Energy Implications of Photonic Networks With Speculative Transmission

Philip M. Watts, Simon W. Moore, and Andrew W. Moore

***Abstract*—Speculative transmission has been proposed to overcome the high latency of setting up end-to-end paths through photonic networks for computer systems. However, speculative transmission has implications for the energy efficiency of the network, in particular, control circuits are more complex and power hungry and failed speculative transmissions must be repeated. Moreover, in future chip multiprocessors (CMPs) with integrated photonic network end points, a large proportion of the additional energy will be dissipated on the CMP. This paper compares the energy characteristics of scheduled and speculative chip-to-chip networks for shared memory computer systems on the scale of a rack. For this comparison, we use a novel speculative control plane which reduces energy consumption by eliminating duplicate packets from the allocation process. In addition, we consider photonic power gating to reduce processor chip energy dissipation and the energy impact of the choice between semiconductor optical amplifier and ring resonator switching technologies. We model photonic network elements using values from the published literature as well as determine the power consumption of the allocator and network adapter circuits, implemented in a commercial low leakage 45 nm CMOS process. The power dissipated on the CMP using speculative networks is shown to be roughly double that of scheduled networks at saturation load and an order of magnitude higher at low loads.**

***Index Terms*—Assignment and routing algorithms; Networks; Optical interconnects.**

I. Introduction

The energy consumption of computing systems has become a major focus of attention in recent years in terms of both total energy requirements and the power dissipation on chip multiprocessors (CMPs). Reduction of the former helps to restrict the growing contribution of information technology to total energy usage as well as having important economic benefits for the running of large facilities such as data centers and high performance scientific computers. On the other hand, CMP power dissipation is seriously restricting the ability of designers to fully exploit the increased number of transistors provided by Moore's law [1]. Increasingly, network interfaces such as Ethernet and PCI are being integrated on CMPs in order to reduce latency and overall power [2]. Integrated photonic networks on CMPs have been widely proposed as a solution for reducing the energy consumption of interconnects for both on-chip and chip-to-chip networks [3,4], driven by rapid developments in silicon photonics [5], polymer waveguides in standard printed circuit boards (PCBs) [6] and 3D integration [7]. However, as more cores are integrated on CMP, the energy consumption of all circuits is under scrutiny.

It is well known that the exploitation of photonics in short range computer networks is challenging due to the lack of practical photonic memory. Gaining the maximum advantage from photonics requires the use of an edge buffered network in which end-to-end paths must be created, favoring large message sizes to reduce the overhead of setting up each communication. This is particularly limiting in the case of high performance shared memory systems in which the largest messages consist of cache lines (typically 8–128 B). Speculative transmission schemes have been proposed as a way of overcoming the control latency overhead, as the sole means of arbitration [8] or in parallel with a scheduler [9]. However, no previous work has compared the energy implications of speculative versus scheduled control.

Electrical power gating is a well established technique to optimize the energy consumption of complex integrated circuits such as CMPs. While hybrid integration of lasers on chips has been demonstrated [10], in order to keep large, power hungry optical devices off chips, many photonic network proposals use an off-chip photonic power supply (PPS), as shown in Fig. 1. A large proportion of the power supplied by the PPS is absorbed on the processor chip and contributes to the thermal load even when no communications are taking place. We consider the viability of photonic power gating schemes which do not damage the latency characteristics of the photonic network.

Finally, the choice of photonic switching technology has a major impact on the energy characteristics of the network and whether that energy is dissipated on the CMP. In this work we consider two switching technologies with contrasting energy characteristics, physical dimensions and states of development: semiconductor optical amplifiers (SOAs) [11] and ring resonators [12].

The concept investigated in this paper, shown in Fig. 1, uses photonic transceivers integrated with the processor as the basis of a chip-to-chip network which can potentially interconnect large numbers of shared memory CMPs. Multi-wavelength photonic switching technologies are now available with nanosecond switching times. However, it is unlikely that these technologies will scale beyond 64 ports in an integrated device with a reasonable power budget [11,13]. This network

Manuscript received December 2, 2011; revised April 6, 2012; accepted May 1, 2012; published May 22, 2012 (Doc. ID 159364).

Philip M. Watts (e-mail: pwatts@ee.ucl.ac.uk) is with the Electronic and Electrical Engineering Department, University College London, United Kingdom.

Simon W. Moore and Andrew W. Moore are with the Computer Laboratory, University of Cambridge, United Kingdom.

Digital Object Identifier 10.1364/JOCN.4.000503

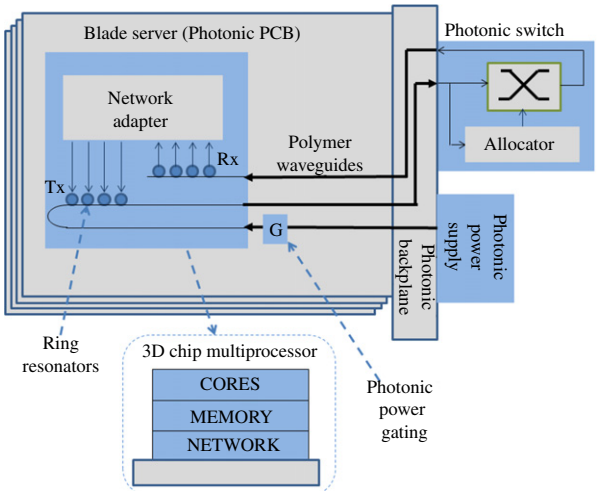

Fig. 1. (Color online) Rack-scale network of 3D integrated chip multiprocessors with distributed shared memory communications. For clarity the control path between network adapter and allocator is not shown.

size allows, for example, the interconnection of a rack of blade servers with two sockets per board and one network port per socket. Even today, such a system could contain over 1000 cores (16 cores per socket). Larger networks can be constructed by connecting these units in a hierarchical manner with optical–electrical–optical (OEO) units and electronic buffering. We investigate the energy characteristics of an edge buffered 32-port photonic network using a PPS and either scheduled or speculative transmission. We modify previously reported speculative allocator designs to minimize the performance and energy impact of duplicated speculative transmissions. Message sizes of 32 B are used, consistent with communication within a shared memory computer system. A system approach is taken, assessing the energy impact of the PPS, integrated transceivers, switches and network control electronics as well as exploring strategies for power gating individual circuits. We consider both total network power and the power dissipated on the CMP. The rest of the paper is organized as follows. Section II discusses speculative and scheduled networks proposed for optical switching in computer networks. Section III describes the methodology used for network simulation and power modeling. Results showing power and performance characteristics of scheduled and speculative networks are presented in Section IV. Finally the results are compared with current shared memory networks and future research is discussed in Section V.

II. Scheduled and Speculative Networks

Scheduling traffic across a crossbar is well understood. Algorithms such as iSLIP ensure fairness with 100% throughput under random traffic [14]. The hardware required for iSLIP (as we have implemented it in this work) is shown in Fig. 2. iSLIP is a separable, output port first, round robin allocator. However, it differs from other round robin schemes in the method for updating the priority state to avoid priorities becoming synchronized, which reduces throughput at high loads. A maximal matching is not produced, but the result converges to the best available after N iterations. In practice, however, the number of iterations is determined by the time available

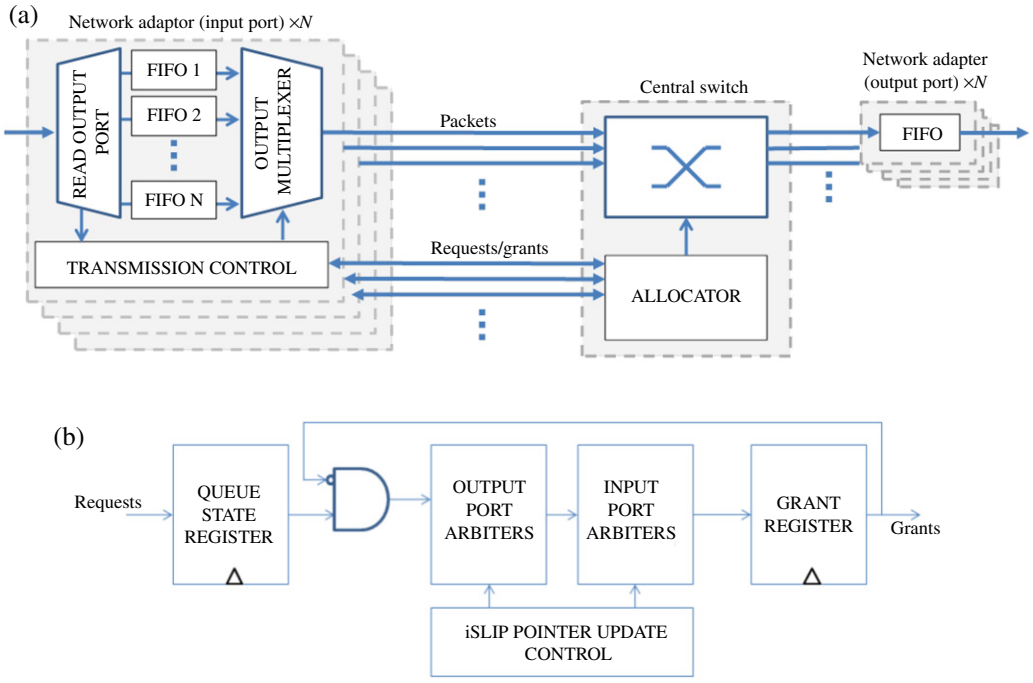

Fig. 2. (Color online) N-port iSLIP scheduled network model: (a) network schematic, (b) allocator detail.

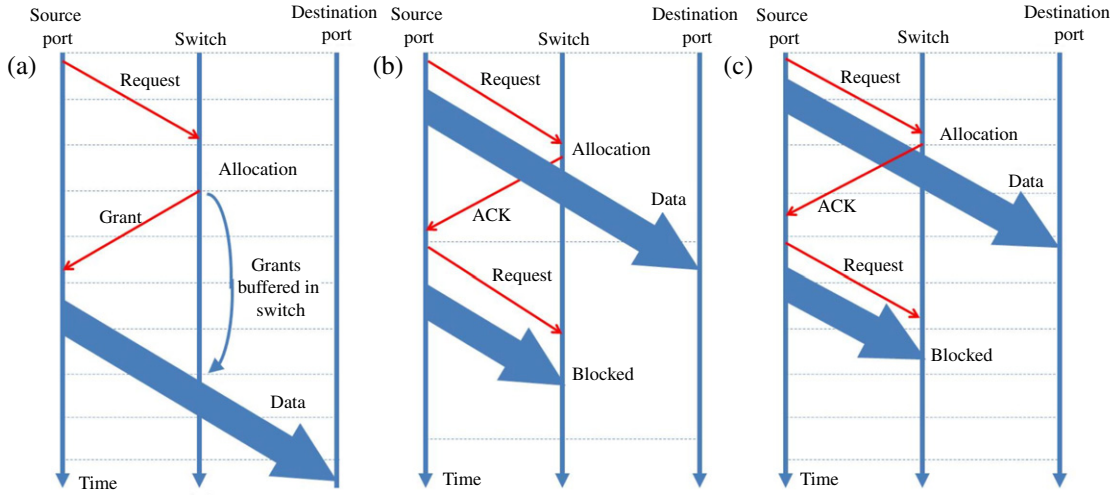

Fig. 3. (Color online) Timing diagrams for (a) scheduled transmission, (b) speculative transmission using the SPINet scheme and (c) pipelined speculative transmission. In each case, the dotted lines show the slot period boundaries defining the switching period. Request transmissions are timed to arrive at the switch at the start of the arbitration period.

for allocation. Each input port must maintain separate FIFOs, usually known as virtual output queues (VOQs) for each destination. To gain maximum power advantage in an optical network, signals must be maintained in the optical domain from the source to the final destination, requiring that the switch and scheduler are remote from the nodes. This in turn creates an additional control latency overhead as the request and grant signals must make a round trip to the scheduler in addition to the time required for allocation, as shown in Fig. 3(a). In this paper, we assume that the allocation is allowed to take only one slot period. Parallel and pipelined allocation schemes allowing a large number of iterations to be performed in one slot period independent of the logic delays have been developed [15]. However, the schemes are very complex and were estimated to require four ASICs (or 36 FPGAs) for a 64-port implementation using the CMOS technology available in 2006. As this paper is focused on low power networks, we use iSLIP as our scheduled allocation scheme.

In the OSMOSIS project [16], the control overhead of the request and grant process was reduced by adding a speculative allocator operated in parallel with the scheduled allocator. Thus if any network adapter does not have a grant in a given time slot, it sends out a request and packet anyway which is dropped if there is no path available on the switch in that slot period (Fig. 3(c)). Adding speculative transmission was shown to eliminate the latency overhead due to the request/grant cycle for loads up to 50% [9].

Pipelined speculation was implemented in OSMOSIS, allowing the acknowledgment (ACK) message to be received by the source in a subsequent slot period as shown in Fig. 3(c). However, buffer organization and guaranteed in-order delivery become more complex in this case. Alternatively, the SPINet network used only speculative transmission and was designed to have a low complexity control plane [8]. A multi-stage SOA network was used and each packet was sent speculatively with the path request for each stage of the switch encoded onto a separate wavelength. An ACK, received in the same slot period, was sent back to the source for packets winning the final round of arbitration as shown in Fig. 3(b). Therefore, the arbiter logic for each network stage was very simple and low power and only a single FIFO queue was required in the transmitter network adapter. However, the slot periods must be greater than the network diameter (or the round trip time from the network adapter to the switch), which is a serious drawback for shared memory computer networks (considered in this paper) which must efficiently handle packet lengths of a single cache line. However, it does offer low power and low latency for very large messages of the type observed in virtual machine migration or at the synchronization points in scientific algorithms running on message passing architectures. Another viable approach is to reserve the optical network for large messages with a backup electronic network to handle short messages [17,18], although this limits the potential for energy savings.

In this paper, we compare the energy implications of using scheduled (iSLIP) and pipelined speculative transmission for shared memory computer systems. We assess the scheduled and speculative schemes separately in order to understand the energy characteristics of each as the network load is varied.

III. Method

The performance and power estimates for the photonic chip-to-chip network are based on SystemVerilog models of scheduled and speculative networks. To obtain network latency results, behavioral simulation was carried out using Bernoulli distributed random packet arrivals and uniformly distributed random destinations with fixed packet sizes of 32 B (one cache line). A slot time of 6.8 ns was used, consisting of a data serialization latency of 3.2 ns (32 B striped across eight wavelengths of 10 Gb/s), a preamble for clock recovery of 1.6 ns, a switching time of 1 ns and an additional 1 ns to cover slot timing uncertainty between different CMPs, serialization, deserialization and optoelectrical conversion delays. Therefore we transmitted 256 data bits in each 6.8 ns slot, giving an

effective bandwidth per port of 37.6 Gb/s. Detailed timing diagrams for both the scheduled and speculative models are given in Fig. 4. We consider a network on the scale of a single rack of servers with a maximum distance from each port to the switch of 2 m. Hence, the time of flight between port and switch over polymer waveguides with effective refractive index of 1.5 is 10 ns.

Key components (allocator and network adapter) were then synthesized using four-element VOQs and clock gating with a commercially available low leakage 45 nm CMOS standard cell library. Topological synthesis using Synopsys Design Compiler was used for greater accuracy, in which layout is performed in parallel with synthesis, avoiding the use of wire-load models. Simulation of the synthesized designs was used to generate power estimates for each module under various network loads using Synopsys PrimeTime. The network model was also used to generate activity data, which, along with power parameters for photonic components extracted from the published literature, were used to generate accurate power figures for the full network. Care must be taken in comparing the allocator and network adapter power with the optical component power as there are substantial variations between CMOS processes and design styles. However, we believe the use of a 45 nm standard cell library gives representative power estimates. Dynamic power will not be substantially lower in future CMOS processes, while the high cost of full custom design is only justified in the most high volume, high performance applications, e.g., the data path of a microprocessor. The main aspects of the network control plane designs and network power model are described in the following sections.

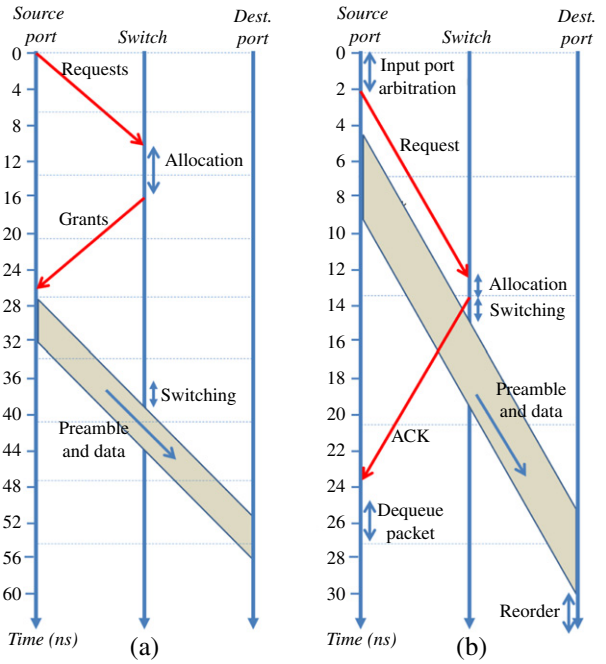

Fig. 4. (Color online) Detailed timing diagrams for successful packet transmissions using (a) scheduled and (b) speculative networks. The time of flight between the source port and the switch is 10 ns. In the speculative case, input port arbitration and packet dequeue take one clock cycle, which is 2.26 ns.

A. Control Plane Models

1) Scheduled Control Plane: The scheduled model is based on iSLIP allocation [14] and is shown schematically in Fig. 2; its timing diagram is shown in Fig. 4(a). The source port VOQs are four-element implemented using flip-flops. In each slot period, the source transmitter sends any new requests to the allocator, which maintains a record of the state of all VOQs [19]. Each allocation iteration (including both output and input port arbitration) is carried out in a single clock cycle. The arbiter circuits use the fast priority encoder described in [20] with a look-ahead of four to reduce the length of the carry chain. The number of iterations performed in each slot period is adaptive. One clock period per slot must be reserved for pointer update and maintaining request queues. In this work, we do not consider the details of control message transmission (requests, grants and ACKs) except that parallel signaling at low speed is used to avoid additional control latency and power. This could be implemented as a WDM optical link, as in SPINet [8], or with electronic links.

2) Speculative Control Plane: The model for the speculative technique is shown in Fig. 5 and its timing diagram is shown in Fig. 4(b). The design is based on [9], but we have made modifications in order to reduce power by ensuring that switch paths are not turned on for duplicate packets. As for iSLIP, separate VOQs are maintained for each output port. In the network adapter transmitter, the incoming packet is routed to the appropriate queue, has a sequence number added and is placed in a four-element random-out queue (ROQ). The ROQ differs from an FIFO in that entries can be dequeued in any order. A round robin arbiter selects a candidate for transmission from the valid entries in the ROQ which have a sequence number of no more than the lowest stored sequence number plus three (for the four-element queue). Restricting the extent to which packets can be transmitted out-of-order in this way ensures that there will always be an available location in the receiver's reorder queues for every received packet. Another round robin arbiter selects from the ROQs which have packets waiting, and the winning packet from this arbitration is transmitted in the next slot. This process of selecting the packet for transmission is labeled input port arbitration in Fig. 4(b). ACKs returning from the allocator (four slots after transmission in the rack-scale network considered here) cause the relevant ROQ entry to be invalidated. The retransmission policy for failed speculative packets is based on selective retry (SR) [9] as this has higher throughput and a reduced number of retransmissions compared with the alternatives such as go-back-*N*. SR does require reorder queues in the receiver. However, the results presented in Subsection IV.C demonstrate that the reorder queues consume a relatively small proportion of the adapter power. The adapter does not attempt to retransmit a packet until four slot periods (1 RTT to the switch) have elapsed, in order to avoid unnecessary transmissions and the associated power consumption. The speculative allocator only performs output port arbitration and hence is much simpler than the iSLIP allocator. Our implementation also tracks the last sequence number granted

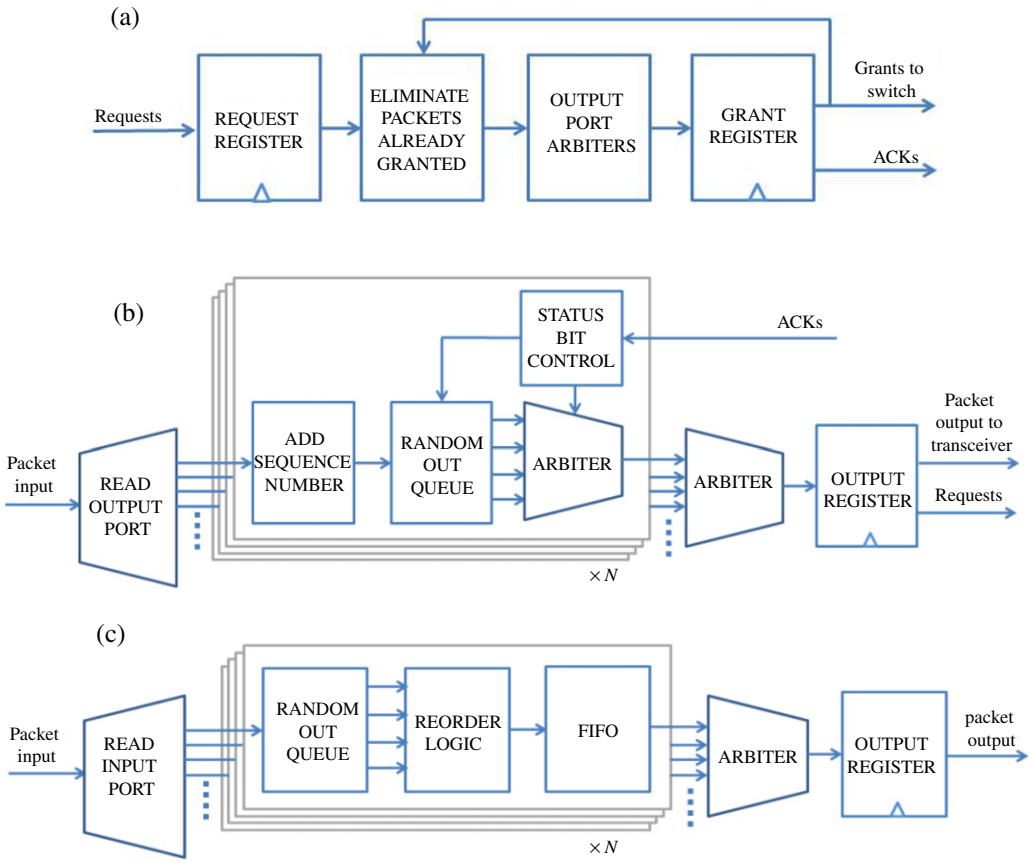

Fig. 5. (Color online) Schematic diagrams of speculative network control circuits: (a) allocator, (b) network adapter transmitter side and (c) network adapter receiver side.

for each combination of input port and output port and ensures that the switch is not turned on for duplicate packets, thus reducing switch drive energy. In the network adapter receiver, incoming packets are placed in a four-element ROQ for packets from the same input port. The reorder circuit places the packets into the FIFO in order of sequence number, waiting for out-of-order packets arriving in later slots if necessary.

B. Network Power Model

Figure 1 shows the photonic path we assume in estimating the optical power, and a summary of parameters and the literature which they are taken from is given in Table I. We have taken the best reported values for these parameters. While noting that these technologies are unproven on the scale required here, these values will probably be conservative in the long term as silicon photonic integration matures.

1) Photonic Power Supply: We assume the use of an off-chip PPS feeding processor chips with integrated photonic transceivers. The processor chips and a central switch chip are interconnected with polymer waveguides over boards and backplanes. We assume that the chip-to-chip optical path over the rack-scale network consists of 2 cm of silicon waveguide and 4 m of polymer waveguide. The upper bound on the efficiency of the PPS is determined by the efficiency of a laser, which can be up to 60%. Further losses due to temperature control, amplification, combining/splitting and distribution could be included depending on the type of PPS employed. In this work we assume a bank of lasers per port with 50% efficiency combined with a WDM multiplexer with 3 dB insertion loss and neglect temperature control power. This gives an efficiency of 25% for the PPS, compared with 80% for a typical electrical power supply. However, the actual efficiencies of practical PPSs supplying multiple ports with temperature control could be substantially below our assumed value, leading to increased values of the total network power. The figures for on-chip dissipation are not affected by the PPS efficiency. The optical power requirement per wavelength at the input to the processor chip is calculated from the receiver sensitivity and loss parameters listed in Table I as well as the switch characteristics discussed below.

2) Photonic Transceivers: On each processor chip, we assume the use of silicon waveguides and ring resonators used as modulators (at the transmitter) and wavelength selective elements (at the receiver) [26]. However, it has to be noted that ring resonator technology requires considerable development for use in the harsh temperature environment of a processor chip and it is difficult to simultaneously achieve high speed, low drive voltage and high extinction ratio from CMOS compatible devices [5]. As current research is aimed at athermalizing ring resonator devices [27], we assume

TABLE I
ASSUMED PARAMETERS FOR POWER MODELING

Parameter	Value
Bit rate	10 Gb/s
Number of wavelengths per waveguide	8
Receiver sensitivity at 10 Gb/s	−18 dBm [21]
Chip interface loss	0.5 dB [22]
Modulator loss	3 dB (ideal)
Polymer waveguide	4 m at 0.03 dB/cm [23]
Silicon waveguide (Tx)	1 cm at 0.5 dB/cm [13]
Silicon waveguide (Rx)	1 cm at 0.5 dB/cm [13]
Silicon waveguide crossing	0.05 dB [13]
Ring resonator switch through loss	0.3 dB [24]
Ring resonator switch drop loss	1.6 dB [24]
Payload bits per packet	256
Overhead bits per packet	128
Ring resonator drive energy	80 fJ/bit [21]
Transceiver electronics energy	1.1 pJ/bit[25]

that no temperature control of the ring resonators will be required. If thermal wavelength tuning is required, even with the best reported tuning efficiencies [28], the tuning power is likely to be substantially higher than the ring resonator drive power and not amenable to power gating on short time scales. High confinement silicon waveguides also enable high efficiency SiGe detectors with capacitances of around 1 fF [29], adding negligible power consumption if directly integrated with CMOS electronics [4]. A recent hybrid integrated receiver based on ring resonators consumed 260 fJ/bit, higher than the modulator drive circuit, but only increasing the total transceiver energy by 22% when taking into account the energy of the transceiver electronics. Power figures for the transceiver electronics (SERDES and clock recovery, etc.) were estimated from [25].

3) Photonic Switches: We consider two switching technologies with nanosecond switching times which have contrasting power characteristics: SOA and silicon ring resonators. Active devices such as SOAs provide broadband gain blocks forming the basis of multi-wavelength switches. The gain of SOA switches can reduce the CMP power dissipation by reducing input power requirements. In this paper, we conservatively assume that the SOA switch can overcome its own internal losses, but does not provide overall gain. Although each SOA stage is of the order of 1 mm long, a 16 port rearrangeably non-blocking Clos SOA switch integrated onto a 6×6 mm substrate has been demonstrated [30], consuming 300 mW per path. It is unlikely that this power consumption can be significantly reduced due to cross talk considerations [31]. However, unused ports can be switched off when not in use. Due to the losses between each stage, 64-port devices carrying 10×10 Gb/s per port appear to be the limit for this type of switch [11], although much larger networks of 2×2 SOA switching stages are possible [32]. Silicon ring resonator switches, while in a less mature state of development, potentially offer greatly reduced power consumption and area. Multiple wavelength switching is possible by matching the free spectral range of the resonator with the signal wavelength spacing [12]. As with the modulator and wavelength selective ring resonators, we assume no temperature control. Although only small scale demonstrations of ring resonator switching have been shown [33], it is unlikely that this technology will scale beyond 64 ports with around 10 wavelengths per port [13]. A third order ring resonator switch using an integrated PIN diode, sized for eight wavelengths centered on 1500 nm, consumes 150 μW in the on state [24] and a 32-port crossbar built from these devices is estimated to have losses of 9.8 dB. Overcoming these losses substantially increases the input photonic power requirements and hence the processor chip dissipation. As this paper is concerned with low power networks, we assume crossbar and Clos topologies for both switch technologies in order to minimize the number of switching elements per path and hence the drive power and optical losses. The power consumption figures could be considerably higher for topologies with an increased number of switching stages. The network power results are sensitive to these assumptions, particularly in the case of ring resonator switched networks where the losses and temperature characteristics of large scale integrated systems are speculative at the current time. As in the discussion of modulators and filters above, thermal tuning would considerably increase the ring resonator switch power.

4) Power Gating: In deriving power figures for the entire network, it is important to consider which circuits can be power gated. We assume that the transceiver electronics, modulator drivers and photonic switches can be put into a zero energy state when not in use. In practice, front end transmitter and receiver circuits may require continuous bias in order to avoid high startup latency, but this is likely to be a small proportion of the total transceiver power. However, we assume that the allocator and network adapters must be powered continuously.

Consideration of power gating the PPS is speculative as it has not been investigated in the literature. In order to rapidly power gate the PPS on the order of a network slot (6.8 ns), the lasers would have to be biased just above threshold and therefore would not be in a zero power state. In addition, the PPS is likely to be an expensive device to be shared among many network ports. A more realistic scenario, considered in this paper, may be to provide constant low power from a central PPS and amplify and power gate using an SOA very close to the processor chip (as shown in Fig. 1). In this case, the optical power dissipated on chip is proportional to the load and the constant PPS output power can be kept at a low level. As with the SOA switches, we take the on-state power of the SOA power gater to be 100 mW with a gain of 15 dB and a switching time of 1 ns. Both the scheduled and speculative adapter designs can generate control signals for power gating on a slot by slot basis.

The network total power figures are derived by summing all the above mentioned energy sources. Similarly, the processor chip power dissipation is found by summing the power of the network adapter, transceiver electronics plus the proportion of optical power which is absorbed on the chip.

IV. RESULTS

A. Network Simulation Results

Figure 6 compares the latencies of 32-port networks with scheduled and speculative control. 100% offered load is 1.47×10^8 packets/s/port. The average latencies at low loads are 32.4 ns and 62.2 ns in the speculative and scheduled cases

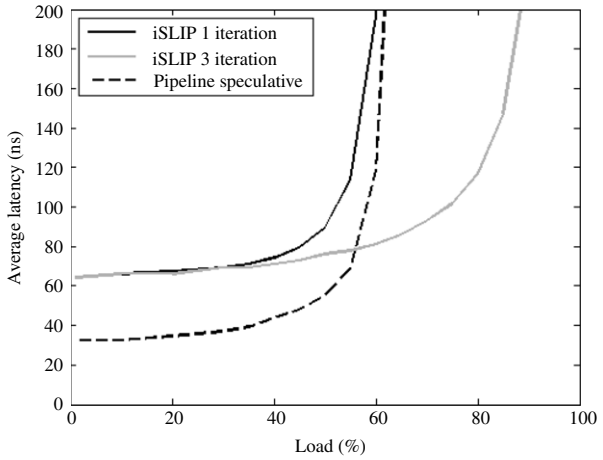

Fig. 6. Comparison of the latency characteristics of scheduled and speculative networks with random traffic.

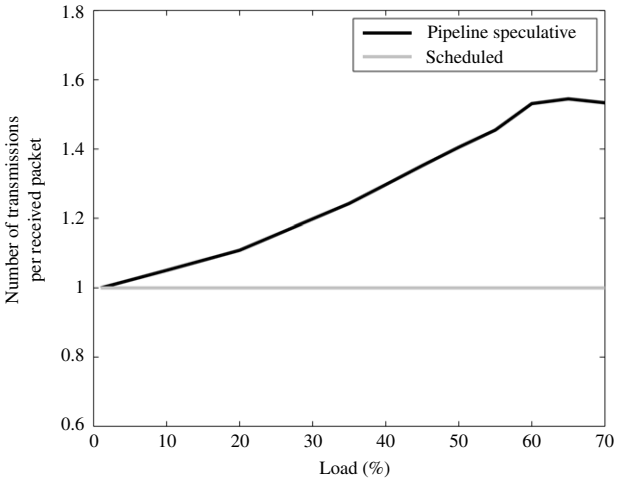

Fig. 7. The number of transmissions per received packet for scheduled and speculative networks.

respectively, compared with the single trip time of flight latency of 10 ns between each port and the switch. Speculative transmission is shown to have a similar maximum throughput to single iteration iSLIP under random traffic (60% of full load). This is expected as the speculative scheme is a single iteration separable allocator with input port arbitration performed at the network adapter. The maximum throughput of iSLIP is increased when further iterations are performed but, at least for random traffic, does not improve significantly beyond three iterations. Due to the energy focus of this work, only random traffic was considered, which is well known to be benign.

Speculative transmission increases the use of network resources which could otherwise be power gated. Figure 7 shows the number of transmissions made per received packet as the load is varied. For the scheduled network, this remains 1 for all loads. In the speculative case, the number of transmissions per received packet increases from 1 up to 1.6 at the saturation load. Although our speculative allocator design does not turn the switch on for duplicate packets, the proportion of time for which the transceivers and PPS can be power gated decreases with increasing load.

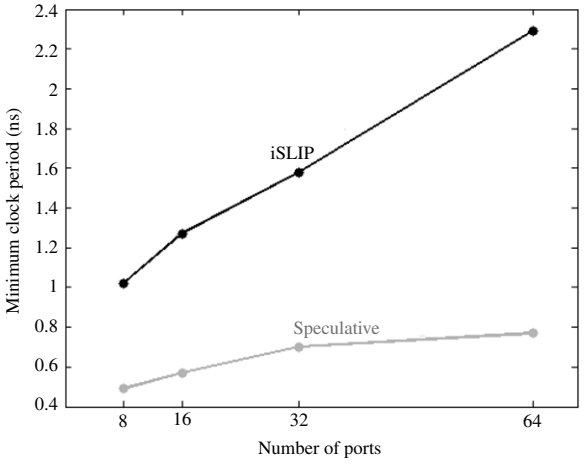

Fig. 8. Minimum clock periods for the iSLIP and speculative allocator circuits using a commercial, low leakage 45 nm CMOS standard cell library.

B. Allocator and Network Adapter Timing Results

The minimum clock period of iSLIP allocation circuits determines the number of iterations which can be performed in a single slot period. The minimum clock period of the speculative allocator has a minor effect on latency by reducing the time that the request has to be sent ahead of the data. Figure 8 shows the minimum clock periods for the speculative and iSLIP allocators for networks with between 8 and 64 ports. For iSLIP, one clock period per slot is required for updating pointers. Hence, 5, 4, 3 and 2 iterations can be performed in a single slot period of 6.8 ns for 8-, 16-, 32- and 64-port networks respectively. It should be noted that for iSLIP we chose to perform both input and output port arbitration for each iteration in one clock cycle. A modest improvement in timing and power could have been achieved by performing input and output port arbitration in separate clock cycles. By comparison, the speculative allocator has a minimum clock period of 0.77 ns even for a 64-port network, as only a single iteration of output port allocation is required. In the speculative case, the input port allocation effectively occurs in the network adapter (i.e., the adapter decides which packet to send to the output ports).

C. Allocator and Network Adapter Power Results

Figures 9 and 10 show the power of the allocator and network adapter circuits respectively for a 32-port network as the load is varied. The power was not measured beyond the saturation load of the network. The maximum clock frequency achievable was used for each allocator circuit, but in order to reduce power, a clock frequency of 441 MHz (three clock periods per 6.8 ns slot period) was used for the network adapters. It can be observed for the allocator circuits that, even with clock gating of all registers, the power at low loads (static power) is a significant proportion of the power at full load, i.e., the circuits

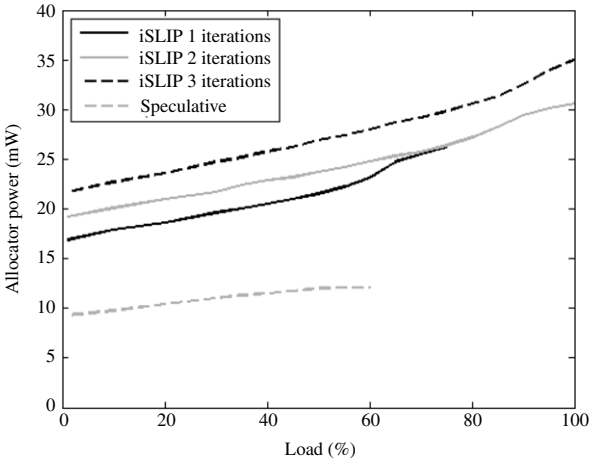

Fig. 9. Power consumption of 32-port allocator circuits under random traffic of varying load.

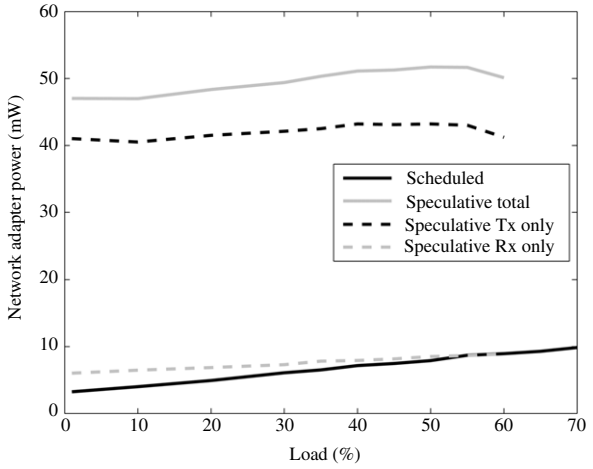

Fig. 10. Power consumption of a single network adapter (for a 32-port network) under random traffic of varying load.

display poor energy proportionality. The iSLIP allocator has greater power consumption than the speculative allocator for all loads. However, the single central allocator power is small compared with the combined power consumption of the 32 network adapters for both the scheduled and speculative cases. The speculative network adapter has power consumption that is greater than four times that of the scheduled adapter for all loads and is relatively invariant with load. This is mainly due to the additional complexity of managing random-out memories in the transmitter. In addition, the speculative adapter requires reordering circuits at the receiver, although these consume a small proportion of the overall speculative adapter power. By comparison, the scheduled receiver consists of a single small FIFO (not included in these figures).

D. Network Power Results

Figure 11 shows the contributions to total network power for a 32-port network versus load. The overall power consumption

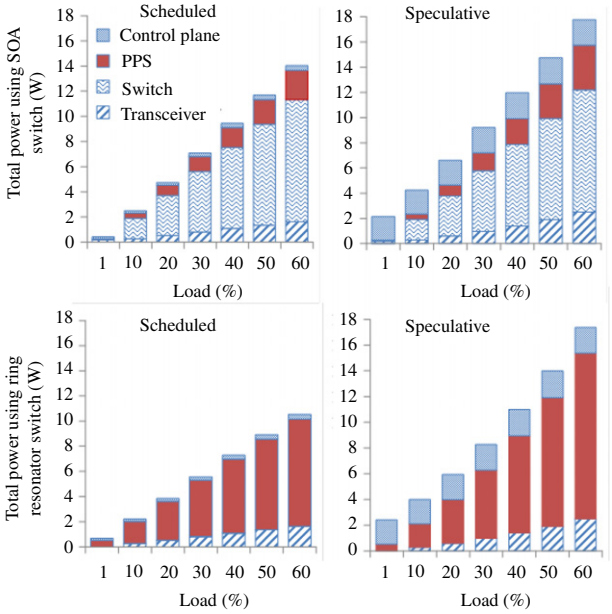

Fig. 11. (Color online) Contributions to total power consumption of a 32-port network. PPS is the photonic power supply, which includes the power gating SOA. The control plane power includes the allocator and adapters.

of the speculative network is greater than the scheduled network for all loads and both switch types due to repeated speculative transmissions and the increased network adapter power. Using both electronic and photonic power gating, all scenarios achieve good energy proportionality with load. However, the sources of power consumption vary with switch type. The gain of the SOA switch means that the PPS is operated at relatively low power. On the other hand, the ring resonator drive power is negligible compared with the overall system power. Using a single SOA for power gating (and noting that we have made optimistic assumptions about the PPS efficiency), the eight wavelength PPS power consumption for the 32-port network is 23 mW in the SOA case against 347 mW in the ring resonator case.

Figure 12 shows the power dissipated on each CMP due to the integrated photonic network port. Repeated transmissions and the higher (and ungated) adapter power cause the power dissipation to roughly double in the speculative network at 60% load compared with the scheduled network. At low loads, due to electronic and photonic power gating, the adapter power dominates, with the scheduled adapter consuming only 3.2 mW and the speculative adapter consuming 47.0 mW. The absorption of the PPS is only significant when using ring resonator switching. Without photonic power gating, the CMP dissipation due to the PPS would be 55.3 mW, constant with load, for ring resonator switching, but only 3.7 mW for SOA switching.

V. Conclusions and Further Work

The results presented above compare the energy characteristics of scheduled and speculative networks for shared

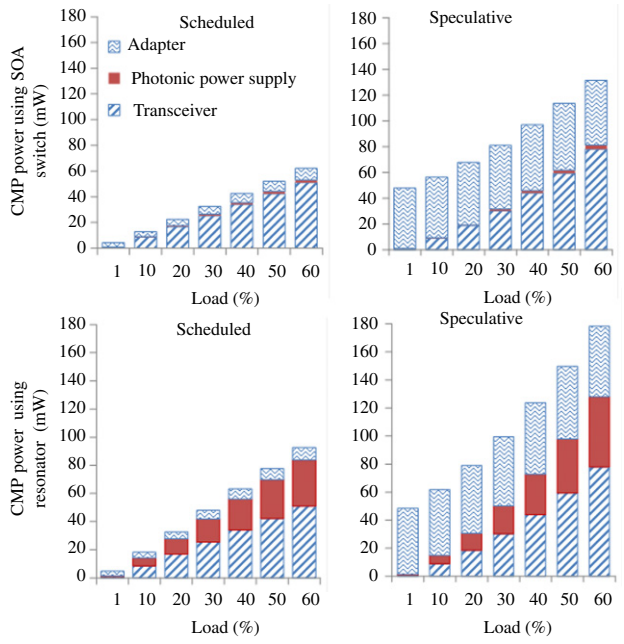

Fig. 12. (Color online) Contributions to the network power dissipated on the chip multiprocessor (CMP) for 32-port scheduled and speculative networks.

memory communications on the scale of a single rack. We find that the advantage of reduced latency at low loads offered by speculative transmission comes at the cost of higher total power consumption and processor chip dissipation. This is due to an increased number of transmissions per packet and a more complex network adapter. In particular, the processor chip dissipation of the speculative network using ring resonator switches at 60% load is 178 mW compared with 93 mW for the scheduled network. In the shared memory scenario that this paper investigates, this network would interconnect caches on multiple CMPs to allow cache line accesses to remote locations and maintain memory coherence. For example, the 16-core shared memory CMP described in [2] has 402 Gb/s coherence links for this purpose in a four-CMP system (not including direct links to DRAM memory). As the photonic networks considered in this paper have an effective bandwidth per port of 37 Gb/s, the processor chip dissipation for a CMP with similar coherence link bandwidth would be 1.9 W for the speculative network, compared with 0.9 W for the scheduled network. This is significant given that each core consumes 2.1 W. By comparison, the (unswitched) electronic coherence links in [2] consume 8.8 W. Future systems may require significantly higher coherence link bandwidths.

We find that the power consumption of the allocator circuits is insignificant compared with other power sources. Therefore, more complex, higher performance allocators can be tolerated on energy grounds. However, the OSMOSIS 64-port parallel and pipelined allocator [15], estimated to require four ASICs for implementation in 2006, would certainly dominate the total network power. Allocation algorithms are required which balance performance against power consumption and, in particular, avoid complexity at the network adapter. At 60% load, the speculative adapter accounts for 28–38% of the CMP dissipation (depending on switch type). Given the additional CMP power dissipation of speculative transmission, full system simulations are being carried out to determine whether the latency advantage translates into significant overall performance gains for real shared memory workloads. This study is also considering adapter FIFO depth requirements for the bursty traffic observed in real applications. For larger networks on the scale of a machine room or data center, the latency benefits of speculation are greater in absolute terms (due to the higher time of flight) and more complex network adapters can be considered as the adapter power is not dissipated in the critical area of the processor chip.

It has to be noted that the power results in this paper are highly dependent on the assumptions; in particular, higher optical losses, lower PPS efficiency and the need for temperature control of silicon photonic devices could substantially increase the energy figures. We also made aggressive assumptions on the slot time in order to maximize the bandwidth, minimize the latency and efficiently handle short messages. Such a system requires burst mode receivers with fast clock recovery. We included estimates of the SERDES, clock buffers, samplers and phase shifters as well as a 16-bit preamble per wavelength for clock recovery. However, the power cost of coding and protocol for burst mode operation was not included. This is an area which is only starting to receive attention [34,35] and where there is significant potential for energy savings.

Acknowledgments

The authors thank Robert Mullins and Yury Audzevich of the University of Cambridge for useful discussions. This work was supported by the United Kingdom Engineering and Physical Sciences Research Council (EPSRC) grant EP/I004157/2 and the Royal Commission for the Exhibition of 1851.

References

[1] H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, and D. Burger, "Dark silicon and the end of multicore scaling," in *Proc. of the 38th Annu. Int. Symp. on Computer Architecture*, 2011.

[2] J. L. Shin, H. Dawei, B. Petrick, H. Changku, K. W. Tam, A. Smith, H. Pham, H. Li, T. Johnson, F. Schumacher, A. S. Leon, and A. Strong, "A 40 nm 16-core 128-thread SPARC SoC processor," *IEEE J. Solid-State Circuits*, vol. 46, pp. 131–144, 2011.

[3] U. Vlasov, "Silicon photonics for next generation computing systems," in *European Conf. on Optical Communications (ECOC)*, 2008.

[4] D. A. B. Miller, "Device requirements for optical interconnects to silicon chips," *Proc. IEEE*, vol. 97, pp. 1166–1185, 2009.

[5] G. T. Reed, G. Mashanovich, F. Y. Gardes, and D. J. Thomson, "Silicon optical modulators," *Nat. Photonics*, vol. 4, no. 8, pp. 518–526, 2010.

[6] I. H. White and R. V. Penty, "Optical interconnects for backplane and chip-to-chip photonics," in *2nd ACM/IEEE Int. Symp. on Networks-on-Chip (NOCS '08)*, 2008.

[7] J. U. Knickerbocker, P. S. Andry, B. Dang, R. R. Horton, M. J. Interrante, C. S. Patel, R. J. Polastre, K. Sakuma, R. Sirdeshmukh, E. J. Sprogis, S. M. Sri-Jayantha, A. M. Stephens, A. W. Topol, C. K. Tsang, B. C. Webb, and S. L. Wright, "Three-dimensional

silicon integration," *IBM J. Res. Dev.*, vol. 52, no. 6, pp. 553–569, 2008.

[8] A. Shacham and K. Bergman, "Building ultralow-latency interconnection networks using photonic integration," *IEEE Micro*, vol. 27, no. 4, pp. 6–20, 2007.

[9] I. Iliadis and C. Minkenberg, "Performance of a speculative transmission scheme for scheduling-latency reduction," *IEEE/ACM Trans. Netw.*, vol. 16, no. 1, pp. 182–195, 2008.

[10] L. Schares, J. A. Kash, F. E. Doany, C. L. Schow, C. Schuster, D. M. Kuchta, P. K. Pepeljugoski, J. M. Trewhella, C. W. Baks, R. A. John, L. Shan, Y. H. Kwark, R. A. Budd, P. Chiniwalla, F. R. Libsch, J. Rosner, C. K. Tsang, C. S. Patel, J. D. Schaub, R. Dangel, F. Horst, B. J. Offrein, D. Kucharski, D. Guckenberger, S. Hedge, H. Nyikal, C.-K. Lin, A. Tandon, G. R. Trott, M. Nystrom, D. P. Bour, M. R. T. Tan, and D. W. Dolfi, "Terabus: Terabit/second-class card-level optical interconnect technologies," *IEEE J. Sel. Top. Quantum Electron.*, vol. 12, no. 5, pp. 1032–1044, 2006.

[11] I. White, A. E. Tin, K. Williams, H. B. Wang, A. Wonfor, and R. Penty, "Scalable optical switches for computing applications," *J. Opt. Netw.*, vol. 8, no. 2, pp. 215–224, 2009.

[12] B. G. Lee, A. Biberman, D. Po, M. Lipson, and K. Bergman, "All-optical comb switch for multiwavelength message routing in silicon photonic networks," *IEEE Photon. Technol. Lett.*, vol. 20, no. 10, pp. 767–769, 2008.

[13] A. Biberman, G. Hendry, J. Chan, H. Wang, K. Bergman, K. Preston, N. Sherwood-Droz, J. S. Levy, and M. Lipson, "CMOS-compatible scalable photonic switch architecture using 3D-integrated deposited silicon materials for high-performance data center networks," in *Proc. Optical Fiber Communications Conf.*, Mar. 2011.

[14] N. McKeown, "The iSLIP scheduling algorithm for input-queued switches," *IEEE/ACM Trans. Netw.*, vol. 7, no. 2, pp. 188–201, 1999.

[15] C. Minkenberg, I. Iliadis, and F. Abel, "Low-latency pipelined crossbar arbitration," in *IEEE Global Telecommunications Conf. (GLOBECOM)*, 2004, vol. 2, pp. 1174–1179.

[16] R. Luijten, C. Minkenberg, R. Hemenway, M. Sauer, and R. Grzybowski, "Viable opto-electronic HPC interconnect fabrics," in *Proc. of the ACM/IEEE Supercomputing Conf.*, 2005.

[17] K. J. Barker, A. Benner, R. Hoare, A. Hoisie, A. K. Jones, D. K. Kerbyson, D. Li, R. Melhem, R. Rajamony, E. Schenfeld, S. Shao, C. Stunkel, and P. Walker, "On the feasibility of optical circuit switching for high performance computing systems," in *Proc. of the ACM/IEEE Supercomputing Conf. 2005*.

[18] A. Shacham, K. Bergman, and L. P. Carloni, "Photonic networks-on-chip for future generations of chip multiprocessors," *IEEE Trans. Comput.*, vol. 57, no. 9, pp. 1246–1260, 2008.

[19] C. Minkenberg, "Performance of i-SLIP scheduling with large round-trip latency," in *Workshop on High Performance Switching and Routing (HPSR)*, 2003.

[20] P. Gupta and N. McKeown, "Designing and implementing a fast crossbar scheduler," *IEEE Micro*, vol. 19, no. 1, pp. 20–28, 1999.

[21] X. Zheng, F. Liu, J. Lexau, D. Patil, G. Li, Y. Luo, H. Thacker, I. Shubin, J. Yao, K. Raj, R. Ho, J. E. Cunningham, and A. V. Krishnamoorthy, "Ultra-low power arrayed CMOS silicon photonic transceivers for an 80 Gb/s WDM optical link," in *Proc. Optical Fiber Communications (OFC) Conf.*, Mar. 2011.

[22] V. R. Almeida, R. R. Panepucci, and M. Lipson, "Nanotaper for compact mode conversion," *Opt. Lett.*, vol. 28, no. 15, pp. 1302–1304, 2003.

[23] Y. Kuwana, S. Takenobu, K. Takayama, S. Yokotsuka, and S. Kodama, "Low loss and highly reliable polymer optical waveguides with perfluorinated dopant-free core," in *Optical Fiber Communication Conf. (OFC)*, Mar. 2006.

[24] A. W. Poon, X. S. Luo, F. Xu, and H. Chen, "Cascaded microresonator-based matrix switch for silicon on-chip optical interconnection," *Proc. IEEE*, vol. 97, no. 7, pp. 1216–1238, 2009.

[25] J. Poulton, R. Palmer, A. M. Fuller, T. Greer, J. Eyles, W. J. Dally, and M. Horowitz, "A 14 mW 6.25-Gb/s transceiver in 90 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 42, no. 12, pp. 2745–2757, 2007.

[26] B. G. Lee, B. A. Small, Q. F. Xu, M. Lipson, and K. Bergman, "Characterization of a 4 × 4 Gb/s parallel electronic bus to WDM optical link silicon photonic translator," *IEEE Photon. Technol. Lett.*, vol. 19, no. 5, pp. 456–458, 2007.

[27] W. N. Ye, R. Sun, J. Michel, L. Eldada, D. Pant, and L. C. Kimerling, "Thermo-optical compensation in high-index-contrast waveguides," in *5th IEEE Int. Conf. on Group IV Photonics*, 2008.

[28] J. E. Cunningham, I. Shubin, X. Zheng, G. Li, H. Thacker, Y. Luo, J. Yao, K. Raj, B. Guenin, T. Pinguet, and A. V. Krishnamoorthy, "Compact, thermally-tuned resonant ring muxes in CMOS with integrated backside pyramidal etch pit," in *Proc. Optical Fiber Communication Conf. (OFC)*, 2011.

[29] M. R. Reshotko, B. A. Block, B. Jin, and P. Chang, "Waveguide coupled Ge-on-oxide photodetectors for integrated optical links," in *5th IEEE Int. Conf. on Group IV Photonics*, 2008, pp. 182–184.

[30] A. Wonfor, H. Wang, R. Penty, and I. White, "Large port count high-speed optical switch fabric for use within datacenters [Invited]," *J. Opt. Commun. Netw.*, vol. 3, no. 8, pp. A32–A39, Aug. 2011.

[31] R. S. Tucker, "Green optical communications—Part II: Energy limitations in networks," *IEEE J. Sel. Top. Quantum Electron.*, vol. 17, no. 2, pp. 245–260, 2011.

[32] O. Liboiron-Ladouceur, B. A. Small, and K. Bergman, "Physical layer scalability of WDM optical packet interconnection networks," *J. Lightwave Technol.*, vol. 24, no. 1, pp. 262–270, 2006.

[33] N. Sherwood-Droz, H. Wang, L. Chen, B. G. Lee, A. Biberman, K. Bergman, and M. Lipson, "Optical 4 × 4 hitless silicon router for optical networks-on-chip (NoC)," *Opt. Express*, vol. 16, no. 20, pp. 15915–15922, 2008.

[34] T. W. Y. Chen and R. Katz, "Energy efficient Ethernet encodings," in *33rd IEEE Conf. on Local Computer Networks (LCN)*, Oct. 2008, pp. 122–129.

[35] Y. Audzevich, P. M. Watts, S. Kilmurray, and A. W. Moore, "Efficient photonic coding: A considered revision," in *GreenNets 2011 (SIGCOM workshop)*, Aug. 2011.

Philip M. Watts (M'04) obtained his B.Sc. in applied physics from the University of Nottingham, UK, in 1991. He obtained his M.Sc. in technologies for broadband communications in 2003 and a Ph.D. investigating electronic dispersion compensation for high speed optical communication in 2008, both from University College London (UCL), UK.

From 1991 to 2000, he worked at BAE Systems Advanced Technology Centre (Chelmsford, UK), and from 2000 to 2002, he was a senior optical hardware engineer with Nortel Networks (Harlow, UK, and Ottawa, Canada). While a Ph.D. student, he was a researcher at Intel Research and a consultant to Azea Networks and Huawei Technologies. From 2008 to 2010, he was a Research Fellow at the Computer Laboratory, University of Cambridge. He is currently an EPSRC Research Fellow and Lecturer in the Department of Electronic and Electrical Engineering at UCL where his research interests cover

optical interconnects and electronic signal processing, control and coding circuits for optical communications.

Dr Watts was awarded the IEEE Photonics Society Postgraduate Student Fellowship in 2006, the Royal Commission for the Exhibition of 1851 Brunel Research Fellowship in 2008 and the EPSRC Career Acceleration Fellowship in 2010.

Simon W. Moore (M'98–SM'08) received his M.Eng. degree from the University of York, UK, in 1991 and his Ph.D. degree in multithreaded processor design from the University of Cambridge, UK, in 1995, wherein his thesis was published by Kluwer in 1996.

As a student, he worked at Smiths Industries on aerospace systems (hardware and software) and at the DEC Western Research Centre, Palo Alto, CA, on processor design. In 1998, he was appointed as a University Lecturer at the Computer Laboratory, University of Cambridge, where he is currently a Reader in Computer Architecture and heads the Computer Architecture Group. His research interests span massively parallel computer architectures and associated algorithm issues.

Andrew W. Moore received Bachelors and Masters degrees from Monash University, Melbourne, Australia, in 1992 and 1994. He completed his Ph.D. with the Cambridge University Computer Laboratory in 2001.

Previous appointments include Lecturer at Monash University, Intel Research Fellow and Roberts Fellow in the Department of Computer Science at Queen Mary, University of London, UK. He is currently a Senior Lecturer at the University of Cambridge, Computer Laboratory. His interests lie in addressing the scalability, usability and reliability of the Internet, coding in photonic networks and computer architectures which capitalize on the features of photonics in their design. He is a Chartered Engineer and Member the Association of Computing Machinery (ACM) and the Institute of Engineering and Technology (IET).