On the Energy Efficiency of MapReduce Shuffling Operations in Data Centers

Sanaa Hamid Mohamed, Taisir E. H. El-Gorashi, Jaafar M. H. Elmirghani
School of Electronic & Electrical Engineering, University of Leeds, LS2 9JT, United Kingdom

ABSTRACT

This paper aims to quantitatively measure the impact of different data centers networking topologies on the performance and energy efficiency of shuffling operations in MapReduce. Mixed Integer Linear Programming (MILP) models are utilized to optimize the shuffling in several data center topologies with electronic, hybrid, and all-optical switching while maximizing the throughput and reducing the power consumption. The results indicate that the networking topology has a significant impact on the performance of MapReduce. They also indicate that with comparable performance, optical-based data centers can achieve an average of 54% reduction in the energy consumption when compared to electronic switching data centers.

Keywords: Data Center Networking (DCN), MapReduce, energy efficiency, completion time.

1. INTRODUCTION

The MapReduce programming model and its widely-used platform, Hadoop, are enabling several cost-effective cloud-based big data services [1]. These services typically require extensive all-to-all communications between hosting servers leading to increased congestion and power consumption in data centers. Moreover, they result in the East-West traffic dominating over the South-North traffic. This new traffic trend has become the focus in designing state-of-art production data centers [2]. These challenges are increasingly motivating the consideration of all-optical networking in future data centers to cope with the increasing demands of big data applications while improving the data centers performance and decreasing their power consumption [3].

The processing in MapReduce is composed of map, shuffle, and reduce phases. The input data is stored in several servers' local disks and is globally managed by a distributed file system (DFS) [1]. The processing starts by assigning map slots according to the number of input data chunks and available computing resources, and reduce slots according to the user's configurations. If chunks are more than map slots, the map phase will run in several waves according to their scheduling [4]. Each map slot processes it assigned chunks, preferably available locally, and generates intermediate results in the form of <key,value> pairs. The intermediate results are shuffled to reduce slots according to their keys where each slot is assigned to process a unique set of keys [1]. Finally, each reduce slot sorts its inputs, calculates final results, and saves them in the DFS.

Several optimization studies have been carried out by both academia and industry to enhance the performance and energy efficiency of big data applications (e.g. [2], [4]-[21]). The performance of big data applications and frameworks such as MapReduce is associated with a wide range of factors and parameters such as the cluster specifications (e.g. CPU, memory, networking, and disk I/O resources [9]), framework used or version, in addition to selected configurations and mechanisms for data and jobs placements and tasks scheduling [4]-[8]. Moreover, as the deployments of big data applications are evolving from dedicated clusters to public clouds where multi-tenants simultaneously request different big data services, additional challenges are encountered to orchestrate virtualized resources and lease of hybrid clusters at reduced cost and power consumption while avoiding under or over utilization [9]-[13].

Most of big data applications optimization studies neglected the impact of data centers networking and topologies on the performance by considering single-rack setups to reduce costs and to focus on optimizing the framework based on CPU, I/O, and memory resources (e.g. [4]-[7]). However, with growing data volumes and increasing need for clusters scaling, networking resources are expected to become a severe bottleneck [8]. Few studies tackled optimizing data center networking and topologies for big data applications [2], [17]-[21]. The authors in [2] considered utilizing the Energy Efficient Ethernet (EEE) feature enabled in modern data centers switches to achieve up to 60% reduction in MapReduce energy consumption. The network influence on Hadoop performance in multi-rack clusters is considered in [17]. The energy efficiency of several switch and server centric data centers is evaluated in [18] under MapReduce traffic. MRPerf which is a widely-used MapReduce simulator considered network configurations and compared the performance of double-rack clusters with DCell architecture [19]. Camdoop is proposed in [20] as a MapReduce-like system to exploit in-network aggregation of intermediate results in CamCube which is a server-centric architecture. In [21], the performance of MapReduce is experimentally examined in hybrid electronic/optical switching data centers namely c-Through and Helios. However, their optical links were not fully utilized.

This study aims to quantitatively measure the impact of different data centers topologies on the performance and energy efficiency of MapReduce shuffling operations. Different switching technologies for data centers are considered which are electronic, hybrid and all-optical. The rest of this paper is organized as follows. Section 2 briefly describes the data centers architectures considered. Section 3 explains the proposed optimization models and shows their results, while Section 4 provides the conclusions and future work.

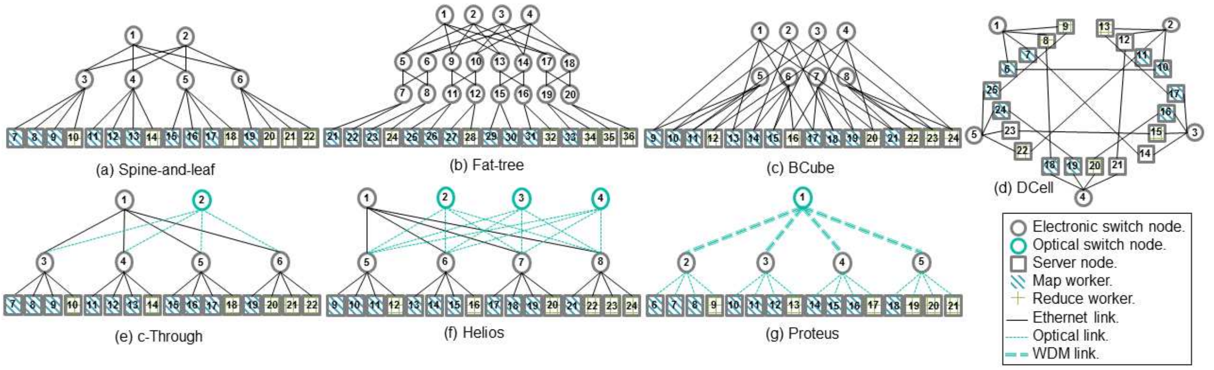


Figure 1: Electronic, hybrid, and all-optical-based data centers models for MapReduce shuffling optimization.

2. DATA CENTERS ARCHITECTURES:

Data Center Networking (DCN) is an important design aspect that determines the performance and power consumption of data centers and defines traffic routing requirements between their servers. Data centers have been hosting legacy web, data management, and content distribution applications, and are currently challenged by the increasing need to host big data applications [14]-[16]. These applications introduce different computing and routing requirements with denser server-to-server communications which cause increased congestion and power consumption [22], and [23].

2.1 Electronic Switching Data Centers:

State-of-art production data centers were characterized by having 3-tiers of switches; core, aggregation, and access that typically interconnect the servers under large oversubscription ratios that cause poor performance [24]. To reduce the oversubscription and increase the bisection bandwidth, the Spine-and-leaf architecture that utilizes modern spine, and leaf switches with high number of ports was introduced and is commercially used as in [24]. To achieve the same goals while utilizing commodity switches, novel architectures such as the Fat-tree [25], BCube [26], and DCell [27] were proposed. The Fat-tree architecture is composed of a layer of core switches and a pods layer composed of aggregation and access switches connected in a folded-Clos topology. The interconnections in Fat-tree provide multipath routes between servers and hence, Fat-tree provides higher fault-tolerance [25]. BCube and DCell are recursive architectures proposed to provide scalability and fault-tolerance in modular data centers. Both architectures utilize Network Interface Cards (NIC) attached to the servers in addition to few commodity switches to route traffic [26], and [27]. Example of small deployments for Spine-and-leaf, Fat-tree, BCube, and DCell data centers are illustrated in Figure 1 (a, b, c, and d) respectively.

2.2 Hybrid and All-Optical Data Centers:

To improve the performance of existing electronic switching data centers, several studies have suggested supporting the switching fabric with Optical Circuit Switching (OCS) capabilities such as in c-Through [28] and Helios [29] topologies. The high bandwidth availability of OCS is utilized for bulky transfers and slowly varying traffic, while the electronic packet switching is kept to handle bursty and fast-varying traffic. In these architectures, the OCS is realized through Micro-Electro-Mechanical System (MEMS) switches [3]. As MEMS switches control input to output connections by moving the mirrors mechanically, high reconfiguration time in the scale of few milliseconds is required. To avoid the inherit complexities of managing two systems in hybrid data centers, and to further exploit the capacity and speed capabilities of optical networks, several all-optical DCN architectures were proposed [3]. These architectures utilize optical passive and active components for networking such as couplers, splitters, Arrayed-Waveguide Grating (AWG), Tunable Wavelength Converters (TWC), and Wavelength Selective Switches (WSS). An example of all-optical data centers that uses MEMS and WSS technologies is Proteus [30]. Proteus is malleable to traffic as it adapts its optical connections and links capacities between top-of-rack (ToR) switches according to the demands. Examples of small designs based on c-Through, Helios, and Proteus are illustrated in Figure 1 (e, f, and g) respectively.

3. OPTIMIZING SHUFFLING OPERATIONS IN DATA CENTERS

To quantitatively assess the impact of the data center topology on the completion time and the energy efficiency of shuffling operations in MapReduce, we utilized Mixed Integer Linear Programming (MILP) models. The MILP models aim to optimize the routing required for intermediate data shuffling in different data centers while maximizing the throughput (i.e. served traffic) and reducing the power consumption.

3.1 Methodology

Four electronic switching data centers: Spine-and-leaf, Fat-tree, BCube, and DCell, in addition to two hybrid architectures: Helios and c-Through and an all-optical data center: Proteus are considered. The topologies of the aforementioned data centers are modelled as graphs where both switches and servers are considered as graphs nodes, while the bi-directional links are considered as graphs edges as illustrated in Figure 1. All bidirectional links are considered to have 10 Gbps capacity, except for the Proteus architecture where the links connected to the MEMS are 30 Gbps WDM links with three wavelengths (i.e. 10Gbps per wavelength) [30]. The power consumption and details of the electronic and optical equipment used in each topology are summarized in Tables 1 and 2. In all topologies, 16 servers (allocated in 4 racks or 8 racks in Fat-tree) are considered to accommodate the map and reduce workers. In the DCell model, 4 extra servers are considered due to the requirements of constructing the topology but are not assigned any additional workers. To simplify the evaluations of hybrid and all-optical data centers, 12×12 MEMSs are used to ensure all-to-all connectivity between ToRs without the need for reconfigurations. Also, fixed wavelength assignments with all-to-all ToR connectivity are assumed for Proteus while keeping the WSS components.

Table 1. Electronic-switching data centers characteristics and parameters.

Topology	No. of Servers	No. of links	Networking Devices characteristics		
			Equipment	Units	Power Consumption per unit in Watts
Spine-and-leaf	16	24	Nexus 3048	6	120 [24]
Fat-tree	16	48	SG500XG-8F8T	20	95 [32]
BCube	16	32	SG500XG-8F8T	8	95 [32]
			NIC	16	3 [22]
DCell	20	30	SG500XG-8F8T	5	95 [32]
			NIC	20	3 [22]

Table 2. Hybrid and all-optical switching data centers characteristics and parameters.

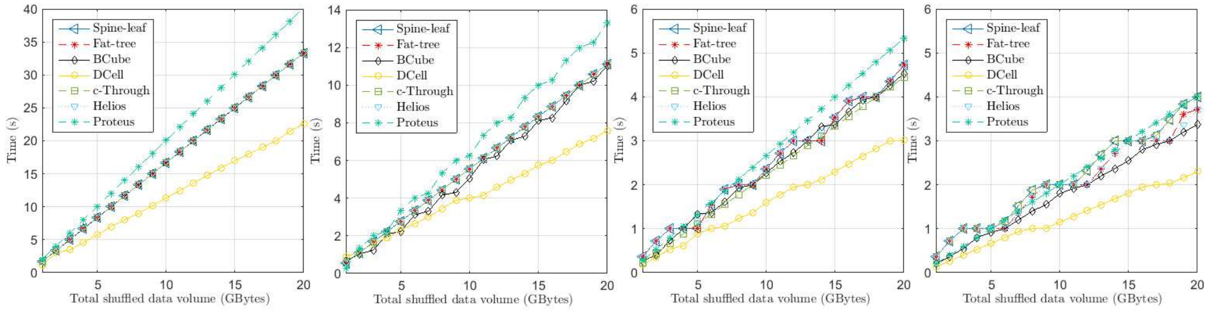
Topology	No. of Servers	No. of links	Networking Devices characteristics		
			Equipment	Units	Power Consumption per unit in Watts
c-Through	16	24	SG500XG-8F8T	5	95 [32]
			12×12 MEMS	1	2.88 (0.24 per port [30])
Helios	16	32	SG500XG-8F8T	5	95 [32]
			12×12 MEMS	3	2.88
Proteus	16	20	12×12 MEMS	1	2.88
			WSS	4	3 [30]

In this evaluation, 10 servers are dedicated for map functions and 6 for reduce functions which resembles a typical tasks ratio in original Google's MapReduce [1]. In this work, the placement of map and reduce workers is assumed to be as depicted in Figure 1. To effectively examine network bottlenecks, sort workloads are assumed. Sorting via MapReduce utilizes identity map functions to generate <word,1> pairs from large text files. The entire intermediate data is to be shuffled according to words to reduce workers in order to be sorted and finally saved. Hence, input, intermediate, and output data are all equal in size. The volume of total data to be sorted is varied from 1 GBytes to 20 GBytes and is equally distributed between map workers. The workloads are assumed to be from the Indy GraySort benchmark which have uniform intermediate key distributions due to balanced words count [31], and hence symmetric traffic is transferred from each map worker to 6 reduce workers.

The MILP models are utilized to minimize the energy consumption of the data center while prioritizing maximizing the throughput. The data write and read speeds from the servers are considered in the models where four maximum rate/server values are used. The considered rates are 100, 300, 750, and 1000 MBytes/s, which can be realized by HDD, SSD, with technologies such as Redundant Array of Independent Disks (RAID), or with any caching capabilities in NICs. For simplicity, all map tasks are assumed to finish at the same time, and the ON/OFF power model is used for the switches. The models optimize routing the traffic through minimum set of core and aggregation switches if the network is bottlenecked at the links to/from the ToRs. Increasing the rate/server reduces this bottleneck and allows for higher throughput. Hence, the models optimally utilize higher number of core and aggregation switches resulting in lower shuffling completion time.

3.2 Results

Figure 2 summarizes the shuffling completion time results, while Figure 3 provides the average power consumption results for different rate/server values. For 100 MBytes/s rate, the results indicate that the completion time is mainly determined by the I/O bottleneck at the servers to ToR switches links leading to a linear relation between completion time and data volume. As a result, further energy saving can be achieved except for BCube, DCell, and Proteus because they call for the use of all switches to accommodate the traffic at all rate/server values. As the rate/server increases, the DCNs increasingly become the bottleneck, and to achieve maximized throughput with lower completion time, more switches have to be used which leads to increased power consumption. For all rate/server values, the best achieved performance was for DCell, which has several server-to-server connections with properly placed map and reduce workers.



(a) Max rate/server = 100 MBytes/s (b) Max rate/server = 300 MBytes/s (c) Max rate/server = 750 MBytes/s (d) Max rate/server = 1000 MBytes/s

Figure 2: Shuffling completion time for different maximum rate/server.

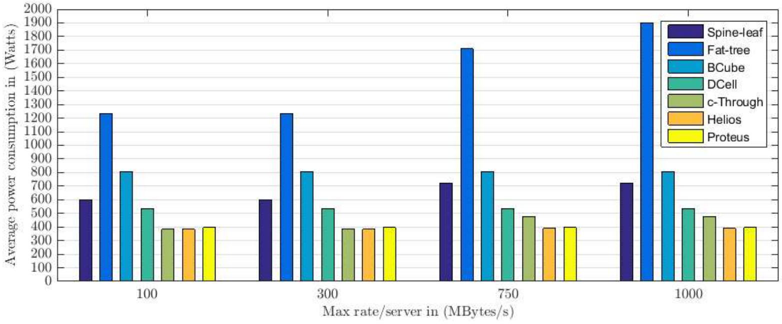


Figure 3: Average power consumption for different maximum rate/server.

4. CONCLUSIONS AND FUTURE WORK

In this paper, the impact of DCN topologies on the completion time and energy efficiency of MapReduce shuffling is quantified through MILP models under increasing shuffle traffic. The completion time results indicate that the topology of the data center and the maximum rate/server have a significant impact on the performance of MapReduce. Moreover, utilizing optical networking technologies in DCNs can achieve an average of 54% reduction in the power consumption when compared to electronic switching DCNs with comparable performance. The best performance was obtained by DCell, however, with larger DCell realizations, the performance superiority is expected to disappear as multi-hop connections are required between servers in different cells. Future work includes considering optimizing the power consumption and completion time while considering different workloads and more data center topologies.

ACKNOWLEDGEMENTS

The first author would like to acknowledge her PhD funding from the Engineering and Physical Sciences Research Council (EPSRC) and the School of Electronic and Electrical Engineering, University of Leeds. All data are provided in full in the results section of this paper.

REFERENCES

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

[2] R. F. e Silva and P. M. Carpenter, “Exploring interconnect energy savings under east-west traffic pattern of mapreduce clusters,” in Local Computer Networks (LCN), 2015 IEEE 40th Conference on, Oct 2015, pp. 10–18.

[3] C. Kachris and I. Tomkos, “Power consumption evaluation of all-optical data center networks,” Cluster Computing, vol. 16, no. 3, pp. 611–623, 2013.

[4] A. Verma, L. Cherkasova, and R. H. Campbell, “ARIA: Automatic Resource Inference and Allocation for Mapreduce Environments,” in Proceedings of the 8th ACM International Conference on Autonomic Computing, ser. ICAC ’11. New York, NY, USA: ACM, 2011, pp. 235–244.

[5] M. Zaharia, D. Borthakur, J. Sen Sarma, K. Elmeleegy, S. Shenker, and I. Stoica, “Delay Scheduling: A Simple Technique for Achieving Locality and Fairness in Cluster Scheduling,” in Proceedings of the 5th European Conference on Computer Systems, ser. EuroSys ’10, 2010, pp. 265–278.

[6] J. Polo, C. Castillo, D. Carrera, Y. Becerra, I. Whalley, M. Steinder, J. Torres, and E. Ayguadé, “Resource-aware Adaptive Scheduling for Mapreduce Clusters,” in Proceedings of the 12th ACM/IFIP/USENIX International Conference on Middleware, Middleware’11, (Berlin, Heidelberg), pp. 187–207, Springer-Verlag, 2011.

[7] J. Leverich and C. Kozyrakis, “On the Energy (in)Efficiency of Hadoop Clusters,” SIGOPS Oper. Syst. Rev., vol. 44, no. 1, pp. 61–65, Mar. 2010.

[8] Z. Ren, J. Wan, W. Shi, X. Xu, and M. Zhou, “Workload Analysis, Implications, and Optimization on a Production Hadoop Cluster: A Case Study on Taobao,” IEEE Transactions on Services Computing, vol. 7, pp. 307–321, April 2014.

[9] D. Xie, Y. C. Hu, and R. R. Kompella, “On the performance projectability of MapReduce,” in Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference on, Dec 2012, pp. 301–308.

[10] K. Chen, J. Powers, S. Guo, and F. Tian, “CRESP: Towards Optimal Resource Provisioning for MapReduce Computing in Public Clouds,” Parallel and Distributed Systems, IEEE Transactions on, vol. 25, no. 6, pp. 1403–1412, June 2014.

[11] A. M. Al-Salim, H. M. M. Ali, A. Q. Lawey, T. El-Gorashi and J. M. H. Elmirghani, "Greening big data networks: Volume impact," 2016 18th International Conference on Transparent Optical Networks (ICTON), Trento, 2016, pp. 1-6.

[12] A. M. Al-Salim, A. Q. Lawey, T. El-Gorashi and J. M. H. Elmirghani, "Energy Efficient Tapered Data Networks for Big Data processing in IP/WDM networks," 2015 17th International Conference on Transparent Optical Networks (ICTON), Budapest, 2015, pp. 1-5.

[13] L. Nonde, T. E. H. El-Gorashi and J. M. H. Elmirghani, "Energy Efficient Virtual Network Embedding for Cloud Networks," in Journal of Lightwave Technology, vol. 33, no. 9, pp. 1828-1849, May1, 1 2015.

[14] N. I. Osman, T. El-Gorashi, L. Krug and J. M. H. Elmirghani, "Energy-Efficient Future High-Definition TV," in Journal of Lightwave Technology, vol. 32, no. 13, pp. 2364-2381, July1, 1 2014.

[15] A. Q. Lawey, T. E. H. El-Gorashi and J. M. H. Elmirghani, "Distributed Energy Efficient Clouds Over Core Networks," in Journal of Lightwave Technology, vol. 32, no. 7, pp. 1261-1281, April1, 2014.

[16] A. Q. Lawey, T. E. H. El-Gorashi and J. M. H. Elmirghani, "BitTorrent Content Distribution in Optical Networks," in Journal of Lightwave Technology, vol. 32, no. 21, pp. 4209-4225, Nov.1, 1 2014.

[17] J. Han, M. Ishii, and H. Makino, “A Hadoop performance model for multi-rack clusters,” in Computer Science and Information Technology (CSIT), 2013 5th International Conference on, March 2013, pp. 265–274.

[18] Y. Shang, D. Li, J. Zhu, and M. Xu, “On the Network Power Effectiveness of Data Center Architectures,” Computers, IEEE Transactions on, vol. 64, no. 11, pp. 3237–3248, Nov 2015.

[19] G. Wang, A. R. Butt, P. Pandey, and K. Gupta, “A simulation approach to evaluating design decisions in MapReduce setups,” in Modeling, Analysis Simulation of Computer and Telecommunication Systems, 2009. MASCOTS ’09. IEEE International Symposium on, Sept 2009, pp. 1–11.

[20] P. Costa, A. Donnelly, A. Rowstron, and G. O’Shea, “Camdoop: Exploiting In-network Aggregation for Big Data Applications,” in Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation, ser. NSDI’12, 2012, pp. 3–3.

[21] H. H. Bazzaz, M. Tewari, G. Wang, G. Porter, T. S. E. Ng, D. G. Andersen, M. Kaminsky, M. A. Kozuch, and A. Vahdat, “Switching the Optical Divide: Fundamental Challenges for Hybrid Electrical/Optical Datacenter Networks,” in Proceedings of the 2Nd ACM Symposium on Cloud Computing, ser. SOCC ’11, 2011, pp. 30:1–30:8.

[22] László Gyarmati and Tuan Anh Trinh. 2010. How can architecture help to reduce energy consumption in data center networking?. In Proceedings of the 1st International Conference on Energy-Efficient Computing and Networking (e-Energy '10). ACM, New York, NY, USA, 183-186.

[23] A. Hammadi, T. E. H. El-Gorashi, and J. M. H. Elmirghani, “high performance AWGR PONs in data centre networks,” in 2015 17th International Conference on Transparent Optical Networks (ICTON).

[24] Pall Beck, Peter Clemens, Santiago Freitas, Jeff Gatz, Michele Girola, Jason Gmitter, Holger Mueller, Ray O'Hanlon, Veerendra Para, Joe Robinson, Andy Sholomon, Jason Walker, Jon Tate, “IBM and Cisco: Together for a World Class Data Center”, IBM Redbooks, July 31, 2013.

[25] M. Al-Fares, A. Loukissas, and A. Vahdat, “A Scalable, Commodity Data Center Network Architecture,” SIGCOMM Comput. Commun. Rev., vol. 38, pp. 63–74, Aug. 2008.

[26] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y.Shi, C. Tian, Y. Zhang, and S. Lu., “BCube: a high performance, server-centric network architecture for modular data centers”. In Proceedings of the ACM SIGCOMM 2009 conference on Data communication (SIGCOMM '09). ACM, New York, NY, USA, 63-74.

[27] C. Guo, H. Wu, K. Tan, L. Shi, Y. Zhang, and S. Lu. 2008. Dcell: a scalable and fault-tolerant network structure for data centers. SIGCOMM Comput. Commun. Rev. 38, 4 (August 2008), 75-86.

[28] Guohui Wang, David G. Andersen, Michael Kaminsky, Konstantina Papagiannaki, T.S. Eugene Ng, Michael Kozuch, and Michael Ryan. 2010. c-Through: part-time optics in data centers. SIGCOMM Comput. Commun. Rev. 40, 4 (August 2010), 327-338.

[29] N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, “Helios: a hybrid electrical/ optical switch architecture for modular data centers,” SIGCOMM Comput. Commun. Rev., vol. 41, no. 4, Aug. 2010.

[30] A. Singla, A. Singh, K. Ramachandran, L. Xu, and Y. Zhang, “Proteus: A Topology Malleable Data Center Network,” in Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks, ser. Hotnets-IX, 2010, pp. 8:1–8:6.

[31] A. Rasmussen, G. Porter, M. Conley, H. V. Madhyastha, R. N. Mysore, A. Pucher, and A. Vahdat, “TritonSort: A Balanced and Energy Efficient Large-Scale Sorting System,” ACM Trans. Comput. Syst., vol. 31, no. 1, pp. 3:1–3:28, Feb. 2013.

[32] Cisco 500 Series Stackable Managed Switches Data Sheet. (Accessed on 2017, April). [Online]. Available: http://www.cisco.com/c/en/us/products/collateral/switches/small-business-500-series-stackable-managed-switches/c78-695646_data_sheet.html