A Low Latency Optical Switch for High Performance Computing with Minimized Processor Energy Load [Invited]

Shiyun Liu, Qixiang Cheng, Muhammad Ridwan Madarbux, Adrian Wonfor, Richard V. Penty, Ian H. White and Philip M. Watts

***Abstract*— Power density and cooling issues are limiting the performance of high performance chip multiprocessors (CMP) and off-chip communications currently consume over 20% of power for memory, coherence, PCI and Ethernet links. Photonic transceivers integrated with CMPs are being developed to overcome these issues, potentially allowing low hop count switched connections between chips or data center servers. However, latency in setting up optical connections is critically important in all computing applications and having transceivers integrated on the processor chip also pushes other network functions and their associated power consumption onto the chip. In this paper, we propose a low latency optical switch architecture which minimizes power consumed on the processor chip for two scenarios: multiple socket shared memory coherence networks and optical top-of-rack switches for data centers. The switch architecture reduces power consumed on the CMP using a control plane with a simplified send and forget server interface and the use of a hybrid Mach-Zehnder Interferometer (MZI) and semiconductor optical amplifier (SOA) integrated optical switch with electronic buffering. Results show that the proposed architecture offers a 42 % reduction in head latency at low loads compared with a conventional scheduled optical switch as well as offering increased performance for streaming and incast traffic patterns. Power dissipated on the server chip is shown to be reduced by over 60% compared with a scheduled optical switch architecture with ring resonator switching.**

***Index Terms*—Assignment and routing algorithms; Networks; Optical Interconnects**

I. Introduction

Research efforts in optical networking for data centers are aimed at both lower latency and lower energy consumption. Although total energy consumption is a critical issue for the largest data centers, networking equipment only accounts for 5% of this with the majority consumed at the server or chip multiprocessor (CMP) level [1]. In addition, CMP power density and thermal management issues are seriously limiting processor performance [2]. High performance server chips require >1Tb/s of off-chip bandwidth including Ethernet, PCI, main memory and coherence links which are consuming >20% of total power [3]. In parallel, there is a major research effort aimed at packaging optical communications components within the CMP, for example using silicon photonics [4-9] to minimize latency and energy consumption and eliminate communications bottlenecks. However, on-chip optics also necessitates integrating the PHY and MAC layers and their associated energy consumption onto chip and also requires an optical power supply. Previous work has shown that a large proportion of the network energy consumed on the processor chip is due to buffering, transmission control and absorbed optical power in on-chip and chip-to-chip networks [10, 11]. Therefore, network architectures are required which provide low latency but reduce the energy dissipated on the processor chip.

A large proportion of the bandwidth of a high performance server is used for point-to-point main memory links (200 Gb/s in [3]) and low latency and power optical replacement options have been studied [12, 13]. However, in this paper we focus on applications which can benefit from optical switching in particular chip-to-chip memory coherence, Ethernet and PCI networks. Chip-to-chip coherence networks are used in high performance servers which share the memory space across multiple chips to

Manuscript received July 1 2014.

S. Liu was with the Electronic Engineering Department, University College London, London WC1E 7JE, UK. She is now with Barclays Investment Bank, London, UK.

M.R Mardarbux and P.M. Watts are with the Electronic Engineering Department, University College London, London WC1E 7JE, UK (philip.watts@ucl.ac.uk).

Q. Cheng, A. Wonfor, R.V. Penty and I.H. White are with the Centre for Advanced Photonics and Electronics, University of Cambridge, Cambridge CB3 0FA, UK.

improve parallel application performance (Fig. 1) by exchanging control (typically 8B) and data (16-256B) messages to ensure that memory is consistent across all caches [14]. Coherence network latency has a critical effect on multiprocessor performance as processors must stall until coherence transactions on the network have completed. This low latency requires high bandwidth (460 Gb/s in [2]) and hence coherence networks consume a significant proportion of processor chip power. Switched photonics potentially reduces the latency and power consumption of coherence networks spanning multiple chips by providing a single network connecting every core or cluster of cores (Fig. 1b) rather than the separate on-chip and chip-to-chip networks used in current servers (Fig. 1a).

Ethernet and PCI server interfaces can both benefit from switched photonic connections. Although data centers can have $>10^5$ servers, scaling optical switches to these port counts is challenging as is the associated global allocation problem. Semiconductor optical amplifier (SOA) switching has been shown to allow large switching fabrics at the physical layer [15]. Optical switches on a single integrated circuit are limited by losses, but have been shown to be viable with 64 ports or more using SOAs [16], Mach-Zehnder interferometers (MZI) [5] and silicon ring resonators [6]. Switches of this radix can replace the electronic top-of-rack (ToR) switch in leaf and spine data center architectures as shown in Fig. 2 [17] providing sufficient links to connect a rack of servers as well as uplinks to core electronic routers. This architecture offers two hop connections to any other server within the data center, keeps the allocation problem manageable and avoids the issues of multiple stage switching [18]. Current servers feature a small number of 10 Gb/s Ethernet interfaces and therefore consume a much smaller proportion of server off-chip communications power than memory and coherence links. However, in this case, optical switching has the potential to reduce latency and total network power consumption and provide higher bandwidth without electronic pin or front panel limits. Higher bandwidth can also mitigate the performance issues of data center workloads such as those caused by incast traffic.

Optical networks for shared memory [19-22] and data center [23-26] application have been previously proposed. In contrast, we propose a low latency optical switching architecture supporting at least 64 ports which specifically minimizes power consumed and dissipated on the CMP in these applications. Low latency is provided using speculative transmission, in which messages are sent before a switch path is established, combined with fast electronic allocators and electronic buffers at the switch for packets which fail allocation. Energy consumption on the server chip is minimized by (1) providing a simple server-side send and forget network interface with minimal buffering and control logic (2) use of hybrid MZ/SOA switching which reduces the optical power absorbed at the server transmitter and (3) avoiding any significant receiver side buffering by ensuring in-order delivery. Initial results were presented in [10]. This paper provides a more detailed description and power models for the proposed switch architecture including practical allocation circuits for the MZI/SOA switch and characterization of the latency in problem workloads such as incast traffic.

The rest of the paper is organized as follows: Section II describes the recirculation network control plane and the hybrid MZI/SOA switch architecture. Section III presents latency results for common data center workloads taking into account the operating clock frequency of the control plane. Section IV presents the power model for the network along with results showing the reduction in power dissipated on the processor chip. Section V discusses the results and their impact on future computing systems focusing on the multi-socket shared memory network and optical top-of-rack applications described above. Finally, section VI concludes.

II. Network Architecture

A. Control Plane

Baseline Virtual Channel Switch

Figure 3a shows a high performance input queued virtual channel (VC) scheduled switch connecting multiple compute cores. In an N port switch, the source port contains N-1 first in first out (FIFO) queues, known as virtual channels (VC), one for each destination. The source ports queue new messages in the appropriate FIFO and send requests for a switch path to the allocator. The allocator (also known as scheduler or arbiter) attempts to find the best switch configuration to serve all requests and sends grants back to the ports which have been successful. Unsuccessful requests will be served in future allocation cycles. The use of VCs and the iSLIP allocation algorithm has been shown to achieve 100% throughput under random traffic in a fair manner [27]. iSLIP is a separable (arbitrates separately for output and input ports of all outstanding requests) round robin allocator which updates priority states in a way which avoids individual arbiters becoming synchronized. In this work, a broadband switch and wavelength striped transmission is used in order to achieve high bandwidth per port and hence low serialization latency (compared with alternative approaches using wavelength selective elements, e.g. [18, 23]). As shown in Fig 4a, control signaling between the port and allocator (requests and grants) increases arbitration latency. For this reason, optical-electrical-optical (OEO) conversion to allow queuing at the every switch port has been proposed [18]. However, to get full energy and latency advantage of optical switching, data should remain in the optical domain from the source port through to the destination port. In this work, we use the VC switch shown in Fig 3a with latency characteristics shown in Fig 4a as the baseline. Our proposed switch deals with the issue of control latency while using OEOs for only the packets which fail allocation.

Proposed Send and Forget Interface with Buffered Switch

In the proposed scheme, shown in Fig. 3b hereafter described as the buffered switch, speculative transmission is used to minimize control signaling latency. Speculative transmission of messages, in which data is sent without waiting for a grant, has been previously proposed either operating independently [4] or in parallel with a scheduled allocator [28]. However, our previous work showed that high performance speculative schemes require complex logic and buffering at the transmitter (and also at the receiver if in-order delivery is required) which increases power consumption on the server chip [10]. Our speculative implementation which simplifies the server side of the network operates as follows. Each transmitter has a simple FIFO queue which considerably reduces the power and area of buffering resources on the processor chip. To meet the aim of providing a low energy send and forget interface at the server, the switch must not drop packets. When the channel is free, the transmitter controller first checks that there is a free slot in the switch buffers (and hence there is no chance that the packet may be dropped). This single bit full control signal, the only connection required back to the server from the switch, is asserted when there is one free slot in the switch buffers to allow for the fact that a packet may already be in transit. If there is buffer space available, the controller sends a switch path request to the allocator for the packet at the front of the FIFO and then several clock cycles later, speculatively sends the packet in a wavelength striped format. The number of clock cycles between request and data transmission is determined by the allocation time and the switch reconfiguration time and is discussed further in section III. Contention resolution is handled entirely at the switch using electronic buffers. Although fiber delay line buffers have been studied for all-optical packet switching, single chip integrated WDM transceivers [7] and fast dense electronic memory provide improved area and timing characteristics for low latency networks [23]. If allocation is successful, the switch is reconfigured and the packet is delivered with low latency (Fig. 4b). If allocation is unsuccessful, the packet is sent to the switch buffers after conversion back to the electronic domain (Fig. 4c). Packets are queued by source port (mapping input 1 to buffer 1, input 2 to buffer 2 etc) and transmitted through the switch in a later allocation cycle. In contrast to [23, 26], this direct mapping of buffers to source ports considerably simplifies the allocation problem and switch architecture and is essential for the send and forget protocol to ensure that no packet will be dropped. Also in contrast to [23, 26] these buffers store wavelength striped messages rather than a single serial message per wavelength increasing optical transmitter and receiver count but reducing serialization latency and increasing throughput. Strict in-order delivery is adopted by always giving priority in allocation to packets in the switch buffers over new packets from the servers as our previous work showed that there is a significant power cost in reordering packets at the receiver [10].

B. Optical Switch Architecture

Several optical broad-band (multiple wavelength) integrated switching technologies with ns reconfiguration times have been demonstrated based on semiconductor optical amplifiers (SOA) [16], Mach-Zehnder interferometers (MZI) [**5**] and ring resonators [6, 9]. In this work we use the hybrid MZI and SOA dilated switch architecture [29] which has been shown to scale to 128 ports by using an 8x8 port switch in a recirculating loop experiment [30]. The MZIs in this device have been designed to operate over the wavelength range 1540 – 1560 nm providing a large bandwidth for wavelength striped transmission. Operation using 10 wavelengths of 10 Gb/s each has been demonstrated for an 8-port switch [31]. In this architecture, the SOAs overcome the main limitation of pure MZI devices to increase crosstalk suppression to over 50dB and also provide gain to reduce the overall switch insertion loss which significantly reduces input optical link power (and hence the power absorbed on the processor chip as described in section IV A).

Figure 5a shows the hybrid dilated switch architecture which is a type of butterfly network [32] based on 4-port switch building blocks each using 4 MZI and 8 SOAs. These blocks are interconnected with passive shuffle networks which comprise passive waveguides, bends and waveguide crossings. An NxN switch, such as that required for the baseline VC switch, requires an array of $(N\log_2(N))/2$ 4-port switching blocks arranged as $\log_2(N)$ stages (or columns) and N/2 rows. The hybrid switch design uses a dilated scheme to achieve a lower crosstalk ratio. The purpose of dilation is to ensure that each individual MZI switching element only carries one signal at a time and hence the maximum usage of the total switch fabric is 50%. This is the reason that the input/output stages only use 2 of the 4 ports. This architecture significantly reduces component count and waveguide crossing losses compared with a crossbar architecture. For further details of the hybrid MZI/SOA switches refer to [29-31].

Although the switch buffer scheme requires 2N ports (N ports for processor links and N ports for the buffers), as buffer ports never need to send to other buffer ports, a full 2N x 2N switch is not required and the internal architecture can be simplified. We consider two cases as shown in Fig. 5b and 5c. Case 1 uses two NxN switches, N 1x2 input switches and (N/2) 4x2 output switches. The input and output switches are constructed from the 4-port switch blocks described above and shown in Figure 5a. Overall, the case 1 switch uses Nlog2(N) + (3N/2) 4-port blocks. Failed speculative packets are routed to the buffers by the input switch. These packets are routed across the second NxN switch to the output switches when the output port is free.

Further simplified structures are possible such as the case 2 architecture shown in Fig. 5c. As with case 1, an input switch determines whether the packet is routed to the main NxN switch or the buffers. However, in this case failed speculative packets stored in the buffers are routed

back through the 2x2 input switch and the main NxN switch when both the input and output ports are free. The number of 4-port blocks is reduced to Nlog2(N)/2 + N, but the limitation is that packets from the transmitter and recirculation buffer on the same input port but destined for different output ports cannot pass through the switch simultaneously. In the following sections, we evaluate the performance and power characteristics of the two buffered switch designs relative to the VC switch.

Another key advantage of the proposed architecture is that variations in power and optical signal to noise ratio which could affect physical layer performance are expected to be very small and, if necessary, can be calibrated out by adjusting the gain of individual SOAs. In the ToR scenario, all transmissions, whether server to server or server to/from core router, take one pass through the switch. Within the switch itself, all transmissions pass through the same number of 4-port switching blocks. In addition, the differences in attenuation due to fiber transmission distance between server-to-server traffic and server-to-router traffic will be minimal on the data center scales (<1 km). However, future work is required to define fabrication variations for the integrated photonic components and calibration procedures for their mitigation.

III. Latency Results

We have modeled the control planes of the baseline scheduled VC network and the proposed buffered switch network in SystemVerilog including buffering, transmission control and switch allocation. Delays are inserted into the control plane model to account for the time of flight of optical data and control signals between network ports, allocator and switch. The SystemVerilog model allows us both to simulate the network control plane to obtain latency under various traffic patterns but, in addition, key circuits can be synthesized using an application specific integrated circuit (ASIC) design flow to obtain the minimum clock period, area and power consumption possible in a real CMOS circuit. In our previous work, latency values for a 64-port ToR switch were reported in clock cycles [10]. However, as discussed in section III B below, the allocation circuit depends on both the switch and control plane architecture, each having a different achievable clock period. The allocator clock period in turn determines the clock periods of other circuits and hence has a major impact on overall latency. Section III A describes allocation circuits required for the different control plane and switch cases and their timing characteristics in a 45 nm CMOS process. Then Section III B reports latency without congestion while sections III C – D show the relative performance under load with random, streaming and incast traffic.

A. Allocation Circuits

Figure 6a shows a separable allocator such as iSLIP suitable for a VC crossbar switch consisting of a two stage process of output port arbitration followed by input port arbitration. In previous work [10], it was shown that the clock period of this circuit increases rapidly with number of ports reaching 2.3 ns for a 64-port switch in a 45nm CMOS process. The input and output arbitration stages can be pipelined to reduce the clock period [27] and although this does not reduce allocation latency, it reduces the latency of other control plane functions which use the same clock as the allocator. Allocators for speculative transmission using crossbar switches (Fig. 6b) only require output port arbitration (as the input port decision has been made at the transmitter) and hence are more scalable having a clock period of 0.75 ns for 64-ports [10]. However, as discussed above, crossbar switches have a high component count leading to more scalable switch architectures such as the hybrid dilated structure described in Section II B. The hybrid dilated switch requires more complex allocation because there are multiple paths through the switch for each pair of input and output ports. This switch is a type of butterfly network, for which destination tag routing can be used [32]. Destination tag routing is an oblivious/deterministic routing technique which can suffer from poor load balancing. On the other hand it is simple to implement and fast so is often used in practice. Here, we adopt destination tag routing to obtain minimum latency at low loads. Adaptive routing may reduce latency at high network loads and will be considered in future work. The allocator for the hybrid dilated switch (using either VC or buffered switch architectures) is shown in Fig. 6c. It consists of an array of 4-port arbiters, one for each 4-port block. The 4-port arbiter has been synthesized in the same 45nm CMOS process and found to have a critical path length of 0.34 ns, giving a minimum clock period including sequencing overheads of 2.15 ns for the 6 cascaded arbiters of a 64-port switch. Pipelining can easily be applied between stages of arbitration.

A final important point to be made about allocator circuits is that the results in [10] showed that, despite their critical effect on latency, the allocator power consumption is not significant compared with other network power sources. However, the allocator synthesis power results are included in the energy analysis of Section IV.

B. Latency without contention

Figure 7 shows the head latency for various server-to-switch distances and switch configurations without contention (the case in which all speculative transmissions in the recirculation case are successful and the switch buffers are therefore not used). The head latency is defined as the time between new data arriving in the input buffers until the first bit arrives at the receiver and does not include serialization latency to remove the effect of the difference in message sizes between applications. Table I summarizes the clock period and allocation pipelining used in each case. Other control plane functions shown in Fig. 4 such as sending requests, processing grants and synchronization of requests and grants with the local clock domain take one clock cycle each (at the allocator clock period) based on timing results from synthesis. The discontinuities in Figure 7 are caused by rounding up request, grant and data transmission to the nearest clock

cycle. Using a linear fit on these results and adding the serialization latency (assuming 10 wavelengths of 10 Gb/s), the no contention latency of the buffered MZI/SOA switch in ns as a function of distance from port to switch, x (in m), and packet length, p in (B), is:

$$L = 8.9 + 20.0x + 0.08p \quad (1)$$

compared with:

$$L = 7.1 + 10.0x + 0.08p \quad \textbf{(2)}$$

for the VC switch and crossbar. It can be observed that the latency advantage of the recirculation switch increases with network dimensions, from several ns for a chip-to-chip coherence network (typical dimensions 10 - 30cm) to 20 – 40 ns in the case of a rack scale network (2 – 4 m).

C. Latency with Random Traffic

The SystemVerilog model was used to characterize the performance of the switch under load using the techniques described in [32]. Figure 8a shows the comparison between the VC switch and the two buffered switch cases with uniform random packet inter-arrival times and random destinations for the ToR application case. The latencies include the optical time of flight for data and control signals between servers and ToR (assuming a 2m fiber connection) and the serialization latency of 128B packets using 10 wavelengths of 10Gb/s. In practice, packets in the ToR switch application could be up 9000B long (assuming Ethernet). However, as our SystemVerilog allocator and buffer designs are currently limited to fixed packet sizes, we simulate for 128B packets. Larger packets would need to be split up and routed separately in this scenario. All FIFOs in the transmitter and switch buffers can contain 4 packets. Unlike the results in [17], realistic clock periods and synchronization overheads are included as discussed in section III A. It can be observed that the case 1 buffered switch maintains its latency advantage over the VC switch up to the saturation load of 65% load despite having approximately 32 times lower buffering requirements. The simpler case 2 recirculation switch saturates at 50 % load.

The allocation algorithms used in this work are designed to be fast rather than achieving a maximal matching between requests and grants. The saturation load or maximum throughput of the VC network could be increased using multiple iterations of iSLIP to approach maximal matching at the expense of a latency penalty at low loads [10, 27]. However, increasing the number of iterations in the current buffered switch allocator using a fast deterministic routing algorithm will not provide any further benefit. Therefore the proposed buffered switch architecture trades off throughput to achieve minimum latency. Further research is required to investigate adaptive allocation and routing algorithms to increase throughput in the MZI/SOA buffered switch.

D. Latency with Streaming and Incast Traffic

Random traffic is well known to be benign [27, 32]. We also tested the switch control planes using streaming and incast traffic. In the streaming case, one source port sends all its traffic to a single destination port with random destinations for traffic from all other source ports. This simulates the transmission of packetized video or large segmented messages. In the incast case, all source ports send to a single destination port. This traffic pattern is common in data center workloads, for example in large scale search algorithms and is well known to stress data center networks. Figures 8b and 8c show the performance for streaming and incast traffic respectively for the same ToR scenario. Both buffered switch cases have a higher saturation load than the VC switch for both streaming and incast traffic. Round robin arbitration used in both the VC and buffered switch cases, will not give priority to the streaming or incast packets. However, in the buffered switch case, failed speculative packets are stored close to the switch for rapid retransmission whereas in the VC case additional control latency is incurred reducing the streaming port utilization and hence maximum throughput. It can be observed that the saturation loads are very low in the incast case, as expected, due to stressing a single receiver. The saturation load was found to be very sensitive to the number of incast ports but independent of the switch buffer depths due to the strict in-order delivery policy.

IV. Energy Analysis

To assess the energy consumption in each network and demonstrate that the send and forget interface combined with MZI/SOA switching can reduce power consumed in future processor chips with integrated optical transceivers, we have modeled the power consumption of each network component. This section describes the energy models and gives results for the total network power and the power consumed on the processor chip. This analysis is for transceivers which are integrated on chip with the processor elements with optical power supplied by off-chip lasers as shown in Fig. 9. Lasers are not power gated. Other assumed parameters with references are given in Table II.

A. Optical Power and Switch Power Requirements

As previously discussed, one of the key advantages of the MZI/SOA switch architecture in the ToR or shared memory applications is the low insertion loss due to the gain of the SOA elements. However, increasing the SOA length and bias current to increase gain and reduce insertion loss has to be balanced against the increased spontaneous emission noise (and hence a higher receiver power penalty) and higher power consumption. Although only 8-port MZI/SOA switches have been fabricated to date [13], the architecture has been shown to operate with 64-ports with 1.9 dB receiver penalty using 20 mA bias current for each SOA by measurement of a 2x2 hybrid dilated switch in a recirculating loop [20]. In this configuration, each 4-port switch block has a loss of 1.2 dB

giving 7.2 dB and 8.4 dB overall losses respectively for the two buffered switch cases. This represents a good tradeoff between low insertion loss, optical signal to noise ratio (OSNR) and drive power consumption. The drive power of each SOA is 20 mW which dominates the overall power consumption of the MZI/SOA switch with the MZI drive power being negligible by comparison [29]. For comparison, we use a silicon photonic silicon micro ring resonator (MRR) switch connected in a 3-stage Clos configuration. Silicon ring resonator switches are attractive due to low area, drive powers and potential cost, but have relatively high losses. Using figures extrapolated from published literature (see Table II) we calculate that a 64-port MRR switch will have loss of 17.7 dB. Note that in practice the hybrid MZI/SOA switch, ring resonator switches and laser sources require temperature control. However, as the power consumption of temperature control will be similar for all cases, we do not include this in the energy comparison.

Input optical power requirements were calculated using the switch insertion losses discussed above and other component loss parameters given in Table II. Figure 9 shows the power budgets for the MZI/SOA switch and the ring resonator crossbar switch, demonstrating the reduction in optical power absorbed on the processor chip in the former case. It is important to note that all chip-to-chip links are assumed to use fiber which has negligible loss on these network scales and, hence, there is no significant difference in the power budgets between the ToR and shared memory network applications.

B. Electronic Control and Transmission Power

Power models for the control plane circuits are obtained by synthesizing the SystemVerilog models of the transmitter controller, allocator and recirculation buffers using in a 45nm standard cell ASIC flow and Synopsys Design Vision. Activity data is captured from the SystemVerilog simulations using Modelsim and power is estimated using Synopsys Primetime.

The power consumption of transmitter and receiver front ends is taken from measurements on a recently reported transceiver [7]. Serialization and deserialization (SERDES) power is found using the CONTEST open source transceiver design toolkit [37]. SERDES and transmitter front ends are assumed to be power gated; receiver front ends, control plane circuits and optical power supplies are always on.

C. Power Dissipated on Server Chip

Figure 10a shows the power dissipated on the processor chip at 30% network load for the MMR and MZI/SOA switches and the VC and proposed buffered switch architectures. The gain of the MZI/SOA switch substantially reduces the optical power absorbed on the server chip due to a reduction in the power budget from 26.8 dB with the MMR switch down to 16.3 dB. The simplified send and forget interface used for the buffered switch also, significantly reduces the network adapter (transmitter control) power due to reduced FIFO storage requirements (reduced from 55.7 mW at in the VC case to 0.9 mW at 30 % load for the ToR application). However, these figures are for 128 B packets. Greater packet lengths will increase FIFO memory requirements and hence adapter power consumption. For example, providing storage for four 1500 B Ethernet packets in the ToR case will increase the power consumption of the send and forget adapter to 8.0 mW at 30% load. However, in this case, the VC adapter will increase to 304 mW. In the shared memory case, maximum packet lengths are fixed by the cache block size. The remaining power consumption in the buffered switch is dominated by receivers and SERDES. Receivers could be power gated at the expense of a latency penalty using a reservation scheme [19]. SERDES is an inevitable consequence of operating at the high bit rates of optical links, but is energy proportional with a fixed energy per bit [37]. Overall, at 30% network load, the buffered switch architecture reduces the power dissipated on the processor chip by 64 % from 171.0 mW to 61.1 mW in the ToR application and by 60 % from 150.6 mW to 60.8 mW in the shared memory application. These results are for the case 2 switch. There is an additional power dissipation of 1.7 mW, constant over all load levels using the case 1 switch due to the loss of the output switch.

In all cases, the power dissipated on the server chip scales linearly with load as shown in Fig 10b. The gradient of dissipated power against load is greater for the VC architectures, due to the more complex adapter, particularly for the larger packets of the ToR application.

D. Total Network Power

Figure 11a shows contributions to the total power of the 64-port switch networks at 30% load. For MMR switches, the power is dominated by optical power due to high optical losses. Assuming MZI/SOA switches are used, the buffered switch architectures have increased power consumption over the VC case as the power of the additional transmitters, receivers and adapters at the switch and the effect of the additional input/output switches outweighs that of the larger VC transmitter adapter. As shown by Figure 11b, the low required transmitter powers combined with the gain provided by the SOAs means that the MZI/SOA switch cases are also more energy proportional with power consumption of 4.8 – 6.6 W at low loads. The power of the MZI/SOA switches approaches that of the MRR switch at high loads as the SOA power dominates. MMR switches would require transmitter based optical power gating, not easy to apply without a latency penalty, to achieve the same levels of energy proportionality. The increase in the buffered case 1 switch compared with the VC switch is 2.4 W or 48 %. This increases to 4.3 W (20%) at 60% load as the switch buffers are used more often. The energy proportionality of the SOA based switches also means that there is only a small increase in total power for the more complex case 1 buffered switch compared with the case 2 switch. The power differences between the two switch cases is reduced at high loads as more packets use the buffers in case 2.

V. Discussion

In section I, two potential applications of the buffered switch architecture in future high performance servers were described: optical top-of-rack replacement and multiple socket shared memory networks.

For the top-of-rack switch application, optical switching using wavelength striped (WDM) links provides high bandwidth without pin or front panel limitations or the requirement for power hungry electronic switching fabrics. Store and forward 10G Ethernet switches can introduce latencies from 100ns up to 10 us plus processing depending on the packet length. High performance cut through routers can start to forward the packet after receiving the first 54B (MAC addresses, Ethertype and IPv4 layer 3 and 4 headers) taking on the order of 100 ns before starting to forward packets of any length. By comparison, the optical ToR bypasses the buffering and processing in electronic switches but introduces an overhead due to optical switch allocation and reconfiguration. The optical buffered switch proposed in this paper mitigates this overhead using speculative transmission. To accurately compare the optical buffered switch with an electronic cut through switch independently of packet length and distance, the 100ns cut through forwarding time should be compared with the sum of the request synchronization, allocation and switching times which is 7 clock cycles or 5 ns (see table I). The 100 Gb/s bandwidth of the optical ToR also reduces serialization latency compared with current 10 Gb/s Ethernet ToRs to reduce incast issues without pin or front panel bandwidth limits. It has to be noted however, that applications running in a data center environment have a wide range of end-to-end latency requirements down to a few microseconds and not all applications will benefit from the reduced latency. From an energy point of view, the Ethernet ports on current CMPs represent only a small proportion of chip power consumption, so the reduction in CMP dissipation for the proposed architecture is a relatively minor advantage. Total power consumption comparisons with electronic Ethernet switches are difficult; however, the energy proportionality demonstrated by the MZI/SOA switch is an important advantage over electronic equivalents [1]. It has to be noted however that the energy savings through reduced buffering in the send and forget interface are near the lower bound as we consider relatively small packets of 128B.

In the shared memory coherence network case, while energy proportionality is also an important advantage, the power dissipated on the CMPs is critical in order to reduce the large proportion of power consumed by off-chip communications in such chips. The server chip described in [3] has total coherence bandwidth of 460 Gb/s using electronic SERDES consuming 11.1 mW/(Gb/s) giving a power consumption of 5.1W, significant compared with the 120W total processor power envelope. By comparison, the processor chip power dissipation of our architecture (at 30% load) is 0.5 mW/(Gb/s), consuming only 0.23W for the same coherence bandwidth. Such comparisons are difficult, for example the electronic SERDES power includes other physical layer functions such as clock recovery, coding and equalization (some of which are not required in the optical case) whereas our buffered switch power figures includes buffering not included in the electronic case. However, the more than order of magnitude reduction suggests that the proposed architecture can make significant reductions in CMP dissipation. As discussed in section I, latency is a key factor in shared memory networks. The proposed architecture has the ability to connect each core over an optical switch, avoiding the two stage network of current multiple-socket systems. The results demonstrate that cores on different chips can be connected with similar latency to cores on the same chip. For example, for an electronic 16 core mesh network-on-chip using single cycle routers operating at 1 GHz clock frequency, the head latency (ignoring messages size) is between 3 and 13 ns depending on the position of source and destination cores [32]. Figure 7 shows that networks with <30cm distance between port and switch have a head latency of <10ns.

In both applications, scalability in both port count and bandwidth per port is important to support future increase in compute capacity and density. Our ongoing research into hybrid switch design aims to build very large port count optical switches. We believe that integration of larger than 128 port count optical switches is feasible in the future. We have demonstrated 10×10Gb/s operation with the hybrid MZI/SOA switch [31] and we are now aiming at demonstrating higher bit rate operations. The large operating wavelength range also allows operation with more than 10 wavelengths.

Finally, we do not consider the latency or energy implications of data synchronization in this work which will be an important issue in future chip-to-chip optically switched interconnects. Source synchronous wavelength striped optical links have been demonstrated operating at up to 4 Gb/s [38] and due to the fundamentally lower delay variation in photonic compared with electronic links [39] are a possible candidate for higher bit rates. Injection locking clock recovery, either electronic [40, 41] or optical [42] is another promising solution to the synchronization problem and recovery times below 25 ns have been demonstrated in both cases.

VI. Conclusions

We have proposed a low latency optical switch architecture for data center top of rack and shared memory coherence network applications and compared it with a high performance optical VC switch and electronic alternatives. The proposed architecture has the important property of minimizing the power consumed and dissipated in future server chips with integrated photonic transceivers thus mitigating the dark silicon effect. SOA based switching is often thought to be a high power option. However, the results shown in this paper demonstrate that it gives greater energy proportionality and allows effective power management. The speculative control plane with electronic buffering at the switch both reduces latency and further reduces the complexity and power consumption of

the server side circuits.

ACKNOWLEDGMENT

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) INTERNET program grant and an EPSRC Fellowship grant to Philip Watts. Both University College London and the University of Cambridge are members of GreenTouch.

REFERENCES

[1] L.A.Barroso, J.Clidaras, U.Hölzle, "The Datacenter as a Computer", 2nd edition, (Morgan Claypool 2013).

[2] H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, D. Burger, "Dark Silicon and the End of Multicore Scaling", IEEE Micro **32**(12), 2012

[3] J. L. Shin, D. Huang, B. Petrick, C. Hwang, K.W. Tam, A. Smith, H. Pham, H. Li, T. Johnson, F. Schumacher, A.S. Leon, A. Strong, "A 40 nm 16-core 128-thread SPARC SoC processor", IEEE J. Solid-State Circuits 46(1), pp. 131–144, 2011.

[4] A. Shacham and K. Bergman, "Building Ultralow Latency Interconnection Networks Using Photonic Integration," IEEE Micro 27(4), 2007.

[5] B. Lee, A. Rylyakov, W. Green, S. Assefa, C. Baks, R. Rimolo-Donadio, D. Kuchta, M. Khater, T. Barwicz, C. Reinholm, E. Kiewra, S. Shank, C. Schow, and Y. Vlasov, "Monolithic silicon integration of scaled photonic switch fabrics, cmos logic, and device driver circuits," J. of Lightwave Technology, vol. 32, pp. 743-751, Feb 2014

[6] A. Biberman, G. Hendry, J. Chan, H. Wang, K.B Preston, N. Sherwood-Droz, J.S. Levy, M. Lipson, "CMOS-Compatible Scalable Photonic Switch Architecture Using 3D-Integrated Deposited Silicon Materials for High-Performance Data Center Networks", Proceedings of Optical Fiber Communications (OFC) Conference, Los Angeles, March 2011.

[7] 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, A.V. Krishnamoorthy, "Ultra-efficient 10Gb/s hybrid integrated silicon photonic transmitter and receiver", Opt. Express 19(6), 5172-5186, 2011

[8] Y. Liu, J. M. Shainline, X. Zeng, and M. Popovic, "Ultra-low-loss waveguide crossing arrays based on imaginary coupling of multimode bloch waves," in Advanced Photonics 2013.

[9] A. Poon, X.S. Luo, F. Xu, H. Chen, "Cascaded Microresonator-Based Matrix Switch for Silicon On-Chip Optical Interconnection," Proc. of the IEEE, vol. 97, no. 7, 2009.

[10] P.M. Watts, A.W. Moore, S.W.Moore, "Energy implications of photonic networks with speculative transmission", J. Optical Comms and Networking 4(6), 2012

[11] M. Ortin Obon, L. Ramini, V. Viñals, D. Bertozzi, "Capturing Sensitivity of Optical Network Quality Metrics to its Network Interface Parameters", Workshop on Exploiting Silicon Photonics for energy-efficient heterogeneous parallel architectures (SiPhotonics'14), Vienna, Jan 2014.

[12] C. Batten, A. Joshi, J. Orcutt, A. Khilo, B. Moss, C.W. Holzwarth, M.A. Popovic, H.Q. Li, H.I. Smith, J.L. Hoyt, F.X. Kartner, R.J. Ram, V. Stojanovic, K. Asanovic, , "Building Manycore Processor-to-DRAM Networks", IEEE Micro, Vol. 29, pp. 8-21, 2009

[13] S. Beamer, C. Sun,Y-J. Kwon, A. Joshi, C. Batten, V. Stojanovi, K. Asanovi, "Re-Architecting DRAM Memory Systems with Monolithically Integrated Silicon Photonics", Proceedings of the International Symposium on Computer Architecture (ISCA), June 2010.

[14] J. L. Hennessy and D. A. Patterson, Computer Architecture, A Quantitative Approach. Morgan Kaufmann, 4th ed., 2007.

[15] O. Liboiron-Ladouceur, B.A. Small, K. Bergman, "Physical layer scalability of WDM optical packet interconnection networks", Journal of Lightwave Technology, Vol. 24, pp. 262-270, 2006

[16] I. White, A.E. Tin, K. Williams, H.B. Wang, A. Wonfor, R. Penty, "Scalable optical switches for computing applications", Journal of Optical Networking 8(2), pp. 215-224, 2009.

[17] S. Liu, Q. Cheng, A. Wonfor, R. Penty, I. White, P.M. Watts, "A Low Latency Optical Top of Rack Switch for Data Centre Networks with Minimized Processor Energy Load", Proceedings of Optical Fiber Communications (OFC), San Francisco, March 2014.

[18] R. Luijten, C. Minkenberg, R. Hemenway, M. Sauer, R. Grzybowski, "Viable opto-electronic HPC interconnect fabrics", Proceedings of the ACM/IEEE Supercomputing Conference 2005

[19] Y. Pan, P. Kumar, J. Kim, G. Memik, Y. Zhang, A. Choudhary, "Firefly: Illuminating future network-on-chip with nanophotonics," in Int. Symp. on Comput. Archit., 2009.

[20] D. Vantrease, R. Schreiber, M. Monchiero, M. McLaren, N.P. Jouppi, M. Fiorentino, A. Davis, N. Binkert, R.G. Beausoleil, J.H. Ahn, "Corona: System implications of emerging nanophotonic technology," in Int. Symp. on Computer Architecture (ISCA), 2008.

[21] A. Krishnamoorthy, R. Ho, X.Z. Zheng, H. Schwetman, J. Lexau, P. Koka, G.L. Li, I. Shubin, J.E. Cunningham, "Computer systems based on silicon photonic interconnects," Proc. of the IEEE, vol. 97, no. 7, 2009.

[22] S. Beamer, K. Asanovic, C. Batten, A. Joshi, and V. Stojanovic, "Designing multi-socket systems using silicon photonics," in Proceedings of the 23rd International Conference on Supercomputing (ICS), 2009.

[23] X. Ye, Y. Yin, S. Yoo, P. Mejia, R. Proietti, V. Akella, "DOS: A scalable optical switch for datacenters," In Proc. Symp. Arch. for Networking and Comms. Systems (ANCS), 2010.

[24] N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, "Helios: a hybrid electrical/optical switch architecture for modular data centers," ACM SIGCOMM Computer Communication Review, vol. 41, no. 4, pp. 339-350, 2011.

[25] A. Singla, A. Singh, K. Ramachandran, L. Xu, and Y. Zhang, "Proteus: a topology malleable data center network," In Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks, Oct. 2010.

[26] L. Liu, Z. Zhang, Y. Yang, "Packet Scheduling in a low – latency optical interconnect with electronic buffers", J. of Lightwave Technology, Vol. 30, No. 12, pp. 1869-1881, June 2012

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

[28] I. Iliadis, and C. Minkenberg, "Performance of a speculative transmission scheme …" IEEE Trans. Networking 16(1), 2008

[29] Q.Cheng, A. Wonfor, R.V. Penty, I.H. White, "Scalable, low energy hybrid photonic space switch", Journal of Lightwave Tech. 31 (18), pp. 3077-3084, 2013

[30] Q.Cheng, A. Wonfor, J.L. Wei, R.V. Penty, I.H. White, "Demonstration of the feasibility of large port count optical switching using a hybrid MZI-SOA switch module in a recirculating loop", Optics Letters 39 (18), pp. 5244-5247, Sept 2014.

[31] Q Cheng, A Wonfor, J L Wei, R V Penty, and I H White, 'Monolithic MZI-SOA Hybrid Switch for Low-power and Low-penalty Operation', Optics Letters, Vol. 39, Issue 6, pp 1449-1452, 2014.

[32] W.J.Dally and B.Towles, “Principles and Practices of interconnection networks”, Morgan Kaufmann, 2004

[33] P. Koka, M. O. McCracken, H. Schwetman, X. Zheng, R. Ho, and A. V. Krishnamoorthy, “Silicon-photonic network architectures for scalable, power-efficient multi-chip systems,” SIGARCH Comput. Archit. News, vol. 38, pp. 117–128, June 2010.

[34] Y. Liu, J. M. Shainline, X. Zeng, and M. Popovic, “Ultra-low-loss waveguide crossing arrays based on imaginary coupling of multimode bloch waves”, Optics Letters, Vol. 39, No. 2, pp. 335-338, 2014

[35] D. Livshits, A. Gubenko, S. Mikhrin, V. Mikhrin, C.H. Chen, M. Fiorentino, R. Beausoleil, “High efficiency diode comb laser for DWDM optical interconnects”, IEEE Optical Interconnects Conference, May 2014.

[36] V.R. Almeida, R.R. Panepucci, M. Lipson, “Nanotaper for compact mode conversion”, Optics Letters, vol. 28, pp. 1302-1304, 2003

[37] Y. Audzevich, P.M. Watts, A. West, A. Mujumdar, S.W. Moore, A.W. Moore, "Power Optimized Transceivers for Future Switched Networks," IEEE Trans. on VLSI, Vol. 22, No. 10, pp. 2081-2092, 2013.

[38] C.E Gray, O. Liboiron-Ladouceur, D.C. Keezer, K. Bergman, “Test electronics for a multi-Gb/s optical packet switching network,” in Electronics Packaging Technology Conference (EPTC), December 2006.

[39] G. Q. Chen, H. Chen, M. Haurylau, N.A. Nelson, D.H. Albonesi, P.M. Fauchet, E.G. Friedman, “Predictions of CMOS compatible on-chip optical interconnect,” in Integration, the VLSI journal, vol. 40, 2007.

[40] B. Li, L.S. Tamil, D. Wolfe, J. Plessa, “10 Gb/s burst-mode optical receiver based on active phase injection and dynamic threshold level setting”, IEEE Communications Letters, Vol. 10, No.10, pp. 722 -724, Oct 2006.

[41] J. Lee, M. Liu, “A 20 Gb/s Burst-Mode CDR Circuit Using Injection-Locking Technique”, Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC), 2007.

[42] L. Jun, J. Parra-Cetina, P. Landais, H.J.S Dorren, N. Calabretta, “Performance Assessment of 40 Gb/s Burst Optical Clock Recovery Based on Quantum Dash Laser”, IEEE Photonics Technology Letters, Vol. 25, No. 22, pp. 2221-2224, 2013.

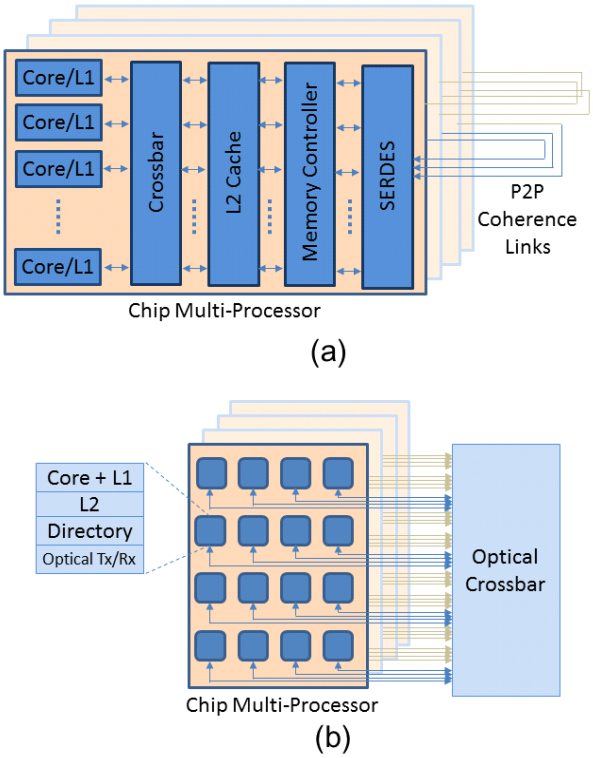

Fig. 1. Shared memory coherence networks for multiple socket servers (a) Due the fundamental difference between electronic communications for on-chip (wide buses of small wires) and off-chip (serial transceivers driving transmission lines), separate networks are currently used for on-chip and chip-to-chip coherence. (b) Optical switching could provide a single network connecting all cores on multiple chips.

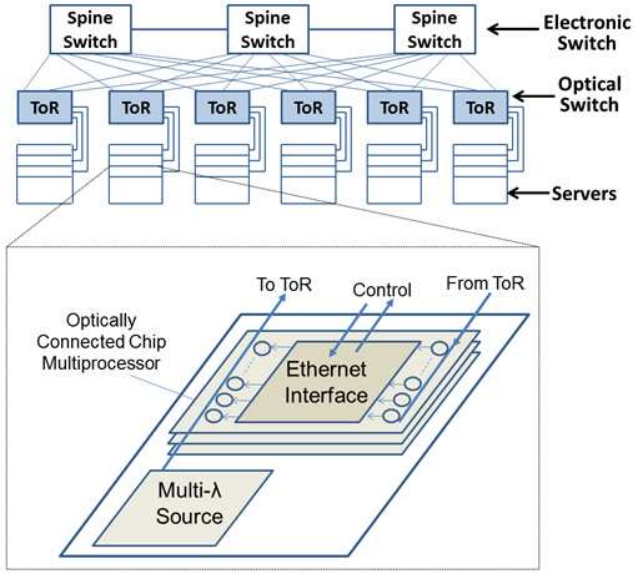

Fig. 2. Integrated optical transceivers packaged with the chip multiprocessor and an optical top-of-rack switch connecting to spine Ethernet switches can provide 2 hop connections between any two processors in a data center. The optical top-of-rack switch replaces the conventional Ethernet switch used for this purpose with lower power consumption and latency.

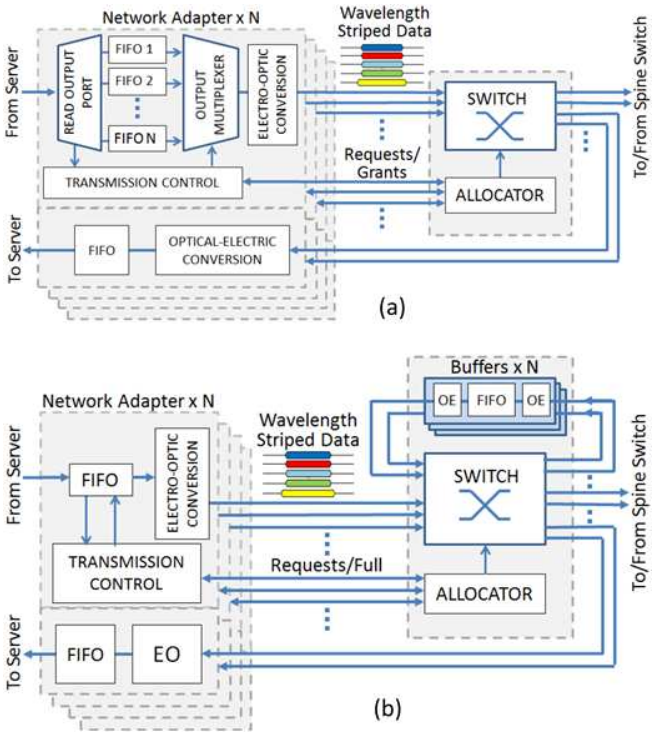

Fig. 3. Control plane architectures (a) baseline input queued VC switch (b) proposed send and forget interface with electronic buffers at the switch. All data transmission is wavelength striped consisting of 10 wavelengths at 10 Gb/s each. The network adapter is the server side interface. The switch and allocator are located in the top-of-rack switch. OE = optical to electronic conversion, EO = electronic to optical conversion.

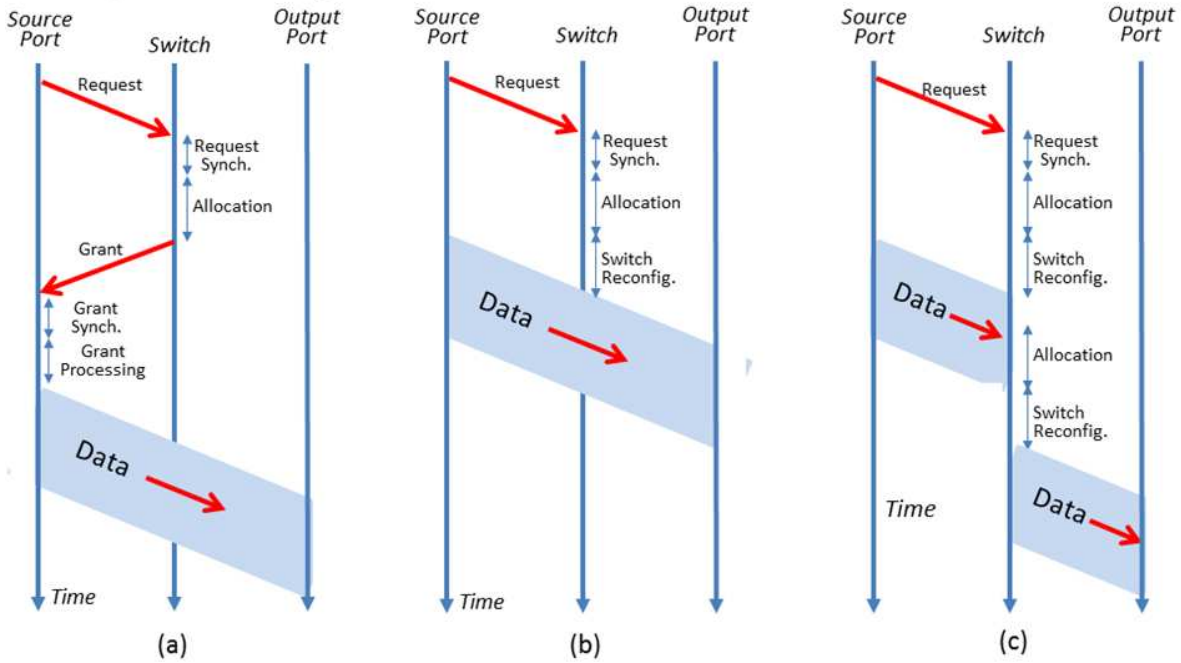

Fig. 4. Latency comparison of (a) VC Switch (b) Successful allocation in a speculative or buffered switch and (c) failed allocation in the buffered switch with retransmission from the switch buffers.

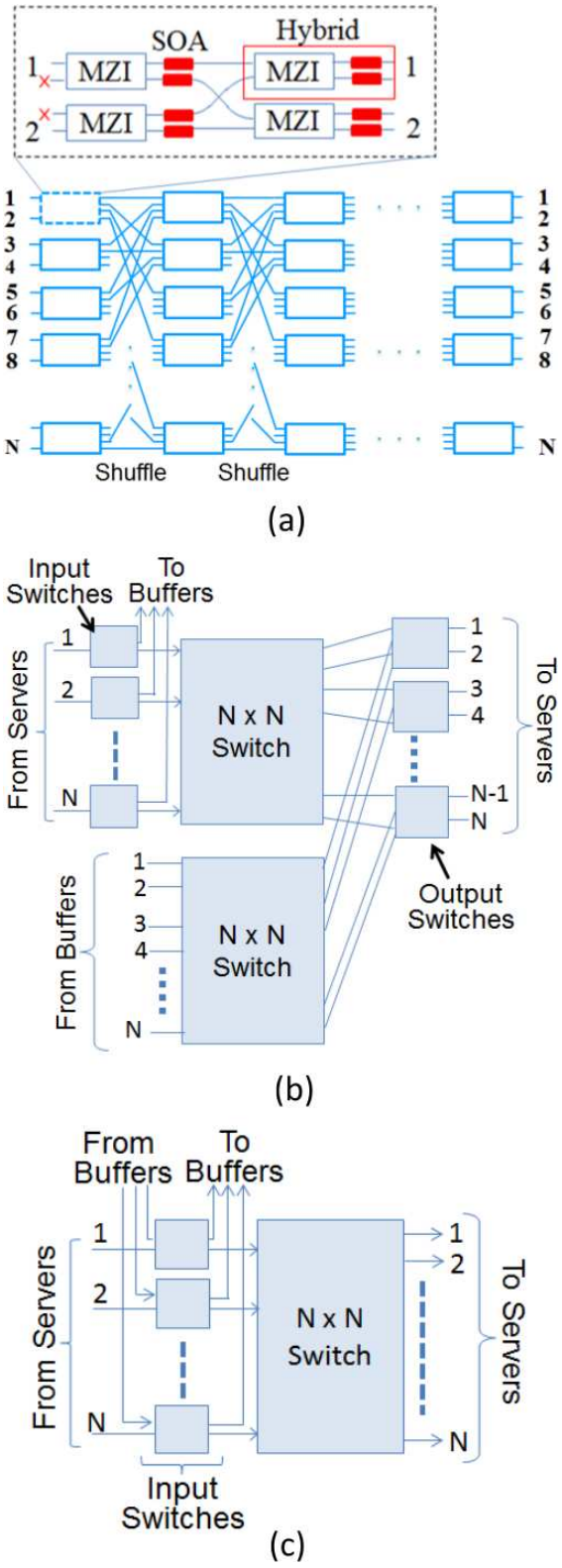

Fig. 5. Hybrid MZI/SOA switch architectures (a) An NxN switch consists of a matrix of 4-port switch blocks as shown in the callout. The input and output stages of the NxN switch use only 2 ports due to dilation. The two options for connecting the buffer ports in the buffered switch architectures are shown in (b) case 1 and (c) case 2. The input and output switches use the same 4-port switch blocks.

TABLE I
CONTROL PLANE LATENCY PARAMETERS

Parameter	Value
Synchronization of requests/grants	1 cycle
VC allocator pipelining	2 cycles
VC allocator clock period	1.2 ns
Buffered xbar switch allocator pipelining	1 cycle
Buffered xbar switch allocator clock period	0.75 ns
Buffered MZI/SOA allocator pipelining	3 cycles
Buffered MZI/SOA allocator clock period	0.8 ns
Switching time	2 ns

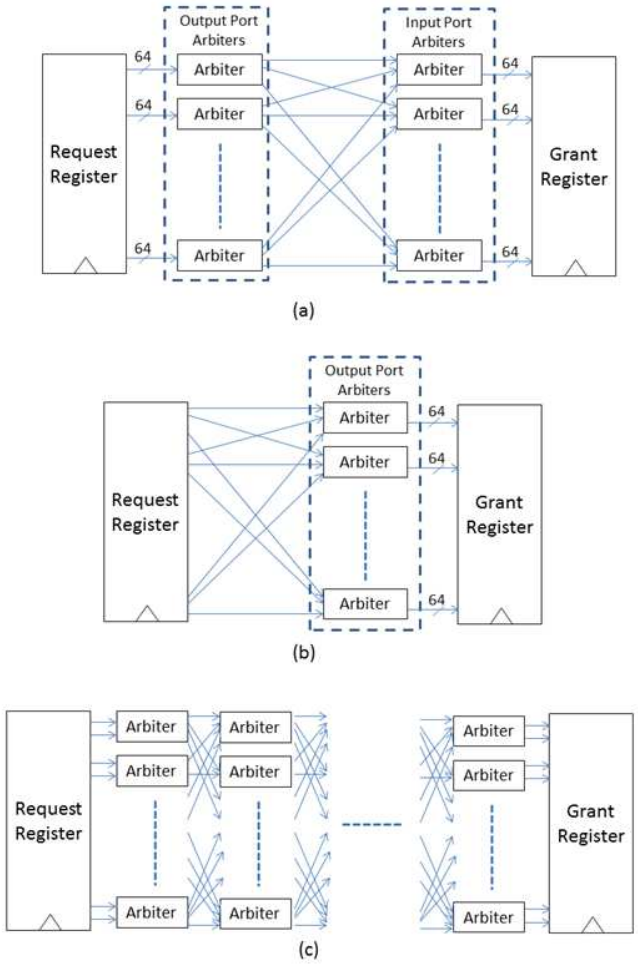

Fig. 6. Allocators for optical switches used in this work. (a) A separable VC allocator for a crossbar optical switch. (b) An allocator for an optical crossbar using the buffered switch control plane (c) An allocator for the hybrid MZI/SOA optical switch using the buffered switch control plane.

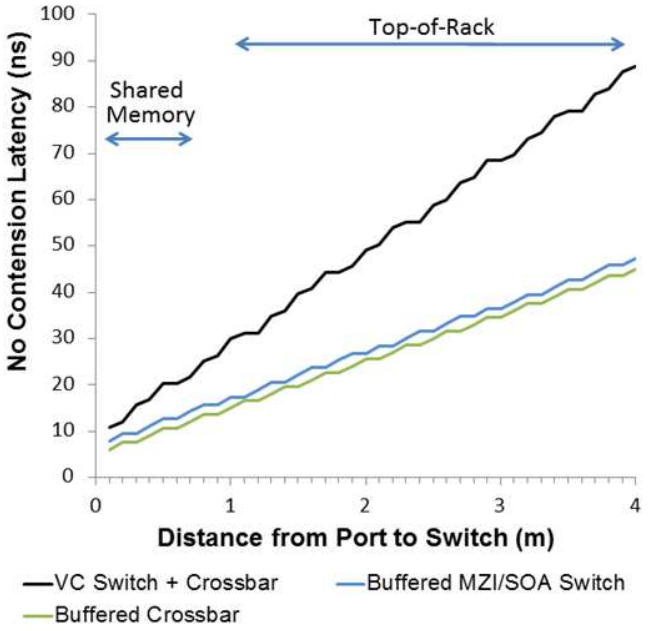

Fig. 7. Head latency without contention taking into account allocator clock period differences and network dimensions. Typical scales for the shared memory coherence and top-of-rack networks are indicated.

TABLE II
ENERGY MODELING ASSUMPTIONS

Parameter	Value
Bit rate per wavelength	10 Gb/s
No. of wavelengths per port	10
Loss of 4-port MZI/SOA switch block	1.2dB[30]
SOA drive current at 1V bias voltage	20 mA [30]
Ring resonator modulator loss	4 dB [33]
Silicon waveguide loss	1 .3 dB/cm
Off-chip waveguide loss	negligible
Waveguide crossing loss	0.04 dB [34]
Ring resonator through loss	0.33 dB [9]
Ring resonator drop loss	1.6 dB [9]
Power Consumption of ring resonator per Circumference	1.3 W/m [9]
Receiver sensitivity	-18 dBm[7]
Receiver front-end power	2.6 mW [7]
Transmitter front-end power	0.66 mW [7]
Laser Efficiency	30% [35]
Loss at silicon/fibre interface	0.5 dB [36]
Packet size for ToR application	128 B

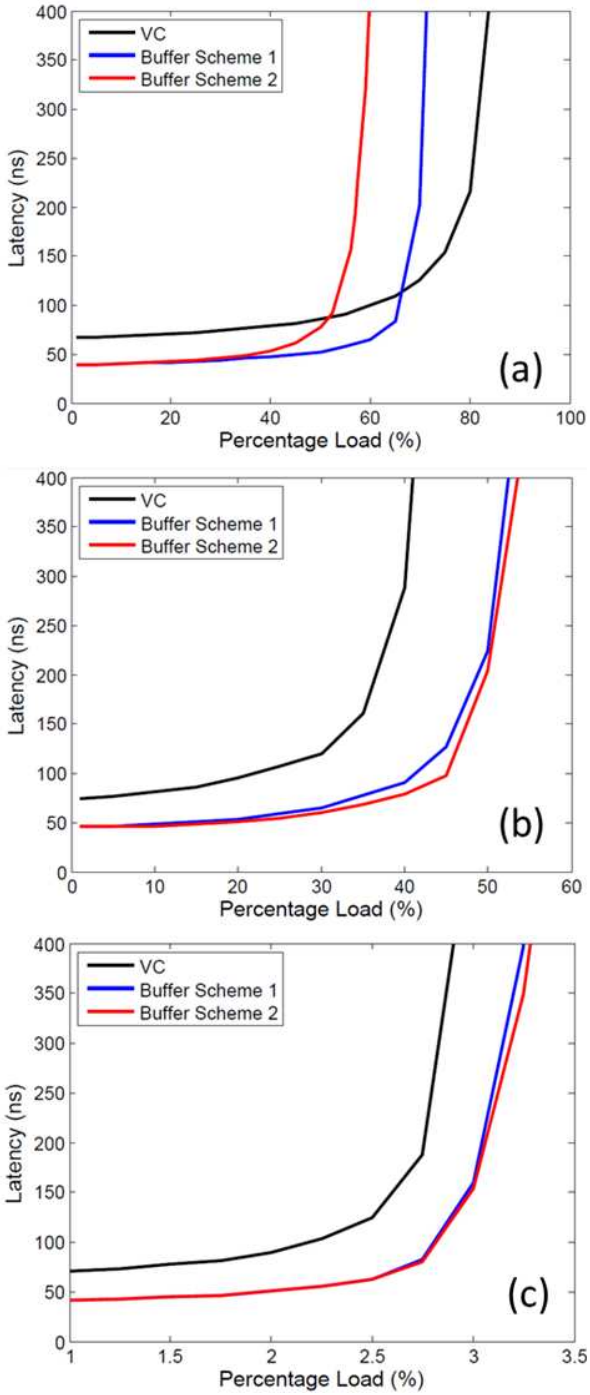

Fig. 8. Latency vs load for (a) uniform random traffic (b) streaming traffic between two ports with random traffic on other ports (c) incast traffic.

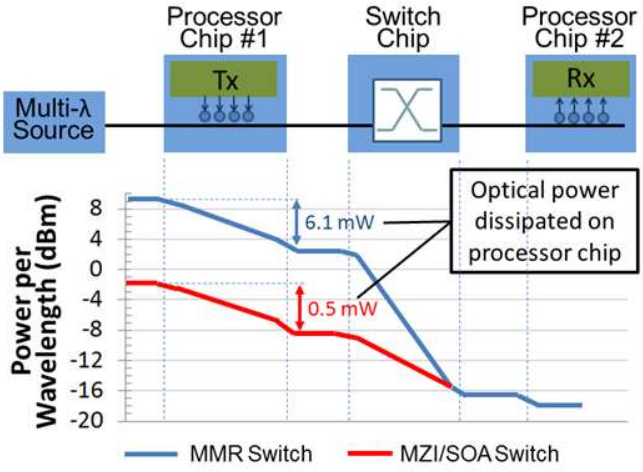

Fig. 9. Power budgets for links using a 64-port micro-ring resonator (MMR) crossbar and a 64-port MZI/SOA switch.

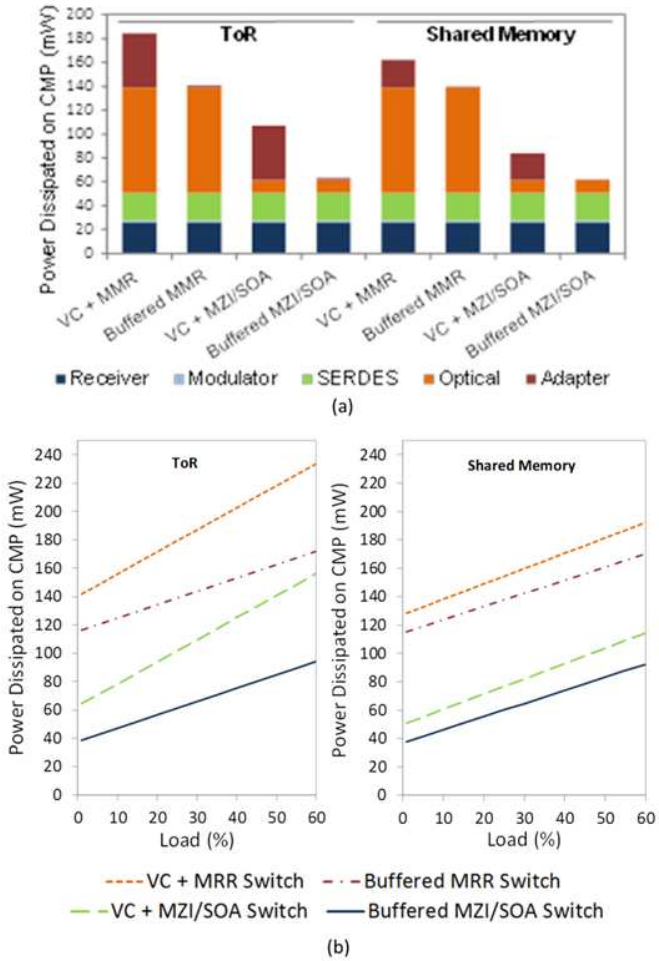

Fig. 10. Power dissipated on the server chip (a) sources of power dissipation at 30% load (b) power dissipation versus load. The adapter contains all the server based FIFOs and transmission control.

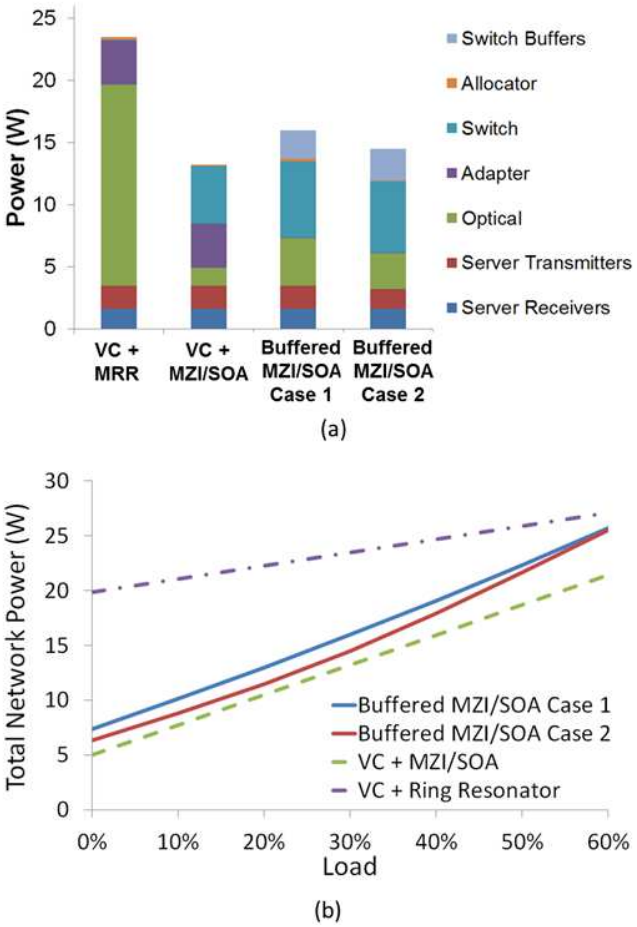

Fig. 11. Total network power (a) showing breakdown by component at 30% load and (b) power against network load. The switch buffers include receivers, electronic FIFOs, modulators and SERDES. Server transmitters include modulators and SERDES. The adapter contains all the server based FIFOs and transmission control.