Scalable Photonic Packet Switch Test-bed for Datacenters

Hamid Mehrvar*, Yan Wang*, Xiaoling Yang*, Mohammad Kiaei*, Huixiao Ma*, Jianchao Cao*
Dongyu Geng*, Dominic Goodwill*, Eric Bernier*

*Huawei Technologies Canada Co. Ltd., Ottawa K2K 3J1, Canada
†Communication Technology Lab, Huawei Technologies Co. Ltd., Longgang, Shenzhen 518129, China

hamid.mehrvar@huawei.com

Abstract: We demonstrate scaling photonic switch for a datacenter using arrays of fast silicon photonic space switches that are time-slot synchronized, despite the varying distance of interconnected nodes. A low latency scheduling algorithm allows high throughput, with latency comparable to electronic switches.

OCIS codes: (060.4259) Networks, packet-switched, (060.6719) Switching, packet, (250.5300) Photonic Integrated Circuits

1. Introduction

The steep increase of datacenter traffic has motivated research into photonic switches that are potentially high capacity and energy efficient. Examples include: arrayed waveguide grating router (AWGR) plus fast-tuning lasers [1], and a hybrid of MEMS or liquid crystal optical circuit switching with electrical packet switching [2]. AWGR systems require a single well-controlled wavelength per signal, which is not consistent with low-cost 40G/100G optical interconnects. On the other hand, the slow switching of MEMS and liquid crystals limits their use to circuit connections. Therefore, we focus on wavelength-agnostic photonic packet switches, implemented in silicon photonic carrier injection technology with nanoseconds switch time. In [3], we implemented a 40GE synchronous time-slot photonic packet switch system that handles all packet types using a photonic framing, but we did not at that time have silicon photonic switches available. Use of photonic packet framing improved the efficiency of the photonic packet switch, and separation of the control and data-path wavelengths allowed simplification of the control system.

In this paper, we report a scalable architecture for photonic packet fabric using a set of fast switch matrices, implemented as carrier injection (CIOS) silicon photonic Mach-Zehnder matrices. The architecture uses a time-slot synchronous centralized control to perform scheduling of packets through the stack of photon switches, with no re-ordering. The controller uses an efficient low latency algorithm to assign connections to the aggregation nodes via the photonic fabric. A low latency scheduling algorithm achieves packet latency through the photonic fabric that is comparable to the latency of commercial store-and-forward electronic packet switches.

Scaling photonic switches to larger sizes is challenging. The size of a silicon photonic switch is limited by the number of 2x2 CIOS cells that can fit on a die, the number of optical I/O, and the insertion loss. The insertion loss limits fast photonic packet switches to 16x16 or 32x32 currently, until on-chip amplification becomes practical. In this paper, we use two 4x4 switches in parallel, aggregated by the control to deliver 8 channels capacity.

2. Scalable Photonic Switch Architecture

Our approach to scaling the photonic switch is a stack of \( M \) small \( N \times N \) photonic switch matrices, all controlled by a time-synchronous controller that makes the stack of switches act as one switch with capacity of \( (M \times N) \times (M \times N) \). We observe that there are multiple possible paths from a given source to a given destination – at least one path through each photonic switch. The stacking can be off chip or, preferentially, on-chip or both. Compared to a single large matrix, the stacked design has many fewer waveguide crossings and therefore has better insertion loss. For instance, a stacked on-chip photonic switch with I/O capacity of \( N \) can be implemented as \( P \) \( G \times G \) switches, each with \( N=6 \times P \) I/Os that are coupled together. In this paper, we assume only off-chip stacking, hence, \( P=1 \) and \( N=G \), and we use \( M=2 \) physical chips each with \( N=4 \) I/O, to achieve switch with \( M \times N=8 \) capacity. Fig 1(a) illustrates how the synchronous scalable photonic packet switch would be deployed in a datacenter. It is a 3-tier hierarchy with \( K \)

---

Fig 1 (a) Scaled datacenter using M bufferless photonic switch chips. (b) Control and Data path diagram. (c) Main controller block diagram.
servers connected to each top of rack switch (TOR), L electronic TORs interconnected to each electronic aggregation switch, and N aggregation switches inter-connected by a core photonic switch that consists of M chips of NxN photonic switches. For K=48, L=32 and N=32, the total number of servers is approximately 50,000. The aggregation switches are slightly different than conventional aggregation switches for an electronic packet network, to implement the buffering and photonic framing and to interface to the photonic controller. Alternatively, a two-tier architecture can be implemented by moving the new functions inside the TORs. The core switch comprises a stack of M buffer-less NxN silicon photonic switches controlled synchronously on time-slot space packet switch, and a controller for connection management, scheduling and synchronization. For N=32 and M=16, the total switch capacity in and out of the core fabric is 100Tb assuming 100G interconnectivity.

Each aggregation node has M photonic wrapper/un-wrapper interfacing with all M switches. Aggregation nodes perform buffering, and report to the main controller (via control channel) the status of the virtual output queues during each time slot. Using these reports, the main controller (Fig 1c) decides on the connectivity map of the photonic interfaces for the next time slot. Fig 1b show the timing diagram on both control and data channels. When an aggregation node receives a grant to send a photonic frame from a virtual queue, it de-queues a photonic frame amount of data, and wraps it with a preamble and start-delimiter sequence to create a photonic frame. It sends the photonic frame to the physical interface selected by the main controller. The photonic frame length is equal to the slot time. Using synchronization, each aggregation node adjusts its frame transmit boundary clock to ensure that the photonic frames arrive at the inputs of the photonic packet switch at the same time, despite diverse fiber lengths.

3. Low latency Control Algorithm
To achieve packet latency comparable to those of electronic switches, a low latency control algorithm is required. There have been many studies for control algorithms of slotted networks with equal length packets or cells[4]. We consider a scenario with variable size packets wrapped into a photonic frame that fit into a fixed timeslot. During a time slot, each aggregation node examines all N virtual output queues and reports occupancy status of top R queues (M ≤ R ≤ N). The control scheme based on reporting the number of bytes in the queue is called Longest Queue First (LQF) while the scheme based on reporting the number of packets is called Largest Number of Packets First (LNPF). After receipt of reports from all N aggregation nodes; the controller sort them in descending order and grants connection if both receiving and transmitting aggregation nodes have available interfaces. The number of interfaces allocated to each aggregation node depends on traffic condition.

Both LQF and LNPF can cause starvation as a packet in a short queue may have to wait a long time before its transmission is granted. Therefore, we introduce starvation avoidance (SA) to LQF and to LNPF, by including packet delay in the calculation. Let us define $Q_{ij} = q_{ij} / Q_{th} + d_{ij} / D_{th}$ and $P_{ij} = p_{ij} / P_{th} + d_{ij} / D_{th}$ as the normalized report for LQF and LNPF, respectively, for queue $j$ of node $i$. $q_{ij}$ and $d_{ij}$ represent the number of packets, the length of packets and the delay of oldest packet for queue $j$ of node $i$, respectively. $Q_{th}$, $P_{th}$ and $D_{th}$ are the normalization thresholds for the same quantities. The second term in $Q_{ij}$ and $P_{ij}$ addresses starvation. If $d_{ij}$ becomes larger than the threshold delay $D_{th}$, then the corresponding queue index is increased and the request has a higher chance of getting a grant. Note that use of $d_{ij}$ alone in reporting leads to a method called oldest packet first (OPF).

We performed simulation for N=64 aggregation nodes each with M=8 interfaces, with each interface connected to one of M silicon photonic chips. For uniform loading, Fig2(a,b) show the average and maximum delay, in number of time slots, under the assumption that the controller allocates no more than one interface to each queue in every time-slot. By allowing the allocation of more interfaces to a single queue, the delay performance improves as shown in Fig2(c), which shows delay bounds for LQF/SA with uniform and non-uniform traffic loading. In this figure, $\omega=0$ refers to uniform distribution of traffic loading of a node to all other nodes, while $\omega=1$ refers to full connectivity of one node to another. Assuming a photonic frame time slot of $T_s=1\mu s$, the modeling results of low latency LQF/SA algorithm are comparable with several microsecond measured delays of electronic packet switches.

Fig2. Simulation results for average delay (a) and maximum delay (b) for uniform traffic when there is a maximum of one allocation per queue per slot. (c) Simulation results for both uniform and non-uniform for LAF/SA when a queue can have many allocations in a slot.
4. **Test-bed implementation and experimental results**

In our test-bed, we implemented the aggregation nodes of Fig.1(a) using FPGA cores with external off-the-shelf optical transceivers. Fig.3(a) and (b) shows the experimental setup of our photonic packet switch test-bed. It uses two 4x4 silicon photonic switches \((M=2)\) that interconnect four aggregation nodes each with two interfaces. Each interface includes a photonic transmitter and a photonic receiver. We used 10Gb/s interfaces as these are simple to implement in FPGA’s. Fig.3(c) shows our packaged photonic switches that operate in the 1550nm band. The die are each packaged in a ceramic pin grid array, with a fiber ribbon attached. When an aggregation node receives a packet from any of the two 10GigE client lanes, it extracts the MAC destination address and en-queues the packet to the virtual queue of the destination egress card. The queue report is sent in each time-slot and the main controller assigns bandwidth based on the reports. When an aggregation node receives a grant in a time-slot, it wraps many MAC frames of the same queue into a photonic frame as shown in Fig.1(b). The length of each photonic frame, hence \(T_s\), is 12.7us with inter-frame gap of 3.17us.

The frames are sent to photonic switches using commercial XFP’s at rate of 11.3Gb/s. Polarization control is used at the switch input as this generation of silicon is polarization sensitive. The 1550nm light is amplified by an EDFA after switching. The modifications of the aggregation node are implemented in external Xilinx Virtex-7 FPGAs that shared a common clock. This clock simplified synchronization at the photonic Rx, as there is no frequency offset among the 4 aggregation nodes [3]. For a more practical implementation, burst-mode photonic Rx will be required.

Fig.3(d) shows the waveforms of switched packets, captured at an output port of the time-slot photonic switch. It illustrates the synchronous operation. The photo-receiver voltage polarity is inverted. Fig.3(e,f) show the average packet delay and throughput results for three control schemes when uniform traffic loading is used. Scheme 1 considers no dependency between the interfaces, and assesses the report from each interface separately. If there is an output contention, only one of the requesting interfaces is granted. Contention resolution in scheme 2 considers that each aggregation device can send their top \(M=2\) queue traffic using any of the two interfaces. As seen low latency scheme of LQF/SA scheme, discussed in the previous section outperforms the other two approaches. We observed no bit-error rate among any of the 8 traffic lanes that were applied to the 8-port switch system, comprising two 4x4 silicon photonic switch die. We demonstrated that it is possible to scale photonic switches using smaller scale photonic switches when efficient control is applied.

![Fig.3. Block diagram (a) and photo (b) of experimental testbed using stack of 2 silicon photonic switches. (c) A packaged 4x4 fast silicon photonic switch. (d) Packet capture at the output of synchronous switch. (e) Throughput of each 10G interface. (f) Measured average delay.](image)

5. **Conclusion**

We report scaling buffer-less silicon photonic switch fabric using a stack of \(M\) photonic switches, a synchronous control system and an efficient low latency scheduling scheme. The capacity of silicon photonic packet fabric is \(M \times C\) where \(C\) is capacity of each photonic switch. The results show that, in comparison with electronic packet switches, silicon photonic fabric can potentially offer comparable QoS performance in terms of delay and throughput, in addition to the higher rate, power, cooling and footprint advantages.

6. **References**


