Energy Efficient And Low Latency Interconnection Network For Multicast Invalidates In Shared Memory Systems

Muhammad Ridwan Madarbux
Optical Networks Group
Electronic and Electrical
Engineering Department
University College London
m.madarbux@ucl.ac.uk

Anouk Van Laer
Optical Networks Group
Electronic and Electrical
Engineering Department
University College London
anouk.vanlaer@ucl.ac.uk

Philip M. Watts
Optical Networks Group
Electronic and Electrical
Engineering Department
University College London
philip.watts@ucl.ac.uk

Timothy M. Jones
Computer Architecture Group
Computer Laboratory
University of Cambridge
timothy.jones@cl.cam.ac.uk

ABSTRACT

Optical network-on-chip (NoC) are being investigated to reduce the latency and power consumption of networks for multicore processors. Our previous work has shown that switched optical networks can achieve lower latency for a given power consumption and component count in shared memory processors compared with arbitration-free networks such as single writer multiple reader. We have also shown the advantage of leaving optical circuits open after being generated to capture multiple memory transactions. However invalidation processes, where numerous cores are sharing a memory block, need to establish a large number of very short lived circuits and this increases the average message latency and overall on-chip contention.

In this paper, a low power broadcast architecture is proposed which deals specifically with multicast messages. Separating multicast messages from unicast ones shows an improvement in average arbitration latency of up to 88.2% for the Vips benchmark while the Swaptions benchmark shows the highest improvement in average memory access time (up to 21.1%). Vips also sees an increase of 147% in the average number of messages passing through an open optical circuit. Obtaining these advantages requires an additional broadcast network which consumes only 66.1mW power.

CCS Concepts

•Networks → Network on chip;

1. INTRODUCTION

Multicore processes and networks-on-chip (NoC) have been introduced to increase performance using the continually increasing transistor counts offered by Moore's law while reducing design complexity and power consumption.

However, this progress is limited by the latency and power consumption properties of the electrical wires and network elements connecting the cores. Hence, optical interconnection networks are being investigated as an alternative with lower crosstalk, end-to-end latency and power consumption together with a much higher bandwidth for message transmission.

Because of the absence of a viable optical buffer, switched optical networks require an initial arbitration stage for each new connection and, considering that the messages to be sent in shared memory networks are quite small (typically 8B for a control message and 72B for a data message), this represents a significant overhead.

There are several means of avoiding this initial arbitration overhead. One is to eliminate the arbitration process entirely by having dedicated waveguides/wavelengths for each source/destination pair such as in the Single Writer Multiple Reader (SWMR) scheme [13]. However, this scheme requires extensive resources both in terms of optical components and power consumption. Another method is to predict the need for a circuit between two cores and opening it before the request for the circuit is even sent [11, 15]. The work discussed in this paper focuses on the network described in [11] whcih uses a central optical crossbar and in which bidirectional optical circuits are opened for the full duration of a memory transaction and are kept open unless another circuit is required. This is a simple control mechanism which does not require much additional logic in the arbiter and previous work has shown that, due to temporal and spatial locality in memory accesses, it provides a good performance with a large percentage the messages passing through circuits already opened beforehand. However, one issue is that some memory transactions, namely invalidations with a large number of sharers, require a number of very short lived circuits to be generated which disrupts the overall performance of this system. Ideally efficient handling of invalidations requires broadcasts. Therefore, this paper proposes an architecture which is capable of handling normal memory transactions and these invalidations on separate networks without drastically increasing the optical power consumption of the system. Section 2 describes the baseline network and viable broadcast networks. Section 3 discusses the different parameters involved in the simulations while section 4 shows the results and discussion of these simulations. Finally, section 5

AISTECS 2016, January 18 2016, Prague, Czech Republic
© 2016 ACM. ISBN 978-1-4503-4084-7/16/01...$15.00
DOI: http://dx.doi.org/10.1145/2857058.2857065

concludes the paper.

2. OPTICAL NETWORK-ON-CHIPS INTERCONNECTION NETWORKS

Figure 1 shows the baseline configuration considered in this paper. It consists of tiled node structures consisting of a processing core, a private L1 cache, a slice of the shared L2 cache, a memory controller to interface with the main memory and a network interface to send messages onto the network. All tiles have point to point connections (which could be optical or electrical) to the central arbiter which is responsible for establishing optical paths in the central optical crossbar. Messages are exchanged among the tiles optically across that crossbar. In order to establish a specific optical circuit, the source node needs to send an electrical path request message to the arbiter. If the destination node is not communicating with any other nodes at that particular time, the arbiter issues a path grant which is also sent back electrically to the source node. After a serialisation stage, the message can now be sent optically to the destination node.

Communications among the tiles are series of control and data messages generated by the finite state machine of the cache coherence protocol. Most of these memory transactions involve only two tiles. Consequently, for such transactions, establishing a bidirectional optical circuit between the original source and destination tiles can accommodate all the messages [10, 11].

However, there is also the issue of invalidations of memory blocks involving many sharers. This is illustrated by Figure 2. The transaction starts with the node requiring exclusive access to a memory block sending a request message to the node containing the L2 slice and directory of the memory block. The latter node sends a response message back saying that it will have the exclusive access only after this block has been invalidated in all of the other sharer nodes which have the memory block in their private L1 cache. All the sharers then receive an invalidation unblock message from the directory node. These are multicast messages and in this type of circuit switched optical network, it involves creating a large number of circuits consecutively to send only a single control message. Considering that all control messages for our system are just 8B and that both serialisation and transmission of the message take one clock cycle each, the arbitration of these individual circuits represent a significant overhead which reduces the overall system performance. Furthermore, each of the tiles receiving the invalidation message needs to acknowledge the process by sending another control response message to the tile which had initially asked for exclusive access.

Figure 3 shows the occurrence of such multicast messages in the eight PARSEC benchmarks running on a 16 core system. Unicast messages or two simultaneous messages which are more numerous, have not been represented here. The number of multicast messages amounts up to 4.4% of total on-chip traffic for the freqmine benchmark. The graph can also be seen to have some components with greater than 16 simultaneous messages. This represents the overlapping of several memory transactions and because of its statistical nature, these are relatively few in number. In spite of the low percentage of multicast messages, the number of single message circuits that they generate can adversely affect the average number of messages passing per any given optical circuit. One potential solution is to have a separate broadcast network which can manage these invalidations.

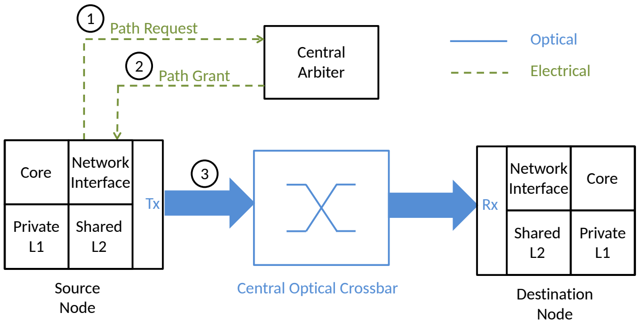


Figure 1: Components in a Chip Multiprocessor (CMP)

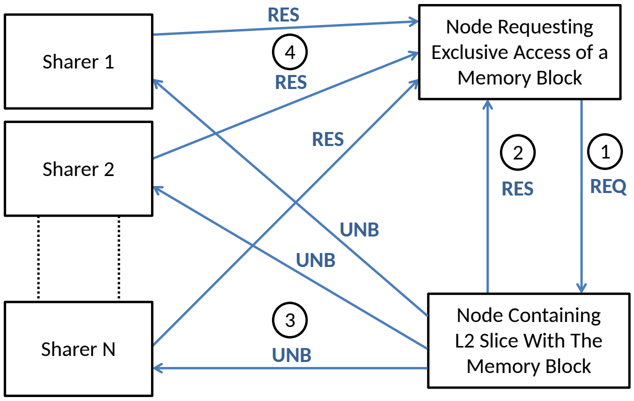


Figure 2: Sequence Of Messages Exchanged In An Invalidation Involving Numerous Sharers

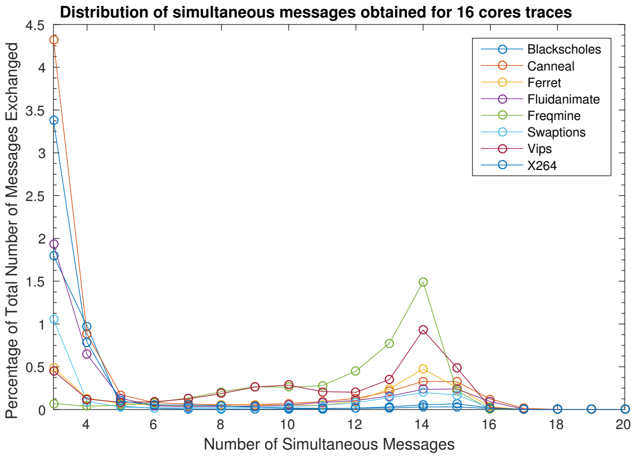


Figure 3: Number Of Simultaneous Messages Sent In A 16 Cores NoC System For The PARSEC Benchmarks

2.1 Broadcast Networks

Broadcast networks have been proposed before as a solution to such traffic [6, 17, 7]. However, they generally involve a large number of optical components and consume significant optical power considering that they use the same broadcast network to send both unicast and multicast messages. Hence, as stated in [6], their solutions were only viable for cache coherence protocol involving a large percentage of multicast messages. Also, they do not address the issue of the many-to-one messages seen when each of the sharers respond to the invalidations. In the MESI cache coherence protocol, the occurrence of multicast messages is much smaller and

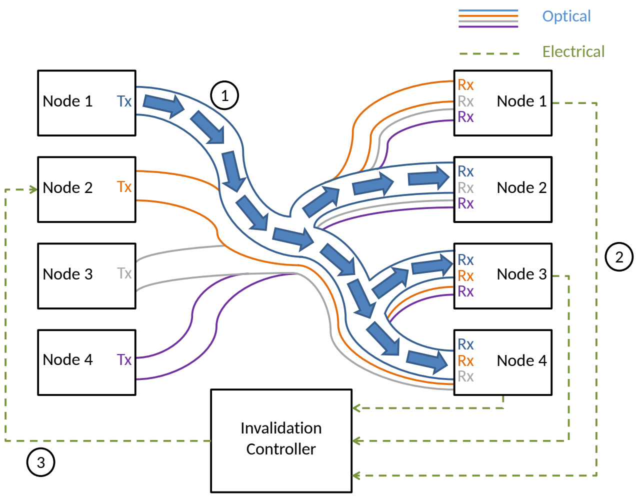


Figure 4: Single Writer Multiple Reader Broadcast Network

therefore, in order to make the system viable, the optical power consumption needs to be very small. Two architectures are proposed in this work. The first one showed in Figure 4 is based on the SWMR topology where each node has its own dedicated waveguide where it can transmit and each of the other nodes need to have a receiver connected to that waveguide. Although having the advantage of no arbitration and the possibility of all the nodes transmitting (both unicast and multicast messages) at the same time, this arrangement still has the drawback of requiring a large number of optical components. The connection of N nodes requires N waveguides, N transmitters and $N(N-1)$ receivers. However, using the combination of an arbiter based crossbar for the unicast network and the SWMR scheme for the broadcast network requires much less optical power and components than using only the SWMR network to send both unicast and multicast messages.

The second proposed solution is a broadcast scheme requiring an arbitration stage as shown in Figure 5. This ensures that there can be at most one node broadcasting at any one time and avoids the situation where N waveguides are required. Such a system requires only 1 waveguide, N transmitter and N receivers to work on N nodes. The only drawback is the 3 clock cycles of arbitration overhead necessary and the fact that only one node can broadcast at any one time. Figure 6 shows the average inter-arrival time between consecutive multicast messages for the eight benchmarks considered and since they are very large, it can be assumed that this arbiter based broadcast scheme would incur little additional contention as compared with the SWMR based broadcast scheme. Furthermore, the arbitration latency is small compared to the latency saved not having to send each of these multicast messages via individual optical circuits.

For both broadcast schemes, the responses from the sharers are sent to the central crossbar arbiter via their dedicated point to point links. Once the confirmation from all the sharers have been received, the arbiter then sends an electronic message to the node requiring exclusive access in order to notify it that the invalidation process has successfully completed. Hence, both multicast and many-to-one messages associated with the invalidation process are not transmitted across the central optical crossbar.

3. METHODOLOGY

The simulations were performed using the full system simulator gem5 [2] together with the PARSEC benchmark [1] in an x86 instruction set architecture. A 16 cores system was simulated with each core having a private 4 way associative 16kB L1 cache and a shared slice of a 32 way associative 32kB L2 cache with their consistency regulated by the MESI cache coherence protocol. The clock speed used was 2GHz with a cache line size of 64B. The optical circuit switching was modelled in a Matlab based trace simulator using the communication trace files generated by gem5. A modulation speed of 25Gbps per wavelength with 8 wavelengths was assumed for the switched crossbar [16] and for both broadcast schemes, only one wavelength at the same modulation speed of 25Gbps was used. The parameters used for the calculation of the optical power consumption are given in Table 1.

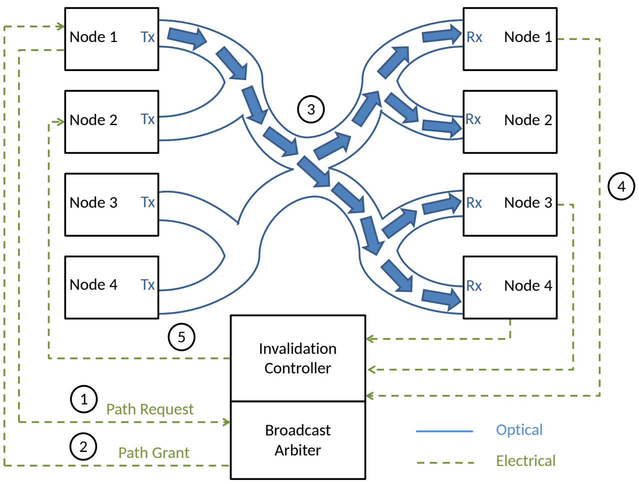


Figure 5: Broadcast Network Requiring Arbitration

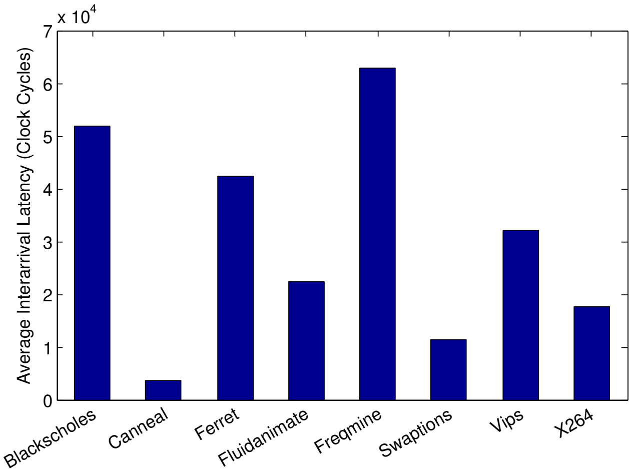


Figure 6: Average Inter-arrival Time Between Consecutive Multicast Messages In The PARSEC Benchmark Suite

4. EVALUATION

The simulations were performed on an interconnection network with the unicast messages passing via the central optical crossbar using bidirectional circuits and the multicast messages via the respective broadcast network. The results were then compared with a baseline in which all the messages pass through the central optical crossbar.

Optical Power Parameters	
Chip Size	400 mm^2 [12]
Propagation Loss	1.3 dB/cm [12]
Splitter Loss	0.015 dB [18]
Crossing Loss	0.05 dB [9]
Modulator Loss	4 dB [5]
Micro Ring Resonator Drop Loss	1.6 dB [14]
Micro Ring Resonator Through Loss	0.33 dB [14]
Receiver Sensitivity	-18 dBm [8]
Modulator Power	0.66 mW [19]
Receiver Power	2.6 mW [19]
Micro Ring Resonator Heating Power	0.1 mW [3]

Table 1: Optical Power Parameters

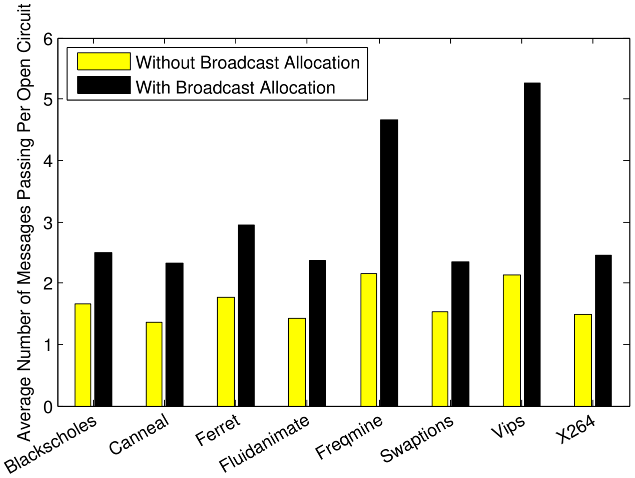
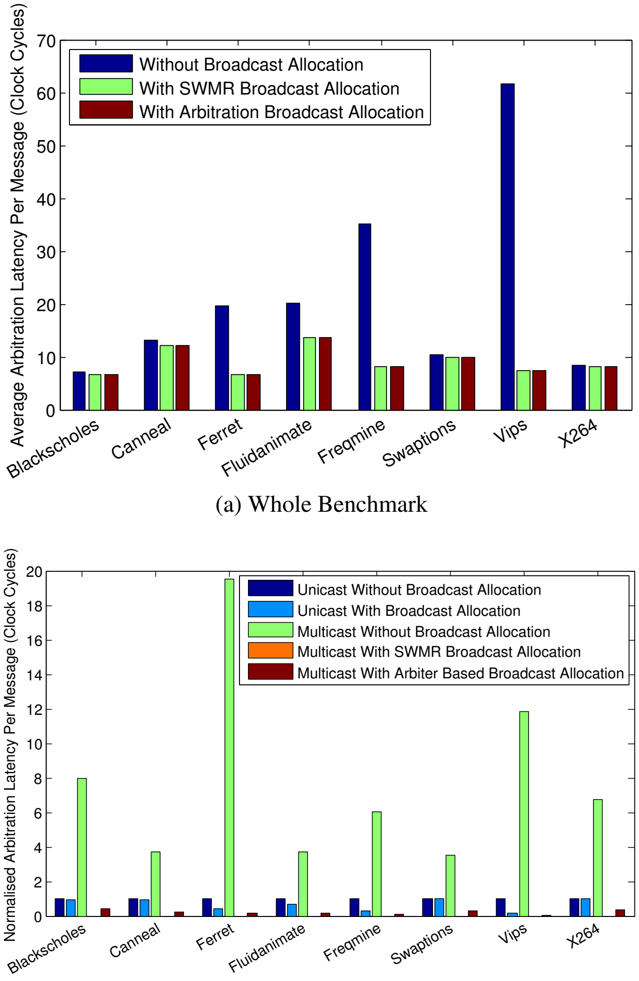


Figure 7: Average Number Of Messages Per Circuit With And Without Allocating For Broadcast Messages

The number of messages that can pass through an optical circuit once it is opened in the central optical crossbar is expected to increase since the occurrence of single message circuits have been greatly reduced. This is shown in Figure 7 where the percentage increase in the number of messages per circuit vary between 48.9% in the Blackscholes benchmark and 147.0% for the Vips benchmark. This shows that the additional broadcast network can exploit the benefits of optical switching by causing a rise in the percentage of messages transmitted in the network without requiring the arbitration step.

The average arbitration latency is calculated by considering the additional latency faced by messages other than the serialisation latency and the time of flight in the network. It consists mostly of latency spent doing the arbitration process and by the messages delayed because a circuit was not readily available. With the reduction of the number of messages doing arbitration and the number of messages delayed in the ingress FIFO, together with the reduction in the average waiting latency, average arbitration latency is expected to be much smaller when dealing separately with the multicast messages. Furthermore, a smaller arbitration latency is expected for the SWMR based broadcast network than the arbiter based broadcast network since the former does not require any arbitration. The results are shown in Figure 8a with the SWMR broadcast allocation scheme shows a reduction of the average head latency ranging from 3.8% for the X264 benchmark to up to 88.1% for the Vips benchmark as compared to the baseline system. Such a large range in variation is observed because the percentage of multicast messages present and their interaction with the unicast traffic are different for each benchmark. However, considering that arbitration is still required for the arbiter based broadcast network, this scheme sees a surplus in average head latency of between 0.1% for the Blackscholes benchmark and 1.6% for the Freqmine benchmark. Nevertheless, with the latency savings obtained by both schemes allocating for broadcasts as compared to the baseline scheme, the difference between the two broadcast schemes can be considered as negligible.



(a) Whole Benchmark

(b) Unicast and Multicast Messages Considered Separately

Figure 8: Average Arbitration Latency Per Message In The Ingress FIFO

To better understand the effect on the individual messages, Figure 8b shows the effect on unicast messages and multicast messages separately. Having a separate broadcast network clearly benefits all the multicast messages since the arbitration latency is reduced to zero or three clock cycles in all cases. The benefit for unicast messages is also significant with decreases in average arbitration latency ranging from 1.8% in the Swaptions benchmark to up to 84.5% in the Vips benchmark being observed. This shows that the presence of the broadcast network does not solely benefit the broadcast messages but impacts also significantly on the unicast messages.

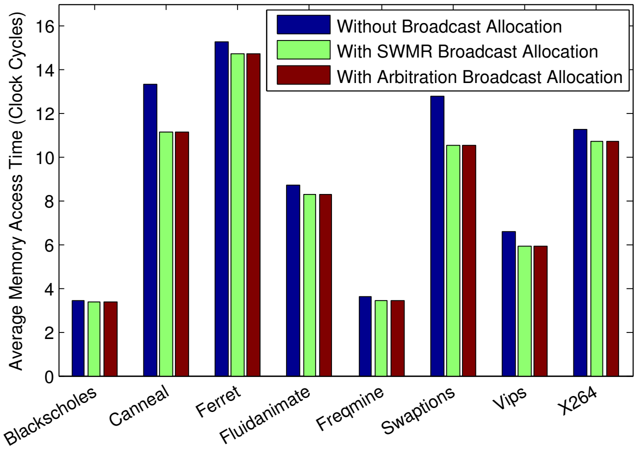


Figure 9: Average Memory Access Times Comparison Of The Different Schemes

Optical Components Involved In Proposed Broadcast Schemes		
Components	**SWMR Broadcast**	**Arbiter Broadcast**
Wavelengths	1	1
Waveguides	16	1
Micro Ring Resonators	16	16
Transmitters	16	16
Receivers	240	16

Table 2: Optical Components Involved In Proposed Broadcast Schemes

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

To better understand the impact of this arbitration latency savings on overall system performance, the Average Memory Access Time (AMAT) can be calculated as shown in the equation above [4]. This variable takes into consideration how often such a decrease in latency is experienced by the running system and what the overall benefit of it is to the benchmark memory processes. The results are shown in Figure 9 where it can be seen that AMAT speedups vary between 1.4% for the Blackscholes benchmark and up to 21.1% for the Swaptions benchmark for the SWMR broadcast allocation scheme as compared to when not allocating for multicast messages. It can also be seen that the although the Vips benchmark has a very good improvement in terms of arbitration latency savings, it has a low miss rate of around 2.0% and this brings down the overall AMAT speedup to only 11.2%. On the other hand, the Canneal benchmark, which had an average arbitration latency saving of only 8.8% experiences a relatively high miss rate of 4.4%. Consequently, this brings up its overall AMAT speedup to 19.5%. Also, very similar to the arbitration latency results, there is little variations between the SWMR broadcast allocation scheme results and the arbiter based broadcast scheme.

Table 2 shows the number of optical components required in order for the broadcast network to function. The SWMR broadcast scheme is shown to require more components as compared to the Arbiter based broadcast scheme. This is also impacted on the total power consumption required by these two systems. A detailed breakdown of the optical power required is shown in Table 3. It is shown that the SWMR broadcast scheme requires around 10.6 times as much power as the Arbiter based broadcast scheme whilst providing no apparent advantage in terms of AMAT or arbitration latency savings. Hence, the Arbiter based broadcast scheme is the best solution which provides a significant performance gain with only a small increment in overall power consumption.

Optical Power Consumption In Proposed Broadcast Schemes		
Power	**SWMR Broadcast**	**Arbiter Broadcast**
Laser Power (mW)	61.5	12.3
Ring Heating Power (mW)	1.6	1.6
Modulator Power (mW)	10.6	10.6
Receiver Power (mW)	624.0	41.6
Total Optical Power (mW)	697.6	66.1

Table 3: Optical Power Consumption In Proposed Broadcast Schemes

5. CONCLUSION

In this paper, we proposed an optical interconnect network which separates unicast and multicast messages. The unicast messages are transmitted via a central optical crossbar while two schemes are proposed for the multicast messages: an arbiter based broadcast network and a SWMR based broadcast network. Using the results obtained from the PARSEC benchmark suite, it is shown that separating the traffic offers increases of up to 147% in the vips benchmark for the average number of messages passing per open circuit and up to 88.2% decrease in the average arbitration latency for the same benchmark. These observations can be translated into AMAT speedups of up to 21.1% (for the swaptions benchmark). Although the two broadcast schemes show very similar performance in terms of AMAT speedup, they differ significantly in terms of optical power consumption, with the SWMR broadcast scheme consuming more than ten times the power required by the arbiter based scheme. Hence, the latter scheme is the best solution with only 66.1mW of additional optical power required.

6. 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).

7. REFERENCES

[1] C. Bienia, S. Kumar, J. P. Singh, and K. Li. The parsec benchmark suite: Characterization and architectural implications. Technical Report TR-811-08, Princeton University, January 2008.

[2] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish, M. D. Hill, and D. A. Wood. The gem5 simulator. *SIGARCH Comput. Archit. News*, 39(2):1–7, Aug. 2011.

[3] 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*, 30(10):1507–1520, Oct 2011.

[4] J. L. Hennessy and D. A. Patterson. *Computer Architecture, Fifth Edition: A Quantitative Approach (The Morgan Kaufmann Series in Computer Architecture and Design)*. Morgan Kaufmann, 2011.

[5] 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*, 38(3):117–128, June 2010.

[6] G. Kurian, J. E. Miller, J. Psota, J. Eastep, J. Liu, J. Michel, L. C. Kimerling, and A. Agarwal. Atac: A 1000-core cache-coherent processor with on-chip optical network. In *Proceedings of the 19th International Conference on Parallel Architectures and Compilation Techniques*, PACT '10, pages 477–488, New York, NY, USA, 2010. ACM.

[7] S. Le Beux, H. Li, I. O'Connor, K. Cheshmi, X. Liu, J. Trajkovic, and G. Nicolescu. Chameleon: Channel efficient optical network-on-chip. In *Design, Automation and Test in Europe Conference and Exhibition (DATE), 2014*, pages 1–6, March 2014.

[8] 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*, pages 124–125, Feb 2013.

[9] L. Liu and Y. Yang. Energy-aware routing in hybrid optical network-on-chip for future multi-processor system-on-chip. *Journal of Parallel and Distributed Computing*, 73(2):189 – 197, 2013.

[10] M. R. Madarbux, A. Van Laer, and P. M. Watts. Low latency scheduling algorithm for shared memory communications over optical networks. In *High-Performance Interconnects (HOTI), 2013 IEEE 21st Annual Symposium on*, pages 83–86, Aug 2013.

[11] M. R. Madarbux, A. Van Laer, P. M. Watts, and T. M. Jones. Towards zero latency photonic switching in shared memory networks. *Concurrency and Computation: Practice and Experience*, 26(15):2551–2566, 2014.

[12] 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*, 25(1):83–92, Jan 2014.

[13] Y. Pan, P. Kumar, J. Kim, G. Memik, Y. Zhang, and A. Choudhary. Firefly: Illuminating future network-on-chip with nanophotonics. In *Proceedings of the 36th Annual International Symposium on Computer Architecture*, ISCA '09, pages 429–440, New York, NY, USA, 2009. ACM.

[14] A. Poon, X. Luo, F. Xu, and H. Chen. Cascaded microresonator-based matrix switch for silicon on-chip optical interconnection. *Proceedings of the IEEE*, 97(7):1216–1238, 2009.

[15] A. Van Laer, C. Ellawala, M. R. Madarbux, P. M. Watts, and T. M. Jones. Coherence based message prediction for optically interconnected chip multiprocessors. In *Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition*, DATE '15, pages 613–616, San Jose, CA, USA, 2015. EDA Consortium.

[16] 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*, page OTh1A.2. Optical Society of America, 2013.

[17] D. Vantrease, R. Schreiber, M. Monchiero, M. McLaren, N. Jouppi, M. Fiorentino, A. Davis, N. Binkert, R. Beausoleil, and J. Ahn. Corona: System implications of emerging nanophotonic technology. In *Computer Architecture, 2008. ISCA '08. 35th International Symposium on*, pages 153 –164, june 2008.

[18] 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.

[19] X. Zheng, F. Liu, J. Lexau, D. Patil, G. Li, Y. Luo, H. Thacker, I. Shubin, J. Yao, K. Raj, R. Ho, J. E. Cunningham, and A. Krishnamoorthy. Ultra-low power arrayed cmos silicon photonic transceivers for an 80 gbps wdm optical link. In *Optical Fiber Communication Conference/National Fiber Optic Engineers Conference 2011*, page PDPA1. Optical Society of America, 2011.