An Integrated Photonic Network for Multi-Processor Applications

Assaf Shacham and Keren Bergman
Department of Electrical Engineering, Columbia University, New York, NY, assaf@ee.columbia.edu

Introduction

The scaling in speed and integration density of semiconductor technology following Moore's law [1] has created an emerging performance gap between the elements comprising high performance computing systems. While commercial processors and memories that continue to benefit from amortized design cost and economies of scale shrink in size and accelerate processing speeds, the interconnecting medium fails to scale at an appropriate rate. Power consumption, signal distortion and pin density problems are aggravated as the data rates and port counts grow, and electronic interconnection networks have been identified as performance bottlenecks in systems based on high capacity communications between processors and memory elements [2]. As Moore's law begins to reach its fundamental limits, it is widely believed that future high-performance computing systems will be based on multiple processors that are expected to exchange large amounts of data at very high speeds [2]. An interconnect solution that provides a high-bandwidth, low-latency communication medium may be found in emerging technologies such as integrated photonic packet switching networks [3].

In this abstract, we present SPINet (Scalable Photonic Integrated Network), a new optical packet switching architecture, specifically designed for implementation with photonic integrated circuits. In a photonic integrated network, where the entire roundtrip time of light through the network is very small (<1ns), packet storage in fibers or waveguides is not practical and an original solution must be found for resolving packet contentions. The concepts of SPINet are discussed and simulation and experimental results are shown.

Architecture Overview

A SPINet network is a binary butterfly-class multistage interconnection network, comprised of 2×2 photonic switching nodes. The network is indirect, and each user has access to one input and one output terminal.

In this abstract we demonstrate the SPINet ideas on a modified Omega network, which is one possible implementation. An $N\times N$ Omega network [4] consists of $N_S=\log N$ identical stages. Each stage consists of a perfect shuffle interconnection followed by $N/2$ switching elements, as shown in fig. 1a. The modified Omega differs from the Omega in the switching node structure. Whereas in the original Omega the switching node has four allowed states (straight, interchange, upper broadcast, and lower broadcast), in the modified Omega, we remove the broadcast states and introduce four new states (upper straight, upper interchange, lower straight, lower interchange) in which only data from one input port is passed to an output port while the data from the other port is dropped (see fig. 1b).

In designing the photonic network for possible integration on a single photonic chip we consider the routing of messages that are longer in duration than the network roundtrip propagation delay. The short optical messages (tens of ns) will appear as lightpaths stretching across the network. The messages are wavelength-parallel (fig. 1c): the message header (address, priority, etc.) is encoded on several dedicated wavelengths, a single bit per wavelength, and remains constant over the entire duration of the message. Other wavelengths carry the payload at a capacity which is the product of the data rate and the number of wavelengths used, thus utilizing the immense bandwidth offered by WDM. The header and payload wavelengths are routed together as a unit.

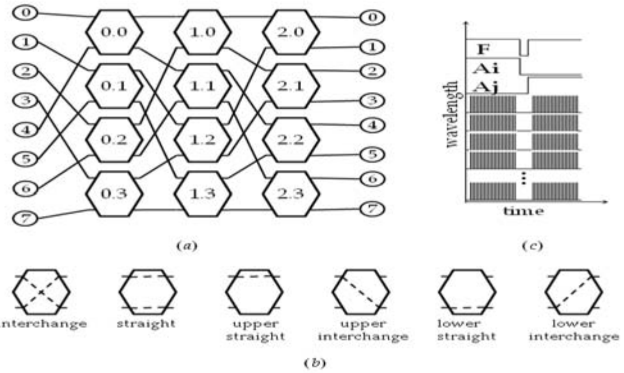

Figure 1: The modified Omega network (a). Each node has six switching states (b). The wavelength parallel messages (c).

The system is synchronous and slotted, so that all the terminals that transmit in a given slot, start transmitting at the same time and the messages start propagating simultaneously through the network.

As the leading edges of the messages propagate through the network, relevant bits of the header are decoded and a routing decision is made at every switching node. The wavelength-parallel structure greatly simplifies the header decoding process, as the required control bits can be extracted using wavelength filters and slow photodetectors (operating at the packet rate). Since only a single address bit is required for the 2×2 switching decision, other control bits can be used, such as an existence bit or a priority bit. The detected signals are then used by a fast and simple electronic circuit to determine routing decision, which is carried out by semiconductor optical amplifiers (SOAs) organized as drawn in fig. 2. In the case of a port contention within a node (i.e. two messages that are addressed to the same port) one message, chosen in a random or alternating fashion or according to a priority bit, is dropped while the other message is routed correctly.

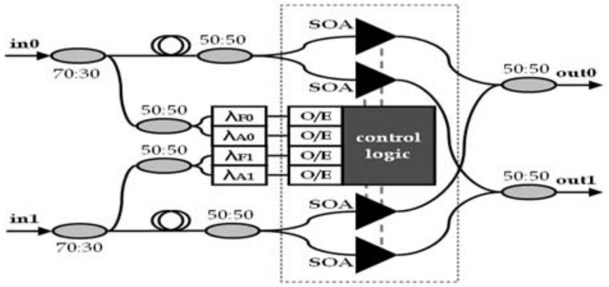

Figure 2: The switching node is comprised of SOAs, optical couplers (ellipses, with coupling ratios), wavelength filters (λ), p-i-n receivers (O/E), and an electronic control circuit.

As the messages propagate through the network, different header wavelengths are filtered at different stages, and the messages are routed to their destination (unless they are dropped in the network). When the leading edges reach the end of the network, an acknowledgement optical pulse is generated by the receiving terminals and transmitted in the reverse direction. The *acks* propagate through the open lightpaths, while the rest of the payload is still being transmitted, to let the successful sources know that their messages were received. Through this *physical layer acknowledgements* mechanism the sources know, before the end of the slot, whether their transmission was successful and can choose to retransmit in case wasn't, in a fashion similar to the retransmission of colliding messages in the CSMA/CD channel [5].

Performance Study

The acceptance rate, defined as the probability that a message successfully reaches its destination, is a key performance metric. In order to measure the acceptance rate, the network is simulated under uniform Bernoulli traffic. Figure 3 shows the acceptance rate as a function of the offered load (the Bernoulli parameter).

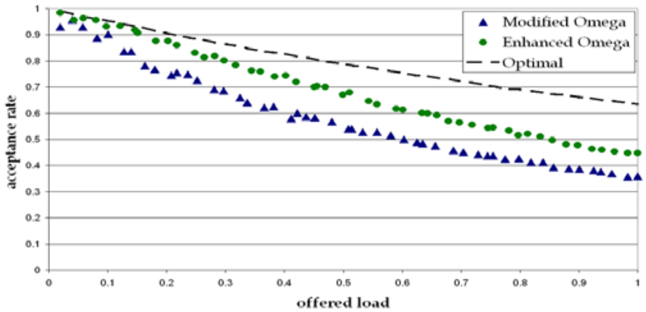

Figure 3: Acceptance rate for 64-port modified Omega and Enhanced Omega networks, at different offered loads. The dashed line denotes the upper bound set by port contentions.

The degradation of the acceptance rate with the load is noticeable and has to be addressed. The Enhanced Omega topology addresses this issue by placing *scattering stages* before the routing stages. A scattering stage is comprised of nodes that try to identify messages that are expected to contend in the following routing stage and scatter them to different nodes in an effort to mitigate *path contentions* (a state where two messages that are addressed to different system ports are dropped due to a contention for an internal path) and reduce the contention probability. Fig. 3 shows that the Enhanced Omega topology yields a 15%-20% performance improvement under the same simulation conditions. Other improvement methods are currently being studied.

The immense instantaneous bandwidth offered by the wavelength-parallel messages allows us to trade some of it to improve that acceptance rate and the throughput. E.g. if we choose to limit the offered load to 0.3 (i.e. each port is permitted to transmit during 30% of the slots) then an acceptance rate of 0.8 can be attained in an Enhanced Omega, 64-port network (fig. 3). Transmission of 16×10 Gb/s payload wavelengths yields an average bandwidth of: 160Gb/s•0.3•0.8 = 38.4 Gb/s per port and nearly 2.5 Tb/s total average bandwidth on chip for a 64-port system. Priority encoding can be used on messages to assure a nearly 1.0 acceptance rate for critical messages.

Experimental Results - Node Prototyping

Although the system is designed for integrated photonic networks, most of its concepts can be modeled on experimental systems that use currently commercially available technologies. A prototype switching node was fabricated where the passive optical elements are packaged together and the SOAs, the photodetectors, and the control electronics were mounted on a custom designed printed circuit board. The prototype design was used to verify the correct routing of messages with optically encoded addresses and the data integrity of the multi-wavelength messages after the routing.

Fig. 4 demonstrates transmission of messages from both input ports to output port #1 and the reception of the ack pulse. In the third slot, messages are sent from both ports, but an ack is received only in port #0, as expected. A power penalty smaller than 0.15 dB was measured for the switching node and 16 wavelengths of 10 Gb/s were routed with a BER or 10^{-12} or better.

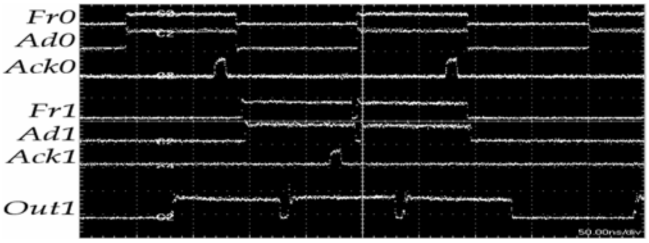

Figure 4: The waveforms show the transmission of messages from both input ports to Out1 and ack reception.

References

[1] G.E. Moore, *Electronics*, **38** (1965), 114-117.

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

[3] Papadimitriou, *J. Lightwave Technol.*, **21** (2003), 384-405.

[4] D. H. Lawrie, *IEEE Trans. Comput.*, **24** (1975), 1145-1155.

[5] D. P. Bertsekas and R. Gallager, *Data Networks*, Englewood Cliffs, NJ: Prentice Hall, 1992.