Optical Space Switches in Data Centers: Issues with Transport Protocols

MDPI

Article

Optical Space Switches in Data Centers: Issues With Transport Protocols

Mamun Abu-Tair [1,†*], Philip Perry [2,‡], Philip Morrow [1,†], Sally McClean [1,†], Bryan Scotney [1,†], Gerard Parr [3,‡] and Md Israfil Biswas [1,]

(1) School of Computing, Ulster University, United Kindgom
(2) Radio and Optical Communications Laboratory, Dublin City University, Ireland
(3) School of Computing Sciences, University of East Anglia, United Kindgom
* Correspondence: philip.perry@DCU.ie

Version January 26, 2019 submitted to Photonics

Abstract: A number of new architectures for data centre networks employing reconfigurable, SDN controlled, all-optical networks have been reported in recent years. In most cases, additional capacity was added to the system which unsurprisingly improved performance. In this study, a generalised network model that emulates the behaviour of these types of network was developed but where the total capacity is maintained constant so that system behaviour can be understood. An extensive emulated study is presented which indicates that the reconfiguration of such a network can have a detrimental impact on TCP congestion control mechanisms that can degrade the performance of the system. A number of simple scheduling mechanisms were investigated and the results show that an on-demand scheduling mechanism could deliver a throughput increase of more than $\sim 50\%$ without any increase in total installed network capacity. These results, therefore indicate the need to link the network resource management with new datacentre network architectures.

Keywords: Optical Space Switch; Transport layer; SDN

1. Introduction

A considerable number of research studies have investigated the network architecture and topologies of Datacenters (DCs). This is mainly motivated by the huge amount of data exchanged within DCs and between different DCs. By 2019, the global DC IP traffic is expected to reach 10.4 ZettaBytes per year [1]. According to Cisco Global Cloud Index, 73.1% of traffic is exchanged within DCs, 18.2% between DCs and the users and 8.7% between DCs and other DCs [1].

Nowadays the most two common network topologies deployed in DCs are the Fat Tree and Leaf-Spine network topologies due to their ability to interconnect large numbers of servers efficiently[2,3]. However, with increasing demands on the computational power of the DC, the DC network is one of the main bottlenecks that affects the overall performance of the DCs. To handle this demand, several research studies have introduced network extensions to the DC backbone network using optical space switches. These extensions deploy optical space switches to directly connect the Top-of-Rack(ToR) switches between each other. This approach provides a set of reconfigurable links that can be used to offload traffic between hot spots in the DC and reduce the load on the DC's backbone network.

1.1. Motivation and Approach

The provision of additional capacity would be expected to improve networking performance in DCs, but there remains a question about whether reconfiguration alone will improve performance. In this paper, we explore how higher layer protocols such as Transmission Control Protocol (TCP) will react to such reconfigurations that may impact the potential performance gains. In this work, we will analyse the performance of a network that uses a hybrid networking approach that uses a standard electro-optical Ethernet network and an additional all-optical network that can be reconfigured using Software Defined Network (SDN) techniques that have been previously demonstrated in [4] and [5].

The related work section of this paper reviews a number of these approaches where the all-optical network provides additional network capacity. In this work, however, the goal is to observe network behaviour and performance when the total capacity of the combined hybrid network remains constant. That is, the all-optical network is comprised of capacity that has been diverted from the electro-optical network. Moreover, the capacity of each of the links in the two segments of the hybrid network are equal so that the performance of each type of network can be fairly compared. Clearly, such an approach would not be used in a real deployment, but is used here to reveal system behaviour.

In particular, the work investigates network behaviour by addressing the following questions:-

- Can the provision of reconfigurable all-optical paths enable better usage of existing network capacity?
- Are there demonstrable benefits from providing an all-optical link to connect two nodes/hosts that have sufficient traffic demands to utilise the link efficiently, rather than contending for bandwidth in the traditional leaf-spine Ethernet network?

In order to explore these questions, the network architecture is abstracted to capture the essential behaviour while keeping the total network capacity constant. That is to say, if a direct network link is created between two ToRs, then a corresponding link is removed from the leaf-spine network.

Our approach is based on measurements of performance of emulated DCs exchanging different sizes of virtual machines. We use open-source software to allow our results to be validated / reproduced easily. Additionally, by using open- source software we make it possible to apply our methodology easily to other similar scenarios.

Possible applications that could benefit for fast network reconfiguration are: VM migration, Hadoop data storage re-optimisation, data mirroring/backup, data movement for disaster mitigation. Here we will consider a number of concurrent VM migrations from a rack in one part of the DC to another rack.

1.2. Contribution and structure of this paper

This paper presents an evaluation study of a Data Centre Network (DCN) that uses reconfigurable all-optical connections between ToR switches. The experiments have been designed carefully to reflect current DCN traffic scenarios that capture and reflect real world situations. Our experiments show that deploying an optical space switch to extend the DCN requires an "application aware" traffic management regime if the performance benefits are to be maximised.

In particular, considerable attention needs to be paid to other network components such as the network's transport layer and flow scheduling mechanisms. This will make sure that the new optical space switch works in harmony with other existing network components. We first present some related work in Section 2. We then describe the experimental design of our study in Section 3. In Section 4 and Section 5, we explain and discuss our results. We conclude with a summary in Section 6.

2. Related Work

The work presented in this paper is necessarily interdisciplinary and this section is therefore organised in three parts: Typical DCN topologies in Section 2.1, DCNs with optical space switch extensions in Section 2.2, and TCP in Section 2.3.

2.1. *Typical DCN topologies*

A DC consists of three main layers: DC foundation, DC services and user services layers [6]. The context of this study is the foundation layer, which deals with the basic network aspects such as: routing, switching and network topology alongside different elements such as computing power and data storage. In terms of network topologies: the DCN topology should provide a seamless connection between the servers, storage devices and interacting users. There are several network topologies for DCNs such as Fat Tree and Leaf-Spine topologies.

The Fat-Tree topology has been shown to be cost effective in terms of low power consumption and heat emission [7]. Additionally, Fat-Tree allows the connected servers to communicate at line speed. However, the Fat-Tree suffers from issues such as bottlenecks and wiring complexity.

The Leaf-Spine topology is currently one of the most widely used DCN topologies. This is due to several reasons such as the ease of adding extra hardware to extend the network without changing the topology. Moreover, Leaf-Spine creates Equal-Cost Multipathing (ECMP) between the peers which leads to a stable communications system with predictable end-to-end propagation delays.

2.2. *Reconfigurable All-Optical DCN*

The use of all-optical switching capabilities for DCNs has been the subject of considerable research attention lately. The ProjecToR project at the Center for Integrated Access Networks (CIAN), University of Arizona [8] explored the possible use of free space optical communications between Top-of-Rack (ToR) switches to provide opportunistic optical links between ToRs to accommodate large flows of data between compute clusters in a DC.

In the Helios system [9], a Micro-Electro- Mechanical Systems (MEMs) switch is used to connect links with higher capacity between ToRs to solve problems of network congestion and over subscription. This approach adds new capacity to the system which will result in better performance. The C-Thru [10] system uses a MEMs switch and explores the performance of the system when it is constrained by normal protocols used in Ethernet switched networks.

The HyDra system described in [4] analyses the delays caused by a Floodlight SDN controller when reconfiguring a DC network with the ability to connect ToR switches via a MEMs Space switch.

The above studies have paid particular attention to the relatively long setup time for these optical paths using MEMs switches which can be of the order of milliseconds. Other work has explored the use of wavelength switching on the order of tens of nanoseconds combined with MEMs switching to enable fast setup of optical interconnects using the set of possible paths established through the MEMs switch [11], [5].

Many other studies have introduced a range of different ways to include reconfigurable paths in a DCN context[12]. In this paper we will not introduce a new architecture, but, rather we will develop a model that is a genaralised abstraction of these types of architectures that use optical space switches.

2.3. *Transmission Control Protocol (TCP)*

TCP is the most widely used Transport protocol on the Internet as it is used for email, file transfer, web access, video streaming and many other applications. The main aim of TCP is to provide end-to-end

reliable data transmission with a *congestion control* mechanism in order to adjust the transmission rate, avoiding end-to-end loss [13].

In the context of this work, TCP's dynamic behaviour impacts on the system performance, thus, a summary of TCP's congestion control behaviour is presented here. As described in [14], TCP operates by sending segments of data. When the segment is received, the recipient responds with an acknowledgment message (ACK) to the sender. TCP is layered on top of the Internet Protocol (IP) which ships the TCP segments between end-systems (clients, hosts) using (IP) packets.

In order to make more efficient use of the end-to-end path between sender and receiver, the sender may transmit more than one segment before it receives an ACK. This number of segments is known as the TCP sender's *Congestion WiNDow (cwnd)*. Careful control of the cwnd is important to maintain high throughput, without causing congesting in the network. TCP senders rely on the receipt of ACKs as a natural 'clocking' mechanism – when a sender gets an ACK it can send more data.

In order to set an optimum size for cwnd, a TCP sender 'probes' for available capacity by sending out a single segment initially, and then doubles the number of segments in the cwnd every time that an ACK is received. Eventually, the time taken to receive an ACK exceeds a threshold and the TCP sender assumes that congestion has occurred and it must reduce cwnd to ease congestion by effectively throttling its sending rate.

In the context of this work, the behaviour means that TCP will have a dynamic transmission rate and can have a large amount of unacknowledged data in the network. If a link is reconfigured (moved) then data can be lost or stuck in a buffer and TCP will wait for a timeout and then begin the slow-start transmission process again."

3. Network Configuration

The goal here is not to present a new network architecture, but rather to explore the interaction of an SDN reconfigurable optical link and the higher layer transport protocols. In particular, if the all-optical part of the network uses fast tunable lasers to implement the wavelength switching, then it becomes possible for these all-optical paths to be shared between nodes in a burst mode. This approach may lead to lower cost network infrastructure. The network architecture studied here is rather similar to the Helios system [9] where an Optical Space Switch (OSS) is used to create circuits between ToR switches.

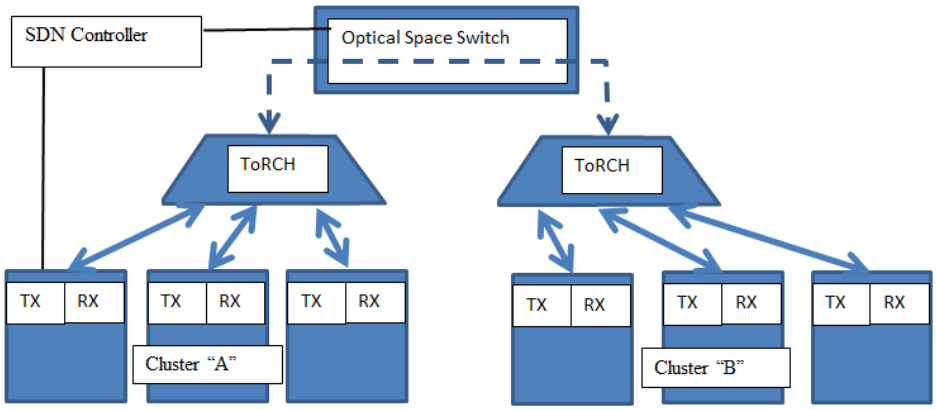


Figure 1. Direct-Optical Connection part of the DCN.

In this paper, though, we consider a case where the ToR switches are equipped with a number of optical interfaces, one of which can be diverted from the normal Leaf-Spine network and directed towards

our new network based on the OSS. This special interface is also equipped with a fast tunable laser that can be set to any of the possible Wavelength Division Multiplexing (WDM) channels in the all-optical network. The ToRs are arranged in clusters, where each ToR Cluster (ToRC) contains a few tens of switches and the OSS can be reconfigured to create a transparent optical path between a pair of these clusters.

A simplified segment of the all-optical network architecture is shown in Figure 1, where six of the ToR switches are organised into two clusters. It is important to note that each ToR has other interfaces that connect only to the Spine switches, so this diagram only shows the special interfaces on the ToR that can be redirected from the Spine switch to the OSS. Each cluster has a ToR Cluster Head (ToRCH) which in this case is an optical coupler which combines the signals on different wavelengths from each ToR and routes them towards the OSS. The OSS is then configured to route the optical signals towards a specific receiver ToRCH. The ToRs contain wavelength a tunable TX and the receiver ToRCH contains an arrayed waveguide grating (AWG), so that a specific TX wavelength will effectively select the receiver ToR within the receiving cluster. The SDN controller sends commands to reconfigure the OSS and also controls the wavelength switching in the ToRs so that any ToR in cluster "A" can communicate with any ToR in cluster "B".

To explore the impact of transport protocols and the usefulness of an SDN controller in such a system, we first abstract the network topology to be represented purely by Ethernet links so that it can be emulated in the Mininet environment [15]. The reconfigurable optical link at each ToR then becomes one of these Ethernet links that can be moved from the Spine and replaced by a direct path to the desired ToR.

This approximation allows us to explore the performance constraints caused by the OSS switching events in isolation by using an idealised (instantaneous) fast wavelength switching mechanism. With the addition of an ideal scheduler and no switching transient, the abstracted architecture can be modelled as a set of logical channels in a Leaf-Spine topology where each source ToR can switch one special link to directly connect to the destination ToR switch as shown in Figure 2.

This study considers two scenarios: the normal Leaf-Spine topology scenario and a Leaf-Spine topology equipped with a direct optical link between the ToRs where access to the direct optical link is shared between multiple ToRs within the ToR cluster rather than giving dedicated access for only one ToR. As shown in the figure the topology consists of 2 spine switches connecting 6 leaf (ToRs) switches together. Each ToR serves 24 hosts (servers) that have different sizes of files/VMs to be transferred using TCP.

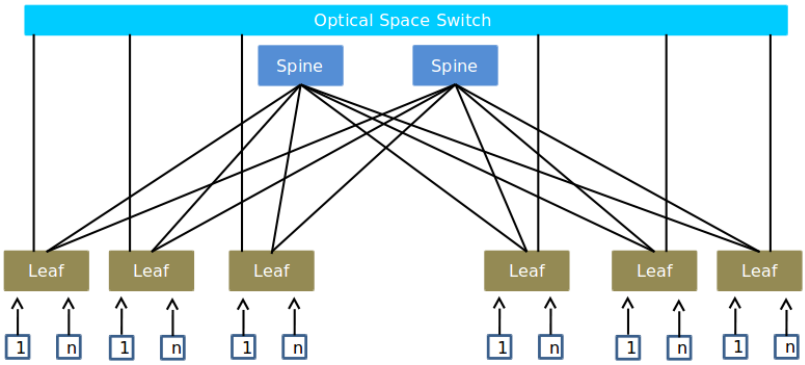


Figure 2. Schematic of emulated network used in the study showing Hosts/Switches connectivity. All Links have a 1Gbps bandwidth limit.

3.1. Scheduling Techniques

A real network implementation would typically use links with 100Gbps links and have OSS reconfiguration events every few tens or hundreds of milliseconds. These events will establish new

All-Optical circuits between a pair of ToRCH's and that circuit will exist for a Space Switch Epoch (SSE) before the next reconfiguration event. During each SSE, the bandwidth of the link is shared between the ToRs within the connected clusters using Burst-Mode (BM) techniques using fast wavelength switching. The abstracted network topology is emulated in Mininet which has a maximum line rate of 1Gbps, so we must scale the SSE to be of the order of seconds.

The basic idea of deploying an OSS is to provide a reconfigurable link between the ToRs to improve the overall performance of the DCN and hence improve the performance of the DC. However, sharing this additional resource among different ToRs in the DCN needs an efficient mechanism. In this study we consider three simple regimes to assign traffic to these new direct links. The first does not differentiate between the traffic flow sizes, while the second case, only elephant flows can access the direct path, but the path is shared between clusters in a time-division regime. The third scheme assigns the direct path to a number of elephant flows and these flows have exclusive access to the path until all flows are completed. In detail, the three allocation regimes are:-

- Undifferentiated Traffic (UT) link sharing: A cluster of racks would like to communicate with another cluster of racks and the physical link is shared in bursts of data exchange on the logical links between individual ToR pairs. The proportion of each SSE that is given to each logical link is a fixed proportion of the epoch. Here the duration of the SSE is explored to assess the impact of this approach on the TCP flows that will be carrying the VM migrations.
- Differentiated Traffic (DT) link sharing with traffic engineering: this is similar to the UT approach above however the all-optical links will only be for the use of very large file/VM transfers (so-called elephant flows). Such a differentiation of flows is well known in the literature [16] and is known to have beneficial impact on system performance.
- On-demand dedicated links: The ideal scheduler instructs the SDN controller to reconfigure the optical links to connect racks that need to transfer large VMs/files, i.e. only elephant flows use the direct link and they use them for the duration of that flow.

The duration of the SSE was varied for the UT scheme and the average throughput of 24 concurrent VM transfers was measured. The results shown in Figure 3 indicate that the maximum throughput is achieved when the SSE is set to 10 Seconds. This value is used throughout the rest of this study.

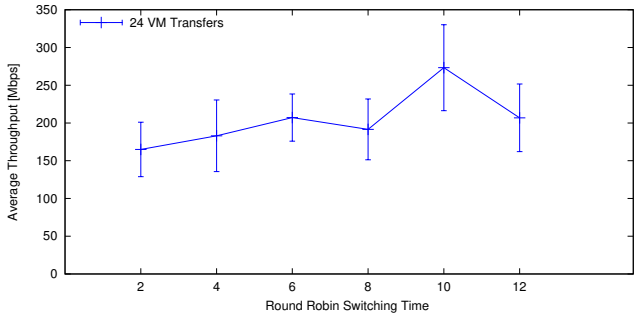


Figure 3. Average Throughput of 24 VM transfers against different switching epochs.

The results presented in this paper are intended to explore the interaction between transport protocols and reconfigurable optical networks under SDN control. The results for these 1Gbps experiments can be scaled to the 100Gbps links by dividing completion times by 100, so that the SSE of 10 seconds for the 1Gbps links can be approximately scaled to 100 ms for the 100Gbps links. If we can assume a MEMs reconfiguration tranient time of the order of 1 ms, then this transient time corresponds to less than 1% of the SSE.

3.2. Workloads: Flow configuration

We measured performance of TCP flows emulating VM transfers transmitted over a small segment of a DCN. This small segment is used to help isolate the behavioural interaction between TCP and the SDN reconfiguration scheme in a well controlled environment. Different VM exchange scenarios have been considered including 24, 48 and 72 simultaneous VM transfers. Two different VM sizes have been used to represent Mice (250 MB) and Elephant (3 GB) VM transfers. The VMs are equally distributed to the three sender ToRs with a ratio of 7 mice VMs to 1 elephant VM.

A packet size of 1460 byte packet is chosen as that is a common size used for Internet-wide communication and to avoid the effects of IP-level fragmentation[17]. Each measurement was performed 15 times for each DC network topology and each scheduling technique used.

3.3. Performance Metrics

We have used a number of directly observed measurements in order to evaluate the performance of the direct optical link topologies in a DCN. For TCP scenarios, the following metrics are considered to be important performance indicators: End-to-end data rate and VM transfer completion time.

4. Results

In this section we provide a comprehensive analysis of the results obtained from our experiments. Firstly, in subsection 4.1, we compare the performance of a DCN with and without the direct optical link using the UT scheduling mechanism. Subsection 4.2 provides analysis of the DCN when the DT scheduling technique with traffic engineering is deployed. Subsection 4.3 provides analysis of the DCN when an on-demand scheduling technique is deployed and subsection 4.4 discusses some limitations of our experiments.

Since variability in performance can be high with TCP traffic, each experiment is repeated 15 times and the results graphs show the following statistics drawn from those 15 experiments:

- All 15 measurements summarised as a standard boxplot (minimum whisker, 25th-percentile, median, 75th-percentile, maximum whisker).
- Offset to the right of the boxplot, for each set of 15 measurements, we plot a point for the mean value, with a whisker showing the 95th-percentile and 99th-percentile. It is worth noting that very small 95th-percentile and 99th-percentile were calculated in some cases, so they may not always be easily visible even though they have been plotted.

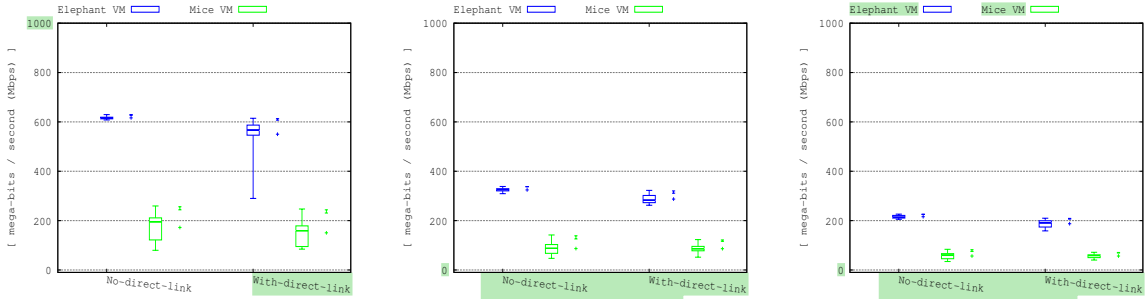
4.1. Undifferentiated Traffic Scheduling Results

This subsection discusses the results obtained from the network emulator using a UT scheduling mechanism, where the ToRs equally share access to the direct optical link connecting the ToRC's in a round robin basis.

When a ToR has been granted access to the direct optical link, all its traffic will go through this direct optical link.

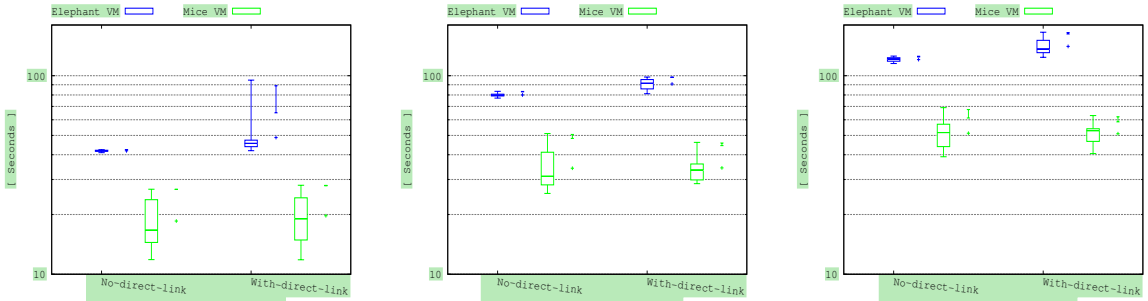
The results shown in Figures 4 and 5 show the throughput and the transfer completion time for both mice and elephant VMs in 24, 48 and 72 VM transfers.

As shown in the figures, the elephant VM transfers always performed worse in the case of deploying the direct optical link extended topology compared to the DCN without the direct optical link extension. In the mice VM transfers scenarios, the performance is almost the same as the error bars of the two cases overlap. A detailed inspection of the results indicated that this behaviour was due to packets being held in a buffer for a link that had been removed due to the movement of links in the reconfigurable



(a) 24 VM transfers **(b)** 48 VM transfers **(c)** 72 VM transfers

Figure 4. Throughput for elephant and mice VMs in the 24, 48 and 72 VM transfers scenarios. To the right of each boxplot, we show the mean and a whisker marking the 95th and 99th percentiles.



(a) 24 VM transfers **(b)** 48 VM transfers **(c)** 72 VM transfers

Figure 5. Completion time for elephant and mice VMs in the 24, 48 and 72 VM transfers scenarios. To the right of each boxplot, we show the mean and a whisker marking the 95th and 99th percentiles.

network case. This resulted in repeated slow-start TCP behaviour on both types of flows, which had an disproportionately severe impact on the elephant flows as stalled transmission of the mice flows could cause additional delays to the elephant flows.

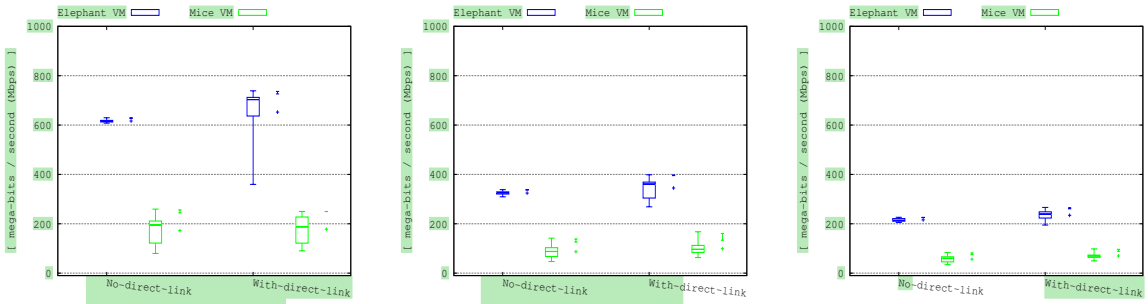
4.2. Differentiated Traffic Scheduling Results

The DT scheduling mechanism distinguishes between the mice and elephant flows and assigns only elephant flows to the direct all-optical link while the mice flows are directed to the conventional Ethernet links. The Elephant flows from the various hosts in each rack are given access to the all-optical network in a round robin manner.

Figures 6 and 7 show the throughput and the completion time of the 24, 48 and 72 VM transfer scenarios.

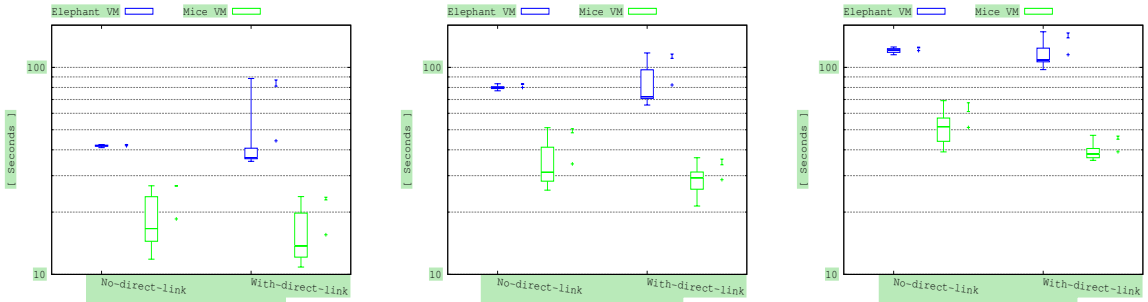
If we consider the mean values, there is a small performance improvement for mice VM transfers when the ToRs shared the direct optical link. However, the whole statistical picture of the results show that the error bars overlap between each other which indicates the performance improvement is not always the case. The improved performance for the elephant flows is also quite modest with a large spread of results. Deeper investigation confirmed that separating the two types of flows has helped isolate the dynamic TCP behaviour of the mice flows from the elephant flows and most importantly any mice flows that become stalled due to the link reconfigurations cannot cause any additional delay to the elephant flows.

Nonetheless, the performance improvements are not sufficiently significant to justify the increased cost of the all-optical network.



(a) 24 VM transfers **(b)** 48 VM transfers **(c)** 72 VM transfers

Figure 6. Throughput for elephant and mice VMs in the 24, 48 and 72 VM transfers scenarios. To the right of each boxplot, we show the mean and a whisker marking the 95th and 99th percentiles.



(a) 24 VM transfers **(b)** 48 VM transfers **(c)** 72 VM transfers

Figure 7. Completion time for elephant and mice VMs in the 24, 48 and 72 VM transfers scenarios. To the right of each boxplot, we show the mean and a whisker marking the 95th and 99th percentiles.

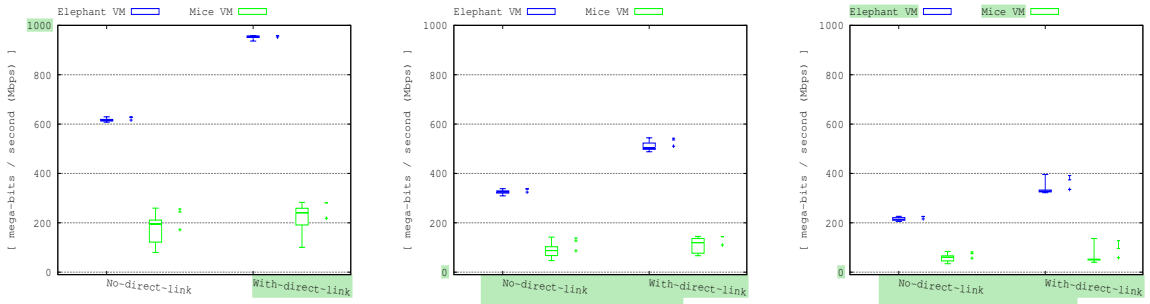
4.3. ON Demand Scheduling Results

The results in this subsection consider the on-demand scheduling technique in which the ToR/host requests access to use the direct optical link for transferring an elephant VM so that there will be no interruption to the end-to-end connection through the network during the transfer.

Figures 8 and 9 show the results of the mice and elephant VM transfers for 24, 48 and 72 VMs transfer scenarios. As shown in the figures, there is no significant performance improvement for the mice VM transfers. This is as expected as the total network capacity remains the same, so although the elephant flows have been offloaded from the leaf-spine network, the capacity of one of the links in the network has been moved to the all-optical network.

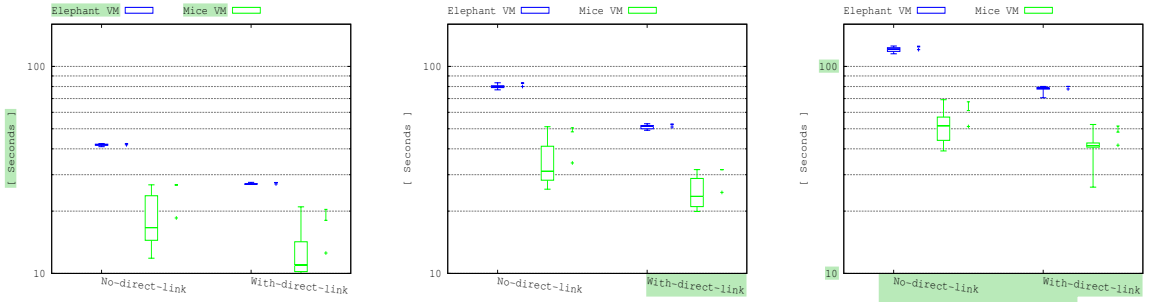
However the results show a significant improvement for the elephant flows. For example in Figure 8b the average throughput of the elephant VM transfers in the 48 VM transfer scenario is improved by ∼35% when the direct optical link is deployed. The performance improvement varies according to the number of elephant VMs competing to use the direct optical link and the normal TCP dynamic behaviour. These results show that by keeping the mice flows out of the all-optical network, the switching overhead

is reduced and by removing the switching during each elephant flow, the network can be used much more efficiently and the TCP behaviour becomes more stable and allows the flows to complete much more quickly that otherwise as shown in figure 9.



(a) 24 VM transfers **(b)** 48 VM transfers **(c)** 72 VM transfers

Figure 8. Throughput for elephant and mice VMs in the 24, 48 and 72 VM transfers scenarios. To the right of each boxplot, we show the mean and a whisker marking the 95th and 99th percentiles.



(a) 24 VM transfers **(b)** 48 VM transfers **(c)** 72 VM transfers

Figure 9. Completion time for elephant and mice VMs in the 24, 48 and 72 VM transfers scenarios. To the right of each boxplot, we show the mean and a whisker marking the 95th and 99th percentiles.

4.4. Limitations of our experiments

Our testbed presents an emulated DC of a small size with 1 Gbps link capacity to enable a tightly controlled set of experiments to be conducted using the mature emulation environment of Mininet. Real DCs are much larger and use network links with rates the order of 100Gbps. The scaling of time and bitrates here is approximate as the impact of different VM sizes, different packet sizes and different implementations of TCP will different numerical results. However, the results presented here are intended to show the scale and nature of the impact of the interaction between TCP and SDN-based reconfiguration of all-optical networks that can alert practitioners to design and implementation considerations when developing such networks.

5. Discussion

The graphical results presented in the previous section showed that the use of an all-optical network capability in a DCN can result in degraded performance with strict time-based sharing of the resource.

Table 1. Summary of the Average throughput [Mbps] for elephant VMs

	24 VM Transfers			48 VM Transfers			72 VM Transfers		
	Min	Ave.	Max	Min	Ave.	Max	Min	Ave.	Max
No direct Line	607.8	616.3	629.5	309.1	324.4	338.3	205.4	215.7	226.1
UT	289.7	550.0	614.6	262.1	287.1	322.5	158.2	187.4	209.6
DT	359.5	652.4	738.7	268.8	344.8	398.5	194.6	234.4	266.1
On Demand	936.4	952.7	958.4	487.8	510.4	544.4	321.9	335.2	396.0

Table 2. Summary of the Average Completion time [seconds] for elephant VMs

	24 VM Transfers			48 VM Transfers			72 VM Transfers		
	Min	Ave.	Max	Min	Ave.	Max	Min	Ave.	Max
No direct Line	40.9	41.8	42.4	77.0	79.9	83.4	115.1	120.5	125.5
UT	41.9	48.8	94.9	81.1	90.8	98.4	123.7	140.4	165.5
DT	35.1	44.2	88.4	65.8	82.2	117.4	97.5	115.2	148.7
On Demand	26.9	27.0	27.5	49.1	50.9	52.8	70.6	77.8	80.0

The main reasons behind that is TCP's congestion mechanism and the differing behaviour with different sizes of flows sharing the optical network. The results for the on-demand scheduling mechanism indicated an improved performance due to the separation of mice flows from elephant flows and removing any possibility of reconfiguring the network during a long-lived elephant flow.

A summary of the throughput results is presented in Tables 1 and 3 while Tables 2 and 4 summarise the completion time results. The results for the throughput show that the on-demand scheduling mechanism yielded an average increase in throughput of $\sim 54\%$, $\sim 57\%$ and $\sim 55\%$ for the scenarios with 24, 48 and 72 simultaneous VM transfers. These significant increases in throughput have arisen purely due to the improved usage of the network resources as the total capacity of the new network is the same as that of the leaf-spine network against which it is compared.

Although the network segment used in this work is relatively small, the network topologies being investigated are typically aimed at providing these direct ToR to ToR links within clusters or targetted at hotspots. Their deployment is, therefore focussed on small segments of the DCN at any given time. We therefore anticipate that the performance figures reported in this work are broadly valid for a larger DCN, although each DCN implementation will have its own specific behaviour that will modify the details of the results presented for the generalised case here.

As link speeds increase into the Terabit per second region, it is possible that newer variants of TCP will need to be deployed. These will need to address the congestion window adjustment mechanism to account for even the shortest of interruptions to a reconfigurable link. An extensive study of various TCP parameters would be required to assess the details of the behavioural interactions and may need to be tuned to each specific DCN deployment. Nonetheless, the results presented here indicate that such tuning needs to include a mechanism to separate mice and elephant flows and to minimise the occurrences of network reconfiguration.

The deployment of such all-optical networks therefore requires careful design and configuration of the end point behaviour through the examination of different TCP parameters. It also motivates the creation of an agent to broker the requests for access to the network that are generated by the applications running on the servers. These requests will require moderation between the agent and the SDN controller

Table 3. Summary of the Average throughput [Mbps] for mice VMs

	24 VM Transfers			48 VM Transfers			72 VM Transfers		
	Min	Ave.	Max	Min	Ave.	Max	Min	Ave.	Max
No direct Line	79.6	171.9	259.3	46.9	86.8	141.7	34.1	56.5	83.6
UT	84.3	150.4	246.8	51.5	86.3	123.1	40.1	56.5	71.5
DT	90.1	177.9	249.8	63.0	99.0	167.6	49.6	69.4	98.2
On Demand	100.4	218.2	282.8	66.6	109.7	144.7	40.1	58.7	136.2

Table 4. Summary of the Average Completion time [seconds] for mice VMs

	24 VM Transfers			48 VM Transfers			72 VM Transfers		
	Min	Ave.	Max	Min	Ave.	Max	Min	Ave.	Max
No direct Line	11.8	18.5	26.7	25.5	34.1	51.1	39.0	51.3	69.0
UT	11.8	19.7	28.0	28.5	34.3	46.1	40.5	51.0	63.0
DT	10.8	15.5	23.7	21.3	28.6	36.6	35.5	39.1	46.9
On Demand	9.6	12.5	20.9	19.9	24.6	31.7	26.1	41.6	52.4

to optimise resource sharing and prioritise jobs by their criticality and potential benefit from using the direct optical connection.

6. Conclusion

A generalised abstraction of a DCN with a reconfigurable all-optical network and a regular Leaf-Spine network was modelled and used to explore the interaction between such a network and TCP under a number of scheduling regimes.The results for time-based scheduling regimes with and without traffic differentiation yielded disappointing results, while an on-demand scheme yielded significant improvements in performance without increasing the installed capacity of the network.

The results show that by deploying an on demand scheduling mechanism to share the all-optical network to deliver elephant flows in DCs could improve the overall performance of the DC. More specifically, the throughput for elephant VM transfers increased by more than $\sim 50\%$ for all of the three workloads examined. It is likely that the most beneficial way to implement such a system would involve a mechanism to connect the demands of the application that manages the VM migrations, the SDN controller and the schedulers at the network edges.

Acknowledgments: This work has been supported by the Agile Cloud Service Delivery Using Integrated Photonics Networking project, funded under the US-Ireland R&D programme supported by National Science Foundation (US), Science Foundation Ireland (RoI) and Department for the Economy (NI) (Grant number NSI-085).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cisco Systems. (2015, February) Cisco global cloud index: Forecast and methodology, 2014-2019 white paper. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.html
2. Hamedazimi, N.; Qazi, Z.; Gupta, H.; Sekar, V.; Das, S.R.; Longtin, J.P.; Shah, H.; Tanwer, A. FireFly: A Reconfigurable Wireless Data Center Fabric Using Free-space Optics. Proceedings of the 2014 ACM Conference on SIGCOMM; ACM: New York, NY, USA, 2014; SIGCOMM '14, pp. 319–330. doi:10.1145/2619239.2626328.

3. Saha, S.; Deogun, J.S.; Xu, L. HyScaleII: A high performance hybrid optical network architecture for data centers. 2012 35th IEEE Sarnoff Symposium, 2012, pp. 1–5. doi:10.1109/SARNOF.2012.6222725.
4. Christodoulopoulos, K.; Lugones, D.; Katrinis, K.; Ruffini, M.; O'Mahony, D. Performance evaluation of a hybrid optical/electrical interconnect. *IEEE/OSA Journal of Optical Communications and Networking* **2015**, *7*, 193–204. doi:10.1364/JOCN.7.000193.
5. Samadi, P.; Gupta, V.; Xu, J.; Wang, H.; Zussman, G.; Bergman, K. Optical multicast system for data center networks. *Opt. Express* **2015**, *23*, 22162–22180. doi:10.1364/OE.23.022162.
6. Cisco Systems. (2014, August) Data center technology design guide. [Online]. Available: http://www.cisco.com/c/dam/en/us/td/docs/solutions/CVD/Aug2014/CVD-DataCenterDesignGuide-AUG14.pdf
7. Al-Fares, M.; Loukissas, A.; Vahdat, A. A Scalable, Commodity Data Center Network Architecture. Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication; ACM: New York, NY, USA, 2008; SIGCOMM '08, pp. 63–74. doi:10.1145/1402958.1402967.
8. Ghobadi, M.; Mahajan, R.; Phanishayee, A.; Devanur, N.; Kulkarni, J.; Ranade, G.; Blanche, P.A.; Rastegarfar, H.; Glick, M.; Kilper, D. ProjecToR: Agile Reconfigurable Data Center Interconnect. Proceedings of the 2016 ACM SIGCOMM Conference; ACM: New York, NY, USA, 2016; SIGCOMM '16, pp. 216–229. doi:10.1145/2934872.2934911.
9. Farrington, N.; Porter, G.; Radhakrishnan, S.; Bazzaz, H.H.; Subramanya, V.; Fainman, Y.; Papen, G.; Vahdat, A. Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers. Proceedings of the ACM SIGCOMM 2010 Conference; ACM: New York, NY, USA, 2010; SIGCOMM '10, pp. 339–350. doi:10.1145/1851182.1851223.
10. Wang, G.; Andersen, D.G.; Kaminsky, M.; Papagiannaki, K.; Ng, T.E.; Kozuch, M.; Ryan, M. c-Through: Part-time Optics in Data Centers. Proceedings of the ACM SIGCOMM 2010 Conference; ACM: New York, NY, USA, 2010; SIGCOMM '10, pp. 327–338. doi:10.1145/1851182.1851222.
11. Vujicic, V.; Anthur, A.P.; Gazman, A.; Browning, C.; Pascual, M.D.G.; Zhu, Z.; Bergman, K.; Barry, L.P. Software-Defined Silicon-Photonics-Based Metro Node for Spatial and Wavelength Superchannel Switching. *J. Opt. Commun. Netw.* **2017**, *9*, 342–350. doi:10.1364/JOCN.9.000342.
12. A. S. Thyagaturu, A. Mercian, M. P. McGarry, M. Reisslein and W. Kellerer, Software Defined Optical Networks (SDONs): A Comprehensive Survey IEEE Communications Surveys Tutorials, vol. 18, no. 4, pp. 2738-2786, Fourthquarter 2016.
13. Postel, J. Transmission Control Protocol. RFC 793 (Standard), 1981. Updated by RFCs 1122, 3168.
14. Jacobsen, V. Congestion avoidance and control. Proc. ACM SIGCOMM 1988, 1988.
15. Mininet: [http://www.mininet.org].
16. M. Ruffini and D. O'Mahony and L. Doyle A cost analysis of Optical IP Switching in new generation optical networks. 2006 International Conference on Photonics in Switching; Heraklion, Crete, 2006
17. M. Abu-Tair, M. I. Biswas, P. Morrow, S. McClean, B. Scotney, and G. Parr, "Quality of service scheme for intra/inter-data center communications," in *2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA)*, March 2017, pp. 850–856.