

# Experimental Demonstration of an Optical Interconnection Network with Asynchronous Transmission

Assaf Shacham, Caroline P. Lai, and Keren Bergman

Department of Electrical Engineering, Columbia University, New York, NY 10027, assaf@ee.columbia.edu

**Abstract** *Asynchronous routing of optical packets experimentally demonstrated on a 4x4 3-stage interconnection network testbed, offering increased flexibility and efficiency. 6x10 Gb/s wavelength-stripped packets with optically encoded addresses routed correctly. Error-free transmission is verified.*

## Introduction

In contemporary multiprocessor high-performance computing systems (HPCS), the processor–memory interconnect is often the single most constraining bottleneck in overall system performance [1]. This problem originates from the limitations of electronic hardware to provide sufficient bandwidths and low latency communications. Optical interconnection technologies and networks have been proposed as a means of eliminating the severity of this bottleneck [2]. The SPINet architecture [3]-[5] uniquely employs wideband photonic switching nodes to provide a ultra-broad bandwidth, low-latency interconnection among a large number of ports by setting up end-to-end optical paths across a multistage interconnection network (MIN) [5]. Contentions are resolved by dropping contending messages; however, optical acknowledgement pulses are exchanged on the acquired optical paths, providing a physical layer indication for message reception. Thus, the performance penalty associated with the loss of messages is mitigated.

Originally developed as a synchronous slotted network, the SPINet architecture straightforwardly lends itself to asynchronous operation. By incorporating simple changes to the switching nodes' electronic control circuitry, asynchronous operation can be achieved, offering improved flexibility with respect to network performance and management. This critical modification alleviates the need to synchronize a large number of terminals which may be physically distributed across a large room. Secondly, it enables the use of variable-length packets. This is especially important in HPCS interconnection networks in which many messages are control packets containing little data. When these messages occupy full slots, the network is clearly underutilized [6].

In this paper, we experimentally demonstrate the transmission of asynchronous messages of arbitrary lengths across a 3-stage SPINet testbed [4]-[5]. Correct routing and error-free transmission ( $BER < 10^{-12}$ ) are verified for wavelength-stripped messages.

## Architecture Overview

A SPINet network is comprised of 2x2 non-blocking switching nodes where semiconductor optical amplifiers (SOAs) are used as switching gates,

**Fig. 1:** Network block diagram and photograph (left) and wavelength striped messages (right)

organized as a MIN (Fig. 1). The network is designed to be integrated on a PIC. The optical messages, with durations in the range of tens of nanoseconds, span across several centimetres or more and are longer than the path in the PIC. Thus, the nodes have no storage capability. The messages are wavelength-stripped (Fig. 1); control information such as 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 (e.g. 10 Gb/s per wavelength) on the rest of the band [3]-[5]. Facilitated by the relatively short reach requirement of the application, this structure leverages WDM to offer very high transmission bandwidths. The architecture also 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 (Fig. 1).

The switching elements in the node are four SOAs, organized in a gate-matrix structure. Once the framing and address signals are recovered from the incoming packet 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). Banyan networks are preferred in implementing the architecture as they offer mapping of a large number of ports using only  $\log_2 N$  stages of  $N/2$  nodes.

Messages which are successfully routed (i.e. not dropped) create transparent lightpaths that extend across the entire network. When the leading edges reach the output ports (while the messages are still being transmitted), optical *ack* pulses are sent in the reverse direction. This leverages the bidirectional transparency of the switching nodes, thereby acknowledging to the sources that their messages

have been received. Sources of dropped messages do not receive *ack* pulses and can attempt retransmission at a later time. Due to the low latency of the integrated network, *ack* pulses are received at the sources within minimal time, facilitating retransmission with very low latency penalty.

### Adapting for Asynchronous Operation

The 6-node, 3-stage experimental testbed is constructed using switching-nodes fabricated from individually packaged elements (SOAs, passive optics, optical receivers, and digital electronics). While it uses macro-scale elements, unlike the envisioned integrated implementation of the SPINet architecture, the testbed serves to demonstrate critical network concepts such as address encoding and decoding, correct routing, and error free transmission of wavelength-striped messages in the presence of *ack* pulses [4,5].

The testbed can be adapted to asynchronously route variable length messages by the insertion of a 2-bit register into the routing logic of each switching node. Under the synchronous operation assumption, where all the messages are received simultaneously (at the beginning of the time slot), combinatorial logic is sufficient to process the header information, to route the messages, and to resolve contentions. However, when messages arrive at different times under an asynchronous operation, the 2-bit memory is required to encode state information in order to give priority to paths that are already established and to avoid interference from new messages.

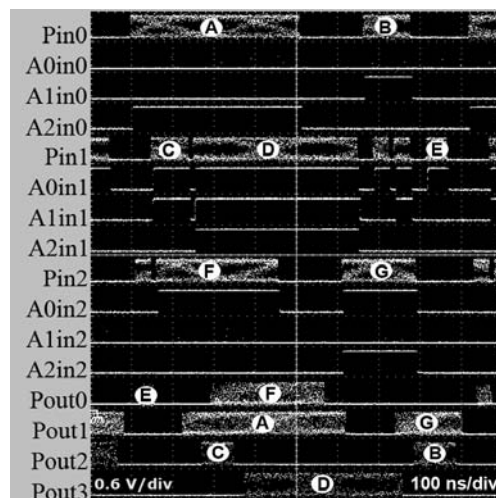
### Experimental Results

To verify the ability of the network to asynchronously route messages of arbitrary length, a pattern of wavelength-striped packets is injected via three independent input ports. The optical waveforms of the input and output signals are shown in Fig. 2.

The packets (annotated A through G in Fig. 2), with lengths varying between 53.3 ns and 409.6 ns are injected from 3 input ports (*in0*, *in1* and *in2*), with optically encoded addresses denoting the requested output port (*out0*, *out1*, *out2*, or *out3*). No relationship is assumed between the start or end times of individual packets. Each packet is comprised of a 4-wavelength header and 6 payload wavelengths, each modulated at 10 Gb/s. The wavelengths used for payload modulation range from 1539.9 nm to 1560.2 nm. The optical header of each packet is comprised of a *frame* signal (not shown in Fig. 2), a distribution address (*A0*, selecting a path between the two possible ones), and a routing address (*A1*, *A2*).

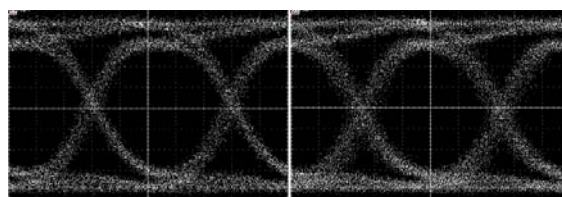
For example, packet C (length=89.2 ns) is injected from *in1*, addressed to *out2* (*A1*=1, *A2*=0), and emerges from *out2* after a latency of 133 ns.

After injection, a packet may be blocked by other



**Fig. 2:** Experimental waveforms of asynchronous traffic packets whose transmission began earlier. In this situation, the packet is dropped and no *ack* pulse is received at the originating source. In this experiment, *ack* pulses are not used since the round-trip latency of the testbed is large (approximately 270 ns) compared to the envisioned integrated implementation. It is therefore assumed that *ack* pulses are received (or not received) after 10 ns. In the case of a missing *ack* pulse, the source recognizes that the packet has been blocked and attempts to retransmit via a different path by changing the packet's distribution address. For example, the first attempt from *in2* is blocked. The source then retransmits with a different distribution address (change *A0* from 0 to 1) and another path is found in the network (packet *F*).

A Bit Error Rate Tester (BERT) is used in the experimental setup to verify error-free routing of the payload. A BER better than  $10^{-12}$  is verified on all six wavelengths. Eye diagrams at the input and the output of the network are shown in Fig. 3.



**Fig. 3:** Eye diagrams at 10 Gb/s of the input (left) and output (right) signals ( $\lambda=1560.2$  nm)

We acknowledge support from the NSF under grant CCF-0523771 and the U.S. DoD under subcontract B-12-664.

### References

- 1 Dally and Towles, *Principles and Practices of Interconnection Networks*, Morgan Kaufmann, 2004.
- 2 Hemenway et al., *JON*, 3(2004), pp. 900-913.
- 3 Shacham et al., *PTL*, 17 (2005), pp. 2742-2744
- 4 Shacham et al., *OFC 2007*, Paper OThF7
- 5 Shacham et al. *JLT*, submitted for peer review.
- 6 Dai and Panda, *Lect. Notes in Comp. Sci.*, 1417 (1998), pp. 171-184