Comparison of SDM and WDM on Direct and Indirect Optical Data Center Networks

Yifan Liu[(1)], Hui Yuan[(1)], Adaranijo Peters [(1)], Georgios Zervas[(1)]

[(1)] High Performance Networks Group, University of Bristol, UK, georgios.zervas@bristol.ac.uk

Abstract *We benchmark Data Centre topologies under SDM and WDM transport in terms of network capacity, utilization, blocking probability, cost and power consumption. SDM offers cost and power benefits than WDM while Spine-Leaf demonstrates all-round best performance among all topologies.*

Introduction

Exponentially increasing demand of network traffic drives the necessity of exascale data centers. Optical interconnects are expected to support such Data Center network requirements [1]. Wavelength Division Multiplexing (WDM) and advanced modulation formats have been used traditionally on terrestrial networks to stretch the capacity limit of single mode fibre (SMF). However the complexity, cost, size, power consumption and heat dissipation of WDM transmission and switching systems over C+L-Band might deem them unsuitable for Data Center networks. Space Division Multiplexing (SDM) [2] by use of multi-core fibers has been introduced showing huge potentials to improve the network performance and can utilize cost effective and energy efficient integrated technologies (i.e. integrated VCSEL array [3]). Studies have been conducted on how to optimize and allocate spectral and/or spatial resources while considering particular constraints, i.e. inter-core crosstalk (XT) [4, 5] on backbone networks. However, there hasn't been any study to investigate SDM for Data Centers and which of the two multiplexing approaches best suit such short-reach networks. Other than multiplexing technologies, Top-of-Rack (ToR) and computing node interconnection is decisive when estimating the network performance. Thus, investigating data center topologies, direct or indirect, with different number of nodes, links and switch nodes using either SDM or WDM is of great importance.

This paper proposes and investigates the use of either SDM-only with MCFs or WDM-only using SMFs on five topologies: direct (2D Torus) and indirect (Star, Spine-Leaf, Facebook, Data Vortex) (Fig. 1a). All support 16 Racks each with 37 compute nodes. Each server interconnects to ToR with a single channel (spatial or spectral) at 400 Gb/s. In case of SDM, we consider a 37-core homogeneous MCF (Fig. 1b) carrying one channel per core whereas in case of WDM we use 37 wavelengths for a fair comparison. Using developed resource allocation algorithms, we evaluate topology and SDM/WDM performance in terms of network utilization, blocking probability, capacity under 10% blocking probability as well as cost, switching devices and ports per node, and power consumption.

Resource Allocation for SDM-only and WDM-only Data Center networks

A Matlab simulator was developed to evaluate the performance of investigated Data Center networks with routing, spectrum and core assignment algorithms as seen on Fig. 1c. The resource for WDM takes C+L-band to transmit data signals represented as 148 frequency slots. Each frequency slot occupies 50 GHz and the requests simulated in this paper are regarded as 400 Gb/s over 4 frequency slots making it 37 spectral channels (Fig 1.b). In case of 37-core SDM, a single channel per core is considered. The characteristics of the homogeneous 37-core MCF used to calculate the crosstalk are 30μm core-pitch (Λ), 6×10^{-2} m^{-1} coupling coefficient (κ), transmission distance (L), bending radius $50\times10^{-}$

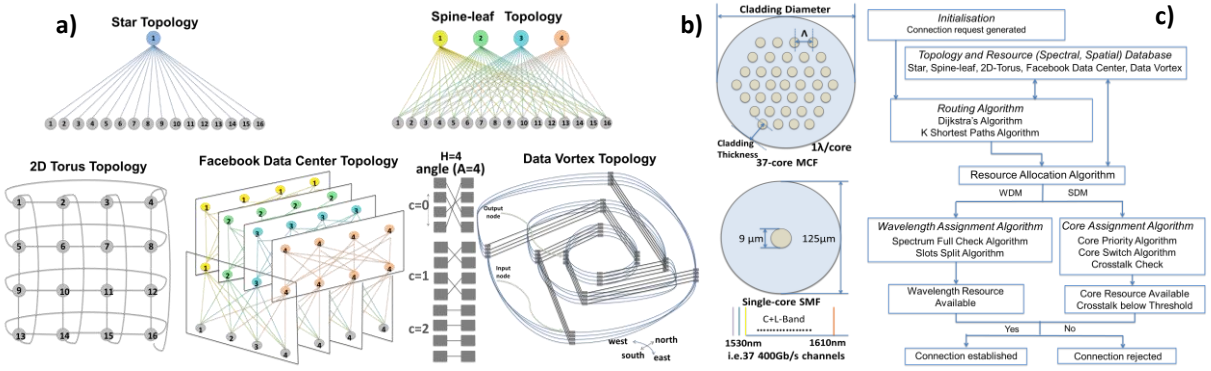

Fig. 1: a) Topologies under investigation, b) WDM and SDM resource considerations, c) simulation process

^{3} m (R), propagation constant 4x10^{6} (β) [2]. Two link distances of 25m and 100m are considered to evaluate the crosstalk impact. Each link uses two fibres one per direction. The threshold for network crosstalk value is set to -24 dB for both new and existing connections prior to accepting any new request. Fig.1a illustrates the topologies investigated with 16 end nodes (i.e. ToR) for a fair comparison.

Fig. 2a and Fig. 2b illustrate the procedure of the proposed allocation assignment algorithms for WDM-only and SDM-only cases. The requests that follow a random distribution with no holding time (incremental traffic load) are first generated with the source node, the destination node and the required bandwidth (either 4 frequency slots for WDM or 1 channel for SDM). K-Shortest Path routing algorithm then provides 3 alternative paths for the request. For WDM, the spectrum allocation algorithm combines Spectrum Full Check algorithm and Slots Split algorithm where the requests are split into smallest pieces and allocated according to the available slots. After connection request of 4 frequency slots is generated and the routing algorithm finds the path(s), the request is separated into four 1-slot bandwidth pieces (Slot-Split algorithm). The 4 groups are consequently allocated in order. The request will be rejected once all the 148 slots are checked across all links of the path(s) and there are not enough available resources (Fig. 2a).

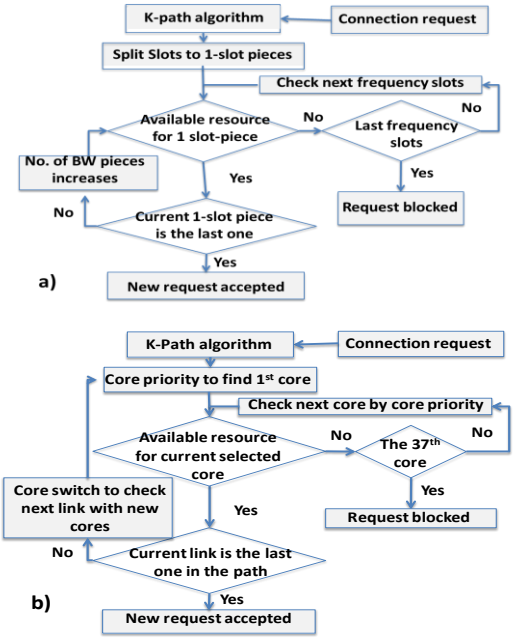

Fig. 2: a) Route & Spectrum allocation for WDM, b) Route & Core allocation for SDM

In case of SDM (Fig. 2b) XT occurs between adjacent cores, which results in blocking. Therefore, Core Priority algorithm [4] is used as a pre-defined policy to reduce the crosstalk between adjacent cores by setting the sequence of core usage for transmission. In addition, core switch algorithm is proposed due to the inherent flexibility to switch the allocated cores freely between two links and mitigates the spectrum continuity issues existing in WDM and SDM networks. The core priority, core switch and crosstalk check process will be repeated until all resources are checked. The request will only be accepted if there is available resource and the crosstalk is below threshold (i.e. -24 dB).

Performance Evaluation

We evaluate the performance of the investigated topologies while benchmarking WDM and SDM in the form of network behavior, network capacity, cost and power consumption. In order to find the best performance among all the options, the network behavior is looked into first. The network behavior is plotted as the blocking probability versus the network utilization when considering 25m-link distance (Fig. 3a). Higher network utilization with a relatively low blocking probability is highly desirable. The maximum blocking probability of 0.1 (10%) is selected as a typical maximum acceptable value. Performance of 2D Torus, the only direct topology is worst of all since the random selection of source-destination and load imposed to each node from both bypass and add-drop traffic causing an elevated blocking probability even for low load. Star, Spine-Leaf, and Facebook appears to perform better when using SDM rather than WDM. This is due to multi-hop and/or multi-route ability that is enhanced by core switch and deteriorated by spectrum continuity constraint in SDM and WDM cases respectively. Out of these, Spine-Leaf and Facebook topologies offer very low blocking probability i.e. 0.01 even under very high network utilization >80%. Data Vortex topology shows the opposite behavior (WDM performs better than SDM) since the path distances are considerably longer (average 4 links per path) and SDM suffers from XT.

Regarding the network capacity, as shown in Fig. 3b Spine-Leaf and Facebook topologies perform best with values of up to 0.49 Pb/s under 0.1 blocking probability. Small differences exist on maximum capacity under 0.1 blocking probability when using SDM or WDM across Star, Spine-Leaf and 2D Torus. Compared to WDM, SDM offers ~15% capacity improvement in Facebook topology due to increased number of path options whereas SDM offers 60% less capacity in Data Vortex due to long paths and XT. However, the impact of XT in SDM case is reflected on capacity reduction (Fig. 3b) when the link distance is set to 1km. To compare topologies and transport in terms of cost, Table

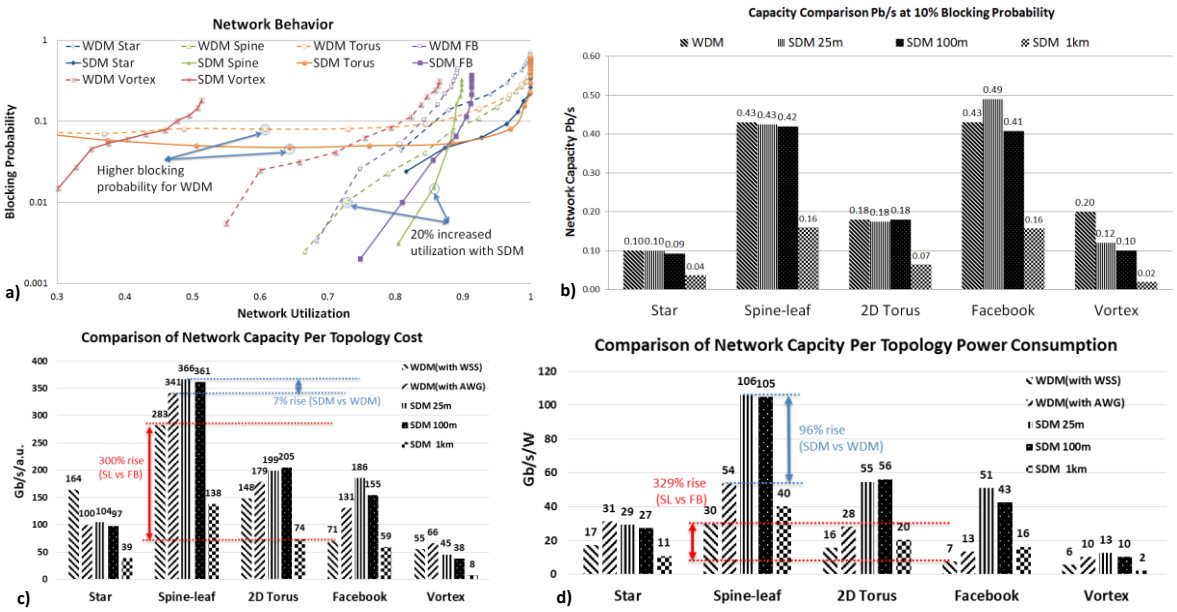

Fig. 3: Topology and transport benchmark, a) blocking probability vs. network utilization, b) network capacity under 10% blocking probability, c) network capacity per topology cost and d) network capacity per topology power consumption.

1 (left) lists the arbitrary units for different switching devices considering as reference a 1x40 array waveguide grating (AWG).

Manufacturers provide the costs that could change depending on volume, market and fabrication process. Note that since a 600x600 (1200 ports) 3D-MEMS Fibre switch is not commercially available (yet required for SDM case on Star and Spine-leaf intermediate nodes) we have considered its cost 40% higher than the 320x320 one. Table 1 (right) shows the number of switching devices and ports per switch required. The node design assumption for SDM is the use of a single fibre switch and passive MCF-to-SMF fan-in/out devices. In case of WDM we have considered two alternatives, one using AWGs and fibre switch in route, switch and select architecture or wavelength selective switch in route and select configuration.

Conclusions

Although Spine Leaf and Facebook have similar maximum network capacity, Spine Leaf delivers 97%-300% higher capacity per topology cost (Fig. 3c) and 108%-329% energy efficiency improvement (Fig. 3d). However, Facebook topology is considerably more modular and scalable. Fig 3c and Fig 3d clearly indicates that SDM provides better cost (except Star) even without considering the reduced cost of Tx/Rx required for SDM and power performance across all topologies than that of WDM.

Acknowledgements

This work was supported by EC H2020 dRedBox project, and EPSRC grant EP/L027070/1

References

[1] C. Kachris, et al., "A Survey on Optical Interconnects for Data Centers", IEEE Com. Surveys Tuts, Vol.14, No. 4, Oct. 2012

[2] G. Saridis, et al., "Survey and Evaluation of Space Division Multiplexing: From Technologies to Optical Networks", IEEE Com. Surveys Tuts, Is,99, Dec. 2015

[3] B. G. Lee, et al., "120-Gb/s 100-m Transmission in a Single Multicore Multimode Fiber Containing Six Cores Interfaces with a Matching VCSEL Array", Photonic Society Summer Topical Meeting, August 2010

[4] S. Fujii, et al., "On-demand spectrum and core allocation for reducing crosstalk in multicore fibers in elastic optical networks", JOCN, vol. 6, no. 12, pp. 1059–1071, 2014.

[5] A. Muhammad, et al., "Routing, spectrum and core allocation in flexgrid SDM networks with multicore fibers", ONDM, May 2014, pp. 192–197.

Table 1 : Cost, power assumptions and resource requirements per topology and technology (SDM and WDM)

Device	**Cost (a.u.)**	**Power (W)**		**SDM**: Total # of switching devices (left) and port/switch (*2*C*L*) for end and intermediate nodes (right)		**WDM**: Total # of switching devices (WSS & AWG - left) and ports/WSS_device (end and intermediate node - right)	
Common equipment	-	50	**Star**	17	148 (*E-n*) /(*Int-n*) 1184	64(*WSS*)/64 (*AWG*) + 1 *f-sw*	2 / 17
λ Mux (AWG 1x40)	1	0	**Spine-Leaf**	20	370 / 1184	160 / 160+1 *f-sw*	2 / 17
λ Switch (WSS 1x20)	9.5	40	**2D Torus**	16	222 / 222	128 / 128+1 *f-sw*	6 / 6
Fibre Switch 3D-MEMS (320x320)	55	150	**Facebook**	48	370 / 444	640 / 640+1 *f-sw*	5 / 9
Fibre Switch piezo-electric (384x384)	66	100					
Fibre Switch 3D-MEMS (600x600)*	77	200	**Vortex**	48	185 / 148	384 / 384+1 *f-sw*	5 / 5

C: # cores per MCF, L: # links per node, E-n: End node, Int-n: Intermediate node, f-sw: Fibre switch. *Cost of 600x600_switch =1.4xCost of 320x320_switch