On the Performance of a Fast Optically Switched Network for Machine-Learning Accelerator Clusters

M.P.G. Rombouts,[1,*] and N. Calabretta,[1]

[1] *Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands*

[*] *m.p.g.rombouts@tue.nl*

Abstract: We investigate the viability of optically switched network for ML accelerator clusters and compare it to a leaf-spine network with 256/1024 GPUs. Results show almost ideal throughput, sub-µs latency and zero packet-loss for ¡0.6 traffic-load. © 2022 The Author(s)

1. Introduction

The integration of Machine Learning (ML) models in our daily lives has caused an increasing demand in compute power. With the yearly doubling of compute for large scale models, the model size and trained parameters increase proportionally [1]. Developments in accelerator hardware support this growth, but large-scale models such as Google's PaLM [2] with 540 billion parameters, require a large-scale distributed approach involving significant traffic streams through the network. Interconnection technologies such as PCI/e, InfiniBand and Ethernet are often identified as a bottleneck in ML accelerated cluster networks [3].

Combining network bottlenecks with the limited I/O density seen in electronic systems and chips [5] and the power efficiency of electronic interconnects, there is a need to improve off-chip bandwidth using efficient integrated optical technologies. Various promising efforts exist, e.g., in [6], a 1.6 Tbps chiplet has been demonstrated which can be integrated next to processing units. Designing an optical network using transparent optical switches not only increases the bandwidth, but also reduces latency and power consumption by removing O/E/O conversions and its associated latency. OPSquare [7] is a network providing these features, as its flat architecture enables high-throughput and scalability.

In this paper we assess whether an optically switched network would be an alternative to existing ML distributed network designs. We will adapt the optical network OPSquare for ML applications and compare it to a state-of-the-art GPU cluster. Important characteristics for ML clusters are tested using an event-based network simulator called OMNeT++ [8]. Scalability and system performance have been investigated with 256 and 1024 GPUs. Results show that the proposed architecture achieves a near ideal throughput until a load of 0.6, with zero packet loss and sub-microsecond latency.

2. Network architectures and simulation setup

The OPSquare network from [7] is shown in Fig. 1a. The network contains N pods, and each pod has M nodes, grouped in P groups. Each node is directly connected to the optical network and communicates to one $N \times N$ inter-pod switch and P $M \times M$ intra-pod switches. The architecture can connect $M \cdot N$ nodes. Thus, with $M = N = 16$ up to 256 nodes can be connected and with $M = N = 32$ up to 1024 nodes. This implies that with relatively low (and

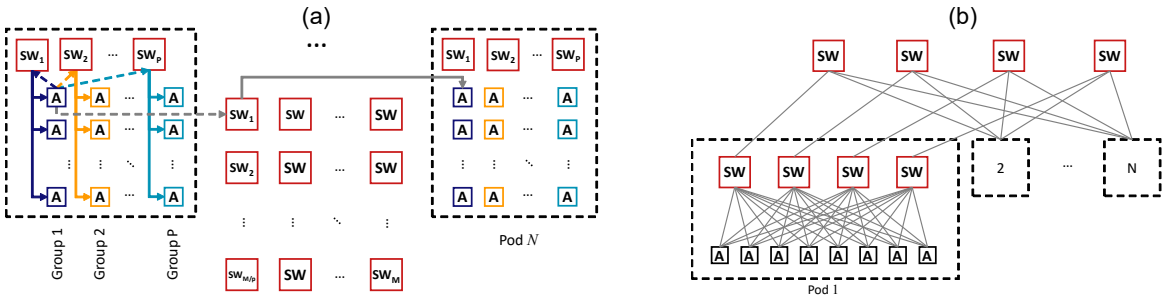

Fig. 1. (a) The OPSquare for ML architecture, with N pods containing M nodes (or accelerators) divided in P groups. Outgoing connections are shown for node 1 only. (b) Leaf-Spine network architecture as presented in [4] with N pods.

feasible) port count of the switch (M or N), large number of nodes can be interconnected. Both switches are based on a broadcast-and-select design: The input is split by a factor F ($F = M/P$), after which the output is passed by one of the F gating semiconductor optical amplifiers (SOAs). Each node is equipped with P optical transceivers, one for each group. The number of transceivers is compensated by a reduction in the switching losses, as the split reduces from M to F. As a result, less cascaded SOA gain is required, reducing possible signal distortion and added noise. Next to the P optical transceivers, each node has an additional transceiver that interfaces with one of the N inter-pod switches such that all m-th accelerators are connected. In this way, at most two hops are required for the entire network. The transceiver forward latency is set to 80 ns, the switch's label processing delay to 10 ns and the optical preamble to 30 ns. Each transmitter has an input buffer of 1 MB. More implementation details on the switch and measured parameters can be found in the original paper [7]. To test the scalability, a larger 1024 node version is created with a pod size M of 32 nodes, a group size P of 8, and 32 pods (N).

The performance of OPSquare is compared to a state-of-the-art GPU cluster by NVIDIA [4]. The reference network is based on a leaf-spine architecture shown in Fig. 1a, where 8 GPUs are electrically interconnected to four parallel non-blocking NVLink switches using four interconnects of 900 Gb/s per link (3.6 Tb/s aggregated traffic). Each pod is interconnected using four 800G transceivers to four 32-port spine-switches. The switch latency is set to 1 µs, estimated by Tomahawk 4 switch ASICs (450 ns [9]) and characterization of older NVLink devices ($\approx$ 1.5 µs [10]). The actual buffer size of each switch is unknown, but for this experiment the leaf-switches have a buffer size of 16 MB and the spine-switches 64 MB. Similarly, a large scale network is tested with 1024 nodes, using 128 pods.

The traffic in the network is generalized according to common patterns found in ML distributed networks. The patterns for data parallelism (to distribute gradients) are mostly based around an all-reduce operation, but can be structured using a central parameter server, ring topology or a tree topology. Model parallelism and pipelining are fixed point-to-point streams [11]. The various software and architectural implementations make it hard to properly generalize the traffic, hence, we choose to use the worst-case traffic. In both networks, the nodes will generate uniform random traffic with fixed length packets of 1500 Bytes. The destination is uniform random, but a locality factor shapes the traffic to be intra-pod for a certain percentage to model communication to neighboring nodes.

All simulations are warmed up for 20 µs prior to sampling the notable events for an additional 20 µs to fill the buffers and get a converged steady-state measurement. The number of sent and received packets is recorded per node, together with the end-to-end latency of every received packet. Discarded packets are labeled as lost and recorded at the destination nodes. Data from all nodes are summed, averaged or normalized whenever applicable.

3. Results

The OPSquare network shows almost ideal performance when the majority of the traffic is local, as shown in Fig. 2. The 3.6 Tb/s links to other pods and nodes keep the throughput high and latency low. With strong outgoing traffic, the throughput is limited by the inter-pod capacity of a single 3.6 Tb/s channel. The packet loss in Fig. 2 shows that 1 MB of input buffers is enough to reliably deliver packets at low traffic load. Increasing the buffer size might decrease the packet loss, but at a cost of increased latency. Since retransmission is included, as it is necessary to accommodate rejected packets by the optical switch, the buffer size is the main reason for packet loss. The packet latency is dictated mainly by the buffer queue for outgoing traffic. Intra-pod packets are delivered within 0.5 µs, inter-pod packets are delivered within 6 µs. Scaling the network worsens the performance by less than 1 µs, as the splitting ratio of the optical switches and hence the wait time for transmission is increased.

Fig. 3 demonstrates the performance of the reference network for both 256 and 1024 GPUs. The throughput in Fig. 3a shows that the network is designed with the assumption that local traffic dominates. Only when the traffic is for two-thirds local, the total throughput of each node exceeds half the capacity. From the packet loss in Fig. 3b it can be seen that the spine-switch buffer size is causing a step in the number of lost packets from a load of 0.8 for 33% local traffic. For a large network, even local traffic faces packet loss. The packet latency in Fig. 3c shows

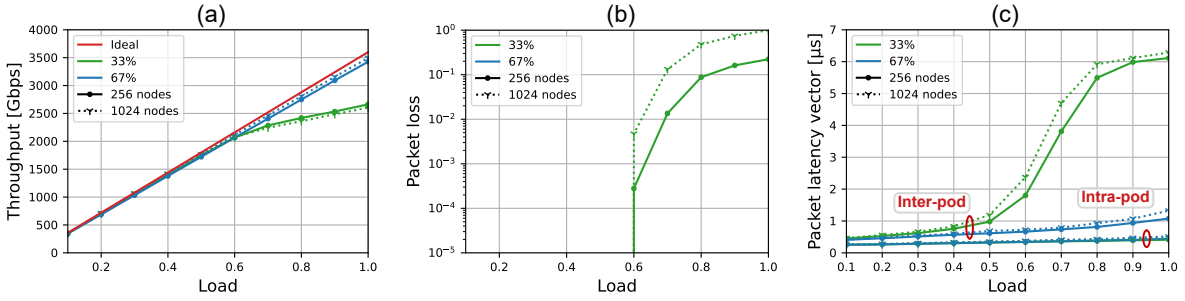

Fig. 2. OPSquare for ML: (a) Throughput for various traffic locality, load intensities and network sizes. (b) Packet loss. (c) Packet latency.

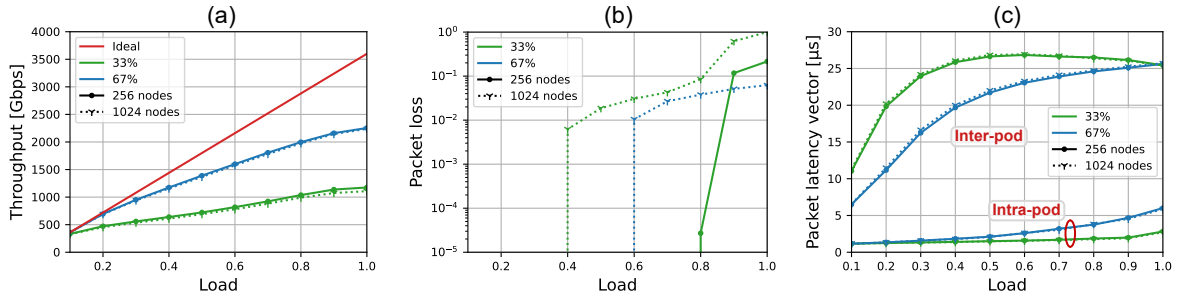

Fig. 3. Leaf-spine network: (a) Throughput for various traffic locality, load intensities and network sizes. (b) Packet loss. (c) Packet latency.

that local traffic is delivered within 6 µs, but at low traffic loads around 1 µs, being equal to the switch latency. Inter-pod traffic saturates to 25 µs. The scalability has no big impact on the performance, as the network is limited already with 256 nodes. With four, high port-count switches, the total throughput of the switches must go up to 102.4 Tbps and 128 ports with 1024 nodes, which is double of the current state-of-the-art with the Tomahawk 5. A second spine-layer could reduce the extreme switch throughput at the cost of increased latency.

High-capacity, node-based networks are a good fit for optical interconnects due to its low latency and high throughput. The transparency of the switches allows for a ten times lower latency for local packets, as information does not have to be buffered in the switch, no O/E/O conversion takes places and no additional protocol related overhead occurs. In terms of scalability, using a two-layer leaf-spine solution does require a doubling in possible throughput for the switches with 128 ports. Whereas the OPSquare network require a feasible 32×32 switch. The node grouping reduces switch complexity, but also distributes the traffic. Nevertheless, to provide 3.6 Tb/s optically using a single fiber, 18 WDM channels running at 100 GBaud PAM-4 are necessary. Optical switches will have to be designed to realize such interconnect using broadband, polarization insensitive, low-loss and high port count optical components.

4. Conclusion

By modeling an optical network and comparing it to the state-of-the-art accelerator cluster of NVIDIA, we have demonstrated that optical networks are a good fit for high-capacity devices such as GPUs. With latencies in the order of single microseconds, and almost ideal throughput for local traffic, OPSquare has the potential to outperform today's standards. State-of-the-art electrically switched networks are at the boundary of their performance, requiring high port count electrical switches with high throughput. OPSquare will require some investigation regarding the high capacity interconnects, and its compatibility with readily demonstrated switches. The predictable traffic patterns make control of the optical switches easier and their transparency allows for future scaling of capacity.

References

1. J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, and P. Villalobos, "Compute Trends Across Three Eras of Machine Learning," (2022). ArXiv:2202.05924 [cs].
2. A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts *et al.*, "PaLM: Scaling Language Modeling with Pathways," (2022). ArXiv:2204.02311 [cs].
3. Y. Ren, S. Yoo, and A. Hoisie, "Performance Analysis of Deep Learning Workloads on Leading-edge Systems," in *2019 IEEE/ACM Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS),* (IEEE, Denver, CO, USA, 2019), pp. 103–113.
4. "NVIDIA H100 Tensor Core GPU Architecture Overview," https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper (2022).
5. J. Shalf, "HPC Interconnects at the End of Moore's Law," in *Optical Fiber Communication Conference (OFC) 2019,* (OSA, San Diego, California, 2019), p. Th3A.1.
6. K. Hosseini, E. Kok, S. Y. Shumarayev, C.-P. Chiu, A. Sarkar, A. Toda *et al.*, "8 Tbps Co-Packaged FPGA and Silicon Photonics Optical IO," p. 3 (2021).
7. F. Yan, W. Miao, O. Raz, and N. Calabretta, "Opsquare: A flat DCN architecture based on flow-controlled optical packet switches," J. Opt. Commun. Netw. **9**, 291–303 (2017). Conference Name: Journal of Optical Communications and Networking.
8. "OMNeT++ Discrete Event Simulator," https://omnetpp.org/.
9. B. Wheeler, "TOMAHAWK 4 SWITCH FIRST TO 25.6TBPS," p. 3 (2019).
10. A. Li, S. L. Song, J. Chen, J. Li, X. Liu, N. Tallent, and K. Barker, "Evaluating Modern GPU Interconnect: PCIe, NVLink, NV-SLI, NVSwitch and GPUDirect," IEEE Trans. on Parallel Distributed Syst. **31**, 94–110 (2020). ArXiv: 1903.04611.
11. W. Fedus, B. Zoph, and N. Shazeer, "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity," p. 39 (2022).