Training Giant Neural Networks Using Weight Streaming on Cerebras Wafer-Scale Clusters

Stewart Hall, Rob Schreiber, Sean Lie, Cerebras Systems, Inc.

Abstract

State-of-the-art language models are extremely challenging to train; they require huge compute budgets and complex distributed compute techniques. As a result, few organizations train large language models (LLMs) from scratch.

In this paper, we present a new training execution flow called *weight streaming*. By disaggregating parameter storage from primary compute, weight streaming enables the training of models two orders of magnitude larger than the current state-of-the-art. Because weight streaming runs in strictly data parallel form on Cerebras CS-2 systems, it avoids the complex and time-consuming distributed computing techniques that bedevil ML practitioners. Weight streaming demonstrates near perfect linear scaling across clusters of Cerebras CS-2 systems.

We present experimental results showing scaling of large GPT-style large language models across clusters of up to 64 CS-2s, containing 54 million AI cores. We also show how the weight streaming architecture enables the harvesting of dynamic, static, structured and unstructured sparsity.

Contents

1. Trends in Large Neural Network Models

Large language models are growing extremely quickly. While there has been some debate about whether the best approach is to increase parameter count or to increase the number of tokens on which the model is trained, there can be no debate that the computational challenge posed by training the largest models has grown exponentially. And there is no end in sight. Moreover, the fact that computational complexity has increased faster than the performance of traditional compute elements (CPUs, GPUs etc), is shown by the fact that the cluster size used to train these models has grown past thousands of processors to in some cases tens of thousands.

With the computational complexity of training outpacing the capabilities of individual compute units, we must distribute training to multiple compute units. The computational workload for LLM training is commonly distributed by splitting the batch or the model over many compute units, such as GPUs. NVIDIA® estimates that GPT-3 could be scaled to 1,000 GPUs, reducing the training time to about 34 days[1]. There is no data available for either the number of GPUs used or the parameter size for GPT-4.

Distributing training compute over a clusters of thousands of compute elements is enormously challenging, but it is the only choice; but training on a small cluster is impossible due to the memory required to store model parameters and activations. This is because common approaches to distributed training require the model to be stored entirely in the memory of the compute units, what we refer to as stored weight training. Even if it were possible to work around memory constraints, a training time lasting years would be prohibitive.

In this paper, we describe a new paradigm for training giant neural networks, called weight streaming, where model parameters are not stored in the memory of the compute units. Weight streaming allows the cluster size to scale independent of model size. Using weight streaming to harness the power of the Cerebras CS-2 system, we can reduce training time and simplify training at scale.

2. Concepts in NLP Model Training

This section gives a brief overview of the structure and process of training NLP models and defines related terminology used throughout the paper. Readers who are already familiar with these concepts may skip this section.

NLP models process sequences of tokens which commonly represent text as a series of words. We use S to represent the number of tokens (words) in each sequence. Each token is encoded as an integer value corresponding to a word in the model's vocabulary. An NLP model converts a source sequence into a target sequence through a series of transformations, beginning with an embedding table lookup.

The embedding table contains a vector of features, called a hidden state vector, for each word in the model's vocabulary. The number of features in each hidden state vector, which we refer to as H, is a property of the model and typically numbers in the thousands. The per-token hidden state vectors propagate through the model's layers, each of which transforms the input vector to an equal-length output vector.

Layers use sets of trainable parameters to transform the hidden state vectors with operations such as matrix multiplications and element-wise additions. We represent the number of layers in the model as L, also referred to as the depth. The total number of parameters in the model is represented as P, which includes the embedding table and each layer's parameters. Throughout this paper, model size is used to refer to the number of parameters. The hidden state of an input sequence at any point in the model is represented by an activation tensor, which consists of a hidden state vector for each token of the input sequence. As a shorthand, we use T to represent the activation tensor size for one sequence:

$$T = H \times S$$

Training an NLP model is performed by feeding labeled inputs, called training samples, through the model to compute gradients which can be used to update the model's parameters. The output of the model for each sample is compared with the label to compute an error measure, which is differentiated to compute an activation gradient. The gradient is backpropagated through the layers of the model to compute both a weight gradient and new activation gradient at each layer.

A training job consists of multiple steps, called training iterations. In each training iteration a subset of the training dataset, called a batch, is used to compute weight gradients for each layer. These weight gradients are consumed by an optimizer algorithm to update the model's weights between each training iteration. Stochastic gradient descent (SGD) and Adam are commonly used optimizer algorithms. Training iterations are performed until the model has converged, i.e. when an error measure levels off.

3. Methods for Scaling Training with Stored Weights

Training can be distributed to many compute units using data parallelism or model parallelism (Figure 1). Data parallelism splits the batch of training samples over N_d compute units with each compute unit processing all layers of the model for its subset of the batch. Model parallelism splits the model over multiple compute units, with each compute unit processing a subset of the layers for all samples in the batch. The model can either be split by placing a subset of each layer on each of N_m compute units, called tensor model parallelism, or by placing a subset of the layers on each of N_p compute units, called pipeline model parallelism. These modes of parallelism can be combined, for example by creating N_d instances of the model, sharding the batch into N_d shards, and distributing each model instance across a set of N_m compute units, thereby using $N_d \times N_m$ compute units; or one might use other combinations.

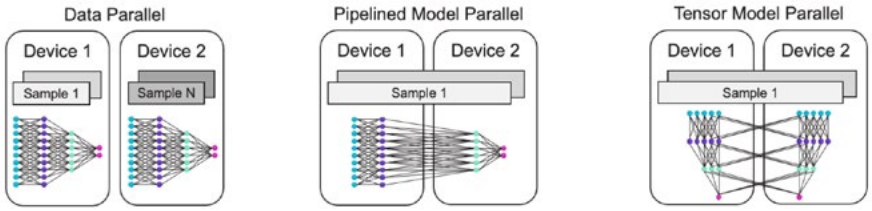


Figure 1. Approaches to parallelism in neural network training.

This section quantifies memory and communication needed by each of the scaling approaches. The memory and communication requirements of these common scaling approaches are summarized in Table 1. Memory requirements are calculated based on the size of model weights and per-layer activation tensors stored on a single compute unit. The communication models account for transfers of weight gradients and per-layer activation tensors into and out of a single compute unit. Although communication is also needed to transfer each batch of training samples to the compute units, we do not account for this in our models because the inputs to NLP models, which are the focus of this paper, are very small relative to gradients and activation tensors. Inputs consist of a single integer per token, representing a word ID in the vocabulary, whereas activation tensors consist of a H-length vector per token.

3.1. Data Parallelism

Data parallelism places a copy of the model on each of N_d compute units and shards each batch of training samples over the compute units. The number of training samples processed by each compute unit, b_d, is:

$$b_d = \frac{B}{N_d}$$

After each unit has computed partial weight gradients using its shard of the batch, gradients are summed between all compute units to get the gradient for the full batch. Each compute unit uses the summed gradient to update its copy of the weights, guaranteeing that the replicas of the model stay in sync.

Memory is needed on each compute unit to store model parameters and activations. Each compute unit stores an entire replica of the model and an activation tensor per-layer for each sample in its shard of the batch, needed for gradient computations in the backward pass. The total memory requirement per compute unit is:

$$Memory = P + LTb_d$$

At the end of every training iteration, compute units communicate in order to reduce gradients prior to each weight update. They perform an all-reduce operation, where each compute unit receives a copy of the sum of the original, partial gradient tensors. Commonly used algorithms for all-reduce require each node to send and receive the entire gradient for each layer's tensor once or twice, so total communication per compute unit is:

$$Communication\ volume\ = O(P)$$

This amount of data is communicated bidirectionally at each compute unit, for sending and receiving weight gradients, after each batch.

3.1.1. Fully-Sharded Data Parallelism (FSDP)

A major limitation of the standard data parallel implementation is that each compute unit must have enough memory to store all the model parameters. A popular modification designed to address this, which allows for training very large models with stored weights, is called fully-sharded data parallelism (FSDP). This approach shards the weights and optimizer parameters between the compute units, such that each weight is stored persistently in the memory of only one compute unit. Prior to each layer's forward or backward pass computations, the compute unit owning the layer's weights broadcasts them to all other compute units. After the layer's computations have been completed, the compute units which do not own the weights can reclaim the memory that was temporarily used to store them. This approach makes much more efficient use of each compute unit's memory, avoiding replication of the weights for the duration of the training iteration. It also allows data parallelism to scale with model size as long as the model's weights fit in the aggregate memory of the set of compute units.

$$Memory\ = \frac{P}{N_d} + \frac{P}{L} + LTb_d$$

This memory efficiency comes at the cost of extra communication to broadcast the weights during the forward and backward passes, but the communication volume is asymptotically the same as the traditional implementation. These communication operations can be pipelined and overlapped with other layer computations if the interconnect between compute units has sufficient bandwidth.

3.2. Model Parallelism

Model parallelism involves distribution of the model's layers over multiple compute units. Individual layers can be split over multiple compute units, or each compute unit can contain groups of entire layers, also known as pipeline model parallelism. Parameters are not replicated when model parallelism is used alone: each compute unit stores a subset of the model parameters corresponding to the computations which it performs. Activations are communicated between compute units as a batch of training data progresses through the model.

3.2.1. Tensor Model Parallelism

One method of model parallelism, known as tensor model parallelism, splits each layer over N_m compute units. For example, a fully connected layer can be split on its input features, its output features, or both. When input features are split, each compute unit computes partial sums for all output features, then a reduction is required to compute output features. When output features are split, input activations are broadcast to each compute unit where a subset of output features are computed for each sample. Storage of parameters and activations, for gradient computations,

is distributed, so memory required per compute unit is:

$$Memory = \frac{P}{N_m} + \frac{LTB}{N_m}$$

Communication among compute units is required, at the transition between finishing a layer and beginning the next layer, to reduce partial sums or redistribute the output tensor so that it matches the distribution of weights used by the next layer. When weights are split on either the input feature or output feature dimension, the communication required for each layer is proportional to the activation tensor size:

$$Communication\ volume\ = O(LTB)$$

If weights are split on both input and output feature dimensions, then the minimal communication required decreases proportional to the square root of compute units:

$$Communication\ volume\ = O\left(\frac{LTB}{\sqrt{N_m}}\right)$$

3.2.2. Pipeline Model Parallelism

Pipeline model parallelism is a different method of model parallelism where each compute unit is responsible for a subset of the model's layers. Each subset of layers constitutes a stage in a large pipeline. The training batch is split into shards, of size b_p, to keep the pipeline full, with each stage operating on a different shard of the training batch at any given time. We denote the number of pipeline stages as N_p, so the number of layers allocated to each compute unit is roughly equal to:

$$Layers\ per\ compute\ unit = \frac{L}{N_p}$$

There is a latency to fill and drain the pipeline between each weight update, so the number of batch shards, equal to B / b_p, should be high relative to N_p. Cerebras has employed pipeline model parallelism effectively for moderate sized networks, since the company's Wafer-Scale Engine (WSE) can operate efficiently on a batch size of one in each pipeline stage ($b_p=1$). Memory is distributed across compute units and is used to store parameters and activations which are stored for each layer processed by the compute unit. Activations need to be stored for the duration of time between when a batch shard is processed in the forward pass and when it is processed in the backward pass. This means that the number of samples buffered for each layer depends on the layer's location in the network pipeline, with the first pipeline stage storing activations for $2N_p-1$ batch shards and the last pipeline stage storing activations for only one batch shard. Increasing the number of compute units increases the pipeline depth, which increases the aggregate memory required for activation buffers. As a result, the percentage of memory used for activation buffers increases with the number of compute units. The worst-case memory required on a single compute unit for parameters and activations is:

$$Memory = \frac{P}{N_p} + \left(2N_p - 1\right)\frac{L}{N_p}Tb_p$$

The bandwidth required to move activations between pipe stages for the full batch is:

$$Communication\ volume\ = O(TB)$$

Scaling Approach	Activation Memory	Parameter Memory	Communication Volume	Main Limitations
Data Parallel (DP)	$\frac{LTB}{N_d}$	P	P	Parameter memory does not decrease with N.
Tensor Model Parallel (TMP)	$\frac{LTB}{N_m}$	$\frac{P}{N_m}$	$\left(\frac{LTB}{\sqrt{N_m}}\right)$	Communication does not scale as well as compute.
Pipeline Model Parallel (PMP)	$\left(2N_p - 1\right)\frac{L}{N_p}Tb_p$	$\frac{P}{N_p}$	TB	Activation memory and communication do not decrease with N.
DP+TMP	$\frac{LTB}{N_d N_m}$	$\frac{P}{N_m}$	$\frac{P}{N_m} + \frac{LTB}{\sqrt{N_m}}N_d$	Communication does not scale as well as compute. Complexity.
DP+PMP	$\left(2N_p - 1\right)\frac{L}{N_p}Tb_p$	$\frac{P}{N_p}$	$\frac{P}{N_p} + \frac{TB}{N_d}$	Sub-linear activation memory and communication scaling. Complexity.
DP+TMP+PMP	$\left(2N_p - 1\right)\frac{L}{N_p}\frac{T}{N_m}b_p$	$\frac{P}{N_m N_p}$	$\frac{P}{N_m N_p} + \frac{LTB}{\sqrt{N_m}}N_d N_p$	Complexity.

Table 1. Overview of memory and communication requirements of common scaling approaches. The last column highlights the primary limitation when scaling for a giant model.

3.3. Limitations of Scaling Stored-Weight Training

When scaling to multiple compute units, data parallelism is commonly chosen for its simplicity: each compute unit performs the same computations and participates in one communication step, to synchronize gradients, once per training iteration. However, data parallelism alone fails for giant models due to its memory requirement per compute unit. The activation component of this memory requirement can be reduced by increasing N_d, which decreases b_d. The model parameter component of this memory requirement is not reduced as N_d increases, which means model size is limited by the compute unit's memory capacity. For GPT-3, the weights alone require 700GB of memory, which is an order of magnitude greater than the capacity of a typical GPU.

Using FSDP removes this limitation, since it distributes weight storage over all compute units, but has the drawback of coupling the batch size to the parameter count. N_d must be increased proportional to P, so that the compute units have enough memory to store all parameters. Since compute is only parallelized over the batch dimension, B must be increased proportionally to N_d to attain high utilization on each compute unit.

Tensor model parallelism allows for the memory requirement per compute unit to be reduced as N_m grows, but it introduces significant communication overhead. Activation tensors must be communicated between compute units for each layer. The compute requirement per compute unit for a fully connected layer is $O(H^2SB/N_m)$ and the amount of communication required for activations is $O(HSB/\sqrt{N_m})$. Training on a large cluster with tensor model parallelism alone is likely to be communication bottlenecked since compute units often provide orders of magnitude more FLOPS than network bandwidth. Unlike data parallelism, where batch size can be increased to amortize this overhead, this bottleneck cannot be as easily overcome because the batch size also factors into the

volume of communication needed for activations. Due to the frequency of communication, latency can also become a critical component of training time if the cluster is large enough.

Both the frequency of communication and required parameter memory are addressed by pipeline model parallelism, but the aggregate memory needed for activations grows with N_p. Earlier layers in the network need to store more activation buffers as the number of pipeline stages increases, which counters the effect of distributing the layers to more compute units. Since the percentage of memory used for activations increases with the number of compute units, scaling out with pipeline model parallelism for the purpose of fitting a larger model has diminishing returns. Pipeline model parallelism also introduces implementation complexities. Each compute unit may be executing different computations and communicating different types of tensors. The network needs to be distributed such that each pipeline stage completes in the same amount of time. Furthermore, as the pipeline depth increases, the overhead of filling and draining the pipeline between weight updates becomes more significant, requiring a larger batch size.

These three approaches to parallelism can be combined to make training of giant neural networks feasible, both in terms of compute time and memory required per compute unit. However, the complexity introduced by these hybrid approaches puts them out of reach for many users. Furthermore, all of these stored-weight solutions share one problem: that the number of compute units required is partially dictated by the number of parameters in the network. The total number of compute units used by a specific scale-out configuration is:

$$N = N_d \times N_m \times N_p$$

For a neural network like GPT-3 with its large parameter count P, there is a smallest value of N below which the model cannot be trained in the stored-weight paradigm due to per-compute unit memory requirements. This is problematic because the compute requirement does not always scale with the model's parameter count. Workloads such as fine-tuning require storage for the entire set of parameters but involve far less compute than pre-training. Allocating the same amount of compute power to both workloads does not make sense. Architectural changes to the neural network, such as the use of sparse attention, similarly reduce compute without reducing storage required for parameters.

3.4.A Survey of Parallelism Approaches

Figure 2, gathered from recent publications, shows the number of graphic processing units required to run the largest of the LLMs. So large and complicated are these clusters that successful training takes months of work, tens of millions of dollars of hardware, tens of megawatts of power, and the efforts often result in a publication. In fact, as the model size gets larger, the number of GPUs needed to train the model increases so rapidly that the Y-axis is on a logarithmic scale.

Today, the fundamental limiting factor in running LLMs is not the AI. It is the distributed compute challenge of putting LLMs on thousands of GPUs, and the scarcity of the distributed compute expertise necessary to do so.

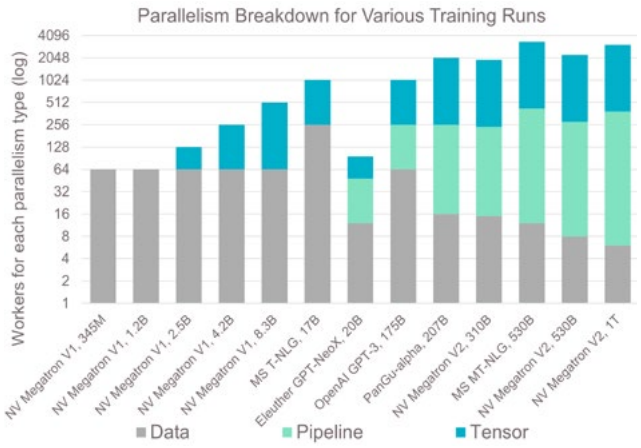


Figure 2. Number of GPUs and types of parallelism required to train the largest LLMs.

3.5. A New Solution Is Needed

Clearly a new solution to scaling is needed to enable efficient training of giant neural networks. The new solution should:

1. Not impose constraints on model size based on the memory available on individual compute units,
2. Be able to scale compute throughput with the computational requirements of the model, and
3. Achieve scaling without complicated hybrid approaches to parallelism.

To accomplish this decoupling of compute performance from model size, we have developed a radical new approach to training giant neural networks. It is based on three key points:

1. The replacement of CPU and GPU processing by wafer-scale accelerators such as the Cerebras CS-2 system. This change reduces the number of compute units needed to achieve an acceptable compute speed.
2. To meet the challenge of model size, we employ a system architecture that disaggregates compute from model storage. A compute service based on a cluster of CS-2 systems (providing adequate compute bandwidth) is tightly coupled to a memory service (with large memory capacity) that provides subsets of the model to the compute cluster on demand. As usual, a data service serves up batches of training data to the compute service as needed.
3. An innovative model for the scheduling and coordination of training work across the CS-2 cluster that employs data parallelism, layer at a time training with sparse weights streamed in on demand, and retention of activations in the compute service.

We will give below the details of an implementation of this new approach, based on the Cerebras CS-2 system, and we will assess its effectiveness for training GPT-3 as well as order-of-magnitude smaller and larger models.

4. Introducing Weight Streaming

Our new approach to training is called weight streaming. It is based on the disaggregation of compute and storage. Similar solutions have been developed which spread model weights over multiple compute units and insert communication as necessary[2], but our approach takes this a step further. In weight streaming, model weights are not stored in the memory of the compute units at all, as has been traditionally done. Instead, we move the weights to a separate memory service

characterized by a high storage capacity with relatively low compute capabilities. This means that each component of the solution can be optimized for its specific role. Compute units can be optimized for high floating-point throughput on linear algebra operations and the memory service can be optimized for capacity and bandwidth. Specialization is commonly applied in existing distributed training solutions for storage of the training dataset. In like manner, we store the weights of a giant model in a separate memory service and stream them to the compute units as needed. The dataset service and memory service have different requirements, so they are separate components of the solution. The most important difference is the bandwidth requirement, which is generally lower for the dataset service when training NLP models. As a result, we focus only on the compute unit and memory service for the remainder of this paper.

Weight streaming can be combined with existing approaches for parallel training. The training batch can be sharded over N_d compute units to leverage data parallelism, requiring weights to be broadcast from the memory service to all compute units and weight gradients to be reduced between compute units. Tensor model parallelism can be used with weight streaming by splitting activations on their feature dimension over N_m compute units and pairing separate memory services with each compute unit. Pipeline model parallelism can similarly be combined with weight streaming by splitting activations by layer over N_p compute units and pairing separate memory services with each compute unit. Weight streaming, in all cases, requires bandwidth proportional to P for streaming weights to the compute units and gradients from the compute units. This matches the bandwidth required for stored-weight data parallelism, so does not introduce extra bandwidth requirements. Combining weight streaming with model parallelism increases bandwidth requirements since data movement is required for activations, between compute units, in addition to weights and gradients between the memory service and compute units. For this reason and reduced implementation complexity, we have focused on data parallelism *($N_m = N_p = 1$ and $N = N_d$)* in the Cerebras weight streaming implementation. Using this approach to scaling, an interconnect is needed to link the single memory service to the cluster of compute units.

5. The Parameters of the Problem and Requirements on the Services

We designed the Cerebras weight streaming solution to enable scalable training of the largest existing neural network models and to power the development of future models. Specific examples of these models inform the requirements for the compute units, memory service, and interconnect. Observing trends in the sizes and parameters of these models allows us to further refine the requirements to best support the needs of future research. NLP models have seen the most growth in parameter counts recently, requiring large-scale training, so these are our initial focus. Table 2 details several of the largest transformer-based language models introduced over the past three years.

Model	Parameters (P)	Layers (L)	Features (H)	Sequence Length (S)
GPT-2 XL [3]	$1.5x10^9$	48	1,600	1,024
Megatron-8.3B [4]	$8.3x10^9$	72	3,072	1,024
Turing-NLG [5]	$1.72x10^{10}$	78	4,256	1,024
GPT-3 [1]	$1.75x10^{11}$	96	12,288	2,048
Megatron T-NLG [6]	$5.3x10^{11}$	105	20,480	2,048

Table 2. Evolution of large transformer-based language models over the past 3 years. Note that the features (H) is also sometimes referred to as d_{model}.

5.1. The Required Performance

The compute units in the cluster should be capable of training these large models in a reasonable amount of time. To estimate the compute requirement for the cluster, we calculate a target floating-point operation rate which the cluster would need to deliver to train each model in one week. The number of floating-point operations required to train each model can be estimated given the number of tokens needed to train each model to convergence. Forward propagation uses each model parameter once per token. Accounting for weight and activation gradient computations, each model parameter is used three times per token. In all three cases, the weight is used for a multiply-accumulate operation, resulting in 6 total floating-point operations per parameter per token. The number of operations needed to train the model is dominated by the fully connected layers, so we can approximate the operation count for each model, shown in Figure 3 and Table 3, by multiplying 6 × tokens × parameters. Dividing the estimated total operations by 604,800, the number of seconds in one week, gives a target for the massive floating-point performance required. For comparison, the target for Megatron T-NLG of 1,420 petaFLOPS is more than three times greater than the performance, 1.1 exaFLOPS, of Frontier, currently the world's largest supercomputer!

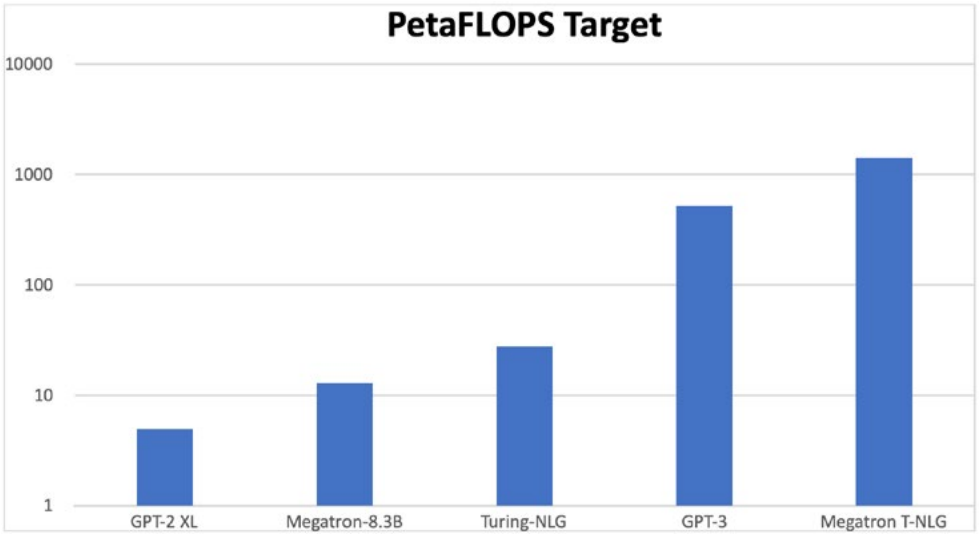


Figure 3. PetaFLOPS required to train each model in 1 week.

Model	Tokens to Train	Total Operations	PetaFLOPS Target
GPT-2 XL	3.00×10^{11} (est.)	2.7×10^{21}	5
Megatron-8.3B	1.57×10^{11}	7.82×10^{21}	13
Turing-NLG	1.57×10^{11}	1.62×10^{22}	28
GPT-3	3.00×10^{11}	3.15×10^{23}	520
Megatron T-NLG	2.7×10^{11}	8.59×10^{23}	1,420

*Table 3. From left to right, (1) tokens to train the model to convergence or batch size multiplied by total training steps (2) total floating-point operations used to train the model based on 6*tokens*parameters (3) target petaFLOPS for a cluster to train in one week.*

5.2. The Required Memory Service Capacity

The main requirement of the memory service is capacity. Enough capacity is needed to hold the weight, gradient, and optimizer state for each parameter in the model. Adam is a popular optimizer choice, which requires two momentum terms per parameter. Assuming that weight updates are computed in single-precision floating point format (FP32), the memory service must store 16 bytes per model parameter to contain the weight, gradient, and two momentum terms. We further round this up to 20 bytes per weight to account for a sparse working copy of the nonzero weights, which includes a 2-byte column index and FP16 copy of the weight. Capacity required of the memory service can then be computed by multiplying a model's parameter count by 20 bytes (Figure 4 and Table 4).

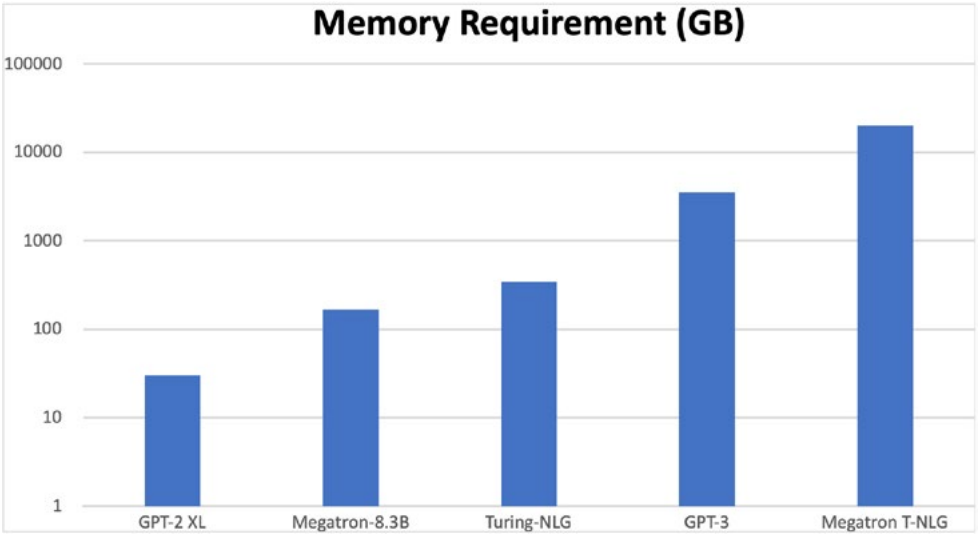


Figure 4. Memory required to store model parameters and optimizer state.

Model	Total Parameters	Memory Requirement
GPT-2 XL	1.5×10^{9}	30GB
Megatron-8.3B	8.3×10^{9}	166GB
Turing-NLG	1.72×10^{10}	344GB
GPT-3	1.75×10^{11}	3.5TB
Megatron T-NLG	5.3×10^{11}	10.6TB

Table 4. Total parameters and storage size of model parameters and optimizer state.

5.3. The Required Interconnect Bandwidth

The interconnect between the memory service and the compute units should provide enough bandwidth to support both the flow of weights to the compute units and the flow of weight gradients from the compute units at a rate determined by the compute speed. Compute units must be supplied with weights fast enough to avoid bubbles (unused processor cycles). During the backward pass, weight gradients must be streamed out of the compute units as fast as they are computed. We compute the bandwidth requirement by dividing the volume of data fed into and out of the compute units by the target compute time of one week. Weights are represented by half-precision floating-point (FP16) values in our implementation. They are fed into the compute units once during the forward pass, for computation of activations, and once during the backward pass, for computation of activation gradients. Weight gradients are sent out of the compute units in FP32 format during the backward pass. In total, 32 bits of data are sent in each direction per parameter in the model per training iteration (Figure 5 and Table 5). (As is now commonplace, gradients are computed and weights are updated at 32-bit precision, which preserves accuracy over the many learning steps during training. On streaming to the compute units, they are rounded to 16-bit precision for use there.)

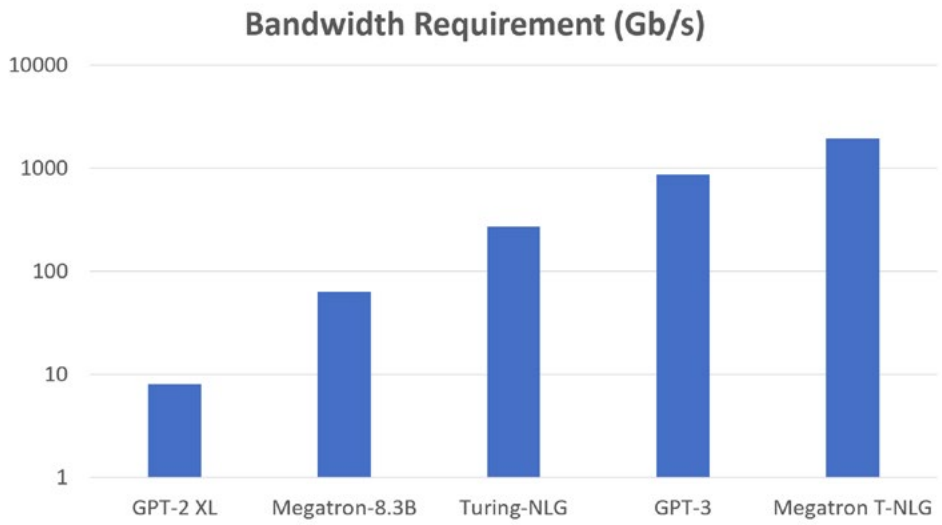


Figure 5. Bandwidth required for weight and gradient communication to train each network in 1 week.

Model	Total Parameters	Tokens to Train	Batch Size	Training Iterations	Bandwidth Requirement
GPT-2 XL	1.50×10^{9}	3.00×10^{11} (est.)	5.24×10^{5}	1×10^{5} (est.)	8Gb/s
Megatron-8.3B	8.30×10^{9}	1.57×10^{11}	1.1×10^{6}	1.43×10^{5}	63Gb/s
Turing-NLG	1.72×10^{10}	1.57×10^{11}	5.24×10^{5}	3.00×10^{5}	273Gb/s
GPT-3	1.75×10^{11}	3.00×10^{11}	3.2×10^{6}	9.37×10^{4}	868Gb/s
Megatron T-NLG	5.30×10^{11}	2.70×10^{11}	3.9×10^{6}	6.92×10^{4}	1.94Tb/s

Table 5. From left to right, (1) Total parameters in the model (2) Total tokens needed to train the model (3) Batch size in tokens used to train the model based on published values (4) Total training iterations calculated by dividing tokens to train by batch size (5) bandwidth required in each direction computed by multiplying number of parameters by number of iterations by 32 bits and dividing by seconds per week.

The estimates in Tables 3, 4, and 5 provide bounds for the compute, memory, and bandwidth capabilities required of our weight streaming implementation. For example, the compute units should scale from a combined compute throughput of tens to thousands of petaFLOPS based on the floating-point operations required to train each model. The memory service should support a minimum of 4TB capacity to enable training of GPT-3 sized models.

6. Components of the Cerebras Solution: Wafer-Scale Engine, MemoryX, SwarmX

The Cerebras weight streaming implementation is built around our second-generation Wafer-Scale Engine (WSE-2), which lies at the heart of the CS-2 system. Storage for model parameters and optimizer state is disaggregated from compute into our memory service, called the MemoryX service. The MemoryX service provides persistent storage for the parameters. It accepts weight gradients and uses stored optimizer parameters to compute weight updates between training iterations. The Cerebras weight streaming architecture allows for training to be scaled, without replication of the model, to a Cerebras Wafer-Scale Cluster, consisting of an arbitrary number of CS-2 systems, each served by a single MemoryX service. Weights and gradients are shuttled between the single MemoryX service and multiple CS-2 systems by our SwarmX interconnect fabric.

6.1. Compute: The Wafer-Scale Engine

The Cerebras weight streaming implementation harnesses the power of the WSE-2 to compute activations, activation gradients, and weight gradients for each batch of training samples. The WSE-2 is an optimal compute unit for training giant neural networks because it can parallelize computation of massive layers over its 850,000 cores. The on-wafer network of the WSE-2 offers an aggregate 220Pb/s of bandwidth, which is three orders of magnitude more than the aggregate bandwidth between GPUs in an NVIDIA DGX™ server. This bandwidth is uniform between cores, meaning that the WSE-2 can truly be treated as a single massive compute unit. This allows giant layers in our weight streaming implementation to be executed extremely quickly on a single compute unit, avoiding the need to employ tensor model parallelism.

Activations computed during the forward pass and intermediate activation gradients computed during the backward pass are stored in the WSE-2's on-wafer memory. With 40GB of SRAM distributed across 850,000 cores, the WSE-2 has enough activation storage for massive layers and provides 20PB/s of aggregate memory bandwidth for ultra-fast access to activations during computation. The activation tensors stored on-wafer are used for computation of the subsequent layer's activations as weights arrive from the MemoryX service. During the backward pass the stored activation and activation gradient tensors are used for computation of weight gradients which are sent back to the MemoryX service. The WSE-2's 1.2Tb/s of I/O bandwidth is used for receiving weights from and transmitting gradients back to the MemoryX service.

6.2. Exploiting weight sparsity

While the WSE-2 allows each layer to be distributed over hundreds of thousands of cores, reducing computation time, parallelism is not the only way to reduce training times. In response to the explosive growth in model size, researchers have been developing smarter models which make more efficient use of their parameters. Weight sparsity is one promising way to reduce parameter count, and in recent tests large drops in training floating-point operations were reported, as shown in Table 6. The Lottery Ticket Hypothesis[7] for example, showed that model parameters can be pruned by 90% without a reduction in accuracy.

Technique	Sparsity	FLOP ↓	Reference
Fixed Sparse Training	90%	**8x**	Lottery Ticket [MIT CSAIL]
Dynamic Sparse Training	80%	**2x**	Rig the Lottery [Google Brain, DeepMind]
Scaling-up Sparse Training	90%+	**10x+**	Pruning scaling laws [MIT CSAIL]
Monte Carlo DropConnect	50%	**2x**	DropConnect in Bayesian Nets [Nature]

Table 6. Some Existing Sparsity Research Examples.

Realizing a performance improvement from this kind of unstructured sparsity, while historically difficult due to hardware limitations, will be essential to affordably training larger models. The WSE-2's unmatched performance on sparse linear algebra operations is key to achieving our targeted training times.

Using the WSE-2 for weight streaming allows us to reduce training time by automatically taking advantage of weight sparsity. The WSE-2 is the only processor that handles unstructured sparsity at the silicon level. The WSE accelerates sparse computation using dataflow scheduling where computational work is triggered by the arrival of data over the fabric. Zeros are omitted when a tensor is transmitted over the fabric, which results in a reduction of computation time proportional to the sparsity of the tensor. In the weight streaming execution mode, model weights are transmitted over the fabric, reducing compute proportional to the sparsity of the weights. Cores in the WSE take advantage of the enormous memory bandwidth available to operate at full utilization on a single weight value at a time, providing a speedup even for weights with unstructured sparsity. Sparsity can reduce wall-clock time of both weight communication and all four computational phases: activation, activation gradient, weight gradient, and weight update. This allows the Cerebras weight streaming solution to train giant networks faster without a bottleneck.

6.2.1. Sparsity in Action: Sparse Pre-Training and Dense Fine-Tuning

The Cerebras CS-2 system derives its performance from hundreds of thousands of fast processors that are not speed limited by memory bandwidth, because the local per-processor memory (and that is the only memory in the system) is as fast as the processor. For this reason, vector operations run at full performance. On most other machines, matrix multiplication, but not vector operations, can run full out because they can exploit a cache hierarchy whereas a vector operation cannot.

Learn more about Cerebras sparsity research

- Accelerating Large GPT Training with Sparse Pre-Training and Dense Fine-Tuning
- Creating Sparse GPT-3 Models with Iterative Pruning
- Can Sparsity Make AI Models More Accurate?

Why does this matter here? The answer is that there are two important facts. First, networks in which most of the synaptic connections – the weights – are constrained to be zero can learn and perform nearly as well as unconstrained networks, and they can learn faster, with less work. These networks exhibit weight sparsity (Figure 6). And second, to train a weight-sparse network, the hardware spends almost all its time multiplying a sparse matrix by a dense matrix. The Cerebras architecture can do this at very close to full speed; other machines struggle with it.

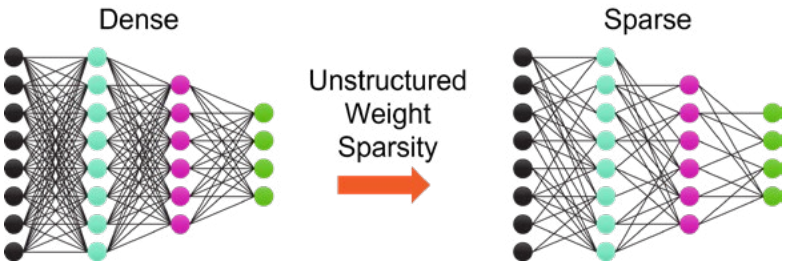


Figure 6. Applying weight sparsity to a dense neural network by zeroing weights effectively prunes neuron connections within the network.

To exploit this idea, we built an implementation of GPT-3 that is dense and fully accurate, but we sped up training by 2.5X using weight sparsity[8]. How's that? Well, we first trained a version of the network that has 75% zero weights. This was a pre-training phase, on the Pile[9], an 800 GB generic language dataset. Then we allowed all the weights to be nonzero and did fine tuning on a more specialized dataset, Curation Corpus[10], which is a set of news item summaries. The result: It took 40% of the FLOPs that training the dense would have taken, and the accuracy was comparable (Figure 7).

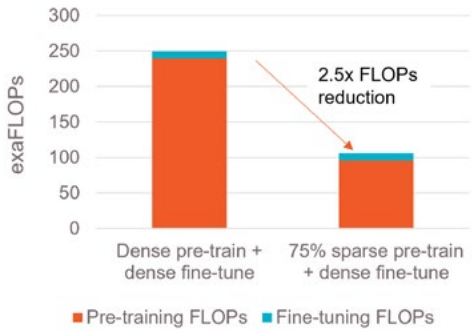


Figure 7. FLOPs spent in pre-training GPT-3 XL on the Pile dataset, followed by fine-tuning on Curation Corpus. FLOPs spent during pre-training dominate the overall FLOPs. Sparse pre-training at 75% sparsity leads to 2.5x reduction in overall FLOPs.

6.2.2. Sparsity in Action: Creating Sparse GPT-3 Models with Iterative Pruning

Large language models are costly to setup, train, and deploy. Indeed, inference costs are often high enough to preclude the use of the most performant, largest models. Weight sparsity, when coupled with hardware which accelerates unstructured sparsity, is a promising way to cut inference time, speed up training, and reduce memory requirements.

We have shown that large-scale language models can be pruned to high levels of sparsity while maintaining accuracy competitive with their dense counterparts.

In this work, we applied pruning and learning rate rewinding on a GPT-3 1.3 billion parameter model. We first trained a dense model from scratch on the Pile and then iteratively pruned and trained to find sparse models (Figure 8). Our 83.8% sparse model represents a 3x reduction in inference FLOPs without any degradation in validation loss. All models are trained on a single CS-2 device, without requiring user level distributed code or setup and are competitive with the original dense model.

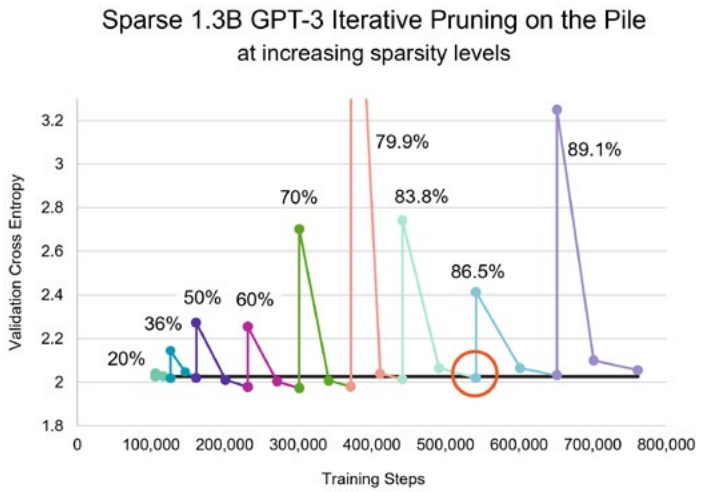


Figure 8. Cross entropy on the Pile validation set vs. training steps (lower is better). Each vertical line represents the increase of loss when a model is sparsified (note that this is recovered as we train the model). Sparsity levels refer to average sparsity percentages on projections and feed forward layers in GPT-3. Evaluation is done for 380M tokens, roughly the full validation set of PILE.

We measured the effectiveness of iterative pruning by evaluating the resulting sparse models using several metrics: Pile loss, zero-shot LAMBADA[11] accuracy, and zero-shot WikiText103[12] perplexity (Figure 9). We show that the GPT-3 1.3B dense model can be iteratively pruned to high degrees of sparsity without significant reduction in quality across these metrics.

Sparsity	FLOPs (reduction ↑)	The PILE (cross entropy ↓)	LAMBADA (accuracy ↑)	WikiText103 (adjusted ppl ↓)
Dense	1x	2.055	40.79	20.19
20%	1.2x	2.046	41.96	20.02
36%	1.4x	2.048	40.54	19.99
50%	1.7x	2.005	44.17	18.74
60%	1.9x	1.999	43.96	18.56
70%	2.3x	2.005	42.91	18.76
79.9%	2.8x	2.038	39.49	19.65
83.8%	3.0x	2.044	42.60	19.94
86.5%	3.2x	2.054	41.39	20.29
89.1%	3.5x	2.078	40.40	20.97

Figure 9. PILE cross entropy on validation set, accuracy on LAMBADA, and adjusted perplexity on WikiText103 across trained sparse models, along with inference FLOPs reductions compared to baseline dense model. The direction of the arrow indicates better result (e.g., up indicates higher is better). Training done on a Cerebras CS-2.

6.2.3. Sparsity in Action: Increasing accuracy without increasing compute using Sparse-IFT

To achieve higher accuracy generally requires more compute, most commonly by increasing the model and dataset size. However, this trend is not sustainable because even today's models are already prohibitively expensive to train in terms of time, cost, and energy.

We believe the way forward is to break the link between accuracy and compute, by training with sparsity. As discussed briefly above, sparsity is already often used in inference, but two factors have hindered widespread use in training: a lack of hardware that can train with sparsity and a lack of accessible systematic ML techniques to improve sparse training.

At Cerebras, we are targeting both issues with hardware-ML co-design. The Cerebras CS-2 system is explicitly designed to accelerate training using sparsity. In fact, it is the only hardware architecture capable of accelerating unstructured sparse training at scale today.

Most traditional sparsity techniques primarily aim to reduce model compute, as shown in point B of Figure 10. Sparsity can also be used to improve model accuracy, as shown in point C, but that has been relatively little studied.

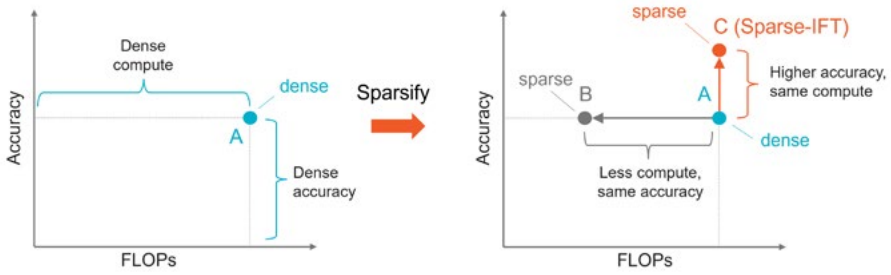


Figure 10. Traditional sparsity techniques primarily aim to reduce compute for a given accuracy, moving point A to point B. Sparse-IFT increases accuracy for the same compute budget, moving from point A to point C.

Building on existing sparsity techniques as a foundation, we have developed a sparse transformation developed Sparse-IFT[16], or *Sparse Iso-FLOP Transformation*, that can be easily applied to *any* existing model. As detailed in our paper, Sparse-IFT creates a larger versions of existing dense layers (Figure 11) by making them sparse while preserving the compute requirement of the original dense layers, hence the term "Iso-FLOP". By increasing the model size, Sparse-IFT increases the model's representational capacity which, in turn, increases accuracy. And by using sparsity, it does not incur the significantly higher compute requirement traditionally resulting from larger models.

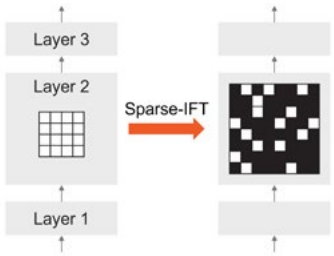


Figure 11. Sparse-Wide Iso-FLOP Transformation of a layer's fully-connected matrix to a larger sparse matrix with the same number of parameters, resulting in the same FLOPs.

Sparse-IFT has shown to increase accuracy on computer vision models (e.g., ResNet, MobileNet) by up to 3.5% (Figure 12) and improve perplexity in language models (e.g., GPT) by up to 0.4, all without significantly increasing compute requirements.

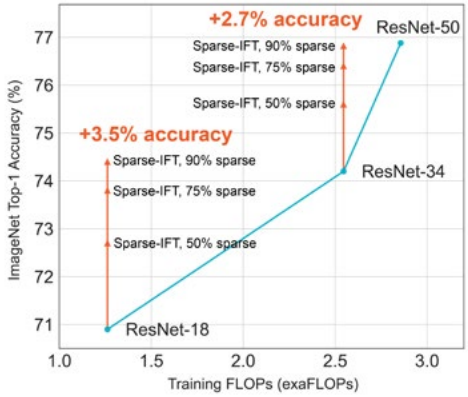


Figure 12. Using Sparse-IFT, we increase accuracy by up to 3.5% on ImageNet for ResNet models that are Iso-FLOP to ResNet-18 and ResNet-34.

Sparse-IFT is also easy to use. It's a simple drop-in replacement for existing dense layers and can be used without extensive hyper-parameter tuning.

6.3. Weight Storage: The MemoryX Service

Model parameters and optimizer state are stored using the MemoryX service, where they are also updated between each training iteration. The capacity of the MemoryX service can scale from 4TB to 2.4PB, allowing the solution to support models with up to 120 trillion parameters. Internally, the MemoryX architecture uses both DRAM and flash storage in a hybrid fashion to achieve both high performance and high capacity. Achieving full compute utilization requires enough network and memory bandwidth to feed weights into the compute units as fast as they are consumed for computations. Both the storage and I/O interface of the MemoryX service can match or exceed the I/O bandwidth of a CS-2 system.

It is important that parameters can be accessed by the compute units with minimal latency, to avoid bubbles during the training process. The process of streaming weights for each layer can be pipelined since weights for most layers can be accessed before computations of the previous layer complete. The one exception to this occurs at the boundary between training iterations, when weights are updated. This means that latency can be hidden for most of the training iteration and has a minimal impact on performance.

Weight updates are performed by the MemoryX service, which provides flexible compute capable of supporting any optimizer algorithm, such as SGD or Adam. The amount of compute required for weight updates is relatively low compared to the compute used to calculate activations and gradients. This is because the number of weight update operations is proportional to the number of parameters, O(P), but activation and gradient compute increases linearly with the batch size, O(BSP). For the same reason, it is possible for the compute provided by the MemoryX service to support any size of CS-2 cluster. Weight updates must be computed at least as fast as weights are streamed out to the CS-2s to avoid a compute bottleneck. Each weight is streamed out of the MemoryX service twice between each weight update, once in the forward pass and once in the backward pass. The MemoryX service delivers a FLOP/s rate three orders of magnitude greater than its I/O bandwidth, which allows for execution of thousands of floating-point operations per weight on each training iteration. This is plenty of compute power to support any commonly used optimizer algorithm.

6.4. Linking Weights to Compute: The SwarmX Fabric

For each training iteration, for each network layer, a copy of the layer weight tensor is sent from the MemoryX service to each CS-2 system, once for computing activations in forward propagation, and again for computing activation gradients during backwards propagation. Weight gradients computed by each CS-2 system are also sent back to the MemoryX service where they are used for weight updates. We use the SwarmX fabric to connect the MemoryX service to each CS-2 system, facilitating the broadcast of weights and aggregation of gradients (Figure 13). The MemoryX service sends a single copy of the weights to the SwarmX fabric, which handles the broadcast to each CS-2 system. On the backward pass, the MemoryX service receives a single copy of the gradients from the SwarmX fabric.

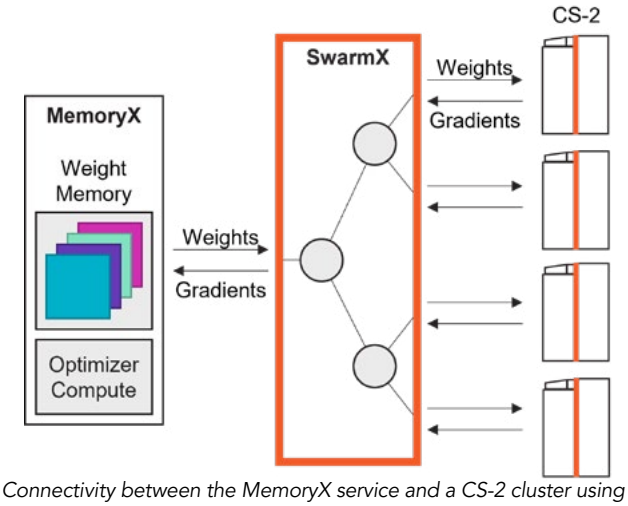


Figure 13. Connectivity between the MemoryX service and a CS-2 cluster using the SwarmX fabric.

Since each CS-2 system is computing a partial gradient for the entire training batch, the gradients from each CS-2 system need to be accumulated prior to the weight update. We chose to perform this reduction inside of the SwarmX fabric, which requires the fabric to perform just one add per element per CS-2. Reducing along the way has the benefit of making the bandwidth requirement symmetrical for weight and gradient communication. One copy of the gradient tensor propagates back through the SwarmX fabric, over each link, to the MemoryX service. Moving the reduction into the SwarmX fabric also has the benefit of providing a clean abstraction. From the perspective of the MemoryX service, the SwarmX fabric looks just like a single CS-2 and from the perspective of the CS-2, the SwarmX looks just like the MemoryX service. This abstraction allows us to use the same compute unit and memory service regardless of cluster size.

The SwarmX fabric is composed of broadcast-reduce nodes each containing a set of 100Gb/s network interfaces. Each broadcast-reduce node provides enough bandwidth to perform either 1:4 broadcast-reduce operations or a pair of 1:2 broadcast-reduce operations. The nodes can be configured in several different modes, which provides flexibility to meet the needs of a particular cluster. Each node provides enough compute to perform the floating-point gradient reductions at line-rate, allowing reductions to occur as gradients flow back to the MemoryX service.

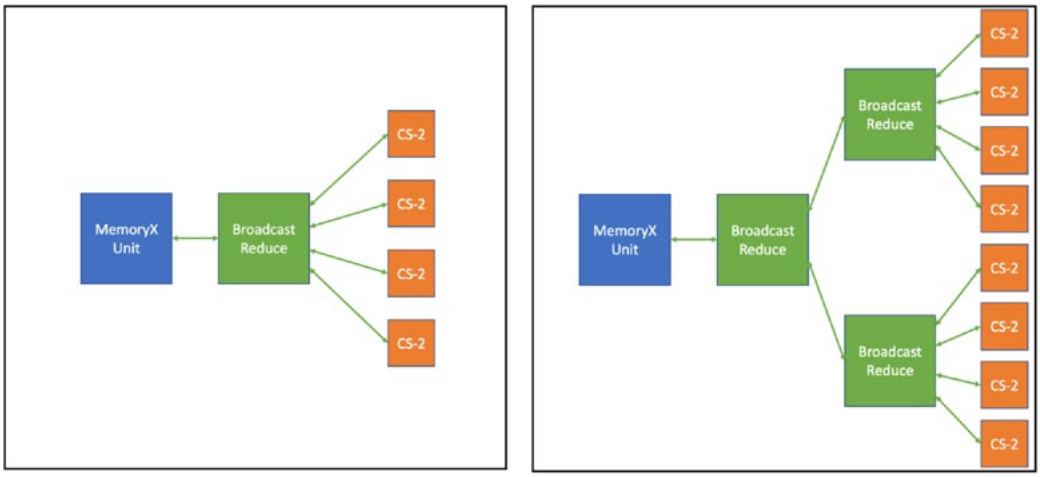


Figure 14 SwarmX fabric connectivity for clusters of 4 and 8 CS-2 systems.

SwarmX nodes are connected in a bidirectional tree topology which minimizes the overall bandwidth and latency required to perform the broadcast and reduction operations (Figure 14). Each CS-2 system has 1.2Tb/s of I/O bandwidth and, in the worst case, needs weights to be delivered at this rate to keep the compute units busy, so the aggregate bandwidth required from the SwarmX fabric increases linearly with N, the number of CS-2 systems. To satisfy this requirement, the number of nodes composing the SwarmX fabric scales linearly with N. Since

tree reductions are work-efficient, the compute required also increases linearly with N, and is delivered by the compute in each broadcast-reduce node. A tree topology also has the benefit of reduced latency, with the latency between the MemoryX service and the CS-2 systems growing logarithmically with N.

7. Principles of Operation

The Cerebras Graph Compiler (CGC) integrates with machine learning frameworks, such as TensorFlow and PyTorch, to compile user models into binaries that can execute on a CS-2 system and supporting cluster. The CGC now natively supports the new weight streaming execution mode. All details of the workload mapping and distribution are handled by the CGC, allowing users to easily bring up existing models on a cluster of CS-2s. As explained above, the weight streaming approach streams weights (one layer at a time) from the MemoryX service to a CS-2 cluster, which computes weight gradients and streams them back to the MemoryX service where weights are updated. One batch of activations remains resident in the CS-2 cluster while the forward and backwards propagation passes occur; then this process is repeated with a new batch streaming in from the data storage service.

7.1. Execution on the Wafer-Scale Engine

At the start of each training iteration, a unique shard of the training batch is downloaded from the data service to each CS-2 system, and then to the WSE-2. This shard serves as the activation input to the first layer in forward propagation. The WSE-2 is responsible for computing activations, activation gradients, and weight gradients for its local shard. Layers run one at a time in the forward and then in the backward pass. For each layer, the sparse weights arrive from the MemoryX service twice, first to trigger compute operations with the activations in the forward pass, and again for the backward pass when the activation gradient from the following layer becomes the input activation gradient for the previous layer. The activations for the local shard are stored on the wafer throughout the training iteration. Stored activation shards, one per layer, that are generated during the forward pass are consumed in the backward pass to compute weight gradients. Weight gradients computed by the WSE-2 in the backward pass are sent to the MemoryX service through the SwarmX fabric, where they are added to the weight gradients of all the other shards, so that the sum of the per-shard weight gradients is finally presented to the MemoryX service.

7.2. Data Layout on the Wafer

Activation and activation gradient tensors are stored in the memory of the WSE-2. Each of the 850,00 cores on the WSE-2 contains 48kB of memory, which is used to store these tensors. The method of distribution of these tensors over the wafer, what we refer to as *data layout*, is an important aspect of execution on the wafer. This is because the arrangement of data determines how much work each core performs, and which communication operations are required. On the WSE-2, we use a rectangular array of interconnected cores to perform the tensor operations found in neural network layers. The matrix multiplication of a fully connected layer is an example. A WSE-2 implementation of a tensor operation specifies its own layout for its weight and activation tensors as well as the local computation and the inter-core communication needed for the implementation. Activation tensors are often distributed across the cores such that an individual activation tensor element is stored on one core. Weight tensors are received from the IO interface along one edge of the rectangle of cores and trigger local computations on each core. The communication often entails broadcast of input tensors across rows or columns of cores and the sum-reduction of partial outputs, also across rows or columns of cores. The software stack allows

for each implementation of a tensor operation to specify its desired data layout, but it is important to use a consistent strategy to reduce data movement as the network executes.

Activation tensors in models like GPT-3 have three dimensions: batch, sequence, and hidden (feature). Our general method of distributing tensor data on the wafer involves splitting one or more of these tensor dimensions over the wafer's x and/or y dimension. For example, if we split the tensor's hidden dimension over the wafer's x dimension, it means that each core in a row of cores has a contiguous chunk of features from the logical tensor, with adjacent cores containing adjacent chunks of features. One goal of this activation layout strategy is to reduce data movement during and between compute operations. Compute operations involving activations or activation gradients are performed on the core containing the corresponding elements of the tensor. This means that elementwise operations can be performed without any data movement. Elements are read from local memory, the compute operation is executed, then the resulting elements are written back to local memory. However, reductions over dimensions of these tensors do require communication between cores containing each chunk of the reduced dimension. Our data layout strategy for NLP models, depicted in Figure 15, optimizes for reductions over the hidden and sequence dimensions since these are the most common. The feature dimension is split over the wafer's x dimension, and the sequence and batch dimensions are split over the wafer's y dimension, with elements from each sequence mapped to adjacent sets of cores. Reductions over the batch dimension are needed for computation of weight gradients, but these operations also require a reduction over the sequence dimension, both of which can be efficiently accomplished with a single sequential reduction over the wafer's y dimension.

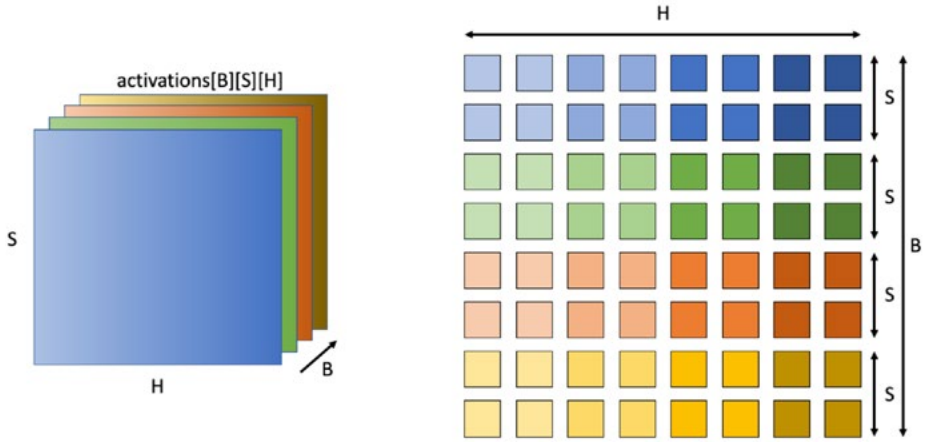


Figure 15. Layout of an activation tensor on the wafer. Chunks of the H dimension are distributed over rows of cores. Chunks of the S and B dimensions are distributed over columns of cores with sequential chunks of the S dimension on adjacent cores.

Activation and activation gradient computations for fully connected layers involve matrix multiplication operations which reduce over the hidden dimension. For an activation computation, where X_0 is the input activation tensor, W is the weight matrix, and X_1 is the output activation tensor:

$$X_1[B \times S][H_1]^T = W[H_1][H_0] \times X_0[B \times S][H_0]^T$$

Where H_0 is the hidden dimension of the input and H_1 is the hidden dimension of the output. This data layout allows for reductions over the hidden dimension to occur between adjacent sets of cores in each row, reducing the bandwidth requirement. Weight gradient computations for

fully connected layers involve matrix multiplication operations which reduce over the batch and sequence dimensions:

$$dW[H_1][H_0] = dX_1[B \times S][H_1]^T \times X_0[B \times S][H_0]$$

Our chosen data layout also allows for reductions over these dimensions to occur between adjacent sets of cores in each column.

Our array of cores is roughly square, but these tensor dimensions are not; the feature dimension used by GPT-3 is 12k, and batch shards typically contain hundreds of thousands of tokens. To better map these disparate dimensions to the wafer, we can also split the batch dimension over the fabric x dimension as shown in Figure 16. This allows us to evenly spread the activation and activation gradient tensors to the entire wafer, while giving each core a roughly square subregion of the tensor. Features from the same sequence in the batch are still placed on adjacent cores within the row, but each row of cores contains data from more than one sequence in the batch.

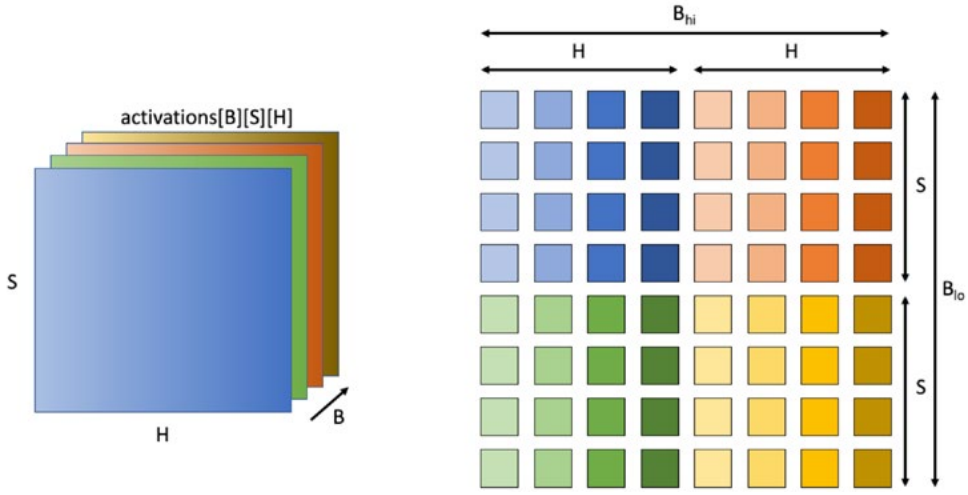


Figure 16. Layout of an activation tensor which splits the B dimension over both spatial dimensions on the wafer. This allows the per-core sub-tensor to be roughly square even when the number of tokens is much greater than the number of features.

7.2.1. Data Layout for Weights and Gradients

Weight and gradient tensors are not stored persistently on the WSE-2 but are streamed on and off the WSE-2 from and to the MemoryX service. Weight data arrives over the on-wafer network connections which link each core to its four adjacent neighbors. Weight data is split over these links such that each core only receives weights needed for computations on its local activations. The number of columns, H_0, in a weight matrix corresponds to the hidden dimension of the input activations. When computing activations for the next layer, the H_0 dimension of the weight matrix is split over the wafer's x dimension to produce the same distribution as is used for the H dimension of the input activations. This is depicted in Figure 17. During the backward pass computations of activation gradients, the number of rows, H_1, in the weight matrix corresponds to the hidden dimension of input activation gradients. As a result, we split the H_1 dimension over the fabric X dimension for this operation. When weight gradients are computed in each column of cores, they are transmitted out of the column with similar distribution to the weights. Gradients are computed one row at a time to match the order of activation computations, so each row of gradients is split along its H_0 dimension over the wafer's x dimension. Computing gradients in the same order as forward pass activations means that the weights of the first layer can be updated in

the same order as they are streamed out, allowing for more efficient pipelining.

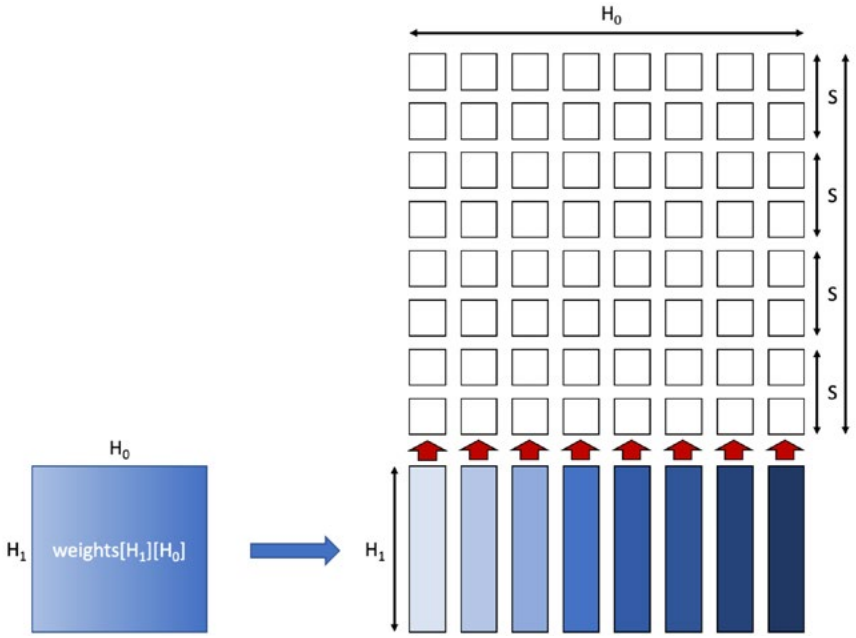


Figure 17. Data layout used for weights arriving during activation computations.

7.3. Wafer-Scale Matrix Multiplication

Matrix multiplication operations dominate the compute used by NLP models such as GPT-3. They are used for all three phases of fully connected layer execution: activation, activation gradient, and weight gradient. We have achieved high utilization on these operations by avoiding bandwidth bottlenecks, through our data layout strategy, and minimizing overheads. There are two flavors of matrix multiplication that are needed to train sparse models: one which supports a single sparse input and dense output and another which supports two dense inputs and a sparse output.

7.3.1. One Sparse Input

The first flavor of matrix multiplication is targeted at activation and activation gradient computations. A dense activation tensor, stored in local memory, is multiplied with a sparse weight tensor streamed into the WSE-2 from the MemoryX service. The resulting dense activation tensor is stored to local memory.

This matrix multiplication operation is broken down into a series of matrix vector multiplications. Each matrix vector operation multiplies one row of the weight tensor with the entire activation tensor in local memory and results in a single vector of the output tensor. As described in the previous section, a row of sparse weights is streamed onto the wafer with the row split such that each weight arrives on the column of cores containing all corresponding activations. Weights are broadcast to all cores in the column using the on-wafer network whose routers support multicast in hardware. As weights arrive at each core, they are multiplied with the corresponding feature for each token in the core's subset of the batch and accumulated into a temporary buffer. After all weights for the row have been processed, each core contains a partial sum which must be reduced with all cores in the row to compute the result. Partial sums are reduced over a ring using the on-wafer network, with the result landing on the column of cores which should store the feature corresponding to the received row of the weight matrix. These broadcast and reduction communication patterns are shown in Figure 18.

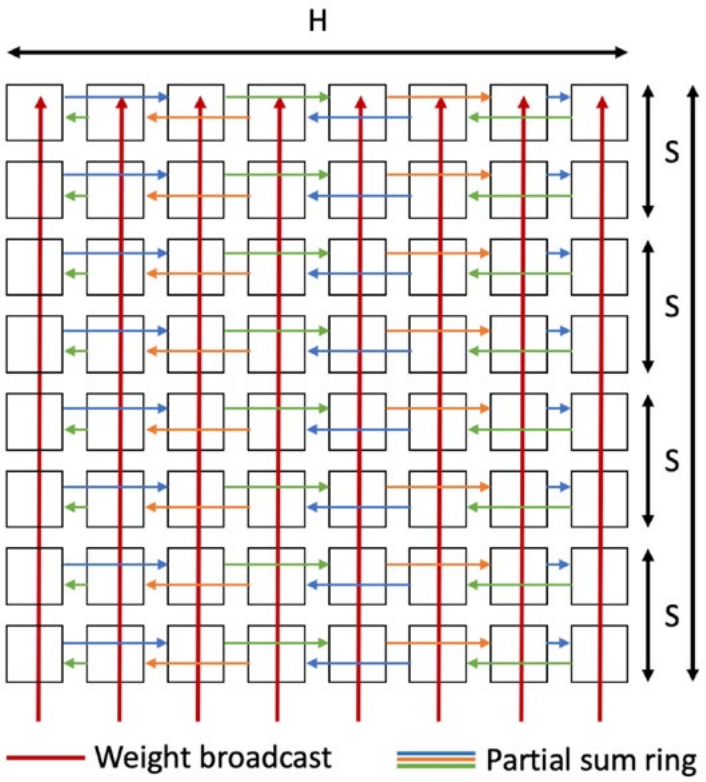


Figure 18. Inter-core communication patterns used for matrix multiply.

Tensor operations on the wafer are driven by the arrival of weight data or commands. The arrival of weight data triggers a floating point multiply-accumulate operation between the received weight data and corresponding row of activations. Commands are sent by the MemoryX service's coordinator, as shown in Figure 19, which drives execution of each tensor operation, and are used to trigger other operations such as the partial sum reduction. When the cores receive a partial sum command, they initiate communication with their upstream neighbor, reducing incoming partial sums with the values in their local accumulators and transmitting the results to their downstream neighbor. The fully reduced values are received and stored on one column of cores indicated by a special argument to the partial sum command sent to that column. Commands can be used to trigger other computations on the wafer such as nonlinear functions or normalization operations. This system allows us to support all tensor operations required by NLP models.

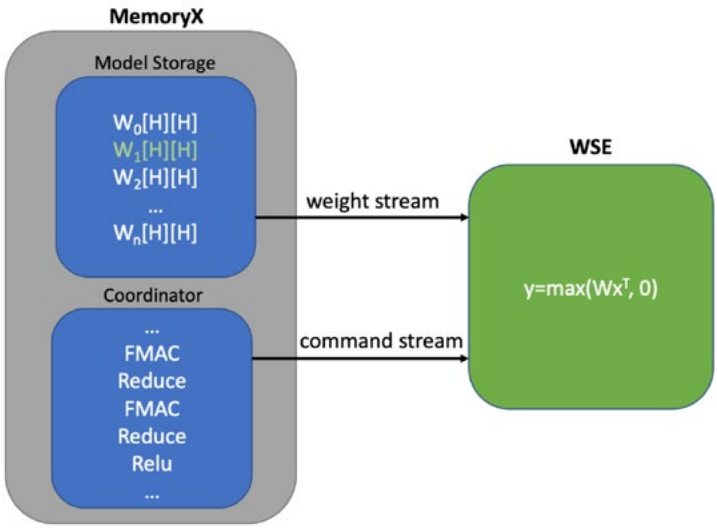


Figure 19. Inputs from the MemoryX service to the WSE which drive execution of each tensor operation.

7.3.2. Sparse Output

The second flavor of matrix multiplication is intended for computing weight gradients. A dense activation tensor and dense activation gradient tensor, both stored in wafer memory, are multiplied to compute a sparse weight gradient tensor which is sent off-wafer.

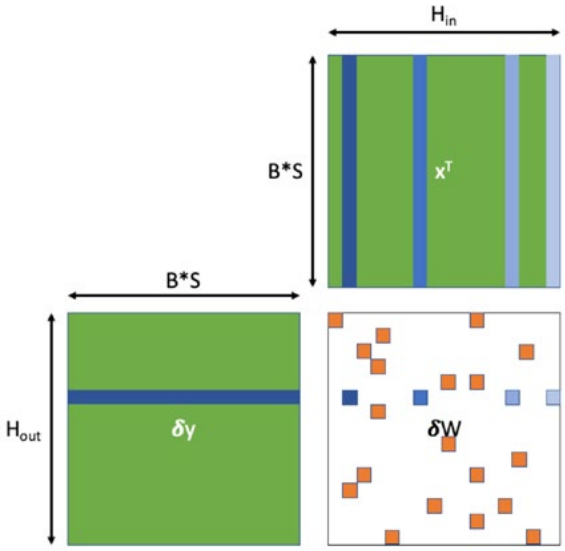


Figure 20. Matrix multiplication with a sparse output.

Since the Cerebras weight streaming implementation supports unstructured weight and gradient sparsity, this matrix multiplication operation is computed one output element at a time on each core. Computation of each weight gradient involves a dot product between the corresponding

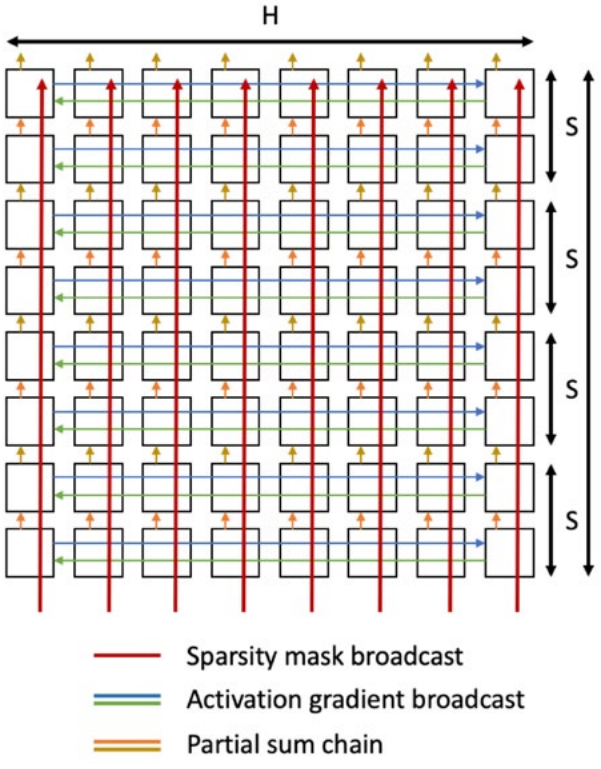


Figure 21 Inter-core communication patterns used for gradient variant of matrix multiply.

activation and activation gradient vectors, as shown in Figure 20. The MemoryX service sends a sparsity mask to the WSE-2 as input, which instructs the wafer to compute a gradient for each masked position. Gradients are computed for each row of the weight matrix in order, so this mask is received one row at a time. Each row of the weight matrix corresponds to a single output feature. The row of the activation gradient tensor corresponding to this feature is broadcast to all cores in each row of cores using the blue and green routes shown in Figure 21. An element of the output gradient tensor is then computed by multiplying this broadcasted vector with the feature from the activation tensor corresponding to the weight gradient's column index. The sparsity mask input is split over the fabric the same way that the weight matrix is split, such that masked elements are received on the core containing the corresponding activation feature data. This means that the broadcasted gradient vector can be multiplied with activation data from local memory to compute a partial sum of the gradient element. A chain routing pattern, also shown in Figure 21, over the on-wafer network is used to accomplish the reduction of these partial sums between cores in the same column, containing the same feature for different tokens in the batch.

7.4. Execution on the MemoryX Service

Model parameters and optimizer state are stored persistently on the MemoryX service during training in weight streaming mode. The MemoryX service performs three primary functions corresponding to the three phases of training: weight streaming, gradient receipt, and weight update. These operations and the weight and gradient streaming interfaces are shown in Figure 22. During the forward and backward passes, the MemoryX service streams weights out to the WSE-2. In the backward pass, the MemoryX service sends a sparsity mask for each layer as a request for the wafer to compute gradients. It then receives gradients back from the WSE-2 and updates the weights before the next training iteration begins.

The MemoryX service stores weights in both dense and sparse formats. A dense copy of the weight and optimizer state is needed for most sparsity algorithms, since they need the ability to regrow the weights. The dense format includes FP32 values for all weights and per-weight optimizer states such as momentums. A compressed sparse row (CSR) representation is used for the sparse copy of the weights, using FP16 for the weight values and a 16-bit delta-encoded integer for the index. The dense FP32 format is used for performing weight updates, but only the active sparse weights are updated. Updated weights are converted into the sparse format by applying a sparsity mask and rounding values from FP32 to FP16. During the forward and backward passes, it is the sparse format of the weights which is sent to the WSE-2.

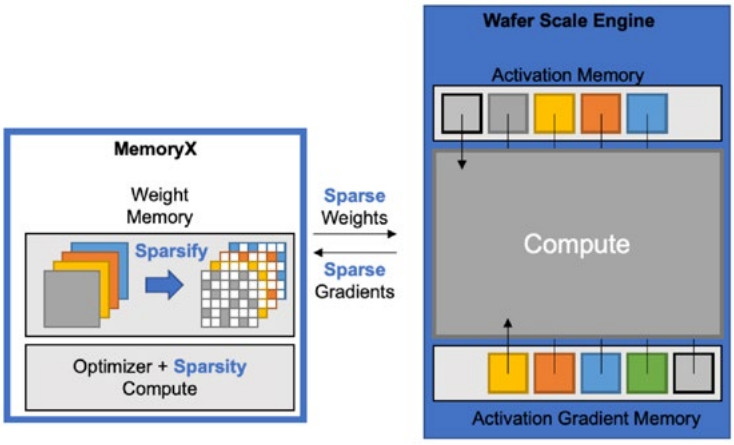


Figure 22. Data stored on the MemoryX service and the WSE and communication between the two.

Weights can be streamed out of the MemoryX service in either forward or backward order. As described in the previous section, the WSE-2 computes GEMM operations, corresponding to activation or activation gradient computations, as a series of GEMV operations. One feature of the output tensor is computed at a time. In the forward pass, the output features correspond to rows of the weight matrix, so the MemoryX service streams sparse weights in row-major order so that the wafer receives full rows of the matrix one at time. This order is flipped in the backward pass when the WSE-2 is computing one row of the activation gradients, corresponding to the previous layer's features, at a time. Supporting both orders requires the MemoryX service to transpose the sparsity pattern of the weight matrix each time the sparsity pattern is updated. The forward and transpose orders both provide views to the same underlying data, so that only one copy of weights needs to be updated.

During the backward pass, weight gradients are requested and received from the WSE-2 by the MemoryX service. Computation of weight gradients on the WSE-2 is triggered by arrival of a sparsity mask indicating which gradient values should be computed. The sparsity mask is sent by the MemoryX service and has the same format as the index component of the forward-order weights. However, the MemoryX service can choose to use a different sparsity mask than that of the weight matrix to trigger computation of gradients corresponding to zero-valued weights. This can be useful when trying to change the sparsity of the weight matrix. For example, the MemoryX service can change a zero-valued weight to a non-zero when its gradient value exceeds some threshold. FP32 gradient values are sent back from the WSE-2 to the MemoryX service in the order that the sparsity mask was sent.

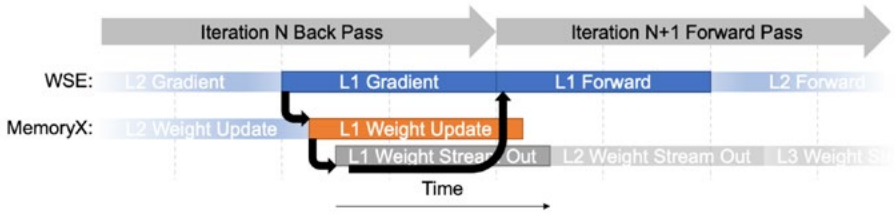


Figure 23. Pipelining of weight and gradient communication between the MemoryX service and the WSE-2.

The MemoryX service performs weight updates after gradients have been received. Scheduling in the MemoryX service is designed to avoid a latency bottleneck in the training loop. While storing model parameters farther from the compute nodes results in a higher communication latency, this is hidden by pipelining which is depicted in Figure 23. In the forward pass, the MemoryX service begins streaming weights for a layer before the WSE-2 has completed activation computations for the previous layer. The same approach is used in the backward pass, but in reverse with transposed weights. While the MemoryX service is streaming weights in transpose order for the backward pass, it is also collecting gradients being streamed back from the WSE-2. Weight updates can be applied as gradients are received, such that the new weights are ready for the next training iteration as soon as gradients are received for the first layer. The critical path is for the weights of the first layer, as these are the last to receive gradients and the first needed in the forward pass of the next training iteration. To mitigate this, the WSE-2 computes gradients for weights in the same order as weights are transmitted for forward pass operations. This means that updated weights for the next training iteration can be streamed out before the WSE-2 has completed gradient computations for the first layer. The round-trip latency between the MemoryX service and the WSE-2 can be completely hidden when the compute time of the first layer is greater than the round-trip latency.

8. Experimental Results

Since Cerebras weight streaming technology was announced in 2021, considerable progress has been made in building reference systems, measuring performance and joint work with partners.

This section contains summaries of some of the recent work in cluster scaling performance characterization, image segmentation for computer vision and award-winning research in COVID full-genome models.

Learn more about published weight streaming results

- Linear Scaling Made Possible with Weight Streaming
- Unlocking High-Resolution Computer Vision with Wafer-Scale Technology
- Cerebras Wafer-Scale Cluster Trains Large Language Models on the Full COVID Genome Sequence
- Andromeda: a 13.5 Million Core AI Supercomputer

8.1. Near-Linear Scaling Demonstrated Across a Range of GPT-Style Models

As we've discussed, our wafer-scale processor, working in tandem with MemoryX and SwarmX technologies, enables weight streaming by disaggregating memory and compute. We have demonstrated that we achieve near-linear scaling for Cerebras Wafer-Scale Clusters up to 64 CS-2s – that's perfect scaling up to 54 million cores – across a range of GPT-style model configurations (Figure 24). So, to go 10 times as fast as a single CS-2, you need exactly 10. and huge performance gains without needing to think about complex distributed compute challenges.

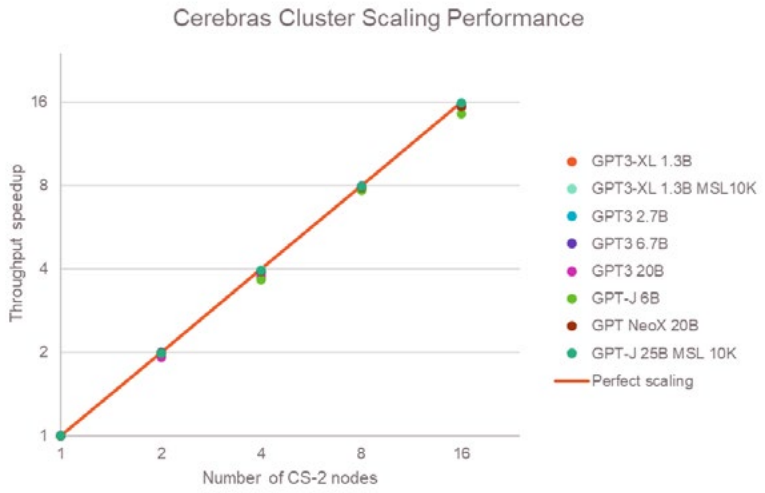


Figure 24. We have shown that weight streaming allows the Cerebras Wafer-Scale Cluster to achieve almost-perfect linear scaling performance for a range of GPT-style models.

8.1.1. The Andromeda AI Supercomputer

To produce the remarkable scaling results reported above, Cerebras constructed a Cerebras Wafer-Scale Cluster called Andromeda (Figure 25). Andromeda consists of 64 CS-2 nodes, for a total of more than 54 million AI-optimized compute cores capable of more than 1 ExaFLOP of sparse AI compute. As reported above, unlike any known GPU-based cluster, we measured near-perfect scaling across GPT-class large language models, including GPT-3, GPT-J and GPT-NeoX.

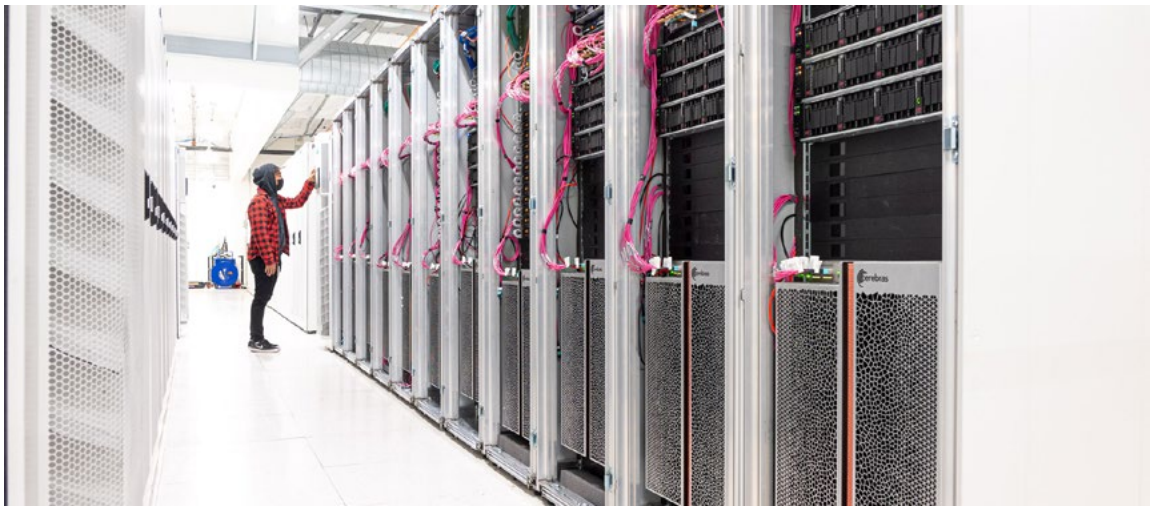


Figure 25. Part of Andromeda, one of the largest AI supercomputers yet constructed.

8.2. Image Segmentation on 25 Megapixel Images

The same characteristics that enable weight streaming to accelerate training of large, autogressive large language models also have applications in the computer vision field.

Breakthrough performance achieved with convolutional neural network models (CNNs) for computer vision is what triggered the current "AI summer". The triggers were larger networks (more layers), new network designs (like ResNet), and fast hardware with which to train these.

However, segmentation and classification for large images or volumes stress today's AI compute systems. They need enough memory to store the weights and activations. For example, the popular 3D U-Net model, when training on 512^3 sized volumes requires 64 gigabytes of storage for intermediate activations from the forward pass per image.

To grow memory and compute capacity, implementers have used large clusters of GPUs. But it has proven difficult to split a training job across large clusters and achieve an efficient, stable, successful training run.

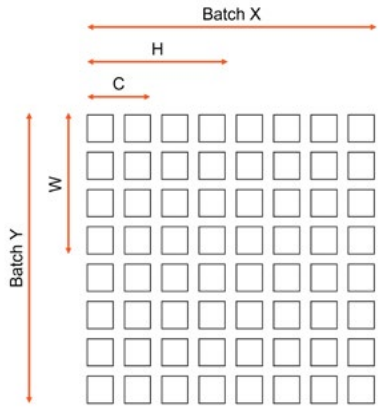


Figure 26. Spatial Data Layout for Multi-Dimensional Convolution tensor on the chip. This is shown for a you example with 8×8 grid of PEs. Actual Cerebras Wafer Scale Chip comprises of 850,000 PEs arranged in a rectangular grid.

In our weight streaming architecture, convolution has been parallelized by mapping these dimensions either spatially, meaning across different axes of the wafer, or temporally, performing operations one after another rather than in parallel due to the streaming nature of weights. Figure

26 shows the mapping of the spatial dimensions to the two wafer axes. The Cerebras Graph Compiler takes care of picking the functionally correct and best performing layout choice for all layers in a CNN.

We have demonstrated that weight streaming enables us to overcome the limitations of GPUs and allow users to easily and rapidly train large models on high-resolution, 25-megapixel images (Figure 27). With the exploding availability of high-quality, high-resolution data, researchers now have an easy way to train deep neural networks on large images and take advantage of their rich contextual information in fields as diverse as medical imaging and remote sensing via satellite.

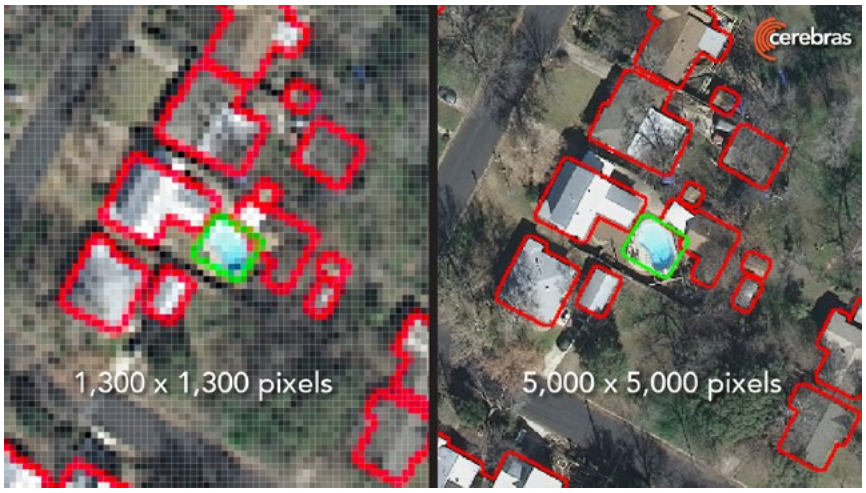


Figure 27. Illustration showing the benefits of high-resolution imagery for computer vision tasks.

8.3. Training Large Language Models on the Full COVID Genome Sequence

Over the last several years, Cerebras has been engaged in a portfolio of ongoing work using AI models to accelerate COVID molecular modeling[13] and microbial gene Transformers[14] with Argonne National Laboratory (ANL). Of particular interest is a project to train full-genome models enabled by the unique long sequence length capabilities of the CS-2. The challenge of training these complex models was addressed using weight streaming to accelerate training using a Cerebras Wafer-Scale Cluster.

In genomics, a physical characteristic (or phenotype) can be coded for by genes that are far apart in the genetic sequence. The SARS-CoV-2 virus genome, which causes COVID-19, can have a sequence of 30,000 nucleotides or 10,000 codons. LLMs can make sense of genomic sequences, but their self-attention mechanism, which can relate different positions of a sequence, is severely limited by conventional computer systems. Fortunately, the Cerebras CS-2 has the capacity to train state-of-the-art LLMs with input sequences longer than the SARS-CoV-2 genome. That allows researchers to study genomics from a whole new perspective.

Cerebras engineers worked with the lab of Dr. Arvind Ramanathan at Argonne National Laboratory (ANL) to develop genome-scale language models (GenSLM) that reveal evolutionary dynamics of SARS-CoV-2. For the first time, we trained LLMs using between 123 million and 25 billion parameters with the full SARS-CoV-2 genome of sequence length 10,240, on a single CS-2 system, and to speed up the training we used as many as 16 CS-2s in a Cerebras Wafer-Scale Cluster (Figure 28). This work resulted in a paper, "*GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics*"[14], which was presented at SC22.

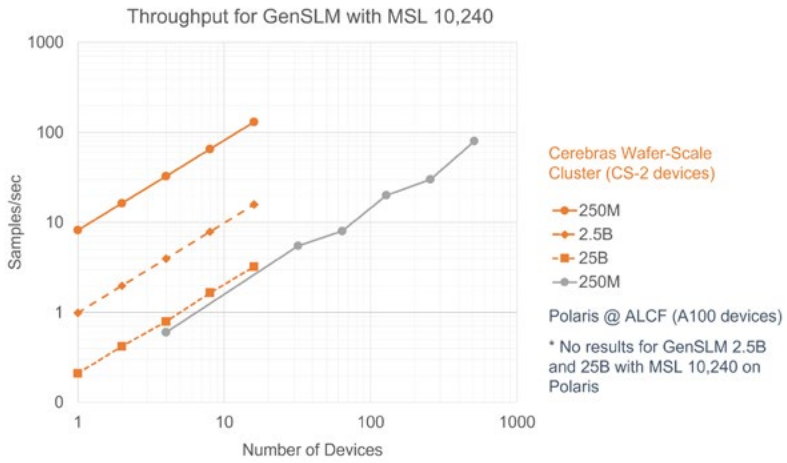


Figure 28. Linear throughput scaling of GPT-J up to 25B parameters with MSL 10,240 on Cerebras Wafer-Scale Cluster, and throughput comparison of GPT-J 250M between Cerebras and the GPU-equipped Polaris supercomputer at the Argonne Leadership Computing Facility.

This joint work with Argonne National Laboratory (ANL) and NVIDIA was honored with the 2022 Gordon Bell Special Prize for HPC-Based COVID-19 Research (Figure 29).[15]

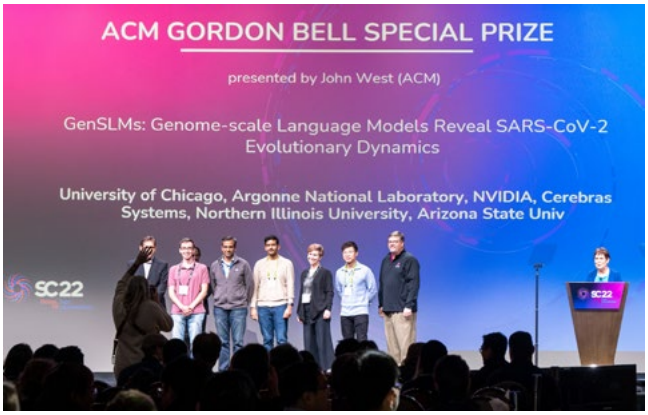


Figure 29. The awards ceremony at SC22 where the GenSLMs paper, for which weight streaming was an Important enabling technology, was honored with the 2022 ACM Gordon Bell Special Prize.

9. Summary

We have presented weight streaming as a new paradigm for training giant models. Weight streaming disaggregates the storage of parameters from the compute units. We described an implementation, i.e. the Cerebras weight streaming architecture, that is based on wafer-scale compute units; a new system, the MemoryX service, for parameter storage and update; and a novel interconnection, SwarmX fabric, between parameter memory and compute.

The architecture is designed around the Cerebras Wafer-Scale Engine processor which provides enough compute power and on-wafer SRAM to support layer sizes an order of magnitude greater than those used in today's state-of-the-art models. Since each WSE can support massive layers, our architecture is able to use a scale-out model based on pure data parallelism, which can support a cluster delivering more floating-point performance than the current largest supercomputer in the world for this class of workloads.

The WSE is also a preferred platform because, unlike units that prefer dense matrix multiplication, it can fully exploit sparsity in the weight tensors, for a one to two orders of magnitude reduction in computational work and runtime. We achieve runtime reduction nearly linear in number of nonzero weights for unstructured weight sparsity.

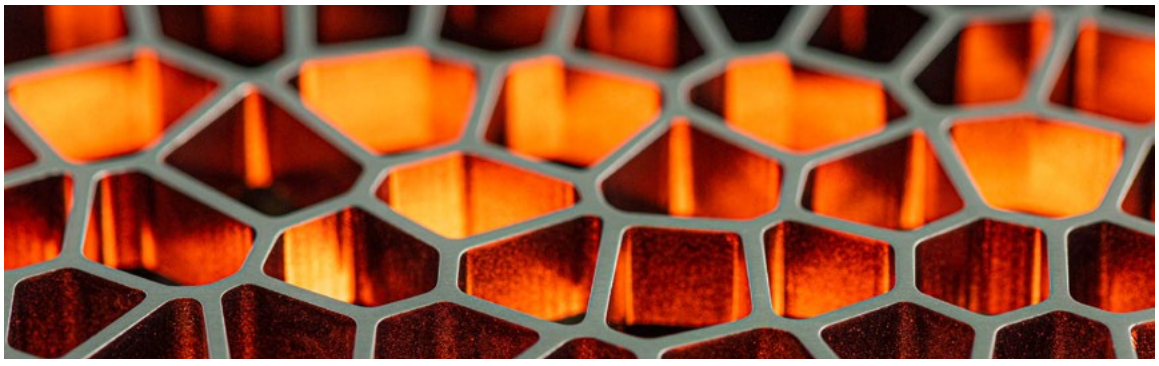
We contend that the combination of effective sparsity, compute units capable of storing full layers, and memory disaggregation gives ML practitioners the only practical way to train models with trillions of parameters.

10. References

1. Deepak Narayanan, Mohammad Shoeybi, Jared Casper, et al, "Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM", 2021, https://arxiv.org/abs/2104.04473
2. Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, et al, "ZeRO-Offload: Democratizing Billion-Scale Model Training", 2021, https://arxiv.org/abs/2101.06840
3. Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language Models are Unsupervised Multitask Learners", 2019, https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe
4. Mohammad Shoeybi, Mostofa Patwary, Raul Puri, et al. "Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism, 2019, https://arxiv.org/abs/1909.08053
5. Corby Rosset. "Turing-NLG: A 17-billion-parameter language model by Microsoft", https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/, 2020
6. Paresh Kharya and Ali Alvi. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model. https://developer.nvidia.com/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/ , 2021
7. Jonathan Frankle and Michael Carbin. "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks", ICLR 2019, https://arxiv.org/abs/1803.03635
8. Vithursan Thangarasa, "Accelerating Large GPT Training with Sparse Pre-Training and Dense Fine-Tuning", Cerebras blog, 2022, www.cerebras.net/blog/accelerating-large-gpt-training-with-sparse-pre-training-and-dense-fine-tuning/
9. Leo Gao, Stella Biderman, Sid Black, et al, "The Pile, an 800GB Dataset of Diverse Text for Language Modeling", 2020, https://arxiv.org/abs/2101.00027
10. Curation Corpus, GitHub, 2020, https://github.com/CurationCorp/curation-corpus
11. Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, et al, "The LAMBADA dataset: Word prediction requiring a broad discourse context." Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics 2016, https://aclanthology.org/P16-1144/
12. Stephen Merity, Caiming Xiong, James Bradbury, Richard Socher, "Pointer Sentinel Mixture Models", arXiv 2016, https://arxiv.org/abs/1609.07843
13. Anda Trifan, Defne Gorgun, Zongyi Li, et al, "Intelligent Resolution: Integrating Cryo-EM with AI-driven Multi-resolution Simulations to Observe the SARS-CoV-2 Replication-Transcription Machinery in Action", bioRxiv, 2011, https://www.biorxiv.org/content/10.1101/2021.10.09.463779v1
14. Maxim Zvyagin, Alexander Brace, Kyle Hippe, Yuntian Deng, Bin Zhang, Cindy Orozco Bohorquez, et al, "GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics", SC22, 2022, https://www.biorxiv.org/content/10.1101/2022.10.10.511571v2
15. ACM Press Release "ACM Gordon Bell Special Prize for HPC-Based COVID-19 Research Awarded to Team for Modelling How Pandemic-Causing Viruses, Especially SARS-CoV-2, are Identified and Classified", 2022, https://www.acm.org/media-center/2022/november/gordon-bell-special-prize-covid-research-2022

11. Starting your collaboration with Cerebras

To explore how Cerebras systems can accelerate your research or to see a demo, please contact us at www.cerebras.net/get-demo.



cerebras

CEREBRAS SYSTEMS, INC.
1237 E. ARQUES AVE, SUNNYVALE, CA 94085 USA
CEREBRAS.NET