Ultra-low Latency Optical Switching for Short Message Sizes in Cluster Scale Systems

Gouri Dongaonkar[1], Sébastien Rumley[1], Qi Li[1], Keren Bergman[1] and Madeleine Glick[2]
1: Department of Electrical Engineering, Columbia University, 500 W. 120th St., New York, NY 10027, USA
2: APIC Corporation, 5800 Uplander Way, Culver City, CA 90230, USA
gpd2115@columbia.edu

***Abstract*—A key performance criteria for many applications is the low latency delivery of small messages. We develop the SPINet (Scalable Photonic Interconnection Network) architecture for rack scale systems that can deliver near time-of-flight latencies.**

I. Introduction

Optical interconnects have clearly emerged as a promising technology for addressing the growing need for communications bandwidths among the increasing number of compute and memory elements in cluster-scale systems. Beyond the high bandwidth and energy efficient communications, photonic interconnects can be leveraged to enhance latency sensitive applications, such as graph visiting algorithms. Many messages generated in these algorithms are critical small packets accessing disparate parts of the graph [1,2].

In this paper, we propose a silicon photonics based architecture able to interconnect a cluster of communicating nodes that provides low latency switching for small messages. This architecture – based on the previously introduced chip-scale SPINet: Scalable Photonic Interconnection Network [3] is a self-routed packet-like transmission scheme which alleviates path setup time present in circuit switching. A novel physical layer retransmission protocol is implemented to compensate contention. The performance of the architecture and its scalability are investigated here.

II. Architecture and Network Protocol

SPINet is designed to optically interconnect the boards of a rack. We assume each board consists of computation and memory resources that send inter-board traffic to a silicon photonic network interface. This interface emits source-routed packets, which are delivered to the destination board through a silicon photonic switch fabric. This fabric consists of 2x2 switching elements connected hierarchically in an omega topology consisting of log(N) stages where N is the system radix. The feasibility of such a fabric has been previously assessed by a 4x4 experimental implementation with complex programmable logic devices (CPLDs) and discrete components. This implementation showed correct address encoding and decoding, routing and switching, and error-free transmission of high bandwidth messages [4]. It also included an acknowledgement mechanism to alleviate contention.

Packets destined to other boards are first dispatched to the inter-board network interface(NIF), where they are placed in a message buffer. In a first-in first-out manner, the messages are modulated onto header and data wavelengths utilizing WDM (Wavelength Division Multiplexing). The message header is encoded on separate header wavelengths, along with one additional wavelength indicating packet presence. Each header wavelength includes routing information for one stage of the omega topology, i.e. one 2x2 switch. At each switch, the corresponding header wavelengths is filtered, and following data wavelengths are routed just-in-time, enabling near time-of-flight latencies.

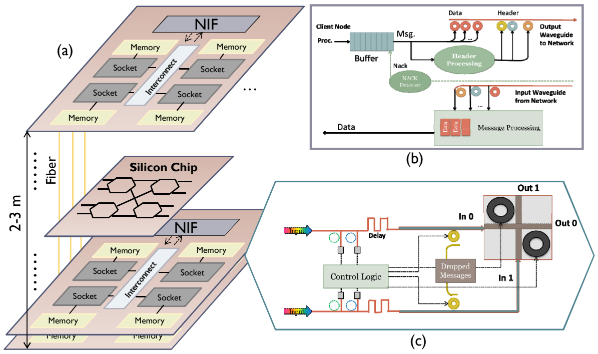

Fig. 1. (a) Rack scale SPINet architecture. (b) Deatils of Silicon Photonic network interface and (c) 2x2 switching element.

The network interface consists of a FIFO message buffer, modulator bank and waveguide to the network as seen in Fig. 1b. The header and data are modulated using ring resonators onto its respective wavelengths. On the reception side, incoming data payload striped over the multiple wavelengths are recomposed in packets, which are dispatched to the final destination on the board.

The switch fabric which interconnects all the boards in the rack can be located at the middle of the rack to minimize the average distance to the end nodes. As depicted in Fig. 1c, each switching element consists of a filter ring for header detection, comb switch rings for routing the data wavelengths (and remaining header wavelengths), and a packet dropping mechanism. The header wavelength is filtered and analyzed by a control logic that decides the appropriate routing or dropping for the packet. The state of the switch is tracked through a state register. An incoming packet is dropped through the broadband comb ring if its destination is incompatible with the switch state.

We implement a novel no-acknowledgement (NACK) mechanism in the switch fabric and network interface. In case of contention, a NACK is reflected

978-1-4673-5063-1/13/$31.00 ©2013 IEEE

back to the sender network interface. This is accomplished via a grating reflector on the switch node that reflects some of the packet energy back to the originating network interface [5]. We also assume a minimum packet size, such that in case of dropping the same packet is still at the head of the queue. This prevents buffer operations upon NACK reception. For further diminishing latency associated with retransmission, upon NACK reception we allow the sender to reset the current emission and restart it right away. We call this scheme FastNACK. Both the NACK and FastNACK protocols perform better than the original SPINet's ACK based scheme since the cluster scale distances are larger than chip scale (Fig. 2a).

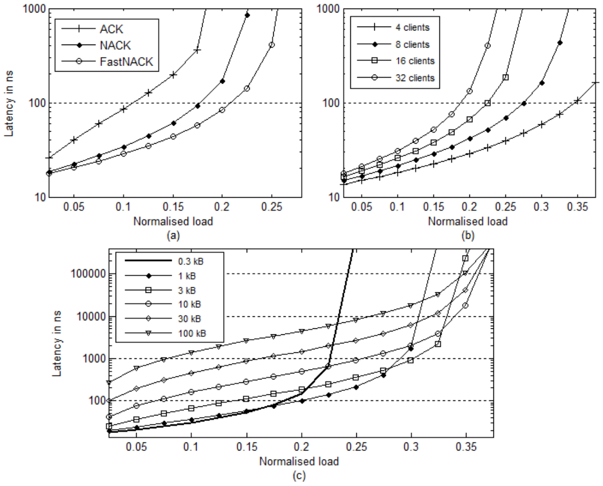

Fig. 2. Latency in ns: (a) Comparing protocols ACK, NACK, FastNACK for 32 clients and 0.5 kB message size (b) Scaling number of clients for 0.5 kB message size with FastNACK and (c) Scaling message size for 32 clients with FastNACK.

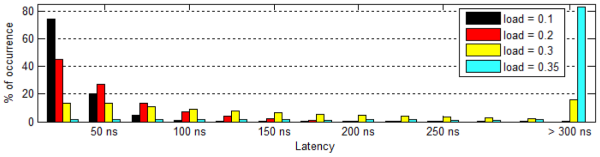

Fig. 3. Latency distribution for 8 clients for 0.5kB message size with FastNACK – most messages have a very low latency.

III. Simulation setup

In testing the SPINet architecture through simulation, previous experimental results [3]-[9] are used for photonic device parameters. We assume that data is modulated onto 16 wavelengths, each having a rate of 10Gbit/s. Poisson traffic is generated at each client node, at a rate corresponding to a fraction of the 160Gbit/s link capacity. As mentioned, we assume that the transmission time at the network interface is at least twice as long as signal propagation to the last stage in the switch fabric. For instance, for a propagation time of 5 ns, equivalent to 1m of fiber, and five switching stages (assuming 1ns as the switching time), the packet transmission time must be larger than 15ns. With 16 wavelengths at 10Gb/s, this implies a minimum packet size of 300 bytes.

IV. Performance Results

Results (Figure 2) show that for a 32 nodes implementation, the architecture is maintaining the latency very close to the time-of-flight up to a load of about 20-30%. Above this point, the number of retransmissions increases, limiting system bandwidth and further increasing the blocking probability. In scaling the network, the main effect is seen in the critical load supported by the architecture. The zero-load latency (that can be deduced by extrapolating Fig. 2) is also marginally increased, as longer headers are required.

Increasing packet sizes cause a rightward shift in the critical load (Fig. 2c). First, the time spent sending the header is better amortized with larger packets. Second, in presence of a NACK, only the data corresponding to the first 15ns (using the above numerical values) is lost, for any packet size. Once a packet finds a way through the fabric, the network resources are then optimally used for the packet duration. With larger packets, fewer packets are sent for the same load, which translates to fewer retransmissions that each utilizes the same amount of resources.

Finally, the latency distribution in Figure 3 shows that most messages have an ultra-low latency below 100 ns for loads under 30%.

V. Conclusion

We propose a transport network architecture that utilizes cut-through, all-optical switching and guarantees close to time of flight latency when the system has low loads(< 0.2). This also holds for small message sizes of 0.3 kB. SPINet guarantees packet transmission, which eliminates the need for a higher-level protocol like TCP and further improves the latency of the system. The results show architectural scalability up to 32 nodes, beyond which the blocking probability increases.

The scalability limitation related to physical layer properties is the next step of this study.

Acknowledgment

We acknowledge support from APIC Corp. under contract N00421-03-9-0002.

References

[1] P. A. Dinda *et al.*, *Proc. 2001 Intl. Conf. Parallel Processing*, pp. 175-184.

[2] M. Muller-Hannemann and S. Schirra (Eds.). *Algorithm Engineering: Bridging the Gap between Algorithm Theory and Practice*. Berlin, Heidelberg: Springer-Verlag, 2010.

[3] A. Shacham *et al.*, IEEE Micro **27**, 6-20 (2007).

[4] A. Shacham *et al.*, Proc. of OFC 2007, OThF7 .

[5] J. M. Foley *et al.*, Opt. Lett. 37, 1523-1525 (2012)

[6] B. G. Lee *et al.*, Proc. of OFC 2009, OMJ4.

[7] G. Li *et al.*, Opt. Express 19 (21) pp. 20435-20443 (2011).

[8] S. Assefa *et al.*, CLEO, PDPB11 (2011).

[9] G. L. Wojcik *et al.*, SPIE Pwest, 7230-21 (2009).