Programmable Fabrics with Optical Switches in AI Supercomputers

Nikos Terzenidis[1], Giannis Patronas[1], Dimitris Syrivelis[1], Eitan Zahavi[2], Athanasios Fevgas[1], Nikos Argyris[1], Prethvi Kashinkunti[3], Louis Capps[3], Zsolt-Alon Wertheimer[2], Chen Avin[2], Julie Bernauer[3], Elad Mentovich[2], Paraskevas Bakopoulos[1]

[1] NVIDIA, Ermou 56, Athens 10563, Greece
[2] NVIDIA, 2788 San Tomas Expressway, Santa Clara, California 95051, USA
[3] NVIDIA, Hakidma 26, Ofer Industrial Park, Yokneam 2069203, Israel
*nterzenidis@nvidia.com

Abstract: *We explore the integration of Optical Circuit Switches (OCSs) in AI/HPC clusters' fabrics, to enhance resiliency in case of failures and enable dynamic topology reconfiguration for optimized deep-learning training, using a Layer-1 Software-Defined Network approach.*

Keywords: *Optical packet/burst/flow switching networks, Software defined networking (SDN) and OpenFlow/GMPLS, Optical network design, control, management, and security, Large-scale switching technologies for data center (DC) and high-performance computing (HPC) systems*

I. INTRODUCTION

The rise of Artificial Intelligence and Machine Learning (AI/ML) applications have gradually transformed traditional DataCenters (DC), into specifically-built AI factories, in order to support the computational demands of deep learning (DL) workloads. At the same time, soaring model sizes and computational requirements of state-of-the-art models [1], necessitate large scale-out systems, where the network fabric is an integral part enabling the exchange of enormous amount of data across the GPUs [2] and directly affecting system performance, power, cost and availability. However, architectural innovations both in terms of system-, as well as network-architecture, emerge as a prerequisite to sustain scaling at this pace, as DC deployments are currently capped by the maximum power available on-site.

To this end, Optical Circuit switches (OCS) [3-5] are being deployed in volume in Data Center and AI systems [6-7] transforming network cabling from a rigid infrastructure into a fully programmable resource that enables physical topology reconfiguration at runtime and delivers improved energy efficiency, hardware resiliency, incremental deployment and physical isolation for the network fabric. To realize this innovative functionality portfolio, extensions are needed on the Software-Defined-Networking (SDN) stack, that currently supports programmability only down to Layer-2 of the network.

This paper extends our previous work [8-9] and presents the latest findings of our research towards integrating OCSs in AI/HPC clusters' networks, via a Layer-1 (L1) SDN approach. We investigate two use-cases empowered by OCS-enabled networks; i) enhanced fabric resiliency in case of network hardware failures, and ii) optimized DL training enabled by dynamic topology reconfiguration. Finally, we investigate the use of BiDirectional (BiDi) transceivers in OCS fabrics towards reducing the number of required OCS ports.

II. INTEGRATING OPTICAL SWITCHES INTO AI/HPC NETWORKS

With Fat-Tree (FT) architectures offering full bisection bandwidth across all interconnected nodes, 2- or 3-level FT fabrics have been designated as the preferred network choice in AI clusters [2]. Figure 1 presents the main integration points of optical switches in a generic 3-level FT topology. By intercepting the fiber connections between the different node and packet switch layers, respective OCS layers are introduced, providing a L1 programmable dataplane for the attached network endpoints. When accompanied by redundant hardware (packet-switches, transceivers and servers), OCS-enabled architectures can provide resiliency against HW and SW failures in the fabric, minimizing their significant impact on utilization and efficiency of computing clusters [11-13]. Moreover, OCS layers can be exploited towards creating flatter networks, in conjunction with the elimination of respective packet switch layers. For example, the OCS Core Layer can establish direct connections between the Spine packet switches, eliminating the need for a Core packet switch layer.

From the control plane perspective, the incorporation of OCSs in the fabric introduces a new set of challenges on the network's control plane that needs to update the physical network topology to mitigate HW failures, or to continuously adapt it to the workload/traffic-pattern changes. In this context, appropriate resource management extensions need to be

developed for OCS-enabled networks that will ensure establishment of the required circuits and optimal bandwidth allocation for all the workloads that co-reside on the cluster at any given time.

It is important to note that despite most of the work and demonstrations presented here focus on the scale-out part of the network (Ethernet/Infiniband), the underlying principles can also be applied in other networks (eg NVLINK).

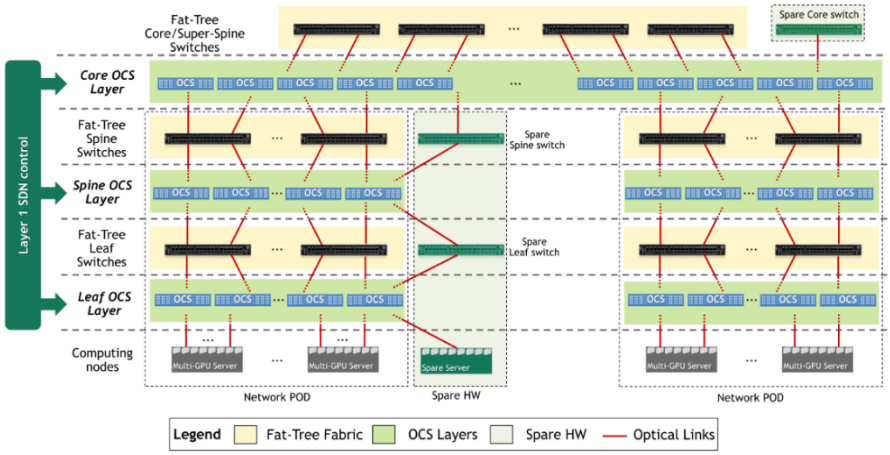

Fig. 1. Integration points of optical switches in a 3-layer Fat-Tree network topology

III. OPTICALLY-ENABLED FABRIC RESILIENCY

In order to demonstrate hardware resilience over an L1 programmable dataplane we deployed a small-scale OCS testbed, that is depicted in Fig. 2 (a). The setup consists of 2 general-purpose servers and 4 Infiniband Quantum2 switches, connected as 2 Leaves, 1 Spine and 1 redundant Spine (RS). The servers and switches were populated with 400Gb/s NDR FR4 transceivers, extending our previous evaluation [10] with the latest generation of Infiniband devices, while a single commercial OCS [3] was used to accommodate all the Leaf-to-Spine connections.

The OCS is controlled by a custom L1 control plane software that serves as an SDN stack extension for physical layer resources. The developed software framework leverages a representation of the cluster's physical network that includes also the optical switching elements - meant to be transparent to L2 and above. Moreover, it introduces a collection of concepts and algorithms that allow the described SDN L1 controller to identify the different topology configurations for a given deployment, carry out the physical topology changes through the OCSs and signal the L2 layer SDN controller to adapt to the changes of the physical network, enabling this way seamless automatic failover to redundant hardware.

We orchestrate an IB Spine switch failure and trigger the L1 SDN control plane to initiate the failure mitigation process towards the redundant Spine switch. Figure 2 (b) shows the results of our tests with OSU benchmarks [14]. The telemetry measurements show the bandwidth of the IB interfaces of a server over time, for an MPI bandwidth benchmark. Such a failure would normally result in an application crash, since the Spine switch provides the only available path between the two Leaf switches, and the IB interface going offline until the failure is fixed. When the resilience functionality is enabled, the full performance of the cluster is restored, in both cases, within a few seconds, as shown in the graph

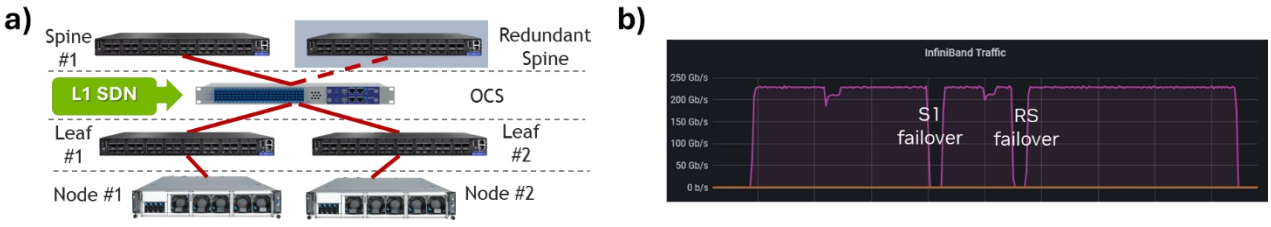

Fig. 2. a) Resilience architecture and experimental setup. b) Bandwidth recovery demonstration with OSU BW benchmark on Spine switch failover.

IV. OPTIMIZED DEEP LEARNING TRAINING VIA DYNAMIC TOPOLOGY RECONFIGURATION

The second application we explored for the L1 programmable data plane is training in DL clusters. The structured nature of DL training jobs allows for cross layer (compute-networking) optimization by tailoring the network topology to the application traffic patterns, towards unlocking significant benefits in terms of network power, cost and latency.

The underlying principle of operation leverages the programmable data plane to adjust the network topology according to the expected communication pattern for each job [15-16]. Our approach assumes knowledge of basic information on the jobs arriving to the cluster, such as the collectives and parallelization approach followed. For each job and allocation of ranks, the programmable data plane implements a suitable topology (e.g. torus, full graph etc.) that aims to provide full bandwidth for the expected traffic pattern. We have studied heuristics for resource management targeting LLM (3D Megatron) and DLRM jobs. For the current study, we rely on NVIDIA DGX servers, with multiple GPUs on each server interconnected through NVLink, along with separate connections to the programmable data plane through InfiniBand DPUs. The result is the architecture presented in Fig 3 a): a flat, low cost, low power and low latency network.

An example of adapting the topology to job requirements is given in Fig. 3 b). A job arrives requiring 1024 GPUs (i.e. 128 DGXs) and exhibiting an all-to-all (e.g. DLRM) pattern among them. For ease of presentation, the example allocates the 128 nodes under fully populated Leaf switches. In this case, the optimal topology is a full graph (per rail): equal distribution of connections from every Leaf to all the other Leaf switches that are part of this job. By reconfiguring the topology to a full graph, we provide full bandwidth for the all-to-all traffic pattern and ensure low latency (since there are at most 2-hops). We have extended the same principle to LLM jobs that use 3D parallelism [9].

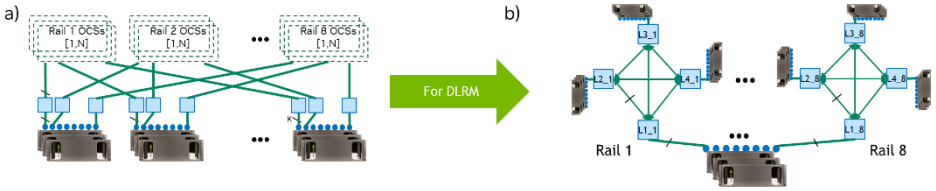

Fig. 3. a) Flat, reconfigurable network using OCSs, b) reconfiguring the topology in a full graph for all-to-all DLRM jobs

V. COMBINING OPTICAL SWITCHES WITH BIDIRECTIONAL TRANSCEIVERS

Large-scale fabrics necessitate the use of multiple OCS devices, hindering this way the cost-effectiveness of the solution, while the limited port count of the OCSs can potentially impact the performance due to inefficient link-allocation schemes. The first line of defense towards reducing the required OCS ports is the use of wavelength-multiplexed optics, e.g. FR4 optical transceivers, that multiplex 4 wavelengths into a single fiber, reducing the number of ports by 4x compared to parallel DR4 optics. Despite this reduction, FR4 transceivers still use 2 fibers, 1 per direction (Tx/Rx), and thus require 2 ports/transceiver on the OCS side, leaving room for significant improvement. To this end, BiDi optics emerge as a necessity [6], halving the required OCS ports compared to FR4 and leading to a total 8x reduction against DR4 transceivers. However, the use of a single fiber for both Tx and Rx signals introduces a new set of limitations on the connectivity scheme of endpoints to OCS devices that need to be addressed both through the architecture of the fabric, as well as through modifications on the respective cluster control-plane algorithms.

Figure 4 illustrates a comparison analysis, in terms of cluster utilization over time, between an OCS-enabled cluster with 32K GPU nodes and unidirectional optics against an identical cluster with BiDi optics. Both systems are evaluated using a job trace that includes 10,000 DLRM-only jobs, where each job requires a random number of nodes uniformly distributed between 64-1024 (steps of 64) GPUs. The performance is evaluated under various link allocation algorithms that distribute the available uplinks between the multiple OCS devices towards minimizing the blocking factor from optical circuit allocation. We show that the integration of BiDi optics induces no practical performance impact, as system utilization remains almost identical and in certain cases slightly better to the unidirectional optics system.

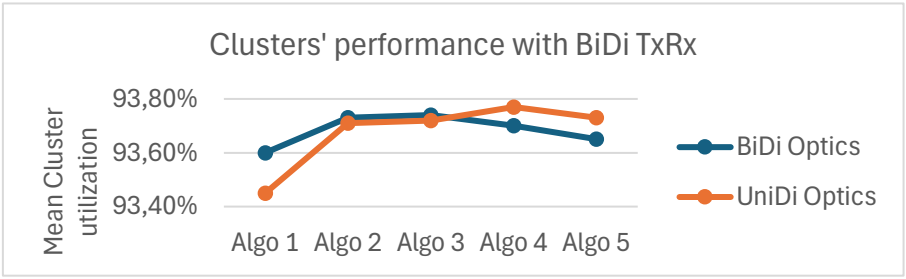

Fig. 4. Cluster utilization results comparing an OCS-enabled cluster with 32K GPU nodes and unidirectional optics against an identical cluster with BiDi optics, under 5 different uplink allocation algorithms.

ACKNOWLEDGMENT

This work has received funding from the European Union under the Horizon Europe projects: ALLEGRO (101092766), PHORMIC (101070332), photonixFAB (101111896), and MLSysOPS (101092912), PUNCH(101070560).

REFERENCES

[1] Epoch AI, “Machine learning trends,” Epoch AI, https://epochai.org/trends (accessed May 30, 2024).

[2] Nvidia DGX superpod: Next generation scalable Infrastructure for AI Leadership, https://docs.nvidia.com/https:/docs.nvidia.com/dgx-superpod-reference-architecture-dgx-h100.pdf (accessed May 30, 2024).

[3] “Series 7000 - 384X384 port software-defined optical circuit switch,” Polatis, http://www.polatis.com/series-7000-384x384-port-software-controlled-optical-circuit-switch-sdn-enabled.asp (accessed May 30, 2024).

[4] “Optical Circuit Switch for Data Centers - live demo at OFC 2024 based on Ult,” Optical Circuit Switch for Data Centers - Live Demo at OFC 2024 Based on Ultrareliable DLX Technology, https://www.coherent.com/news/press-releases/optical-circuit-switch-for-data-centers-live-demo-at-ofc-2024-based-on-ultrareliable-dlx-technology (accessed May 30, 2024).

[5] “Photonic optical circuit switching: Calient Technologies,” CALIENT Technologies -, https://www.calient.net/ (accessed May 30, 2024).

[6] U. Ryohei, *et al*. "Mission Apollo: landing optical circuit switching at datacenter scale." *arXiv preprint arXiv:2208.10041* (2022).

[7] H. Liu *et al.*, “Lightwave fabrics: AT-scale optical circuit switching for Datacenter and machine learning systems,” *2024 IEEE 37th International Conference on Micro Electro Mechanical Systems (MEMS)*, Jan. 2024. doi:10.1109/mems58180.2024.10439411

[8] G. Patronas *et al.*, “Software-defined, programmable L1 dataplane: Demonstration of fabric hardware resilience using optical switches,” *OFC 2023*, 2023. doi: 10.1364/ofc.2023.th2a.15

[9] P. Bakopoulos, *et al.*, "Photonic switched networking for data centers and advanced computing systems," in *Optical Fiber Communication Conference (OFC) 2024*, Technical Digest Series (Optica Publishing Group, 2024), paper M2G.1.

[10] G. Patronas *et al.*, “(OFC 2024) Optical Switching for data centers and advanced computing systems,” *Journal of Optical Communications and Networking*, Oct. 2024, doi: https://doi.org/10.1364/jocn.534317.

[11] S. Zhang *et al.*, “OPT: Open Pre-trained Transformer Language Models,” *arXiv (Cornell University)*, May 2022, doi: https://doi.org/10.48550/arxiv.2205.01068.

[12] J. Meza, T. Xu, K. Veeraraghavan, and O. Mutlu, “A Large Scale Study of Data Center Network Reliability,” *Proceedings of the Internet Measurement Conference 2018*, Oct. 2018, doi: https://doi.org/10.1145/3278532.3278566.

[13] R. Singh, M. Mukhtar, A. Krishna, A. Parkhi, J. Padhye, and D. Maltz, “Surviving switch failures in cloud datacenters,” *ACM SIGCOMM Computer Communication Review*, vol. 51, no. 2, pp. 2–9, Apr. 2021, doi: https://doi.org/10.1145/3464994.3464996.

[14] “MVAPICH :: Benchmarks,” *Ohio-state.edu*, 2025. https://mvapich.cse.ohio-state.edu/benchmarks/

[15] N. Jouppi *et al.*, “TPU V4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings,” *Proceedings of the 50th Annual International Symposium on Computer Architecture*, Jun. 2023. doi:10.1145/3579371.3589350

[16] W. Weiyang, *et al.* "{TopoOpt}: Co-optimizing Network Topology and Parallelization Strategy for Distributed Training Jobs." *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. 2023.