Network Synthesis of a Topology Reconfigurable Disaggregated Rack Scale Datacentre for Multi-Tenancy

Adaranijo Peters[1], Georgios Zervas[1]
[(1)]Department of Electrical and Electronic Engineering, University of Bristol, United Kingdom
adaranijo.peters@bristol.ac.uk

Abstract: A performance analysis of a hybrid reconfigurable disaggregated datacentre is presented. It offers substantial benefits in terms of network blocking, power consumption and cost when compared to pure circuit switched and statistical hybrid architectures.
OCIS codes: (060.4253) Networks, circuit-switched; (060.4259) Networks, packet-switched, (060.4230) Multiplexing

1. Introduction

Traditional datacentre (DC) architectures are experiencing significant drawbacks in scalability, power consumption, modularity and resource utilization to effectively manage the exponential growth of internet traffic. To this end, novel DC architectures and technologies must be explored. Resource disaggregation in DCs has been proposed as a solution to improve resource utilization, modularity, customization and upgradeability in DC structures and technology [1]. In more detail, each server which is composed of compute and memory resources are split up as standalone resource pools attached to the DC network. In addition, hybrid DC architectures employing optical and electrical technologies have been proposed to offer significant benefits in reduction of power consumption, high bandwidth and low latency [2, 3]. Therefore, the concept of resource disaggregation combined with hybrid technologies has a great potential to avert the current under-utilized DCs. The dRedbox vision aims to materialize of the concept of resource disaggregation with the state of the art software plane, optical and electrical technologies to create a customizable low power DC which can deliver ultra-low latency, high throughput and modularity [4]. In this paper, a hybrid reconfigurable disaggregated rack scale DC architecture is presented. A simulator is developed to investigate the performance of multiple switching technologies and disaggregated resource architectures when supporting multi-tenancy via Virtual Machine (VM) deployment. Finally a comparison of the proposed architecture with classical hybrid architecture is analyzed in terms of network cost and power showcasing its potential benefits.

2. Disaggregated rack scale datacentre architecture

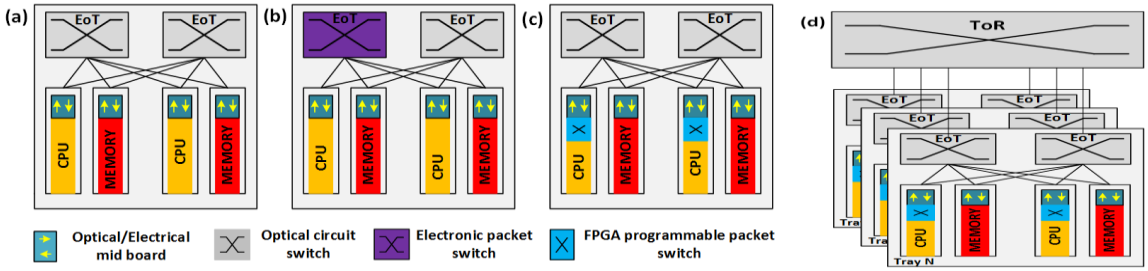

Fig. 1. (a) Pure OCS tray architecture (b) classical statistically dimensioned and configured hybrid tray architecture (c) dRedbox tray architecture (d) dRedbox disaggregated rack architecture

Figure 1(c) presents the dRedbox tray architecture. The tray contains embedded CPU on Multiprocessor System on Chip (MPSoC) bricks, memory bricks and compact optical switches called edge of tray (EoT) which provide pure optical interconnect for low latency intra-tray and inter-tray communication between Bricks. Each MPSoC Brick has embedded CPU cores and FPGA programmable switch interface card that drives mid-board optics. The programmable logic can configure each of its ports to provide layer 2 packet switch functionality while the memory bricks have only interfaced with mid-board optics. The dRedbox disaggregated rack scale architecture is illustrated in figure 1(d), the top of the rack (ToR) is a higher degree optical switch which provides intra-rack and inter rack communication. Figure 1(a) and 1(b) illustrates the tray architecture of the pure OCS and classical statistically dimensioned hybrid respectively. In the pure OCS tray architecture all bricks are only interfaced with mid-boards to the OCS based EoTs while in the classical hybrid tray architecture all bricks are connected partly to one EPS based EoT and partly to one OCS based EoT. Different configurations of the proposed architecture can be achieved by varying numbers of EoTs per tray, ports/transceivers per brick, trays in a rack, port configurations of optical switches and the arrangement of MPSoC and Memory resource pools per tray. Heterogeneous rack has both MPSoC and memory bricks on one tray and a homogenous rack has MPSoC only or memory only bricks on one tray.

3. Proposed working principle and Algorithm

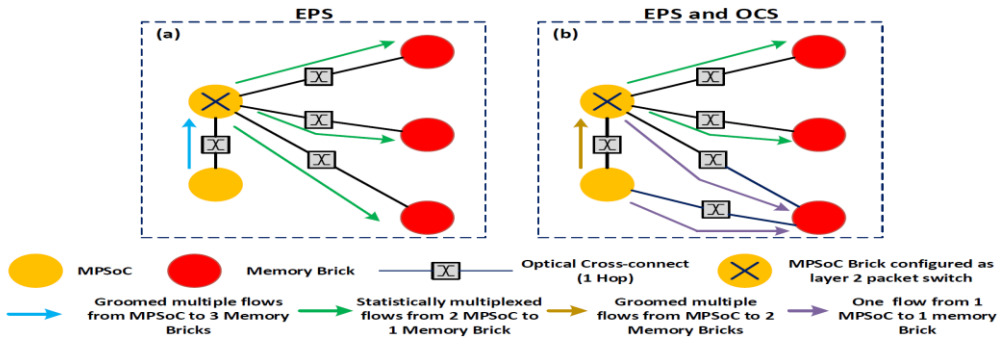

Fig. 2. Illustration of possible topologies using (a) EPS (b) a combination or EPS and OCS

A simulator was developed in Matlab for optical network function and topology synthesis of the proposed DC architecture to investigate the best combination of multiplexing and switching techniques to process network traffic. VM requests consisting of 2 MPSoC bricks and 3 Memory bricks arrive one after the other dynamically specifying the location of each brick in the rack and the required bandwidth of each flow. The first step of the simulator is to analyze the VM requests and required bandwidth of each flow to decide the routing, switching and multiplexing strategy i.e. EPS or EPS and OCS or OCS and how to select the MPSoC brick that should be configured as layer 2 packet switch. We build a strategy matrix by calculating the total bandwidth of flows from each of the 2 MPSoC bricks to the 3 Memory bricks and the total bandwidth of flows from the 2 MPSoC brick to the same Memory brick. For pure EPS scenario, if the total bandwidth of all flows from 1 or 2 MPSoC to 3 memory is less than or equal to the capacity of a transceiver (i.e. can form an EPS channel) and the total bandwidth of flows from the 2 MPSoC bricks to the same Memory brick can also form an EPS channel. Then the multiple possible end to end topologies and network service chain combinations between the MPSoC and Memory bricks for the VM request are created. Figure 2(a) illustrates one of the possible topology combination using one MPSOC brick as a packet switch. For the EPS and OCS scenario, if the total bandwidth of flows from each of the 2 MPSoC to the 3 Memory brick is greater than the capacity of a transceiver, then an EPS channel of flows from the 2 MPSoC bricks cannot be formed thus MPSoC to MPSoC path for 3 Memory bricks is not possible. The simulator then checks if MPSoC to MPSoC path is possible for different combinations of the 2 Memory bricks. If the condition is satisfied i.e. total bandwidth of flows from 1 or 2 MPSoC to 2 Memory bricks can form an EPS channel. A topology and network function service chain of EPS path is built between the 2 MPSoC and 2 Memory, while the third Memory brick is accessed through OCS from both MPSoC bricks as illustrated in figure 2(b). This scenario also has multiple possible end to end topologies between the MPSoC bricks and Memory bricks. Thirdly for the Pure OCS scenario if none of the previous mentioned criteria is satisfied, a pure OCS brick to brick topology is built. After all possible topologies for the VM request are built. For each brick to brick path to be established, the simulator searches for already established connections in the DC network to check whether available bandwidth exist to perform grooming services. If no established brick to brick path exists it then allocates available network resources in a first fit approach. Once resources are found for all links in a VM request, the request is established. If no resources are found for any link in a VM requests, all other possible topology combinations of the VM requests are searched and if there are no resources for all the possible topologies, the VM request is rejected. Each accepted VM request has a lifetime and once the lifetime expires the simulator releases all resources assigned to that VM request.

4. Simulation scenario and results

The dRedbox and Pure OCS rack architecture simulated each has 1 rack with 4 trays. Each tray contains 24 bricks i.e. 12 MPSoC, 12 Memory and each bricks is interfaced with mid board optics of eight 10G transceivers corresponding to eight I/O ports. The bricks in each tray are interconnected to 2 EOT optical switches with 192 ports each and the 4 trays are interconnected by 2 TOR optical switches with 384 ports each. We assume that VM requests arrive dynamically following a poison process with a mean inter-arrival rate of 10 time units and an increasing holding time range of 100-1000 time units with incremental steps of 100 time units. Each VM request consist of 2 MPSoC bricks (source) to interconnect to 3 Memory bricks (destination). The link bandwidth requirement on each VM request varies between mice flow (1Gb to 5 Gb) and elephant flow of (6Gb to 10Gb). We vary the percentage of mice and elephant flows for range of network request. For 1 to 200 we assume a mice flow: elephant flow ratio of 0%:100%, 201 to 400 a ratio 25%:75%, 401 to 600 a ratio of 50%:50%, for 601to 800 a ratio of 75%:25% and for 801 to 1000 a ratio of 100%:0%. As shown in figure 3(a) the dRedbox architecture demonstrates a lower blocking probability than pure OCS architecture (7.3 % of blocking probability lower than

OCS at 1000 holding time unit). This translates to a resource savings of 16.5 % which can be utilized to handle more VM requests. Figure 3(b) shows the blocking probability of the dRedbox homogenous and heterogeneous disaggregated rack architectures. It is noted that the blocking probability of the two architectures are almost overlapping suggesting similarity in terms of acceptance of VM requests, However figure 3(c) shows that the homogenous tray architecture utilizes more switch ports the heterogeneous tray architecture for all cases. This is because MPSoC to Memory Brick communication must pass through three hops of switches i.e. two EoT and one ToR unlike the heterogeneous scenario were intra-tray communication is used to process some request. This leads to an increase in network resources and power consumption which in turn implies that the heterogeneous architecture for disaggregation is more cost efficient.

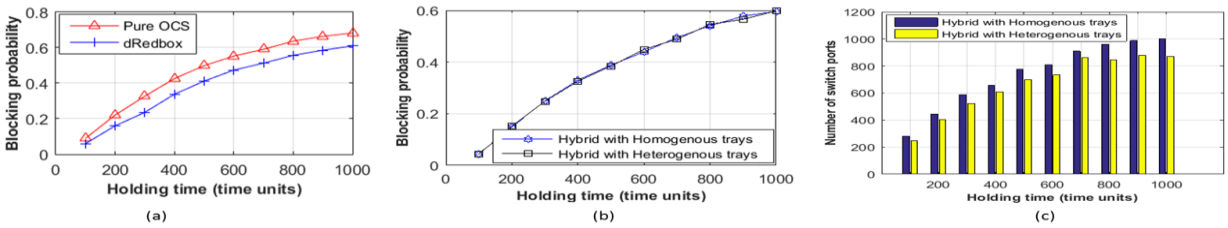

Fig. 3. Blocking probability of (a) dRedbox vs pure OCS (b) Heterogeneous vs Homogenous tray (c) Number of used switch ports

We then compare the performance of the dRedbox tray architecture presented in figure 1(c) to the classical statistically dimensioned hybrid architecture presented in figure 1(b). For the both dRedbox and classical hybrid tray architecture, the tray has 12 MPSoC bricks and 12 Memory bricks and each brick is interfaced with 32 transceivers that can be realized by four mid board optic each supporting 8 channels at 1310nms. The dRedbox tray is connected to two optical switches with 384 ports each while the classical hybrid tray is connected an EPS switch 384 ports with equivalent transceivers and an optical switch with 384 ports. The cost and power of each tray architecture is calculated using parameters listed in table 1 and are based on values in [2, 3]. As presented in figure 4(a), the dRedbox tray architecture demonstrates a considerable lower blocking probability (12.6 % of blocking probability lower at 1000 holding time unit) than the classical traditional architecture, this is because the classical hybrid architecture has fixed I/O ports for EPS and OCS while the in dRedbox architectures all I/O ports can be used for either for EPS and OCS, hence adapting to any kind of traffic flow. Figure 4(b) and 4(c) illustrates the total capacity per cost and watt respectively. It can be clearly observed that the dRebox tray architecture performs substantially better in terms of Cost (37.4% improvement) and power consumptions (873% improvement. This is due to power hungry electronic switch and additional transceivers required for EPS on the classical hybrid architecture.

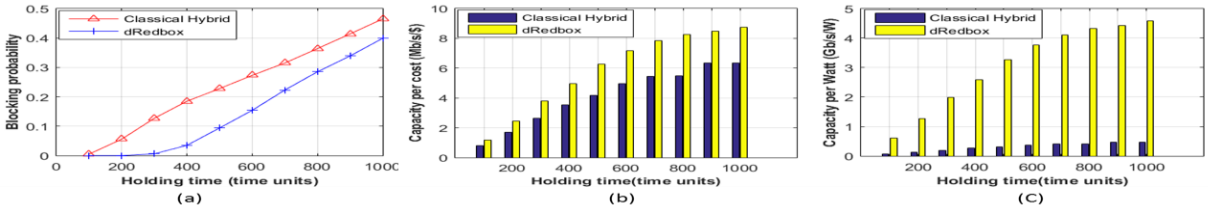

Fig. 4. dRedbox vs Traditional Hybrid (a) Blocking probability (b) Capacity per cost (c) Capacity per Watt

Table 1: Cost and Power consumption of switches and transceivers

Component	Cost ($)	Power(W)
Optical circuit switch port	500	0.05
Electronic packet switch port	500	12.5
10G transceiver	50	1

5. Conclusion

The dRedbox heterogeneous disaggregated rack architecture have shown potential benefits in terms of resource savings when compared homogenous and OCS disaggregated rack architecture. The dRedbox architecture offers substantial benefits in terms of CAPEX and OPEX has when compared to the classical hybrid architecture.

6. Acknowledgements

The work was supported by European Union's H2020 funded dRedBox project with grant agreement No.687632.

7. References

[1] S. Han et al., "Network Support for Resource Disaggregation in Next-Generation Datacenters", ACM (2013)
[2] N. Farrington, et al., Helios: A hybrid electrical/optical switch architecture for modular data centers. SIGCOMM, (2010)
[3] M. Imran, et al "Performance evaluation of hybrid optical switch architecture for data center networks," Opt. Switching Netw. (2016)
[4] K. Katrinis et al., "Rack-scale disaggregated cloud data centers: The dReDBox project vision," (DATE), Dresden, (2016).