Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2010-009 February 11, 2010

Efficient Cache Coherence on Manycore Optical Networks

George Kurian, Nathan Beckmann, Jason Miller, James Psota, and Anant Agarwal

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Efficient Cache Coherence on Manycore Optical Networks

George Kurian, Nathan Beckmann, Jason Miller, James Psota and Anant Agarwal
{gkurian,beckmann,jasonm,jim,agarwal}@csail.mit.edu
Massachusetts Institute of Technology, Cambridge, MA

***Abstract*—Ever since industry has turned to parallelism instead of frequency scaling to improve processor performance, multicore processors have continued to scale to larger and larger numbers of cores. Some believe that multicores will have 1000 cores or more by the middle of the next decade. However, their promise of increased performance will only be reached if their inherent scaling challenges are overcome. One such major scaling challenge is the viability of efficient cache coherence with large numbers of cores. Meanwhile, recent advances in nanophotonic device manufacturing are making CMOS-integrated optics a reality—interconnect technology which can provide significantly more bandwidth at lower power than conventional electrical analogs.**

The contributions of this paper are two-fold. (1) It presents ATAC, a new manycore architecture that augments an electrical mesh network with an optical network that performs highly efficient broadcasts. (2) It introduces ACKwise, a novel directory-based cache coherence protocol that provides high performance and scalability on any large-scale manycore interconnection network with broadcast capability. Performance evaluation studies using analytical models show that (i) a 1024-core ATAC chip using ACKwise achieves a speedup of 3.9× compared to a similarly-sized pure electrical mesh manycore with a conventional limited directory protocol; (ii) the ATAC chip with ACKwise achieves a speedup of 1.35× compared to the electrical mesh chip with ACKwise; and (iii) a pure electrical mesh chip with ACKwise achieves a speedup of 2.9× over the same chip using a conventional limited directory protocol.

I. Introduction

As silicon resources become increasingly abundant, massive multicore chips are on the horizon. But will current processor architectures, especially their interconnection networks and cache coherence mechanisms, scale to thousands of cores? This paper argues that they will not. It presents ATAC [1], a new processor architecture that leverages the recent advances in CMOS-integrated optics, and ACKwise, a novel cache coherence protocol, to address these scalability issues.

State-of-the-art multicore chips employ one of two strategies to deal with interconnection costs. Small-scale multicores typically interconnect cores using a bus. This simple design does not scale to large numbers of cores due to increasing bus wire length and contention. More scalable interconnection strategies use point-to-point networks. For example, the Raw microprocessor [2] uses a mesh interconnect. This avoids long global wires but communication between distant cores requires multiple hops. Furthermore, contention will become prohibitive as processors are scaled to thousands of cores. Global communication operations (e.g., broadcasts to maintain cache coherence) will also be highly inefficient on these networks because each global operation ties up many resources and consumes a lot of energy.

The ATAC processor architecture addresses these communication issues using on-chip optical communication technologies to augment electrical communication channels. In particular, ATAC leverages the Wavelength Division Multiplexing (WDM) property of optical waveguides to allow a single waveguide to simultaneously carry multiple independent signals on different wavelengths and thus provide high bandwidth while simultaneously achieving lower power consumption.

ACKwise enables large-scale cache coherence by exploiting the broadcast capabilities of the underlying interconnection network. ATAC is particularly well suited, since optical interconnects provide a cheap broadcast capability in hardware. Preliminary results show that ACKwise enables large-scale cache coherence, allowing programmers to use widespread sharing of data.

The remainder of this paper is organized as follows. Section II provides an overview of the ATAC architecture, including the optical background and its processing and communication mechanisms. Section III introduces the ACKwise cache coherence protocol. Section IV evaluates the ATAC architecture and the ACKwise protocol and provides a preliminary set of results, focusing on cache coherence across 1,024 cores. Section V follows with a detailed discussion of related work, and Section VI concludes the paper.

II. Architecture Overview

The ATAC processor architecture is a tiled multicore architecture combining the best of current scalable electrical interconnects with cutting-edge on-chip optical communication networks. The tiled layout uses a 2-D array of 1,024 simple processing cores, each containing a single- or dual-issue, in-order RISC pipeline, private L1 data and instruction caches, and private L2 caches. The ATAC architecture is targeted at an 11nm process in 2019.

A. Optical Technology Background

The key elements in a nanophotonic network such as the one employed by the ATAC chip include: the offchip "optical power supply" light source; waveguides to carry optical signals; modulators to place signals into the waveguides; and detectors and filters to receive signals from the waveguides. The modulator couples light at its pre-tuned wavelength λ from the optical power source and encodes either a 0 or 1 onto the data waveguide as directed by its driver. The optically-encoded data signal traverses the waveguide at approximately one-third the speed of light and is detected by a filter that is also tuned to wavelength λ. Photons are detected by the photodetector and received by a flip-flop on the receiver side. Further details about ATAC's optical technology is available in [3].

B. ATAC

The cores in an ATAC processor are connected through two networks: the electrical EMesh and the optical/electrical ANet. The EMesh is a conventional 2-D point-to-point electrical

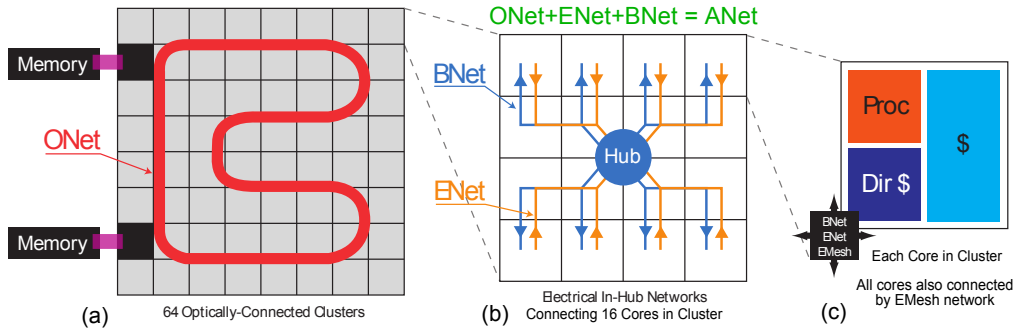

Fig. 1: ATAC architecture overview. (a) The layout of the optical waveguide (ONet) across a chip. (b) The architecture within a cluster. (c) The architecture of a core within a cluster.

mesh network and is ideal for predictable, short-range communication. The ANet employs state-of-the-art optical technology to enable low-latency, energy-efficient, contention-free global communication. The core of the ANet is the all-optical ONet shown in Figure 1. The ANet also contains two small electrical structures called the ENet and BNet that are used to interface with the ONet. The ANet is especially useful for long-distance communication or global operations such as broadcasts. The remainder of this section focuses on ANet.

The ONet provides a low-latency, contention-free connection between a set of optical endpoints called hubs. Hubs are interconnected via waveguides that visit every hub and loop around on themselves to form continuous rings (see Figure 1). Each hub can place data onto the waveguides using an optical modulator and receive data from the other hubs using optical filters and photodetectors. Because the data waveguides form a loop, a signal sent from any hub will quickly reach all of the other hubs. Thus every transmission on the ONet has the potential to be a fast, efficient broadcast. The ONet consists of a bundle of waveguides: 64 for data, 1 for backwards flow control, and several for metadata. The metadata waveguides are used to indicate a message type (e.g., memory read, barrier, raw data) or a message tag (for disambiguating multiple messages from the same sender).

To avoid the interference of these broadcasts with each other, the ONet uses wavelength division multiplexing (WDM). Each hub has modulators tuned to a unique wavelength to use when sending and contains filters that allow it to receive signals on all the wavelengths. This eliminates contention and the need for arbitration in the optical network. In addition, the improved propagation speed of optical signals eliminates the heterogeneous, distance-dependent cost of communication between cores; any pair of hubs on the chip can communicate with low, fixed latency instead of the one-cycle-per-hop delay found in point-to-point networks. Taken together, these features mean that the ONet is functionally similar to a fully-connected, bi-directional point-to-point network with an additional broadcast capability.

Due to a variety of constraints, including power at the photodetectors and the off-chip power source, ATAC is limited to 64 hubs. Because of this limit, the set of 1024 cores is broken into 64 clusters of 16 cores. All the 16 cores within a cluster share the same optical hub. The ONet interconnects the 64 symmetric clusters with a 64-bit wide optical waveguide bus. Within a cluster, cores communicate electrically with each other using the EMesh and with the hub using two networks called the ENet and BNet. The ENet is an electrical network that is used only to send data from cores within a cluster to the hub for transmission on the ONet. The BNet is an electrical broadcast tree that is used to forward data that the hub receives from the ONet down to the cores.

C. Baseline Architecture

To help evaluate ATAC and ACKwise, we compare against a baseline electrical mesh architecture (pEMesh). pEMesh is identical to ATAC, except that the ANet is replaced with a 64-bit 2-D electrical mesh network similar to [2]. Additionally, the pEMesh network supports broadcast messages by copying and forwarding packets in a simple tree structure as described in [4].

III. ACKwise

This section introduces *ACKwise*, a novel cache coherence protocol that leverages the broadcast capabilities of the underlying network, either ANet or pEMesh, to provide highly efficient and scalable cache coherence on manycore processors. ACKwise makes one novel addition to a MOSI limited directory with broadcast protocol [5]. In a limited directory with broadcast protocol, when the number of sharers of a cache block exceeds the capacity of the sharer list, the protocol assumes that all cores share that cache block. Hence, on an exclusive request to that cache block, invalidation requests are sent to all the cores and acknowledgements received from each. In a processor with hundreds or thousands of cores, this is highly detrimental to performance.

ACKwise solves this problem by intelligently keeping track of the number of sharers of a cache block once it exceeds the capacity of the sharer list. On an exclusive request to that cache block, ACKwise broadcasts an invalidation request to all the cores and expects acknowledgements only from those cores that are actual sharers of the cache block, not from all the cores as in a limited directory with broadcast protocol.

IV. Evaluation

This section shows performance results for the ATAC network (ANet) versus a pure electrical mesh (pEMesh), and the ACKwise protocol versus a limited directory with broadcast protocol (Dir_iB)[5]. We evaluate a synthetic shared memory

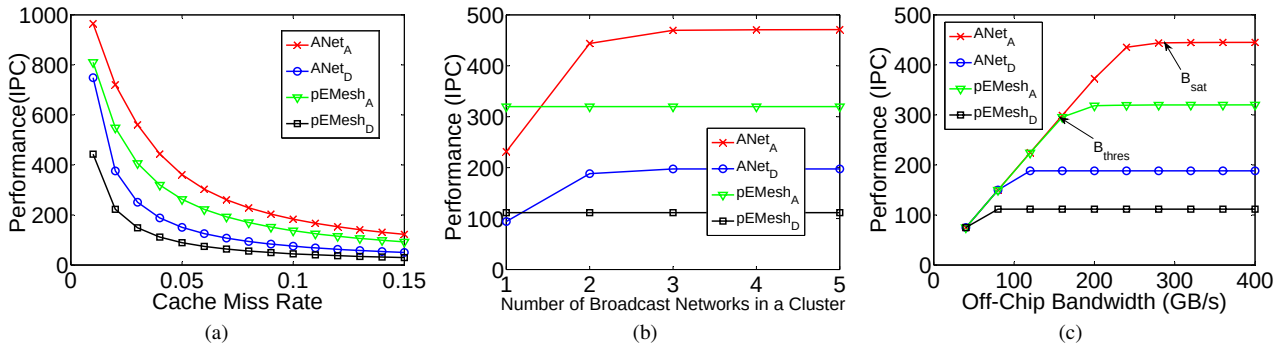

Fig. 2: Performance of ANet vs pEMesh as a function of (a) Cache Miss Rate; (b) Number of Electrical Broadcast Networks(BNet); and (c) Off-Chip Bandwidth

System Parameters	Value
CPI of Non-Memory Instructions	0.6
Number of Cores	1024
Number of Clusters	64
Frequency of a Core	1 GHz
Cache Access Time	1 ns
Cache Line Size	64 bytes
Capacity of the Sharer List	6
Memory Access Time	0.1 μs
Single Hop Latency through Electrical Mesh	1 ns
Propagation Time though Optical Waveguide	2.5 ns
Link Width of the Optical Network (ONet)	64 bits
Link Width of an Electrical Broadcast Network (BNet)	32 bits
Number of Electrical Broadcast Networks	2
Link Width of the pure Electrical Mesh (*pEMesh*)	64 bits

TABLE I: Baseline system configuration

Parameter	Value
Frequency of Data References	0.3
Fraction of Reads in Data References	2/3
Fraction of Writes in Data References	1/3
Cache Miss Rate	4%
Average Number of Sharers	4
Fraction of Memory Requests going off-chip	0.7
Fraction of Memory Write Requests that cause Invalidation Broadcasts	0.1

TABLE II: Benchmark characteristics

benchmark on the following 4 combinations of networks and protocols: (i) $ANet_A$, (ii) $ANet_D$, (iii) $pEMesh_A$, and (iv) $pEMesh_D$, where A denotes ACKwise$_i$ and D denotes Dir$_i$B (i being the number of sharers supported in hardware). Results demonstrate the advantages of both ANet and ACKwise.

A. Methodology

Due to the impracticality of simulating manycore systems such as ATAC with current simulators, we built an analytical model of processor performance. The model is based on an in-order processor model focusing on latency of memory requests. It takes into account queueing delay in the on-chip network as well as off-chip. All network traffic generated by cache coherence messages is modeled and contributes to queueing delay. Refer to [3] for further information about the analytical model.

B. Results

The system parameters used are shown in Table I and the synthetic benchmark characteristics are shown in Table II. The characteristics of the synthetic benchmark have been derived using the PARSEC benchmark suite [6]. Unless otherwise mentioned, these parameters are used in the performance studies conducted in this section.

Figure 2a shows the miss rate of the synthetic benchmark varied from 1% to 15%. As expected, performance is highly sensitive to application miss rate, as the network is quickly saturated by 1,024 cores generating memory requests. It is clear that ACKwise$_i$ outperforms Dir$_i$B. This is primarily due to the high queueing delays in the network created by the enormously high traffic generated when a write miss occurs at an address that is widely shared. We observe that, on average, $ANet_A$ outperforms $ANet_D$ and $pEMesh_D$ by a factor of 2.3× and 3.9× respectively.

The results also show that ACKwise$_i$ significantly improves performance. $ANet_A$ outperforms $ANet_D$ by a factor of 1.35×. Although ACKwise was originally developed to leverage the strengths of the ATAC network, it is also beneficial on any network that supports broadcast. Results show that $pEMesh_A$ outperforms $pEMesh_D$ by a factor of 2.9×. This is greater than the improvement seen on the ANet due to the fact that the pEMesh is lower bandwidth and is more congested when using Dir$_i$B. However, note that ANet outperforms pEMesh in absolute terms for both protocols.

The number of electrical broadcast networks ($BNet$) in $ANet$ significantly impacts $ANet$'s performance. The number is important because the traffic through $BNet$ exceeds that through the optical network ($ONet$). This is due to messages generated by the cache coherence protocol that have more than one receiver — multicast and broadcast messages consume more bandwidth on $BNet$ than on $ONet$. Using the analytical model and the above parameters we find that the traffic is on an average 15% larger on $BNet$.

Figure 2b shows the effect of varying the number of electrical broadcast networks. Near-peak performance is achieved with 3 networks. However, since the performance drop for 2 networks is small, it seems reasonable to use 2 $BNet$s to conserve cost and area. Finally note that with a single $BNet$, pEMesh outperforms ANet due to the high queueing delays at receiving clusters.

Figure 2c shows the effect of varying off-chip bandwidth.

Type	$ANet_A$	$pEMesh_A$
AMAT	6.26	9.26
On-chip base latency	2.71	5.12
On-chip queueing delay	0.78	1.37
Off-chip queueing delay	2.77	2.77

TABLE III: Breakdown of average memory access latency (AMAT) (in processor cycles) for ANet$_A$ and pEMesh$_A$

Performance is dominated by off-chip bandwidth at low bandwidths, and all configurations perform equally. Till a certain threshold B_{thres}, which depends on the network and protocol, performance stays sensitive to the off-chip bandwidth only. Once this threshold is exceeded, differences can be observed between networks and protocols. There is another point B_{sat} after which performance levels off due to minimal queueing going off-chip. Stated another way, the off-chip bandwidth exceeds application demands. (Figure 2c shows B_{thres} and B_{sat} for ANet with the ACKwise protocol).

In our experiments, we find that the ANet outperforms pEMesh due to its higher bandwidth, lower latency and broadcast capabilities. As seen from our studies, this is true with the ACKwise$_i$ protocol or with the Dir$_i$B protocol. To demonstrate this fact, we measure the contribution to the average memory latency due to on-chip base latency, on-chip queueing delay, and off-chip latency. The on-chip base latency ignores all queueing delays. On-chip queueing latency is the delay introduced by buffers in the on-chip network. Off-chip queueing delay is queueing delay going to DRAM. Results are shown in Table III.

$ANet_A$ outperforms $pEMesh_A$ both in terms of both on-chip bandwidth and base latency. Compared to $pEMesh_A$, $ANet_A$ exhibits 47.1% lower on-chip base latency and 43.1% lower on-chip queueing delay. $ANet_A$ also outperforms $pEMesh_A$ in its broadcast capability, but could do so more significantly if the number of broadcast networks were increased in proportion to the amount of broadcast traffic (as illustrated in Figure 2b).

V. Related Work

CMOS-compatible nanophotonic devices are an emerging technology. Therefore there have only been a few architectures proposed that use them for on-chip communication: Corona [7], the optical cache-coherence bus of Kirman et al [8], and the switched optical NoC of Shacham et al [9].

The Corona architecture primarily differs from ATAC in the way that it assigns communication channels. While Corona assigns a physical channel to each receiver and uses WDM to send multiple bits of a dataword simultaneously, ATAC assigns a physical channel to each sender and uses WDM to carry multiple channels in each waveguide, thereby eliminating contention and the need for arbitration. Kirman et al [8] design a cache-coherent hierarchical opto-electronic bus, consisting of a top-level optical broadcast bus which feeds into small electrical networks connecting groups of cores. The design of their network is similar to ATAC but is limited to snooping cache coherence traffic whereas ATAC is composed of a network supporting a general communication mechanism and a coherence protocol (i.e, *ACKwise*) designed to scale to hundreds of cores. Shacham et al [9] propose a novel hybrid architecture in which they combine a photonic mesh network with electronic control packets. Their scheme is still partially limited by the properties of electrical signal propagation since they use an electronic control network to setup photonic switches in advance of the optical signal transmission. It only becomes efficient when a very large optical payload follows the electrical packet. ATAC, on the other hand, leverages the efficiencies of optical transmission for even a single word packet.

Batten et al. [10] take a different approach and use integrated photonics to build a high-performance network that connects cores directly to external DRAM. However, their design does not allow for optical core-to-core communication. An ATAC processor could leverage their design to connect its memory controllers to DRAM.

Previous limited directory schemes invoke software support [11] or assume all cores as sharers [5] when the sharing degree exceeds the capacity of the sharer list. However, some schemes always ensure that the number of sharers is less than the capacity of the sharer list [5]. The *ACKwise* protocol, on the other hand, intelligently keeps track of the number of sharers once the capacity of the sharer list is exceeded.

VI. Conclusion

The recent advances of optical technology have inspired confidence in computer architects that optics will continue to make their way into smaller and smaller packages; just as optical interconnect has moved from connecting cities to connecting data centers to connecting peripherals, it seems likely that it will soon connect chips and on-chip components. Overall, this paper presented a novel manycore architecture that scales to 1024 cores by embracing new technology offered by recent advances in nanophotonics. This paper also introduced ACKwise, a novel directory-based cache coherence protocol that is tailored to on-chip networks with broadcast support.

References

[1] J. Psota et al., "ATAC: All-to-All Computing Using On-Chip Optical Interconnects," in *BARC*, 1/2007.

[2] M. Taylor et al., "Evaluation of the Raw Microprocessor: An Exposed-Wire-Delay Architecture for ILP and Streams," in *ISCA*, 2004.

[3] J. Miller, J. Psota, G. Kurian et al., "ATAC: A Manycore Processor with On-Chip Optical Network," MIT, Technical Memo, 2009.

[4] N. E. Jerger, L.-S. Peh, and M. Lipasti, "Virtual circuit tree multicasting: A case for hardware multicast support," in *ISCA*, 2008.

[5] A. Agarwal et al, "An evaluation of directory schemes for cache coherence," in *ISCA*, 1988.

[6] C. Bienia et al., "The PARSEC Benchmark Suite: Characterization and Architectural Implications," in *PACT*, 2008.

[7] D. Vantrease et al, "Corona: System Implications of Emerging Nanophotonic Technology," in *ISCA*, 2008.

[8] N. Kirman et al., "Leveraging Optical Technology in Future Bus-based Chip Multiprocessors," in *MICRO*, 2006.

[9] A. Shacham et al., "Photonic NoC for DMA Communications in Chip Multiprocessors," in *Hot Interconnects*, Aug 2007.

[10] C. Batten et al., "Building manycore processor-to-dram networks with monolithic silicon photonics." in *Hot Interconnects*, Aug 2008, pp. 21–30.

[11] D. Chaiken et al., "Limitless Directories: A scalable cache coherence scheme," in *ASPLOS*, 1991.