Throughput Prediction of Asynchronous SGD in TensorFlow

Zhuojin Li*
University of Southern California
Los Angeles, California
zhuojinl@usc.edu

Wumo Yan*
University of Southern California
Los Angeles, California
wumoyan@usc.edu

Marco Paolieri
University of Southern California
Los Angeles, California
paolieri@usc.edu

Leana Golubchik
University of Southern California
Los Angeles, California
leana@usc.edu

ABSTRACT

Modern machine learning frameworks can train neural networks using multiple nodes in parallel, each computing parameter updates with stochastic gradient descent (SGD) and sharing them asynchronously through a central parameter server. Due to communication overhead and bottlenecks, the total throughput of SGD updates in a cluster scales sublinearly, saturating as the number of nodes increases. In this paper, we present a solution to predicting training throughput from profiling traces collected from a single-node configuration. Our approach is able to model the interaction of multiple nodes and the scheduling of concurrent transmissions between the parameter server and each node. By accounting for the dependencies between received parts and pending computations, we predict overlaps between computation and communication and generate synthetic execution traces for configurations with multiple nodes. We validate our approach on TensorFlow training jobs for popular image classification neural networks, on AWS and on our in-house cluster, using nodes equipped with GPUs or only with CPUs. We also investigate the effects of data transmission policies used in TensorFlow and the accuracy of our approach when combined with optimizations of the transmission schedule.

KEYWORDS

Distributed Machine Learning; TensorFlow; SGD; Performance

ACM Reference Format:
Zhuojin Li, Wumo Yan, Marco Paolieri, and Leana Golubchik. 2020. Throughput Prediction of Asynchronous SGD in TensorFlow. In *Proceedings of the 2020 ACM/SPEC International Conference on Performance Engineering (ICPE '20), April 20–24, 2020, Edmonton, AB, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3358960.3379141

1 INTRODUCTION

Deep learning [8] has achieved breakthrough results in several application domains, including computer vision, speech recognition, natural language processing. In contrast with traditional machine learning, Deep Neural Networks (DNNs) discover internal representations suitable for classification from training data, without the need for manual feature engineering. This approach requires very large amounts of training data and computation: for example, popular DNNs for image classification include millions of model parameters (*DNN weights*) trained using datasets of millions of labeled images. Training examples are grouped in small batches and used for optimization steps with Stochastic Gradient Descent (SGD), which is computationally expensive (gradients are computed by propagating output errors back to each model parameter, through a sequence of matrix multiplications [8]).

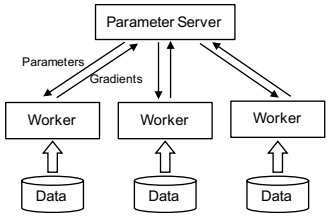


Figure 1: Parameter Server Architecture

Training performance can be improved by using more powerful hardware, such as GPUs and FPGAs. To improve performance even further, machine learning frameworks such as TensorFlow [1] can use multiple *worker nodes*, each performing SGD steps on a shard of the training data. A popular architecture to share model updates between worker nodes is the *parameter server* [9], illustrated in Fig. 1. The parameter server holds a global version of model parameters (the DNN weights): each worker receives these parameters (*downlink phase*), computes an update from a batch of labeled examples (*computation phase*), and transmits its update to the parameter server (*uplink phase*), where it is applied to the global model (*update phase*). In *asynchronous SGD* (the focus of our work), workers proceed independently; in contrast, in synchronous SGD the parameter server waits for updates from all the workers before sending an updated model, introducing blocking at the workers.

As the number of worker nodes increases, network traffic at the parameter server also increases, resulting in sublinear scaling of *training throughput* (examples/s processed by all the workers). For

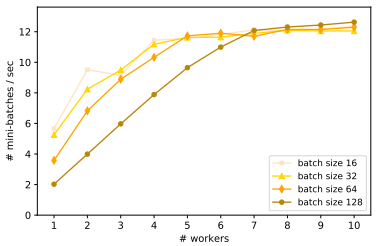
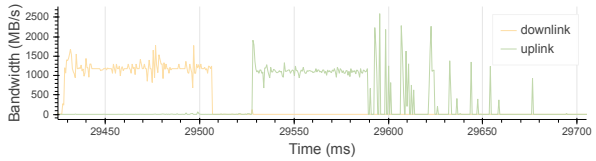


Figure 2: Training throughput of Inception-v3 on AWS `p3.2xlarge` GPU instances for different batch sizes

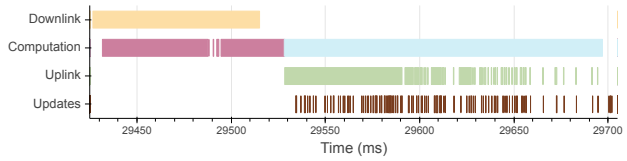
*Authors with equal contribution.

ICPE '20, April 20–24, 2020, Edmonton, AB, Canada

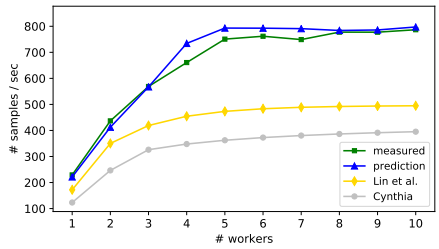
ACM ISBN 978-1-4503-6991-6/20/04…$15.00
https://doi.org/10.1145/3358960.3379141



(a) Downlink/uplink transmission bandwidth (from `tcpdump`)



(b) Summary of profiling data captured in TensorFlow



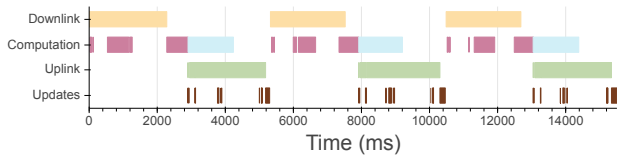
(c) Throughput prediction with different methods

Figure 3: Analysis of an SGD step (batch of 64 examples) of Inception-v3 training using 1 worker and 1 parameter server in TensorFlow (on AWS `p3.2xlarge`) and prediction results

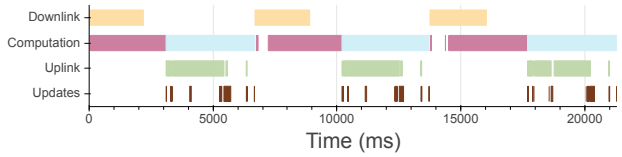
example, Fig. 2 illustrates the training throughput measured for the Inception-v3 model [17] when training on AWS `p3.2xlarge` instances (each equipped with NVIDIA V100 GPU) with TensorFlow and asynchronous SGD, for batch sizes of 16, 32, 64, 128. Throughput saturates at 4 workers for batch sizes 16 and 32; adding more workers yields only marginal improvements. In contrast, for batch sizes 64 and 128 throughput saturates at 5 and 7 workers, respectively: in this case, workers access the network less frequently (it takes longer to compute a model update), reducing network load.

The goal of our work is to provide an approach to predicting training throughput of asynchronous SGD for any number of workers W, from quick job profiling performed in TensorFlow using a single worker node. This would allow users to avoid testing multiple configurations and manually checking throughput; cost savings with respect to manual benchmarks are particularly important in large cloud environments with GPU nodes, where users submit multiple jobs and schedulers need to decide how many nodes to assign to each job based on its size and ability to scale.

Existing approaches for performance prediction of asynchronous SGD are based on very coarse models of computation and communication, not accounting for dependencies and overlaps of fine-grained operations. For example, previous work [10] infers the duration of each SGD phase from profiling information collected using network analysis tools such as `tcpdump` [19]: the time interval between the end of the downlink transmission and the start of the uplink transmission is interpreted as the single worker's computation, and the durations of these phases are used as parameters



(a) Batch size = 4



(b) Batch size = 16

Figure 4: Summary of TensorFlow profiling data for 3 training steps (1 worker/1 server) on a private CPU-only cluster

of a queueing model to predict throughput. An even coarser model, proposed in [25], estimates throughput with W workers and batch size K from the network utilization U_1 measured for a single worker as $WK/(T_P \max(1, WU_1) + 2T_C)$, where T_P is the time required to process a batch and T_C is the model/updates transmission time.

Overview We argue that these models overlook the complexity of computation and communication in asynchronous SGD. As illustrated in Fig. 3(a) for Inception-v3, the uplink phase (green) is spread out over a long time interval. By looking at the trace information collected with TensorFlow (Fig. 3(b)), we observe that computation overlaps with both uplink and downlink communication: the first part of the computation (forward propagation, in red, computing the output error) overlaps with the downlink, while the second part (backward propagation, in cyan, propagating the error back to model parameters) overlaps with the uplink. As illustrated in Fig. 4, these overlaps are not limited to a specific DNN model, batch size, or platform. In fact, TensorFlow starts each operation in an SGD step as soon as its dependencies are satisfied: forward propagation at the worker can start as soon as the initial layers are received; similarly, as soon as backward propagation completes for one of the final layers, its uplink transmission can start.

Our proposed approach collects fine-grained profiling information using TensorFlow traces, recording dependencies of each operation in a training step. For example, for the simple 4-layer DNN model of Fig. 5, we collect 1-worker profiling such as in Fig. 6 (but real-world DNNs include thousands of operations): each DNN layer results in multiple transmissions (uplink/downlink) and computations (forward/backward propagation at the worker, updates at the server). Specifically, we collect 1-worker profiling steps (e.g., 100 steps) and then sample from them with replacement to generate synthetic traces for multiple workers (Fig. 7). To obtain accurate throughput estimates from synthetic traces, we need to account for network sharing between workers and also for limitations of recorded traces, which track only the start/end of operations but not their active transmission intervals. Fig. 3(c) compares our prediction results for Inception-v3 with existing methods, clearly indicating the need for a more fine-grained approach (like ours) rather than the coarse-grained approach in the current literature.

The *contributions* of our work are as follows.

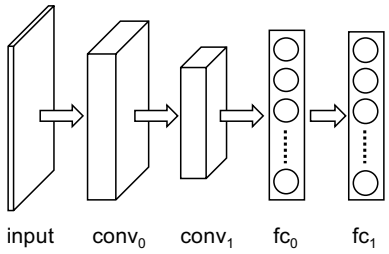


Figure 5: A simple DNN model

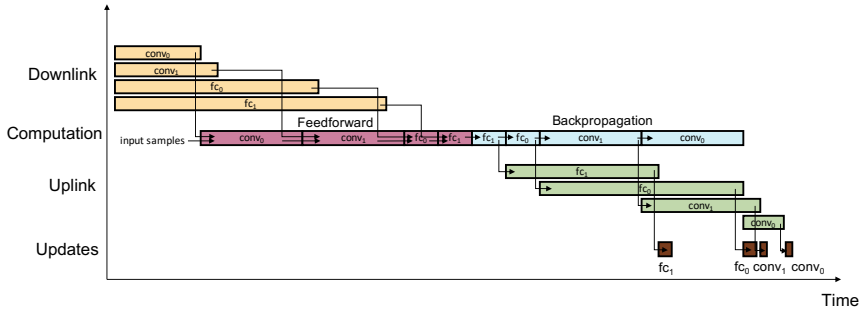


Figure 6: Training steps profile: four phases of a training step

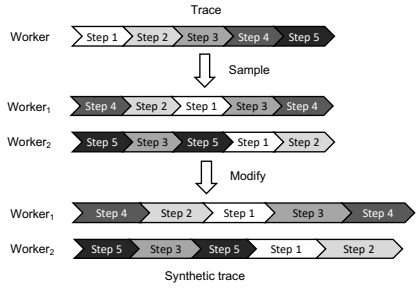


Figure 7: Synthetic trace generated for 2 workers sampling and modifying steps from 1-worker profiling

- We propose an approach for throughput prediction of asynchronous SGD based on fine-grained tracing information collected from 1-worker profiling. We account for the type and dependencies of each operation in a discrete-event simulation with multiple workers, which allows us to predict delays caused by network congestion with great accuracy. To enable this approach, we overcome many limitations of TensorFlow trace profiling data, discussed in Section 2. In particular, we provide a model for communication overhead due to message parsing (Section 3.2.1) and for HTTP/2 multiplexing of multiple streams (Section 3.2.2) in TensorFlow.
- We include these models in a simulation algorithm (Section 3.4) for throughput prediction and we validate our approach using a large set of experiments for multiple DNN models, with many batch sizes, on private clusters and public cloud platforms, using CPU or GPU resources for each node, for different network speeds. The results highlight that our approach can accurately predict throughput and bottleneck points (Section 4).
- We investigate the performance effects of state-of-the-art optimizations on the communication strategy of TensorFlow (Section 3.3) and show that our approach can be modified to account for such changes, accurately predicting throughput in these settings (Section 4).
- We investigate a model of bandwidth sharing in configurations with 2 parameter servers and evaluate the accuracy of our approach (Section 5); extension to a larger number of parameter servers is part of our ongoing efforts.

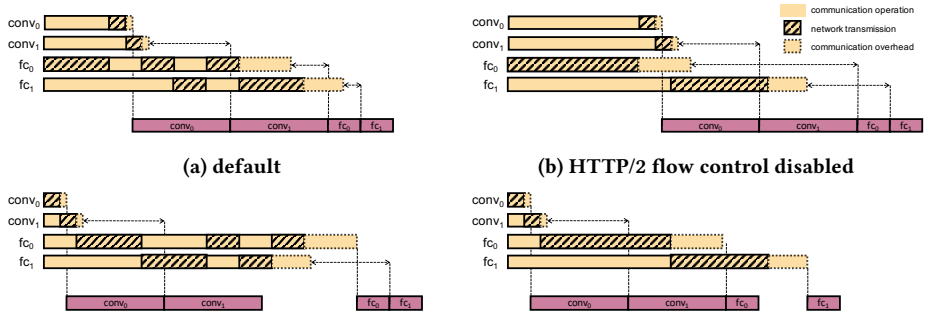
2 PROFILING

Dependencies among operations of DNN training (adjusting weights to fit a dataset of input/output pairs) or inference (computing output classifications for new inputs) are represented in TensorFlow as a *computation graph*: different operations (e.g., matrix multiplications and activation functions in a layer) are nodes of this graph, while tensors (multidimensional arrays) flow along directed edges between nodes (i.e., the output of one operation is the input of the next). Execution is triggered by feeding data (e.g., an input image) into the input nodes of the graph.

Fig. 6 illustrates the computation graph of a step of *distributed* training for the simple 4-layer DNN model of Fig. 5. There are two types of operations: (1) *computation operations* (e.g., addition, matrix multiplication, convolution) producing an output tensor from one or more input tensors (in red, cyan, and brown); and (2) *communication operations*, transferring data between nodes (in orange and green). During *forward propagation*, a worker can compute the output of each layer in order, after receiving its weights from the parameter server; during *backward propagation*, updates to each layer are computed in reverse order and immediately scheduled for transmission to the parameter server. Note that dependencies are much more complex than shown in Fig. 6: the computation of a single layer breaks down into many basic operations that are scheduled for execution on the CPU and GPU.

We generate profiling information for a training job (a specific DNN processing batches of training examples of a given size) by running distributed TensorFlow for a few SGD steps using one parameter server and one worker. Each step can be described as a set of operations with mutual dependencies: in the following, for each operation *op* we denote by *op.waiting_for* and *op.dependent_ops* the set of operations that *op* depends on, and the set of operations that can start only after the completion of *op*, respectively. Note that the dependencies are the same for all training steps since training steps are generated from the same DNN model.

Each operation *op* in the collected profiling traces uses exactly one resource, $op.res \in \{\text{DOWNLINK}, \text{WORKER}, \text{UPLINK}, \text{PS}\}$ where: DOWNLINK models the transmission channel of the parameter server (used by the workers to receive up-to-date model parameters), WORKER models the computation unit at the worker (CPU cores or GPU used for SGD), UPLINK models the receiving channel of the parameter server (used by the workers to transmit model updates), and PS models the computation unit at the parameter server (used



(a) default

(b) HTTP/2 flow control disabled

(c) communication ordering enforcement

(d) communication ordering enforcement, HTTP/2 flow control disabled

Figure 8: Different communication mechanisms

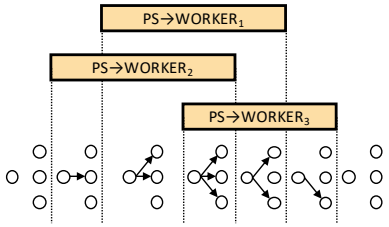


Figure 9: Active connections and bandwidth allocation change over time as transmissions start/end

to apply model updates). For communication operations, the size *op.size* of the transmitted tensor is also recorded.

Unfortunately, communication operations recorded by TensorFlow profiling tools do not accurately represent the exact timings for the transmission of the corresponding parameters. In fact, TensorFlow uses gRPC [3], a framework for remote procedure calls (RPC) over HTTP/2, to transfer tensors and manage connections. Tensor transfers are triggered at the beginning of each training step: first, TensorFlow finds all tensors that need to be transferred to another device; then, it starts a communication operation for each such tensor through gRPC.

Fig. 8(a) shows that, when training the DNN model from Fig. 5, each layer triggers a communication operation. For each transfer, TensorFlow creates a corresponding RPC request: the recorded start time of the communication operation corresponds to when the tensor is ready on the sender side, while the recorded end time tracks the time when the data is available to the receiver. In fact, these start/end times do not correspond to the underlying network transmissions: (1) Transmission can start after the recorded start time, since gRPC API calls are asynchronous and start times only indicate when transmissions are *requested* by the sender; for example, Fig. 8(a) shows that, at the beginning of each step, all tensors are available to be transmitted from parameter server to the workers, so that the profiler records the beginning of their transmission at the same time, although only one starts transmitting data. (2) The duration of each recorded transmission does not necessarily represent transmission time; not only can transmissions start well after their recorded start time, but they can be performed in parallel or suspended, since each gRPC transfer is assigned to a different HTTP/2 stream subject to multiplexing. (3) In addition, the recorded end time is also increased due to the latency introduced by parsing operations performed after the data has been transferred to the receiver (including deserialization and memory copies).

In order to perform accurate predictions, we need to infer the real start/end times of network transmissions (i.e., which communication operation is being served at each time) based on the limited information provided by TensorFlow profiling. To account for (1) and (2), we propose a model of HTTP/2 multiplexing in gRPC. To account for (3), we propose a linear model of communication overhead. We present and evaluate these models in Section 3.2.

3 PREDICTION

From profiling information collected in a single-worker configuration, we extract detailed information on the communication and computation operations of each SGD step. In this section, we use profiling information to construct synthetic traces for multiple SGD steps in a configuration with an arbitrary number of workers W. To do so, we perform a discrete-event simulation of the operations at each worker, accounting for the reduction in bandwidth due to the presence of multiple workers transmitting or receiving data. In turn, extended communication times at a worker can delay dependent operations in an SGD step. First, we address bandwidth sharing between multiple workers; then, we analyze the effects of HTTP/2 multiplexing of concurrent transmissions at each worker.

3.1 Bandwidth Sharing among Workers

During our profiling phase, only a single worker communicates with the parameter server: in this case, the uplink/downlink operations of the worker can use the entire network bandwidth in each direction. In contrast, in a distributed SGD configuration with multiple workers networking resources are shared.

In order to adapt networking of single-worker profiling traces for multiple-workers prediction, we need to track the number of workers *currently active* in the uplink/downlink direction. In fact, as illustrated by Fig. 3(b), communication with the parameter server is intermittent: the worker sends updates for each layer/tensor of the model as soon as they are ready. With multiple workers, many network states (active/inactive links) are possible (Fig. 9). We keep track of the number of workers n active in each direction (uplink/downlink) and assume that each active worker receives a fraction $\frac{1}{n}$ of the available bandwidth. Our model assumes that network capacity is shared equally, without significant background traffic and with similar round-trip times (RTTs). Although in practice workers may split network bandwidth unevenly because of background traffic and heterogeneous RTTs, we find this model to be accurate for throughput prediction (Section 4).

3.2 Analysis of Single-Worker Profiling Traces

As noted in Section 2, communication operations recorded in TensorFlow during the profiling phase do not accurately represent the

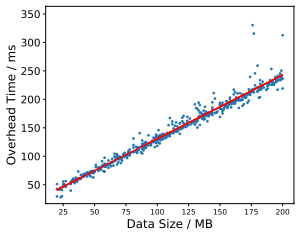


Figure 10: Overhead for each transmitted tensor size

actual transmission times because HTTP/2 flow control alternates the transmission of different streams. This is a major obstacle to predicting communication times in configurations with multiple workers, since extending recorded communication operations according to the available bandwidth yields inaccurate results. We illustrate these challenges and our proposed solution below.

3.2.1 Parsing Overhead of Received Tensors. We benchmark the parsing overhead of communication operations for data transfers of different sizes. The results, illustrated in Fig. 10, suggest a linear model $op.overhead = \alpha \times op.size + \beta$ with respect to the size of the data transferred by the operation *op*. The parameters α and β are independent of the specific DNN model, and they can be estimated once for the nodes used in the cluster. We remove this parsing overhead from the duration of a communication operation and assign it to a dependent computation operation.

3.2.2 Downlink and Uplink Multiplexing. HTTP/2 achieves better performance than HTTP/1.1 (especially for web browsers) due to the introduction of multiplexing, so that multiple streams (e.g., images of a web page) can be transmitted simultaneously within a single connection between client and server without "head-of-line blocking" due to large files being requested before smaller ones.

Stream multiplexing is the mechanism used in HTTP/2 for flow control. The receiver side of every stream advertises a flow control window *WIN*, which is a credit-based value that specifies the amount of data it is prepared to receive, in order to prevent the stream from overwhelming the receiver and blocking other streams. HTTP/2 defines only the format and semantics of the flow control window, while implementations are free to decide how it should adapt over time to current network and memory conditions (usually based on the bandwidth-delay product and memory pressure), and how to switch between multiple streams.

In gRPC, the remote procedure call (RPC) library used by TensorFlow for distributed SGD training, there are two connections (one in each direction) between each parameter server and worker. TensorFlow creates a gRPC request, which initiates an HTTP/2 stream for each tensor that needs to be transferred to a different node. To illustrate the stream multiplexing behavior of gRPC, we perform an experiment that transmits concurrent HTTP/2 streams in TensorFlow: we perform a training step of the model AlexNet [7] with one parameter server and one worker, capture the packets of downlink transmission using `tcpdump` [19] and analyze HTTP/2 frames using Wireshark [22].

The results, presented in Fig. 11, illustrate that (in a single worker scenario) HTTP/2 switches between gRPC streams intermittently. We observe that streams smaller than the flow control window *WIN* finish without switching, while for streams larger than *WIN*, HTTP/2 transmits *WIN* bytes and then switches to another stream. Furthermore, stream preemption happens only once for each stream: when a stream is selected again for transmission, it will transmit to completion, even if its remaining size is larger than *WIN*.

We adopt the following model for the multiplexing mechanism of HTTP/2 in gRPC. We define a *scheduler* for each link (i.e., for each sender/receiver pair), multiplexing streams of multiple transmissions. Each stream (which carries data of a communication operation) is assigned to the scheduler as soon as its corresponding operation starts. While the scheduler is not empty, a *chunk* (e.g., the initial portion of a stream) is selected from one of the active streams for transmission. The first time that a stream is selected, a chunk of size up to *WIN* is selected by our scheduler modeling HTTP/2 multiplexing. If the remaining size of the stream is less than the current *WIN*, or if the stream is selected for the second time, the entire stream is consumed as a chunk scheduled for transmission. Once transmission of the stream completes, another stream is selected by the scheduler in our synthetic trace generation. An example of this model is illustrated in Fig. 12.

This model allows us to predict HTTP/2 stream multiplexing, which is crucial for our trace generation algorithm because of the dependencies between different operations in an SGD step. We extract the flow control window *WIN* (on average, 28 MB) from HTTP/2 headers captured by `tcpdump`. When multiple streams are multiplexed within an HTTP/2 connection between a worker and the parameter server, given the start time of each stream and *WIN*, the model infers which stream is transmitted at each time, eventually predicting its end time. We validate the estimation accuracy of our model on various platforms, by comparing the end times of downlink streams resulting from our model with those measured during single-worker profiling. Results are presented in Table 1 with error statistics obtained from 100 training steps; most HTTP/2 multiplex transmissions are modeled correctly and this model works for different DNNs.

We observe three sources of modeling error: (1) Our model is based on the assumption that *WIN* does not change over time; in fact, *WIN* fluctuates as network conditions change. If a stream is slightly larger than *WIN*, its remaining portion may be transmitted at a much later time; thus, errors in the estimation of *WIN* can greatly affect end time predictions. (2) The estimation of parsing overhead can be inaccurate because our simple linear model does not account for fluctuations of CPU usage on the worker. (3) Another source of error is network instability: transmission times may be affected by background traffic, especially on cloud platforms.

3.2.3 Worker Computation and Model Updates. We assume that, for each additional worker, computation operations run on independent resources not shared with other workers. In particular, processing of a batch of training examples uses local resources at the worker (CPU cores or a GPU); similarly, model updates run at the parameter server independently (on separate cores).

3.3 TensorFlow Networking Optimizations

We observe that HTTP/2 stream multiplexing may introduce delays in training steps, reducing DNN training throughput.

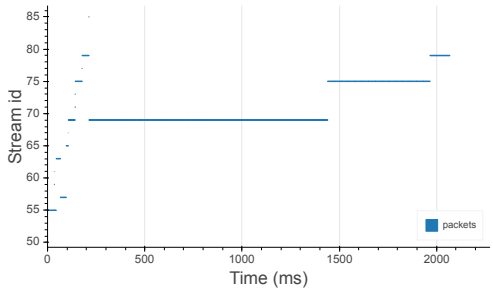


Figure 11: Multiplexing behavior of HTTP/2 streams during the downlink phase for AlexNet using a single worker

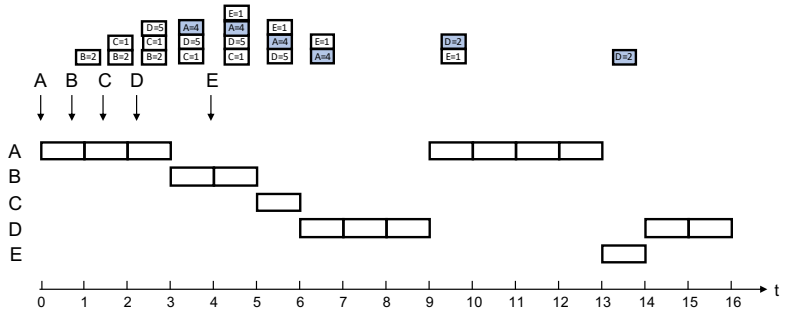


Figure 12: Multiplexing model of streams A, B, C, D, E with $WIN = 3$. Streams waiting to be transmitted (and remaining sizes) are represented on top; each is allowed to transmit up to *WIN* during the first service; streams waiting for their second service (to completion) are marked in blue.

Table 1: Error of endtime prediction of downlink streams

DNN Model		CPU Cluster	AWS Cloud
AlexNet	Average	1.82%	2.89%
	Median	1.18%	0.76%
	95th Percentile	3.35%	9.71%
	Maximum	47.48%	45.44%
GoogLeNet	Average	1.69%	3.43%
	Median	1.07%	2.76%
	95th Percentile	3.74%	9.14%
	Maximum	75.83%	64.24%
Inception-v3	Average	1.02%	9.23%
	Median	0.35%	8.01%
	95th Percentile	3.92%	20.98%
	Maximum	87.88%	98.19%
ResNet-50	Average	1.26%	4.36%
	Median	0.93%	3.78%
	95th Percentile	2.32%	9.70%
	Maximum	72.07%	97.10%

One feasible approach to improving training throughput is to maximize the overlap between computation and communication. Ideally, once a computation operation is completed (e.g., forward propagation of the first DNN layer), the next computation should start immediately, without being delayed by network transfers of required inputs. For simple DNN models in which layers are connected sequentially (i.e., without skip connections or branching [16]), layers should be transmitted in order during the downlink phase to reduce blocking of computation operations; for example, the optimal transmission order of layers in the model of Fig. 5 is $\text{CONV}_0 \rightarrow \text{CONV}_1 \rightarrow \text{FC}_0 \rightarrow \text{FC}_1$ (shown in Fig. 8(d)). For more complex DNN models, the research prototype TicTac [4] was proposed as a heuristic to derive efficient schedules for parameter transfers by analyzing the *critical path* of the computation; these schedules achieve performance improvements by enforcing communication ordering in TensorFlow.

However, this type of optimization cannot be fully implemented because of the multiplexing features of HTTP/2 and of the inevitable switching between pending communication operations. Fig. 8(c) shows an example where HTTP/2 can suspend transmission of the current layer, causing the computation to block.

We find that HTTP/2 stream switching can be eliminated from distributed TensorFlow training by disabling HTTP/2 flow control in gRPC, as in Fig. 8(b); in this case, there is no multiplexing of downlink and uplink transmissions, as also illustrated in Fig. 8(d), where the time required for a training step is reduced with respect to Fig. 8(c). In Section 4.2, we evaluate the accuracy of our predictions when flow control is disabled, under multiple scheduling policies (including TicTac). To perform predictions in this setting, we modify the schedulers used in our synthetic trace generation to process entire streams as a single chunk (i.e., without interruptions), in the order in which they are scheduled.

3.4 Trace Generation for Multiple Workers

We generate a synthetic trace for each system configuration (network bandwidth B, workers W, parameter servers M) through discrete-event simulation.

A sequence of N SGD steps is sampled with replacement for each worker from the set of steps S collected during job profiling (which is performed only once, with a 1-server/1-worker configuration). As illustrated in Algorithm 3.1, for each worker $w \in W$ and resource $r \in \{\text{DOWNLINK}, \text{WORKER}, \text{UPLINK}, \text{PS}\}$, a separate scheduler $scheduler[w, r]$ keeps a queue of pending operations. Operations are split into smaller chunks by the scheduler; when a chunk of worker w is completed with resource r, another chunk is selected by the $scheduler[w, r]$ and added to Q (Line 31). This approach allows us to represent the scheduling policies observed in gRPC when HTTP/2 multiplexing of multiple streams is enabled (as in Fig. 11): first, a chunk of each stream is selected by the scheduler; then the remaining data is transmitted until completion.

When the last chunk of an operation is completed (Line 18), dependent operations may become available for execution (Line 22) and are added to schedulers for their required resources (Line 24). If worker w is not using the required resource, the first chunk of the operation can be processed (Line 29). At any time, $active[\text{DOWNLINK}]$ and $active[\text{UPLINK}]$ track the number of workers using the downlink and uplink resources, respectively: in our bandwidth sharing model (summarized by $\text{SHARE}(r, active)$ in Algorithm 3.1), each

Algorithm 3.1: Simulation for synthetic trace generation. Each operation *op* from the profiling traces uses a resource *op.res*, has some prerequisite operations *op.waiting_for*, and operations *op.dependent_ops* depend on it.

```
StartRandomStep(S, Q, w)
1  step = SampleWithReplacement(S)
2  for op in step.copy()    // each step starts with downlinks
3      if op.res == downlink
4          scheduler[w, downlink].add(op)
5  // scheduler splits worker ops, runs ≤ 1 chunk/worker/resource
6  Q.add( scheduler[w, downlink].remove_chunk() )

Share(r, active)              // fraction of r assigned to each worker
1  if r in {downlink, uplink}
2      return 1/active[r]    // uplink and downlink shared equally
3  else return 1    // processing is independent for each worker

GenerateTrace(S, W)
 1  Q = ∅    // set of scheduled operation chunks
 2  for w in W    // setup for each worker
 3      completed_steps[w] = 0
 4      for r in {downlink, worker, uplink, ps}
 5          scheduler[w, r] = Scheduler(r)    // empty scheduler
 6      StartRandomStep(S, Q, w)  // add first chunk of downlink
 7  trace = Trace()    // empty trace
 8  active = {downlink : |W|, uplink : 0}
 9  while Q ≠ ∅
10      sort chunks x ∈ Q by x.remaining / Share(x.res, active)
11      chunk = Q.remove_min()
12      w, r = chunk.worker, chunk.res
13      duration = chunk.remaining / Share(x.res, active)
14      // chunk.main_op is the operation that the chunk is part of
15      trace.add(w, r, chunk.main_op, duration)
16      for chunk x in Q    // update remaining times
17          x.remaining -= duration × Share(x.res, active)
18      if chunk.is_last    // this is the last chunk of the operation
19          // dependent ops can be assigned to scheduler if ready
20          for d in chunk.main_op.dependent_ops
21              d.waiting_for.remove(chunk.main_op)
22              if d.waiting_for == ∅    // no other dependency
23                  if scheduler[w, d.res] != ∅  // w already using d.res
24                      scheduler[w, d.res].add(d)  // just queue d
25                  else    // start running the first chunk of d
26                      if d.res in {downlink, uplink}
27                          active[d.res] += 1 // w becomes active
28                      scheduler[w, d.res].add(d)
29                      Q.add( scheduler[w, d.res].remove_chunk() )
30      if scheduler[w, r] != ∅    // w has more chunks to run on r
31          Q.add( scheduler[w, r].remove_chunk() )
32      else    // no more chunks for w to run on resource r
33          if r in {downlink, uplink}
34              active[r] -= 1    // become inactive
35          if scheduler[w, i] = ∅ ∀i    // no more pending chunks
36              completed_steps[w] += 1    // step is over
37              if completed_steps[w] < n
38                  StartRandomStep(S, Q, w)
39  return trace
```

worker receives a fraction $1/active[r]$ of networking resource $r \in \{\text{downlink}, \text{uplink}\}$ to transmit consecutive chunks of tensors (in some order defined by the scheduler). In contrast, computations (forward/backward propagation at the worker and model update at the server) run on resources (worker CPU/GPU and parameter server cores) reserved exclusively for each worker (i.e., $\text{Share}(\text{worker}, active) = \text{Share}(\text{ps}, active) = 1$). Execution times of running chunks are extended according to the fraction of resource available to the worker (Lines 10 and 13). When no more chunks are pending for the operations of a step (Line 35), a new step is sampled and scheduled on the worker (Line 38).

The profiled SGD steps S used by the simulation are pre-processed to remove overhead from recorded transmission times and adjusted for the network bandwidth B available in the cluster. In particular, each communication operation is transformed into a new communication operation (with initial duration determined by B) and a computation operation (the overhead, which depends on the amount of transmitted data). Overlaps between communication and computation are modeled by the simulation algorithm using different resources for parameter server uplink/downlink and for computation on each node. The simulation algorithm can be extended to M parameter servers by introducing distinct resources downlink_i, uplink_i, ps_i for each parameters server $i = 1, \ldots, M$ and using the model of Section 5 to share downlink_i and uplink_i among the workers of the configuration.

We note that the queue Q is sorted at every iteration only for ease of presentation at Line 10 of GenerateTrace; our implementation uses priority queues. Simulation time is proportional to the number of steps N simulated for each worker; multiple runs can be performed in parallel on separate cores.

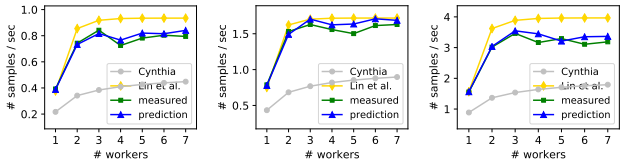
4 RESULTS

4.1 Experimental Setup

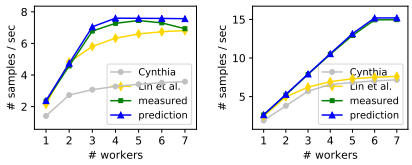
All the experiments are performed using TensorFlow 1.13 and the official TensorFlow benchmarks[1] from the v1.13 branch, with slight modifications (less than 5 lines of code) to turn on trace recording (used to acquire profiling information). The prediction algorithm is validated on the following platforms: (1) *private CPU cluster*: 8 nodes equipped with quad-core AMD Opteron Processor 2376 and 16 GB of memory, and connected by Gigabit Ethernet; (2) *cloud CPU cluster*: AWS `c4.8xlarge` instances (36 vCPUs, 2.9 GHz, Intel Xeon E5-2666v3, 60 GB memory) connected by 10 Gbps networking; (3) *cloud GPU cluster*: AWS `p3.2xlarge` instances (8 vCPUs, 2.7 GHz, Intel Xeon E5-2686v4, 61 GB memory, 1 NVIDIA Tesla V100 GPU with 32 GB memory), connected by 10 Gbps networking.

To perform throughput prediction, we collect the following information: (1) For each platform, we use `iperf` to measure network bandwidth and estimate the parameters α and β of our communication overhead model (Section 3.2.1), using the time difference between TCP packets captured by `tcpdump` and the end of communication operations recorded by TensorFlow. (2) We profile each training job (which specifies a DNN model and hyperparameters such as the batch size) for *100 steps with one parameter server and one worker* to obtain a trace of operations within an SGD step (and their dependencies).

[1] https://github.com/tensorflow/benchmarks

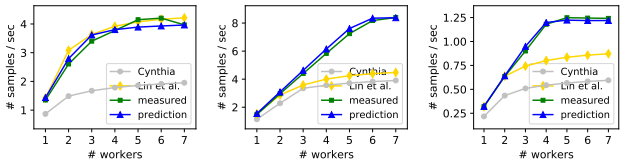


(a) AlexNet, batch size = 2 (b) AlexNet, batch size = 4 (c) AlexNet, batch size = 8

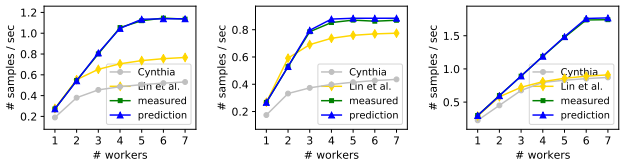


(d) AlexNet, batch size = 16 (e) AlexNet, batch size = 32

Figure 13: Results on private CPU cluster, varying batch size



(a) GoogLeNet, batch size = 1 (b) GoogLeNet, batch size = 2 (c) Inception-v3, batch size = 1



(d) ResNet-50, batch size = 1 (e) VGG-11, batch size = 4 (f) VGG-11, batch size = 8

Figure 14: Results on private CPU cluster, different models

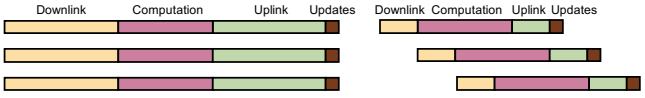
For each target configuration of W workers, we run our trace simulation procedure to generate a synthetic trace and use it to evaluate training throughput (total number of examples/s processed by the workers). In practice, we find that a trace of 1000 steps is sufficient to obtain a consistent estimate. Since it requires some time for asynchronous SGD workers to get out of the initial synchronization (training starts at the same time for all workers) and generate stable training throughput, we exclude the first 50 simulated steps and compute a time-average over the remaining steps. The predicted throughput is compared with the throughput measured in a real cluster with W workers and M parameter servers, as the time-average over the last 50 SGD steps out of 100 measured.

4.2 Private CPU Cluster

First, to illustrate the ability of our approach to accurately predict throughput with different batch sizes on our local CPU cluster, we consider a fixed DNN model (AlexNet [7]) and vary the batch size (batch sizes are small compared to GPU experiments of Section 4.3 because of the limited processing power of CPUs). Fig. 13 presents the results, showing that prediction error is within 10% for all batch sizes (the figure also includes quantitative comparisons to related work detailed in Section 4.4).

Next, we evaluate the accuracy of our throughput prediction method across different DNN models, including GoogLeNet [17], Inception-v3 [18], ResNet-50 [5], VGG-11 [15]. Fig. 14 shows that, in this case, prediction error is also within 10%.

By inspecting the measured traces, we find that all workers start the downlink phase of the first SGD step at the same time, as depicted in Fig. 15(a). As time advances, their training steps gradually start at different times due to small variations in computation and communication times. Ideally, workers continue to interleave and eventually stabilize when their downlink and uplink transmissions completely get out of synchronization, leading to higher throughput (Fig. 15(b)). Such communication pattern could be enforced through *network traffic control*, where the parameter server exchanges parameters with the workers in a strictly sequential order [6]. However, in TensorFlow different workers contend



(a) Completely overlap (b) Completely interleave

Figure 15: Workers start at the same time and gradually get out of synchronization; training steps are fastest when workers alternate using the network.

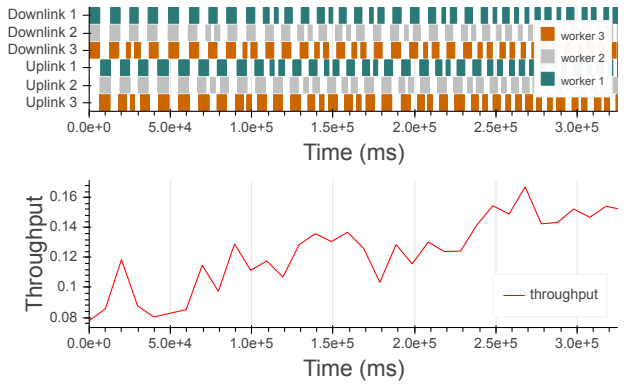
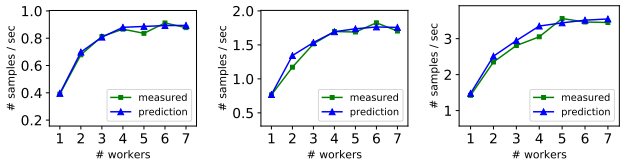


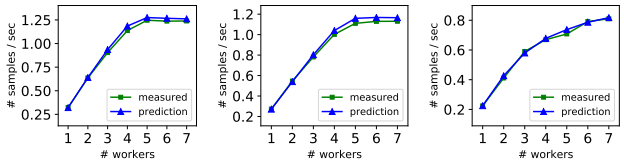
Figure 16: Distributed training of AlexNet with batch size of 8 and 3 workers on private CPU cluster. As downlinks/uplinks start interleaving, throughput increases.

for bandwidth without being regulated, potentially resulting only in partial interleaving of communication operations (Fig. 16), leading to a more challenging environment for predicting throughput, as explored also in Section 4.3.

Next, we explore the effects of HTTP/2 stream multiplexing. In Section 3.3, we motivated approaches to disable HTTP/2 flow control and enforce a specific communication ordering. To evaluate the prediction accuracy of our method in these scenarios, we disable flow control and repeat the experiments of Figs. 13(a) to 13(c) and Figs. 14(c) to 14(e); corresponding results, presented in Figs. 17(a)

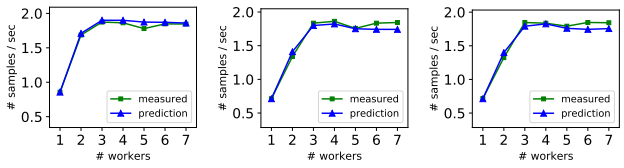


(a) AlexNet, batch size = 2 (b) AlexNet, batch size = 4 (c) AlexNet, batch size = 8



(d) Inception-v3, batch size = 1 (e) ResNet-50, batch size = 1 (f) VGG-11, batch size = 4

Figure 17: Prediction of different models on CPU cluster, with flow control disabled



(a) TIC Order (b) TIC Reverse Order (c) Random Order

Figure 18: Prediction of AlexNet, batch size = 4, with flow control disabled, enforcing different orders on CPU cluster

to 17(c) and Figs. 17(d) to 17(f), highlight good prediction accuracy, with errors of at most 10% in all cases but Fig. 17(b) for $W = 2$ workers, where the error is 20%; we observed that this was due to lower than expected interleaving in the measurements.

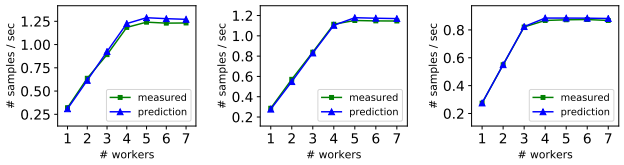
We also enforce different stream transmission orderings on the model in Figs. 13(b) and 17(b): the TIC ordering suggested by TicTac [4], the reverse of such order, and a random order. Results, presented in Fig. 18, illustrate predictions within 10% error. Finally, in Fig. 19 we explore prediction accuracy for the TIC ordering on the DNN models in Figs. 14(c) to 14(e); prediction error is less than 5% for these experiments.

In conclusion, our approach can accurately predict throughput for different stream selection orders, batch size, DNN model, and number of workers. This indicates that the prediction algorithm can adapt to different communication settings, and it has potential of accurately predicting further optimizations and modifications of the current TensorFlow implementation.

4.3 Public Cloud

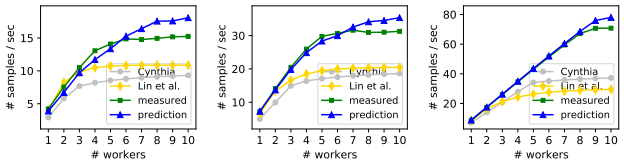
We evaluate our approach on the Amazon Web Services (AWS) cloud platform. This environment is less stable than our private CPU cluster, as networking performance may be affected by background traffic and by the deployment of virtual machines to racks with different latency.

In addition, communication overhead due to parsing of received data (Section 3.2.1) plays a much more important role, since networking is 10× faster (10 Gbps). For example, if overhead accounts

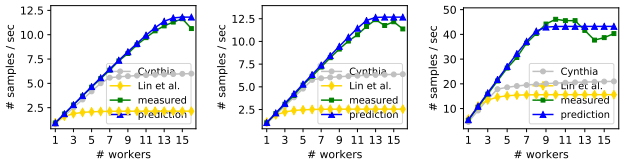


(a) Inception-v3, batch size = 1 (b) ResNet-50, batch size = 1 (c) VGG-11, batch size = 4

Figure 19: Prediction of different models with flow control disabled, enforcing TIC order on CPU cluster



(a) AlexNet, batch size = 4 (b) AlexNet, batch size = 8 (c) AlexNet, batch size = 16



(d) ResNet-50, batch size = 1 (e) Inception-v3, batch size = 1 (f) GoogLeNet, batch size = 1

Figure 20: Prediction of training on AWS cloud (CPU cluster)

for 10% of the duration of communication operations recorded in profiling traces on a 1 Gbps network, it will account for 52.6% of communication operations recorded on a 10 Gbps network.

4.3.1 CPU-Only Instances. As shown in Fig. 20, in most cases the error in throughput predictions on the AWS CPU cluster is within 20% for various DNN models and batch sizes, except for Fig. 20(f) with $W = 14$ workers, where the error is 22.8%. The prediction error on AWS CPU cluster is larger than that on our private CPU cluster, mainly because the network on the cloud is less predictable. In fact, intermittent background traffic can cause the HTTP/2 flow control window to change over time, leading to prediction errors.

4.3.2 GPU Instances. We also validate our approach on AWS GPU training. Fig. 21 illustrates that prediction error is within 20% across most DNN models and configurations. Prediction error is within 30% in Fig. 21(a) for $W = 2, 4$, Fig. 21(f) for $W = 4$, and Figs. 21(g), 21(j) and 21(k) for $W = 3$; and within 40% in Figs. 21(e), 21(l) and 21(m) for $W = 3$. We believe that larger errors occur in scenarios with few workers and smaller computation times (relative to communication), where it is more difficult to accurately predict the interleaving of data transfers between the parameter server and the workers; critical scenarios have 2 to 4 workers and smaller batch sizes (e.g., the error is higher in Fig. 21(a) than Fig. 21(b)).

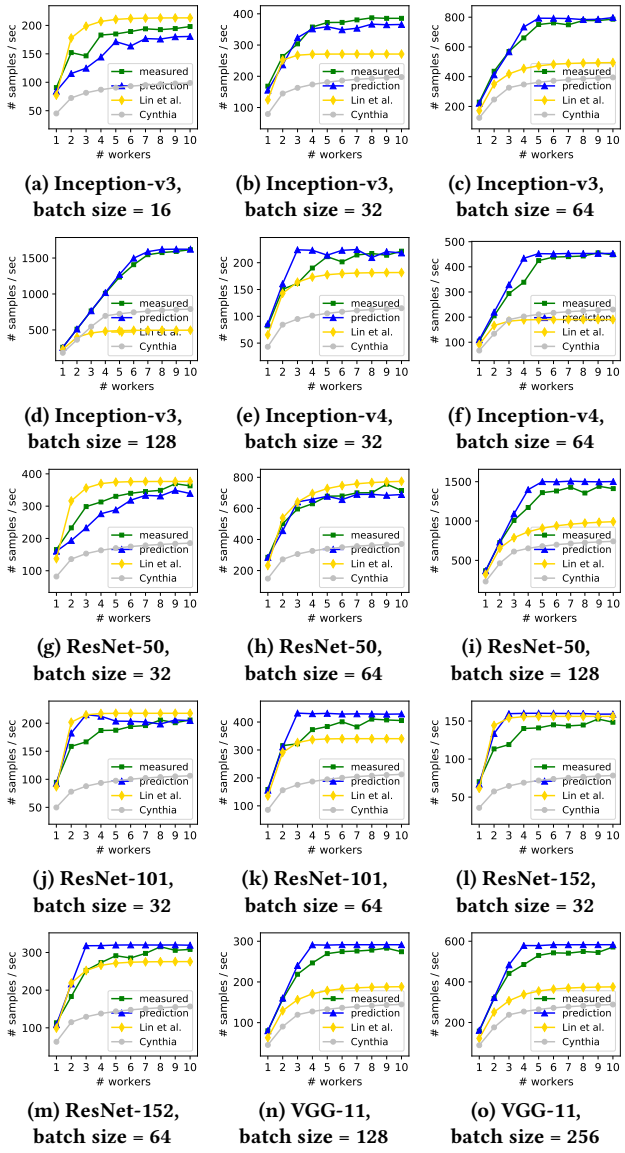


Figure 21: Prediction of training on AWS cloud (GPU cluster)

4.4 Prediction Accuracy Comparison

In Figs. 13, 14, 20 and 21, we compared prediction accuracy of our method with Lin et al. [10] and Cynthia [25]. We observe that the model of Lin et al. predicts throughput accurately in DNN models with small batch sizes, while predicted throughput saturates much earlier than real measurements for larger batch sizes, where the overlap between communication and computation is large. Throughput predicted by Cynthia is lower than measured. To investigate whether a simple change could improve predictions, we modified Cynthia's model by reducing communication times T_C (which affect utilization U_1 and throughput) in half, accounting for separate uplink/downlink network resources; in a small set of test cases, this modification seemed to improve Cynthia's prediction, although still with large errors in some scenarios.

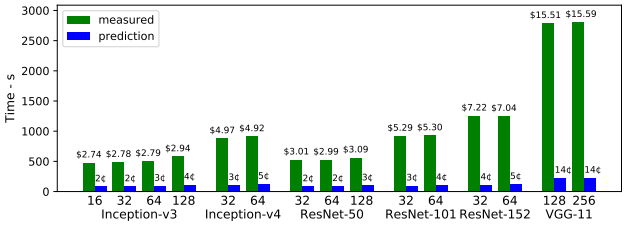


Figure 22: Cost and time comparison between throughput measurement and profiling/simulation-based prediction

4.5 Runtime Evaluation

Fig. 22 compares the execution time of 100 SGD steps on a cloud GPU cluster with that of our prediction algorithm. The algorithm execution time is evaluated using a single CPU core of an AWS `c4.1arge` instance (2 vCPUs, 2.9 GHz, Intel Xeon E5-2666v3, 3.75 GB memory). For example, directly measuring throughput of Inception-v3 with batch size 128 using 1 to 10 workers (100 steps for each setting) took 581 seconds. Our prediction method, including throughput measurement for 100 steps using 1 worker and prediction of 1000 steps for settings with 2 to 10 workers took only 117s. The prediction algorithm runs faster than actual training, and it could be further optimized by executing simulation runs in parallel over multiple cores or CPUs.

While the time required to obtain throughput measurements depends on the batch size and network (training with many workers can reach network bottlenecks), the execution time of the prediction algorithm depends on the number of operations in an SGD step and on the number of workers in the cluster. Predicting throughput (instead of direct measurements) allows not only considerable savings (each GPU instance is over 35× more expensive than the only CPU instance used for simulation) but also shorter evaluation times. Evaluation is particularly fast for DNNs with few operations (e.g., AlexNet, VGG), when network speed is slow (e.g., private CPU cluster with 1 Gbps Ethernet), when computation is slow (e.g., slow CPU or GPU), and when batch size is large.

5 MULTIPLE PARAMETER SERVERS

When a single parameter server becomes a bottleneck, more parameter servers can be added to the cluster with TensorFlow. In this case, model parameters are partitioned among parameter servers: for each part of the model, workers send updates and receive new parameters from a specific parameter server. We observe that the partition of model parameters among parameter servers is often uneven: since a DNN layer is the minimum unit of model parameters assigned in TensorFlow, parameters of entire DNN layers are assigned to parameter servers so as to balance the amount of data and networking load due to all workers. Since the size of different layers can vary greatly, this split can be uneven, as illustrated in Fig. 23 for VGG-11 when model parameters are partitioned among 2 parameter servers PS_1 and PS_2. Each layer is assigned to the parameter server that is currently holding parameters with smallest total size (in bytes); therefore, PS_1 receives model parameters of a larger size than PS_2 (407 MB instead of 100 MB): as a result, during training each worker will exchange more data with PS_1 than

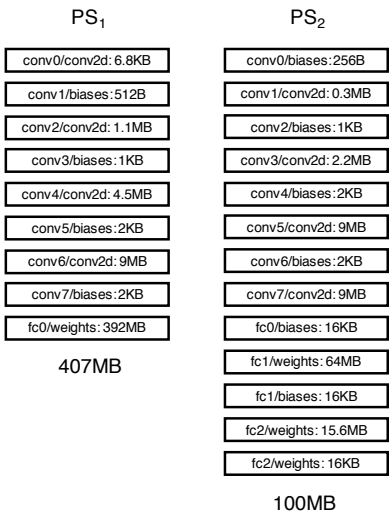
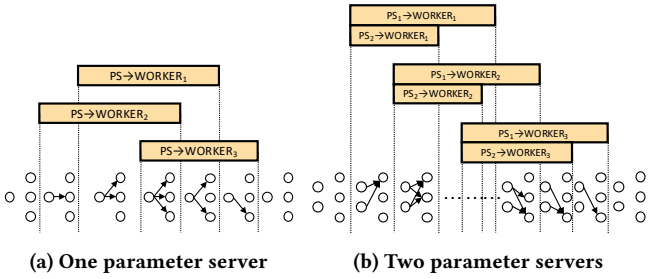


Figure 23: Partition of VGG-11 parameters among 2 PS

with PS_2. This asymmetry complicates network communication patterns in the multiple parameter server scenario. Furthermore, Fig. 24(b) shows that for 2 parameter servers, there are many more network states than for a single parameter server: with W workers and M parameter servers, there are WM distinct links that can be used to download model parameters; each can be active or inactive depending on whether the worker is currently downloading parameters from the specific server, resulting in 2^{WM} downlink states during simulation (and, similarly, 2^{WM} uplink states).

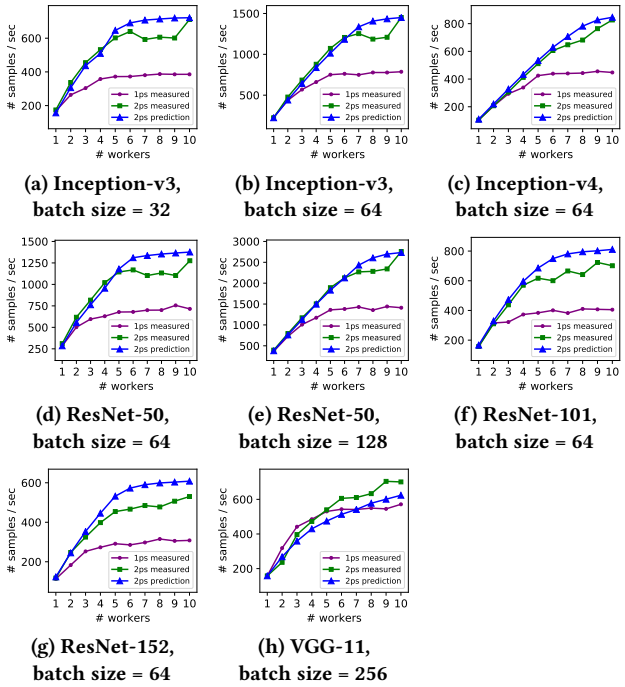
Based on our observations from running `iperf` benchmarks, we adopt a simple model for the case of 2 parameter servers: all active connections with the same parameter server equally share its bandwidth (for each uplink/downlink direction). In addition, we account for configurations of active links where some worker is the only worker exchanging data with PS_1 but has to contend with $n-1$ other workers to exchange data with PS_2. In this case, we assign bandwidth $\frac{1}{n}$ to the connections with PS_2 (equal sharing), but only up to $1-\frac{1}{n}$ to those with PS_1 (because the worker is already using $\frac{1}{n}$ of its transmission bandwidth for PS_2).

Using this model of bandwidth sharing, we modify the trace generation approach presented in Section 3.4: first, we collect profiling traces with 2 parameter servers and 1 worker using `p3.2xlarge` AWS instances (1 × NVIDIA V100 GPU); then, we run the simulation algorithm (Algorithm 3.1) for W workers and resources DOWNLINK_i, UPLINK_i, PS_i for $i = 1, 2$. We also run a real cluster with W workers with GPUs and compare throughput measured with 2 parameter servers (green curve) with our predictions (blue curve), and with throughput measured with 1 parameter server (purple curve). The results, presented in Fig. 25, illustrate that the error of throughput predictions on AWS with 2 parameter servers is within 20%, for most DNN models and batch sizes. Prediction error is within 25% in Fig. 25(d) for $W = 7, 9$, Fig. 25(f) for $W = 6, 8$, and Fig. 25(g) for $W = 6, 7, 8$. Measurements with 1 and 2 parameter servers illustrate that throughput is improved because of the additional parameter server. Note that, due to the uneven split of parameters in VGG-11, adding a second parameter server offers



(a) One parameter server **(b) Two parameter servers**

Figure 24: Possible states (active/inactive) of download links



(a) Inception-v3, batch size = 32 **(b) Inception-v3, batch size = 64** **(c) Inception-v4, batch size = 64**

(d) ResNet-50, batch size = 64 **(e) ResNet-50, batch size = 128** **(f) ResNet-101, batch size = 64**

(g) ResNet-152, batch size = 64 **(h) VGG-11, batch size = 256**

Figure 25: Prediction with 2 PS on AWS GPU cloud

only a marginal improvement in Fig. 25(h); since our method is based on real profiling traces collected in a configuration with 2 parameter servers and 1 worker, we can account for the uneven split and accurately predict throughput.

6 RELATED WORK

Performance models of machine learning and data processing jobs have been proposed in the literature with different levels of granularity for computation and communication operations. Ernest [21] and Optimus [12] use black-box analytical models of throughput and perform, for each job, test runs with different numbers of workers to estimate model parameters; in contrast, our work uses only one profiling run, with a single worker. Jockey [2] predicts performance of data processing jobs based on execution times and dependencies of each phase, but operations in DNN training jobs are more fine-grained than in data processing jobs.

At a lower level of granularity, works predating modern machine learning frameworks [24] estimate speedups of distributed

DNN training jobs through detailed analysis of elementary computations (such as matrix operations) which are profiled through micro-benchmarks; Paleo [13] models DNN computations layer-wise, while [14] and [11] analyze individual computations executed on a GPU. A common characteristic of these works is that throughput predictions are made as a function of computing speed (measured in FLOPS) and complexity of each layer (modeled as number of operations). However, additional factors including optimization strategies of the machine learning framework, operation scheduling, transmission overheads, can potentially affect training performance. Moreover, new optimization strategies are being implemented in machine learning frameworks; hence, such fine-grained models need to be adapted over time. In this work, we base our model on the profiling of individual operations in a computational graph, which is the intermediate representation in TensorFlow: any optimization in the framework is reflected and accounted for in the profiling information, so that our approach is not limited to specific hardware platforms or framework implementations.

While the majority of analytical approaches focus on synchronous SGD [20], modeling asynchronous SGD is more difficult because communication patterns between parameter servers and workers are more complex and can change over time. Previous work [10] builds a queueing model to estimate throughput of asynchronous SGD, using this model for scheduling of heterogeneous training jobs; coarse model parameters (the duration of downlink/uplink and worker/server computation) are estimated from single-worker scenario profiling. Cynthia [25] predicts training time through an analytical model based on network and CPU utilization, in order to provision cloud instances. These models are also at a high-level of granularity and model communication and computation as *sequential* phases, which is contradictory to our findings: overlaps between communication and computation are significant and play an important role in the optimization of training performance, as illustrated by Fig. 3(a). We compared our prediction results to those obtained with these methods in Section 4.4 (Figs. 13, 14, 20 and 21), highlighting major accuracy improvements.

7 CONCLUSIONS

We proposed an approach to predicting training throughput of asynchronous SGD in TensorFlow that extracts operation-level tracing information from minimal single-worker profiling data and performs discrete-event simulation to generate synthetic traces. Experimental results show good prediction accuracy across DNN models, batch sizes, and platforms as well as variants of TensorFlow, including optimizations of the training process. Our experiments also indicate that the more challenging cases are those with small number of workers and smaller computation (relative to communication) times. Although our simulation-based approach outperforms existing analytical modeling approaches, it is still of interest to develop fine-grained analytical models that address the shortcomings of previous work.

Future directions include heterogeneous hardware settings, multiple (> 2) parameter servers (with more complex bandwidth sharing models), and other (than gRPC) communication mechanisms, including TensorFlow with MPI and RDMA support.

ACKNOWLEDGMENTS

This work was supported in part by the NSF CNS-1816887 and CCF-1763747 awards.

REFERENCES

[1] M. Abadi et al. 2016. TensorFlow: A System for Large-Scale Machine Learning. In *OSDI'16*. 265–283.
[2] A. Ferguson, P. Bodík, S. Kandula, E. Boutin, and R. Fonseca. 2012. Jockey: guaranteed job latency in data parallel clusters. In *EuroSys'12*. 99–112.
[3] gRPC. 2019. Homepage. https://grpc.io.
[4] S. Hashemi, S. Jyothi, and R. Campbell. 2019. TicTac: Accelerating distributed deep learning with communication scheduling. In *SysML'19*.
[5] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR'16*. 770–778.
[6] X. Huang, A. Chen, and T. Ng. 2019. Green, Yellow, Yield: End-Host Traffic Scheduling for Distributed Deep Learning with TensorLights. In *IPDPSW'19*. 430–437.
[7] A. Krizhevsky, I. Sutskever, and G. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS'12*. 1106–1114.
[8] Y. LeCun, Y. Bengio, and G. Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.
[9] M. Li, D. Andersen, J. Park, A. Smola, A. Ahmed, V. Josifovski, J. Long, E. Shekita, and B. Su. 2014. Scaling Distributed Machine Learning with the Parameter Server. In *OSDI'14*. 583–598.
[10] S. Lin, M. Paolieri, C. Chou, and L. Golubchik. 2018. A Model-Based Approach to Streamlining Distributed Training for Asynchronous SGD. In *MASCOTS'18*. 306–318.
[11] Z. Pei, C. Li, X. Qin, X. Chen, and G. Wei. 2019. Iteration Time Prediction for CNN in Multi-GPU Platform: Modeling and Analysis. *IEEE Access* 7 (2019), 64788–64797.
[12] Y. Peng, Y. Bao, Y. Chen, C. Wu, and C. Guo. 2018. Optimus: an efficient dynamic resource scheduler for deep learning clusters. In *EuroSys'18*. 3:1–3:14.
[13] H. Qi, E. Sparks, and A. Talwalkar. 2017. Paleo: A Performance Model for Deep Neural Networks. In *ICLR'2017*.
[14] S. Shi, Q. Wang, and X. Chu. 2018. Performance Modeling and Evaluation of Distributed Deep Learning Frameworks on GPUs. In *DASC/PiCom/DataCom/CyberSciTech'18*. 949–957.
[15] K. Simonyan and A. Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR'15*.
[16] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. 2017. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *AAAI'17*. 4278–4284.
[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going deeper with convolutions. In *CVPR'15*. 1–9.
[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *CVPR'16*. 2818–2826.
[19] TCPDUMP. 2019. Homepage. http://www.tcpdump.org.
[20] A. Ulanov, A. Simanovsky, and M. Marwah. 2017. Modeling Scalability of Distributed Machine Learning. In *ICDE'17*. 1249–1254.
[21] S. Venkataraman, Z. Yang, M. Franklin, B. Recht, and I. Stoica. 2016. Ernest: Efficient Performance Prediction for Large-Scale Advanced Analytics. In *NSDI'16*. 363–378.
[22] Wireshark. 2019. Homepage. http://www.wireshark.org.
[23] K. Wongsuphasawat et al. 2018. Visualizing Dataflow Graphs of Deep Learning Models in TensorFlow. *IEEE Trans. Vis. Comput. Graph.* 24, 1 (2018), 1–12.
[24] F. Yan, O. Ruwase, Y. He, and T. Chilimbi. 2015. Performance Modeling and Scalability Optimization of Distributed Deep Learning Systems. In *SIGKDD'15*. 1355–1364.
[25] H. Zheng, F. Xu, L. Chen, Z. Zhou, and F. Liu. 2019. Cynthia: Cost-Efficient Cloud Resource Provisioning for Predictable Distributed Deep Neural Network Training. In *ICPP'19*. 86:1–86:11.