Dynamic Optical Networks for Data Centres and Media Production

Adam Christopher Funnell

A thesis submitted to University College London (UCL) for the degree of Doctor of Philosophy (PhD)

Optical Networks Group
Department of Electronic and Electrical Engineering
University College London (UCL)
2019

I, Adam Christopher Funnell confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Contents

Abstract

This thesis explores all-optical networks for data centres, with a particular focus on network designs for live media production.

A design for an all-optical data centre network is presented, with experimental verification of the feasibility of the network data plane. The design uses fast tunable (< 200 ns) lasers and coherent receivers across a passive optical star coupler core, forming a network capable of reaching over 1000 nodes. Experimental transmission of 25 Gb/s data across the network core, with combined wavelength switching and time division multiplexing (WS-TDM), is demonstrated.

Enhancements to laser tuning time via current pre-emphasis are discussed, including experimental demonstration of fast wavelength switching (< 35 ns) of a single laser between all combinations of 96 wavelengths spaced at 50 GHz over a range wider than the optical C-band. Methods of increasing the overall network throughput by using a higher complexity modulation format are also described, along with designs for line codes to enable pulse amplitude modulation across the WS-TDM network core.

The construction of an optical star coupler network core is investigated, by evaluating methods of constructing large star couplers from smaller optical coupler components. By using optical circuit switches to rearrange star coupler connectivity, the network can be partitioned, creating independent reserves of bandwidth and resulting in increased overall network throughput. Several topologies for constructing a star from optical couplers are compared, and algorithms for optimum construction methods are presented.

All of the designs target strict criteria for the flexible and dynamic creation of multicast groups, which will enable future live media production workflows in data centres. The data throughput performance of the network designs is simulated under synthetic and practical media production traffic scenarios, showing improved throughput when reconfigurable star couplers are used compared to a single large star. An energy consumption evaluation shows reduced network power consumption compared to incumbent and other proposed data centre network technologies.

Impact statement

The work in this thesis has potential impact for implementation in future data centre networks and as a bridge to future research in dynamic optical networks for data centres.

The advances made in this thesis in experimental demonstration of fast switching lasers open up new avenues of academic research, including the use of fast switching lasers with coherent data transmission. The novel network designs presented in chapters 4 and 5 are presented in this thesis in sufficient detail for future experimental implementation of both the network physical connectivity and the network controller.

During this PhD studentship, an interruption was taken for a placement with Microsoft Research, to create a functional prototype network with fast switching lasers integrated with commodity electronic hardware. This internship allowed direct knowledge transfer from the work of this thesis into industry, with a successful demonstration achieved of key optical components and subsystems from this research in an operational prototype.

The work in this thesis was also carried out in partnership with BBC Research and Development, who are researching future network structures to support next-generation broadcast production. The dynamic optical network designs in this thesis have been circulated not only within the BBC but also presented at the European Broadcasting Union, leading to interest from public service broadcasters from across Europe. All of the network architectures researched have been influenced by broadcasters' use cases and day-to-day operations, and are ideal candidates for future data centres supporting multicast, high-bitrate video flows. The results of this work will feed directly into future plans for broadcast network designs.

Some chapters of this thesis have already contributed to journal and conference publications, as listed in section 1.12, and it is envisaged that at least two further submissions will be made of currently unpublished material.

Table of Figures

Acknowledgements

First and foremost, my thanks must go to my supervisors, Benn Thomsen and Polina Bayvel. The guidance and mentorship shown by Benn throughout this work has been superb and I can't thank him enough. His skilled lab expertise, clarity of explanation, and great humour have been an inspiration. I also have many thanks for Polina for taking on my supervision, for her continued support, advice and kindness, and her patient dedication to instilling in me a thorough and rigorous scientific method.

I have been privileged to work in so many partnerships during this work. It has been an exciting and fulfilling experience to work with BBC Research and Development, and to be part of their world of future media that I have long respected. Many thanks to Chris for bringing me into the group and for his enthusiasm for the project, and to David for so readily "adopting" my industrial supervision, with many interesting and helpful discussions. The support of Phil, Peter and the APMM team is also greatly appreciated.

It has also been a real pleasure to work with Microsoft Research on parts of this project. From the early development of concepts, to translating parts of this work into real prototypes, this work has advanced enormously thanks to their support. My thanks go to Hitesh, Paolo, Hugh, Krzysztof and David for supporting me in this work and especially for such an inspiring and fun experience during my time at the Microsoft Research laboratory.

A special mention must also be made to colleagues at Oclaro for supplying both tunable lasers and technical insight which enabled this work.

Thanks are due also to the Royal Commission for the Exhibition of 1851, for the support of an Industrial Fellowship, not only providing funding but also welcoming me into a stimulating and engaging network of fellows. The UCL-Cambridge Centre for Doctoral Training in Photonic Systems Development has had a similar impact; I am very grateful to the CDT and its staff not just for funding this work but for collecting and nurturing a cohort of great scientists, engineers and friends.

Finally to the world outside the laboratories, which has been just as crucial to supporting my success. Thank you to everyone in the Optical Networks Group for their friendship and support, especially Daniel (almost 10 years as lab partners!). And to Fulham Brass Band, the Symphonic Wind Orchestra of North London, my friends, my housemates, Sandra and my family, I thank you all.

1. Introduction

1.1. Data centre size and applications

Increasing worldwide communications connectivity has enabled new models of computing services for businesses and consumers. Computers have traditionally been located on premises with end users, but the advent of fast internet connections means that resource-intensive tasks such as large scale data storage, interactive web applications and processing of complex data sets can now be carried out at remote facilities. By placing many computing resources in a central location, economies of scale can be exploited.

Computing resources are aggregated in data centres, which are vast buildings holding tens of thousands of individual servers. Commercial data centre owners are reluctant to divulge information on their operations, but academic studies have observed 10-15,000 individual servers in commercial data centres, and 147-1093 servers within university data centres [1]. The total number of servers across all data centres has been growing rapidly year on year; by 2008 Google, Microsoft and Yahoo! had each already deployed 4-500,000 servers worldwide [2], and by 2013 this had doubled to over 1,000,000 servers spread across at least 15 sites worldwide for each company [3]–[5].

The data centre computing model provides challenges not just for the long haul data transport between data centres and end users, but also information transfer within the data centre itself. Over 75% of data traffic remains within the data centre [6], while total data centre traffic is doubling every 18-36 months [7]. The predicted growth in data centre traffic (including both the traffic within data centres and the traffic between data centres and consumers) is shown in Figure 1. As data centre sizes and scales increase, the distances over which intra-data centre traffic must be carried are also increasing, to the order of 2 km or more [3].

Data centre energy consumption is a major concern for operators users alike. 2% of US energy usage is directly attributable to data centres [8], and data centres operated by Google alone account for 0.01% of the global energy consumption [3]. Historically, data centre energy consumption has risen at 4% per year [8], resulting in a doubling of total energy usage over 18 years. The network infrastructure is currently only responsible for 10% of the overall power consumption in data centres [9]. However, this is projected to rise drastically to up to 50%, due to greater improvements being predicted in the power efficiency and effective utilisation of servers than any improvements to the power efficiency of networking hardware [9], [10].

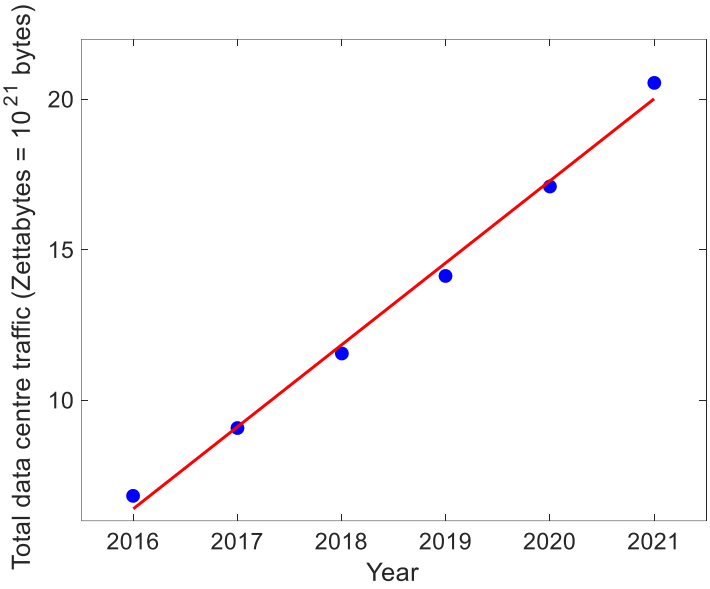

Figure 1: The growth in worldwide data centre network traffic, including traffic within data centres, between data centres, and between data centres and end users. Data from [6], including real traffic sampling for past years to aid future predictions.

A major driver of data centre growth is data analytics, where the parallel nature of the multiple computing resources in data centres can be exploited to perform analysis on datasets of petabits (10^{15} bits) or greater in size. By splitting up datasets and processing them simultaneously using multiple servers, analysis can be performed much faster than by using individual computers. Common algorithms for data processing include MapReduce [11], Hadoop (a specific implementation of MapReduce) [12], and Spark [13], all of which rely on parallel operations across many servers simultaneously to increase efficiency, which in turn is driving data centres to continue growing in size.

A second catalyst for growth in cloud computing is the virtualisation of computing resources. This involves the creation of virtual machines (VMs), which are self-contained units of software that can imitate the performance of dedicated hardware, regardless of the actual underlying hardware that the software runs on [14]. By using VMs, individual processes are no longer tied to specific hardware, and a single server can run multiple VMs concurrently. The connectivity between VMs can be arbitrarily defined; VMs can be isolated or networked as required. VMs allow any software to be run on any hardware in any data centre, and VMs can be created and destroyed on demand, matching the agility and scalability requirements of web services.

Consumer facing services are also key users of data centre resources. Accessing remotely stored data from data centres such as web pages and streamed or downloadable video currently makes up 85% of global internet traffic [15]. Document storage, including photo and video storage, is also handled well by data centres, since

data can be stored in multiple locations across one or multiple data centres simultaneously, for reliability and redundancy. The expected increase in Internet of Things (IoT) devices is also a contributor to data centre growth. IoT devices are often low power, with minimal processing and storage capacity on board the device. Data processing and storage of data from IoT devices can be efficiently handled by cloud services in data centres [16].

Consumer and enterprise cloud services have already been implemented into commercially available data centres. The network architectures of commercial data centres are generally similar, and are based on proven principles from telecommunications networks.

1.2. Evolution of network architectures

Early data centre architectures were generally based on the Clos topology, originally proposed to reduce the total number of crosspoints (2x2 switching elements) required in a telephony switch [17]. The design allowed switches to provide strictly non-blocking paths (allowing all possible input-output connections to occur at any time) without any oversubscription (where oversubscription is defined as the potential maximum input data flow into a switch being greater than the finite total switch capacity). Figure 2a shows the two possible states of a 2x2 crosspoint, which form the basic building block of higher port count switches. In Figure 2b, four separate 2x2 crosspoints are combined to form a 4x4 switch. However, using the switch layout configuration in Figure 2b, some connectivity paths are not always possible simultaneously. For instance, in Figure 2c, if A is already connected to F, there is no possible path for B to connect to E. When not all paths can be simultaneously provided, a switch can be described as “blocking”.

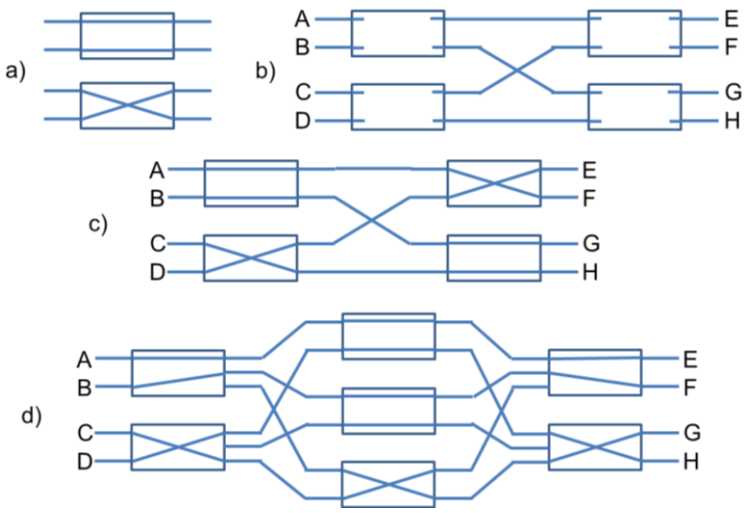

Figure 2: a) The two possible states of a 2x2 crosspoint switch; b) 4x4 network built from the 2x2 switches shown in a); c) demonstration of blocking in the 4x4 network - when A is communicating with F, there is no path for B to reach E; d) Clos network which allows A to reach E and B to reach F without blocking.

To build a non-blocking network connecting 4 inputs to 4 outputs would require sixteen separate 2x2 crosspoints in a square array (not shown in this thesis, but described in [17]). However, Clos's innovation was to increase the output port count of some crosspoints and to add a layer of "middle switches" at the centre of the network, as shown in Figure 2d. In [17] and in the example presented here, the middle switches have a port count of 3x3 and some crosspoints become 2 input x 3 output switching elements. The addition of the centre switching layer permits all possible combinations of input and output connectivity, without blocking any paths. Clos's work in [17] showed the general case for a network of N endpoints, which reduced the number of crosspoints required from $O(N^2)$ in a simple square array, to an upper bound of $O\left(6N^{\frac{3}{2}} - 3N\right)$ crosspoints or fewer. The exact number of layers, and the port count of each crosspoint in each layer, determines whether a network requires fewer crosspoints than this upper bound.

The design shown in Figure 2d provides strictly non-blocking communication; that is, communication between any two nodes can always be arranged, without needing to change the existing connectivity across the network. Clos also proved in [17] that it is possible to provide rearrangeably non-blocking communication, where all possible connections between input and output could be arranged, although existing connectivity may need to be adapted to achieve some connectivity scenarios.

The work by Clos on rearrangeably non-blocking architectures is the basis of "fat tree" networks, also known as "folded Clos" networks, developed by Leiserson in [18]. Given that modern electronic communication switches can simultaneously switch both their input and output connectivity, and that Ethernet links between switches and/or servers are bidirectional, the "middle switches" introduced by Clos can be removed and diagrammatically placed at the top of a hierarchical structure of layers. The "middle switches" in the top layer must have a higher port count than the lower switches to ensure non-blocking throughput (no quantified numerical bound is specified by either Clos or Leiserson other than that it must be higher – the minimum port count required depends on the specific implementation of the topology). This structure, shown in Figure 3, forms the basis of most modern hierarchical data network architectures.

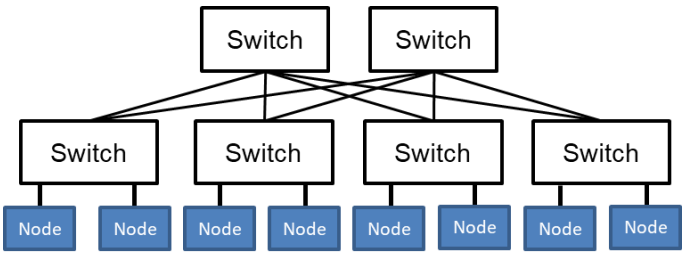

Figure 3: A fat tree (also known as folded Clos) network architecture.

1.3. Electronically switched data centre networks

Present data centre networks are generally constructed using hierarchical layers of electronic switches with fewer than 128 ports per switch, due to the convenience of deploying small units and their compatibility with legacy networking systems [9]. When discussing data centre architectures, there is a convention to describing and drawing network topologies; individual nodes (also known as servers) are generally depicted at the bottom of the hierarchy, with links “upwards” to horizontal layers of switches. A typical data centre layout is shown in Figure 4, with servers at the base of the diagram linked “upwards” to switches shown in “higher” layers.

Individual servers are located in racks, which are physical structures each holding 20-100 servers. Each rack also contains a Top-of-Rack switch (ToR switch), which is an electrical packet switch connecting all of the servers in a single rack, and providing links upwards to other switches in the data centre. ToR switches are connected upwards to aggregation switches, which provide a first layer of traffic aggregation by connecting several racks into a group. Each ToR switch may be connected to several different aggregation switches; this provides additional data pathways for both increased data throughput and reliability through redundancy.

Each aggregation switch is then connected to a layer of core switches, which link together aggregated groups. Additional layers join together core switches to connect an entire data centre site (over distances up to a few kilometres), and to access gateways to other data centre sites and the wider internet (over international distances). In large data centres, multiple layers of core switches would be required to provide connectivity between all of the aggregation switches.

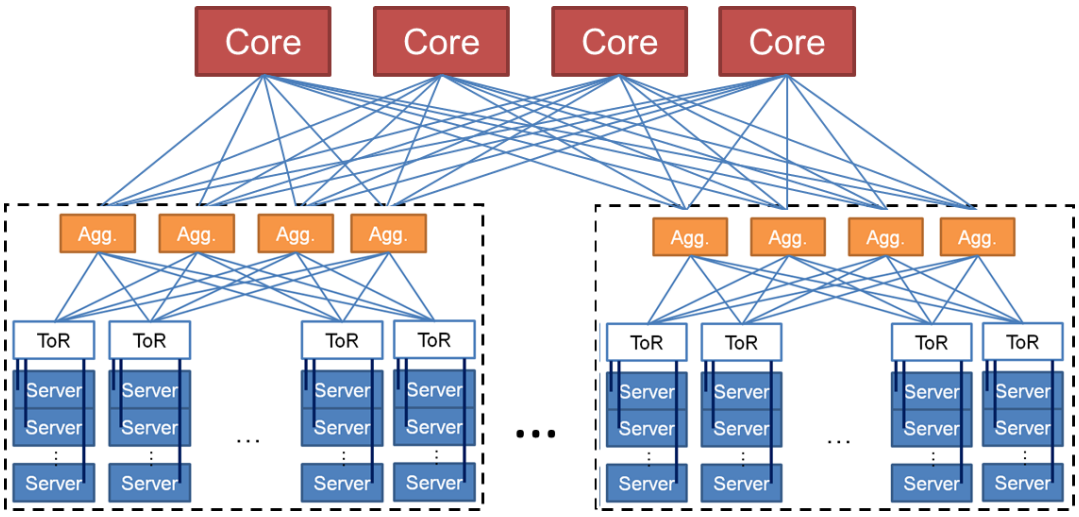

Figure 4: A representative data centre architecture, showing individual servers in racks, top of rack switches (ToR), aggregation switches (Agg.) and core switches. Each switch has multiple links to higher layers of the network.

Hierarchical designs show several benefits. Firstly, oversubscription can be used at lower layers of the network structure. This technique takes advantage of the low utilisation rates of individual network links from servers to ToRs (for example, < 1% in a Facebook data centre [19]) to combine data from multiple servers onto a single network link, reducing the total number of links and switch ports required. In addition, each switch from the ToR layer upwards in the hierarchy is connected to the rest of the network by multiple links, providing redundancy should any individual network link fail.

However, hierarchical structures also have some flaws. When performing a parallel data processing task, the traffic pattern across a data centre shows little communication “locality” - a measure of whether data transfer occurs between servers that are co-located, or physically separated across the data centre. Traffic patterns without locality are not served well by hierarchical structures, particularly in terms of bisection bandwidth (defined as the minimum bandwidth between the two halves of a network, after considering all possible ways to split the total number of servers into two equally sized groups) [20].

Building high performance hierarchical structures without oversubscription also quickly becomes expensive in cost and energy consumption due to the number of switches required even at modest (low Tbit/s) total throughput levels – to increase the switch port counts, number of links and link transmission rates across a network to completely remove oversubscription would cost 14 times as much as a typical oversubscribed network [21].

Latency can also become an issue in hierarchical topologies, as traffic may pass through many switches between source and destination servers; each switch adds both buffering and processing time delays to packets passing through. Switch latency increases exponentially with port count, with typical values of latency from 0.5-4.5 µs per switch [22]. Different paths through the layered structure may incur different delays dependent on the load and buffer size of each switch, resulting in a continuous stream of data from a source having variable packet arrival times at the destination should packets not all follow the same route through the network.

Commercial implementations of data centre networks are currently built using hierarchical layers of electrical switches. For example, Google published details of their existing and historic datacentre architectures in [7]. Clos structures were favoured by Google to provide high bisection bandwidth of over 1.3 Pbit/s using link rates of 10Gbit/s and 40Gbit/s in their 2012 network design (the most recent published work). However, the high throughput came at the cost of latency (unquantified in published work) due to intermediate buffering in hierarchical layers. Non-standard routing

protocols were also required to efficiently route traffic through Google's multi-layered network. Google do not mention any novel optical technology; instead the network structure relied on custom silicon switching processors, with commodity optical links.

Creating one single electrical switch with greater than 128 ports would also be preferable to hierarchical layered designs, by reducing latency, complexity, blocking and oversubscription. However, it is challenging to make efficient high port count electrical switches, since high port count switching chips are fabricated by forming an internal Clos topology of low port count chips. In addition, only small future increases to both the total pin count per chip and the bandwidth per pin are expected, due to the power constraints at each pin [23].

The choice of network topology is often driven by the application which will be run on the data centre hardware. One such application, underserved by current technologies and explored in work to be described in this thesis, is live media production.

1.4. Live broadcast production workflows

Live television and video broadcast production currently uses point-to-point data link protocols to send and receive video, audio and ancillary data [24]. Uncompressed video footage requires bitrates from 1.5 Gbit/s up to 192 Gbit/s; the bitrate required for several common uncompressed video formats is shown in Table 1. For future live production systems, the basic video capture format will be an ultra-high definition (UHD) format requiring at least 24 Gbit/s. Uncompressed video capture is preferable to compressed media streams in live productions, to avoid delays incurred in compression and decompression, which can slow down editorial and creative workflows. In addition, compression at the point of capture can introduce visual artefacts to the media during later processing stages, degrading the video quality.

The most common standard for links of uncompressed video is the Serial Digital Interface (SDI). SDI is a suite of standards for continuous streams of video data (including embedded audio), ranging from 270 Mbit/s up to 24 Gbit/s depending on the video resolutions required [25]. The broadcast industry at present relies on point-to-point links using this standard, generally over electrical coaxial cable, although with point-to-point optical connectivity when required for long distance links [26], [27].

Live media environments predominantly use legacy technologies rather than embracing modern network designs. Production facilities currently use IT network architectures and protocols to configure point-to-point network routing infrastructure [28]. Media transmission protocols up until 2010 have focussed on creating new data rate standards from multiples of the SDI point-to-point link standards, rather than fully embracing commodity IT network technology [29].

Table 1: A comparison of the data rates required for uncompressed video capture and transmission. Data from [30]**.**

Video format	Height (pixels)	Width (pixels)	Bits per pixel (dependent on colour sampling quality)	Frame rate (Hz)	Video data rate (Gbit/s)
1080p or Full HD (High Definition)	1080	1920	12	60	1.5
			24		3
			36		4.5
			48		6
			12	120	3
			24		6
			36		9
			48		12
UHD 4K (Ultra-High Definition)	2160	3840	12	60	6
			24		12
			36		18
			48		24
			12	120	12
			24		24
			36		36
			48		48
UHD 8K (Ultra-High Definition)	4320	7680	12	60	24
			24		48
			36		72
			48		96
			12	120	48
			24		96
			36		144
			48		192

There is now movement across the broadcast industry to use commodity IP (Internet Protocol) networks for programme production rather than point-to-point streaming links [31]. This type of network will allow advanced live production techniques such as remote production (the separation of editorial gallery functions from the event venue), and interactive and immersive experiences (e.g. selection of alternative viewpoints, overlaid information, and virtual reality content). This relies on elemental “grains” as the basis of programme making, rather than streams. Each grain is an individual unit of video, audio or data, and is encapsulated in one or more IP packets in an “IP Studio” ecosystem [32]. For example, a grain could be a single frame of video, or a 10 ms sample of audio.

A critical component in live production is timing synchronisation across all media sources captured. Video sources are currently locked together using a legacy signal from analogue video production, where an electrical pulse signals the start of each frame of video. The frequency of each camera’s frame capture is matched to this signal, and phase adjustment can be performed when video frames from multiple cameras arrive at a central switching point to ensure clean switching between sources. Although precise alignment of IP packets is not required to ensure clean switching, due to buffering within all network nodes, video streams can now be switched by changing the routing of IP media streams across a network [32]. To avoid delays or buffering, a

200 ns bound to network switching time is required, so that switching does not cause video or audio break-up while a switch takes place.

One solution to timing synchronisation for IP studio use is the Precision Time Protocol (PTP) standard (formalised as IEEE Standard 1588). This standard was developed to synchronise clocks across IP networks, and has been adopted to synchronise clocks at every node in an IP media system [33]. Using PTP, the capture of each grain can be carefully timestamped to µs accuracy, so that media flows can be reassembled in the correct order and at the correct frame rate. Specialist hardware can reduce timing uncertainty to 500 ps or less, provided that the synchronisation frequency at which PTP timestamps are shared is high enough [34].

Live media production is sensitive to latency, due to the need for human operators to react to remote situations viewed only through cameras. Although a numerical bound on latency can only be determined by the speed of the specific action being filmed (e.g. fast moving sports action is more sensitive to latency than landscapes and static beauty shots), delays of more than 30 ms cause problems for the interactions between remotely located production contribution sites [35]. This is mostly a problem for the long-haul network links between remote sites and broadcast galleries, but any undue delay in a data centre could cause problems to a production team.

Delay arises from network transmission and processing of uncompressed video and audio data, due to the high bitrate of the media streams. This difficulty can be overcome by using lower quality proxies of video streams, with linked timing metadata between all of the proxies [36]. However, this adds complexity to the production and means that programme makers may not be able to work with the highest quality media streams throughout the creative process. An example remote production system is shown in Figure 5; a sporting event in Glasgow is captured by cameras at the venue, with vision mixing (the choice of camera angles by the director) performed in Salford, and audio commentary and processing produced in London. A field trial of this system had audio latency of 950 ms and video latency of 1400 ms between Glasgow and London, which resulted in difficulties in the creative process of vision mixing [37].

There is a difference between the required end-to-end latency at the application layer (defined above as 30 ms) and the target network physical layer reconfiguration time. When media flows are present across a network, it is essential to switch between the flows without any buffering or packet loss causing audio or video break-up. Due to the high bitrates of uncompressed media, a single node may only be able to receive a single media flow at any time, as the network interface card will be limited in total throughput.

Copyright image removed – available in [37] and downloadable (at time of submission) from https://www.bbc.co.uk/rd/publications/whitepaper289

The relevant image is Figure 2 of the referenced work, on page 6.

Figure 5: A remote production scenario of a sports broadcast from Glasgow with production systems in Salford and audio processing and contribution in London. IP = Internet protocol; HEVC = High efficiency video coding, a compression method for video distribution to the public. Figure reproduced from [37]**.**

It is therefore required to switch media flows for programme production within the network, rather than at the end nodes themselves. Using existing SDI technology, “switching points” are defined within video streams as timeslots where time-aligned media streams can be safely switched [38]. By extrapolating the standard switching points defined in [38] for low resolution video to the high resolution formats shown in Table 1, the “safe” switching timeslot is approximately 200 ns. This provides a target bound for network reconfiguration time which is much smaller than the latency required for data transfer from node to node.

Complete flexibility is required in the connectivity between all 1000+ nodes in a media production network. In most media production facilities, the application requirements vary from day-to-day dependent on the type of programme being filmed, the required video format of the programme and the video and audio processing requirements. This necessitates constant changes to network configurations and routing, in contrast to a data centre with generally fixed node locations and fixed link bitrates. In addition, a single media flow from a camera or other source will need to be simultaneously transmitted to multiple galleries for editorial programme construction, alongside edit suites, storage archives and transcoders. This one-to-many data transmission pattern is known as multicast. Other nodes will need to simultaneously receive multiple media flows for quality control, automated video enhancement and editorial decision making. This requires many-to-one data reception, known as incasting.

The specific network requirements for live media production, which are the target for the network designs presented in this thesis, can be summarised as:

- All-optical network with transparency to bitrate and modulation format, achieving bitrates of > 24 Gbit/s with a clear upgrade pathway to 200 Gbit/s
- Network scalable to support at least 1000 nodes
- End-to-end latency below 30 ms
- Packet switching delay below 200 ns (to allow switching of data streams without noticeably buffering media flows)
- Support for both multicast and incast traffic
- Completely variable multicast and incast group sizes, from 2 nodes up to and including a single multicast group connecting all 1000+ nodes

The complex patterns of streaming media flows in one-to-many (multicast) and many-to-one (incast) patterns, which are constantly changing over timescales of seconds to hours, are crucial to live media production. Support for these traffic patterns across existing data centre topologies is explored in the next section.

1.5. Multicast and incast in data centres

When sending data to and from multiple network nodes simultaneously, some terminology can be defined to represent different types of traffic flow. Multicast is defined as the transmission of data from a single source to multiple destinations simultaneously. This is in contrast to unicast (the dominant method of network data transfer), defined as data transmission from a single source to a single destination. The opposite of multicast is incast, defined as multiple sources simultaneously sending data to a single destination. Both multicast and incast traffic patterns are essential for the live media application requirements discussed in the previous section. If multicast and incast traffic are both present in a network (whether involving the same nodes performing dual functions as transmitters or receivers, or different sets of sending and receiving nodes), the resulting overall traffic pattern is described as anycast.

When the same data from a single source is to be transmitted to multiple destinations, it is inefficient to set up multiple unicast links, and congestion can occur in parts of the network. Consider the topology shown in Figure 6a, which is an arbitrarily chosen simple network design showing a single source and two destination nodes connected via multiple switches. If the same data is to be sent from the source to both destinations by unicast transmissions, the data must be sent twice over some of the same links, causing congestion as shown in Figure 6b. To avoid this, a multicast transmission can be prepared, so that the network can efficiently send a single stream of the data wherever possible, and only split the single stream at the last possible switch to minimise the unicast traffic. This reduces congestion, as shown in Figure 6c.

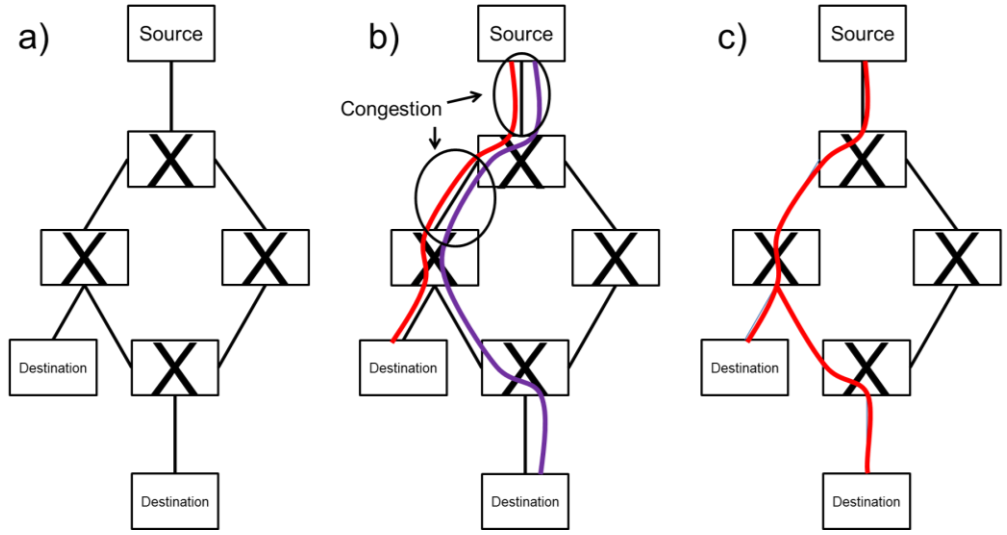

Figure 6: a) an example network topology, with a source, two destinations, and electronic packet switches (labelled "X"); b) unicast transmission of the same data from the source to both destinations; c) multicast transmission of the same data from the source to both destinations.

However, multicast does not scale across hierarchical switch networks, or indeed any electrical packet switched networks, due to the limits of multicast addressing. IP networks use a single “multicast address” to identify a “multicast group” of servers. To route multicast data to the correct destinations within the multicast group, each switch must use lookup tables to identify the membership of the groups, and then ensure that optimum multicast paths are formed across the network. There is a limited size to the address lookup table of each switch (generally in the low 100s of addresses, using memory only available for “non-standard” addresses), forcing a limit on the total number of multicast groups that a network could support. Each server could be a member of several multicast groups simultaneously, resulting in more multicast group addresses being required than servers can store and process.

In addition, protocols for switches to communicate routing paths to their neighbours do not efficiently cover multicast traffic. Switches may discover routing paths through the wider network by exchanging messages with their immediate neighbours, and can assume that the addressing and connectivity of the neighbouring servers and nodes remains static. Multicast groups can be created and destroyed on demand, and entries must be created on all switches on any path between multicast group members. Without a central multicast address registry (which is generally the case), the complexity of keeping every switch up-to-date with multicast routing becomes unmanageable, and especially intolerant of link or switch failures if connectivity must be rapidly remapped.

There is also lack of topological structure to multicast group IP addressing, resulting in impractical complexity [39],[40]. IP addresses in data centres are usually allocated so

that areas of a data centre are addressed using the same subnet (the first digits of an IP address) e.g. 192.174.xxx.xxx would correspond to one region of interconnected switches and nodes under a core switch, while 192.175.xxx.xxx would correspond to another. This simplifies routing decisions made at each switch and maps the virtual address space to the physical data centre layout. However, there is no such planning in multicast group addressing; IP multicast addresses are allocated from a reserved block with no topological hierarchy. Moreover, the closed protocols of network transport between virtual machines may not follow strict IP address structures, further limiting the feasibility of IP multicast implementations [41].

The lack of support for IP multicast over the hierarchical switch network designs described in section 1.3 means that physical layer multicast, including optical multicast, is a promising solution to provide multicast at data centre scales. To move media production services requiring high data rate optical transmission into data centres, it is desirable to ensure efficient multicast support using optical technologies. The following sections describe the main categories of optical switching which can be used as building blocks to create data centre networks.

1.6. Optical Circuit Switching

Optical circuit switching (OCS) describes networks which can be reconfigured by physically repositioning the paths (“circuits”) over which light flows through the network. The paths are generally constructed from optical fibres, although free-space optics can also be used to route signals in the optical domain. OCS also encompasses light-paths in wavelength division multiplexed (WDM) optical networks, where circuits are set up across a network between multiple nodes, using dedicated wavelengths over a single fibre [42]. OCS can build future-proof optical networks, since circuit switches are agnostic to the optical modulation format and data rate of the optical signals, passing all signals equally. In addition, OCS switching units draw little power; for example, a 320x320 port OCS consumes only 3.3% of the power of a 224 port 10 Gbit/s electrical Ethernet packet switch [43].

OCS units typically use microelectromechanical systems (MEMS) to steer beams between input and output fibre ports; switches with up to 384 x 384 ports are already commercially available, with a switching speed of the order of 25 ms [44]. Commercial products generally switch on a ms timescale, with a few laboratory prototypes in the µs timescale [45]. Given that data packets transmitted at 10 Gbit/s or higher speeds are only of ns duration, allowing time for OCS units to switch each packet individually would impose a large overhead on network throughput. It is, therefore, more efficient to

switch optical paths for flows or bursts of multiple packets, so that the reconfiguration time is a shorter proportion of the data flow.

In this thesis, OCS is used in chapters 4 and 5 to provide connectivity between optical passive star couplers to create multicast-capable networks. The slow speed of OCS switches (10s ms) compared to fast laser tuning times (< 200 ns) means that it may be necessary to reconfigure the OCS units at a different rate to the tunable lasers, to avoid excessive network reconfiguration times (see section 1.10) The low power consumption of OCS units is reviewed at the end of chapter 5 to justify adding OCS units to passive star networks – the potential increases in network throughput from using OCS units are greater than

1.7. Optical Burst Switching

Optical burst switching (OBS) assembles groups of data packets at the edge of the network that are destined for the same output node (or a group of nodes), before transmitting them together as a “burst” to provide efficient routing of data through the network core [46]. Once a burst is assembled at the edge of a network, control data is transmitted ahead of the data burst, so that a dedicated optical path can be set up through the network.

There are several methods of controlling the network to set up the required optical paths, with a trade-off between causing a delay in transmitting the burst onto the network, and guaranteeing that the path reserved for a burst will be collision-free. In “just-in-time” signalling, a control packet requesting an optical path is sent on an out-of-band control channel just ahead of the burst data. The burst itself is sent soon afterwards, without waiting for confirmation that the path has been reserved through the network core [47]. Although acknowledgements can be sent by the receiver to confirm successful reception of the transmitted data, waiting for acknowledgements may cause network latency to grow larger than the 30 ms specified for live media production in section 1.4.

Using a central controller to co-ordinate burst access to the network was shown to remove the possibility of collisions within the network core in [48]. In addition, with global oversight of the assigned wavelengths in a WDM network, wavelengths were flexibly reallocated to serve bursts from multiple sources. However, this came at the cost of reduced admission of bursts to the network, and a greater (albeit deterministic) delay experienced by bursts when waiting at the network edge. While nodes gather packets for assembly into bursts, the network interface is not idle and does not needlessly seize a wavelength, but can transmit bursts that have previously been assembled. However, the work in [48] showed that although increasing the burst

assembly time to 50 ms guaranteed the quality of service by removing collisions and limiting latency, it reduced overall network utilisation to 50%. This was an inefficient use of network resources, and would result in unacceptably high latencies for media applications (see section 1.4).

In this thesis, the operation of the WS-TDM network introduced in chapter 2 operates at a mid-point between OBS and optical packet switching (see section 1.8). In the WS-TDM network, transmitters request wavelengths and timeslots which are allocated by a central scheduler. Similarly to OBS networks with fixed time periods to assemble bursts, the WS-TDM network reconfigures transceiver wavelengths every 2 µs. However, the WS-TDM network does not require a transmitter to wait to collect multiple packets into a burst before sending data – transmitters can request any number of timeslots within any 2 µs epoch.

Although OBS makes more efficient use of network resources than OCS when transferring short lived data flows, the indeterminate latency (caused by waiting for acknowledgements or transmission rights to avoid collisions), and the overall latency involved in collating bursts, mean that OBS is not suitable for transferring real-time media streams. An alternative method of optical switching is to switch individual packets, rather than bursts of data.

1.8. Optical Packet Switching

Optical packet switching (OPS) describes systems where individual data packets are switched entirely in the optical domain. This can be achieved using a single centralised optical switch, or a network of optical switching units.

Semiconductor optical amplifiers (SOAs) can be used in OPS sub-systems due to their fast on-off switching time (ns), high extinction ratio (20-40 dB), low power operation (< 0.2 W) and broadband performance across the optical C-band. Early SOA switches exploited the refractive index change within SOA devices due to carrier injection to provide switching through interferometric effects [49]. However, it is much more common for SOA switching devices to be formed from a power splitter, which splits an optical signal in a “broadcast-and-select” fashion [50]. As shown in Figure 7, an input signal can be split using a passive optical power splitter, and each output of the splitter can pass through an independent SOA; an applied electric current enables the optical route through an SOA, while other SOAs are reverse biased to attenuate light approaching their paths. Optical power combiners (functionally identical to splitters in reverse) combine the signals from multiple SOAs to ensure that all paths are available.

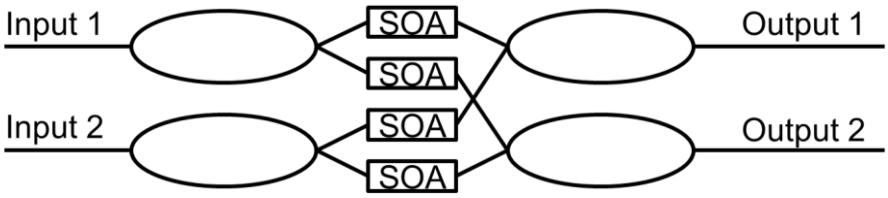

Figure 7: A simple design for a broadcast-and-select semiconductor optical amplifier (SOA) based optical switch. In this example each of two inputs are split into two outputs using a passive splitter. Each output can be independently enabled or disabled, allowing unicast or multicast from either input.

The high optical power gain of SOAs has also been exploited to overcome power splitting losses when combining multiple SOAs into non-blocking switching units [51].To date, only low port count (from 8x8 up to 32x32 ports) SOA switches have been experimentally demonstrated, due to the impact of cascaded ASE noise from multiple successive SOAs adding to the original signal. SOA switches can be efficiently constructed by using the 2x2 switch design of Figure 7 in a Clos topology (i.e. for an 8x8 port network, 3 layers of SOAs, requiring 64 SOAs in total). By measuring the parameters of multiple optical passes through an 8x8 SOA switch construction block, a 64x64 SOA switch was simulated in [52] which allowed a 6dB input power dynamic range with 2 dB output power penalty.

Alternatively, Mach Zehnder intensity modulators (MZIs) can be used as optical switching elements. The crosstalk from the use of cascaded MZIs as switches was shown to be reduced by using SOAs to partially block crosstalk from passing through the switch stages [53]. However, the effect of accumulated ASE noise from the SOAs on high bitrate signals has yet to be fully assessed in a practical implementation. SOA switches have also been integrated into full optical packet processing systems [54], but the scalability of this design to greater than 64 outputs has not been demonstrated, so the application of SOA switches to data centres requiring at least 1000 outputs is not clear.

SOA switches utilise active switching at the core of the network, but it is also possible to route packets in the optical domain using passive network components. An arrayed waveguide grating router (AWGR) is a passive device, which routes each different wavelength at an input to different destinations. AWGRs use a cyclic wavelength routing pattern so that every input can transmit to every output by using a particular wavelength [55].

An example of the cyclic wavelength pattern of a 4x4 AWGR is shown in Figure 8; each of 4 different wavelengths (labelled λ_1 to λ_4) is present at each input (labelled A-D). Each wavelength is routed to a different output, following a mapping which is

dependent on the input port. For example, λ_1 transmitted into input A is routed only to output A, but λ_2 transmitted into input A is routed only to output B.

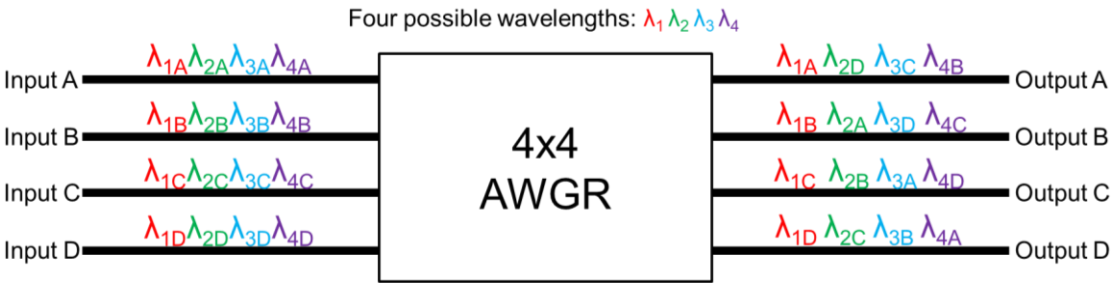

Figure 8: An example wavelength routing pattern of an arrayed waveguide grating router (AWGR). Each input wavelength is routed to only one output fibre, showing that AWGRs do not readily support multicast.

However, when using AWGRs in switching networks, laser tuning times (often ms or higher, as described in section 2.2.2) dominate packet switching latency, and the coloured nature of AWGRs does not allow any form of multicast [56]. This is shown in Figure 8, where each individual wavelength input to any port of an AWGR is only routed to a single output. This one-to-one mapping between input and output ports, determined by wavelength, forbids broadcast across AWGRs.

Experimental realisations of AWGR based networks have achieved large port counts – using 512x512 AWGRs in hierarchical systems has been proposed to connect data centre scale networks [57]. Despite physical layer feasibility, the latency across large scale AWGR systems has yet to be fully quantified, especially when including laser tuning times into the total latency experienced by each packet.

Both SOA and AWGR network designs require a central scheduler, using an independent control network to receive requests from nodes and grant allocations. This increases the network complexity and cost, by requiring either a parallel electrical switching network, or a secondary optical control network, albeit at a lower bitrate. To avoid this, it is also possible to use optical header recognition to route data packets [58]. A correlator can be used to match addresses contained within the optical data packet with known bit patterns, formed from a fibre Bragg grating etched with a reflection pattern to match a binary bit sequence. Optical header recognition has been shown to operate across several WDM channels simultaneously [59]. However, the correlation technique would not scale to a network comprising > 1000 nodes, due to the need for a unique optical comparator (i.e. a unique address) for each output. Optical header recognition is, therefore, not recommended for use at data centre scale.

The work in this thesis predominantly concerns “packet-like” switching rather than true OPS. Due to the 200 ns reconfiguration time of the WS-TDM network in chapter 2, it is not possible to switch every packet individually without the network reconfiguration

overhead being larger than the data transmission time. The network is instead reconfigured by wavelength every 2 µs, and each wavelength could carry a single large packet from a single transmitter up to 50 packets from up to 26 different transmitters. However, work in chapter 3 reduces the laser retuning time to 35 ns, which is approaching packet-level timescales (analysis of packet durations and switching overheads for a range of data transfer rates is shown in Figure 38).

Having described the three major categories of optical switching, data centre network designs incorporating some or all of these technologies can be reviewed to assess their suitability for all-optical networks supporting multicast for media production.

1.9. Data centre network designs

1.9.1. Literature review criteria

The following literature review is a study of both theoretical network architectures and specific experimental implementations of network topologies, and it is not always possible to find numerical metrics by which to compare designs. For instance, quantifying the number of distinct multicast groups available in designs based on OCS may depend on the port count of each OCS, which is a design choice in any specific implementation of an architecture; comments could then only be made on the general scalability of a network topology. However, there are some clear metrics which can be used to compare architectures; where numerical comparisons are not possible, the design choice which would limit the metric is specified e.g. if the only barrier to increasing the total number of supported nodes in an OCS network is the OCS port count, this will be noted.

The comparison metrics are as follows:

Total node count. Data centres for media production contain over 1000 servers as end nodes of the network, so any network design should be able to support at least that many end nodes.

End-to-end latency. As specified in section 1.4, live media production places strict bounds on end-to-end communication latency between any pair of servers on the network of 30 ms.

Reconfiguration time. Although a network may have high flexibility and reconfigurability to allow new routes or connectivity, it may take time for the physical layer reconfiguration (whether wavelength tuning, fibre connectivity via an OCS, or any other reconfiguration to take place). For media applications, optical packet switching in < 200 ns is essential to ensure that switching operations can take place in inter-packet gaps and video or audio streams do not need to be paused or buffered.

Multicast capability. In optical network designs with wavelength routing and segregation, or limited pathways between central core nodes, multicast is not always possible. A network architecture suitable for media production must be multicast capable.

Unlimited number and size of multicast groups. Even though a system may support multicast, it may not be possible to create multicast groups of arbitrary sizes. A network architecture suitable for media production must be able to create multicast groups of any size up to and including all nodes simultaneously.

Designs excluded from this review include:

- Any design requiring intermediate optical-electrical-optical conversion
- Light-tree approaches using wavelength tuning and/or splitting targeting long-haul mesh networks rather than data centres

The metrics are restricted by the following conditions:

- Where the end nodes are ToRs, a value for total node count only includes the number of ToRs stated in the published work, since the ToRs themselves could have any number of ports to connect servers.
- End-to-end latency does not include control plane latency, but does include all propagation time for data transfer.

Total throughput is not useful as a metric, due to the lack of objective comparison traffic patterns and recognised benchmarks [60]. Attempts were made in [60] to define a "worst case" traffic matrix by which to compare topologies, although network simulations in [60] using the defined worst case traffic matrix contradicted many previous findings in the literature.

A comparison of prior work to the defined metrics is presented in Table 2 at the end of this chapter, with analysis of the works which most closely meet the application demands applications presented below.

1.9.2. Non-hierarchical electrical switching networks

Before reviewing all-optical networks, there are some notable electrical switch architectures which do not employ hierarchical layered structures. It is challenging to quantify numerical metrics to compare the performance of the network topologies for several reasons: a lack of experimental data; the wide range of variables in constructing the networks; the impracticably high number of possible network connectivity arrangements; and the inherent dependence of the system performance on the underlying network traffic pattern. Nevertheless, useful (even if qualitative)

comparison metrics include the bisection bandwidth, latency and support for diverse traffic patterns.

DCell [2] used a recursive structure of “cells” to form a network which was highly fault tolerant due to the partitioning of servers and switches into a cellular structure. However, the bisection bandwidth of DCell was only 15% of the bisection bandwidth of a fat tree architecture (for a 1000 node network), most likely due to the reliance on congested interconnects between cells. Latency was also high (unquantified in published work) due to many hops required between switches, and broadcast traffic from a single source to multiple destinations was not supported.

BCube [61] used servers to forward packets on to other servers, which reduced the number of switches required; layers of switches were then used to connect groups of servers. Further layers of switches were added in a recursive pattern to provide full connectivity between the groups, although switches only ever connect to servers, no aggregation is provided over switch-to-switch links. Custom routing software or hardware was required, which reduced server performance. Queues of buffered packets at intermediate servers would also cause latency (unquantified in published work).

VL2 [21] used an enhanced fat tree topology to increase the connectivity within each layer of the hierarchical structure. By adding more links between switches than the minimum required for a Clos network, the overall throughput and bisection bandwidth were increased compared to the Clos architecture (the exact amount of increase depends on the number of additional links added). However, this came at the cost of increasing the total number of switches required, and the queueing/buffering latency that the additional switches incur.

1.9.3. All-optical data centre networks

The “Archon” architecture [62] used a central OCS capable of switching multi-element fibre to provide switching between groups of ToR switches in a data centre. The OCS was also linked to both a time division multiplexing (TDM) switch for localised traffic within the data centre, and wavelength converters to provide WDM links between data centres. The overall network throughput of up to 320 Gbit/s per link and the combination of an OCS with TDM allowed multicast and incast traffic. However, the network scalability was poor due to the large number of complex multi-element fibre switching and conversion units required to reach a high port count. Additionally, the switch reconfiguration speed was 25 ms (limited by MEMS OCS units), which is greater than can be tolerated by media applications.

The “Optical Multicast System” [63] constructed a hybrid architecture of OCS switching and EPS switching, and permitted multicast flows by switching optical couplers and splitters in and out of optical paths, as shown in Figure 9. Although multicast and incast are well supported, and the end-to-end latency of 30-50 ms meets the application needs, there is not necessarily support for arbitrary and variable multicast group sizes. This is because the possible multicast group sizes are set by the sizes of the optical couplers and splitters attached to the OCS units, which is set at the time of commissioning the network and cannot be changed in real time as application demands change.

“LIGHTNESS” [22], [64] used a hybrid combination of OCS and OPS to connect data centre rack switches. The “LIGHTNESS” network showed end-to-end latencies of 5 µs and the OPS switching units could reconfigure within around 200 µs (although SOAs with ns rise times were used as OPS switching elements, the delay time arose from the processing required on FPGAs used to control the SOAs). The OPS units only allowed multicast in limited circumstances (dependent on the locations of servers within the data centre that are members of a multicast group), so OCS can be used to provide multicast via optical power splitters. To reconfigure the OCS units of the “LIGHTNESS” design incurred a large delay of 300 ms, due to the configuration of wavelength selective switches (WSS). Although beyond the scope of this review, the complexity of the “LIGHTNESS” software control plane brought the total reconfiguration latency to a very high 970 ms, which is far in excess of the 30 ms required by media applications.

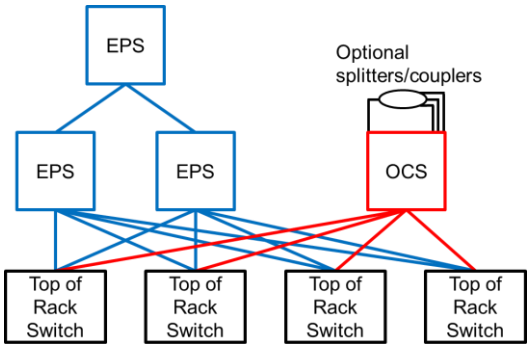

Figure 9: An example hybrid electrical packet switch (EPS)/optical circuit switch (OCS) network, connecting to top of rack switches.

“*-cast” [65] demonstrated an all-optical network capable of multicast and incast data traffic patterns. The system architecture used OCS units to switch “photonic gadgets” in and out of fibre paths between transmitters and receivers; the possible “gadgets” included splitters, couplers, amplifiers and wavelength filters. Some experimental results showed the latency reaching 199 ms when trying to serve 1000 racks with the “*-cast” design, which is far larger than the acceptable latencies for media production outlined above. The network design also only partially meets the design requirements

for multicast traffic, since fully flexible group sizes are not feasible. This is due to the possible group sizes being fixed by the gadgets attached to the OCS, which cannot be changed in real time as application demands change.

The limits to the network reconfiguration speed in “Optical Multicast System”, “LIGHTNESS” and “*-cast” are the OCS units. Highly scalable OCS units (with port counts > 64x64) generally use MEMS to change the path which the light beams follow, which require milliseconds to precisely and securely align optical beams in 3D space. OCS units are therefore best suited to switching long flows of data, rather than individual packets, so that the network reconfiguration time overhead is small (a few %) compared to the data transmission time [43]. Only a few laboratory prototype systems with low port count (such as a 24x24 port switch in [45] or a 64x64 port switch in [66]) have shown microsecond switching times.

“LAMBDANET” [67] showed a network centred around an optical star coupler, connecting nodes over 40 km distances. Each transmitter in the network was allocated a fixed wavelength, and each node contained a WDM demultiplexer followed by an individual receiver for each wavelength, so that each receiver could directly receive data from all transmitters simultaneously. The “LAMBDANET” network design only targeted up to 128 nodes (limited by power budget across the optical star coupler core), which would require 128 receivers at each node – neither the power budget nor number of receivers would permit scaling the network design to > 1000 nodes. However, the star coupler network core permitted fully flexible multicast groups, up to and including all nodes joining a single multicast group.

“OvS” [68] created “optical virtual switches” (OvS) by using switchable wavelength routing in the network core to direct packets to their destinations. Although “OvS” showed end-to-end latency of 114 µs which meets the design criteria for low latency data transfer, the switching time of 7 ms is too long to provide packet level switching that would not interrupt media flows. In addition, although multicast is supported in “OvS” by placing optical couplers and splitters in the optical circuit paths, some multicast group arrangements are only feasible using intermediate buffering at switching nodes, which may incur queuing and therefore increase latency.

“Splitter and Delivery” [69] and “Mixed Splitter and Delivery” [70] networks used optical splitters followed by optical switches to provide a fully flexible multicast network. The published work presented only architectural designs so no quantitative comparisons of switching speed can be made from the published results. However, since the designs use tunable filters and tunable wavelength converters to ensure that all signals from the network core can reach all outputs, the switching time would be limited by the

wavelength tuning speed of tunable lasers (the technical feasibility of fast wavelength switching is discussed further in section 3.2).

The network architectures described in this section were all designed for generic data centre use, however, some designs have been designed and/or constructed specifically targeting live media production applications.

1.10. Optical networks for broadcast production

To date, point-to-point links are the only major optical transport innovations for media to be commercially realised. The use of specialist fibre optic links for media production has been limited to the parallel transmission of video data, breaking down a single high bitrate stream into smaller flows and using either space multiplexing across bundles of fibre [71], or through coarse wavelength division multiplexing (CWDM) in a proprietary protocol [72].

Some prior work has already considered applications of all-optical networks specifically for media production. In [73] an optical star coupler network supporting broadcast workflows was demonstrated as scalable to 200 nodes at a bitrate of up to 2.29 Gbit/s per media flow. This network was implemented in a media production centre, but was not scalable or futureproof to increased data rates, and would not permit arbitrarily sized and variable multicast groups due to the use of a parallelised star coupler topology to reach all nodes.

A further media production network design aimed to connect up to 200 nodes at bitrates of up to 1.2 Gbit/s per node [74]. Although this network was also based on a star coupler, the design would not scale to the requirements of today's media data centres (>1000 nodes and 25-200 Gbit/s bitrates) due to the direct detection optical receivers allowing insufficient optical power budget.

It is notable that both networks described in this section, and most networks described in section 1.9 were generally based on optical star couplers. It therefore appears that star couplers are a promising optical technology to meet the application demands of live media streaming.

1.11. Star coupler networks

The information in section 1.9 highlighted that prior work on optical networks has not led to a design capable of meeting all of the demands of media production applications, and the networks that have the best performance against the defined metrics are all based on passive optical star couplers. To consider the specific properties of star coupler networks in detail, further literature has been reviewed, and the distinct

contribution of the work presented in this thesis highlighted compared to prior publications.

An early design known as LOCNET split wide and narrow band services via an optical star coupler [75]. The network was designed to support only 64 end nodes (not the >1000 targeted by this work for data centre scale), and switching functionality was only performed using TDM, which limited the bitrate that could be used for data transmission by each node (to 4 Mbit/s in the published implementation, and to the transmitter line rate divided by the number of nodes in a generalised case).

The Knockout star coupler switch described in [76], [77] showed a star coupler switch capable of broadcasting all inputs to all outputs for up to 1000 nodes, however the switch design assumed that packet loss was unavoidable in any practical switching implementation. The effects of packet loss were mitigated using parallel buffers at the receivers, but for a 70% network load, 12 receiver buffers each capable of holding 40 packets were required. This approach is not suitable for media streaming, since buffering would incur latency, and if buffers overflowed, dropped packets would cause loss of media which could not be resent in real time.

Work published in [78] considered the use of tunable transmitters and receivers over a single large optical star coupler, despite fast tunable lasers not having been developed at the time of its publication. The design in [78] was focussed on suitable control protocols for scheduling traffic over a star coupler network rather than the network design itself; this thesis does not consider the optimum control protocol but instead provides physical layer verification of the feasibility of a star coupler data centre network.

A suite of publications considered fibre star couplers to be “shared directional multichannel” (SDM) devices, capable of connecting multiple nodes in a single-hop architecture[79]–[82]. However, the general case of the SDM topology required multiple transceivers per node to guarantee flexible connectivity, and the separated nature of the star couplers meant that efficient multicast (using a single transmitter per multicast flow) could not be supported between all nodes. Additionally, the total throughput of the network was limited, as the work explicitly ruled out sharing the fibre bandwidth through wavelength or time division multiplexing. The work described in this thesis explored star coupler architectures which require only a single transmitter per node to permit any-to-any multicast and incast traffic, which reduces the overall data centre network complexity and cost.

Other prior work has considered the arrangement of multiple star couplers connected to have the same connectivity as a single star coupler. The work in [83] used multiple

small star couplers to reduce the total amount of cabling while maintaining the abstraction of a single star coupler connecting all network nodes, while the work in [84] considered the construction of larger stars from smaller stars to increase the total node count ([84] showed only a general network topology so no specific numbers were given). However, the work in both [83] and [84] did not use reconfigurable star coupler combinations as a tool to increase the total capacity. The work in chapters 4 and 5 of this thesis, where stars are constructed from combinations of smaller star couplers via switching units, is focussed on increasing the total network throughput without losing the connectivity properties of star couplers.

Star coupler properties will be explored further in section 2.2.1, which describes each of the subsystems necessary to build an all-optical network capable of meeting all of the design metrics described above.

1.12. Conclusions and thesis outline

The application requirements of live media production, alongside the prior work reviewed in this chapter outlined the following design requirements which are not all met in a single design by any work to date:

- All-optical network with transparency to bitrate and modulation format, achieving bitrates of at least 24 Gbit/s with a clear pathway to upgrade to 200 Gbit/s
- Network scalable to support at least 1000 nodes
- End-to-end latency below 30 ms
- Packet switching delay below 200 ns (to allow switching of data streams without noticeably buffering media flows)
- Support for both multicast and incast traffic
- Completely variable multicast and incast group sizes, from 2 nodes up to and including a single multicast group connecting all 1000+ nodes

The designs reviewed here which showed performance closest to these desired characteristics used optical star couplers to enable fully flexible multicast groupings. However, none of the prior work also met all of the bitrate, node count scalability, latency and switching speed requirements, which the work presented in this thesis targets:

- Most prior work on star coupler networks reviewed in this chapter was performed at low transmission rates per node (Mbit/s or lower); chapter 2 describes the feasibility of 25 Gbit/s transmission across an optical star, while

the network designs of chapters 4 and 5 are transparent to the bitrate and transmission formats used.

- To integrate optical star coupler networks into data centres suitable for live media production requires networks with over 1000 ports, which presents challenges for signal integrity due to high power splitting ratios; chapter 2 demonstrates the physical layer feasibility of a 1000 port star coupler network at both 10 Gbit/s and 25 Gbit/s.
- End-to-end latency was not measured during the experiments described in this thesis, due to the focus of the work being on the network physical layer rather than network drivers in the nodes and/or the control plane. However, the data plane latency will only be due to the speed of light in the optical fibre between the transmitter and receiver, due to the single-hop nature of the network. This is described in chapter 2.
- Packet switching times of 200 ns can be achieved using fast tunable lasers; full system demonstrations of 200 ns laser tuning speed (measured from the start of a laser tuning to error-free data recovery) are described in chapter 2, and further reductions of laser tuning time to 35 ns are described in chapter 3.
- When using an optical star network, multicast and incast are inherently supported for arbitrary numbers of casting groups and group sizes. Chapters 4 and 5 present methods of increasing the input across star networks, by using reconfigurable circuit switches to configure optical star couplers while maintaining the same multicast and incast connectivity of a single passive optical star.

In summary, Chapter 2 outlines the WS-TDMA physical layer network design based on a star topology, and describes experiments carried out to explore and verify the physical layer feasibility of the design. Chapter 3 suggests improvements to the tunable laser and line coding subsystems within the WS-TDMA network, to reduce the network reconfiguration time and increase overall throughput. Chapter 4 presents methods to construct passive optical star networks on-demand, while chapter 5 presents a method to split a larger star; both chapters are seeking to use dynamic optical components to increase the throughput of star networks without losing the desirable characteristics. Chapter 6 both concludes the thesis and discusses future work which may develop from the concepts and results that it describes.

1.13. Publications related to work described in this thesis

- Alistarh, D., Ballani, H., Costa, P., Funnell, A., Benjamin, J., Watts, P. and Thomsen, B., 2015, August. A high-radix, low-latency optical switch for data

centers. In *ACM SIGCOMM Computer Communication Review* (Vol. 45, No. 4, pp. 367-368). ACM.

- Funnell, A., Benjamin, J., Ballani, H., Costa, P., Watts, P. and Thomsen, B.C., 2016, March. High port count hybrid wavelength switched TDMA (WS-TDMA) optical switch for data centers. In *Optical Fiber Communications Conference and Exhibition (OFC), 2016* (pp. 1-3). IEEE.
- Benjamin, J.L., Funnell, A., Watts, P.M. and Thomsen, B., 2017, August. A high speed hardware scheduler for 1000-port optical packet switches to enable scalable data centers. In *High-Performance Interconnects (HOTI), 2017 IEEE 25th Annual Symposium on* (pp. 41-48). IEEE.
- Funnell, A.C., Shi, K., Costa, P., Watts, P., Ballani, H. and Thomsen, B.C., 2017. Hybrid Wavelength Switched-TDMA High Port Count All-Optical Data Centre Switch. *Journal of Lightwave Technology*, *35*(20), pp.4438-4444.

Table 2: A comparison of prior work on all-optical networks for data centres. Prior work which almost, but not entirely, meets the data centre application requirements is denoted in the right-hand column with an asterisk (*), and analysed further in section 1.9.

Prior Work	Ref.	Network Type	Total Node Count (N)	Reconfigur-ation time	End-to-end latency	Multicast	In-cast	Variable group multicast	All-to-all multicast	*
Archon	[60]	OCS	Unknown	25 ms limited by OCS	Unknown	Yes	Yes	No	Yes, only if OCS has a single NxN star attached	*
RHODA	[83]	Clustered Hybrid OCS/OPS	106 with restrictive traffic assumptions	Unknown	Unknown	Limited	Limited	No	No	
Optical Multicast System	[61]	Hybrid OCS/EPS	512 - limited by OCS ports	25 ms limited by OCS	30-50 ms limited by control plane	Yes	Yes	No, limited by splitters at OCS	Yes, only if OCS has a single NxN star attached	*
LIGHT-NESS	[21], [62]	Clustered Hybrid OCS/OPS	> 128 per cluster – limited by OPS	300 ms limited by WSS (970 ms limited by control plane)	2-3 µs OCS, 33-200 µs OPS	Yes	No	Yes via OPS only, not via OCS	Yes via OPS only, or if OCS has a single NxN star attached	*
Work in this thesis		**Hybrid OCS/OPS**	**> 1000**	**< 200 ns**	**< 1 µs**	**Yes**	**Yes**	**Yes**	**Yes**	

Prior Work	Ref.	Network Type	Total Node Count (N)	Reconfiguration time	End-to-end latency	Multi-cast	In-cast	Variable group multicast	All-to-all multicast	*
*-cast	[63]	Hybrid OCS/EPS	10 to 10000 implementation dependent	0.1 ns to 1 ms	199 ms	Yes	Yes	No, limited by splitters at OCS	Yes, only if OCS has a single NxN star attached	*
LAMBDA-NET	[65]	Passive Star	Unknown – 18 demonstrated	Unknown	Unknown	Yes	Yes	Yes	Yes	*
OvS	[66]	Hybrid OCS/EPS	1089	7 ms limited by wavelength selective switches	114 µs	Yes	Yes	Yes	Yes	*
Work in this thesis		**Hybrid OCS/OPS**	**> 1000**	**< 200 ns**	**< 1 µs**	**Yes**	**Yes**	**Yes**	**Yes**	

Prior Work	Ref.	Network Type	Total Node Count (N)	Reconfiguration time	End-to-end latency	Multi-cast	In-cast	Variable group multicast	All-to-all multicast	*
POPI	[85]	All-optical	1000s	> 456 µs limited by fixed TDMA schedule	500 µs below 80% load, 3 ms above	No	No	No	No	
TWIN	[86], [87]	All-optical	Unknown	Unknown	Unknown,	No	No	No	No	
Quartz	[88]	All-optical	1056	Unknown	< 20 µs	No	No	No	No	
Data Vortex	[89]	Al-optical	10000	1 µs	Few ns	No	No	No	No	
Helios	[90]	Hybrid OCS/EPS	65536	57 ms	Unknown	No	No	No	No	
Mordia	[91]	Hybrid OCS/EPS	Unknown	2.8 µs	Unknown	No	No	No	No	
Work in this thesis		**Hybrid OCS/OPS**	**> 1000**	**< 200 ns**	**< 1 µs**	**Yes**	**Yes**	**Yes**	**Yes**	

Prior Work	Ref.	Network Type	Total Node Count (N)	Reconfiguration time	End-to-end latency	Multi-cast	In-cast	Variable group multicast	All-to-all multicast	*
c-Through	[92]	Hybrid OCS/EPS	Unknown	Unknown	Unknown	No	No	No	No	
WDM /TDM PON	[93], [94]	OBS	512	10 µs	0.4-2.8 ms	No	No	No	No	
DOS	[95]	OPS with buffering	512	2-10 ns	400 clock cycles	No	No	No	No	
Proteus	[96]	OCS	2560	Unknown	Unknown	No	No	No	No	
Work in this thesis		**Hybrid OCS/OPS**	**> 1000**	**< 200 ns**	**< 1 µs**	**Yes**	**Yes**	**Yes**	**Yes**	

2. The hybrid WS-TDM star network

This chapter describes an all-optical network design which meets the system requirements specified in chapter 1, for all-optical, low-latency, fully flexible multicast communications across a 1000 node data centre. Components and subsystems necessary to realise construction of the network are presented, and experimental results showing the feasibility of the physical layer of the design are discussed.

2.1. High level system design

At a high level, the star network design proposed in this chapter can be abstracted as a single broadcast-and-select switch, with a finite total throughput capacity to be shared between all connected nodes. By creating a single shared pool of bandwidth, the network can flexibly allocate transmission rights to nodes that request connectivity, before reconfiguring as the desired traffic pattern changes. All network nodes are at the same topological layer, connected to just one switch, without a hierarchical structure. The network therefore has lower absolute latency, and lower variability in latency compared to transmitting data in several short “hops” between multiple electrical switches. This is due to buffering and queuing being implemented independently at each switch within a layered electrical switching network.

The core element of this network design is a passive star coupler, as shown at the centre of Figure 10. The broadcast-and-select property of a passive optical star coupler means that all input optical signals are split in power and distributed across all outputs. The broadcast nature of a passive optical star coupler means that multicast is inherently supported. Multicast groups can be formed up to and including all network nodes simultaneously, and there are no limits to the flexibility of multicast group creation.

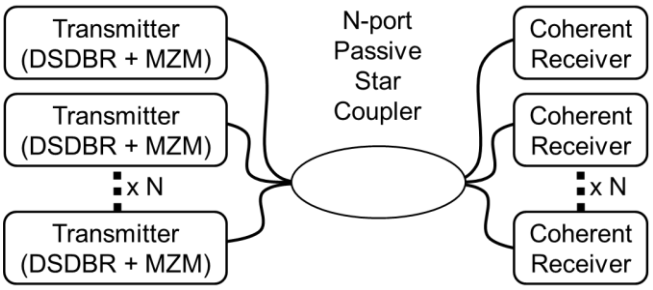

Figure 10: The star coupler network design, supporting N nodes. DSDBR = Digital Supermode Distributed Bragg Reflector. MZM = Mach-Zehnder Modulator.

Each node is equipped with an optical transmitter, comprising a tunable DSDBR laser and a Mach-Zehnder modulator (MZM). If a different wavelength is used at each transmitter, all wavelengths will combine in the core of the star coupler, and every

output port will simultaneously carry all of the input wavelengths. Tunable filtering at each output is necessary to ensure that only a single channel is received from the star coupler core without interference from other channels. In this design, a tunable local oscillator to a coherent receiver provides tunable wavelength filtering at each output.

This network design is highly reconfigurable; wavelengths can be used to partition the network dependent on the connectivity requests between transmitters and receivers. During a fixed time period defined as an “epoch”, all transmitters and receivers are each allocated a wavelength on which to send or receive data, and they remain at the same wavelength for the duration of the epoch. This effectively partitions the network by wavelength into distinct groups of senders and receivers, as illustrated in Figure 11.

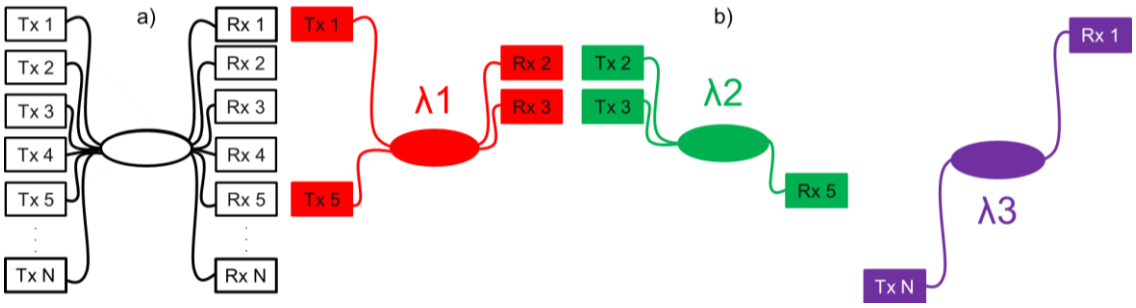

Figure 11a): An example star network connecting N nodes to a central passive optical star; b): The same physical network partitioned by wavelength during an epoch.

In Figure 11a, the physical connectivity of the fibre network shows that all senders and receivers are connected to a single optical star coupler. In Figure 11b, the star connectivity has been partitioned by wavelength during an epoch; the tunable receivers each only receive transmissions on the wavelength that they are tuned to. In between each epoch, tuning time is allowed for all transmitters and receivers to tune to new wavelengths simultaneously; no data is transmitted during the tuning time. This removes the possibility of spurious wavelengths emitted during the tuning process interfering with data transmissions through the network.

Each epoch has a fixed 2 μs duration, before transmitters and receivers all retune simultaneously over the subsequent 200 ns. The epoch duration is selected to meet a trade-off between several competing factors:

- It is desirable that the reconfiguration downtime causes at most a 10% overhead to reduce total throughput; 200 ns is a conservative estimate of the maximum time required for the receive and transmit lasers to retune between any two wavelengths and for error free data recovery. Fixing reconfiguration time at 200 ns therefore means at least an epoch duration of at least 1.8 μs $(\frac{0.2}{1.8+0.2} = 10\%)$.

- The network requires a separate control plane to perform scheduling of transmissions through the star coupler core. Work in [85] described the implementation of a control plane for the WS-TDM star network, requiring 0.5 µs to collect requests from transmitters, 1 µs to calculate and allocate transmission rights, and 0.5 µs to send allocated grants to nodes. This requires 2 µs total; a 2 µs epoch would permit scheduling to be performed just one epoch in advance. A shorter epoch time would mean a reduction in the number of nodes supported, a reduction in the number of requests or grants that nodes can send/receive, or a less optimal allocation of requests due to less computation time available.
- The data plane must be able to efficiently support large Ethernet frame sizes e.g. the maximum Ethernet packet size is 1542 bytes (= 12336 bits). At 25 Gbit/s, this packet would require 0.49 µs to transmit. A 2 µs epoch would therefore permit 4 maximum size packets to be transmitted at 25 Gbit/s in a single epoch, allowing multiple transmitters to share a single wavelength during an epoch even when the individual units of data are large.
- For 1000 nodes on the network, and assuming completely equal sharing of the total available bandwidth (i.e. 89 wavelengths each carrying 25 Gbit/s), every node could transmit 564 bytes per epoch. This could be anything from a single 564 byte packet, up to 8 minimum sized (64 byte) packets per epoch, per transmitter. For a fixed data rate, an increase in the number of nodes would result in a decrease in the mean data per epoch that a node can transmit. For 1000 nodes, a 2 µs epoch is a sensible trade-off between maximising the average amount of data that each node can transmit per epoch and minimising the time between retuning to allow high flexibility in sharing the total pool of available bandwidth.

The tunable lasers of the transmitters and receiver local oscillators can be tuned to any of 89 possible wavelengths on the ITU 50 GHz grid (an international standard to define absolute frequency of optical channels) in the optical C-band (defined as 1530-1565 nm). The available tuning range of the DSDBR lasers is limited to the C-band to match the useful range of erbium doped fibre amplifiers (EDFAs) used in long-haul optical transmission systems.

To avoid interference between transmissions within the coupler, only 89 transmissions (one per wavelength) can be launched into the star coupler simultaneously, regardless of the total number of network nodes. However, using wavelength alone to split the network would cause inefficient bandwidth allocation. Each transmitter and receiver is tuned to a single wavelength for the entire duration of an epoch, but individual packets

are of short duration relative to the 2 µs epochs (the minimum 64 byte Ethernet packet size is 50 ns at 10 Gbit/s or even 5 ns at 100 Gbit/s). If the traffic pattern and wavelength allocation only permits a single transmitter to send data for a duration much shorter than the 2 µs epoch, a large amount of potential transmission time may be wasted.

By using time division multiplexing (TDM), each wavelength can be shared by multiple transmitters during an epoch. Each epoch can be further split in the time domain into 50 “timeslots” and a transmitter can be allocated specific timeslots to transmit on its allocated wavelength. The number 50 was chosen to allow high flexibility in transmission rights allocation but without excessive controller complexity. Each transmitter can request multiple adjacent timeslots to transmit any packet size (from the length of a single timeslot up to the full epoch duration).

To avoid conflicting wavelength usage and ensure consistent alignment of timeslots, it is necessary to synchronise all transceivers to a central controller. This could be achieved using a separate star coupler to connect dedicated control transceivers at each node to a centralised network controller and clock. Details of the controller/scheduler are beyond the scope of this thesis, which focuses on the architecture and feasibility of the data plane. However, work elsewhere in [86] has shown the feasibility of computing schedules for optical star networks of 1000 nodes, where demands were collected by a centralised controller, transmission rights allocated, and grants returned from the controller to all transceivers, all within a 2 µs epoch. Further work elsewhere in [87] has demonstrated clock synchronisation and receiver phase alignment on sub-nanosecond timescales for data centre networks, verifying the feasibility of an optical packet-like switching approach.

2.2. Enabling technologies and sub-systems

2.2.1. Optical stars

Optical star couplers are simply optical power combiners and splitters - all input optical power is split equally between all output ports. Optical star couplers are directional i.e. power can flow from any input port to any output port, or vice versa, but power does not flow between inputs or between outputs. Couplers can be formed by fusing bundles of fibre together, creating a mixing region where power from each input fibre is coupled into all other fibres. Star couplers can also be formed on planar lightwave circuits, where multilayer designs can efficiently reach port counts in excess of 100x100 [88].

The key characteristics of a star coupler are the uniformity of power splitting between ports (values of 0.1 to 0.5% power variation between output ports are typical, mainly due to misalignment of the fibres during fabrication), and the uniformity of power

splitting by wavelength (±0.2 dB power variation across the C-band is typical). Both of these properties can be tuned at the point of fabrication. For use in an all-optical star network, minimum power variation between ports and with wavelength is ideal, so that each port can be considered identical.

A fibre star coupler formed from a single fused bundle of fibres could lead to high levels of optical power at the centre of the star at the point where the optical power from all transmitters is combined, as in Figure 12a. High optical powers within fibre results in non-linear interactions due to the intensity dependent refractive index (or Kerr effect), such as self-phase modulation, cross-phase modulation and inter- and intra-channel four wave mixing [89]. All of these effects would degrade data transmission performance across the coupler.

Assuming 10 dBm transmitted power per transmitter, 1000 x 10 dBm = 40 dBm at the centre of the star coupler, although assuming only 89 transmitters can be active simultaneously (one per C-band wavelength) and the rest are attenuated by 15 dB, this total power is reduced to 29.5 dBm. This is still excessively high, so to counteract this, a Banyan topology of couplers can be formed, which uses a connected network of smaller couplers to produce the same connectivity as a single large star, as in Figure 12b [90]. Since input power is split by the small couplers at every stage of a Banyan topology, the total power at any coupler is not high enough to cause non-linear effects. For example, a topology using three layers of 10x10 couplers to reach 1000 ports in total would result in a total power of 20 dBm in the centre of each coupler, which should be tolerable without signal degradation or fibre damage.

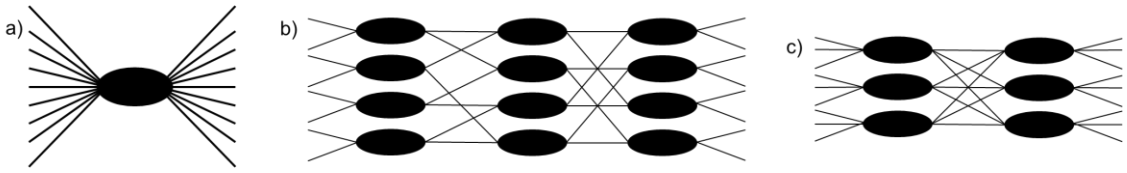

Figure 12a): A star coupler formed from the fusion of fibre tails at a single central point; b): a Banyan network formed from 2x2 fibre couplers which abstracts to a single non-blocking passive star; c): a Banyan network based on 3x3 couplers.

In practical manufacturing terms, each coupler incurs additional power losses, through both splicing and excess loss. These losses are accumulated per coupler, regardless of the coupler split ratio. Therefore, if a higher split ratio coupler is used as a building block, the total system loss can be reduced. For example, using 3x3 couplers as a building block as shown in Figure 12c requires fewer couplers on each end-to-end path than an equivalent port count network built from 2x2 couplers. This in turn requires fewer splices, resulting in a lower total system loss.

2.2.2. DSDBR lasers

Digital Supermode Distributed Bragg Reflector (DSDBR) lasers are used in this network design to enable rapid network reconfigurability, through fast wavelength switching [91]. The DSDBR laser is capable of switching across the entire optical C-band (1530-1570 nm) due to a combination of DBR gratings; one chirped grating at the front to select the coarse lasing “supermode”, and one at the rear with a finer pitch to precisely adjust the wavelength. Extra-fine tuning control is provided by a “phase” diode section, adjacent to the main gain region. Tuning of all sections is performed by current injection to the relevant section, which adjusts the effective grating pitch. The front grating is broken up into a series of smaller sections, and each smaller section has a slightly different grating pitch (i.e. the grating as a whole is “chirped”), and an independent current supply. Application of current into a particular front section causes its spectral reflectivity to match that of an adjacent section, which selects a broad range of wavelengths to reflect most strongly. Figure 13 shows the internal structure of the DSDBR laser sections.

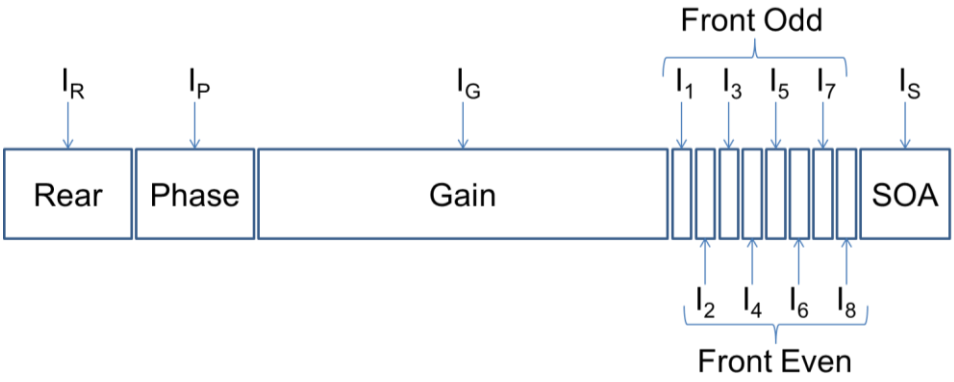

Figure 13: The internal structure of a DSDBR laser, showing the independent current injection points to enable wide ranging tunability.

The average linewidth across all tuning channels of a DSDBR laser is below 1.5 MHz, comparable with fixed wavelength DFB sources used for telecom applications [92]. However this linewidth broadens to more than 3.8 MHz during the 200 ns following a tuning event, and has been shown to take microseconds to settle, even when applying pre-emphasis to the step changes in tuning currents [93]. In a steady-state wavelength configuration, the lower the current applied to the rear section of a DSDBR laser, the lower the linewidth observed [94]. It was found that the linewidth can vary from 1.8 MHz for 60 mA current in the rear section down to 0.4 MHz for zero current in the rear section [94]. This is due to the increased carrier density reducing reflectivity in the rear section and thus reducing the coherence of the output. Although these high linewidths may cause an issue for future coherent transmission applications, they are not of great concern for the intensity-only modulation formats used in this work.

During a wavelength switching event, DSDBR lasers pass through several intermediate lasing wavelengths, before reaching the desired output wavelength. When retuning while a laser is coupled into a WDM system, the intermediate wavelength emissions can interfere with transmissions on other channels. It is possible to shutter the output of the laser during a tuning event using the integrated SOA on the output of the device, by removing the bias to the SOA so that the semiconductor junction absorbs photons [95]. This also reduces the momentary frequency drift observed during wavelength channel switching, which can be problematic in WDM systems. However, using the SOA to attenuate the optical output can also increase the time taken for a tuning event, due to the thermal changes in the front grating induced when the adjacent SOA section experiences changes in current [95]. It has been shown that SOA blanking on a microsecond tuning timescale can reduce the output power without impairing the tuning time and accuracy through thermal drift [96].

For additional attenuation during tuning times, the SOA can be operated in reverse bias. This was shown to provide more than 40 dB of attenuation, however it was most effective over millisecond timescale switching due to the thermal instabilities caused [97]. In addition, due to the packaging constraints of the DSDBR laser design, there is a shared ground plane between the SOA section and all other laser sections (front, rear and phase tuning sections). When switching high currents (~250 mA) through the SOA section, there is undesirable current leakage via the ground contacts onto other tuning sections, which affects the stability of the wavelength output, causing intermittent mode-hops.

Adding pre-emphasis to the current waveforms which drive the DSDBR laser sections, can reduce the impact of the thermal effects of switching, and remove spurious wavelength emissions during switching events [98]. It is also possible to add pre-emphasis to a heating/cooling element adjacent to the active laser elements, which can reduce the overall variation in tuning time between channel switches [99]. However, this has not been shown to reduce the total tuning time for all possible channel switching combinations. Additionally, a wavelength locking circuit has been implemented to reduce the wavelength drift during an optical burst (where data is transmitted over a few microseconds to hundreds of milliseconds duration). A wavelength feedback loop was shown to be critical to ensure low bit error rates when using a DSDBR laser to transmit optical data bursts of ms length [96].

DSDBR laser tuning stability is highly sensitive to temperature, with only 0.1° temperature change sufficient to cause mode hops, but it is possible to operate a DSDBR laser without active cooling and still maintain a constant wavelength [100]. An open loop feedback system can inject current to the laser rear section to counteract

changes in laser temperature and maintain a constant output wavelength [101]. This could reduce the power consumption of the laser by not using the integrated cooling element, but the long timescales (seconds) over which the open loop feedback operates means that the technique is not feasible for rapid wavelength switching. Additionally, stability is only maintained within a 0.1 nm range, making coherent reception difficult due to the potential for variable transmitter-LO frequency offsets.

The speed of wavelength switching in DSDBR lasers is discussed further in section 2.3.2, after an introduction to the receivers, data modulation and other sub-systems.

2.2.3. Coherent receivers

To detect an intensity modulated optical signal, the lowest complexity and cheapest solution is a square-law photodiode. The intensity envelope of the optical signal is converted by the photodiode to an electrical signal, and a decision circuit can operate on the electrical signal to decode the intensity data. Although this technique is simple to implement and low cost, intensity-only data has low spectral efficiency and requires high bandwidth photodetectors to receive high bit-rate optical signals.

An alternative method of receiving optically transmitted data uses a coherent receiver, as shown in Figure 14. An optical signal has 4 possible dimensions that can be modulated – independent signals can be modulated on the in-phase and quadrature components of both polarisation states of the light. A coherent receiver is able to split received light into these 4 dimensions, so that each can be processed independently and the data modulated onto each successfully recovered.

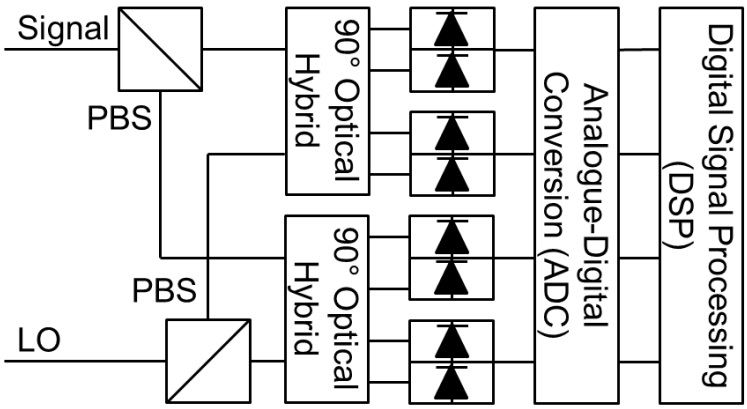

Figure 14: The internal components of an optical coherent receiver. LO = Local Oscillator. PBS = Polarising Beam Splitter.

Coherent receivers are a mature technology for optical communications and the general operating model is as follows: an optical signal entering the “signal” port of a coherent receiver is initially split into two orthogonal polarisation states. Each polarisation of the signal then enters a 90° optical hybrid, which mixes the signal with

the local oscillator (LO, an unmodulated laser tuned close to the signal wavelength) and splits out the 4 independent components of signal polarisation and phase to separate photodiodes [102]. The output of each hybrid falls on a balanced differential photodiode pair for optical-electrical conversion.

Consider the electric field from an optical signal (denoted $E_S(t)$) and the electric field from an optical local oscillator (denoted $E_{LO}(t)$), where the optical carrier frequencies are ω_s and ω_{LO} respectively. The electric fields mix in the optical hybrids of the coherent receiver, and considering only a single polarisation (i.e. either the top two pairs or bottom two pairs of balanced photodiodes in Figure 14), the optical power (P) falling on each photodiode is

$$\begin{aligned}
P_{I+} &= \left|E_{LO}(t)\right|^2 + \left|E_S(t)\right|^2 + 2\Re\left[E_s(t)E_{LO}^*(t)\exp\left(i(\omega_s-\omega_{LO})t\right)\right] \\
P_{I-} &= \left|E_{LO}(t)\right|^2 + \left|E_S(t)\right|^2 - 2\Re\left[E_s(t)E_{LO}^*(t)\exp\left(i(\omega_s-\omega_{LO})t\right)\right] \\
P_{Q+} &= \left|E_{LO}(t)\right|^2 + \left|E_S(t)\right|^2 + 2\Im\left[E_s(t)E_{LO}^*(t)\exp\left(i(\omega_s-\omega_{LO})t\right)\right] \\
P_{Q-} &= \left|E_{LO}(t)\right|^2 + \left|E_S(t)\right|^2 - 2\Im\left[E_s(t)E_{LO}^*(t)\exp\left(i(\omega_s-\omega_{LO})t\right)\right]
\end{aligned} \qquad 1$$

where $+$ and $-$ refer to the two members of the differential pairs of photodiodes, I denotes the in-phase component and Q denotes the quadrature component, and $\Re$ and $\Im$ denote the real and imaginary components of a complex number respectively.

The responsivity of each photodiode (denoted R), is the conversion factor between optical and electrical signal power. By subtracting the current signal from one photodiode from the other, a differential signal is obtained for the in-phase and quadrature components of the received optical signal (denoted I_I and I_Q respectively):

$$\begin{aligned}
I_I &= I_{I+} - I_{I-} = 4\Re\left[RE_S(t)E_{LO}^*(t)\exp\left(i(\omega_S-\omega_{LO})t\right)\right] \\
I_Q &= I_{Q+} - I_{Q-} = 4\Im\left[RE_S(t)E_{LO}^*(t)\exp\left(i(\omega_S-\omega_{LO})t\right)\right]
\end{aligned} \qquad 2$$

This derivation is identical for each polarisation, resulting in a fully diverse receiver which can demodulate data on both polarisations of the optical signal. DSP techniques can be used to rotate the received polarisation states; each received polarisation can be processed independently to recover the data modulated on the signal lasers [103].

To consider how coherent receivers provide frequency selectivity, consider I_I from equation 2 (the following derivation holds whether I_I or I_Q are considered). In an ideal system, the frequency offset between the optical signal and the optical local oscillator will be almost zero, i.e. $(\omega_S - \omega_{LO}) \to 0$. This removes the oscillating term from I_I:

$$I_I = 4R\Re\left[E_S(t)E_{LO}^*(t)\right] \quad 3$$

Consider a second optical signal also incident on the coherent receiver, at a frequency offset of $\Delta\omega$ from the original signal i.e. the absolute optical carrier frequency of this new optical signal is $\omega_{s2} = \omega_s + \Delta\omega$. The coherent receiver has only a single optical local oscillator, at a carrier frequency ω_S. By substituting ω_{S2} in place of ω_S in equations 1 to 3, the electrical signal observed at the output of the in-phase photodiode due to the second optical signal is

$$I_2 = 4RE_{S2}(t)E^*{}_{LO}(t)\exp(i\Delta\omega t) \quad 4$$

Note that in equation 4 the oscillating term remains, since the local oscillator is only aligned with the original signal and not with the second optical signal. However, assuming that $\Delta\omega$ is greater than the bandwidth of the photodiodes, the entire I_2 signal is filtered out by the low pass response of the photodiode circuitry. Common wavelength grid spacings of $\Delta\omega$ = 50-100 GHz are greater than generally available photodiode bandwidths (of the order of 30 GHz maximum), and selection of appropriate photodiodes per application can thus ensure wavelength selectivity.

In addition to wavelength selectivity, coherent receivers provide advantages in low signal power regimes through enhanced signal-to-noise ratio (SNR) performance compared to a direct detection photodiode receiver. It is assumed that only shot noise is significant in an optical coherent detection system (thermal noise is far smaller in magnitude than shot noise at laboratory and data centre temperatures, and high LO powers mean that shot noise dominates). The signal to noise ratio at the coherent receiver can be calculated for a shot noise limited system, where the following derivation is adapted from [104].

Recalling that the power output of a photodiode is proportional to the square of the current passing through the diode circuit, and combining all constant factors from equation 3 into a new constant (R), consider the square of the photodiode current:

$$\overline{i_{circuit}{}^2} = R\left(E_S(t)E^*{}_{LO}(t)\right)^2 = RP_SP_{LO} \quad 5$$

The dominant noise source of the system (assuming no optical pre-amplification) is the shot noise on each photodiode:

$$\overline{i_{shot}{}^2} = 2eR\frac{P_{LO}}{2}\frac{B}{2} \quad 6$$

for electron charge e and photodiode bandwidth B. Here it is assumed that the LO is of substantially greater optical power than the signal, so LO shot noise dominates. However, this noise level should be doubled to find the total, taking into account both the I and Q photodiodes:

$$\overline{i_{total-noise}^{\ 2}} = 2\overline{i_{shot}^{\ 2}} \qquad 7$$

The final signal to noise ratio can then be found:

$$SNR_{coh.} = \frac{\overline{i_{circuit}^{\ 2}}}{\overline{i_{noise}^{\ 2}}} = \frac{RP_S}{eB} \qquad 8$$

However, a similar derivation for a pre-amplified direct detection system would result in a signal to noise ratio of [105]:

$$SNR_{DD} = \frac{RP_S}{2eB} \qquad 9$$

resulting in a factor of two increase in achieved signal-to-noise ratio when using coherent detection, in the shot noise limited regime.

Coherent receivers have been selected for use in this system for the two reasons described above: they can provide increased sensitivity compared to direct detection, and provide fast switching filtering across WDM channels. Although the cost of coherent receiver hardware is currently higher than equivalent bandwidth direct-detection solutions, it is expected that this cost will decrease as the manufacturing volume of coherent receivers increases. Coherent receiver hardware is now commercially available, but is generally implemented in tandem with digital signal processing (DSP), which this work aims to avoid, as described in the next section.

2.2.4. DSP-free coherent reception of OOK

To successfully receive high order modulation formats, coherent receivers require complex digital signal processing (DSP) to overcome dispersion, signal-LO frequency offset, phase noise and non-linearity, by employing techniques such as digital back-propagation [106]. The complexity of operating DSP algorithms in real-time means that even dedicated DSP hardware resources such as FPGAs and ASICs are energy intensive and expensive. Equally challenging is the contribution that DSP makes to end-to-end communication latency. Signal processing can contribute to network delays, especially so in network designs where there are otherwise no buffers between

transmitter and receiver. To avoid both latency and scalability power constraints, the system design in this work implements coherent reception without DSP.

Approaches to DSP-free coherent receivers have been previously proposed: an analytical derivation was described in [107], where 2x2 and 3x3 optical hybrids were used to recover only intensity information of an OOK signal. Other experiments, using a data signal only modulated on one polarisation, showed an improved receiver sensitivity of over 20dB compared to a direct-detection receiver formed from the same specification photodiodes [108]. Polarisation diversity was ensured in [108] through the use of a polarising beam splitter on the signal at entry to the hybrid, although this resulted in a 2dB power penalty on each polarisation compared to a single polarisation signal. Using full coherent receivers with reduced complexity DSP, particularly the adaptive equaliser, was also experimentally shown to have no implementation penalty in received signal quality, and is a promising area for future research [109].

The work described in this thesis proposes the recovery of only optical intensity data at the receiver, by summing and squaring the individual electrical outputs of the coherent receiver that correspond to the electric field amplitude of the received light. This technique requires no DSP, resulting in a simple and low-complexity receiver signal path. Figure 15 shows the proposed signal processing flow after the electrical outputs of the coherent receiver (i.e. the currents shown in equation 2 for each polarisation are XI, XQ, YI and YQ in Figure 15). All of the filtering, summing and squaring blocks shown in Figure 15 could be performed in real time using passive components. This means that DSP would not be required, bringing down the cost, and energy consumption of the transceivers. However, in all experiments described in this thesis, the processing was performed offline in software, for simplicity of laboratory trials.

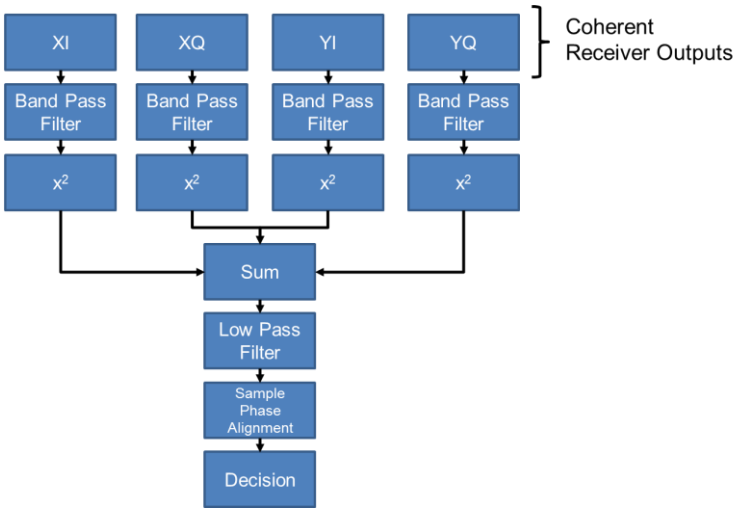

Figure 15: The signal flow following the outputs of the coherent receiver. The entire signal processing flow could be performed using passive electronic components, but in this thesis the processing was performed offline in software.

Coherent receivers can also be used to receive complex modulation formats (e.g. QPSK and higher orders of QAM), and by deploying coherent receivers in data centres, it may be possible to upgrade transmitters in future to support increased data rates. A clear upgrade path to higher data rates makes investing in coherent receiver hardware a promising and attractive concept to data centre operators. However, to increase the flexibility in sharing throughput capacity, the system design presented here uses TDM in addition to wavelength switching.

2.2.5. TDM

To enable finer grained sharing of the available bandwidth through the star core than would be possible using wavelength switching alone, the wavelength tuning epochs are further split into “timeslots”. The granularity of these timeslots can be high, serving the smallest possible Ethernet packet size in a single timeslot. Since the smallest possible Ethernet transmission is 64 bytes [110], which at a transmission rate of 25 Gbit/s would require 20.5 ns, up to 97 timeslots (since $2000 \div 20.5 = 97.7$) can be defined in each epoch. To reduce timeslot allocation complexity, in this work an epoch is defined as containing 50 timeslots, each of 40 ns. Each transmitter can request multiple adjacent timeslots, which can be combined to make a larger slot to accommodate data awaiting transmission.

A major challenge for systems with multiple inputs sequentially reaching the same output is that of phase locking and clock recovery. Phase locking is the process a receiver undertakes to match the sampling instant of the receiver to the optimum point in the bit period, for lowest errors in the received signal. To maintain a constant and well-locked phase, a clock signal must also be recovered from the transmitted data to ensure that the sampling clock of the receiver is aligned with the transmitter’s clock. Any deviation in the frequency of the two clocks will result in a phase mismatch between transmitter and receiver, and thus a reduction in system performance – even 40 parts per million frequency offset between transmitter and receiver can cause performance degradation equivalent to 20 dB signal power penalty [111].

To operate a network connecting thousands of nodes at scale, it is necessary to distribute not just high frequency clocks to align the transmitter and receivers, but lower frequency clocks (ns to µs) so that timeslots and wavelength changes can be aligned across all transceivers. A deep exploration of methods to achieve this is beyond the scope of this work, but there are several promising developments elsewhere. These include:

- Precision time protocol (PTP) [33], a standardised protocol based on measuring the round trip time of packets between two nodes using accurate

clocks at each node. PTP can reach nanosecond clock synchronisation accuracy, but can be non-deterministic if the network is fully loaded.

- White Rabbit [112], an extension of the PTP protocol, using Synchronous Ethernet between neighbouring nodes to reach sub-nanosecond clock synchronisation. White rabbit delivers high accuracy but is only compatible with tree network topologies, and as it also uses PTP to communicate between nodes, the performance degrades under high network loads similarly to standalone PTP.
- Datacenter time protocol (DTP) [113], which uses the physical layer of links between nodes within a data centre to synchronise clocks to nanosecond precision. DTP requires specialist networking hardware with adaptions to the physical layer, but is a fully deterministic protocol.

Given the feasibility of sub-nanosecond clock synchronisation across the data centre, it is receiver phase alignment which presents the lower bound to the guard bands between each TDM timeslot. Further work elsewhere in [87] has already demonstrated clock synchronisation and receiver phase alignment on sub-nanosecond timescales for data centre networks. The guard bands between each timeslot must therefore have a minimum duration of a single bit period, to account for the worst possible phase mis-alignment between two transmitters. Although TDM increases the flexibility in sharing the available network capacity, it introduces interference between transmitters at the same wavelength. The effect of this interference, and how it can be mitigated using line coding, is considered in the next section.

2.2.6. Line coding

During each epoch, multiple transmitters can be allocated the same wavelength, even though only one transmitter can be granted data transmission rights in each timeslot. All lasers that are not transmitting data must be attenuated, so that optical power does not leak into the star network core and interfere with the data signals. An SOA is integrated on the output of each DSDBR chip, and SOAs can be switched on and off in nanosecond timescales. This would allow fast switching between providing gain and attenuation respectively. However, the resulting temperature change causes wavelength instability [98], and even a nanosecond switching timescale is greater than the desired single bit time domain guard bands between each timeslot (0.04 ns for 25 Gbit/s transmission).

To perform partial shuttering, the Mach-Zehnder modulator (external to the laser package) can be set to continuously transmit a “zero” level during all timeslots where the node is not allocated transmission rights. This reduces the transmitter output power

by the modulator extinction ratio (10 to 14dB). Despite this attenuation, each unmodulated transmitter continues to send a finite optical power into the star coupler. A signal from unmodulated transmitters will, therefore, still appear on any receiver which is tuned to the same wavelength channel, interfering with valid data. In addition, each unmodulated transmitter will be tuned to a slightly different wavelength within a ±500 MHz range of the local oscillator, due to the limited accuracy of fast wavelength tuning. Each signal from an unmodulated transmitter will be received at the coherent receiver as a “beat signal” oscillating at the frequency difference between the transmitter laser and the local oscillator.

To explore the impact of the finite modulator extinction ratio, a simulation was performed of a single transmitter connected directly to a single receiver. The transmitter consisted of a DSDBR laser modulated with 10 Gbit/s bipolar on-off keyed (OOK) data. The extinction ratio of the modulator was simulated at an experimentally realistic value of 12 dB. The receiver consisted of a coherent receiver with a DSDBR local oscillator, at a frequency 200 MHz higher than the signal, using the DSP-free signal processing described in Figure 15. The power spectral density of the simulated received signal is shown in Figure 16a, with a peak at 0 Hz from the unattenuated signal power due to the finite extinction of the modulator.

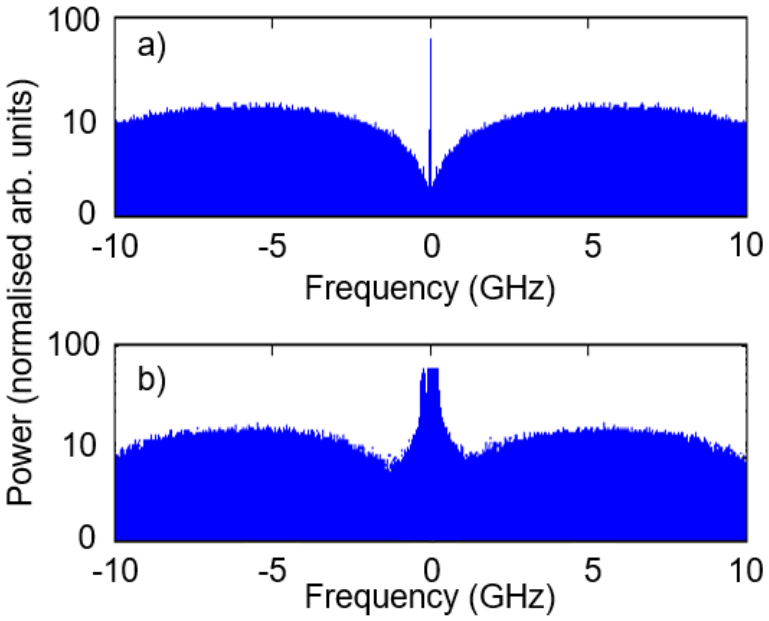

Figure 16a) Power spectral density of a single transmitter modulated with data at 10 Gbit/s; b) A single 10 Gbit/s modulated transmitter combined with 25 unmodulated transmitters at notionally the same wavelength

A further simulation was performed with the same parameters as for Figure 16a, but with an additional 25 unmodulated transmitters at the same wavelength coupled into the same fibre before reaching the coherent receiver. The additional unmodulated transmitters are all simulated with power attenuated by the attenuation of the modulator

in the “off” state (i.e. 12 dB attenuation) relative to the signal. The unmodulated transmitters were each simulated with a random frequency offset between ± 500 MHz of the LO frequency. Figure 16b shows the simulated received signal. There is a signal observed at each of the frequency offsets between the local oscillator and each of the unmodulated lasers, which has a peak power approximately 5 dB greater than the signal peak power.

For the intensity modulated data to be received without errors, the interfering DC signals must be removed from the data stream. If the signal-LO and signal-interference frequency offsets are always kept to below 1 GHz (an achievable target while maintaining fast tuning capability) and the data from the signal of interest can be manipulated to contain no low frequency components < 1 GHz, a high pass filter can be used to remove the interfering terms without corrupting the transmitted data. A band-pass filter could also be used to filter the signal, which would remove both the low-frequency terms from transmitters at the same wavelength, and any high frequency noise outside the signal bandwidth.

To shape the frequency spectrum of a data stream and remove any data signals below 1 GHz, it is necessary to apply a line code immediately prior to modulation. Line coding is defined as the choice of pattern in light or voltage at the transmitter to represent digital signals over a transmission line. Line coding of transmitted data has several purposes, including spectrally shaping transmitted data to match the properties of the transmission channel, enhancing the number of transitions between symbols to assist clock recovery, and the detection and/or correction of errors at the receiver.

It is often desirable to use line coding to reduce the DC component of any transmitted data stream, to avoid channel-specific signal degradation. For example, when using digital regenerative repeaters powered by a DC voltage over the same transmission line as the AC signals, it is necessary to avoid DC and low frequency components in the data stream to avoid ripples in the repeater power supply [114]. Optical recording systems also require line coding of the stored data for two reasons: the separation of AC components from DC to simplify the hardware required to decode the information stored on optical discs; and to overcome manufacturing imperfections when mass producing copied discs by using error-correcting line codes [115].

A further use of DC suppression line codes is in passive optical networks (PONs). When PONs are already installed and operational at a fixed bitrate, they can be difficult to upgrade, due to the reluctance of users to change from legacy low bitrate services. New services transporting high bitrate data could be spectrally shaped to remove any low frequency components [116]. This allows a higher bitrate stream to be carried over

the same fibre network and wavelength channel as the original low bitrate legacy service. Passive filtering at the receiver could separate the two different data rates.

An example line code which provides DC suppression is 8B10B, which uses a codeword system to convert 8-bit source data blocks into 10-bit coded sequences for transmission [117]. 8B/10B coded sequences are DC-balanced (meaning they contain no DC component through long-term equal likelihood of a 1 or 0 being transmitted), and have bounded disparity (the maximum possible number of consecutive identical symbols). No more than 5 consecutive ones or zeros are permitted, which results in reduced power at low frequencies compared to an uncoded data stream. An implementation of a partitioned 8B10B coder, which used 5B/6B and 3B/4B subordinate coders, reached very close to the theoretical performance limits for 8B10B encoding [117]. However, 8B10B encoding comes at the expense of a 25% data overhead lost to the additional data bits. Despite this high overhead, Gigabit Ethernet and serial AT attachment (SATA, used for PC hard drive connections) signals, among other formats, use 8B/10B line coded signals.

To improve on 8B10B line coding, 64B66B line coding only increases the size of a 64-bit data source packet by 2 bits, which is only a 3.1% overhead on the raw data. The two additional bits added to the uncoded data identify the packet as either raw data or a mixed data/control packet, and provide a guaranteed transition between 1 and 0 or vice versa. This results in a maximum run-length of 65 continuous 1s or 0s. However, the 64B66B code has almost no impact on the low frequency power of the signal, unlike 8B10B. Additionally, 64B66B provides no guarantee of DC-balance, short run lengths or high bit transition densities; they are statistically likely due to a scrambling operation carried out after the addition of the two extra bits, but have no guarantees. Nevertheless, due to the greatly reduced overhead in comparison to 8B10B encoding, 64B66B encoding is used for signal formats such as 10 Gigabit Ethernet and Fibre Channel [118].

In this system proposed in this thesis, it is essential to remove frequency components below 1 GHz, and desirable to reduce the required line coding overhead to a minimum, to maximise the total throughput of useful data traffic. Therefore, the aim is to select line coding that does not require any additional redundant bits added to the data stream while still achieving suppression of low frequency spectral content.

Bipolar line codes remove DC components from signals by translating binary data to a three level signal: two non-zero values of equal magnitude but opposite polarity (referred to as + and -), and the zero level. A simple bipolar code is alternate mark inversion (AMI), whereby source data zeros are transmitted as zero levels, and

alternate source data ones are transmitted as + or -, regardless of the number of zeros in between each one [119]. This code provides DC balance, but cannot guarantee a clock signal from repeated transitions in the data stream, meaning patterns of data with long runs of zeros can result in poor reception if a clock frequency is being derived from the received signal.

Although bipolar line codes applied to binary data produce a three-level transmission signal, it is possible to bias a Mach-Zehnder modulator such that the received intensity signal is only two levels. Receiving a three-level signal would require better SNR performance than a two-level signal, due to smaller differences in intensity between each signal level. However, by biasing the Mach Zehnder modulator (MZM) at the null-power point, the received two-level signal in power/intensity would retain the spectral properties of the line code, but not suffer from SNR degradation.

Figure 17 shows the MZM transfer function between voltage and both electric field and power, where Vπ is defined as the voltage required for a phase shift of π in the optical power transmitted. In general, MZMs for intensity modulation are voltage biased at a midpoint in power transfer. For example, biasing the modulator at a voltage of $-\frac{V\pi}{2}$ in Figure 17 would increase transmitted power for increased voltage (up to a voltage of zero), and decrease power for decreased voltage (down to a voltage of Vπ). However, for the bipolar coding used in this work, the MZM is biased at a null-power point e.g. $-V\pi$ in Figure 17. Increasing the voltage from this point would increase electric field and power, but decreasing the voltage would decrease electric field while still increasing power. Applying the bipolar modulation around this null-power point allows positive, negative and zero electric fields (or more practically $\pm\pi$ phase shifted electric field or zero electric field, given the oscillating nature of light waves) to be transmitted, and two levels of intensity (zero and one) to be received.

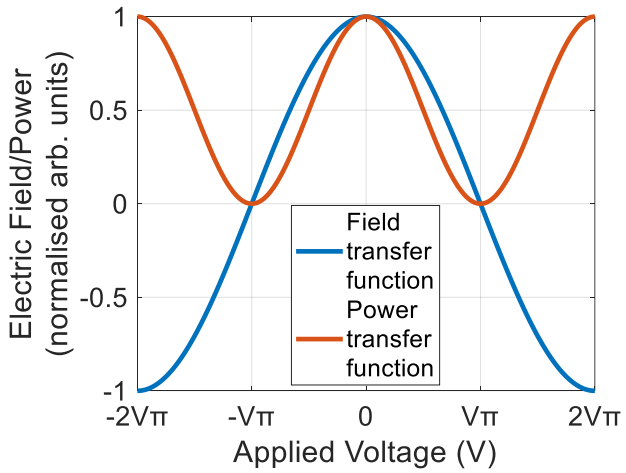

Figure 17: Mach Zehnder Modulator (MZM) transfer functions between voltage and electric field/power

By interleaving the encoding, the high frequency spectral properties of the encoded signal can also be manipulated. To perform interleaving, the incoming bit stream is split alternately into “odd” and “even” bit sets, and the coding is the performed individually on each set by independent encoders, each running at half of the full clock rate [120]. After recombining and interleaving the two encoded signals, the frequency spectrum of the full signal is altered e.g. for interleaved bipolar coding, a spectral null is also created at around 50% of the bitrate. Interleaving also increases the likelihood of transitions, but a data scrambler is still required to reduce the likelihood of long runs of consecutive symbols in a data stream, as the two encoders operate independently. The interleaved bipolar line coding outlined above can be implemented as a series of linear logical operations on an input binary data stream; this is shown in Figure 18.

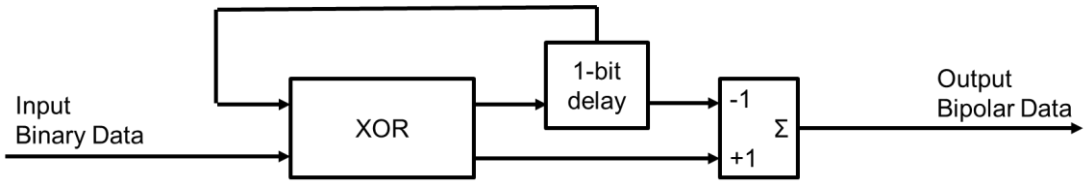

Figure 18: Interleaved bipolar line coding (IBLC) performed on a continuous stream of input binary data as a series of linear operations.

The target required for line coding in this system design is to remove any DC component and any signal bandwidth below 1 GHz. Figure 19 shows a comparison of the power spectral density of a simulated 25 Gbit/s binary signal encoded as NRZ OOK binary, and three different line codes: IBLC, 64B664B and 8B10B. In Figure 19a, which plots the power spectral density over 25 GHz, it is shown that IBLC is the only line code out of those studied which requires a lower overall bandwidth than the uncoded binary signal. In the IBLC coded signal stream, all of the useful signal information is encoded at frequencies below the null at 12.5 GHz; the total signal bandwidth is contained within half of the bandwidth of comparable 8B10B or 64B66B streams [120].

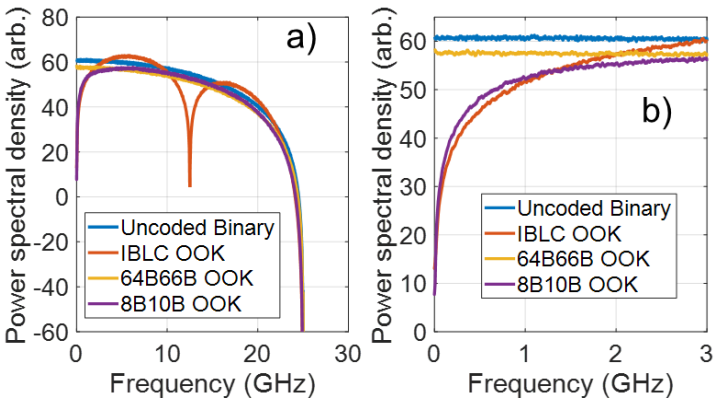

Figure 19: Comparison of the spectral power density of 2^{20} random binary bits encoded using uncoded binary (NRZ OOK), IBLC OOK, 64B66B OOK and 8B10B OOK line coding schemes, where a) shows the full bandwidth of a 25 Gbit/s signal, and b) shows the detail in the low frequency region below 3 GHz.

Figure 19b shows the detail of the low frequency (below 3 GHz) spectral shaping of the data for each of the line codes studied. For the application in this thesis, the aim is to suppress signal power in the low frequency region (< 1 GHz) as much as possible. IBLC has the most suppression, even better than 8B10B, despite the 20% overhead of the 8B10B code. The 64B66B code does not provide any low frequency suppression compared to the raw data stream, so is not useful for this network design.

Figure 20 shows the IBLC line coding in the time domain: Figure 20a) shows an input random stream of binary data; b) shows the IBLC encoding of this random stream as optically modulated onto the electric field; and c) shows the optical power eye diagram, as would appear after the summing and squaring stage of the coherent receiver output signal chain, averaged over many received bits. Figure 20c shows how the three level signal in Figure 20b reverts to the same eye diagram and detection requirements as binary NRZ when intensity detection is used at the receiver.

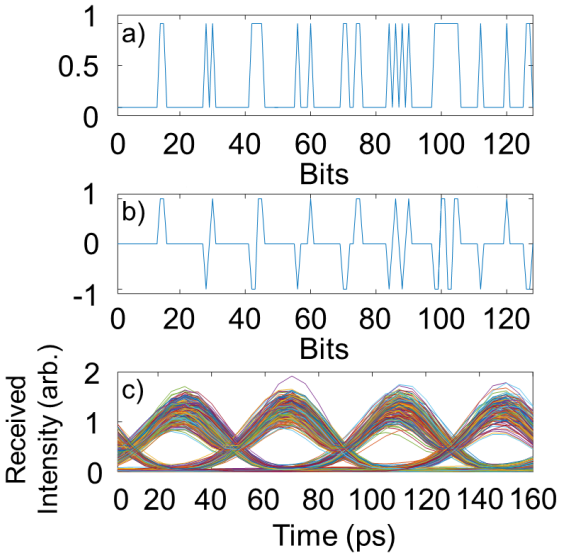

Figure 20a): A stream of random binary data with no line coding applied; b): the stream of random data from a) after IBLC line coding onto an optical electric field; c): an eye diagram showing how IBLC can be decoded as binary NRZ after square law power detection at an optical receiver.

The IBLC line code was chosen as best meeting the system requirements, and alongside the transmitter and receiver subsystems as well as the star coupler core, can be used in experiments to verify the feasibility of the entire network physical layer.

2.3. Experimental demonstrations

2.3.1. 10G and 25G receiver sensitivity

Initial characterisation of this system design was carried out at 10 Gbit/s. To assess the performance of the coherent receiver under low received optical power (< -20 dBm due to the 30 dB loss of a 1000-port star coupler), a measurement of receiver sensitivity to

signal power was made. A variable attenuator was used to gradually decrease the optical signal power from a 10 Gb/s transmitter reaching a coherent receiver. The experimental setup is shown in Figure 21, where DSDBR lasers were used both as a transmission source and as a Local Oscillator (LO) for a coherent receiver.

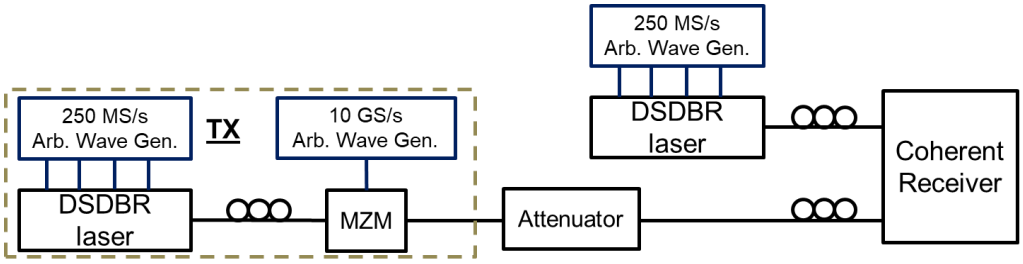

Figure 21: Experimental setup for 10 Gbit/s data plane demonstration. Arb. Wave Gen. = arbitrary waveform generator. MZM = Mach-Zehnder modulator.

Only a single data polarisation was used in these experiments, as a data centre transmitter module is only likely to include a single polarisation modulator, to reduce the cost compared to manufacturing a dual-polarisation capable device. In the experiments described in this thesis, a polarisation controller immediately before the coherent receiver was used to align the polarisation to arrive with equal power on the optical hybrid for each polarisation. This minimised the signal power arriving on all 4 pairs of photodiodes, and effectively provided the worst-case noise performance. In practice, no polarisation control is envisaged in the receiver subsystem to reduce cost, so a worst-case demonstration scenario verifies the performance in a real environment.

Each DSDBR laser was supplied with constant front, rear and phase tuning currents by a 250 MS/s arbitrary waveform generator (Arb. Wave Gen.). The gain and SOA currents of each laser were held also constant using continuous current. A continuous stream of a known sequence of 2^{20} random data bits using a 2^{13} psuedo-random bit sequence (PRBS) pattern was modulated onto the transmitter laser, captured at the coherent receiver electrical outputs, and processed offline following the signal flow in Figure 15. A swept decision threshold measurement was made on the final NRZ binary output to determine the Q-factor of the received data, and from the measured Q-factor, the BER was calculated using the relationship $BER = \frac{1}{2} erfc\left(\frac{Q}{\sqrt{2}}\right)$ where $erfc$ is the complementary error function [121].

The target received BER was 10^{-12}, as this can be considered error-free in the physical layer of Ethernet systems [110]. It is likely that a data centre implementation of this system would operate a standard Ethernet network stack, justifying this choice of target. By reaching this BER level in the data plane without any receiver DSP and/or error correction coding, the latency, network overhead and power requirements of each transceiver can be kept to a minimum.

The BER was measured as the variable attenuator was varied to adjust the signal power at the receiver, and the measurements are plotted in Figure 22. A received power of -24.6 dBm was required to achieve a received BER of 10^{-12}. Following the calculations of the loss across a 1000 port star coupler in section 2.2.1 (30 dB), and considering an achievable transmitter output power of +10 dBm, the total loss budget of 10 – (–24.6) = 34.6 dB gives sufficient power budget to support a 1000 port star coupler, with 4.6 dB implementation margin. For received powers of -23 dBm and greater, the BER approached an error floor where received optical power was no longer the limiting factor to bit errors.

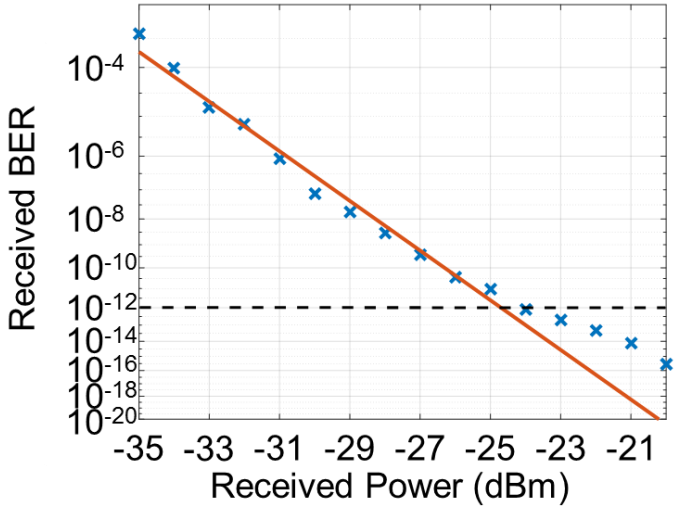

Figure 22: The sensitivity of the BER to received optical power for a stream of 2^{20} random data bits at 10 Gbit/s.

Given the feasibility of transmission and reception at 10 Gbit/s, the data rate was increased to 25 Gbit/s for a further sensitivity measurement. To carry out this experiment, the 10 GS/s Arb. Wave Gen. supplying the electrical data for modulation onto the optical carrier was changed to a 25 Gbit/s pulse pattern generator (PPG), and an external cavity laser was used as the local oscillator laser at the receiver (due to only a single DSDBR laser being available at the time of this experiment). The experimental setup is shown in Figure 23, with all other experimental details the same as those for the 10 Gbit/s experiment above.

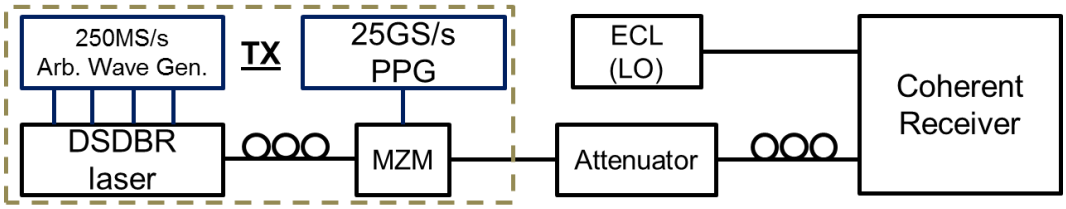

Figure 23: The experimental setup for 25 Gbit/s sensitivity testing of the data plane. PPG = Pulse Pattern Generator. Arb. Wave Gen. = arbitrary waveform generator. MZM = Mach-Zehnder modulator. ECL = external cavity laser. LO = local oscillator.

To successfully receive data at 25 Gbit/s, it was necessary to incorporate an adaptive equaliser in the received signal pathway. To keep the receiver circuit low complexity, a 10 tap T/2 fractionally spaced, decision directed equaliser was selected as an ideal candidate for this work, as it is feasible in low cost hardware. Similar processing is already found in commodity SFP transceivers, albeit in the analogue domain e.g. a continuous time linear equaliser (CTLE) based on a two-pole analogue filter. In this experiment, the equalisation was performed offline, in between the sample phase alignment and decision blocks shown in Figure 15.

As per the earlier experiment at 10 Gbit/s, the BER was measured while varying received optical power, both with and without an equaliser. These measurements are plotted in Figure 24.

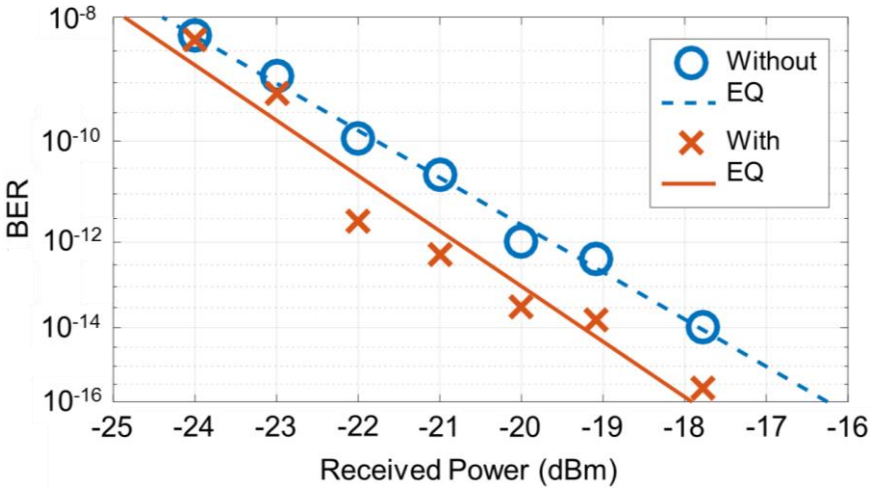

Figure 24: The sensitivity of the BER to received optical power for a stream of 2^{20} random data bits at 25 Gbit/s. The performance is plotted both with and without a low complexity equaliser incorporating in the signal flow.

The addition of an equaliser showed a performance gain of 1.2 dB in power sensitivity at the 10^{-12} BER limit, relative to data recovery without an equaliser. For a BER of 10^{-12}, the required optical power was -20.8 dBm when using an adaptive equaliser. As per the discussion for the 10 Gbit/s demonstration, assuming a transmitter power of +10 dBm, a power loss budget of 30.8 dB was feasible across the network core. This means that a star coupler supporting over 1000 nodes may be used between 25 Gb/s transmitters and coherent receivers (30 dB coupler loss with 0.8 dB excess implementation margin).

2.3.2. Wavelength switching

Fast switching has previously been studied in detail in DSDBR lasers, where the fundamental limit to tuning speed is the spontaneous carrier lifetime in the diode sections of a few nanoseconds [122], [123]. DSDBR lasers in particular can switch rapidly between wavelengths when driven by custom high-speed electronics, with

lasers shown to tune between all combinations of pairs of 32x32 channels at 100 GHz spacing within 50 ns [124] and all combinations of pairs of 64x64 channels at 50 GHz spacing within 5 ns [125]. However, these fast switches are limited in accuracy as to the wavelength that they reach, with tuning only guaranteed to within ±8 GHz or ±12 GHz of the target wavelength in [124] and [125] respectively.

An early implementation of a widely tunable fast switching laser showed the feasibility of reaching 100 distinct wavelength channels on a 50 GHz grid, but only demonstrated switching to and from 2 of the available channels [126]. By reducing the grid spacing to 2 GHz, 2000 distinct channels were shown to be available from a single laser, but with reduced side-mode suppression (often < 30 dB) and limited accuracy (tuning errors of over 300 MHz) [127]. A challenge in fast current switching of laser diode sections is the parasitic capacitance of the package, which when coupled with the high dynamic resistance of the diode at low currents, results in long settling times when switching from high to low current regimes or vice versa [128].

In prior work on integrating fast tuning lasers into full systems, often only a few wavelength switches are presented, and it is not clear if the presented switches are representative of the full laser performance. For example, despite a headline figure of 130 ns switch time between wavelength change and error-free data recovery in [129], only a single wavelength switch was demonstrated, and the switch chosen required minimal change to the current through the diode sections. It was described in [130] that wavelength switching is slowest for switching currents below 5 mA. Given that there is no direct linear relation between current and wavelength across the entire C-band, wavelength switching demonstrations must choose the wavelengths used to demonstrate switching based on diode currents rather than wavelengths, to ensure that small scale demonstrations are representative of full range performance. This work aims to present full and thorough laser characterisations, wherever possible.

For the WS-TDM star system, three separate characterisations were performed such that the total system performance is verified through their combination: the time taken to reach 90% of steady state intensity when switching from one wavelength channel to another; the time taken for laser frequency to stabilise within ±1 GHz of the target wavelength after switching (±1 GHz range was chosen such that the IBLC line code can be used to remove all interfering terms at the receiver); and the time elapsed before error-free data can be received following a switch event.

An initial experiment was performed to verify the laser switching speed by the measuring the time for the laser to reach 90% of the steady-state intensity. A single unmodulated DSDBR laser was supplied with tuning currents by arbitrary waveform

generators (Arb. Wave Gen.), so as to switch repeatedly between a pair of wavelengths. The laser tuning currents were switched using simple step functions (i.e. no pre-emphasis was applied), and the temperature of the laser was held constant.

The laser optical output passed through an attenuator into the signal port of an optical coherent receiver, and two commercially available integrated tunable laser assemblies (ITLAs) were used as LOs at fixed wavelengths (λ_1 and λ_2) into the same coherent receiver. The ITLAs comprised DSDBR lasers with pre-set tuning current drive electronics, in an integrated transceiver form factor for use in optical transmission equipment manufacture. This experimental setup is shown in Figure 25 (note that the optional data modulation was not used in this initial experiment).

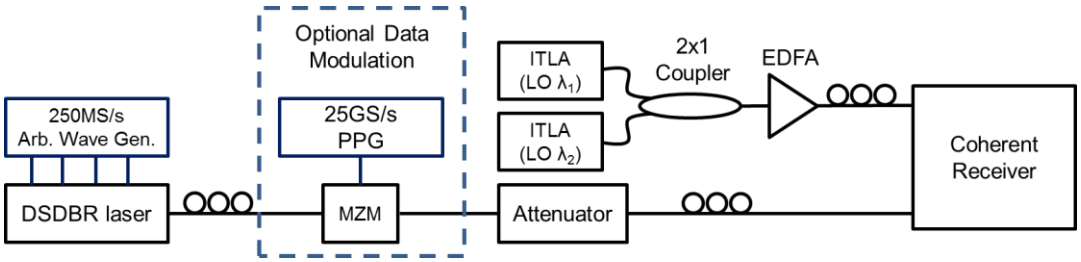

Figure 25: Experimental setup for measuring the switching time of a DSDBR laser between λ_1 and λ_2. Arb. Wave Gen. = Arbitrary Waveform Generator. PPG = Pulse Pattern Generator. MZM = Mach Zehnder Modulator. ITLA = Integrated Tunable Laser Assembly. EDFA = Erbium Doped Fibre Amplifier.

To measure the switching time by the intensity of the received signal alone, the two LOs were tuned to match the wavelength channel that the signal switched away from, and the wavelength channel that the signal switched to, respectively. A signal was observed on the coherent receiver electrical outputs when the optical signal was aligned in frequency with either of the LOs, within the bandwidth of the photodiodes (26 GHz). The coherent receiver signal outputs were summed and squared to return an intensity signal (as described in section 2.2.4), which was processed using a one-dimensional Canny edge detector [131]. A Canny edge detector calculates the convolution of an input signal with the derivative of a Gaussian function; where the convolution produces peaks, edges are found. Edge detection thresholds were set to record times where the intensity at the original wavelength fell to below 10% of the steady state intensity, and where the intensity of the new target wavelength reached 90% of the steady state intensity. The switch time duration was captured 10 times for each of the possible wavelength pairs, and an average was taken.

Figure 26 shows the cumulative distribution function (CDF) of wavelength switching times between all 89×89 possible pairs of wavelengths, as measured using the steady state intensity criterion. A CDF shows the normalised fraction of all wavelength switches tested which are shorter in duration than the switch time on the x-axis. 90% of

laser channel switches complete within 40 ns, and the median laser switching time is 12 ns. Some channel switches take noticeably longer than others, as shown by the long tail towards the top of the CDF. The channels with the longest switch times require the largest changes in the current applied to the rear grating; this effect will be explored further later in this thesis, in section 3.2.

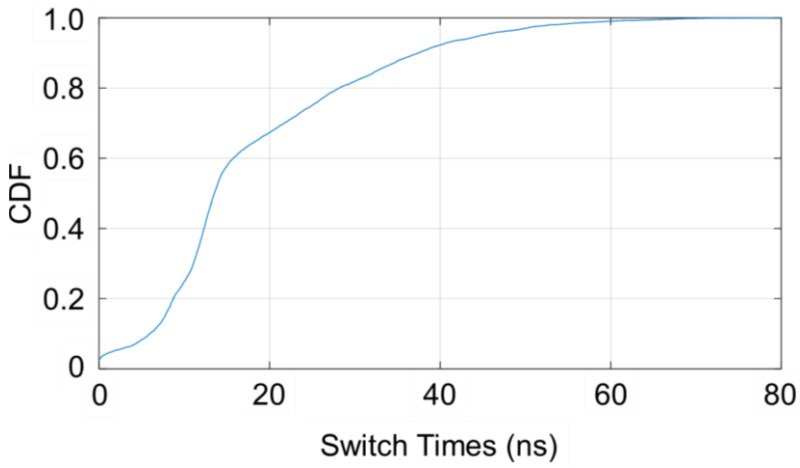

Figure 26: Cumulative distribution function (CDF) of switching time between all possible pairs of wavelengths within the C-band on an ITU 50 GHz grid spacing.

During some wavelength switches, the laser underwent mode hops, which caused the desired lasing mode to be reached briefly, before the laser mode temporarily jumped to a wavelength far from the target for a few ns, and finally returned to the desired wavelength channel. In Figure 26, tuning times are only reported upon reaching the final steady state, after all mode hops and instabilities have ceased. Figure 26 shows that it is possible to tune between any pair of channels in the C-band in less than 90 ns.

The wavelength switching time can also be measured by the time that it takes for the wavelength to stabilise following a switching event. This can be measured using the same experimental setup as in Figure 25, but instead of processing the coherent receiver outputs to calculate intensity and find edges, a time domain windowed Fast Fourier Transform can be taken of the coherent receiver outputs, processed offline. The DSDBR remained unmodulated during this experiment.

Figure 27 shows an example of the measured frequency offset as a function of time, demonstrating 3 switch events between ITU 50 GHz grid channels 46 and 44 (i.e. 100GHz total switch distance). The frequency offsets between the DSDBR and each LO are reduced to within ±1 GHz after an average of 113 ns, and after a further 50 ns settling period remain in a stable ±500 MHz range for the rest of each 2 µs epoch.

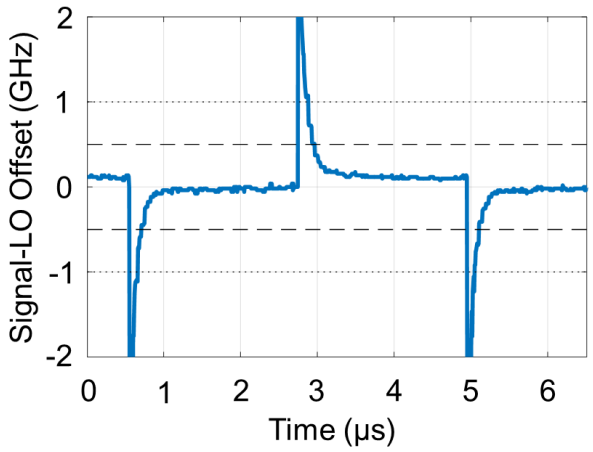

Figure 27: The frequency offset between a fast tunable DSDBR (signal) and a pair of fixed wavelength ITLA lasers at ITU grid channels 44 and 46. The DSDBR laser is switched between the two wavelength channels every 2.2 μs.

Figure 28 shows the time-resolved frequency offset when switching between ITU grid channels 2 and 87 (i.e. 4.25 THz total switching distance). The frequency offset still settles within 147 ns of the switch event, but the frequency is no longer stable and drifts over the remainder of the 2 μs epoch duration, due to thermal recovery from the large change in switching current in the rear section (this current change was not directly measured, but is estimated at around 35 mA). These frequency instabilities may cause issues for higher complexity coherent modulation formats, for instance in the tracking of the offset between the signal and an LO. However, these frequency drifts should not affect the performance of amplitude modulation, as used in this system design.

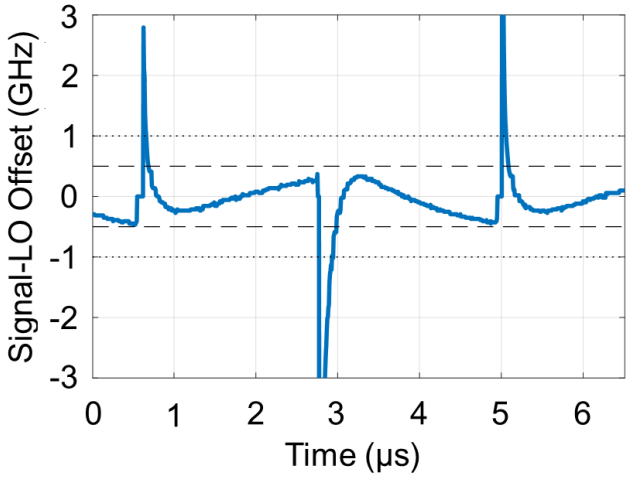

Figure 28: The frequency offset between a fast tunable DSDBR (signal) and a pair of fixed wavelength ITLA lasers at ITU grid channels 2 and 87. The DSDBR laser is switched between the two wavelength channels every 2.2 μs.

The DSDBR laser permits full and continuous use of the C-band, and wavelength tuning is not restricted to any particular grid. It should, therefore, be possible to reduce the grid spacing in future to allow elastic grids. This could permit flexibility in the modulation format used to match the achievable signal to noise ratio (SNR) e.g. lower

SNR over long distances or noisy channels. Given that the total wavelength drift observed in Figure 28 is around ±1 GHz, it is imperative that sufficient guard bands are allowed between wavelength channels to avoid cross-talk if adjacent channels drift towards each other. For example, experimental optimisation of the bandpass filter used at the receiver gave the best BER performance for a low cut-off filtering at 22.5 GHz. This means that no smaller grid spacing than 24.5 GHz should be used, to allow for a worst case laser drift of two notionally adjacent channels drifting by 1 GHz each towards each other.

To examine how wavelength switching directly impacts on the integrity of the modulated data, a continuous stream of 25 Gbit/s IBLC data was modulated onto an optical signal using a pulse pattern generator and a Mach-Zehnder modulator, as shown in Figure 25. The optical power source came from a DSDBR laser switching between a pair of wavelengths on the ITU grid every 2.2 µs; several different pairs of wavelengths were used, including switching distances from 1 to 88 ITU grid channels.

The modulated data signal passed through an optical attenuator into the coherent receiver, which was set to attenuate the signal such that -20.5 dBm signal power was present at the receiver; this power level would ordinarily allow error-free reception of data (see section 2.3.1). Both of the LOs shown in Figure 25 were also enabled into the coherent receiver, with their wavelengths statically set to match the pair of wavelengths being switched between by the signal.

There was no synchronisation between the continuous stream of data and the wavelength switching events. It was expected that the coherent receiver should recover a continuous stream of data, which is error-free. However, during a switch event, a burst of bit errors were observed in the data stream captured at the receiver, and the duration of this burst of errors was recorded as the switching time.

Figure 29 shows the recorded durations of bursts of errors i.e. the times taken to return to error-free data reception after a switching event. The subset of wavelength switches in Figure 29 is a selection of all possible wavelength switches, including full range changes of tuning current in the front, rear and phase sections, and switches between wavelength pairs spaced from 1 to 88 ITU 50 GHz grid channels apart. By selecting test cases with these properties, the performance in Figure 29 is representative of all accessible channels of any DSDBR laser. In all cases error-free reception was observed within 200 ns of the switch event.

For switching between adjacent channels, requiring only small changes to the phase and rear tuning currents of the laser, the tuning time was zero, as no errors were observed in the received data. There is no direct correlation between switching time

duration and the change in laser wavelength during the switch; this is due to tuning time being primarily influenced by the large (up to 60 mA) changes in switching current in the rear grating, which do not directly correlate with wavelength when crossing multiple laser modes. This topic is explored further in section 3.2.

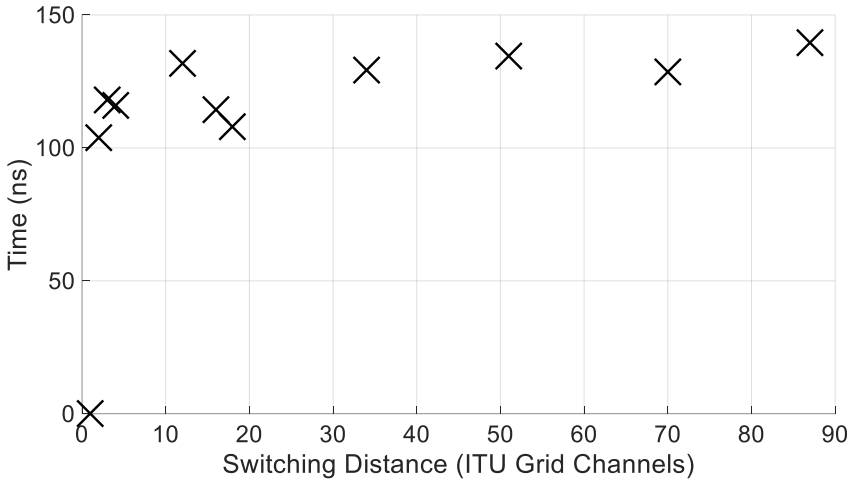

Figure 29: Switching time for a selection of pairs of switches between ITU grid channels, as measured by the time taken for data to return to error-free performance after a burst of errors caused by a wavelength switch event.

2.3.3. Maximum number of transmitters per wavelength

When using TDM to share the available capacity on each wavelength, there is a finite amount of optical power entering the network core from optical transmitters that do not have data transmission rights at any given time. Section 2.2.6 described how line coding can remove data components at low frequencies, allowing filtering to remove interfering signals. However, there is a limit to the total interference signal power, above which a filter cannot fully remove all of the interference. This limit provides a bound on the number of transmitters that can share a wavelength during any epoch.

An experiment was performed to determine the maximum number of transmitters that can simultaneously share a wavelength, before the receiver filtering can no longer remove the interference from unmodulated lasers. Due to the fast tuning of the lasers, the absolute wavelength of each transmitter laser can only be guaranteed within ±1 GHz of the target. This means that although multiple transmitters are notionally tuned to exactly the same wavelength, there will be a normal distribution of absolute wavelengths around ±1 GHz of the desired target. To simplify the experimental setup, it was preferable to use just one single laser to emulate all of the interfering transmitters, rather than requiring multiple lasers, each at a slightly different wavelength. However, this would only be a valid simplification if the same effect on BER is observed when a single interfering laser is used compared to multiple lasers each at a different wavelength.

Two simulations were performed to monitor the impact on BER of unmodulated interference on data transmission, modelling the experimental setup shown in Figure 30, where the interfering channels are shown as “1 or more ECLs”. The first simulation used between 1 and 100 interfering lasers, each with a wavelength distributed around a normal distribution between ±1 GHz of the data signal; the second simulation used a single interfering laser at a static wavelength 300 MHz greater than the data signal, at a range of power levels to emulate integer numbers of interfering transmitters. The effect on BER of the received signal when increasing the number of interfering transmitters in both simulations is plotted in Figure 31.

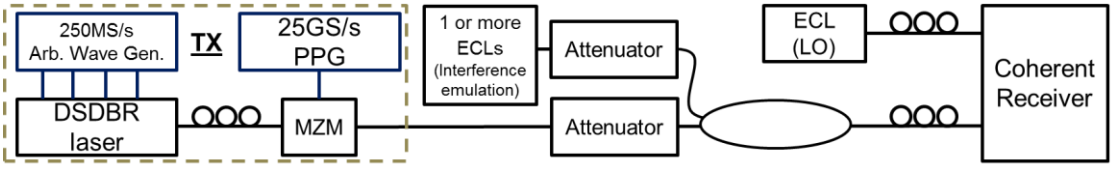

Figure 30: Experimental setup to determine the maximum number of transmitters that can share a wavelength at any time without degrading the BER performance of the single channel that is granted data transmission rights. PPG = Pulse Pattern Generator. Arb. Wave Gen. = arbitrary waveform generator. MZM = Mach-Zehnder modulator. ECL = external cavity laser. LO = local oscillator.

There is no noticeable deviation in Figure 31 between the BER performance for the system using multiple interfering transmitters, and the system using a single interfering transmitter. This result verifies that it is a reasonable experimental simplification to use a single unmodulated laser to emulate additional transmitters, rather than needing a bank of unmodulated lasers all at slightly different frequencies. In this simulation, where data was recovered without using an equaliser, any number of interferers caused the BER of the received data to drop below the error-free rate of 10^{-12}, despite the signal power at the receiver being -20 dBm, verified in section 2.3.1 to be error free. This informed the subsequent experimental procedure: an equaliser was always required to ensure an error-free system when multiple transmitters share a wavelength.

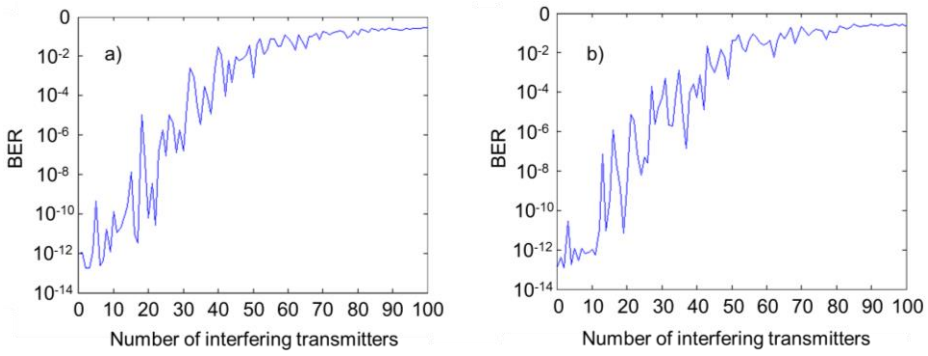

Figure 31: Simulation of the impact on BER of interfering transmitters, where in a) the interfering transmitters are individually modelled; and b) the interfering transmitters are modelled as a single laser with the power of the sum of all interfering laser powers.

An experiment was performed to assess the maximum number of interfering transmitters that could be tolerated by a receiver, with the received data remaining error-free. An ECL was used to emulate transmitters at the same wavelength as the data transmitter, and was coupled into the signal path before the coherent receiver, as shown in Figure 30.

The attenuator following the signal laser was set to allow a signal power level of -20 dBm at the coherent receiver; Figure 24 showed that this power level would normally be received error-free. By varying the attenuation of the attenuator following the interference emulating ECL, the power levels of interfering lasers were emulated. The maximum number of nodes sharing a single wavelength is defined as M. As measured at the receiver, each interfering transmitter has the average power of an operational data transmitter (-20 dBm), attenuated by the steady-state shuttered attenuation of the Mach Zehnder Modulator (measured to be 14.2 dB). The total power entering the coherent receiver due to the interfering lasers would therefore be $(M-1)\times(-20-14.2)$ dBm. This equation was used to set the received power at the coherent receiver for integer values of M. The measured BER, both with and without the simple equaliser outlined in section 2.3.1, is plotted in Figure 32.

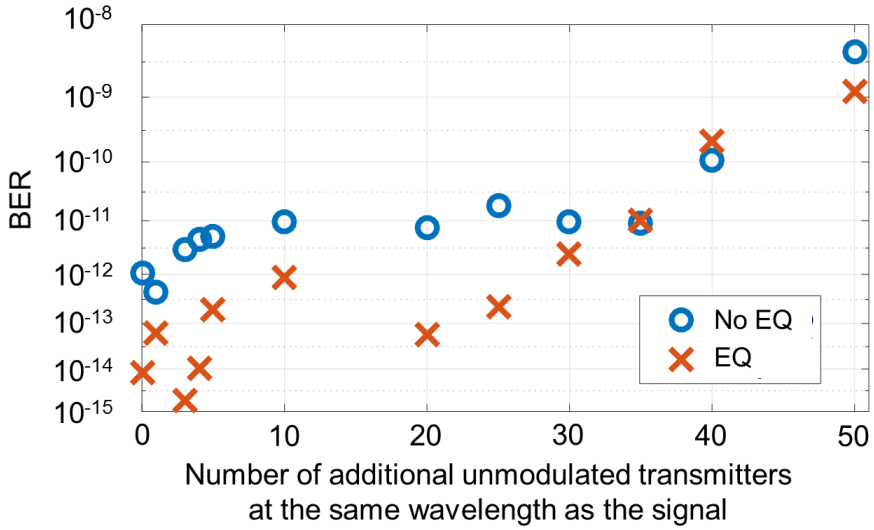

Figure 32: The impact of additional unmodulated transmitters only attenuated by the Mach Zehnder modulator on the received signal BER, with and without an equaliser applied to the signal chain.

Figure 32 shows that to retain a BER below 10^{-12}, it was necessary to limit the number of additional unmodulated transmitters to 25 or fewer, and to use an equaliser. Above this number, the power from the interfering transmitters can no longer be filtered, and the data can no longer be considered error-free. The plot on Figure 32 of the data without an equaliser is also in reasonable agreement with Figure 31, as the simulation was performed without an equaliser.

This result has implications for the scheduling of data across the network. When the central scheduler is allocating wavelengths to transmitters and receivers, this result shows that no more than 26 transmitters should be allocated the same wavelength during the same epoch. It does not matter how many timeslots each of the transmitters is allocated, as it is only the cumulative effect of unmodulated transmitters at any instant which impacts on data integrity. There are no limits to the number of receivers that can be tuned to the same wavelength simultaneously, since each local oscillator laser is completely independent of all others.

Section 2.2.2 described that 89 wavelengths at 50 GHz grid spacing are available in the C-band from DSDBR lasers, and the results of this section showed that 26 transmitters can share a wavelength in each epoch. This implies the potential for a maximum of 89 x 26 = 2314 unique transmitters to all be allocated data transmission rights within an epoch. Although the total number of nodes in the network is limited to 1000 by loss budget constraints (see section 2.2.1), it is convenient that all of those nodes could transmit some data within each epoch, should the traffic pattern permit.

2.3.4. Full system demonstration

To verify the performance of the system, a full experimental demonstration of the data plane was performed. The aims of the system demonstration were to simultaneously verify the feasibility of wavelength switching, the feasibility of TDM with minimal guard bands between packets from different sources, the BER performance of the system through attenuation equivalent to a 1000 port star coupler, and the BER performance of the system when multiple transmitters are sharing the same wavelength (using the simplified emulation of a single interfering transmitter as described and verified in section 2.3.3).

The full experimental setup is shown in Figure 33. The system has two transmitters (labelled TX); TX 1 used a fast tunable DSDBR laser, with tuning currents supplied by an arbitrary waveform generator (Arb. Wave Gen.). The tuning currents were set to switch the laser between λ_1 (1564.68 nm) and λ_2 (1524.89 nm) every 2.2 µs. 200 ns were allowed for the wavelength switching time, before each 2 µs data transmission epoch. TX 2 emulated two further optical transmitters, using a pair of external cavity lasers (ECLs), one tuned to λ_1 and the other tuned to λ_2, with both lasers coupled into a 2x1 fibre coupler. The output of the DSDBR (TX 1) and the coupler (TX 2) were connected to Mach-Zehnder modulators (MZMs), each independently driven with 10 Gbit/s IBLC PRBS data streams. Following the modulator, TX1 and TX2 passed through independent variable optical attenuators, with attenuation set to ensure that the total system loss emulated the loss incurred by a 1000 port star coupler.

At the coherent receiver, a second fast tunable DSDBR was used as a local oscillator (LO), with pre-set tuning currents supplied by a second Arb. Wave Gen. to enable fast wavelength switching. In between the two transmitters and the receiver was a 4x1 port coupler. Two external cavity lasers (ECLs), one at each of λ_1 and λ_2, were connected to the remaining ports of the 4x1 coupler, via attenuators to provide the equivalent optical power of 24 additional transmitters at the same wavelength (section 2.3.3 showed that 26 is the maximum transmitters supported at the same wavelength without impacting on the system BER, and this experimental setup already included TX1 and TX2, making 26 in total).

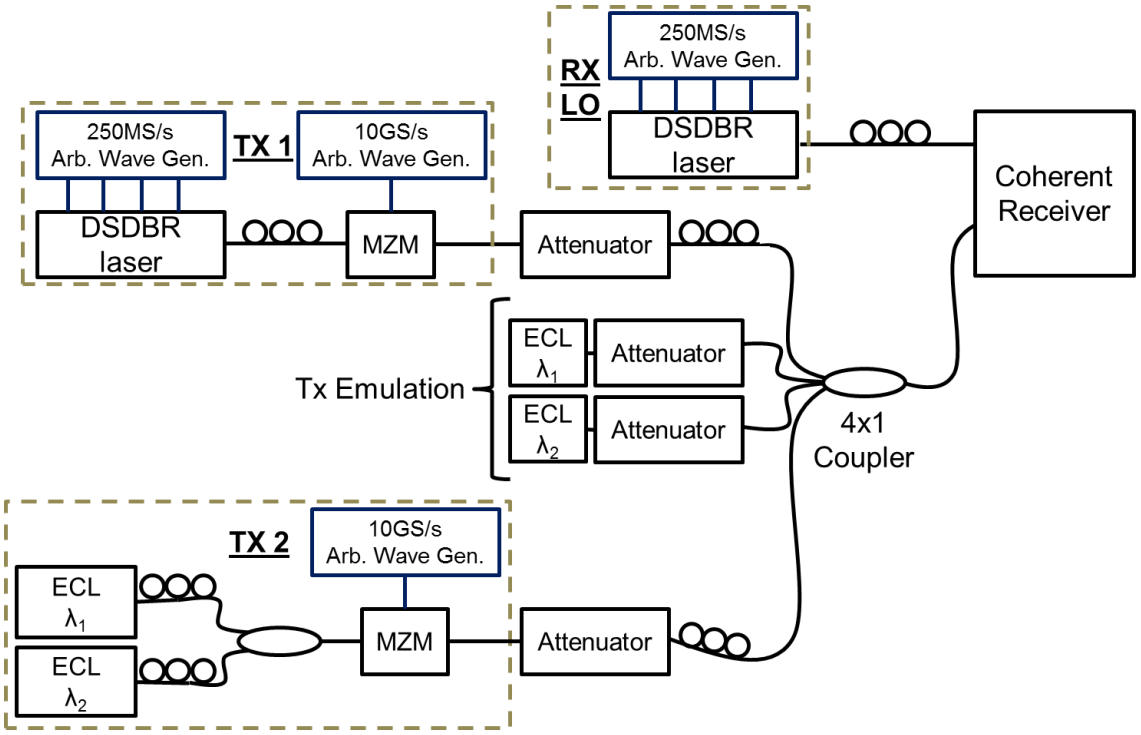

Figure 33: Experimental setup for a full demonstration of the wavelength switched TDM star network. A fast wavelength switching DSDBR laser was used as a transmitter and receiver local oscillator (LO). Other transmitters (including unmodulated transmitter emulation) used external cavity lasers (ECLs). Arb. Wave Gen. = arbitrary waveform generator. MZM = Mach-Zehnder modulator.

To verify the performance of the network in a wavelength switching regime, a pattern of switching was developed for the fast tunable lasers in TX 1 and the RX, alongside a fixed allocation of data transmission rights across both transmitters, to avoid collisions. The fixed switching pattern is shown is shown in Figure 34. All lasers were tuned to a constant wavelength for the entire duration of an epoch (2 µs) before retuning over the subsequent 200 ns. The tuning time was synchronised across all lasers in this work, through the use of an external clock and triggering system connected to each Arb. Wave Gen. TX 1 and the RX LO were both tuned to the same wavelength in each epoch, either λ_1 or λ_2. Since TX 2 comprises two constant wavelength ECLs, it was always transmitting on both λ_1 and λ_2 in all epochs.

Each epoch was divided into 50 timeslots, each of which was 40 ns in duration. At a data rate of 10 Gbit/s, this permitted 399 useful data bits to be transmitted, followed by a single zero level guard bit, to allow for phase misalignment between the transmitters. A fixed pattern of data transmission rights in each timeslot was allocated to the transmitters, with each transmitter being granted data transmission rights in alternate timeslots. This pattern resulted in the receiver during each epoch receiving a constant stream of data packets which originated alternately from TX 1 and TX 2.

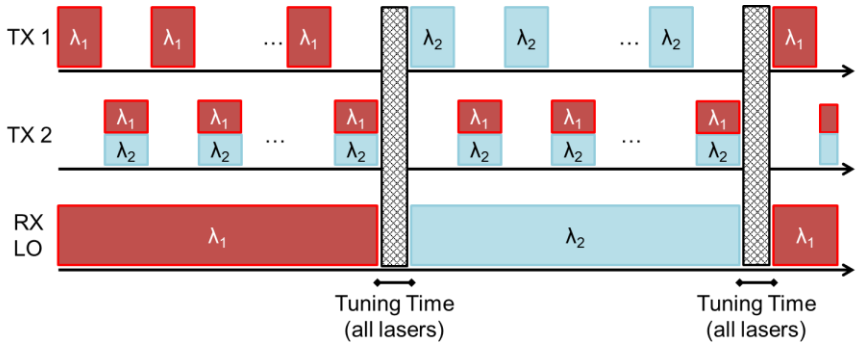

Figure 34: The wavelength switching pattern of the two transmitters (TX 1 and TX 2) and the receiver local oscilator (RX LO).

To investigate the system operation, the signals from the coherent receiver were captured and processed off-line, following the signal flow outlined in Figure 15. The receiver captured a constant stream of incoming data, including multiple repeats of the pattern shown in Figure 34.

The received signal power is shown in Figure 35 for two epochs, showing TX 1 and RX LO tuned to the two different wavelengths with a 200 ns reconfiguration gap in between epochs. Note that there was a power mis-match between the two transmitters, which shows in the figure as alternate timeslots having varying average powers at the receiver. This did not corrupt the received signal in this experiment, as the offline processing separated out each timeslot to recover data.

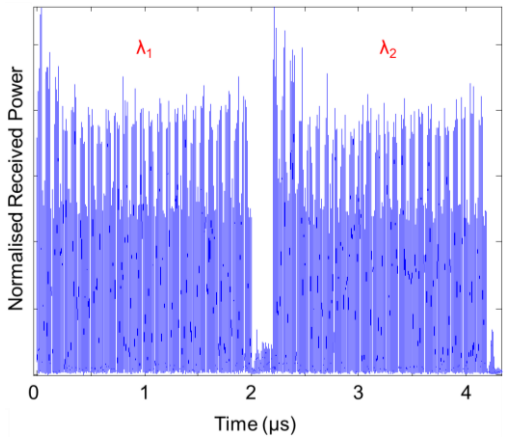

Figure 35: The received signal power, after summing and squaring at the coherent receiver. TX 1 and RX LO are tuned to λ_1 and λ_2 in each epoch.

In a real system where a receiver would continuously be receiving data from a variety of transmitters, this would cause a problem for the automatic gain control of amplifiers following the coherent receiver photodiodes, and for equalisers and decision circuits to correctly interpret the received signals. Solutions to this problem could include varying the current through the SOA that is part of the DSDBR laser package to change the transmitter power, or varying the drive signal to the MZM at each laser to increase or decrease the modulation depth, which in turn changes the optical power transmitted.

To verify the performance of the wavelength switching, the laser intensity during the 200 ns wavelength switch time in between each epoch was observed, as shown in detail in Figure 36. There was a clear drop in laser power to almost zero at 2000 ns (0 ns of the switch interval), before a small peak at 2050 ns (50 ns) followed by a drop in laser power before recovering at 2100 ns (100 ns). At 2050 ns, the small peak in power signifies that both the TX 1 and RX LO DSDBR lasers were tuned to the new target wavelength, before a brief mode-hop, followed by settling at the desired wavelength 100 ns after the wavelength switch was initiated.

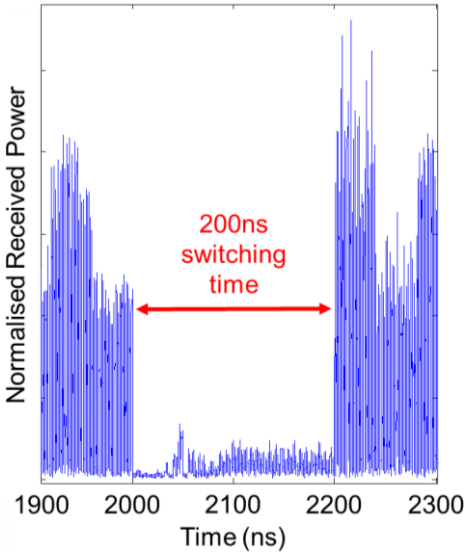

Figure 36: Detail of the received optical power during the 200 ns wavelength switching time from Figure 35.

To verify the performance of the TDM guard band between individual timeslots, the transmitter output during the gap between two adjacent timeslots was observed, and is plotted in Figure 37. It is not possible to reduce the guard interval further than the time interval of a single bit, else the worst case phase misalignment between any pair of transmitters could not be accommodated, as discussed in section 2.2.5. Figure 37 shows a single bit guard interval between the two timeslots; all data from all timeslots in this experiment was received error-free.

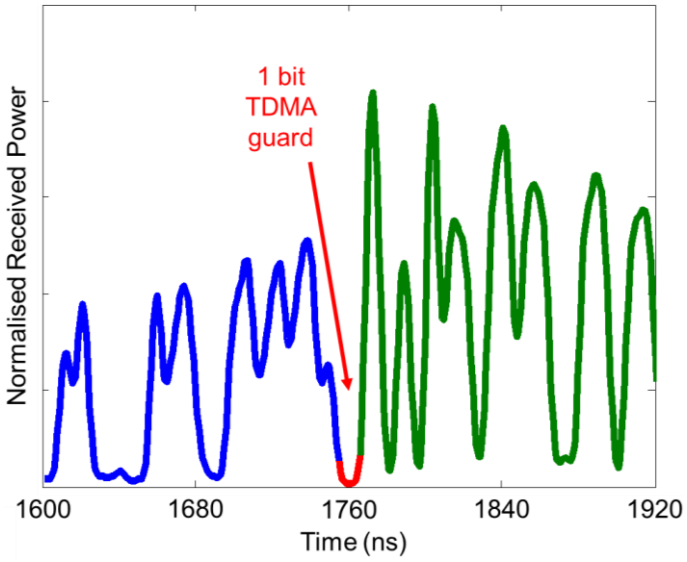

Figure 37: Detail of the single bit TDM guard band between two adjacent timeslots. Each timeslot shown was transmitted by an independent transmitter, verifying the feasibility of a single bit guard interval.

All data was received at a BER below 10^{-12} throughout this experimental demonstration, and wavelength switching within 200 ns and TDM with a single bit guard interval were both successfully verified. The combination of these results verifies the overall feasibility of the physical layer of the WS-TDM switch design. Further work, described in the next chapter, can enhance the performance of some of the sub-systems, to improve the total data throughput and network reconfigurability.

2.4. Conclusions

This chapter has introduced the WS-TDM network design, described the key subsystems required to construct the network, and presented experimental results showing the feasibility of the network data plane.

By using fast tunable optical transmitters and receivers across a star coupler core, an inherently multicast and incast capable network is formed, including full flexibility in constructing and adapting multicast group memberships. All network nodes are connected via a direct optical link, which removes hierarchical layers from data centre networks to minimise data transmission latency and reduce power consumption. Although the WS-TDM designs demonstrated in this thesis use intensity-only modulation formats at 10 Gbit/s to 25 Gbit/s, by equipping data centre nodes with optical coherent receivers there is a clear upgrade path to higher data rates per wavelength, by using higher complexity modulation formats. The WS-TDM design therefore meets all of the network design requirements listed in section 1.12.

Experimental tests and simulations of individual subsystems have shown the feasibility of connecting 1000 nodes across a single star while achieving error-free (10^{-12} BER)

transmission. Fast laser retuning was shown to complete for error-free data recovery within 200 ns, and up to 26 transmitters can share a wavelength to transmit data using TDM. A full data plane experimental demonstration showed both wavelength switching and TDM simultaneously, verifying the operation of the complete network.

Although the experimental results demonstrated the feasibility of the WS-TDM network design, there are some limits to the performance of the topology. The following chapter will outline some limits to both the total network throughput and the transceiver reconfiguration time, and suggest improvements to physical layer subsystems to overcome the challenges.

3. Physical layer improvements to the WS-TDM star network design

Chapter 2 described the WS-TDM star network design, which can support multicast traffic over a network reaching more than 1000 nodes while switching traffic within 200 ns. However, the network design as experimentally tested has some disadvantages which limit the overall network performance. These include the proportion of data transmission time lost to wavelength tuning, and the high oversubscription ratios when the network core is at full capacity. In this chapter these problems are quantified and potential solutions explored.

3.1. Drawbacks to the WS-TDM star network

3.1.1. Transmission time overhead due to laser tuning

In the WS-TDM star network design, all transmitters retune their wavelengths simultaneously at the end of each 2 µs epoch, and no data can be transmitted during the 200 ns tuning time. As a proportion of the total available transmission time, the tuning time causes a 9.1% reduction in overall throughput. The epoch and tuning time durations were set so that the tuning time caused less than a 10% overhead on the network throughput. By increasing the length of each epoch, the downtime during tuning events is shared over longer data transmission periods, so that the tuning time overhead percentage is reduced. However, this comes at the cost of reduced granularity in sharing the network bandwidth, as wavelength reconfiguration of the network is only performed once per epoch.

Depending on the connectivity requests made by the network nodes, it may be impossible to serve all requests within the same epoch that they arrive, due to the limited combinations of wavelengths and timeslots available. If a request cannot be served by the network immediately, the data packet must be held in a queue at the transmitting node, incurring latency until transmission can be scheduled. Shorter epochs are thus preferable to increase reconfigurability and thus potentially serve more requests.

The shortest possible epoch duration would contain a single packet per epoch, and would therefore allow wavelength tuning before every individual packet transmission. Assuming a single packet per epoch, Figure 38 shows the target laser tuning time, defined here as 10% of a packet duration, for a range of transmitter data rates and for both the minimum and maximum packet sizes (64 and 1542 bytes) according to the Ethernet specification [110]. Studies of real data centre traffic have shown packet sizes are likely to be in one of two ranges close to the minimum or maximum i.e. either 64-

400 bytes, or 1300-1400 bytes, but unlikely to take a value in between these two ranges [1], hence why the maximum and minimum cases are considered here.

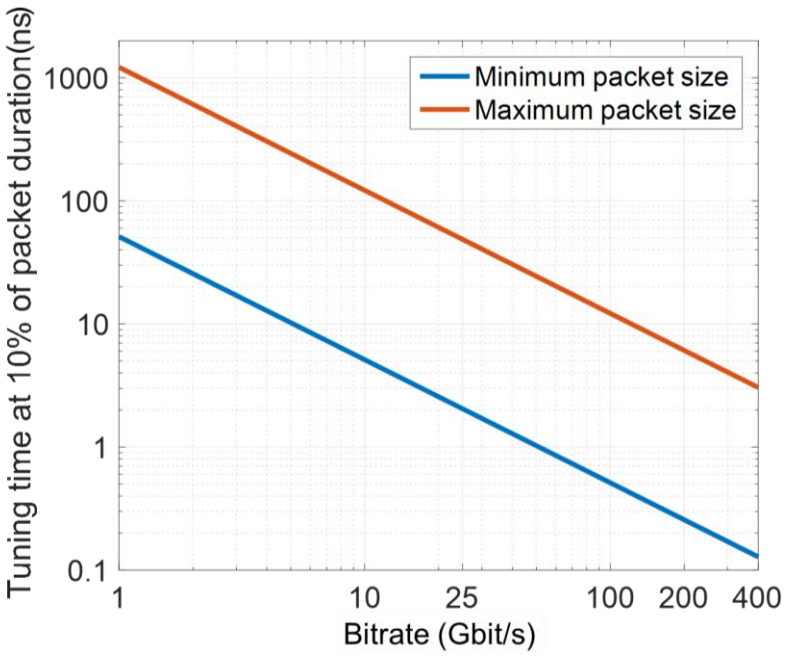

Figure 38: Laser tuning times at 10% of packet duration by link bitrate.

The 200 ns tuning time demonstrated in Section 2.3 is larger than the desired tuning time for the maximum packet size at any bitrate above 6 Gbit/s, as shown in Figure 38. For the minimum packet size at 25 Gbit/s, the desired tuning time is just 2 ns, 1% of the experimentally demonstrated tuning time in chapter 2. Therefore, it is desirable to reduce the laser wavelength tuning time, and this is explored further in section 3.2.

3.1.2. High oversubscription due to fewer wavelengths than nodes

Although the WS-TDM network design provides full single-hop connectivity between all nodes, it has a finite limit in total capacity. At the centre of a passive optical star network all wavelengths are combined, so each wavelength can only be used for a single transmission at any time (avoiding interference). This means that the total capacity of the star network, which can be shared by all network nodes, is:

$$C = WB \qquad 10$$

for total capacity C, number of wavelengths W, and bit rate per wavelength B.

By dividing equation 10 by the total number of nodes N, the expected bit rate per node (assuming equal sharing of the total capacity) can be calculated, denoted B_{EXP}.

$$B_{EXP} = \frac{W}{N}B \qquad 11$$

Equation 11 shows that for any passive optical star network where the total number of connected nodes is greater than the number of wavelengths through the star, the expected bit rate per node is less than the bit rate per wavelength. In networks where the number of wavelengths exceeds the number of nodes, B_{EXP} would not exceed B. This is due to the line rate of the transmitter providing an upper bound to B_{EXP}.

Figure 39 shows the theoretical expected bandwidth per node for a single passive optical star system with 89 wavelengths, as the number of nodes on the star is increased. The expected bandwidth per node assumes that all nodes on the star are granted an equal share of the total capacity. As per equation 11, there is a $\frac{1}{N}$ relationship between expected bitrate and the number of nodes (N) for any number of nodes $N > 89$. For 10^3 nodes, the expected bitrate per node is reduced to 8.9, 2.23 and 0.89 Gbit/s despite a bitrate per wavelength of 100, 25 and 10 Gbit/s respectively.

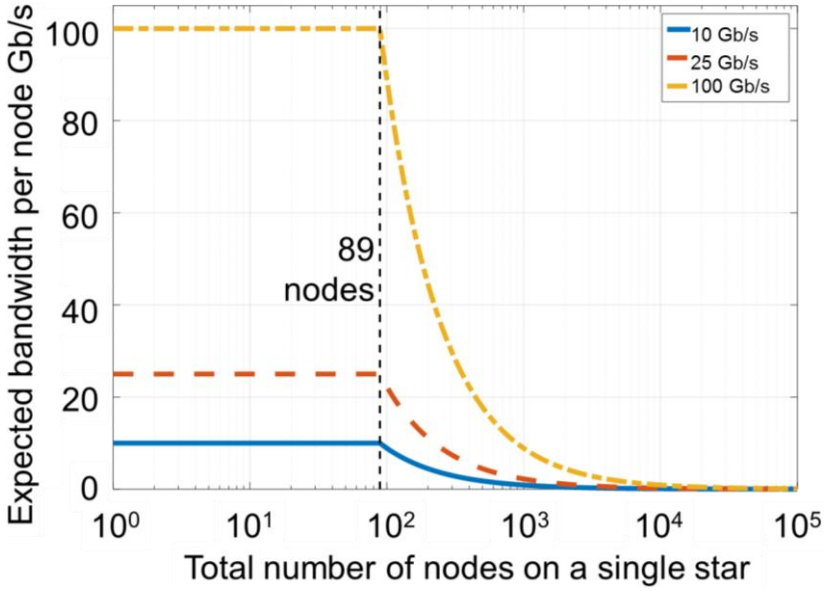

Figure 39: The expected bandwidth per node as the number of nodes on a single star network varies, where 89 wavelengths are available in the network.

Given that 50 GHz spaced channels in the optical C-band provide only 89 wavelengths, and data centre networks require at least one order of magnitude more than 89 nodes, this limit provides a major problem for the WS-TDM star design. Even though the network design provides a single, global pool of capacity to be shared between all connected nodes, when the offered network load is greater than the fraction $\frac{W}{N}$, each node has reduced effective throughput compared to the transmitter line rate.

To overcome this limit, there are a number of possible solutions. The first is to increase the bit rate B. To increase B beyond 25 Gbit/s using only an OOK format (including line coded approaches such as that described in section 2.2.6) would require high speed electronics with high bandwidth signal paths. Additionally the increased signal-

to-noise ratio requirements of higher line rate signals (required SNR is proportional to bit rate) would reduce the maximum port count of the star coupler, due to power splitting losses. To increase B beyond 40 Gbit/s would require the use of higher complexity modulation formats to transmit multiple bits per symbol. This would increase the complexity of the transmitter hardware compared to the design in section 2.2.4, as it would require digital signal processing (DSP) to successfully demodulate received signals. In a data centre environment where the number of nodes is high, the cost and energy consumption of each transceiver is a key concern, making DSP unattractive.

The second solution is to increase the number of wavelengths W. This could be achieved by spacing the channels closer together i.e. rather than 50 GHz spacing between channels, perhaps 37.5 GHz or even 25 GHz could be used. This places strict requirements on the stability of the wavelength that the tunable laser must maintain during an epoch. In section 2.3.2, Figure 28 showed how the central frequency of a transmission can vary by ±0.6 GHz after a tuning event. This drift places a lower bound on the amount of guard interval required between adjacent channels, to avoid crosstalk between channels as wavelengths drift over time.

To avoid issues with wavelength drifts across tight frequency grids, an alternative way to find more wavelengths is by extending the operational range of the lasers (both within transmitters and receiver local oscillators). Despite the DSDBR laser itself only being certified for use within the optical C-band, it is possible to produce extra wavelengths at either end of the C-band using customised drive electronics and tuning current mapping. This is explored further in section 3.2.

A final solution explored in this work involves the splitting of the star coupler network so that the expected bitrates above are a worst-case bound on the network performance. By intelligently designing stars formed from smaller couplers, sub-stars could be created, each with their own finite capacity. The sub-star topologies can be reconfigured to match the traffic patterns as they change over time, maximising throughout. Split-star designs are considered further in chapters 4 and 5 of this thesis.

3.2. Fast wavelength tuning

3.2.1. Laser characterisation and setup

DSDBR lasers require full characterisation before use as tunable optical sources, due to their large tuning range using multiple tuning diode sections, as described in section 2.2.2. A “tuning map” can be created to record how the laser output wavelength varies with applied current [92], [95]. To create a tuning map, a pair of adjacent front sections is selected, and the lower number of the front section pair (corresponding to the lowest wavelength grating) is held at a constant current of 5 mA. The adjacent front section

has its injection current incrementally increased from 0 mA up to a maximum of 5 mA, while the current to the rear section is also incrementally increased (between 0 mA and approximately 60 mA). The peak wavelength output is measured for each current combination using an optical spectrum analyser (OSA). This process is then repeated for all pairs of front sections and all possible rear currents. To perform a tuning map measurement is time consuming (up to 18 hours), although the process can be automated. However, it is essential to perform tuning characterisation for every individual laser diode, due to manufacturing tolerances in the laser diode structure.

A tuning map was recorded for a single DSDBR laser, using FPGA-controlled digital-to-analogue converters (DACs) to independently drive current into each tuning section of the laser. The laser optical output was connected to an OSA via an optical attenuator, as shown in Figure 40; the tuning map of the wavelengths recorded by the OSA for the full range of all tuning currents are shown in Figure 41.

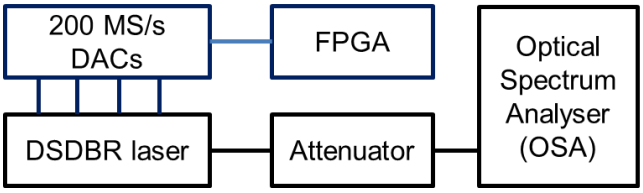

Figure 40: Experimental setup to record a tuning map for a single DSDBR laser.

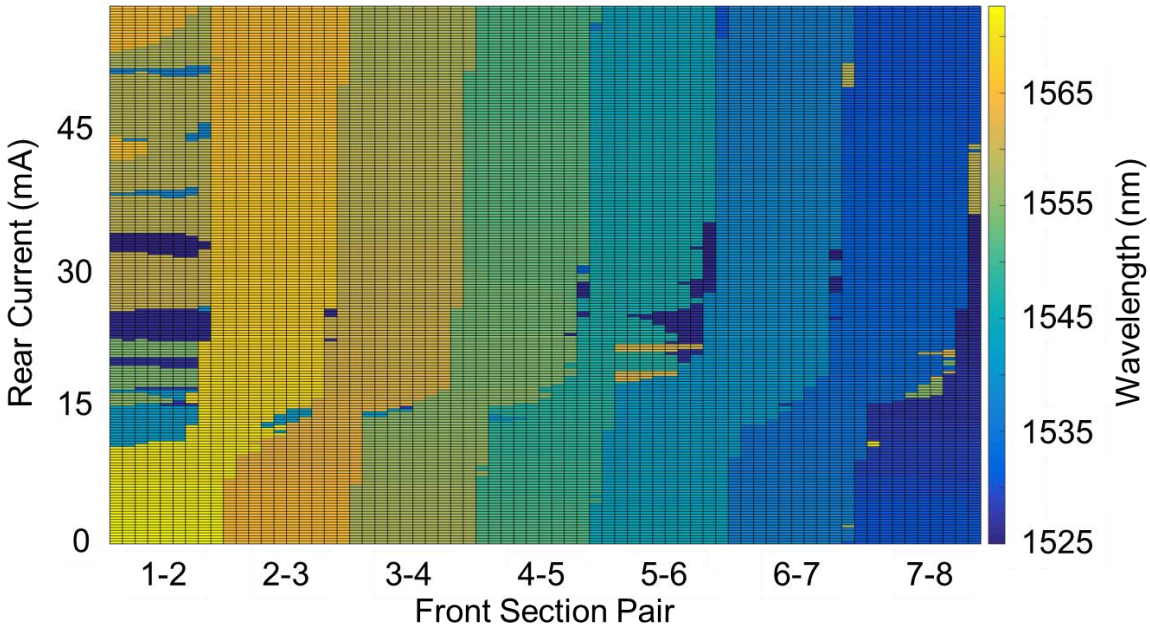

Figure 41: An example tuning map for a DSDBR laser, showing the variation of the laser output wavelength as measured by an OSA while the currents through the front and rear tuning sections were varied. The laser gain, SOA and phase currents were all held constant.

Figure 41 shows that a single laser can reach all wavelengths in a continuous range throughout the optical C-band (1530-1565 nm) and 5-7 nm either side. A grid of

wavelength channels can be selected as operating points for transmissions across a network, so that there is no interference between data transmitted on adjacent wavelengths. The international standard grid system was developed by the International Telecommunication Union (ITU) which specifies a set of frequency spacings from 12.5 GHz – 200 GHz [132]. Each frequency on the ITU grid is called a "channel". Throughout the remainder of this thesis, the channel numbers used will refer to the ITU 50 GHz grid, where channel 1 is defined as 191.35 THz (1566.723 nm), and frequencies increase by 50 GHz sequentially with channel number.

At manufacture, DSDBR lasers are only certified to provide 89 channels, spaced 50GHz apart, within the optical C-band. By using more current combinations than are suggested in the device datasheets, up to 120 channels at 50 GHz spacing can be found. Front sections 7-8 are not recommended for use, due to the small current range over which the lowest wavelength supermode is seen (front currents at 5mA in both of the front section pair, or rear currents below 15 mA), but using a precision current driver could enable the use of that front pair and therefore extend the device capabilities for 5 nm below the optical C-band. Front section pair 1-2 can only be used in the low rear current regime (< 12 mA), as any wavelengths at higher rear currents while using that front section pair can be unstable. This can be seen in Figure 41 where the wavelength jumps between several supermodes (large changes in wavelength) for small rear current changes (< 5 mA) when using the front section pair 1-2 and greater than 12 mA of rear section current. However, the wavelengths available using low currents could still increase the laser tuning range by a further 3 nm above the optical C-band, making a total tuning range gain of $\frac{8\ nm}{35\ nm} = 22.9\%$.

To produce any wavelength within the laser's tuning range, the rear section current, front pair selection and front section currents can all be read from Figure 41. A lookup table can be produced of the DAC values required to provide the currents for every wavelength grid channel, and electronic switches can configure the current flow from the DACs to the required pair of front sections. Further fine tuning to precisely adjust the laser wavelength can be performed by varying the laser phase current.

The output wavelength of a DSDBR laser has a non-linear response to the applied rear section current. This can be observed by holding all tuning currents at a constant value and varying only the rear section current; a measurement of the variation of output wavelength with rear current can then provide the required resolution of DACs to directly drive DSDBR lasers, to ensure that sufficient resolution bits are available in the low current regime. Figure 42 shows a vertical cross-section through Figure 41, for the

variation in laser wavelength due to a variation in rear section tuning current alone (gain, SOA and all other tuning section currents are constant).

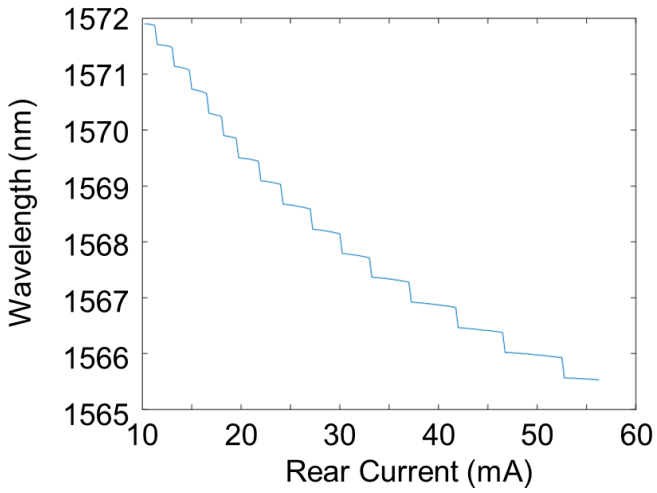

Figure 42: The variation of laser wavelength with current applied to the rear section with all other currents held constant.

Although in the high rear current regime the longitudinal modes of the laser are separated by ~4 mA steps in rear current, in the low current regime the longitudinal modes are separated by only ~1.5 mA steps in current. It is necessary to use a DAC resolution of at least 0.05 mA, as the wavelength can vary slightly with the applied rear current even within each longitudinal mode. This is shown by the negative gradient throughout Figure 42, with no horizontally flat regions of constant wavelength.

Having specified the DAC resolution required and observed the tuning pattern for the DSDBR laser, an automated method can be found to calibrate the FPGA lookup table to the wavelength grid, optimised for fast wavelength switching and based on the laser side-mode suppression ratio (SMSR).

3.2.2. Automatic wavelength tuning with SMSR optimisation

The side-mode suppression ratio (SMSR) of a laser is defined as the difference in power between the optical spectrum peak power and the second highest peak power, and is usually measured in dB. SMSR is a measure of the quality of the laser output; a high quality laser beam would have a single wavelength output with low or no power (at least 30-40 dB SMSR) in any competing modes. The optimisation of SMSR is therefore a useful tool for laser wavelength calibration, to ensure that high quality laser beams are selected for all desired operating wavelengths.

After creating a tuning map, the conventional method for finding laser wavelengths on the ITU grid is to trace linear diagonal lines on a semilog-plotted tuning map, at the centre of each super-mode. These diagonal lines exponentially relate the front and rear currents [92]. By incrementing the front and rear currents along the traced diagonal

lines, super-mode jumps can be minimised, while reaching all channels within the laser's wavelength range. The laser gain and SOA currents can also be varied if it is necessary to ensure a constant output power across all channels [91].

When the laser is to be used in a steady state configuration this method works well to accurately find wavelength channels on ms or longer timescales. However, when performing fast wavelength switching, the laser current dynamics on a nanosecond to microsecond timescale mean that the tuning map is not an accurate representation of the wavelength expected for any current combination. In addition, this method of calibrating laser wavelengths is slow, due to the need to create a full tuning map for each individual laser chip. Each data point on a tuning map requires a full measurement sweep of an OSA, which can take 10 seconds (assuming a resolution of 0.02 nm over a bandwidth of 50 nm). Multiplying this single sweep duration by 1024 steps in rear current and 64 steps in front current results in more than 18 hours being required to produce a tuning map for each new DSDBR device.

An alternative method to find steady state tuning currents is presented here, which only requires a tuning map to be produced for a single DSDBR laser chip. This should both reduce the time taken to calibrate the laser tuning, and more importantly, improve the fast tuning characteristics of the laser through the choice of tuning currents. Most wavelengths can be reached by several possible tuning current combinations, but not all of those combinations are sufficiently far from mode boundaries that fast tuning can be successfully realised. By selecting tuning currents which are furthest from mode boundaries, fast switching should be easier to optimise.

When the tuning map in Figure 41 was recorded, measurements were made of both the laser wavelength and the side mode suppression ratio (SMSR). It is possible to achieve an SMSR of 45 dB or more across all channels [133]. A "mask" was created from the measured SMSR data, by plotting a binary pattern of current combinations which result in an SMSR greater than 45 dB (the mask contained "1" for current combinations giving SMSR > 45 dB and "0" for SMSR < 45 dB). By multiplying the SMSR "mask" by the wavelength tuning map shown in Figure 41, a new tuning map was created on which only tuning current combinations which result in SMSRs of 45 dB or greater were displayed. The masked tuning map is shown in Figure 43.

Having removed the low SMSR regions of the tuning map by using the mask, the laser wavelengths that were closest to the desired wavelength channels were located. To minimise the likelihood of mode hops when fast tuning the laser, regions of the tuning map close to low SMSR regions were avoided, because poor SMSR is an indication of the boundary between modes. This resulted in a set of candidate tuning currents being

selected for each wavelength channel. The candidate tuning currents are shown in Figure 43, where 50 GHz spaced ITU grid channels are highlighted in yellow on top of the masked tuning map. Note that the yellow highlighted locations do not border any currents where a low SMSR was recorded, and that the pattern of yellow locations do not fit an exponential relationship between front and rear current, in contrast to the established calibration methods in [92].

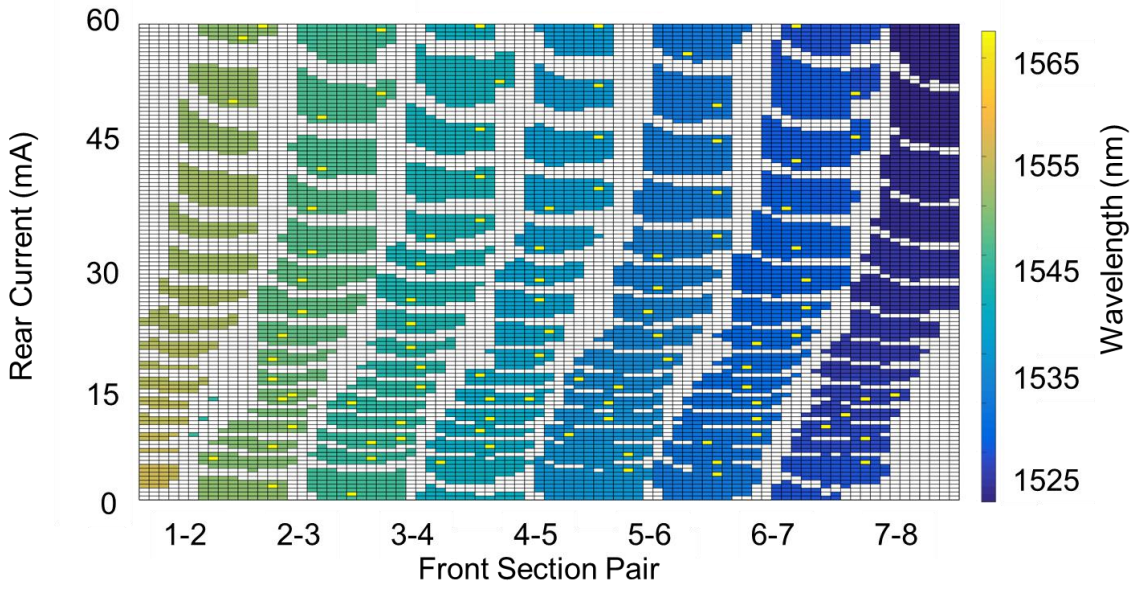

Figure 43: A DSDBR laser tuning map of wavelength against front and rear currents, masked by a side-mode suppression ratio of 44 dB or better.

The candidate tuning currents for one laser chip can be used as initial seeds for an alternative algorithm (compared to that described in section 3.2.2) to find optimum fast tuning currents for any new laser chip, improving on the fast tuning wavelength stability and reducing the time taken to calibrate each individual laser chip. For optimum fast tuning performance, the target wavelengths should be accessed using tuning currents that are as far as possible from poor SMSR regions, i.e. at the centre of the coloured regions in Figure 43. This can be achieved by using the initial seed currents found at the centre of coloured regions in Figure 43 (these initial seed currents are highlighted in yellow) and iteratively adjusting the laser phase and rear section currents while monitoring the laser optical output on an OSA to simultaneously maximise SMSR and minimise offset between the laser output wavelength and grid wavelength. An automated algorithm for this process calibrated the laser to 96 channels on the 50 GHz ITU grid in 90 minutes – a reduction of 91% compared to using the method from previously published work. The static tuning currents found using this process also form an optimised basis for the following work on fast wavelength switching.

3.2.3. Challenges in observing fast switching

When characterising the fast switching performance of the laser, it is difficult to precisely observe the wavelength of the optical output at nanosecond resolution. To aid in laser optical output monitoring, the DSDBR laser package is equipped with two photodiodes which measure the optical transmission through 100 GHz spaced gratings. A fractional tap of the laser optical output is coupled into both gratings, and each photodiode current is proportional to the frequency offset between the laser output and the nearest multiple of 100 GHz. These photodiode signals can be used to determine how close the laser wavelength is to a grid channel. However, they have low bandwidth (~500 MHz, limited by poor packaging) so cannot provide nanosecond resolution, and they do not provide information of the absolute wavelength that the laser is currently emitting, only the proximity of the wavelength to a grid channel.

To experimentally investigate the wavelengths emitted over time during a laser switching event, a single DSDBR laser was repeatedly switched between channels 38 and 21, with the laser output connected via an attenuator to a coherent receiver. A tunable laser assembly (containing a DSDBR second DSDBR laser) was used as a local oscillator (LO), as shown in Figure 44. The LO laser was statically set at each of 96 wavelengths in turn, and the coherent receiver electrical outputs were captured using an 80 GSample/s oscilloscope. These outputs were processed to record the optical intensity at each wavelength during a switching event, as shown in Figure 45.

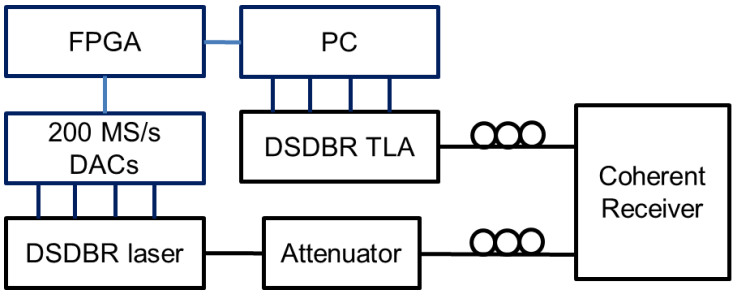

Figure 44: Experimental setup to observe laser fast switching properties. TLA = Tunable laser assembly.

The optical power emitted at intermediate wavelengths when switching (shown between 3.8 and 9.5 ns in Figure 45) would interfere with other transmissions across an all-optical network if the switching laser was not attenuated sufficiently that the spurious wavelengths do not enter the fibre. Figure 45 also illustrates the challenge in using the photodiodes on the DSDBR laser package. Due to packaging constraints, the signals from the photodiodes have a bandwidth of approximately 500 MHz. The rise and fall times of the successive wavelengths produced by the laser are of such short duration (< 0.2 ns), that the photodiode signal path does not have sufficient bandwidth to show the gaps. It could therefore be erroneously recorded that a wavelength switch

has completed while the optical output is still passing between intermediate wavelengths, if the photodiodes alone are used to monitor the laser output.

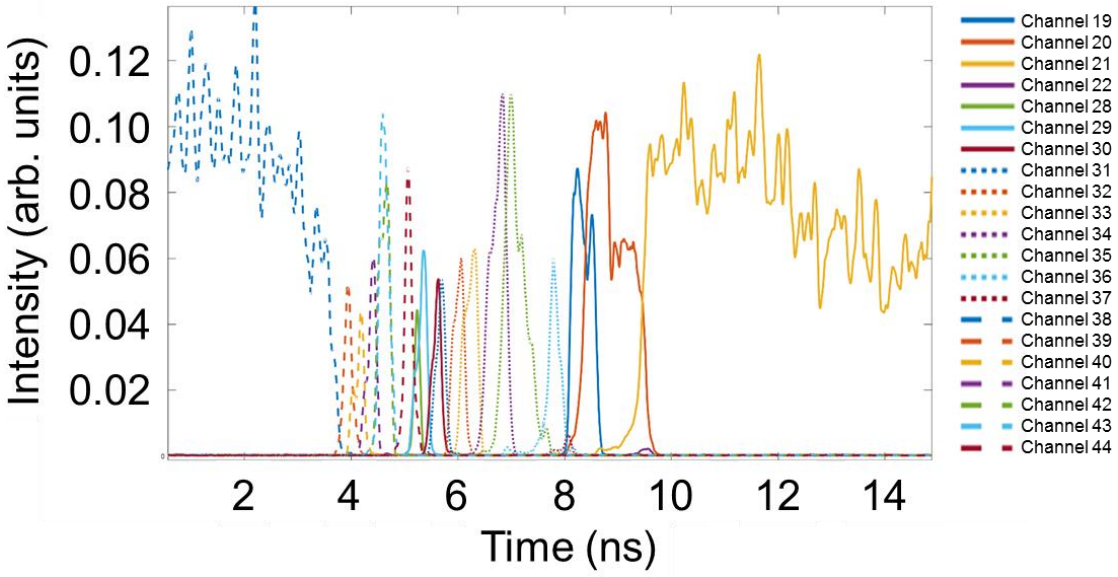

Figure 45: Time-resolved laser wavelengths when switching between C-band channels 38 and 21.

Observing the laser wavelengths at sub-nanosecond time resolution provides conflicting wavelength information compared to observations of the laser using the OSA. The steady state wavelengths which are observed on an OSA are time averaged measurements over ms timescales, meaning that the steady state values are only reached after a much longer duration than the minimum Ethernet packet duration at 10 Gbit/s or higher (51 ns or smaller). For example, a time resolved measurement of the wavelength of a DSDBR laser immediately after a wavelength switch is shown in Figure 46.

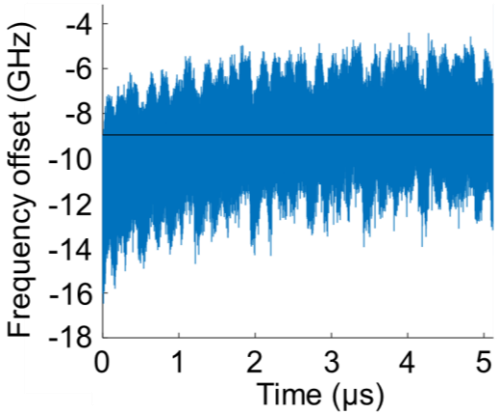

Figure 46: The frequency offset between a switching DSDBR laser and a statically tuned DSDBR laser, immediately after the DSDBR laser has switched from channel 21 to 85.

Figure 46 shows the frequency offset over a 5 μs period after a fast switching event from channel 21 to channel 85. The black line through the middle of the figure shows

the steady state frequency offset value of -8.8 GHz, which is reached within 2.3 µs; immediately after the switch, the offset is -11.5 GHz

Given that the epoch duration proposed in chapter 2 was 2 µs, and that the minimum Ethernet packet at 10 Gbit/s has a duration of only 51 ns, the steady state wavelength would not be reached on either of these timescales. An alternative method of using current to rapidly tune the laser must be found, and despite the challenges in observing DSDBR fast switching performance, a reduction can still be made to the wavelength switching time, using current pre-emphasis.

3.2.4. Current drive pre-emphasis

Current drive pre-emphasis has previously been experimentally demonstrated to reduce the wavelength tuning time of multi-section DBR lasers [93], [98]. However, few details are presented in the literature which quantify the amount and type of pre-emphasis required. Most prior work demonstrates only the speed of only a few wavelength switches rather than all possible combinations. Additionally, the switches selected for publication were often adjacent wavelengths or deliberately chosen for minimal current injection or draining required to change the laser wavelength.

In this work, an effective pre-emphasis method is found for not only a select few wavelength switches, but all possible combinations across the optical C-band. By observing the pattern of front and rear currents required for laser wavelength switching across a sub-set of wavelength switches and applying an iterative search, an effective pre-emphasis model was determined. This model is related only to the change in rear current between the two channels being switched.

To control the laser wavelength, 4 independent DACs were used to supply current to odd numbered front sections, even numbered front sections, the phase section and the rear section respectively. Circuit switches were used to connect the even and odd front section DACs to one laser diode section each, enabling the full tuning range to be reached with only 4 DACs (rather than 10 independent DACs if each DAC was directly connected to a laser section). Each DAC operated at 200 MSample/s, and was controlled by an FPGA to produce up to 4 samples per wavelength switch event. The 4^{th} sample was then held for the duration of a wavelength tuning epoch to produce the steady state wavelength. The experimental setup used to measure the laser switching time is shown in Figure 44.

To assess which laser tuning sections have the largest effect on the laser tuning time, the time resolved wavelengths observed during a tuning event can be overlaid on a tuning map. Figure 47 shows the steady state currents for each of the intermediate

channels seen during the switching event previously shown in Figure 45, with lines connecting the channels observed during the switch event in chronological order.

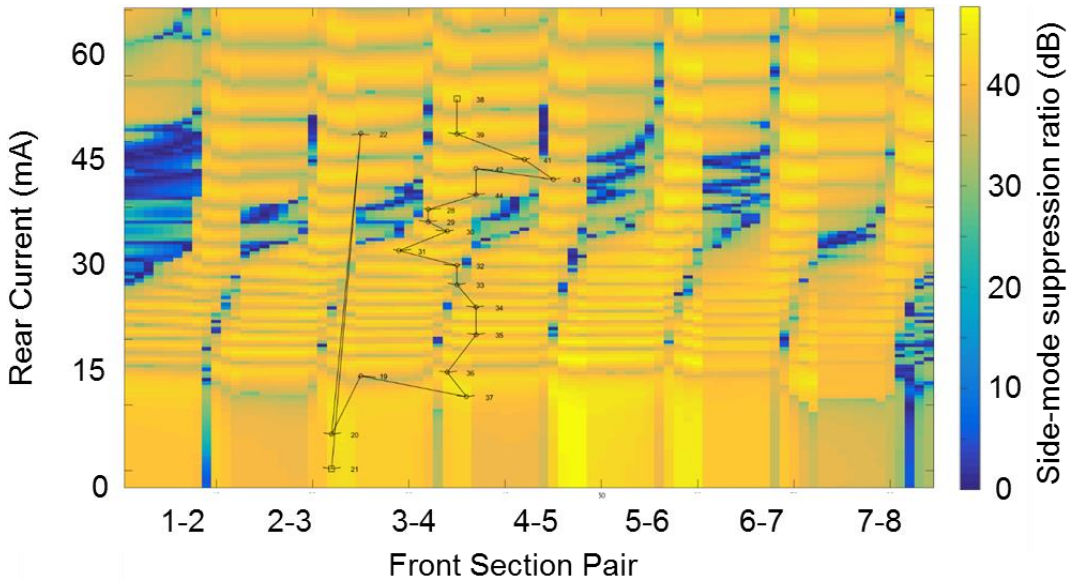

Figure 47: The intermediate wavelength channels reached during a wavelength switch between channel 21 and 38, overlaid on a map of side-mode suppression ratio against applied tuning currents.

The general trend for the switch from channel 21 to 38 is for the wavelengths to initially track a path following changes in front section pair selection and front section current, before a gradual increase in rear current. There is an outlier to this general trend for wavelength 22, however this is most likely an artefact of the plotting process. Wavelength 22 (and most other wavelengths) can be reached with several combinations of tuning currents, but the current combinations chosen using the method outlined in section 3.2.2 were selected specifically for their fast switching properties rather than by following a strict relationship between channel number and current. It is therefore more likely that channel 22 is reached at an alternative switching current combination with low rear current, than the steady state location plotted in Figure 47.

Given that the laser has been shown in Figure 47 to take more time to move through wavelengths on the rear current axis than on the front current axis, it follows that the rear current injection has a larger impact on laser tuning time than front or phase current injection. This is compounded by the rear current section requiring more current overall (a maximum of 58.5 mA was used in these experiments) than the front (4.8 mA) or phase (9.5 mA) sections.

To reduce the impact of the rear current injection time on the wavelength switching time, the DACs can create pulse shapes with pre-emphasis rather than step functions. The novel pre-emphasis model presented here, compared to prior work, is asymmetric depending on whether the rear current increases or decreases during the switch, to

attempt to increase the carrier injection (or removal) rate at the start of a wavelength switching event. For a switch requiring an increase in current in the rear section, by initially injecting more than the desired steady state current level during the first sample of the switch event, a faster current rise time can be achieved. The inverse pattern can be used for switches requiring decreases in the rear current, to draw current out from the diode junctions as quickly as possible. Three further DAC samples were used to counteract any overshoot or oscillations due to parasitic capacitive and inductive effects in both the laser packaging and the DAC/laser interfaces and circuitry.

The steady state rear section current required for the desired wavelength after the switch was defined as I_{SS}, and the change in steady state rear section current between the two wavelengths of the switch was defined as ΔI. Table 3 shows how the 4 samples to be played out by the rear DAC can be calculated if both I_{SS} and ΔI are known. This pattern can be applied to all laser switches to create a look-up table of 4 samples for every possible wavelength switch.

Table 3: Current pre-emphasis applied to laser rear section for fast tuning without mode-hops. The values in this table are the values written to the DACs, where the minimum value is 2050 (0 mA) and the maximum is 4095 (58.5 mA).

Decrease of ΔI, where steady state current is I_{SS}				
	Sample 1	Sample 2	Sample 3	Sample 4
ΔI > 400 and I_{SS} > 2900	$I_{SS} - \frac{\Delta I}{8}$	$I_{SS} - \frac{(1500 - \Delta I)}{50}$	$I_{SS} - \frac{(1500 - \Delta I)}{50}$	I_{SS}
All other cases	$I_{SS} - \Delta I$	$I_{SS} - \frac{(1500 - \Delta I)}{80}$	$I_{SS} - \frac{(1500 - \Delta I)}{80}$	I_{SS}
Increase of ΔI, where steady state current is I_{SS}				
	Sample 1	Sample 2	Sample 3	Sample 4
ΔI > 700 and I_{SS} > 2900	I_{SS}	I_{SS}	$I_{SS} - \frac{\Delta I}{15}$	I_{SS}
All other cases	$I_{SS} - \frac{2\Delta I}{3}$	I_{SS}	$I_{SS} - \frac{\Delta I}{15}$	I_{SS}

The lookup table values explicitly depend on both the absolute steady state rear current that is required after the switch event, and the difference between the steady state rear currents before and after the switch. The values presented here were found

using an iterative method of trials for a select number of wavelength switches, and further experiments to verify the switching times across all possible channel combinations are described in the next section.

3.2.5. Optimising laser tuning speed across all wavelength pair combinations

The current patterns to be produced by the DAC for each switching event, shown in Table 3, do not depend on the absolute wavelength that the laser is set to produce, and only on the rear currents. This means it is possible to assess the improvement in tuning time based upon applying the DAC current patterns for a subset of laser channel switches, and to then extrapolate the pattern across the full range of the DSDBR laser.

Figure 48 shows the switching time between all possible pairs of wavelength channels from 1 to 22, selected as a subset of the full laser tuning range including the full range of rear current from 0 to 58 mA (as section 3.2.3 identified that the rear current has most impact on tuning speed). There is a clear diagonal line through Figure 48 which corresponds to fast adjacent channel switches below 3 ns, which can be achieved with small (< 5 mA) rear current changes. All switches in this sample completed (measured by 90% rise-time of the laser intensity) within 28 ns. The yellow shaded regions were the slowest wavelength switches tested (> 20 ns), and correspond to switching from a high rear current (> 50 mA) to a low rear current (<5 mA).

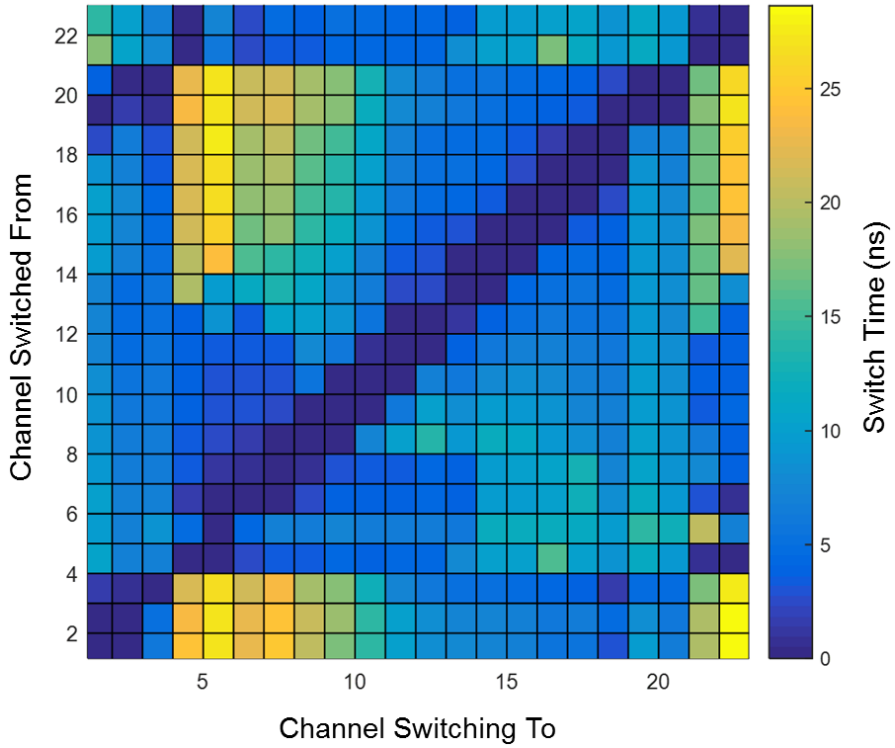

Figure 48: Switching time between any of a 22 channel subset of wavelengths for a single DSDBR laser.

Figure 48 shows that current pre-emphasis can be applied to the rear section currents such that switching between all possible pairs of wavelengths for channels 1-22 can be achieved within 23 ns. It has been asserted so far in this section that the pre-emphasis has no dependence on wavelength; this can be experimentally verified by applying the same pre-emphasis which successfully reduced the tuning time for channels 1-22 across all available channels. Due to equipment limitations with the tunable local oscillator DSDBR, 96 wavelength channels were available; this is greater than the total number of 50 GHz spaced channels than within the optical C-band alone, but less than the full 120 channels that a DSDBR laser could potentially reach as shown in section 3.2.1. A measurement of switching time for all 96x96 channel pair combinations was performed, and the resulting switching times are shown in Figure 49.

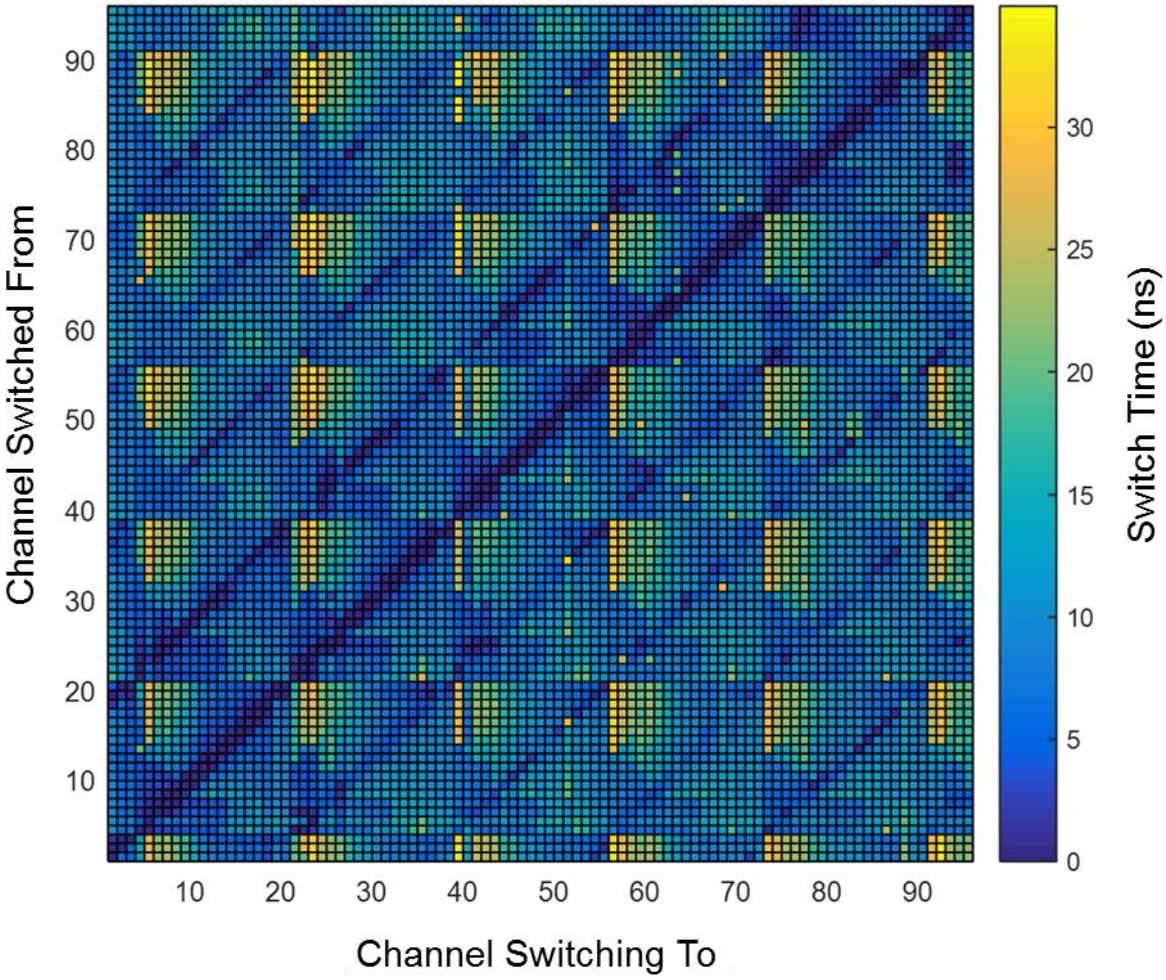

Figure 49: Switching time for all possible pair combinations of 96 available wavelength channels.

The clear periodicity of the switch times in Figure 49 matches the periodicity of the rear current i.e. increasing rear current from channel 21 to channel 38. The longest switch times are consistently observed for switches from high rear current to low rear current, for instance, the region around switches from channel 21 (y-axis) to channel 38 (x-axis), is yellow, corresponding to the longest switch times (> 25 ns). A summary of the

recorded switch times is shown in Figure 50, which presents a cumulative distribution function of the switch times across all 96x96 = 9216 possible wavelength switch pair combinations. Figure 50 shows that in 60% of the possible wavelength switches, the laser tunes to the new wavelength within 10 ns, and more than 90% of the possible wavelength switches are complete within 20 ns.

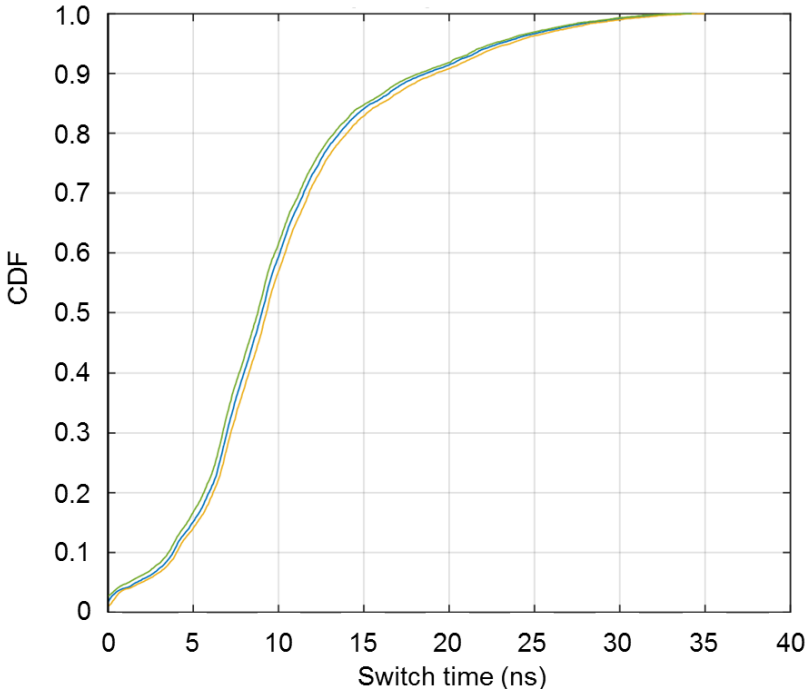

Figure 50: Cumulative density function (CDF) of the switch times between all possible pair combinations of 96 wavelength channels. Each switch was performed 5 times; the blue trace is the mean switch time from the 5 attempts while the green and yellow traces provide maximum and minimum bounds respectively.

The significance of Figure 50 is that a DSDBR laser has been shown to be capable of rapidly switching between all possible wavelength pairs across 96 channels, with a switching time always less than 35 ns. By using pre-calculated pre-emphasis on the laser rear section, the wavelength switch time of a DSDBR laser is now bounded at 35 ns for all possible wavelength switches, which is an 82.5% reduction on the previous bound of 200 ns shown in section 2.3.2. The most comparable prior work showed 64x64 channel switching within 5 ns [125]; 64x64 channel pairs corresponds to 44% of the available switches from 96x96 channel pairs, and Figure 50 shows that 44% of switches complete around 7 ns. The novelty of this work lies in extending the range of the laser fast tuning demonstration, and the asymmetric current pre-emphasis design used to achieve all fast tuning results.

The switching time bound of 35 ns can be compared back to Figure 38, where the required laser tuning times for a range of data rates was presented. A tuning time of 35 ns is sufficiently small to allow laser wavelength retuning between every packet at

25 Gbit/s. This in turn allows increased flexibility and reconfigurability of the star network design, supporting increased data throughput for varied traffic patterns if it could be integrated into future optical network implementations.

By reducing the laser wavelength tuning time, but retaining a constant Epoch duration, the data transmission time lost to tuning events could also be a shorter fractional overhead of the usable data transmission time. This in turn would increase the overall network throughput. An alternative route to increasing the total available pool of bandwidth is to increase the bitrate carried on each wavelength channel by using a higher complexity modulation format, such as PAM.

3.3. Higher bitrate using line coded PAM

Pulse amplitude modulation (PAM) signals are ideal for increasing the throughput across the WS-TDM star network design. PAM signals use multiple amplitude levels to provide an increase in the number of bits transmitted per symbol, which results in an increased data transfer rate per wavelength compared to binary OOK formats. The increase in data rate is achieved without increasing the symbol transmission rate or adding DSP to receivers, since PAM can be received using only intensity detection.

The WS-TDM star network requires the transmitted data signals to be spectrally shaped to remove data content below 1 GHz (see section 2.2.6).This was demonstrated using IBLC, which is an example of many spectral shaping codes for binary data. However, limited progress has been made to date in developing line codes which can be applied directly to PAM signals to achieve spectral shaping for low frequency suppression.

Prior work has explored the use of line codes for multilevel signal formats. In [134] the bipolar signal format derived from AMI presented in [135] is extended to multilevel signal formats. This encoding gives the benefit of DC balanced encoding and suppression of signal spectrum around DC, but the coding relies on both the transmitter and receiver being capable of producing more levels than in the original signal. For example, a PAMn signal (multilevel PAM with n distinct levels) is precoded before transmission as a $(2n-1)$ level signal, and a receiver must be able to distinguish between all $(2n-1)$ levels to correctly decode the data. Increasing transmitter resolution and receiver sensitivity to $(2n-1)$ levels from n levels would increase the required complexity and cost of the communications link, due to the increase in transmitter and receiver SNR required to differentiate between tightly spaced intensity levels.

In the WS-TDM star network, the optical receiver is a square-law intensity detector (after the squaring and summing operation of the coherent receiver outputs), only capable of detecting the power envelope of a received signal. However, when the electric field of an optical signal is modulated, it can take both positive and negative values. At the receiver, a square-law detector converts this to power; the electric field amplitude is squared and it is this value that is recorded by the slicer. For the multilevel PAM line code in [134], the levels $-q$ and q would both be received as $+q^2$ and transmitted information is lost. The receiver design used in the WS-TDM network design rules out the use of any coding schemes that need to distinguish between positive and negative values of signal amplitude.

The naïve application of line coding to a stream of binary data prior to PAM mapping does not necessarily give the same spectral properties to the mapped PAM signal as it does to the binary coded signal alone. The spectral properties of a PAM signal can also be affected by the binary to PAM mapping, and the number of signal levels (n) in the PAMn format.

Four line codes for PAMn are proposed and simulated in this section, that all meet the objectives from section 2.2.6 of producing a spectral null around DC. All four provide increased spectral efficiency compared to binary OOK; three of them are suitable for optical intensity reception of electric field modulation. The four line codes proposed here, in increasing complexity of implementation, are a block-based inverting code (LC1), an asymmetric block-based inverting code (LC2), an asymmetric block-based rotating code (LC3), and a running disparity symbol-wise monitoring code (LC4).

3.3.1.1. LC1: Block inverting code

This encoding is an extension of that presented in [114], which considered a method of encoding binary symbols for fixed block lengths b. By integrating the signal amplitudes over each block, the DC component of the block is found. Each block could then be inverted (swapping 0s for 1s and vice versa) if necessary to minimise the long-term accumulation of DC offset from all blocks. An additional bit must be added to each block to describe whether an inversion took place for that block.

When mapping binary data to PAMn symbols, the optimal mapping scheme in the presence of additive white Gaussian noise is a Gray coding scheme [136]. Gray coding schemes maintain only a single binary bit difference between any two adjacent symbols, minimising bit errors in the presence of noise. Gray coded symbols form a cyclic pattern; a binary bit pattern of all zeros need not be paired to the lowest amplitude level of a PAMn system but could correspond to any of the amplitude levels. It is, therefore, possible to create multiple mappings from binary data to Gray coded

PAMn symbols for any given block of b binary bits. For any block of binary bits, up to n independent PAMn symbol mappings are possible using Gray coding, and the LC1 code uses that property to choose between two “inverted” mappings.

The full coding mechanism is outlined in Figure 51 (using PAM4 as an example) and detailed here. An incoming stream of binary data is split into blocks of b bits; $\log_2 n$ blocks are taken from this binary bit stream in parallel and passed through two independent symbol mappers, which also calculate the disparity (the maximum possible number of consecutive identical symbols) of each mapping. Table 4 shows an example of two possible mapping sets for PAM4 (similar patterns can be designed for higher order PAMn formats). The polarity of the symbols in one mapping is the opposite of the polarity in the other mapping for each block of binary bits. This is the “inversion” of the code, whereby it is possible to invert the polarity of each symbol as it is coded (albeit with a different magnitude).

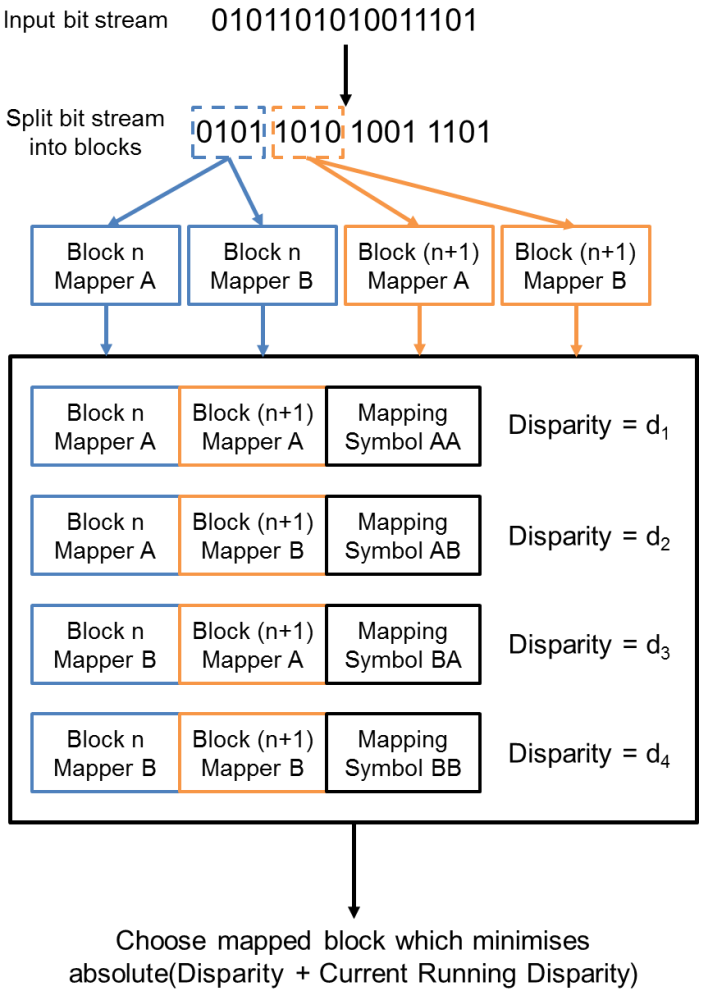

Figure 51: The symbol mapping for binary data, for LC1 using PAM4 coding.

Without any additional markers to denote the mapping, it would be impossible to know which mapping (A or B) had been used for any given block. Thus it would not be possible for the receiver to recover the original binary data, as two parallel de-mappers

would be required but the receiver could not possibly know which one recovered the data as transmitted.

Table 4: An example of the two possible PAM4 mappings in LC1.

Binary Bit Pattern	PAM4 Mapping A	PAM4 Mapping B
00	-3	1
01	-1	3
10	3	-1
11	1	-3

Given that PAMn can carry n binary bits per symbol, a single PAMn marker inserted into the data stream can denote which of the two mappings was used for up to $\log_2 n$ blocks. Therefore after every $\log_2 n$ blocks, a single additional symbol is inserted which denotes which mapping was used in the preceding $\log_2 n$ blocks. For example, for PAM4, appending mapping symbol “-1” may represent that mappings “AB” were used in the preceding two blocks. The symbols representing the mappings are predetermined, and are consistent for all blocks at both the transmitter and receiver.

The total disparity of the $\log_2 n$ blocks and mapping symbol is found for all mapping combination, and compared to the running disparity (cumulative sum of all previous symbols in the transmission) of the data stream. The combination of mapped blocks (including the mapping symbol) is selected which would result in minimising the absolute value of the running disparity.

The number of binary bits transmitted per block (b) has no relation to the size of the PAMn symbol space (n), and can be chosen based on several factors: the spectral properties produced by the choice of b; the overhead imposed by the insertion of the mapping flip symbols every $\mathrm{b}\log_2 n$ bits; and the ease of parallel implementation of the code e.g. the transmitter’s capacity to calculate the disparity of all block/mapper combinations simultaneously. These concepts are explored further in section 3.3.2

However, this code using symmetric symbol levels could not be implemented in the WS-TDM network proposed in chapter 2, as there is no way for a square-law receiver to distinguish between signal levels of the same amplitude but opposite polarity. For instance, consider symbols transmitted from the set $[-3\ -1\ 1\ 3]$; a square law receiver would square the amplitude signals to produce the received signal set $[9\ 1\ 1\ 9]$. Information is irrecoverably lost in the squaring operation which makes symbols of the same magnitude appear identical, regardless of their polarity. This code thus provides

a convenient model for introducing block inversion PAMn coding, but is of no practical use for the WS-TDM system.

3.3.1.2. LC2: Asymmetric block inverting code

To overcome the issue with the symmetric symbol set in LC1, where symbols of the same amplitude but opposite polarity appear identical at a square law receiver, a further line code is proposed. LC2 uses an asymmetric PAM symbol set, so that when the symbol amplitudes are squared, each symbol retains a unique value. In addition, for application to the WS-TDM star network system, it is desirable to use a PAM symbol set which includes a zero level at the transmitter. This would simplify the control interface for an optical modulator to extinguish optical power when a transmitter did not have transmission rights; a transmitter could simply be instructed to transmit a continuous run of zeros.

An example of a mapping for PAM4 transmission meeting the above constraints would be the levels $[-3, 0, 1, 2]$. For statistical DC balance of this symbol set it would be necessary to transmit as many "-3" symbols as "1" and "2" symbols combined. However, any code designed to achieve this would reduce the information capacity of the link compared to uncoded transmission. Instead, bounded DC balance can be achieved by using the mapping inversion technique to minimise running disparity, in exactly the same manner as described for LC1 in section 3.3.1.1. As with LC1, blocks of binary bits are compiled with a symbol identifying the mapping used, and the mappings are chosen to minimise the disparity within the block and mapping symbol combination. An example pair of PAM4 mappings for LC2 is shown in Table 5.

Table 5: An example of the two possible PAM4 mappings in LC2: asymmetric block inverting line code.

Binary Bit Pattern	PAM4 Mapping A	PAM4 Mapping B
00	0	2
01	1	-3
10	-3	1
11	2	0

If the absolute value of the signal magnitude is taken at the receiver, a direct translation into evenly spaced symbols is easily achieved. This is ideal for simple receiver decision threshold structures with noise affecting all symbols equally. However, if the square of the signal magnitude is measured at the receiver, the received symbols are not evenly spaced (e.g. the equally spaced transmitter symbol set $[0\ 1\ 2 - 3]$ would be received

as [0 1 4 9]). Either the transmitter could be adapted to send unevenly spaced levels (i.e. the square root of the PAM levels shown in Table 5), or a receiver could be constructed which maintains performance for uneven level spacing. Both of these methods impose stricter signal-to-noise ratio requirements on either the transmitter or the receiver (but not both) than would be expected for a standard PAMn system.

3.3.1.3. LC3: Asymmetric block rotating code

Both codes LC1 and LC2 are based on the principle of "inverting" the mapping of blocks of symbols, so that the chosen mapping would reduce the running disparity of the code stream. This principle was developed from the binary coding in [114], in which only two possible inversions exist for any given set of bits. However, when symbols are Gray coded to PAMn, it is possible to "rotate" between n possible mappings, each of which would produce a different disparity over a coded block. This method forms the basis of LC3.

A possible set of symbol mappings, retaining the symbol asymmetry outlined for LC2 above, is shown in Table 6. The coding operation proceeds as outlined above for LC1 and as shown in Figure 51, by assembling blocks of binary bits with a mapping symbol, and choosing the mapping which minimises the disparity of the data stream..

Table 6: An example of the four possible PAM4 mappings in LC3: asymmetric block rotating line code.

Binary Bit Pattern	PAM4 Mapping A	PAM4 Mapping B	PAM4 Mapping C	PAM4 Mapping D
00	0	2	1	-3
01	1	-3	2	0
10	-3	1	0	2
11	2	0	-3	1

There is one difference in the symbol coding process between LC1 and LC3: the insertion of symbols to identify the mapping used in each block. In LC3, there are n possible mappings, so a single mapping symbol can now only identify the mapping used within one block. This means that the overhead due to mapping symbol insertion for LC3 is $\frac{n}{2}$ times larger than the overhead for LC1 and LC2.

3.3.1.4. LC4: Symmetric running disparity monitoring code

In LC2 and LC3 above, the method of measuring at the receiver either the absolute magnitude or square of a transmitted symbol was described. Positive and negative symbols can be transmitted over the channel, before all symbols are converted to an

absolute positive value at the receiver. By alternating positive and negative values appropriately at the transmitter, the overall DC component of the signal can be reduced.

For the final line code studied in this work, LC4, a transmitter is used which can transmit all symbol levels in the set -$(n-1)$ ... 0 ... $(n-1)$ i.e. the total number of distinct symbols in the transmitted code is $2n-1$. A receiver is used which rectifies the received signal to its absolute value, and a conventional n level PAMn signal can be recovered. Given that an uncoded PAMn transmitter would only require n levels, this requires additional complexity (i.e. an additional DAC resolution bit) than would be required for uncoded transmission. However, this line code requires no overhead since there is no need to insert symbols to communicate which mapping was used per block at the receiver.

Each incoming n binary bits of the data stream are Grey coded by two independent symbol mappers, each mapping to the same amplitude of PAM symbol, but with one mapping positive and the other mapping negative. An example for PAM4 is shown in Table 7. The mapping is then chosen that would minimize the magnitude of the running disparity of the whole signal stream.

Table 7: An example of the two possible PAM4 mappings in LC4: symmetric running disparity monitoring line code.

Binary Bit Pattern	PAM4 Mapping A	PAM4 Mapping B
00	0	0
01	1	-1
10	3	-3
11	2	-2

There is no need for the transmitter to communicate to a square law receiver which mapping was used, since the binary bit pattern always maps to the same magnitude of PAMn symbol (polarity is ignored by a square law receiver). This results in no coding overhead in LC4 due to the insertion of mapping symbols, and no need to consider a block-wise approach to coding.

3.3.2. Performance of the proposed line codes

3.3.2.1. Symbol transition frequency

Serial communications systems require frequent transitions between symbols, so that clock recovery can be performed at the receiver. This process aligns the receiver

sampling frequency with the transmitted clock rate, so that received data is sampled at the optimum symbol phase. It is, therefore, important to consider the number of adjacent identical symbols that are permitted in each of these codes, as a long run of identical symbols precludes any symbol transitions from taking place.

For LC1, there are no limits directly imposed on the number of identical adjacent symbols within each block. However, if multiple consecutive blocks were to contain the same identical symbols, block inversion would be necessary to ensure that there is no continuous build-up of running disparity. This limits the maximum run length without a transition to the length of a single block.

To discuss this by example, again considering PAM4, the highest disparity that a single block of $\frac{b}{2}$ symbols could introduce is $\pm\frac{3b}{2}$. A subsequent block of b identical binary bits could not be mapped to the same symbols, as this would increase running disparity rather than reduce it. Since both mappings from binary to PAM symbols in LC1 are always of opposing polarity, it is always possible to select a mapping that works to undo the contribution to disparity made by the previous block. It therefore follows that there must be a transition at least every block, and thus that the maximum distance between transitions is $\frac{b}{\log_2 n}$ symbols.

For LC2 and LC3, a similar argument can be followed to LC1, to find that again the maximum number of symbols without a transition is $\frac{b}{\log_2 n}$. However, an exceptional case for these codes is a block containing only a continuous run of "0" symbols, which could be followed by a further block containing only "0" symbols should the running disparity before the two blocks be ≥ 0. To attempt to introduce a transition in the coding layer, an exception could be triggered which causes a block which would cause successive blocks of all "0" symbols to be coded using the "inverted" mapping, even though that works against the primary aim of reducing the running disparity. This is not a viable solution, as this could theoretically lead to continuously increasing disparity due to the asymmetry of the symbol set when a "0" symbol is included. For example, using LC2 or LC3 for PAM4 encoding, it could be possible to repeatedly only choose symbols from the set $[0\ 2]$, which would not allow the disparity to return below zero. To use LC2 or LC3 therefore requires limits to the symbol run-length to be placed at a higher layer of the transmission protocol. This could be through statistical means via scrambling, or by imposing limits on the packet structure and assembly at higher levels of the network stack to remove the possibility of adjacent symbols.

For LC4, there is generally a transition between every symbol. Even if two consecutive sets of n binary bits entering the two mappers are identical, the second mapped

symbol would be allocated the same magnitude but opposite polarity from the first, to continuously minimise the running disparity. However, as for LC2 and LC3, issues arise for long streams of "0" symbols in the data stream. Without any limits imposed by higher layers, there is nothing to prevent a continuous run of zeros, especially as this code does not require markers to be sent signifying mapping inversions (or the lack thereof) for each block. This code therefore also requires scrambling or run-length limits to be imposed higher up in the network stack.

The maximum distance between symbol transitions for each of the four codes, as calculated and discussed in the previous four paragraphs, is summarised in Table 8.

Table 8: A comparison of the maximum distance between transitions for each of the three proposed line codes.

Line code	Maximum Distance Between Transitions
LC1: Symmetric block inverting	$\frac{b}{\log_2 n}$
LC2 and LC3: Asymmetric block inverting/rotating	$\frac{b}{\log_2 n}$ (except whole blocks of zeros)
LC4: Asymmetric running disparity monitoring	1 symbol (except runs of zeros)

3.3.2.2. Coding overhead

The coding overhead is dependent on both the block length chosen and the number of symbols in the PAM symbol set for line codes LC1, LC2 and LC3. In a PAMn modulation format, each symbol represents $\log_2(n)$ binary bits. For LC1 and LC2, one binary bit is required per block of b binary bits to denote the mapping used within that block. This is regardless of whether the information is encoded as a definitive statement of the mapping used within the block, or a flag to denote whether the mapping changed in each block compared to the previous block. It follows that for LC1 and LC2 a single symbol must be added to the data stream every $\log_2(n)$ blocks to inform the receiver of the mapping used by the transmitter in the prior $\log_2(n)$ blocks. This results in a percentage overhead imposed on the input data rate of $\frac{1}{b}$. For LC3, one symbol is required per block to denote the mapping used within that block, and

thus the imposed overhead is $\frac{\log(n)}{b}$. LC4 incurs no overhead, since no symbols need to be inserted into the output symbol stream to identify mappings. By inserting values for block length b and PAM amplitude levels n into these equations, the coding overheads can be calculated for a range of scenarios, to determine which codes may be suitable for implementation in the WS-TDM star network.

Figure 52 shows the variation of coding overhead with the block length b for LC1, LC2 and LC3 encoded PAM4 and PAM8.

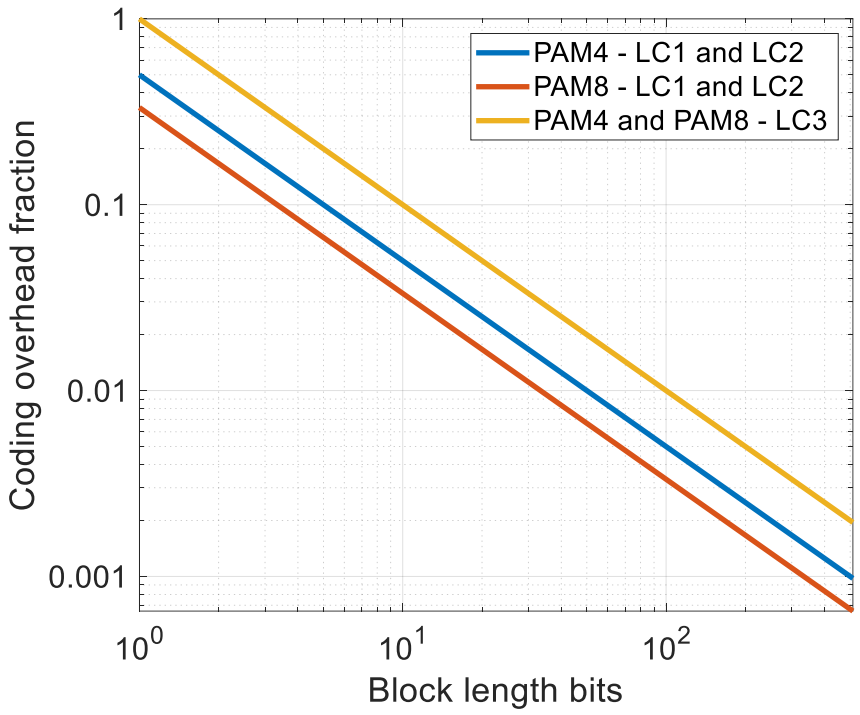

Figure 52: The fractional coding overhead introduced by LC1, LC2 and LC3, as the block size and modulation format are varied.

The coding overhead is identical for LC1 and LC2, and for PAM4 and PAM8 using LC3. The coding overhead discussed in this section does not include any additional overhead that may be incurred by higher layers in removing the possibility of long zero run lengths in LC2, LC3 and LC4. The maximum overhead calculated was 50% for a block length of $b = 2$ bits, using LC3. For an overhead of 1% (a limit chosen to minimise the impact of coding on the overall throughput of the WS-TDM star network), the block sizes required are 33 (for LC1 and LC2 PAM8), 50 (for LC1 and LC2 PAM4) and 100 (for LC3). Larger block sizes reduce the overhead, but also impact on the low frequency suppression of the code, as discussed in the next section.

3.3.2.3. Low frequency suppression

To assess the suitability of the four codes presented here for the WS-TDM star network and other applications described in section 2.2.6, an important characteristic is the suppression of signal power at low frequencies. Specifically for the WS-TDM star

network, the suppression of frequencies below ± 1 GHz is required to use TDM with multiple transmitters at the same wavelength, as outlined in section 2.2.6.

A simulation was performed for a stream of 2^{21} random bits of binary data, encoded with the four different line codes outlined above, using a PAM4 symbol set. The results obtained here with PAM4 are representative of similar designs for higher order PAMn formats (where n is any power of 2), without loss of generality. For LC1, LC2 and LC3, the block size of each code was varied to be a power of 2 between 2 and 512 bits. The coded data rate was simulated as 25 GBaud, and the power spectral density of each signal was calculated. For comparison, IBLC at 25 Gbit/s was also simulated, as applied to the same input binary data stream.

Figure 53 shows the simulated relative signal power of the four proposed line codes and IBLC, at 1 GHz. Power spectral density was normalised so that the peak of the power spectral density was aligned at 0 dB, for ease of comparison. Figure 53 shows that LC4 has the highest suppression of low frequency data at 1 GHz (-13.19 dB relative to the spectral peak), even more than IBLC (-10.96 dB). The best suppression for LC1 and LC2 is for a block length of 2 bits, with suppression of -7.69 dB and -6.82 dB respectively compared to the peak power. The best suppression for LC3 is for a block length of 4 bits, at -6.28 dB compared to peak power.

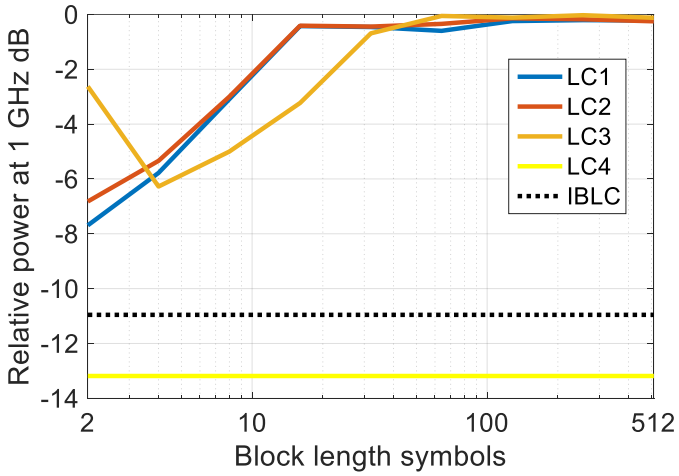

Figure 53: Comparison of the relative power at 1 GHz of the proposed line codes with IBLC as coding block length is varied.

There is a general trend across LC1-LC3 for higher power at 1 GHz with increased block length, with the exception of LC3. Given that in LC3 a mapping symbol must be attached to every block, using a binary block length of 2 bits (= 1 PAM4 symbol) results in each block consisting of a data symbol and a mapping symbol. Although the impact of the mapping symbol is included in the coding method, the assignment of mapping symbols to mappings is arbitrary and predetermined. Therefore, the impact of a 1:1 ratio of mapping symbols to data symbols actually reduces the performance of the

code. For longer block lengths, LC3 has better suppression of power at 1 GHz than both LC1 and LC2, as would be expected given the greater range of mappings available to reduce the running disparity at each block.

To further explore the low frequency power suppression of the line codes, the full power spectral densities of the simulated signals are shown in Figure 54 for block sizes of 2, 8 and 128 bits. The LC4 line code (which has no blocks) shows a low frequency roll-off within 2 dB of IBLC below 3 GHz, while for all block sizes tested, the LC1 and LC2 code show a low-frequency roll-off between 0.5 dB of each other below 3 GHz.

For a block size of 128, as shown in Figure 54e and f, the low-frequency roll-off only begins at 220 MHz for LC1 and LC2, and 350 MHz for LC3. This is likely due to the high suppression of the DC component through mapping inversions/rotations, but no removal of long streams of adjacent symbols. At long block sizes such as 128 bits, a full block of identical symbols (including a mapping symbol identical to the data symbols) could result in frequency components all the way down to 385 MHz, which is in agreement with the simulations. This further emphasises the need for higher layer scrambling or run-length limits, to remove the possibility of long runs of identical symbols.

The implications of the simulations for the WS-TDM system are that LC4 is the most optimal line code (of those studied) to use with PAM transmissions and thereby increase the total network throughput. This conclusion is reached due to the lack of overhead using the line code, and better suppression of signal power at 1 GHz than the IBLC code used on binary data. The drawbacks to implementing LC4 are the increased transmitter complexity required to transmit additional symbol levels compared to the other line codes studied. However, given that LC1 cannot be used in the WS-TDM system, and LC2/LC3 do not have low frequency suppression comparable to IBLC for any block length, there are no other choices. Having chosen LC4 as the optimum solution, the next section describes attempts to implement LC4 into a practical demonstration of the WS-TDM network.

3.3.3. RIN Modelling

The properties of the LC4 line code discussed in the previous section show promise for increasing the overall throughput across the WS-TDM design by increasing the bitrate per wavelength. To transmit any PAM format the transmitter would still only require an amplitude modulator, and DSP-free coherent reception can provide the intensity information required to demodulate PAM signals. Given that better low frequency suppression performance can be obtained using LC4 than with the IBLC OOK, the high-pass filter requirements of the WS-TDM system should be easily met.

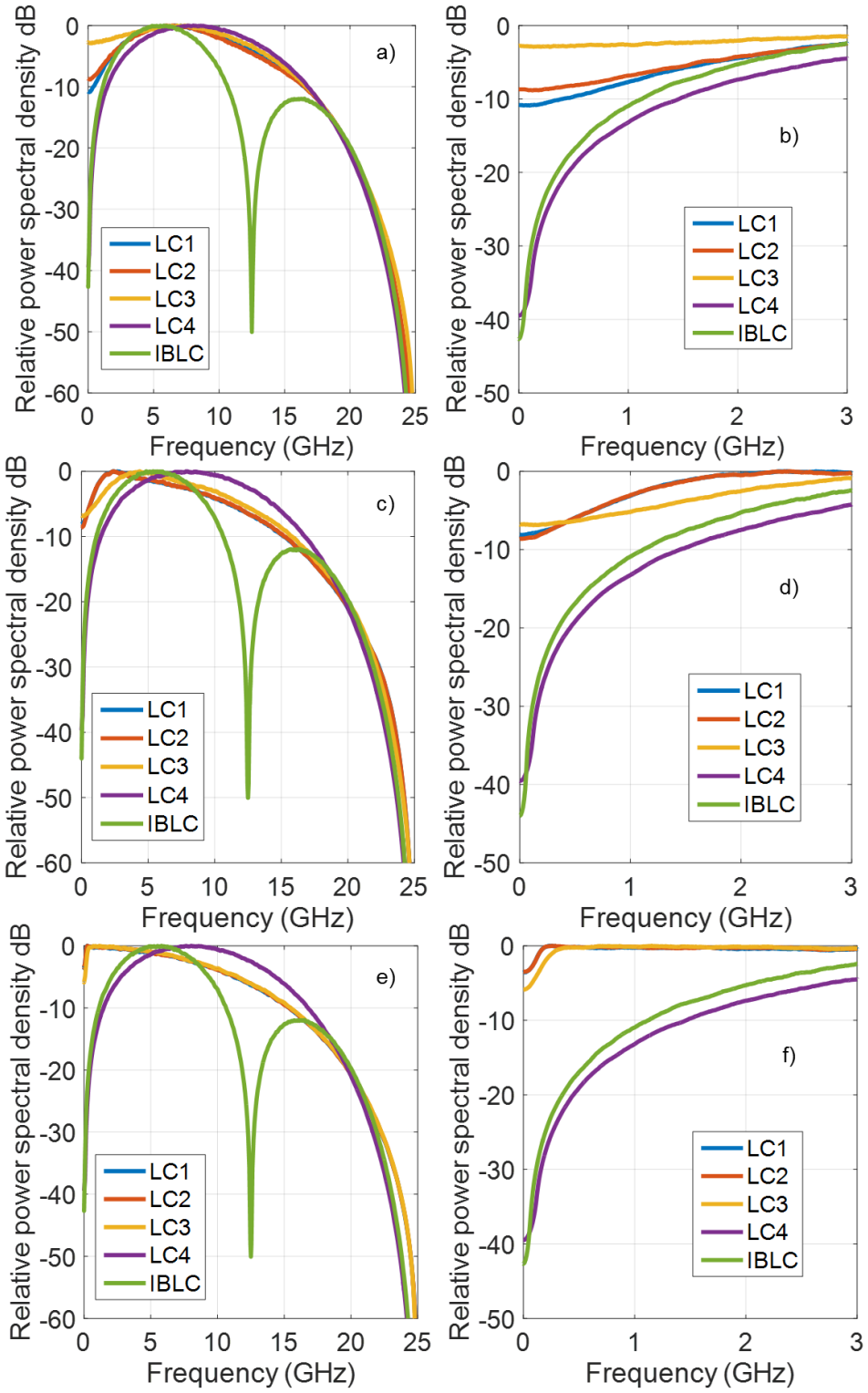

Figure 54a): Power spectral density for 2 bit block size; b): Detail of a) from 0-3 GHz; c): Power spectral density for 8 bit block size; d): Detail of c) from 0-3 GHz; e): Power spectral density for 128 bit block size; f): Detail of f) from 0-3 GHz.

A BER power sensitivity experiment was performed using the same experimental setup as in Figure 21, with the exception of the 10 GS/s Arb. Wave Gen., which was replaced with a 50 GS/s Arb. Wave Gen. to create the PAM4 electrical drive waveforms for the Mach-Zehnder modulator. The data waveform was a 22 GBaud PAM4 signal, encoded

from raw binary data using the LC4 method described in the previous section. The variable attenuator following the transmitter was set to provide attenuation such that the received optical signal power was -14 dBm. This power level was chosen based on the error-free optical power level for 25 Gbit/s IBLC of -24 dBm, with an additional 10 dB optical signal power to accommodate the theoretically required 9.5 dB increase in SNR when moving from an OOK to a PAM4 signal format.

The received data was captured from the oscilloscope and processed offline, using the same signal workflow as in Figure 15. The decision circuit at the end of the signal flow was adapted to be capable of 4 level decisions using three decision thresholds, and Q factor was calculated from statistical measurements and swept decision thresholding following the method in [137]. An eye diagram of the received signal is shown in Figure 55, and there is no clear eye opening at any sampling instant. The BER calculated from the measured Q factor was approximately 0.16. Given that this experimental result was beyond the correction capability of common error correction codes [138] the experimental investigation was not continued further.

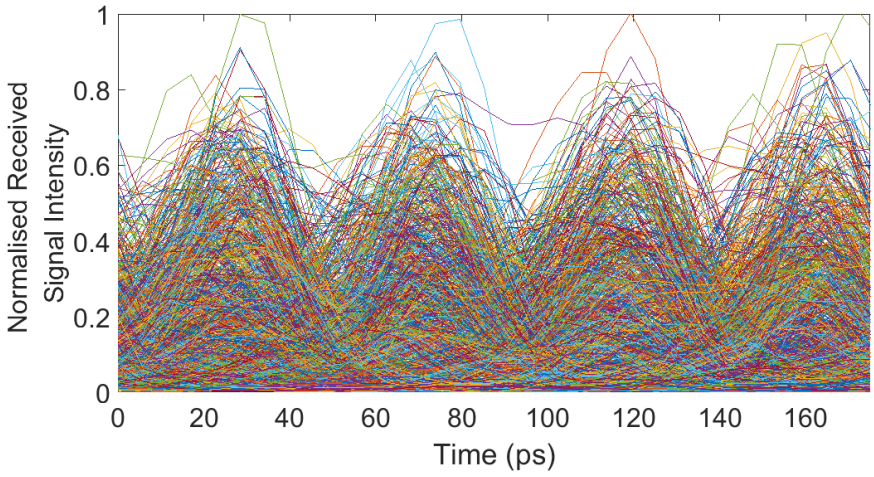

Figure 55: Eye diagram of 22 GBaud line coded PAM4 experimentally received using the DSP-free coherent receiver from the WS-TDM system in chapter 2.

In Figure 20c (which showed the eye diagram for 25 Gbit/s OOK at the same receiver), the amplitude of “1” symbols had a high variance, due to the squaring of the electrical signals produced by the coherent receiver squaring the received optical noise as well as the signal. The squaring operation has also caused all of the PAM4 symbols to spread in amplitude so that there are no clear decision thresholds between them. The noise source which contributed most significantly to the poor BER performance of the system was most likely the laser relative intensity noise (RIN), which is the variation of the laser intensity over time due to the spontaneous emission of photons. DSDBR lasers have poor RIN characteristics; measurements of the RIN spectrum of DSDBR lasers in [139] show a RIN of -140 dB/Hz or higher for bandwidths up to 200 MHz, and

between -140 and -160 dB/Hz for bandwidths greater than 200 MHz. This is due to noise on all of the electrical contacts (all 12 tuning diode sections) translating into RIN on the laser optical output. In [140], simulation results showed that RIN imposed at least a 3dB penalty on PAM4 reception for a laser RIN of -135 dB/Hz. It was further proposed in [141] that a single RIN value would not fully specify a multilevel PAM system in the same way as it would a binary OOK system, so the RIN performance of DSDBR lasers when transmitting PAM modulated data may be even worse than the value quoted in [139].

A simulation was performed of the experimental setup of Figure 21, with the exception of an arbitrary waveform generator simulated to modulate the laser with data (instead of a 10 Gbit/s pulse pattern generator). The simulation transmitted data using 50 GBaud PAM4 encoded with the LC4 line code outlined in this chapter, for an ideal laser without RIN range of RIN levels between -170 and -130 dB/Hz. The simulation was also performed without any RIN, to verify the error floor without the effects of RIN. In the simulation, the same RIN was assumed on both signal and LO lasers, although the contribution from the LO dominates given the low signal power at the receiver. The receiver BER sensitivity to received power for various RIN levels is plotted in Figure 56.

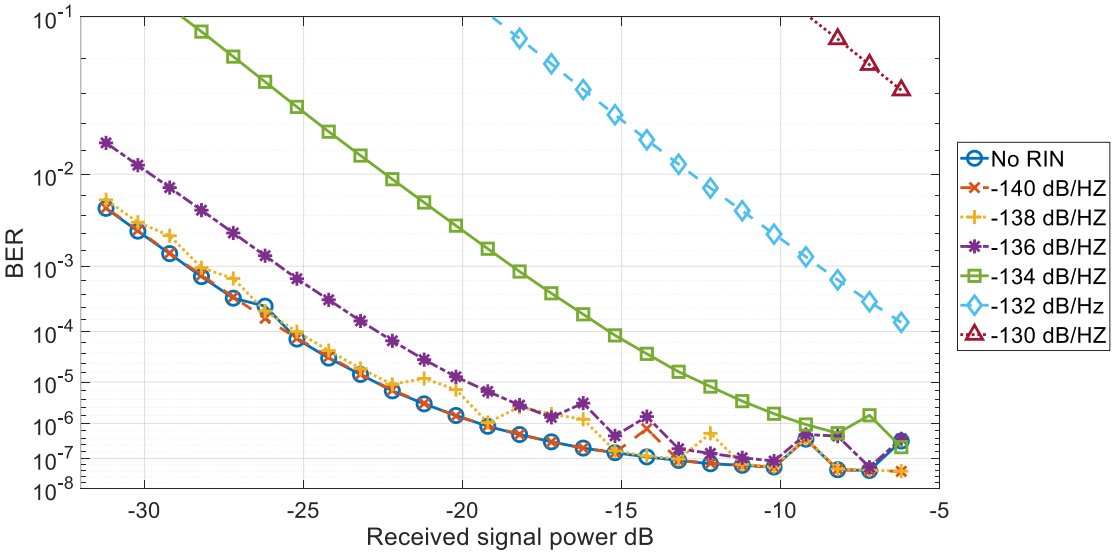

Figure 56: Simulated receiver BER sensitivity to received power for varying RIN levels on both signal and LO lasers.

The simulation was performed for a selection of RIN values spaced by 2 dB/Hz between -170 dB/Hz to -130 dB/Hz. However, for all RIN levels between -170 dB/Hz and -140 dB/Hz, the BER performance of the system was identical to the performance simulation without RIN, implying bit error performance was limited by receiver sensitivity rather than RIN. For clarity of the figure, only RIN levels of -140 dB/Hz or higher are plotted on Figure 56.

In [139], the RIN power spectral density of a DSDBR laser was measured to be above -130 dB/Hz for frequencies up to 10^8 Hz, and in the region of -160 to -135 dB/Hz for frequencies above 10^8 Hz. The simulation results in Figure 56 show that error-free transmission will not be feasible without the application of an error correction code. Forward error correction (FEC) codes can correct data with bit error rates of the order of $4x10^{-3}$ to have error rates of the order of 10^{-15}. Considering the simulation results at a BER of $4x10^{-3}$, there is a sensitivity penalty of 2.9 dB for RIN of -136 dB/Hz compared to no RIN, increasing to a sensitivity penalty of 18.8 dB for RIN of -132 dB/Hz compared to no RIN.

The non-linear increase in sensitivity penalty for increasing values of RIN can be explained by the receive signal flow of the WS-TDM receivers (see Figure 15 in section 2.2.3). To convert the electric field signal components produced by the coherent receiver back to optical intensity, a squaring operation takes place. Assuming a RIN model of additive Gaussian white noise to the electric field amplitude, the RIN power will also be squared.

In [139], a method for removing the impact of LO RIN was proposed, using a linear FIR high pass filter. The experimental demonstration in [139] was implemented to reach BER values in excess of 10^{-3}, where FEC is necessary to recover error-free data; error floors were seen for lower BERs where the RIN cannot be completely removed. However, the effect of the PAM line coding presented in this work is already to enable a high pass filter to be used to remove interfering transmitters at the same wavelength. A high pass filter was already integrated in the signal path for the experiments in thesis, with a larger roll-off at a higher frequency than that in [139], and no improvement in system performance was observed due to this filtering. This concurs with the findings of [139], which observed an error floor to receiver sensitivity attributed to the incomplete removal of RIN – as seen in Figure 56.

For all of the RIN levels simulated, it is clear from Figure 56 that a BER of 10^{-12} corresponding to error-free data cannot be reached using PAM4. Given the results of this simulation, and despite the excellent properties of LC4 for the application, it is therefore not suggested that PAM signals are used with the WS-TDM optical network design, and alternative methods of increasing the total system throughput should be sought.

3.4. Conclusions

This chapter has outlined the major limits to the performance of the WS-TDM network, including the restriction in total bandwidth across an optical star coupler due to the

finite number of available wavelengths, and the data transmission overhead and lack of packet-by-packet switching due to the laser tuning time.

By using pre-emphasis in the laser diode current drivers, 35 ns switching between all pairs of 96 wavelength channels (at 50 GHz spacing) was demonstrated using a single laser. In the nanosecond tuning regime of a DSDBR laser, the most important factor was shown to be the rear tuning current and ensuring that longitudinal mode hops are avoided. This can be achieved by selecting tuning currents for the target wavelengths which are far from longitudinal mode boundaries, in contrast to established slow tuning techniques. Creating lookup tables of tuning currents based on fast tuning properties alone allows for greater wavelength stability over nanosecond timescales.

To increase the bit rate per wavelength, which in turn increases the total network throughput, the modulation format of the WS-TDM network can be changed from binary OOK to a PAM format which encodes multiple data bits per symbol. However, in chapter 2, line coding (IBLC) was required to suppress low frequency signal components, which in turn permitted up to 26 transmitters to share a wavelength using TDM. Similar line codes are required for higher order PAM formats, and this chapter presented candidate line codes which match or beat the low frequency suppression of IBLC e.g. more than 13 dB suppression of power at 1 GHz for LC4, compared to 10 dB suppression at 1 GHz for IBLC. Although experimental work to optically transmit line coded PAM4 formats did not show success due to laser RIN and the noise enhancement of the squaring operation in the DSP-free coherent receiver, the work in this chapter can inform future implementations of low frequency suppressed PAM transmission.

The physical layer improvements presented in this chapter can be integrated with the WS-TDM network design of chapter 2 to increase the total throughput. A further method to increase the overall available bandwidth is to exploit traffic locality and split the single star coupler into multiple physically separated sub-stars. By forming sub-stars, each sub-star can independently allocate wavelengths and timeslots from the full range. Details of sub-star construction and operation are described in chapters 4 and 5.

4. Reconfigurable star networks by sub-star construction

In section 3.1.2, the limit to network throughput across the WS-TDM star system was stated – an optical star coupler has a finite total capacity, limited by the total number of wavelengths that can exist simultaneously in the network core without interference. To overcome the throughput limit, it is possible to partition the network into independent pools of transmission capacity. This chapter explores a method of separating a single star into multiple sub-stars, while still maintaining the attractive properties of star couplers for media networks (single hop latency, flexible multicast and in-cast support, and the ability to connect all network nodes into a single multicast group).

This chapter introduces criteria for sub-star designs to meet the needs of the applications running in the data centre, before reviewing four possible topologies which at least partially meet the criteria. Three data centre traffic scenarios are then defined (random traffic, hotspot traffic, and zonal media production traffic), so that flow level simulations can be carried out across each topology for a range of network traffic loads. This allows quantification of the increased throughput per network node for the reconfigurable topologies presented here compared to a single large passive star.

4.1. Motivation and design criteria

The WS-TDM star network design introduced in chapter 2 and enhanced in chapter 3 is an ideal architecture for data centres with predominantly multicast traffic, such as live media production data centres. There are two key design criteria for media data centres, which will be used to assess the suitability of network designs:

1. It is required for one input data flow to be multicast from a single transmitter to an arbitrarily sized group of receivers on the network, from 2 receivers up to all N receivers simultaneously (where N is at least 1000). For instance, a video flow of static colour bars or a visual interference pattern can be sent to multiple monitors to ensure that colour grading and contrast settings are matched. This presents the first network design constraint: any transmitter must be capable of multicasting to any number of receivers, including all receivers simultaneously.
2. When signal flows already exist across the network between transmitters and receivers, it is required that these flows are not interrupted by network reconfiguration, as this would result in video or audio break-up. This is especially important in an all-optical star coupler network with no central buffering, as there are no network locations where packets can be temporarily stored while waiting for the network to reconfigure. This presents the second

constraint: “truly non-blocking” optical circuit switching [142], where any optical switching activity must not disrupt existing circuits through a network.

A solution which meets both of these design constraints is the WS-TDM star network of Chapter 2, where a single passive star coupler forms the network core. The broadcast-and-select nature of a star coupler always permits multicasting to arbitrarily sized groups of receivers, including all receivers being part of the same multicast group. Wavelength switching occurs only at the transmitters and receivers, with no active switching elements in the network core, so network reconfiguration is always non-blocking. The design thus meets both of the design criteria, but the total throughput of a single optical star is limited by the number of wavelengths that can be simultaneously supported through the central core, as discussed in section 3.1.2.

It is proposed here to create star networks on demand using optical circuit switches (OCSs). Instead of using just one large passive optical coupler capable of connecting all nodes in the system via a single star, smaller passive optical couplers can be joined via an OCS to form multiple sub-stars. Sub-stars are physically partitioned from each other, and are formed as required to meet the connectivity demands at any given instant. Each sub-star would independently have the same total throughput capacity as a single large star, thus linearly increasing the available capacity with the number of sub-stars formed. The abstraction of each sub-star is a single passive star directly connecting all nodes on that sub-star by a single hop, which minimises network latency.

The OCS connectivity must be reconfigurable to add and remove passive couplers from each sub-star, permitting connectivity and optimising the resource usage of the network as traffic demands change. Sub-stars must be designed such that they could all be instantaneously joined to provide the topology of a single star connecting all nodes; enforcing this requirement meets design constraint 1. If the creation, growth, shrinking and joining of sub-stars can be performed without blocking any existing flows across the network, then design constraint 2 is also met.

This chapter presents four possible designs for constructing sub-star networks from smaller couplers, aiming to meet the criteria outlined above: centred stars; rings; meshes; and centred meshes.

4.2. Network design topologies

4.2.1. Design and modelling assumptions

All of the network designs analysed in this section assume the following physical construction: a set of N nodes, each with a single fibre coupled transmitter and

receiver, are initially not connected to any sub-star. Each transmitter and receiver is attached to an OCS (for example, a MEMS switch with lenses or mirrors moved in 3D by mechanical actuators, such as those described in [143]–[145]). The OCS also has connections to multiple passive optical couplers and wavelength filters, which can be used to construct star networks. This design is shown in Figure 57. For the purposes of this architectural design comparison, it is assumed that the OCS port count does not limit the connectivity of the network. The OCS units are transparent and agnostic to the modulation format used to transmit and receive data, but are actively switching in the network core, in contrast to the designs presented in the earlier chapters of this thesis.

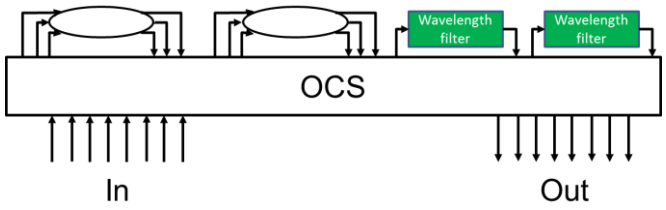

Figure 57: The building blocks of the sub-star networks discussed throughout this section. OCS = optical circuit switch.

Where the port counts required to create each topology are calculated below, the numbers quoted do not include the ports required to connect nodes to the sub-stars, which is constant across all architectures studied. The connection power loss over each OCS link is defined as L, and in loss budget simulations, fixed at 0.2 dB. The wavelength filter power loss is defined as F, and is simulated as 0.2 dB loss in the pass-band, and infinite loss in the stop-band.

Within each sub-star, the WS-TDM combination outlined in chapter 2 can be used to share the transmission capacity. Given that each sub-star is physically separated from all others, each sub-star can independently share the full bandwidth of the WS-TDM system (WB as described in section 3.1.2). This results in a greater expected bandwidth per transmitter when multiple sub-stars are formed within a network, compared to all N nodes connected to a single large star.

Should a node connected to one sub-star request connectivity to a node connected to a different sub-star, the two sub-stars must join together to accommodate the connectivity request; this is an essential design requirement met by all designs proposed in this thesis. After joining multiple sub-stars together, the WS-TDM network wavelength and timeslot allocation algorithm (beyond the scope of this thesis) must reallocate transmission rights across the new combined sub-star as a whole.

The major drawback to these designs is the relatively long time (ms) for a device to connect to the network, and for reconfiguration of the sub-stars when traffic demands change. This time is imposed by the speed of switching of the OCS units which

connect the nodes and couplers. However, it is envisaged that the reconfiguration of the OCS units occurs on a timescale of seconds or minutes, over which some traffic patterns within data centres, particularly media facilities, can be considered stable [1], [146]. The millisecond-scale switching time of OCS units would then only impose a reconfiguration downtime of at most 1 % of the duration of stable traffic flows.

There are several characteristics of the four network designs studied that allow comparisons with both the design criteria and the other designs, including:

- The method of adding a new node to an existing sub-star;
- The method of connecting sub-stars together;
- The maximum and minimum power losses across a sub-star;
- The number of wavelength filters required per sub-star;
- And the total number of OCS ports required to create the network design.

4.2.2. Centred star

The centred star network design is constructed as follows: a “central” optical passive star coupler, of size $b \times b$ is at the centre of each sub-star. Further “outer” couplers of size $c \times c$ connect nodes to the central coupler. The topology of a single sub-star is shown in Figure 58; the OCS is omitted for clarity, but all links between couplers (shown in red lines) are via an OCS.

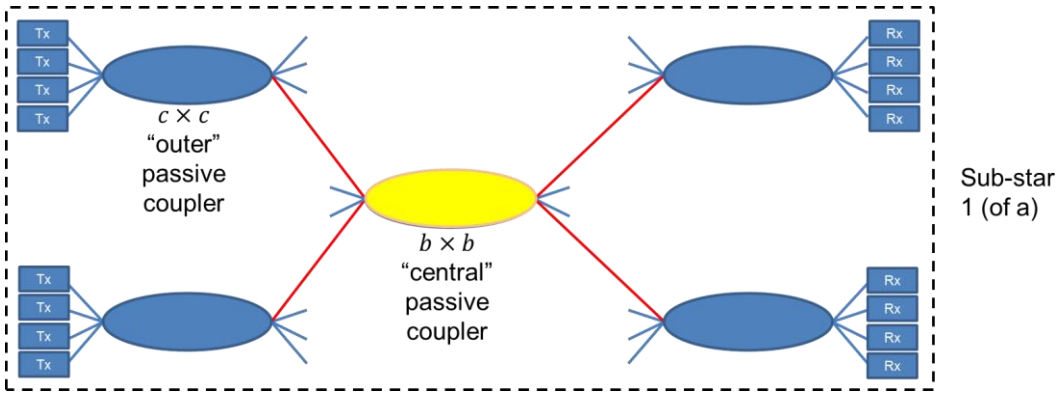

Figure 58: A central star network design, where an initial central $b \times b$ coupler has $c \times c$ passive stars connected to it, creating a single large star.

For a single sub-star to connect all N nodes, the inequality $bc \geq N$ must be obeyed. However, the network may also grow to connect all N nodes by combining multiple sub-stars together. The maximum possible number of sub-stars that can be simultaneously formed is defined as a, and is limited only by the number of central $b \times b$ passive optical couplers attached to the OCS. To allow any permutation of sub-stars to combine to connect all nodes into a single sub-star network, the inequalities $c \geq a$ and $b \geq a$ must also be obeyed.

When connectivity is required between nodes located on two different sub-stars, the sub-stars are combined by cross-connecting the outer transmit couplers and central stars of both sub-stars. This can be explained by an example, shown in Figure 59. When a data flow is requested between a transmitter on sub-star 1 and a receiver on sub-star 2, sub-stars 1 and 2 must join together. To join the two sub-stars requires the outer couplers on the transmitter side of sub-star 1 (top left in Figure 59) to be connected to the central coupler of sub-star 2, and the outer couplers on the transmitter side of sub-star 2 (bottom left in Figure 59) to be connected to the central coupler of sub-star 1. The additional connectivity to join the sub-stars is shown by dashed red lines on Figure 59. No reconfiguration is required on the receiver side of either sub-star. This is because connectivity between the sub-stars on the receiver side would create multiple paths through the network, causing interference where optical signals are superimposed on themselves.

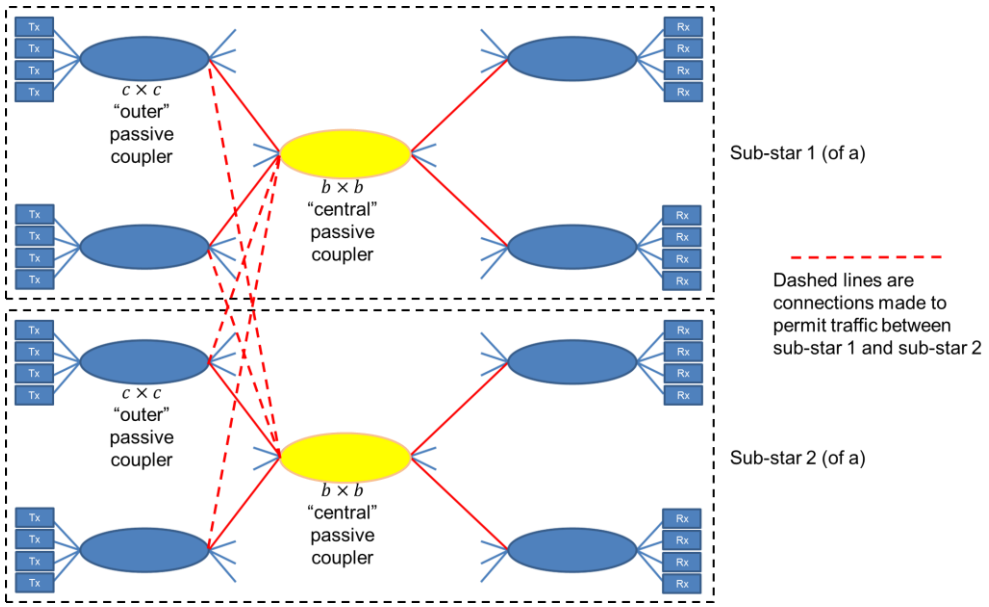

Figure 59: An example of two central star networks joining to provide bidirectional connectivity between transceivers on sub-stars 1 and 2. The dashed lines are the connections that must be made by the OCS to support traffic without creating multiple optical paths through the combined network.

The maximum throughput that could be obtained through each individual sub-star is limited to the WB product, as described in section 3.1.2. However, the total network throughput is equal to the sum of the throughput of all sub-stars, up to aWB. This upper bound to throughput is achieved only if all sub-stars remain separated.

The maximum power loss through a centred star sub-star network is $20\log_{10} c + 10\log_{10} b + 2L$ dB. However, this is also the minimum power loss, and this remains constant even after combining multiple sub-stars together to form a new sub-star network. This is appealing, since a constant power loss independent of network path

avoids power equalisation issues at the receiver (such as misaligned decision thresholds or equaliser inaccuracy), which occur when signals experience path-dependent attenuation. The power loss also does not depend on the potential number of sub-stars that can be formed, which is constrained only by the number of central optical couplers available to the network, a.

For a target network size of 1024 nodes, the system power loss budget for varying sizes of optical coupler b and c is plotted in Figure 60. The region coloured white does not meet the target of 1024 or more nodes. The lowest loss of 37.5dB is achieved for $b = 205, c = 5$ (marked with a ☆ on Figure 60); the general trend is for lower loss for large central coupler sizes (b) combined with small outer coupler sizes (c).

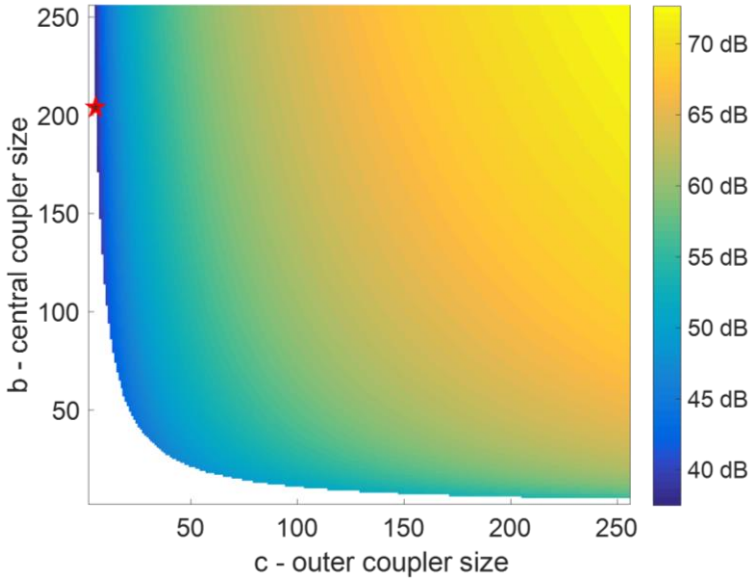

Figure 60: Network optical power loss for a centred star optical network of varying central and outer coupler sizes. The operating point for the lowest power loss of 37.5 dB is marked with ☆.

The centred star topology for sub-stars does not require wavelength filtering. The abstraction of a single large optical star can always be achieved with no optical path loops where optical signals at the same wavelength could potentially interfere.

Two extreme cases of connectivity set bounds to the number of OCS ports required to form the network. When all nodes are attached to a single sub-star, the number of OCS ports required within the network is bounded at $2b$. This is because the only connections that need to be made are from the central $b \times b$ coupler to b couplers on the transmitter side, and b couplers on the receiver side. However, if all a central couplers form separate sub-stars, with only a single outer coupler per sub-star, the total number of OCS connections required within the network is bounded by $a^2 + a$. Whether these bounds form upper or lower limits to the full network OCS port count depends on the choice of both a and b.

4.2.3. Ring

The ring sub-star topology is constructed by connecting a passive optical star couplers in a "ring", as shown for $a = 4$ in Figure 61. Each "ring" coupler has a port count of $b \times b$ (shown in blue in Figure 61), and one output from each "ring" coupler is connected to the input of just one other "ring" coupler, via a wavelength filter. Each of the ring couplers can then be attached to up to $b - 1$ optical passive star couplers, each of size $c \times c$ (shown in yellow in Figure 61).

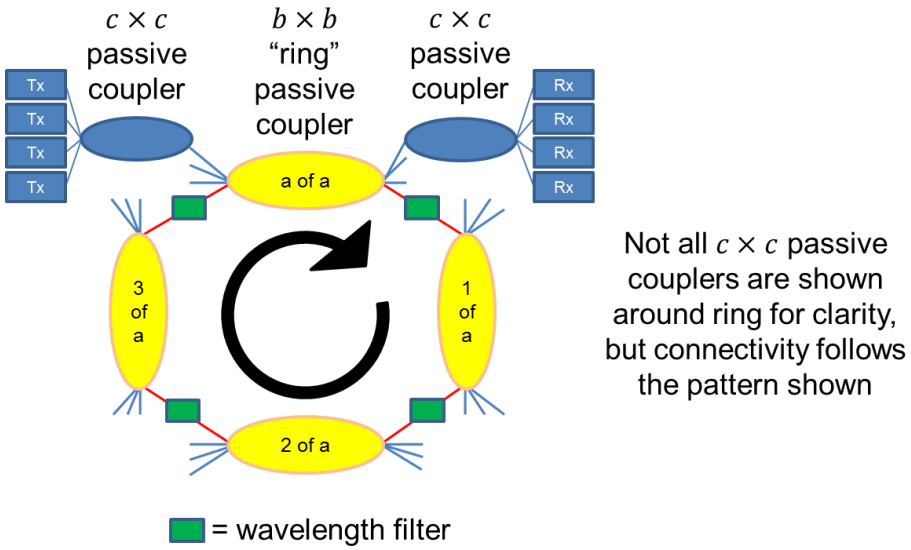

Figure 61: A ring network topology which provides all-to-all connectivity. In this example, $a = 4$ central passive optical star couplers each of port count $b \times b$ are connected in a ring. Note the fixed direction of transmission flow around the ring, and the need for wavelength tunable filtering on every link between couplers on the ring.

The splitting of a ring network (in the same fashion as splitting a network into sub-stars) was considered here by using wavelength filtering alone – it is not possible to add and remove ring couplers without temporarily blocking transmission around the ring, which would disobey the design requirement for fully non-blocking switching (see section 4.1). To ensure that the ring design abstracts to a single large star connecting all N nodes, the number of optical star couplers and their port counts must obey the equation $a(b-1)c \geq N$ (derived by combining the total number of ports on each coupler to reach a total node count).

Wavelength filtering is essential at all links between central couplers on the ring, so that transmissions launched into the ring do not continue indefinitely in a loop. Three possible options are considered to configure the pass-band of the wavelength filters:

1. Fixed wavelength assignment: each ring coupler is assigned a set of wavelengths, which can be freely allocated to any transmitter which is attached to that ring coupler. A simple assignment would provide $\frac{W}{a}$ wavelengths to each

central coupler on the ring, so that each ring coupler has an equal share of the available bandwidth. This means that each filter should block $\frac{W}{a}$ wavelengths and transmit $1 - \frac{W}{a}$ wavelengths, with a different set of wavelengths blocked at each filter. Since the wavelengths allocated to each ring coupler are fixed, fixed wavelength filters can be used. The bandwidth can also be shared unequally (but still in a fixed configuration) by varying the number of wavelengths assigned to each ring coupler, if this matches the connectivity pattern.

2. Dynamic wavelength assignment (coupler assigned): each ring coupler is assigned a number of wavelengths to allocate freely between transmitters attached to that ring coupler, but the number of wavelengths assigned to each coupler can be varied as the application demands change. Each wavelength can only be assigned to a single ring coupler, to avoid contention if the same wavelength were used multiple times, causing interference. This assignment method requires tunable wavelength filters.
3. Dynamic wavelength assignment (flow assigned): the wavelengths allocated to each ring coupler can be varied on demand, and the same wavelength can be assigned to multiple ring couplers simultaneously. The wavelength filter at the input to each ring coupler can reject not only wavelengths that originated at that ring coupler to avoid interference and continuous loops, but can also reject wavelengths that are not required at destinations further around the ring. An example of this approach is shown in Figure 62, where the same wavelength is used by two transmitters simultaneously. By blocking the dark blue wavelength at the filters marked “B” in Figure 62, but transmitting it through the other filters, the wavelength can be used for two flows simultaneously.

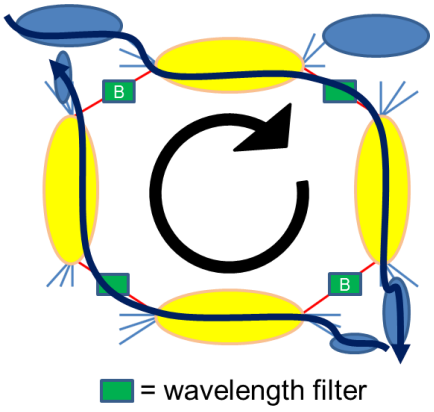

Figure 62: An example of dynamic wavelength filter adjustment across the network, permitting wavelength reuse. By setting the filters marked B to block the dark blue wavelength, and transmitting the dark blue wavelength through the other two filters, this dark blue wavelength can be re-used around the ring, increasing the overall transmission capacity.

The ring topology described above still abstracts to the same connectivity as a star coupler. All nodes are directly connected by a single hop, and the multicasting from one node to all other nodes is always permitted.

In the ring topology, some paths would experience vastly different losses compared to others, for example the maximum path length would pass through all other couplers on the ring to reach all destinations. The minimum loss between a receiver and transmitter is $20\log_{10} c + 10\log_{10} b + 2L$ dB, but the maximum loss is $20\log_{10} c + 10a\log_{10} b + (a-1)F + 2aL$ dB. The number of ring couplers, a, has a linear impact on the maximum loss, assuming that the port count of each ring coupler, b, remains constant.

To assess the impact of the number of ring couplers on the network optical power loss, alongside the impact of the port count of both the ring couplers and the outer couplers, a simulation was performed. For fixed values of the number of central couplers $a = 5,10,15$ and 20, the central and outer coupler size was varied and the maximum system optical power loss calculated. The results are plotted in Figure 63, where the white regions of the plot are areas where a single sub-star supports fewer than 1024 nodes.

The general trend for any number of central couplers (a) is for the lowest loss budget to be achieved with smaller ring couplers (lower b) and larger outer couplers (higher c), as per the darkest blue regions of Figure 63. Intuitively this is reasonable: the loss equation above is defined for the worst-case scenario, where a data flow will have to pass through all ring couplers to ensure that the transmission reaches all nodes.

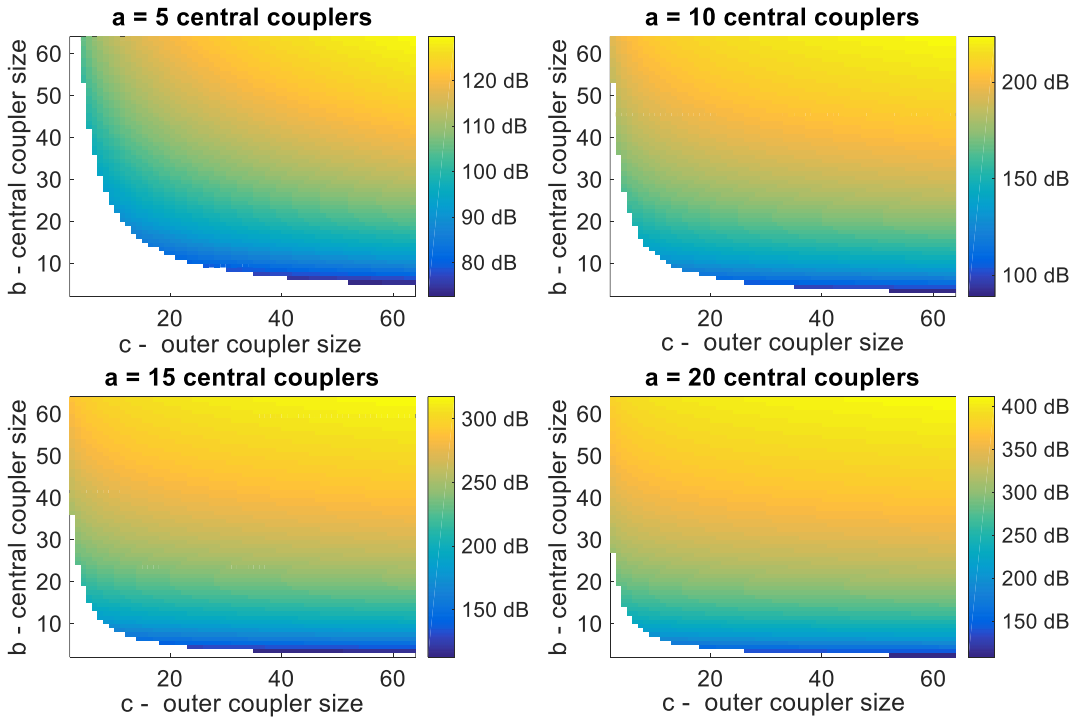

Figure 63: Network optical power loss for a ring optical network of four different numbers of central couplers, as well as varying central and outer coupler sizes.

For any number of ring couplers, the power loss is much higher than the centred star topology, with a minimum loss of 64.5 dB for $a = 5$ ring couplers ($b = 5, c = 52$). The power loss increases linearly with the number of ring couplers (a), assuming constant values of b and c. In this simulation, the lowest power loss observed for $a = 20$ central couplers was 108.2 dB ($b = 3, c = 53$), and all power losses for $a = 20$ were unreasonably high (from 108.2 dB to more than 400 dB). All of these losses are impracticably high for full implementation, considering the tolerable loss budgets across the WS-TDM design of only 30 dB shown in chapter 2 - Table 9 shows a comparison of the loss budget for each topology assuming a 1000 node network.

The minimum total throughput is the same as for a single large star i.e. WB, and this throughput is constant for the first two wavelength allocation scenarios. Considering wavelength allocation scenario 3 (dynamic filters with wavelengths assigned to flows rather than couplers), the total throughput can increase. The exact value of the increase is traffic dependent, rather than coupler size dependent, but if transmitters only require connection to receivers on the same ring coupler (i.e. all wavelength filters block all wavelengths), the potential maximum throughput is aWB. The upper bound to the number of OCS ports required is $2ab$; this bound may not be reached depending on the uniformity of the distribution of outer couplers around the ring.

4.2.4. Mesh

This topology is constructed as follows: individual optical passive star couplers of size $b \times b$ are connected such that one output of each coupler is connected to an input of all other couplers, via a wavelength selective filter. An example of this topology is shown in Figure 64 for three couplers. The wavelength filters are required to attenuate any wavelengths that originate from the coupler that they precede, so that continuous loops are not formed. All other wavelengths are transmitted through the filter, attenuated only by the filter's pass-band loss of F dB.

Should connectivity be required between two sub-stars, direct connectivity must be made between all couplers in both sub-stars, via wavelength filters. The overall topology remains identical after performing the connection; each coupler has a direct connection to all other couplers in the combined sub-star.

The number of couplers forming the network (i.e. including all sub-stars) is defined as a. Each coupler has a port count b obeying the inequality $b > (a - 1)$, so that each coupler can connect to all other couplers in the mesh, while also still connecting $(b - a + 1)$ nodes. The inequality includes the limiting case $a = b$, which would mean each coupler in the sub-star connecting only a single network node to all other nodes. To meet the abstraction of a single large star, the variables must also obey the inequality

$a(b-a+1) \geq N$ so that all of the nodes across the whole network can be connect to a single sub-star.

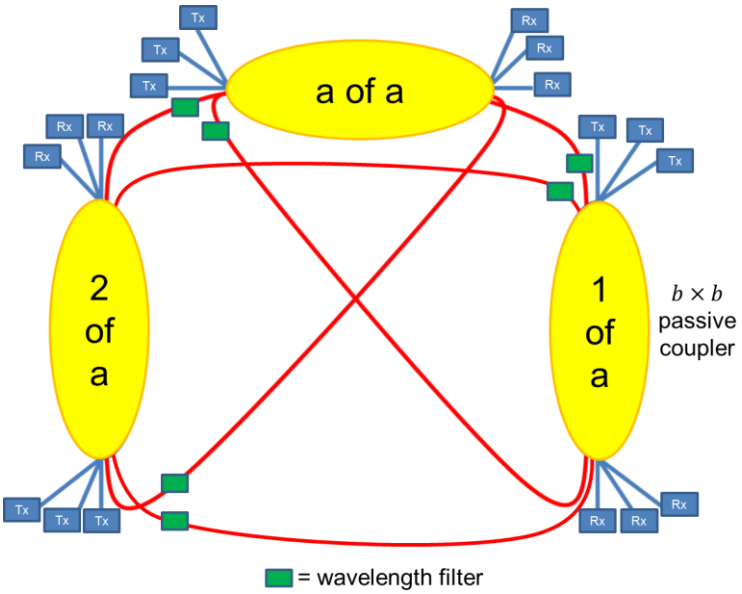

Figure 64: A mesh sub-star topology constructed from smaller couplers. Note that each coupler has a direct connection to all other couplers.

The minimum optical path loss occurs when nodes communicate across just a single coupler, with a loss of $10 \log_{10} b$ dB. The maximum loss occurs when nodes attached to different couplers communicate, with a loss of $20 \log_{10} b + F + 2L$ dB (where F is the insertion loss of each wavelength filter and L is the optical path loss of each connection between couplers). These are the only two possible power loss values in this topology.

A simulation was performed of the total maximum power budget of a single sub-star as both the number of couplers and their port count was varied. The results are shown in Figure 65. The lowest system power loss for $a = 19$ couplers each of size $b = 72$ is 38.1 dB. There is a general trend for the lowest system loss for higher numbers of couplers (higher a) and smaller coupler port counts (lower b). This is reasonable given that the loss budget of the mesh network has no dependence on the number of couplers a. The value of a is only important in determining valid values of b to meet the total node count requirement (N, defined above).

Since it is essential to provide a wavelength filter at every input to a coupler, the number of wavelength filters required in this topology is $a(a-1)$. However, each coupler has a direct connection to every other coupler via only a single wavelength filter, so the high number of filters required (compared to the other topologies described in this chapter) does not result in excessive insertion loss penalties. Although fixed optical filters are inexpensive, tunable wavelength filters are expensive and complex devices, especially fast tunable filters (e.g. silicon photonic devices have been

demonstrated to tune in 10 µs across the C-band, but are not yet commercially available [147]). The total number of OCS ports required in this network topology is $2a(a-1)$.

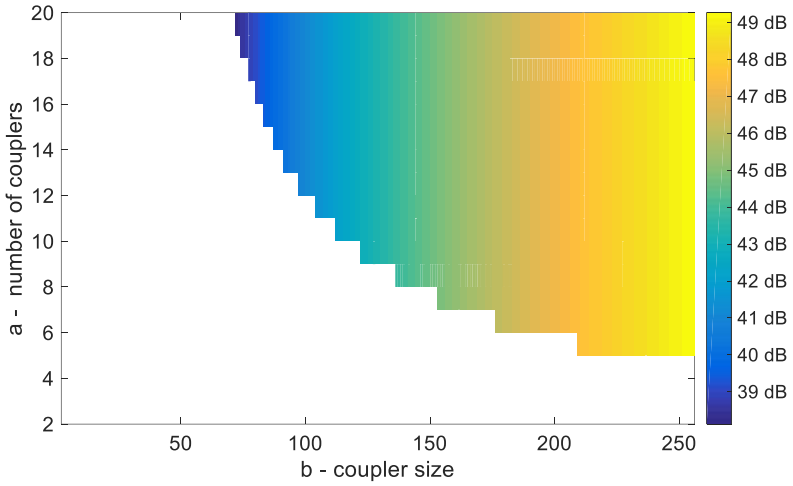

Figure 65: Sub-star power loss for a mesh of varying coupler sizes and numbers.

4.2.5. Centred mesh

This topology is formed as follows, and shown diagrammatically in Figure 66: a coupler of size $b \times b$ is defined as the "central" coupler of the sub-star (yellow in Figure 66). Other couplers of size $c \times c$ (defined as "outer" couplers, and shown blue in Figure 66) are attached to the central coupler by two links: an output of each outer coupler is connected to an input of the central coupler via a wavelength filter; and an input of each outer coupler is connected to an output of the central coupler. Node transceivers are connected to outer couplers. The wavelength filter preceding each outer coupler should be set to block all wavelengths that originated from that outer star, and to transmit all other wavelengths.

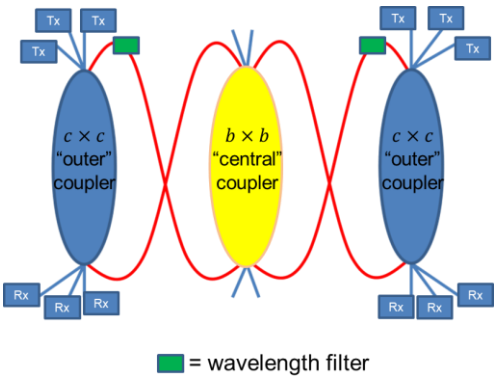

Figure 66: A centred mesh network topology. All outer couplers (blue) are connected to the central coupler, (yellow), providing the abstraction of a single large star connecting all points. Wavelength filters are necessary to stop transmissions continuing around the star structure indefinitely.

To enable this topology to connect all N network nodes into a single large star, the inequality $c(b-1) \geq N$ must be obeyed. There are no further restrictions on the size of the two types of coupler, although the total number of central couplers available at the OCS would define the number of sub-star networks that could be simultaneously formed. New nodes joining a sub-star can be attached to any available ports on an outer coupler; if no ports are free on any outer couplers, a new outer coupler is added to the central coupler. Should two sub-stars need to combine to meet connectivity requirements, the central couplers of each sub-star should be joined output to input, via a wavelength filter, as shown in Figure 67. The resulting topology has increased loss budget compared to the two separate sub-stars, due to the potential for communications to need to traverse multiple central couplers. Once two central couplers have been combined into a single sub-star network, one of the central couplers should be designated the network master central coupler. If a sub-star network which already contains two central couplers is required to merge with a further sub-star network, the network master central coupler should be connected to the merging central coupler, as per Figure 67. This minimises any further optical power loss due to sub-star merging.

The minimum loss budget between two nodes on a single sub-star is $20\log_{10} c + 10\log_{10} b + F + 3L$ dB. The maximum loss budget is reached when multiple central couplers are attached directly i.e. after the combination of two or more sub-stars. For any link that requires crossing two central couplers (i.e. source and destination are attached to outer couplers that were initially part of different sub-stars), the maximum loss budget occurs, which is $20\log_{10} c + 20\log_{10} b + 2F + 5L$ dB (where F is the insertion loss of each wavelength filter and L is the optical path loss of each connection between couplers).

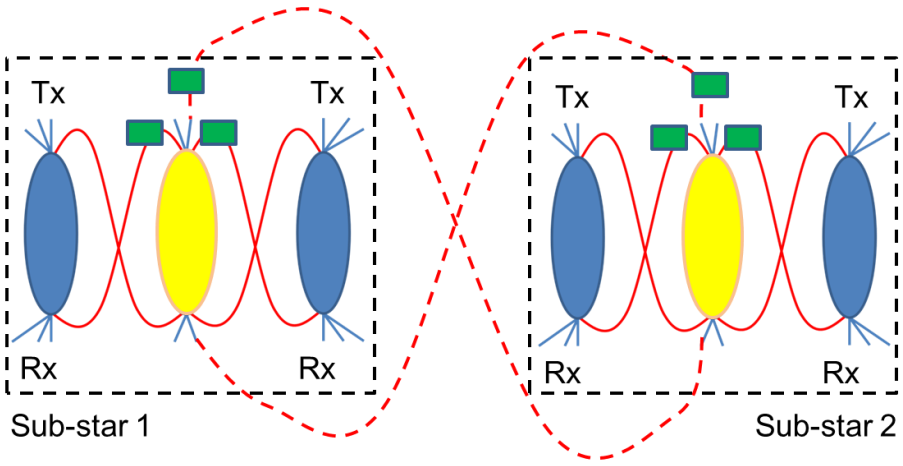

Figure 67: Two sub-star centred meshes combining to form a single sub-star. Dashed red lines show the new connectivity between the two original sub-stars.

The impact on the total optical power loss of varying the port count of both the central couplers and the outer couplers has been simulated, and the results are plotted in Figure 68. The optical power loss was affected equally by the size of the central couplers and the outer couplers used in the network, as shown by the diagonal symmetry of Figure 68. However, to reach the target number of 1024 nodes, the lowest loss simulated here of 61.9 dB is observed for $b = 205,\ c = 5$, which matches the general trend across all topologies in this chapter, of lowest loss for a smaller outer coupler size (lower c) and a larger central coupler size (larger b).

The number of wavelength filters in this design also depends on the number of central star couplers used to construct the topology. Defining the number of central couplers used in the entire network as a, and the number of outer couplers as d, the total number of wavelength filters required is $a + d$. The parameters for each of the topologies can now be summarised and compared, and conclusions drawn as to the optimum architecture to meet the design requirements.

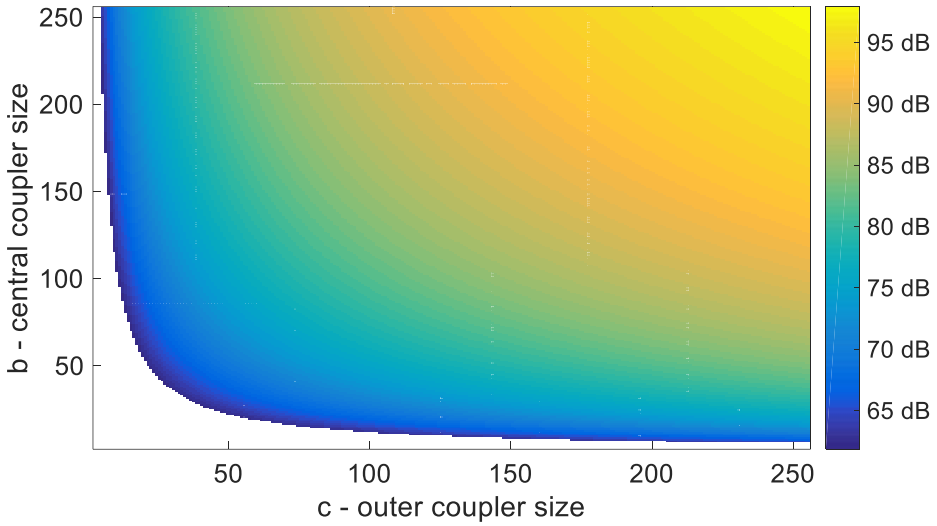

Figure 68: Network optical power loss for a centred mesh optical network as the coupler port counts (*b* and *c*) are varied.

4.2.6. Summary and comparison

A summary of the characteristics of the four sub-star networks is presented in Table 9. For each topology, generalised formulae for physical attributes are given, alongside absolute values for a target of 1024 nodes and a target of 20 central/ring couplers.

Table 9: A summary of the star construction topologies and relevant parameters. *Italic* values correspond to a target of 1024 nodes and 20 central/ring couplers.

Topology	Minimum loss budget (dB)	Maximum loss budget (dB)	Wavelength filtering	Sub-star throughput
Centred Star	$20\log_{10} c + 10\log_{10} b + 2L$ *b = 205, c = 5* *37.5 dB*	$20\log_{10} c + 10\log_{10} b + 2L$ *b = 205, c = 5* *37.5 dB*	None	WB
Ring	$20\log_{10} c + 10\log_{10} b + 2L$ *a = 5, b = 5, c = 52* *41.3 dB*	$20\log_{10} c + 10a\log_{10} b + 2aL + (a-1)F$ *a = 5, b = 5, c = 52* *72.1 dB*	a filters required *a = 5*	$> WB$ if flexible wavelength filters used
Mesh	$10\log_{10} b$ *a = 19, b = 72* *18.6 dB*	$20\log_{10} b + 2L + F$ *a = 19, b = 72* *38.1 dB*	$a(a-1)$ filters required *a(a-1) = 342*	$> WB$ if flexible wavelength filters used
Centred Mesh	$20\log_{10} c + 10\log_{10} b + F + 3L$ *b = 205, c = 5* *37.9 dB*	$20\log_{10} c + 20\log_{10} b + 2F + 5L$ *b = 205, c = 5* *61.4 dB*	$(a+d)$ filters required	WB

4.3. Flow-level traffic simulations

To compare the throughput of these network designs, simulations have been performed at “flow-level” - that is, considering the capacity of the network by simulating flows rather than individual packets. A flow is defined as a continuous stream of data, with no gaps; this means that the maximum flow rate in these simulations is higher than a practical data rate with inter-packet gaps. It is assumed in these simulations that the higher layers of the network stack at each node are capable of sending and receiving

flows at the full transceiver line rate, defined throughout these simulations as 25 Gb/s (matching the WS-TDM experiments of chapter 2).

A sub-star is defined here as an isolated group of nodes that can have their connectivity abstracted to a passive optical star network. All nodes can behave as sources (transmitting a flow), destinations (receiving one or more flows), or both. Nodes request flows, between a source and a destination, by contacting a central controller to ensure that connectivity is available between them. There are then two basic principles to the flow level simulations in this work:

1. All flows that are requested are always allocated connectivity. This means that physical layer connections are always made between a source and destination pair that request a flow between them, even if this means combining two sub-stars together into a single sub-star. In a practical implementation, the source node would request a specific bandwidth from the controller. If the network only has the capacity available to allocate less bandwidth to a flow than was requested, the network controller must instruct the source node to reduce the flow rate to meet the allocation e.g. by sending a lower resolution video or changing the media codec.
2. All flows within a sub-star are allocated the same bandwidth. This allows for meaningful comparisons between the network topologies simulated and the network traffic scenarios. It is feasible for the wavelength/timeslot allocation controller to offer different bandwidths per flow, based on priority or other flow metadata, so long as the total throughput does not exceed the limit per sub-star. However, the implementation of complex controllers capable of variable bandwidth allocation is beyond the scope of this thesis, which considers only the data plane of the network.

Despite the differences between the four topologies described in the previous section, the network designs reduce to just two possible models when performing flow level simulations:

- For the centred star, mesh and centred mesh topologies, the flow level simulation is identical for all designs. Sub-stars are formed to accommodate network requests as they are received, connecting sub-stars together as necessary to meet the traffic demands. A flow level simulation does not need to be aware of the precise mechanics of splitting and connecting sub-stars and/or wavelength filtering, but only needs to be aware that splitting and combining of sub-stars is always possible, which is true for all three of the centred star, mesh and centred mesh designs.

- For the ring topology, the links between couplers are always fixed, so the reconfiguration of the tunable wavelength filters is the only method of varying connectivity. The links between ring couplers provide more of a bottleneck than the couplers themselves, since each link must carry transmissions that originate from many sources across multiple stars. A flow level simulation must determine the number of flows per link, which restricts the available bandwidth. The third wavelength allocation scenario from section 4.2.3, dynamic allocation of wavelengths at flow level, is used in these simulations.

4.3.1. Simulation parameters and workflow

The throughput performance of the network is always dependent on the network traffic, as it is the source-destination connectivity which defines the sub-star formation. It is defined that destinations request connectivity from sources in this simulation model, matching practical media workflows (high bitrate media streams are not injected into a network unless a receiver requests them). To assess the potential increase in total network throughput by splitting the network into sub-stars (relative to a single star network), three distinct traffic patterns were used to model the flows through the system:

1. Random
   Each destination selects any source on the network, at random, to receive a data flow from.
2. Hotspot
   A group of g sources out of the N total sources is designated a "hotspot". Each destination has an $h\%$ likelihood of requesting a data flow from any source in the hotspot, and a $(1-h)\%$ chance of requesting a data flow from any source that is not in the hotspot. In this work g = N/10 and h = 50%.
3. Zonal Media Production
   Each node is defined as a member of one of five "zones", designed to separate nodes according to their function at the application layer (a node can function as a source, destination or both simultaneously). A traffic probability matrix defines the probability that a destination would request a source from each of the zones; the probability values are taken from a survey of a media production network traffic within a medium sized live production centre.

A visualisation of a sample dataset of each of the three traffic patterns is shown in Figure 69. The random traffic shown in Figure 69a has no locality, as there is the same

chance that a destination will request a data flow from a specific source as from any other source. In Figure 69b, data flows from the “hotspot” source nodes are shown in red, while data flows from all other source nodes are shown in blue. The zonal media production traffic is shown in Figure 69c, where sources in each of the five zones are each allocated a colour. There is clear locality shown in Figure 69c, where most data flows follow distinct patterns of connectivity across the network.

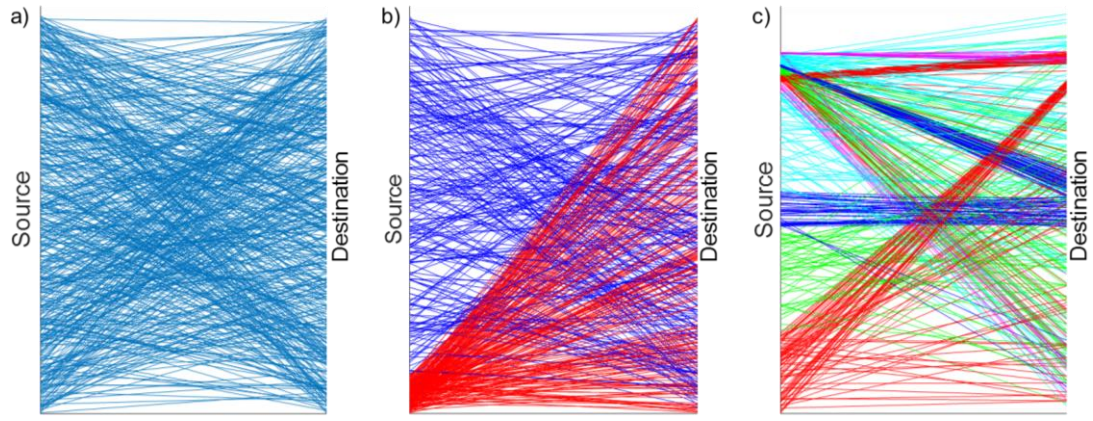

Figure 69: Visualisations of the three traffic patterns used in simulation: a) random traffic; b) hotspot traffic; and c) zonal media production traffic.

To perform each simulation, the following workflow was simulated:

- Generate a list of 10^6 source-destination pairs which follow the traffic distribution under simulation (from the three definitions above). The source-destination pairs are listed in a random order.
- Considering each source-destination pair in turn, provision the fibre connectivity required to transmit a flow between the source and the destination i.e. connect the source and the destination to the same sub-star via the OCS (including connecting sub-stars if necessary).
- It is possible that multiple destinations will request the same source i.e. multicast. It is also possible that multiple sources will be requested by the same destination i.e. in-cast. This is because each pair is added to the network sequentially, with no consideration of prior connectivity, and is a realistic method of provisioning connectivity in a media production data centre.
- Count the total number of unique sources which are transmitting flows into the network. Defining the load percentage of the network as l, after lN unique sources are transmitting into the network, the load level has been reached and no further flows are added from the source-destination pairs list. Note that the total number of active destinations has no direct impact on the network load.
- Count the number of sources in each sub-star (denoted n) and find the expected transmission rate per source for each sub-star $\left(\frac{WB}{n}\right)$.

- Calculate the median transmission rate per source, considering all sources on all sub-stars in the network.

The median transmission rate per source, considering all sub-stars, is selected in preference to other averages to more accurately reflect network performance. A sub-star with few active sources would have a disproportionately strong effect on the mean transmission rate per node of all sub-stars in the network. For example, the network could form one sub-star of $n = W$ sources, each transmitting at a transmission rate of B, and one sub-star of $n = (N - W)$ sources, each transmitting at a transmission rate of $\frac{WB}{N-W}$. In this example, the mean bandwidth per transmitter, considering all transmitters on all sub-stars, is $\frac{2WB}{N}$, while the median is $\frac{WB}{N-W}$. Given that $N \gg W$, the mean transmission rate could be approximately double the median transmission rate. However, there are far more flows with the lower transmission rate, so the median is more representative of transmission rates on average. Presenting the median in this work, therefore, gives a realistic representation of the expected transmission rates. In addition, the transmission rate of each source has an upper bound, as transmitters cannot transmit faster than their maximum line rate B when $n < W$.

The results in this section, starting with Figure 70, are presented in terms of the probability distribution function (PDF) of the median transmission rate per node.

- For a single star network, all nodes can transmit at the same rate (as described in section 3.1.2). This rate is fixed, regardless of the network traffic, and is displayed on the PDF plots as vertical dashed lines. Single star transmission rates are a minimum bound to the split star transmission rate per node.
- For network designs which can split into sub-stars, the transmission rate per node can vary, based on the connectivity required. The network seeks to maximise the transmission rate of all nodes by optimally splitting into sub-stars wherever possible. Transmission rates directly depend on the traffic, and as it is not practical to simulate all connectivity combinations, a Monte Carlo simulation finds the probability of the network achieving each transmission rate.

If the peak of a PDF is at a higher transmission rate than the single star worst-case transmission rate, the sub-star network design has improved upon the single star design of Chapter 2 by exploiting traffic locality to increase the throughput per node. Where a percentage comparison is made between the expected transmission rates of the split sub-star topologies and the single star topology of chapter 2, it is quoted for the mean of the median transmission rates. This means that the percentage increases in throughput will be achieved on average 50% of the time. For instance, if it is stated

that there is a 26% increase in throughput for a given traffic pattern and network load, this means that each transmitter on average can transmit 26% more data per epoch in the split star configuration compared to a single large star. Because the percentage increase is calculated as an average over the Monte Carlo simulation, in a real implementation the throughput increase has a 50% likelihood of being more than the value quoted, and a 50% likelihood of being less than the value quoted.

The specific parameters used in these simulations were:

- All transmitters have a maximum transmission rate of 25 Gbit/s;
- For the random and hotspot traffic, the network had 1024 nodes (all nodes could function as a source or destination);
- For the zonal media production traffic, 1280 nodes were destinations, and 960 nodes were sources (matching the layout of a real media production centre);
- Every transmitter and receiver can access 120 wavelengths (matching the achievable extended range of the DSDBR lasers, as described in section 3.2.1);
- And the number of timeslots per epoch is not relevant, since in flow level simulations, it is assumed that sufficient timeslots will be allocated to provide the allocated transmission rate on aggregate.

4.3.2. Centred star, mesh and centred mesh simulation

To perform the simulation, instructions were defined to determine the formation of sub-stars as flows were added to the network, which in turn can be translated to connectivity at the OCS for each of the three network topologies: This is presented as an algorithm in Appendix A and described as follows:

- In general, pre-existing active connectivity is used to determine which sub-star a new flow should be connected through; if either source or destination are already active, they cannot be moved to another sub-star without interrupting existing flows.
- If neither the source nor destination of a flow are already members of any sub-stars, a preferred sub-star is chosen which has the lowest number of active sources out of all existing sub-stars in the network.

The most efficient use of bandwidth is made when each sub-star has the same number of active nodes as wavelengths (i.e. $n = W$). In this situation, each node can transmit at the full transmitter line rate on a dedicated wavelength, and no capacity is spare. The algorithm, therefore, targets this optimal state by only attaching flows for newly active sources and destinations to sub-stars with less than W active sources. This does not

preclude sub-stars ever reaching $> W$ active sources when sources are added to the same sub-star as existing destinations, but does aim to reduce wasted bandwidth.

Figure 70 shows PDFs for a split star network formed using the centred star, mesh, or centred mesh topologies, for random traffic. For traffic loads of 80% or 100%, there is no improvement to the median bandwidth per node compared to a single passive star network (shown by the alignment of the PDF with the dashed line at the single star transmission rate). For a 40% or 20% network load, there is a 6.6% and 27.3% increase in the median transmission rate compared to a single star network.

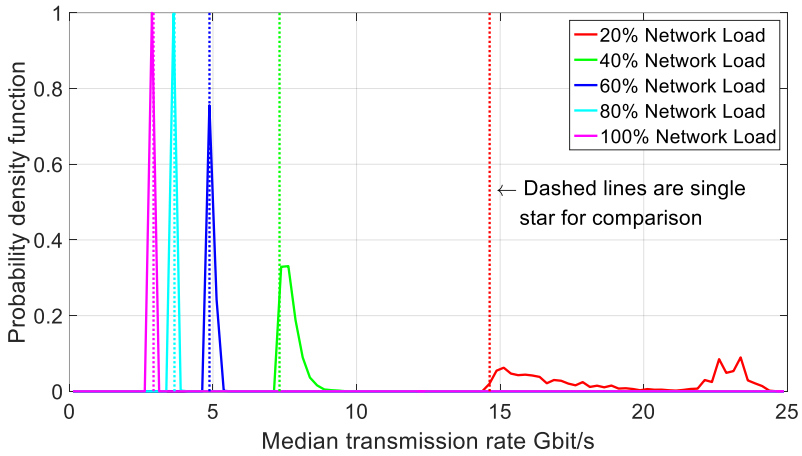

Figure 70: Probability density function (PDF) of the median transmission rate for random traffic. Results are shown for a split star network using centred star, mesh, or centred mesh topologies, compared to a single passive star network.

Figure 71 shows PDFs of the median transmission rate per node for a split star network formed using the centred star, mesh, or centred mesh topologies, for hotspot traffic. There is a clear increase in the expected transmission rate per node compared to a single large star for all network loads simulated, from a 258% increase at 100% network load, down to a 136% increase in throughput for 20% network load.

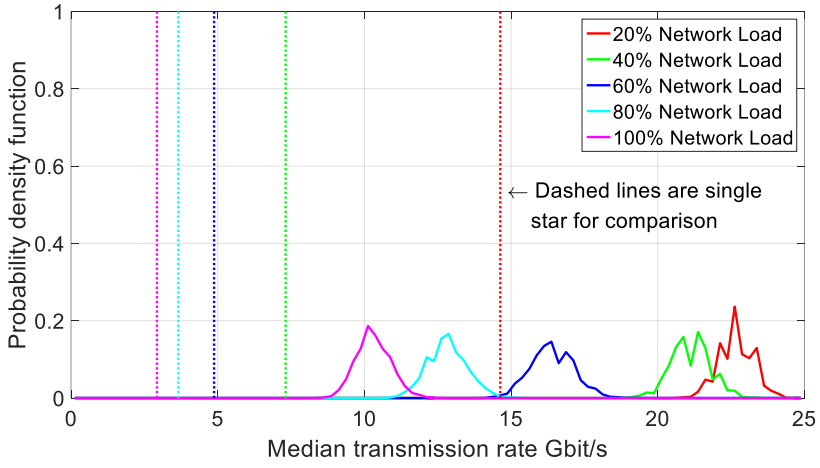

Figure 71: Probability density function (PDF) of the median transmission rate for hotspot traffic. Results are shown for a split star network using centred star, mesh, or centred mesh topologies, compared to a single passive star network.

Hotspot traffic is a good fit to network topologies that grow sub-stars as network requests arrive, due to the clustering effect of the hotspot on the network connectivity. In this model, there is a 50% likelihood that new destinations becoming active on the network will request a source from the hotspot group. If the same destination then requests an additional flow from another source (possible due to expressly including in-cast in the network traffic feasibility), there is a 50% likelihood that the requested source will also be a member of the hotspot group, and if it is not already active, it will join the same sub-star as the originally requested source. This pattern repeats as more and more flows are set up across the network, resulting in a highly clustered network.

Figure 72 shows PDFs for a split star network formed using the centred star, mesh, or centred mesh topologies, for zonal media production traffic. For a 100% and 80% network load, on average there is a 10.1% and 10.5% increase in the median throughput per node. Since the zonal traffic displays locality, it is more likely that when a new source or destination becomes active, flows will be requested which match the existing physical connectivity. If new flows can be added without requiring new connectivity between sub-stars, it is more likely that each sub-star can remain physically separated. This separation between individual sub-stars allows the increased throughput compared to a single large star.

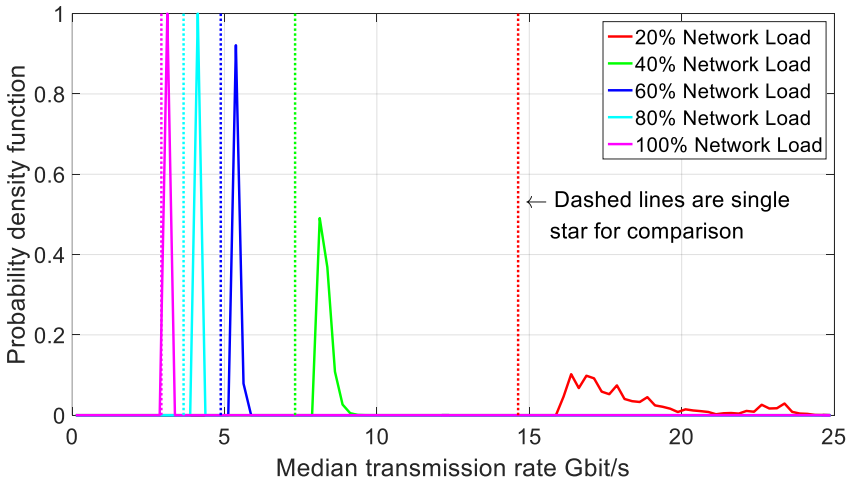

Figure 72: Probability density function (PDF) of the median transmission rate for zonal media production traffic. Results are shown for a split star network using centred star, mesh, or centred mesh topologies, compared to a single passive star network.

The combination of Figure 70 to Figure 72 show that the traffic pattern across the data centre is highly influential on the achievable data transmission rates. The aim of this work is to increase the data transmission rates when using a reconfigurable set of couplers to form sub-stars compared to a single large passive star network. Assuming a high network load (> 80%), there is only increased throughput for the hotspot or zonal traffic patterns, with the hotspot traffic showing the largest increase in expected

bandwidth per node. Traffic patterns with locality (including hotspot, which shows extreme locality), allow the construction of segregated sub-stars where it is unlikely (but not impossible) that the sub-stars would need to join together to add new connectivity.

Figure 73 shows the mean number of sub-stars formed over all 10,000 simulation trials performed. Although the number of sub-stars formed must practically be an integer, the mean is presented here for ease of comparison. The hotspot traffic consistently splits the network into more sub-stars than the other two scenarios. The extreme locality of hotspot traffic means that many destinations request sources which are already on a sub-star, and few connections are necessary to combine sub-stars together.

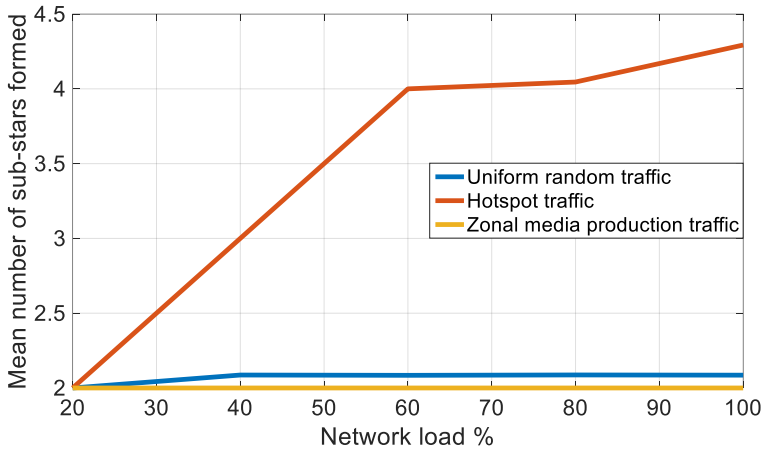

Figure 73: The mean number of sub-stars formed for each of the three traffic scenarios, varying with network load.

Hotspot traffic allows partitioning of the network into more sub-stars than the other scenarios; more sub-stars means that each sub-star connects fewer nodes. This is shown in Figure 74, where the mean number of nodes per sub-star at 100% network load is only 239 for hotspot traffic, compared to 491 and 465 for random and zonal media respectively. For all network loads studied, the hotspot traffic formed sub-stars with relatively fewer nodes; resulting in turn in a higher transmission rate per node.

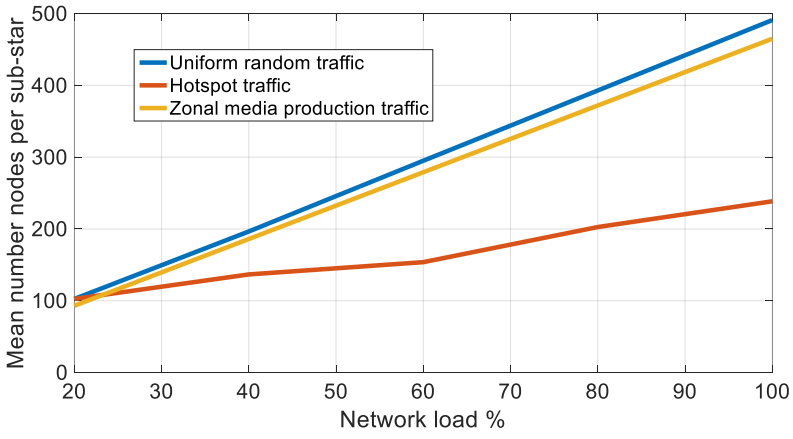

Figure 74: Mean number of nodes on each sub-star, varying with network load, for each of the three traffic scenarios studied.

4.3.3. Ring simulation

In the ring topology, the number of links between ring couplers is fixed, and is the same as the number of ring couplers. In contrast to the other topologies, the connectivity of the couplers cannot be changed to create separate sub-stars. Instead, the tunable wavelength filters between ring couplers are reconfigured to partition the network. Although section 4.2.3 showed that the ring topology could not be realised in practice for a 1000 node network due to impracticably high power losses, throughput simulations are included here for completeness.

For flows that are transmitted from a source to a destination on the same ring coupler, only the total throughput through the ring coupler would limit the bandwidth allocation per flow, assuming that wavelength filtering can be used to segregate a wavelength for use within that coupler alone. However, flows could also be transmitted from a source on one ring coupler to a destination on a different ring coupler. The links between couplers then set the limit to the maximum throughput.

For the ring network topologies, many wavelength allocation algorithms have been proposed in prior work [148]–[151], and no new proposals for wavelength allocation are presented here. Instead, it is assumed that an optimal wavelength allocation could be found, which ensures that every link between couplers is saturated, and all wavelengths and timeslots are fully utilised. By measuring the desired connectivity requests for each of the three traffic patterns, the number of unique flows over each link of the ring can be calculated. Given that the maximum bandwidth per link is known (WB, assuming optimal allocation), the expected bandwidth per flow is the total bandwidth divided by the number of flows per link. The most congested link limits the expected bandwidth of all flows passing through that link, providing a worst case bound to expected bandwidth per flow around the whole network.

The simulation proceeded as follows:

```
1       for all source-destination pairs
2               find central coupler of source
3               find central coupler of destination
4               determine path from source coupler to destination coupler
5               reserve a flow request on each link on the path
6       end
```

It was shown in [148] and [152] that the most important parameter in determining capacity bounds around optical ring networks is the maximum number of flows that

must share any link around the ring, which is defined as L_{max}. In [148] it is shown that to serve all connection requests around a ring network without sharing any wavelengths, a minimum of $0.5L_{max}\log_2 N + L_{max}$ wavelengths are required. It is then always possible to assign wavelengths and meet connection requests without blocking, provided that L_{max} does not exceed $\frac{W}{3\log_2 N}$. Given that the data centre scales that this thesis targets are on the order of $W = 120$ and $N = 1024$, non-blocking throughput would only be guaranteed if $L_{max} < 4$, which is a low number of flows through any link.

It is assumed that each sub-star uses all available wavelengths and timeslots of a WS-TDM system. This is achieved by sharing wavelengths via TDM, to allocate timeslots on the same wavelength to multiple transmitters. The maximum load of any link around the ring is therefore likely to exceed $L_{max} = 4$, given that there are 1024 nodes and network load levels of at least 70%. It is assumed in the simulations that a wavelength assignment can always be found which allows all nodes to transmit some data, that all flows are assigned the same bandwidth, and that the link with maximum congestion provides a limit to the maximum bandwidth of all flows around the ring.

Figure 75 shows how the median transmission rate per node varies under random traffic for the ring topology. The ring allows the transmission rate to increase compared to the passive star for low network loads i.e. a 70.8% and 73.7% increase for 20% and 40% traffic loads respectively. However, the increase for a high traffic load is more modest, at only 11.3% for a 100% network load, and still only providing a 3.26 Gbit/s flow rate despite a 25 Gbit/s transmitter line rate.

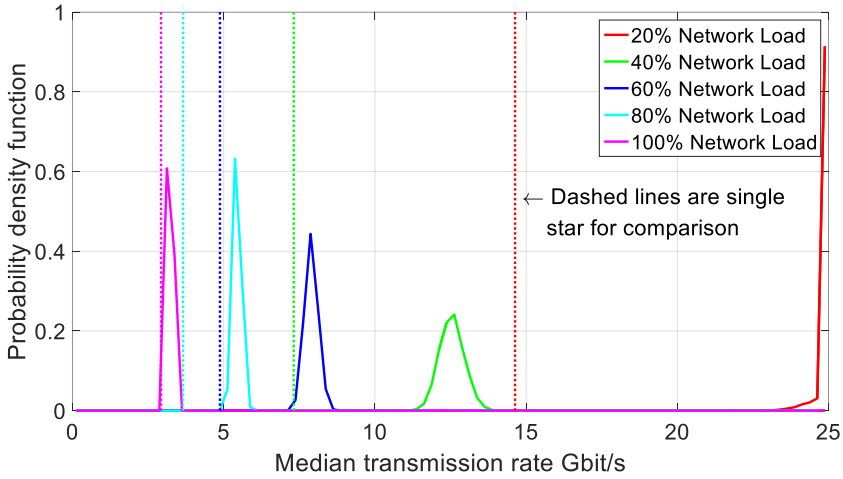

Figure 75: Probability density function (PDF) of the median transmission rate for the ring network topology, with a random traffic distribution.

Figure 76 shows the PDFs of the median bandwidth per source node for a hotspot traffic model over the ring topology. There is an increase in throughput compared to a single passive star for all network loads simulated, however this increase is smaller

than for random traffic. At 100% network load, the increase is only 5.9%, and at 20% network load the increase is only 54.1%. Given that traffic flow around the ring is unidirectional, and that the hotspot nodes are co-located, all flows from hotspot sources must use the same link, causing congestion. It is assumed that the most congested link limits the bitrate of all flows around the network. The hotspot traffic scenario is not as well suited to the ring topology as the other two traffic patterns studied here.

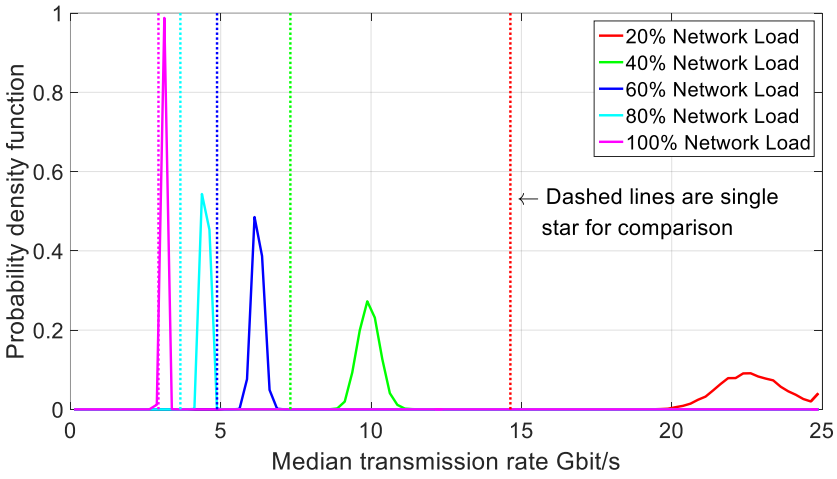

Figure 76: Probability density function (PDF) of the median transmission rate for the ring network topology, with a hotspot traffic distribution.

For zonal media production traffic across the ring topology, there is an increase in median transmission rate compared to a single passive star, except for 100% network load. Figure 77 shows that for low network loads the median transmission rate per source is similar to the random traffic scenario, displaying increases over a single passive star network of 67.1% and 68.7% at network loads of 20% and 40% respectively. However, for 100% network load there is no increase at all in median transmission rate compared to the single passive star network.

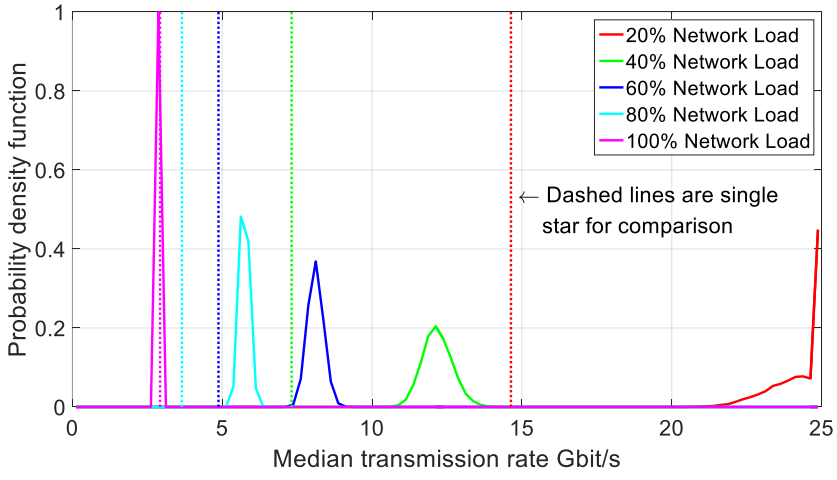

Figure 77: Probability density function (PDF) of the median transmission rate for the ring network topology, with a zonal media production traffic distribution.

4.4. Discussion and conclusions

This chapter has shown that it is possible to meet the design criteria for live media data centre networks by building optical networks from passive couplers on-demand. The topologies analysed in this chapter can provide the same connectivity as a passive star while retaining the multicast and in-cast properties of a single large star network.

For the target node count of $N = 1024$, only the centred star or mesh networks are potentially viable for use with the WS-TDM network design, as the optical path losses across the other topologies are too high. The ring topology in particular is excluded from further analysis, due to the excessive optical power losses of 65-400 dB calculated here. Chapter 2 showed a tolerable loss budget across the WS-TDM network of only ~35 dB at 25 Gbit/s, although low complexity FEC could increase this figure by a further 10 dB, albeit at the cost of a 6% overhead on data throughput [153]. Only the centred star or mesh networks have optical path losses of less than 45 dB.

The centred star is preferable to the mesh for two reasons: the loss budget is the same between all pairs of nodes (i.e. minimum = maximum), so a node receiving in-cast signals from multiple transmitters will not have to tolerate wide variations in received signal power; and no wavelength filtering is required, so no active optical components are required in the network core other than OCS units to connect the couplers.

Considering the achievable gains in transmission rate, particularly at high network loads (80-100%), the traffic pattern to be served across the network has a major impact on whether these split-star network designs provide any significant gains. For the centred star, mesh and centred mesh topologies, when traffic follows a random pattern without locality, there is no increase in transmission rate per node. However, when traffic follows a hotspot pattern, the transmission rate per node increases by 252% (under 80% network load). Sub-stars can form to efficiently serve the hotspot clustering, creating separated sub-stars for traffic both within and outside the hotspot.

Hotspot traffic is an extreme case of traffic locality, and a realistic model of zonal media production traffic shows more modest improvements when using a sub-star topology over a single large star. A 13.7% improvement in transmission rate per node is expected for the centred star, mesh, or centred mesh topologies at 80% network load.

In conclusion, of all the split star designs tested, the centred star design is the most promising for practical implementation alongside the WS-TDM network concept. The following chapter presents an alternative topology for forming split star networks, while maintaining optimal properties for media-centric connectivity.

5. Reconfigurable star networks by splitting

This chapter presents a second approach to partitioning a passive star network into sub-stars. In chapter 4, sub-stars were grown from zero initial connectivity, with couplers joined together to form stars connecting up to N nodes. As an alternative, sub-stars could be constructed by initially creating a fully connected star network of N nodes. The star network could then be partitioned in the centre to remove connectivity that is not required, while maintaining a low and bounded optical power loss budget across all network paths. The design presented here uses switchable optical attenuation at the core of the network, alongside an efficient control algorithm to always split the star wherever possible. This makes it feasible to always split the network into the maximum number of sub-stars possible and thereby maximise throughput, while still meeting the design constraints outlined in section 4.1.

5.1. High level topology design

Figure 78 shows an example of an $N \times N$ port star network, that has been split at the centre into 2 layers: a layer of $\sqrt{N} \times \sqrt{N}$ "input" couplers, and a layer of $\sqrt{N} \times \sqrt{N}$ "output" couplers. Each input coupler has a direct connection to all of the output couplers, via an optical power switching element (shown as "S" in Figure 78).

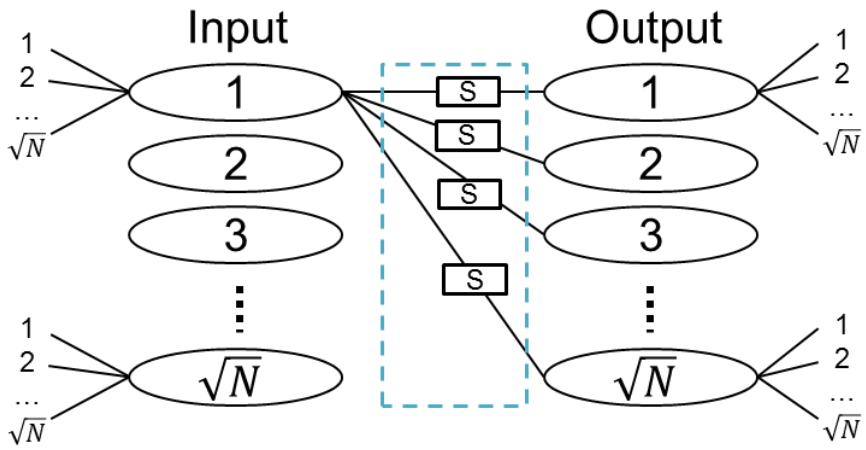

Figure 78: A dual layer split star system connecting a total of N nodes. Each element marked S is an optical switching unit, capable of switching between transmissive and blocking states.

The central power switching elements have two states: blocking (at all wavelengths), or transmissive (at all wavelengths). If all of the central switching elements are set to the transmissive state, then all optical signals in all input couplers will be broadcast to all of the output couplers. The network is then functionally identical to a single large passive star connecting all nodes. However, if some of the switching elements are set to block the optical signal, the network is partitioned into sub-stars. As in chapter 4, each sub-

star can independently allocate wavelengths and timeslots from the full range of the tunable lasers, resulting in increased total throughput across the whole network.

Figure 79 shows an example of how a two-layered star network could be partitioned into sub-stars in different ways. In Figure 79a, all of the input and output couplers require connectivity, meaning that all central optical switches are in the transmissive state. This all-to-all connectivity pattern results in a topology identical to a single passive star connecting all nodes (meeting the design criteria presented in section 4.1 for a single, all-to-all multicast group).

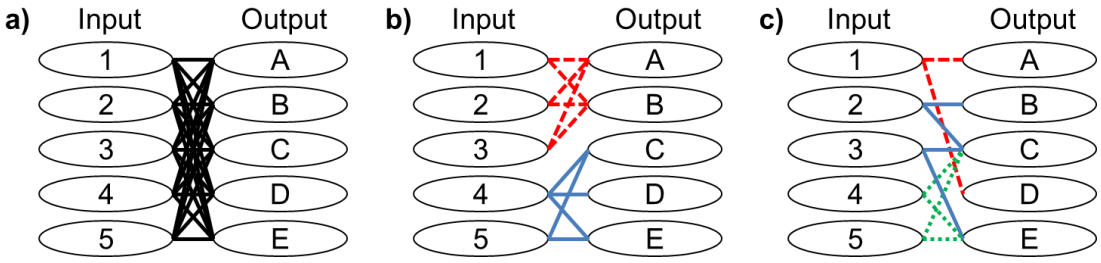

Figure 79: a) An example star network where connections are transmissive between all input and output couplers, which can be abstracted as a single large star; b) the same network but with connectivity enabled only in two distinct sub-stars, allowing full reuse of wavelengths and timeslots in the two groups; c) more complex connectivity pattern across the same network, with splitting into two distinct and disconnected sub-stars feasible.

Should simultaneous all-to-all connectivity not be required, the network can be partitioned into several sub-stars using the central switches. The switches are set to be transmissive only where connectivity is required e.g. for Figure 79b, the connection requests between nodes only require optical transmission between couplers 1-A, 1-B, 2-A, 2-B, 3-A, 3-B,4-C, 4-C, 4-E, 5-C and 5-E. All other switches are set to attenuate light so that the optical signals between input and output couplers are not transmitted. Two sub-stars are then clearly formed: the red, dashed links in Figure 79b form one sub-star, while the blue solid links form another. Given that there is full optical separation of the two sub-stars, the full range of wavelengths and timeslots can be shared in each sub-star independently i.e. the same wavelengths can be used in both sub-stars simultaneously. This effectively doubles the total possible throughput compared to a single passive star connecting all nodes.

Complex network connectivity requests can be analysed so that the network can always be optimally partitioned into sub-stars. Figure 79c shows an example of a complex connectivity pattern through the central switches of the network. Input couplers 2, 3, 4 and 5 all have a connection in common to an output coupler with at least one other input coupler. It is, therefore, not possible to further split the sub-star containing input couplers 2, 3, 4 and 5 due to the mixing of optical signals at the

common output couplers. However, input coupler 1 does not share any output couplers with any of the other input couplers. This means that input coupler 1 and output couplers A and D together form a sub-star, separated from the rest of the network. The full network has again split into two sub-stars, resulting in a potential doubling of total network throughput compared to a single large star.

If a star splits such that the number of nodes connected to each sub-star is less than the number of accessible wavelengths, then all transmitters can operate at the full line rate. However, if the number of nodes in each sub-star is greater than the number of wavelengths, then the network is oversubscribed. The expected transmission rate per node is given by equation 11, but is now calculated independently for each sub-star.

The improvement in network throughput for split stars compared to a single large star is, therefore, entirely dependent on the traffic pattern across the network. It is possible that the required connectivity would force all of the switching units to be in the transmissive state, connecting all input and output couplers together. In this case the network throughput reduces to the worst-case, matching the throughput of a single large star. In practice, considering that the central switches could be switched at the same speed as the transceiver wavelengths and thus on a packet or few-packet timescale, sufficient locality is likely to exist in the traffic to permit splitting the star into multiple sub-stars, even if only instantaneously on a packet-by-packet basis. The maximum possible throughput capacity in this network design occurs if all input couplers do not share any output couplers i.e. each input coupler is connected to only one output coupler. This upper bound on network throughput is given by $\sqrt{N}WB$.

System loss budget is a major consideration for optical star networks. To avoid wasting power by splitting the optical signal more than is necessary, each individual optical coupler should have the same number of output ports as input ports. This means that for a network of N nodes to be served by a dual layer system of couplers, the optimum number of input and output ports on each coupler is $\sqrt{N}$, with $\sqrt{N}$ couplers in both the input layer and output layer. This can be verified by considering the insertion power loss of a single $N \times N$ port coupler, of $10\log_{10}\frac{1}{N}$ dB. The total insertion loss after light has travelled both input and output couplers in this layered design is then found to be $2 \times \left(10\log_{10}\frac{1}{\sqrt{N}}\right) = 20\left(\log_{10}N^{-\frac{1}{2}}\right) = 20(\frac{1}{2}\log_{10}N^{-1}) = \ 10\log_{10}\frac{1}{N}$ dB, which is the same as the insertion loss of a single $N \times N$ coupler. In practice, optical switches incur additional insertion loss even in the transmissive state; the choice made of switching technology is key to minimise these additional losses, as discussed in the next section.

5.2. Optical switching hardware

Careful choice must be made of the hardware which provides switchable optical attenuation at the centre of the star network. An ideal switching technology would provide: high extinction ratio (> 40 dB), so that no optical power can flow between sub-stars when blocking; low insertion loss (< 1 dB) in the transmissive state, so as not to increase total system power loss; and fast switching time (~ 35 ns), so that the central switches can be reconfigured at the same rate as the tunable transceivers.

Semiconductor optical amplifiers (SOAs) meet all of these switching technology requirements. Additionally, SOAs could provide optical power gain at the centre of the star network, which would help to overcome power loss due to the high power splitting ratio. However, they are not suitable for use with WDM signals, due to the non-linear cross-gain modulation effect, whereby channels interfere with each other inside the SOA [154]. Given that the connections between the two layers of couplers each carry a WDM signal, SOAs cannot be used as switches in this design.

Wavelength selective switches (WSS) could be used as switches to create the network partitions. This would allow additional flexibility in allocating wavelengths to traffic flows, since a connection between sub-stars would not need to carry all wavelengths but only those necessary to meet connectivity requirements. However, WSSs are comparatively slow to switch (µs or greater) [9], and current technologies for implementing WSS functionality such as liquid crystal reconfigurable filters incur a high insertion loss (often > 4 dB) [155]. These drawbacks to WSS technology would not permit packet timescale switching, meaning that the partitioning at the centre of the network would have to operate over longer epochs than the edge transceivers.

Instead, the use of Acousto-Optic Modulators (AOMs) is proposed to provide switchable blocking functionality at the centre of the network. AOMs operate over a wide range of wavelengths with both low insertion loss (around 1.5 dB across the full C-band, as shown in Figure 80) and low variation in insertion loss (< 0.03 dB across the full C-band, as shown in Figure 80). The acousto-optic effect can operate on the order of nanoseconds [156], matching the switching times of the transceivers. This permits the centre of the star to switch at the same rate as the tuning speed of the transmitters and receivers, resulting in a packet-scale reconfigurable network fabric.

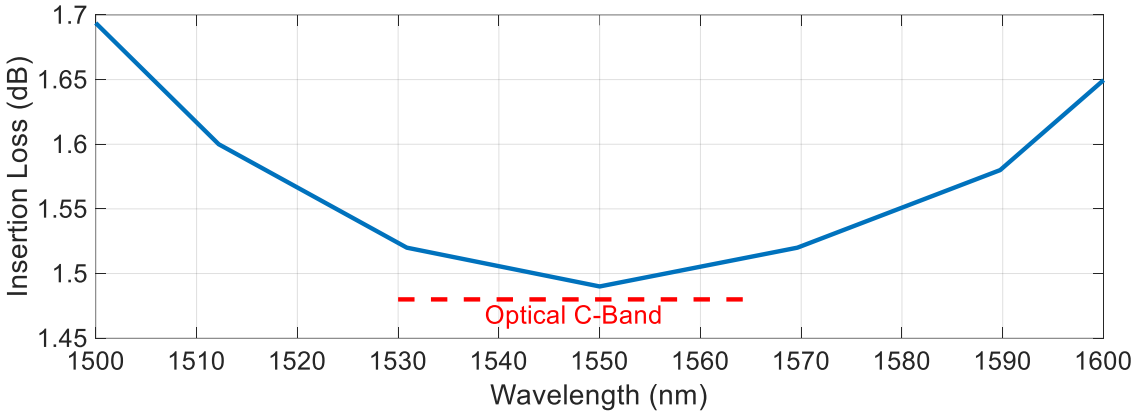

Figure 80: The insertion loss variation with wavelength of an AOM capable of 35 ns switching time between transmissive and blocking states [157]**.**

5.3. Computing the optimum partitions into sub-stars

This section presents an algorithm which finds the sub-stars which can be formed within the network, based on the required network connectivity at any moment in time. A network controller is essential for this network design, to analyse the connectivity requests between all nodes and ensure that the switch states are correctly set to link all nodes that require connectivity. It is not necessary to prescribe whether the network controller is centralised or distributed, but given that the controller must have knowledge of all connectivity across all input and output couplers, a centralised controller which processes all communication requests is an efficient method.

A naïve interpretation of the operation of the switch might assume that there must be no connectivity whatsoever between two (or more) sets of input and output couplers to enable partitioning of the network into two (or more) sub-stars. However, it is only necessary for any two input couplers not to share connectivity to any output couplers, to define that the same wavelengths can be used by transmissions into those input couplers without interference. By assessing whether each pair of input couplers shares connectivity to any output couplers, the partitioned sub-stars can be defined, and in addition, opportunities for wavelength re-use within each sub-star identified. This analysis must be performed before wavelength and timeslot allocation so that each sub-star can independently allocate the full balance of wavelengths and timeslots.

The algorithm seeks one optimisation target, bounded by two operating principles:

Target: maximise the overall pool of wavelengths and timeslots across the whole network, by partitioning the large star into multiple sub-stars.

Operating principles:

1. All flows that request a connection are offered some transmission slots i.e. every data flow request must be granted. If fewer slots are offered to a

transmitter than were requested, the application layer is expected to reduce the flow rate to match the allocation (e.g. send a lower resolution video stream).

2. All flows across the same sub-star are offered equal transmission rates. Although the wavelength and timeslot scheduler could allocate variable bandwidth per node, dependent on priority and/or flow size, this is beyond the scope of this work. This work targets increasing the overall throughput capacity, which is best shown as increases to average flow sizes.

The algorithm then operates as follows (described in pseudo-code in Appendix C):

A. Map the connectivity between input and output couplers, by finding which switches must be set to the transmissive state to accommodate all connectivity requests
B. Find pairs of input couplers that do not have connections in common to any output couplers
C. Disregard any input couplers which are not connected to any output couplers
D. Identify which combinations of input coupler pairs form sub-stars, and any potential sharing of wavelengths within a sub-star

The outputs of the algorithm are:

- A matrix M which denotes the central switches set to the transmissive state;
- A matrix D which denotes the input coupler pairs that do not have any output coupler connections in common;
- A list L which lists all input coupler pairs that are members of any sub-star;
- And a list of sub-stars, where each entry of the list contains a list of input couplers which are members of that sub-star.

The outputs can be passed to the wavelength and timeslot allocation algorithm (such as that in [85], designed for use in a single star system), so that each sub-star can independently allocate wavelengths and timeslots. Whenever the network splits into multiple sub-stars, each sub-star has access to the full W wavelengths, which results in greater total throughput (up to WBS where S is the number of sub-stars formed within the network). The maximum possible number of sub-stars is $\sqrt{N}$, although this can only be achieved if each input coupler connects to only one output coupler.

To present a specific example of the algorithm output, consider the connectivity pattern shown in Figure 81. This pattern is deliberately chosen to demonstrate splitting into two independent sub-stars, wavelength re-use feasible within one sub-star, and ignoring couplers that do not contain any active nodes.

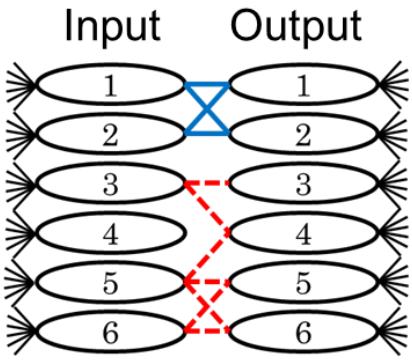

Figure 81: An example connectivity pattern across the split-star network, including two possible sub-stars, possible wavelength sharing within a sub-star, and an input coupler without any active nodes.

Implementing the splitting algorithm (outlined above and described in full detail in Appendix C) on the connectivity pattern shown in Figure 81, gives the following results:

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

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

$$L = [1,3 \quad \sout{1,4} \quad 1,5 \quad 1,6 \quad 2,3 \quad \sout{2,4} \quad 2,5 \quad 2,6 \quad 3,1$$

$$3,2 \quad \sout{3,4} \quad 3,6 \quad \sout{4,1} \quad \sout{4,2} \quad \sout{4,3} \quad \sout{4,4} \quad \sout{4,5} \quad \sout{4,6}$$

$$5,1 \quad 5,2 \quad \sout{5,4} \quad 6,1 \quad 6,2 \quad 6,3 \quad \sout{6,4}];$$

$$sub\text{-}stars\ [1,2];\ [3,5,6];$$

$$sharing\ potential\ within\ sub\text{-}star\ [3,6].$$

Regarding optimality, the algorithm will always find the optimal locations to split the network into sub-stars, in that it will always split the network as many times as possible, without blocking any requested flows. This meets the goal of the algorithm, to maximize the possible pool of transmission slots to be shared across the entire network. This target is constrained by the design principle of always guaranteeing connectivity to any transmitter that requests it. It is always possible to offer all-to-all connectivity if required, as the algorithm could collapse the design to match a single $N \times N$ port star by setting all switches to the transmissive state.

In terms of complexity, algorithm stages A-C must be performed sequentially, in that order, but each stage can be independently parallelised. Stages A-C always take a

deterministic number of operations, scaling linearly with the number of network nodes. The complexity of stage D depends on the length of L and cannot be parallelised. This is because L is entirely dependent on the traffic pattern; in the simplest case L may have zero entries, and in the most complex case L could have a length AB, requiring $2AB$ sequential calculations to be performed.

This algorithm assumes that when a connectivity request is made, the nodes to be connected are in fixed locations, and when a request is made for a connection between two nodes, it is those specific nodes that must communicate. If services at the application layer can run on any one of multiple nodes, and an application can be directed to use a particular node based on its physical location and the overall coupler connectivity, the network performance could be improved. Such load balancing techniques at the application layer are beyond the scope of this thesis, but would be a worthwhile topic for further investigation.

5.4. Flow-level traffic simulations

To assess the potential increase in network throughput that is achievable by splitting the network into sub-stars, flow-level simulations were performed. The traffic models and simulation workflows that were described in section 4.3.1 were also used to produce the results shown in this section i.e. the split star network was simulated for random, hotspot and realistic zonal media traffic flows, over 10,000 simulation trials at each network load percentage. Where a percentage increase in transmission rate relative to the single star network is described, the increase is measured at the mean of the median transmission rate per node i.e. where the integral of the probability distribution function (PDF) is 0.5.

Figure 82 shows the probability distribution functions (PDF) of median transmission rate per node after splitting the network optimally under random traffic connectivity. For all network loads higher than 50% there is no noticeable difference between the performance of the single large star network and the reconfigurable star network. When the network is under low load (network load of 50% or 30%) there is some deviation between the single large star case and the reconfigurable split star case, with an increase in median transmission rate of 3.9% and 10.0% for the 50% and 30% loads respectively.

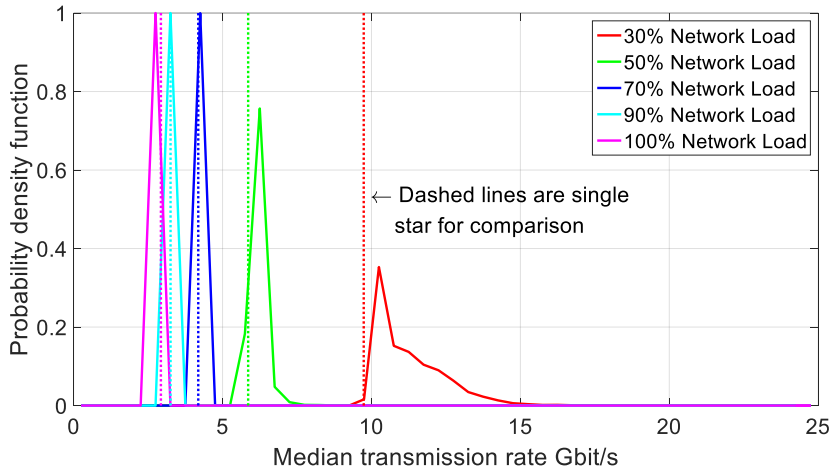

Figure 82: PDF for random traffic over the split-star network, showing the median transmission rate, assuming that the total throughput is shared equally after optimally splitting the star using the central switches.

The connectivity simulation was repeated for the hotspot traffic scenario. The results are shown in Figure 83; there is no advantage to using a reconfigurable star network rather than a single large star for any network load of greater than 50%. At 50% network load, an improvement on the median transmission rate is even less probable than for random traffic, with only an 8% likelihood of the split star network design increasing transmission rate (this is difficult to ascertain from Figure 83 as plotted, but can be found from the cumulative distribution function, not shown here). At 30% network traffic load, there is only a 12.7% increase in the median transmission rate.

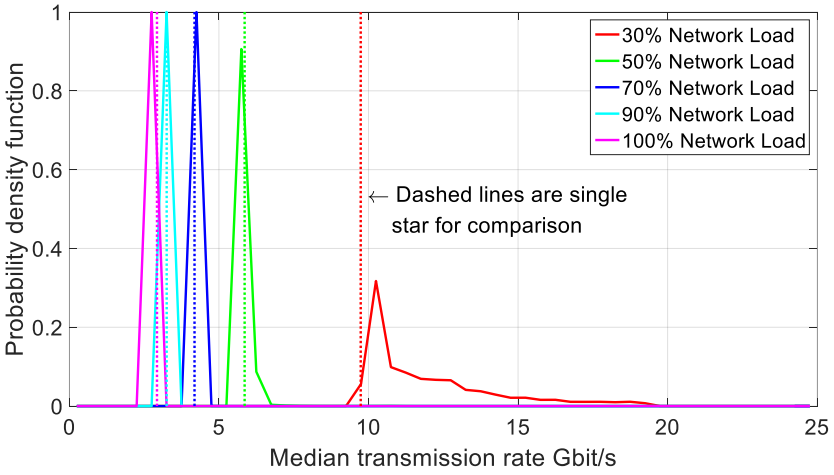

Figure 83: PDF for hotspot traffic over the split-star network, showing the median transmission rate, assuming that the total throughput is shared equally after optimally splitting the star using the central switches.

The split star network design performs worse under hotspot cluster traffic compared to random traffic due to the high likelihood of multiple output couplers connecting to the same input star(s) where the hotspots are located. A marginal improvement could be achieved by spreading the nodes which are members of the “hotspot” over several input couplers, reducing the contention at a single (or few) input couplers which

effectively joins groups together. This could be achieved by physically attaching the hotspot nodes to multiple different input and output couplers, rather than aligning the input and output stars to physical rack structures. However this would require prior knowledge of the hotspot location when cabling the nodes into the network; this is not feasible when launching new services or when demands change over time [19].

The final traffic connectivity scenario simulated over this network design is the zonal media production traffic, and the median transmission rates are shown in Figure 84. This traffic pattern scenario shows the greatest improvement in transmission rate per node for the split star over a single large star. The zonal traffic connectivity pattern makes it likely that traffic will be clustered, since some zones are more likely to communicate than others, as shown in the traffic diagram in Figure 69. For zonal or clustered traffic, if the physical connectivity between nodes and input/output couplers is also aligned as far as possible with particular zones, it is more likely that the network can split into multiple distinct groups, increasing the transmission rate per node.

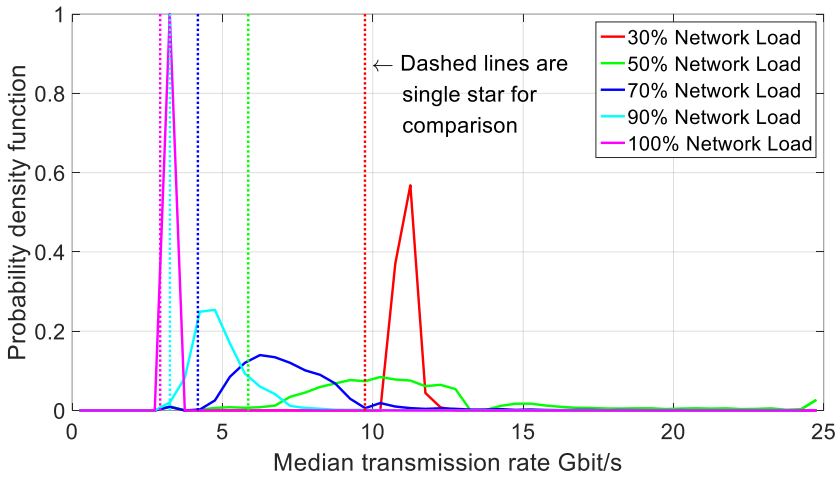

Figure 84: PDF for zonal media production traffic over the split-star network, showing the median transmission rate, assuming that the total throughput is shared equally after optimally splitting the star using the central switches.

Figure 84 shows that for even 90% load, the expected transmission rate increases by 26% compared to that expected from a single large optical star. For 70% load, the expected transmission rate increases by 48% – on average the median bandwidth simulated per port is 6.8 Gbit/s, compared to the 4.6 Gbit/s expected in the case of a single large star.

To emphasise the importance of the traffic pattern on the effectiveness of this network design, the number of sub-stars formed can be compared between the three traffic patterns. Figure 85 shows the mean number of sub-stars that are formed over all 10,000 simulation trials for each traffic scenario and network load. Zonal media production traffic forms a higher number of sub-stars than the other two traffic patterns,

consistently above 2 on average, which in turn results in a higher total network bandwidth which can be shared. For high network loads (greater than 70% for random or 50% for hotspot) the random and hotspot traffic scenarios on average form a single sub-star. This corroborates the data seen in Figure 82, Figure 83 and Figure 84: that there is no discernible benefit to introducing this network design under those traffic scenarios and network loads.

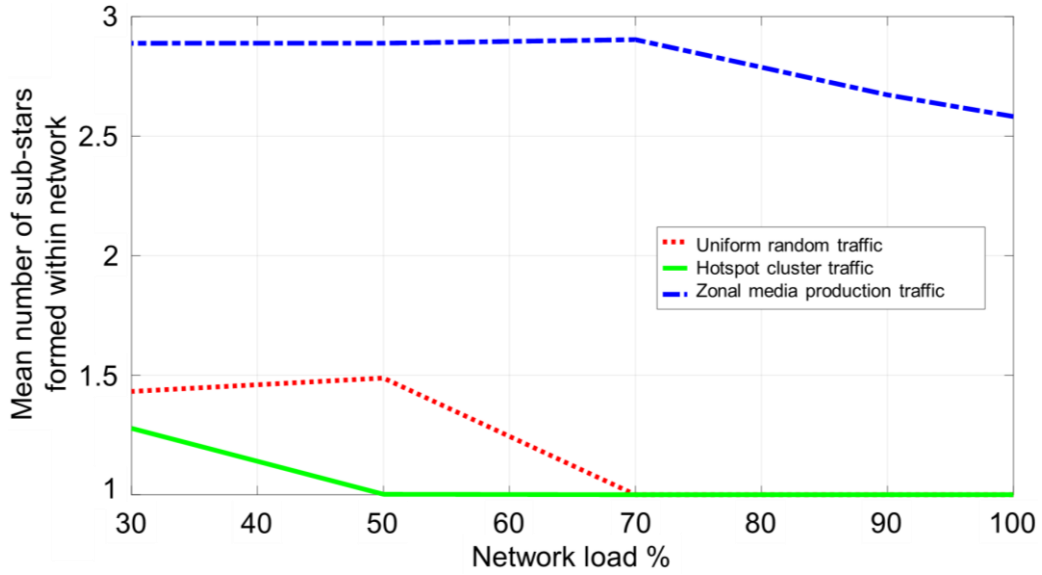

Figure 85: Comparison of the number of sub-stars formed in the network for the three traffic scenarios. Zonal media production traffic always splits into more sub-stars than random or hotspot traffic, regardless of the network load.

Figure 86 shows the mean number of nodes attached to each sub-star within the network. The total bandwidth that can be shared within in each sub-star is identical, as it is known that each sub-star can independently allocate all wavelengths and timeslots. A lower number of nodes per sub-star is thus advantageous, as this means the total available bandwidth is split between a smaller number of nodes, increasing the bandwidth expected by each node.

The zonal media production traffic consistently produces a lower number of nodes per sub-star compared to the other two traffic scenarios. For the zonal media traffic scenario, a mean value of 370 nodes per sub-star (as shown in Figure 86 for 100% network load) means an expected transmission rate of 8.1 Gbit/s per node within the group, an increase of 273% compared to a single, static star. For high network loads (greater than 70% for random or 50% for hotspot cluster) the random and hotspot cluster traffic show 1024 members per sub-star. This is the same as if no splitting of a single large star supporting 1024 nodes has occurred. For a network load of 70%, random traffic again performs slightly better through the network than hotspot cluster traffic, with on average 40.8% fewer nodes per group and thus the same percentage increase in expected bandwidth per node. This is likely due to the hotspot nodes being

in high demand for connectivity, and large numbers of input couplers with connectivity to a few hotspot output couplers.

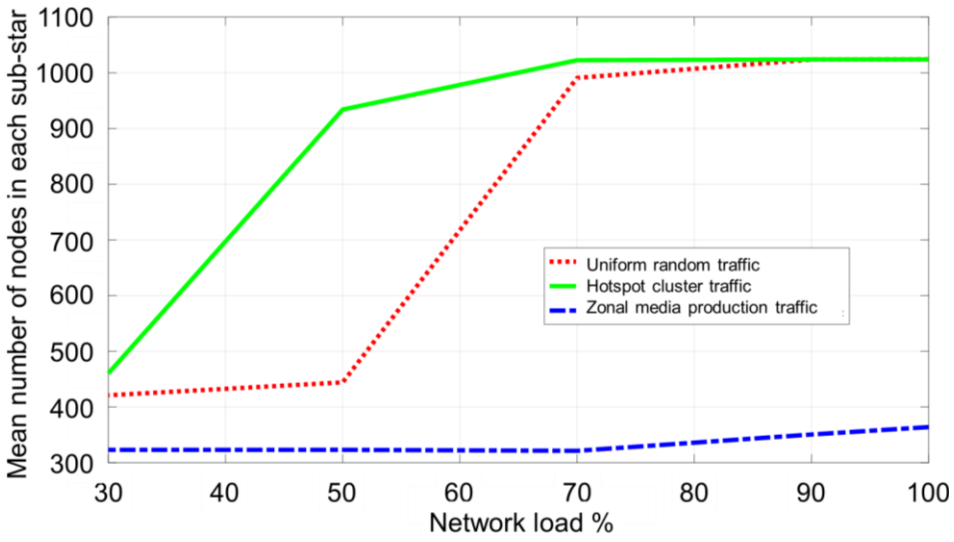

Figure 86: A comparison of the mean number of nodes attached to each sub-star for all three traffic scenarios and network loads. Zonal media production traffic displays a lower number of nodes per sub-star compared to random and hotspot traffic patterns.

5.5. Power consumption analysis

Conventional data centre networks formed from hierarchical layers of electronic packet switches consume vast amounts of energy. Each switch in the network can consume several hundreds of Watts, and given that the port count of each switch is limited, the total number of switches required to span an entire data centre scale network is also high. To ensure full redundancy of the network should any key components fail, multiple redundant switches are often used in key locations; for media production networks every element in the entire network is duplicated and a robust protocol used to establish instant failover in the event of network failure [19]. This results in the total network throughput being doubly overprovisioned, and double the bandwidth of each application layer request being reserved across the network at all times. Such a comprehensive redundancy protocol has high associated costs in capital, complexity and energy consumption, but is necessary to remove all possible single points of failure within the network.

Reducing the total number of switches required in the network topology is one method to design data centre network cores which consume as little energy as possible. In the optical network designs presented in this thesis, far fewer electrical switches are required compared to Clos or Leaf-Spine architectures. Only the control plane of these architectures requires electrical switching, and although sufficient electrical switches must be deployed in the control plane to reach the full port count of N nodes, each

control plane switch can run at a lower data rate than the main data plane. This reduces the total network power consumption drastically due to the lower speed electronics and processing required in the control plane switches.

A comparison is presented in Table 10 of the power requirements of the networks proposed in this thesis, alongside a conventional hierarchical EPS network topology, and a hybrid EPS/OCS network (suitable for limited multicast usage as in [158], although not meeting all of the design constraints for a media production centre).

The total power consumption was evaluated for each of the network architectures, using the parameters of a real world media production centre i.e. 1260 nodes all co-located at a single facility. In this analysis, all packet switched connections in EPS networks are assumed to use active optical cables (AOCs), consuming 1 W per link. The increased complexity of the transceivers for the optical star system (i.e. fast tunable lasers with current driving DACs) is wrapped into a 5.6 W power consumption per transceiver. It is assumed for this power consumption analysis that both of the network topologies presented in both chapters 4 and 5 are using optical MEMS units in the network core to provide reconfigurability of the couplers.

Although there is a 8.8 % increase in power consumption to upgrade from a single passive star network to the reconfigurable switching design presented in chapters 4 and 5, the increased network throughput means that the power consumption per bit transferred will be lower overall. An EPS network would use 49 % more energy than the single passive star network. The comparatively high power consumption of the leaf-spine switches in the incumbent EPS network designs far outweighs the issue of increased transceiver power, resulting in lower overall energy consumption in the all-optical networks.

A hybrid EPS/OCS solution is attractive for providing fast packet switching alongside some limited multicast support. The analysis presented here assumes that a full EPS network is constructed alongside a reconfigurable optical plane for multicast traffic i.e. each node has two independent transceivers, although neither transceiver is fast tunable.

However, the need for both a full EPS network and a series of optical circuit switches (including overprovision in the port count of the OCS to enable greater flexibility in the multicast group sizes from passive splits etc. attached to the OCS) results in high total power consumption. Such hybrid network designs requiring effectively two independent data plane implementations would consume 82 % more power than the single passive star network of chapter 2.

Table 10: A comparison of the power consumptions of the optical networks described in chapters 2-5 with other data centre network topologies.

Single Passive Star Network (Chapter 2)	**Power**	**Reconfigurable Star network (Chapters 4/5)**	**Power**	**EPS Network**	**Power**	**Hybrid EPS/OCS e.g.** [63]	**Power**
CORE:		CORE:		CORE:		EPS Network:	33577 W
46 x control plane packet switch	5980 W	10 x 384 port optical circuit switch	2000 W	2 x 72 port spine switch	7534 W	OCS Network:	6520 W
		46 x control plane packet switch	5980 W	36 x leaf switch	23040 W	(includes 2520 AOC transceivers and 20 OCS units)	
				138 x Leaf-spine AOC	483 W		
ACCESS:		ACCESS:		ACCESS:		SDN Controller:	1000 W
2520 x fast tunable optical transceiver	14112 W	2520 x fast tunable optical transceiver	14112 W	2580 x node-leaf AOC	2520W		
2520 x control AOC	2520 W	2520 x control AOC	2520 W				
TOTAL:	**22612 W**	**TOTAL:**	**24612 W**	**TOTAL:**	**33577 W**	**TOTAL:**	**41097 W**
		Compare to single passive star:	**+8.8 %**	**Compare to single passive star:**	**+48.5 %**	**Compare to single passive star:**	**+81.7 %**

In practice, the reduction in energy consumption will be even greater than this per-device analysis implies, due to the passive nature of optical star components. Each EPS within a data centre network requires cooling, however the passive components in

the optical network core would not require any active cooling. A MEMS or AOM based optical cross connect would require some active heat management, the amount of which is hard to quantify. However, even making a conservative and unlikely estimate that a MEMS fibre circuit switch produces as much excess heat as an EPS, there is a lower total number of MEMS switches and low speed EPSs in a network control plane, compared to the total number of high speed EPSs of an all-electrical network. The overall effect is that less cooling is required in all-optical networks compared to electrical EPS networks, which in turn means a reduced network energy consumption.

5.6. Summary and conclusion

This chapter has shown that by splitting a star network, using attenuating switches between two layers of optical star couplers, the achievable transmission rates per node can be increased compared to a single large star network. This has been achieved using a design which could still provide any-to-any multicast and single-hop connectivity between all nodes, should the traffic pattern require it. In addition, the optical loss budget of a split star network remains acceptable (< 35 dB to support 1024 nodes).

For a realistic network traffic load level of at least 70%, the split-star network shows up to a 40% increase in the throughput per node compared to a single star, but the total achievable throughput is entirely dependent on the network traffic pattern. Traffic in a zonal pattern, such as that observed in a live media production centre, showed greater transmission rates per node compared to random or hotspot traffic; at high network nodes (> 90 %), only the zonal traffic pattern showed an improved throughput per-node compared to a single star network (26 %). It is therefore critical to match the network traffic pattern to the choice of topology.

An advantage to the reconfigurable split-star network is that changes in traffic pattern can be immediately met with changes to the connectivity, to maximise the network throughput. This could have applications in flexible data centre networks where nodes are FPGAs rather than fixed servers and can be reprogrammed to meet application layer demands [159].

Considering the power consumption of four different network topologies which meet the design criteria, the lowest power consumption is found for a single star topology at 23 kW. A 9 % increase in power is required to convert a single star network to a split-star design (due to the central circuit switching elements). However, both of the all-optical topologies discussed in this thesis (single star and split-star) consume vastly less power than an EPS-only network and an EPS-optical hybrid network (34 kW and 41 kW respectively). Energy consumption is one of the largest ongoing costs for data

centre operators, as well as a major environmental concern; by moving away from EPS networks towards all-optical designs, the overall data centre energy consumption requirements could be reduced.

6. Conclusions and future work

The network designs presented in this thesis investigated all-optical switching within data centres to enable multicast and incast traffic patterns. These traffic patterns are often overlooked by conventional network designs, not due to a lack of desire from network operators and applications, but due to incompatibility with current implementations of electrical switching networks. The lack of support for multicast over hierarchical switch networks means that physical layer multicast, such as the capability offered by the optical network designs in this thesis, is a promising solution to provide multicast at data centre scales.

The hybrid WS-TDM network design presented in chapter 2 described a network design capable of connecting over 1024 nodes at 25 Gbit/s, with fully flexible multicast and incast group creation, of group sizes from 1 node up to the full 1024 nodes. By using fast wavelength tunable transceivers and sharing wavelengths using TDM, packet-like switching allowed a total throughput capacity of 2.03 Tb/s to be flexibly shared by all connected nodes. The total capacity figure quoted here incorporates overhead due to laser tuning (200 ns after every 2 µs of data transmission), and overhead due to TDM guard interval (error-free data recovery was observed after only a single bit guard interval between subsequent data packets).

To improve the physical layer performance of the WS-TDM network design, chapter 3 presented enhancements to key transceiver subsystems. DSDBR lasers were selected for use in the network due to their fast wavelength tuning capability, but calibration of the lasers to find optimal tuning currents has previously only been described for optimal use in steady state operation. Chapter 3 described a method of selecting tuning current operating points optimised for fast switching feasible for up to 120 wavelength channels at 50 GHz spacing. By applying pre-emphasis to the current waveforms driving the laser tuning sections, tuning times were measured below 35 ns for switching between any pair of laser wavelengths across 96 channels at 50 GHz spacing.

Having reduced the laser tuning speed to 35 ns, the new methods for laser tuning could be implemented in the WS-TDM network in two ways: to increase the overall throughput across the whole network (potentially to 2.95 Tb/s, assuming 120 channels switching within 35 ns every 2 µs); or to move the network to a truly packet switching model, rather than combining multiple packets into epochs (35 ns network reconfiguration time is comparable to the packet durations of 20-500 ns at 25 Gbit/s).

To further increase the total throughput across a WS-TDM network, chapter 3 also presented line codes suitable for transmitting PAM signals, using only the same transmitter and receiver hardware as the binary data format used in chapter 2. A line

code is essential to remove signal frequencies below 1 GHz, so as to allow TDM as well as wavelength switching. Using PAM4 instead of a binary modulation format could double the potential total throughput to 5.9 Tb/s, due to the increased spectral efficiency of PAM4 modulation. A line code was developed for implementation on generalised PAM symbols which provides 2.23 dB more suppression of signal power below 1 GHz than the IBLC scheme used in chapter 2. It was not feasible to implement the line codes developed in Chapter 3 into a fully functioning optical network due to the relative intensity noise (RIN) of the DSDBR lasers. However, the line codes developed may have applications in data storage or network transmission, or future laser designs with lower RIN may permit line coded PAM to be used alongside fast wavelength switching.

The finite limit to throughput across an optical star network can also be increased by splitting a single large star into multiple sub-stars. By partitioning the wider network into physically separated sub-stars using circuit switching or switchable attenuation, each sub-star can independently allocate the full throughput capacity of a single large star. Chapter 4 described four methods of constructing sub-stars by starting with no connectivity and making connections as the network requests them, while chapter 5 described a method of splitting a fully connected single star into sub-stars wherever possible, to maximise the number of sub-stars formed. All sub-star topologies were analysed to ensure that all-to-all multicast was always feasible, and that network reconfiguration never interrupted existing network traffic flows – key criteria for media production networks.

The traffic pattern across the network has a large effect on the achievable throughput per node of all designs. For instance, the expected throughput per node for the sub-star network design presented in chapter 5 is 48% larger for zonal media production traffic than for hotspot traffic. Given the current trend in data centre design for reconfigurable components rather than fixed, dedicated infrastructure, future data centre traffic patterns will be unpredictable and vary widely on timescales from minutes to days. This trend is compounded by the recent interest in disaggregated data centres, where compute, storage and short term memory units are physically separated within the data centre [160]. Given that hardware resources could be freely allocated as applications request them, the traffic pattern in disaggregated data centres is likely to be even less predictable than in current scenarios. Network designs based on star couplers, which allow flexible sharing of the total throughput capacity between all transceivers, are ideal to support the rapidly changing connectivity requirements.

In terms of the maximum number of network nodes supported by the optical network designs discussed in this thesis, the single star design in chapter 2 experimentally

showed that 1000 nodes is feasible without FEC, and the data plane is only limited by optical power loss in the star splitter. The designs in chapters 4 and 5 incur additional optical loss due to the insertion of optical switching devices. However, the overall loss due to these switches is sufficiently small (< 5 dB total), that if FEC were used in addition to the physical layer design of chapter 2, the network could still communicate error-free (assuming that 5 dB additional optical loss in the network core would result in the signal at the detector being attenuated from -19 dBm to -24 dBm, and an associated change in BER from 10^{-12} to 10^{-8}). Low complexity hard decision FEC such as BCH codes could be used to return BER below 10^{-12}, with only a 7% implementation overhead [153].

An energy consumption analysis comparing the network designs presented in this thesis to existing EPS networks and other network designs in the literature showed that the WS-TDM network consumes 32.7% less power than a hierarchical EPS topology supporting the same node count and throughput capacity. The passive nature of a star coupler network core results in all network power consumption at the transceivers, and even when active OCS units are added to the network core to enable sub-star formation, the resulting topology consumes 26.7% less power than the incumbent EPS hierarchical structures.

The work presented in this thesis has concentrated on the physical layer design of several network topologies for media production data centres, meeting the design criteria set out in chapter 1. However, further improvements are possible and additional challenges remain, as discussed in the next section.

6.1. Future work

6.1.1. Flexible node-coupler allocation

The network topology proposed in chapter 5 allowed the connectivity between input and output couplers to be switched on demand, forming sub-stars dependent on the traffic pattern. However, the allocation of nodes to input and output couplers was fixed, and not flexible to changes in the traffic demand patterns. For example, a set of nodes that are physically adjacent (e.g. within the same rack or cluster of racks) may not all frequently request connectivity to the same nodes, but may be attached to the same optical coupler(s). It is therefore suggested to add optical circuit switching to the outside of the split-star network design of chapter 5. This allows the best of both designs from chapters 4 and 5: flexibility to assign nodes to couplers to match the traffic pattern, with bounded and tolerable optical power losses.

The central core of a more flexible network design remains the same as that in chapter 5: a dual layer system of couplers (input and output), with switchable optical

attenuation elements in between the two layers. However, the node transmitters and receivers no longer have a direct and fixed connection to a coupler, but instead are connected to the input of OCS units. The number of OCS units in the network is defined as T, where each OCS unit is of size $C \times C$ i.e. each OCS unit connects C nodes to the network and $CT \geq N$. The outputs of the OCS units are then connected to the input couplers. For the largest flexibility in connecting nodes to couplers, each OCS has connectivity to all $\sqrt{N}$ input couplers; each OCS therefore has $\frac{C}{\sqrt{N}}$ links (rounded down to an integer) to each coupler. This network design is shown in Figure 87.

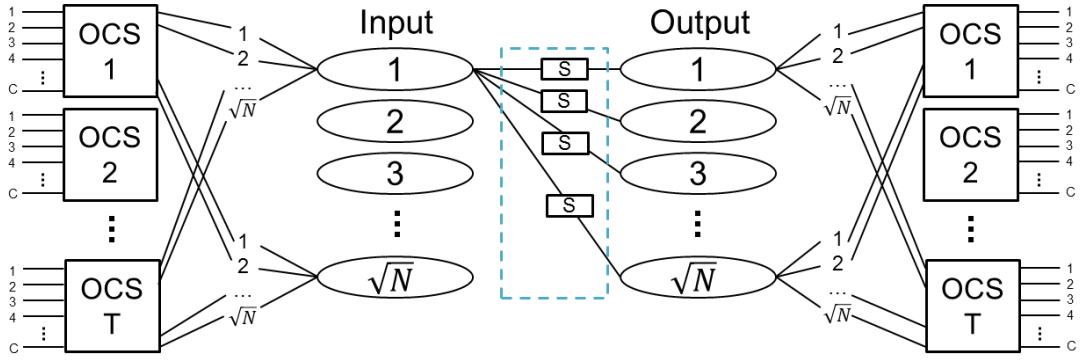

Figure 87: The split-star network with outer optical circuit switch (OCS) units, allowing flexibility in the allocation of nodes to couplers.

This network design therefore offers the best attributes of the previous designs while permitting locality of the traffic pattern to be exploited. The OCS units on the outside of the network can be reconfigured on a slow timescale, such as seconds or minutes, to match the long-term traffic pattern to the physical allocation of nodes to couplers. The central switches are switchable at packet level (ns) timescales if formed from AOMs.

If the traffic flow pattern changes at packet timescales such that the allocation of nodes to couplers via the OCS switches is no longer optimal, this only makes the network less efficient, but does not limit connectivity. Reconfiguration of the outer OCS units to better match the desired connectivity pattern would interrupt existing network flows and thus not meet the criteria, but sub-star partitioning could always be reconfigured using the central switches without affecting existing traffic flows. Optimisation of the parameters C and T for maximal throughput could reveal a trade-off between total network throughput and network cost or energy consumption.

When a node requests a new flow across this new network design, the choice of which input coupler that the node connects to can be influenced by two aims for the network:

- To split the network into as many sub-stars as possible via the central switches;
- To target W nodes per sub-star, so that each sub-star avoids oversubscription.

Meeting both of these aims requires a holistic view of the network including the switches at the centre of the coupler layers and the outer OCS units simultaneously. A possible algorithm aiming to meet the two aims is outlined in Figure 88 – this algorithm is not claimed to be optimal but is a candidate for future development.

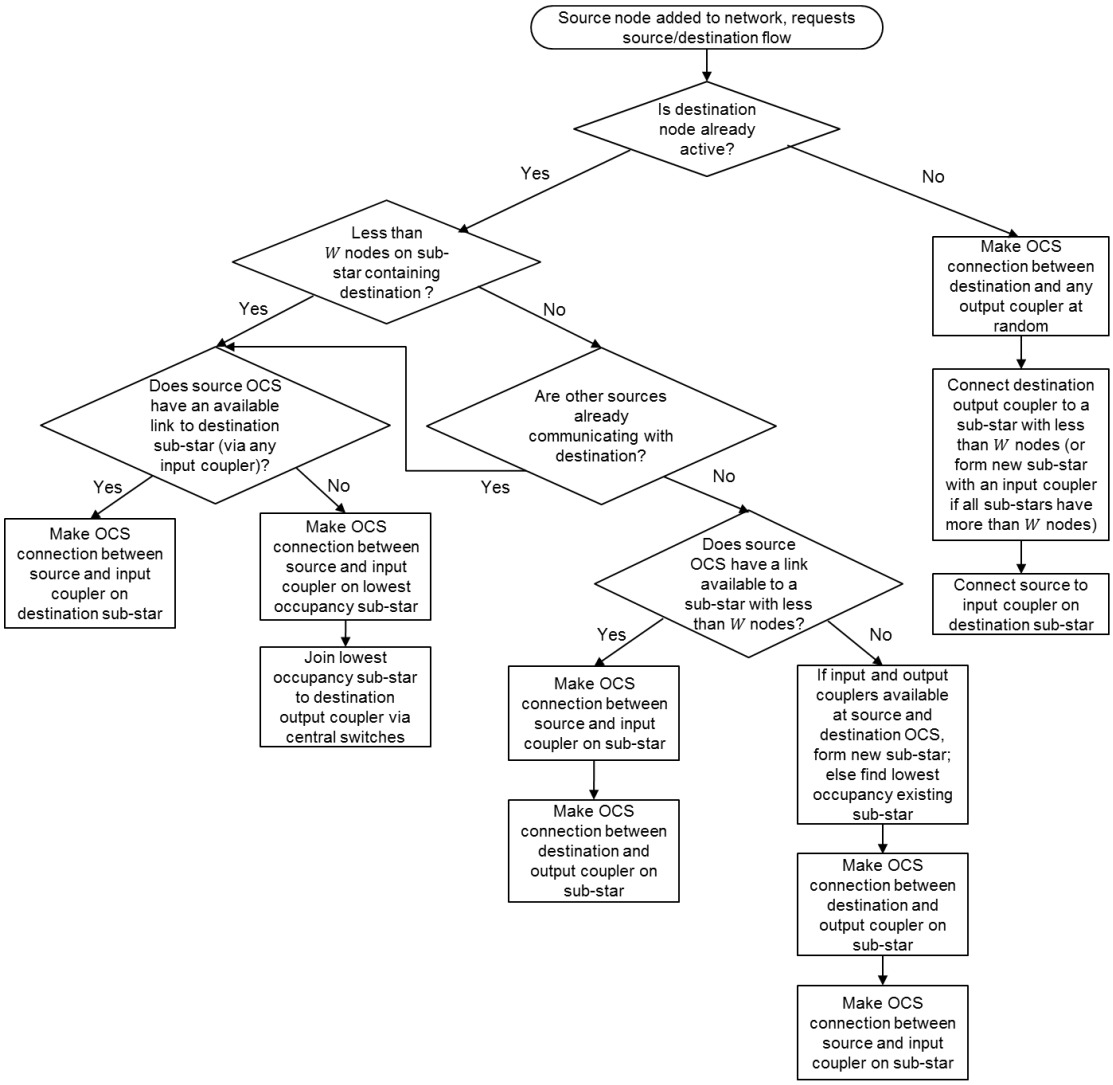

Figure 88: Flow chart of an algorithm to determine how to connect a source and destination via the outer OCS split star network.

Further future work could consider the pre-allocation of nodes to couplers, using a precomputed schedule assigned by a central controller. The schedule could be optimised based on past traffic demands, and machine learning could be used to predict the likely future demands and optimise the network topology to meet them.

6.1.2. Network control plane

A key challenge to network scalability, not considered so far in this thesis, is the network control plane. When scheduling transmissions across shared media (such as an optical star), all requests from nodes must be collected by a controller, schedules constructed, and grants returned to nodes without undue delay. Considering the

timescale suggested in chapter 2 of a 2 µs epoch (limited by only allowing laser tuning time to impose a maximum of 10% overhead), it is feasible to schedule transmissions with only a single epoch delay for a 1000 node network [97]. However, any increase in the number of nodes would reduce the optimality of the schedule solution that the controller could offer [97] – quantification of the reduction in optimality is left for future analysis.

A further challenge in operating a network control system for an all-optical network lies in the transparent nature of the network core between transmitter and receiver. It is essential for the network to know the physical location of all nodes when using optical circuit switching, so that physical connections can be made between servers. In an all-optical network, there are no electrical switches in the path of the data through the network core, and it is not possible for the main network data plane to create a connectivity map through the interrogation of neighbouring nodes. This means that standardised network discovery protocols to discover network topology by mapping neighbours, such as the simple network management protocol (SNMP) [161] and proprietary mapping algorithms such as Cisco discovery protocol (CDP), are only feasible across an electrically switched network control plane.

Networks with both static connectivity and static locations of nodes can overcome the problem of topology discovery, by defining a fixed network map when the network is commissioned. However, when nodes are connected and removed dynamically (as is often the case in media production or flexible optical networks), the network topology frequently changes.

A possible physical layer design which provides out-of-band control plane signalling (so as not to reduce data throughput by using the main network bandwidth for control data) while minimising cabling complexity within the data centre is shown in Figure 89. A separate wavelength band is used for control plane communications (e.g. 1310 nm), compared to the data plane, which operates at a range of tunable wavelengths across the C-band. Both wavelength bands could share a single fibre connection from individual nodes to gateways, which could use passive filters to multiplex and demultiplex the signals.

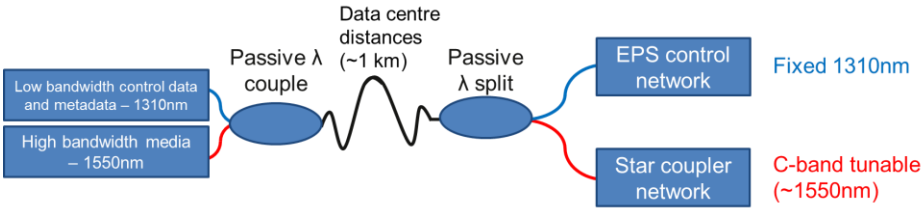

Figure 89: A possible network architecture with a distinct control plane and data plane operating in different wavelength bands.

The design shown in Figure 89 also solves the problem of identifying the physical connectivity of the nodes to the optical network; a one-to-one mapping can be created between each port of the EPS control network and each port of the optical network. This would require some customised protocols such that a single EPS port (which would normally support bidirectional communication over multicore copper cable) can provide identification for both input and output optical network ports.

A further challenge in realising this design lies in the additional power loss that a filter system may introduce, of the order of 1-3 dB per filter, and requiring filters in both the transmitters and receivers. Chapter 2 showed that the node count of an optical passive star network is limited only by the star coupler power loss and receiver sensitivity. This means that any additional optical power loss due to filters at the edges of the network would limit the total node count further.

6.1.3. Connecting Star Networks

The network designs in this thesis all describe a single, self-contained network, and assume that the end nodes attached to the network are each individual servers (or top-of-rack switches). No work to date in this thesis or in prior literature has considered the joining of a star network to another network (star or otherwise) over a long distance (e.g. 40 km or more), where amplification would be required.

Connecting star networks at a distance presents two key challenges: timing synchronisation and optical noise arising from amplification. The synchronisation of clocks between remote nodes can be performed using global positioning system (GPS) clock receivers, or by using protocols such as those described in section 2.2.5.

Optical amplification from erbium doped fibre amplifiers (EDFAs) can increase the power of WDM signals across the entire optical C-band simultaneously. However, due to spontaneous emission, optical noise is added to the signal over the full amplifier bandwidth. In the optical star coupler network presented in this thesis, the full bandwidth of the optical C-band is carried across the entire network, with no wavelength filters to remove noise outside the signal bandwidths from reaching receivers. Without adding filtering, the signal-to-noise ratio of the received signal would be reduced, and the feasible network data throughput lowered.

Future work could explore the feasibility of connecting two star networks in multiple locations via an amplified link, using the application scenario of a broadcast production with video and audio capture from multiple sites simultaneously.

References

[1] T. Benson, A. Akella, and D. A. Maltz, “Network traffic characteristics of data centers in the wild,” in *Proceedings of the 10th Annual Conference on Internet Measurement - IMC ’10*, 2010.

[2] C. Guo, H. Wu, K. Tan, L. Shi, Y. Zhang, and S. Lu, “DCell: A Scalable and Fault-Tolerant Network Structure for Data Centers,” in *Proceedings of the ACM SIGCOMM 2008 conference on Data communication - SIGCOMM ’08*, 2008, pp. 75–86.

[3] A. Ghiasi and R. Baca, “Overview of Largest Data Centers - 802.3bs Task Force, May 2014 Interim Meeting Presentation.” 2014.

[4] Microsoft Corporation, “Microsoft’s Cloud Infrastructure - Datacenters and Network Fact Sheet,” *http://download.microsoft.com/download/8/2/9/8297f7c7-ae81-4e99-b1db-d65a01f7a8ef/microsoft_cloud_infrastructure_datacenter_and_network_fact_sheet.pdf*, 2015. .

[5] Google, “Data center locations - Data Centers - Google.” [Online]. Available: https://www.google.com/about/datacenters/inside/locations/index.html.

[6] Cisco Systems, “Cisco Visual Networking Index: Forecast and Methodology, 2016–2021,” 2017.

[7] A. Singh *et al.*, “Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google’s Datacenter Network,” *Proc. 2015 ACM Conf. Spec. Interes. Gr. Data Commun.*, pp. 183–197, 2015.

[8] J. Koomey, “Growth in Data Center Electricity use 2005 to 2010 (A report by Analytics Press, completed at the request of The New York Times),” 2011.

[9] W. Xia, P. Zhao, Y. Wen, and H. Xie, “A Survey on Data Center Networking (DCN): Infrastructure and Operations,” *IEEE Commun. Surv. Tutorials*, vol. 19, no. 1, pp. 640–656, 2017.

[10] D. Abts, M. R. Marty, P. M. Wells, P. Klausler, and H. Liu, “Energy Proportional Datacenter Networks,” in *ISCA’10*, 2010, pp. 338–347.

[11] J. Dean and S. Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters,” *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[12] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, “The Hadoop Distributed File System,” in *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, 2010, pp. 1–10.

[13] B. Y. M. Zaharia *et al.*, “Apache Spark : A Unified Engine for Big Data Processing,” *Commun. ACM*, vol. 59, no. 11, pp. 56–65, 2016.

[14] J. E. Smith and R. Nair, “The Architecture of Virtual Machines,” *Computer*, vol. 38, no. 5, pp. 32–38, 2005.

[15] A. Berl *et al.*, “Energy-Efficient Cloud Computing,” *Comput. J.*, vol. 53, no. 7, pp. 1045–1051, 2010.

[16] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, “A Survey on Internet of Things: Architecture, Enabling Technologies, Security and Privacy, and Applications,” *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1125–1142, 2017.

[17] C. Clos, “A Study of Non-Blocking Switching Networks,” *Bell Syst. Tech. J.*, vol. 32, no. 2, pp. 406–424, 1953.

[18] C. E. Leiserson, “Fat-Trees: Universal Networks for Hardware-Efficient

Supercomputing,” *IEEE Trans. Comput.*, vol. c–34, no. 10, pp. 892–901, 1985.

[19] A. Roy, H. Zeng, J. Bagga, G. Porter, and A. C. Snoeren, “Inside the Social Network’s (Datacenter) Network,” in *SIGCOMM 2015 - Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, 2015, pp. 123–137.

[20] M. Al-Fares, A. Loukissas, and A. Vahdat, “A scalable, commodity data center network architecture,” *ACM SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 4, p. 63, 2008.

[21] A. Greenberg *et al.*, “VL2: A Scalable and Flexible Data Center Network,” *SIGCOMM 2009 - Proc. 2009 ACM Conf. Spec. Interes. Gr. Data Commun.*, pp. 51–62, 2009.

[22] F. Yan, W. Miao, H. Dorren, and N. Calabretta, “On the cost, latency, and bandwidth of LIGHTNESS data center network architecture,” *2015 Int. Conf. Photonics Switch. PS 2015*, vol. 1, no. Topic 4, pp. 130–132, 2015.

[23] N. Binkert *et al.*, “The role of optics in future high radix switch design,” *ACM SIGARCH Comput. Archit. News*, vol. 39, no. 3, p. 437, 2011.

[24] L. Montalvo, G. Macé, C. Chapel, S. Defrance, T. Tapie, and J. Le Roux, “Implementation of a TV Studio Based on Ethernet and the IP Protocol Stack,” in *2009 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, BMSB 2009*, 2009.

[25] J. Hudson and E. Frlan, “Towards a Hierarchy of SDI data Rates,” in *SMPTE 2013 Annual Technical Conference & Exhibition*, 2013.

[26] L. Sliwczynski, M. Lipinski, P. Krehlik, and A. Wolczko, “Fiber-optic Transmission of SMPTE 259M and SMPTE 292M SDI Signals,” *SMPTE J.*, pp. 213–220, 1999.

[27] A. Bond and S. Pirritano, “Optical SDI Networks : Evaluating Robustness in Your SDI Network,” *SMPTE Motion Imaging J.*, vol. 120, no. 7, pp. 44–48, 2011.

[28] R. Conrod, “The Convergence of Networking and Broadcasting,” *SMPTE J.*, vol. 104, no. 12, pp. 779–787, 1995.

[29] H. Hoffmann, “Evolution in Studio Interconnectivity,” *SMPTE Motion Imaging J.*, vol. 121, no. 3, pp. 10–11, 2012.

[30] H.-R. Shao, “Video Data Rate for HEW (A report for the IEEE802.11 High Efficiency WLAN Study Group),” 2013.

[31] T. Edwards and M. Bany, “Elementary Flows for Live IP Production,” *SMPTE Motion Imaging J.*, vol. 125, no. 2, pp. 24–29, 2016.

[32] P. J. Brightwell, J. D. Rosser, R. N. J. Wadge, and P. N. Tudor, “The IP Studio,” *SMPTE Motion Imaging J.*, vol. 123, no. 2, pp. 31–36, 2014.

[33] IEEE, “IEEE Standard 1588-2008, IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems.” 2008.

[34] P. Loschmidt, R. Exel, A. Nagy, and G. Gaderer, “Limits of synchronization accuracy using hardware support in IEEE 1588,” in *2008 IEEE International Symposium on Precision Clock Synchronization for Measurement, Control and Communication, ISPCS 2008, Proceedings*, 2008.

[35] Z. Kurtisi, X. Gu, and L. C. Wolf, “Enabling Network-centric Music Performance In Wide-area Networks,” *Commun. ACM*, vol. 49, no. 11, pp. 52–54, 2006.

[36] E. Calverley, “Time-Compensated Remote Production Over IP,” *SMPTE Motion*

Imaging J., vol. 127, no. 5, pp. 51–57, 2018.

[37] B. B. Corporation, “Research & Development White Paper: Covering the Glasgow 2014 Commonwealth Games using IP Studio,” 2015.

[38] SMPTE, “RP168-2009: Definition of Vertical Interval Switching Point for Synchronous Video Switching,” 2009.

[39] X. Li and M. J. Freedman, “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.

[40] C. Diot, B. N. Levine, B. Lyles, H. Kassan, and D. Balendiefen, “Deployment Issues for the IP Multicast Service and Architecture,” *IEEE Netw.*, vol. 14, no. 1, pp. 78–88, 2000.

[41] H. Wu, Y. Ding, C. Winer, and L. Yao, “Network security for virtual machine in cloud computing,” *2010 5th Int. Conf. Comput. Sci. Converg. Inf. Technol.*, no. 60803057, pp. 18–21, 2010.

[42] I. Chlamtac, A. Ganz, and G. Karmi, “Purely Optical Networks for Terabit Communication,” in *IEEE INFOCOM*, 1989, pp. 887–896.

[43] L. P. Barry, J. Wang, C. Mcardle, and D. Kilper, “Optical Switching in Data Centers: Architectures Based on Optical Circuit Switching,” in *Optical Switching in Next Generation Data Centers*, 2018, pp. 23–45.

[44] HUBER+SUHNER Polatis, “Product Datasheet: Series 7000n Network Optical Matrix Switch.” 2017.

[45] G. Porter *et al.*, “Integrating microsecond circuit switching into the data center,” *Proc. ACM SIGCOMM 2013 Conf. SIGCOMM - SIGCOMM ’13*, p. 447, 2013.

[46] Y. Chen, C. Qiao, X. Yu, and C. Science, “Optical Burst Switching (OBS): A New Area in Optical Networking,” *IEEE Netw.*, vol. 18, no. 3, pp. 16–23, 2004.

[47] J. Y. Wei and R. I. McFarland, “Just-In-Time Signaling for WDM Optical Burst Switching Networks,” *J. Light. Technol.*, vol. 18, no. 12, pp. 2019–2037, 2000.

[48] M. Düser and P. Bayvel, “Bandwidth Utilisation and Wavelength Re-Use in WDM Optical Burst-Switched Packet Networks,” in *IFIP 5th Working-Conference on Optical Network Design and Modelling (ONDM 2001)*, 2001.

[49] S. Diez, R. Ludwig, and H. G. Weber, “All-optical switch for TDM and WDM/TDM systems demonstrated in a 640 Gbit/s demultiplexing experiment,” *Electron. Lett.*, vol. 34, no. 8, p. 803, 1998.

[50] J. D. Evankow and R. A. Thompson, “Photonic Switching Modules Designed with Laser Diode Amplifiers,” *IEEE J. Sel. Areas Commun.*, vol. 6, no. 7, pp. 1087–1095, 1988.

[51] H. Wang, A. Wonfor, W. K. A, R. V Penty, and I. H. White, “Demonstration of a Lossless Monolithic 16x16 QW SOA Switch,” in *2009 35th European Conference on Optical Communication*, 2009.

[52] Q. Cheng, M. Ding, A. Wonfor, J. Wei, R. V. Penty, and I. H. White, “The feasibility of building a 64×64 port count SOA-based optical switch,” in *2015 International Conference on Photonics in Switching, PS 2015*, 2015, pp. 199–201.

[53] Q. Cheng, A. Wonfor, R. V. Penty, and I. H. White, “Scalable, low-energy hybrid photonic space switch,” *J. Light. Technol.*, vol. 31, no. 18, pp. 3077–3084, 2013.

[54] N. Kataoka *et al.*, “4K uncompressed streaming over multicast-capable 80 (8λ x 10) Gbps colored optical packet switching network using SOA switch and

stacked OC-label processing," in *OFC/NFOEC Technical Digest. Optical Fiber Communication Conference, 2009.*, 2009, p. OWK3.

[55] S. J. B. Yoo *et al.*, "Rapidly Switching All-Optical Packet Routing System With Optical-Label Swapping Incorporating Tunable Wavelength Conversion and a Uniform-Loss Cyclic Frequency AWGR," vol. 14, no. 8, pp. 1211–1213, 2002.

[56] Y. Yin, R. Proietti, X. Ye, R. Yu, V. Akella, and S. J. B. Yoo, "Experimental Demonstration of LIONS : A Low Latency Optical Switch for High Performance Computing," *Photonics Switch. (PS), 2012 Int. Conf.*, no. February 2015, pp. 1–3, 2012.

[57] Z. Cao, R. Proietti, and S. J. B. Yoo, "Hi-LION: Hierarchical Large-Scale Interconnection Optical Network With AWGRs [Invited]," *J. Opt. Commun. Netw.*, vol. 7, no. 1, pp. A97–A105, 2015.

[58] H. J. S. Dorren *et al.*, "Optical Packet Switching and Buffering by Using All-Optical Signal Processing Methods," vol. 21, no. 1, pp. 2–12, 2003.

[59] J. E. McGeehan, M. C. Hauer, A. B. Sahin, and A. E. Willner, "Multiwavelength-channel header recognition for reconfigurable WDM networks using optical correlators based on sampled fiber Bragg gratings," *IEEE Photonics Technol. Lett.*, vol. 15, no. 10, pp. 1464–1466, 2003.

[60] S. A. Jyothi, A. Singla, P. B. Godfrey, and A. Kolla, "Measuring and Understanding Throughput of Network Topologies," *arXiv:1402.2531v4*, pp. 1–15, 2016.

[61] C. Guo *et al.*, "BCube: A High Performance, Server-centric Network Architecture for Modular Data Centers," in *SIGCOMM 2009 - Proceedings of the 2009 ACM Conference on Special Interest Group on Data Communication*, 2009, pp. 63–74.

[62] S. Yan *et al.*, "Archon: A function programmable optical interconnect architecture for transparent intra and inter data center SDM/TDM/WDM networking," *J. Light. Technol.*, vol. 33, no. 8, pp. 1586–1595, 2015.

[63] P. Samadi, V. Gupta, J. Xu, H. Wang, G. Zussman, and K. Bergman, "Optical multicast system for data center networks," *Opt. Express*, vol. 23, no. 17, p. 22162, 2015.

[64] G. M. Saridis *et al.*, "Lightness: A Function-Virtualizable Software Defined Data Center Network with All-Optical Circuit/Packet Switching," *J. Light. Technol.*, vol. 34, no. 7, pp. 1618–1627, 2016.

[65] H. Wang and T. S. E. Ng, "Rethinking the Physical Layer of Data Center Networks of the Next Decade : Using Optics to Enable Efficient * -Cast Connectivity," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 3, pp. 53–58, 2013.

[66] S. Han, N. Quack, T. J. Seok, M. C. Wu, and R. S. Muller, "Large-scale broadband digital silicon photonic switches with vertical adiabatic couplers," *Optica*, vol. 3, no. 1, p. 64, 2016.

[67] M. S. Goodman, H. Kobrinski, M. P. Vecchi, R. M. Bulley, and J. L. Gimlett, "The LAMBDANET Multiwavelength Network: Architecture, Applications, and Demonstrations," *IEEE J. Sel. Areas Commun.*, vol. 8, no. 6, pp. 995–1004, 1990.

[68] Z. Zhu, S. Zhong, L. Chen, and K. Chen, "Fully programmable and scalable optical switching fabric for petabyte data center," *Opt. Express*, vol. 23, no. 3, p. 3563, 2015.

[69] W. S. Hu and Q. J. Zeng, "Multicasting optical cross connects employing splitter-

and-delivery switch," *IEEE Photonics Technol. Lett.*, vol. 10, no. 7, pp. 970–972, 1998.

[70] H. Du *et al.*, "Separated Unicast/Multicast Splitter-and-Delivery Switch and Its Use in Multicasting-Capable Optical Cross-Connect," *IEEE P*, vol. 21, no. 6, pp. 368–370, 2009.

[71] ARIB, "INTERFACE FOR UHDTV PRODUCTION SYSTEMS (English Translation of Association of Radio Industries and Businesses Working Paper)." 2014.

[72] M. Abe *et al.*, "Development of Super Hi-Vision Compact Cameras and Recording System," *SMPTE Motion Imaging J.*, vol. 120, no. 8, pp. 32–43, 2011.

[73] T. Shiozawa, I. Makita, M. Murakami, N. Shimosaka, and M. Fujiwara, "Optical Video/Audio Signal Distribution Network for a Broadcast Center Using WD/TD/SD Hybrid Multiplexing," *IEEE Trans. Broadcast.*, vol. 45, no. 3, pp. 276–282, 1999.

[74] K. J. Hood *et al.*, "Optical Distribution Systems for Television Studio Applications," vol. 1, no. 516, pp. 680–687, 2001.

[75] T. Hermes, B. Hoen, J. Saniter, and F. Schmidt, "LOCNET-A Local Area Network Using Optical Switching," *J. Light. Technol.*, vol. LT-3, no. 3, pp. 467–471, 1985.

[76] K. Y. Eng, "A Photonic Knockout Switch for High-Speed Packet Networks," *IEEE J. Sel. Areas Commun.*, vol. 6, no. 7, pp. 1107–1116, 1988.

[77] Y. S. Yeh, M. G. Hluchyj, and A. S. Acampora, "The Knockout Switch: A Simple, Modular Architecture for High-Performance Packet Switching," *IEEE J. Sel. Areas Commun.*, vol. 5, no. 8, pp. 1274–1283, 1987.

[78] I. Habbab, M. Kavehrad, and C. Sundberg, "Protocols for very high-speed optical fiber local area networks using a passive star topology," *J. Light. Technol.*, vol. 5, no. 12, pp. 1782–1794, 1987.

[79] Y. Birk, "Power-Efficient Layout of a Fiber-Optic Multistar that Permits Log2 N Concurrent Baseband Transmissions Among N Stations," *J. Light. Technol.*, vol. 11, no. 5/6, pp. 908–913, 1993.

[80] Y. Birk, "Fiber-optic Bus-Oriented Single-Hop Interconnections among Multi-Transceiver Stations," *J. Light. Technol.*, vol. 9, no. 12, pp. 1657–1664, 1991.

[81] Y. Birk, "Power-Optimal Layout of Passive, Single-Hop, Fiber-Optic Interconnections Whose Capacity Increases with the Number of Stations," in *IEEE INFOCOM '93 The Conference on Computer Communications*, 1993, pp. 565–572.

[82] Y. Birk, N. Linial, and R. Meshulam, "On the Uniform-Traffic Capacity of Single-Hop Interconnections Employing Shared Directional Multichannels," *IEEE Trans. Inf. Theory*, vol. 39, no. 1, pp. 186–191, 1993.

[83] M. Irshid and M. Kavehrad, "Distributed Optical Passive Star Couplers," *IEEE Photonics Technol. Lett.*, vol. 3, no. 3, pp. 247–249, 1991.

[84] M. E. Marhic, "Hierarchic and combinatorial star couplers," *Opt. Lett.*, vol. 9, no. 8, pp. 368–370, 1984.

[85] J. L. Benjamin, A. Funnell, P. M. Watts, and B. Thomsen, "A high speed hardware scheduler for 1000-port optical packet switches to enable scalable data centers," in *Proceedings - 2017 IEEE 25th Annual Symposium on High-Performance Interconnects, HOTI 2017*, 2017.

[86] J. L. Benjamin, A. Funnell, P. M. Watts, and B. Thomsen, "A High Speed

Hardware Scheduler for 1000-port Optical Packet Switches to Enable Scalable Data Centers," in *2017 IEEE 25th Annual Symposium on High-Performance Interconnects, HOTI 2017*, 2017.

[87] K. Clark *et al.*, "Sub-Nanosecond Clock and Data Recovery in an Optically-Switched Data Centre Network," in *2018 European Conference on Optical Communication (ECOC)*, 2018.

[88] K. Okamoto, H. Okazaki, Y. Ohmori, and K. Kato, "Fabrication of Large Scale Integrated-Optic N × N Star Couplers," *IEEE Photonics Technol. Lett.*, vol. 4, no. 9, pp. 1032–1035, 1992.

[89] A. D. Ellis, J. Zhao, and D. Cotter, "Approaching the Non-Linear Shannon Limit," *J. Light. Technol.*, vol. 28, no. 4, pp. 423–433, 2010.

[90] L. R. Goke and G. J. Lipovski, "Banyan networks for partitioning multiprocessor systems," *25 years Int. Symp. Comput. Archit. (selected Pap. - ISCA '98*, no. 8, pp. 117–124, 1998.

[91] N. D. Whitbread, A. J. Ward, L. Ponnampalam, and D. J. Robbins, "Digital wavelength selected DBR laser," in *SPIE Photonics West 2003*, 2003.

[92] D. J. Robbins *et al.*, "A high power, broadband tuneable laser module based on a DS-DBR laser with integrated SOA," in *Optical Fiber Communication Conference 2004*, 2004, p. TuE3.

[93] J. E. Simsarian, J. Gripp, S. Chandrasekhar, and P. Mitchell, "Fast-Tuning Coherent Burst-Mode Receiver for Metropolitan Networks," *IEEE Photonics Technol. Lett.*, vol. 26, no. 8, pp. 813–816, Apr. 2014.

[94] A. J. Ward, G. Busico, N. D. Whitbread, L. Ponnampalam, J. P. Duck, and D. J. Robbins, "Linewidth in widely tunable digital supermode distributed bragg reflector lasers: Comparison between theory and measurement," *IEEE J. Quantum Electron.*, vol. 42, no. 11, pp. 1122–1127, 2006.

[95] L. Ponnampalam *et al.*, "Dynamically controlled channel-to-channel switching in a full-band DS-DBR laser," *IEEE J. Quantum Electron.*, vol. 42, no. 3, pp. 223–230, 2006.

[96] B. Puttnam *et al.*, "Burst mode operation of a DS-DBR widely tunable laser for wavelength agile system applications," in *2006 Optical Fiber Communication Conference and the National Fiber Optic Engineers Conference*, 2006, p. OWI86.

[97] L. Ponnampalam *et al.*, "Dynamic control of wavelength switching and shuttering operations in a broadband tunable DS-DBR laser module," in *OFC/NFOEC Technical Digest. Optical Fiber Communication Conference, 2005.*, 2005, p. OTuE3.

[98] R. Maher, D. S. Millar, S. J. Savory, and B. C. Thomsen, "SOA Blanking and Signal Pre-Emphasis for Wavelength Agile 100Gb/s Transmitters," in *17th Opto-Electronics and Communications Conference (OECC 2012) Technical Digest*, 2012, pp. 905–906.

[99] H. Matsuura *et al.*, "Suppression of Channel-by-Channel Variation in Wavelength Switching Time of TDA-CSG-DR Laser," in *OECC/ACOFT 2014*, 2014, pp. 296–298.

[100] R. Cush, A. J. Seeds, C. C. Renaud, M. J. Wale, R. Turner, and L. Ponnampalam, "Simplified wavelength control of uncooled widely tuneable DSDBR laser for optical access networks," in *39th European Conference and Exhibition on Optical Communication (ECOC 2013)*, 2013, no. 1, pp. 1212–1214.

[101] S. H. Lee *et al.*, "Self-Configuring Athermal Tunable DS-DBR Laser for Passive

Optical Networks," *Lasers Electro-Optics Quantum Electron. Laser Sci. Conf. (QELS), 2010 Conf.*, p. CWN5, 2010.

[102] H. Guan *et al.*, "Compact and low loss 90° optical hybrid on a silicon-on-insulator platform," *Opt. Express*, vol. 25, no. 23, p. 28957, 2017.

[103] K. Kikuchi, "Fundamentals of Coherent Optical Fiber Communications," *J. Light. Technol.*, vol. 34, no. 1, pp. 157–179, 2016.

[104] K. Kikuchi and S. Tsukamoto, "Evaluation of sensitivity of the digital coherent receiver," *J. Light. Technol.*, vol. 26, no. 13, pp. 1817–1822, 2008.

[105] G. P. Agrawal, *Fiber-Optic Communications Systems, Third Edition.* 2002.

[106] L. Galdino *et al.*, "Amplification Schemes and Multi-Channel DBP for Unrepeatered Transmission," *J. Light. Technol.*, vol. 34, no. 9, pp. 2221–2227, 2016.

[107] L. G. Kazovsky, "Phase- and polarization-diversity coherent optical techniques," *J. Light. Technol.*, vol. 7, no. 2, pp. 279–292, 1989.

[108] M. Artiglia, R. Corsini, M. Presi, F. Bottoni, G. Cossu, and E. Ciaramella, "Coherent Systems for Low-Cost 10 Gb / s Optical Access Networks," *J. Light. Technol.*, vol. 33, no. 15, pp. 3338–3344, 2015.

[109] S. Faruk, D. Lavery, R. Maher, and S. J. Savory, "A Low Complexity Hybrid Time-Frequency Domain Adaptive Equalizer for Coherent Optical Receivers," *Opt. Fiber Commun. Conf. 2016*, p. Th2A39, 2016.

[110] IEEE, "IEEE Standard 802.3 for Ethernet." 2016.

[111] T. Pollet, P. Spruyt, and M. Moeneclaey, "The BER Performance of OFDM Systems using Non-Synchronized Sampling," in *1994 IEEE GLOBECOM. Communications: The Global Bridge*, 1994, pp. 253–257.

[112] P. Moreira, J. Serrano, T. Wlostowski, P. Loschmidt, and G. Gaderer, "White rabbit: Sub-nanosecond timing distribution over ethernet," in *IEEE International Symposium on Precision Clock Synchronization for Measurement, Control and Communication, ISPCS '09 - Proceedings*, 2009, pp. 58–62.

[113] K. S. Lee, H. Wang, V. Shrivastav, and H. Weatherspoon, "Globally Synchronized Time via Datacenter Networks," *Proc. 2016 Conf. ACM SIGCOMM 2016 Conf. - SIGCOMM '16*, pp. 454–467, 2016.

[114] R. Carter, "Low-disparity binary coding system," *Electron. Lett.*, vol. 1, no. 3, pp. 3–4, 1965.

[115] K. A. S. Immink, "A Survey of Codes for Optical Disk Recording," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 4, pp. 756–764, 2001.

[116] Y. Hsueh, M. S. Rogge, W. Shaw, J. Kim, L. G. Kazovsky, and S. Yamamoto, "Spectral shaping line codes for instant upgrade of existing Passive Optical Networks," in *OFC 2004*, 2004, p. FG2.

[117] P. A. Franaszek, "A DC-balanced, partitioned-block, 8B/10B Transmission code," *IBM J. Res. Dev.*, vol. 27, no. 5, 1983.

[118] O. Ishida, "40/100GbE Technologies and Related Activities of IEEE Standardization," *2009 Opt. Fiber Commun. Conf. Expo. Natl. Fiber Opt. Eng. Conf.*, p. OWR5, 2009.

[119] P. J. Winzer, A. H. Gnauck, G. Raybon, S. Chandrasekhar, Y. Su, and J. Leuthold, "40-Gb/s REturn-to-Zero Alternate-Mark-Inversion (RZ-AMI) Transmission Over 2000 km," *IEEE Photonics Technol. Lett.*, vol. 15, no. 5, pp. 766–768, 2003.

[120] A. Croisier, “Introduction to Pseudoternary Transmission Codes,” *IBM J. Res. Dev.*, vol. 14, no. 4, pp. 354–367, 1970.

[121] E. W. Weisstein, “Erfc,” *MathWorld -- A Wolfram Web Resource. http://mathworld.wolfram.com/Erfc.html*. .

[122] B. Broberg, P.-J. Rigole, S. Nilsson, M. Renlund, and L. Andersson, “Widely tunable semiconductor lasers,” in *LEOS’98: IEEE Lasers and Electro-Optics Society 11th Annual Meeting*, 1998, pp. 151–152.

[123] J. Buus and E. J. Murphy, “Tunable Lasers in Optical Networks,” *J. Light. Technol.*, vol. 24, no. 1, pp. 5–11, 2006.

[124] J. E. Simsarian *et al.*, “Fast switching characteristics of a widely tunable laser transmitter,” *Photonics Technol. Lett. IEEE*, vol. 15, no. 8, pp. 1038–1040, 2003.

[125] J. E. Simsarian, M. C. Larson, H. E. Garrett, H. Xu, and T. A. Strand, “Less than 5-ns wavelength switching with an SG-DBR laser,” *IEEE Photonics Technol. Lett.*, vol. 18, no. 4, pp. 565–567, 2006.

[126] C. K. Chan, K. L. Sherman, and M. Zirngibl, “A fast 100-channel wavelength-tunable transmitter for optical packet switching,” *IEEE Photonics Technol. Lett.*, vol. 13, no. 7, pp. 729–731, 2001.

[127] R. O’Dowd, S. O’Duill, G. Mulvihill, N. O’Gorman, and Y. Yu, “Frequency plan and wavelength switching limits for widely tunable semiconductor transmitters,” *IEEE J. Sel. Top. Quantum Electron.*, vol. 7, no. 2, pp. 259–269, 2001.

[128] J. E. Simsarian *et al.*, “A Widely Tunable Laser Transmitter with Fast, Accurate Switching Between All Channel Combinations,” in *Optical Communication, 2002. ECOC 2002. 28th European Conference on*, 2002.

[129] R. Maher, D. S. Millar, S. J. Savory, and B. C. Thomsen, “Fast Wavelength Switching 100Gb/s Burst Mode Transceiver for Coherent Metro Networks,” in *2012 International Conference on Photonics in Switching (PS)*, 2012.

[130] B. Puttnam, B. C. Thomsen, R. Muckstein, A. Bianciotto, and P. Bayvel, “Nanosecond tuning of a DS-DBR laser for dynamic optical networks,” in *CLEO/Europe - EQEC 2009 - European Conference on Lasers and Electro-Optics and the European Quantum Electronics Conference*, 2009.

[131] J. Canny, “A Computational Approach to Edge Detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, 1986.

[132] International Telecommunication Union, “Spectral grids for WDM applications: DWDM frequency grid - Recommendation ITU-T G.694.1.” 2012.

[133] A. J. Ward *et al.*, “Widely tunable DS-DBR laser with monolithically integrated SOA: design and performance,” *IEEE J. Sel. Top. Quantum Electron.*, vol. 11, no. 1, pp. 149–156, Jan. 2005.

[134] H. Kobayashi, “A Survey of Coding Schemes for Transmission or Recording of Digital Data,” *IEEE Trans. Commun. Technol.*, vol. 19, no. 6, pp. 1087–1100, 1971.

[135] M. R. Aaron, “PCM Transmission in the Exchange Plant,” *Bell Syst. Tech. J.*, vol. 41, no. 1, pp. 99–141, 1962.

[136] E. Agrell, J. Lassing, E. G. Ström, and T. Ottosson, “Gray coding for multilevel constellations in Gaussian noise,” *IEEE Trans. Inf. Theory*, vol. 53, no. 1, pp. 224–235, 2007.

[137] M. Chagnon *et al.*, “Experimental study of 112 Gb/s short reach transmission employing PAM formats and SiP intensity modulator at 13 μm,” *Opt. Express*, vol. 22, no. 17, p. 21018, 2014.

[138] E. Agrell and M. Secondini, “Information-Theoretic Tools for Optical Communications Engineers,” in *2018 IEEE Photonics Conference (IPC)*, 2018.

[139] D. Lavery, R. Maher, D. S. Millar, B. C. Thomsen, P. Bayvel, and S. J. Savory, “Digital coherent receivers for long-reach optical access networks,” *J. Light. Technol.*, vol. 31, no. 4, pp. 609–620, 2013.

[140] L. Tao, Y. Ji, J. Liu, A. Tao Lau, N. Chi, and C. Lu, “Advanced modulation formats for short reach optical communication systems,” *IEEE Netw.*, vol. 27, no. 6, pp. 6–13, 2013.

[141] S. K. Pavan, J. Lavrencik, and S. E. Ralph, “Experimental demonstration of 51.56 Gbit/s PAM-4 at 905nm and impact of level dependent RIN,” in *ECOC 2014*, 2014.

[142] C. J. Smyth, “Nonblocking Photonic Switch Networks,” *IEEE J. Sel. Areas Commun.*, vol. 6, no. 7, pp. 1052–1062, 1988.

[143] X. Zheng *et al.*, “Three-dimensional MEMS photonic cross-connect switch design and performance,” *IEEE J. Sel. Top. Quantum Electron.*, vol. 9, no. 2, pp. 571–578, 2003.

[144] P. De Dobbelaere, K. Falta, L. Fan, S. Gloeckner, and S. Patra, “Digital MEMS for optical switching,” *IEEE Commun. Mag.*, vol. 40, no. 3, pp. 88–95, 2002.

[145] E. Ollier, “Optical MEMS devices based on moving waveguides,” *IEEE J. Sel. Top. Quantum Electron.*, vol. 8, no. 1, pp. 155–162, 2002.

[146] R. Cartwright, “An Internet of Things Architecture for Cloud-Fit Professional Media Workflow,” *SMPTE Motion Imaging J.*, vol. 127, no. 5, pp. 14–25, 2018.

[147] M.-E. Ganbold *et al.*, “A Large-Scale Optical Circuit Switch Using Fast Wavelength- Tunable and Bandwidth-Variable Filters,” *IEEE Photonics Technol. Lett.*, vol. 30, no. 16, pp. 7–8, 2018.

[148] O. Gerstel and S. Kutten, “Dynamic Wavelength Allocation in All-Optical Ring Networks,” in *Proceedings of ICC’97 - International Conference on Communications*, 1997, pp. 432–436.

[149] J. Luo *et al.*, “Performance and energy aware wavelength allocation on ring-based WDM 3D optical NoC,” in *Proceedings of the 2017 Design, Automation and Test in Europe, DATE 2017*, 2017, pp. 1372–1377.

[150] A. Shukla, L. P. Singh, and R. Datta, “Wavelength Assignment in a Ring Topology for Wavelength Routed WDM Optical Networks,” in *Proc. 8th International Conference on Information Technology*, 2005.

[151] R. Ramaswami and G. Sasaki, “Multiwavelength optical networks with limited wavelength conversion,” *IEEE/ACM Trans. Netw.*, vol. vol, no. 914, pp. 6pp744-754.

[152] O. Gerstel, G. H. Sasaki, and R. Ramaswami, “Dynamic Channel Assignment For WDM Optical Networks With Little Or No Wavelength Conversion,” in *Proc. 34th Allerton Conference on Communication, Control, and Computing*, 1996.

[153] G. Tzimpragos, C. Kachris, I. B. Djordjevic, M. Cvijetic, D. Soudris, and I. Tomkos, “A Survey on FEC Codes for 100 G and Beyond Optical Networks,” *IEEE Commun. Surv. Tutorials*, vol. 18, no. 1, pp. 209–221, 2014.

[154] M. Oberg and N. A. Olsson, “Crosstalk Between Intensity-Modulated Wavelength- Division Multiplexed Signals in a Semiconductor Laser Amplifier,” *IEEE J. Quantum Electron.*, vol. QE-24, no. I, pp. 52–59, 1988.

[155] G. Baxter *et al.*, “Highly Programmable Wavelength Selective Switch Based on Liquid Crystal on Silicon Switching Elements,” *Opt. Fiber Commun. Conf.*, no.

Lc, p. OTuF2, 2006.

[156] G. & Housego, “FIBER-Q ® 1550 nm Fiber Coupled Acousto-Optic Modulator,” no. T-M080-0.4C2J-3-F2S Product Data Sheet. .

[157] “Personal commmunication with Gooch and Housego.” .

[158] N. Farrington *et al.*, “Helios: a hybrid electrical/optical switch architecture for modular data centers,” *ACM SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 4, p. 339, 2010.

[159] N. Tarafdar, T. Lin, E. Fukuda, H. Bannazadeh, A. Leon-Garcia, and P. Chow, “Enabling Flexible Network FPGA Clusters in a Heterogeneous Cloud Data Center,” pp. 237–246, 2017.

[160] Y. Yan *et al.*, “All-Optical Programmable Disaggregated Data Centre Network Realized by FPGA-Based Switch and Interface Card,” *J. Light. Technol.*, vol. 34, no. 8, pp. 1925–1932, 2016.

[161] J. Case, M. Fedor, M. Schoffstall, and J. Davin, “Network Working Group RFC 1157: A Simple Network Management Protocol (SNMP).” 1990.

Appendices

A – Algorithm for adding new flows to centred star, mesh, and centred mesh topologies

```
1   if source is active
2       if destination is active
3           if source and destination are on same sub-star
4               do nothing
5           else
6               join together sub-stars of source and destination
7           end
8       else // destination is NOT active
9           add destination to same sub-star as source
10      end
11  else // source is NOT active
12      if destination is active
13          add source to same sub-star as destination
14      else // destination is NOT active
15          if a sub-star exists with fewer than W active sources
16              add source and destination to that sub-star
17          else
18              add source and destination to the same new sub-star
19          end
20      end
21  end
```

B – Algorithm for adding new flows to ring topologies

```
1       for all source-destination pairs
2               find central coupler of source
3               find central coupler of destination
4               determine path from source coupler to destination coupler
5               reserve a flow request on each link on the path
6       end
```

C – Algorithm for splitting a two–layered optical star coupler network into sub-stars

```
1       A = [a1,a2,…,ai]; // input couplers
2       B = [b1,b2,…,bj]; // output couplers
3
4       // Stage I
5       // Map connectivity between input and output couplers.
6       // This stage should produce an i x j connectivity matrix
7       // denoted M, showing 1 where links between input and
        8 // output are required and 0 for no connectivity required.
9
10      for i = 1:length(A)
11              for j = 1:length(B)
12                      if coupler ai∈A requested connectivity to bj∈B
13                      M(i,j) = 1 // connectivity matrix
14                      else
15                              M(i,j) = 0 // connectivity matrix
16                      end
17              end
18      end
19
20       // Stage II
21       // Find input coupler pairs that do not share outputs.
22      // This stage should produce an i x j matrix denoted D,
23      // showing 1 if input coupler i and input coupler j do not
24      // share connectivity to any output couplers, or 0 if input
25      // couplers i and j are connected to one or more of the
26      // same output couplers.
27      // A list, denoted L, is made of (i, j) pairs which have no
28      // shared connectivity to output couplers.
29
30      for i = 1:A
31              for j = 1:A
32                      if maximum(M(i,:)+M(j,:)) < 2
33                              D(i,j) = 1 // no shared output couplers
34                              Append [i,j] to L
```

```
35                      else
36                      D(i,j) = 0 // shared output couplers
37                      end
38              end
39      end
40
41      // Stage III
42       // Remove entries from L where the input coupler
43       // does not have active connections to any outputs.
44
45      for i=1:length(L)
46              if sum(M(i,:)) < 1
47                      discard any entries in L containing i
48              end
49       end
50
51       // Stage IV
52      // Find larger sub-stars from the list of coupler pairs.
53      // Each member of list L is a pair of input couplers (i,j)
54      // which share an output coupler. By definition, input
55      // couplers i and j must be in the same sub-star.
56      // This stage checks connectivity between all possible
57      // pairs of members of L by finding the four
58      // combinations of i and j from each pair of members.
59      // D from stage II can be used to find connectivity
60      // between each combination.
61
62      for i = 1:length(L)
63              for j = 1:length(L)
64                      if D(L(i,1),L(j,1)) == 0
65                              input couplers L(i,1) and L(j,1) are in a sub-star
66                      end
67                      if D(L(i,1),L(j,2)) == 0
68                              input couplers L(i,1) and L(j,1) are in a sub-star
69                      end
70                      if D(L(i,2),L(j,1)) == 0
71                              input couplers L(i,1) and L(j,1) are in a sub-star
72                      end
```

```
73                      if D(L(i,2),L(j,2)) == 0
74                              input couplers L(i,1) and L(j,1) are in a sub-star
75                      end
76              end
77      end
78      // where L(i,1) = the 1st entry of the ith pair of L;
79      // and L(i,2) = the 2nd entry of the ith pair of L
80
81      // Stage V
82       // at this point, all sub-stars have been identified
83       // but some wavelengths can be shared within sub-stars
84
85      for i = 1:length(L)
86              if (both members of L(i,:) are in the same sub-star)
87                      allow the same wavelengths and timeslots
88                      to be used simultaneously from both
89                      input couplers L(i,:)
90              end
91      end
92
93      // where L(i,:) denotes the ith pair of L
```