Low Latency, Rack Scale Optical Interconnection Network for Data Center Applications

Sébastien Rumley[(1)], Madeleine Glick[(2)], Gouri Dongaonkar[(1,3)], Robert Hendry[(1)], Keren Bergman[(1)], Raj Dutt[(2,3)]

[(1)] Dept. of Electrical Engineering, Columbia University, New York, NY 10027, USA, sr3061@columbia.edu
[(2)] APIC Corporation, 5800 Uplander Way, Culver City, CA 90230, USA
[(3)] PhotonIC Corporation, 5800 Uplander Way, Culver City, CA 90230, USA

Abstract *Warehouse scale datacenters running complex applications involving many servers require low latency interconnects to avoid excessive delays to the user. The SPINet(Scalable Photonic Interconnection Network) architecture can dynamically support ultralow latencies for packetized light loads and high-bandwidth long flows under heavy traffic.*

Introduction

The bandwidth bottleneck and growing power requirements have become central challenges for interconnection networks in high performance data centers. To address these challenges there has been considerable focus on deploying optical interconnection networks within these systems[1-3]. However, in addition to the increasing need for low cost, power efficient, and high bandwidth networks, there is also a growing requirement for data center architectures that will reduce latency.

For various commercial applications, user-end responsivity should be within 100 ms to be perceived as timely and natural. In [4] the authors point to experiments at Amazon in which every 100 ms increase in page load time decreased sales by 1% and experiments at Google in which a 500 ms increase in the search results display time reduced revenue by 20%. The strain on achieving an adequate response time grows with increasing complexity of the application and the size of the network. Web search results and web pages, from web sites such as Facebook or Amazon, query many servers, assembling pages from many different sources. As the applications become more complex (e.g. Google Instant) and as the size of the data center grows, the latency requirements become more challenging.

Congestion in the network can further cause data center round trip times to increase by two orders of magnitude forming a *long tail distribution*[5,6]. Reducing the long flow completion tail improves the worst case completion time. In [6], the authors point out that the network can help alleviate the long tail latency problem by providing: low latency for short flows, high burst tolerance and high utilization for long flows.

In this paper, we exploit the SPINet (Scalable Photonic Interconnection Network) architecture to address latency challenges in the data center[7]. SPINet leverages silicon photonics ring resonators to create an ultra-low latency / high-bandwidth optical switch, which can be used to realize intra and inter-rack connections required in the data center. By using wavelength striped transmissions, SPINet achieves link bandwidth scaling to up to hundreds of gigabits per second. These high-throughput connections can be reconfigured in a few nano-seconds, using an optical signaling scheme that controls the optical multistage ring network (Figure 1).

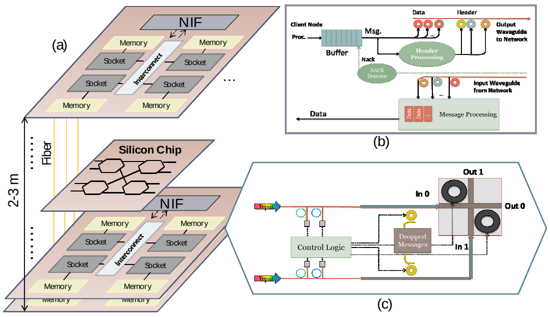

Fig. 1: (a) Rack scale SPINet architecture. (b) Details of silicon photonic network interface and (c) 2x2 switching element

In order to maintain low latency and simplify design, SPINet is a bufferless switch with all buffering occurring at the clients input interface. Applying a centralized arbitration to the switch would significantly increase latency due to the round-trip time of handshaking. Therefore, we let the clients independently control the switch resources by means of signaling wavelengths that are setup in-advance of data transmission. Once the data flow begins, the communication is cut-through. Each switching element is controlled by the simple presence or absence of control wavelengths (ON or OFF)[7].

Due to the use of the self-routing Omega topology, and the lack of output queues, the network can suffer from blocking within the switch if operated without central arbitration. As retransmission of dropped packets increases latency when left to the responsibility of the higher layer protocols, we include a retransmission protocol in our proposed architecture. To address this, in [8], we introduced a novel optical collision detection mechanism (the FastNACK protocol) that exploits optical bi-directionality of optical links: upon collision, the optical signal is reflected back to the emitting clients. The presence of reflected optical power thus notifies the source that the current transmission has been interrupted, and must be reinitialized. Therefore, the injected traffic is kept low for transactions requiring ultra-low latencies.

Although the FastNACK protocol enables a faster response to dropped packets, it also consumes available bandwidth. Here we alleviate this limitation with no changes to the SPINeT switch architecture (which must be kept simple for the reasons described above). We achieve this by introducing synchronization among the switch clients.

Baseline - Asynchronous Case

We first present the baseline SPINet architecture asynchronous operation. In this configuration, clients emit traffic in packet formats. Header wavelengths must be emitted a time *t* in advance (where *t* is the product of the stages in the switch topology and the switch configuration time). We take this configuration time to be of the order of nanoseconds for the silicon photonic, resonant ring based switch. The ratio of the packet emission time to the header time must be kept as high as possible in order to not lose efficiency due to header overhead. We thus consider a minimal packet duration of 200 ns duration. A base link bandwidth of 160 Gb/s (16 wavelengths at 10 Gb/s) thus yields a packet size of 4 kbytes. The minimal packet size is also dependent on the link distances, since the FastNACK protocol expects the packet to still be in the transmission buffer at NACK reception. Note that in presence of frequent small packets, SPINeT packets of sufficient duration can be composed of smaller packets through aggregation achieved at the client SPINeT interface. In presence of rare small packets, padding can be used to extend the packet size.

Extended - TDM Case

Packet collisions can be avoided by employing a form of Time Division Multiplexing (TDM), with clients transmitting to a given destination only during defined intervals.

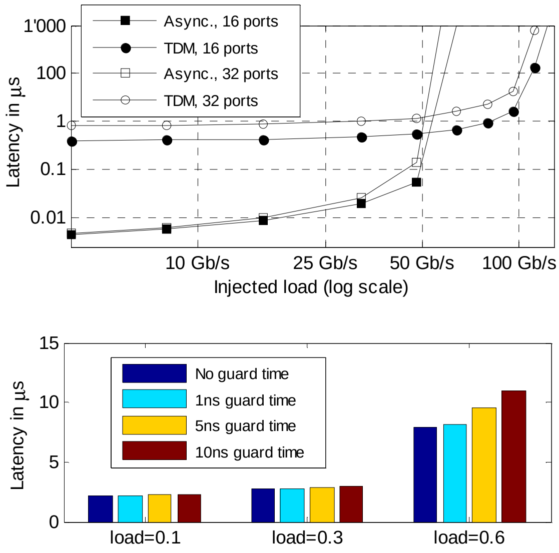

Fig. 2: TDM performance compared to the baseline case and effect of longer slot times (to account for loose synchronization)

As shown in Figure 2 (simulation conditions detailed hereafter), the TDM implementation almost fully leverages the high bandwidth provided by the SPINeT links (some bandwidth is lost due to the advance header mechanism). This advantage, however comes at a price of a much higher zero-load latency. When using TDM, messages must wait until the next designated slot to be transmitted. The number of slots depends on the topology and the routing but roughly scales with the number of clients. For an Omega topology with 16 ports, 16 slots are required. Under these conditions, each packet will spend an average of 1600 ns (200 ns x 16 slots / 2) simply waiting for the slot. Moreover, if multiple packets are scheduled to send during the same slot and form a queue, the delay scales with the inter-slot period. Latency distributions for the TDM case with various loads are displayed in Figure 3.

Note that in this TDM scheme the clients must know which communications (between source-destination pairs) should not occur simultaneously. This configuration can be made known to the clients *a priori*, or established by a negotiation process at system startup. Clients must also have a common clock to avoid slot mismatch. The synchronization can however be kept relatively loose, although tighter synchronization would improve performance, as illustrated in Figure 2.

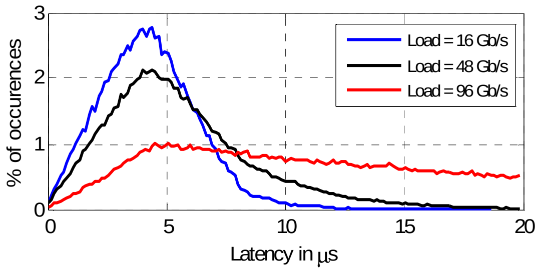

Fig. 3: Distribution of packet latencies using the TDM scheme, 16 ports.

Hybrid - Enhanced TDM Case

Thus far, we have presented an asynchronous operation mode achieving low latency for light loads but subject to unacceptable high latencies under heavy load, and a TDM mode able to support high bandwidth but showing high latency penalty for low bandwidth. In order to improve utilization, we propose an Enhanced TDM operation mode. Packets are organized in different queues at the client side (one queue per destination). When a slot begins, the corresponding queue is tested. If a packet is present, it is sent *with high priority*. If no packet is present, a packet from another queue is sent opportunistically *with low priority*. Priorities are attached to packets by means of an additional wavelength. At the SPINet switches, high priority packets pre-empt the switch if a low priority packet is already flowing. In this way, priority packets are guaranteed to get through the switch. The sender is notified of low priority dropped and pre-empted packets using the same FastNack mechanisms. As transmission always begin at slot start, packets are pre-empted early which leaves sufficient time for the NACK to flow to the transmitter.

Figure 4 compares the performance of the proposed Enhanced TDM scheme with the asynchronous FastNack and TDM schemes. The introduction of slots affects the latency for very low loads compared to the asynchronous case. However, as each slot can potentially be used for a packet (the low load increasing the probability of finding a free one), this surge is kept much more limited than with the TDM case. Conversely, for high loads, most packets are sent with high priority according to the slots, which then prevent collision and retransmission. Note that the SPINeT receivers must be changed accordingly to detect pre-empted packets.

Simulation Conditions

Performance analysis is achieved by mean of an ad-hoc discrete-event simulator, built on top of the Javanco framework^9. For each configuration we simulated the injection of 100,000 packets. Packets of uniform size (4 kB) are generated with exponential inter-arrival times. A configuration time of 1 ns is used for the optical rings. The latency of the links to and from the SPINeT switch is set to 15 ns (3 meters). Distances between the switching elements are neglected. All measurements consider head-to-head latencies.

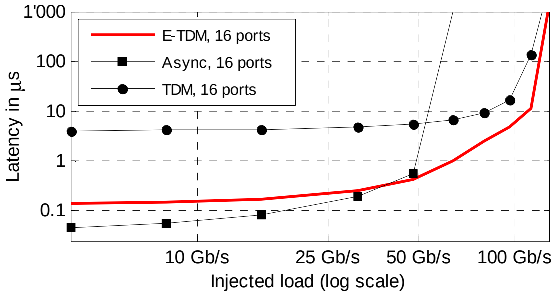

Fig. 4: Latency performance of the Enhanced TDM compared to the other operation modes

Conclusions

We extend the SPINeT architecture with a hybrid packet/TDM operation scheme which provides significantly improved latency and bandwidth characteristics. We believe that such an architecture can be of interest to address the both the high bandwidth and latency challenges of the warehouse scale datacenters.

References

[1] A. Vahdat et al., IEEE Micro, pp. 29-41, (2010).

[2] N. Farrington et al., ACM SIGCOMM Computer Communication Review, pp. 339-350, (2010).

[3] G. Wang et al., Hot Topics in Networks (HotNets-VIII) October 22-23, (2009).

[4] R. Kohavi and R Longbotham, “Online Experiments: Lessons Learned,” Computer, pp. 85-87, September 2007.

[5] D. Zats et al., “DeTail: Reducing the Flow Completion Time Tail in Data Center Networks,” Sigcomm’12, (2012).

[6] M. Alizadeh et al., “Data Center TCP (DCTCP),” Sigcomm’10 (2010).

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

[8] G. Dongaonkar et al., IEEE Optical Interconnects Conference (2013).

[9] S. Rumley et al., accepted for publication, IEEE ICTON 2013.