Parallel Star-coupler OCS Architectures using Distributed Hardware Schedulers

Joshua L Benjamin
Electronic & Electrical Engineering
University College London
London, United Kingdom
joshua.benjamin.09@ucl.ac.uk

Georgios Zervas
Electronic & Electrical Engineering
University College London
London, United Kingdom
g.zervas@ucl.ac.uk

***Abstract*— WDM/TDM star-coupler architectures have been proposed for building high-radix optical switches to enhance data centre network scalability. Here we propose a 1000-server star coupler network, which uses parallel OCS sub-stars to scale the network capacity to 256 Tbps. The harmony of distributed ns-timescale hardware schedulers and parallel OCS resource management achieves a throughput of 82.5% (200 Tbps) while incurring an average latency of 9.6 µs at 100% network load.**

Keywords—Star coupler, optical circuit switch, scheduling algorithms, scalability, WDM/TDM network

I. Introduction

The global demand for information technology is constantly increasing, demanding continuous growth of connectivity within data centre networks (DCNs). The scalability of a DCN is limited by the total capacity of the switching devices it employs. However, high capacity electronic packet switches are limited in scalability by their low port count and high energy consumption per bit per port. High port-count, high capacity optical switches can potentially offer scalability and a substantial reduction in DCN power consumption. Recent research work has proposed high radix optical switches for high-speed circuit switching [1]. We also previously proposed a 1000-port star-coupler based optical circuit switch (OCS) that uses dynamic wavelength tuning (i.e. WDM) and synchronised timeslots (i.e. TDM) to establish communication [2]. Nevertheless, reconfiguring the switch requires a fast and scalable controller. Hence, we proposed a novel parallel hardware scheduler design for a 1000-server optical switch that computes schedules in microsecond timescales [3].

Although a pool of sharable bandwidth is created by the star-coupler OCS, the ‘bandwidth per server’ remains low, which leads to node (server) starvation in skewed DCN traffic. In this paper, we propose a star-coupler network architecture that scales the capacity and increases bandwidth per server by using multiple and parallel star coupler OCSs (sub-stars). Using a distributed scheduler for each of the stars, we aim to optimize connectivity and resources by using the optimal number of wavelength channels.

II. Proposed Network Architecture

A. Parallel Star Architecture

The physical layer of the proposed *N*-server (1000) OCS-based network has passive star-couplers at its core, each creating a broadcast and select network. To mitigate loss, improve switch efficiency and boost overall network throughput, sub-networks are created by grouping N/x nodes into clusters, as shown in Fig. 1. As in Fig. 1, each source server in the sub-network has x (1, 2, 4, 8 or 16) transceivers which are connected to x sub-stars. Each of the x sub-stars is connected to a different sub-network, connecting the source server to all N destination servers in the network. Every sub-star has a local scheduler, which dynamically computes and allocates wavelength and timeslot. A total of x^2 parallel stars (and x^2 distributed hardware schedulers) are required to completely connect all N nodes, where the size of each star reduces with x. This architecture simplifies the synchronization problem by reducing the number of nodes across which an optical clock must be distributed [4]. Each node sends requests to the relevant scheduler one epoch (synchronous non-tuning communication time) in advance and awaits response before reconfiguring wavelengths at transceivers to communicate in the subsequent epoch.

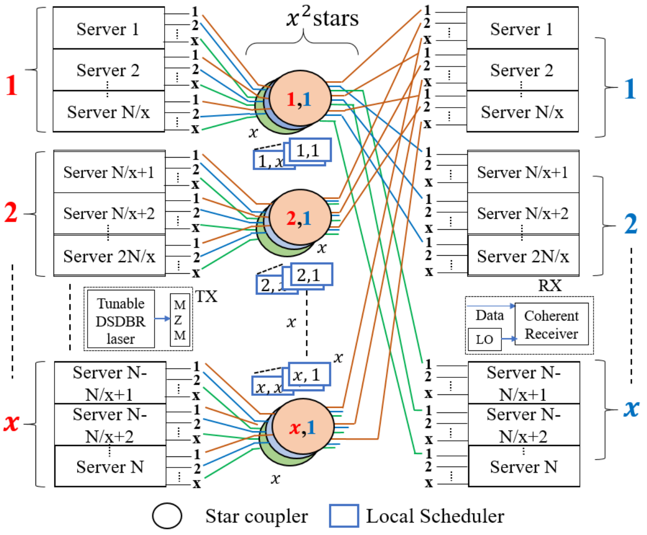

Fig. 1: Proposed switch architecture with star couplers and distributed scheduling

B. Transceiver Technology

Each node is equipped with one optical transmitter per star, comprising a Mach-Zehnder Modulator (MZM) and tunable DS-DBR laser, and one coherent receiver with an independently tunable DS-DBR local oscillator laser. This enables fast wavelength selectivity at both the transmitter and receiver, and the high sensitivity of the coherent receiver enables a larger star [5]. For data transmission across the switch, transmitter and receiver must both tune to the same wavelength . TDM is also adopted to better share the total available bandwidth between the nodes and maximize throughput. The use of DP-QPSK modulation for transmission can achieve a line rate of 100 Gbps (assuming a symbol rate of 25 GBaud). A dedicated synchronous tuning time of 20 ns, based on transition of wavelength channels between transceiver pair shown in [8], is fixed after every epoch to tune all transceivers to their respective scheduled wavelengths.

The work was partially supported by the EU's H2020 funded dRedBox project (grant agreement no. 687632) and EPSRC project TRANSNET (EP/R035342/1).

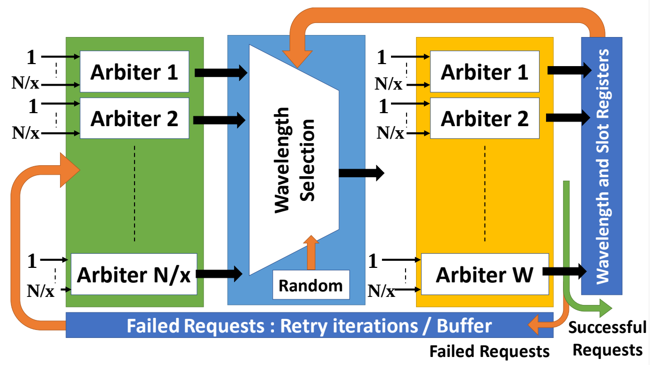

Fig. 2: Scheduler Architecture

C. *Control Communication*

The control network also uses fixed wavelength transceivers for its WDM/TDM sub-star coupler network, with each scheduler having 64 transceivers. Each source server must use a fixed sub-star, wavelength and timeslot to communicate with the scheduler appropriately. A key requirement for both data and control plane communication is slot-level synchronization.

The source sends the request to the relevant scheduler and awaits response. Each request structure contains the destination node and the timeslot size it requires. The scheduler can take up to R requests per server per epoch, as shown in Table 1. The scheduler computes the request and instructs the server with the relevant wavelength and timeslot it must use to communicate its message. Once the grant is issued, the relevant destination and the timeslots the server must use are communicated. The transceivers tune later with an independent tuning matrix, making the network transparent for the servers.

D. *Scheduler Architecture*

We propose the use of distributed parallel hardware scheduler, based on scalable round-robin arbiters [6]. Each scheduler dynamically allocates up to W wavelength channels and up to T timeslots per wavelength for its local star, within the epoch time (1μs in this case). The scheduler deals with N/x requests (one request per server attached in sub-star) in any given clock cycle. Firstly, as shown in Fig. 2, N/x parallel N/x-port arbiters resolve contention between nodes for destination requests. Next, a random seed initiates wavelength selection; a simple feedback logic from registers dominates wavelength selection based on previous allocations. Finally, W parallel N/x-port arbiters resolve contention in wavelength selection (W < Max W in Table 1). At each clock cycle, a maximum of W grants is generated. Once wavelengths are assigned for a source and destination node pair, the scheduler grants as many slots as requested in sequence, based on slot availability. The network targets the transfer of 250 bytes per time-slot, 20ns at 100 Gbps line rate, which corresponds to the overall median packet size across various data centre workflows [7]. The 20ns slot duration corresponds to 50 timeslots per wavelength in a 1μs epoch. Once granted, the timeslot grants are sent to the nodes, while the slot pointers and wavelength registers are updated. At the end of every epoch, failed requests are stored in a 128kB buffer and the switch configuration matrix is sent to the nodes to tune transceivers to relevant wavelengths. The 1000-port hardware was shown to occupy an ASIC area of 52.7 mm^2, when synthesized on 45nm NanGate Opencell CMOS library, meeting timing at a clock period of 7.2ns [3].

TABLE I. NETWORK CHANNEL AND CAPACITY SCALABILITY

Parameter	Values for transceivers per server (x)				
	1	***2***	***4***	***8***	***16***
Sub-star size (N/x)	1024	512	256	128	64
Requests/server /epoch (R)	2	4	8	16	32
Number of sub-stars (x^2)	1	4	16	64	256
Maximum wavelength[a] per sub-star (Max W)	164	81	41	20	10
Total channels (W*x^2)	164	328	656	1280	2560
Maximum Capacity[a] (Tbps)	16.4	32.8	65.6	128	256

[a.] Maximum Unsaturated Channels and Capacity at **Max W** =W/N*star size, where W/N = 0.16

The scheduler algorithm has the following improvements compared to previous work presented in [3]. A novel buffer system is introduced and in every epoch, scheduler gives priority to buffers than requests. The number of clock cycles allocated for buffers are based on the size of buffered requests. To maximize scheduler run-time, control communication is pipelined to avoid idling. The scheduler works back-to-back on subsequent epochs, cutting down latency by half. In the proposed network architecture, as each scheduler needs to only serve one sub-star, more iterations are available for each star.

E. *Capacity*

Increasing transceivers per server (x) reduces sub-star size and maximum unsaturated wavelengths per transceiver (Max W). This reduces tunable laser complexity, permitting a coarser tuning. Another advantage of the sub-star size reduction is that the optical loss is less for smaller sub-stars. Although the permitted wavelengths per transceiver (W) decrease, the number of sub-networks (x^2) increases, increasing the total number of available channels and overall network capacity. The scheduled network throughput saturates if wavelengths per transceiver (W) is increased above a limit (Max W) as mapping requests on different wavelengths creates increased conflicts. This saturation limit was found to be 0.16, after which point the network throughput drops substantially. Table 1 shows how the network capacity scales under this limit.

III. PERFORMANCE ANALYSIS

A. *Simulation setup*

A uniform random request generator sends the demand matrix to distributed schedulers with a 100% network load. Poisson distribution was used to model request arrival intervals with a median of epoch time / R (where epoch time is 1μs and R is the number of requests per server per epoch, as shown in Table 1). In each of the R requests, the nodes can request up to 7 slots (out of 50 slots/channel in 1μs). The scheduled network throughput, for 60μs worth of epochs, was calculated by measuring the network slot usage over slots available. For all epochs, each request generation is stamped with a birth timestamp, attached within each request structure. After every grant, a timestamp including the epoch number and the allocated slot number is stamped. The timestamps have a resolution of 20ns. The latency shown in this work highlights latency without propagation delay. Propagation delay introduces a linear vertical shift to Fig. 3b proportional to fibre length. Finally, for measuring the total network energy, the number of transceiver pairs used by the scheduler and the nodes for different values of x

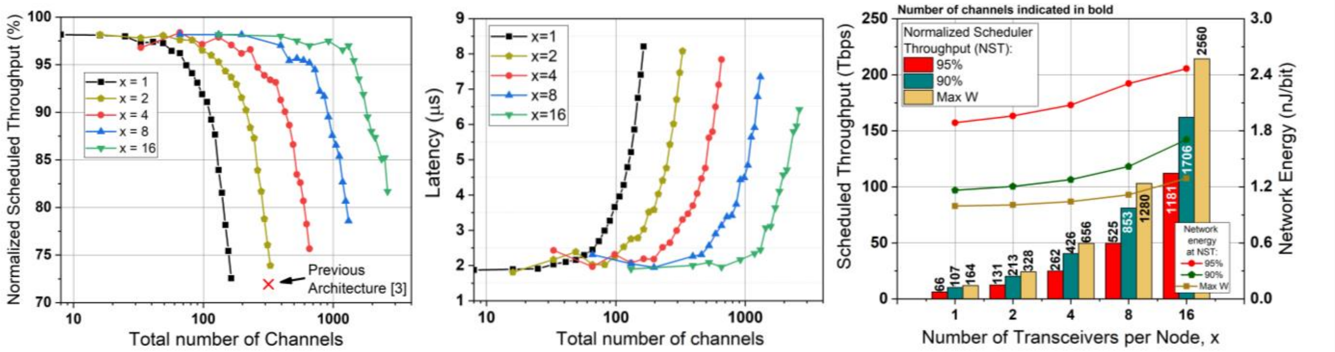

Fig. 3 (a) Channel and Throughput scaling with x (b) Latency as x increases (c) Channels required to maintain 95%, 90% scheduled throughput and using all unsaturated channels (Max W) for different values of x

were calculated. A power estimation of 5.6 W for a fast tunable DS-DBR laser and coherent receiver pair was used. The power consumption of 1000-port pipelined scheduler was estimated by 45 nm CMOS library as 5.2 W; the same value was used for all scheduler sizes in these results to show the worst-case energy/bit, as the network scales.

B. Results

Fig. 3a shows the effect of increasing wavelengths per transceiver (W) and transceivers per server (x) on normalized scheduled network throughput (NST) for a 1µs epoch. By normalized, we mean that the total requested slot size is kept the same as the total slots available across all wavelengths and sub-stars. As wavelengths within a sub-star (W) increases, capacity increases but NST drops due to increased conflicts. As transceivers per server (x) increases, although the maximum unsaturated wavelength that can be supported by a sub-star (Max W) is lower (as shown in Table 1), more channels can be supported as there are x^2 stars. With the aid of distributed scheduling, each of sub-stars are scheduled with a high NST. As marked in Fig. 3a, our previous architecture showed the scheduler serving a 320-channel network at 72% NST. Here we show that up to 82.5% NST is achieved when using all 2560 channels (at x = 16). Using 1000-port sub-stars, x=16 and distributed scheduling, this architecture can potentially scale to support 16000 servers and achieve 72% NST for a 4.2 Pbps network.

Fig. 3b shows the implication of increasing wavelengths per transceiver (W) and transceivers per server (x) on latency. Latency within a sub-star increases as wavelengths per transceiver (W) increase. As transceivers per server (x) increases, latency remains the same for higher number of channels. For all values of x, the latency range is 2.5 - 3 µs at 95% NST and 4 - 4.5 µs at 90% NST. When all channels are connected a worst-case application latency of 9.6µs is experienced, assuming a server-to-server distance of 100m in a clustered network.

The effective scheduled throughput (in Tbps) when using wavelength channels corresponding to 95% and 90% NST are shown in Fig. 3c. For comparison, we also include the scenario where all unsaturated channels (Max W) are used. At x = 16, we show that up to 211 Tbps is achieved with Max W channels. The total network energy consumed per bit, when using network channels corresponding to 95% NST, 90% NST and all unsaturated channels (NST) are also shown in Fig. 3c. As channels increase, the capacity increases, reducing the energy consumption per bit. In the all channel (Max W) case, we show that we have a total network energy consumption of 1.2 nJ/bit.

IV. Conclusion

In this paper, we have showcased a novel star-couple network architecture, which uses multiple WDM/TDM-based OCS sub-stars to scale a 1000-port network to a capacity of 256 Tbps. We have shown that using efficient distributed scheduling and using the write number of wavelength channels per sub-star a scheduled network throughput of up to 200 Tbps (82.5%) consuming 9.6 µs latency at 100% network load for a 1µs epoch. Achieving an order of magnitude higher 'bandwidth per server' compared to previous work, a pure OCS-based network is formed. A network energy consumption of 1.2 nJ/bit is shown for the star-coupler network.

References

[1] S. Cheung, T. Su, K. Okamoto, S.J.B, Yoo, "Ultra-compat silicon photonic 512 x 512 25 GHz arrayed waveguide grating router", IEEE J.Sel.Top. Quantum Electronics, vol. 20, no. 4, pp. 310–316, 2014.

[2] D. Alistarh, H. Ballani, P. Costa, A. Funnell, J. Benjamin, P. Watts, B. Thomsen, "A high-radix, low-latenct optical switch for data centers", SIGCOMM, 2015, pp. 37-368.

[3] J. L. Benjamin, A. Funnell, P. M. Watt, B. Thomsen, "A high speed hardware scheduler for 1000-port optical packet switches to enable scalable data centers," IEEE HOTI, 2017, pp.41-48.

[4] D. H. Hartman, P. J. Delfyett, and S. Z. Ahmed, "Opticalclock distribution using a mode-locked semiconductor laserdiode system," in OFC. OSA, 1991, p FC3K.

[5] A. Funnel, J. Benjamin, H. Ballani, P. Costa, P. Watts, B. Thomsen, "High port count hybrid wavelength switched TDMA (WS-TDMA) optical switch for data centers" OFC, OSA 2016.

[6] N. McKeown, "The iSLIP scheduling algorithm for input-queued switches", IEEE/ACM Transactions on Networking, vol. 7, no. 2, April 1999.

[7] A. C. Funnell, K. Shi, P. Costa, P. Watts, H. Ballani, B. C. Thomsen, "Hybrid wavelength switched-TDMA high port count all optical data centre switch," JLT, vol. 35, no. 20, pp. 4438–4444, Oct 2017.

[8] A. Roy, H. Zeng, J. Bagga, G. Porter, and A. C. Snoeren, "Inside the social network's (datacenter) network," SIGCOMM Comput. Commun. Rev., vol. 45, no. 4, pp. 123–137, Aug. 2015.