

# SPARCML: High-Performance Sparse Communication for Machine Learning

Cédric Renggli  
ETH Zurich

Dan Alistarh  
IST Austria

Torsten Hoefler  
ETH Zurich

## Abstract

One of the main drivers behind the rapid recent advances in machine learning has been the availability of efficient system support. This comes both through faster hardware, but also in the form of efficient software frameworks and programming models. Despite existing progress, scaling compute-intensive machine learning workloads to a large number of compute nodes is still a challenging task. In this paper, we address this challenge, by proposing SPARCML,<sup>1</sup> a general, scalable communication layer for machine learning applications. SPARCML is built on the observation that many distributed machine learning algorithms either have naturally sparse communication patterns, or have updates which can be sparsified in a structured way for improved performance, without any convergence or accuracy loss. To exploit this insight, we design and implement a set of communication-efficient protocols for sparse input data, in conjunction with efficient machine learning algorithms which can leverage these primitives. Our communication protocols generalize standard collective operations, by allowing processes to contribute sparse input data vectors, of *heterogeneous sizes*. We call these operations *collectives on sparse streams*, and present efficient practical algorithms with strong theoretical bounds on their running time and communication cost. Our generic communication layer is enriched with additional features, such as support for non-blocking (asynchronous) operations and support for low-precision data representations. We validate our algorithmic results experimentally on a range of large-scale machine learning applications and target architectures, showing that we can leverage sparsity for order-of-magnitude runtime savings, compared to existing methods and frameworks.

## 1 Introduction and Motivation

A key enabling factor behind the tremendous progress in machine learning over the past decade has been *efficient system support*, in the form of efficient hardware—customized accelerators, e.g. [27], but also through *efficient software platforms* specialized for machine learning computation, e.g. [5, 10, 52]. Due to the sheer size of the datasets and models, production-scale machine learning workloads are usually *distributed* across multiple computing nodes, such as clusters of CPUs or GPUs in a datacenter environment. The arguably standard distribution strategy in machine learning is *data parallelism*, in which nodes partition the dataset, and maintain consistent copies of the set of model parameters by exchanging messages, either all-to-all or through a coordinator node, called a parameter server [32]. The high *bandwidth* and *latency* requirements of these workloads put pressure on system scalability. For example, when training a deep neural network such as AlexNet [30] through stochastic gradient descent (SGD), nodes perform *all-to-all* exchanges of their gradient updates upon every batch of examples: the message size per each node is  $> 200$  MB, exchanged every few milliseconds. This communication can easily become the system bottleneck.

<sup>1</sup>Stands for Sparse Communication layer for Machine Learning, to be read as *sparse ML*.

Given the significant practical impact of this problem, it is not surprising that significant effort has been invested into identifying scalable solutions. Virtually all major frameworks optimize for efficient communication [5, 10, 28, 41, 52], while GPU vendors are developing specific communication layers for this goal [37]. The research community proposed several techniques to reduce communication, such as *quantized updates* [7, 41], *asynchronous communication* [53], *structured sparsification* [6, 14, 43], or *large batch methods* [16, 51]. However, scaling machine learning applications remains a complex process, requiring non-trivial system insight.

**Conceptual Contribution.** We propose SPARCML, a scalable, general communication layer for machine learning. SPARCML builds on the idea that, to reduce communication and synchronization in machine learning applications, we should exploit the *relaxed consistency conditions* supported by such applications. This property, which we unite under the umbrella term of *stochastic delayed consistency*, allows individual nodes to compute with an *inconsistent* view of the set of model parameters, which can be corrupted with noise from various sources, such as communication quantization or asynchrony. The immediate system implication, which we exploit in SPARCML, is that the updates which nodes wish to communicate are either *naturally sparse*, or can be *sparsified in a principled manner* by exploiting stochastic delayed consistency properties. To illustrate, consider the SGD algorithm, a standard tool in machine learning. When performing regression on large models and sparse data—which is common in practice [48]—the gradient updates are *sparse*, since they only act on non-zero data entries. Moreover, when training large networks, one can *sparsify* gradient updates by magnitude, without loss of convergence [6, 7, 14, 43].

**Technical Contribution.** Our thesis is that *exploiting sparsity and compression* should be standard when scaling machine learning applications. Surprisingly, support for efficient sparse communication or compression is currently not available in standard communication libraries such as MPI [15], nor in specialized machine-learning communication libraries [37]. One possible reason is the fact that designing and implementing general sparse collective operations is non-trivial, as sparsity adds a whole new dimension to the already complex system trade-offs arising when implementing collective operations efficiently at scale [45].

We take on this challenge in SPARCML. Our implementation is efficient both in theory and in practice: for some workload parameters, it can be shown to be within constant factors of optimal in terms of bandwidth and latency cost. At the same time, our implementation achieves order-of-magnitude speedups versus highly optimized *dense collective* implementations, or over naive sparse implementations, both in synthetic tests and in real application scenarios. SPARCML has several additional features. It has efficient support for *reduced-precision collectives* and for *non-blocking* operations. For example, we can perform all-to-all sparse reductions for gradient exchange at 4bits of precision per coordinate, overlapping computation and communication.

**Target Applications and Architectures.** Our main target applications are two large-scale distributed machine learning tasks: training of state-of-the-art deep neural networks for image classification and machine translation, and large-scale regularized regression tasks, including linear and logistic regression, and Lasso.

Our target systems are multi-node computing clusters. We study two interesting scenarios: the first is *supercomputing*, where nodes are connected by a high-powered, extremely well optimized network. For this, we execute on CSCS Piz Daint, currently Europe’s most powerful supercomputer, which has a state-of-the-art interconnect. The second scenario we consider is *datacenters*, where the network is *relatively* slower, such as InfiniBand or Gigabit Ethernet. We target Amazon EC2 instances, as well as clusters with lower background traffic.

**Challenges.** The main algorithmic contribution behind our layer is a set of techniques for implementing collective communication operations, such as all-reduce sum, over a large number of nodes having input vectors that are *sparse*. The principal difficulty for designing and analyzing such algorithms lies in the

unknown overlap of non-zero indices, and hence the size of the reduced result. We provide an adaptive set of techniques which can systematically handle all cases and their trade-offs.

This algorithmic work is backed by optimizations and additional features. One ingredient is a custom data representation, which we call *sparse streams*, designed to perform *automatic switches* between sparse and dense vector representations. We integrate sparse streams into efficient communication operations, allowing for additional optimizations such as non-blocking transmission of data (non-blocking collectives), and we support data formats of low bitwidth.

**Experimental Results.** We validate SPARCML on a wide range of benchmarks, including real-world applications, as well as synthetic benchmarks aimed to confirm our theoretical analysis. Synthetic benchmarks show that SPARCML can bring order-of-magnitude speedups with respect to highly-optimized dense implementations, with little to no overhead in the dense case. Further, we incorporate SPARCML into two machine learning frameworks: Microsoft CNTK, to train state-of-the-art deep neural networks on large image classification and natural language processing datasets, and MPI-OPT, a framework we developed to perform large-scale data processing on clusters, which we compare against Spark MLlib [36]. In the supercomputing scenario, SPARCML reduces convergence time of a state-of-the-art network for digit recognition by  $3.65\times$ , and completes a large-scale URL classification task  $63\times$  faster than Spark MLlib. On cloud-grade networks, SPARCML speeds up the same neural network training task by  $19\times$ , and completes the URL classification task  $86\times$  faster than Spark MLlib. While these are some of the larger speedups we obtained, we note that SPARCML regularly yields non-trivial speedups on a wide variety of machine learning applications, in both scenarios. Thus, our experimental findings suggest that existing frameworks can still significantly speed up communication by leveraging sparsity and relaxed consistency guarantees.

## 2 Preliminaries

*Data-parallelism* is a classic distribution strategy for machine learning algorithms: nodes partition a large dataset, and each maintains its own copy of the model. Model copies are kept in sync by frequently exchanging local model updates between nodes, either via global averaging of updates or via a central coordinator called a parameter server [32]. For instance, in the context of the classic stochastic gradient descent (SGD) algorithm, each node has a dataset partition, and, in each *iteration*, it proceeds to process a randomly chosen set of samples (a *mini-batch*), and computes the model updates (gradients) locally using the classic SGD rule

$$\vec{x}_{t+1} = \vec{x}_t - \eta \nabla F(\vec{x}_t),$$

where  $\vec{x}_t$  is the value of the model at time  $t$ ,  $\eta$  is the learning rate, and  $\nabla F$  is the *stochastic* gradient of the current model with respect to the chosen set of samples. This ensures that nodes have a consistent version of the model at the beginning of each iteration. The trade-off is between the parallelism due to the fact that, given  $P$  nodes, we are processing  $P$  times more samples per iteration, and the overhead of maintaining a consistent model. Data parallelism is, arguably, the standard distribution method for large-scale parallel training of neural networks, e.g. [5, 52].

**Delayed Consistency.** When distributing to large node counts, the above trade-off can overwhelm the benefits of parallelism. Recent work, e.g. [34, 39], suggested that stochastic algorithms such as SGD or stochastic coordinate descent (SCD) can withstand a bounded amount of asynchrony, and still converge. More precisely, the following condition, which we call *delayed consistency*, is sufficient:

**Remark 2.1** (Bounded Staleness [39], [34]). *SGD and SCD can still converge under asynchronous iterations, under standard assumptions, as long as there exists a bound  $\tau$  on the delay between the time at which*

an update is generated at a node and the time at which the update is applied to the shared model.

The value of  $\tau$  can adversely impact the convergence rate of various algorithms [26, 34, 39]. Precisely characterizing these staleness-convergence-speed trade-offs is a topic of active research.

**Top- $k$  SGD.** Recent work proposes the following Top- $k$  communication-reduced SGD variant [6, 14]: each node communicates only the  $k$  largest (by magnitude) components of its gradient vector  $\nabla F(\vec{x}_t)$  instead of all values in the traditional method. Usually,  $k$  is fixed to represent some percentage of the components, which can be  $\leq 1\%$ , which enforces high gradient sparsity at each node. The smaller components are *accumulated*, and added to the gradient vector of the next iteration. Upon closer inspection, we notice that Top- $k$  SGD can be seen as special case of *asynchronous* SGD, where the components which do not make the top- $k$  at an iteration are *delayed* by being accumulated locally.

**Stochastic Consistency.** An orthogonal approach for reducing the communication cost of machine learning algorithms has been to *quantize* their updates, lowering the number of bits used to represent each value, e.g. [7, 12, 41, 49]. Such quantization techniques can also be shown to preserve correctness, as long as the quantization noise can be shown to be zero-mean [7]. We summarize this *stochastic consistency* condition as follows:

**Remark 2.2** (Stochastic Consistency [12], [7]). *SGD converges even quantized noisy updates, under standard assumptions [7, 12], as long as the noise in the quantized gradient updates is zero in expectation.*

**Relaxed Consistency in SPARCML.** Our framework is designed to leverage both delayed and stochastic consistency. We provide an efficient non-blocking implementation of top- $k$  SGD, as well as an implementation of state-of-the-art quantization methods [7].

**The MPI-OPT Framework.** MPI-OPT is a framework we developed to run distributed optimization algorithms such as SGD or SCD. It is written in native C++11, and can link external libraries such as SPARCML and MPI for communication. MPI-OPT implements parallel stochastic optimization algorithms, like gradient and coordinate descent, on multiple compute nodes communicating via any MPI library, with minimal overhead. It implements efficient distributed partitioning of any dataset converted in the predefined format using MPI-IO, data-parallel optimization on multiple compute nodes, with efficient multi-threading inside each node, parametrized learning rate adaptation strategies, as well as customizations to use SPARCML as the communication layer between nodes allowing for sparse, dense, synchronous, and asynchronous aggregation variations.

**The Microsoft Cognitive Toolkit (CNTK).** For large-scale neural network training, we slightly modified CNTK [52] v2.0 to use SPARCML as its communication layer. CNTK is a computational platform optimized for deep learning. The general principle behind CNTK is that neural network operations are described by a directed computation graph, in which leaf nodes represent input values or network parameters, and internal nodes represent matrix operations on their children. CNTK supports several popular network architectures, such as feed-forward DNNs, convolutional nets (CNNs), and recurrent networks (RNNs/LSTMs). To train such networks, CNTK implements stochastic gradient descent (SGD) with automatic differentiation. CNTK supports parallelization across multiple GPUs and servers, with efficient MPI-based communication. We exploit the structure of CNTK for computational tasks, and plug in SPARCML as the communication layer.

**Notation.** Throughout this paper, we use the following notation for input parameters:

Var	Description
$P$	Number of nodes
$N$	Problem dimension
$p_i$	Node $i$ , $1 \leq i \leq P$
$H_i$	Set of non-zero indices at $p_i$
$k$	Max number of non-zero (nnz) elements: $\max_i  H_i $
$\mathcal{K}$	Total nnz in global sum: $ \cup_{i=1}^P H_i $
$d$	Density of non-zero elements: $\frac{k}{N}$
$M$	Number of training samples
$B$	Mini-batch size per node

### 3 Data Representation: Sparse Streams

We now describe the data types used to store sparse and dense vectors, which we call *sparse streams*. Sparse streams allow for efficient computation and communication of the data. Our implementation is in C++11, and we follow this standard in our description. For simplicity, we focus on the case where the binary operation executed upon two or multiple streams is *summation*, but the same discussion would apply for other component-wise operations such as XOR or products.

Initially, we assume that each node is assigned a subset of non-zero elements from a universe of size  $N$ . Let  $H_i$  denote the set of non-zero elements given at node  $p_i$ . We assume that these sets are *sparse* with respect to  $N$ , i.e., that  $k = \max_i |H_i| \ll N$ . We further denote by  $d_i$  the density of each set given by  $d_i = \frac{|H_i|}{N}$  and define  $d = \max_i d_i = \frac{k}{N}$ . Define the total number of non-zero elements after having performed the reduction as

$$\mathcal{K} = |\cup_{i=1}^P H_i|.$$

For simplicity, we omit the unlikely possibility of cancellation of indices. We have that

$$k \leq \mathcal{K} \leq \min\{N, P \times k\}.$$

**Vector Representations.** We start from a standard representation, storing a sparse vector as a sequence of non-zero indices, together with the actual scalar values of each dimension. (While more complex representations exist, they come at the cost of additional computational effort.) The stream is stored in an array of consecutive index-value pairs. The datatype of the values yields the number of bits needed for every non-zero value. We either work with single (32 bits) or double precision floating point values (64 bits). We discuss lower precision support in Section 5. Given the dimension  $N$  of the vector, we need at least  $\lceil \log_2(N) \rceil$  bits for storing an index value.

**Auto-Switching to a Dense Format.** So far, the representation is straightforward. However, although we are interested in *sparse* problems, namely  $k \ll N$ , the size and non-zero index distribution of the input vectors can be such that the algorithm may not benefit from the sparse representation after some intermediate point in the summation process: if the density of the intermediate result vector reaches the universe size  $N$ , the sparse representation becomes wasteful.

Let *isize* be the number of bytes needed to represent a non-zero input value and *nnz* the number of non-zero elements. We further define  $c \geq \left\lceil \frac{\log_2(N)}{8} \right\rceil$  to be the number of bytes needed to store an index. Thus, the sparse format will transmit  $nnz(c + isize)$  bytes while the dense format transmits  $N \times isize$  bytes. Our sparse representation only reduces the communication volume if  $nnz \leq \delta = \frac{N \times isize}{(c + isize)}$ . Yet, this volume estimation does not capture the fact that summing sparse vectors is computationally more expensive than summing dense vectors. Thus, in practice,  $\delta$  should be even smaller, to reflect this trade-off.

Moreover, as inputs sum up, for large node counts  $P$ ,  $\mathcal{K}$  is almost certainly larger than  $\delta$ . To address this dynamic fill-in, we add an extra value to the beginning of each vector that indicates whether the vector is dense or sparse. In fact, when allocating memory for vectors of dimension  $N$ , we request  $N \times isize$  bytes. It is therefore never possible to store more than  $\delta$  sparse items. This threshold is used to automatically switch the representation from sparse to dense. SPARCML checks the format prior to every size-changing operation.

**Efficient Summation.** The key operation is summing two vectors  $u_1$  and  $u_2$ , which could be either sparse or dense. To implement this operation efficiently, we first distinguish the case when  $u_1$  and  $u_2$ 's indices come from any position between 1 and  $N$ , and can potentially overlap, from the case where the index sets are disjoint, which arises in instances where we partition the problem by dimension. This latter case is handled via concatenation.

If input indices can overlap, we distinguish the following cases, depending on whether inputs are sparse or dense. Denote by  $H_1$  and  $H_2$  the sets containing the sparse indices of non-neutral elements for the vectors  $u_1$  and  $u_2$ , respectively. If indices are overlapping, and both vectors are sparse, we first check whether the result might become dense. Theoretically, one needs to calculate the size of the union of non-zero indices  $|H_1 \cup H_2|$ . This is costly, and thus we only upper bound this result by  $|H_1| + |H_2|$ . (The tightness of this upper bound will depend on the underlying sparsity distribution, on which we make no prior assumptions.) If this value is bigger than  $\delta$ , we switch to a dense representation. If one of the inputs is dense, whereas the other is sparse, we iterate over all the index-value pairs stored in the sparse vector and set the value at the corresponding position in the dense vector. Finally, if both vectors are already dense, we simply perform a (vectorized) dense vector summation in either  $u_1$  or  $u_2$ , and do not allocate a new stream.

## 4 Efficient Collectives on Sparse Streams

We now define collective operations over a set of sparse vectors located at different nodes. We focus on the AllGather and AllReduce operations [17]. We support arbitrary coordinate-wise associative reduction operations for which a neutral-element can be defined,<sup>2</sup> and outline the algorithms for summation, for simplicity.

**Analytical Model.** We assume bidirectional, direct point-to-point communication between the nodes, and consider the classic Latency-Bandwidth ( $\alpha$ - $\beta$ ) cost model: the cost of sending a message of size  $L$  is  $T(L) = \alpha + \beta L$ , where both  $\alpha$ , the latency of a message transmission, and  $\beta$ , the transfer time per word, are constant.  $L$  represents the datum size in words or bytes. When sending sparse items, we denote  $\beta_s$  to be the transfer time per sparse index-value pair.  $\beta_d$  represents the transfer time per value, and is smaller than  $\beta_s$ .

Given this setting, the goal is to perform a collective operation over the elements present initially at every node. That is, each node should obtain the correct result locally, i.e., the element-wise sum over the  $N$  dimensions in the AllReduce case, while minimizing the total communication costs, measured in the  $\alpha$ - $\beta$  model.

**Assumptions.** We make the following assumptions, which we relax in our actual implementation:

1. each node contributes exactly  $k$  elements:  $\forall i : |H_i| = k$ ;
2.  $P$  is a power of 2,  $P > 4$ ; and
3.  $N$  is divisible by  $P$ .

---

<sup>2</sup>By neutral element we mean an element which does not change the result of the underlying operation, e.g. 0 for the sum operation.

## 4.1 Algorithms

Our main technical contribution consists of a set of algorithms to solve the sparse AllReduce problem efficiently under various common input scenarios. An *AllReduce* operation performs a global reduction operation on inputs located at *each node*, and distributes the result to *all nodes*. To implement AllReduce, we utilize the *AllGather* operation, an operation in which the data contributed by *each node* is gathered at *all nodes* (as opposed to a single root node). Both operations can be (inefficiently) implemented by performing a *Reduce* or *Gather* operation to a dedicated root, followed by a *Broadcast* to all the other nodes.

We emphasize that none of the following algorithms we propose requires knowledge about the amount of data contributed by each node, nor about the distribution of the non-zero indices. Nevertheless, depending on the size of the result  $\mathcal{K}$ , we will differentiate two types of instances: In *static sparse AllReduce* (SSAR),  $\mathcal{K}$  remains below  $\delta$ , such that we will never switch to a dense representation. In *dynamic sparse AllReduce* (DSAR), we have  $\mathcal{K} \geq \delta$ , we will start with a sparse and switch to a dense representation. We begin by discussing the most significant insights we gained to guide the reader through the section.

**Lower Bounds.** Sparse AllReduce with summation as its reduction operator is a generalization of both the well known AllReduce and AllGather collectives.

The key distinction is that each node has some subset of non-zero (non-neutral) elements assigned initially. We can obtain instances of the classical problems as follows: First, if none of these non-zero index sets  $H_i$  overlap, we obtain an *AllGather* instance on a subspace of  $N$ . The resulting set of non-zero indices therefore would have size  $\sum_{i=1}^P |H_i|$ . Second, the *AllReduce* problem is obtained if  $H_i = H_j$  for all nodes  $i$  and  $j$ , that is, the sets fully overlap and therefore the reduced result has  $|H_1|$  elements. If  $|H_i| = N$  for all  $i$ , this is just dense AllReduce. If  $|H_i| = k < N$ , we say that the problem is equivalent to a dense AllReduce on a subspace of dimension  $k$  rather than  $N$ . Those two instances are illustrated in Figure 1.

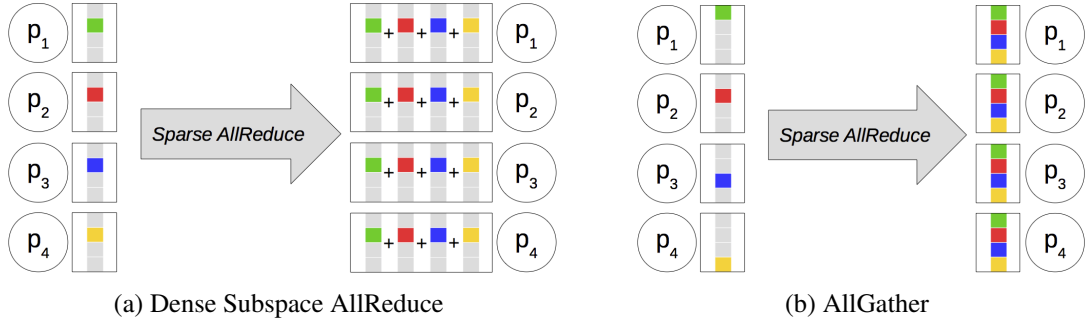


Figure 1: Specializations of sparse AllReduce. The gray items are neutral elements in the vectors.

Using this, we can lower bound on the runtime of every sparse AllReduce algorithm. Performing an AllGather operation, where each node contributes exactly  $k$  elements, results in a performance lower bound in this cost model [9]

$$\log_2(P)\alpha + (P-1)k\beta_d.$$

The lower bound in terms of the  $\alpha$ - $\beta$  model for the AllReduce collective with on vectors of size  $k$  is [9]

$$\log_2(P)\alpha + 2\frac{P-1}{P}k\beta_d,$$

if we assume that the computational cost of the reduction is minimal. Under this assumption, we obtain that:

**Lemma 4.1.** Any algorithm solving the SSAR problem needs time at least  $\log_2(P)\alpha + (P - 1)k\beta_d$  if  $\mathcal{K} = k \times P$ , and  $\log_2(P)\alpha + 2^{\frac{P-1}{P}}k\beta_d$  assuming  $\mathcal{K} = k$  and computation for reduction is perfectly parallelized.

*Proof.* The proof follows directly from the fact that the lower bound is known for both specializations of the sparse AllReduce problem: AllGather if  $\mathcal{K} = k \times P$  and subspace dense AllReduce if  $\mathcal{K} = k$ .  $\square$

We note that this lower bound is no longer relevant in a setting where the computational cost is non-trivial. In that case, a pipelining algorithm on a ring topology might perform better, e.g. [24]. We consider these algorithms in the experimental section.

**Latency-Bandwidth Trade-Offs.** A survey of algorithms for dense collectives including AllReduce [24] highlights the fact that single algorithms are not able to achieve both optimal latency and optimal bandwidth. Efficient MPI libraries such as MPICH or Open MPI therefore distinguish between small message sizes and long messages, and switch between algorithms as appropriate [45]. We adopt a similar approach.

**The Latency Dominated Case.** When the overall reduced data is small, latency dominates the bandwidth term. In this case, we will adopt a *recursive doubling* technique: in the first round, nodes that are a distance 1 apart exchange their data and perform a local sparse stream reduction. In the second round, nodes that are a distance 2 apart exchange their reduced data. Following this pattern, in the  $t$ -th round, nodes that are a distance  $2^{t-1}$  apart exchange all the previously reduced  $2^{t-1}k$  data items. This behavior is illustrated in Figure 2. We use the same algorithm for solving a sparse AllGather with non-overlapping indices, with the difference that local sparse summation can be executed more efficiently as described in the previous section. The recursive doubling technique can also be used for solving *dense* AllReduce and AllGather problems [24].

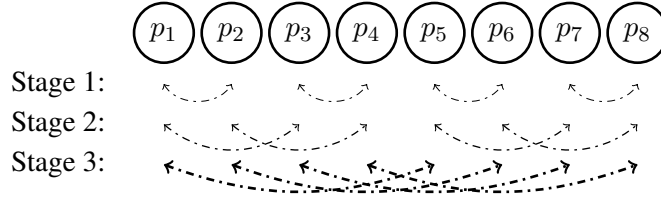


Figure 2: Static Sparse AllReduce: Recursive doubling - Increasing amount of sparse data in every stage

The resulting latency for the SSAR\_Recursive\_Double algorithm will be

$$L_1(P) = \log_2(P)\alpha,$$

as there are  $\log_2(P)$  stages. This is optimal, and data-independent. The runtime will lie in the range

$$L_1(P) + \log_2(P)k\beta_s \leq T_{ssar\_rec\_dbl} \leq L_1(P) + (P - 1)k\beta_s.$$

The lower bound is reached when the  $k$  indices fully overlap. Therefore, at every stage,  $k$  items need to be transmitted as the intermediate results maintain constant size. The upper bound is given when the indices do not overlap at all. Therefore, at stage  $t$ , the number of items transmitted is  $2^{t-1}k$ . Taking the sum, we get

$$\sum_{i=1}^{\log_2(P)} 2^{i-1}k = k \frac{2^{\log_2(P)} - 1}{2 - 1} = k(P - 1).$$



**The Bandwidth Dominated Case.** When the data is large, standard dense AllReduce implementations make use of Rabenseifner’s algorithm [38], which has two steps. The first is a ReduceScatter step, which partitions the result vector across nodes, assigning a partition to each node. This is implemented by a *recursive halving* technique [38]. In the second step, the reduced answers are gathered to all other nodes by calling a recursive doubling algorithm as described above. This algorithm has a total runtime of

$$T_{ar\_rab} = 2 \log_2(P) \alpha + 2 \frac{(P-1)}{P} k \beta_s,$$

which reaches the lower bound on the bandwidth term and is off by a factor 2 on the latency term.

We will start from a similar idea for Sparse AllReduce. We split the algorithm into two steps: In the first Split phase, we uniformly split the space dimension  $N$  into  $P$  partitions, and assign to each node the indices contained in the corresponding partition. We split each sparse vector at its node, and directly send each subrange of indices to the corresponding recipient, in a *sparse* format. This direct communication comes at a theoretical price of higher latency cost, but by using non-blocking send and receive calls, the computation and communication can be overlapped. Figure 3 depicts this communication pattern. Each node then reduces the data it received, and builds the result for its partition. In the second phase, the data has to be gathered to all other nodes. This sparse AllGather is executed by a recursive doubling algorithm with efficient sparse summation as described in the previous section.

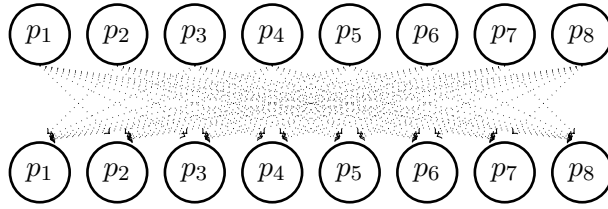


Figure 3: Direct Sparse Send AllToAll

Obtaining runtime bounds for SSAR.Split.AllGather is slightly more involved. The Split part takes time

$$(P-1)\alpha + 0\beta_s \leq T_{split} \leq (P-1)\alpha + k\beta_s.$$

Notice that both extremes imply that each node has  $k$  items for the sparse AllGather, and thus  $\mathcal{K} = k \times P$  is reached. For this second step in the algorithm to be optimal, every node must have an intermediate result of size  $\frac{k}{P}$ , as we want the final result to have a size  $\mathcal{K} = k$  and the communication to be equally distributed.

For every node to have an intermediate result of the desired size, we know that each node has to send at least  $\frac{P-1}{P}k$  items to other nodes. Otherwise, if every node has exactly  $k$  items, we reach the upper bound for the result size of  $\mathcal{K} = k \times P$ . So we get

$$L_1(P) + \frac{P-1}{P}k\beta_s \leq T_{sparse\_ag} \leq L_1(P) + (P-1)k\beta_s.$$

The algorithm latency is again data-independent:

$$L_2(P) = (P-1)\alpha + L_1(P).$$

Combining these terms yields

$$L_2(P) + 2 \frac{P-1}{P}k\beta_s \leq T_{ssar\_split\_ag} \leq L_2(P) + Pk\beta_s.$$

**The Dynamic Case: Switching to Dense.** The discussion so far focused on the case where maintaining a sparse representation is efficient. However, as we gather data, the size of the result  $\mathcal{K}$  might become larger than the sparsity-efficient threshold  $\delta$ , in which case we switch to a dense representation. This is the *dynamic* version of the problem (DSAR). In this case, the following lower bound on the efficiency of the algorithm will hold, whose proof is left for the full version of our paper.

**Theorem 4.2.** *Any algorithm solving the DSAR problem needs at least time  $\log_2(P)\alpha + N\beta_d$ .*

*Proof.* As every node needs to communicate to every other node directly or indirectly, there is at least one node communicating to  $\log_2(P)$  other nodes. For the bandwidth term, every node needs to send its  $k$  items. As the size of the resulting vector increases such that it eventually has no sparse representation anymore, each node has to receive or send at least  $N - k$  items. Taking the sum results in  $N\beta$  on the bandwidth term.  $\square$

This implies the following result, which says that, in this dynamic case, one can only hope to get a speedup of at most factor 2 compared to an optimal *dense* AllReduce algorithm, when focusing on the bandwidth term.

**Lemma 4.3.** *The bandwidth required by any algorithm for the DSAR problem is at least  $\frac{1}{2}$  that of a bandwidth-optimal dense AllReduce algorithm.*

*Proof.* Notice that the dense AllReduce with  $k = N$  has a lower bound of  $2\frac{P-1}{P}N\beta_d$  on the bandwidth if computation is equally distributed. From Theorem 4.2 we know that every DSAR algorithm has a minimum bandwidth term of  $N\beta_d$ , which is obviously bigger than  $\frac{P-1}{P}N\beta_d$  for any  $P$ .  $\square$

Based on these insights, our solution for DSAR adapts the previous two-stage algorithm to exploit the fact that every reduced split will become dense. DSAR\_Split\_AllGather hence receives the data in a sparse format from all the other nodes in the first phase, then switches the representation and performs a dense AllGather in the second stage. Here, we can leverage existing implementations, which are highly optimized to perform this second step with dense data. The Split part of DSAR\_Split\_AllGather remains identical to the first phase of SSAR\_Split\_AllGather. As we force every split result to become dense, there is no data dependency after having performed this first part of the algorithm. Based on the known times needed by a dense AllGather, we derive the running time for our algorithm given both extremes. The latency is again  $L_2(P)$ . Combined, we get

$$L_2(P) + \frac{P-1}{P}N\beta_d \leq T_{dsar\_split\_ag}$$

and

$$T_{dsar\_split\_ag} \leq L_2(P) + k\beta_s + \frac{P-1}{P}N\beta_d.$$

## 5 Artifact and Additional Features

**Interface and Code.** The SPARCML library provides an interface that is similar to that of standard MPI calls, with the caveat that the data representation is assumed to be a sparse stream. Given this, the changes needed to port MPI-enabled code to exploit sparsity through SPARCML are minor. The library implementation consists of around 2,000 lines of native C++11. (Counting infrastructure such as benchmarks and tests raises the line count by an order of magnitude.) Adding SPARCML to CNTK required changing around 100 lines of code.

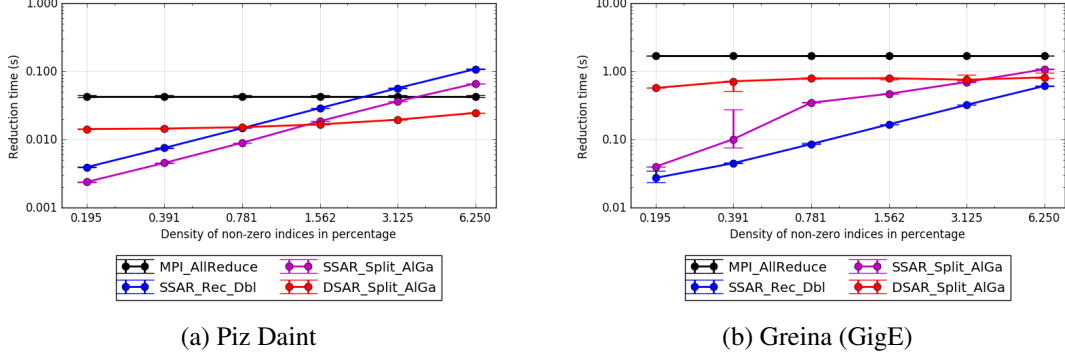


Figure 4: Data density versus reduction time for various algorithms, dimension  $N = 16M$  and  $P = 8$  nodes.

**Non-Blocking Operations.** We also implement the previous algorithms in a *non-blocking* way. Specifically, we allow a thread to trigger a collective operation, such as AllReduce, in a nonblocking way. This enables the thread to proceed with local computations while the operation is performed in the background. For deep networks, we can nicely overlap communication and computation during the gradient aggregation phase by calling the aggregation per layer in a non-blocking fashion. As of MPI-3, implementations support non-blocking collective operations. However, rendering a custom operation implementation non-blocking is not entirely straightforward [22, 23] and needs to consider subtle message progression issues [21].

**Low-Precision Support.** To further reduce bandwidth cost, SPARCML supports lower-precision data representation for the updates, using QSGD quantization [7], which provably preserves convergence. In brief, in this variant, each vector is split into *buckets* of size  $B$  (in the order of 1,024), and each bucket is quantized independently and stochastically. Thus, each bucket corresponds to  $B$  *low-precision data items*, e.g., 4-bit integers, packed to reduce space, and a full-precision *scaling factor*. We focus on low-precision to reduce the bandwidth cost of the *dense* case. Hence, in the low-precision implementation, we either omit sparsity, or we use the low-precision data representation for the second part of the DSAR\_Split\_AllGather algorithm, where the data becomes dense. Interestingly, by making use of this functionality, we are able to achieve speedup higher than  $2\times$  compared to a fully dense AllReduce, which was the upper bound when working with full-precision data.

## 6 Experiments

**Setup.** We now validate SPARCML on real world applications and synthetic experiments. Complete code and experimental logs are available at [3]. Our experiments target the supercomputing and cloud computing scenarios. For the first setting, we execute on the CSCS Piz Daint supercomputer [4], with Cray CX50 nodes, each of which has a 12 cores HT-enabled Intel Xeon E5-2690 v3 CPU with 4GB RAM and an NVIDIA Tesla P100 16GB GPU. Piz Daint is currently the most powerful supercomputer in Europe (3rd in the world), and has a high-performance Cray Aries interconnect with a Dragonfly network topology.

For the second setting, we use Amazon EC2 instances with a similar CPU configuration, but a relatively older NVIDIA K80 GPU, and are connected through Gigabit Ethernet. We will also perform additional tests on a cluster called CSCS Greina, with CX50 nodes and an InfiniBand FDR or selectively Gigabit Ethernet interconnect. This serves as a small but high-performance research cluster, without background traffic.

We measure the MPI point-to-point bandwidth of each system with standard ping-pong techniques. Piz Daint reaches a bandwidth of up to 50 Gbps, Greina achieves roughly 1 Gbps using the Gigabit Ethernet

Name	# Classes	# of samples	Dimension
URL [35]	2	2 396 130	3 231 961
Webspam [48]	2	350 000	16 609 143
ImageNet [40]	1000	1,2 million	256x256x3
MNIST [31]	10	60 000	28x28
CIFAR-10 [29]	10	60 000	32x32x3
ATIS [19]	128	4 978 <i>s</i> / 56 590 <i>w</i>	-
Hansards [1]	-	948K <i>s</i> / 15 657K <i>w</i>	-

Table 1: Real World Application Datasets. *s* stands for sentences (or pairs) and *w* for words.

(GigE) interconnect, and around 40 Gbps when using the InfiniBand FDR (IB) interconnection. The Amazon cluster bandwidth saturates at 4 Gbps. In all our experiments, the baseline will be the MPI AllReduce implementation on the fully dense vectors. On EC2, we make use of the default Open MPI installation optimized by Amazon. On Piz Daint, we compare against the custom Cray-MPICH installation, highly optimized by Cray, which provides an extremely high-performance baseline. Since most of our problems have dimension  $N > 65K$  (maximum unsigned short), we fix the datatype for storing an index to an unsigned integer.

**Micro-Benchmarks.** We begin by validating our theoretical analysis on synthetic data, on the Piz Daint and Greina (GigE) clusters. We vary the data dimension  $N$ , and the data density  $d$ , as well as the number of nodes  $P$ . Based on the defined density,  $k$  indices out of  $N$  are selected uniformly at random at each node, and are assigned a random value. We run our sparse AllReduce algorithms in order to validate both correctness and the relative ordering of the derived analytical bounds. The choice of parameters to generate the data, is within reasonable ranges, seen in real world datasets. Graphs are in a log-log scale. As execution times are non-deterministic, we conduct five experiments with newly generated data, while running each one for ten times. Based on those 50 resulting runtime values, we state the 25 and 75 percentage quantiles.

Following the theoretical analysis of the proposed algorithms, we expect SSAR\_Recursive\_Double to perform best for a small amount of data, when latency dominates over the bandwidth term. At higher node count  $P$ , data becomes larger, which leads to less improvement of the algorithm SSAR\_Recursive\_Double at the same number of non-zero entries over the other variants. Furthermore, the algorithm SSAR\_Split\_AllGather dominates over the DSAR\_Split\_AllGather variant as long as the number of non-zero indices is relatively low compared to the overall reduced size. To show the impact of the network on performance, we run identical tests on both Piz Daint and Greina (GigE) in Figure 4. The results confirm our analysis.

We have also compared our approaches against the ring-based MPI dense all-reduce (not shown). On the a fast Aries network and with a relatively small number of nodes, the ring-based algorithm is faster by a  $< 30\%$  margin. Our sparsity-based approach is consistently superior on the Gigabit Ethernet interconnect.

**Large-Scale Regression Instances.** We use MPI-OPT to train linear classifiers (Logistic Regression, SVM) on large-scale classification datasets using SGD and SCD. The goal of this benchmark is to examine the runtime improvements by just exploiting the sparsity inherently present in the datasets and algorithms. Hence, we do not apply any relaxed consistency techniques.

The datasets are specified in Table 1. We make use of the binary classification datasets URL and Webspam. For SGD, the samples (and hence, gradients) have high sparsity since the features are trigrams: while many such combinations exist, an item, e.g., a sentence, can only have a very limited set of them present. This is extremely common in text-based datasets. Since communication is lossless, convergence is

System	Dataset	Model	# of nodes	Baseline Time (s)	Algorithm	Time (s)	Speedup
Piz Daint	Webspam	LR	32	2.40 (2.16)	SSAR_Recursive_Double	0.68 (0.35)	<b>3.53 (6.17)</b>
		LR	128	0.64 (0.58)		0.18 (0.10)	<b>3.56 (5.80)</b>
		SVM	32	1.62 (1.42)		0.65 (0.44)	<b>2.49 (3.23)</b>
Piz Daint	URL	LR	32	2.64 (2.58)	SSAR_Recursive_Double	0.75 (0.70)	<b>3.52 (3.69)</b>
		LR	128	0.57 (0.56)		0.29 (0.28)	<b>1.97 (2.00)</b>
		SVM	32	1.98 (1.93)		0.56 (0.53)	<b>3.54 (3.64)</b>
Piz Daint	Webspam	LR	8	4.67 (3.79)	SSAR_Split_AllGather	2.56 (1.58)	<b>1.82 (2.40)</b>
	URL	LR	8	3.77 (3.53)		2.09 (1.50)	<b>1.80 (2.35)</b>
Greina (IB)	Webspam	LR	8	6.52 (4.67)	SSAR_Split_AllGather	3.63 (1.90)	<b>1.80 (2.46)</b>
	URL	LR	8	8.14 (4.47)		6.11 (2.49)	<b>1.33 (1.80)</b>
Greina (GigE)	Webspam	LR	8	76.80 (75.95)	SSAR_Split_AllGather	3.79 (2.95)	<b>20.26 (25.75)</b>
	URL	LR	8	104.50 (100.46)		8.26 (4.22)	<b>12.65 (23.81)</b>

Table 2: Distributed optimization using MPI-OPT. The times are averages for a full dataset pass, with the communication part in brackets. Speedup versus dense MPI is shown end-to-end, with communication speedup in brackets.

preserved and we only report speedup of the communication and overall training time. We run SGD with large batches ( $1,000 \times P$ ) for various combinations. The achieved speed of MPI-OPT with the best sparse reduction algorithm is reported in Table 2.

Additionally, we run SCD incorporated in MPI-OPT following the distributed implementation of [34]. We run the optimization on the logistic regression loss function for the URL dataset distributed on 8 nodes of Piz Daint to achieve identical convergence compared to SGD. Every node contributes 100 indices after every iteration. As the values calculated by each node lie in different slices of the entire model vector, we compare the runtime of a sparse AllGather from SPARCML to its dense counterpart. MPI-OPT with a dense AllGather has an average epoch time of 49 seconds, with 24 seconds dedicated to communication. The sparse AllGather executes a dataset pass (epoch) in 26 seconds on average, with 4.5 seconds spent in the communication layer. This implies an overall speedup of factor  $1.8\times$ , due to a  $5.3\times$  speedup in communication time.

**Comparison with Apache Spark.** We compare MPI-OPT with Apache Spark v1.6, which is officially supported by CSCS [2]. Comparison is performed on the same datasets, with the note that Spark uses its own communication layer, and does not exploit sparsity. On the Piz Daint supercomputer, using 8 nodes, MPI-OPT with SPARCML reduces the time to convergence on the URL dataset by  $63\times$ . This is largely due to the reduction in communication time, which we measure to be of  $185\times$ . Concretely, the average epoch time is reduced from 378 seconds, with 319 seconds spent for communication, to an average of 6 seconds per epoch, whereof 1.7 seconds represent the communication time. Compared to Spark, MPI-OPT with the standard Cray-optimized *dense* AllReduce has a  $31\times$  speedup to convergence, due to a  $43\times$  speedup in communication time. An epoch is executed in 13 seconds on average, with 8.6 seconds spent on communication.

We further investigated these speedups on an 8-node research cluster with a Gigabit Ethernet interconnection. This mimics an Amazon EC2 scenario, with no background traffic. Using MPI-OPT, the average training time per epoch drops from 1,274 seconds (Spark) to 14 seconds, representing a speedup of  $86\times$ . On the communication part, the time per epoch drops from 1,042 seconds to 6 seconds. The communication time and overall speedup of a *dense* AllReduce over Spark’s communication layer are both of factor  $12\times$ .

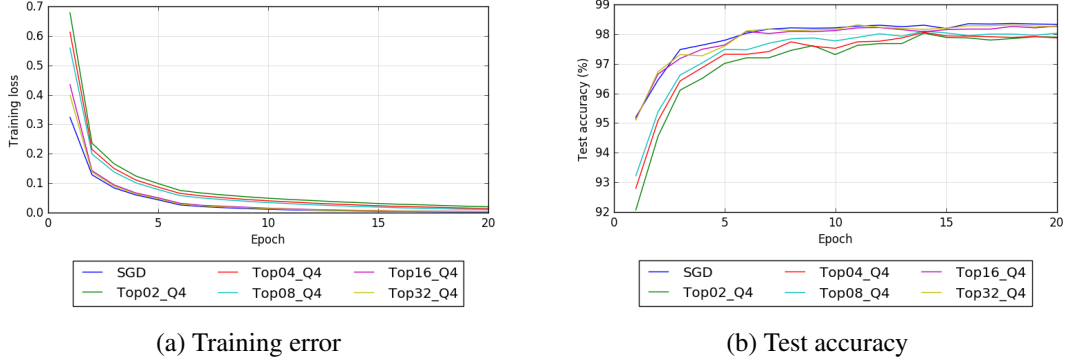


Figure 5: MNIST convergence for various algorithm variants.

System	Dataset	Model	# of nodes	Baseline Time (m)	Algorithm	Time (m)	Speedup
Piz Daint	ImageNet	VGG19	8	61.23 (32.72)	Q4	39.49 (9.87)	<b>1.55 (3.31)</b>
Piz Daint	ImageNet	AlexNet	16	33.69 (26.82)	Q4	25.83 (19.77)	<b>1.30 (1.36)</b>
Piz Daint EC2	MNIST	MLP	8	1.78 (1.65)	Top16_Q4	0.49 (0.36)	<b>3.65 (4.53)</b>
				24.41 (24.18)	Top16_Q4	1.28 (1.05)	<b>19.12 (22.97)</b>

Table 3: Training neural networks using CNTK. Times represent average time in minutes for a full dataset pass, with the communication part in brackets. Speedup versus dense MPI is shown end-to-end, with communication in brackets.

## 6.1 Training Deep Neural Networks

We turn to training state-of-the-art deep neural networks in CNTK, using SPARCML. To exploit sparsity, we use the Top- $k$  SGD algorithm [6, 14, 43], combined with low-precision support [7]. We execute three types of tasks: *image classification* on the ImageNet, CIFAR-10 and MNIST datasets, *natural language understanding* on the ATIS corpus, and *machine translation* on the Hansards dataset. The datasets are specified in Table 1. For vision, we run experiments on the following deep networks: AlexNet [30], VGG [42], ResNet [18] and multi layer perceptrons (MLPs) with fixed numbers, two hidden layers of dimension 4,096 each. For natural language understanding and machine translation we use an encoder-decoder LSTM [20].

For this, we have interfaced SPARCML into CNTK. We make use of the non-blocking version of the sparse AllReduce algorithm when working with full precision. Selecting the biggest elements in magnitude is implemented efficiently in CNTK by building “buckets” with 512 consecutive tensor elements each (reshaping the tensor when necessary), and implementing a fast randomized Top- $k$  algorithm. This enables fast computation on GPUs and allows for significant speedup over the fully dense AllReduce variant. We omit selecting the Top- $k$  values and quantization of gradient matrices with small ( $< 10K$ ) size, as the computational effort of either operation exceeds the overall reduction in time given by reducing the amount of data transmitted. We use standard batch sizes and default hyper-parameters for 32-bit full accuracy convergence in all our experiments. These values are given in the open-source CNTK 2.0 repository. We have also experimented with supporting the 1-bit SGD quantization algorithm [41] and asynchronous (ASGD) training. Unfortunately, for 1-bit SGD, the way tensors in convolutional layers are handled, can actually *increase* overall training time with respect to the MPI baseline in networks with lots of convolutions, such as ResNet. ASGD training requires extremely careful hyperparameter tuning to maintain accuracy [53], and

is therefore left for future work.

**Accuracy.** As stated by in the preliminaries, working with delayed consistency might affect convergence speed, as the sparsity-accuracy trade-offs are not yet fully understood. On the one hand, similarly to [6, 14], experiments on MNIST with multi-layer perceptrons (MLPs), CIFAR-10 on ResNet110 and ATIS with LSTMs show almost identical convergence speed even when forcing high sparsity. When combining Top- $k$  and quantization with selecting  $k = 16$  out of every bucket of size 512, and quantizing the dense values using 4-bit precision we achieve a top-1 accuracy 0.48% below the full precision, and even increase top-5 accuracy by 0.12% on the CIFAR-10 dataset. For MNIST, the training error fluctuates by less than 0.1%. The impact of different  $k$  values for this dataset is visible in Figure 5.

We note that training convolutional networks with Top- $k$  SGD to full convergence on ImageNet requires additional hyper-parameter tuning. This is because, as discussed in Section 2, Top- $k$  SGD is essentially an instance of asynchronous SGD. Optimal convergence can be achieved for this algorithm, but requires careful tuning of the momentum hyper-parameter [53]. For this reason, we only employ low precision (quantization to 4bits) in SPARCML when executing our benchmarks on the ImageNet dataset, in which case we achieve convergence.

**Speedup.** In Table 3, we provide the overall epoch time, and the time spent communicating for various image classification models and training algorithms. *Q4* denotes the 4-bit quantized version on all the gradient values. *Top16* represents the version where the biggest 16 values in magnitude are selected per buckets of 512 consecutive tensor values. We give the speedup for those values of  $k$  in combination with 4-bit low precision on the dense values in the second part of the algorithm based on the previous convergence tests. We emphasize the difference in performance between the supercomputer network on Piz Daint and the Amazon EC2 network, due to the faster network, in the MNIST experiment.

Further, we conduct neural machine translation experiments on the Hansards dataset using an encoder-decoder network with two LSTM cells each. By selecting the 4 largest values out of each 512 consecutive elements, we achieve an overall speedup of factor  $1.54\times$  with slightly worse convergence speed on 8 nodes of Piz Daint. For natural language understanding problems on the ATIS dataset, using an identical network, we obtain a overall speedup of  $2.6\times$  to full convergence, compared to a full dense AllReduce, by only selecting the biggest 2 values out of buckets of 512 elements.

## 6.2 Related Work

There has recently been a tremendous surge of interest in distributed machine learning [5, 10, 52] and communication-reduction methods [13]; due to space constraints, a complete survey of such frameworks and techniques is infeasible. We will focus on approaches and techniques that are closely related.

**Reduced Communication Techniques.** Seide et al. [41] was among the first to propose quantization to reduce the bandwidth and latency costs of training deep networks. More recently, Alistarh et al. [7] introduced a theoretically-justified distributed SGD variant called Quantized SGD (QSGD), which allows the user to trade off compression and convergence rate. We implement QSGD as a default quantization method.

Dryden et al. [14] and Aji and Heafield [6] considered an alternative approach to communication reduction for data-parallel SGD, *sparsifying* the gradient updates by only applying the top- $k$  components, taken at every node, in every iteration, for  $k$  corresponding to  $< 1\%$  of the update size. Since then, other references [33, 43] explored this space, showing that extremely high gradient sparsity ( $< 0.1\%$ ) can be supported by convolutional and recurrent networks with preserved accuracy, although maintaining accuracy requires very careful tuning of hyperparameters.

Our paper complements this line of the work by providing highly efficient sparsity support, with consis-

tent runtime gains in large-scale settings, both for supercomputing and cloud computing scenarios.

**Lossless Methods.** One loss-less communication-reduction technique is *factorization* [11, 50], effective in deep neural networks with large fully-connected layers. This is less applicable in networks with large convolutional layers, which is the case for many modern architectures [18, 44]. Poseidon / Petuum [50] is a complete distributed machine learning framework built on the idea of reducing communication through factorization. A second such method is executing *extremely large batches*, thus hiding the cost of communication behind larger computation [16, 51]. Although promising, large-batch methods require careful per-instance parameter tuning, and do not eliminate communication costs.

**Communication Frameworks.** Several frameworks have been proposed for reducing communication cost of distributed machine learning. One popular example is NVIDIA’s NCCL framework [37], which significantly reduces communication cost when the nodes are NVIDIA GPUs, and the proprietary NVLINK interconnect is available, which is not the case in multi-node settings, such as supercomputing. Further, NCCL currently only implements a very restricted set of reduction operations. In addition, there is a non-trivial number of frameworks customized to specific application scenarios, such as the Livermore Big Artificial Neural Network Toolkit (LBANN) [47] or S-Caffe [8]. While very efficient in specific instances, these frameworks do not usually leverage reduced-communication techniques, or sparsity.

**Sparse Reduction.** Efficient MPI support for reductions over sparse input vectors was considered by [25], and from the algorithmic perspective in [46]. The first reference proposes and evaluates a direct runlength encoding approach; we significantly extend this approach in the current work, including the observation that data might become dense during the reduction process, and that an efficient and flexible data representation must be provided in this case. Kylix [54] considers sparse many-to-many reductions in the context of computation over large scale distributed graph data on community clusters. However, Kylix assumes knowledge of the data distribution, and performs multiple passes over the reduction, which make it not applicable to our scenario. Dryden et al. [14] implement a sparse variant of the classical AllReduce algorithm via a pairwise reduce-scatter followed by a ring-based AllGather. The amount of data is kept constant at every stage of their algorithm by re-selecting the top  $k$  values and postponing the other received values, which is different than AllReduce.

## 7 Conclusions and Further Work

We have described and analyzed SPARCML, a high-performance communication framework that allows the user to leverage sparse and low-precision communication in the context of machine learning algorithms. Due to its compatibility with MPI semantics, SPARCML should integrate easily into existing computational frameworks, and can provide order-of-magnitude speedups in several real-world applications.

One aspect of SPARCML which we aim to further develop is compatibility with other frameworks and fundamental machine learning algorithms. In particular, we aim to provide support for Google TensorFlow via its recently-added MPI support [5], Apache MXNet [10], and direct support for Apache Spark MLlib [36]. We also aim to investigate other distributed machine learning algorithms which can benefit from sparsity, such as distributed principal component analysis (PCA) and k-means clustering (Lloyd’s algorithm). We believe that the simple but effective sparsity schemes we explore in our work will play a deciding role in avoiding communications in future distributed machine learning systems.



## References

- [1] Aligned Hansards of the 36th Parliament of Canada. <https://www.isi.edu/natural-language/download/hansard/>. Accessed: 2017-10-25.
- [2] Apache Spark on the CSCS Cluster. [https://user.cscs.ch/scientific\\_computing/supported\\_applications/spark/](https://user.cscs.ch/scientific_computing/supported_applications/spark/). Accessed: 2018-1-25.
- [3] Complete implementation of SparCML and benchmarks. <https://gitlab.com/crenggli/MPIML>. Accessed: 2018-2-6.
- [4] The CSCS Piz Daint supercomputer. [http://www.cscs.ch/computers/piz\\_daint](http://www.cscs.ch/computers/piz_daint). Accessed: 2018-1-25.
- [5] ABADI, M., BARHAM, P., CHEN, J., CHEN, Z., DAVIS, A., DEAN, J., DEVIN, M., GHEMAWAT, S., IRVING, G., ISARD, M., ET AL. Tensorflow: A system for large-scale machine learning. In *OSDI* (2016), vol. 16, pp. 265–283.
- [6] AJI, A. F., AND HEAFIELD, K. Sparse communication for distributed gradient descent. *arXiv preprint arXiv:1704.05021* (2017).
- [7] ALISTARH, D., GRUBIC, D., LI, J., TOMIOKA, R., AND VOJNOVIC, M. QSGD: Randomized quantization for communication-efficient stochastic gradient descent. In *Proceedings of NIPS 2017* (2017).
- [8] AWAN, A. A., HAMIDOUCHE, K., HASHMI, J. M., AND PANDA, D. K. S-caffe: Co-designing mpi runtimes and caffe for scalable deep learning on modern gpu clusters. In *Proceedings of the 22nd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming* (2017), ACM, pp. 193–205.
- [9] CHAN, E., HEIMLICH, M., PURKAYASTHA, A., AND VAN DE GEIJN, R. Collective communication: theory, practice, and experience. *Concurrency and Computation: Practice and Experience* 19, 13 (2007), 1749–1783.
- [10] 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. *arXiv preprint arXiv:1512.01274* (2015).
- [11] CHILIMBI, T. M., SUZUE, Y., APACIBLE, J., AND KALYANARAMAN, K. Project adam: Building an efficient and scalable deep learning training system. In *OSDI* (2014), vol. 14, pp. 571–582.
- [12] DE SA, C., ZHANG, C., OLUKOTUN, K., AND RÉ, C. Taming the wild: A unified analysis of Hogwild. *Style Algorithms. In NIPS* (2015).
- [13] DEMMEL, J. Tutorial on communication-avoiding algorithms. [http://people.eecs.berkeley.edu/~demmel/SC16\\_tutorial\\_final](http://people.eecs.berkeley.edu/~demmel/SC16_tutorial_final).
- [14] DRYDEN, N., JACOBS, S. A., MOON, T., AND VAN ESSEN, B. Communication quantization for data-parallel training of deep neural networks. In *Proceedings of the Workshop on Machine Learning in High Performance Computing Environments* (2016), IEEE Press, pp. 1–8.
- [15] FORUM, M. P. I. MPI: A Message-Passing Interface Standard Version 3.0, 09 2012.
- [16] GOYAL, P., DOLLÁR, P., GIRSHICK, R., NOORDHUIS, P., WESOŁOWSKI, L., KYROLA, A., TULLOCH, A., JIA, Y., AND HE, K. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677* (2017).
- [17] GROPP, W., HOEFLER, T., THAKUR, R., AND LUSK, E. *Using Advanced MPI: Modern Features of the Message-Passing Interface*. MIT Press, Nov. 2014.
- [18] HE, K., ZHANG, X., REN, S., AND SUN, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778.
- [19] HEMPHILL, C. T., GODFREY, J. J., DODDINGTON, G. R., ET AL. The atis spoken language systems pilot corpus. In *Proceedings of the DARPA speech and natural language workshop* (1990), pp. 96–101.
- [20] HOCHREITER, S., AND SCHMIDHUBER, J. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [21] HOEFLER, T., AND LUMSDAINE, A. Message Progression in Parallel Computing - To Thread or not to Thread? In *Proceedings of the 2008 IEEE International Conference on Cluster Computing* (Oct. 2008), IEEE Computer Society.
- [22] HOEFLER, T., LUMSDAINE, A., AND REHM, W. Implementation and performance analysis of non-blocking collective operations for mpi. In *Supercomputing, 2007. SC'07. Proceedings of the 2007 ACM/IEEE Conference on* (2007), IEEE, pp. 1–10.
- [23] HOEFLER, T., LUMSDAINE, A., AND REHM, W. Implementation and Performance Analysis of Non-Blocking Collective Operations for MPI. In *Proceedings of the 2007 International Conference on High Performance Computing, Networking, Storage and Analysis, SC07* (Nov. 2007), IEEE Computer Society/ACM.

- [24] HOEFLER, T., AND MOOR, D. Energy, Memory, and Runtime Tradeoffs for Implementing Collective Communication Operations. *Journal of Supercomputing Frontiers and Innovations* 1, 2 (Oct. 2014), 58–75.
- [25] HOFMANN, M., AND RÜNGER, G. Mpi reduction operations for sparse floating-point data. *Lecture Notes in Computer Science* 5205 (2008), 94.
- [26] JIANG, J., CUI, B., ZHANG, C., AND YU, L. Heterogeneity-aware distributed parameter servers. In *Proceedings of the 2017 ACM International Conference on Management of Data* (New York, NY, USA, 2017), SIGMOD '17, ACM, pp. 463–478.
- [27] JOUPPI, N. P., YOUNG, C., PATIL, N., PATTERSON, D., AGRAWAL, G., BAJWA, R., BATES, S., BHATIA, S., BODEN, N., BORCHERS, A., ET AL. In-datacenter performance analysis of a tensor processing unit. *arXiv preprint arXiv:1704.04760* (2017).
- [28] KETKAR, N. Introduction to pytorch. In *Deep Learning with Python*. Springer, 2017, pp. 195–208.
- [29] KRIZHEVSKY, A., AND HINTON, G. Learning multiple layers of features from tiny images.
- [30] KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (2012), pp. 1097–1105.
- [31] LECUN, Y., AND CORTES, C. MNIST handwritten digit database.
- [32] 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 *OSDI* (2014), vol. 1, p. 3.
- [33] LIN, Y., HAN, S., MAO, H., WANG, Y., AND DALLY, W. J. Deep gradient compression: Reducing the communication bandwidth for distributed training. *arXiv preprint arXiv:1712.01887* (2017).
- [34] LIU, J., AND WRIGHT, S. J. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. *SIAM Journal on Optimization* 25, 1 (2015), 351–376.
- [35] MA, J., SAUL, L. K., SAVAGE, S., AND VOELKER, G. M. Identifying suspicious urls: an application of large-scale online learning. In *Proceedings of the 26th annual international conference on machine learning* (2009), ACM, pp. 681–688.
- [36] MENG, X., BRADLEY, J., YAVUZ, B., SPARKS, E., VENKATARAMAN, S., LIU, D., FREEMAN, J., TSAI, D., AMDE, M., OWEN, S., ET AL. Mllib: Machine learning in apache spark. *The Journal of Machine Learning Research* 17, 1 (2016), 1235–1241.
- [37] NVIDIA. The nvidia collective communications library (nccl). <https://developer.nvidia.com/nccl> (2016).
- [38] RABENSEIFNER, R. Optimization of collective reduction operations. In *International Conference on Computational Science* (2004), Springer, pp. 1–9.
- [39] RECHT, B., RE, C., WRIGHT, S., AND NIU, F. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in neural information processing systems* (2011), pp. 693–701.
- [40] RUSSAKOVSKY, O., DENG, J., SU, H., KRAUSE, J., SATHEESH, S., MA, S., HUANG, Z., KARPATHY, A., KHOSLA, A., BERNSTEIN, M., ET AL. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.
- [41] SEIDE, F., FU, H., DROPPA, J., LI, G., AND YU, D. 1-bit Stochastic Gradient Descent and its Application to Data-parallel Distributed Training of Speech DNNs. In *Fifteenth Annual Conference of the International Speech Communication Association* (2014).
- [42] SIMONYAN, K., AND ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [43] SUN, X., REN, X., MA, S., AND WANG, H. meprop: Sparsified back propagation for accelerated deep learning with reduced overfitting. *arXiv preprint arXiv:1706.06197* (2017).
- [44] SZEGEDY, C., IOFFE, S., VANHOUCKE, V., AND ALEMI, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI* (2017), pp. 4278–4284.
- [45] THAKUR, R., AND GROPP, W. D. Improving the performance of collective operations in mpich. In *European Parallel Virtual Machine/Message Passing Interface Users Group Meeting* (2003), Springer, pp. 257–267.
- [46] TRÄFF, J. Transparent neutral element elimination in mpi reduction operations. *Recent Advances in the Message Passing Interface* (2010), 275–284.
- [47] VAN ESSEN, B., KIM, H., PEARCE, R., BOAKYE, K., AND CHEN, B. Lbann: Livermore big artificial neural network hpc toolkit. In *Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments* (2015), ACM, p. 5.

- [48] WEBB, S., CAVERLEE, J., AND PU, C. Introducing the webb spam corpus: Using email spam to identify web spam automatically. In *CEAS* (2006).
- [49] 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 *Advances in Neural Information Processing Systems* (2017), pp. 1508–1518.
- [50] XING, E. P., HO, Q., DAI, W., KIM, J. K., WEI, J., LEE, S., ZHENG, X., XIE, P., KUMAR, A., AND YU, Y. Petuum: A new platform for distributed machine learning on big data. *IEEE Transactions on Big Data* 1, 2 (2015), 49–67.
- [51] YOU, Y., GITMAN, I., AND GINSBURG, B. Scaling sgd batch size to 32k for imagenet training. *arXiv preprint arXiv:1708.03888* (2017).
- [52] YU, D., EVERSOLE, A., SELTZER, M., YAO, K., HUANG, Z., GUENTER, B., KUCHAIEV, O., ZHANG, Y., SEIDE, F., WANG, H., ET AL. An introduction to computational networks and the computational network toolkit. *Microsoft Technical Report MSR-TR-2014-112* (2014).
- [53] ZHANG, J., MITLIAGKAS, I., AND RÉ, C. Yellowfin and the art of momentum tuning. *arXiv preprint arXiv:1706.03471* (2017).
- [54] ZHAO, H., AND CANNY, J. Kylix: A sparse allreduce for commodity clusters. In *Parallel Processing (ICPP), 2014 43rd International Conference on* (2014), IEEE, pp. 273–282.