Reconfigurable Optical Star Network Architecture for Multicast Media Production Data Centres

Adam C. Funnell[1], David Butler[2], Georgios Zervas[3]

[1] Work performed at Optical Networks Group, University College London, Torrington Place, London, WC1E 6BT. Currently at Multidisciplinary Engineering Education, University of Sheffield, Western Bank, Sheffield, S10 2TN.
[2] BBC Research and Development, The Lighthouse, 201 Wood Lane, London, W12 7TQ
[3] Optical Networks Group, University College London, Torrington Place, London, WC1E 6BT.

Passive optical star networks have attractive properties for multicast traffic in data centres, but are limited in transmission bandwidth per node due to sharing a finite total throughput capacity. By adding reconfigurable switching elements to the core of an optical star topology, simulations show that the expected transmission rate per node can be increased by 26 to 40% (at 90% and 70% network load respectively). The proposed architecture shows no loss of multicast functionality compared to a single passive optical star, and only 7.1% increase in power consumption. Network throughput is shown to be highly dependent on the network traffic pattern, with simulations of multicast zonal media production traffic showing 6 times greater throughput than random or hotspot traffic models.

Introduction

Data traffic generated by data centres is rapidly growing year-on-year, with more than 71% of this traffic remaining inside the data centre [1]. At present, data transfer between the racks of servers in data centres is carried over hierarchical structures of electronic packet switches (EPS). These architectures have high power consumption due to frequent optical-electrical-optical conversion of transmitted data and the large total number of switches required (for example, 400-800 switches for 10-15,000 servers in commercial data centres [2]). EPS networks also have high and inconsistent latency, due to the independent packet buffering queues at each switch, and the possibility of multiple network paths between any pair of servers.

Data centres provide remote transcoding, multi-channel synchronisation and even real-time video editing for live media production operations [3]. A typical data centre located on a broadcaster's premises would have of the order of 1000 to 10000 servers, to include all media ingress and egress points alongside storage, processing and backup nodes. For efficient media production, it is necessary to multicast (transmit in a one-to-many pattern) media across a network, so that a single media stream reaches several destinations simultaneously. It is also required for several media flows to simultaneously reach a single destination, for editorial decisions, comparisons or combined processing. This necessitates many-to-one traffic (e.g. inverse multicast), known as in-cast. The combination of the two flow types results in a complex many-to-many multicast pattern.

Multicast is not deployed in most data centre networks, and this is not necessarily through a lack of desire to efficiently serve multicast traffic patterns, but due to a lack of hardware and software support [4]. Beyond media streaming [5], multicast traffic is also seen in data replication, web cache servicing and virtual machine migration [6]. Algorithms such as Hadoop and MapReduce have multicast phases for data sharing and shuffling [7], as does GoogleFileSystem [8]. However, multicast does not scale across EPS networks due to the limited capacity for multicast addresses in switch forwarding tables, alongside a lack of topological structure to multicast group IP addressing [6]. The lack of support for multicast over hierarchical switch networks means that physical layer multicast is a promising solution.

In [9], a passive optical star network capable of supporting 1024 nodes (individual servers for media storage or processing) at 25 Gb/s line rate using fast tunable transceivers was experimentally demonstrated. Optical passive star networks, such as the design in [9], natively support many-to-

many cast traffic. In [10], scheduling algorithms for the 1024 port passive optical star network were presented, including a design that scheduled traffic in real time with a throughput close to the theoretical optimum. However, although the combination of [9] and [10] showed the feasibility of a single-hop optical star switch supporting 1024 nodes, the worst-case throughput per node was just 2.17 Gb/s despite a 25 Gb/s transceiver line rate. This was due to the fixed and finite total network capacity, limited by the number of wavelengths that could simultaneously pass through the star coupler core.

This paper presents the design of flexibly reconfigurable, high port count optical stars, targeting increased per-node throughput compared to a single large star. Competing solutions for multicast optical networks are presented, to motivate the reconfigurable star design. A method of connecting optical star couplers to make larger stars on demand is presented, and an algorithm for optimum reconfiguration of the couplers is described. The network performance was simulated under two synthetic traffic scenarios and under realistic media production traffic matrices from the network plans of a major national broadcaster. Finally, the power consumption and cost of a reconfigurable optical star network was calculated and compared to competing solutions. The combined results show that the reconfigurable optical star coupler networks is an ideal architecture for multicast data centres.

Prior work on optical multicast

Prior work has developed some network designs which provide all optical multicast. Designs such as light trees for wide-area networks are discounted for data centre applications, as the required number of complex switching nodes (add-drop nodes, wavelength filters and converters etc.) would be impracticably high in a data centre with > 1000 endpoints [11]. The remaining candidate architectures can be broadly grouped into three categories:

(1) Photonic gadgets via optical circuit switching (OCS)
(2) Broadcast and select matrices
(3) Passive optical star

A typical photonic gadget network is shown in Fig. 1a. Using OCS, photonic gadgets can be switched in and out of the optical paths to adapt the network physical layer. These gadgets can include power splitters (to support multicast), power combiners (in-cast), wavelength filters (add-drop) and amplifiers (increased transmission distance). For the flexible multicast applications targeted by this work, only gadgets based on power combiners and splitters are considered in Fig. 1; any wavelength filtering elements would restrict the flexibility of multicast traffic patterns. Algorithms can determine the most efficient gadgets to deploy to meet the desired traffic pattern; by moving complex patterns such as multicast into the optical layer, connectivity can be provided more efficiently than by using hierarchical EPS structures. Examples of multicast photonic gadget switches include many hybrid optical-EPS architectures [12, 13, 14, 15].

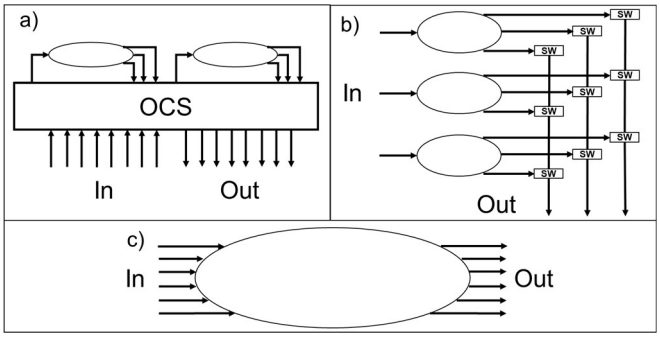

Figure 1: a) Photonic gadget/optical circuit switching (OCS) architecture. Each oval represents a 1:N passive optical splitter; b) Broadcast and select architecture, where each oval represents a passive optical splitter and each rectangle marked “SW” represents a transmissive or blocking optical switch; c) Passive star architecture, formed from a single N:N passive optical star.

Once a photonic gadget is placed into an optical circuit, network routing cannot be reconfigured without interrupting the flow through the gadget. If network requirements change, such as more receivers joining an existing multicast group, the multicast flow must be paused to reconfigure the gadgets. This makes gadget-based architectures unsuitable for live media production - continuous streams of media cannot be paused or interrupted, as this would cause packet buffering and audio-visual break up. Furthermore, photonic gadget designs have a hard limit to both the number and size of the multicast groups that can be supported, dependent on the number and size of the splitter/combiner gadgets available at the OCS.

An example photonic gadget network is the hybrid EPS/OCS network, a competing solution to supporting optical multicast [14, 16]. The hybrid EPS/OCS network design directly connects each network node to both an EPS network and an optical circuit switch (OCS) network. It fully supports all-optical multicast traffic flows, using star couplers inserted into optical-end-to-end paths via a MEMS switch. The hybrid network structure is shown in Fig. 2.

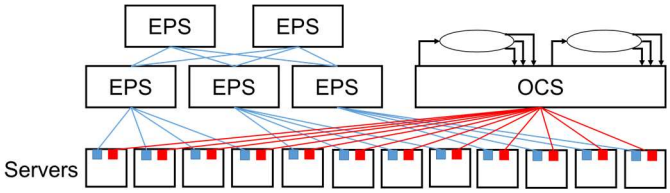

Figure 2: A hybrid EPS/OCS network supporting optical multicast, as described in [14, 16].

The second competing architecture is based on broadcast and select matrices, as shown in Fig. 1b. Input signals are split to multiple outputs, and a switch at each output either transmits or blocks the signal. The switches could be semiconductor optical amplifiers (SOAs), Mach-Zehnder interferometers, or MEMS switches to either transmit (“select”) or block the optical path. Examples of this architecture include [16, 17].

Broadcast-and-select is highly flexible and permits arbitrarily sized multicast groups. However, the physical layer requires either multiple receivers per node (one per transmitter), or power combiners to permit fully flexible in-cast. Neither option is practically feasible: multiple receivers per node is impractical in a network with large numbers (e.g. > 1000) of transmitters; and power combiners would result in double the total power split loss across the network (e.g. a 1024 way power split incurs 30 dB power loss, and double this is 60dB, which is too high for low complexity optical communications).

The final competing architecture is the passive optical star coupler network, shown in Fig. 1c. Using a passive optical star, a data centre network can scale to a high node count (> 1000) and can minimize latency by providing single-hop connectivity between all nodes. Passive star networks with independently tunable transmitters and receivers allow multicast and in-cast groups of any size, up to a single group containing all network nodes. Strong resilience to failure is also observed in passive star networks, since the network core contains no active electronic or mechanical components. Failures therefore only affect individual transceivers, minimizing and localizing network downtime.

A passive optical star permits both Wavelength Division Multiplexing (WDM) and Time Division Multiplexing (TDM), resulting in a hybrid WDM-TDM network design. An example WDM-TDM network was shown in [18]. All network nodes were equipped with fast tunable transceivers which reconfigured every 2 μs (defined as an epoch) to set up wavelength paths through the star. To share the total available bandwidth with fine granularity, TDM was also used to split the 2 μs epochs into timeslots, allowing multiple transmitters and receivers to share a wavelength allocation during each Epoch.

However, when using a single passive optical star to connect many transceivers, there is a limit to the maximum total network throughput. This is the product of the number of distinct wavelengths that the system can allocate (bounded by the tunability range of the transceivers), denoted W, and the maximum line rate of each transceiver, denoted B_{LR}. The maximum total throughput (denoted T_{MAX}) is therefore $T_{MAX}=WB_{LR}$ (1). For a total number of network nodes, denoted N, and assuming an equal share of the total throughput for every transmitter, the expected transmission bandwidth per node is

$B_{EQUAL}=T_{MAX}/N$. When substituting in T_{MAX} from Eqn. 1, the expected bandwidth is therefore $B_{EQUAL}=WB_{LR}/N$ (2).

Once the number of nodes N increases above the number of wavelengths W, the expected bandwidth per node rapidly decreases. Even if 100 Gb/s transceivers are used at every node in an N=1000 node network, each transmitter is allocated below 10 Gb/s if the bandwidth is shared equally, for W≤100.

To increase the expected bandwidth per node across a passive star for a fixed number of nodes N, it is necessary to increase either the number of available wavelengths or to increase the transmitter line rate. However, laser designs with fast switching properties across a tuning range wider than the optical C band (1530-1565 nm) are difficult to fabricate. To increase the transmission line rate beyond 25 Gb/s requires high-speed and high bandwidth electrical drive circuitry to run on-off keying at a higher baud rate, or coherent transmission and receiver digital signal processing (DSP) to transfer multiple bits per symbol. All of these enhancements incur increased cost and system complexity.

A similar problem was studied as the drop-and-waste principle of passive filterless networking [19], where all parts of a wide-area fibre network share all wavelengths. Passive filterless networks transparently pass all wavelengths, even those which only contain data that is not useful for downstream nodes, resulting in bandwidth wastage. Programmable and reconfigurable filterless networks were proposed to reduce the drop-and-waste problem and in turn increase throughput [20]. However the increased complexity of the hardware at each node is impractical for data centre networking, where the node count is high and the network connectivity demands change frequently.

Early work in reconfigurable star architectures included the development of a wideband all-optical network based on hierarchical layers of optical couplers [21]. The network design in [21] showed all-optical routing, multipoint-to-multipoint communication, and splitting of the wider network into sub-sections to increase wavelength utilisation across the whole network. However, full flexibility to create multicast groups of any size was impossible due to both the fixed wavelength allocations to each layer and the use of wavelength routing.

The design requirements, which are better met by the system proposed in this work than competing solutions, are summarized as follows: reconfigurable multicast group sizes; arbitrary group sizes and total number of groups; a single transceiver per node; and increased total network throughput compared to a single passive star.

System design and operation

To enable an increased number of available wavelengths across the star coupler core, this work proposes the flexible reconfiguration of a large port count star into physically separated sub-stars. Each sub-star can independently allocate wavelengths from all other sub-stars, potentially increasing the total network throughput by the number of sub-stars formed.

The design consists of two layers of optical star couplers. To connect N nodes in total, a layer of "input" couplers, each of port count $\sqrt{N} \times \sqrt{N}$, are connected to a layer of "output" couplers (also $\sqrt{N} \times \sqrt{N}$ port count). Every input coupler has a connection to each output coupler via a switch that can "transmit" or "block" the optical signals. If all input couplers and output couplers are connected in the "transmit" state, the network reverts to a single large star. A toy example network supporting $N = 9$ nodes is shown in Fig. 3. Each input coupler connects $\sqrt{N} = 3$ transmitters to the network, and each output coupler connects $\sqrt{N} = 3$ receivers to the network. Given that each input coupler has one connection to each output coupler via a block/transmit switch, a total of N switches are required in the central switching cross connect (e.g. $N = 9$ switches in Fig. 3).

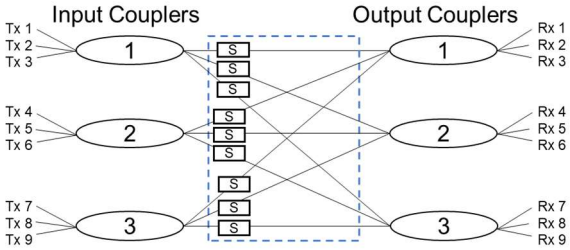

Figure 3: An example of the full connectivity required for $N = 9$ nodes. Each coupler connects $\sqrt{N} = 3$ transmitters or receivers into the network. The central optical cross-connect is inside the dashed box, where N optical switches (denoted "S") are required to connect N nodes.

The central switches can separate couplers into distinct sub-stars, to allow wavelengths to be reused across the network. Figure 4 shows some possible transmission patterns across a dual-layer star system. Figure 4a shows the worst-case scenario, where all input and output couplers are connected together, effectively forming a single large star with N ports. In this case, no wavelengths can be reused and the expected throughput per node is limited to Eqn. (1). In Fig. 4b, there is complete separation between two distinct, isolated sub-stars i.e. input couplers 1-3 only connect to output couplers A and B, and input couplers 4 and 5 only connect to output couplers C-E. The full range of wavelengths can be completely reused in the two distinct sub-star groups, doubling the overall maximum throughput compared to a single large star. Figure 4c shows a more complex mesh of connectivity, however, there are potential groups where wavelengths can be re-used. It is therefore essential to analyse the required connectivity across the full network, to ensure that all separable groups are identified and separated using the central switches, to maximise overall throughput.

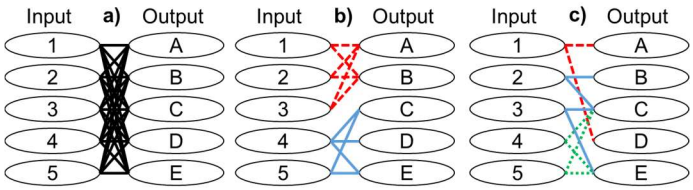

Figure 4: a) A fully connected mesh of input and output couplers – this arrangement would not permit wavelength reuse and results in the worst-case equal bandwidth allocation. b) A mesh of smaller couplers with separation between two distinct groups. c) A complex mesh of input and output couplers.

An ideal switch for this application would provide binary switching between transmission and blocking of optical signals at all wavelengths, but common optical switching technologies are not suitable for this application. Semiconductor optical amplifiers (SOAs) do not perform well when loaded with multiple wavelengths due to non-linear interference and cross-talk [22]; and Mach-Zehnder interferometers are polarisation sensitive - impractical to achieve at data centre scale due to expensive polarisation maintaining fibre and couplers.

Two alternative technologies are thus proposed for the central optical cross-connect: a micro-electro-mechanical system (MEMS) fibre circuit switch, or an array of acousto-optic modulators (AOMs). MEMS switches have the advantage of proven technology and integration with SDN workflows to provide integration with a scheduler, but would only be suitable for long-lived flows due to their slow reconfiguration time (µs to ms) which would reduce throughput if used to reconfigure the network more frequently than a s to ms timescale. AOMs have fast switching capability (down to 35 ns) and low insertion loss (down to 1.6 dB for AOMs compared to 3 dB for MEMS units) [23] and thus could be reconfigured at packet level timescales. A possible downside to AOMs compared to MEMS is their high power consumption (~10 W per AOM switch compared to < 1W per port in a MEMS device). An array of AOMs scaling to $N = 1000$ ports would require a complex custom design of oscillators and RF amplifiers, but this could be designed as a convenient single circuit to serve all N AOMs and may provide improved power efficiency.

In previous work on the physical layer of a single star supporting 1000 nodes, in [9], 4.6 dB of system power margin was available. This margin is more than sufficient to incorporate the additional 1.6 dB loss of AOM switching elements. The network in [9] was designed for a single data centre building, with maximum fibre lengths of up to 1 km. The power variation between nodes connected to the network by differing fibre lengths is therefore limited to approximately 0.2 dB. This could be overcome using variable current supplies to the SOA integrated on each tunable laser transmitter, which would vary the transmitter optical power outputs to equalise the power at the receiver.

The reconfiguration may not be hitless, depending on the switching hardware used in the central optical switches. Individual AOM switches could be hitless, as switching them from the block state to the transmit state would not affect any other connectivity. However, high port count MEMS switches may incur crosstalk between channels during switching. Switch reconfigurations should therefore be scheduled to occur simultaneously in defined timeslots. To accommodate this switching, buffering should occur at the network nodes where necessary to maintain synchronous switching timeslots when no data is transmitted. Each node can then adjust the media streaming quality based on the overall effective transmission rate, including the switching downtime. This would require all network nodes to maintain a synchronised clock, which could be provided using a parallel star coupler infrastructure to broadcast a time locking signal.

It is desirable for network switching downtime to be less than 10% of network transmission time. For the MEMS hardware design which can switch within 30 ms, this corresponds to reconfiguring the sub-stars on second to minute timescales, to optimize throughput at a flow level. It is envisaged that this network serves only media flows with durations of seconds, minutes or hours, with no bursty traffic. Any control information, and small packets of media metadata, can be carried by the parallel control plane network. For the AOM hardware design, which can switch within 200ns, the star splitting optimization could be performed every 2 μs Epoch for packet-level reconfiguration at the same rate as the wavelength retuning in [9]. The two approaches are functionally identical when considering only the impact of traffic pattern locality on network performance, as in the simulations in this work.

An algorithm was constructed to meet a single objective: given the required connectivity mapping between input and output couplers, partition the network into the maximum possible number of sub-stars, and determine the input coupler membership of all sub-stars, and any wavelength reuse feasible within each sub-star. The algorithm is summarised in Fig. 5.

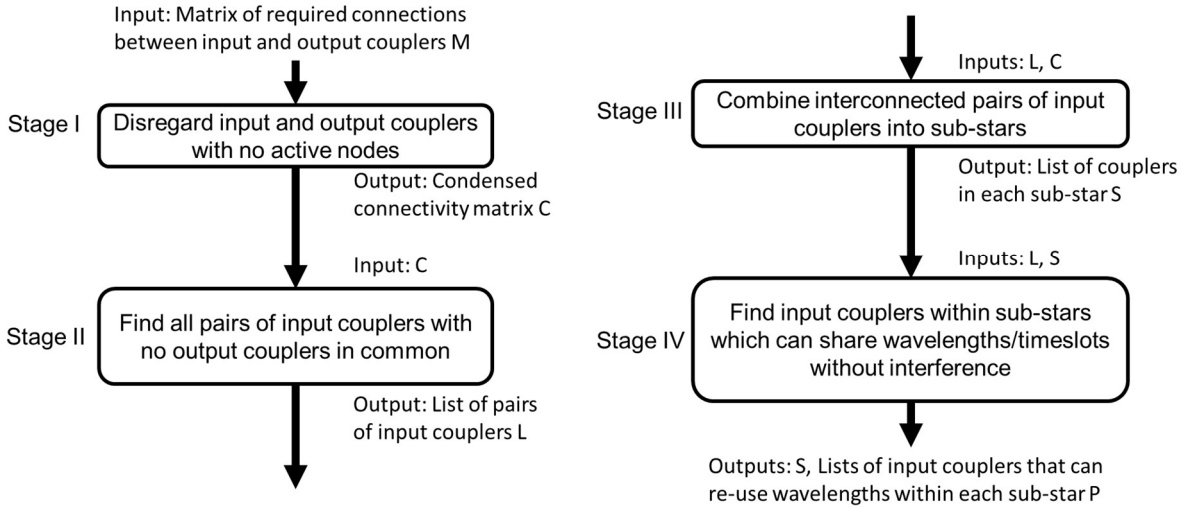

Figure 5: Summary of the sub-star formation algorithm.

An example network under light load is shown in Fig. 6, with the required connectivity between couplers shown at the centre of the network map. This example includes two distinct sub-stars (the blue solid lines including input couplers 1 and 2, and the red dashed lines including input couplers 3, 5

and 6), alongside an unused input coupler (4) and the potential for wavelength re-use within a star coupler (couplers 3 and 6 can be allocated the same wavelengths).

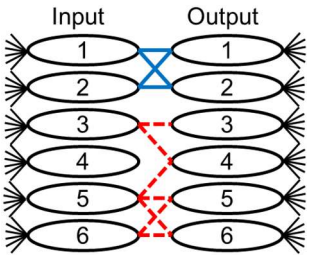

Figure 6: Example connectivity across a split-star network supporting $N = 36$ nodes using two layers each containing $\sqrt{N} = 6$ couplers.

By following the algorithm summarised in fig. 5, the following key results are obtained for the network shown in fig. 6:

Algorithm input:

$$M = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Stage I output:

$$C = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Stage II output:

$$L = [1,3 \quad 1,5 \quad 1,6 \quad 2,3 \quad 2,5 \quad 2,6 \quad 3,1 \quad 3,2 \quad 3,6 \quad 5,1 \quad 5,2 \quad 6,1 \quad 6,2 \quad 6,3]$$

List L contains all possible pairs of input couplers with no output couplers in common, but this would not build up a full picture of sub-star connectivity. For example, naïve inspection of L suggests that input couplers 1 and 3 should be in separate sub-stars (no output couplers in common), as should input couplers 3 and 5. This would suggest S has at least three member sub-stars, containing 1, 3 and 5 respectively. However, C shows that input couplers 3 and 5 have an output coupler in common (coupler 4), so they must be part of the same sub-star; S must include [1] and [3,5]. Stage III of the algorithm analyses all possible pairs of members of list L, alongside C, so that a full map of sub-star connectivity can be constructed.

Stage III output:

$$S = [1,2]; [3,5,6]$$

Stage IV output:

$$P = [3,6].$$

Once both the sub-star members (S) and the potential to share bandwidth within sub-stars (P) are known, an independent scheduler allocates wavelengths and timeslots to the requests. The scheduler is out of the scope of this work as wavelength/timeslot allocation has already been explored in [10]. For this simulation work, transmission bandwidth is allocated following two simple principles:

1. All requested flows are always guaranteed connectivity.
2. All flows across the same sub-star will be allocated the same bandwidth.

However, there is no barrier to introducing prioritisation and flow control techniques in future, using existing techniques well-studied elsewhere to share bandwidth unequally [24, 25].

The limit to the total throughput across any star coupler network is the total number of wavelengths (W) that can be used simultaneously without interference. The total maximum throughput capacity is therefore WB_{LR}. Each sub-star is effectively separated from the other sub-stars, and can allocate the

full range of wavelengths. If n sub-stars are formed, each can allocate the full range of wavelengths, giving a total throughput of $nB_{LR}W$. The more sub-stars that are formed, the higher the total throughput, up to $\sqrt{N}B_{LR}W$.

Traffic scenarios

The network performance was simulated for three network traffic scenarios, summarized in Table I. In the first scenario, the network comprises 1024 nodes. A traffic matrix was generated containing randomly matched transmit and receive pairs of nodes. The traffic matrix included multicast requests i.e. multiple receivers requesting the same transmitter.

TABLE I
Summary of the three scenarios studied

Scenario	Source nodes	Destination nodes	Traffic Mapping
1: Random	1024	1024	Random source-destination pairing
2: Hotspot Cluster	1024	1024	Destination 50% likely to request 1 of 102 sources in the hotspot cluster, 50% likely to request 1 of 922 other sources
3: Live Media Production	930	1260	Destinations and sources split into 5 zones, 5x5 traffic matrix between zones

In the second scenario, the network comprises 1024 nodes, with 10% of the nodes designated to be part of a “hotspot cluster”. A traffic matrix was generated such that every transmitter has a 50% probability of sending data to any of the 102 nodes in the hotspot cluster, and a 50% probability of sending data to any of the 922 nodes elsewhere in the data centre. This is typical of some data centre traffic patterns such as front-end web services and MapReduce clusters [26]. Multicast traffic was also included i.e. each transmitter could send to multiple receivers both inside and outside the hotspot cluster.

In the third scenario, the network comprises 1260 nodes, of which 930 are transmitters of data flows and all 1260 are receivers. Following the network commissioning plans of a live broadcast production centre, the nodes are split into two main groups (audio and video), and within those two groups, split into five further zones based on their function (TV, Radio, Production, Playout and Core). A traffic matrix was produced using real data centre measurements of the likelihood of connectivity requests within and between zones. A video receiver will only ever be paired with a video transmitter, and similarly for audio. Multicast was again included in the scenario i.e. multiple receivers can request the same transmitter.

Additionally in the third scenario, connection requests followed a practical production workflow. All nodes were initially unconnected to the network (or physically connected but inactive). Transmitters and receivers were activated one-by-one at random. Transmitters joining the network were added to a list of available media sources. When a receiver was added to the network, it requested a connection to any available source within a zone chosen based on the traffic matrix. The sub-star splitting algorithm was run after sufficient transmitters were activated to reach the specified load level.

Simulation results

Each network node was simulated with an optical transmitter line rate of 25 Gbit/s and a total tunable range of 120 wavelengths. The effective transmission rate per transmitter is governed by the number of transmitters attached to the sub-star. A sub-star with a large number of transmitters connected (> W) will require multiple transmitters to share each wavelength, using a granular timeslot approach and a TDM controller [9]. When a sub-star contains more than W transmitters, multiple transmitters must share each wavelength, but when a sub-star contains fewer than W transmitters, each transmitter can transmit at the full line rate (25 Gb/s in this work). The expected transmission rate per node takes account of this wavelength sharing when calculated, by dividing the total transmission capacity of each sub-star by the number of transmitting nodes attached to the sub-star.

If the network splits into sub-stars each with greatly different numbers of nodes (e.g. a sub-star of 950 nodes and a sub-star of 74 nodes), the expected transmission rate per node will be vastly different in each sub-star (e.g. 3.16 Gbit/s and 25 Gbit/s in this example). It is only realistic to calculate the median transmission rate per node across the whole network, as the mean can be unfairly distorted by small sub-stars (e.g in this example the mean is 4.74 Gbit/s while the median is 3.16Gbit/s).

The achievable throughput depends entirely on the traffic pattern across the network, which directly determines the transmit/block states of the central switches which form the connectivity pattern input to the algorithm. However, for a 1024 node network, there are 2^{1024} possible combinations of switch states. The simulations presented here therefore take a Monte Carlo approach, sampling 10,000 possible connection loads per traffic scenario. The simulation was written in Matlab to consider only connectivity requests at flow level, to assess only the impact of the traffic pattern on the network performance. Assessment of the control plane latency and the performance for non-uniform flows are left for future work. The potential throughput increases quoted in the subsequent text give the average of the median transmission rates per node for each traffic scenario.

The simulation results are presented as probability density functions (PDFs) of the median transmission rate per node, in Figures 7-9. The PDF shows the likelihood of each median transmission rate across the whole network, for each network load level simulated. For example, if the PDF curve has a value of 0.5 at 7.6 Gbit/s, there is a 50% chance that the median transmission rate of all transmitters on the network will be 7.6 Gbit/s.

Throughout this work, “load” denotes the number of nodes that are active within the network. For example, at 50% network load in scenario 1, only 512 of the 1024 nodes are actively transmitting a data flow. Although load is usually measured as a function of data transmitted per node over time, in the live media applications targeted by this work, data flows last for minutes to hours. Flow completion time, which measures the time taken for individual data transactions between nodes, is only a useful metric when each individual flow transaction is of short duration, rather than a continuous real-time stream [27]. Flow completion time is therefore neglected as a metric in this work, in favour of the number of simultaneously active nodes.

The performance of the network under traffic scenario 1 is shown in Fig. 7. When 90% or 100% of the network nodes were connected across the network, it was not possible to split the larger star into any smaller sub-stars. This is due to the lack of locality in a uniform random traffic pattern, meaning that clusters do not form which the star can split to accommodate. For 30% load the median transmission rate per node increases by 15.4% on average compared to a single star. The tail to the right of the PDF peak for both 50% and 30% load shows that small increases to transmission rate per node are possible but thoroughly improbable, particularly for 50% load.

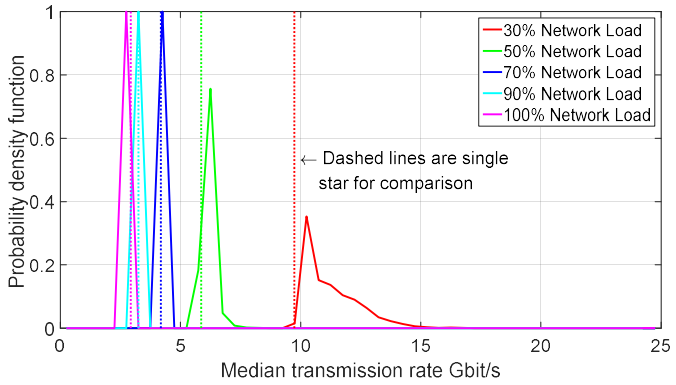

Figure 7: The median transmission rate per node for scenario 1, random traffic.

Figure 8 shows that the performance of the network under hotspot traffic is slightly worse than under the random traffic of scenario 1. For network traffic loads of greater than 50%, there is no advantage to using a reconfigurable star network compared to using a single large star. At 50% network load, only 8% of trials showed any improvement in median transmission rate per node, which is less probable than in the random traffic scenario. For 30% network traffic load, there is only on average a

12.7% increase in transmission rate per node. These results are due to the high likelihood of multiple input couplers connecting to the same output coupler(s) where the hotspots are located. This links several input couplers into the same sub-star, meaning that wavelengths cannot be reused within that sub-star and no improvement in bandwidth can be achieved.

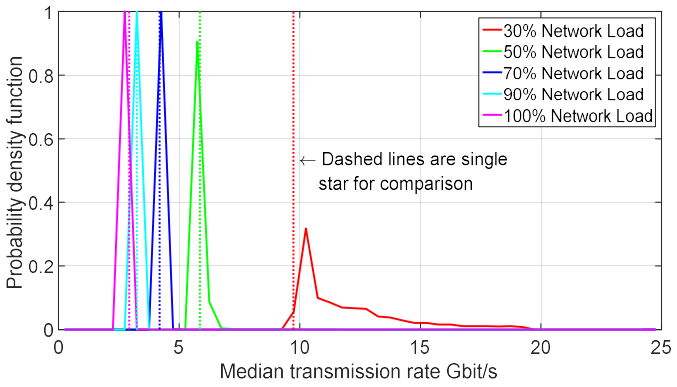

Figure 8: The median transmission rate per node for scenario 2, hotspot traffic.

Improved throughput could be achieved by spreading the “hotspot” over several output couplers, reducing the contention at the single (or few) output couplers which make up the hotspot. However, this would require prior knowledge of the hotspot location when cabling the network; this is not feasible when demands change over time [26].

Zonal media production traffic, modelled in scenario 3, showed the greatest increase in transmission rate per node compared to a single large star. Figure 9 shows that for even 90% load, the expected bandwidth per port increases by 26% on average compared to that expected from a single large optical star. For 70% load, the expected bandwidth per node increases by 48% – on average the median bandwidth simulated per port is 6.8 Gb/s, compared to the 4.6 Gb/s expected for a single large star.

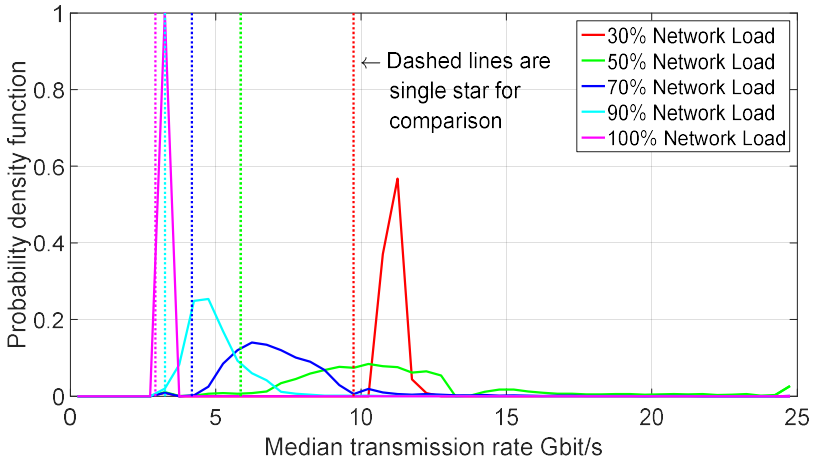

Figure 9: The median transmission rate per node for scenario 3, zonal media production traffic.

The large increases in throughput per node were due to the zonal traffic pattern, with each zone more likely to communicate with some zones than others. When traffic is clustered (i.e. zonal), if the input and output couplers are also aligned as far as possible with particular zones, it is more likely that the star can split into multiple sub-stars, increasing the overall network capacity.

To emphasise the impact of the traffic pattern on the effectiveness of this network design, the three traffic scenarios simulated can be compared. Figure 10 shows the number of sub-stars that are formed across the network, for each traffic pattern scenario and network load. Zonal traffic consistently forms more than twice as many sub-stars on average, which in turn results in a higher total throughput. For high network loads the uniform random and hotspot traffic scenarios only form a

single star, and random traffic gives only marginally better performance than hotspot traffic under low loading.

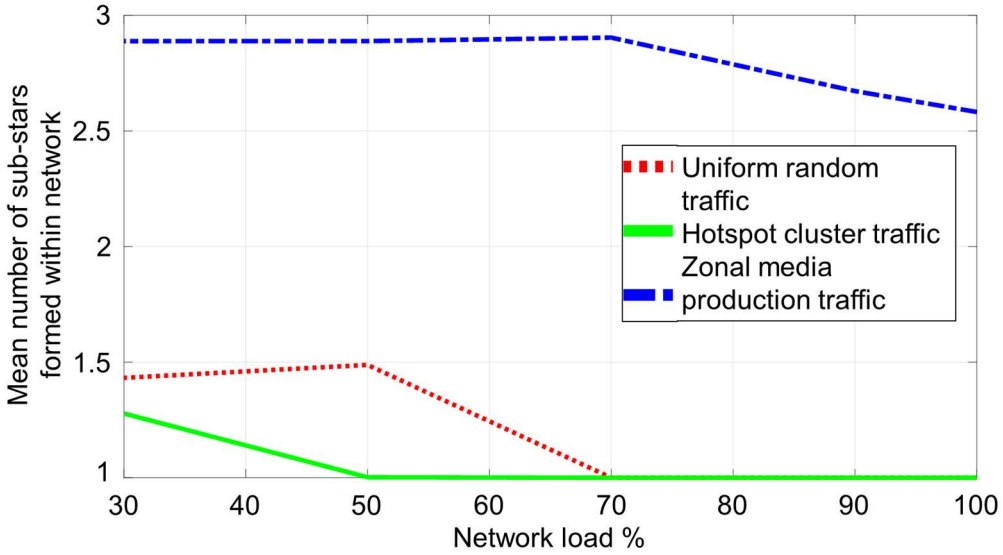

Figure 10: The mean number of sub-stars formed in the network for each traffic scenario.

The mean number of nodes per sub-star across all network simulations is shown in Fig. 11. The media production zonal traffic consistently produces a lower number of nodes per group compared to the other two traffic scenarios, which in turn implies increased transmission rates per node. For zonal media traffic, the mean number of nodes per sub-star across all network load percentages was 370. This implies an expected transmission rate of 8.1 Gb/s per node, an increase of 273% compared to a single large star. For high network loads above 70%, the uniform random and hotspot traffic scenarios show 1024 nodes per sub-star, matching a single large star as described in figures 7-9. For network loads of 70% or below, random traffic again performs slightly better than hotspot traffic, with on average 40.8% fewer nodes per group. This is due to the hotspot nodes being in high demand for connectivity, and large clusters of inputs forming around a few hotspot outputs.

The combination of Figs. 10 and 11 demonstrate that the data centre traffic pattern has a large impact on the effectiveness of this network topology. Considering all network loads simulated, the average throughput increase is over 6 times larger for zonal media traffic relative to random or hotspot traffic.

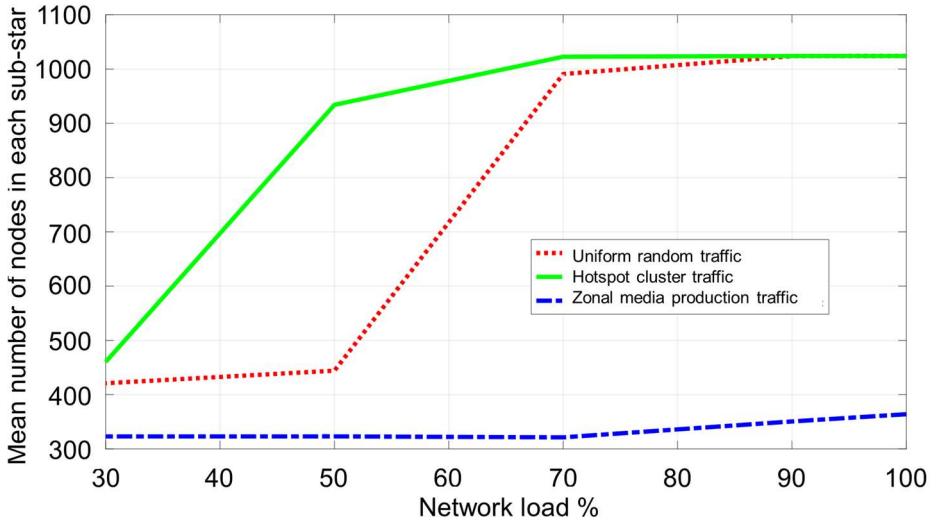

Figure 11: The mean number of nodes per sub-star for each traffic scenario.

Energy consumption and cost comparison

Conventional data centre networks comprising hierarchical layers of electronic packet switches consume vast amounts of energy. Each switch can consume several hundreds of Watts, and many switches are required to create a full data centre scale network. To ensure full redundancy of the network should any key components fail, multiple switches are often used in key locations. For media production networks, every active element is duplicated and a robust protocol can establish instant recovery in the event of equipment failure [5].

Reducing the number of switches in the network can reduce the total energy consumption. In the optical network design presented here, far fewer electrical switches will be required than in a Clos or Leaf-Spine EPS architecture. Only the control plane of this architecture requires electrical switching, and although sufficient switches must be used to connect all N nodes, each control plane switch can run at a lower data rate than the data plane. This reduces the power consumption drastically due to the lower speed electronics required.

A comparison is presented in Table II of the power requirements of this proposed network alongside the incumbent hierarchical EPS network design, a single passive star network, and a hybrid EPS/OCS network capable of limited multicast transmission, such as in [28]. The network design and layout of the real-world media network described in simulation scenario 3 was evaluated for each network architecture i.e. 1260 nodes co-located at a single facility. All packet switched connections in EPS networks were assumed to use active optical cables (AOCs), consuming 3.5 W per leaf-spine connection and 1 W per node-leaf connection. The increased complexity of the tunable transceivers for the optical star system was wrapped into a 5.6 W power consumption per transceiver, to include all electronic components necessary for the experimental setup in [9]. Two transceivers per node were required for redundancy, as outlined above.

TABLE II - Comparison of energy consumption and cost of the proposed network to EPS, passive star and EPS/OCS hybrid networks

Proposed Split Star Network	Power (W)	Cost ($1000)	EPS Network	Power (W)	Cost ($1000)	Single Passive Star Network	Power (W)	Cost ($1000)	Hybrid EPS/OCS [14,16]	Power (W)	Cost ($1000)
CORE: 8 x optical circuit switch (OCS) [29]	1600	1405	**CORE:** 2 x 72 port spine switch [32]	7534	459	**CORE:** 46 x control plane packet switch [30]	5980	885	EPS Network	33577	1568
CORE: 46 x control plane packet switch [30]	5980	885	**CORE:** 36 x leaf switch [33]	23040	977	**ACCESS:** 2520 x fast tunable optical transceiver	14112	1134	OCS Network (includes 2520 AOC transceivers [31] and 20 OCS units [29])	6520	3891
ACCESS: 2520 x fast tunable transceiver	14112	1134	**CORE:** 138 x leaf-spine AOC [34]	483	21	**ACCESS:** 2520 x control AOC [31]	2520	111	SDN Controller	1000	1
ACCESS: 2520 x active optical cable (AOC) [31]	2520	111	**ACCESS:** 2520 x node-leaf AOC [31]	2520	110						
TOTAL:	**24212**	**3535**	**TOTAL:**	**33577**	**1567**	**TOTAL:**	**22612**	**2130**	**TOTAL:**	**41097**	**5460**
			Compare to split star:	**+38.7 %**	**-55.7%**	**Compare to split star:**	**-6.6 %**	**-39.7%**	**Compare to split star:**	**+69.7 %**	**+54.5%**

Although Table II shows an 7.1% increase in power consumption to upgrade from a single passive star network to the reconfigurable switching design presented in this work, the potential for at least 28% increased network throughput means that the power consumption per bit transferred will be lower overall. A reconfigurable optical star system compares favourably to an EPS network which would use 39% more energy, since the high power consumption of all switches in an EPS network far outweighs the increased power consumption of tunable transceivers.

Comparison is also made in Table II to a hybrid EPS/OCS solution, such as [14] or [16] as described in the Prior Work section. Hybrid EPS/OCS network architectures are attractive for providing fast packet switching alongside some limited multicast support. However, hybrid network designs requiring two independent data plane implementations would consume 70% more power than the reconfigurable split star network presented here. This is assuming that a full EPS network was constructed alongside a reconfigurable optical plane for multicast traffic i.e. each node has two independent transceivers, although neither transceiver was fast tunable. The need for both a full EPS and OCS network (including overprovision in the port count of the OCS to enable flexibility in the multicast group sizes) resulted in high total power consumption. Although the split-star network targets only all-optical multicast of long duration media production flows, the hybrid EPS/OCS network allows highly variable and bursty traffic alongside longer flows. The increased power consumption of the hybrid EPS/OCS may be worthwhile dependent on the traffic pattern required across the network; bursty traffic could make efficient use of the EPS network plane.

A cost comparison between network architectures has also been performed. Using the unit prices and switch specifications in [35], Table II shows approximate costs for each topology. To account for the increased cost of fast tunable transceivers incorporating DSP-free coherent receiver hardware (matching the experimental demonstration in [9]), transceivers for the star networks are estimated to cost three times single wavelength transceivers.

A single passive star network costs 39.8% less than the split star network presented in this work, but a hybrid EPS/OCS network as in [14, 16] would cost approximately 54.5% more than the split star. The split star cost is dominated by both the transceivers and requiring multiple optical circuit switches (4 OCS units of 320 ports each are required to provide sufficient switching at the centre of a split star, and each unit is 76.5% of the cost of an EPS network spine switch). Neither of these devices are currently available as volume products and future prices are likely to decrease once adopted in production data centres. However, the hybrid EPS/OCS architecture is better suited to bursty and short-lived flows than the split-star design presented here, and could allow multicast over both the EPS and OCS network planes. The increased capital expenditure could be justified by the increased flexibility offered of high throughput across a range of different network traffic scenarios.

Conclusion

To support the increased demand for multicast network capacity within data centres, it is necessary to move away from networks comprising hierarchical layers of electronic switches and use optical end-to-end networks with fewer but higher port count switches. This work has presented a reconfigurable optical star network architecture which can achieve increased bandwidth per node compared to a single, large optical star while maintaining flexible multicast capability.

Increases in per-node throughput compared to a single large star of 26-40% can be achieved for applications such as live media production centres, where traffic follows a zonal pattern, allowing the network to be partitioned. However, there is no significant increase in the throughput in data centres with uniform random traffic or hotspot traffic compared to a single large optical star. Zonal traffic is common in the application areas that this work targets, and a reconfigurable optical star network could increase the available bandwidth per transceiver compared to a single large optical star, even at 90% network load.

Potential future work could assess the achievable control plane latency, comparing both hardware and software control plane solutions, and compare this to any latency introduced through optical element reconfiguration time. Additionally, the physical distribution of nodes to each input and output star could be optimized, or even reconfigured on-demand, to maximize the potential network

throughput. There is further scope for applying machine learning techniques to analyse the applications and services running across the network, and to suggest physical distribution of nodes or allocation of application tasks to particular nodes for optimal use of the available network bandwidth.

Acknowledgment

This work was supported by an Industrial Fellowship from the Royal Commission for the Exhibition of 1851, the UCL-Cambridge Doctoral Training Centre in Photonic Systems Development, and BBC Research and Development. This work is a continuation of a project initiated with Microsoft Research, whom the authors thank, particularly Benn Thomsen, for their continued support. The authors additionally thank Polina Bayvel for discussions, proof-reading and her support of this work.

References

[1] Cisco Global Cloud Index: Forecast and methodology, 2016-2021, Cisco Systems, Inc., 2018.
[2] Benson, T. et al., “Network traffic characteristics of data centers in the wild.” In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, 2010, pp. 267-280.
[3] Brightwell, P.J. et al., “The IP Studio”. SMPTE Motion Imaging Journal, 123(2), 2014, pp.31-36.
[4] Diot C. et al., “Deployment issues for the IP multicast service and architecture,” IEEE Network, vol. 14(1), 2000, pp. 78–88.
[5] Kojima, T. et al., “A Practical Approach to IP Live Production.” SMPTE Motion Imaging Journal, 124(2), 2015, pp. 29-40.
[6] Li, X. and Freedman, M. J., “Scaling IP Multicast on Datacenter Topologies.” In proceedings of the ninth ACM conference on Emerging networking experiments and technologies (CoNEXT '13), 2013, pp. 61-72
[7] Dean, J. and Ghemawat, S., “MapReduce: simplified data processing on large clusters”. Communications of the ACM, 51(1), 2008, pp.107-113.
[8] Samadi, P. et al., “Accelerating Incast and Multicast Traffic Delivery for Data-intensive Applications using Physical Layer Optics.” Proceedings of the 2014 ACM conference on SIGCOMM (SIGCOMM '14), 2014, pp. 373-374.
[9] Funnell A.C. et al., “Hybrid Wavelength Switched-TDMA High Port Count All-Optical Data Centre Switch”, Journal of Lightwave Technology 35 (20), 2017, pp 4438-4444
[10] Benjamin J.L. et al., “A high speed hardware scheduler for 1000-port optical packet switches to enable scalable data centers”, High-Performance Interconnects (HOTI), 2017 IEEE 25th Annual Symposium on, 2017, pp 41-48.
[11] Zhou, Y. and Poo, G-S., “Optical multicast over wavelength-routed WDM networks: A survey”, Optical Switching and Networking, no. 2, 2005, pp 176-197.
[12] Xia, Y. et al., “Blast: Accelerating high-performance data analytics applications by optical multicast.” In Computer Communications (INFOCOM), 2015 IEEE Conference on, 2015, pp. 1930-1938.[13] Wang, H. et al, “Rethinking the physical layer of data center networks of the next decade: Using optics to enable efficient*-cast connectivity.” ACM SIGCOMM Computer Communication Review, 43(3), 2013, pp.52-58.
[14] Samadi, P. et al., "Optical multicast system for data center networks," Opt. Express 23, 22162-22180 (2015)
[15] Samadi, P. et al., “Experimental demonstration of one-to-many virtual machine migration by reliable optical multicast”. In Optical Communication (ECOC), 2015 European Conference on 2015, (pp. 1-3).
[16] Saridis, G. M. et al. “Lightness: A Function-Virtualizable Software Defined Data Center Network With All-Optical Circuit/Packet Switching.” Journal of Lightwave Technology 34 (2016): 1618-1627. [17] Miao, W. et al., “Novel flat datacenter network architecture based on scalable and flow-controlled optical switch system”. Optics Express, vol. 22, no. 3, pp 2465-2472.
[18] Funnell A.C. et al., "High port count hybrid wavelength switched TDMA (WS-TDMA) optical switch for data centers," 2016 Optical Fiber Communications Conference and Exhibition (OFC), 2016.
[19] Tremblay C. et al., “Filterless optical networks: A unique and novel passive WAN network solution,” Proc. OECC/IOOC, 2007, pp. 466-467.

[20] Furdek M. et al., "Programmable filterless network architecture based on optical white boxes," 2016 International Conference on Optical Network Design and Modeling (ONDM), Cartagena, 2016, pp. 1-6.
[21] Kaminow I. P. et al., “A Wideband All-Optical WDM Network”. IEEE Journal on Selected Areas in Communications, 14(5), 1996, pp 780-799.
[22] K Oberg, M.G. and Olsson, N.A., “Crosstalk between intensity-modulated wavelength-division multiplexed signals in a semiconductor laser amplifier”. IEEE Journal of Quantum Electronics, 24(1), 1988, pp.52-59.
[23] FIBER-Q® 1550 nm Fiber Coupled Acousto-Optic Modulator T-M080-0.4C2J-3-F2S product datasheet, Gooch & Housego, Ilminster, UK, December 2016.
[24] Specht, J. and Samii, S. “Synthesis of Queue and Priority Assignment for Asynchronous Traffic Shaping in Switched Ethernet,” 2017 IEEE Real-Time Systems Symposium (RTSS)
[25] Fan F., et al., “Distributed and Dynamic Multicast Scheduling in Fat-tree Data Center Networks”. 2016 IEEE International Conference on Communications (ICC)
[26] L Roy, A. et al., “Inside the social network's (datacenter) network.” In ACM SIGCOMM Computer Communication Review, 45 (4), 2015, pp. 123-137.
[27] Munir, A et al., “Minimizing flow completion times in data centers”, 2013 Proceedings IEEE INFOCOM, 2013, pp. 2157-2165.
[28] Farrington, N. et al., “Helios: a hybrid electrical/optical switch architecture for modular data centers,” ACM SIGCOMM Comput. Commun. Rev., 40(4), 2010, p.339.
[29] “Series 7000n Network Optical Matrix Switch”, Polatis, Inc., 2016.
[30] “Cisco Catalyst 3560 Platform Switches Data Sheet”, Cisco Systems, Inc., 2018.
[31] “Cisco 25GBASE SFP28 Modules Data Sheet”, Cisco Systems, Inc., 2018.
[32] “Cisco Nexus 9500 Platform Switches Data Sheet”, Cisco Systems, Inc., 2018.
[33] “Cisco Nexus 9200 Platform Switches Data Sheet”, Cisco Systems, Inc., 2018.
[34] “Cisco 100GBASE QSFP-100G Modules Data Sheet”, Cisco Systems, Inc., 2018.
[35] Reyes, R. and Bauschert, T., “Infrastructure Cost Comparison of Intra-Data Centre Network Architectures”. Proc. 2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC, 2018, pp 265-271.