2012 IEEE 26th International Parallel and Distributed Processing Symposium

Cross-layer Energy and Performance Evaluation of a Nanophotonic Manycore Processor System using Real Application Workloads

George Kurian, Chen Sun, Chia-Hsin Owen Chen, Jason E. Miller, Jurgen Michel, Lan Wei
Dimitri A. Antoniadis, Li-Shiuan Peh, Lionel Kimerling, Vladimir Stojanovic and Anant Agarwal
Massachusetts Institute of Technology, Cambridge, USA
{gkurian, sunchen, owenhsin, jasonm, jmichel, lanwei, antoniadis, peh, lckim, vlada, agarwal}@mit.edu

***Abstract*—Recent advances in nanophotonic device research have led to a proliferation of proposals for new architectures that employ optics for on-chip communication. However, since standard simulation tools have not yet caught up with these advances, the quality and thoroughness of the evaluations of these architectures have varied widely. This paper provides the first complete end-to-end analysis of an architecture using on-chip optical interconnect. This analysis incorporates realistic performance and energy models for both electrical and optical devices and circuits into a full-fledged functional simulator, thus enabling detailed analyses when running actual applications.**

Since on-chip optics is not yet mature and unlikely to see widespread use for several more years, we perform our analysis on a future 1000-core processor implemented in an 11nm technology node. We find that the proposed optical interconnect can provide between 1.8x and 4.8x better energy-delay product than conventional electrical-only interconnects. In addition, based on a detailed energy breakdown of all processor components, we conclude that athermal ring resonators and on-chip lasers that allow rapid power gating are key areas worthy of additional nanophotonic research. This will help guide future optical device research to the areas likely to provide the best payoff.

***Index Terms*—on-chip networks; photonics; cache coherence**

I. Introduction

The trend in modern microprocessor architectures is clear: multicore is here to stay. As process technologies continue to scale down, processor designers are able to place exponentially more cores on a single chip. If current trends continue, processors containing hundreds, if not thousands, of cores will be available in five to ten years. However, it is not yet clear how to make these cores work together to provide scalable performance, energy efficiency, and tractable programmability. One of the key components in this quest is the on-chip interconnection network.

Emerging nanophotonic technology [1] has yielded a rich design space for on-chip optical-electrical architectures [2], [3], [4], [5], [6]. Since standard simulation tools have not yet caught up with these advances, however, the quality and thoroughness of architectural evaluations have varied significantly. The goals of this paper are to provide the first complete end-to-end analysis of an architecture using on-chip optical interconnect and help guide future optical device research to the areas likely to provide the best payoff.

This paper extends the field of CMOS-integrated nanophotonic-based processor architectures through the following contributions (discussed briefly below):

1) Presents the first realistic at-scale evaluation of a 1000-core processor at the 11 nm technology node, running actual applications and accurately capturing the interplay between applications, multicore hardware architecture, and the underlying electronic and photonic devices.
2) Performs the first full-fledged performance, power and area analysis for an optical many-core architecture (the ATAC architecture [6]) using our simulation infrastructure.
3) Proposes novel optimizations of the ATAC architecture to better leverage opto-electronic technology for improved performance and power consumption.
4) Guides future nanophotonic device research by identifying the components that impact system performance and power the most.

The evaluation in this paper is the first to fully integrate all levels of the system from electrical and optical device models to applications. We have augmented an architectural simulator with performance, area, and energy models for key electrical and optical components based on models of a projected 11 nm process technology. This enables a complete end-to-end evaluation using real application workloads to exercise network, cache and core models, including effects of back pressure on the application and the cache coherence protocol. Previous studies [7], [4], [5], [8] have used coarse, higher-level models and/or unrealistic traffic patterns like synthetic workloads or captured traces. These studies have the following weaknesses:

1) Synthetic traffic and trace-driven approaches do not propagate network delay back to the application and thereby other processor components.
2) They estimate only network energy and do not consider how it relates to total system energy.
3) They do not capture the impact of network delays on

This work was funded by NSF and the U.S. Government under the DARPA UHPC program. The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

1530-2075/12 $26.00 © 2012 IEEE
DOI 10.1109/IPDPS.2012.103

non-data-dependent (consumed regardless of usage, such as leakage and ungated clocks) core/cache energy.

In this paper, we build upon our previous work on the ATAC architecture [6], which employs optics to achieve fast and energy-efficient global connectivity. Compared to conventional electrical networks, ATAC promises better performance and programmability by providing uniform communication costs and efficient cache coherence [9] for massively parallel processors. Though these properties are appealing, our previous work evaluated only performance and had very few results for 1000-core processors.

This paper presents a thorough analysis of power, performance and area for a complete cache-coherent 1000-core multicore processor. Based on our results, we propose changes to the ATAC architecture designed to improve energy efficiency without compromising performance. This includes the ability to adaptively switch network links between broadcast and unicast modes using integrated on-chip lasers. We also make changes to the network topology and routing algorithm to maximize performance and minimize energy. We call the new architecture ATAC+. Our results show that the ATAC+ architecture can provide between 1.8x and 4.8x better energy-delay product than competitive electrical-only architectures.

CMOS-integrated nanophotonics is still an immature technology and therefore the exact capabilities of these devices have not been fully established. Although basic functionality of the different optical devices is well known (lasers, modulators, waveguides, filters, *etc.*), the exact properties of those devices are still somewhat uncertain. This is especially true when considering what CMOS-integrated nanophotonics will ultimately be capable of in the future rather than just what has been demonstrated today. Furthermore, different tradeoffs can be made when designing these devices, resulting in a variety of properties. By evaluating ATAC+ with multiple sets of assumptions about future optical technology, we provide insight into the most critical device properties and help guide nanophotonic researchers towards the improvements that provide the greatest benefit to architects. In particular, we find that athermal ring resonators and lasers that allow for rapid power gating are important areas for future research, while ultra-low-loss optical devices are less valuable.

The remainder of this paper is structured as follows. Section II provides some background on the nanophotonic technology assumed in this work. Section III recaps the basic ATAC [6] architecture. Section IV presents the proposed improvements and the ATAC+ architecture. Section V explains the evaluation framework and performs a wide range of experiments to evaluate the system-level area, energy and performance implications of on-chip photonic networks, including comparing variants of the ATAC+ architecture with two alternative electrical-only architectures. Section VI discusses related work and Section VII concludes.

II. Silicon Photonic Technology

Advances in electronic-photonic integration have enabled tighter electrical-optical integration and dense wavelength division multiplexing for ever-higher bandwidth densities [10], [11], [12], [13]. Recent research has shown that optical devices can be built using standard CMOS processes [14], [1], leading to the possibility that optics can replace global electrical wires and on-chip buses [15], which scale poorly with technology.

The components of a nanophotonic communication fabric include a multi-wavelength laser source, waveguides for routing the optical signals on-chip, modulators for imprinting bit streams onto wavelengths, resonant rings for wavelength-selective filtering, and photodetectors/receivers for converting the optical signal back into the electrical domain. In this paper, we assume the same basic optical components as those in the original ATAC paper but also consider two emerging optical technologies: on-chip lasers and athermal ring resonators. Both of these technologies are immature and would require a substantial additional research investment to fully realize. One of the key goals of this paper is to evaluate the potential impact of these technologies to determine if that investment is warranted.

A. On-chip Germanium Lasers

Recent research in nanophotonic technology has made on-chip Ge lasers a possibility [16]. On-chip lasers have the advantage that they avoid coupling and distribution losses (suffered when bringing in light from an off-chip source) and are close enough to be shut down and restarted very quickly. Off-chip lasers, on the other hand, are difficult to turn on/off due to the communication latency from the sender to the laser as well as the energy needed to communicate off-chip. Hence, their power dissipation is static and irrespective of the utilization of the optical channel. Static laser power is a sizable fraction of the network power [8] and prior works [5], [8], [17] aim to reduce this overhead by efficiently using a small number of optical channels or by using laser guiding/sharing techniques. Based on current progress in laser technology, we estimate that on-chip lasers that can be turned on and off within 1 ns will be available in the 11 nm technology node we are targeting in this work.

B. Athermal Ring Resonators

A significant source of power consumption in many proposed optical architectures is the heaters used to thermally tune the frequency response of the ring resonators. These heaters are necessary because traditional resonators shift their frequency as their temperature changes. Recent advances have shown that it is possible to design CMOS-compatible ring resonators that have less than 1 pm/K shift over a 100 C range [18]. The approach employed uses a polymer with negative thermo-optic coefficient to compensate for the positive thermo-optic coefficient of the Si ring. An alternative approach [19] uses the ring resonator in connection with a Mach Zehnder interferometer to compensate for the temperature shift. This approach shows athermal behavior over a 40 C range.

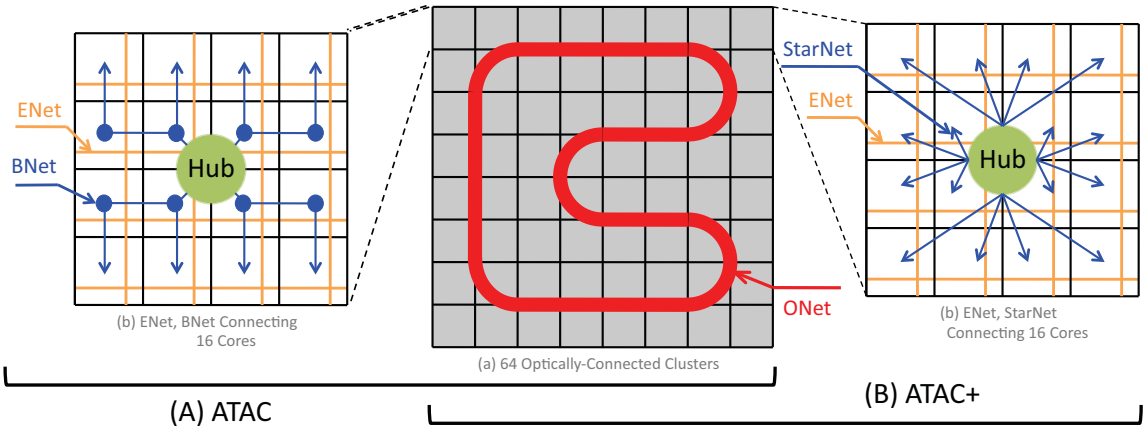

Fig. 1. (A) ATAC architecture. (B) ATAC+ architecture. Changes over the ATAC architecture include the adaptive SWMR link, the point-to-point StarNet, and the distance-based routing protocol for unicasts.

III. ATAC Architecture Recap

The original ATAC processor uses a tiled multicore architecture with an electrical mesh network augmented with an optical network. Here we briefly review the network architecture and memory system. See [6] for more detail.

A. Network

The underlying electrical architecture consists of a 2-D array of processing cores connected by a conventional point-to-point, packet-switched mesh network (called the *ENet*) like those seen in other multicore processors [20], [21], [22]. Each core in ATAC contains a single-issue, in-order RISC pipeline with data and instruction caches.

To this electrical baseline, a global optical interconnect is added (the *ONet* shown in Figure 1(A)(a)) based on state-of-the-art optical technology. Whereas the ENet is ideal for predictable, short-range point-to-point communication, the ONet provides low-latency, energy-efficient global and long-distance communication. In the 1024-core ATAC architecture, cores are grouped into 64 "clusters", each containing 16 cores. Each cluster contains a single ONet endpoint called a *hub*. The hub is responsible for interfacing between the optical components of the ONet and the electrical components within a cluster.

Individual cores are connected to the hub in two ways: data going from a core to the hub uses the standard ENet; data going from the hub to the cores uses the BNet, a small electrical broadcast network (shown in Figure 1(A)(b)). Because the BNet is dedicated to broadcasts, it is essentially a fanout tree and requires no routers, crossbars, or internal buffering. The ATAC network uses a 64-bit wide ONet (64 optical waveguides for data); one 64-bit wide electrical ENet; and two parallel 64-bit wide BNets.

The key to efficient global communication in a large ATAC chip is the optical ONet. The ONet provides a low-latency, contention-free connection between the hubs in each cluster. Hubs are interconnected via waveguides that visit every hub and loop around on themselves to form continuous rings (shown in Figure 1(A)(a)). 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. Each hub's filters are tuned to extract approximately 1/64th of the signal, allowing the rest to pass on to the downstream hubs. Thus every transmission on the ONet is a fast, efficient broadcast.

The ONet uses wavelength division multiplexing (WDM) to avoid contention. Each hub's modulators are tuned to send on a unique wavelength. Each hub also contains receive filters tuned to all of the other wavelengths. This eliminates the need for arbitration in the optical network. This makes the ONet functionally similar to a broadcast bus, but without any bus contention.

Besides broadcasts, optical technology also allows efficient long-distance point-to-point communication. Initiating an optical signal (*i.e.,* switching the modulator) requires more energy than switching a short electrical wire. However, once generated, the optical signal can quickly travel anywhere on the chip without the need for repeaters. To avoid wasting excessive power and resources delivering these unicast messages to all cores, ATAC includes filtering at the receiving hubs and cores. Packets labeled as intended for a single core are only rebroadcast on the BNet of the cluster containing that core. In addition, the other cores in that cluster will drop the packet immediately, rather than process it.

To summarize, in ATAC, all unicast messages between cores on different clusters use the ONet (as well as the ENet and BNet to get to and from the ONet) while those between cores on the same cluster use only the ENet.

B. Memory Subsystem

Each core in ATAC contains private L1 and L2 caches. The data in the L2 caches across all cores on the ATAC

chip are kept coherent using a modified limited directory-based coherence protocol called *ACKwise*. The directory is distributed evenly across all the cores. Furthermore, each core is the "home" for a set of addresses (the allocation policy of addresses to homes is statically defined).

Each entry in the directory managed by ACKwise can hold a maximum of k hardware pointers. ACKwise$_k$ operates like a full-map protocol when the number of sharers is less than or equal to k (like all other limited directory-based protocols). When the number of sharers exceeds k, a global bit is set and the sharer list is replaced by the total number of sharers. On an exclusive request to such an address, an invalidate request is broadcast but acknowledgements are received only from the actual sharers of the cache block.

When cores need to communicate with external memory, they do so via 64 on-chip memory controllers. Each cluster has one core replaced by a memory controller. After receiving requests through the ATAC network, the memory controller communicates with external DRAM modules through I/O pins. Replies are then sent back to the processing cores through the ATAC network. The choice of I/O bus technology is independent of the on-chip network architecture since the memory controller is performing a translation. However, to support the large number of memory controllers needed for a 1000-core chip, we assume that the connection to memory is optical as well (built using technologies such as those described in [17]).

IV. ATAC+ Architecture

The ATAC architecture presented above is optimized for broadcast traffic and performs sub-optimally using more energy than necessary for unicasts due to the following three reasons.

First, ATAC uses off-chip lasers provisioned for broadcast traffic. This causes the laser to operate at full power even when the link is idle or serving unicasts. Note that the laser power provisioned for broadcasts is approximately a linear function of the number of receivers on the optical link.

Second, the broadcast BNet, being a fanout tree, sends a unicast message to all the cores in a cluster while it needs to be received by only one, thereby causing unnecessary energy consumption.

Finally, the cluster-based routing policy in ATAC dictates that all inter-cluster unicast communication flows over the optical network (ONet). This leads to a lot of links in the electrical mesh network (ENet) remaining unutilized even though the bisection bandwidth of the ONet and the ENet are identical (64 flits/cycle) for unicast traffic.

We introduce three changes to the ATAC architecture to solve the problems discussed above: (1) an adaptive SWMR (Single Writer Multiple Reader) optical link to address the static laser concerns, (2) a point-to-point StarNet electrical network to solve the problem posed by the BNet and (3) a distance-based unicast routing protocol to achieve good performance for unicast traffic. The ATAC architecture incorporating these modifications is called the ATAC+ architecture (shown in Figure 1(B)). The impact of each of these modifications is evaluated in Section V.

A. Adaptive SWMR Link

Section II introduced on-chip Ge lasers that could be switched on/off within a time interval of 1 ns. In addition to this capability, the bias current in the Ge laser can be adjusted at runtime to generate varying output power, the variation also occuring within 1 ns. The above two capabilities of the laser are ideal for building an optical link that operates in idle, unicast and broadcast modes and switches dynamically between them. Since the laser can be controlled so as to generate only enough power for the operation to be performed, there is no energy wastage.

Since the laser modes and recipients of a message can be changed dynamically, a mechanism is needed to notify the receivers on the optical link when a message is intended for them. This is done using a select link (in addition to the already existing data link). Before a message is sent, the appropriate receiver(s) are notified via the select link to tune-in. The receiver(s) remained tuned-in until they receive the last flit of the message at which point they tune-out. The throughput of the optical link depends on the rate at which the ring resonators at the receiver can "tune-in"/"tune-out". Fortunately, this can be done within 1 ns electrically using a charge depletion structure. This implies that the notifications on the select link should be sent exactly 1 ns before the message is transmitted on the data link.

The operation of the optical link in the idle, unicast and broadcast modes is shown in Figure 2 and explained briefly below. In the idle mode, the lasers associated with both the select and data links are turned off. The ring resonators on the select links are tuned-in while those on the data link are tuned-out. In fact, the ring resonators on the select links always remain tuned-in since the laser modes and the recipients of a message can be changed dynamically. To switch to unicast mode, the laser is turned on and its power is adjusted to that for one receiver. A notification is then sent on the select link to cause the intended receiver to tune-in. To switch to broadcast mode, the laser power is adjusted accordingly and a notification is sent on the select link to cause all receivers to tune-in.

Once the receiver(s) are tuned-in, the message is sent on the data link. The ring resonators on the data link remain tuned-in until the tail flit of the message is received. This includes any time spent stalled waiting for the downstream buffers to free-up. The laser, on the other hand can switch to idle mode during the stall phase, thereby expending energy only while doing useful work. In this paper, we implement a wormhole flow control scheme with a single virtual channel. With this scheme, messages to different receivers from the same sender cannot be interleaved on the data link and hence setup (via the select link) and teardown (when the tail flit is received) of a connection only occur at the start and end of a message.

The select link is log(C) bits wide where C is the number of hubs on the adaptive SWMR link (note that a hub is not

allowed to unicast a message to itself).

The unicast mode of the adaptive SWMR link is identical in operation to the reservation-assisted link proposed in Firefly [4].

The adaptive SWMR link's broadcast mode is used for sending invalidation broadcasts required by the ACKwise [6] coherence protocol and the unicast mode is used for sending long-distance coherence and data messages. The broadcast mode is extremely useful because our previous work [6] revealed significant performance improvements when using the ACKwise protocol with a network that has native broadcast support.

B. StarNet Electrical Network

As mentioned earlier, the broadcast BNet is not an ideal network for unicast communication. Hence, it is replaced with a point-to-point electrical network called the StarNet shown in Figure 1(B)(b).

The StarNet consists of a 1-to-16 demultiplexer and 16 point-to-point links that forward data from the hub to the cores in the receiving cluster. A unicast uses only one StarNet link while a broadcast uses all 16 StarNet links. The dynamic energy consumption of a unicast on the StarNet is much lower than that on the BNet ($\sim \frac{1}{8}$th). Although the energy consumption of a broadcast is twice that on the BNet, this can be tolerated since broadcasts are relatively rare compared to unicasts.

The area overhead of replacing the BNet with the StarNet is negligible (evaluated later in Section V-D). The performance of the StarNet is exactly the same as the BNet. Both the StarNet and BNet have single-cycle latencies since at the 11 nm node, the clusters are small enough that a flit can travel from the hub to all cores within one cycle. The system-level energy impact of using the StarNet over the BNet is evaluated in Section V-E.

C. Distance-Based Unicast Routing

In the ATAC network architecture, all broadcast communication as well as unicast communication between cores on different clusters takes place over the optical network (ONet). Only unicast communication between cores in the same cluster takes place purely over the ENet. Note that the ENet is still used to send messages to the hub to be forwarded over the ONet for both inter-cluster and broadcast communication. While the optical ONet is optimized for both energy and performance in the case of broadcasts due to its high receive-weighted bisection bandwidth, it is not the optimal path for all inter-cluster unicast communication. Arriving at the optimal routing policy requires a careful energy and performance analysis of the network.

Performance Analysis: For benchmarks with low network demands, it is optimal to send all inter-cluster unicasts over the ONet due to its low zero-load latency but for benchmarks with high network demands, the routing policy has to balance load between the ENet and the ONet so as to maximize the saturation throughput of the network.

Since the attractiveness of ENet decreases with an increase in communication distance, we use a distance-based routing scheme to decide whether to send a unicast over the ONet or just use the ENet. This routing scheme has a parameter called r_{thres} which is the distance below which a packet is sent completely over the ENet. At r_{thres} or above it, a unicast packet is sent over the ONet. Here, distance is defined as the manhattan distance between the sender and receiver as measured over an electrical mesh network. A distance-based routing scheme with an r_{thres} of i hops is referred to as *Distance-i*.

The optimal value of r_{thres} is small at low network loads due to the low zero-load latency of the ONet. As the network load increases, the optimal value of r_{thres} also increases until it reaches a value that maximizes the saturation throughput of the network. Further increases in the network load do not change the optimal value of r_{thres}.

This is illustrated in Figure 3. In the figure, *Cluster* refers to the routing scheme where all inter-cluster communication is sent over the ONet. A careful look at the figure reveals that the optimal value of r_{thres} changes from 5 to 15 to 25 as the offered load increases (the transition from 5 to 15 is not significant). An r_{thres} of 25 maximizes the saturation throughput of the network and so values above it are never optimal. *Distance-All* refers to the routing scheme where all unicast communication takes place completely over the ENet and the ONet is used only for broadcast communication. *Distance-All* and *Distance-35* are never optimal routing schemes for the network as observed in Figure 3. From the above observations and analysis, we conclude that the the routing scheme that purely optimizes performance of the network is adaptive.

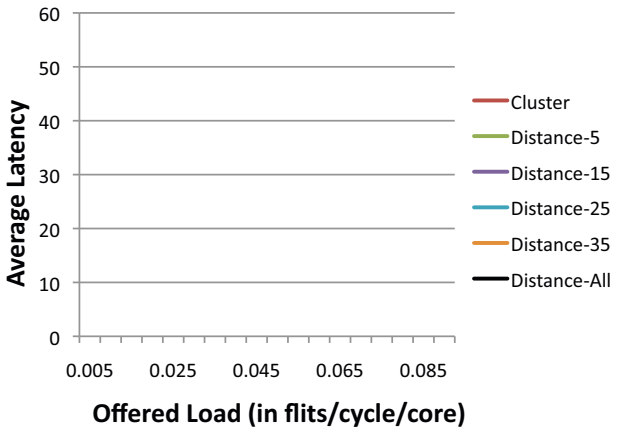

Fig. 3. Latency vs Offered Load graph with uniform random unicast traffic and 0.1% broadcast injection. *Distance*$_i$ refers to the the distance-based routing scheme with an r_{thres} of i hops and *Cluster* refers to the routing scheme where all inter-cluster unicast communication takes place over the ONet.

Energy Analysis: The dynamic energy expended to communicate over the ONet is constant (independent of the receiver) while that expended to communicate over the ENet

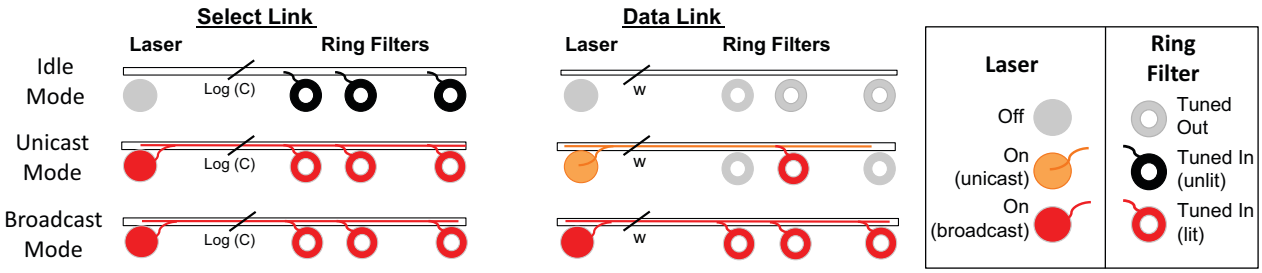

Fig. 2. Architecture of an adaptive SWMR link. The data link is w bits wide (w being the flit size). The select link is $log(C)$ bits wide (C being the number of hubs connected to the link).

is directly proportional to the number of hops between the sender and receiver. When considering purely dynamic or data-dependent energy consumption, our analysis shows that the crossover distance is 8 hops; ENet-only routing is data-dependent energy-optimal when the sender and receiver are less than 8 hops apart, ONet usage is optimal for distances that are 8 hops or more.

The non-data-dependent (NDD) energy, on the other hand, is directly tied to performance. The distance-based routing scheme that maximizes performance will minimize runtime, reducing the time which the network spends burning clock power and leaking. NDD energy consumption of other structures, such as cores and caches, is similarly affected by the runtime, hence a careful system-level study is required to arrive at an optimal value of r_{thres}. For simplicity reasons, we assume an oblivious routing scheme which keeps the value of r_{thres} constant irrespective of the network load. Section V-E performs a study to find an optimal r_{thres}.

1) Sequence Numbers: The above unicast routing algorithm change creates a problem for cache coherence. This is because most coherence protocols require that the coherence messages belonging to the same address flow in FIFO order. This is no longer true since broadcast invalidates potentially take a route different from that of unicasted coherence messages. Note that broadcasts and unicasts still flow in FIFO order among themselves.

To address this problem, we introduce a unique sequence number per directory that is incremented every time a broadcast invalidate is sent. The unicasted coherence messages from the directory carry the same sequence number as the previous broadcast. If a unicast message arrives at its receiver before a previously sent broadcast message, the receiver can detect this scenario and buffer the unicast message until all previous broadcast messages are processed.

On the other hand, the scenario where a broadcast message arrives before a previously sent unicast message cannot be detected since the sequence numbers are incremented only on a broadcast. In this scenario, correctness is maintained by modifying the cache coherence protocol.

A complete proof of correctness of the protocol would involve showing all the states, coherence requests and responses in the protocol and is omitted here due to lack of space. Instead, only a few important points are discussed.

In a directory-based coherence protocol, requests for *exclusive* and *shared* copies of an address from the cache are processed serially at the directory to maintain sequential consistency. In response to *exclusive* and *shared* requests from the cache, the directory sends out *invalidate*, *flush* and *write-back* requests to remote caches and waits for acknowledgements. An *invalidate* request is sent on an *exclusive* request for an address in *shared* state. A *flush* request is sent on an *exclusive* request for an address in *modified* state and a *write-back* request is sent on a *shared* request for an address in *modified* state. The cache line is either piggy-backed with one of the acknowledgements or fetched explicitly from main memory. Once the directory receives all the acknowledgements and the cache line, it sends an *exclusive* or *shared* response containing the cache line to the requester and moves onto processing the next request. An *exclusive* response sets the address in *modified* state and a *shared* response sets the address in *shared* state.

The only broadcast message in this protocol is the *invalidate* broadcast sent by the directory on an *exclusive* request for an address that is cached in multiple locations in *shared* state. An address can be cached only in a single location in *modified* state. The *invalidate* broadcast cannot arrive ahead of a previous *write-back* or *flush* request since these have to be explicitly acknowledged before any further messages can be sent from the directory for the same address. It also cannot arrive ahead of a previous *exclusive* response since the only messages that can be sent directly after an *exclusive* response are the *write-back* and *flush* requests that have to be explicitly acknowledged. So, the *invalidate* broadcast can only arrive ahead of a previously sent *shared* response. In the event that this happens, the miss status holding register (MSHR) of the cache will have an entry indicating an outstanding *shared* request for that address. If such an entry exists, the *invalidate* broadcast has potentially arrived out-of-order and is buffered for later processing. Buffering the *invalidate* broadcast does not create a deadlock since the ACKwise coherence protocol only requires acknowledgements from the actual sharers of an address in response to a broadcast invalidation. When the *shared* response finally arrives, it is determined whether the *invalidate* broadcast arrived out-of-order by comparing their sequence numbers. If it did not arrive out of order, the *invalidate* broadcast is simply dropped. If it did arrive out-of-order, the *invalidate* broadcast is processed one cycle after

processing the *shared* response. This ensures that the messages are processed in the order in which they were sent from the directory, hence maintaining correctness.

Storage Overhead: The storage overhead for the sequence numbers is negligible. Each core has to store a sequence number for every directory slice. There are a total of 64 directory slices in the system. Each sequence number is 2 bytes long. This leads to a storage overhead of 128 bytes per core or a total of 128 KB for the entire chip.

Network Traffic Overhead: There is no network traffic overhead since the 2 bytes can be accommodated in a network packet without creating additional flits. A coherence message requires 64 bits for the address, 20 bits for the sender and receiver IDs' and 4 bits for the message type, a total of 88 bits. A data message, on the other hand, requires 512 bits for the data (cache block size is 64 bytes), 64 bits for the address, 20 bits for the sender and receiver IDs' and 4 bits for the message type, a total of 600 bits. The flit size employed in the network is 64 bits. Hence, adding 16 bits for the sequence number does not create any additional flits.

Sequence Number Overflow: When the sequence numbers overflow, they simply wrap around. Ordering between packets is calculated in the same way as with TCP/IP sequence numbers. Correctness is only affected in the event that there are more than 2^{15} broadcast messages in transit simultaneously from the same directory slice. This is theoretically impossible due to the buffering limits of the interconnection network.

V. Evaluation

In this section, we perform application-driven full-system evaluations of the ATAC+ architecture. First we describe our experimental methodology and technology assumptions in Section V-A. Next, in Section V-B, we look at performance and application characteristics for the benchmarks we use. We then describe the various scenarios considered for future photonic technology and compare the costs of the architecture under each in Section V-C. In Section V-D, we evaluate the adaptive SWMR link in more detail and perform bandwidth sensitivity studies on the ATAC+ network. Previously, we motivated both distance-based routing and the cluster-level StarNet. The impact of these improvements is detailed in Section V-E. Section V-F explores the synergy between the ATAC+ architecture and cache coherence protocols and, finally, Section V-G uses a first-order core model to demonstrate the role of the core in overall system-level energy consumption.

A. Experimental Setup

We compare the ATAC+ architecture against two electrical baselines: EMesh-Pure and EMesh-BCast. EMesh-Pure is a plain electrical mesh and EMesh-BCast allows multicasting at each router, providing native hardware support for broadcasts. Wormhole flow-control and oblivious routing is assumed for all networks. Cache sizes are chosen to fit a 1024-core architecture on a single die at the 11 nm node.

We choose seven representative applications from the SPLASH-2 [23] benchmark suite along with a dynamic graph application [24] that finds strongly connected components. We adopt the Graphite [25] distributed multicore simulator to run applications and model the performance tradeoffs between various network configurations and coherence protocols. The primary performance-related parameters are shown in Table I. With the exception of Section V-F, all experiments use the ACKwise$_4$ [6] coherence protocol.

Common Parameters	**Value**
Frequency (Cores and Network)	1 GHz
Core Type	in-order, single-issue
L1-I Cache	Private, 32 KB
L1-D Cache	Private, 32 KB
L2 Cache	Private, 256 KB
Total Memory Controllers	64
Bandwidth per Mem. Controller	5 GBps
Memory Latency	100 ns
Electrical Mesh Parameters	**Value**
Router Delay	1 cycle
Link Delay	1 cycle
Flit Size	64 bits
ATAC+ Network Parameters	**Value**
ONet Link Delay	3 cycles
ONet Select - Data Link Lag	1 cycle
ENet Link Delay	1 cycle
StarNet Link Delay	1 cycle
Total StarNets per Cluster	2
Electrical Router Delay	1 cycle
Flit Size	64 bits

TABLE I
Network Parameters

For power and area evaluations of on-chip electrical routers, links, and hubs, we use the DSENT [26] tool. Power and area estimates for the L1, L2, and directory caches are obtained using McPAT [27]. We use the built-in photonic models of DSENT, which are based on the modulator, receiver, and thermal tuning models of [28], and expand the existing point-to-point link model to that of the adaptive SWMR link used in the ATAC+ network. Our key photonic technology parameters are summarized in Table II. Parameters not shown are assumed to be those used in the full link evaluation of [28].

Parameter	**Value**
Laser Efficiency	30 %
Waveguide Pitch	4 μm
Waveguide Loss	0.2 dB/cm
Waveguide Non-linearity Limit	30 mW
Ring Through Loss	0.0001 dB [17], [7]
Ring Drop Loss	1.0 dB
Ring Area	100 μm^2
Photodetector Responsivity	1.1 A/W

TABLE II
Optical Technology Parameters

As the ATAC+ architecture assumes fairly mature photonic devices that are likely not available until the end of the

decade, it is necessary to compare with a suitably advanced electrical technology. To this end, we derive models for a tri-gate 11 nm electrical technology node using the virtual-source transport models of [29] and the parasitic capacitance model of [30]. These models are used to obtain electrical technology parameters (Table III) used by both McPAT and DSENT. As clock frequencies are relatively slow, high threshold (HVT) transistors are assumed for lower leakage.

Parameter	Value
Process Supply Voltage (V_{DD})	0.6 V
Gate Length	14 nm
Contacted Gate Pitch	44 nm
Gate Cap / Width	2.420 fF/μm
Drain Cap / Width	1.150 fF/μm
Effective On Current / Width (N/P)	739/668 μA/μm
Off Current / Width	1 nA/μm

TABLE III
PROJECTED TRANSISTOR PARAMETERS FOR 11 NM TRI-GATE

The overall toolflow is as follows. Graphite runs a benchmark for the chosen network (ATAC+, electrical mesh, etc.), cache configuration, and cache coherence protocol, producing event counters and performance results. The specified cache and network configurations are also fed into McPAT and DSENT to obtain area, static power, and dynamic per-event energies for each component. Event counters and completion time output from Graphite are then combined with per-event energies and static power to obtain the overall energy usage of the benchmark.

B. Application Performance

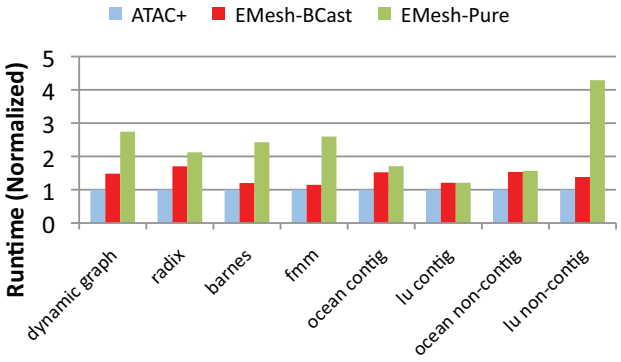

Fig. 4. Application runtime comparison

We first compare performance profiles of the 8 evaluated applications. The runtime of each application is shown in Figure 4 for the architectures of interest and the distribution of broadcast vs. unicast traffic is shown in Figure 5. All traffic is due to cache coherence so the differences are the result of different sharing patterns. Average network injection is shown in Figure 6 for ATAC+ as a metric of network utilization and loads for each application.

In all cases, ATAC+ commands a sizable lead over both EMesh-Pure and EMesh-BCast. Note that without hardware broadcast support, EMesh-Pure performs broadcasts by sending multiple unicast messages in succession, severely degrading performance for broadcast-heavy applications (`dynamic graph`, `radix`, `barnes`, `fmm`). Though EMesh-BCast improves upon EMesh-Pure for broadcasts, ATAC+ retains its overall latency advantage. For applications with frequent cache misses and higher network loads (`radix`, `ocean contig` `ocean non-contig`), the latency advantage translates to a large runtime advantage over EMesh-BCast for ATAC+.

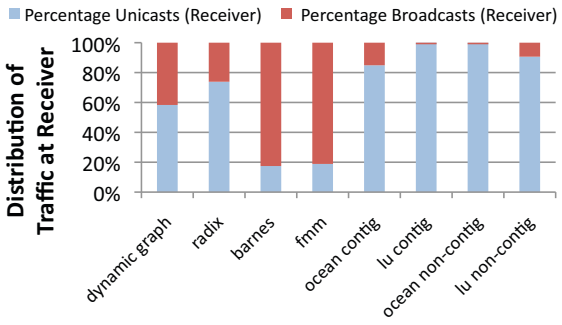

Fig. 5. Percentage of Unicast and Broadcast Traffic (as measured at the receiver)

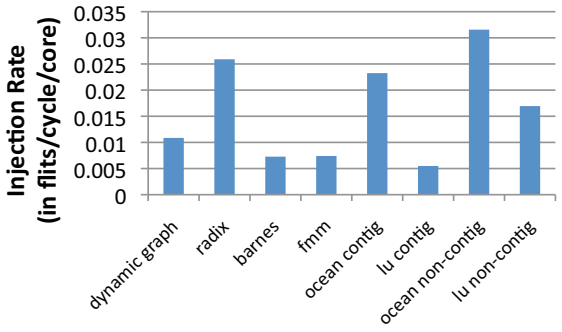

Fig. 6. Offered Network Load (measured as flits/cycle/core)

C. Technology Scenario Comparison

Though reasonable nanophotonic maturity is expected by the thousand-core timeframe, it is critical for architects to understand the relative importance of various optical device features in order to guide nanophotonic device research towards those that impact system performance and power the most.

Laser Power Gating and Athermal Rings

Laser power gating and athermal ring resonators are interesting device features that can potentially lower network power consumption. However, fast-switching on-chip lasers have not been demonstrated with current technology nor is it clear whether athermal rings can overcome process mismatches in a commercially scalable way [28]. This study tries to understand how critical these device features are and evaluates the energy characteristics of a nanophotonic network without them.

We consider four different flavors of the ATAC+ architecture, shown in Table IV, each representing different technology

	Optical Devices	Laser	Ring Temperature Dependence
ATAC+ (Ideal)	Ideal	Power-Gated	Athermal
ATAC+	Practical	Power-Gated	Athermal
ATAC+ (RingTuned)	Practical	Power-Gated	Tuned
ATAC+ (Cons)	Practical	Standard	Tuned

TABLE IV
ATAC+ ARCHITECTURE FLAVORS

features. ATAC+(Cons) represents a conservative estimate, for which both laser power-gating and athermalized rings are unavailable. Under this scenario, the laser must be kept on at worst-case power (the power needed to reach all receivers during a broadcast), even when the channel is idle. Rings also require tuning and we assume the electrically-assisted tuning strategy as outlined in [28]. ATAC+(RingTuned) allows laser power-gating, while rings still require tuning. Both power-gating and athermal rings (no tuning required) are enabled for ATAC+. We include ATAC+(Ideal), which assumes ideal (zero-loss) optical devices and a 100% efficient laser, as another comparison point.

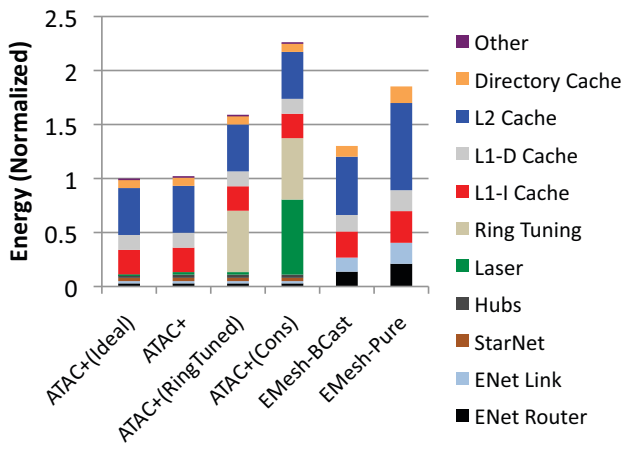

Fig. 7. Total Energy Breakdown of the ATAC+ and Electrical Mesh Networks averaged across all benchmarks [normalized with respect to ATAC+(Ideal)]. *Other* includes energy consumed by modulators and receivers on the optical link.

Figure 7 shows the energy breakdown of the ATAC+ variants, and the electrical mesh networks, averaged across all eight evaluated benchmarks. We observe that the *Laser* is a significant energy consumer should power-gating be unavailable (ATAC+(Cons)). Lacking any form of output power control, the laser is always at full-blast, wasting energy when idle or when only unicasts are desired. With a large number of rings (∼260K) present in ATAC+, both ATAC+(Cons) and ATAC+(RingTuned) suffer from high ring tuning costs due to ring heating. With both laser power-gating and athermal rings, ATAC+ takes the energy-efficient lead against EMesh-BCast. Interestingly, ATAC+ has about the same energy as ATAC+(Ideal), as idealized devices help to reduce the laser power which is already a tiny fraction of ATAC+ (∼2%).

For ATAC+ and the baseline mesh networks, the cache energy dominates (>75%) the combined total energy. The majority of the cache energy is by the private L2 caches, evenly split between the leakage and dynamic components. The L1-I and L1-D caches, though many times smaller, have a larger fraction of dynamic energy consumption since they receive many more accesses than the L2. The directory cache consumes the least since it is small (the ACKwise$_4$ protocol tracks only 4 unique hardware sharers) and receives fewer accesses than the L1-I and L1-D caches.

From our results, we highlight the importance of *non-data-dependent* (NDD) energy consumption, which is power consumed regardless of usage or activity. Leakage, ungated clocks, ungated laser and ring heating are all sources of non-data-dependent energy consumption. Under low flit injection rates or when broadcasts are used sparingly, (the case with the shown applications) any sources of NDD energy consumption will quickly dominate. Similarly, if an application takes longer to complete, the system components idle longer and consume more NDD energy in the process. The fact that ATAC+ can complete applications faster than the mesh baselines (thus preventing additional NDD energy from being consumed) results in significant energy savings. For example, in Figure 7, the higher energy consumption of the L2 cache in the mesh baselines is due to the longer runtimes of the application when compared to the ATAC+ network.

Figure 8 shows the normalized energy-delay product metric for the ATAC+ variants and mesh baselines running each application. As expected, ATAC+ architecture has an almost identical E-D product as ATAC+(Ideal) architecture. The EMesh-BCast and EMesh-Pure networks have on average a 1.8x and 4.8x higher E-D product than ATAC+ respectively.

Waveguide Loss

Another important parameter of optical networks is the waveguide loss. The higher the waveguide loss, the higher the laser power that needs to be supplied at the input of the waveguide to yield the same power at its output. We vary the waveguide losses from 0.2 dB to 4 dB and show the resulting energy consumption in Figure 9. The results are normalized to the energy consumption of the EMesh-BCast network. It is found that the ATAC+ network can tolerate a loss of upto 2 dB before its energy consumption exceeds that of the EMesh-BCast network. Hence, if the optical technology is mature enough to enable laser power gating and athermal rings, the ATAC+ architecture can tolerate moderate waveguide losses.

D. ATAC+ Evaluation

In this section, we discuss the adaptive SWMR link in greater detail, specifically the time it spends in its three modes (idle, unicast and broadcast). Next, we evaluate the area of the ATAC+ architecture and finally, we perform a bandwidth sensitivity study to determine the flit width of the network.

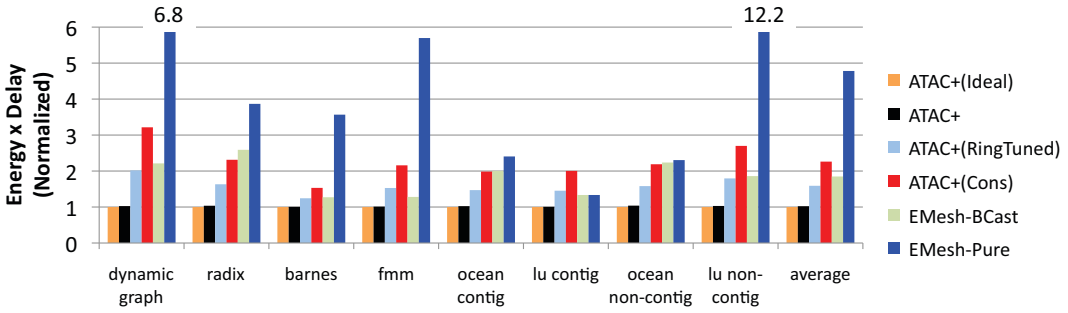

Fig. 8. Normalized Energy-Delay Product using eight benchmarks when using the ACKwise$_4$ protocol. Normalization is done with respect to ATAC+(Ideal).

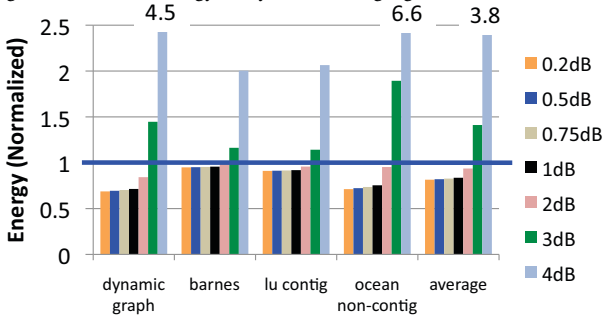

Fig. 9. Sensitivity to waveguide loss

Benchmark	Adaptive SWMR Link Utilization	# of Unicasts to Broadcasts
dynamic graph	12%	505
radix	25%	1086
barnes	9%	92
fmm	8%	95
ocean contig	20%	1812
lu contig	6%	30705
ocean non-contig	29%	13731
lu non-contig	19%	1324

TABLE V
ADAPTIVE SWMR LINK UTILIZATION; AND AVERAGE NUMBER OF UNICAST PACKETS SENT ON THE ADAPTIVE SWMR LINK BETWEEN SUCCESSIVE BROADCAST PACKETS

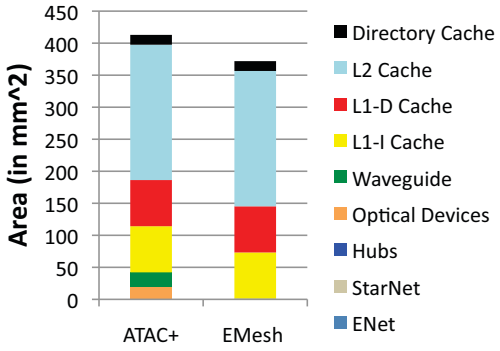

Fig. 10. Area of chip with ATAC+ and electrical mesh networks (includes caches and network)

Adaptive SWMR Link - Mode Transitions

First, we look at how frequently the laser transitions between its idle, unicast and broadcast modes. Table V shows the link utilization (the percentage of time in unicast or broadcast modes) and the average number of unicast packets between successive broadcast packets. Since the link is idle 70%-90% of the time, the capability to power-gate the laser yields high energy gains as seen in Section V-C.

Applications with fewer unicast messages between successive broadcasts (`dynamic graph`, `barnes`, `fmm`) yield higher performance gains with a broadcast enabled network (as seen in Section V-B). In the absence of broadcast support on the optical link, each broadcast would have to be converted into 64 unicast messages and serialized over the optical link. Such a scheme may make sense in the ATAC+(Cons) network (in Section V-C) where a broadcast-enabled SWMR is expensive since the laser must remain at maximum (broadcast) power even when not broadcasting. In ATAC+, the ability to down-throttle the laser during unicasts makes a broadcast-enabled SWMR implementation much more efficient.

Area

In Figure 10, we plot the area of the cache and network components in ATAC+ and compare that against the electrical mesh baseline. We observe that the caches dominate the total area ($\sim$ 90%). The area footprints of the electrical networks/components (ENet, StarNet and Hubs) are negligible. The waveguides and optical devices present in ATAC+ occupy $\sim$ 40mm^2 on the die.

Network Bandwidth Sensitivity

Figure 11 plots the runtime of the ATAC+ architecture as the flit width of the network is varied from 16 to 256 bits. The performance is poor at 16 bits and improves with flit width before starting to flatten out at 64 bits. On an average, the runtime improves by 50% from 16 bits to 64 bits and by 10% from 64 bits to 256 bits. A 256-bit flit width offers the best runtime. However, we choose a 64-bit flit width as extra waveguides and photonic devices consume active area on the die. At 256 bits, the waveguides and photonic devices occupy $\sim$ 160mm^2 of the die, which is unacceptable. Though higher link data-rates and SerDes can be used to decrease the number of photonic devices (and hence area) for wide flit-widths, the

SerDes power overhead and latency overcomes the marginal gain in performance.

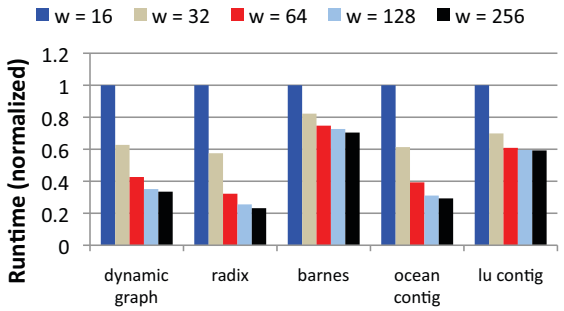

Fig. 11. Application runtime of the ATAC+ architecture when the flit width is varied from 16 to 256

E. Architectural Improvements Over Baseline ATAC Network

As mentioned in Section IV, three architectural improvements are made to baseline ATAC architecture: (1) the optical broadcast link in ATAC is replaced with the adaptive SWMR link; (2) the broadcast BNet of ATAC is replaced with a point-to-point electrical network called StarNet to reduce the energy of a long-distance unicast (3) the cluster-based unicast routing protocol in ATAC is replaced with a distance-based routing protocol that balances the network load between the adaptive SWMR link and the ENet mesh network, thereby improving the overall performance of the system.

The first improvement arises mostly from advances in optical device technology and has already been evaluated in the previous sections. In this section, the effect of the second and third improvements are evaluated.

Figure 12 demonstrates the effect of replacing the broadcast BNet with the point-to-point StarNet network. In the figure, the "ReceiveNet" portion in the first bar of each application represents the broadcast BNet while that in the second bar represents the point-to-point StarNet. The experiment is conducted with a cluster-based routing protocol in order to quantify just the reduction in energy obtained by replacing the BNet with the StarNet. The overall energy consumption is reduced by an average of 8%. More energy improvements are observed in benchmarks with high rates of unicast traffic such as `radix` and `ocean contig` than in benchmarks like `barnes` which have low rates of unicast traffic (refer to Figure 6).

Figure 13 demonstrates the effect of improving the unicast routing protocol. Recall that *Distance-i* refers to the distance-based unicast routing protocol with an r_{thres} of i hops. *Cluster* refers to the cluster-based routing protocol of ATAC where all inter-cluster unicast communication takes place over the ONet. Over all the evaluated benchmarks, the *Distance-15* routing protocol has the lowest energy-delay product. The *Distance-15* protocol shows an average of 10% reduction in the energy-delay product when compared to the *Cluster* protocol. Here

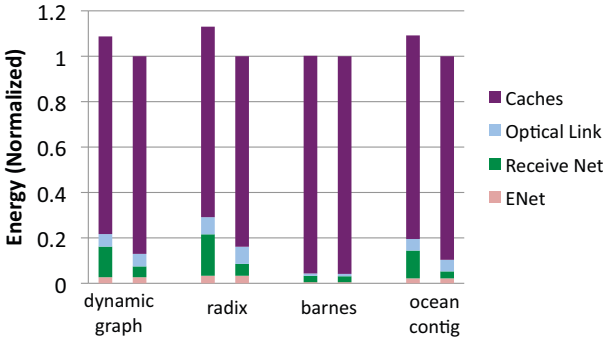

Fig. 12. Energy consumption analysis of broadcast BNet and point-to-point StarNet networks with a cluster-based unicast routing protocol. The first bar in each application corresponds to BNet and the second bar corresponds to StarNet.

again, the gains are higher for benchmarks with high rates of unicast traffic like `radix` and `ocean contig`.

In all the other evaluation sections, the *Distance-15* routing protocol and the StarNet network are assumed.

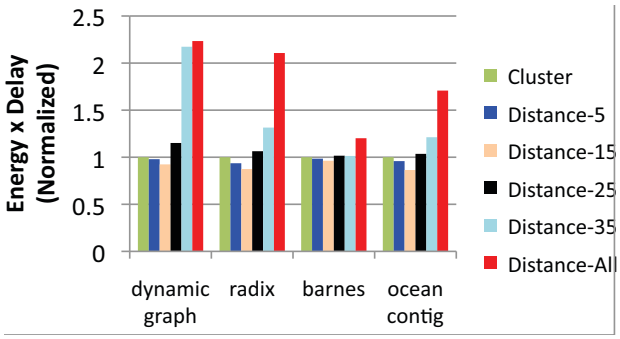

Fig. 13. Energy-delay product of cluster-based and distance-based unicast routing protocols (normalized with respect to the *Cluster* routing protocol).

F. Cache Coherence Protocols

Our previous work [6] evaluated the performance of three different cache coherence protocols on the ATAC and electrical mesh networks using two SPLASH-2 benchmarks at 1024 cores. In this section, we start with a larger set of benchmarks and evaluate the energy and area of the architectures as well. For the evaluation, we pick the two best performing protocols reported in [6], (a) the ACKwise$_k$ and (b) the Dir$_k$B protocols. *Dir$_k$B* is a limited directory-based protocol which broadcasts an invalidate request once the capacity of the sharer list is exceeded and collects acknowledgements from all the cores in the system. *ACKwise$_k$* on the other hand, tracks the number of sharers once the capacity of the sharer list is exceeded and needs acknowledgements from only the actual sharers of the data on a broadcasted invalidation. ACKwise$_k$, however, cannot support silent evictions while the Dir$_k$B protocol can support them. (Note that k denotes the number of hardware sharers in both the above protocols).

We evaluate the ACKwise$_k$ and Dir$_k$B protocols on the ATAC+ and EMesh-BCast networks. The EMesh-Pure net-

work is dropped in this evaluation since it does not handle broadcast traffic efficiently. Figure 14 plots the energy-delay product of the four configurations. The performance difference between the ACKwise$_k$ and Dir$_k$B protocols is proportional to the frequency of broadcast invalidations since the Dir$_k$B protocol expects 1024 acknowledgements in response. In benchmarks like `barnes`, `fmm` and `radix` with moderate-to-high broadcast traffic, the Dir$_k$B protocol suffers performance degradation. The performance degradation is felt to a greater extent on the EMesh-BCast network.

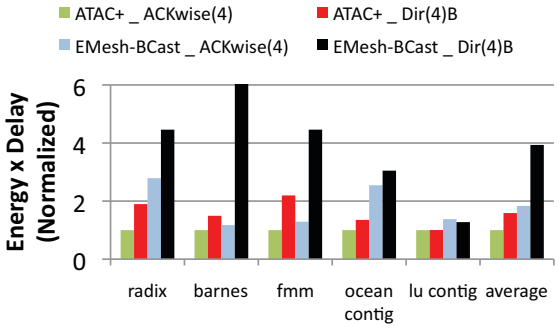

Fig. 14. Normalized energy-delay product when simulating the ATAC+ and EMesh-BCast network with the ACKwise$_4$ and Dir$_4$B protocols.

We now measure the dependence of the runtime of the ACKwise protocol on the ATAC+ network as the number of hardware sharers is varied. From Figure 15, it is clear that there is little runtime variation from 4 to 1024 sharers. Runtime is also found to not increase or decrease monotonically with the number of sharers. This can be explained by two factors that operate in opposite directions. An increase in the number of sharers causes a broadcast invalidate message to be replaced by multiple unicast invalidation messages. Multiple unicast messages cause higher contention in the ENet near the sending core. They however, lead to little or no contention at the receive hub near the entry point to the StarNet. A broadcast message, on the other hand, causes little contention on the ENet but causes higher contention at the receive hub near the entry point to the StarNet due to replication. Since these two factors work in opposite directions, there is no monotonic increase or decrease in runtime.

Both the energy consumption and the area of the ATAC+ architecture increases as the number of hardware sharers is increased from 4 to 1024. Figure 16 plots the variation of energy with the number of hardware sharers. There is a 2x increase in energy from 4 to 1024 sharers. The increase in energy is due to the directory cache, whose area and energy overheads are directly proportional to the number of hardware sharers. The total system area follows similar trends and increases by 2x from 4 to 1024 sharers.

Hence, the ACKwise protocol on the ATAC+ architecture provides the performance of a full-map directory protocol with much less area and energy overhead.

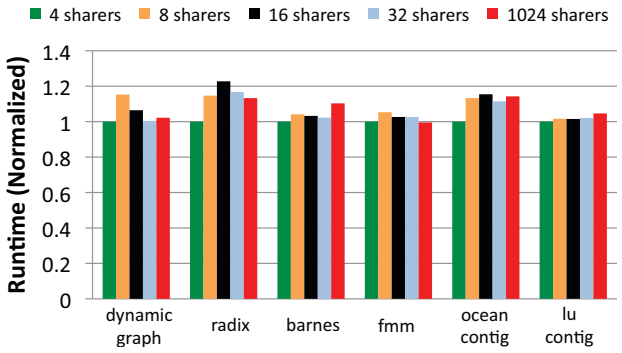

Fig. 15. Delay (Completion Time) of the ATAC+ architecture when the number of sharers in the ACKwise protocol are varied as (4,8,16,32,1024).

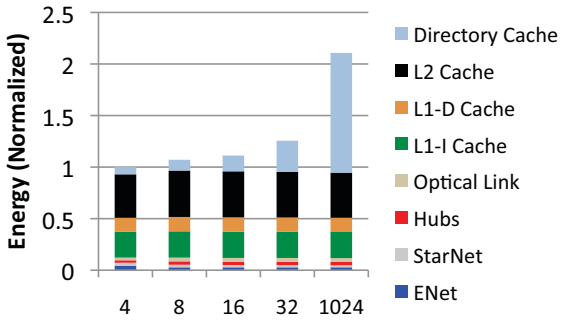

Fig. 16. Energy Breakdown of the ATAC+ architecture when the number of sharers in the ACKwise protocol are varied from 4 to 1024.

G. Core Power

Though we use the same core model to evaluate all architectures, the contribution of the non-data-dependent core power (leakage, ungated clocks) to total system energy changes, owing to the difference in application completion times. As highlighted in previous discussions, architectures that finish faster yield greater NDD energy savings, with the NDD energy of the core being no exception. Core data-dependent energies, on the other hand, are roughly identical between architectures since the same instructions are executed regardless of the network.

We evaluate the impact of core power by using a simple first order power model of the in-order single-issue core used in this paper. We assume a peak power of 20 mW for the core, which we obtain by scaling the energy/flop for the FPU in [31] to 11 nm and then dividing that number by the average fraction of power consumed by an FPU in a core. We consider two scenarios where core NDD power consumption is 10% and 40% of the peak power. The data-dependent power consumption is scaled using the measured IPC from architectural simulations. Hence, if the IPC is 0.25, the runtime data-dependent power is 25% of the peak data-dependent power.

From Figure 17, core NDD energy for EMesh-BCast is larger than that of ATAC+ as a result of the performance difference, shown clearly by `radix` and `ocean non-config`.

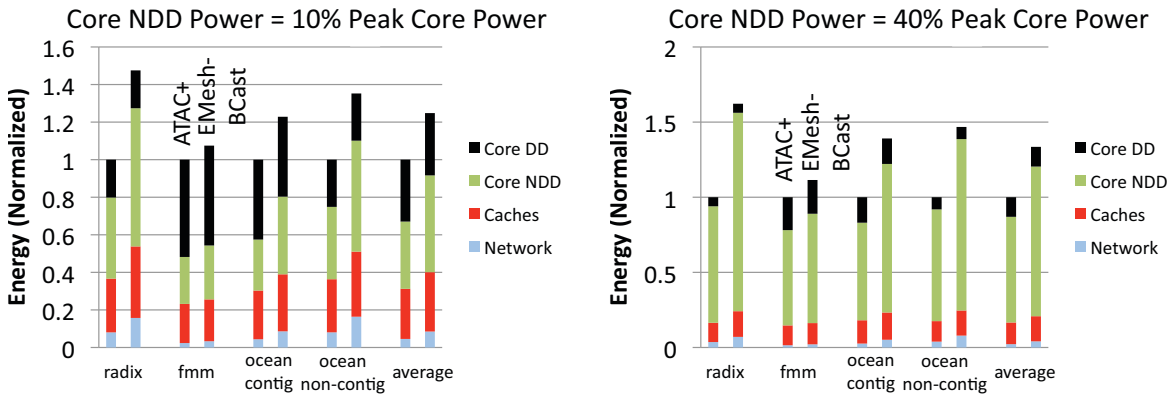

Fig. 17. Chip energy breakdown into core, cache and network components (for 10% and 40% core NDD power). The first bar for each benchmark corresponds to the ATAC+ network and the second bar corresponds to the EMesh-BCast network. NDD stands for non-data-dependent and DD stands for data-dependent.

`fmm`, shows no difference for core NDD energy as there is almost zero performance difference. As expected, when the percentage of NDD power increases, the contribution of the core to total system energy increases as well, since NDD power is burnt regardless of core utilization or IPC. In all cases, the cache and network are dwarfed by the core.

We end this section with an important insight. Even if an "uncore" component, such as a cache coherence protocol or an on-chip network, is not energy-efficient by itself, it may still achieve a significant system-level energy win. By allowing applications to complete faster, non-data-dependent energy of the core (the dominant energy consumer) can be effectively reduced.

VI. Related Work

As an emerging technology, photonics has been widely regarded as a potential solution to relieve the interconnect bottleneck of chip multiprocessors. However, standard simulation tools have not been used with the previous proposals for on-chip photonic interconnects, hence, the quality and thoroughness of architectural evaluations have varied significantly. Previous proposals [4], [32], [5], [7], [33], [8], [34], [2] either used synthetic benchmarks or trace-driven simulation to study the performance and power of photonic interconnection networks in isolation from the rest of the system. This paper presents the first at-scale evaluation of a 1000-core processor at the 11 nm technology node capturing the interplay between applications, multicore hardware and interconnection networks.

Chan et al [35] propose PhoenixSim, a simulator that incorporates detailed models of optical devices to perform the analysis of photonic interconnection networks both at the physical scale and at system scale. They also do not incorporate power models or accurate performance models incorporating feedback from the network for "non" network components and instead, rely on trace-driven and synthetic benchmarks for evaluation.

The ATAC+ architecture proposed in this paper is most similar to the ATAC [6] architecture since it preserves its ACKwise cache coherence protocol and broadcasting capabilities. It differs from ATAC in its adaptive SWMR link and its improved handling of unicast traffic. The unicast mode in the adaptive SWMR link employed by ATAC+ is similar to the reservation-assisted SWMR link proposed in Firefly [4] which also uses a global optical bus to communicate between clusters of cores.

Kirman et al [2] propose a hierarchical opto-electronic bus to support snooping cache coherence traffic. The scalability of their network is tied, however, to the scalability of snooping cache coherence protocols which are not expected to scale beyond tens of cores. Vantrease et al [32] propose a multiple writer single reader (MWSR) optical crossbar for communication between clusters of cores and use an optical mechanism for arbitration. They however do not tackle the problem posed by cache coherence and their global arbitration strategy limits the saturation throughput of the MWSR crossbar. ATAC+ borrows the ACKwise protocol to solve the coherence problem.

Shacham et al [34] propose a switched photonic network that uses electrical control packets to set-up the switches in the photonic network. This requires massive amounts for data transfer to amortize the overhead of the setup time and hence is not suitable for networks in cache coherent multicore processors where communication is short and frequent.

Vantrease et al. [5] and Yan Pan et al. [8] propose to mitigate the static laser and ring tuning power overheads by reducing the total number of waveguides in the system and making efficient use of them using similar novel flow control techniques. ATAC+, on the other hand, is targeted at a timeframe where athermal ring resonators and the technology to do rapid power gating of lasers are expected to be mature hence, does not have to deal with these overheads.

Joshi et al. [7] propose a photonic clos network, using point-to-point optical links to efficiently bridge long link distances characteristic of low-diameter networks such as the clos. We note that while the clos topology can also be used to perform broadcasts, port counts of each clos router will also continue to grow, hence requiring clustering.

High-diameter networks, such as Phastlane [33], maintain

a mesh-like grid of waveguides, using ring filters at cross-points to route optical packets. As noted by the authors of Phastlane, however, large losses due to waveguide crossings occur in mesh-like photonic networks. These losses will increase with the number of cores, making it very difficult for such architectures to scale to a thousand cores and beyond. These crossing losses are not present in ATAC+, as we employ a global ring bus. Our approach, however, does involve longer waveguides but with laser power gating technology, ATAC+ can tolerate moderate waveguide losses.

We however, note that the adaptive SWMR link proposed in ATAC+ can improve the energy efficiency of all the previously proposed photonic network architectures.

VII. Conclusion

This paper provides the first complete end-to-end analysis of an architecture using on-chip photonic interconnect. This analysis incorporates realistic performance and energy models for both electrical and optical devices and circuits into a full-fledged execution-driven functional simulator, thus enabling detailed analyses when running actual applications. Using this framework, we evaluate the ATAC+ architecture and observe that it provides 1.8x and 4.8x better energy-delay product than conventional electrical-only interconnects.

In addition, based on a detailed energy breakdown of all processor components, we conclude that athermal ring resonators and on-chip lasers that allow rapid power gating are key areas worthy of additional nanophotonic research, while ultra-low-loss optical devices are less valuable.

The full-system evaluation framework provides new insights not provided by earlier photonic network evaluations, the most important one being the impact of the network performance on the non-data-dependent (NDD) energy of "non-network" components like the core and caches. If the network is fast but not energy-efficient by itself, it may still achieve a significant system-level energy win by reducing the NDD energy of major system components.

References

[1] R. Kirchain and L. Kimerling, "A roadmap for nanophotonics," in *Nature Photonics, 1 (6): 303-305*, 2007.

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

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

[4] Y. Pan, P. Kumar, J. Kim, G. Memik, Y. Zhang, and A. N. Choudhary, "Firefly: illuminating future network-on-chip with nanophotonics," in *ISCA*, 2009, pp. 429–440.

[5] D. Vantrease, N. L. Binkert, R. Schreiber, and M. H. Lipasti, "Light speed arbitration and flow control for nanophotonic interconnects." in *MICRO'09*, 2009, pp. 304–315.

[6] G. Kurian, J. E. Miller, J. Psota, J. Eastep, J. Liu, J. Michel, L. C. Kimerling, and A. Agarwal, "Atac: a 1000-core cache-coherent processor with on-chip optical network," in *PACT '10*, 2010, pp. 477–488.

[7] A. Joshi *et al.*, "Proceedings of nocs '09," may 2009.

[8] Y. Pan, J. Kim, and G. Memik, "Flexishare: Channel sharing for an energy-efficient nanophotonic crossbar," in *HPCA*, 2010, pp. 1–12.

[9] J. Psota et al., "Improving performance and programmability with on-chip optical networks," in *ISCAS*, 2010.

[10] J. F. Liu and J. Michel, "High Performance Ge Devices for Electronic-Photonic Integrated Circuits," in *ECS Transactions, Vol 16, p 575-582*, 2008.

[11] M. Beals et al., "Process flow innovations for photonic device integration in CMOS," in *Proc. of the International Society for Optical Engineering (SPIE) 6898, 689804*, 2008.

[12] J. Michel et al., "Advances in Fully CMOS Integrated Photonic Circuits," in *Proc. of the International Society for Optical Engineering (SPIE) 6477, p64770P-1-11*, 2007.

[13] C. Schow, "Optical Interconnects in Next-Generation High-Performance Computers," OIDA 2008 Integration Forum, 2008.

[14] J. S. Orcutt *et al.*, "Nanophotonic integration in state-of-the-art CMOS foundries," *Optics Express*, vol. 19, pp. 2335–2346, 2011.

[15] "The International Technology Roadmap for Semiconductors (ITRS) Technology Working Groups," 2008.

[16] J. F. Liu, X. Sun, R. Camacho-Aguilera, L. C. Kimerling, and J. Michel, "Ge-on-si laser operating at room temperature," in *Optics Express*, vol. 35, 2010, pp. 679–681.

[17] S. Beamer *et al.*, "Re-architecting dram memory systems with monolithically integrated silicon photonics," June 2010.

[18] V. Raghunathan, W. N. Ye, J. Hu, T. Izuhara, J. Michel, and L. Kimerling, "Athermal operation of silicon waveguides: spectral, second order and footprint dependencies," in *Optics Express*, vol. 18, 2010, pp. 3487–3493.

[19] B. Guha, B. B. C. Kyotoku, and M. Lipson, "Cmos-compatible athermal silicon microring resonators," in *Optics Express*, vol. 18, 2010, pp. 3487–3493.

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

[21] Intel Corporation, "Intel's Teraflops Research Chip," http://techresearch.intel.com/articles/Tera-Scale/1449.htm.

[22] D. Wentzlaff et al., "On-chip interconnection architecture of the Tile Processor," *IEEE Micro*, vol. 27, no. 5, pp. 15–31, 2007.

[23] S.Woo et al., "The SPLASH-2 Programs: Characterization and Methodological Considerations," 1995.

[24] "DARPA UHPC Program BAA," https://www.fbo.gov/spg/ODA/DARPA/CMO/DARPA-BAA-10-37/listing.html, March 2010.

[25] J.Miller et al., "Graphite: A Distributed Parallel Simulator for Multicores," *HPCA*, 2009.

[26] C. Sun *et al.*, "DSENT - a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling," *NOCS '12*, May 2012.

[27] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, "Mcpat: an integrated power, area, and timing modeling framework for multicore and manycore architectures," in *Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture*, ser. MICRO 42. New York, NY, USA: ACM, 2009, pp. 469–480.

[28] M. Georgas *et al.*, "Addressing link-level design tradeoffs for integrated photonic interconnects," *Custom Integrated Circuits Conference*, September 2011.

[29] A. Khakifirooz, O. Nayfeh, and D. Antoniadis, "A simple semiempirical short-channel MOSFET current-voltage model continuous across all regions of operation and employing only physical parameters," *Electron Devices, IEEE Transactions on*, vol. 56, no. 8, pp. 1674 –1680, aug. 2009.

[30] L. Wei, F. Boeuf, T. Skotnicki, and H.-S. Wong, "Parasitic capacitances: Analytical models and impact on circuit-level performance," *Electron Devices, IEEE Transactions on*, vol. 58, no. 5, pp. 1361 –1370, may 2011.

[31] S. Galal and M. Horowitz, "Energy-efficient floating-point unit design," *IEEE Trans. Computers*, vol. 60, no. 7, pp. 913–922, 2011.

[32] D. Vantrease and others, "Corona: System Implications of Emerging Nanophotonic Technology," in *ISCA*, 2008.

[33] M. J. Cianchetti, J. C. Kerekes, and D. H. Albonesi, "Phastlane: a rapid transit optical routing network," *SIGARCH Comput. Archit. News*, vol. 37, pp. 441–450, June 2009.

[34] A. Shacham, B.G. Lee, A. Biberman, K. Bergman, and L.P. Carloni, "Photonic NoC for DMA Communications in Chip Multiprocessors," in *Hot Interconnects*, Aug 2007.

[35] J. C. et al., "Phoenixsim: A simulator for physical-layer analysis of chip-scale photonic interconnection networks," *DATE*, 2010.