Experimental Demonstration of a Complete SPINet Optical Packet Switched Interconnection Network

Assaf Shacham, Howard Wang, and Keren Bergman
Department of Electrical Engineering, Columbia University, 500 W. 120th St., New York, New York 10027
assaf@ee.columbia.edu

Abstract: A 4×4 fully implemented photonic interconnection network is experimentally demonstrated. The network routes 60 Gb/s wavelength-striped packets (6×10 Gb/s) error-free in the optical domain, resolves contentions, and detects dropped packets via a unique acknowledgement protocol.
OCIS Codes: (060.4250) Networks; (200.4650) Optical interconnects

1. Introduction

SPINet (Scalable Photonic Integrated Network) is an architecture for optical packet switched interconnection networks with applications in local area networks, storage area networks and especially as a multiprocessor interconnect in high-performance computing (HPC) systems [1–3]. A SPINet switching fabric is comprised of a set of wideband 2×2 photonic switching nodes [2] organized as a multistage interconnection network (MIN) and implemented on a photonic integrated circuit (PIC). Messages are switched using semiconductor optical amplifier (SOA) gates, allowing for wideband transmission, packet-rate granularity and integrability, which are necessary to facilitate the high bandwidth, ultra-low latency, and high utilization required for future interconnection networks [4].

Because it is targeted for photonic integration, SPINet cannot employ optical buffering of any kind and messages have to be dropped upon contention. A novel *physical layer acknowledgement* protocol is used to provide a drop-detection mechanism. Integration of the network on a PIC facilitates fast transmission of these optical *ack* pulses and their reception by the source while the message is still being transmitted. As such, re-transmission can occur with minimal latency and the penalty incurred for message dropping is greatly diminished.

Previous work on the SPINet architecture included performance study under traffic patterns commonly used in HPC systems [1] and an experimental demonstration of a prototype switching node [2]. Methods of increasing the network throughput by utilizing its path diversity and the acknowledgement protocol were also investigated [3]. In this paper we present, for the first time, a fully implemented 3-stage SPINet network demonstrator with 4 input ports and 4 output ports. The network is comprised of a single-stage distribution network, used to increase the network throughput and its immunity to adversarial traffic patterns [3], and a 2-stage routing network providing full 4×4 mapping. While the macro-scale network reported here is not the envisioned fully-integrated SPINet, this implementation nonetheless demonstrates all the critical architectural concepts including optical address encoding/decoding, photonic end-to-end payload transmission and backward transmission of *ack* pulses. Experiments performed on the demonstration network verify its correct functionality in routing and switching of both optical packets and *ack* pulses. Error-free transmission (BER<10^{-9}) of 60 Gb/s payload wavelength-striped packets (6×10 Gb/s) is also confirmed.

2. SPINet Architecture Overview

A SPINet network is comprised of 2×2 SOA-based non-blocking switching nodes (Fig. 1a) organized as a MIN (fig. 1c). Since the network is designed to be integrated on a PIC, the optical messages are assumed to be longer than the

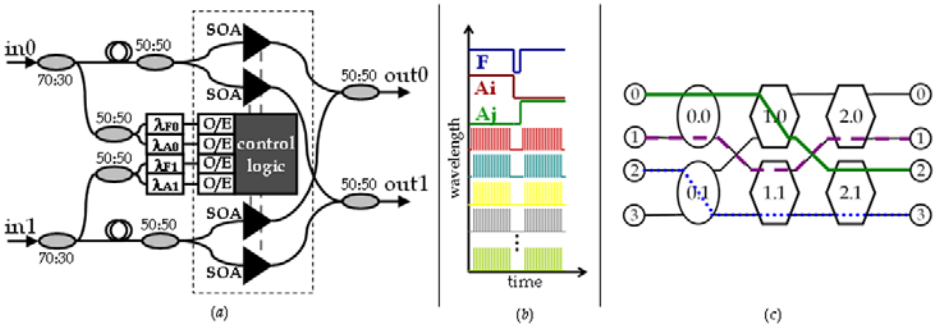

Fig. 1. (a) The wideband, non-blocking 2×2 switching node [2]. (b) The wavelength-striped packets: control information (*F*, *A*) and high-speed payload modulated on separate wavelengths. (c) The implemented, Omega-topology, 3-stage network. Distribution nodes are ovals and routing nodes are hexagons. Three lightpaths crossing the network are marked.

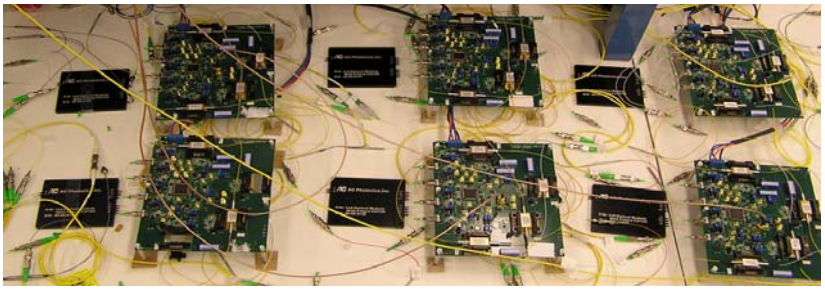

Fig. 2. The implemented 3-stage network demonstrator.

switching nodes (by an order of magnitude or more), so the nodes have no storage capability. The messages used are wavelength-striped (fig. 1b) where control information (framing and address) is encoded on dedicated wavelengths, a single bit per wavelength, and the payload is segmented and modulated at a high data rate on the rest of the band [1,2]. This structure, facilitated by the short reach requirement of the application, leverages WDM to offer very high transmission bandwidth and allows the switching nodes to decode the control information immediately upon the reception of the leading edges using a wavelength filter and a low-speed optical receiver. Once the framing and address signals are recovered from the incoming packets and processed by the high-speed electronic circuit the appropriate SOAs are switched on and the optical messages are routed to their desired destinations (or dropped when necessary) [2]. Banyan networks are preferred in the implementation of this architecture because they offer mapping of a large number of ports using only $\log_2 N$ stages of $N/2$ nodes.

The system is slotted so that the leading edges of the messages start propagating through the network simultaneously. Routing decisions are made in the switching nodes and messages are dropped when contentions occur. The successfully routed messages form transparent lightpaths that extend across the entire network. When the leading edges reach the output ports (while the messages are still being transmitted) optical pulses are sent in the reverse direction, leveraging on the bidirectional transparency of the switching nodes, acknowledging to the sources that their messages are being received. Sources of dropped messages do not receive *ack* pulses and can therefore retransmit. Owing to the low latency of the integrated network, the *ack* pulses are received at the sources before the end of the slot, facilitating the retransmissions with very low latency penalties.

The acceptance rate (i.e. the probability that a transmitted message is not dropped) is identified as the main performance metric. Since the queueing latency depends on the number of attempts required to successfully transmit a message, it is directly affected by the acceptance rate. Simulations show that a substantial increase in the acceptance rate can be achieved by utilizing the network's path diversity and the acknowledgement pulses [1], [3]. Average throughput of 48.6 Gb/s per port and 3.1Tb/s for a 64-port network has been shown to be achievable with an overall latency lower than 400 ns [1].

3. Experimental Network Demonstrator

The building block of the experimental network is a 2×2 wideband switching node reported in [2]. The switching node (Fig. 1a) is comprised of 4 SOA gates, controlled by a high speed Xilinx complex programmable logic device (CPLD). Two header bits are extracted from each input port using fixed 100-GHz wavelength filters: frame (λ_F=1555.75 nm) and address, whose wavelength is stage-dependent, encoding the requested output port.

The network demonstrator is comprised of 6 such switching nodes, connected as an Omega network. The address extraction filters at each stage correspond to the wavelength of the appropriate address bit (λ_{A0}=1535.04 nm, λ_{A1}=1533.47 nm, λ_{A2}=1550.92 nm). Stage 0 serves as a single-stage distribution network whose application is to mitigate contentions so the single-bit distribution address is encoded on λ_{A0}. Messages are not dropped in the distribution stage but are rather deflected on contention. The destination address is encoded on λ_{A1} and λ_{A2} which are used as the MSB and LSB, respectively.

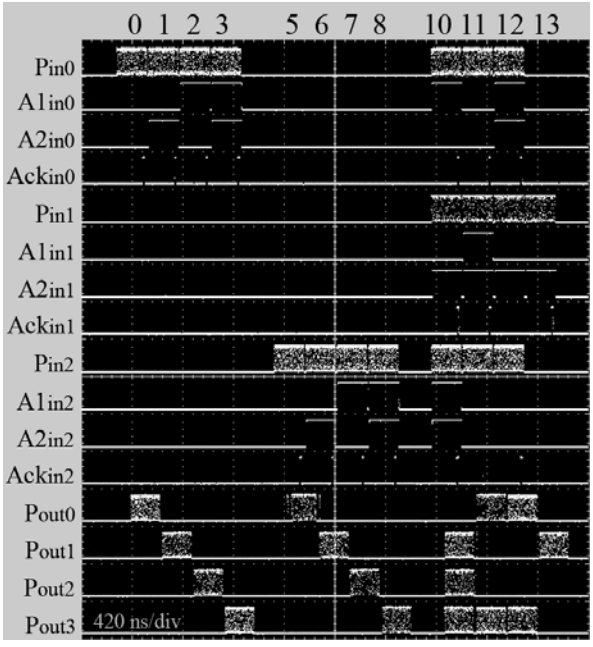

Fig. 3. Optical waveforms of the input and output signals: *payload*, *address*1, *address*2, and *ack* for each input port, and the packets' payload for each output port.

To test the network's functionality, an experimental setup is assembled to generate wavelength-striped optical packets independently for three input-ports and inspect the ejected packets at the system outputs. Six DFB-CW lasers are multiplexed, modulated with 2^{15}-1 PRBS at 10 Gb/s by a $LiNbO_3$ modulator, and then decorrelated using approximately 20 km of a SMF-28 fiber. The decorrelated data streams are then split using a 1×3 optical coupler, each stream is segmented into packets using an SOA gate and is then multiplexed with the header wavelengths, which are also modulated by SOAs according to the packet pattern. The packets are 250-ns long, spaced by a deadtime of 16.6 ns, thus forming 266.6-ns long slots. The payloads and addresses of the injected packets are shown in Fig. 3, where address 0 (the distribution address) is omitted for brevity.

4. Experiment I: Functionality

In the first experiment, the correct functionality of the network is verified. Packets are injected into 3 input ports (*input0*, *input1*, and *input2*). The 16-slot pattern is programmed so that initially (slots 0-3) all output ports are addressed from *input0*, then (slots 5-8) all output ports are addressed from *input2*. Finally (slots 10-13) several cases of contentions and packet-dropping are shown when the packets are injected simultaneously from three input ports. Whenever a packet is received at any output port, a 16-ns *ack* pulse is injected, using an optical circulator, into the output fiber in the opposite direction (λ_{ack}=1547.90 nm). Optical circulators are also placed at the inputs, to extract the *ack* pulses. The routed packets, as appearing at the outputs, and the *ack* pulses, as received at the inputs, are shown in Fig. 3, verifying correct routing functionality in both directions.

To assist in the interpretation of Fig. 3, the following example is given: in slot 11, 3 packets are injected: *input0*→*output0*, *input1*→*output3*, and *input2*→*output0*. The contention on *output0* is resolved by dropping the message from *input2* as can be seen by the fact that *ack*s are received only in *input0* and *input1*. The packet from *input2* to *output0* is retransmitted successfully in slot 12.

5. Experiment II: Payload Integrity

Signal integrity and error-free transmission are critical in any optical switching system, and require experimental verification in the presence of noise and non-linear effects that occur in SOAs. A bit error rate (BER) tester is used in the experimental setup to measure the BER of the routed packets. A BER of 10^{-9} or better is measured on all 6 payload wavelengths at the output. The spectrum of the packets, as they appear at the output of the network, and eye diagrams, at 10 Gb/s, at the input and at the output are shown in Fig. 4.

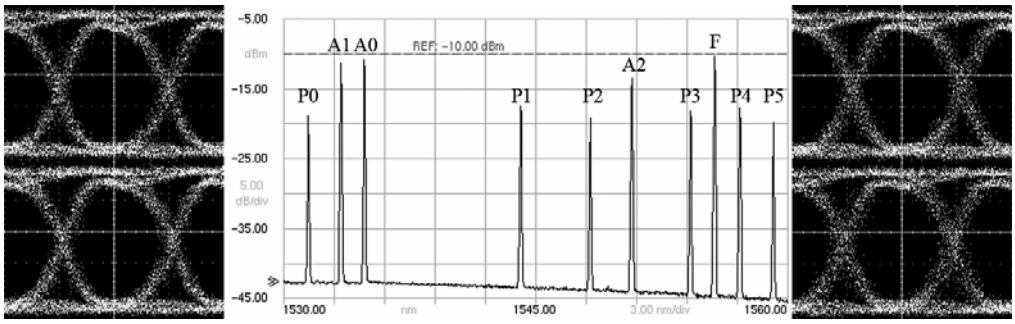

Fig. 4. The packets' spectrum at the outputs: *frame* (F), *address* (A0-A2), and *payload* (P0-P5) wavelengths are annotated. Eye diagrams at 10 Gb/s: inputs (left): 1531.56 nm (top) and 1558.59 nm (bottom), and outputs (right): 1531.56 nm (top) and 1558.59 nm (bottom).

6. Conclusions

We report, for the first time, the construction of an experimental network demonstrator of a 3-stage SPINet optical packet switched interconnection network. The macro-scale prototype implements the critical architecture concepts: optical address encoding/decoding, contention resolution, end-to-end photonic path, physical layer acknowledgement transmission and path distribution. We experimentally prove the feasibility of these concepts, the correct functionality of the network and the integrity for the optical signals routed through it.

We acknowledge support for this work from the NSF under grant CCF-0523771 and the U.S. Department of Defense under subcontract B-12-664.

7. References

[1] A. Shacham, K. Bergman, "Building Ultra Low Latency Interconnection Networks Using Photonic Integration", accepted for publication in *IEEE Micro*.

[2] A. Shacham, B. G. Lee, K. Bergman, "A Wideband, Non-Blocking, 2x2 Switching Node for a SPINet Network," *IEEE Photon. Technol. Lett.* **17**, Dec. 2005.

[3] A. Shacham, K. Bergman, "Utilizing Path Diversity in Optical Packet Switched Interconnection Networks," OFC 2006, OTuN5.

[4] NRC, *The Future of Supercomputing: An Interim Report*. Washington, DC: National Academies Press, 2003.