Utilizing Path Diversity in Optical Packet Switched Interconnection Networks

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

Abstract: The concept of *path adjustments* is introduced as a mean of increasing the utilization of optical packet switched networks. Simulation results show substantial performance improvement under uniform and non-uniform traffic, and an experimental demonstration proves feasibility.

OCIS Codes: (060.4250) Networks; (200.4650) Optical interconnects

1. Introduction

SPINet (Scalable Photonic Integrated Network) was recently introduced as an optical packet switched interconnection architecture for local area applications such as high-performance computing systems and storage-area networks [1,2]. A SPINet switching fabric is comprised of a set of wideband 2×2 photonic switching nodes [2] organized in a multistage interconnection network (MIN). Semiconductor optical amplifiers (SOAs) are used as switching elements to allow for all-optical wideband transmission and packet-rate granularity, facilitating high bandwidth, ultra-low latency, and high utilization that are required for future interconnection networks.

Mainly targeted for photonic integration, SPINet does not employ optical buffering of any kind and messages are dropped upon contention. A novel *physical layer acknowledgement* protocol is used to provide a drop-detection mechanism. Integration of the network on a photonic integrated circuit (PIC) will facilitate fast transmission of these optical *ack* pulses and their reception in the source while the message is still being transmitted. As such, re-transmission can occur in with minimal latency and the penalty incurred for message dropping is greatly diminished.

Performance analysis demonstrating an average bandwidth exceeding 40 Gb/s per port has been reported [1] and error free routing of 160 Gb/s peak bandwidth has been experimentally verified in a prototype switching node [2]. In this paper we investigate the benefits of adding a distribution network before the routing network to enable path diversity and serve two purposes: (1) *load-balancing* to battle adversarial traffic patterns; and (2) *path adjustments* that increase network utilization. Simulations results showing the benefits of the distribution network and an experimental demonstration of the modified switching node functionality are presented.

2. SPINet Architecture Overview

A SPINet network is comprised of 2×2 SOA-based non-blocking switching nodes organized as a MIN (fig. 1a). Since the network is designed to be integrated on a PIC, the optical messages are assumed to be longer than the switching nodes by orders of magnitude, so the nodes have no storage capability. The messages used are wavelength-parallel (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, offers high bandwidth and allows the switching nodes to decode the control information immediately upon the reception of the leading edge. The optical messages are then routed accordingly [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. Their deficiency lies in their blocking property and lack of path diversity – a single path connects each input-output pair.

The system is slotted so that the leading edges of the messages start propagating down 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

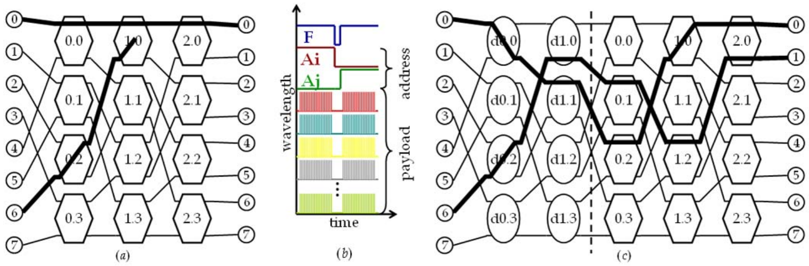

Fig. 1. (a) Omega network with 2 messages: In0→Out0 and In6→Out1 (dropped). (b) The wavelength parallel messages: frame, address and payload simultaneously encoded on dedicated wavelengths. (c) An omega routing network with a 2-stage distribution network, separated from the routing network by a dashed line. Both messages are routed successfully.

leading edges reach the output ports (while the messages are still being transmitted) optical pulses are sent in the reverse direction, utilizing the bidirectional transparency of the switching nodes, acknowledging to the sources that their messages were 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 (viz. the probability that a transmitted message is not dropped) is identified as the main performance metric. Since the queueing latency is measured as the number of attempts required to successfully transmit a message, it is directly affected by the acceptance rate. A low acceptance rate also translates to an underutilized network. Means of increasing the acceptance rate have therefore been sought, and the Enhanced Omega network [1] was introduced as an improvement of the Banyan style topology.

3. Path Diversity, Load Balancing, and Path Adjustments

The low roundtrip latency across the integrated network and the ability to integrate a large number of switching nodes on a single PIC [3] suggest the insertion of network elements for increasing the acceptance rate, leading to a latency reduction and improved network utilization. Scattering nodes, where contending messages are not dropped but are deflected to different output ports, were introduced in [1] as building blocks in the Enhanced Omega network. In this paper we investigate the insertion of a network of scattering nodes that precedes the routing network (fig 1c), and serves as a *distribution network* [4]. The distribution network routes messages to different input ports of the routing network in a manner that minimizes contentions. The routing network follows the distribution network to ensure correct routing functionality is intact.

The advantages of the distribution network are twofold: (1) A random route can be chosen in the distribution network by encoding a random *distribution address*, balancing the load on the routing network regardless of the real traffic pattern [4]. (2) Exploiting the SPINet physical layer acknowledgement protocol and the now-available path diversity, path adjustment can be made by changing the distribution address if the message is dropped in the first attempt. These *path adjustments* can be made in several iterations within the same timeslot, during the guardband that precedes the payload transmission. Each iteration takes as long as the sum of the roundtrip time across the network and the response time of the *ack* generation modules so the number of iterations can be a design parameter balancing the added utilization gained from multiple iterations and time that can be allocated to path adjustments.

In order to demonstrate these two features we consider the following example of an 8x8 network. Four messages are transmitted: messages A (from In0 to Out0), B (In1→Out6), C (In5→Out7), and D (In6→Out1). In an Omega network, two messages (C & D) are dropped due to internal path contentions (fig. 2a). By adding a distribution network and routing the messages through it according to a random distribution of addresses, message D takes a different path and is transmitted successfully (fig. 2b). Message C is still dropped, so when the *ack* pulse doesn't arrive on time, source #5 encodes a new distribution address on message C forcing it to take a different path that resolves the contention (fig. 2c). In this manner, the path adjustment technique uses the ultra-low latency of the integrated interconnection network and the distributed computing power of the switching nodes, to increase the network utilization by resolving contentions in the space domain.

The new configuration requires minor modifications to the message structure and to the switching nodes. It also requires the addition of input modules to attach the distribution address and re-encode that address when necessary.

4. Performance Study

The acceptance rate of the path-diversified SPINet is investigated on a representative 64×64 Omega network with a distribution network. Simulations are run using Bernoulli *iid* traffic with a varying p parameter (offered load). In the first performance study the immunity for adversarial traffic patterns is shown by simulating bit-reversal traffic patterns, chosen as adversarial patterns for the Omega network [5]. Fig. 3a shows the acceptance rate for varying sizes of distribution networks (0 to 6 stages). In these simulations no path adjustment iterations are allowed. As expected, adding distribution stages increases the path diversity and performance, but with a diminishing rate.

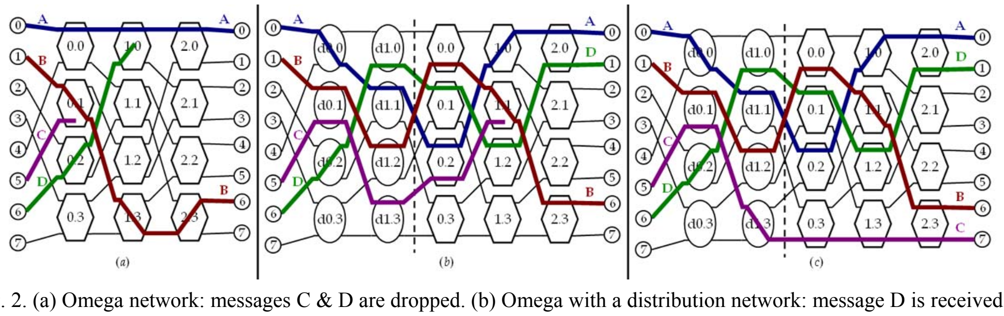

Fig. 2. (a) Omega network: messages C & D are dropped. (b) Omega with a distribution network: message D is received, C is dropped. (c) With path adjustments: message C's path is adjusted and it is now successfully received.

The effect of the path adjustments is investigated in the study. The network is simulated with 0 to 4 path adjustment iterations, under uniform traffic. The results lead to two interesting conclusions: (1) the performance improvement is substantial, effectively pushing the performance curve closer to the upper boundary represented by a non-blocking network (that would require 4096 switching nodes). (2) The performance improvement beyond 2 iterations is diminishing, and considering the time limitation, 2 iterations seem to be a good trade-off point.

5. Experimental Demonstration

The abovementioned modifications to the switching nodes' control logic are implemented on the programmable SPINet 2×2 switching node demonstrator [2]. Whereas the switching node in [2] only requires a gate array as the logic circuit, we now require the node to toggle between three states (*idle*, *cross*, and *bar*) to avoid interference between messages whose paths are being adjusted and messages that are already successfully routed. When the node is in *idle* state it can switch to *cross* or *bar* states according to the address encoded on the input message(s). Once it is in *cross* or *bar* state, it will remain in that state until the message that triggered that state (the original message) is over. During that time, it can only route messages that are not contending with the original message, while messages that are contending are blocked.

In Fig. 4 waveforms of the input and output optical signals are given. In the first slot, two messages are routed successfully without contention according to the optically encoded addresses (In0→Out1, In1→Out0). In the second slot, the messages are initially contending for Out0, and *ack* is received only at In0. A different message, which also contends with the first message, is then driven into In1 (an experimental setup emulates the distribution network) and is blocked. In the 2nd path adjustment iteration a message is received which doesn't contend with the first message and it is therefore passed successfully. Only then the payload transmission begins. Error-free transmission of 16×10 Gb/s wavelength-parallel messages has been confirmed for the switching node [2].

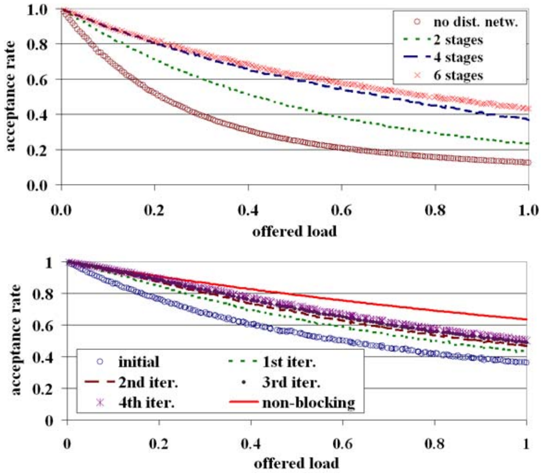

Fig. 3. Simulation results demonstrating the load balancing property (top) and path adjustments (bottom)

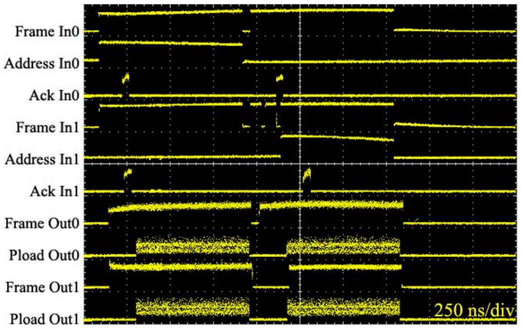

Fig. 4. Optical signals at the inputs of the node (*Frame*, *Address* and *Ack*) and at the outputs (*Frame* and Payload at 10 Gb/s) during two 880-ns timeslots. In the second slot, both messages are contending for Out0, so In1 receives path-adjusted messages. The 2nd message (to Out1) is successfully routed. The *ack* waveforms are delayed by 50 ns by the experimental setup.

6. Conclusions

Path diversity is presented as a mean of reducing the latency and increasing the utilization of SPINet-based interconnects while making them more immune to adversarial traffic patterns. Simulation results and an experimental demonstration of the modified switching operation are reported. Future research will include mathematical analysis of the acceptance rate in various traffic scenarios as well as a complete implementation of a multiport SPINet network and its supplemental modules.

This work was supported by the National Science Foundation under grant CCF 05-23771.

7. References

[1] A. Shacham, B. G. Lee, K. Bergman, "A Scalable, Self-Routed, Terabit Capacity Photonic Interconnection Network," in *Proc. Hot Interconnects, 13th Annual IEEE Symposium on High Performance Interconnects*, Stanford, Aug 2005, pp. 147-150.

[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] R. Nagarajan et al., "Large-Scale Photonic Integrated Circuits," IEEE Select. Topics Quantum Electron., **11**, pp. 50-65, Jan. 2005.

[4] F. Tobagi, "Fast Packet Switch Architectures for Broadband Integrated Services Digital Networks," *Proc. IEEE*, **78**, Jan. 1990, pp. 133-167.

[5] W.J. Dally and B. Towles, *Principles and Practices of Interconnection Networks* (Morgan Kaufmann, 2004)