Integrating Nanosecond Optical Switching in Deep Distributed Learning System

Cen Wang,[1,*] Noboru Yoshikane,[1] and Takehiro Tsuritani[1]

[1] *Photonic Transport Network Laboratory, Photonics Division, Advanced Technology Laboratories, KDDI Research, Inc., Saitama, Japan, 356-8502*

[*] *xce-wang@kddi.com*

Abstract: We propose $O(10ns)$ optical switching structure with wavelength-elastic and topology-flexibility in support of distributed deep learning platform. Our experimental demonstration shows accelerations for three use cases.

1. Introduction

To accelerate distributed deep learning (DDL) and similar computing systems which have massive intermediate result interactions (i.e., from *MB* to *GB*) and complex computing modes, it is the mainstream optimization trend to deeply integrate the current key bottleneck, i.e., the networking. In DDL jobs, computing and communication have deterministic time granularity (i.e., typically $O(100ms)$), and the communication patterns between nodes have special topology or present periodic dynamics. Related works have paid attention to using optical switching thick pipes and topology construction capabilities to provision independent slices for DDL jobs, thereby avoiding slowdown or tailing caused by resource competition. Our previous work [1] provided ring-shape topology for *Ring AllReduce* patterns, and multi-job computing/communication phases are interleaved in the slices. The topology is only adjusted when an old jobs completed and a new job arrives, and this slow adjustment method is a compromise to the slow speed ($O(1ms)$) of analog micro-electro-mechanical system (MEMS)-based optical switch. Khani et al. [2] used the micro-ring resonator (MRR) to slicing a wavelength division topology for a DDL job, and the wavelength switching speed is $O(1\mu s)$.

However, the aforementioned approaches cannot fully meet the emerging demands of DDL. First, the communication patterns are diverse. Although *Ring AllReduce* is the most bandwidth-efficient, it has many communication cycles and poor scalability, which means high probability of congestion and tailing. In contrast, *Rabenseifner's AllReduce* and *Hierarchical Ring AllReduce* promise to be the alternatives as DDL platform scaling. Both patterns are with periodic topology changes. Second, although data parallelism (DP) is widely deployed, as the number of model parameters surges, it puts a huge pressure on the memory of a single GPU. The model parallelism (MP) of splitting a model into different GPU nodes and DP + MP solutions are worth being considered. Introducing MP into DP increases the frequency of intermediate results interaction. Final, the calculation and communication of a DDL job do not overlap (as shown in Fig. 1 (a)), thus, providing a whole wavelength slice for a single job is not resource-efficient. Precise two-job interleaving is not practical, since the computing time of one job is rarely equal to the communication time of another job. Therefore, a granularity-adaptive interleaving is more efficient, i.e., when a large job is computing, the network is scheduled to provide topologies for other concurrent small jobs. As a result, the frequency of (wavelength) topology adjustments may increase accordingly.

All of above limitations are pointing to a technology as a solution, that is topology-flexible fast optical switching. Previous PULSE [3] and Sirius [4] take semiconductor optical amplifier (SOA)-based fast tunable transmitter ($O(1ns)$) plus arrayed waveguide grating router (AWGR) as their architecture solution, however, the fixed connections in the core layer restrict the topology flexibility and wavelength elasticity. In this work, we propose a switching structure composed of high-speed wavelength switching ($O(10ns)$) and high-speed space switching ($O(10ns)$). Our approach has structural inheritance compared with the MEMS-based optical switching network solutions [5–7] in the past 10 years. Fast wavelength selective switch (FWSS) enables wavelength elasticity in lightpath, and fast optical switch array (FOSA) enables fine-grained topology reconfiguration. With these advantages, we further discuss how the proposed structure can support deep learning platforms with (1) diverse and complex communication patterns; (2) DP-MP hybrid parallelism; (3) concurrent jobs interleaving requests. We experimentally demonstrate the proposed FOSA architecture and the former three use cases running over it. The results show stable/fast switching speed, stable/high throughput, and effective accelerations for DDL jobs.

2. Architecture and Scheduling

As shown in Fig. 1 (b), currently we still require PCIe to interconnect a GPU and an FPGA-based network interface card (NIC). To improve the throughput, the TCP/IP protocol stack is bypassed [8] and the intermediate results of a DDL job are directly written onto the NIC. Fixed TX/tunable RX (FTTR) or tunable TX/tunable RX (TTTR) optical modules are inserted into the NICs. The optical modules can be connected to the FWSS first or directly

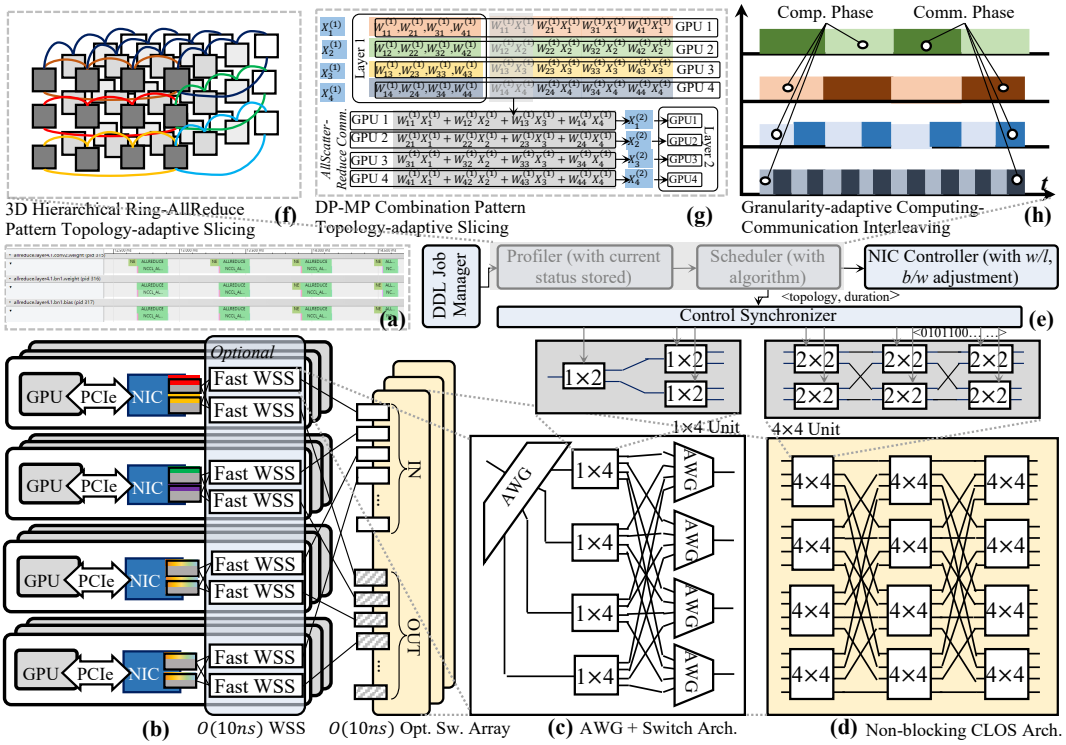

Fig. 1. Architecture and scheduling: (a) timelines in current commercial DDL platform (Horovod); (b) the proposed switching architecture; (c) FWSS; (d) FOSA; (e) control plane; (f) dynamic patterns adapting; (g) DP-MP combination pattern slicing; (h) granularity-adaptive multi-job interleaving.

connected to the FOSA. The FWSS adopts a combined structure of AWG and $1 \times N$ optical switches (constructed by multiple 1×2 optical switch units (OSUs)), as shown in Fig. 1 (c), and the FOSA adopts a clos non-blocking structure (constructed by multiple 2×2 OSUs), as shown in Fig. 1 (d). These OSUs are digital MEMS-based, thus to have the same profile, which can realize precisely synchronous control through the entire network. In addition, the structure is modularized, where the FWSS and the FOSA can be independently deployed according to the scale, and the type of transmitters is not specific.

As shown in Fig. 1 (e), using the proposed switching structure to adapt DDL jobs primarily requires profiling of the DDL jobs (i.e., to obtain the communication patterns and computing and communication duration of each batch). The profiling method can be testing a batch. Note that the communication time depends on the allocated bandwidth (data rate), however, it is hard to exactly know the bandwidth demand as prior knowledge. We measure the amount of generated intermediate results in a job over a period to estimate the maximum bandwidth demand. The scheduler generates wavelength (bandwidth) allocation and OSUs configuration (formed topology and topology duration) according to the input DDL job profiles, and sends them to the NIC controller and the OSU control (signal) synchronizer, respectively. We discuss the following three adaptive use cases.

Dynamic Patterns Adapting. As shown in Fig. 1 (f), the switching structure needs to periodically establish and dismantle the ring topology in order in the three directions of the GPU cube. **DP-MP Slicing.** As shown in Fig. 1 (g), a GPU only processes a piece of input data and a slice of a weight tensor, after the multiplication operation, *AllScatter-Reduce* is conducted among the GPUs to exchange intermediate results, and the topology duration is $O(10\ 100\mu s)$, much less than that in pure DL. **Multi-job Interleaving.** As shown in Fig. 1 (h), granularity-adaptive interleaving alleviates the constraints on the homogeneity of concurrent jobs in computing time and communication time. This enables the switching structure to support several smaller jobs while one large job is in its computing phase. To avoid the communication conflict of multiple jobs, the scheduler needs to properly control the start of computing phases and communication phases of a job (i.e., delaying). The $O(1ms)$ delaying can be implemented in the software layer by the job manager, whereas $O(1\mu s)$ delaying of a DP-MP job is better to be implemented by controlling the NIC.

3. Demonstration

We leave the demonstration with FWSS + FOSA in the near future. The experiment setup is shown in Fig. 2 (a), 4 GPU nodes are connected to 4×4 (6 OSUs) through NICs (with 10Gbps FTTRs). The NICs push the raw data onto optical resources bypassing the TCP/IP protocol stack. The jobs to be injected are running PCam dataset [9]

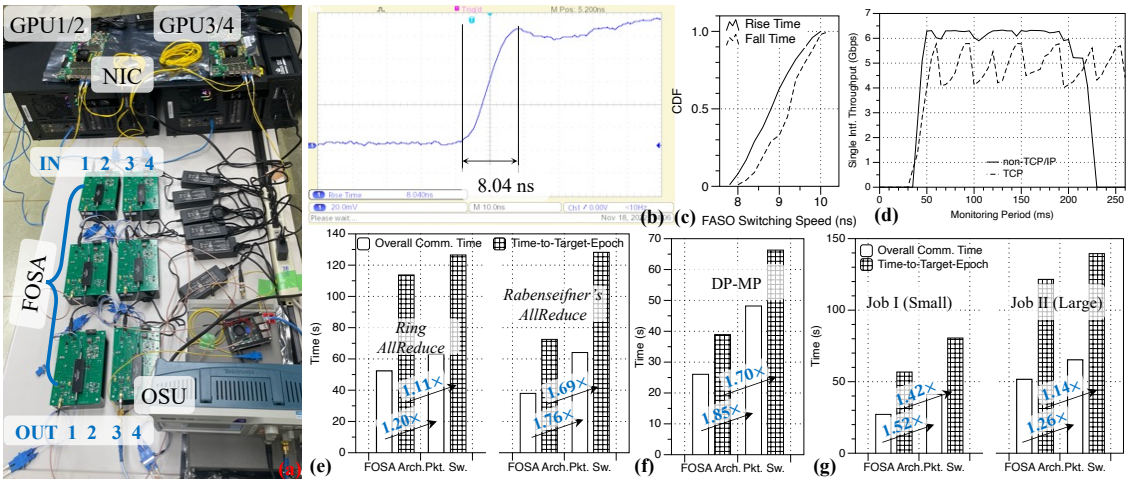

Fig. 2. Architecture and scheduling: (a) the experiment setup; (b) single rise capturing; (c) cumulative distribution function (CDF) of 100 rises and falls; (d) througput (FOSA Arch. with non-TCP/IP vs. Eth. Pkt. Sw. with TCP/IP); (e) pattern adapting accelerations; (f) DP-MP slicing accelerations; (g) granularity-adaptive multi-job interleaving accelerations.

using convolutional neural networks (CNNs). To shield the complex and unnecessary functions of the commercial platforms, we implemented the communication process of intermediate result interaction according to the principle (i.e., *Rabenseifner's AllReduce*, *Ring AllReduce*, and DP-MP).

We first measure the speed of rise and fall, as shown in Fig. 2 (b) and (c). It shows that the switching speed of FOSA is stable at $O(10ns)$. We also monitor the throughput of non-TCP/IP (over FOSA architecture) and TCP/IP (over packet switch) in a batch, as shown in 2 (d), and the throughput in our system is higher and stable. This suggests that optically interconnected DDL systems can facilitate protocol overhead reduction, or simplify protocol functions.

We compare the overall communication time and time-to-target-epoch of jobs on the packet switch (Dell S4128F-ON) and our system, respectively. We do not take time-to-accuracy [2] as our metric because that the time to the same accuracy is uncertain and affected by model initialization, input data size, and other hyper parameters, etc. We demonstrate three use cases. The acceleration is calculated by using the performance over packet switched network to devide performance over FOSA architecture. As for dynamic topology adapting, as shown in 2 (e), the acceleration of *Rabenseifner's AllReduce* is higher than that gained by *Ring AllReduce*. This shows that topology adaptation is necessary for a job with a single large communication similar to *Rabenseifner's AllReduce*. The intermediate results volume of DP-MP (only forward propagation) is larger than that of pure DP. From the obtained acceleration, it shows that our architecture can well support relatively fine-grained DP-MP. Granularity-adaptive interleaving accelerates Job I with a small amount of data better than large Job II. This is because two jobs collide without interleaving, and the communication time tailing of the small one is more serious.

Conclusion

We propose the FWSS + FOSA switching architecture composed by digital MEMS-based OSU, in support of DDL platforms. This architecture inherits the features of the optical circuit switched data center for last 10 years, in contrast, the speed reaches $O(10ns)$. The experimental demonstration with FOSA shows stable rise and fall and throughput. Additionally, the FOSA architecture effectively accelerates the DDL jobs under three use cases: dynamic pattern adapting, DP-MP hybrid slicing and multi-job interleaving.

References

1. C. Wang, N. Yoshikane, F. Balasis, and T. Tsuritani, *Computer Networks*, vol. 214, p. 109191, 2022.
2. M. Khani, M. Ghobadi, Z. Alizadeh, Mohammad Zhu, M. Glick *et al.*, in *ACM SIGCOMM 2021*.
3. J. L. Benjamin, T. Gerard, D. Lavery, P. Bayvel, and G. Zervas, *JLT*, vol. 38, no. 18, pp. 4906–4921, Sep 2020.
4. H. Ballani, P. Costa, R. Behrendt, *et al.*, in *ACM SIGCOMM 2020*.
5. N. Farrington, G. Porter *et al.*, *SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 4, p. 339–350, 2010.
6. A. Singla, A. Singh, and Y. Chen, in *NSDI 12*, 2012, pp. 239–252.
7. L. Poutievski, O. Mashayekhi, J. Ong *et al.*, in *ACM SIGCOMM 2022*.
8. C. Wang, X. Xue, N. Yoshikane, F. Balasis, T. Tsuritani, and H. Guo, in *OFC 2022*, pp. 1–3.
9. B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, *arXiv cs.CV 1806.03962*, 2018.