

Towards Universal Performance Modeling for Machine Learning Training on Multi-GPU Platforms

Zhongyi Lin¹, Ning Sun, Pallab Bhattacharya¹, Xizhou Feng, *Member, IEEE*, Louis Feng, and John D. Owens¹, *Fellow, IEEE*

Abstract—Characterizing and predicting the training performance of modern machine learning (ML) workloads on compute systems with compute and communication spread between CPUs, GPUs, and network devices is not only the key to optimization and planning but also a complex goal to achieve. The primary challenges include the complexity of synchronization and load balancing between CPUs and GPUs, the variance in input data distribution, and the use of different communication devices and topologies (e.g., NVLink, PCIe, network cards) that connect multiple compute devices, coupled with the desire for flexible training configurations. Built on top of our prior work for single-GPU platforms, we address these challenges and enable multi-GPU performance modeling¹ by incorporating (1) data-distribution-aware performance models for embedding table lookup, and (2) data movement prediction of communication collectives, into our upgraded performance modeling pipeline equipped with inter-and intra-rank synchronization for ML workloads trained on multi-GPU platforms. Beyond accurately predicting the per-iteration training time of deep learning recommendation models (DLRM) models with random configurations with a geomean error of 5.21% on two multi-GPU platforms, our prediction pipeline generalizes well to other types of ML workloads, such as Transformer-based natural language processing (NLP) models with a geomean error of 3.00%. Moreover, even without actually running ML workloads like DLRMs on the hardware, it is capable of generating insights such as quickly selecting the fastest embedding table sharding configuration (with a success rate of 85%).

Index Terms—DLRM, machine learning, model training, multi-GPU, NLP, performance modeling.

I. INTRODUCTION

MODERN machine learning (ML) workloads tend to grow in size and computation, usually beyond the capability

of one single GPU to host and train. Therefore, platforms with multiple compute devices such as GPUs are critical in undertaking training jobs for these workloads. Modern industrial recommendation models, such as DLRM [1], consist of embedding tables of hundreds of gigabytes and are trained in a distributed way on hierarchical compute systems, e.g., 16 nodes $\times$ 8 GPUs, as an advanced software-hardware co-designed system [2]. For large language models (LLM), the rapidly growing number of model parameters requires an equally rapid increase in the number of compute devices used to train them. Examples include GPT3 [3] (2020, 175B (parameter count, same below), 1024 NVIDIA V100 GPUs (estimated)) Megatron-Turing NLG [4] (2022, 530B, 2240 80-GB NVIDIA A100 GPUs), PaLM [5] (2022, 540B, 6144 Google TPU v4), OPT-175B [6] (2022, 175B, 992 80-GB NVIDIA A100 GPUs), and LLaMA [7] (2023, 65B, 2048 80-GB NVIDIA A100 GPUs). For such daunting multi-GPU training jobs, ML practitioners are interested in not only the training performance (e.g., iteration time, query-per-sec (QPS), FLOPS, etc.) but also how to speed up training and use the hardware most efficiently. Characterizing the performance behavior of such multi-GPU jobs is the key to identifying and optimizing performance bottlenecks, and subsequently helping ML practitioners save development time and budget, improve ML model quality provided to users, and benefit the environment by avoiding excessive emissions of CO₂.

However, achieving the above goals is particularly challenging in the multi-GPU platform space for two primary reasons:

- Communication collectives, like *all-to-all* and *all-reduce* across various network media (e.g., NVLink, PCIe, network cards) and topologies that connect multiple compute devices, are essential operations in multi-GPU training and commonly the performance hotspot. The performance modeling of these operations is missing in previous single-GPU and compute-op-focused work [8], [9], [10], [11], [12], [13], [14].
- More importantly, the synchronization behaviors of multiple GPU streams on the same device or across multiple devices are complicated. Previous works have only inadequate or premature modeling for these behaviors, making them insufficient to account for all the GPU idle time on each rank caused by data dependencies and multi-GPU synchronization, and as a result, cannot accurately model workload execution time.

Received 4 March 2024; revised 21 November 2024; accepted 24 November 2024. Date of publication 28 November 2024; date of current version 30 December 2024. This work was supported by an unrestricted grant/gift from Meta, Inc (then Facebook). Recommended for acceptance by D. Li. (*Corresponding author: Zhongyi Lin.*)

Zhongyi Lin is with the Advanced Micro Devices, Inc, San Jose, CA 95054 USA (e-mail: zhongyi1@amd.com).

Ning Sun, Xizhou Feng, and Louis Feng are with the Meta, Inc, Menlo Park, CA 94025 USA (e-mail: nsun@meta.com; fengx@meta.com; lofe@meta.com).

Pallab Bhattacharya is with the NVIDIA, Santa Clara, CA 95051 USA (e-mail: pallabb@nvidia.com).

John D. Owens is with the Department of Electrical and Computer Engineering, University of California, Davis, CA 90095 USA (e-mail: jowens@ucdavis.edu).

Digital Object Identifier 10.1109/TPDS.2024.3507814

¹Code is open-sourced at https://github.com/owensgroup/ml_perf_model.

Also, even in single-GPU setups, the limitations of the performance modeling of certain operators (ops) in prior work inhibit us from extending it to a broader range of workloads. As an important example, embedding lookups in real-world DLRM pipelines have unknown input distributions and loads across batches and devices, resulting in unpredictable memory access patterns and thus difficulty in predicting their performance. Existing embedding lookup performance models are rigid and can only cover limited problem sizes. Besides, extending the performance model to other workloads such as natural language processing (NLP) models requires the support of extra element-wise or minor ops.

To summarize, the performance modeling problems of communication collectives, multi-GPU stream synchronization, embedding lookups with randomly distributed input data, and additional minor ops such as *layer_norm* and *dropout* remained unsolved until we address them in this paper. We extend our previous performance-modeling work [14] and inherit its critical-path-based execution trace simulation method to model the training performance of DLRM and NLP workloads on multi-GPU platforms. The novel and major contributions include:

- (Section IV-A) Added performance models for communication operations (*all-to-all* and *all-reduce*) by improving a basic heuristic model with straightforward and efficient sigmoid curve fitting.
- (Section IV-B) An enhanced critical-path-based end-to-end (E2E) performance modeling algorithm.
- (Section IV-C) Improved embedding-table-lookup performance modeling with flexible lookup numbers and patterns (i.e., input data distribution) through an ML-based approach.
- (Section IV-D) Supported extra minor ops such as *layer_norm* and *dropout* for NLP models.

We claim that *the modeling of both inter- and intra-rank synchronizations are the keys to accurately modeling ML workloads training performance on multi-GPU, or even more broadly, all types of workloads running on multi-heterogeneous-device platforms*. Therefore, the proposal of the critical-path-based end-to-end (E2E) performance modeling algorithm (second item above) is the most significant contribution of this paper. Our performance model obtains 5.21% and 3.00% geomean prediction errors on predicting the per-iteration performance of randomly generated industrial-scale DLRM workloads and Transformer-based NLP models based on BERT [15], GPT2 [16], and XLNet [17], respectively, on two different multi-GPU platforms. In addition, in a use case of choosing a (DLRM) embedding table sharding configuration without running the workloads, our performance model achieves an 85% success rate in selecting either the fastest config or predicting the execution time with less than 10% error, showcasing our performance model's ability to generate insights for multi-GPU training optimization.

II. RELATED WORKS

A. Performance Modeling for Both Single-Device and Distributed Training

There is abundant prior art in predicting the execution time of per-batch training and/or inference of ML workloads on a single

GPU. A majority of them [8], [9], [10], [11], [12], [13], [14] focus on convolutional neural networks (CNNs), while some also cover NLP [12] and recommendation models like DLRM [14]. These works developed methodologies for accurately modeling performance on the op/kernel levels, though they did not address the challenges of multi-GPU or distributed training in general.

Early performance models for distributed training are mostly analytical and consider computation and communication separately. Yan et al. [18] studied data and model parallelism of CNNs on CPU clusters and modeled their performance and scalability. Oyama et al. [19] predicted per-batch execution time for GPU-equipped supercomputers with low error based on statistics of mini-batch size and staleness for an asynchronous SGD algorithm. It also provided some generality in terms of precision and communication standards. As a use case, both approaches claimed to be able to search for the best system configurations for distributed training of CNNs. Qi et al. [20] introduced PALEO for per-layer CNN execution time estimation and demonstrated its ability to accurately model CNN training performance at scale on cloud clusters.

However, these prior successes do not extend well to modern ML workloads. Today, the implementation of ML workloads as well as the environment and setup of training clusters have evolved significantly. This outdates features such as sub-batch size (fewer than 16 [19] versus considerably larger sizes in modern settings, e.g., 2048 to 65536), model workload size and diversity, and hardware settings. Prior work either lacks support on GPUs [18] or has a complex methodology that inhibits generalizing to other pipelines [19]. As decentralized distributed training such as DDP (distributed data-parallel) with multiple GPUs is increasingly preferred to parameter servers, modeling the communication among GPUs to account for interaction and overlap between GPU streams must also be addressed for accurate prediction. This rules out per-layer methods such as Qi et al. [20].

Wang et al. [21] implied a possible speed-of-light (SOL) model that the total execution time is the maximum among the time of data, computation, and communication, assuming these operations are perfectly overlapped. Inspired by this, and thanks to the detailed information we obtain from the execution trace [22], we further predict the training performance in general when these operations are not perfectly overlapped. Yang et al. [23] proposed the computational-graph simulation-based PerfEstimator running on a single device for distributed training of CV models, with computation-communication overlap estimated by an analytical scaling factor. In our methods, this overlap is estimated by the critical-path-based method which is less error-prone especially when workloads across devices are imbalanced.

B. Model Parallelism and Sharding

As modern ML workloads have grown rapidly in size, exploiting model parallelism has become a critical tool in scaling up ML workloads. Lepikhin et al. [24] summarize the main challenges of model parallelism, which include: 1) device underutilization due to the sequential dependency of the network and gradient-based optimization; 2) superlinear compute cost versus model

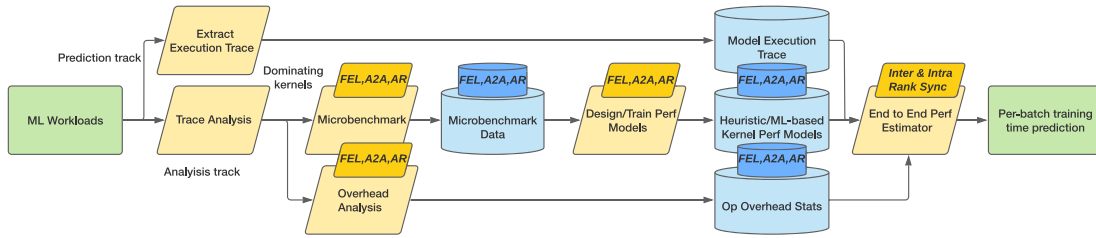


Fig. 1. An overview of our prediction pipeline, based on the one proposed by Lin et al. [14]. We mark new components with small darker shapes and italic text, such as microbenchmark and kernel-level performance models of FBGEMM embedding lookup (Section IV-C), all-to-all (A2A), and all-reduce (AR) (Section IV-A), as well as inter- and intra-rank synchronization mechanism in the critical-path algorithm (Section IV-B) to handle multi-GPU end-to-end performance prediction.

size; 3) poor infra scalability; 4) non-trivial implementation of partitioning strategies. As an example of one increasingly common use of model parallelism, modern recommendation models such as industrial DLRMs often have hundreds of embedding lookup tables with up to tens of millions of rows, which are too large to be stored on a single GPU and thus must be sharded and dispatched to multiple GPUs. Balancing the load of these sharded tables (to achieve uniform access latency and memory usage) thus becomes a critical problem that previous researchers have tried to solve in different ways. Lui et al. [25] first applied baseline (e.g., dimension-based, row-based, and size-based) heuristics to shard embedding tables in DLRMs. Sethi et al. proposed RecShard [26] that uses integer linear programming (ILP) to optimize the sharding problem. Zha et al. further improved the sharding efficiency and balance by the reinforcement-learning-based AutoShard [27] and Dreamshard [28]. In our paper, we demonstrate through a case study that our performance model can evaluate the performance of multiple sharding algorithms and quickly select the one that leads to the best E2E execution time for DLRM training. Notice that we are *not* proposing a smart sharding algorithm here as these previous works did.

III. METHODOLOGY

Fig. 1 provides an overview of our multi-GPU performance modeling pipeline with the modules that exist in our previous work [14] marked by light blocks and extended modules of this current work marked by dark blocks and italic text. The previous work builds a solid foundation for the pipeline. In the analysis track, we profile ML workloads written with PyTorch and analyze the profiler traces to collect dominating kernels/ops. Then we collect microbenchmark data for these kernels/ops, design/train and verify performance models using this data, and extract the overhead statistics for these kernels/ops. The kernel benchmark data, kernel performance models, and overhead statistics are saved as the assets of the pipeline (blue cylinders in Fig. 1)). In the prediction track, the pipeline extracts the execution trace (renamed from the “execution graph”) of an input ML workload, simulates it by traversing its ops with the pipeline assets, and generates the per-iteration E2E training time prediction of the input ML workload within several seconds.

We leverage the modularity of the previous work and extend it with a focus on performance modeling on *single-node*

multi-GPU platforms. This makes sense for two reasons: 1) the pipeline assets are reused and shared across different input ML workloads, and 2) the E2E time prediction by simulating the execution trace of the target workload and traversing ops in it is dictated by the same critical-path-based logic that is both *model-architecture-agnostic* and *platform-agnostic*. In our current work, we add kernel performance models for dominating kernels/ops of ML workloads running on multi-GPU platforms and address multi-GPU execution in the E2E simulation. *The updated pipeline inherits the main modules and procedures and reuses all kernel performance models (for GEMM, memcopy, transpose, etc) and the overhead estimator from the previous version, while the critical-path-based simulation algorithm is enriched to adapt to multi-GPU execution.* We focus on DLRM and Transformer-based NLP workloads in this work, but we expect our techniques can seamlessly extend and apply to other ML workloads (details discussed in Section VI-B).

A. Dataset Exploration

The open-source DLRM dataset [29] contains synthetic embedding lookup (EL) data that resembles the memory-access reuse pattern of Meta’s production data. The 2021 data includes one batch of 65536 EL data samples from 856 tables, while the 2022 data includes one batch of 131072 samples from 788 tables. Due to the unpredictable lookup pattern in production data, the pooling factors (average number of lookups per data sample, denoted with L) vary significantly across all tables. This is very different from the previous work [14], [30] which considers L as a fixed value across data batches and even tables in the same workload. Our exploratory analysis (Fig. 3) shows that average L values of tables in the dataset tend to concentrate in the range of $[0, 10)$, especially $[0, 1]$, while there are still some tables with average L value spreading to a few hundred. We define tables with an average L equal to greater than 20 as “heavy tables”. As shown in Fig. 3, there are 103 and 166 heavy tables in the 2021 and 2022 data respectively.

B. Multi-GPU Training Benchmark and Analysis

We first benchmark several ML workloads that will be further analyzed in later sections to understand the behavior of multi-GPU training performance. Fig. 2 shows the per-GPU-stream execution time breakdown of three workloads: *DLRM_Heavy_EL* with a batch size of 4096 (Task_14 in Section V-B1, involving

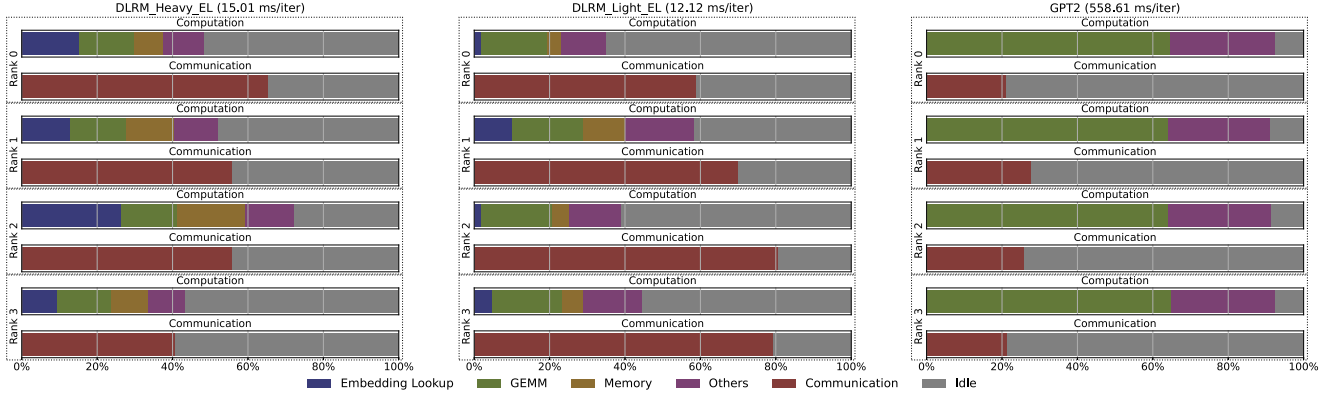


Fig. 2. Per-GPU-stream training execution time breakdown of selected ML workloads on a 4-GPU platform. The per-iteration time of each workload is provided for reference. We discuss these results in Section III-B.

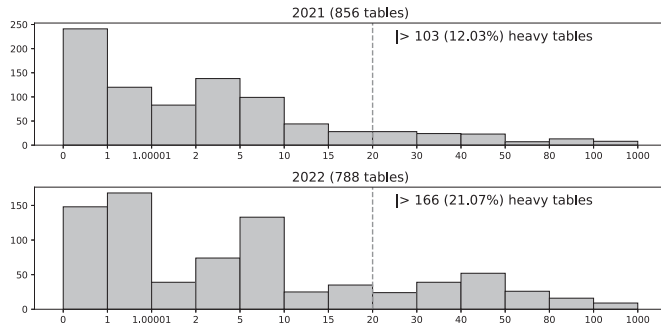


Fig. 3. Histogram of average L values of EL tables in the dataset. All bins but the last one are right-open.

heavy embedding tables only), *DLRM_Light_EL* with a batch size of 4096 (Task_5in Section V-B1, involving random embedding tables), and *GPT2* with a batch size of 64 on each rank of the 4xGV100 platform (details shown in Section V). The runtime is contributed by a few groups of kernels and types: EL (forward and backward), GEMM (forward and backward), memory operations (*concat*, *copy*, *transpose*, etc), others (normalizations, elementwise, etc), communication (*all-to-all* and *all-reduce*), and idle time. We summarize the following insights from our benchmark:

- *Communication operations could be but are not always a performance bottleneck*: Obviously, communication operations dominate the two DLRM models, mainly because in addition to *all-reduce* commonly caused by accumulating gradients in the backward pass, training the EL module in DLRM with model-parallelism introduces *all-to-all* for merging data with that from the multilayer perceptron (MLP) modules trained with data parallelism. In most pairs of streams shown here (except for *DLRM_Heavy_EL*'s rank 2), the communication stream has a longer GPU active time than the computation stream. This means it cannot be fully overlapped by the computation, and thus similar workloads tend to thrash the communication network of the interconnected multi-GPU platform. In contrast, *GPT2* is quite compute-dominated; *all-reduce* time is minor and can theoretically be fully overlapped.

- *Load balance substantially varies across the workloads*: DLRM models experience load imbalance across ranks due to the existence of EL tables with various compute and communication loads. Notice that instead of the forward and backward computation of EL itself, it is the *all-to-all* it introduces that dominates, *especially when the EL is light*. Examples such as *DLRM_Light_EL*'s rank 3 and 4 show that communication can occupy up to 80% of the execution time, in which case GPU computation resources are wasted by sitting idle. In contrast, we see an almost-perfect load balance for *GPT2* as it is purely DDP-trained and the load is evenly distributed.
- *The significance of GPU stream idle time is way more than "being non-negligible"*: What we see in single-GPU training is that idle time is contributed mostly by CPU op calls and overheads that block the scheduling of GPU kernels; data dependencies play a less important role given there is only one stream. On the contrary, in the case of multi-GPU training data dependencies of ops lead to streams that wait on each other and thus dominate waiting time. This time cannot be predicted in the same way (i.e., statistically) as the previous work such as Lin et al. [14] did, because the kernel execution time on each stream varies and thus the overlap of these kernels on different GPU streams is complicated and case-by-case. Therefore, to understand and *simulate* how GPU streams interact and synchronize with each other is the key to accurately predicting the per-iteration time of training.
- *As we support more kinds of ML workloads such as (Transformer-based) NLP models, we will see an increasing number of ops that must be modeled accurately to get an accurate result*: We see that "other" ops (*layer-norm*, *dropout*, *gelu*, *tanh*, etc.) contribute to more than 20% of execution time on *GPT2*, although each one of them only contributes a tiny amount of time every time it is executed. This might not be a surprise, given the performance of Transformer-based models has been profoundly studied these years. Techniques such as advanced layer fusion may be able to effectively optimize and reduce their execution time, but that is beyond the scope of this paper. Under our

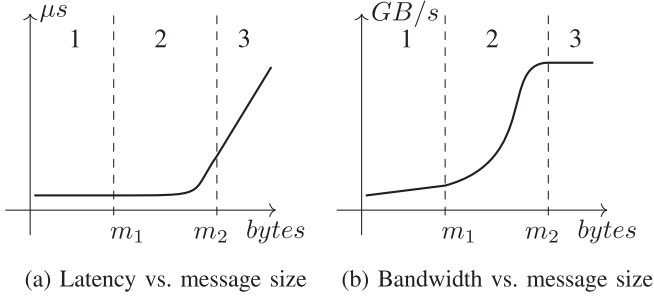


Fig. 4. Typical characteristic curves for data movement. X-axes are in log scale. m_1 and m_2 are boundary message sizes that separate the three regions.

experiment settings and from the performance-modeling point of view, this observation means more ops from this category should be supported than the previous work did.

The lessons we learn guide the direction of our research, which includes the following novel components or improvements of our performance model covered in the next section:

- Communication operations like *all-to-all* and *all-reduce*;
- E2E performance modeling for multi-GPU training.
- Embedding table lookup with flexible lookup numbers and patterns (input data distribution);
- Additional ops in (Transformer-based) NLP models, such as *layer-norm*, *dropout*, *gelu*, and *tanh*, etc;

IV. KERNEL AND END-TO-END PERFORMANCE MODELING FOR ML TRAINING ON MULTI-GPU

A. Communication Collective Performance Modeling

All-to-all and *all-reduce* are two communication collectives commonly seen in multi-GPU training of recommendation workloads and contribute significantly to the per-iteration time of multi-GPU training. The difficulty of modeling the performance of these two ops is that their performance highly depends on how multiple devices are connected (e.g., network connection pattern and medium such as NVLink and PCIe), while this configuration differs from platform to platform.

We devise a simple way to model the performance of *all-to-all* and *all-reduce* based on the message passing cost model defined by Grama et al. [31]:

$$t_{comm} = t_{cold_start} + t_{per_word} m_{size} \quad (1)$$

with t_{cold_start} representing the fixed overhead of data transaction, and t_{per_word} being the per-word data transaction time of the communication medium. Distilled from a comprehensive study on multi-GPU communication on platforms by Li et al. [32] and a prior performance model work (LogCA [33]), we make a critical observation that eventually leads to the improvement of the above model:

Regardless of the ops, network connection pattern, and network medium, as shown in Fig. 4 the curve of message size versus bandwidth can always be divided into three regions as the message size increases: 1) linear bandwidth with constant latency; 2) S-shape bandwidth curve with non-linear latency; 3) constant (saturated) bandwidth with linear latency.

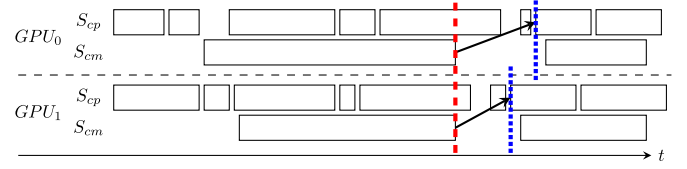


Fig. 5. Inter-rank synchronization (red dashed line) and intra-rank/inter-stream synchronization (blue dotted line). For simplicity, we assume two GPUs and two streams (S_{cp} and S_{cm} , for compute and communication respectively) per GPU, while CPU op calls are omitted in the plot. Rectangles represent GPU kernels, and arrows indicate the data dependency between compute and communication kernels.

Therefore, we model the data movement latency as a piecewise function of message size m with m_1 and m_2 being boundary sizes separating the three regions:

$$t_{comm} = \begin{cases} t_s & \text{if } m \leq m_1 \\ f(m, param) & \text{if } m_1 \leq m \leq m_2 \\ t_s + \frac{m}{BW_{max}} & \text{if } m \geq m_2 \end{cases} \quad (2)$$

where

$$f(m, param) = \frac{\log_2^m}{10^{sigmoid(m, param)}} \quad (3)$$

and $param$ include 4 parameters L , x_0 , k , and b that define a standard sigmoid function:

$$sigmoid(x) = \frac{L}{1 + e^{-k*(x-x_0)}} + b. \quad (4)$$

This method only involves 8 parameters for each op (4 for sigmoid, 2 for region boundaries, 1 for startup latency, and 1 for maximum bandwidth). It is fast and simple for both curve fitting and prediction; it is also *topology-agnostic* and thus able to generalize to any communication pattern. We utilize the PARAM benchmark [34] and generate microbenchmark data to fit curves for both ops with equal message size starting from 4 bytes on each device and doubling it until exceeding the device memory. Test data is generated with random per-device message size for *all-to-all* and equal per-device message size for *all-reduce*. We calculate the input message size to the performance model as “the maximum of (the maximum of sent/received message size per device) across all devices”.

B. Multi-GPU E2E Performance Modeling

As we mentioned in Section III-B, the main difference between single-GPU and multi-GPU performance modeling is that in multi-GPU scenarios, communication collectives like *all-to-all* and *all-reduce* simultaneously execute on multiple devices and trigger synchronizations and waiting across ranks and streams. This additional behavior significantly impacts runtime. This subsection describes how we simulate synchronizations and waiting during the execution analysis.

Before we dive deep into the multi-GPU E2E performance modeling algorithm, we need to first understand two types of synchronization in distributed training (depicted in Fig. 5):

- *inter-rank synchronization* that occurs at the termination of a communication collective kernel, and

- *intra-rank or inter-stream synchronization* that happens at the launch of a compute/memory kernel that depends on the last communication kernel.

We expect communication kernels' launch time to differ across ranks because each rank's latency to reach these ops due to unbalanced prior loads, such as data movement and computation, is different. However, since communication kernels are synchronous, theoretically they must terminate simultaneously. In reality, they do terminate almost simultaneously with negligible variance. This is the *inter-rank synchronization*, marked by the red dashed line in Fig. 5. When it occurs, the time fronts of the communication streams on different ranks should be set to the same during analysis. Nevertheless, since the output data of the communication ops are used as inputs to certain successive ops, successive communication kernels, e.g., the last communication kernels on S_{cm} on both ranks in Fig. 5, should be launched at least after the *first dependent kernel* of the previous communication kernel (marked in blue dashed lines) rather than right after the previous communication kernel. This is called *intra-rank/inter-stream synchronization*. To reflect it, we set the communication stream time front to the same as the launch time of the first dependent kernel.

We claim that *the modeling of both inter-rank and intra-rank synchronizations are the keys to accurately modeling ML workloads training performance on multi-GPU, or even more broadly, all types of workloads running on multi-heterogeneous-device platforms*. This is because the best modeling method without considering them is to sum the kernel time per GPU stream and take the maximum as the E2E prediction result. It might work for a few workloads dominated by one certain type of stream, e.g., GPT2 being dominated by the compute stream as shown above. However, in most cases, this will miss a large amount of GPU idle time (e.g., the gap between the red and blue lines in Fig. 5) created by data dependencies and other possible waiting conditions during the execution, leading to a significant and unacceptable underestimation of execution time, i.e., up to 60% as we show later in Section V-B1.

To consider these synchronizations, we need the information on data dependency between ops in the ML model/workload, which is thankfully provided by the PyTorch execution trace (ET). We extend the critical-path-based algorithm proposed by Lin et al. [14] for single-GPU performance modeling to adapt to multi-GPU scenarios by incorporating inter-rank and intra-rank synchronizations, as shown in Algorithm 1. The new algorithm has a similar basic structure to the single-GPU version. Based on our observation, we assume two streams (one for compute and memory kernels and one for communication kernels) on each rank as that is the case for most ML workloads. Instead of running a single process for training time prediction on a single GPU, the new algorithm runs N processes in parallel, processing N EGs and predicting execution time for N ranks simultaneously (line 4). In addition to CPU time and GPU active time, we also track communication stream time T_{cm} and compute/memory stream time T_{cp} per process (line 6). At line 7, we also initialize a variable *last_comm_op* to keep track of the dependency between compute/memory ops and the last communication op. To recap, T1–T5 in line 9 represents

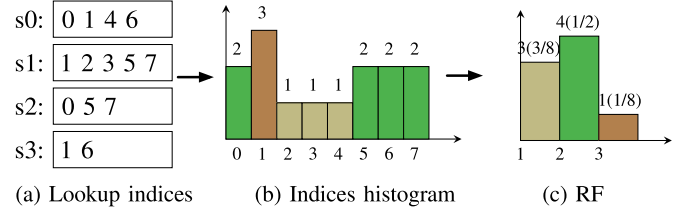


Fig. 6. Generation of reuse factors (RF) from EL indices. (a) is an example of lookup indices of a batch of 4 samples on one embedding table. In (b), the x-axis is the indices and the y-axis is the count of these indices. In (c), the x-axis is the count of accesses and the y-axis is the number of such indices.

the five types of execution overhead on the CPU of each op, which are summarized in our previous work [14]. During the traversal of ETs, we detect intra-rank synchronization when an op that has data dependency on the last communication op or a new communication op is encountered (lines 12–15); when the current op is a communication op, we do inter-rank synchronization after processing all kernels of it (lines 29–31) and update *last_comm_op* for the next iteration (lines 36–38). At the end (lines 40–41), we calculate the local total time and GPU active time and do an *all_gather* operation across all ranks to collect their values, the maximums of which are returned as the predicted total time and GPU active time.

C. Embedding Lookup Kernel Modeling

The performance modeling of *EL*, an indispensable and commonly dominating [14] operator in many recommendation models including DLRM, is challenging in a few aspects. First, the dimensions and parameters of every single embedding table, specifically the *number of embeddings* (E), the *embedding dimension* (D), and the *pooling factor* (L) can all be different in real workloads and data. Second, the input data distribution of each table varies considerably across each batch. Both factors significantly increase the difficulty of applying a heuristic-based performance model to the *EL* op, since it is almost impossible to estimate the L2 cache hit rate and data movement accurately and thus predict the latency of the op. Previous work such as Lin et al. [14] only considers modeling the performance of the case when D , E , and L are all fixed for each table with uniform input data distribution, making their solution insufficient to deal with real-world workloads and data.

To address this issue, we adopt a straightforward and clear expression of describing *EL*'s input data distribution using *reuse factors* (RF), as introduced by Meta's open-source DLRM dataset [29]. Fig. 6 shows an example of generating reuse factors (RF) from EL indices. First, the histogram of per-table index counts in a batch of input data is calculated ((a) to (b)). Then, the "histogram of histogram", i.e., how many indices are accessed once, twice, thrice, etc., is calculated, which is thus the RF of this batch on a certain table ((b) to (c)). In practice, the bins for this step have sizes of exponents of 2, e.g., $[0, 1)$, $[1, 2)$, $[2, 4)$, $[4, 8)$, etc. Finally, the counts of every bin are normalized into the range of $[0, 1]$. In this way, the distribution of lookup indices is described as by what probability any row of a table is hit by

Algorithm 1: Multi-GPU E2E Training Performance Model.

```

1: Input:  $ET_{0,1,\dots,N-1}$  of an ML workload trained on a
   single-node N-GPU platform, one execution trace per rank;
   Kernel performance models  $\{M\}$ ; Overhead statistics  $Ov$ .
2: Output: Predicted per-batch training time  $T$ .
3: Spawn  $N$  processes  $P_{0,1,\dots,N-1}$ .
4: parfor  $i \leftarrow 0, N - 1$  do
5:   Initialize  $cpu\_time = 0$  and  $gpu\_time = 0$  for  $P_i$ .
6:   Initialize communication stream time  $T_{cm}$  and compute
     stream time  $T_{cp}$  for process  $P_i$ .
7:   Initialize  $last\_comm\_op$  as none.
8:   for each  $op$  in  $EG_i$  do
9:     Look up  $T_1, T_2, T_3, T_4, T_5$  from  $Ov$  for  $op$ .
10:    Identify current stream  $s$  (as  $cm$  or  $cp$ ).
11:     $\triangleright$  Intra-rank synchronization
12:    if  $op$  depends on  $last\_comm\_op$ , or  $op$  is an
        communication op then
13:      Set  $last\_comm\_op$  to none.
14:      Synchronize  $T_{cm}$  and  $T_{cp}$  for process  $P_i$ .
15:    end if
16:     $cpu\_time \mathrel{+}= T_1$ 
17:    if  $op$  has kernel calls then
18:       $cpu\_time \mathrel{+}= T_2$ 
19:      for each kernel call  $k$  under  $op$  do
20:        Predict kernel time  $T_k$  with  $M$ 
21:         $T_s = \max(T_s + 1, cpu\_time + T_4/2) + T_k$ 
22:        Update  $gpu\_time$  with  $T_s$  and  $T_k$ .
23:         $cpu\_time \mathrel{+}= T_4$ 
24:        if  $k$  is not the last kernel then
25:           $cpu\_time \mathrel{+}= T_5$ 
26:        end if
27:      end for
28:       $\triangleright$  Inter-rank synchronization
29:      if  $op$  is a communication op then
30:        Synchronize  $T_{cm}$  across all processes.
31:      end if
32:       $cpu\_time \mathrel{+}= T_3$ 
33:    else
34:       $cpu\_time \mathrel{+}= T_5$ 
35:    end if
36:    if  $op$  is a communication op then
37:       $last\_comm\_op = op$ 
38:    end if
39:  end for
40:   $T = \max(T_{cm}, T_{cp}, cpu\_time)$ 
41:  Synchronize  $T$  and  $gpu\_time$  across all processes and
    return their maximums.
42: end parfor

```

a certain range of times, such as 2^m to 2^n times where m and n are both integers and $m < n$. Meta's DLRM dataset provides the RF of 17 bins for each data file, and we follow this convention in our experiments. The advantages of RF are its low overhead in terms of processing time and parameter size. The calculation is relatively simple, and thus it is trivial to obtain RF values of each batch of input data on the fly before the training of a batch is started. Because each table has 17 RF values and the number of tables residing on each rank is usually limited, e.g., around

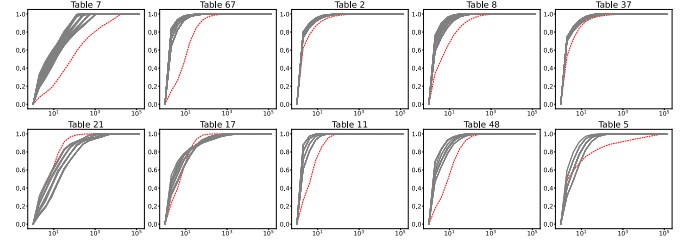


Fig. 7. RF values of the top 10 most visited tables among 6000 random samples from the DLRM open-source dataset for the EL microbenchmark. Red dashed lines represent the cumulative distribution function (CDF) of the overall RF values of each table in the whole dataset, while each thin gray line represents the CDF of RF values of a sample that visits a certain table. It can be observed that the data distribution (represented by RF values) of each data batch is different and sometimes distant from that of the whole dataset. Compared to overall RF values, the RF values of a data batch tend to concentrate towards 0 along the X axis.

10, the total number of RF values for each batch of input data remains acceptable.

We adapt the cost model proposed by Zha et al. [28] and use it in real-time performance modeling of *EL*. Zha et al. create a multi-head MLP cost model to simultaneously predict the cost of the forward time, backward time, and communication time of a sharding scheme of *EL* as a whole on multiple devices. Our method is different in two aspects. 1) We handle these three constituent times separately in the granularity of ops to accommodate our robust E2E algorithm shown in Section IV-B. This better simulates and models the actual execution since considering these three times as a whole misses the opportunity to explain the computation-communication overlaps (described in Section IV-B). 2) Instead of using RF values of each table, we use RF values of each *batch* for better prediction based on the reason stated in the last paragraph. We construct our MLP performance model resembling one of the heads in Zha et al.'s cost model and match those used for other kernels for convenient training and inference. To enable *EL*'s execution incorporating flexible D , E , and L , we use a popular, efficient, and flexible batched-EL implementation provided by FBGEMM [35], an open-source high-performance kernel library for training and inference on both CPUs and GPUs. We first create a microbenchmark dataset by randomly sampling tables and batches (e.g., 10 tables out of 856, and a batch of 1024 lookups out of 65536) from the DLRM open-source dataset, and make sure that the size of the sampled tables in each sample will not exceed the DRAM memory size of the selected GPU. Although Zha et al. trained the cost model with RF values of *the whole dataset*, it is improper to do so for real-time performance modeling because RF values tend to spread in a larger range. This fact is easy to understand intuitively. Suppose the batch size is small, the number of times each row in a table is hit tends to concentrate close to 0 because there are few indices in the batch. If we consider the whole dataset as a huge batch (e.g., size 65536 for the DLRM dataset we use), there are likely some rows in a table to be hit many times, such as 500 times, which could never appear with a small batch size such as 256. In a word, these two distributions are different (as shown in Fig. 7), and thus the distribution of the

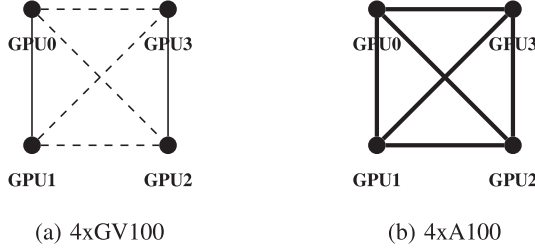


Fig. 8. Communication topologies of the two multi-GPU platforms used in our experiments. Thin lines: 4 NVLinks (NV4); thick lines: 12 NVLinks (NV12); dashed lines: PCIe.

whole dataset is not representative of a batch of data sampled from it. Therefore, we calculate the RF values of *each batch* and store them together with the execution time as part of the training data. Finally, we randomly split the microbenchmark result dataset into a training set and a test set with a split factor of 0.8, and use them to train and validate the performance models.

D. Additional Ops Performance Modeling Support

To support Transformer-based NLP models, We also add kernel performance models for additional ops including *layer_norm*, *dropout*, and element-wise ops such as *gelu* and *tanh*. By applying principles mentioned in our previous work [14], ML-based models suit *layer_norm* and *dropout* for their hidden kernel implementation detail in PyTorch, while the roofline model suits element-wise ops for their simplicity. Therefore we predict the latency of *layer_norm*, *dropout* with ML-based models trained with microbenchmark data of PyTorch ops for and predict *gelu* and *tanh* using the roofline model.

V. EVALUATION AND ANALYSIS

We evaluate our kernel and E2E performance models with PyTorch v2.0, FBGEMM v0.4.1, CUDA 11.7, and Python 3.9. The performance model of FBGEMM’s embedding lookup kernel is evaluated on single NVIDIA GV100 and A100 (40 GB) GPUs, while that of both communication kernels and multi-GPU E2E are assessed on two multi-GPU platforms, including 4xGV100 equipped with 48-core Intel(R) Xeon(R) Gold 6146 CPU @ 3.20 GHz, and 4xA100 equipped with GCP’s a2-highgpu-4g with 48-vCPU. Fig. 8 shows the GPU communication topologies of these two platforms. We use data sampled from the DLRM open-source dataset [29] as both microbenchmark data for FBGEMM embedding lookup kernel performance model training and verification, and input data for embedding lookup of the DLRM models in E2E tests. Multiple pieces (.pt files) of the dataset are merged for later use. The multi-GPU E2E evaluation in this work covers DLRM model training (code adapted from <https://github.com/facebookresearch/dlrm>) and fine-tuning of natural language processing (NLP) models such as BERT [15], GPT2 [16], and XLNet [17] (all model implementations called from HuggingFace’s Transformers library [36]). We use PyTorch’s DistributedDataParallel (DDP) in distributed training of ML workloads in our experiments for no GIL contention and less model replication and data movement overheads [37].

TABLE I
PREDICTION ERROR OF FBGEMM EMBEDDING LOOKUP, ALL-TO-ALL, AND ALL-REDUCE KERNEL PERFORMANCE MODELS

Kernel	4xGV100			4xA100		
	GMAE	MAPE	std	GMAE	MAPE	std
A2A	6.28%	9.42%	7.76%	5.25%	7.14%	4.72%
AR	6.35%	9.17%	8.23%	4.98%	6.77%	4.11%
ELF	4.37%	7.11%	7.44%	5.64%	9.17%	9.52%
ELB	3.08%	4.42%	3.39%	3.63%	5.44%	4.33%

Abbreviations: A2A (all-to-all), AR (all-reduce), ELF (embedding lookup forward), ELB (embedding lookup backward).

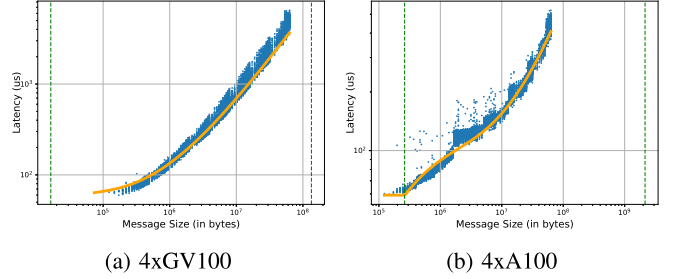


Fig. 9. Fitted curves for *all-to-all* benchmark data on both platforms. Message size section boundaries are plotted as green vertical dashed lines.

In addition, we manually insert a barrier at the beginning of each batch for better profiling of communication ops; we set the bucket size of gradient bucketing [38] to the default value (25) for all our experiments, as we observe that varying it only causes a negligible performance change.

A. Kernel Performance Modeling

We provide prediction errors in two metrics: GMAE (geometric mean absolute error) to reduce the impact of outliers that is common when measuring E2E latency, and MAPE (mean absolute percentage error) as an intuitive measurement of deviation as an (always positive) percentage. We obtain less than 10% both GMAE and MAPE for all kernel performance models shown in Table I. Particularly, the adjustment of *all-to-all* message size yields a low prediction error of latency, implying that the operation is bounded by the biggest per-device data bulk sent from or received by one certain device. Specifically, we present the fitted curves for *all-to-all* benchmark data on both platforms in Fig. 9. We observe that the practical problem size for *all-to-all* in DLRM workloads lies in Section 2 (the transitional section), which justifies our improvement on (1).

B. Multi-GPU E2E Performance Modeling

1) *DLRM on Multi-GPU Platforms*: In our experiments to predict the multi-GPU training performance of DLRM models, we first generate 20 tasks with embedding tables randomly sampled from the DLRM open-source dataset. Among these 20 tasks, each of the 2021 and 2022 parts of the dataset contributes 5 heavy tasks (i.e., all sampled tables are heavy) and 5 normal tasks (i.e., tables can either be heavy or light). The total number of embedding tables per task is within the range of

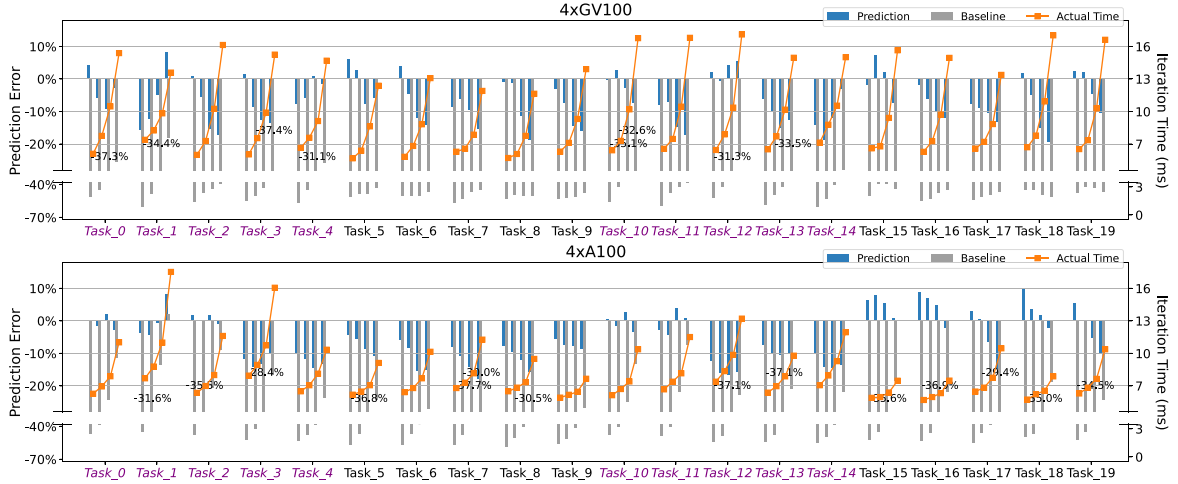


Fig. 10. Prediction, baseline, and reference of multi-GPU training performance of DLRMs on two multi-GPU platforms. Notice that the “Task_x”s on different platforms do not have the same embedding table configuration as the random task generation is platform-specific. Tasks 0–9 and 10–19 on each platform are from the 2021 and 2022 parts of the dataset, respectively. Heavy tasks are marked with purple color and italics. Percentage figures are the error of bars ending in the broken area.

TABLE II
STATISTICS OF DLRM E2E TIME PREDICTION ERRORS ACROSS TWO
MULTI-GPU PLATFORMS

Overall			4xGV100			4xA100		
g.m.	min	max	g.m.	min	max	g.m.	min	max
5.21%	0.05%	19.38%	5.60%	0.27%	19.38%	4.85%	0.05%	17.87%

$(0.7 \sim 1.3) \times \#GPU \times 13$ to resemble real-world workloads. To prevent out-of-memory errors, we limit the total memory footprint of all embedding tables on each rank to less than 80% DRAM size of the GPU for each task. We set the embedding table sharder to `size_lookup_greedy` based, i.e., the cost of each table is estimated as $L \times D \times \log(E)$. For each training iteration in the E2E test of each task, we randomly sample a mini-batch of data (with batch size set to 512/1024/2048/4096) corresponding to each table selected in the task from the dataset and distribute it to each rank based on the sharding scheme. The overhead statistics (mean latency, etc) of PyTorch ops are aggregated from all collected traces and shared by *all* tested workloads. The actual/predicted time is measured/calculated by averaging over 30 iterations. It takes about a day to run the benchmark and analysis for all 80 (20 workloads $\times$ 4 batch sizes) tasks, while it takes less than 20 minutes to predict their performance.

We present the statistics of all E2E performance modeling tests on the two multi-GPU platforms in Table II, and the prediction error and reference time of each task in Fig. 10. The baseline result compared with our prediction in Fig. 10 is given by the maximum sum of kernel active time of each GPU stream. We can see that this baseline prediction, yielding error values higher than 60%, is insufficient to be used as the predicted E2E time per iteration because no idle or waiting time caused by data dependency between the streams is considered. Instead, our enhanced algorithm can accurately predict both

normal and heavy tasks with high accuracy, with an overall geomean prediction error of 5.21%. Most prediction results on both devices underestimate the actual time, possibly because the communication time dominates the per-iteration time when batch size is big, so the syncing and waiting time among all ranks also increases and contributes to the per-iteration time. The remaining tests that overestimate might be explained by the overestimation of GPU idle time caused by CPU overheads when the workload is latency-bound. Device-wise, the behavior of prediction errors, such as geomean/minimum/maximum error and trends when the batch size changes, does not deviate much, which justifies the consistency and stability of our prediction algorithm across platforms.

2) *NLP Models Performance Modeling*: We also test our E2E performance model on Transformer-based NLP models including BERT, GPT2, and XLNet. In Fig. 11, we see that the absolute prediction errors are less than 10% in all tests except two, which are slightly larger. The geomean prediction error of all the presented tests is 3.00%. Also, the variance of the prediction errors is lower than that of DLRM workloads. The reasons are 1) these NLP models are compute (GEMM) dominated with the communication stream being well-overlapped by the compute stream, and 2) the loads across devices are more balanced than the DLRM workloads, thus intra-rank and inter-rank synchronizations are rare and have little interference in the prediction. With our highly accurate kernel performance models, we can precisely predict the aggregation of compute kernel time and subsequently the E2E time of these models.

VI. APPLICATION AND DISCUSSION

A. Case Study: Fast Sharding Config Selection Using Performance Modeling

Industrial DLRM models can take days to train. Therefore, selecting a sharding configuration (i.e., the way to distribute

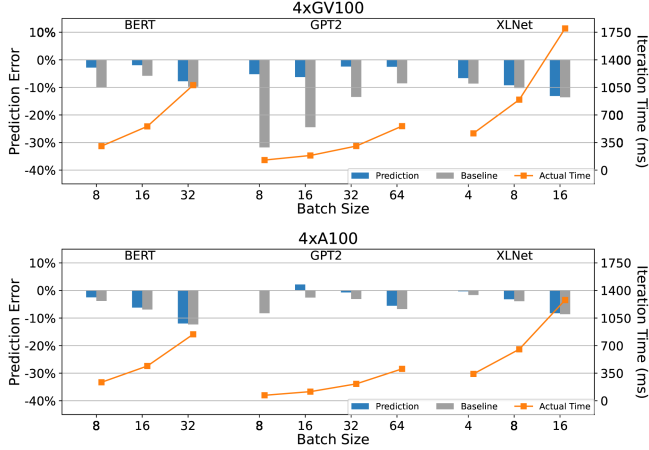


Fig. 11. Prediction, baseline, and reference of multi-GPU training performance of BERT, GPT2, and XLNet on two multi-GPU platforms. Batch sizes greater than 32 on BERT and 16 on XLNet result in out-of-memory errors and these tests are thus skipped.

embedding tables to multiple GPUs) that balances the loads on GPUs and speeds up the E2E per-iteration time is critical to reducing their training costs. Industrial sharding configs might consist of a sharding algorithm (greedy, multi-cost greedy, etc.), cost functions (multi-cost, memory-based, compute-based, etc.), table partition (column-based or row-based), and memory placement of tables (HBM/UVM). Previously, the best configs were selected by benchmarking and grid-searching over a big search space formed by these factors, which can take as long as one day per workload. We consider using our performance model for this task so that *without running the workload*, the selection time can be shortened to around 1 minute, with E2E per-iteration time predicted in seconds for each config. Notice again that we are *not*, like some previous works did [26], [27], [28], proposing a smart sharding algorithm here to achieve optimal performance. Instead, the goal is to quickly evaluate various sharding algorithms or configs and pick the fastest one for a specific problem size *without benchmarking the model*.

We conduct an experiment to demonstrate how our performance model can quickly select the best sharding config for DLRM training on a multi-GPU system. To show our idea with simplicity, we only consider sharding algorithms as the config and omit all other factors mentioned above. We also exclude recently sophisticated sharders [26], [27], [28], although it is straightforward to integrate and test them in the future from an engineering point of view. We use six sharders listed in Table III for the experiment. In addition, we pick the 10 heavy tasks from the 20 tasks on each device generated in Section V-B1 and run them again with the batch size 4096 and all sharders except for size_lookup_greedy. This is to guarantee that embedding lookup latency dominates the E2E time and thus the sharding config is likely to make a difference.

Table IV shows the actual and prediction time of using different sharding configs in selected tasks trained with platforms 4×GV100 and 4×A100. We set the success criterion to be *either the performance model accurately selects the fastest config*,

TABLE III
SHARDERS AND THEIR INDEXING (I) OR COST (C) FUNCTIONS, WHERE x REPRESENTS AN EMBEDDING TABLE

Sharders	Functions
naive	(i) $x.idx \% ngpus$
random	(i) $random(ngpus)$
size_greedy	(c) $x.E$
lookup_greedy	(c) $x.L \cdot x.D$
norm_lookup_greedy	(c) $x.L/x.E$
size_lookup_greedy	(c) $x.L \cdot x.D \cdot \log_{10}(x.E)$

An indexing function assigns a table directly to a certain rank, while a cost function estimates the cost of a table for the greedy algorithm.

or the absolute error between the actual time of the predicted fastest config and the time of the actual fastest config is less than 10%. The reason for using this criterion is that we not only care about whether the fastest config is selected but also how close the actual time of the predicted fastest config is from the actual fastest time when the fastest config is *not* selected. This is because in practice failing to select the fastest config is tolerable as long as the actual time of the selected config is close enough to the actual fastest time. In the rightmost column, we see that the prediction result of our performance model meets the criterion in 17 out of 20 (85%) tasks. In the 3 remaining tasks, the absolute errors are only 13.76% at most. This demonstrates that our performance model is accurate enough to aid multi-GPU training optimization with low time and compute cost. Since we have demonstrated that our performance model has low prediction errors both generally and in individual cases, we are confident that it can also perform well in unseen future cases.

B. Discussion

The biggest advantage of our performance model is its strong adaptability to new ML models. It is essentially an *execution simulator* built in the granularity of kernels and ops. Supporting new ML models is straightforward, requiring only adding the missing performance models and overhead statistics of *any* new kernels/ops. The library of supported kernels/ops grows as the system evolves, making it even easier to support new models later. More importantly, from the performance perspective, how the workload is *executed* is more fundamental than how it is constructed because the same ML model can be trained with different strategies—e.g., data-parallel (DP), DDP, model-parallel (MP), pipelining, etc.—with completely different performance. However, our performance model can still handle all these cases because the easy-to-obtain execution trace (the input to our performance model) has information on both the model architecture and the execution. This means our performance model will extend well to unseen future workloads and execution paradigms.

Currently, this work has the following limitations:

- Our prediction pipeline only supports ML workloads in FP32 precision. However, it can be seamlessly adapted for workloads in other precision types (FP16, INT8, etc.) by preparing kernel performance models for all kernels/ops with these types.

TABLE IV
EMBEDDING TABLE SHARDING CONFIG SELECTION EXPERIMENT RESULTS

Platforms	Tasks	Fastest (P)	Time (P, us)	Time (AP, us)	Fastest (A)	Time (A, us)	Abs Error	Meets Criterion?
4xGV100	Task 0	naive	13.40	15.80	size_lookup_greedy	15.39	2.66%	✓
	Task 1	naive	12.13	15.46	size_lookup_greedy	13.59	13.76%	✗
	Task 2	naive	12.40	15.81	random	14.74	7.26%	✓
	Task 3	size_lookup_greedy	13.13	15.25	random	14.47	5.39%	✓
	Task 4	norm_lookup_greedy	12.26	15.54	lookup_greedy	14.00	11.00%	✗
	Task 10	naive	11.77	15.13	naive	15.13	0.00%	✓
	Task 11	naive	11.75	14.22	naive	14.22	0.00%	✓
	Task 12	naive	12.39	15.07	naive	15.07	0.00%	✓
	Task 13	norm_lookup_greedy	12.61	16.04	size_lookup_greedy	14.96	7.22%	✓
	Task 14	naive	11.87	14.46	lookup_greedy	13.13	10.13%	✗
4xA100	Task 0	naive	8.68	10.62	naive	10.62	0.00%	✓
	Task 1	lookup_greedy	13.15	15.05	random	14.52	3.65%	✓
	Task 2	naive	9.81	11.43	naive	11.43	0.00%	✓
	Task 3	naive	11.31	14.04	random	13.85	1.37%	✓
	Task 4	naive	8.64	10.10	naive	10.10	0.00%	✓
	Task 10	random	8.69	10.85	size_lookup_greedy	10.38	4.53%	✓
	Task 11	naive	9.18	11.40	naive	11.40	0.00%	✓
	Task 12	naive	9.26	11.76	size_greedy	11.54	1.91%	✓
	Task 13	naive	8.08	10.32	size_lookup_greedy	9.79	5.41%	✓
	Task 14	naive	8.13	9.67	norm_lookup_greedy	9.18	5.34%	✓

Abbreviations: P: predicted; A: actual; AP: actual (time) of predicted (fastest config). Abs error is given by $|Time_{AP} - Time_A| / Time_A \times 100\%$, which is 0 when the selected fastest config is exactly the actual fastest config. Notice again that Task_x on different platforms are different workloads.

- This work currently only covers single-node multi-GPU performance modeling. To extend it to multi-node multi-GPU platforms, kernel performance models of *all-to-all* and *all-reduce* for the multi-node communication network must be prepared. Algorithm 1 should also be slightly modified to track multiple processes from all nodes. This will further adapt the performance prediction pipeline to the industrial environment for enormous ML workloads such as large language models (LLM).
- The training data loading from data centers, usually impacted by uncontrollable factors like network speed and data center setups, might cause load imbalance in the industrial environment and is not considered in this work.
- Some additional infrastructure features, such as the support of capturing dynamic tensor sizes and fused ops information in the execution trace, can increase the robustness of this work on various types of ML workloads (such as training NLP models with variable-input-length and no padding) and cooperate with modern ML compilation and optimization techniques.

We plan to extend our code base to support these features.

VII. CONCLUSION AND FUTURE WORK

We extend our previous work on single-GPU performance model with two critical components for more accurate single- and multiple-GPU performance modeling: performance modeling of communication collectives and E2E training time prediction with inter-rank and intra-rank synchronization to enable multi-GPU training performance modeling of ML workloads, and support for input-data-distribution-aware performance modeling of embedding lookup as a general enhancement. We achieve high prediction accuracy on various types of ML workloads such as randomly generated DLRM and NLP models,

and also demonstrate the performance model's ability to speed up DLRM training through the use case of embedding table sharding config selection.

There are a series of future works. Beyond the issues identified in Section VI-B, the data-distribution-aware method we use to model the performance of FBGEMM embedding lookup can be generalized to other sparse ops such as SpMM and SpGEMM to handle the training and inference performance prediction of future sparsified neural networks.

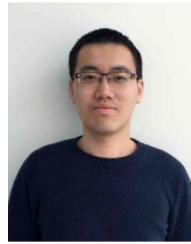
ACKNOWLEDGMENT

The authors sincerely thank Daochen Zha, Valentin Andrei, and Yufei Zhu for their constructive feedback on this work. This work was done when Zhongyi Lin was a Ph.D. student at UC Davis and a research intern at Meta, Inc. This work was done when Pallab Bhattacharya was with Meta, Inc.

REFERENCES

- [1] M. Naumov et al., "Deep learning recommendation model for personalization and recommendation systems," 2019, *arXiv: 1906.00091*.
- [2] D. Mudigere et al., "Software-hardware co-design for fast and scalable training of deep learning recommendation models," in *Proc. 49th Annu. Int. Symp. Comput. Archit.*, New York, NY, USA, 2022, pp. 993–1011.
- [3] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf
- [4] S. Smith et al., "Using deepspeed and megatron to train megatron-turing NLG 530B, a large-scale generative language model," 2022, *arXiv:2201.11990*.
- [5] A. Chowdhery et al., "PaLM: Scaling language modeling with pathways," 2022, *arXiv:2204.02311*.
- [6] S. Zhang et al., "OPT: Open pre-trained transformer language models," 2022, *arXiv:2205.01068*.
- [7] H. Touvron et al., "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

- [8] D. Justus, J. Brennan, S. Bonner, and A. S. McGough, "Predicting the computational cost of deep learning models," in *Proc. IEEE Int. Conf. Big Data*, 2018, pp. 3873–3882.
- [9] Z. Pei, C. Li, X. Qin, X. Chen, and G. Wei, "Iteration time prediction for CNN in multi-GPU platform: Modeling and analysis," *IEEE Access*, vol. 7, pp. 64788–64797, 2019.
- [10] S. Li, R. J. Walls, and T. Guo, "Characterizing and modeling distributed training with transient cloud GPU servers," in *Proc. IEEE 40th Int. Conf. Distrib. Comput. Syst.*, 2020, pp. 943–953.
- [11] Y.-C. Liao, C.-C. Wang, C.-H. Tu, M.-C. Kao, W.-Y. Liang, and S.-H. Hung, "PerfNetRT: Platform-aware performance modeling for optimized deep neural networks," in *Proc. Int. Comput. Symp.*, 2020, pp. 153–158.
- [12] G. X. Yu, Y. Gao, P. Golikov, and G. Pekhimenko, "Habitat: A runtime-based computational performance predictor for deep neural network training," in *Proc. USENIX Annu. Tech. Conf.*, 2021, pp. 503–521. [Online]. Available: <https://www.usenix.org/conference/atc21/presentation/you>
- [13] A. Rajagopal and C. Bouganis, "perf4sight: A toolflow to model CNN training performance on edge GPUs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, Los Alamitos, CA, USA, 2021, pp. 963–971.
- [14] Z. Lin et al., "Building a performance model for deep learning recommendation model training on GPUs," in *Proc. IEEE 29th Int. Conf. High Perform. Comput., Data, Analytics*, 2022, pp. 48–58.
- [15] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, Association for Computational Linguistics, 2019, pp. 4171–4186.
- [16] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, Feb. 14, 2019. [Online]. Available: <https://openai.com/blog/better-language-models/>
- [17] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., 2019, pp. 5753–5763.
- [18] F. Yan, O. Ruwase, Y. He, and T. Chilimbi, "Performance modeling and scalability optimization of distributed deep learning systems," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, New York, NY, USA, 2015, pp. 1355–1364.
- [19] Y. Oyama, A. Nomura, I. Sato, H. Nishimura, Y. Tamatsu, and S. Matsuoka, "Predicting statistics of asynchronous SGD parameters for a large-scale distributed deep learning system on GPU supercomputers," in *Proc. IEEE Int. Conf. Big Data*, 2016, pp. 66–75.
- [20] H. Qi, E. R. Sparks, and A. Talwalkar, "Paleo: A performance model for deep neural networks," in *Proc. 5th Int. Conf. Learn. Representations*, Toulon, France, 2017. [Online]. Available: <https://openreview.net/forum?id=SyVVJ85lg>
- [21] M. Wang et al., "Characterizing deep learning training workloads on Alibaba-PAI," in *Proc. IEEE Int. Symp. Workload Characterization*, Los Alamitos, CA, USA, 2019, pp. 189–202.
- [22] S. Sridharan et al., "Chakra: Advancing performance benchmarking and co-design using standardized execution traces," 2023, *arXiv:2305.14516*.
- [23] C. Yang et al., "PerfEstimator: A generic and extensible performance estimator for data parallel dnn training," in *Proc. IEEE/ACM Int. Workshop Cloud Intell.*, 2021, pp. 13–18.
- [24] D. Lepikhin et al., "Gshard: Scaling giant models with conditional computation and automatic sharding," in *Proc. Int. Conf. Learn. Representations*, Austria, 2021. [Online]. Available: <https://openreview.net/forum?id=qrw7XHTmYb>
- [25] M. Lui et al., "Understanding capacity-driven scale-out neural recommendation inference," in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw.*, 2021, pp. 162–171.
- [26] G. Sethi, B. Acun, N. Agarwal, C. Kozyrakis, C. Trippel, and C.-J. Wu, "RecShard: Statistical feature-based memory optimization for industry-scale neural recommendation," in *Proc. 27th ACM Int. Conf. Architectural Support Program. Lang. Operating Syst.*, New York, NY, USA, 2022, pp. 344–358.
- [27] D. Zha et al., "Autoshard: Automated embedding table sharding for recommender systems," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, New York, NY, USA, 2022, pp. 4461–4471.
- [28] D. Zha et al., "Dreamshard: Generalizable embedding table placement for recommender systems," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 15190–15203. [Online]. Available: https://openreview.net/forum?id=_atSgd9Np52
- [29] Meta, "DLRM open-source datasets," Dec. 2021. [Online]. Available: https://github.com/facebookresearch/dlrm_datasets
- [30] A. Tulloch, Batch embedding lookup GPU kernel and more, 2020, May [Online]. Available: <https://github.com/ajtulloch/sparse-ads-baselines>
- [31] A. Grama and A. Gupta, *Introduction to Parallel Computing*, 2nd ed. Noida, Uttar Pradesh, India: Pearson, 2003.
- [32] A. Li et al., "Evaluating modern GPU interconnect: PCIe, NVLink, NV-SLI, NVSwitch and GPUDirect," *IEEE Trans. Parallel Distrib. Syst.*, vol. 31, no. 1, pp. 94–110, Jan. 2020.
- [33] M. S. B. Altaf and D. A. Wood, "LogCA: A high-level performance model for hardware accelerators," in *Proc. ACM/IEEE 44th Annu. Int. Symp. Comput. Archit.*, 2017, pp. 375–388.
- [34] Meta, "PARAM benchmark," Sep. 2020. [Online]. Available: <https://github.com/facebookresearch/param>
- [35] D. S. Khudia et al., "FBGEMM: Enabling high-performance low-precision deep learning inference," 2021, *arXiv:2101.05615*.
- [36] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process.: Syst. Demonstrations*, Association for Computational Linguistics, 2020, pp. 38–45.
- [37] Meta, "Comparison between DataParallel and DistributedDataParallel," May 2023. [Online]. Available: https://pytorch.org/tutorials/intermediate/ddp_tutorial.html#comparison-between-dataparallel-and-distributeddataparallel
- [38] S. Li et al., "PyTorch distributed: Experiences on accelerating data parallel training," *Proc. VLDB Endowment*, vol. 13, no. 12, pp. 3005–3018, Aug. 2020.



Zhongyi Lin received the PhD degree in electrical and computer engineering from the University of California, Davis. He is a senior software engineer with AMD, responsible for inference optimization of Large Language Models on specialized hardware for artificial intelligence. He is experienced in building high-performance machine-learning software systems for efficient training, inference, and performance modeling of language, recommendation, and vision models on devices such as GPU and AI chips.



Ning Sun is a senior performance engineer with Meta. She has more than 20 years of experience in developing tools to analyze performance and detect performance anomalies in very large distributed applications. At Meta, she is working on scaling AI training stack and HHVM server on multiple generations of hardware. Prior to joining Meta, she was a lead contributor for SPECweb and SPECjAppServer while working in the Performance team with Sun/Oracle.



Pallab Bhattacharya is a distinguished engineer with Nvidia. Prior to joining Nvidia, he helped design and deploy AI Training Cluster at Meta, moving applications from CPU-based training to GPU-based training. He has extensive experience in High-Performance Communications using Infiniband and RDMA, building, debugging, and doing performance analysis of large-scale distributed and multi-thread applications.



Xizhou Feng (Member, IEEE) received the PhD degree in computer science from the University of South Carolina and the JD degree from Marquette Law School. He is currently a software engineer with Meta Platform, Inc. His research interests include model stability and training efficiency of large AI models, scalable parallel algorithms, complex system modeling, high-performance computing, and the interactions between technology and law.



John D. Owens (Fellow, IEEE) received the BS degree in electrical engineering and computer sciences from the University of California, Berkeley, in 1995, and the PhD degree in electrical engineering from Stanford University, in 2003. He is the a child family professor of engineering and entrepreneurship with the Department of Electrical and Computer Engineering, University of California, Davis, where he leads a research group with a focus on GPU computing. He is an AAAS fellow and a distinguished member of the ACM.



Louis Feng is an engineering manager with Meta, where he spearheads a team focused on improving fleet-wide efficiency visibility. His role involves a wide spectrum of responsibilities, including telemetry, performance analysis, and modeling. His notable contributions to the open-source community include his work on PyTorch execution trace and PARAM benchmarks, which have been adopted by the ML-Common's Chakra working group. Before his tenure with Meta, he served as a staff performance engineer with Intel. There, he played key roles in the develop-

ment of DreamWorks MoonRay, Pixar's RenderMan, the Embree ray tracing engine, and the nGraph AI compiler framework. He received 12 patents and published papers at conferences like SIGGRAPH, NeurIPS, KDD, and ISCA. His contributions have sparked innovation across various industries.