



PRIORITY-BASED PARAMETER PROPAGATION FOR DISTRIBUTED DNN TRAINING

Anand Jayarajan^{1,2} Jinliang Wei³ Garth Gibson^{2,3} Alexandra Fedorova¹ Gennady Pekhimenko⁴

ABSTRACT

Data parallel training is widely used for scaling distributed deep neural network (DNN) training. However, the performance benefits are often limited by the communication-heavy parameter synchronization step. In this paper, we take advantage of the domain specific knowledge of DNN training and overlap parameter synchronization with computation in order to improve the training performance. We make two key observations: (1) the optimal data representation granularity for the communication may differ from that used by the underlying DNN model implementation and (2) different parameters can afford different synchronization delays. Based on these observations, we propose a new synchronization mechanism called *Priority-based Parameter Propagation (P3)*. *P3* synchronizes parameters at a finer granularity and schedules data transmission in such a way that the training process incurs minimal communication delay. We show that *P3* can improve the training throughput of ResNet-50, Sockeye and VGG-19 by as much as 25%, 38% and 66% respectively on clusters with realistic network bandwidth.

1 INTRODUCTION

Training DNN is notoriously time consuming because of the high degree of computational complexity involved in tuning billions of parameters and processing large amount data (MLPerf, 2018). Data parallel distribution with synchronous stochastic gradient descent (SGD) is a popular method for accelerating the training by parallelizing the process over a cluster of machines (Chen et al., 2016).

In this paradigm, the input data set is sharded among the worker machines and they train a shared DNN model iteratively by independently computing the parameter updates and synchronizing them at the end of every iteration. A single iteration on a worker machine involves three main steps: (1) *forward propagation* step for calculating the value of a loss function on a subset of local data shard, (2) a subsequent *backward propagation* step for computing the gradients for every model parameter based on the computed loss, and (3) a *parameter synchronization* step for aggregating local gradients from all the worker machines and updating the parameters using the SGD algorithm (Bottou, 2010).

On every iteration, each worker machine generates and syn-

chronizes hundreds of megabytes of gradient values (Alan et al., 2018). This often makes data parallel training a communication-bound workload (Zhu et al., 2018). Handling such huge volume of traffic require high bandwidth networks like Gigabit Ethernet (Cunningham et al., 1999) or InfiniBand (Shanley, 2002). However, these technologies are yet to be adopted widely because of the high deployment cost. Moreover, this problem has been exacerbated with the emergence of faster hardware accelerators and larger DNNs as it leads to increase in data transmission rate and volume (Nvidia, 2018; Shazeer et al., 2017). Most of the major cloud providers and academic clusters are having trouble in catering to such high bandwidth demands (Luo et al., 2018). In this work, we propose solutions to scale data parallel training under limited bandwidth conditions.

Gradient compression is one of the popular approaches aimed at reducing the communication overhead. Since gradient values are generally represented as floating point numbers, it is extremely challenging to get reasonable compression ratios from lossless compression techniques (Burtscher & Ratanaworabhan, 2009). Instead, recent studies propose lossy compression techniques like gradient quantization (Seide et al., 2014; Alistarh et al., 2017; Wen et al., 2017) and sparse parameter synchronization (Aji & Heafield, 2017; Lin et al., 2018). These methods, however, risk affecting the final convergence accuracy of the model because of the information loss that comes with the value approximation and stale parameter updates (Khoram & Li, 2018).

An orthogonal approach is to utilize the available network

¹University of British Columbia, Vancouver, Canada ²Vector Institute, Toronto, Canada ³Carnegie Mellon University, Pittsburgh, USA ⁴University of Toronto, Toronto, Canada. Correspondence to: Anand Jayarajan <anand.indukala@gmail.com>.

bandwidth more efficiently by leveraging domain specific opportunities in DNN training. The traffic generated by the training processes are generally bursty because of its iterative nature. A common practice used in some distributed machine learning (ML) frameworks is to attenuate these traffic bursts by overlapping communication with computation. Since the training computation is performed as a sequence of operations (*layers*), during backward propagation, the gradients for the parameters associated with each layer are generated one after another. Frameworks such as TensorFlow (Abadi et al., 2016), MXNet (Chen et al., 2015) and Caffe2 (Jia et al., 2014), exploit this sequential layered structure to overlap parameter synchronization with backpropagation by issuing synchronization of each layer immediately after its gradients are computed.

In this work, we find new opportunities to better overlap communication and computation. Our first observation is that the domain specific knowledge of DNN training algorithm allows us to better schedule parameter synchronization not only based on when the gradients are generated, but also based on *when the data is consumed*. During training, the gradients of the layers are generated from final to initial layers and subsequently consumed in the reverse order in the next iteration. Figure 1 shows a snapshot of the training process containing the backward propagation of one iteration and the forward propagation of the next one. The temporal gap between gradients generated and consumed per layer are higher for final layers compared to the initial ones. Scheduling parameter synchronization using this information can help to overlap communication with both the forward and the backward propagation.

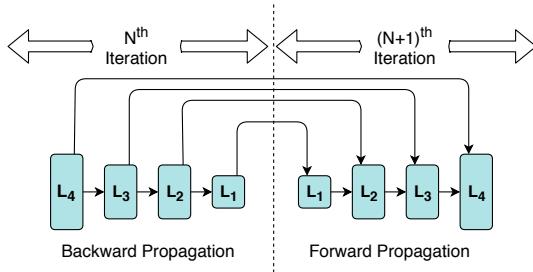


Figure 1. Training iterations

Secondly, the layer-wise granularity used by the underlying neural network implementation may not always be optimal for parameter synchronization. In our experiments, for certain heavy models (e.g., VGG (Simonyan & Zisserman, 2014), Sockeye (Hieber et al., 2017)), parameter synchronization at a finer granularity improves the network utilization and reduces the communication delay.

Based on these observations, we propose a new synchronization mechanism called *Priority-based Parameter Propagation* ($P3$).

1.1 Our Approach

$P3$ consists of two key components: (1) *Parameter Slicing*: $P3$ splits the layers into smaller slices and synchronize them independently. (2) *Priority-based Update*: $P3$ synchronizes the parameter slices based on their priority, where the priority of a slice is defined by when it is required again in the subsequent iteration. During backpropagation, $P3$ always allocates network cycles to the highest priority slices in the queue, preempting synchronization of the slices from a previous lower priority layer if necessary.

$P3$ offers following advantages over state-of-the-art parameter synchronization mechanisms (Zhang et al., 2017; Lin et al., 2018). (1) $P3$ can provide improved training performance under limited bandwidth conditions by better overlapping communication with computation and utilizing the available network bandwidth more efficiently. (2) $P3$ is model-agnostic, its implementation requires minimal programming effort, and all required changes are localized within the framework. (3) $P3$ always communicates full gradients and does not affect model convergence.

In summary, this paper makes the following contributions:

- We show that parameter synchronization at layer-wise granularity can cause suboptimal resource utilization in heavy models (e.g., VGG, Sockeye). We also show that the parameter synchronization can be scheduled better to efficiently use the available network bandwidth by taking into account not only the information on when the gradients are generated, but also when they are consumed.
- We present a new parameter synchronization mechanism called *Priority-based Parameter Propagation* ($P3$), which uses parameter slicing and priority-based updates to reduce communication overhead. We demonstrate that $P3$ has better resiliency towards bandwidth limitations compared to other state-of-the-art synchronization mechanisms (Zhang et al., 2017).
- We implement and open source $P3^1$ on MXNet (Chen et al., 2015), a popular distributed ML framework, and evaluate its performance against standard MXNet implementation as the main baseline. We observe that, $P3$ improves training performance of several state-of-the-art models like ResNet-50 (He et al., 2015), Sockeye (Hieber et al., 2017) and VGG-19 (Simonyan & Zisserman, 2014) by as much as 25%, 38% and 66% correspondingly.

2 BACKGROUND

The fundamental building blocks of DNNs are mathematical operations such as convolution, matrix multiplication,

¹<https://github.com/anandj91/p3>

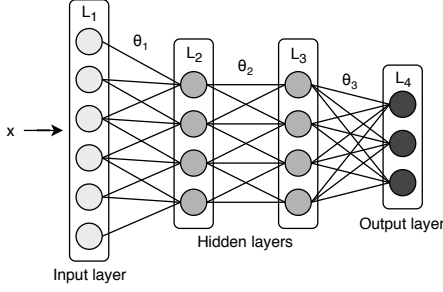


Figure 2. Deep neural network structure

and activation functions. These operations perform certain transformations ($f_{\theta}(x)$) on an input vector (x) using the parameters (θ) associated with it. A DNN is defined by a sequence of such operations (layers). In Figure 2, the initial layer takes the application specific data samples as input and produces a prediction as an output vector at the final layer. The goal of the training algorithm is to find the parameter values which can make the most accurate predictions.

DNN training usually starts from a random parameter initialization and iterates by randomly sampling input vectors from the training dataset. On every iteration, the DNN computes the output vector on the inputs and calculates the error associated with the prediction (loss) by feeding the output vector to a loss function. This step is called a *forward propagation*. After that, a *backward propagation* step is performed that calculates the error contribution of each parameter by computing gradients of all the layers with respect to the loss. The backward propagation method for calculating gradients is based on the chain rule of derivatives and is therefore performed in the reverse order of forward propagation, i.e., gradients of the final layer is calculated first and moves backwards to the initial layers, hence the name backward propagation (Rumelhart et al., 1986). Once the gradients are calculated, the parameters are updated using an optimization algorithm like Stochastic Gradient Descent (SGD) (Bottou, 2010). The iterations are repeated several times over the training dataset until the model converges to an acceptable prediction accuracy.

Each iteration is highly compute-intensive which makes the training process very time consuming. The total training time can be dramatically reduced by distributing the workload into multiple machines by taking advantage of the data parallel nature of the SGD algorithm. Data parallel training (Keuper & Preundt, 2016) involves multiple workers simultaneously training a shared DNN with the training dataset distributed equally among them. On each iteration, workers independently calculate the gradients locally for a common parameter value assignment but on different input data samples. Then the gradients are aggregated in a synchronous fashion for performing parameter updates. This method is

called a *synchronous SGD* algorithm (Chen et al., 2016).

Parameter server architecture (Li et al., 2014) is one of the most popular methods used in practice for parameter synchronization and it is widely supported in most of the distributed ML frameworks (e.g., MXNet (Chen et al., 2015), TensorFlow (Abadi et al., 2016), Caffe2 (Jia et al., 2014)). The parameter server keeps track of the up-to-date values of all the model parameters. Before every iteration, each worker machine reads the latest parameter values (θ) from the parameter server and locally computes gradients for the inputs sampled from its data shard. The workers then send the local gradients (∇) to the parameter server. The parameter server waits until it receives gradient updates from all worker machines, then aggregates the gradients together and updates the parameters for the next iteration.

Figure 3 shows parameter server-based data parallel training in a four-node cluster. The communication between worker machine and parameter server is usually over a network and often becomes the bottleneck in achieving linear scalability in data parallel training (Zhu et al., 2018; Shi & Chu, 2017).

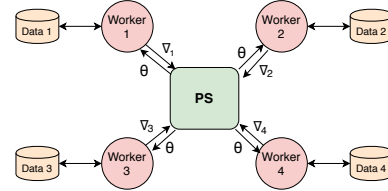


Figure 3. Parameter server architecture

State-of-the-art ML framework MXNet is designed specifically for making data parallel training efficient and easy to execute. It comes with a built-in implementation of parameter server called *KVStore*. In MXNet, worker machines send out gradients of a layer to the *KVStore* as soon as they are calculated, and issue parameter pull requests once all the other workers have finished sending the gradients of that layer. This aggressive parameter synchronization makes data parallel training very efficient on MXNet.

TensorFlow, on the other hand, is designed as a more generic ML framework. Hence it does not have an explicit parameter server implementation. However, a parameter server can be implemented on top of the graph computation framework provided by TensorFlow. Since the parameter server is a part of the computation graph, the communication between the worker subgraph and parameter server subgraph is handled by the framework itself. TensorFlow automatically places *Send* and *Receive* operations on the edges of the computation graph that crosses the device boundaries. Similar to MXNet, the worker subgraph executes the send operation as soon as the gradients are computed. However, since every training iteration is a separate graph execution, the parameter pull request is not issued until start of the

next iteration. This disconnection in sending gradients and receiving parameter updates could cause underutilization of bidirectional network bandwidth.

Despite small differences described above, we observe that state-of-the-art ML frameworks (e.g., MXNet, TensorFlow, Caffe2.) have two common characteristics. For performance reasons, the operations in the DNN implementation usually perform computations on large data representations. Therefore, the gradients for all the parameters within a layer are usually generated in a single shot. We observe that, because the gradients are generated at the layer level granularity, frameworks perform parameter synchronization at the same granularity as well. We also observe that since the DNN implementation is written as a dependency graph in these frameworks, the gradients of the layers are sent out to the parameter server over the network as soon as the backward propagation of that layer has completed. In this work, we address the limitations associated with these two characteristics of ML frameworks.

Apart from parameter server architecture, there are other mechanisms used for gradient aggregation. For example, there are many variations of MPI *all_reduce* operation specifically designed for ML workloads (Daily et al., 2018; Awan et al., 2017). In this work, we implement *P3* over the parameter server architecture in MXNet. However, we believe, *P3* design principles (namely, parameter slicing and priority-based propagation) are general enough to be applied to any gradient aggregation methods.

3 LIMITATIONS OF PARAMETER SYNCHRONIZATION

Current parameter synchronization mechanisms have major limitations in effectively utilizing available network bandwidth due to two main reasons. The first one comes with the aggressive synchronization performed by the frameworks where the gradients of the layers are sent to the parameter server immediately after finishing the backward propagation of that layer. Since the backward propagation progresses from the final to the initial layer, the gradients are also generated and propagated in that order. However, the next forward propagation can only be started after receiving the updated parameters of the *first* layer. We observe that, under limited bandwidth, gradient propagation of the final layers can induce queuing delay onto the gradient propagation of the initial layers and subsequently delay the next iteration. This prevents the communication from being overlapped with the forward propagation.

Figure 4(a) shows the parameter synchronization of a 3-layer DNN, where the forward and backward propagation of each layer takes one time unit and parameter synchronization takes two time units. With aggressive synchronization, the

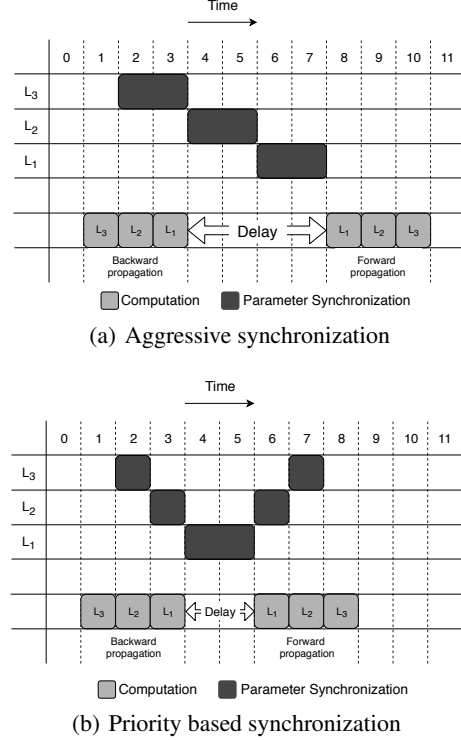


Figure 4. Parameter synchronization

total delay between the two iterations is twice the time taken for synchronizing the first layer because of the additional queuing delay induced by the previous layers. Moreover, during forward propagation the network stays totally idle.

This effect becomes even more noticeable when the communication time required for individual layers vary due to the presence of fully connected (FC) layers in the DNN, as the synchronization time needed for such dense layers is relatively higher. Figure 5 shows the parameter distribution of two popular image classification models: ResNet-50, VGG-19, and a machine translation model: Sockeye. The skewed parameter size distribution is a general trend in image classification models where the final FC layers are usually heavier and can potentially induce higher queuing delays on to the lighter initial convolution layers.

The second limitation is due to the parameter synchronization being performed at layer-level granularity. The communication time of parameter synchronization consists of three major parts: (1) *gradient propagation* time for the worker machine to send the gradients to the parameter server, (2) *parameter update* time taken by the parameter server to aggregate the gradients and update the parameters, and (3) *parameter propagation* time taken by the parameter server to send the updated parameters back to worker machine(s). As we describe in Section 2, current distributed ML frameworks overlap gradient propagation of one layer with the

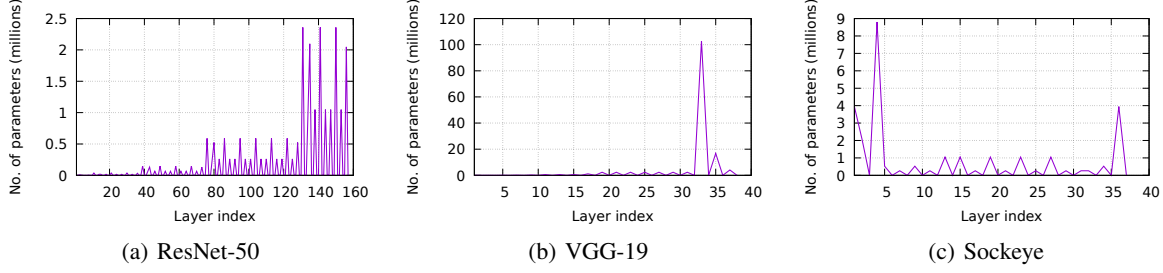


Figure 5. Parameter distribution

backward propagation of the next one. On top of this, at the parameter server side, the gradient propagation of a layer is overlapped with the parameter update of the previous layer. This type of communication-computation pipelining is effective only if the size of the layers are more or less uniform. Unfortunately, this is usually not the case. For example, Figure 5(b) shows that VGG-19 contains a single FC layer which has 71.5% of all the parameters in the entire network. We observe that the disproportionately heavy layers like this could severely affect the efficient utilization of network bidirectional bandwidth.

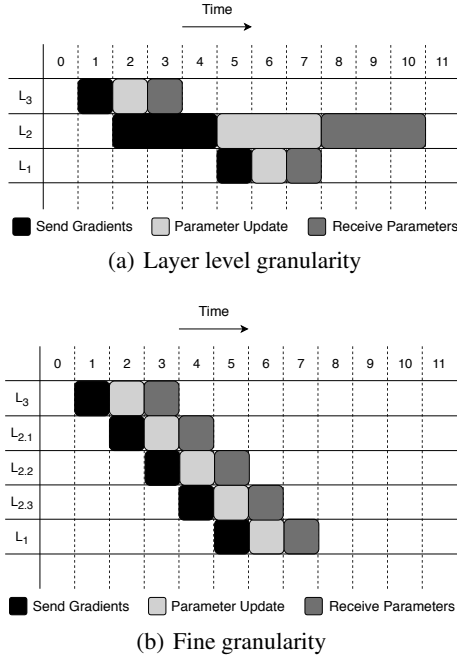


Figure 6. Coarse and fine granularity

This effect is explained in Figure 6(a) using the previous example of parameter synchronization of the 3-layered DNN. In this case, gradient propagation, parameter update, and parameter propagation of the second layer take thrice as much time as that of the first and third layers. Because of this imbalance, the communication delay in this model is mainly dominated by the second layer. The parameter

synchronization of the first and the third layer can only be partially overlapped with the second layer. As seen in the example, this severely underutilizes the computing resources and bidirectional bandwidth by spending the last three time steps just for receiving parameter updates from the parameter server.

From the above observations we draw two major conclusions. First, the application domain-specific knowledge of DNNs can be utilized to schedule communication not only based on the data generation in the backward propagation, but also based on when the data consumption in the subsequent forward propagation. Scheduling parameter synchronization based on this information and sending the gradients conservatively could reduce the delay by better overlapping communication with both the forward and the backward propagation. Second, the optimal granularity required for parameter synchronization may differ from the one used for data representation by the underlying model implementation. Synchronizing parameters at a finer granularity can better utilize the available computing and networking resources as we empirically show in Section 5.4.

4 P3: DESIGN AND IMPLEMENTATION

Based on the above observations, we propose a new method for parameter synchronization called Priority-based Parameter Propagation (*P3*). As explained in Section 1.1, *P3* has two core components: (1) *parameter slicing*, and (2) *priority-based update*.

Once the gradients are computed, gradient aggregation and updates of each parameter can be performed independent to each other. We take advantage of this property of the SGD algorithm for parameter slicing optimization. *P3* splits the layers into smaller slices of parameters and each of these slices are then independently synchronized. In Figure 6(b), applying parameter slicing optimization on the second layer achieves better overlap between data transmission and parameter update. Moreover, the bidirectional bandwidth is completely utilized as the synchronization of slices are perfectly pipelined. In this example, parameter slicing reduces the communication cost by 30%.

After splitting the layers into smaller pieces, *P3* assigns priorities to every slice. The slices inherit priority values from their parent layers. We determine layers' priorities based on the order in which they are processed in the forward propagation. The first layer gets the highest priority and it decrements moving towards to the end, with final layer getting the lowest priority. During backward propagation, parameter synchronization of the slices are issued based on their priorities as illustrated in Figure 4(b). In this example, with the priority-based update, the delay between the two iterations has been reduced by half and the communication is evenly overlapped with both the forward and the backward propagation.

We implemented *P3* by modifying the parameter server module in MXNet called KVStore. Below, we first explain how the baseline KVStore works, and then we describe our modifications to support *P3*.

4.1 KVStore: Baseline system

KVStore is a wrapper implemented on top of the lightweight parameter server called *ps-lite* (Li et al., 2014). KVStore has two components: *KVWorker*, which runs locally to the worker machine as part of the training process and a separate server process, called *KVServer*. *KVServer* receives gradients from *KVWorkers*, aggregates them, and updates the parameters while ensuring data consistency. The parameters are stored as key-value pairs, where the key is the index of a layer and the value is an array of floating points each corresponding to the parameter values of that layer. For load balancing purposes, more than one *KVServer* can be used for the training with the parameters equally sharded between them. For better resource utilization, a common practice is to run one *KVServer* on every machine along with the worker process.

Before starting the training process, KVStore initializes and distributes the parameters of all the layers among *KVServers*. KVStore follows a simple heuristic for fair distribution of parameters. Layers with size smaller than a fixed threshold are assigned to a randomly chosen *KVServer*. Parameters of larger layers are split equally among the *KVServers*. This is different from parameter slicing used in *P3* (explained in Section 4.2). The threshold is a configurable parameter and is set to 10^6 parameters by default.

KVServer exposes two interfaces to *KVWorker* for sending gradients and requesting updated parameters: a *Push* request and a *Pull* request. During training, MXNet issues a parameter synchronization request for a layer to the *KVServer* through the *KVWorker* as soon as the backpropagation of that layer has finished. *KVWorker* serializes (and fragments in case of large layers) the gradient matrix and issues a *Push* request to the corresponding *KVServer(s)*. *KVServer* waits until it has received gradient updates from all the workers

for that key-value pair. Once all the updates have been received, *KVServer* aggregates the gradients and updates the parameters.

Once the parameters are updated, *KVServer* notifies all the workers. When *KVWorker* receives a notification, it immediately issues a *Pull* request to the *KVServer(s)* for the corresponding updated parameter values. *KVServer* then sends the latest parameter values in response, and *KVWorker* (reconstructs the parameter array for large layers and) updates the local parameter values for the next iteration. MXNet overlaps the parameter synchronization with backward propagation by asynchronously issuing *Push* requests for the layers whose gradients are ready to be propagated.

4.2 *P3*: Implementation

In order to implement *P3*, we modify *KVWorker* and *KVServer* into *P3Worker* and *P3Server*. On the worker side, when a parameter synchronization is issued, *P3Worker* splits the gradient matrix of the layer based on a predefined size threshold (explained in more details in Section 5.7). Unlike in KVStore, this threshold defines the maximum granularity with which layers can be split. This is the parameter slicing part in *P3*. Each of these slices are assigned to *P3Servers* in a round-robin fashion.

The priority-based gradient propagation is implemented using a producer-consumer mechanism communicating through a priority queue. After parameter slicing, the producer part of *P3Worker* assigns priorities to the individual slices and pushes them into the priority queue all at once. A separate consumer thread in the *P3Worker* continuously polls the highest priority slice from the queue and sends the slice to the *P3Server* through the network with its priority added to the packet header. The consumer thread uses blocking network calls. Hence the rate at which the priority queue polled is automatically adjusted based on the networking delay. This simple producer-consumer model makes sure that the network does not experience bursty traffic flow from the *P3Worker*, and that the backward propagation is not hindered at the worker side. Also the slice with the highest priority in the queue always gets the first preference for transmission.

We also add a producer-consumer mechanism at the receiver in the *P3Server* in order to deal with the in-network delays. The packets received at the *P3Server* are pushed into a priority queue with the priority assigned by the *P3Worker* as the key. A server consumer thread then polls from this queue and processes the packets the same way as in a *KVServer*. Prioritization at the *P3Server* ensures highest priority parameters are processed first.

Apart from these modifications, we remove the explicit update notification and pull requests from the *KVServer*.

*P3*Server immediately broadcasts the updated parameters to all workers once it has received all of the updates. Since workers always issue a pull request after every push, this change does not affect the correctness of the training algorithm. This modification was necessary, because otherwise MXNet only issues a pull request once it has received the update notification for all the slices of a layer. Eliminating this helped to improve the bidirectional bandwidth utilization.

5 EVALUATION

5.1 Methodology

We have evaluated the *P3* implementation on three image classification models: ResNet-50 (He et al., 2015), InceptionV3 (Szegedy et al., 2015), VGG-19 (Simonyan & Zisserman, 2014) and on a machine translation model, Sockeye (Hieber et al., 2017). We chose the standard MXNet KVSStore implementation described in Section 4.1 as the baseline in all performance evaluation experiments.

Since *P3* implementation does not interfere with the model implementation or the training algorithm, the model convergence is not affected in any way. This means that the baseline and the *P3* would follow the same training curve for a given hyper-parameter set. The improvement in training performance is completely determined by the rate at which input data is processed. Therefore the primary performance comparison metric we use is the training throughput, which is the number of total training samples processed by the worker machines in one second. We measure the throughput after training the models for a few iterations until the throughput has become stable and then averaged over thousand iterations. In all the experiments, we set the number of KVServers/*P3*Servers equal to the number of worker machines.

5.2 Summary of the experiments

We conduct performance evaluation of *P3* in three different experiments. Section 5.3 shows how resilient *P3* is towards the bandwidth constraints in the network. We perform this experiment by training the model on a four machine cluster each equipped with one Nvidia P4000 GPU (Nvidia, 2017) and interconnected with a 100Gbps InfiniBand network (Mellanox, 2018). We measure throughput variation while artificially limiting the network interface transmission rate. This simulates more realistic networking infrastructure in modern cloud services where bandwidth is usually between 1Gbps and 25Gbps (ec2instances.info, 2019). Section 5.4 shows how well *P3* utilizes the available bandwidth and reduces the network idle time. Finally, in Section 5.5 we test the scalability of *P3* on different cluster sizes. This experiment is conducted on the AWS (Amazon, 2019) using g3.4xlarge machine instances on a 10Gbps network.

In Section 5.6, we compare the convergence accuracy for models trained using *P3* and compression based techniques. For this comparison study, we pick the state-of-the-art compression technique Deep Gradient Compression (DGC) (Lin et al., 2018). We implemented DGC on MXNet based on the details provided in the original paper and information collected from the authors. In addition to these experiments, we have also evaluated the effects of different parameter slice sizes on the training throughput in Section 5.7. We have also included additional results in the Appendix B based on our reviewers’ feedback.

5.3 Bandwidth v.s. throughput

In this experiment, we measure the training throughput of ResNet-50, InceptionV3, VGG-19, and Sockeye on a tightly controlled four-machine cluster by setting different transmission rates on the network interface on all the machines using Linux’s *tc qdisc* utility (Alexey N. Kuznetsov, 1999). Figure 7 compares the throughput from *P3* against the baseline system for different network bandwidths between 1Gbps and 30Gbps. We also measured the performance benefits achieved from the parameter slicing optimization alone.

In Figure 7(a) and 7(b), both baseline and *P3* give similar training performance when the network bandwidth is sufficiently high for scaling these models on four machines. However, the baseline throughput starts to drop in ResNet-50 below 6Gbps. At the same time, *P3* maintains the linear throughput until the bandwidth drops below 4Gbps. This is because *P3* reduces the peak bandwidth required for the model by efficiently overlapping communication with the computation. At 4Gbps, *P3* provides 26% more throughput than the baseline. For InceptionV3, the maximum speed up obtained is 18%. It is interesting to note that these models do not benefit from parameter slicing, as the layer sizes are relatively small in these DNNs (see Figure 5(a)).

Figure 7(c) and 7(d) show the throughput of VGG-19 and Sockeye respectively. These models contain one or more very large layers (Figure 5(b) and 5(c)), and because of the presence of these large layers, the parameter slicing optimization alone is giving considerable improvement in performance. At 30Gbps, parameter slicing can provide 49% speedup on VGG-19. The speedup is further improved with *P3* by as much as 66% at 15Gbps. Sockeye is a special case among other models. Unlike image classification models, the heaviest layer in this model is the initial layer. In Figure 7(d), Sockeye performance has improved by a maximum of 38% with *P3*.

We observe that *P3* always performs better than the baseline with relatively higher performance benefits when bandwidth is limited. Since *P3* reduces the peak network bandwidth required for the training, it is more suitable than baseline on a shared network cluster where effective bandwidth avail-

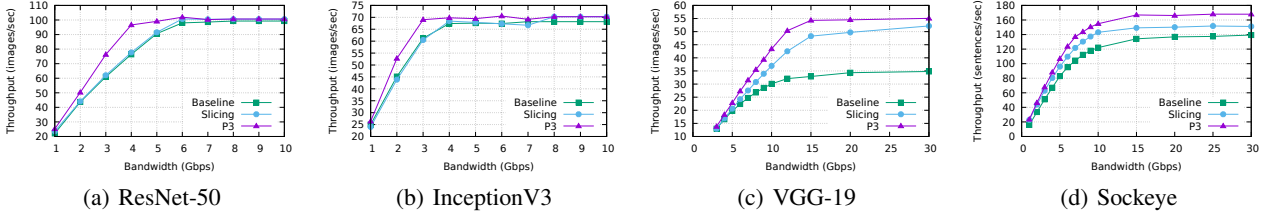


Figure 7. Bandwidth v.s. Throughput

able for a single training process is much lower than the maximum capacity of the network. However, the speed-up diminishes when the network bandwidth is too low. This is because the communication time is significantly higher than computation and there is little room for improvement by overlapping communication and computation.

5.4 Network utilization

This experiment compares the network utilization of *P3* with the baseline system. We conduct this experiment for ResNet-50, VGG-19, and Sockeye and measure the traffic generated from and received by one of the four worker machines. The network utilization is measured at the interface level using Linux’s *bwm-ng* tool (Volker Gropp, 2019) with 10 millisecond precision. Figure 8 shows the network utilization of the baseline system. The baseline implementation has bursty network traffic generated with regular peaks and crests across all models. This pattern is observed in other frameworks like TensorFlow (Abadi et al., 2016) and Poseidon (Zhang et al., 2017) as well (see Appendix B). For Sockeye and VGG-19, the network idle time is extremely dominant because of the presence of heavy layers. Moreover, the inbound and outbound traffics are not overlapped as the baseline fails to fully utilize bidirectional bandwidth.

In contrast, Figure 9 shows the network utilization graph with *P3*. We observe that *P3* significantly improves the network utilization compared to the baseline. In Figure 9(a) and 9(b), the network idle time is seen to be reduced with *P3*. Especially for Sockeye in Figure 9(c), *P3* utilizes bidirectional bandwidth more effectively than baseline system. This is one of the key reasons for the speedup observed for Sockeye despite having heavy initial layers.

5.5 Scalability

We perform scalability analysis on ResNet-50, VGG-19, and Sockeye in order to show how well *P3* can perform on large clusters compared to the baseline system. We conduct this experiment by distributing models on AWS g3.4xlarge clusters of different sizes (2, 4, 8 and 16 machines) over a 10Gbps network.

Figure 10(a) shows that on ResNet-50 both the baseline

and the *P3* achieve similar performance. As shown in Section 5.3, 10Gbps network is more than enough for linearly scaling ResNet-50. The throughput of VGG-19 has been considerably improved with *P3* by as much as 61% on an eight machine cluster (Figure 10(b)).

Figure 10(c) shows the scalability of Sockeye. LSTM-based models like Sockeye are very hard to scale over multiple machines, because of the heavy initial layers and difference in iteration time in worker machines due to the variable sequence length of input data. Nevertheless, with *P3*, we improve throughput of Sockeye by as much as 18% on an 8-node cluster.

5.6 Training accuracy

As we noted in Section 1, there are many compression techniques proposed for improving data parallel training performance. These methods can provide higher performance gains compared to *P3*, however, at the cost of reduction in the model quality. In this section, we compare convergence accuracy of *P3* with the state-of-the-art compression technique called Deep Gradient Compression (DGC) (Lin et al., 2018). DGC reduces the amount of data transferred by taking advantage of the sparsity in the gradient updates. The key idea is to synchronize only those parameters with top-k gradients and accumulate the rest locally. In this experiment, we use a sparsity threshold of 99.9% per layer based on the configurations used in the original paper (Lin et al., 2018).

We trained ResNet-110 on the CIFAR-10 dataset for 160 epochs over a four-machine cluster with both *P3* and DGC using five different hyper-parameter settings. Figure 11 shows the validation accuracy range of *P3* and DGC from these experiments. The dark bands represent the gap between the worst and best accuracy on the five hyper-parameter setting. We observe that the final accuracy obtained with *P3* is always better than DGC. We calculate an average accuracy drop of 0.4% with DGC.

Unlike compression based mechanisms (like DGC), *P3* always communicate the full gradients with other worker machines and does not make any modification in the original SGD algorithm. As a result, the performance benefits from *P3* comes without any penalty on model accuracy.

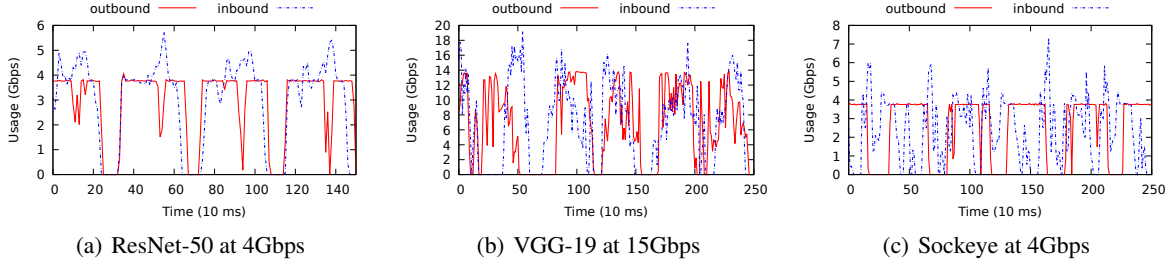


Figure 8. Network utilization of the baseline system

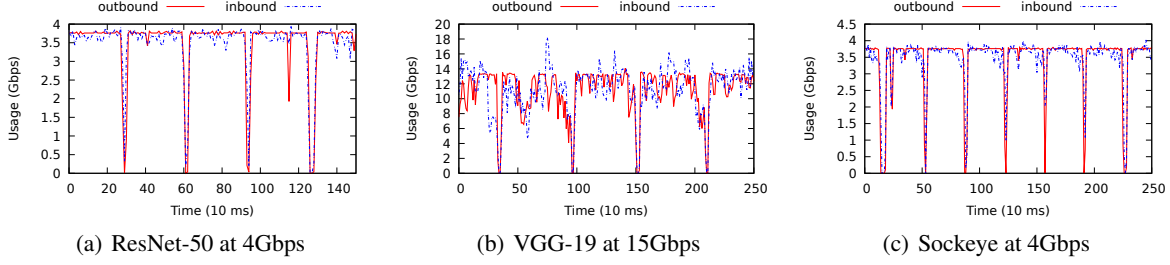


Figure 9. Network utilization of P3

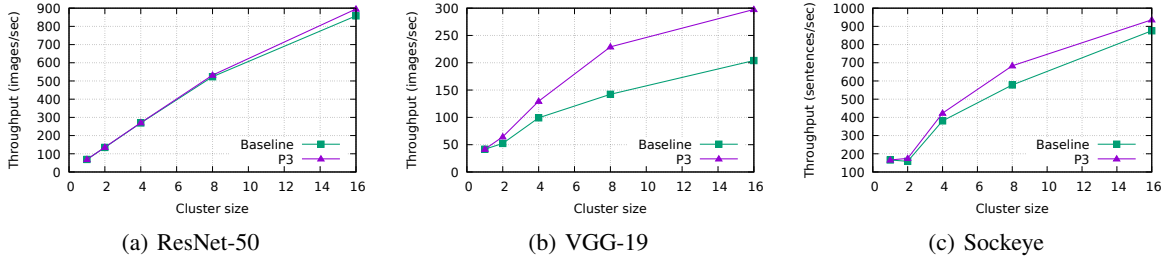


Figure 10. Throughput scaling with different number of machines

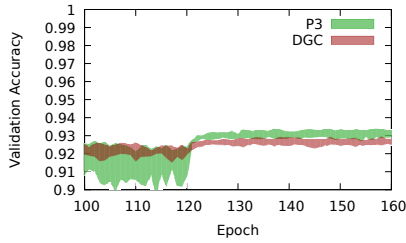


Figure 11. P3 v.s. DGC

5.7 Parameter slice size selection

As we show in Section 4, a small gradient packet size can improve the network utilization and, in turn, can improve overall training throughput. In this section, we show how the size of the parameter slice affects training performance. Figure 12 shows the throughput obtained for ResNet-50 and VGG-19 with P3 on different parameter slice sizes.

Initially, the throughput increases as size decreases, and reaches its peak at 50,000 parameters when it starts to drop. This happens because if the size is too small, the overhead of synchronizing packets at a very small granularity is getting too high, reducing the benefits of parameter slicing. In all our experiments, we use a maximum granularity of 50,000 parameters per slice as it is found to be optimal empirically.

6 RELATED WORK

In this paper, we describe the key limitations in the data parallel deep learning distribution techniques used in popular ML frameworks (e.g., TensorFlow and MXNet), and propose solutions to mitigate these limitations by taking advantage of domain specific characteristics of deep learning models. To the best of our knowledge, this is the first work to summarize and address these limitations.

One notable prior work which proposes domain specific optimizations for data parallel deep learning workloads is

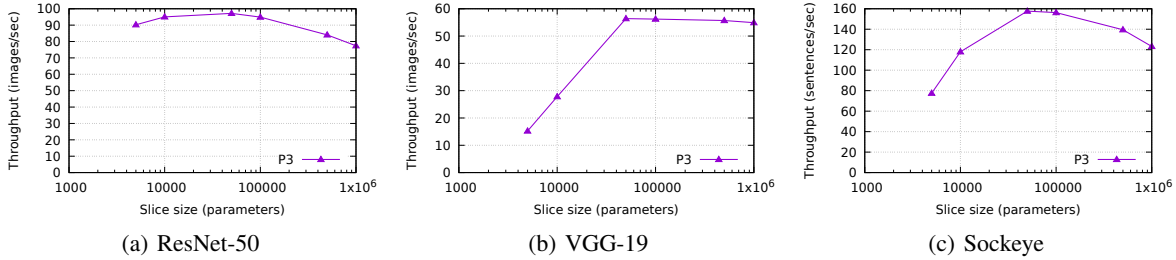


Figure 12. Granularity v.s. Throughput

Poseidon (Zhang et al., 2017). This work introduced the idea of wait-free-back-propagation (WFBP) which hides the communication overhead behind backpropagation by independently synchronizing individual layers in the neural network. We build upon this idea, and show that we can overlap computation with both forward and backward propagation. We further improve this idea by using parameter slicing that utilizes network bandwidth better.

Most of the recent papers in this area try to reduce communication overhead by sending fewer gradients. One popular method to reduce data transmission is gradient quantization (representing the gradient values using fewer bits). For example, 1-bit SGD (Seide et al., 2014) represents a 32-bit floating point gradient value in a single bit. Additionally, an error feedback is added to the SGD algorithm in order to account for the information loss that comes with the value approximation. 1-bit SGD can provide up to $10\times$ speedup for traditional speech recognition applications. Another work called QSGD (Alistarh et al., 2017) proposes a family of compression schemes which balance the trade-off between the accuracy and gradient precision to provide good performance. Similar to QSGD, TernGrad (Wen et al., 2017) uses 3-level numerical compression to reduce data transfer in data parallel training. Both QSGD and TernGrad provide mathematical guarantees on the bounds of final model convergence accuracy. In contrast, *P3* always sends the full 32-bit parameter values.

Another approach is the sparse parameter synchronization. The idea is to synchronize only a few parameters on every iteration instead of the whole model. Gradient dropping method only synchronizes parameters which have gradient values larger than a selected threshold. The threshold is calculated based on a fixed compression ratio (Aji & Heafield, 2017). AdaComp (Chen et al., 2017) automatically tunes the compression ratio depending on the local gradient activity and achieves up to $200\times$ compression.

All the above techniques make the trade-off between training performance and model accuracy because of the information loss introduced by value approximation or stale parameter updates (Khoram & Li, 2018). *P3* on the other hand, does

not introduce any information loss since it always sends full gradient matrix on every iteration.

On the other hand, a more recent work called Deep Gradient Compression(DGC) (Lin et al., 2018) offers up to $600\times$ compression and around $5\times$ speedup in low bandwidth networks while maintaining the same baseline accuracy on several DNN models. DGC use local gradient accumulation and momentum correction techniques to maintain the same accuracy. Even though the authors report no accuracy loss with DGC, there is no formal proof on the convergence guarantees cited in the paper. And as shown in Section 5.6, we find it difficult to reproduce their results despite substantial effort.² In our experiments, *P3* always gives better accuracy than the DGC. Based on these results, we conclude that our mechanism is a safer approach, as *P3* does not introduce information loss in the training algorithm and therefore there is no potential risk of accuracy loss. Moreover, our proposal is an orthogonal approach to the compression techniques and can be used on top of compression mechanisms to further improve performance.

7 CONCLUSION

In this paper, we analyze the data parallel distributed training methods used in current ML frameworks and observe that they fail to fully utilize available network bandwidth and induces high penalty on training performance under network bandwidth limitations. Based on this observation, we propose a new parameter synchronization method called *P3*, which improves the training performance by better utilizing the available network bandwidth. We implement *P3* on top of the state-of-the-art ML framework MXNet and demonstrate it to have higher resiliency towards bandwidth constraints and better scalability than the baseline MXNet implementation. With *P3*, we improve training throughput of ResNet-50 by as much as 25%, Sockeye by 38% and VGG-19 by 66%. We also have made the source code of *P3* publicly available.³

²This includes personal communication with the authors in order to get all their experiments correctly.

³<https://github.com/anandj91/p3>

REFERENCES

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pp. 265–283, Savannah, GA, 2016. USENIX Association. ISBN 978-1-931971-33-1. URL <https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi>.
- Aji, A. F. and Heafield, K. Sparse communication for distributed gradient descent. *CoRR*, abs/1704.05021, 2017. URL <http://arxiv.org/abs/1704.05021>.
- Alan, M., Panda, A., Bottini, D., Jian, L., Kumar, P., and Shenker, S. Network evolution for dnns. *SysML*, doc/182, 2018. URL <http://www.sysml.cc/doc/182.pdf>.
- Alexey N. Kuznetsov. Linux Traffic Control, 1999. URL <https://linux.die.net/man/8/tc>.
- Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. Qsgd: Communication-efficient sgd via gradient quantization and encoding. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 1709–1720. Curran Associates, Inc., 2017.
- Amazon. Amazon Web Services, 2019. URL <https://aws.amazon.com/>.
- Awan, A. A., Chu, C., Subramoni, H., and Panda, D. K. Optimized broadcast for deep learning workloads on dense-gpu infiniband clusters: MPI or nccl? *CoRR*, abs/1707.09414, 2017. URL <http://arxiv.org/abs/1707.09414>.
- Bottou, L. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*, 2010.
- Burtscher, M. and Ratanaworabhan, P. Fpc: A high-speed compressor for double-precision floating-point data. *IEEE Trans. Comput.*, 58(1):18–31, January 2009. ISSN 0018-9340. doi: 10.1109/TC.2008.131. URL <http://dx.doi.org/10.1109/TC.2008.131>.
- Chen, C., Choi, J., Brand, D., Agrawal, A., Zhang, W., and Gopalakrishnan, K. Adacomp : Adaptive residual gradient compression for data-parallel distributed training. *CoRR*, abs/1712.02679, 2017. URL <http://arxiv.org/abs/1712.02679>.
- Chen, J., Monga, R., Bengio, S., and Józefowicz, R. Revisiting distributed synchronous SGD. *CoRR*, abs/1604.00981, 2016. URL <http://arxiv.org/abs/1604.00981>.
- Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C., and Zhang, Z. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *CoRR*, abs/1512.01274, 2015. URL <http://arxiv.org/abs/1512.01274>.
- Cunningham, D., Lane, B., and Lane, W. *Gigabit Ethernet Networking*. Macmillan Publishing Co., Inc., Indianapolis, IN, USA, 1999. ISBN 1578700620.
- Daily, J., Vishnu, A., Siegel, C., Warfel, T., and Amatya, V. Gossipgrad: Scalable deep learning using gossip communication based asynchronous gradient descent. *CoRR*, abs/1803.05880, 2018. URL <http://arxiv.org/abs/1803.05880>.
- ec2instances.info. Easy Amazon EC2 instance comparison, 2019. URL <https://www.ec2instances.info/?region=us-west-2>.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL <http://arxiv.org/abs/1512.03385>.
- Hieber, F., Domhan, T., Denkowski, M., Vilar, D., Sokolov, A., Clifton, A., and Post, M. Sockeye: A toolkit for neural machine translation. *CoRR*, abs/1712.05690, 2017. URL <http://arxiv.org/abs/1712.05690>.
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22Nd ACM International Conference on Multimedia*, MM ’14, pp. 675–678, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-3063-3. doi: 10.1145/2647868.2654889. URL <http://doi.acm.org/10.1145/2647868.2654889>.
- Keuper, J. and Preundt, F.-J. Distributed training of deep neural networks: Theoretical and practical limits of parallel scalability. In *Proceedings of the Workshop on Machine Learning in High Performance Computing Environments*, MLHPC ’16, pp. 19–26, Piscataway, NJ, USA, 2016. IEEE Press. ISBN 978-1-5090-3882-4. doi: 10.1109/MLHPC.2016.6. URL <https://doi.org/10.1109/MLHPC.2016.6>.
- Khoram, S. and Li, J. DNN model compression under accuracy constraints, 2018. URL <https://openreview.net/forum?id=By0ANxbRW>.

- Li, M., Andersen, D. G., Park, J. W., Smola, A. J., Ahmed, A., Josifovski, V., Long, J., Shekita, E. J., and Su, B.-Y. Scaling distributed machine learning with the parameter server. In *Proceedings of the 11th USENIX Conference on Operating Systems Design and Implementation, OSDI'14*, pp. 583–598, Berkeley, CA, USA, 2014. USENIX Association. ISBN 978-1-931971-16-4. URL <http://dl.acm.org/citation.cfm?id=2685048.2685095>.
- Lin, Y., Han, S., Mao, H., Wang, Y., and Dally, B. Deep gradient compression: Reducing the communication bandwidth for distributed training. In *International Conference on Learning Representations*, 2018.
- Luo, L., Nelson, J., Ceze, L., Phanishayee, A., and Krishnamurthy, A. Parameter hub: a rack-scale parameter server for distributed deep neural network training. *CoRR*, abs/1805.07891, 2018. URL <http://arxiv.org/abs/1805.07891>.
- Mellanox. InfiniBand Cards, 2018. URL http://www.mellanox.com/page/infiniband_cards_overview.
- MLPerf. DNN training benchmarks, 2018. URL <https://mlperf.org/results/>.
- Nvidia. Nvidia Quadro P4000, 2017. URL <https://www.pny.com/nvidia-quadro-p4000>.
- Nvidia. cuda c programming guide, 2018. URL <https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html>.
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *Nature*, 323:533–, October 1986. URL <http://dx.doi.org/10.1038/323533a0>.
- Seide, F., Fu, H., Droppo, J., Li, G., and Yu, D. 1-bit stochastic gradient descent and application to data-parallel distributed training of speech dnns. In *Interspeech 2014*, September 2014.
- Shanley, T. *Infiniband*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2002. ISBN 0321117654.
- Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q. V., Hinton, G. E., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *CoRR*, abs/1701.06538, 2017. URL <http://arxiv.org/abs/1701.06538>.
- Shi, S. and Chu, X. Performance modeling and evaluation of distributed deep learning frameworks on gpus. *CoRR*, abs/1711.05979, 2017. URL <http://arxiv.org/abs/1711.05979>.
- Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. URL <http://arxiv.org/abs/1409.1556>.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015. URL <http://arxiv.org/abs/1512.00567>.
- Volker Gropp. Bandwidth Monitor NG (Next Generation), 2019. URL <https://linux.die.net/man/1/bwm-ng>.
- Wen, W., Xu, C., Yan, F., Wu, C., Wang, Y., Chen, Y., and Li, H. Terngrad: Ternary gradients to reduce communication in distributed deep learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 1509–1519. Curran Associates, Inc., 2017.
- Zhang, H., Zheng, Z., Xu, S., Dai, W., Ho, Q., Liang, X., Hu, Z., Wei, J., Xie, P., and Xing, E. P. Poseidon: An efficient communication architecture for distributed deep learning on GPU clusters. In *2017 USENIX Annual Technical Conference (USENIX ATC 17)*, pp. 181–193, Santa Clara, CA, 2017. USENIX Association. ISBN 978-1-931971-38-6.
- Zhu, H., Akrouit, M., Zheng, B., Pelegriis, A., Phanishayee, A., Schroeder, B., and Pekhimenko, G. TBD: benchmarking and analyzing deep neural network training. *CoRR*, abs/1803.06905, 2018. URL <http://arxiv.org/abs/1803.06905>.

A ARTIFACT APPENDIX

A.1 Abstract

Our artifact provides the *P3* source code, benchmarks to regenerate the performance evaluations and scripts to install *P3* and run the benchmarks.

Evaluating *P3* require more than one machine each equipped with at least one GPU and interconnected with high bandwidth network. The artifact includes three image-classification (ResNet-50, InceptionV3 and VGG-19) and one machine translation (Sockeye) benchmarks. For image-classification benchmarks, we use ImageNet1K data set which should be prepared manually. Machine translation benchmark uses IWSLT15 data set and is included in the artifact. The benchmark script launches data parallel training process in the given list of machines and outputs the training throughput.

A.2 Artifact check-list (meta-information)

- **Algorithm:** Priority-based Parameter Propagation (*P3*).
- **Program:** ResNet-50, InceptionV3, VGG-19 and Sockeye. All benchmarks are included.
- **Compilation:** GCC 5.4 or above, CUDA 8 or above, cuDNN 6 or above.
- **Transformations:** No transformation tools required.
- **Binary:** Source code and scripts included to build binaries on Debian/Linux machines.
- **Data set:** ImageNet1K data set need to be prepared manually (download from <http://www.image-net.org/challenges/LSVRC/2012/nonpub-downloads>). IWSLT15 data set is included.
- **Run-time environment:** No restriction on operating systems. Installation scripts are included for Debian based Linux operating systems (Ubuntu 16.04 recommended). Sudo access required. Software dependencies include - OpenCV, OpenBLAS, CUDA, cuDNN and python3.
- **Hardware:** Require more than one machine (four recommended) each equipped with Nvidia GPUs and high bandwidth interconnect (at least 10Gbps).
- **Run-time state:** There is a warm-up period for the training to stabilize. We recommend to take the measurements after skipping first 1000 iterations.
- **Execution:** Run the benchmarks for 1000 iterations once the benchmark execution has stabilized.
- **Metrics:** The primary metric of comparison is the average training throughput.
- **Output:** The benchmarking scripts would output the number of input data points processed per second.
- **Experiments:** Follow the scripts provided below.

- **How much disk space required (approximately)?:** About 200 GB per machine should be enough for running experiments. Data set preparation might require about 500 GB.
- **How much time is needed to prepare workflow (approximately)?:** About one hour to prepare the dataset and compile the source code.
- **How much time is needed to complete experiments (approximately)?:** About 15 minutes per benchmark.
- **Publicly available?:** Yes.
- **Code licenses:** Apache License 2.0
- **Workflow framework used?:** No
- **Archived?:** Yes. <https://doi.org/10.5281/zenodo.2549852>

A.3 Description

A.3.1 How delivered

P3 source code is publicly available on GitHub: <https://github.com/anandj91/p3>. Latest version of *P3* along with benchmarks can be downloaded from <https://doi.org/10.5281/zenodo.2549852>. The unpacked artifact require less than 2 GB disk space.

A.3.2 Hardware dependencies

We recommend to use at least four machines each equipped with Nvidia GPUs and interconnected with at least 10Gbps network.

A.3.3 Software dependencies

We strongly recommend to build the code on Ubuntu 16.04 with GCC 5.4. Building the source code require CUDA, cuDNN, libopenblas-dev, libopencv-dev and python3-dev. We recommend to use CUDA 9 with cuDNN 7.

A.3.4 Data sets

Build and install the source code before preparing the data set. Once prepared, data set need to be copied to the exact same directory location in all the machines.

ImageNet1K: We use ImageNet1K dataset for ResNet-50, InceptionV3 and VGG-19 which can be downloaded from <http://www.image-net.org/challenges/LSVRC/2012/nonpub-downloads>. Once the data set is downloaded it need to be converted to RecordIO format. Rest of the instructions assume that all images are stored as individual image files, and images belonging to the same class are placed in the same directory. All these class directories are then in the same root directory, say *imnet*.

Go to the root directory of the source code and run the following commands for preparing the data set:

```
$ python tools/im2rec.py --list --recursive \
  --train-ratio 0.95 imagenet1k imnet
$ python tools/im2rec.py --resize 480 \
  --quality 95 --num-thread 28 \
  imagenet1k imnet
```

IWSLT15: Used for Sockeye benchmark. Included in the artifact.

A.4 Installation

Install CUDA and cuDNN before installing *P3*. After that, download and unpack the artifact in all the machines at the same location. Go to the root directory and run the following command for building and installing *P3*:

```
$ ./setup-utils/install-mxnet-ubuntu-python.sh
```

Follow the same steps for installing the baseline after switching to the *baseline* branch in the git repository.

A.5 Experiment workflow

Before running the experiments, choose one of the machines as the master machine. Update the *hosts* file at the root directory of the source code in the master machine with the IP addresses of all the participating machines. Make sure that master machine can ssh to all the machines without password.

Run the following command to execute the benchmarks:

```
$ ./run_exp.sh -m (resnet|inception|vgg|sockeye)
```

Following *tc qdisc* command can be used to limit the per interface peak transmission rate:

```
$ sudo tc qdisc add dev <iface> root tbf rate \
<tx_rate>gbit latency 50ms burst 50kb mtu 10000
```

A.6 Evaluation and expected result

The script outputs the training throughput to the standard output. The average throughput can be calculated and compared with the measurements given in the corresponding figures in the Section 5. Depending on the similarity of the hardwares being used for the experiments, speedups of comparable magnitude and trends should be observed.

A.7 Experiment customization

The benchmark scripts can be further customized to adjust the per machine mini-batch size and *hosts* file location. Full usage is given below:

```
$ ./run_exp.sh [-b <mini-batch size>]
                [-h <path to hosts file>]
                -m (resnet|inception|vgg|sockeye)
```

A.8 Methodology

Submission, reviewing and badging methodology:

- <http://cTuning.org/ae/submission-20190109.html>
- <http://cTuning.org/ae/reviewing-20190109.html>
- <https://www.acm.org/publications/policies/artifact-review-badging>

B ADDITIONAL RESULTS

B.1 Network Utilization

In this section, we show that the limitations described in Section 3 exist in other frameworks as well.

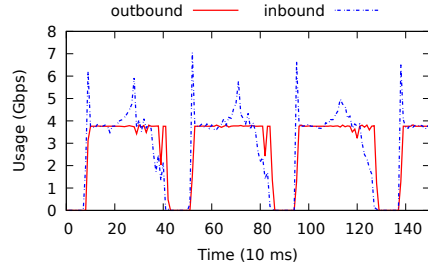


Figure 13. ResNet-50 on TensorFlow at 4Gbps

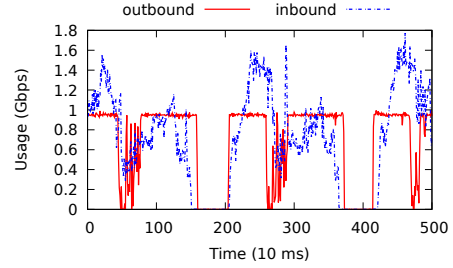


Figure 14. InceptionV3 on Poseidon at 1Gbps

Figures 13 and 14 show the network utilization measurements of TensorFlow and Poseidon taken while training ResNet-50 and InceptionV3 respectively on a 4-node cluster. Similar to MXNet, these frameworks also utilize network poorly even under bandwidth constraints.

B.2 Asynchronous SGD

ASGD algorithm does not perform synchronous update at the parameter server which means each worker machine is only blocked by its on parameter updates in an iteration as opposed to waiting for all the participating workers to finish. ASGD algorithm runs faster than synchronized SGD, however, at the cost of reduced convergence rate because of the stale parameter updates.

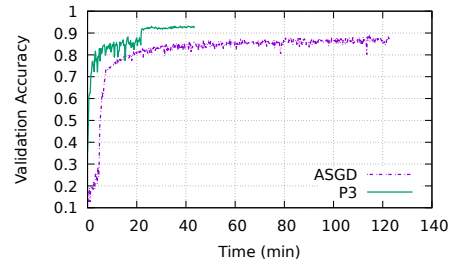


Figure 15. ASGD v.s. P3

We measured the accuracy of ResNet-110 on CIFAR-10 on a 4-machine cluster and 1Gbps network with both *P3* and ASGD. *P3* reaches a final top-1 accuracy of 93% whereas for ASGD, it is only 88%. Additionally, *P3* is able to achieve 80% accuracy roughly 6× faster than ASGD.