Accelerating DML Training in OCS-based DCNs

(Invited Paper)

Xuexia Xie, Binjun Tang, Xiaoliang Chen and Zuqing Zhu†
School of Information Science and Technology, University of Science and Technology of China, Hefei, China
†Email: {zqzhu}@ieee.org

***Abstract*—This paper explores the benefits of in-network computing (INC) empowered all-optical interconnects (AOI) in accelerating distributed machine learning (DML) jobs. We describe the network architecture and service model, and present a large-job-first and a grouping-based interleaved scheduling policy for minimizing job completion time. The results verify the superiority of our scheduling policies and show the effectiveness of INC and AOI in mitigating bandwidth bottlenecking during DML training.**

***Index Terms*—Distributed machine learning, all-optical interconnect, topology engineering, in-network computing.**

I. Introduction

Nowadays, networking technologies have been developing very with numerous innovations [1–9]. Meanwhile, with the rapid growth of artificial intelligence, especially the rapid popularization of large language models (LLMs), the past years has witnessed exponentially expanding model and training data sizes. For instance, Meta's Llama 3 model [10] comprises up to 70 billion parameters and was pretrained on over 15T tokens. Such vast model size and data volume have rendered training on single computing units (*e.g.*, graphics processing unit (GPU)) impractical, and thereby, call for distributed machine learning (DML) training paradigms [11].

In traditional data center networks (DCNs), GPU servers communicate via electrical packet switches (EPS) that interconnect to form multi-tier fabrics (*e.g.*, fat trees) [12–15]. However, unlike legacy DCN workloads, DML applications involve regular collective communications, which significantly transform traffic patterns by producing skewed, periodic, and megaflows. This shift poses great challenges for EPS-based DCNs that rely solely on traffic engineering (TE) to accommodate bandwidth-intensive DML applications, let alone their high energy consumption and end-to-end latency.

To address the aforementioned challenges, DCN operators have begun integrating optical circuit switching (OCS) technologies [16–28] into their infrastructures and building optical DCNs (ODCNs) [29]. Particularly, OCS allows for building low-diameter ODCNs by provisioning large-capacity optical links directly to inter-unit communications and dynamically reconfiguring the all-optical interconnect (AOI) to fit skewed traffic distributions [30–36]. However, AOI reconfiguration often results in service interruptions, considering the fiber-level switching granularity and millisecond-scale reconfiguration latency of current commercial OCS switches. These limitations, if not properly addressed, will diminish the benefits of ODCNs and impede their ability to support the rapid expansion of large-scale DML training, for instance, LLMs that employ hybrid parallelism and may use tens of thousands of GPUs.

On the other hand, recent studies have revealed that in-network computing (INC) on programmable data plane (PDP) can reshape DML traffic patterns and mitigate the bottleneck due to incast communication in parameter aggregation phases [37–40], and hereby, potentially reduce AOI reconfiguration frequency. It is worth exploiting the mutual benefits of INC and AOI to further accelerate DML training. Nevertheless, how to orchestrate the multi-dimensional resources (*i.e.*, GPU, INC and bandwidth) in DML job scheduling and AOI configurations remains an underexplored yet challenging area [41, 42].

In this paper, we investigate how to accelerate DML training in INC-empowered ODCNs. We first detail the ODCN architecture and the principle of DML training aided by INC. Then, we discuss two DML job scheduling policies tailored for INC-empowered ODCNs, namely, large-job-first and grouping-based interleaved scheduling. The former prioritizes jobs with larger bandwidth demand, while the latter groups jobs and interleaves the communication phases of jobs from different groups to ease bandwidth contention. Performance evaluations verify the superiority of our algorithm over benchmarks in terms of job completion time and reconfiguration frequency.

II. Problem Description

A. ODCN Architecture

Fig. 1(a) shows the architecture of an ODCN of racks dedicated for DML training. Each rack houses a few tens of GPU servers wired to a top-of-rack (ToR) switch. The GPUs within each server are directly linked in full mesh via high-speed links (*e.g.*, NVLink) and managed by a job scheduler. Consequently, inter-rack communications present the primary bottlenecks in a DML cluster. In the ODCN, ToRs connect to several OCS switches, which can be dynamically reconfigured to provide diverse connectivity (all-to-all, Torus, *etc.*) among ToRs for serving the time-varying and skewed traffic in DML training. Note that, while such flexibility, also known as topology engineering (TPE) [32], helps ease inter-rack bandwidth bottlenecking, the nonnegligible reconfiguration time of commercial OCS switches (*e.g.*, 50 ms) makes the TPE optimization for DML training a nontrivial task. To this end, we further bring in INC and replace legacy ToRs with the PDP switches that can execute certain parameter aggregation operations of DML training in line rate without adding extra costs or power consumption [43], while effectively reshaping inter-rack traffic to relieve the stress in the OCS plane.

978-3-903176-67-6

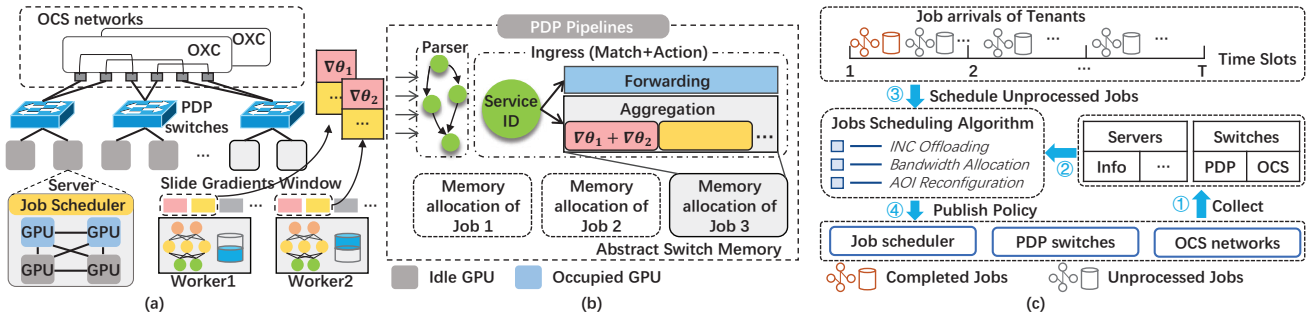


Fig. 1. System overview: (a) ODCN architecture, (b) INC-empowered data-plane, and (c) control plane workflow.

Fig. 1(b) illustrates the aggregation pipeline of an INC-empowered PDP switch for jobs using PS-based data parallelism. Specifically, the switch memory resources are partitioned to serve as aggregators for gradient aggregation. Each worker i of a job first computes its local gradients (denoted by $\nabla\theta_i$), which are then segmented into packets and transmitted to the PDP switch. The switch uses a *service ID* to identify the appropriate aggregator memory block for the job, performs aggregation, and returns the aggregated result to workers for gradient updates. As such, INC accelerates DML training by eliminating the incast communication bottleneck inherent in PS framework and enabling line-rate aggregation operations.

B. DML Service Model

Based on aforementioned ODCN architecture, DML tasks can be serviced according to the control plane workflow depicted by Fig. 1(c). We adopt a time-slotted scheduling mechanism, where both DML job scheduling and all-optical interconnect (AOI) reconfiguration are executed at the granularity of time slots. We also assume that job placement is determined by the service layer in prior and is taken as an input to our scheduling problem. At the beginning of each time slot, the network monitoring module first collects information such as switch status and link bandwidth and forwards it to the job scheduling algorithm module. Then, given a set of DML jobs along with their GPU placements, the scheduling module determines which jobs should leverage INC, when each job's communication should commence, how to allocate bandwidth among these jobs, whether and how to reconfigure the AOI, with the goal of minimizing the overall completion time for the batch of jobs. Finally, the output of the scheduling module is transmitted to the data-plane job scheduler, PDP switches, and OCS for execution and configuration.

III. Algorithm Design

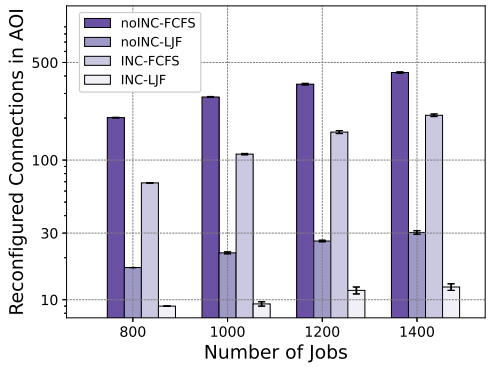
The optimal job scheduling and TPE solution can be derived by solving an integer linear programming (ILP) model. However, ILP is computationally prohibitive in large-scale setting, hindering its applicability to online operations. Alternatively, we resort to time-efficient heuristic algorithms. In particular, we target PS-based data parallelism in this work, which comprises four phases: local computation, push, aggregation, and pull. Local computation is performed by GPUs and persists deterministic durations contingent upon DML configurations (*e.g.*, model size, batch size and optimizer) and GPU computing power. Gradient aggregation, when executed in PDP switches, incurs negligible temporal overhead, as it is done concurrently with push/pull communications at line rate. The durations of push/pull phases depend on the amounts of bandwidth allocated. Next, we will outline two optimization strategies for DML job scheduling.

a) Large-job-first scheduling: Allocating INC resources to jobs with more inter-worker demand cuts down inter-worker traffic, and thereby, reduces AOI reconfigurations. The Large-job-first scheduling strategy operates as follows: 1) prioritize INC resource allocation for jobs with larger worker counts and data transmission volumes; 2) allocate CPU resources as PS aggregators for the remaining jobs when INC resources exhaust; 3) optimize the bandwidth distribution to minimize the overall job completion time; 4) reoptimize the AOI reconfiguration dynamically to mitigate bandwidth bottlenecking.

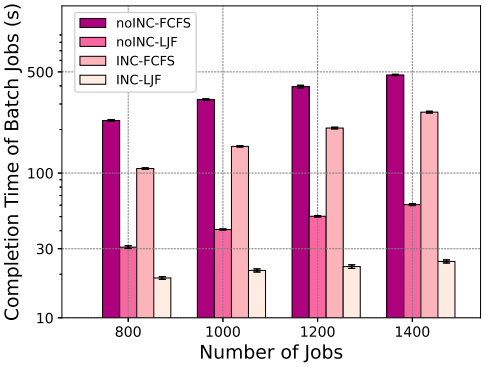
b) Grouping-based interleaved scheduling: Interleaving the communication phases of overlapping jobs can effectively resolve bandwidth contention. Therefore, we devise a grouping-based optimization strategy: 1) group jobs into two cohorts with comparable computation-communication ratios to avoid latency tail effects; 2) interleave the communication phases of the two cohorts by delaying the starting time of one cohort; 3) perform OCS reconfiguration upon the rotation of the communication phases of the two cohorts.

IV. Performance Evaluation

We evaluated the performance of our algorithm through simulations over a 64-rack ODCN setup. Each rack contains 64 servers and one PDP switch with INC capacity that supports up to 4 jobs concurrently. The PDP switches connect to an optical cross-connect (OXC) via 24 40-Gbps optical ports. We assumed DML jobs arrive according to a Poisson process with an average arrival rate of 20 jobs per time slot, each lasts for 1 second. For each job, the number of workers was randomly selected from $\{2, 4, 8, 16\}$ according to the Philly traces [44]. We made the push data size for each worker uniformly distributed within $\{0.552, 3.096, 4.0\}$ GB, to simulate the VGG16, GPT2-large, and Meta Llama3 models, respectively. The workers were randomly placed in each simulation, but we attempted to co-locate the workers of each job in the same rack to



(a) AOI reconfigurations



(b) Completion time of batch jobs

Fig. 2. Results of different numbers of DML jobs.

maximize traffic locality. We compared our algorithm, namely, INC-empowered ODCN with large-job-first (LJF) scheduling (**INC-LJF**), with three benchmark designs:

- **INC-FCFS**: INC-empowered ODCN with first-come-first-served (FCFS) scheduling, which greedily allocates multi-dimensional resources to sequentially arriving jobs to minimize each job's completion time.
- **noINC-LJF**: ODCN without INC, LJF scheduling.
- **noINC-FCFS**: ODCN without INC, FCFS scheduling.

Fig. 2 shows the results of AOI reconfiguration frequency and overall job completion time when the number of DML jobs increases from 800 to $1,400$. INC-LJF consistently delivers the shortest completion time with the fewest number of AOI reconfigurations. Remarkably, INC-LJF respectively reduces the number of AOI reconfigurations and completion time by 96.4% and 93.2% over noINC-FCFS. This strikingly confirms the effectiveness of the symbiosis between INC and our algorithm. The introducing of INC helps reduce the overall job completion time by up to 49.4% for INC-FCFS over noINC-FCFS and 46.8% for INC-LJF over noINC-LJF, demonstrating INC's efficacy in reshaping cross-rack traffic and expediting job training. Regarding the scheduling algorithm, our solution reduces the completion time by an average of 87.3% and 86.3% in the noINC and INC scenarios, respectively, which indicates that our LJF algorithm substantially decreases job completion time by improving overall resource utilization.

V. Summary

In this paper, we presented an INC-empowered AOI architecture and two optimization strategies for accelerating DML jobs. Performance evaluations proved the mutual benefits of INC and AOI in mitigating communication bottlenecking while showing superior performance of our proposal.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2023YFB2903903.

References

[1] "Cisco Annual Internet Report 2024." [Online]. Available: https://www.cisco.com/c/en/us/solutions/executive-perspectives/annual-internet-report/index.html# executive-summaries.
[2] Z. Zhu, W. Lu, L. Zhang, and N. Ansari, "Dynamic service provisioning in elastic optical networks with hybrid single-/multi-path routing," *J. Lightw. Technol.*, vol. 31, pp. 15–22, Jan. 2013.
[3] L. Gong *et al.*, "Efficient resource allocation for all-optical multicasting over spectrum-sliced elastic optical networks," *J. Opt. Commun. Netw.*, vol. 5, pp. 836–847, Aug. 2013.
[4] Y. Yin *et al.*, "Spectral and spatial 2D fragmentation-aware routing and spectrum assignment algorithms in elastic optical networks," *J. Opt. Commun. Netw.*, vol. 5, pp. A100–A106, Oct. 2013.
[5] L. Gong and Z. Zhu, "Virtual optical network embedding (VONE) over elastic optical networks," *J. Lightw. Technol.*, vol. 32, pp. 450–460, Feb. 2014.
[6] W. Lu, Z. Zhu, and B. Mukherjee, "On hybrid IR and AR service provisioning in elastic optical networks," *J. Lightw. Technol.*, vol. 33, pp. 4659–4669, Nov. 2015.
[7] S. Li *et al.*, "Protocol oblivious forwarding (POF): Software-defined networking with enhanced programmability," *IEEE Netw.*, vol. 31, pp. 58–66, Mar./Apr. 2017.
[8] Y. Hu *et al.*, "Polymorphic smart network: An open, flexible and universal architecture for future heterogeneous networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, pp. 2515–2525, Oct.-Dec. 2020.
[9] J. Wu *et al.*, "Theoretical framework for a polymorphic network environment," *Engineering*, vol. 39, pp. 222–234, 2024.
[10] Meta-Llama-3. Accessed: Feb. 6, 2025. [Online]. Available: https://ai.meta.com/blog/meta-llama-3/
[11] J. Duan *et al.*, "Efficient training of large language models on distributed infrastructures: A survey," 2024. [Online]. Available: https://arxiv.org/abs/2407.20018
[12] L. Zhang and Z. Zhu, "Spectrum-efficient anycast in elastic optical inter-datacenter networks," *Opt. Switch. Netw.*, vol. 14, pp. 250–259, Aug. 2014.
[13] P. Lu *et al.*, "Highly-efficient data migration and backup for Big Data applications in elastic optical inter-datacenter networks," *IEEE Netw.*, vol. 29, pp. 36–42, Sept./Oct. 2015.
[14] J. Liu *et al.*, "On dynamic service function chain deployment and readjustment," *IEEE Trans. Netw. Serv. Manag.*, vol. 14, pp. 543–553, Sept. 2017.
[15] P. Lu and Z. Zhu, "Data-oriented task scheduling in fixed- and flexible-grid multilayer inter-DC optical networks: A comparison study," *J. Lightw. Technol.*, vol. 35, pp. 5335–5346, Dec. 2017.
[16] H. Fang *et al.*, "Predictive analytics based knowledge-defined orchestration in a hybrid optical/electrical datacenter network testbed," *J. Lightw. Technol.*, vol. 37, pp. 4921–4934, Oct. 2019.
[17] Z. Zhu, W. Lu, L. Liang, and B. Kong, "Predictive analytics in hybrid optical/electrical DC networks," in *Proc. of OFC 2019*, pp. 1–3, Mar. 2019.
[18] W. Lu *et al.*, "AI-assisted knowledge-defined network orchestration for energy-efficient data center networks," *IEEE Commun. Mag.*, vol. 58, pp. 86–92, Jan. 2020.
[19] S. Zhao and Z. Zhu, "Network service reconfiguration in hybrid optical/electrical datacenter networks," in *Proc. of ONDM 2020*, pp. 1–6, May 2020.
[20] Q. Li *et al.*, "Scalable knowledge-defined orchestration for hybrid optical/electrical datacenter networks," *J. Opt. Commun. Netw.*, vol. 12, pp. A113–A122, Feb. 2020.

[21] S. Zhao and Z. Zhu, "On virtual network reconfiguration in hybrid optical/electrical datacenter networks," *J. Lightw. Technol.*, vol. 38, pp. 6424–6436, Dec. 2020.

[22] X. Pan *et al.*, "Scheduling virtual network reconfigurations in parallel in hybrid optical/electrical datacenter networks," *J. Lightw. Technol.*, vol. 39, pp. 5371–5382, Sept. 2021.

[23] S. Zhao, X. Pan, and Z. Zhu, "On the parallel reconfiguration of virtual networks in hybrid optical/electrical datacenter networks," in *Proc. of GLOBECOM 2020*, pp. 1–6, Dec. 2020.

[24] H. Yang and Z. Zhu, "Application-aware configuration of all-optical interconnects in Hyper-FleX-LION," in *Proc. of OGC 2022*, pp. 1–4, Sept. 2022.

[25] J. Peng *et al.*, "Offloading NFV orchestration to ToR switches: How to leverage PDP to realize agile service function chaining in HOE-DCNs," in *Proc. of ICC 2021*, pp. 1–6, Jun. 2021.

[26] H. Yang *et al.*, "On the bilevel optimization for remapping virtual networks in an HOE-DCN," *IEEE Trans. Netw. Serv. Manag.*, vol. 19, pp. 1274–1286, Jun. 2022.

[27] H. Yang, Z. Zhu, R. Proietti, and B. Yoo, "Which can accelerate distributed machine learning faster: Hybrid optical/electrical or optical reconfigurable DCN?" in *Proc. of OFC 2022*, pp. 1–3, Mar. 2022.

[28] Q. Lv, Z. Ma, and Z. Zhu, "Experimental demonstration of hitless OCS-based DCN reconfiguration to steer multi-class traffic," in *Proc. of ICOCN 2024*, pp. 1–3, Jul. 2024.

[29] N. Jouppi *et al.*, "TPU v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings," in *Proc. of ISCA 2023*, pp. 1–14, Jun. 2023.

[30] H. Yang and Z. Zhu, "Acceleration of coflows with all-optical interconnects in Hyper-FleX-LION," *J. Opt. Commun. Netw.*, vol. 14, pp. 805–814, Oct. 2022.

[31] X. Xie, H. Yang, and Z. Zhu, "P4INC-AOI: When in-network computing meets all-optical interconnect for adaptive and low-latency optical DCN," in *Proc. of OFC 2023*, pp. 1–3, Mar. 2023.

[32] Q. Lv *et al.*, "On the TPE design to efficiently accelerate hitless reconfiguration of OCS-based DCNs," *IEEE J. Sel. Areas Commun., in Press*, pp. 1–13, 2025.

[33] H. Yang and Z. Zhu, "Traffic-aware configuration of all-optical data center networks based on Hyper-FleX-LION," *IEEE/ACM Trans. Netw.*, vol. 32, pp. 2675–2688, Jun. 2024.

[34] X. Dong *et al.*, "On scheduling DML jobs in all-optical DCNs with in-network computing," in *Proc. of GLOBECOM 2024*, pp. 1–6, Dec. 2024.

[35] X. Chen *et al.*, "DRL-TPE: Learning to optimize TPE of optical interconnects to accelerate hitless reconfiguration of OCS-based DCNs," in *Proc. of OFC 2025*, pp. 1–3, Mar. 2025.

[36] W. Wu *et al.*, "Accelerating collective communications with mutual benefits of optical rackless DC and in-network computing," in *Proc. of OFC 2025*, pp. 1–3, Mar. 2025.

[37] Y. Xue and Z. Zhu, "On the upgrade of service function chains with heterogeneous NFV platforms," *IEEE Trans. Netw. Serv. Manag.*, vol. 18, pp. 4311–4323, Dec. 2021.

[38] A. Sapio *et al.*, "Scaling distributed machine learning with in-network aggregation," in *Proc. of NSDI 2021*, pp. 785–808, Apr. 2021.

[39] T. Li, Z. Ma, and Z. Zhu, "st-SFC: Optimizing dynamic deployment of stateful SFCs on P4-based PDP switches," *IEEE Trans. Netw. Serv. Manag.*, vol. 21, pp. 6658–6669, Dec. 2024.

[40] X. Xie, B. Tang, X. Chen, and Z. Zhu, "P4INC-AOI: All-optical interconnect empowered by in-network computing for DML workloads," *IEEE/ACM Trans. Netw., in Press*, pp. 1–16, 2025.

[41] M. Khani *et al.*, "Sip-ml: high-bandwidth optical network interconnects for machine learning training," in *Proc. of SIGCOMM 2021*, pp. 657–675, Aug. 2021.

[42] W. Wang *et al.*, "TopoOpt: Co-optimizing network topology and parallelization strategy for distributed training jobs," pp. 739–767, Apr. 2023.

[43] Intel Intelligent Fabric Processors. Accessed: Feb. 7, 2025. [Online]. Available: https://www.intel.com/content/www/us/en/products/network-io/programmable-ethernet-switch.html.

[44] M. Jeon *et al.*, "Analysis of large-scale multi-tenant gpu clusters for dnn training workloads," in *Proc. of USENIX ATC 2019*, pp. 947–960, Jun. 2019.