# PULSE: SCALABLE SUB-$\mu$s WDM-TDM CIRCUIT SWITCHED DATA CENTER NETWORK

***Joshua L Benjamin[1], Thomas Gerard[1], Polina Bayvel[1], Georgios Zervas[1]***

*[1]Optical Networks Group, Department of Electronic and Electrical Engineering, UCL (University College London), Torrington Place, London WC1E 7JE, United Kingdom*
**E-mail: joshua.benjamin.09@ucl.ac.uk*

**Keywords:** NANOSECOND OPTICAL CIRCUIT SWITCH, DATA CENTER NETWORK, HARDWARE SCHEDULER, SCALABILITY

## Abstract

We propose an OCS data center network and control system that uses distributed hardware schedulers to reconfigure circuits every 40ns. The network is scale-resilient to 8192 servers, achieving >92% sustained throughput, with low median (120ns) and tail (6.6$\mu$s) latencies, while consuming 415 pJ/bit.

## 1 Introduction

Cisco's IP traffic forecast predicts an exponential growth of the cloud to 20.6 ZB by 2021 [1]. Statistics highlight that the number of hyper-scale data centres will double and reach 628 to support the growth of the already enormous cloud [2]. Naturally, by 2021, 95% of global traffic will exist in the cloud. Traditional data centre architectures, based on hierarchical electronic packet switches, cannot sustain high performance for heavy cloud based applications, because of the long tail latency, O(100ms), that they incur [3]. Bursty cloud applications are reported to have 90% of packets that are less than 576 bytes in size [4]. Hence, in today's data centres, there is a requirement for ultra-fast nanosecond speed, energy-efficient optically switched network, that is resilient to traffic loads and can incur low deterministic latency.

Extensive research has been carried out on (optical) packet and (optical) circuit switching technology to achieve fast reconfiguration cycles. Optical packet switches require optical buffer/queue management, congestion control and complex data exchange protocols. They cannot easily replicate the range of complex methods and functionalities that current electronic switch ASICs perform. OPSquare [5], Hipo$\lambda$aos [6] are optical packet switch architectures that aim to limit average latencies to approximately O(1$\mu$s). OPSquare has increased packet loss (10%-50%) at high loads (60%-100%). Hipo$\lambda$aos requires a complex data plane with large number of components to make a 1024-port switch; this leads to increased cost and power consumption. Optical circuit switches like REACToR [7] and RotorNet [8] have been proposed with reconfiguration time of O(10$\mu$s). RotorNet is oblivious to network traffic/load as it establishes optical paths in a cyclic manner, taking almost 1ms to perform a network wide cycle, and it incurs high latency in non-ideal traffic. REACToR has a scheduling period of 1.5ms.

Hence, in this paper, we propose PULSE, an optical circuit switched (OCS) network and control system with sub-$\mu$s reconfiguration cycles O(10ns). PULSE is a single-hop network (diameter = 1) that inherently supports uni-, multi- and broadcast traffic with maximized net throughput, as purely the packet payload is communicated and the need for addressing is removed. We evaluate the effect of reducing the circuit duration of PULSE on throughput and latency. Moreover, we investigate the scalability of PULSE by increasing (1) servers per rack (or servers per sub-network) to 64, 128 and 256 (2) number of racks (or transceivers per server) to 8,16 and 32.

## 2 Network Architecture

PULSE is a synchronous ultra-fast transceiver based architecture with tunable transceivers and passive star-coupler cores. The architecture is reconfigured by tuning the wavelength (WDM) and allocating the timeslot (TDM) at the transceivers to dynamically establish light paths (circuits). Fig. 1 shows the PULSE OCS architecture, supporting up to $x$ $N$-server racks.

In the data plane, top-right of fig.1, each server is equipped with $x$ fast tunable transmitters and $x$ receivers with fast tunable filters or fast local oscillators (LOs) for coherent reception that can tune to one of $N$ wavelengths at sub-nanosecond timescales. Each transceiver connects the server to an $N$-port star-coupler sub-network, which in turn connects the server to one destination rack of servers [9]. The splitting loss of the coupler ($3log_2(N)$ dB) can be compensated by the use of SOAs at the transceivers. A total of $x^2$ $N \times N$ star-couplers are used to build the PULSE OCS architecture, where each star, being disjointed, is completely independent in terms of controller, contention, slot/epoch synchronization, clock/data recovery and communication timeline (epoch and latency overheads).

In the control plane, top-left of fig.1, $x$ local schedulers are hosted within the rack to minimize the round-trip propagation delay for the request-response handshakes. Each scheduler deals with the wavelength and timeslot allocation of one particular sub-network. The schedulers that handle inter-rack communication are equipped with optical transceivers to enable communication with the receivers of different racks.

The communication time-line, as shown by the bottom part of fig.1, is divided in epochs (reconfiguration cycle). Each

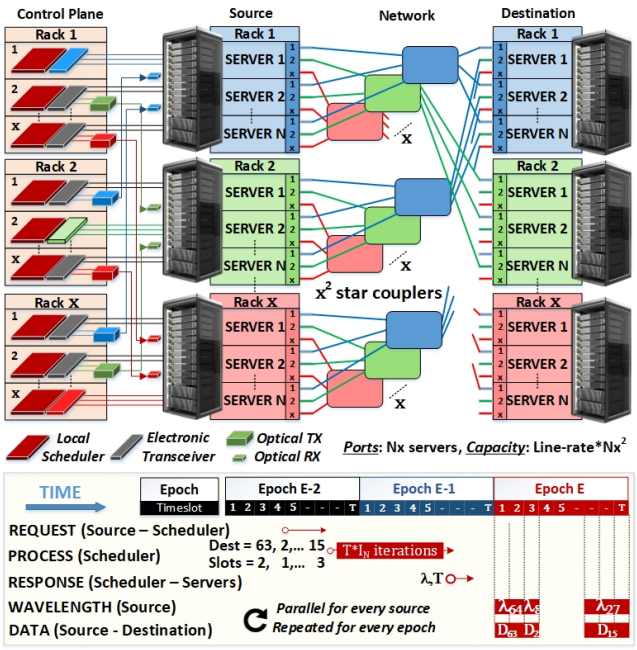


Fig. 1. PULSE architecture: Network and Control

epoch (40-600ns) is composed of $T$ timeslots (20ns). Each server sends its request with the destination server and number of slots in advance to the scheduler. The scheduler takes one epoch to compute the configuration. The response contains the allocated wavelength, timeslot and SOA configuration for each transceiver, which is communicated back to the corresponding source and destination servers.

## 3 Transceiver Architecture

We propose the use of a tunable DS-DBR laser at the transmitter, as shown in fig.2, to achieve tuning across $W$ wavelengths and support a line rate of 100 Gbps with external modulator. Each 20 ns timeslot can carry a 250 byte packet, which corresponds to the overall median packet size across various data centre workflows [10]. Prior experiments have shown that transitions between any pair of 80 wavelength channels (C-band in 1 bank) can complete tuning within 40 ns [9]. Hence, we propose 3 DS-DBRs and 3 SOAs per laser bank, as shown by the timing diagram in fig.2, to achieve 20 ns timeslots. 1,2,4 ($B$) banks are proposed for 64,128 and 256 ($N$) server racks, working in different wavelength bands, to provide $W = N$ channels. The DS-DBRs at the source (and destination for RX2) are instructed with the wavelengths to tune $T$-2 timeslots (40 ns) in advance. In other words, the data plane epoch has an offset of 2 timeslots (plus communication overhead) compared to control plane epoch. Each 20ns timeslot is preceded by a SOA gate reconfiguration overhead of 500ps, determined by the SOA rise/fall time [11]. Only one source or receiver gate is open at a given timeslot. A $3B$ port AWG is used to multiplex the optical signal, which is then modulated.

We propose two options for the receivers, as shown by the bottom of fig.2. The first (RX1) contains $W$ SOAs surrounded by $W$-port AWG de-multiplexer and multiplexer to select the input for the photo-diode in sub-nanosecond timescales. The disadvantage of such a receiver is the requirement of large port-count AWG and number of SOAs needed as $N$ scales. The second receiver (RX2) contains $B$ laser banks, using $3B$ DS-DBR lasers and $3B$ SOAs as LOs for a coherent receiver. The increased sensitivity of the coherent receiver also allows for a larger split in the star coupler as $N$ scales.

The PULSE transceivers require clock/data recovery and picosecond-level timeslot (and epoch) synchronization with respect to the other $N$ servers in the sub-network [4]. Here, we focus on the scheduling aspects and the limits it poses when scaling into larger networks.

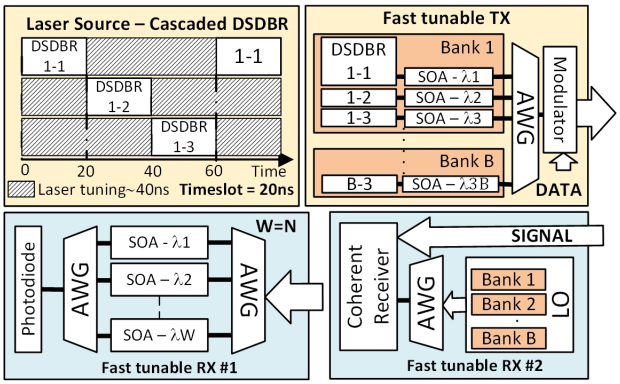


Fig. 2 PULSE transceivers: **TX**: DS-DBR SOA Banks/AWG **RX1**: SOA/AWG, **RX2**:Co-RX with DS-DBR SOA Bank LOs

## 4 Hardware-based Scheduler

We previously reported a scheduler that locked the wavelength prior to each 1$\mu$s epoch [12]. Here, we propose an ultra-fast slot-level hardware scheduler O(10ns), which computes wavelengths for every timeslot within the epoch. The scheduler employs parallelism (spatial and temporal), aiming to minimize epoch length, median and tail latency and maintain high throughput. It has three stages: contention resolution, register sequencer and resource allocation. The first stage has two pipelined $N \times N$-port round robin arbiters that resolve contention between source-destination node pairs in parallel. The second stage has a register sequencer, which checks previously assigned wavelengths (or random if none is assigned) to select available wavelengths in parallel. The third stage uses $W \times N$-port round robin arbiters to resolve wavelength contention and grants wavelength-timeslot pairs. An iteration is the processing of the demand through all stages once. The scheduling algorithm has a state machine, which decides if the current iteration must be used to allocate multiple slots (coarse allocation) or one timeslot (fine allocation) per server per iteration. The initial ($\sim R$) iterations are used for coarse allocation and the later iterations are used for fine allocation. Once the wavelength-timeslot is allocated, the SOA gate configuration at the source

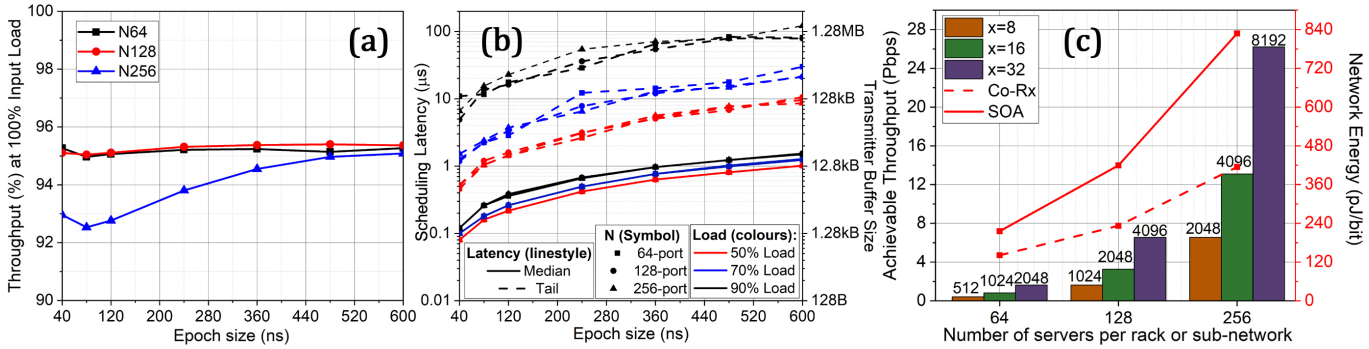


Fig. 3 (a) PULSE epoch size benchmarked against throughput and (b) median/tail latency (c) Network energy, capacity and size

(TX) and destination (RX) is also computed for each timeslot (in parallel) to activate the light path. Failed requests are buffered in the scheduler to retry in consequent epochs.

The efficiency of the scheduler can be maximized by increasing available iterations ($I$). The three stages were synthesized on 45nm CMOS ASIC using OpenCell library; the first stage (slowest of the pipeline) requires a clock period of 2.3, 2.9 and 3.9ns, allowing $I = T\times$ 8, 6 and 5 iterations (slightly less for smaller epochs as first iteration takes 4 cycles), indicated as $T \times I_N$ in fig.1, in an epoch for 64, 128 and 256-port systems respectively [13]. The control plane requires one 2 Gbps transceiver per server to communicate the request/response (9/32 bits) information within 20ns.

## 5 Simulation and Results

The hardware scheduler was modeled in MATLAB to evaluate resource matching performance. The generated demand traffic sends up to 2 requests/server per epoch ($R = 2$) with uniform random destination ($P(1/N)$) and slot demand ($P(R/T)$). A Poisson distribution with a mean inter-packet arrival time of $T/R$ is used.

Fig. 3(a) showcases the throughput achieved by the distributed hardware schedulers at 100% input load. The PULSE network achieves a sustained 95% throughput for 64 and 128 server racks, regardless of epoch size. For a 256 server rack (or sub-network), the throughput is 92.5% for small epoch sizes (40 and 80ns) and it gradually increases to 95% for a 600ns epoch. The PULSE scheduling algorithm achieves sustained throughput of >92%, taking into account a 500ps tuning overhead for every timeslot [11].

Since scheduler duration matches the epoch size, longer epochs also result in latency increase. Fig. 3(b) showcases the median and tail scheduling latency, excluding propagation and transceiver (serialization, coding) delays, of ($N$=) 64, 128 and 256-server PULSE racks for 2000 epochs at 50%, 70% and 90% input loads for different epoch sizes (40-600ns). Sub-$\mu$s median latency is achieved for epoch sizes less than 360ns. At 90% input load ($N$=256), for 40, 80 and 120ns epochs, the median latency is 120, 260 and 383ns respectively. The tail latencies are less than 2 orders of magnitude higher at 6.6, 15.4 and 22.9$\mu$s, which is better than the average latency of [5]. While awaiting the scheduler's response, the transmitters have to buffer the data. Hence, the transmitter buffer size required (Fig. 3(b) right axis) to support these median and tail latencies is less than 2.56 and 512kB (fits on on-chip memory of a network interface) respectively. The control system scales with high tolerance to latency as $N$ scales (upto 256 is shown in Fig. 3(b)) and, being disjointed, has no dependency on $x$.

Fig. 3(c) showcases the network scalability and the power consumption. The achievable throughput for $x = 8$, 16 and 32 (#transceivers/server or #racks) are shown by the bar chart for $N = 64$, 128 and 256 server racks, highlighting almost 24 Pbps for a 8192 server OCS network. The numbers on fig. 3(c) indicate the total number of servers in the OCS network. The network energy consumption is shown by the second y-axis and the red lines in fig. 3(c). The power estimates used for calculating the network energy consumption are: 1W for each tunable DS-DBR laser source [14], 0.26W for each SOA gate [15], 0.4W for modulator [16], 4W for coherent receiver [17] and 0.63W for photodiode [18]. SOA-based receivers (RX1) consume high energy (216 pJ/bit at $N$ =64) due to the number of SOAs needed, also increasing significantly as $N$ increases (828 pJ/bit). Although coherent receiver technology consumes high power, RX2 requires relatively fewer SOAs for high speed wavelength-timeslot selection and consumes lower power (141-415 pJ/bit).

## 6 Conclusion

We introduced PULSE, an ultra-fast OCS network architecture that configures optical circuits at packet timescales. We showcased the control plane to be scale-resilient, specifically to 64, 128 and 256 servers/rack achieving >92% sustainable throughput, and agnostic to the number of racks. PULSE, capable of re-establishing 40 ns circuits, incurs median and tail latency as low as 120ns and 6.6$\mu$s respectively at 90% input load. Coherent receiver technology and/or SOA gates could be used to compensate for the splitting losses of the $N$-port couplers, consuming a network energy of 415 pJ/bit.

## Acknowledgements

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

## References


[1] White Paper, 'Cisco Global Cloud Index: Forecast and Methodology, 2016-2021', Cisco(2018).

[2] 'Cisco - 95% of data centre traffic will come from cloud by 2021', https://www.cloudpro.co.uk/leadership/7304/cisco-95-of-data-centre-traffic-will-come-from-cloud-by-2021, accessed 20 April 2019

[3] Xu, Y., Musgrave, Z., Noble, B. et al.: 'Bobtail: Avoiding Long Tails in the Cloud', 10th USENIX Symposium on Networked Systems Design and Implementation, 2013, pp. 329–341.

[4] Clark, K., Ballani, H., Bayvel, P., et al.: 'Sub-Nanosecond CDR in an Optically-Switched Data Centre', European Conference on Optical Communication (ECOC), Rome, 2018, pp. 1-3, DOI:10.1109/ecoc.2018.8535333

[5] Yan, F., Miao, W., Raz, O., et al.: 'OPSquare: A Flat DCN Architecture Based on Flow-Controlled Optical Packet Switches', IEEE/OSA Journal of Optical Communications and Networking, 2017, 9, (4), pp.291-303, DOI: 10.1364/JOCN.9.000291.

[6] Terzenidis, N., Moralis-Pegios, M., Mourgias-Alexandris, M., et al.: 'High-Port and Low-Latency Optical Switches for Disaggregated Data Centers: The Hipo$\lambda$aos Switch Architecture', IEEE/OSA Journal of Optical Communications and Networking, 2018, 10, (7), pp. 102-116, DOI: 10.1364/JOCN.10.00B102.

[7] Liu, H., Lu, F., Forencich, A., et al.: 'Circuit Switching Under the Radar with REACToR', Symposium on Networked Systems Design and Implementation, 2014, pp. 1-15.

[8] Mellette, W.M., McGuinness, R., Roy, A., et al.: 'RotorNet: A Scalable, Low-complexity,Optical Datacenter Network', SIGCOMM, 2017, pp. 267-280, DOI:10.1145/3098822.3098838.

[9] Funnell, A., Shi, K., Costa, P., et al: 'Hybrid Wavelength Switched-TDMA High Port Count All-Optical Data Centre Switch', Journal of Lightwave Technology, 2017, 35,(20), pp. 38-44, DOI: 10.1109/JLT.2017.2741673.

[10] Roy, A., Zeng, H., Bagga, J., et al.:, 'Inside the Social Network's (Datacenter) Network', SIGCOMM, 2015, 45, (4), pp. 123-137, DOI:10.1145/2785956.2787472.

[11] Figueiredo, R.C., Ribeiro, N.S., Ribeiro, A.M.O., et al.: 'Hundred-Picoseconds Electro-Optical Switching with SOAs Using Multi-Impulse Step Injection Current', Journal of Lightwave Technology, 2015, 33, (1), pp. 69-77, DOI:10.1109/JLT.2014.2372893.

[12] Benjamin, J.L., Zervas, G.: 'Parallel Star-Coupler OCS Architectures using Distributed Hardware Schedulers', Photonic Switching and Computing, Cyprus, 2018.

[13] Benjamin, J.L., Funnell, A., Watts, P.M. et al.: 'A High Speed Hardware Scheduler for 1000-Port Optical Packet Switches to Enable Scalable Data Centers', IEEE 25th Annual Symposium on High-Performance Interconnects (HOTI), Santa Clara, CA, 2017, pp. 41-48, DOI:10.1109/HOTI.2017.2.

[14] Grobe, K., Eiselt, M.: 'Components and Subsystems', in Boreman, G. (Ed.): 'Wavelength Division Multiplexing' (John Wiley & Sons, 2014), pp. 70.

[15] Kobayashi, W., Fujiwara, N., Shindo, T., et al.: 'Ultra low power consumption operation of SOA assisted extended reach EADFB laser', OptoElectronics and Communications Conference (OECC) held jointly with 2016 International Conference on Photonics in Switching (PS), 2016, pp. 1-3.

[16] Zhu, H., Zhou, L., Wang, T., et. al: 'Optimized Silicon QPSK Modulator With 64-Gb/s Modulation Speed', IEEE Photonics Journal, 2015, 7, (3), DOI: 10.1109/JPHOT.2015.2425875.

[17] Perin, J.K., Shastri, A., Kahn, J.M.: 'Design of Low-Power DSP-Free Coherent Receivers for Data Center Links', Journal of Lightwave Technology, 2017, 35, (21), pp. 4650-4662 DOI: 10.1109/JLT.2017.2752079.

[18] Yoshimatsu, T., Nada, M., Oguma, M., et al.: 'Compact and high-sensitivity 100-Gb/s (4 ÃŮ 25 Gb/s) APD-ROSA with a LAN-WDM PLC demultiplexer', Optics Express, 2012, 20, (26), pp. 393-398.