Energy Efficient Modulation for a Wireless Network-on-Chip Architecture

Dominic DiTomaso, Soumyasanta Laha, Savas Kaya, David Matolak, and Avinash Kodi
School of Electrical Engineering and Computer Science
Ohio University, Athens, OH 45701
E-mail: dd292006, sl922608, kaya, matolak, kodi@ohio.edu

Abstract—As both power consumption and leakage currents will limit the scalability of future massively integrated computational systems, research into emerging technologies and devices to replace traditional metallic interconnects has become critical. In this paper we propose an initial implementation for a hybrid wireless network-on-chip (WiNoC) interconnect architecture, named iWISE, for current chip multiprocessors (CMPs). iWISE combines wired interconnects with wireless links that use both frequency and time division multiplexing to offer a balanced, flexible, orthogonal wireless data transfer among cores. We provide a basic description of the iWISE architecture and describe a practical solution for the implementation of wireless interconnects based on an on-off keying (OOK) modulator using ultra-compact Double Gate (DG) CMOS devices. The proposed OOK modulator takes advantage of DG-CMOS devices especially in building compact modulation and tunable amplification circuitry. Real applications from the benchmark suite PARSEC as well as synthetic traffic show an improvement in performance as well as a savings in power.

I. INTRODUCTION

To alleviate the long wire delays and high power consumption of future multi-core integrated circuits (ICs), many prototypes and commercial designs are using Network-on-Chip (NoC) packet switching architectures. Wireless interconnects can improve NoCs by reducing the power dissipation of long global wires while providing high bandwidth and low latency communication [1]. Wireless technologies have the advantage of being a well understood form of communication with many applications already implemented in wireless networking, cell phones, etc. This existing knowledge in the wireless/radio frequency (RF) field will facilitate the integration of wireless interconnects for NoCs. With the on-chip distances of a few centimeters, wireless communication will require less transmitted power than some traditional wireless communication systems. Wireless interconnects can provide some unique benefits which include the following: (1) reduced power dissipation by avoiding multi-hop communication as in traditional metallic interconnects, (2) reduced area overhead and parasitics with the elimination of wires or waveguides, and (3) feasibility of implementation as wireless transceivers are CMOS compatible.

Some recent work has proposed wireless/RF technologies to improve NoCs. Chang et al. [2] proposed using a RF transmission line on top of a wired mesh to quickly propagate RF signals. Lee et al. [3] proposed WCube, a wired/wireless hybrid design, which uses centralized wireless hubs connected on one tier and a wired mesh on another tier. Ganguly et al. [1] proposed using centralized wireless hubs with small subnetworks in a more recent hybrid design. The work in [4] proposed iWISE with distributed wireless hubs, and used frequencies in the 100-500 GHz range. While these designs provide low power and low latency solutions, the design of a practical transceiver design as well as effectively utilizing the limited bandwidth remains a significant challenge.

In this paper, we propose an initial implementation for a wireless interconnect to be used in the hybrid iWISE (inter-router Wireless Scalable Express Channel) NoC architecture for current chip multiprocessors (CMP). iWISE is a low power and area efficient NoC design that uses wireless interconnects to reduce packet latency by avoiding multi-hop communication. The iWISE multiple access (MA) approach employs both time and frequency domains which offers efficient and flexible sharing of the wireless links. In particular, we focus on the 64-core version of iWISE as it represents a current CMP in which our wireless interconnect can be implemented. Using current RF Double Gate (DG) CMOS technology [5], [6], we design an ultra-low power, on-off keying (OOK) modulator in the 100 GHz range. We believe that the design of this modulator with the DG-CMOS technology can be the starting point for future CMPs in which a long-term transceiver design can be implemented with future developments in technologies such as carbon-based electronic materials and nanostructures. Using current technology nodes, our network can reduce the power by up to 34% when compared to leading NoCs. The performance of our network using the real application benchmark suite, PARSEC [7], shows a 2.23 speedup over the Flattened Butterfly (FB) topology and mesh. The major contributions of this work include:

1) We design an ultra-compact, low power modulation technique in a compact wireless transceiver with transmit power control to improve network power and performance.

2) We combine frequency division multiplexing (FDM) and time division multiplexing (TDM) to offer flexible, orthogonal wireless data transfer among cores which improves performance on real applications and synthetic traffic.

II. iWISE ARCHITECTURE

The iWISE architecture is a scalable wireless hybrid NoC. In order limit router area and reduce the bandwidth requirement, iWISE is separated into hierarchical subsections and
uses a token sharing scheme as explained next. Figure 1 shows theses subsections as well as the metal links (solid lines) and the shared wireless links (dotted arrows). Four cores (N=4) are concentrated into one cluster as this has been shown to be an effective technique to reduce the router area overhead as well as serialization latency. Clusters are denoted C(c, s, g) where c=cluster, s=set, and g=group. For iWISE-64, c∈{0, 1, ..., C-1}, s∈{0, 1, ..., S-1}, and g∈{0, 1, ..., G-1} with C=S=4 and G=1. These parameters are chosen as they balance flexible communication and low contention for links. The total number of cores in iWISE is the product N×C×S×G. A wired mesh is used to connect adjacent routers (clusters), and wireless links are connected between every non-adjacent router. Packets with source and destination clusters adjacent to each other use these wired links. Otherwise, packets use wireless links. To share wireless communication links, two sharing schemes are used: token partial and token full. In iWISE with token partial (iWISE-TP), the 4 clusters within a set (dotted box) share the 4 wireless links of that set (same colored arrows). iWISE with token full (iWISE-TF) ignores the set subsections and shares 16 wireless links between all 16 clusters. Since iWISE is a one hop network there can be no deadlocks or livelocks.

The token partial multiple access (MA) for wireless communication allows each cluster within a set the opportunity to transmit to each of the S sets over C consecutive time frames and S frequency channels. Time frame duration is T_f, and the C consecutive frames constitute a cycle of duration T_c=CT_f-this represents the time-division (de-)multiplexing aspect of the MA scheme. During a time frame each of the C clusters within a set can transmit to a unique set using a set-to-set-unique frequency channel, constituting the frequency division duplexing aspect of the MA scheme.

In Figure 2, we show the Set-0 portion of the time-frequency plane to illustrate an example time-frequency allocation. This specific example assumes a uniform traffic demand, in that each cluster has data to transmit on each frame of each cycle. The remaining sets (1, 2, and 3) employ analogous blocks of four frequency channels arranged over the same frame and cycle times.

- \( f_{ij} \) = frequency channel for transmissions from set i to set j; i, j ∈ {0, 1, ..., S-1}; thus there are \( S^2 \) total frequency channels.
- \( \gamma_{c,s,g} = [\gamma_{c,s,g,0}; \gamma_{c,s,g,1}; \gamma_{c,s,g,2}; \gamma_{c,s,g,3}] \) = token vector for cluster c, set s, group g, with \( \gamma_{c,s,g} \in \{0, 1\} \)
- \( C(c,s,g) \rightarrow S_i \) denotes a transmission from the specified cluster to set \( S_i \)

For the uniform allocation of time-frequency (TF) cells in Figure 2, if each frequency channel has bandwidth \( B \) Hz, assuming binary orthogonal modulation (on-off keying), the channel’s data rate \( R_b \approx B \) bps. In this orthogonal and periodic MA example scheme then, any cluster’s data rate to any other cluster (intra- or inter-set) has a maximum value of \( R_b/S=R_b/4 \). Our initial approach has each frame carry 32 bits.

Note that if each cluster has only a single wireless modem—for any specific set-to-set frequency channel (say, \( f_{01} \))-if any of the C clusters within the transmitting set (say \( C(2,0,0) \)) has no data to transmit wirelessly during a given frame (from \( T_f < t < 2T_f \)), that frame can be used by any other cluster in that same transmitting set (\( C(0,0,0), C(1,0,0) \) or \( C(3,0,0) \), provided that the other cluster is not transmitting wirelessly on another frequency channel during that same frame.

III. WIRELESS TRANSEIVER WITH ENERGY EFFICIENT MODULATION DESIGN

Figure 3 shows the generic architecture of the wireless transceiver proposed for iWISE implementation. It uses on-off keying (OOK) since the simplicity of this non-coherent modulation technique leads to a very low power consumption as well as ultra-compact architecture. We target ultra-low power operation (<1pJ/bit) as well as ultra-short (≤2 cm) range, based on 100 GHz RF-CMOS technology and on-chip antennas. In the transceiver blocks, voltage controlled oscillators (VCO) are used to generate the different carrier frequencies to which the data is modulated via DG-CMOS mixers and amplified with tunable gain class A power amplifiers (PA). The gain in the PA must be sufficient to drive poor-efficiency on-chip antennas (<λ/2) via matching networks at the levels provided in Figure 4, which considers transmission losses and 10 dBm error margin. Low-noise amplifiers (LNA) may be needed to boost the incoming signal to an envelope detector required to make bit decisions. The use of RF DG-CMOS
devices provide room for novel tunable (VCO/PA), compact (mixer/modulator), efficient (PA/LNA) circuit blocks, as these transistors will be available in sub-32nm nodes [5]. We assume that magnetic thin films for on-chip inductor design will lead to substantial area savings as argued by Borkar [8]. DG-CMOS circuits [6] are especially suitable for power tuning (core-to-core distance adjustment) and channel selection (frequency hopping) in the iWISE architecture. For instance, Figure 5 shows the simulated response of the DG-CMOS modulator circuit designed to switch the VCO signal via back gate biasing of a single transistor, which would require four active transistors in conventional design. Admittedly, another challenging aspect of the transceiver design is the antenna technology, which may demand compact non-conventional solutions that are beyond the scope of this paper; specific antenna details are left as future work.

The total transmitter and receiver area in 32 nm technology is estimated to be 0.054 mm² and 0.04 mm², respectively. The transmitter area is mostly dominated by the inductors in the PA and VCO; its area was computed from [8]. The receiver area is dominated by the LNA which occupies 0.04 mm². For a 5 Gbps data rate, the total transmitter and receiver power is 12.85 mW and 3.11 mW, respectively, at 32 nm technology node using Synopsys HSPICE. Tx power, dominated by the PA, is held especially high to compensate for the low-efficiency of the on-chip antennas that may only couple via near-field components. Thus, the value of 12.85 mW may be regarded as an upper limit on the transmitter power. According to Figure 4, the signal power is a stronger function of the distance than the data rate, therefore the gain tuning in the PA may substantially lower overall power dissipation. Also, it may be possible to simplify circuits (i.e. lower area) by omitting impedance match circuits if plasmonic antennas are considered along with direct-pulse drive schemes [9]. However, the transmission losses and interference affects must be more precisely evaluated, once antenna technology and full-3D architecture is determined.

IV. PERFORMANCE EVALUATION

In this section we compare iWISE to electrical NoC designs including Flattened Butterfly (FB), mesh, and Concentrated Mesh (Cmesh) architectures. Other wireless/RF networks were not compared against because they do not scale down to 64 cores well. Here, we scale the original 32-bit wireless links in iWISE-64 to 5 Gbps. We scale the wired bandwidth by the same factor for data rate of 20 Gbps. For simulation, a packet size of four flits was used and the bisectional bandwidth was kept the same for all designs by adjusting the link width and adding delays. Here, the unrealistically low wired data rates (which can be as high as 64 or 128 Gbps) may be justified by equally restricted wireless bandwidth around 100 GHz operation in the present implementation. However with larger bandwidth technologies in the future, the data rates for both wired and wireless links will be able to increase equally thus the following simulation results of iWISE relative to mesh, CMesh, and FBly will be the same and are still relevant. For TF and TP, a single cycle delay was used to account for the token delay. In the following sections, we will compare iWISE to other networks by providing power and area estimates along with speed-up simulation results.

A. Power Dissipation and Area Overhead

The energy consumption of a metal link was estimated to be 0.18 pJ/bit for 1 mm at 1 GHz in 32 nm technology. A wireless communication link was estimated to have an energy of 0.160 pJ/bit per mm in 32 nm technology. The buffer write energy of 0.011 pJ/bit and a crossbar power of 0.108 pJ/bit were estimated using Synopsys Design Compiler.

During simulation, the energy consumption of buffer writes, link and crossbar traversals for each packet multiplied by the corresponding number of bits to obtain the power dissipation. The transmitted power was linearly adjusted via gain tuning at the PA according to the required distance the packet must
travel. Figure 6 shows the average power dissipation per packet relative to mesh of 64-core networks for different traffic loads. iWISE under uniform traffic has 80% of the traffic sent on wireless links. This high percentage and the multi-hop nature of the other networks are the major contributors to our overall average power savings of approximately 18%. For traffic patterns such as Complement, iWISE uses only wireless links which results in a saving of 34% over mesh. As mentioned earlier, the savings will increase as technology scales. As the data rates of both wired and wireless links increases, the power will increase for all networks equally and the percent savings will not change.

The area of the 32nm wireless link technology was 0.094 mm². This is significantly higher than the wired link and router areas. The increase in area represents the trade-off of the power and performance gains. Nonetheless, unlike wires and routers, the wireless transceiver area will not increase as the data rates increase. Furthermore, the area of the wireless transceivers will reduce with technology scaling. Additionally, with small transmission distances, it may be possible to remove the major contributors to the transceiver area by eliminating LNA, PA, and impedance matching circuitry as mentioned before. Lastly, recent developments in carbon-based electronic materials and nanostructures may provide even more compact transceivers.

B. Speed-up

The full execution-driven simulator SIMICS from Wind River was used to extract traffic traces from real applications. The PARSEC workload was 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. Using our token partial MA scheme, iWISE-TP has an average speed-up of 2.23 over mesh for the PARSEC applications shown in Figure 7. The one-hop architecture of iWISE allows packets to use wireless links to avoid high latency, multi-hop paths as exists in the 16-hop diameter mesh network. In FBfly, there are long wired links which cause high latency compared to the fast wireless links of iWISE, which result in a 30% - 35% speed up. Comparing iWISE-TF to iWISE-TP shows TP outperforming TF by approximately 30%. TF will have a higher latency as it can wait up to 16 cycles for an opportunity to use a wireless link compared to the 4 cycles of TP.

V. Conclusion

In this paper, we proposed a wireless hybrid interconnect framework, the iWISE NoC architecture, that improves network performance and saves power. We use TDM and FDM to allocate wireless links to efficiently communicate between sets of clusters. Our transceiver for the wireless link uses an OOK technique to achieve an ultra-compact design and low power consumption. We scale down the wired data rate to match our 16 wireless links operating at 5 Gbps since we are limited by the 60-140 GHz bandwidth. Using wireless links reduces network power and improves performance.

ACKNOWLEDGMENT

This research was partially supported by NSF awards, CCF-0915418, CCF-1054339 (CAREER) and ECCS-1129010.

REFERENCES