necessary signal to the memory module to perform the read operation either using the metallic or the wireless link. Receiving the data from the memory module, MC3 sends a response message to the router R9. The L2 cache connected to router R9 updates the cache and sends a new response message to router R0. The response messages also follow DOR protocol. The whole communication takes twelve hops: three hops (R0 to R9), two hops (R9 to MC3), two hops (MC3 to DRAM to MC3), two hops (MC3 to R9), and three hops (R9 to R0).
- **Hybrid-Wireless Interconnects with Omnidirectional Antenna:** Figure 4 (b) shows the communication mechanism of a hybrid-wireless architecture where L2MC communication uses wireless links and MCM communication uses wireless or metallic links. For example, if there is a miss at the L1 cache connected to router R0 and the corresponding address space is serviced by the L2 cache connected to router R5, then R0 sends a request message to R5 asking for the data. The request message uses the metallic links following the DOR protocol. If the L2 cache has the updated data, a response message is sent back to R0. Consider, there is a L2 miss at router R5 and the memory controller MC1 is servicing the corresponding address space. As the transmitter of R5 and the receiver of MC1 are tuned to the same frequency, R5 waits for the token to send a new request message to MC1 using the wireless link. When R5 has the right to transmit using the wireless link of group G0, it broadcasts the request message which is received by

both memory controller MC1 and MC3. MC1 accepts the message while MC3 discards. MC1 collects the data from the memory module it is connected via the off-chip link present. MC1 then broadcasts the response message containing the data to the routers of group G0 and G2. Only R5 accepts the message and sends a new response message to router R0. The new response message follows DOR protocol. The whole communication takes eight hops: two hops (R0 to R5), one hop (R5 to MC1), two hops (MC1 to DRAM to MC1), one hop (MC1 to R5), and two hops (R5 to R0). Therefore, the hops required to access the off-chip memory is reduced. The drawback of this communication mechanism is that router R0 discards the message containing the necessary data which requires R5 to send the data again. This can be a focus for future works.

- **Hybrid-Wireless Interconnects with Directional Antenna:** The basic communication mechanism of hybrid-wireless interconnects with directional antenna is similar to the hybrid-wireless interconnects with omnidirectional antenna. Let us consider the situation described in the previous paragraph. The only difference is that R5 contains two transmitters to talk to MC1 and MC3. Hence, when R5 has the right to transmit, it sends the message using the transmitter pointed towards MC1 and MC3 does not receive any message. Similarly, when MC1 sends the response message, it uses the transmitter that is pointed towards group G0. The number of hops required in this case is also eight and the communication mechanism is shown in Figure 4 (c).

IV. Technological Feasibility Study: Antenna Model and Characterization

In this model, we have conducted full-wave simulations of helical antennas in HFSS. The design consists of four normal mode helical antennas, one at each corner. This design is enclosed in a ceramic casing with a thickness of 100 μm and also consists of a ground plane of size 20 mm by 20 mm, the chip size we simulate. A depiction of the design with all its dimensions is shown in Figure 5(a). The entries d_S, d_D, and d_E in Figure 5(a) stand for the side-to-side separation between the antennas, diagonal separation between the antennas, and the separation of the antenna from the edges of the chip, respectively. We show the insertion and return losses of the helical antenna model in Figure 5(b). The return loss, an impedance mismatch measure, is quantified by the scattering parameter S_{ii}, for i=1, 2, 3, 4 for our four antenna design. The S_{ii} values are lower than -10 dB for the frequency range of 138-165 GHz as seen from Figure 5(b). If we define bandwidth as the range of frequencies where the insertion loss variation is less than 2 dB, we can observe that for the side-to-side helical link, the maximum single-channel bandwidth available is 17 GHz (147-164 GHz), whereas for the diagonal channels the maximum single-channel bandwidth is 21 GHz (147-168 GHz). These single channel bandwidths are the highest we have obtained so far among all the different



Parameter	Dimension (mm)
Helix Diameter	0.1
Helix Spacing	0.112
Feed Pin Height	0.0785
Wire Diameter	0.095

Fig. 5. (a) Helical Antenna Model. (b) Insertion and Return Loss for Helical Antenna Model.

models we have previously considered [12], although at a higher insertion loss (∼15 dB). In particular, the helix side-to-side channels show the lowest insertion loss variation (and hence largest bandwidth) among all antenna types in [12].

V. Performance Evaluation

The proposed hybrid-wireless architectures are compared against the baseline architecture (Table II) to evaluate the performance. We have used Dsent [6] to calculate the area and energy of the metallic links, routers, and memory controllers for 45nm bulk LVT technology. Wireless link energy-efficiency is assumed to be 1 pJ/bit for on-chip communication [13] and estimated for off-chip communication as 2.54 pj/bit considering a linear increase [14]. For an off-chip metallic link, we have considered two cycle transmission time with 10 pj/bit (conservative) and 5 pj/bit (aggressive) energy consumptions. We have used a cycle accurate simulator Multi2Sim [15] to simulate network performance of the proposed architectures for PARSEC 2.1 benchmark [16]. The simulation parameters used are shown in Table III.

A. Execution Time Estimate

Figure 6 shows the execution time of blackscholes benchmark for all the architectures. It can be seen that the proposed architectures, except M-W-O-A and W-W-D-C, require lower execution time than the baseline architecture M-M-X -X. This is due to the fact that the proposed architectures require lower number of hops than the baseline architecture for off-chip memory accesses. Moreover, for the same bandwidth, the off-chip metallic link traversal requires twice the number

TABLE III
SIMULATION PARAMETERS

Core Frequency [17]	2 GHz
MSHR [18]	16
Threads per core [17], [18]	4
Memory Frequency	1 GHz
Cache line [19], [17], [18], [20]	64 Byte
Address Mapping [17]	Interleaving
Page Size [21]	4 KB
Memory Latency [15]	200 Cycle
L1-I (private)[19], [18]	32KB, 4 way, LRU
Memory Controller [20]	4
L1-D (private) [19]	32KB, 4 way, LRU
Trace Length [4], [5]	2 in
L1 Cache Latency [19], [17], [22]	2 cycle
VC per port	4
L2 (shared) [18]	256 KB/core, 8 way, LRU
On-chip Metallic Interconnect Bandwidth	8 GBps
L2 Cache Latency [22]	20 cycle
Channel Width	16 GB/s [19], [20], 8 GB/s [19], [17]

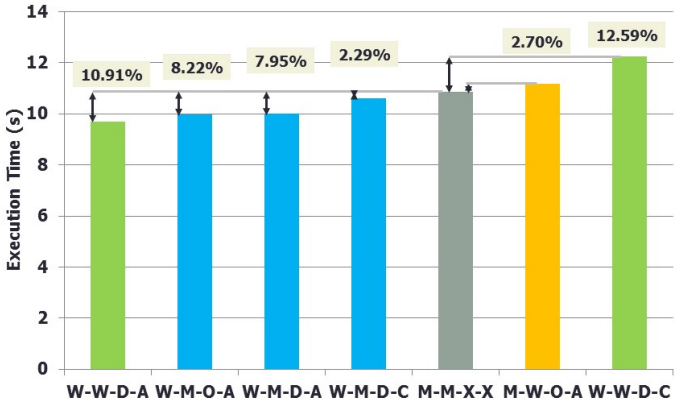


Fig. 6. Execution time of PARSEC 2.1 Benchmark, Blackscholes, for the baseline and proposed architectures.

of clock cycles required by the wireless link due to the RC delay. Therefore, the hybrid-wireless architecture having the highest bandwidth performs the best. As the off-chip link (physical connection between memory controller and DRAM) bandwidth of W-W-D-C is orders of magnitude lower than the baseline, the improvement achieved by hop-count reduction is nullified and it performs the worst. In the case of M-W-O-A, the off-chip wireless link bandwidth is half of the metallic link bandwidth in M-M-X-X but higher than the off-chip wireless link bandwidth in W-W-D-C. Hence, baseline M-M-X-X performs better than M-W-O-A and M-W-O-A performs better than W-W-D-C.

B. Energy per Byte Estimate

The energy per byte requirement of the baseline and proposed architectures is shown in Figure 7. From the figure, we can see that an improvement in energy efficiency is achieved when wireless link is used for off-chip communication instead of metallic link. Figure 7 (a) shows an average 62.07% improvement in energy-efficiency when off-chip metallic link energy consumption is assumed to be 10 pJ/bit; whereas

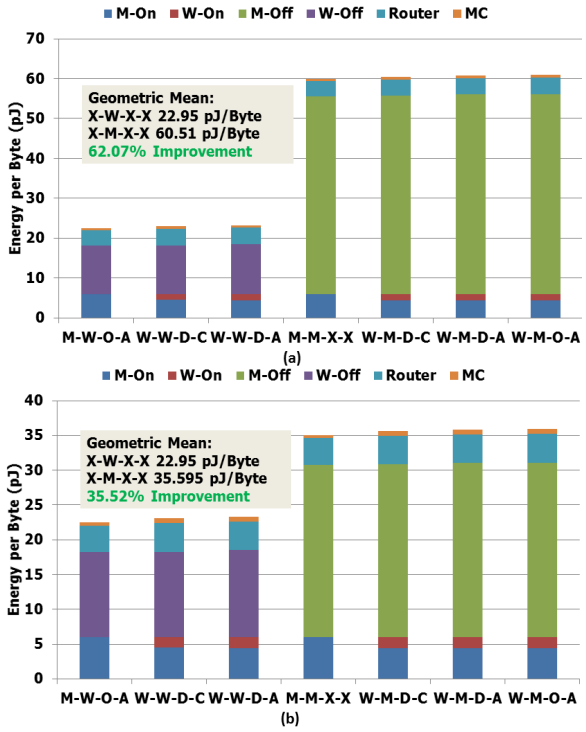


Fig. 7. Energy per byte comparison of the baseline and the proposed architectures considering the on-chip elements such as router, memory controller, and link (metallic/wireless) and the off-chip link (metallic/wireless). Energy per byte with (a) conservative and (b) aggressive off-chip metallic link energy consumption assumption.

Figure 7 (b) shows an average 35.52% improvement in energy-efficiency when off-chip metallic link energy consumption is assumed to be 5 pJ/bit. This improvement in energy-efficiency is expected because the off-chip metallic link energy requirement, which increases quadratically with distance, is an order of magnitude higher than the off-chip wireless link energy requirement. In addition, the number of hops required to access the off-chip memory is reduced in the proposed architectures.

C. Area Estimate

The area of an architecture consists of the link (metallic/wireless) area, router area, and memory controller area. In this paper, we have assumed transmitter area as 0.42 mm^2 and receiver area as 0.20 mm^2 [13]. Figure 8 shows the area comparison of the proposed architectures with respect to the baseline architecture. From the figure, we can see that the hybrid-wireless architectures require more area than the baseline architecture. This is because the wireless transceiver area footprint is higher compared to the metallic link footprint. However, this does not alter the distances between routers since antennas are fabricated on a different layer. Since W-W-D-A requires the highest number of wireless transceiver antennas, it has the highest area requirement. W-W-D-C requires less area than W-W-D-A, even though they require the same number of antennas, because of the reduced memory controller area requirement. The end result is that M-M-X-X

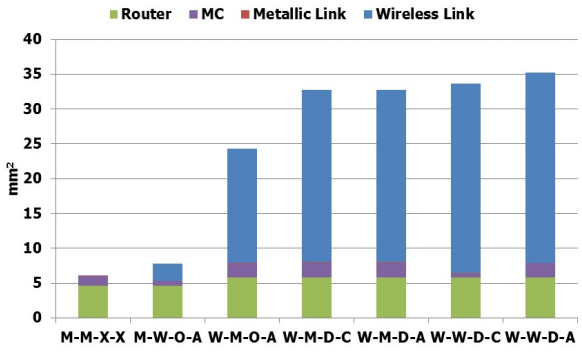


Fig. 8. Area.

and W-W-D-C requires 82.68% and 4.35% less area compared to W-W-D-A.

VI. Conclusions

The proposed hybrid-wireless architectures to access off-chip memory show improved performance compared to the metallic-only baseline architecture. The hybrid-wireless networks require lower hops than the baseline network to access the off-chip memory. However, wireless bandwidth limitation may offset the energy savings achieved via hop reduction. The end result is that we can design a network which shows 10.91% improvement in execution time compared to the baseline architecture. In terms of energy efficiency, networks with off-chip wireless links perform the best. The hybrid-wireless networks can achieve around 62.07% to 32.52% improvement in energy-efficiency when compared with conservative and aggressive energy-efficiency for metallic interconnects. However, the area overhead of the hybrid-wireless networks is higher compared to the baseline architecture because of the higher wireless link area requirement. Nevertheless, since the wireless networks do not require any hard-wired waveguides, it has the advantage of flexibly creating dynamic connections on-demand. This would allow the hybrid-wireless architectures to scale to large core counts without losing performance and improve energy efficiency. This run-time reconfiguration can be explored in future.

VII. Acknowledgment

This research was supported by the grant funding numbers: ECCS-1342657, CCF-1054339 (CAREER), CCF-1420718, CCF-1318981, and CCF-1513606. We would like to thank the reviewers for their valuable feedback.

References

[1] D. Abts, N. D. Enright Jerger, J. Kim, D. Gibson, and M. H. Lipasti, "Achieving predictable performance through better memory controller placement in many-core cmps," in *ACM SIGARCH Computer Architecture News*, vol. 37, no. 3. ACM, 2009, pp. 451–461.

[2] A. Sharifi, E. Kultursay, M. Kandemir, and C. Das, "Addressing end-to-end memory access latency in noc-based multicores," in *Microarchitecture (MICRO), 2012 45th Annual IEEE/ACM International Symposium on*, Dec 2012, pp. 294–304.

[3] T. O. Dickson, Y. Liu, S. V. Rylov, B. Dang, C. K. Tsang, P. S. Andry, J. F. Bulzacchelli, H. Ainspan, X. Gu, L. Turlapati *et al.*, "An 8x 10-gb/s source-synchronous i/o system based on high-density silicon carrier interconnects," *Solid-State Circuits, IEEE Journal of*, vol. 47, no. 4, pp. 884–896, 2012.

[4] I. Micron Technology, "Tn-41-13: Ddr3 point-to-point design support," 2013.

[5] ——, "Tn-46-14: Hardware tips for point-to-point system design," 2006.

[6] C. Sun, C.-H. Chen, G. Kurian, L. Wei, J. Miller, A. Agarwal, L.-S. Peh, and V. Stojanovic, "Dsent-a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling," in *Networks on Chip (NoCS), 2012 Sixth IEEE/ACM International Symposium on*. IEEE, 2012, pp. 201–210.

[7] A. Ganguly, K. Chang, S. Deb, P. P. Pande, B. Belzer, and C. Teuscher, "Scalable hybrid wireless network-on-chip architectures for multicore systems," *Computers, IEEE Transactions on*, vol. 60, no. 10, pp. 1485–1502, 2011.

[8] M. A. I. Sikder, A. K. Kodi, M. Kennedy, S. Kaya, and A. Louri, "Own: Optical and wireless network-on-chip for kilo-core architectures," in *High-Performance Interconnects (HOTI), 2015 IEEE 23rd Annual Symposium on*. IEEE, 2015, pp. 44–51.

[9] M. S. Shamim, J. Muralidharan, and A. Ganguly, "An interconnection architecture for seamless inter and intra-chip communication using wireless links," in *Proceedings of the 9th International Symposium on Networks-on-Chip*. ACM, 2015, p. 2.

[10] S.-B. Lee, S.-W. Tam, I. Pefkianakis, S. Lu, M. F. Chang, C. Guo, G. Reinman, C. Peng, M. Naik, L. Zhang *et al.*, "A scalable micro wireless interconnect structure for cmps," in *Proceedings of the 15th annual international conference on Mobile computing and networking*. ACM, 2009, pp. 217–228.

[11] W. J. Dally and B. P. Towles, *Principles and practices of interconnection networks*. Elsevier, 2004.

[12] W. Rayess, D. Matolak, S. Kaya, and A. Kodi, "Antennas and channel characteristics for wireless networks on chips." Wireless Personal Communications.

[13] A. K. Kodi, M. A. I. Sikder, D. DiTomaso, S. Kaya, S. Laha, D. Matolak, and W. Rayess, "Kilo-core wireless network-on-chips (nocs) architectures," in *Proceedings of the Second Annual International Conference on Nanoscale Computing and Communication*. ACM, 2015, p. 33.

[14] D. DiTomaso, A. Kodi, D. Matolak, S. Kaya, S. Laha, and W. Rayess, "Energy-efficient adaptive wireless nocs architecture," in *Networks on Chip (NoCS), 2013 Seventh IEEE/ACM International Symposium on*. IEEE, 2013, pp. 1–8.

[15] R. Ubal, J. Sahuquillo, S. Petit, P. Lopez, Z. Chen, and D. R. Kaeli, "The multi2sim simulation framework: A cpu-gpu model for heterogeneous computing," 2011.

[16] C. Bienia, S. Kumar, J. P. Singh, and K. Li, "The parsec benchmark suite: Characterization and architectural implications," in *Proceedings of the 17th international conference on Parallel architectures and compilation techniques*. ACM, 2008, pp. 72–81.

[17] I. Bhati, Z. Chishti, S.-L. Lu, and B. Jacob, "Flexible auto-refresh: Enabling scalable and energy-efficient dram refresh reductions," in *Computer Architecture (ISCA), 2015 ACM/IEEE 42nd Annual International Symposium on*, June 2015, pp. 235–246.

[18] J. Ahn, S. Yoo, O. Mutlu, and K. Choi, "Pim-enabled instructions: A low-overhead, locality-aware processing-in-memory architecture," in *Computer Architecture (ISCA), 2015 ACM/IEEE 42nd Annual International Symposium on*, June 2015, pp. 336–348.

[19] Y. Lee, J. Kim, H. Jang, H. Yang, J. Kim, J. Jeong, and J. Lee, "A fully associative, tagless dram cache," in *Computer Architecture (ISCA), 2015 ACM/IEEE 42nd Annual International Symposium on*, June 2015, pp. 211–222.

[20] O. Seongil, Y. H. Son, N. S. Kim, and J. H. Ahn, "Row-buffer decoupling: A case for low-latency dram microarchitecture," in *Computer Architecture (ISCA), 2014 ACM/IEEE 41st International Symposium on*, June 2014, pp. 337–348.

[21] A. Ros and S. Kaxiras, "Callback: Efficient synchronization without invalidation with a directory just for spin-waiting," in *Computer Architecture (ISCA), 2015 ACM/IEEE 42nd Annual International Symposium on*, June 2015, pp. 427–438.

[22] L. Peled, S. Mannor, U. Weiser, and Y. Etsion, "Semantic locality and context-based prefetching using reinforcement learning," in *Computer Architecture (ISCA), 2015 ACM/IEEE 42nd Annual International Symposium on*, June 2015, pp. 285–297.