Disaggregated Optical Data Center in a Box Network using Parallel OCS Topologies

Hui Yuan[1], Arsalan Saljoghei[1], Adaranijo Peters[2], Georgios Zervas[1]
[1] Optical Networks Group, University College London, United Kingdom, [2]University of Bristol
h.yuan@ucl.ac.uk

Abstract: Two parallel OCS topologies are proposed that deliver 95 nsec round-trip latency on disaggregated optical data center in a box system. They offer 40% cost and 68% power consumption efficiency at maximum IT resource utilization.

OCIS codes: (060.4258) Networks, network topology; (200.4650) Optical interconnects; (200.6715) Switching

1. Introduction

Conventional server-centric data center architectures [1], where each server tightly integrates a fixed amount of IT (CPU, memory) and network resources onto a single mainboard present shortcomings in areas such as flexibility, utilization and adaptability. Particularly, IT resources cannot be fully utilized in each server due to the mismatch between fixed proportionalities and diverse set of request requirements [2]. To address these issues, disaggregated data center architecture has been proposed, where each distinct resource element is configured in a standalone baseline disaggregated pool of compute, memory, and accelerators and a network fabric interconnects these resource pools [3]. This new resource-centric architecture can provide an immense level of scalability and flexibility since each specific resource could be dynamically added or removed depending on various application environments, which triggers the expansion on the capacity and optimized utilization of resources through co-operative sharing. Recent research on memory disaggregation at rack level has so far identified approximately a 10-fold increase in performance over server-centric architecture and it also has improved the performance-per-dollar by up to 87% [4].

However, such disaggregated architectures still present a number of fundamental challenges that need to be addressed: a) they require lower latencies compared to the traditional direct-attached modular, b) cost and power consumption need to be reduced whilst supporting a substantially higher bandwidth and bandwidth density, and c) the substrate and orchestration should enable the system to support various specific resource connections (e.g., CPU-RAM, RAM-storage) at low latency and cost. The dReDBox (disaggregated Recursive Datacenter-in-a-Box) architecture as such was proposed showing an advantage in modularity, scalability and IT resource utilization maximization [5]. As shown in Fig. 1(a), this non-parallel architecture utilizes a 3-tier tree topology considering dBOSM (disaggregated Box Optical Switch Module), dROSM (disaggregated Rack Optical Switch Module) and dDOSM, (disaggregated Data Center Optical Switch Module) switches to deliver any to any connectivity.

In this paper, we report on the evaluation of the initial dRedBox architecture and compare it with two proposed parallel optical circuit switching (OCS) topologies. Results show that both of the parallel topologies offer improvements on cost and power consumption while the Box-modular topology shown in Fig. 1 (b) also provides benefits in resource saving and latency reduction compared to the non-parallel dRedBox architecture.

2. Proposed Parallel Disaggregated Optical Data Center Architectures

Fig. 1(a) illustrates the non-parallel architecture, which consists of dRacks comprising of multiple interconnected dBoxes. In each dBoxes, different type of dBricks (i.e. CPU, memory, storage and accelerator) are plugged with arbitrary combinations. Additionally, relatively small port-count optical switches (i.e. 96) with high port-density are adopted at tier-1 to accommodate the dBricks and large port-count switches (i.e. 384x384) at tier-2 and tier-3 to provide any to any dBrick communication within and between dBoxes on one or multiple dRacks.

Compared to the non-parallel architecture, the two proposed parallel architectures are 2-tier topologies, this approach will minimize the routed path distance between any two dBricks and in turn reducing network latency. To realize the parallel structure, the concept of dPlane is developed. As shown in Fig. 1(b) and (c), dPlanes are not connected to each other allowing modular increase per dBox bandwidth. Each dPlane houses the top-tier of dPOSMs (disaggregated Plane Optical Switch Module) connecting to the lower-tier of dBOSMs with a spine-leaf topology. Since the dBOSMs in each dPlane link all the dBricks, any to any dBrick communication can be executed via any individual dPlane. Each dPlane can support and switch one spatial channel (single or multiple wavelengths) per dBrick making it a capacity modular parallel topology.

The only difference between the Box-modular architecture and the Brick-modular architecture is involved with the connectivity between the dBOSMs and the dBricks. For the Box-modular architecture, all dBricks in one dBox are interconnected via a corresponding dBOSM switch in each dPlane. For instance, in Fig. 1(b), dBricks in the grey dBox are only linked to the grey dBOSM in each dPlane. Communication between the dBricks in different dBoxes

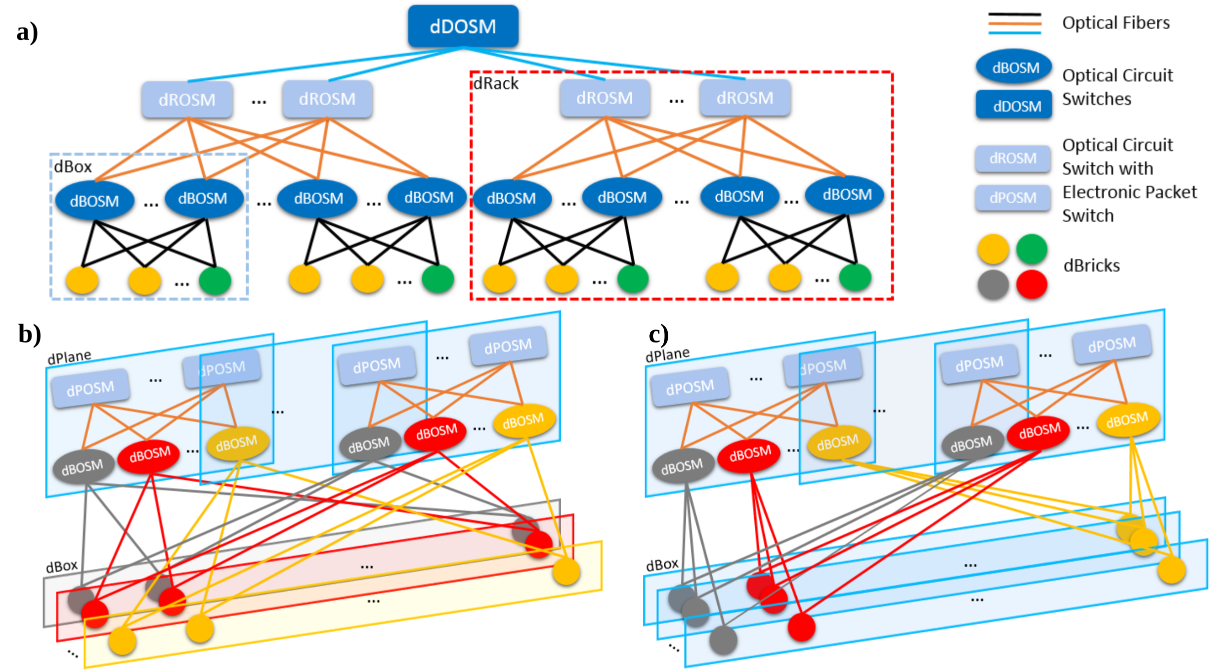

Fig. 1. Disaggregated Data Center Topologies: a) Non-parallel, b) Box-modular, c) Brick-modular

need the participation of the dPOSM. On the contrary, in the Brick-modular architecture as shown in Fig. 1(c), dBricks at the same position of every dBox are interconnected via a corresponding dBOSM switch in each dPlane and dPOSM should be involved for dBrick to dBrick communication in the same dBoxes. For both of the parallel topologies, the dPlane number depends on the number of physical channels per dBrick and link. While the dBOSM number in each dPlane equals to the dBox number for Box-modular architecture and the dBrick number in each dBox for Brick-modular architecture. These fabrics enable data center scaling out by keep adding same small port-count dBOSM and dPOSM switches rather than using large port count switches, which in turn indicates to the advantages realizable by these modulator topologies in terms of scalability, flexibility and switch port utilization.

3. Simulation and results analysis

A simulator has been developed in Matlab to evaluate the performance of the proposed architectures. The overall simulation procedure consists of four main steps: 1) request generation, 2) IT resource allocation, 3) network resource allocation 4) connection establishment. In step 1, VM requests arrive dynamically following a Poisson distribution with a 10-time units average inter-arrival time and they contain the relevant information with regards to the CPU core number, RAM size, storage size, CPU-RAM latency & bandwidth requirements, RAM-Storage latency & bandwidth requirements and the resource holding time. The holding time starts from 6300 time units and increases 360 time units for every 100 requests. The network-aware locality based resource allocation algorithm reported in [5] is applied in step 2 and a modified K-shortest path algorithm is employed in step 3. Resources are reserved for a VM request only when sufficient IT and network resources are identified within the network, otherwise, the request will be dropped. The simulation parameters for the three architectures are depicted in Table 1. Each system has 1/3 ratio of compute, memory and storage resources and each distinct resource element is configured at same position of each dBox.

Fig. 2 depicts the simulation results for all investigated architectures in terms of resource utilization, blocking probability, roundtrip network latency, cost and power consumption. Fig. 2(a) and (b) illustrate that, all the three architectures have very similar blocking probability since all support any-to-any connectivity (no oversubscription) and can fully utilize the IT resource when sufficient bandwidth are provided. Comparing with the non-parallel architecture, the Box-modular architecture shows benefit in network resource saving, as it can be seen it achieves

Table 1 Simulation configurations for the architectures

Architecture	dRacks	dPlanes	dBoxes	dBricks per dBox	Link distance (m)		
					dBrick-dBOSM	dBOSM-dR/POSM	dROSM-dDOSM
Non-parallel	4	-	6 per dRack	8	0.25	3	10
Box-modular	-	8	24	8	0.25	3	-
Brick-modular	-	8	8	24	0.25	3	-
Common setup	8 I/Os per dBrick @ 25Gb/s			heterogeneous dBoxes [5]		5ns pass-through latency per switch	

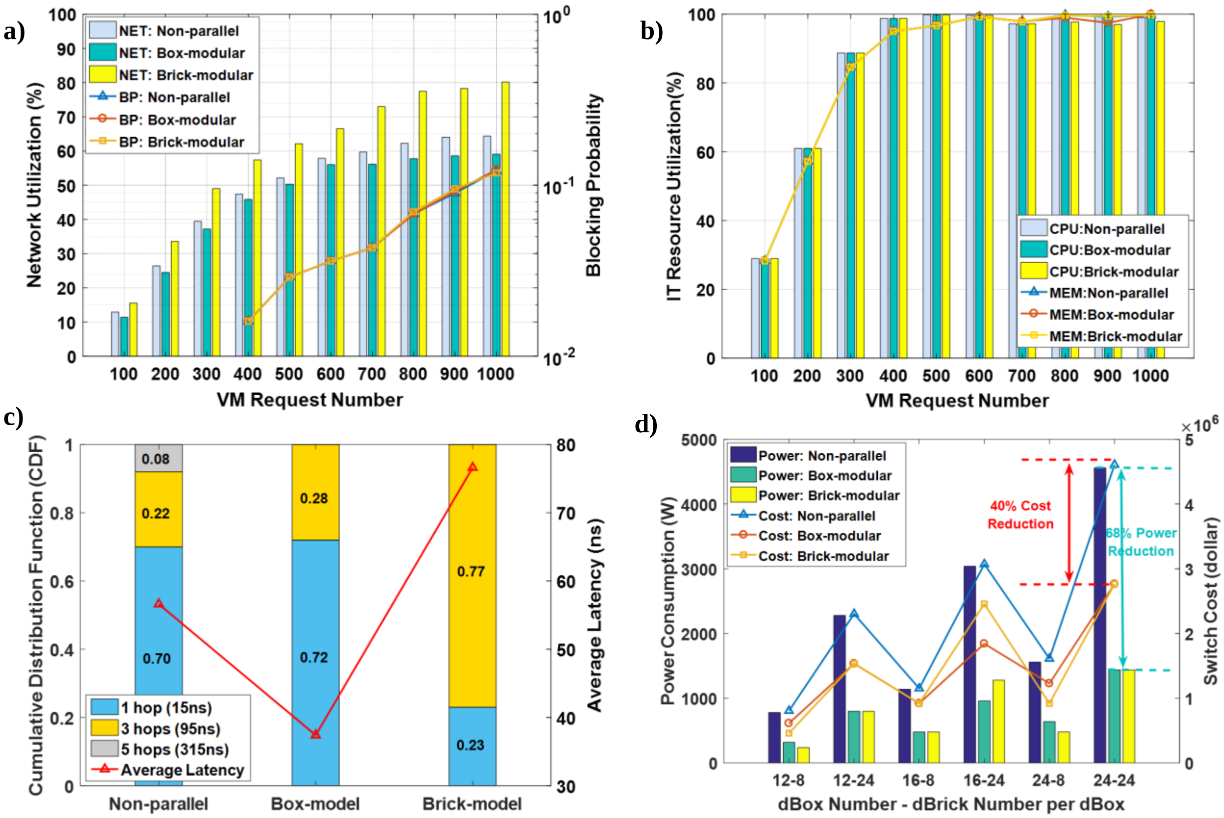

Fig. 2. Overall comparison between non-parallel and parallel topologies in terms of: a) Network utilization and blocking probability, b) IT resource utilization, c) Network latency, d) Switch cost and power consumption. *BP: Blocking Probability

9% reduction in network utilization after processing 1000 requests. However, the Brick-Modular demonstrates a higher network utilization when compared to the Box-modular and non-parallel architecture across all request numbers. This translates into a requirement for more network resources to achieve a similar level of blocking probability. The reason can be found in Fig. 2(c), where the overall traffic/latency distribution for different topologies is shown. Since each type of dBrick is configured at the same position of each dBox, dBOSM to dPOSM links (3-hops connection) as shown in Fig. 1(c) should involve for communication between different type of resources (e.g., CPU-RAM, RAM-storage) in Brick-modular architecture and in turn high network utilization as well as latency. Although Box-modular architecture has similar traffic distribution with the non-parallel architecture, it delivers 34% latency reduction and guarantees 95 nsec round-trip latency between CPU and memory on all established VMs. Cost and power consumption for architectures with various configurations are compared in Fig. 2(d), where we assume 100 dollars per switch port, 100W per 384x384 ports switch and 5W per 96 ports switch. Transparently, the two parallel modulars considerably outperform the non-parallel architecture. Particularly, they can save 40% cost and 68% power for supporting data center with 576 dBricks.

4. Conclusion

We have proposed two parallel architectures: Box-modular and Brick-modular, for a disaggregated optical data center. After comparisons with the 3-tier tree architecture, both of them offer highest 40% cost reduction and 68% power consumption efficiency while delivering the maximum IT resource utilization. Particularly, the Box-modular is the best one which also displays the potential for resource saving and latency decreasing. Moreover, even the 3-tier architecture that can scale to cluster level distances, due to the latency-aware algorithms can deliver 95 nsec round-trip latency for 92% of the VMs.

5. Acknowledgement: This work is supported by European Commission H2020 dRedBox project.

6. References

[1] S. Han, et al, “Network Support for Resource Disaggregation in Next-Generation Datacenters,” in ACM SIGCOMM, Hotnets-XII, 2013
[2] Tencent and Intel, “Tencent Explores Datacenter Resource-Pooling Using Intel® Rack Scale Architecture (Intel® RSA)”, white paper, 2015
[3] K. Katrinis et al., “Rack-scale disaggregated cloud data centers: The dReDBox project vision,” DATE, Dresden, 2016
[4] K. Lim, et al., “Disaggregated memory for expansion and sharing in blade servers,” ISCA, 2009
[5] G. Zervas et al., “Disaggregated compute, memory and network systems: A new era for optical data centre architectures,” OFC, 2017