PULSE: Sub-microsecond Optical Circuit Switched Data Center Network

Georgios Zervas, Joshua L Benjamin
Optical Networks Group, Electronic & Electrical Engineering, University College London, London, UK
g.zervas@ucl.ac.uk

Abstract: *A control and network system is proposed to offer optical circuits of 120 ns duration while achieving 91% throughput between 1000s of Servers. It delivers 270ns median and 4.7μs tail latency under 75% load.*
Keywords: *Data Center network architecture and control, Optical circuit-switched systems and their networking,*

I. INTRODUCTION

According to Google, the demand for bandwidth in data centers doubles every 12-15 months [1]. Electronic packet switches in Data Centers are under constant pressure to scale and be power efficient. However, lots of voices are indicating the slowdown of Moore's Law on silicon devices. Furthermore, packet-based networks have historically led to poor median and tail latency that also tend be load dependent and non-deterministic due to the unpredictability of traffic loads. For example, Amazon EC2 data centers have a median round trip latency of 600 μs and a 100-percentile long tail latency of 100 ms [2]. Also, such networks require complex methods for source/destination addressing, buffer management, admission and congestion control that also introduce overheads. Over the last two decades, there has been extensive research on technologies for optical switching and control that can potentially co-exist or replace electronic switching all together. However, before we reach the crossroad and migrate from electronic to optical switched solutions, we should address one fundamental question that is not widely discussed. Is packet or circuit switching the best solution moving forward for optically interconnected Data Centers?

This paper discusses the pros and cons of optical packet and optical circuit switched networks for Data Centers. It then reports on PULSE, a control and data plane network architecture for sub-microsecond optical circuit switching.

II. PACKETS VERCUS CIRCUITS

Packet switching/routing following one-way reservation is the underlying method of transporting data within Data Centers and over the Internet. It has the advantages of forming highly scalable networks, distribute its forwarding rules to each and every node, handle diverse flows due to variable packet size structures and has the ability to queue and manage flows along the path. However, packet switched networks are heavily influenced by load and distribution of flows and so a range of complex methods such as admission and congestion control, buffer/queue management, complex addressing with numerous header fields are necessary to offer a reasonable quality of service. Electronic chips such as ASICs can reasonably accommodate all these functionalities; however, they will struggle to keep up with demand. On the other hand, optical technologies can't easily replicate such functionalities due to limited and highly rigid data processing and storage functions.

Circuit switched networks can deliver guaranteed and deterministic traffic without the need of any of the packet switching complex methods (admission control, addressing, congestion control, buffer and associated management). However, the flexibility depends on circuit reconfiguration, which is down to switching latency, scheduling time and network size, since it follows a round trip handshake between requestors (end-points) and scheduler/controller. This restricts the flexibility and granularity of a system to deal with diverse flows, unpredictable and highly dynamic traffic. So, forming an optical network, one should aim to create certain design and performance targets:

- Deliver deterministic performance with zero data loss.
- Offer very high throughput with low scheduled median O(100ns) and tail latency O(μs) excluding propagation delays.
- Offer high level of granularity and highly dynamic reconfiguration O(10-100ns).
- Be scalable and energy efficient.
- Offer a simple low-cost solution that has minimum overheads with easy maintenance, control and management.

To satisfy the list, one would assume that an ideal system should have the best of both packet and circuit switch worlds and that is the reason why many studies have explored the use of hybrid packet/circuit switched systems. However, such approaches not only introduce and increase complexity in deployment, control, management, operation and resilience, but also require careful network dimensioning and traffic prediction. Here, we propose PULSE; an ultra-fast optical circuit switched system that can satisfy the design goals and offer the best of both worlds.

III. PULSE: NETWORK ARCHITECTURE

PULSE, is an ultra-fast optical circuit switched network architecture. The data plane architecture (see Fig. 1 right hand side) is based on x ultra-fast wavelength tunable transceivers per server and x^2 star couplers at the core of the network to interconnect x Racks, each populated with N Servers (total of $x \times N$ Servers). Each of the x transmitters of N Servers, within the same Rack, are connected to x *number of* $N \times N$ star couplers. Each transmitter is used and dedicated to connect only one destination Rack. For example, on Rack 1 (blue) we have the transmitter 1 of all N Servers connected to the blue star-coupler (top most star-coupler) and the output of the star coupler connects to the first (out of x) receiver of all N Servers. This creates x^2 parallel and disjoint networks that can be individually controlled since there is no contention between them. Also, the Data Center system can scale by either increasing the number of transceivers per server (x) that will increase the number of Racks supported and/or increasing the number of Servers (N) per Rack. Considering that photonic integration is making substantial progress, accommodating 10s of transceivers on a single package is not unrealistic. The star-coupler network offers the ability to inherently support broadcasting, multicasting, unicasting considering that it's a broadcast and select system. However, care should be taken on designing the physical network topology due to losses of star couplers.

The optical circuit switched system of PULSE is formed around the principle of circuits/epochs that accommodate τ time slots (fig. 1 bottom). Each Server can request a number of connections (destinations) each with a unique number of slots every epoch. Following the principle of having x^2 parallel (disjoint) star-couplers on data plane, the control plane is formed of x^2 hardware-based local schedulers (see on the left-side of Fig. 1). All schedulers are hosted within the Rack to minimize round-trip propagation delay for the request-response handshakes. The Servers transmit their requests to the corresponding scheduler using N port electrical interconnect. The schedulers that handle intra-Rack connection requests have electrical interconnects from scheduler back to the Server receivers. However, the schedulers that handle inter-Rack requests require optical interconnects. Once the hardware-based scheduler receives requests, it processes and grants the requested resources (slots carried over wavelengths), which must be communicated to the corresponding Servers. The transmitter will use the wavelength-slot pair information to configure the tunable wavelength source and the receiver to tune the ultra-fast tunable filter. Due to space restrictions of the paper, we don't describe the method and technologies used for ultra-fast tunable transmitter and receiver. As such, the PULSE architecture doesn't require in-network a) routing/switching, b) buffering and network addressing. However, it requires ultra-fast a) tunable wavelength switching b) filtering, c) clock and data recovery [3], d) synchronization and e) scheduling. However, we focus here on the scheduling aspects.

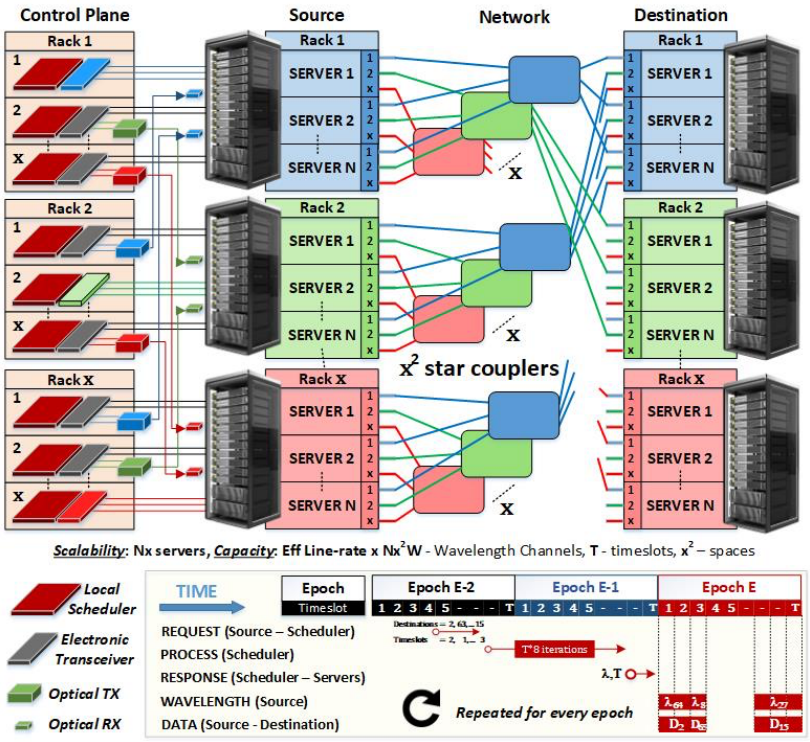

Fig. 1 PULSE network architecture.

A. Hardware-based Scheduler

The authors have developed an ultra-fast hardware based scheduler in order to support the network architecture in order to reduce the epoch length, maximize throughput and minimize median and tail latency. At [4], the authors reported a scheduler that was handing epochs of 1 μs duration and the wavelength switching/tuning was only taking place at the beginning of each epoch. Here, we briefly report on a scheduler that can support epochs of just 120 ns duration (8 times shorter) that allow for substantially more dynamic reconfiguration and is also able to cater for systems tuning wavelength at slot-level (20 ns) rather than epoch-level (120 ns). The hardware implementable scheduler adopts a parallel design with three pipelined stages (node contention resolution, wavelength decision and wavelength-slot selection) that consider R requests from each of the N nodes. Parallelism is used to speed up the scheduling process and round-robin arbiters are used to ensure fair selection of up to W (number of wavelengths supported by a transceiver), with unique source destination ports, per clock cycle. The scheduler can perform I iterations within one epoch to maximize throughout. After I cycles, failed requests are buffered and can be processed again in subsequent epochs. The slot-level scheduler differs from the epoch-level scheduler on the contention resolution and wavelength-slot selection

stages. The first stage has two pipeline sub-stages (instead of one), where one performs contention resolution for the sources and the second for destinations. Each sub-stage is constructed using N parallel N-port arbiters. The third stage uses initial iterations to allocate time-slots in chunks (coarse) and uses later iterations for fine allocation.

B. Simulation and Results

A simulation platform and software model in Matlab was developed to match the functionality of the hardware algorithm. Each stage of the hardware-based scheduler was synthesized on ASIC using 45 nm CMOS Nangate Opencell library. This provided the clock cycle period of 2.3 ns [4] for a 64-port star-coupler sub-network that was used to calculate the total number of iterations (48) possible within a 120 ns epoch. The Data Center simulation assumptions are the following: 1000 Servers across 16 Racks, 64 Servers per Rack, 16 transceivers per Server, 16^2 64-port star-couplers and schedulers, 20 ns slot period, and 6-slots per epoch. Each transceiver can tune across W wavelengths equal to the number of Nodes per Server (64 in this case) and should do so in just 500 ps (leading to 2.5% overhead). At every epoch, the traffic pattern consists of requests with two key items: destination and number of time-slots. A uniform random distribution was used to select the destination. The average size of each request corresponds to the slots available in the epoch divided by the requests per Server per epoch. Up to R requests are generated by each source per epoch and a Poisson distribution forms the inter-arrival rate of the requests. The simulation lasts for 2000 epochs.

Fig. 2 (a) showcases a 91% network throughout achieved by slot-level scheduling, a 60% increase, compared to 53% reached by epoch-level scheduling. This is due to substantially increased wavelength usage (Fig. 2(c)). The slot-level scheduling delivers a median latency of just 270 ns (almost 2-orders of magnitude lower than epoch level (24 μs)) and 4.7 μs (14 times lower than epoch level (66 μs)) 100-percentile tail latency. Fig. 2(c) summarizes all performance indicators and showcases the all-round benefits of the slot-level scheduler with regards to scheduler buffer size, average transmitter (TX) buffer size and wavelength usage, throughput, latency (median and tail).

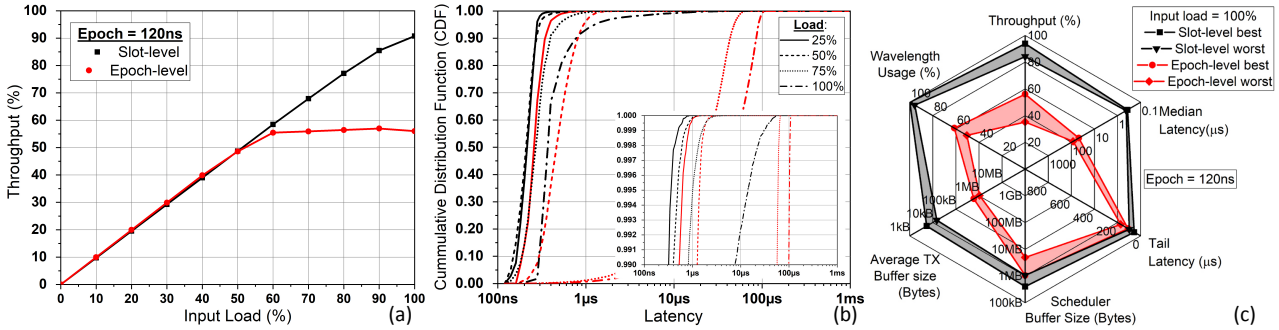

Fig. 2. Slot-level and epoch level schedulers are benchmarked against a) throughput, b) latency and c) numerous performance indicators.

IV. CONCLUSIONS

The PULSE ultra-fast optical circuit switched network architecture, as well as its control and scheduling process reported, showcase its ability to deliver near-identical abilities of packet switched system in terms of fine granularity, dynamicity and statistical multiplexing while it offers guaranteed services (no data loss) with very high throughput, as well as deterministic and ultra-low median and tail latency that also lead to minimal transmitter buffers. It offers a highly flexible, low complexity, low power network since it doesn't require active in-network components a) routers/switches, b) buffers. Since the core is based on star couplers, with the process of broadcast (at source) and select (at destination), it also inherently allows for run-time unicast, broadcast and multicast services. It eliminates the need for network addressing since the slot and wavelength information at source and destination points dictate data exchange. However, apart from the ultra-fast scheduling that is reported here, such architectures will need to address the need for ultra-fast a) tunable wavelength switching b) filtering, c) clock and data recovery, and d) synchronization.

ACKNOWLEDGMENT

The work is supported by EPSRC TRANSNET program (EP/R035342/1) and the UCL-Cambridge CDT program.

REFERENCES

[1] Singh, A et al., "Jupiter rising: A decade of clos topologies and centralized control in Google's datacenter network," SIGCOMM Comput. Commun. Rev., vol. 45, no. 4, pp. 183–197, Aug 2015

[2] Xu, Y et al., "Bobtail: avoiding long tails in the cloud," *Proceedings of the 10th USENIX conference on Networked Systems Design and Implementation* (nsdi'13), USENIX Association, Berkeley, CA, USA, 329-342, 2013.

[3] Clark, K et al., "Sub-Nanosecond Clock and Data Recovery in an Optically-Switched Data Centre Network," ECOC Post-deadline Paper, Sep 2018

[4] Benjamin, J et al., "Parallel star-coupler OCS architectures using distributed hardware schedulers". Proceedings of IEEE Photonics in Switching and Computing 2018