Scaling PULSE Data Center Network Architecture and Scheduling Optical Circuits in Sub-microseconds

Joshua L. Benjamin, Georgios Zervas

Optical Networks Group, University College London, Torrington Place, London WC1E 7JE

joshua.benjamin.09@ucl.ac.uk

Abstract: PULSE, an optical circuit switched data center network, employs custom ASIC schedulers to reconfigure circuits in 240 ns. The revised PULSE architecture scales to 10,000s blades, achieves >95% sustained throughput, with low median (1.23 μs) and tail (145 μs) latencies, while consuming 115 pJ/bit and costing $9.04/Gbps.

OCIS codes: 060.4253 Networks, circuit-switched, 060.4264 Networks, wavelength assignment.

1. Introduction

The network that interconnects data center processing/memory/storage units (or blades) plays a crucial role in computational and application performance. Hence, substantial progress has been made in the development of high bandwidth transceivers (100 GBE, trending towards 400G and 800G) and network switches (Tomahawk, Barefoot Tofino - 12.8 Tbps). However, electronic network solutions suffer from high latency, which tend to degrade performance; median latency in O(100μs) and tail latency in O(100ms) [1]. We propose the scalability of PULSE [2], a wavelength or circuit switched network that establishes paths at packet-timescale granularity (20 ns) and aims to reduce this latency well over 2-3 orders of magnitude. PULSE employs the following to achieve this: (a) A custom ASIC hardware scheduler developed to allocate wavelength and time-slot resources, (b) Transceivers with nanosecond switching speeds, (c) Distributed and scalable transport network, (d) Fast CDR locking with minimal overhead based on [3] and (e) Timeslot-level synchronization. We previously proposed the network architecture and reported low latency optical circuit switched (OCS) reconfiguration [2]. Here, we aim to present a revised architecture and study on the scalability of PULSE to support 32768 blades. The scheduling is also now improved to support multiple sub-networks. Our results showcase a scheduling performance that achieves throughput over 95 % with median and tail latency in O(1 μs) and O(100 μs) respectively. PULSE establishes a flat network that minimizes transceiver/switch power and cost. We compare it against conventional electronic switched networks under three different topologies.

2. OCS Network Architecture

2.1. Operation

PULSE architecture has passive star-couplers in the core, which form sub-networks. Each star broadcasts all input signals across to all receivers at the output. At the edges, PULSE is equipped with fast tunable (<20 ns) (i) DS-DBR laser source at the transmitter and (ii) coherent receiver for wavelength selection and SOA gates (<500 ps) for timeslot selection. As shown in Fig. 1 and [2], each blade contains x transceivers and each transceiver connects a source rack to a destination rack (in a cluster); with x racks, we have x^2 sub-stars. Here, we introduce another dimension of scalability by introducing a split/broadcast and select unit after the transmitter/receiver. A 1 : p splitter makes the transceivers available across p sub-networks instead of one

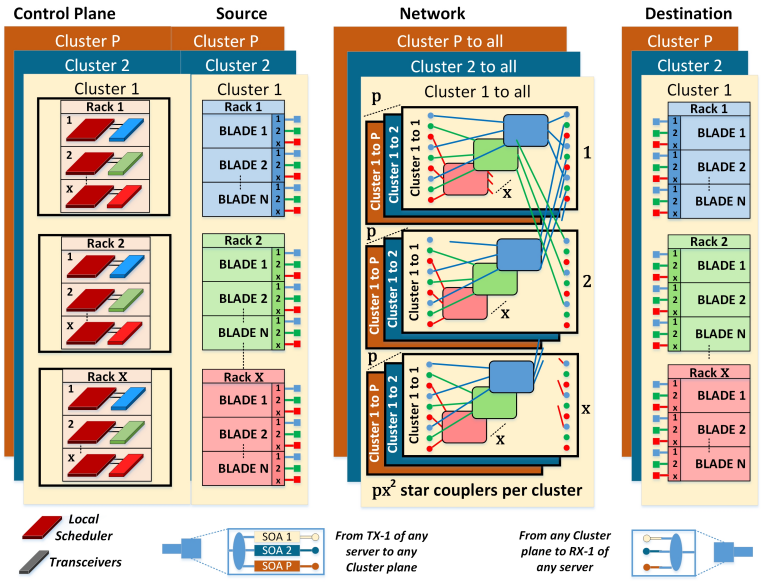

Fig. 1. PULSE: OCS Data and Control Architecture

while a semiconductor optical amplifier (SOA) gain-gate element allows for one of p sub-networks to be used every time-slot and compensate for the added optical loss. A transmitter in cluster p connects to all destination planes (1-to-p) on the 'Cluster P to all' plane. A receiver in cluster P connects to a specific cluster (p) of each 'Cluster P to all' plane. Each transceiver is now equipped with an additional SOA gate to compensate for the split and select a different destination cluster and hence, enhance scalability of network.

2.2. *Scalability*

Table 1. PULSE Scalability (N=64, W=N, linerate (lr) = 100 Gbps, efficiency (e) = 0.976)

Parameters	**Formula**	**Clustering/splitting p (x=16)**			**x=64**		
		2	**4**	**8**	**2**	**4**	**8**
Blades (#)	Npx	2048	4096	8192	8192	16384	32768
Transceivers (#)	Npx^2	32768	65536	0.13M	0.5M	1M	2.1M
Sub-networks (#)	x^2p^2	1024	4096	16384	16384	65536	0.26M
Re-usable Channels (#)	Wx^2	0.26M	1M	4.2M	4.2M	16.8M	67.1M
Capacity (Pbps)	$Npx^2 * lr * e$	3.2	6.4	12.8	51.2	102.3	204.7

The 1:p splitter/coupler that follows/precedes each transmitter/receiver, have a total insertion loss of $6log_2P$. In addition , the splitting loss of a star coupler is $3log_2N$. Assuming 64 blades per rack, N=64 split-sub-star and a split of P=8, a 36 dB loss is experienced. The SOAs that are selecting the path can boost the gain by 27 dB [4] bringing the optical power budget requirement to just 9 dB. This can be supported by either direct detect system and coherent receiver.

Table 1 shows the scalability of PULSE with the new topology when using (x=)16 and 64 transceivers per blade or racks per cluster. The parallel SDM/WDM/TDM creates modular structures, which mean that scalability of data, control plane, synchronization and CDR have to only scale to 64-ports. Integration of large channel bandwidth-dense transceivers as integrated SiP midboard optics (MBOs) has been shown in [5], proving the feasibility of supporting densities of 64 Gbps/mm^2 (as of 2014). In 2018, an ASIC switch with in-package optical transceiver ports was demonstrated [6]. Dense SiP integration of transceivers can enable the accommodation of 64 transceivers on a PULSE blade.

3. Network schedule Processing Unit (NsPU)

3.1. *Modules*

PULSE's NsPU is a parallel and pipelined custom made ASIC that uses three contention-resolving arbiter-based modules to compute schedule with >92 % throughput and 120 ns latency for a 40 ns reconfiguration cycles (epoch) [2]. The input to the scheduler are demands of destination and timeslot requests, while the outputs are wavelength, timeslot and SOA gate assignments with notifications of grant. As parallelism introduces contention, contention is resolved between sources-destination node pairs every iteration. Successful requests are then qualified for wavelength and timeslots request, which are granted based on a round-robin parallel schedule. In [2], we showed that PULSE NsPU can scale and achieve a clock speed of 435 MHz for a N=64-port sub-network, when synthesized on a 45 nm CMOS library.

3.2. *Iteration and Buffer Management*

While the modules discussed above form the core processors in PULSE's NsPU, a smart management of iterations is required to maximize throughput. PULSE NsPU requires 4 clock cycles to boot up every epoch. This corresponds to 9.2 ns in a N=64-port sub-network. Apart from booting up, the scheduler now has to schedule requests across p sub-networks. To increase the QoS, PULSE NsPU employs coarse (multiple slots granted per iteration per node) and fine timeslot allocation (one slot per iteration per node) in coordination with priority iterations for buffered requests, i.e. failed requests from previous epochs, in order to minimise latency.

In order to avoid convergence requirements, the number of schedulers is not increased with each of the p splits. Hence, for all P sub-networks, the scheduler must use the same number of iterations (I) and continue to support a high throughput and low latency. Hence, there is a requirement to manage iterations efficiently; we deal with requests for each sub-star in batches and hence, divide iterations across P networks: I/P iterations per sub-star. In order to maintain fairness, a pseudo-random generator shuffles sub-star priority in a round-robin fashion.

4. Results and Performance Analysis

The resource matching performance of NsPU was evaluated. The generated demand traffic sends up to 2 requests/blade per epoch ($R = 2$) with uniform random sub-network ($P(1/P)$), destination ($P(1/N)$), and slot demand ($P(R/T)$). A

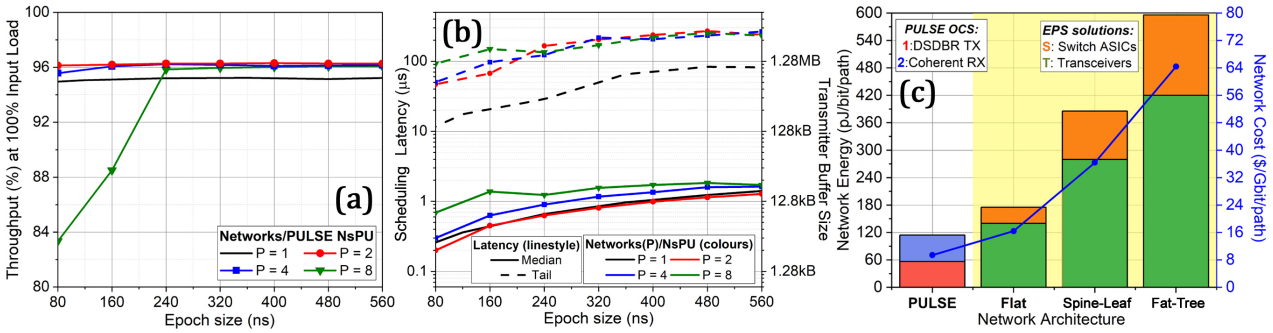

Fig. 2. PULSE NsPU performance: (a) Throughput and (b) Latency impacts of batching iterations for P sub-networks. PULSE Network performance: (c) Normalized network power (pJ/bit per path) and Cost ($/Gbps per path)

poisson distribution with a mean inter-packet arrival time of T/R is used. P=1 represents the latency performance we previously reported in [2].

Figure 2(a) shows the scheduler performance penalty experienced when the same NsPU is used to schedule wavelengths and timeslot resources for P =2, 4 and 8 sub-networks (tuning overhead included). At P=8 and 80 ns epoch, only 3 iterations are available per sub-star per epoch and hence, a 12% throughput penalty is experienced, reducing the matching performance to 83% as expected. However, an epoch size of 240 ns is a better operating point for P=8 with median latency at 1.23 μs and a tail of 145 μs. For P = 1, 2 and 4, throughput is sustained above 95%, even for smaller epoch sizes. In fig. 2(b), the effect of scaling PULSE on scheduling latency (excluding propagation delay) is shown. Scaling PULSE has minimal effect on median latency beyond 240 ns epoch sizes for increasing values of P. However, the tail latency is 4$\times$ at 240 ns epoch compared to P =1 and it decreases to $<3\times$ for 560 ns epoch. A brief comparison of normalized power and cost with equivalent state-of-the-art electronic switches and transceivers is shown in fig. 2(c). While the flat PULSE architecture achieves 115 pJ/bit with a single transceiver per path, electronic networks rely on multiple transceivers, reaching to various levels (Tier 1, 2, 3) of switches. In fig. 2(c), we show that PULSE achieves 65 pJ/bit/path lower power than even the equivalent (in capacity) Flat electronic network that replaces sub-stars with electronic packet switches (EPS) and 3-5x lower compared to Spine-Leaf and Fat-Tree electronic network architectures [7]. While employing coherent transceivers, the cost of PULSE is $9.04/Gbps, achieving 1.25×, 3-7× and 6-11× cost efficiency compared to Flat, Spine-leaf and Fat-tree electronic networks respectively. In electronic architectures, transceivers consume higher energy per capacity and network switches impact cost. While PULSE reduces latency by 3 orders of magnitude, the novel optical architecture also requires lower cost and consumes lower power by reducing the number of switches (to zero) and transceivers (to one pair) required.

5. Conclusion

The scalability of PULSE, a circuit switching architecture that configures optical circuits at packet timescales, was investigated. We showcased that the control plane can generate schedules for multiple sub-stars, specifically to 2, 4 and 8 clusters of 64-blade racks, enabling scalability to 10,000s blades, achieving >95 % sustainable throughput and a low median (1.23μs) and tail latency (145μs) at >240 ns epoch. Coherent receiver technology and SOA gates could coordinate to compensate for the large splitting losses of the coupler and splitting at the transceivers. PULSE consumes a network energy of 115 pJ/bit and a cost of $9.4 /Gbps per path, relatively smaller than equivalent electronic networks.

The work is supported by EPSRC TRANSNET (EP/R035342/1).

References

[1] Y.Xu *et al.*, "Bobtail - Avoiding Long Tails in the Cloud," presented at NSDI, 2013.
[2] J. Benjamin *et al.*, "PULSE: Scalability of a sub-μs wavelength-timeslot based circuit switched Data Center Network," in *ECOC*, 2019.
[3] K. Clark *et al.*,"Sub-Nanosecond Clock and Data Recovery in an Optically-Switched Data Centre Network," in *ECOC, Postdeadline Paper*, 2018.
[4] K. Van Gasse *et al.*, "27 dB gain III–V-on-SOA with > 17 dBm output power,"in *Opt. Express* 27, 293-302, 2019.
[5] G. Zervas *et al.*, "Optically Disaggregated Data Centers with minimal remote memory latency: Technologies, architectures, and resource allocation,"in *JOCN* 10, A270-A285 (2018).
[6] eeNews Analog, "Photonic ASIC directly integrates 100G optical ports" (Jan, 2019), https://www.eenewsanalog.com/news/photonic-asic-directly-integrates-100g-optical-ports.
[7] Lee Benjamin, "Opportunities and Challenges for Optical Switching in the Data Center,"in *OFC Workshop* 2019.