CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. 0000; **00**:1–17
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cpe

Towards Zero Latency Photonic Switching in Shared Memory Networks

Muhammad Ridwan Madarbux, Anouk Van Laer, Philip M. Watts
Dept. of Electronic and Electrical Engineering
University College London
{m.madarbux,anouk.vanlaer,philip.watts}@ucl.ac.uk
Timothy M. Jones
Computer Laboratory
University of Cambridge
timothy.jones@cl.cam.ac.uk

SUMMARY

Optical networks-on-chip based on silicon photonics have been proposed to reduce latency and power consumption in future chip multiprocessors (CMP). However, high performance CMPs use a shared memory model which generates large numbers of short messages, creating high arbitration latency overhead for photonic switching networks. In this paper we explore techniques which intelligently use information from the memory hierarchy to predict communication in order to setup photonic circuits with reduced or eliminated arbitration latency. Firstly, we present a switch scheduling algorithm which arbitrates on a per memory transaction basis and holds open photonic circuits to exploit temporal locality. We show that this can reduce the average arbitration latency overhead by 60% and eliminate arbitration latency altogether for up to 70% of memory transactions. We then demonstrate that this switch scheduling algorithm operating with a central photonic crossbar or Clos switch has significant energy efficiency benefits over arbitration-free photonic networks such as Single Writer Multiple Reader (SWMR) networks. Finally, we demonstrate that cache miss prediction can be used to predict 86% of more complex memory transactions involving multiple nodes or main memory.

Received . . .

KEY WORDS: Photonic interconnection networks; Networks-on-chip; Shared memory architectures

1. INTRODUCTION

Photonic networks on chip (NoC) based on advances in silicon photonics have been widely proposed as one of the solutions to the serious problems of energy consumption and thermal management in chip multiprocessors (CMP) [1–6] due to the fundamentally lower power consumption of photonic communication [7]. In addition, photonic communication enables high bandwidth end-to-end routes for global on-chip paths or for systems spanning multiple chips, without significant power penalties. Figure 1(a) shows a current typical 4-socket high performance shared memory server architecture based on [8]. Due to 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 used for on-chip and chip-to-chip communications with the architecture constrained by the limitations of the electronic interconnect. Furthermore, the SERDES used in off-chip communications consume >20% of total chip power [8]. By contrast, there is no fundamental difference between photonic on-chip and off-chip links, allowing us to build single unified low

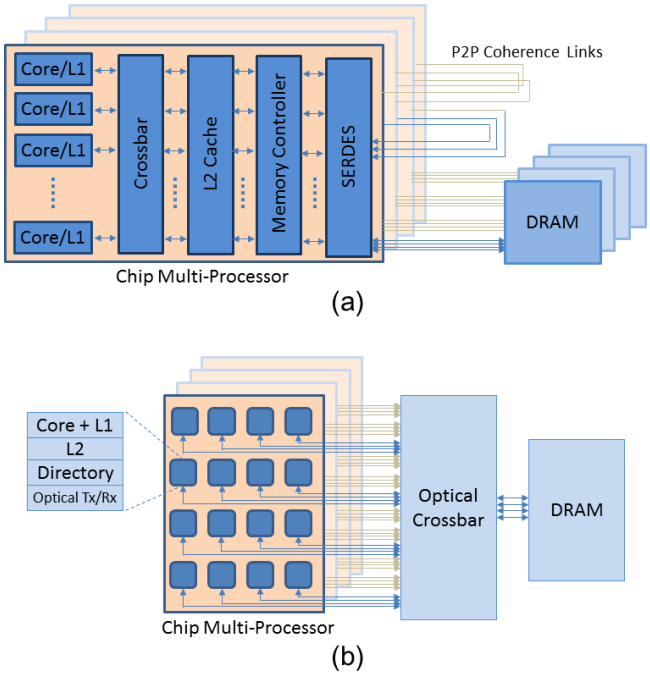

Figure 1. High performance multi-socket servers (a) current architecture with NOC and point-to-point chip-to-chip SERDES links (b) future architecture with unified switched photonic network.

latency photonic networks, as shown in Figure 1(b), to increase performance of shared memory systems spanning multiple chips, or even boards and racks.

NoCs in current systems consist of electronic crossbars [8] or meshes [9] relying on multiple hops between sequential elements. However, photonic NoCs require end-to-end optical paths to be set up in advance of communication meaning that the resulting latency overhead of arbitration and control message transmission between cores and a central switch can be significant. Figure 2 shows the sources of latency in a scheduled photonic switch. Setting up an optical path involves sending a request to the switch arbiter, performing arbitration and returning a grant to the requesting port. We label this time between the transmission of the optical path request and the actual start of the optical transmission, the arbitration latency. The head latency is the time taken for the head of the message to be received at the destination port and includes serialization and deserialization times as well as the time of flight in the waveguide. Note that head latency also applies to the request and grant control messages. Synchronisation latency can be neglected in NoCs in which the transmitter and receiver share the same clock, but can be significant in chip-to-chip networks - we discuss this issue further in the conclusions. Data serialization latency can be very low if a broadband switch is used and messages are wavelength striped to use the high bandwidth of photonic links (high bit rate and multiple wavelengths per waveguide).

This paper focuses on the question of reducing or eliminating arbitration latency with the use of simple optical switch structures. As the majority of traffic in a shared memory system consists of short (8–256 B) coherence messages between caches and directory controllers, arbitration latency can impose a high overhead. Various proposed schemes for overcoming this latency overhead are reviewed in Section 2, but all involve an increase in the number of optical components and/or the complexity of the control plane. In contrast we explore techniques for eliminating arbitration latency by prediction of communication within shared memory systems. Prediction already plays a major role in increasing the performance of modern computer architectures, for example through branch speculation or prefetching in cache hierarchies. The prediction techniques discussed here could also be used in electronic networks, but, due to the hop by hop communication nature of meshes or highly pipelined crossbars, they will have lower impact than in future silicon photonic networks.

DOI: 10.1002/cpe

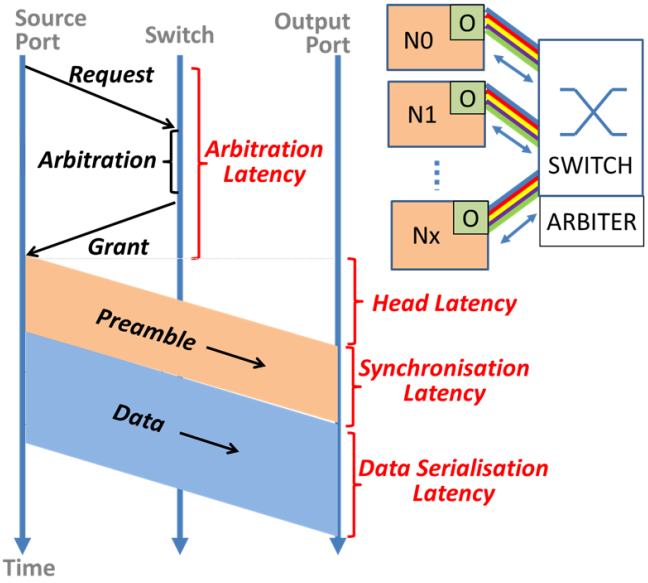

Figure 2. Sources of latency in a wavelength striped photonic switch.

The rest of the paper is organised as follows: following the review of previous work on reducing or avoiding arbitration latency in Section 2, we describe the shared memory system and photonic networks assumed in this work in section 3. Section 4 presents a circuit scheduling algorithm which arbitrates on a per memory transaction basis rather than on individual messages. The benefits in latency and performance (Section 5) and power consumption (Section 6) compared with the Single Write Multiple Read (SWMR) architecture, which avoids arbitration, are presented. Section 7 explores the concept of reducing latency in setting up photonic paths to main memory by prediction of cache misses. Finally Section 8 discusses the system implications of these results and further work.

2. PREVIOUS WORK

In this section we review techniques proposed for reducing control and arbitration latency in photonic computer networks. Speculative transmission, in which messages are transmitted before a grant has been received and either dropped or redirected if there is no path available, has been proposed, either operating in parallel with a centralized arbiter (OSMOSIS [10]) or independently (SPINet [1]). Speculative transmission forces the use of strict time slots and, used independently, suffers from reduced maximum throughput and head of line blocking. High performance speculative schemes also require the additional complexity of reordering in the receiver [10, 11]. SPINet [1] also reduced arbitration latency using a distributed arbitration scheme consisting of a separate wavelength transmitted with the data to determine the configuration of each switching stage, whereas CORONA used an optical token ring arbitration scheme [2]. The single writer multiple reader (SWMR) topology adopted by Firefly [4] avoids arbitration altogether by allowing each node to receive from all other nodes simultaneously but requires flow control to avoid receiver buffer overflow.

Oracle's Macrochip [3] also avoids arbitration using a wavelength and space division multiplexed all-to-all network. Avoiding arbitration in SWMR and all-to-all networks comes at the cost of an increased number of photonic components. SWMR requires N transmitters and $N(N-1)$ receivers and an all-to-all network requires $N(N-1)$ transmitters and $N(N-1)$ receivers. On the other hand, an optical crossbar with the same bandwidth per port requires only N transmitters and N

DOI: 10.1002/cpe

receivers. Both SWMR and all-to-all networks also suffer from high serialisation latency compared with wavelength striped approaches as explained in the following section.

Other architectures reduce the arbitration overhead by splitting up the network into smaller photonic switch sections interspersed with optical-electrical-optical (OEO) conversions to allow electronic buffering, for example [6] in which routing in the x and y directions of an optical mesh are handled separately. However, these schemes reduce the power consumption and latency benefits of introducing photonic networks.

In contrast to the packet switched networks discussed above, the use of relatively long-lived optical circuits to provide low latency transmission of long lived flows or large messages has been investigated in the context of supercomputers [12] and a torus NoC [5]. In this case it is usually necessary to have a backup electronic network to carry small messages. For shared memory systems, the authors have investigated the concept of setting up long lived circuits ($\gg$ message length) between cores which have dense memory sharing requirements. Initial results [13] showed that, with ideal circuit setup decisions made on less than $1\ \mu s$ time periods, a large proportion of traffic from PARSEC applications could be routed onto the circuit switch. However, further investigation has shown that adding background traffic from the operating system considerably reduces the benefits. In addition, overall power consumption is dominated by the backup electronic network, so the power savings from adding the optical circuit switch are proportionally small.

In contrast to the above, this paper discusses techniques for intelligently setting up optical paths by predicting network communication using information from the memory hierarchy. For NoCs, various prediction schemes have been proposed to reduce the latency of the average memory request. In [14], the need for cache-to-cache transfers are predicted based upon the program counter, while caches holding copies of the requested data are predicted using both the program counter and the requested memory address. In [15] prediction is used to forward memory addresses to future readers, thus avoiding L1 misses and the following indirection to the directory. In [16] a cache coherence protocol is proposed which forms a hybrid between a directory and snooping protocol. Coherence messages are forwarded to the predicted sharers of a block (destination-set) and the home node. The home node holds a directory structure which compares the predicted destination set with the actual sharers. While these proposals decrease the latency of memory requests by avoiding unnecessary network transactions, they do not speedup the messages that still need to traverse the NoC.

Other prediction schemes make decisions based upon events in the network. In [17] prediction is used to reduce the setup latency of a hybrid optical circuit/electrical mesh network by using channel prediction in the electrical routers in combination with lookahead routing. In [18], flow control is acheived by predicting congestion in the network and hence controling the injection rate.

3. SYSTEM PARAMETERS AND ASSUMPTIONS

The system we assume for all the results presented in this paper (see Figure 1(b)) consists of 32 processor tiles. Each tile contains an in-order x86 processing core, a private L1 cache (16 kB for instructions,16 kB for data), part of the shared L2 cache (1 MB in total) and part of the directory. The MESI cache coherence protocol is used to keep the physically distributed memory coherent. Coherence messages of 8B for control messages and 72B for data messages (8B + 64B cacheline) are used. Trace files, containing all the coherence messages travelling the network, were generated using the cycle accurate, full system simulator gem5 [19] which is able to boot Linux and run the PARSEC benchmark suite [20]. This benchmark suite contains a collection of financial, animation, routing, compression, server search and online clustering algorithms which provide a realistic workload for a CMP. To remove the effect of the network from the traces, ideal contention free interconnects were implemented in the simulation.

As the work in this paper attempts to reduce the arbitration latency by prediction techniques, we use a central photonic switch with one optical port per tile, thus avoiding the more complex photonic network architectures of arbitration-free networks such as SWMR. This optical crossbar consists of micro-ring resonators which can be activated to switch the light at a waveguide crossing or turned off to allow the light to continue its undeviated path [21]. As shown in Figure 3, two

DOI: 10.1002/cpe

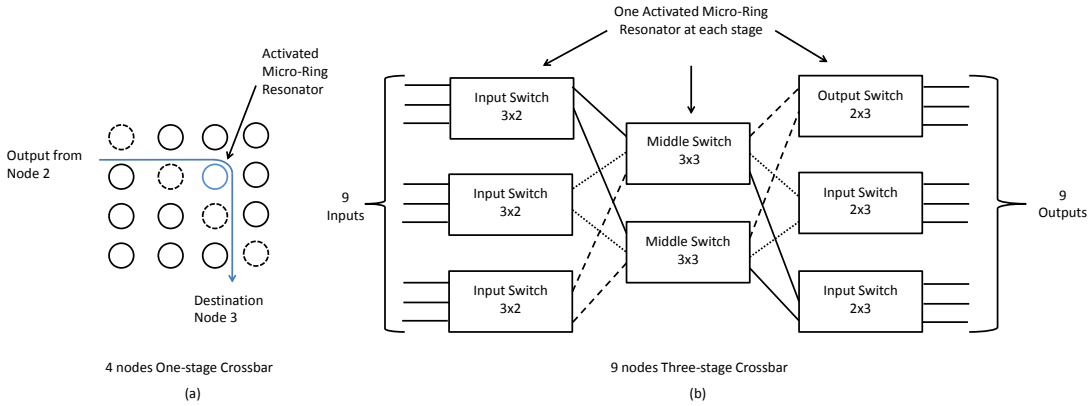

Figure 3. Examples of optical switches used in this work (a) one-stage 4-port crossbar (b) 9-port Clos switch consisting of smaller crossbar elements

optical switches are considered: a one-stage crossbar and a three-stage Clos switch. The crossbar requires only one ring activated to setup an optical path, whist the Clos switch requires activation of three rings. However, the one-stage crossbar uses $N(N-1)$ micro-ring resonators while the three-stage Clos switch, for our system, requires $3N(\sqrt{N}-1)$ for a N-tile system. The ring resonators are sized for switching a wavelength striped message of 16 wavelengths of 10 Gb/s per wavelength to reduce serialisation latency. This bandwidth was previously shown to be optimal in full system gem5 simulations of PARSEC benchmarks [22].

For a single network-on-chip with a clock frequency of 2 GHz and a die size of 400 mm^2, the worst case optical time of flight between any ports and the central switch over silicon waveguides with $n_{eff} = 4.2$ is less than one clock cycle. Including serialisation time and other circuit delays, we can conservatively assume a maximum of 2 clock cycles for communication between tiles and switch. However, for multiple chip systems on a single PCB, such as that shown in Figure 1b, with a maximum distance between tile and switch of 0.5m over polymer waveguides or optical fibre with $n_{eff} = 1.5$, the head latency could be up to 7 clock cycles.

As discussed in section 2, considerable research has been conducted into photonic network schemes which avoid arbitration. In order to demonstrate the latency and power consumption benefits of our approach, we use a SWMR network [4] for comparison as shown in Figure 4. There is no arbitration required for this network. However, given the $N(N-1)$ receivers required, it is expensive in terms of optical component count to reduce serialisation latency by employing wavelength striping in this case. We use SWMR networks with 1–4 transmission wavelengths per node in our latency and power comparisons.

4. ARBITRATION PER MEMORY TRANSACTION

The work presented in this paper exploits the fact that messages in a shared memory network are generated in sequences initiated by transitions in the cache coherence protocol in response to memory requests from the cores. Figure 5 shows some examples of coherence message sequences which commonly occur in the MESI protocol, showing examples of memory transactions which: (a) involve three or more tiles which require additional optical paths; (b, c) involve just two tiles and hence can be served by a bidirectional optical path and (d) involve communication with main memory. We use knowledge of these transactions to efficiently set up optical paths (or circuits) between tiles and main memory.

Figure 6 shows the variation in occurrence and average latency of coherence message sequences of different lengths. The length of a message sequence is defined as the number of messages (transmitted on the NoC) needed to complete a coherence transaction. The average latency

DOI: 10.1002/cpe

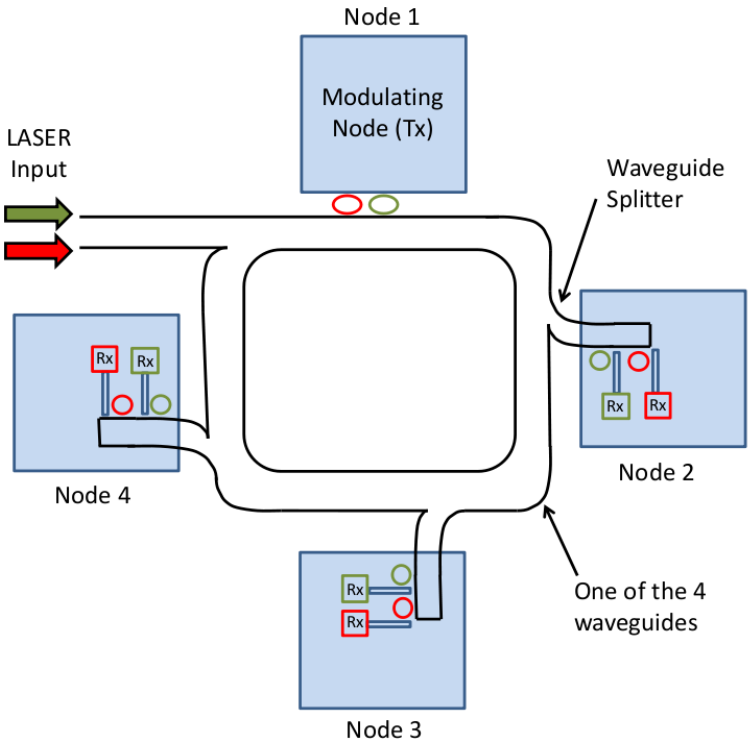

Figure 4. SWMR layout showing one of the 4 waveguides in a 4 node system with two transmission wavelengths per node

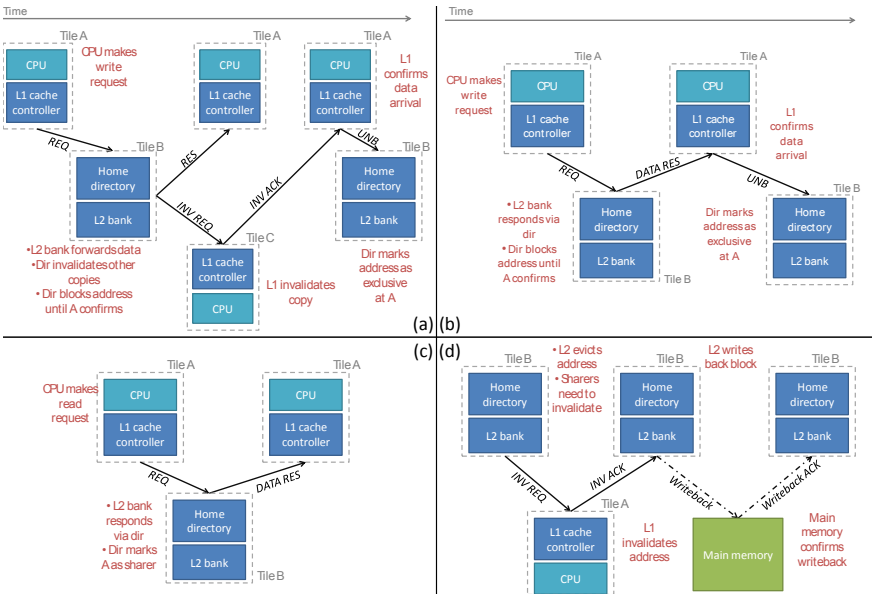

Figure 5. Examples of common coherence message sequences in the MESI protocol (a) CPU A requests *store* access to a memory address cached in other L1 caches (b) CPU A requests *store* access to a memory address cached only in the L2 (c) CPU A requests *load* access to a memory address (d) L2 evicts a block which is cached in a private L1

(Figure 6(a)) depends both on the sequence length and whether or not main memory is involved. The occurrence of each sequence (Figure 6(b)) differs depending on the communication requirements of individual benchmarks. Figure 6(c) shows the resulting weighed latency. The latencies of sequences

DOI: 10.1002/cpe

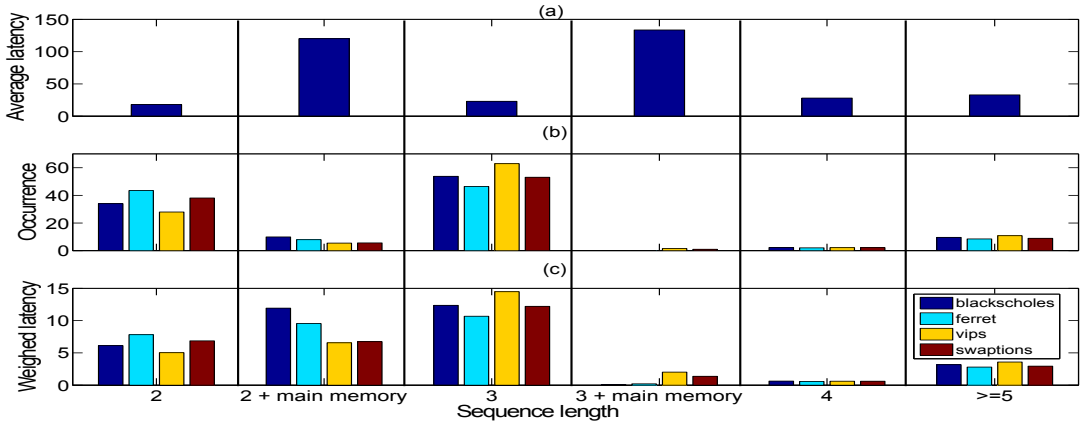

Figure 6. (a) Average latency, (b) probability of occurrence and (c) weighed latency (average latency×occurrence) for memory transactions. Latencies are in clock cycles.

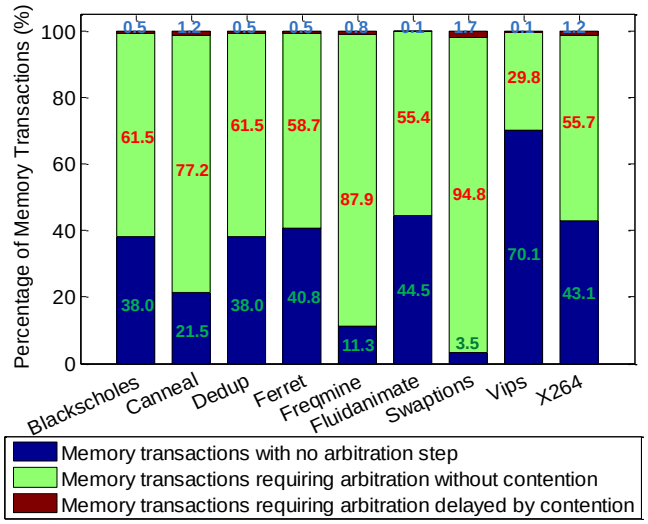

Figure 7. Arbitration outcomes for PARSEC benchmarks using the arbitration per memory transaction algorithm

consisting of 5 or more messages might be longer than pictured as these sequences are often coherence transactions involving the invalidation of memory addresses shared by multiple caches making the latency dependent on the number of sharers. Figure 6 shows the lower bound where there is only one other sharer in the system. This figure shows the performance can be optimised by either focusing on the most common sequences (with a length of 2 or 3 and no main memory access) or the sequences with longest latencies (sequences involving main memory accesses or consisting of 5 or more messages).

In the discussion of arbitration latency in section 1 we assumed that each message goes through the request, arbitration and grant process. This would be appropriate for random and independent messages without temporal or spatial locality. However, in a shared memory coherence network messages are communicated based on the cache coherence protocol finite state machine as shown in Figure 5. For transactions involving just two cores such as the examples in Figure 5 (b), (c) and (d), the memory transaction can be completed by setting up a bidirectional optical path between the two cores [23]. All the information required to set up these bidirectional paths is available in the Miss Status Holding Register (MSHR) at the source port.

DOI: 10.1002/cpe

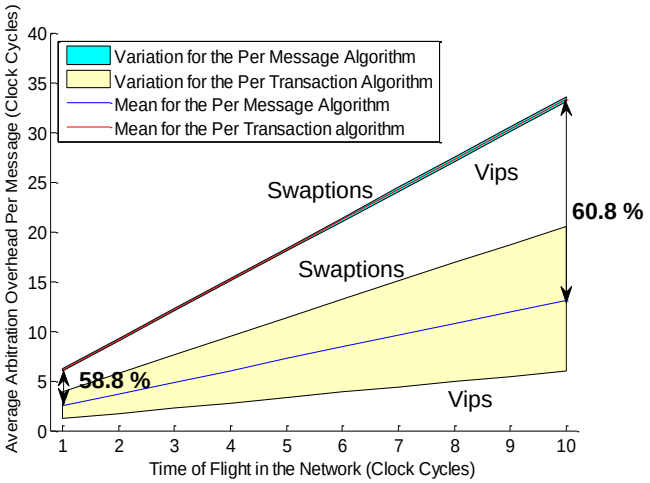

Figure 8. Variation of the average arbitration overhead per message with time of flight between tile and switch

Knowing in advance how many messages would be involved in a memory transaction, arbitration to establish a bidirectional path only needs to be performed once for the initial request message, leaving the optical circuits open for subsequent messages in the same memory transaction, reducing latency compared with arbitrating for every message. There are however two drawbacks of keeping circuits open for extended periods of time: (1) additional energy is consumed in the switches and (2) other communications targeted at either of tiles involved in a memory transaction must wait until the transaction is complete (whereas if arbitration is taking place per message, the communications can be interleaved). Section 6 will address the issue of the energy consumption of the system. On the question of latency, we showed in [23] that only a very small proportion of individual messages have increased latency due to circuit contention for a NoC. This is because the PARSEC benchmarks, as with other applications, load the network very lightly [24]. It therefore makes sense to hold open circuits for the current memory transaction to complete. In addition, using the principles of temporal and spacial locality, it is likely that subsequent memory transactions will involve the same two cores, so optical circuits can be held open unless another request is made to either of the cores. Figure 7 shows the variation of the percentages of memory transactions which benefit from the circuit remaining open for the different benchmarks considered together with the percentage experiencing contention. In the case of `vips`, 70% of memory transactions benefit from no arbitration overhead latency. The proportion of messages experiencing contention is $<2\%$ for all benchmarks.

These simulations were done with 32 cores CMP. Upon scaling to a network with a larger number of cores, it is still expected that the traffic would remain low. Hence, using a 3-stage crossbar, the number of optical paths available would ensure that contention is kept low enough, such that the per transaction scheme is still beneficial. In the case of multiple socket shared memory systems, shown in Figure 1b, circuits must be held open for longer to accommodate longer time of flight latencies, thus increasing the contention probability as well as the head latency of request, grant and data messages. The simulation results in Figure 8 show that per transaction arbitration has greater latency benefits in absolute terms for networks with a longer time of flight between core and switch, although the percentage decrease in the average arbitration overhead per message remains constant at around 60%. In addition, it can be observed from the nearly linear relationship between the arbitration overhead and time of flight that there is no significant increase in contention due to the increased memory transaction times.

DOI: 10.1002/cpe

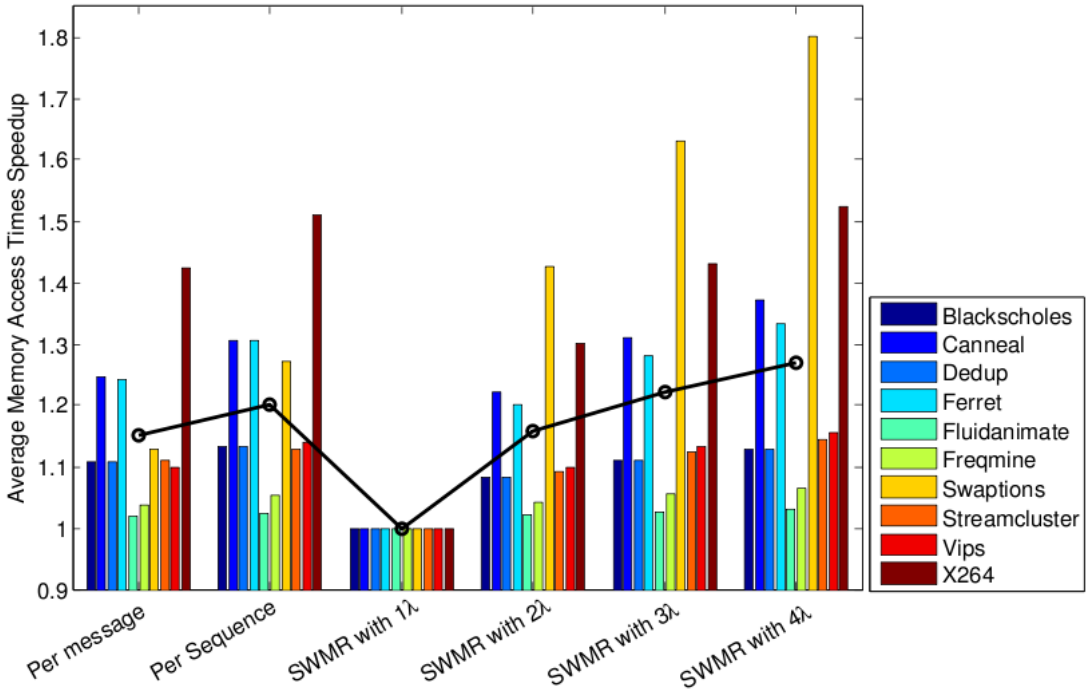

Figure 9. Average Memory Access Time speedup for the PARSEC benchmarks referenced to SWMR with 1 wavelength per core. For SWMR, the numbers refer to the number of wavelengths in the network per core. The black points and line represents the mean speedup averaged across all benchmarks for each of the networks.

5. LATENCY COMPARISONS WITH ARBITRATION-FREE NETWORKS

We will now compare the performance of the photonic switch networks with arbitration per memory transaction with a SWMR network which avoids arbitration. The measure of performance in this work is the Average Memory Access Time (AMAT) which is a good indicator of system performance for in-order cores [25]:

$$AMAT = Hit\ time + (Miss\ Rate \times Miss\ Penalty)$$

The hit time is defined as the time taken to satisfy a memory request by a core if the requested memory address is available in the L1. In the case that the block is either not present or the L1 cache does not have the correct permissions, a miss occurs. The miss penalty is defined as the time taken to correct this situation by either fetching the block or obtaining the permissions needed. All the above parameters were extracted directly from gem5 simulations or from network simulations with the gem5 trace files as input.

Figure 9 shows the AMAT speedup for the PARSEC benchmarks comparing the crossbar networks with SWMR networks with 1–4 wavelengths per link per core. The baseline in the calculation of the speedup is taken to be SWMR with only one wavelength per core. It can be seen that the different schemes favour different benchmarks since the per transaction arbitration scheme has better results for benchmarks with greater numbers of messages per transaction or with high spacial or temporal locality. The SWMR scheme provides better results in benchmarks where the ratio of control messages to data messages is greater since the overall latency is dependent mainly on serialisation latency. The results show that the per transaction arbitration scheme performs on average 20% better varying between 2.4% for `Fluidanimate` and up to 51% for `X264` as compared to the SWMR scheme with only one wavelength. On the other hand, the per transaction arbitration scheme performs on average 4% better than the per message arbitration scheme varying

Prepared using cpeauth.cls
DOI: 10.1002/cpe

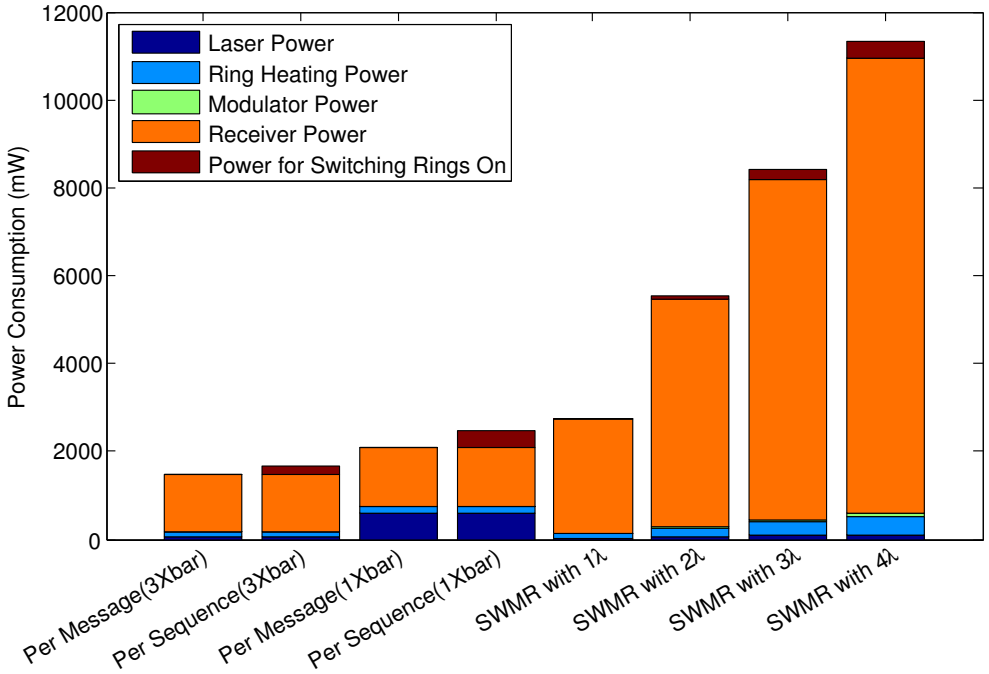

Figure 10. Detailed power consumption variation for the schemes considered

between 0.5% for `Fluidanimate` and up to 13% for `Streamcluster`. This data shows that, on average, the two wavelengths per core SWMR behaves better than per message arbitration and three wavelengths per core SWMR would be enough to perform better than the per sequence arbitration scheme. However, increasing the number of wavelengths has an adverse effect on the energy consumption of this optical system as discussed in the following section.

6. COMPARISONS OF POWER CONSUMPTION

In this section, the power consumed in the optical plane is considered. It can broadly be separated into five categories: the optical power of the laser, the power required for heating the micro-ring resonators for wavelength control, the modulation drive power, receiver power and switching power.

In order to calculate the laser power, the physical layout of the cores was considered and the worst case of power loss was taken together with the receiver sensitivity to calculate the minimum power of the laser that could enable any communication to take place. All the parameters used are provided in Table I.

The receiver, modulator and heating per ring values are taken from experimental data. The power required to resonate a micro-ring resonator depends on its size and is based on the layout, the total power consumption for switching the rings is considered. It is also assumed that power gating is available for modulating the required wavelengths and for switching the rings in the optical path. However, considering that in the per transaction arbitration scheme, one of the improvements has been to keep the optical circuit active even after the memory transaction has been completed, power gating will not be available there. As mentioned in 3, two arrangements of crossbars are considered since they differ in the number of micro-ring resonators put into resonance as well as the total number used.

It can be seen from Figure 10 that the power consumption of the SWMR networks is larger than the other schemes and that the main contribution to the power dissipation comes from the receivers. This arises primarily because the number of through losses in the serpentine path is increasing due to the need to collect information from more wavelengths. This problem would be worse in

DOI: 10.1002/cpe

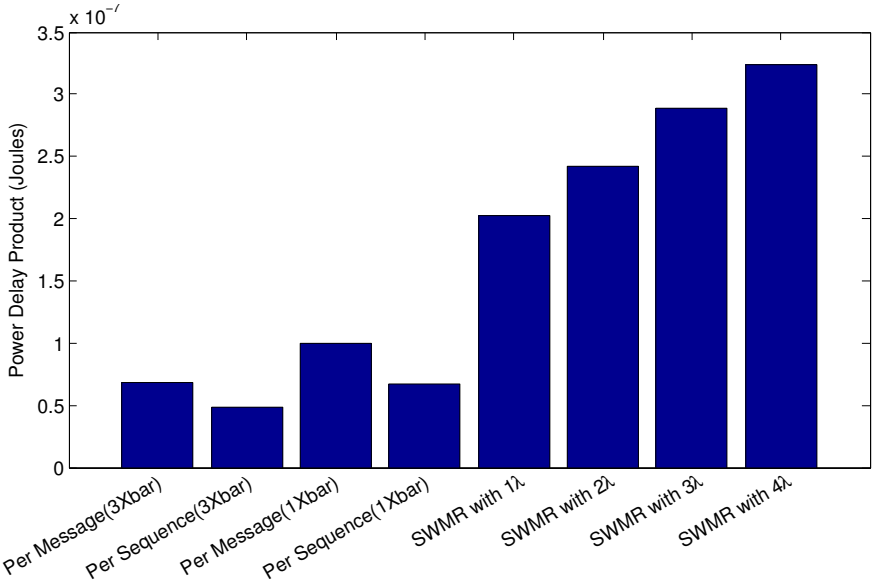

Figure 11. Variation of the power delay product with the different schemes being investigated. The left column considers the power consumption with the receiver always in the on state and the right column considers the power consumption for powergated receivers.

Table I. Parameters involved in Optical Power Calculations

Property	**Parameter**
Modulator Loss	4 dB [26]
Propagation Loss	1.3 dB/cm [27]
Waveguide Crossing Loss	0.04 dB [28]
Micro-Ring Resonator through loss	0.33 dB [21]
Micro-Ring Resonator drop loss	1.6 dB [21]
Splitter Losses	0.015 dB [29]
Receiver Sensitivity	-18 dBm [30]
Power Consumption of Ring per Circumference	1.3 W/m [21]
Effective Silicon Waveguide Refractive Index	4.2
Receiver Power	2.6 mW [31]
Modulator Power	0.66 mW [31]
Ring Heaters	0.1 mW [32]

a Multiple Writer Multiple Reader (MWMR) system where there would be just as many receivers and an increased number of micro-rings due to the need for more modulators. Hence, there will be more through losses from micro-ring resonators on the path both for modulation and for reception. On the other hand, we would get a better power consumption from a Multiple Writer Single Reader (MWSR) scheme where only one receiver would be present per core, but doing that would also introduce the arbitration process again, losing latency savings. In the calculations, it was also assumed that there would be a number of waveguides equal to the number of cores in the system. The modulator power considered here is nearly insignificant considering that it is power gated. Considering that the receivers have no awareness of when they would be receiving data, it would not be practical to introduce powergating yet. If powergating the receivers were to be introduced similar to the reservation scheme, as proposed in [4], there should be a warm-up period allocated for the receivers being turned on which would decrease the latency savings, together with increasing the traffic in the control plane. However, this analysis also does not take into account the additional receiver FIFOs required to implement the SWMR networks. Concerning power consumption, this

DOI: 10.1002/cpe

would drastically decrease the optical power required by all the investigated schemes since most of the power is being used in the receivers. The per transaction scheme will still show better power consumption considering that SWMR consumes more laser power.

The three stage crossbar shows better power performance than the one stage crossbar since the number of micro-ring resonators is reduced and also there are reduced waveguide crossings. On the other hand, it can be seen that the per transaction arbitration requires more power than the per message arbitration by 18.7 % and 12.4 % in the one stage crossbar and three stage crossbar respectively for receivers always in the on state. This is due to the fact that the micro-ring resonators are always switched on in the per sequence arbitration scheme. In order to determine which system offers the best compromise between the power consumed and the latency of the memory transactions a power delay product (PDP) metric was calculated:

$$Power\ Delay\ Product = Power\ Consumption \times Mean\ Time\ To\ Complete\ A\ Memory\ Transaction$$

The system with the best energy efficiency has the lowest value of PDP. Figure 11 shows that the three stage crossbar used with the per transaction arbitration scheme provides best results.

7. PHOTONIC NETWORK SETUP THROUGH CACHE MISS PREDICTION

The technique described in the previous sections reduces arbitration latency by setting up bidirectional circuits based on a knowledge of the communications produced by the MESI protocol. However, arbitration for new circuits cannot begin until the request has reached the MSHR and more complex transactions (such as those shown in Figure 5(a) and (d)) will require two or more arbitrations. Across the PARSEC traces studied, 16% of all transactions take place between three or more tiles and transactions involving main memory access, although relatively rare, have a high impact on AMAT. In this section, we explore the possibility of further latency savings through cache miss prediction, allowing speculative arbitration for optical circuits to start before the message is ready.

A first step towards the prediction of exact message sequences based upon information from the cache controllers is the prediction of coherence requests leaving the L1 cache using a local predictor operated in parallel with the cache access. If a coherence message is predicted, a path request will be sent out to the central arbiter to setup the required optical paths before the actual coherence message reaches the network interface. While this scheme is very easy to implement, the predictor should be faster than the cache access in order to reduce latency. As L1 caches are geared towards low latency operation (1–3 clock cycles), we believe there is limited potential for latency saving using a predictor solely to setup optical paths for messages leaving L1 caches.

Because of the latency constraints imposed on the L1 predictor, we wish to be able to predict L1 requests that cannot be serviced immediately by the L2 bank associated with the directory (such as main memory access (Figure 5(d)) or write requests to shared memory address (Figure 5(a)) in order to setup circuits needed to complete the memory transaction. The predictor proposed in this work predicts the start of such a sequence. In the case of a main memory transaction, this information will allow rapid setup of an optical circuit from the directory to the main memory bank. In other cases such as the write request to a shared memory address, information about the sharers is needed in order to setup optical circuits. Further work is investigating sharer prediction which would also allow setting up optical circuits for these types of memory transactions. While the idea behind the L2 miss predictor is the same as in the L1 predictor case, the actual implementation is more complex as the feedback needed to update the predictor will need to come from a different node. One possibility is piggybacking the outcome of the prediction on coherence messages traveling to the home node of the predictor.

DOI: 10.1002/cpe

7.1. Performing the cache miss prediction

The prediction is made based upon the memory address requested (address based prediction), operating in parallel with the L1 cache access. The predictor used in this work (Figure 12) consists of a lookup table (LUT) with N entries and some peripheral circuits to convert (part of) the address into a key for the LUT, update the entries in the LUT and send out circuit setup requests based upon the prediction that was just made. The lookup table is accessed by hashing a proportion of the requested memory address. Every entry in the lookup table consists of the state of a 2-bit counter, the last prediction, a valid bit and in the case of a set-associative organization a tag. In state 1 and 2, no message will be predicted whereas in state 3 and 4 a message will be predicted. After the prediction the state of the 2-bit counter will be updated based upon the correctness of the prediction. The memory address consists of 7 hexadecimal digits. Using the complete memory address to obtain a key is inefficient as 256M keys would be possible. Sweeping over the *granularity* (number of hexadecimal digits used to obtain the address) shows a higher granularity will result in a lower percentage of messages for which no optical path was setup which comes at the cost of a larger LUT as can be seen in Figure 13 for the `blackscholes` benchmark.

To find the digits in the memory address that carry most of the information, we investigated the effect of the *start digit*. This is the first digit to be included in the address hashing. As Figure 13 shows the various bits in the address do not contain the same information. By carefully choosing the correct start digit and keeping the granularity the same, the missed message rate can be reduced by more than 70%. The three least significant digits of an address (marked in Figure 13 as start digit 5,6 and 7) do not carry a lot of information. This can be explained by the fact the page size is set to 4KB and so these three digits form the page offset.

Although the size of the LUT is quite large, most of the entries are never used: for a granularity of 3 digits and higher less than 30% of the entries are used. This decreases to less than 0.001 % for a granularity of 6. We can reduce the size by changing from the directly mapped setup to a set-associative organization which has a beneficial effect on the misprediction rate as shown in Figure 14. The LUT size of the set-associative predictors was set to 256 entries. For comparison, this is the size of a directly mapped predictor with a granularity of 2. When evicting one address from the LUT, this entry will be reset. The start state of the 2-bit saturating counter is state 4 in which a message will be predicted. By evicting entries from the LUT, the LUT gets slightly biased towards predicting more messages. Increasing the set-associativity will increase the latency of the predictor though as more entries need to be searched. A careful trade-off between the latency and size of the predictor will need to be made.

These results are encouraging: for 4-way associativity, using a granularity of 3 (resulting in a LUT of only 320B) and the start digit set to 4, only 16% of the coherence messages leaving the an L2 bank will not have an optical circuit setup. On the other hand, many circuits (up to 70% in some combinations of granularity and associativity) of setup circuits are not used. As the network load is low, as discussed in Section 4, setting up unused circuits is not necessarily a problem as long as the arbiter can efficiently distinguish between speculative circuits and circuits which are definitely required.

8. CONCLUSIONS

In this paper, we have presented techniques that can significantly reduce the arbitration latency of photonic networks for future shared memory computer systems. Firstly, we have demonstrated that a switch scheduling algorithm which arbitrates on a per memory transaction basis and holds open photonic circuits to exploit temporal and spacial locality can reduce the average arbitration latency overhead by 60% and eliminate arbitration latency altogether for a significant proportion (>70% for `vips`) of memory transactions.

We have compared our proposed scheme with an arbitration-free SWMR network and shown that it would require at least four wavelengths per node to give a better latency performance than our arbitration algorithm. However, such a network has considerably increased optical component count

DOI: 10.1002/cpe

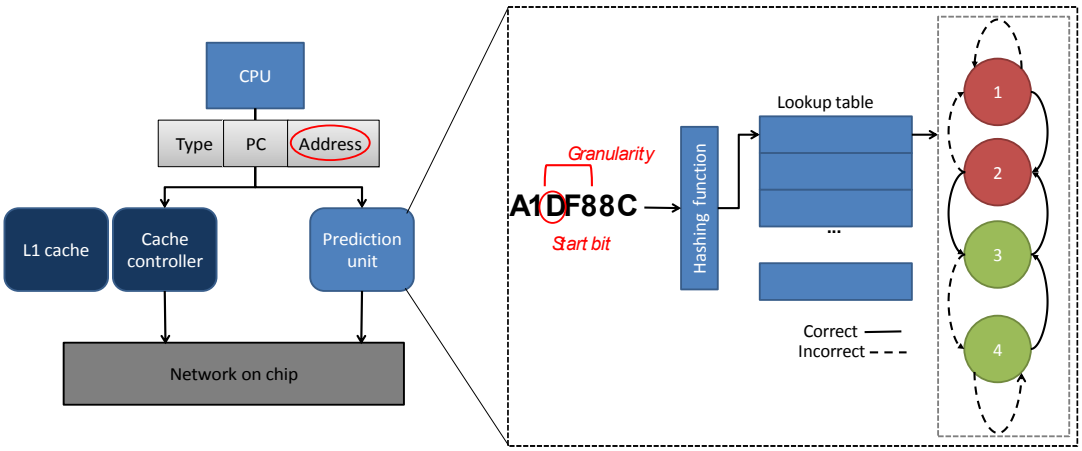

Figure 12. Operation of the prediction unit and interface with the tile.

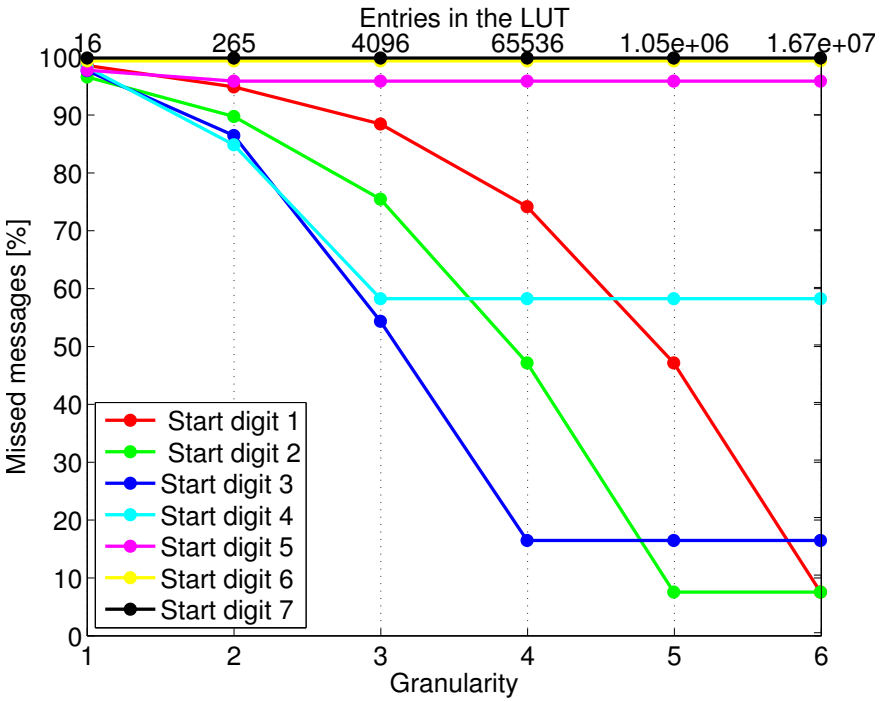

Figure 13. Percentage of coherence messages for which no optical path was predicted.

and hence power consumption compared with a centralised wavelength striped photonic switch. SWMR with four wavelengths consumes nearly seven times the power required for the network using per memory transaction arbitration and a three stage crossbar. It was also found that keeping the optical paths open even after the memory transaction ended, in order to benefit more from spatial and temporal locality, tends to increase the power consumption by 19% and 12% in the one stage crossbar and three stage crossbar respectively. Nevertheless, the power delay product shows that the overall performance is still improved by 33% and 29% in the one stage crossbar and three stage crossbar respectively by arbitrating per memory transaction rather than per message.

Finally, we have also demonstrated that cache miss prediction can be used to predict 86% of photonic circuits to main memory for the `blackscholes` benchmark for further arbitration

DOI: 10.1002/cpe

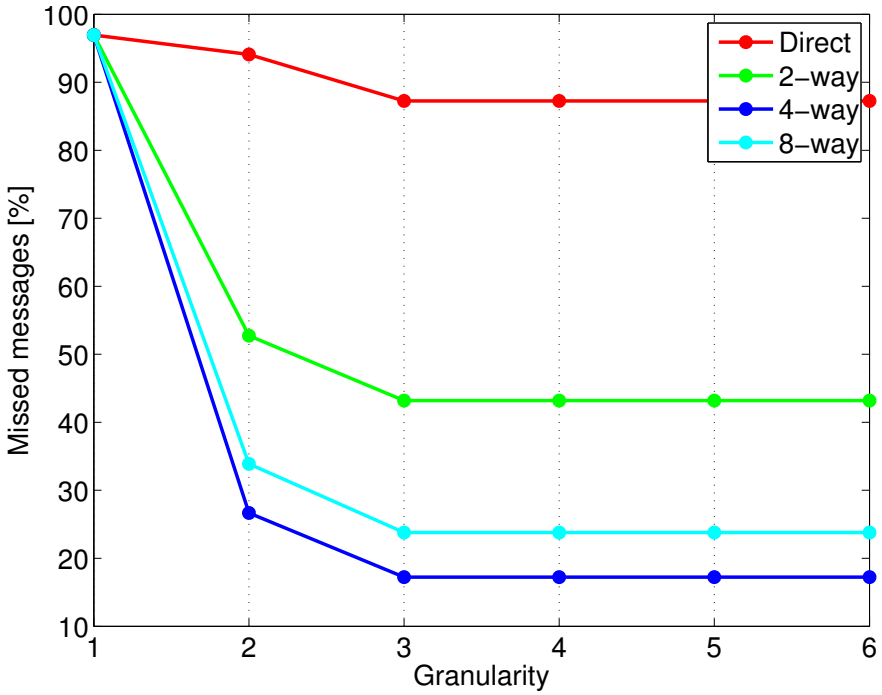

Figure 14. Effect of associativity on predictor performance with start digit = 4.

latency savings. Research is continuing on sharer prediction in order to accelerate setup of circuits for invalidation as well as the optimisation of the L2 miss predictor.

We have shown that arbitration per transaction works effectively for systems spanning multiple chips with longer time of flight between tile and switch. The replacement of separate electonic NoC and off-chip networks with a single photonic network has the potential to reduce both latency and energy consumption in multiple socket servers and could enable efficient larger shared memory systems with increased sockets per card or spanning multiple cards or racks. Nor is there significant power penalty in mutliple chip networks of this kind. Interfaces between on-chip nanophotonic silicon waveguides and larger chip-to-chip polymer waveguides or fiber have been demonstrated with < 0.5 dB loss [33], while polymer and fibre have considably lower transmission losses than silicon waveguides. Employing a separate photonic switch chip as shown in Figure 1b enables a wider range of switching technologies to be considered including semiconductor optical amplifiers (SOA) which can further reduce the processor chip power dissipation [11] while retaining silicon photonic elements for the transmitters and receivers which benefit from tight integration with the processing tiles. However, for multiple chip networks, synchronisation latency (see Figure 2) becomes an important issue as the transmitter and receiver do not share the same clock and the latency savings from the prediction algorithms could be negated by the preamble required for clock recovery at the receiver. Source synchronous wavelength striped photonic links have been demonstrated operating at up to 4 Gb/s [34] and due to the fundamentally lower delay variation in photonic compared with electronic links [35] may also work at higher bit rates.

As well as the latency and energy consumption benefits, the larger shared memory systems with photonic interconnect resulting from this work could promote more efficient programming of emerging applications in big data analysis, media streaming and other large scale data centre operations.

DOI: 10.1002/cpe

9. ACKNOWLEDGMENTS

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) Fellowship grants to Philip Watts (EP/I004157/2) and Timothy Jones (EP/K026399/1).

REFERENCES

1. A. Shacham *et al.*, "Building ultralow-latency interconnection networks using photonic integration," *IEEE Micro*, vol. 27, no. 4, 2007.
2. D. Vantrease *et al.*, "Corona: System implications of emerging nanophotonic technology," in *Int. Symp. on Comput. Archit.*, 2008.
3. A. Krishnamoorthy *et al.*, "Computer systems based on silicon photonic interconnects," *Proc. of the IEEE*, vol. 97, no. 7, 2009.
4. Y. Pan *et al.*, "Firefly: Illuminating future network-on-chip with nanophotonics," in *Int. Symp. on Comput. Archit.*, 2009.
5. G. Hendry *et al.*, "Analysis of photonic networks for a chip multiprocessor using scientific applications," in *Int. Symp. on Networks-on-Chip*, 2009.
6. G. Hendry *et al.*, "Time-division-multiplexed arbitration in silicon nanophotonic networks-on-chip for high-performance chip multiprocessors," *Journal of Parallel and Distributed Comput.*, vol. 71, no. 5, 2011.
7. D. A. B. Miller, "Device Requirements for Optical Interconnects to Silicon Chips," *Proceedings of the IEEE*, vol. 97, no. 7, 2009.
8. J. Shin *et al.*, "A 40 nm 16-core 128-thread SPARC soc processor," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, 2011.
9. D. Wentzlaff and other, "On-Chip Interconnection Architecture of the Tile Processor," *Micro, IEEE*, vol. 27, no. 5, 2007.
10. I. Iliadis and C. Minkenberg, "Performance of a speculative transmission scheme for scheduling-latency reduction," *IEEE/ACM Trans. on Networking*, vol. 16, no. 1, 2008.
11. P. Watts *et al.*, "Energy implications of photonic networks with speculative transmission," *IEEE/OSA Jour. of Opt. Comm. and Netw.*, vol. 4, no. 6, 2012.
12. K. J. Barker *et al.*, "On the Feasibility of Optical Circuit Switching for High Performance Computing Systems," in *Proceedings of the ACM/IEEE Supercomputing Conference*, 2005.
13. P. Watts *et al.*, "Requirements of low power photonic networks for distributed shared memory computers," in *Opt. Fib. Comm. Conf.*, 2011.
14. M. Acacio *et al.*, "Owner prediction for accelerating cache-to-cache transfer misses in a cc-numa architecture," in *Supercomputing, ACM/IEEE 2002 Conference*, 2002.
15. S. Kaxiras and C. Young, "Coherence communication prediction in shared-memory multiprocessors," in *Int. Symp. on High-Performance Computer Architecture*, 2000.
16. M. Martin *et al.*, "Using destination-set prediction to improve the latency/bandwidth tradeoff in shared-memory multiprocessors," in *Int. Symp. on Comput. Archit.*, 2003.
17. C. Adi *et al.*, "An efficient path setup for a photonic network-on-chip," in *Int. Conf. on Networking and Computing*, nov. 2010.
18. U. Ogras and R. Marculescu, "Prediction-based flow control for network-on-chip traffic," in *Design Automation Conference*, 2006.
19. N. Binkert *et al.*, "The gem5 simulator," *SIGARCH Comput. Archit. News*, vol. 39, no. 2, 2011.
20. C. Bienia *et al.*, "The PARSEC benchmark suite: Characterization and architectural implications," tech. rep., Princeton University, 2008.
21. A. Poon *et al.*, "Cascaded Microresonator-Based Matrix Switch for Silicon On-Chip Optical Interconnection," *Proc. of the IEEE*, vol. 97, no. 7, 2009.
22. A. Van Laer, T. Jones, and P. M. Watts, "Full system simulation of optically interconnected chip multiprocessors using gem5," in *Optical Fiber Communication Conference/National Fiber Optic Engineers Conference 2013*, p. OTh1A.2, Optical Society of America, 2013.
23. M. Madarbux, A. Van Laer, and P. Watts, "Low latency scheduling algorithm for shared memory communications over optical networks," in *Symp. on High-Performance Interconnects*, 2013.
24. M. Bhadauria, V. Weaver, and S. McKee, "Understanding PARSEC performance on contemporary CMPs," in *Int. Symp. on Workload Characterization*, 2009.
25. J. L. Hennessy and D. A. Patterson, *Computer Architecture, A Quantitative Approach*. Morgan Kaufmann, 4th ed., 2007.
26. 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.
27. R. Morris, E. Jolley, and A. Kodi, "Extending the performance and energy-efficiency of shared memory multicores with nanophotonic technology," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 25, pp. 83–92, Jan 2014.
28. 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*, p. IM1A.4, Optical Society of America, 2013.
29. P. Wang, G. Brambilla, Y. Semenova, Q. Wu, and G. Farrell, "Design of an extra-low-loss broadband y-branch waveguide splitter based on a tapered mmi structure." 2011.

DOI: 10.1002/cpe

30. C. Li, R. Bai, A. Shafik, E. Tabasy, G. Tang, C. Ma, C.-H. Chen, Z. Peng, M. Fiorentino, P. Chiang, and S. Palermo, "A ring-resonator-based silicon photonics transceiver with bias-based wavelength stabilization and adaptive-power-sensitivity receiver," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2013 IEEE International*, pp. 124–125, Feb 2013.
31. X. Zheng, F. Liu, J. Lexau, D. Patil, G. Li, Y. Luo, H. Thacker, I. Shubin, J. Yao, K. Raj, R. Ho, J. Cunningham, and A. Krishnamoorthy, "Ultra-low power arrayed cmos silicon photonic transceivers for an 80 gbps wdm optical link," in *Optical Fiber Communication Conference and Exposition (OFC/NFOEC), 2011 and the National Fiber Optic Engineers Conference*, pp. 1–3, March 2011.
32. J. Chan, G. Hendry, K. Bergman, and L. Carloni, "Physical-layer modeling and system-level design of chip-scale photonic interconnection networks," *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, vol. 30, pp. 1507–1520, Oct 2011.
33. V. R. Almeida, R. R. Panepucci, and M. Lipson, "Nanotaper for compact mode conversion," *Optics Letters*, vol. 28, no. 15, 2003.
34. C. Gray *et al.*, "Test electronics for a multi-gbps optical packet switching network," in *Electronics Packaging Technology Conference*, dec. 2006.
35. G. Q. Chen *et al.*, "Predictions of CMOS compatible on-chip optical interconnect," in *Integration, the VLSI journal*, vol. 40, 2007.

DOI: 10.1002/cpe