A Scalable, Self-Routed, Terabit Capacity, Photonic Interconnection Network

Assaf Shacham, Benjamin G. Lee, and Keren Bergman
Department of Electrical Engineering, Columbia University, New York, NY 10027
assaf@ee.columbia.edu; (212) 854-2768

Abstract

We present SPINet (Scalable Photonic Integrated Network), an optical switching architecture particularly designed for photonic integration. The performance of SPINet-based networks is investigated through simulations, and it is shown that SPINet can provide the bandwidth demanded by high performance computing systems while meeting the ultra-low latency and scalability requirements. Experiments are conducted on a model SOA-based switching node to verify the feasibility of the SPINet concepts, and demonstrate error-free routing of 160 Gb/s peak bandwidth payload.

1. Introduction

Photonic interconnection technologies are gaining acceptance as a viable solution to the bottleneck created by the fundamental limits of electronic transmission in systems based on intensive data exchange at high-bandwidths and low latencies [1]-[5]. Wavelength division multiplexing (WDM), bit rate transparency, and the low loss offered by the optical medium allow photonic interconnection networks to provide an interconnect solution that can transport unprecedented data bandwidths while mitigating or eliminating the power, distance, and wiring density problems that are common to electronic copper-based systems [2],[6]. The avoidance of costly O/E conversions combined with photonic integration of such a network can allow for ultra-low latencies, approaching the optical time of flight.

With the miniaturization of switching elements afforded by large-scale photonic integration, the propagation delay of light through an entire photonic integrated circuit (PIC), on which a full interconnection network can be implemented, is anticipated to be in the sub-nanosecond scale. A typical optical packet, tens or hundreds of nanoseconds in duration, will stretch across the network and its buffering on optical delay lines within the network will become impractical.

In the SPINet (Scalable Photonic Integrated Network) network architecture proposed in this paper, port-to-port optical packets (messages) are self-routed through an optical multistage interconnection network, constructed from 2×2 photonic switching nodes, while the payload is maintained in the optical domain across the network. The messages are constructed in a manner that takes advantage of WDM to achieve high bandwidth and simplify the node design. Contentions are resolved by dropping one of the contending messages, but a *physical-layer acknowledgement* mechanism allows the source terminals to regard dropped messages as blocked messages and employ immediate retransmissions if necessary. The network's acceptance rate can be increased by modifying the topology and using contention avoidance techniques.

2. Architecture Overview

A SPINet network is a binary butterfly-class multistage interconnection network, comprised of 2×2 photonic wideband switching nodes. The specific topology can vary between implementations (e.g. Banyan, Omega, etc.) [6],[7]. The topology studied in this paper is an Omega network [8], modified to remove the two broadcast states and add four single-path states in which only data from one input port is passed to an output port while the data from the other input port is dropped (Fig. 1).

The system is synchronous and slotted. The messages are constructed in the terminals and are transmitted on optical fibers into the network. Since the entire network can be integrated on a single PIC, even very short messages (tens of ns) are longer than

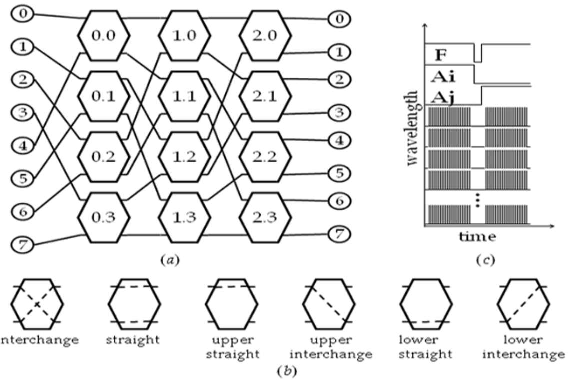

Fig. 1. 8×8 Omega network (a). Switching nodes' 6 states (b). Wavelength parallel messages (c): header bits and payload are encoded on dedicated wavelengths.

Proceedings of the 13th Symposium on High Performance Interconnects (HOTI'05)
1550-4794/05 $20.00 © 2005 IEEE

IEEE COMPUTER SOCIETY

the roundtrip time through the entire network and are perceived as lightpaths in that context.

At every slot messages start propagating simultaneously in the network. At every switching node, when the leading edges of the messages are received, a routing decision is made and the messages continue to propagate to their requested output ports. In the case of output port contention in the switching node, one of the contending messages is dropped. The choice of which message to drop can be random, alternating or priority-based. Since the propagation delay through every stage is identical, all the leading edges of the transmitted messages reach all the nodes of each stage at the same time.

The switching states of the nodes, as determined by leading edges, remain constant throughout the duration of the message so the entire message follows the path acquired by the leading edge, effectively creating a transparent lightpath between the source and destination terminals. When the messages reach their destinations, an acknowledgement optical pulse is generated and sent on the same transparent lightpath in the opposite direction. Owing to the bidirectionality of the switching elements, the acknowledgement pulses reach the appropriate source terminals.

When the slot time is over all terminals cease transmission simultaneously, the switching nodes reset their switching states, and the system is ready for a new slot. The slot duration is set so that the ack pulses are received before the slot ends, letting every terminal know whether its message was accepted and make a timely decision regarding its retransmission. This *physical-layer acknowledgement* mechanism allows the terminals to regard the dropped messages as blocked messages, and avoid the penalty associated with packet recovery at higher layers.

The wavelength domain is used to facilitate a routing mechanism in the switching nodes that can instantaneously determine and execute the routing decision upon receiving the messages' leading edges, and maintain a constant switching state throughout duration of the messages. The messages are constructed in a wavelength-parallel manner, i.e. the routing header and the payload are encoded on separate wavelengths and are received concurrently by the nodes (fig. 1c) [3]. The header is comprised of a *frame* bit, denoting the existence of the message, and a destination *address* tag. Each of the header bits, encoded on a dedicated wavelength, remains constant throughout the message. A single address bit is processed at every stage, so the number of wavelengths required for address encoding is $\log_2$ of the number of ports. The payload is segmented and encoded on multiple wavelengths to utilize the large bandwidth offered by WDM.

Butterfly networks are used in SPINet because their binary nature facilitates the usage of destination tag routing and simple decision rules that are required for ultra-low latency optical switching. The process of *implicit arbitration* between ports through self-routing also eliminates the need for a central arbitration mechanism thus allowing the system to scale to large port-counts. However, the fact that these networks are not non-blocking yields a lower throughput than that of a crossbar, for example. Contention avoidance techniques and topological modifications (as detailed in section III) as well as input speedup can be utilized to increase the message acceptance rate.

3. Performance Study

As mentioned above, the blocking nature of the Omega network leads to an increased dropping probability due to path contentions [6]-[8]. The architecture's efficiency and the performance of different topologies are evaluated through simulations. The simulation model uses Bernoulli uniform traffic at a range of offered loads. The metric chosen to demonstrate and compare the performance of the networks is the acceptance rate, the ratio of the successfully passed messages to the number attempted injections.

The simulation results: (1) acceptance rates of the Omega network over different network sizes at a load of 0.5 and (2) acceptance rates of a 64-port network over a varying offered load are plotted in fig. 2. The performance is compared to a maximum-matched crossbar with input-first separable allocators [6].

Topological modifications in the network can be used to increase the acceptance rate. The Enhanced Omega network (fig. 3) mitigates internal contentions by adding *scattering stages* before the routing stages.

Scattering stages are formed by the insertion of *scattering nodes* before the Omega switching stages (fig. 3). The scattering nodes' task is to identify messages that will contend for the same output port in the subsequent switching stage and to scatter them to different switching nodes. Scattering nodes misroute contending messages rather then drop them, letting the subsequent switching node route them correctly.

The connection patterns between the scattering stage and the routing stage (see fig. 3b.) complete the scattering action while ensuring that even misrouted messages reach their original destinations. Messages are only scattered between switching nodes that lead to the same part of the network in subsequent stages.

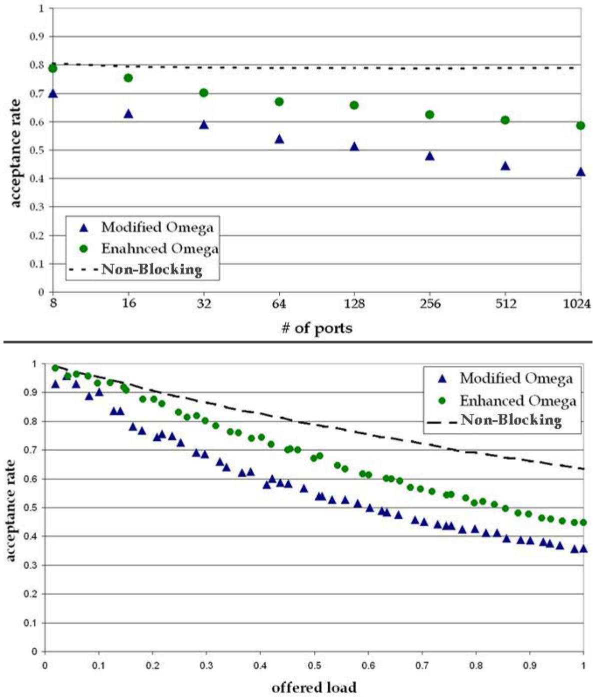

Fig. 2 – Top: Acceptance rate for modified Omega and Enhanced Omega networks of various sizes (load = 0.5). Bottom: Acceptance rate for 64-port modified Omega and Enhanced Omega networks, at different offered loads. Compared to a non-blocking input-arbitrated crossbar.

The scattering nodes cannot be placed before the last stage of the Omega network, because in this stage scattering will cause routing errors. Therefore, a maximum of N_S-1 scattering stages can be added to a network of N_S routing stages, increasing its number of stages to 2N_S-1. Simulation results of the Enhanced Omega network are also plotted in fig. 2.

The average bandwidth routed by a SPINet network can be calculated from the simulation results. For example, a 64-port Enhanced Omega network, operated at 0.8 offered load, attains 0.52 acceptance rate (fig. 2). The normalized throughput is therefore 0.8•0.52=0.42. Operation with a 160 Gb/s wavelength-parallel payload (16 × 10 Gb/s), as shown in the experiments in section 5, yields an average throughput of 67 Gb/s per port and 4.26 Tb/s system total average throughput.

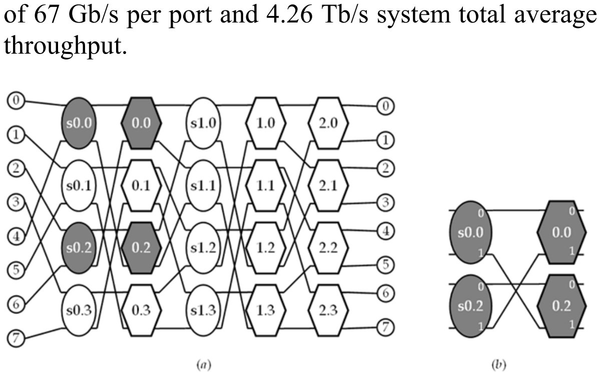

Fig. 3 – Enhanced Omega network (a), scattering nodes (ellipses) and routing nodes (hexagons). The scattering connection pattern is emphasized in (b).

4. Switching Node Design

Recent advances in photonic integration provide several promising technology platforms for the implementation of a large number of switching nodes on a single PIC. Integration of semiconductor photonic elements is an attractive research area recently gaining momentum [9],[10]. SPINet concepts can also be modeled using commercially available optoelectronic elements such as semiconductor optical amplifier (SOA) gates and photodetectors. SOAs offer the uniform gain curve, sub-ns switching time and low latency required from electronically controlled optical switching gates in a SPINet switching node [5].

The feasibility of SPINet's concepts: wavelength parallel messages, decoding of optical addresses and transparent bidirectional high-speed optical routing can be demonstrated using an optical switching node constructed from individually packaged SOAs along with other commercially available optoelectronic, electronic and passive optical elements [11].

An experimental node, based on SOAs, *p-i-n* receivers, passive optical elements and an electronic decision circuit (fig. 5) has been built as a concept demonstrator. When messages enter the node, the relevant header bits (*frame*, denoting message existence, and a relevant *address* bit) are optically extracted from both messages using wavelength filters, detected, and forwarded to an electronic control circuit. The control circuit, a Xilinx complex programmable logic device (CPLD), makes the routing decision and activates the appropriate SOAs (or SOA) to create the required input-output path. The messages, delayed on optical fibers, reach the SOAs exactly when they are activated and are routed appropriately.

5. Experimental Results

The correct functionality of the switching node and the concepts of SPINet are experimentally verified

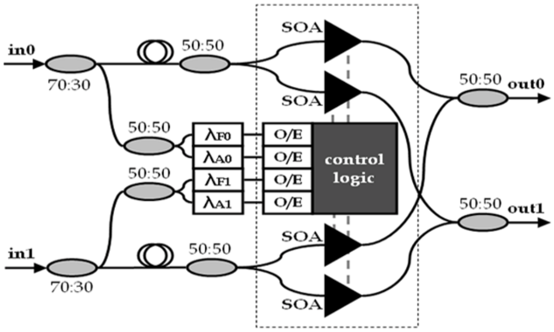

Fig. 4 – The experimental node is comprised of SOAs, optical couplers (ellipses, with coupling ratios), wavelength filters (λ), p-i-n receivers (O/E), optical fibers and a CPLD.

1550-4794/05 $20.00 © 2005 IEEE

using an optical testbed. Wavelength-parallel messages, consisting of 16 wavelengths modulated at 10 Gb/s, are constructed to create a total payload bandwidth of 160 Gb/s. The payload wavelengths span across 29 nm in the C-band, with a minimum spacing of 0.8 nm between adjacent channels, to show that more payload wavelengths can be straightforwardly added to increase the system's bandwidth. The messages are 97.6 ns long, spaced by a 4.8 ns dead time. Once constructed, the messages are multiplexed with the appropriate header wavelengths and injected into the experimental switching node through both input ports. At the node output ports correct routing is verified using an oscilloscope and bit error rate (BER) measurements are conducted on each wavelength individually. Ack pulses (9.6 ns long) are modulated externally on a dedicated wavelength and are injected into the output ports when messages are received.

Full routing functionality of all nine possible input combinations (*no-packet*, *packet-to-out0*, *packet-to-out1* per input port) has been verified in one experiment. Fig. 5 presents a second experiment where the correct routing of the ack pulse is verified for three single-message and contention-resolution cases. The third slot in fig. 5 shows two messages contending for output port #1. Input port #0 wins the contention and receives the appropriate ack pulse, while input port #1 doesn't receive the ack.

Error-free routing of the messages is verified and a BER of 10^{-12} or better is confirmed on all 16 payload wavelengths. It has been shown for SOA based multi-hop networks that after 58 hops, a 10^{-9} bit error rate can still be maintained for 8 wavelengths, spanned across a functional bandwidth of 24.2 nm [12]. As even large SPINet networks are expected to have a significantly lower number of stages, ($N_S \propto \log_2 N$), a larger functional bandwidth can be attained [12].

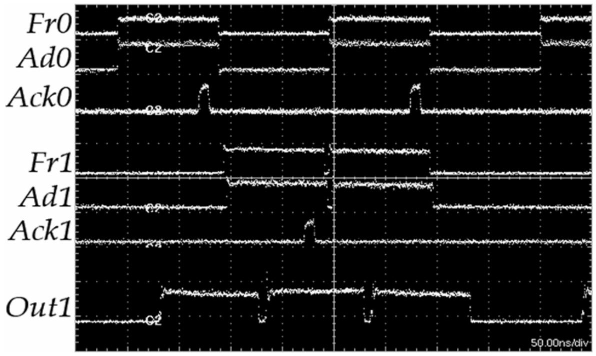

Fig. 5 – Optical signals: frame, address, and ack per input and frame at output #1. Single-packet and contention resolution scenarios demonstrated. The ack pulse is received at the appropriate ports

6. Conclusions

We have presented SPINet, a scalable architecture for photonic interconnection networks, which can be implemented in several different photonic technologies and was specifically designed to take advantage of large scale integration offered by future nanoscale photonics. The fundamental SPINet concepts: (1) implicit arbitration through self routing (2) physical layer acknowledgements and (3) wavelength-parallel messages that allow for ultra-high bandwidth and switching simplicity are presented. The Enhanced Omega topology is suggested as a means of improving performance (viz. acceptance rate) by contention avoidance.

A 2×2 optical self routing switching node is constructed and its functionality and error free transmission of 160 Gb/s payload are experimentally confirmed. The feasibility of the backward propagating acknowledgement pulse concept is also verified.

Future work can be done in performance studies and enhancements as well as in photonic technology research towards the integration of switching elements, optical receivers, and control circuitry.

7. References

[1] "The Future of Supercomputing: An Interim Report," NRC, National Academies Press, 2003.

[2] David A. B. Miller, *Proc. IEEE*, 88, pp. 728-748, Jun. 2000.

[3] Q. Yang, K. Bergman, G.D. Hughes, and F.G. Johnson,., *J. Lightwave Technol.*, vol. 19, pp. 1420-26, Oct. 2001.

[4] R. Hemenway, R. R. Grzybowski, C. Minkenberg, and R. Luijten, *J. Opt. Netw.*, vol. 3, pp. 900-913, Nov. 2004.

[5] G. I. Papadimitriou, C. Papazoglou, and A. Pomportsis, *J. Lightwave Technol.*, vol. 21, pp. 384-405, Feb. 2003.

[6] W.J. Dally and B. Towles, *Principles and Practices of Interconnection Networks*, San Francisco, CA: Morgan Kaufmann, 2004.

[7] C. Wu and T. Feng, *Tutorial: Interconnection Networks for Parallel and Distributed Processing*, IEEE Computer Society Press, 1984.

[8] Duncan H. Lawrie, *IEEE Trans. Comput.*, vol. 24, pp 1145-1155, Dec. 1975.

[9] T. L. Koch, and U. Koren, *IEEE J. Quantum Electron.*, vol. 27, pp. 641-653, Mar. 1991.

[10] K.A. Williams, *et al.*, "Monolithic Integration of Semiconductor Optical Switches for Optical Interconnects," Intel Research Report, October 2004, IRC-TR-04-018, also found at: www.intel-research.net/Publications/Cambridge/101320040239_287. pdf

[11] A. Shacham, B.A. Small, O. Liboiron-Ladouceur, J.P. Mack, K. Bergman, LEOS 2004, WM2, pp. 565-66.

[12] O. Liboiron-Ladouceur, W. Lu, B.A. Small, K. Bergman, LEOS 2004, WM3, pp. 567-68.