eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Impact of Link Failures on the Performance of MapReduce in Data Center Networks

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

ABSTRACT

In this paper, we utilize Mixed Integer Linear Programming (MILP) models to determine the impact of link failures on the performance of shuffling operations in MapReduce when different data center network (DCN) topologies are used. For a set of non-fatal single and multi-links failures, the results indicate that different DCNs experience different completion time degradations ranging between 5% and 40%. The best performance under links failures is achieved by a server-centric PON-based DCN.

Keywords: Data Center Networking (DCN), Passive Optical Networks (PON), Resilience, MapReduce, energy efficiency, completion time.

1. INTRODUCTION

Providing high reliability and continuity is a key requirement for cloud computing services as any disruption or disconnection in the infrastructure typically leads to revenue losses and customer departures. Failures in cloud infrastructure can occur in the transporting networks, data centers, or in the applications [1]. Overcoming the severe impacts of these failures requires the adoption of resilient designs and restoration plans. The resilience and energy efficiency of different IP over WDM networks topologies were examined in [2] under fiber cuts or a core node failure. The work in [3] and [4] considered static and dynamic adaptation to traffic surges resulting from re-routing demands after links failures. In [5], an energy efficient 1+1 protection scheme in IP over WDM networks was proposed based on Network Coding (NC) while considering the impact of the topology and the traffic. The authors in [6] addressed the resilience of geo-distributed transport networks that link cloud data centers by jointly considering the content placement, and anycast routing under failures. At the application level, solutions such as maintaining several replicas of data and software components are considered. For example, MapReduce, which is a widely used framework for big data parallel computations, has a data replication mechanism with default value of 3, and considers speculative execution for straggler tasks [7]. Several studies have considered the resilience of Data Center Networks (DCN) (e.g. [8]-[11]) by addressing the sources of links, switches, and servers' software and hardware failures.

In this paper, we examine single and multi-link non-fatal failures in several DCNs and their effect on the performance of MapReduce shuffling operations. We compare the completion time results in [12] and [13] under no link failures to the results under several link failures for electronic, hybrid, and optical DCNs, in addition to a server-centric Passive Optical Networks (PON)-based DCN design proposed as part of the work in [14]-[19]. The rest of this paper is organized as follows. Section 2 describes the proposed optimization models and shows their results, while Section 3 provides the conclusions and future work.

2. OPTIMIZING SHUFFLING OPERATIONS IN DATA CENTERS UNDER LINKS FAILURES

2.1 Methodology

We utilize Mixed Integer Linear Programming (MILP) models to examine the effects of single and multi-link failures in different DCNs on the optimum routing of intermediate data while balancing the completion time and the energy efficiency. Seven DCNs [12] namely Spine-and-Leaf, Fat-Tree, BCube, DCell, c-Through, Helios, in addition to the server-centric PON-based DCN design in [16] with 4 racks and 1 server per group, are considered. Figure 1 illustrates the assignment of map and reduce workers in 10 and 6 servers, respectively in each DCN model. It also shows the link disconnection cases considered. Indy GraySort benchmark [20] is selected as it is a representative workload for examining the congestion in DCNs due to routing intermediate data that is equal in size to the input data from the 10 map workers to the 6 reduce workers. The power consumption values of the selected electronic and optical equipment are as in [12], and [15].
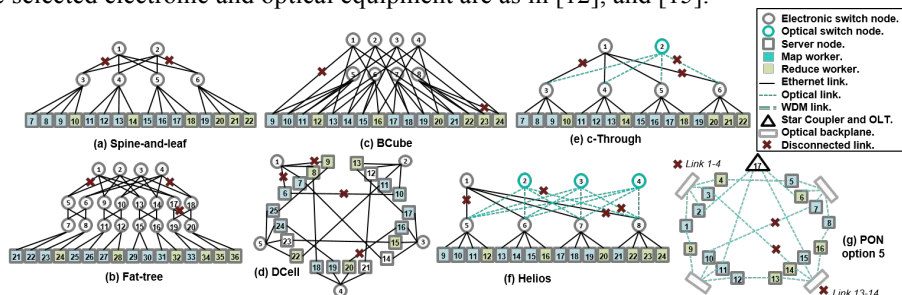

Figure 1: DCN models considered for MapReduce shuffling optimization under different links failures.

2.2 Results

Figure 2 provides the shuffling completion time results for a maximum server rate of 1000 Mbytes/s and input data ranging from 1 to 20 GBytes in different DCNs. The subfigures, show the results in [12], and [13] for the cases where all links are working, in addition to the completion time results under 2 or 3 cases of links failures in each DCN. For the Spine-and-Leaf architecture, the degradation due to links 1-3 disconnection is less than 2-6 because the latter is more utilized due to serving the flows to 3 reduce workers compared to 1 reduce worker by links 1-3. In the Fat-Tree architecture, we examined the impact of disconnecting the most utilized links (i.e. 17-20). Any additional links disconnections did not deteriorate the performance. In both DCNs, failure in any link between the Top of Rack (ToR) and a server is considered a single point-of-failure and requires a software-based resilience mechanism such as replication, where the map computations should be re-performed in another server containing a replica. The BCube and DCell architectures are server-centric DCNs with two links to each server which provides more routes compared to the Spine-and-Leaf architecture and the Fat-Tree architecture, and hence, are more resilient to link failures. Single and multi-link failures including the most utilized link in BCube caused an increase of 24% in the completion time. The DCell architecture exhibited the highest degradation in the completion time (i.e. 40%) with links 1-6 disconnection which are the most utilized due to serving two reduce workers from other racks, which indicates high sensitivity to workload placements. An average increase by 37% in the completion time is experienced in the c-Through architecture and by 5% in the Helios architecture for three link failures because more optical switches are utilized for the same number of ToR switches. Both however have a single-point-of-failure as in the Fat-Tree and Spine-and-Leaf architectures. The server-centric PON-based design exhibited an average degradation of 22% under single and up to 4 link failures. Figure 3 summarizes the overall completion time (i.e. completion time without failures in addition to the extra delay due to failures) for the cases considered in Figure 2.
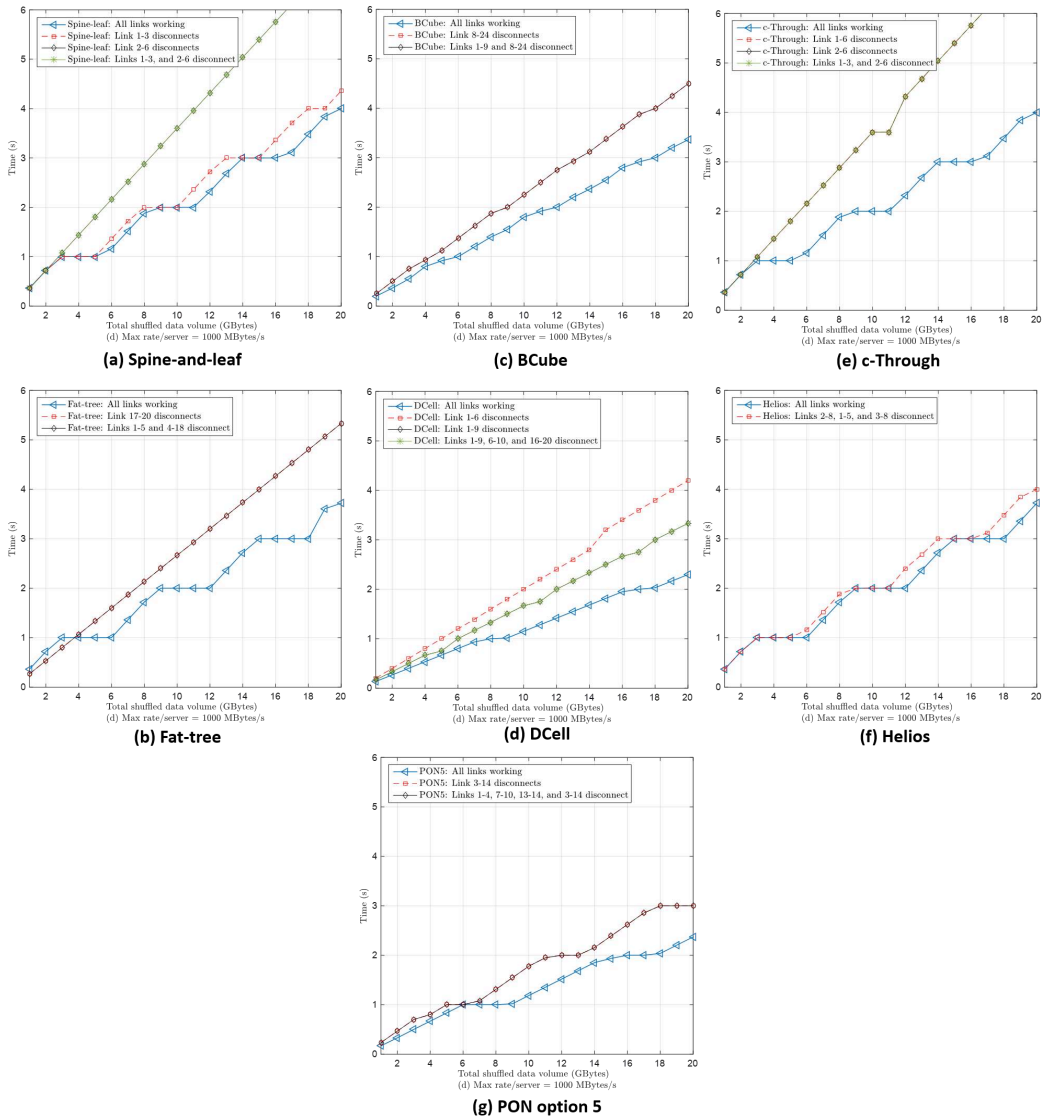

Figure 2: Shuffling completion time in different DCNs under links failure for max. rate/server of 1000 MBytes/s.
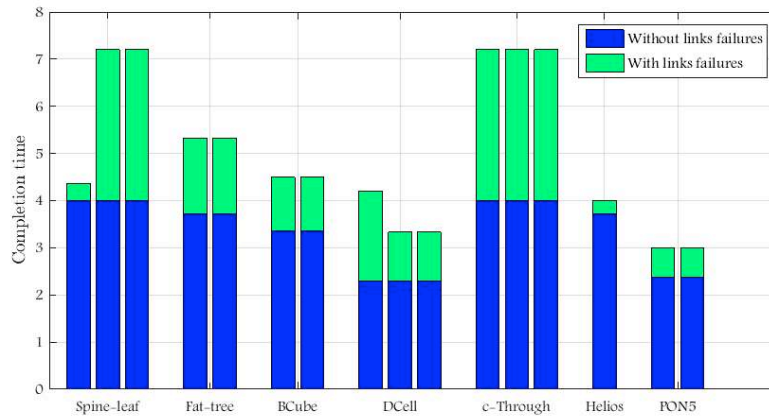
Figure 3: Overall completion time for total shuffled data volume of 20 GBytes.

3. CONCLUSIONS AND FUTURE WORK

In this paper, we examined the impact of single and multi-link failures in different DCN topologies under extensive server-to-server communications. The results for a set of non-fatal single and multi-link failures indicate that different DCNs experience different degradations based on the links redundancy, workloads placement, and links utilizations. The best overall performance was achieved by the server-centric PON-based DCN in [16]. Future work includes considering the switching delay, switches failures, and the data replication factor.

ACKNOWLEDGEMENTS

REFERENCES

[1] C. Colman-Meixner, C. Develder, M. Tornatore and B. Mukherjee, "A Survey on Resiliency Techniques in Cloud Computing Infrastructures and Applications," in *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 2244-2281, thirdquarter 2016.

[2] T. E. H. El-Gorashi, X. Dong, A. Lawey and J. M. H. Elmirghani, "Core network physical topology design for energy efficiency and resilience," 2013 15th International Conference on Transparent Optical Networks (ICTON), Cartagena, 2013, pp. 1-7.

[3] Z. H. Nasralla, T. E. H. El-Gorashi, M. O. I. Musa and J. M. H. Elmirghani, "Routing post-disaster traffic floods in optical core networks," *2016 International Conference on Optical Network Design and Modeling (ONDM)*, Cartagena, 2016, pp. 1-5.

[4] Z. H. Nasralla, M. O. I. Musa, T. E. H. El-Gorashi and J. M. H. Elmirghani, "Routing post-disaster traffic floods heuristics," *2016 18th International Conference on Transparent Optical Networks (ICTON)*, Trento, 2016, pp. 1-4.

[5] M. Musa, T. Elgorashi and J. Elmirghani, "Energy efficient survivable IP-over-WDM networks with network coding," in *IEEE/OSA Journal of Optical Communications and Networking*, vol. 9, no. 3, pp. 207-217, March 2017.

[6] M. F. Habib, M. Tornatore, M. De Leenheer, F. Dikbiyik and B. Mukherjee, "Design of Disaster-Resilient Optical Datacenter Networks," in *Journal of Lightwave Technology*, vol. 30, no. 16, pp. 2563-2573, Aug.15, 2012.

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

[8] Rahul Potharaju and Navendu Jain. 2013. When the network crumbles: an empirical study of cloud network failures and their impact on services. In Proceedings of the 4th annual Symposium on Cloud Computing (SOCC '13). ACM, New York, NY, USA, Article 15, 17 pages.

[9] Rahul Potharaju and Navendu Jain. 2013. An empirical analysis of intra- and inter-datacenter network failures for geo-distributed services. SIGMETRICS Perform. Eval. Rev. 41, 1 (June 2013), 335-336.

[10] Phillipa Gill, Navendu Jain, and Nachiappan Nagappan. 2011. Understanding network failures in data centers: measurement, analysis, and implications. In Proceedings of the ACM SIGCOMM 2011 conference (SIGCOMM '11). ACM, New York, NY, USA, 350-361.

[11] Peter Bodík, Ishai Menache, Mosharaf Chowdhury, Pradeepkumar Mani, David A. Maltz, and Ion Stoica. 2012. Surviving failures in bandwidth-constrained datacenters. In Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication (SIGCOMM '12). ACM, New York, NY, USA, 431-442.

[12] S. H. Mohamed, T. E. H. El-Gorashi and J. M. H. Elmirghani, "On the energy efficiency of MapReduce shuffling operations in data centers," *2017 19th International Conference on Transparent Optical Networks (ICTON)*, Girona, 2017, pp. 1-5.

[13] S. H. Mohamed, T. E. H. El-Gorashi and J. M. H. Elmirghani, "Energy Efficiency of Server-Centric PON Data Center Architecture for Fog Computing," *Submitted to ICTON 2018*.

[14] J. Elmirghani, T. EL-GORASHI, and A. HAMMADI, "Passive optical-based data center networks," 2016, wO Patent App. PCT/GB2015/053,604.

[15] 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), Budapest, 2015, pp. 1-5.

[16] A. Hammadi, T. E. H. El-Gorashi, M. O. I. Musa and J. M. H. Elmirghani, "Server-centric PON data center architecture," *2016 18th International Conference on Transparent Optical Networks (ICTON)*, Trento, 2016, pp. 1-4.

[17] R. Alani, A. Hammadi, T. E. H. El-Gorashi and J. M. H. Elmirghani, "PON data centre design with AWGR and server based routing," *2017 19th International Conference on Transparent Optical Networks (ICTON)*, Girona, 2017, pp. 1-4.

[18] A. Hammadi, M. Musa, T. E. H. El-Gorashi and J. H. Elmirghani, "Resource provisioning for cloud PON AWGR-based data center architecture," *2016 21st European Conference on Networks and Optical Communications (NOC)*, Lisbon, 2016, pp. 178-182.

[19] A. Hammadi, T. E. H. El-Gorashi and J. M. H. Elmirghani, "Energy-efficient software-defined AWGR-based PON data center network," *2016 18th International Conference on Transparent Optical Networks (ICTON)*, Trento, 2016, pp. 1-5.

[20] 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.