

Adaptive Stochastic Gradient Descent for Deep Learning on Heterogeneous CPU+GPU Architectures

Yujing Ma, Florin Rusu
University of California Merced
{yma33,frusu}@ucmerced.edu

Kesheng Wu, Alexander Sim
Lawrence Berkeley National Lab
{kwu,asim}@lbl.gov

Abstract—The widely-adopted practice is to train deep learning models with specialized hardware accelerators, e.g., GPUs or TPUs, due to their superior performance on linear algebra operations. However, this strategy does not employ effectively the extensive CPU and memory resources – which are used only for preprocessing, data transfer, and scheduling – available by default on the accelerated servers. In this paper, we study training algorithms for deep learning on heterogeneous CPU+GPU architectures. Our two-fold objective – maximize convergence rate and resource utilization simultaneously – makes the problem challenging. In order to allow for a principled exploration of the design space, we first introduce a generic deep learning framework that exploits the difference in computational power and memory hierarchy between CPU and GPU through asynchronous message passing. Based on insights gained through experimentation with the framework, we design two heterogeneous asynchronous stochastic gradient descent (SGD) algorithms. The first algorithm – CPU+GPU Hogbatch – combines small batches on CPU with large batches on GPU in order to maximize the utilization of both resources. However, this generates an unbalanced model update distribution which hinders the statistical convergence. The second algorithm – Adaptive Hogbatch – assigns batches with continuously evolving size based on the relative speed of CPU and GPU. This balances the model updates ratio at the expense of a customizable decrease in utilization. We show that the implementation of these algorithms in the proposed CPU+GPU framework achieves both faster convergence and higher resource utilization than TensorFlow on several real datasets.

Index Terms—SGD, fully-connected MLP, adaptive batch size

I. INTRODUCTION

Deep learning has become a disruptive classification technology applied in a wide variety of domains, ranging from image and speech recognition to finance and combustion engines. Building accurate deep learning models is expensive because the training process involves highly-intensive computations, e.g., the multiplication of large matrices. In order to speed up the process, the widely-adopted practice is to use specialized hardware accelerators, e.g., GPUs [7] or TPUs [19], due to their superior performance on linear algebra operations. CPU-only solutions require thousands of cores to achieve similar performance—which is cost-ineffective. However, there is no real system composed only of accelerators—they are add-ons to standard architectures composed of CPUs and memory. In order to use the accelerator, data have to be preprocessed and passed through the system memory – GPUDirect Storage plans to avoid this with a direct data path between fast NVME storage and GPU memory – and the kernels have to be invoked. These procedures are coordinated by the CPU.

Based on the Amazon EC2 instances for accelerated computing, we observe a direct correlation between the number of GPUs – on one side – and the number of CPUs and the memory capacity—on the other. As a practical rule, there are 4-8 CPU cores and 30-60 GB of RAM for every GPU in the system. In addition to the Amazon EC2 instances, the CPU+GPU architecture is already part of some of the most powerful supercomputers on Top500, e.g., Summit and Titan. Due to their flexibility and high-performance, it is likely that CPU+GPU configurations will lead the way to exascale [17] through next-generation systems such as Perlmutter. Thus, provisioning so many resources – CPUs and memory – only to preprocess data and schedule computation on the GPUs is wasteful from the deep learning perspective. Moreover, the number of CPU cores is continuously increasing. The latest CPUs released by Intel [25] and AMD [23] have 112 and 128 hardware threads, respectively. Similar to accelerators, this high degree of CPU parallelism can boost the linear algebra computations common in deep learning.

Problem. We study deep learning training on heterogeneous CPU+GPU architectures. Specifically, our objective is to design heterogeneous SGD algorithms that use efficiently all the resources available in a CPU+GPU system—not only a subpart. The main challenge consists in combining the characteristics of the two architectures – the superior computational power of the GPU with the larger memory on the CPU – into an SGD algorithm with optimal convergence behavior. While heterogeneous architectures have been used for model training before [1], [8], this is done only in the context of the same SGD algorithm and considers how to optimally schedule data and computation across CPU and GPU. Our focus is on optimizing the interaction between the SGD algorithms performed on the two architectures—synchronous SGD on GPU and asynchronous SGD on CPU. The decision to consider architecture-specific algorithms is motivated by their theoretical [5] and empirical [13] characteristics which recommend the use of synchronous SGD with large batches on GPU [7] and asynchronous Hogwild SGD on CPU [16].

Contributions. We introduce an adaptive framework for deep learning on heterogeneous CPU+GPU architectures that maximizes the utilization of each component during the entire execution. We achieve this by concurrent asynchronous coordination, dynamic data partitioning, and architecture-optimized algorithms. CPUs and GPUs are continuously assigned tasks – which they perform concurrently – by a lightweight asyn-

chronous coordinator. The amount of data assigned to a task is dynamically and adaptively determined at runtime based on the current execution state. The CPU+GPU framework is generic and supports the implementation of most existing SGD algorithms [15]. It is an invaluable testbed to evaluate existing algorithms and develop new ones.

We design two heterogeneous SGD algorithms with adaptive batch sizes. They are derived from the scalable asynchronous Hogbatch algorithm [16]. The first algorithm – CPU+GPU Hogbatch – combines small batches on CPU with large batches on GPU, which are both used to update a single shared model asynchronously. While providing better convergence than the single-architecture optimal algorithms, CPU+GPU Hogbatch hinders statistical convergence. The second algorithm – Adaptive Hogbatch – continuously monitors the number of updates performed by every task and changes the batch size dynamically based on the relative speed of CPU and GPU. This balances the model updates ratio at the expense of a customizable reduction in resource utilization.

We implement the two algorithms – together with mini-batch and Hogbatch solutions for CPU and GPU-only – in the heterogeneous CPU+GPU framework. We perform extensive experiments on several deep nets with increasing structural complexity over multiple real datasets. The results show that both algorithms achieve the fastest time to convergence and maximize the CPU and GPU utilization. Moreover, the heterogeneous algorithms outperform TensorFlow – which performs similarly to our GPU-only algorithm – by a significant margin.

II. RELATED WORK

We provide a comparison between the proposed framework and two other classes of systems that support deep learning on heterogeneous CPU+GPU architectures—TensorFlow and Omnivore. TensorFlow [1] – and all the other related systems – use heterogeneity at a smaller granularity. That is, they schedule linear algebra primitives across CPU and GPU. The decision on where to perform a primitive depends on the estimated execution time for each device. Unlike our framework, TensorFlow executes a single instance of the SGD algorithm which updates the unique model synchronously. There are a few problems with this approach. The amount of overlap between CPU and GPU execution is somewhat limited by the sequential structure of the DNN. Since the primitives have order dependencies, it is difficult to schedule more than one at a time. This results in the utilization of a single resource. Scheduling is heavily constrained by previous decisions because switching between CPU and GPU introduces time-consuming data transfers. Moreover, scheduling primitives instead of the complete SGD has more overhead. Similar to our framework, Omnivore [8] splits the training data into batches having size proportional with the speed of the device. However, this size is statically computed and kept constant over the entire execution. The goal is to have perfectly synchronized execution with no delay across devices. The problem is that the actual speed at runtime can be quite different from the estimated one. We address this issue

with dynamic batch sizes and asynchronous model updates. Heterogeneity is also considered in the distributed parameter server setting [10]. The main difference from the centralized CPU+GPU architecture is that training data are statically partitioned to workers. Moving data between workers incurs expensive network traffic and is not viable. Instead, the applied solution uses different learning rates across workers. Similar to our work, the learning rate is computed based on the number of model updates. However, learning rate maintenance is more complex than modifying the batch size.

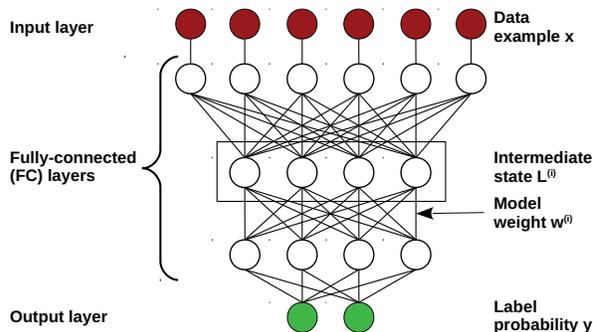


Fig. 1: DNN with three fully-connected (FC) layers.

Although the relationship between the batch size and learning rate – on one side – and the number of updates, convergence, and utilization – on the other – is well-known [7], [11], there is an ongoing debate about the optimal batch size – small or large – and learning rate. Small batches generate more model updates, thus faster convergence. However, they do not saturate the high GPU throughput and result in low utilization. A practical solution is to increase the learning rate proportionally to the batch size [7]. While this increases utilization, it also introduces convergence instability—especially close to the minimum. Our novel approach is to combine small and large batches in a single asynchronous SGD algorithm. The CPU performs a large number of small updates which move the loss function closer to the minimum faster. However, since they are based on a crude estimation of the gradient, they can be quite noisy. This is where the more accurate GPU updates are important—they move the loss in a better direction. Abstractly, we can think of the CPU updates as many small steps in a guessed direction, while the GPU updates are rare jumps using a compass. This combination of updates – albeit sequential – is theoretically proven to enhance SGD convergence [2] and is at the origin of the SVRG family of algorithms [9]. We show empirically that it also improves convergence.

III. SGD FOR DEEP LEARNING

The central component of deep learning is a *Deep Neural Network (DNN)* [3]. As depicted in Figure 1, a DNN is a layered network that takes as input an example given as a feature vector – the input layer – and produces the probability this example belongs to each class in a predefined set—the

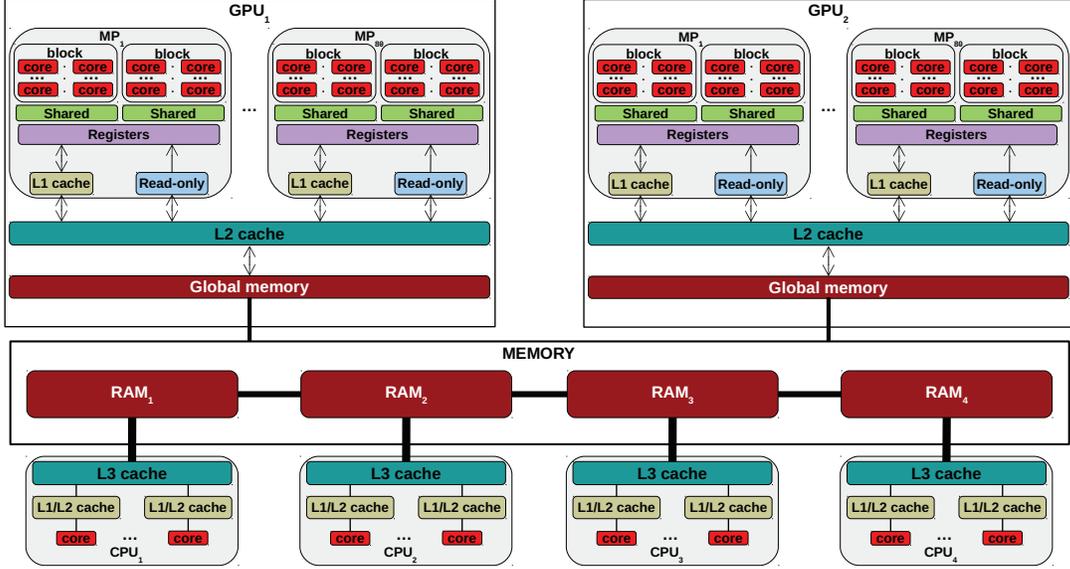


Fig. 2: Heterogeneous CPU+GPU architecture.

output layer. The intermediate layers are hidden to the user—they represent the model to be learned. Each layer contains a set of nodes or vertices. Nodes from two adjacent layers are connected by edges having weights and form a bipartite graph. If the graph is complete, i.e., there is an edge between any pair of nodes, the layer is called fully-connected (FC).

Formally, let the input data be a 2-D matrix $X \in \mathbb{R}^{N \times d_1}$ consisting of d_1 -dimensional vectors x_i for each of the N examples. The vectors x_i propagate through the DNN layers to the output, where their corresponding output labels y_i are produced. Let the intermediate state of x_i at layer l be L_i^l , with $L_i^1 = x_i$, where L_i^l is a d_l -dimensional vector—the input changes its shape through the DNN. In each layer l , a series of linear algebra operations are applied to L_i^l in order to generate L_i^{l+1} . The most intensive such operation is the matrix-vector product between vector L_i^l and matrix $W^l \in \mathbb{R}^{d_{l+1} \times d_l}$ corresponding to the weights on the edges between the nodes in layers l and $(l+1)$ —the other operations are element-wise. If the DNN has P layers and we denote the operation at layer l by \mathcal{F}_l , then the complete processing of x_i can be expressed as:

$$y_i = \mathcal{F}_P(W^{P-1} \cdot \mathcal{F}_{P-1}(\dots \mathcal{F}_1(W^1 \cdot x_i) \dots)) \quad (1)$$

where $L_i^{l+1} = \mathcal{F}_l(W^l \cdot L_i^l)$ are the separate intermediate states. Essentially, a DNN is a composite function of embedded sub-functions over matrix-vector products between data and layer weights. DNN training corresponds to finding the optimal values for the weights in matrices W^l , $1 \leq l \leq P$ —denoted collectively the *model* $W = \{W^1, W^2, \dots, W^P\}$ —that minimize the *loss function* $\ell(X, W, Y)$ for the training dataset X —lower loss indicates high prediction accuracy.

Here, Y represents the set of known labels which are combined in the loss ℓ with the predictions corresponding to a fixed W .

SGD is the most common method to train DNN models [4]. At high-level, SGD iteratively computes the gradient – or derivative – of the loss function over the training dataset and moves the model W in the opposite direction of the gradient—which results in a decrease of the loss. Gradient computation requires a sequence of two passes over the DNN. In the *forward pass*, the predicted labels are computed for the training data based on the current model W —the first model is randomly initialized. The *backward pass* implements the chain rule of calculus for computing the gradient of a composite function starting from the predicted label y_i . If we denote by $\nabla \mathcal{F}_l$ the gradient with respect to model W at layer l , then the back-propagation rule that computes the gradient g_i is:

$$g_i = \nabla \mathcal{F}_1(\dots \nabla \mathcal{F}_{P-1}(\nabla \mathcal{F}_P(\ell(y_i) \cdot W^P) \cdot W^{P-1}) \dots) \quad (2)$$

We observe that the form of the forward and backward expressions in Eq. (1) and Eq. (2) are quite similar, having matrix-vector product as their dominant operation. The update equation at layer l is:

$$W^l \leftarrow W^l - \eta \cdot g_i^l \quad (3)$$

where the *learning rate* η is the scaling factor applied to the magnitude of the gradient. The learning rate is a hyperparameter of SGD—not a parameter of the DNN model. SGD can be stopped either after a fixed number of iterations, i.e., *epochs*, or when there is no significant drop in the loss across iterations. In practice, due to the large dataset size and number of iterations it takes to converge, each SGD iteration is performed only over a randomly selected *batch* of B training examples – not the entire dataset – where

B is another hyperparameter. In this case, the matrix-vector multiplications become matrix-matrix multiplications, which are computationally more intensive, thus, the extensive use of GPUs in DNN training.

IV. CPU+GPU ARCHITECTURE

Figure 2 depicts graphically a heterogeneous CPU+GPU architecture with 4 CPUs and 2 GPUs connected together to the shared memory bus. Each CPU contains multiple cores and cache layers. The L1 and L2 caches are associated with each core, while the L3 cache is shared across all the cores in a CPU node. Each CPU is directly connected to a region of the DRAM memory. The CPUs are connected to each other by high-bandwidth interconnects. To access DRAM regions on other nodes, data is transferred over these interconnects. However, this is slower than accessing the local memory, thus, the non-uniform memory access (NUMA) pattern. NUMA cache-coherency is implemented in hardware, thus implicit.

	CPU	GPU
cores	18	172 per MP
blocks	—	32 per MP
threads	36	2048 per MP
L1 cache	32(I) + 32(D) KB	128 KB
L2 cache	256 KB	6 MB
L3 cache / shared memory	45 MB	96 KB
MEMORY / global memory	488 GB	16 GB

TABLE I: Hardware architecture specifications.

A GPU contains multiple streaming multiprocessors (MP). Each MP consists of a large number of specialized cores targeted at a limited subset of instructions. In the CUDA programming model, work is issued to the GPU in the form of a function, referred to as the kernel. A logical instance of the kernel executed on an MP core is called a thread. The kernel code is parametrized by a logical thread identifier that allows each thread to operate on a different partition of the input data—which has to be moved explicitly between the CPU and GPU memory. Since thousands of threads can be executed concurrently across MPs, global thread synchronization is not available. Nonetheless, synchronization can be enforced at thread block level. Threads can access the various units of the deep memory hierarchy in Figure 2 explicitly in the code. When a global memory address is requested by a thread, aligned successive addresses are converted into a single memory transaction—memory coalescing. Thus, consecutive threads have to access consecutive addresses in order to minimize the number of memory transactions.

Table I gives the hardware specification of the CPU nodes and the NVIDIA Volta V100 GPU used in this paper. While the number of cores and threads is much larger for the GPU, the numbers for the CPU are also quite high, e.g., 36 independent threads can run concurrently on a single CPU. Since the system has two CPUs, there are 72 CPU threads overall. However, this is not sufficient to reduce the gap in performance given by the superior GPU degree of parallelism—there are 80 MPs on the V100 GPU. Although the amount of memory

available on the CPU is 30X larger than on the GPU, the L2 cache on the GPU is 24X larger. This reflects the throughput emphasis of the GPU memory hierarchy as opposed to the latency optimization for CPU.

V. DNN FRAMEWORK ON CPU+GPU

We pursue two main objectives in designing the CPU+GPU framework for deep learning. First, the framework is a generic testbed to evaluate existing SGD algorithms and develop new ones. This is achieved by a modular architecture in which components are assigned independently to hardware resources. An SGD algorithm is expressed by a series of primitive operations and a communication strategy between components. The second objective is to maximize the utilization of every resource during execution. We achieve this by concurrent asynchronous coordination, dynamic data partitioning, and architecture-optimized SGD algorithms. CPUs and GPUs perform concurrent asynchronous SGD algorithms – specialized for their specific architecture – on data assigned dynamically and adaptively at runtime based on the current execution state. In this section, we present the architecture and workflow of the proposed DNN framework on heterogeneous CPU+GPU.

A. Framework Architecture

The architecture of the heterogeneous CPU+GPU framework for deep learning is depicted in Figure 3. It consists of a series of asynchronous worker threads corresponding to each of the CPUs and GPUs in Figure 2, and a central coordinator. In this example, there are four CPU workers and two GPU workers. However, in a shared virtualized environment such as Amazon EC2, the framework can be assigned only a subset of the available hardware resources. The coordinator and workers are implemented as stand-alone system threads that exist over the entire duration of the program. The worker assigned to a hardware component is in charge of managing the resources, e.g., cores, memory, threads, and operation of that component. The coordinator assigns data and tasks to workers, and schedules their interaction. The communication between the coordinator and workers – workers do not communicate directly – is realized through control messages, while data are passed through references in the shared memory space. In the case of deep learning, these data include the model – and its gradient – and the training examples split into batches. The coordinator maintains the global model and prepares the training data. Each worker is assigned a model replica – which can be either a deep or shallow copy of the global model – and a data batch—which is a reference to a range in the training data at the coordinator. Handling the training data is simpler because it requires read-only access. The hyperparameters of the SGD algorithm are maintained by the coordinator.

Coordinator. The coordinator is the equivalent of the parameter server in distributed or multi-GPU [6], [11] settings. Its main role is to control workers’ access to the global model through the model update policy. Since the coordinator thread processes messages sequentially, the default policy is synchronous model updates—the replicas are applied to the

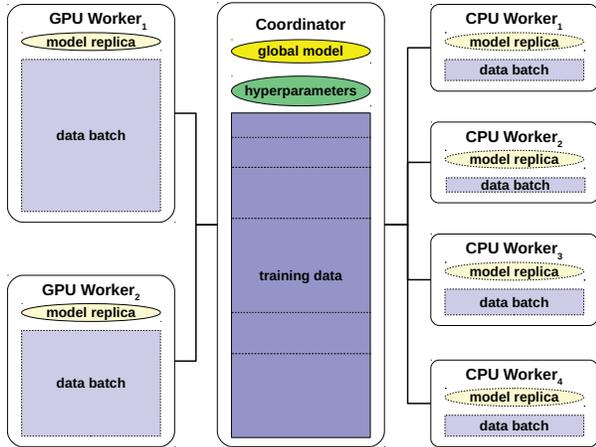


Fig. 3: Framework architecture.

global model one after another, in the order in which they are received. In order to support asynchronous updates, the model update logic has to be moved to the workers. After computing the gradient, the workers apply it to their replica reference – a pointer to the global model – concurrently. In this case, the burden on the coordinator is considerably smaller because it does not execute any part of the SGD algorithm.

In our shared memory framework, the coordinator plays an additional role that is completely missing from distributed parameter servers [6]. The coordinator assigns data batches of different size dynamically and adaptively to the workers based on their processing speed. This is a fundamental feature in a heterogeneous CPU+GPU architecture. If the same batch size is given to a CPU and a GPU worker, the GPU worker would process a number of batches equal to the ratio between the speed of the GPU and that of the CPU [8]. Since this ratio is significant, the GPU would process hundreds or thousands of batches while the CPU processes a single batch. The end result would be that the CPU updates are ignorable. Alternatively, if CPU and GPU are synchronized, the GPU would be stalled most of the time. In order to cope with this issue, the coordinator continuously monitors the number of updates each worker executes and changes the batch size such that the difference between the fastest and slowest worker is bounded. This strategy is implemented inside the model update procedure and requires only a simple reference assignment. As far as we know, this is the first learning framework that supports dynamic batch sizes across concurrent workers. In all the other solutions, the training data is statically partitioned and distributed to workers.

CPU Workers. The workers are statically associated with a computational resource – CPU socket or GPU – and perform an iteration of the SGD algorithm on the assigned batch and model. Since CPU workers share the same address space with the coordinator, they have direct access to the global model and training data. This allows for reference access and avoids deep copies—the dotted lines in Figure 3. However,

due to the uneven NUMA memory access, references can introduce unexpected cache coherency effects [16], [21]. The CPU worker has to consider another level of parallelism – corresponding to the local cores/threads – when performing the SGD algorithm. The alternatives are to compute a single gradient over the entire batch or to split the batch into smaller sub-batches and compute a gradient for each. In the first case, intra-thread parallelism is applied only to the linear algebra operations and is encapsulated in the corresponding library functions, e.g., Intel MKL. In the second case, there are two levels of parallelism—inter-thread parallelism across sub-batches and intra-thread parallelism inside a sub-batch. The inter-thread parallelism has to be implemented in the CPU worker. This can be done with explicit threads or with higher-level constructs such as OpenMP. Based on the level of parallelism and the model update policy, many variations of the SGD algorithm can be designed [21]—supported by the framework with different implementations of the CPU worker.

GPU Workers. A GPU worker is associated with every GPU accelerator in the system—for which it serves as the exclusive interface. The GPU worker coordinates the memory transfers between CPU and GPU, and invokes kernel execution on the GPU—all these happen asynchronously and with minimal interference on the other system components. This allows for advanced GPU features, such as data transfer through the unified memory address space and kernel execution through asynchronous streams, to be isolated in the GPU worker. The execution of the SGD algorithm on GPU follows the standard pattern of first moving the data and the model, and then invoking kernels for the linear algebra operations, e.g., from the cublas library, on the forward and backward DNN passes. By default, the intermediate output of kernel invocations is kept in the GPU memory in order to reduce data movement. However, advanced memory management strategies that work at layer granularity [18] can also be added. The main difference between the CPU and GPU worker is how they handle the model—the model replica in the GPU worker is always a deep copy of the global model. This is because the replica serves as a transition buffer between CPU and GPU, which is accessed only during transfers between the two. Multiple accesses to the global model in the GPU memory may have unexpected consequences.

B. Framework Workflow

Figure 4 illustrates how the framework performs the SGD algorithm for deep learning. The tasks executed by each worker, as well as the messages exchanged with the coordinator, are shown in the figure. During initialization, the coordinator loads the training data in memory and prepares it for the linear algebra operations in SGD. The global model is allocated and initialized with arbitrary/random values. The model configuration is passed to the workers for their initialization of the model replicas. This is necessary only for the GPU workers which have to allocate memory on the device. Since multiple SGD implementations are supported, the coordinator has to select an algorithm and its hyperparameters for each worker,

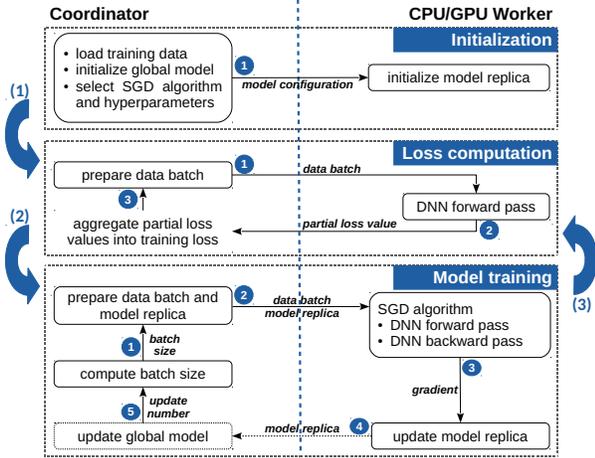


Fig. 4: Framework execution workflow.

and a global model update policy. While currently specified by the programmer, we envision a solution in which the active workers and their algorithms are selected automatically.

The SGD algorithm is performed in the model training stage. At each iteration, the coordinator starts by determining the batch size corresponding to every worker. Initially, the size is proportional to the worker speed. Later, the size changes adaptively based on the number of model updates performed by the worker. The coordinator prepares a batch by selecting a continuous range from the training data and storing a reference to its starting position. The model replica is initialized with the current state of the global model. This can be a reference to the global model or a full deep copy. The batch and the replica are passed to the worker to execute the SGD algorithm selected in the initialization stage. This is the main part of model training and consists of a forward and backward pass through the DNN to compute the gradient. Finally, the gradient is applied to update the model—another DNN forward pass. If the update is applied to the deep replica, this has to be subsequently integrated in the global model—which can be done synchronously at the coordinator or asynchronously at the worker. In the case of reference replicas, the update is directly applied to the global model. The last step in model training – which closes the loop – is the message sent by the worker to the coordinator to inform that the update has been applied. Since these messages are processed sequentially, the next batch size is computed individually for each worker based on its number of updates.

VI. SGD WITH ADAPTIVE BATCH SIZES

The CPU+GPU framework supports the implementation of heterogeneous versions of most – if not all – existing SGD algorithms [15]. Modifications are confined only to the type of messages exchanged between coordinator and the workers, and how they are handled. In this section, we introduce two heterogeneous SGD algorithms derived from the scalable asynchronous Hogbatch [16]. We choose Hogbatch

as our base SGD algorithm because of two reasons. First, it supports asynchronous updates. These are perfectly suited for the speed difference between CPU and GPU. Second, unlike Hogwild [14], Hogbatch operates on batches. This matches better the highly-parallel GPU architecture optimized for throughput—the larger the batch, the higher the utilization [11]. The experimental results in Section VII confirm the necessity to design these specialized algorithms—and their superiority over standard Hogbatch.

Algorithm 1 Hogbatch

Coordinator: ScheduleWork Message Handler

Input:

Worker \mathcal{E} asking for work
Set of batches \mathbb{B} with b training examples per batch

1. **if** $\mathbb{B} \neq \emptyset$ **then**
 2. Extract next batch \mathcal{B} from \mathbb{B}
 3. $\mathbb{B} \leftarrow \mathbb{B} \setminus \mathcal{B}$
 4. Send message *ExecuteWork*(\mathcal{B}) to worker \mathcal{E}
 5. **end if**
-

Worker \mathcal{E} : ExecuteWork Message Handler

Input:

Batch \mathcal{B} with b training examples

1. Gradient: $g \leftarrow \nabla_{\mathcal{F}_1} (\dots \nabla_{\mathcal{F}_P} (\ell(\mathcal{B}) \cdot W^P) \dots)$
 2. Update model: $W \leftarrow W - \eta \cdot g$
 3. Send message *ScheduleWork*(\mathcal{E}) to coordinator
-

A. Hogbatch

The mapping of Hogbatch to our framework is given in Algorithm 1. The main task of the coordinator is to serve work requests from workers. For this, the coordinator extracts a batch of b training examples and sends them to the requesting worker. It is important to notice that *the same batch size b is given to all the workers*. When there are no more batches and all the workers are done, an SGD epoch has finished and the process is restarted with the full training dataset. While the coordinator executes requests serially, the workers process batches concurrently. First, they compute the gradient of the assigned batch on the current DNN model. Then, they update the model with the computed gradient. In Hogbatch, the DNN model is shared across all the workers—the local replicas are references to the global model. Since the workers read and modify the model concurrently without any synchronization primitives, conflicts are unavoidable. However, the speedup provided by parallel processing outweighs the impact of update conflicts and results in faster convergence [14], [16].

B. CPU+GPU Hogbatch

The direct application of Hogbatch in a heterogeneous CPU+GPU architecture raises two problems. First, the model has to be copied to the GPU memory, thus, access by reference does not work. Our solution is to create a deep copy of the global model in the GPU worker *ExecuteWork* message

handler. The GPU kernels operate exclusively on this replica. Once the replica is updated, we push it to the global model asynchronously. If the GPU workers have similar speed, they perform a similar number of updates and their local replicas do not become stale. Otherwise, merging a local stale replica requires careful consideration. In this case, the gradient is computed on a model, while the update is performed on another – most recent – model. Additionally, the learning rate can be decreased to compensate for the stale gradient [10], diminishing the importance of the update.

The second problem is triggered by using the same batch size b across all the workers. Due to the orders of magnitude difference in processing speed between CPU and GPU, the GPU performs considerably more updates. In the worst case, the CPU takes more time to process a single batch than the GPU processing all the others. This behavior is detrimental because the CPU ends up doing useless work and, moreover, it stalls the GPU. Our novel solution is to *use different batch sizes for CPU and GPU*. The CPU batch size is set to t – where t is the number of cores or threads on the CPU – so that each thread processes exactly one example. The rationale for this choice is to ensure that all the threads are utilized. This special case of Hogbatch is the Hogwild algorithm [14]. The GPU batch size is set to a value that satisfies two conditions. First, it guarantees a high enough utilization of the GPU. Second, the time to process a batch on GPU is close to the time on CPU. However, the GPU memory capacity imposes an upper bound on the size. Based on these constraints, the GPU batch size varies from a few hundreds to several thousands, depending on the DNN structure. This idea can be generalized to having a different batch size for every worker. Notice that supporting different batch sizes across workers requires minimal changes to the *ScheduleWork* message handler in Algorithm 1.

While the benefit of using different batch sizes is important to reduce staleness among workers, it may be argued that its impact on convergence is harder to assess. Indeed, there is no theoretical analysis for Hogbatch with different batch sizes. However, the analysis of any SGD asynchronous algorithm makes strong simplifying assumptions [14] that rarely hold in practice. Intuitively, the interaction between small and large batches improves convergence because it combines a large number of model updates based on inaccurate gradients – corresponding to small batches – with updates from accurate gradients computed on large batches. Our conjecture – supported empirically – is that convergence remains superior even when the two types of updates are applied concurrently. Moreover, we set the learning rate to be proportional with the batch size [7]—we have both different batch sizes and different learning rates. This guarantees that the impact of the more accurate gradients on convergence is higher.

C. Adaptive Hogbatch

The problem with CPU+GPU Hogbatch is that the batch sizes have to be determined prior to execution. This can be a lengthy trial-and-error process that adds complexity to hyperparameter tuning. Moreover, the batch sizes are static and

Algorithm 2 Adaptive Hogbatch

Coordinator: *ScheduleWork* Message Handler

Input:

Worker \mathcal{E} asking for work; Number of model updates $u^\mathcal{E}$ performed by worker \mathcal{E} ; Batch size $b^\mathcal{E}$ for worker \mathcal{E} ; Minimum (min_u) and maximum (max_u) number of updates performed by all other workers except \mathcal{E} ; Minimum ($min_b^\mathcal{E}$) and maximum ($max_b^\mathcal{E}$) batch size threshold for worker \mathcal{E} ; Training dataset \mathbb{B}

▷ Update batch size $b^\mathcal{E}$ for worker \mathcal{E}

1. **if** $u^\mathcal{E} < min_u$ **then**
2. $b^\mathcal{E} \leftarrow maximum(b^\mathcal{E} / \alpha, min_b^\mathcal{E})$; $min_u \leftarrow u^\mathcal{E}$
3. **else if** $u^\mathcal{E} > max_u$ **then**
4. $b^\mathcal{E} \leftarrow minimum(b^\mathcal{E} \cdot \alpha, max_b^\mathcal{E})$; $max_u \leftarrow u^\mathcal{E}$
5. **end if**

▷ Prepare and send batch to worker \mathcal{E}

6. **if** $b^\mathcal{E} \leq |\mathbb{B}|$ **then**
 7. Extract batch \mathcal{B} of size $b^\mathcal{E}$ from \mathbb{B}
 8. $\mathbb{B} \leftarrow \mathbb{B} \setminus \mathcal{B}$
 9. Send message *ExecuteWork*(\mathcal{B}) to worker \mathcal{E}
 10. **end if**
-

CPU Worker \mathcal{E} : *ExecuteWork* Message Handler

Input:

Batch \mathcal{B} with b training examples

Number of threads t

1. Split \mathcal{B} into t sub-batches $\{\mathcal{B}_1, \dots, \mathcal{B}_t\}$ of size \mathcal{B}/t
 2. **for** $i = 1$ **to** t **do in parallel**
 3. Gradient: $g_i \leftarrow \nabla \mathcal{F}_1(\dots \nabla \mathcal{F}_P(\ell(\mathcal{B}_i) \cdot W^P) \dots)$
 4. Update model: $W \leftarrow W - \eta \cdot g_i$
 5. **end for**
 6. $u^\mathcal{E} \leftarrow u^\mathcal{E} + t \cdot \beta$ ▷ Increase number of model updates
 7. Send message *ScheduleWork*($\mathcal{E}, u^\mathcal{E}$) to coordinator
-

they do not take into consideration the runtime execution environment. This can lead to unbounded divergence between the number of updates across CPU and GPU, which manifests by loss function instability and, ultimately, slower convergence.

We address these issues in Adaptive Hogbatch—depicted in Algorithm 2. The main idea is to continuously monitor the workers’ status and *update the batch size dynamically based on the number of updates*. This can be done in the *ScheduleWork* message handler at the coordinator. While the relationship between the number of updates and resource utilization is clear, the connection to convergence is not straightforward, especially when the updates are computed over batches with different size. The number of updates has to be large. Due to computational and memory constraints, this can be achieved only with small batches. However, small batches generate inaccurate gradients—which hurt convergence. In order to address these conflicting goals, we apply two criteria when computing the batch size. First, the gap in the number of updates between the slowest and fastest worker has to be bounded. This is achieved by slowing down (i.e., increasing

dataset	#examples	#features	#labels	size	CPU batch size	GPU batch size	DNN architecture
covtype	581,012	54	2	485 MB	[1-64]	[128-8,192]	54-512-512-512-512-512-2
w8a	64,700	300	2	155 MB	[1-64]	[64-8,192]	300-512-512-512-512-512-512-2
delicious	16,105	500	983	128 MB	[1-32]	[64-2,048]	500-512-512-512-512-512-512-983
real-sim	72,309	20,958	2	12.1 GB	[1-64]	[64-8,192]	20,958-512-512-512-512-2

TABLE II: Experimental datasets. CPU and GPU batch size are the minimum and maximum batch size used in the Hogbatch algorithms. DNN architecture corresponds to the number of layers and the number of units per layer.

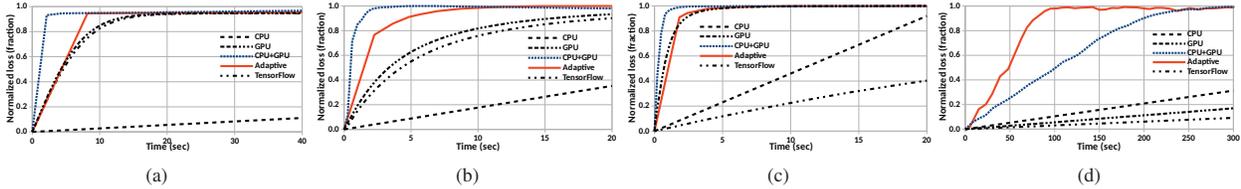


Fig. 5: Normalized loss for time to convergence: (a) covtype, (b) w8a, (c) delicious, (d) real-sim.

the batch size) the worker with the largest number of updates or speeding up (i.e., decreasing the batch size) the worker with the smallest number of updates, respectively. The value of the batch size is scaled up or down by a constant factor α which is a user-defined parameter set by default to 2—the batch size is doubled or halved, respectively. The second goal is to maintain a minimum level of utilization on every worker. For this, we define lower and upper thresholds on the batch size, which we do not allow to be crossed. Alternatively, we can monitor the actual utilization for devices that provide such APIs. The initial batch size is set to the lower threshold for CPU and the upper threshold for GPU. The computation of a new batch size is light and does not incur observable overhead.

The CPU worker in Adaptive Hogbatch (Algorithm 2) has to maintain the number of model updates it performs. This poses some complications because of the asynchrony incurred by the nested Hogbatch execution. While t threads perform updates, these are conflicting, thus, it is not clear how many survive. We quantify this uncertainty through the user-defined parameter β which specifies the fraction of surviving updates. When $\beta = 1$ – the default value determined empirically – the CPU worker performs t updates per batch. The closer β gets to 0, the fewer updates are considered by the coordinator when computing the new batch size.

VII. EXPERIMENTAL EVALUATION

The purpose of the experimental evaluation is to investigate the following questions (see complete details in [12]):

- Are the heterogeneous Hogbatch algorithms improving upon the CPU and GPU-only alternatives in time to convergence and statistical efficiency?
- How does the heterogeneous framework compare with the state-of-the-art TensorFlow?
- What is the ratio of model updates among CPU and GPU?
- What utilization do the implementations of the Hogbatch algorithms in our framework achieve on CPU and GPU?

A. Setup

Implementation. We implement the heterogeneous CPU+GPU framework for deep learning in C/C++ using the pthreads library. The coordinator and each worker is managed by a stand-alone thread. The threads communicate using our custom asynchronous message queue. The CPU worker schedules Hogbatch instances on its corresponding cores using dynamic OpenMP (3.7.0-3) threads. The linear algebra operations on CPU are implemented with Intel MKL (2.187) functions invoked inside OpenMP threads. The GPU worker invokes kernels written in CUDA 10.0 which call the linear algebra primitives from the Nvidia cublas (10.2.1.243-1) library. The TensorFlow (1.13.1) implementation consists of the driver program in which the DNN architecture and the objective function are defined [26]. Then, the mini-batch SGD optimizer is called to perform the training. All the code is available online as open source [24].

Hardware. Although we perform experiments on two systems, due to lack of space, we present only the results for the Amazon AWS p3.16xlarge instance [22] which has 8 Nvidia Volta V100 GPUs and 64 CPU threads. The specification of this computing architecture is given in Table I. We use the standard AWS configuration from March 2020. In order to maximize utilization, we run experiments with a single GPU. We assign 56 out of the 64 threads to a single CPU worker in order to simplify their use in OpenMP. This allows up to 56 threads to perform concurrent model updates. The number of OpenMP threads for linear algebra operations is limited to 60. The remaining threads are the CPU workers and coordinator.

Datasets and DNN configurations. We consider four real data sets – depicted in Table II – that exhibit large variety in size, features, and number of classes. These datasets have been used previously to evaluate the performance of parallel SGD on CPU and GPU [13], [20]—more details can be found therein. We process all the datasets in dense format. The batch size on CPU varies between 1-64 examples per thread, while for GPU it ranges between 64-8,192. The number of hidden

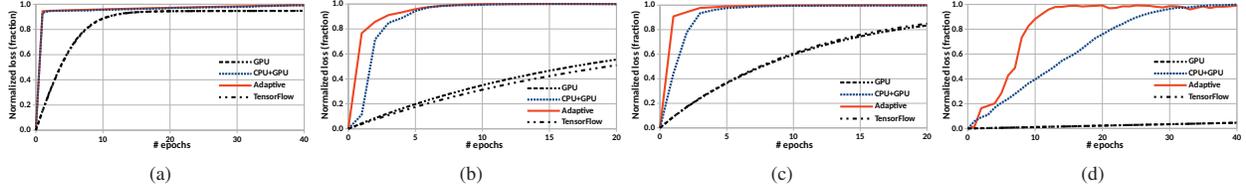


Fig. 6: Normalized loss for epochs to convergence, or statistical efficiency: (a) covtype, (b) w8a, (c) delicious, (d) real-sim.

layers is set inversely proportional to the dataset size, to 4 (real-sim), 6 (covtype), and 8 (w8a and delicious). The number of units in a hidden layer is kept constant at 512. Since all the layers are fully-connected, the processing complexity is proportional to the number of layers.

Methodology. We execute each algorithm for the same fixed amount of time. This is chosen such that the loss converges for at least one algorithm. The minimum loss across all the algorithms is taken as basis for comparison. All the loss values are normalized to this basis. This process measures which algorithm converges fastest to a certain loss—the ultimate goal in practice. The upper and lower thresholds for batch size are varied with the datasets and the DNN architecture. The GPU utilization for the lower threshold is about 50%, while for the upper threshold is close to 100%. The initial batch size is set to the upper threshold on the GPU workers. The CPU worker starts with a batch size of 1 example per thread—it performs Hogwild. The number of model updates is measured as the average over all the epochs. All the algorithms are initialized with the same model, which gives the same initial loss. The initial values of the DNN weights are randomly drawn from a normal distribution with standard deviation equal to the number of units in the current layer. The sigmoid function is used as activation in the hidden layers. Softmax activation is applied to the output layer in order to compute the cross-entropy loss. The SGD learning rate is chosen by gridding its range in powers of 10 and selecting the value that achieves the lowest loss across all the algorithms. The batch size and learning rate are correlated and set according to [7]. We emphasize that the same hyperparameters are used for the same hardware architecture. The time to load the data, output the result, and evaluate the loss are not included.

B. Results

We include four Hogbatch algorithms – CPU, GPU, CPU+GPU, and Adaptive – and TensorFlow in the experiments. Hogbatch CPU is Hogwild on CPU-only, while Hogbatch GPU – and TensorFlow – are mini-batch SGD.

Time to convergence. The normalized loss function is depicted in Figure 5. Hogwild CPU takes considerably longer – from 236X to 317X – to execute an SGD epoch than GPU, thus its loss follows a slope increasing at a much slower – linear – rate in the beginning. In fact, Hogwild CPU did not finish an epoch in the allocated time budget for any of the datasets—it got close for delicious. Nonetheless, the number of

model updates per epoch is the highest among all the methods. The relative performance between CPU and GPU matches perfectly the relationship between Hogwild (CPU) and mini-batch (GPU). On low-dimensional data, mini-batch converges faster. However, as the dimensionality increases, there is a switch between the two, with Hogwild clearly outperforming mini-batch on real-sim. As expected, TensorFlow mirrors almost identically the convergence curve of Hogbatch (GPU). The only exception is delicious, on which TensorFlow has much worse convergence. The reason is the multi-label classification – 983 vs. 2 labels – which is much slower in TensorFlow. It is evident that the heterogeneous Hogbatch algorithms achieve the steepest decrease in loss per unit of time. The mixture of small and large batches combines the best behavior of the CPU and GPU solutions and improves upon them significantly in all the cases. While CPU+GPU outperforms Adaptive Hogbatch on the first three datasets, the results are reversed for real-sim, where Adaptive reaches the minimum loss in less than half of the time. This is because batch sizes having more uniform values, as dictated by the relative performance of CPU and GPU, generate fewer conflicts in the highly-dimensional feature space. We conclude by pointing out that, while 90% of the minimum loss is achieved fast, further improvement is rather slow on covtype.

Statistical efficiency. Figure 6 depicts the statistical efficiency corresponding to the time to convergence in Figure 5. Statistical efficiency – or loss convergence as a function of the number of epochs – is directly proportional with the number of effective model updates per epoch. Mini-batch (GPU) and TensorFlow have the largest batches, thus, they have relatively poor statistical efficiency. The overlapped curves confirm that our implementation and TensorFlow are identical. Since the heterogeneous Hogbatch algorithms combine small and large batches, as expected, their efficiency is a weighted average of the two. The larger the gap between the batch sizes, the higher the deviation from the optimal statistical efficiency. This explains the superiority of Adaptive over CPU+GPU. The curve corresponding to Hogwild CPU is not included in the figures because of the extremely long time – more than 200X – it takes to perform the required number of epochs.

Model updates distribution. The ratio of model updates performed by the CPU and GPU in the heterogeneous Hogbatch algorithms is depicted in Figure 8. In the case of CPU+GPU, the CPU updates are almost exclusive because the gap between the batch size on CPU and GPU is maximized. As

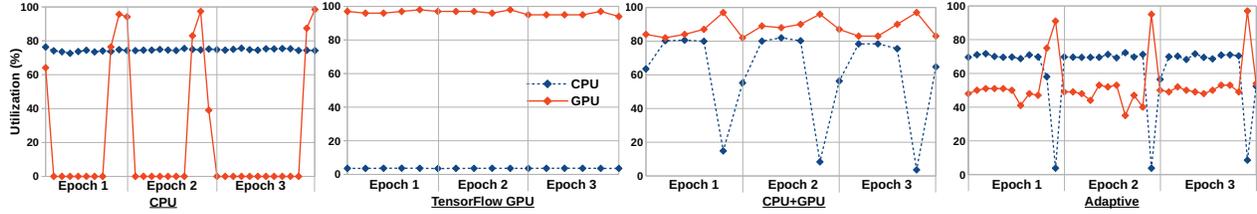


Fig. 7: CPU and GPU utilization for three epochs of the Hogbatch algorithms executed on the covtype dataset.

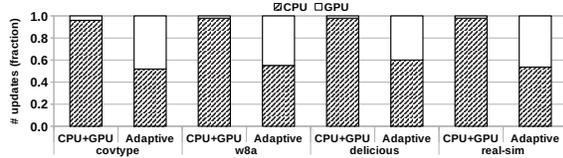


Fig. 8: Ratio of model updates applied by CPU and GPU.

the gap decreases, the distribution moves towards uniformity, with each of the CPU and GPU performing approximately half of the updates in Adaptive. As shown, this enhances convergence in certain cases.

Resource utilization. The CPU and GPU utilization during the execution of three epochs of the Hogbatch algorithms on the covtype dataset are depicted in Figure 7—the results for the other datasets follow a similar pattern. The loss computation is always performed on the GPU at the end of the epoch. This explains the increase in GPU utilization and the decrease in CPU across all the algorithms. The CPU utilization hovers around 80% because only 56 of the available 64 threads are used. The slight decrease on Adaptive is due to the larger batch sizes. The GPU utilization is above 80% in GPU and CPU+GPU since the batch size is 8,192. The batch size in Adaptive decreases to the lower threshold, which triggers the corresponding decrease in utilization. The lower threshold parameter controls the tradeoff between GPU utilization and convergence. In the case of CPU+GPU, utilization is maximized. In Adaptive, the convergence is primordial.

VIII. CONCLUSIONS AND FUTURE WORK

We introduce a generic framework for deep learning on heterogeneous CPU+GPU architectures. We design two asynchronous Hogbatch algorithms. CPU+GPU Hogbatch combines small batches on CPU with large batches on GPU. Adaptive Hogbatch assigns batches with continuously evolving size based on the relative speed of CPU and GPU. These algorithms achieve orders of significantly faster convergence than the corresponding CPU and GPU-only solutions, while using all the resources effectively. In future work, we plan to scale these algorithms to multi-GPU architectures.

Acknowledgments: This work is supported by a U.S. Department of Energy Early Career Award (DOE Career).

REFERENCES

[1] M. Abadi et al. TensorFlow: A System for Large-Scale Machine Learning. In *OSDI 2016*.

[2] D. P. Bertsekas. Incremental Gradient, Subgradient, and Proximal Methods for Convex Optimization: A Survey. MIT 2010.

[3] L. Bottou. Neural Networks: Tricks of the Trade. Springer, 2012.

[4] L. Bottou, F. Curtis, and J. Nocedal. Optimization Methods for Large-Scale Machine Learning. *SIAM Review*, 60(2):223–311, 2018.

[5] J. Chen, R. Monga, S. Bengio, and R. Józefowicz. Revisiting Distributed Synchronous SGD. *CoRR*, abs/1604.00981, 2016.

[6] H. Cui, H. Zhang, G. R. Ganger, P. B. Gibbons, and E. P. Xing. GeePS: Scalable Deep Learning on Distributed GPUs with a GPU-specialized Parameter Server. In *EuroSys 2016*.

[7] P. Goyal, L. Wesolowski, P. Dollár, A. Kyrola, R. Girshick, A. Tulloch, P. Noordhuis, Y. Jia, and K. He. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. *CoRR*, abs/1706.02677v2, 2018.

[8] S. Hadjis, C. Zhang, I. Mitliagkas, and C. Ré. Omnivore: An Optimizer for Multi-device Deep Learning on CPUs and GPUs. *CoRR*, abs/1606.04487, 2016.

[9] Y. Huang, T. Jin, Y. Wu, Z. Cai, X. Yan, F. Yang, J. Li, Y. Guo, and J. Cheng. FlexPS: Flexible Parallelism Control in Parameter Server Architecture. *PVLDB*, 11(5):566–579, 2018.

[10] J. Jiang, B. Cui, C. Zhang, and L. Yu. Heterogeneity-aware Distributed Parameter Servers. In *SIGMOD 2017*, pages 463–478.

[11] A. Kolioussis, P. Watcharapichat, M. Weidlich, L. Mai, P. Costa, and P. Pietzuch. CROSSBOW: Scaling Deep Learning with Small Batch Sizes on Multi-GPU Servers. *PVLDB*, 12(11):1399–1413, 2019.

[12] Y. Ma and F. Rusu. Heterogeneous CPU+GPU Stochastic Gradient Descent Algorithms. *CoRR*, abs/2004.08771, 2020.

[13] Y. Ma, F. Rusu, and M. Torres. Stochastic Gradient Descent on Modern Hardware: Multi-core CPU or GPU? Synchronous or Asynchronous? In *IPDPS 2019*.

[14] F. Niu, B. Recht, C. Ré, and S. J. Wright. Hogwild: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent. In *NIPS 2011*.

[15] S. Ruder. An Overview of Gradient Descent Optimization Algorithms. *CoRR*, abs/1609.04747v2, 2017.

[16] S. Sallinen, N. Satish, M. Smelyanskiy, S. Sury, and C. Ré. High Performance Parallel Stochastic Gradient Descent in Shared Memory. In *IPDPS 2016*.

[17] M. J. Schulte, M. Ignatowski, G. H. Loh, B. M. Beckmann, W. C. Brantley, S. Gurumurthi, N. Jayasena, I. Paul, S. K. Reinhardt, and G. Rodgers. Achieving Exascale Capabilities Through Heterogeneous Computing. *IEEE Micro*, 35(4):26–36, 2015.

[18] S. B. Shriram, A. Garg, and P. Kulkarni. Dynamic Memory Management for GPU-based Training of Deep Neural Networks. In *IPDPS 2019*.

[19] Y. You, J. Li, S. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, J. Demmel, K. Keutzer, and C.-J. Hsieh. Large Batch Optimization for Deep Learning: Training BERT in 76 Minutes. In *ICLR 2020*.

[20] K. Bhatia, K. Dahiya, H. Jain, A. Mittal, Y. Prabhu, and M. Varma. The Extreme Classification Repository: Multi-label Datasets and Code. <http://manikvarma.org/downloads/XC/XMLRepository.html>, 2016.

[21] C. Zhang and C. Ré. DimmWitted: A Study of Main-Memory Statistical Analytics. *PVLDB*, 7(12), 2014.

[22] Amazon EC2 Instance Types. <https://aws.amazon.com/ec2/instance-types/>, 2020.

[23] AMD EPYC 7742. <https://www.amd.com/en/products/cpu/amd-epyc-7742>, 2020.

[24] CPU+GPU SGD. <https://github.com/YMA33/CPU-GPU-SGD>, 2020.

[25] Intel Xeon Platinum 9282 Processor. <https://ark.intel.com/content/www/us/en/ark/products/194146/intel-xeon-platinum-9282-processor-77m-cache-2-60-ghz.html>, 2020.

[26] SLIDE. <https://github.com/keroro824/HashingDeepLearning>, 2020.