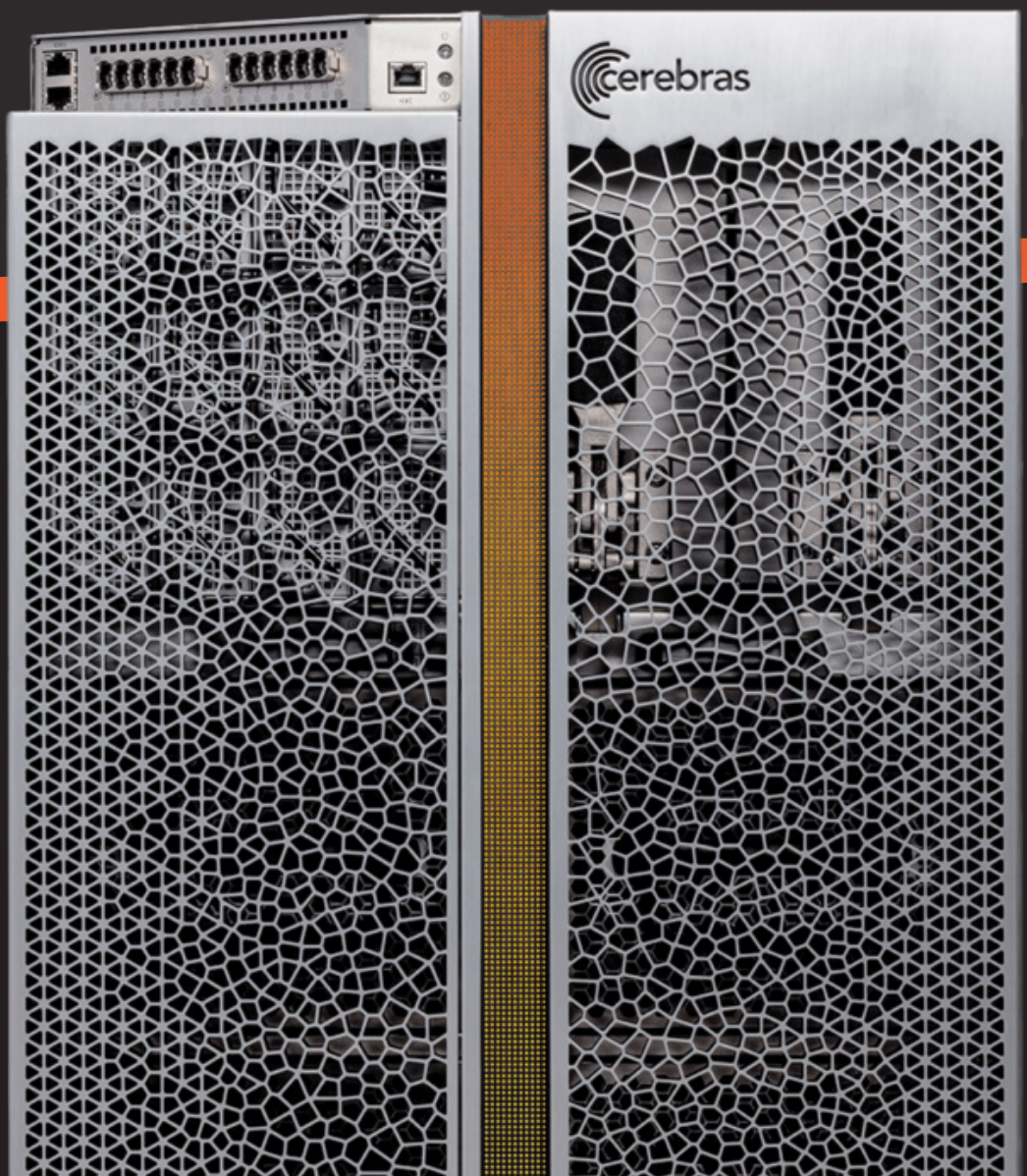
cerebras

Train Large BERT Models Faster

with Cerebras Systems



Scaling BERT Efficiently with Cerebras Systems

Unstructured text is one of the largest human-generated data sources. Web data, academic publications, emails, traditional media, texts, instant messages, digital records, social media — all hold an enormous volume of unstructured text.

This ever-expanding reservoir of text is a treasure trove of valuable data if it can be organized, synthesized, patterns identified and properly mined for insight. This is what Natural Language Processing (NLP) does. NLP enables specialized computers to quickly make sense of this growing volume of data. It enables us to identify key ideas and topics, uncover emerging trends, analyze sentiment and identify correlations that would be impossible for human readers to identify.

The power of NLP is far-reaching and provides valuable results for commercial enterprises and government agencies alike. Today, the dominant deep learning model used for NLP is BERT (Bidirectional Encoder Representations from Transformers[1]). In its original form, it helped computers to understand written texts and understand the meaning of words in the context of a full sentence or paragraph.

1 https://arxiv.org/abs/1810.04805

While many companies use BERT models to better analyze text, the model has a growing list of applications well beyond the natural language domain. BERT can be used to model any text or sequential data, such as numeric time series, nucleic acid sequences, or protein sequences for example. Since its original publication, the artificial intelligence (AI) community has built and used a variety of BERT derivatives for domain-specific applications. These include:

- **BioBERT**[2] for biomedical text mining
- **FinBERT**[3] for financial sentiment analysis
- **SciBERT**[4] for scientific and computer science texts
- **ClinicalBERT**[5] for clinical notes modeling and prediction of hospital readmission
- **GilBERT**[6] for geologically-informed language modeling in oil and gas
- **DNABERT**[7] for genomic sequence analysis
- **PatentBERT**[8] for patent classification and search
- **mBERT**; a variant of the original BERT model for multi-lingual understanding

The idea here was rather than train BERT on a general body of language data, train the model on domain-specific text, and thereby obtain much higher accuracy. Not surprisingly, it was quickly discovered that domain-specific BERT models significantly outperform models trained on general-purpose corpora for domain-specific tasks. Out-performing meant that BERT models trained on the unique dialects or vocabularies, then produced vastly higher accuracy in inference for work in that field. However, BERT models still face many challenges. Cerebras Systems designed solutions that overcome these challenges while improving BERT training and time to solution.

The Need for a Better Solution

Despite overwhelming evidence that training large BERT-type models on enormous, domain-specific datasets produces higher accuracy results, few organizations do it. Why? Because it is challenging, time consuming and expensive. Cerebras Systems took note of these challenges and created a system that makes BERT models more accessible.

Computational Challenges of BERT

Training complex models like BERT on legacy graphics processing units (GPUs) requires setting up clusters of machines, specialized programming expertise and days or weeks of time for each training

2 https://arxiv.org/abs/1901.08746
3 https://arxiv.org/abs/1908.10063
4 https://arxiv.org/abs/1903.10676
5 https://arxiv.org/abs/1904.05342
6 https://openreview.net/pdf?id=SJgazaq5Ir
7 https://www.biorxiv.org/content/10.1101/2020.09.17.301879v1.full
8 https://arxiv.org/abs/1906.02124

run. In one of the landmark works on pre-trained BERT models for biomedical language representation, BioBERT[ii], authors noted the prohibitive computational challenges posed by large NLP models and the implications this has on research in the field:

> Despite our best efforts to use BERT-Large, we used only BERT-Base due to the computational complexity of BERT-Large.

It is noteworthy that this comes from some of the world's most sophisticated AI researchers. This remains a challenge today as evidenced by work forthcoming this year on efficient BERT model training[9] by researchers at Microsoft and the University of Texas who note that:

> Despite the impressive empirical success [of large-scale pre-trained language models such as BERT], their computational inefficiency has become an acute drawback in practice. As more and more transformer layers are stacked with larger self-attention blocks, model complexity increases rapidly. ... Such high model complexity calls for expensive computational resources and extremely long training time.

The long set up, programming, and optimization times involved when clusters of graphic processing units are used for this work mean that ML practitioners spend large portions of their time not doing ML work. They are setting up or tearing down clusters, waiting for others to do so, or carefully tweaking their code to get it to work on a cumbersome array of tiny machines. Frustratingly, models often still take days or weeks to train. When models take this long to train, ML research is no longer iterative; instead, it is slow and episodic.

As a result of the complexity, time-consuming nature, and high cost of clusters of graphic processing units, many organizations that could benefit from domain-specific BERT models for natural language and other sequence data processing applications simply can't afford to do so.

The Cerebras Systems Solution

Cerebras designed solutions to overcome challenges to using BERT models. The new Cerebras CS-2 system dramatically reduces the complexity of setting up and the time it takes to train a large model. It makes high-performance, deep learning compute more accessible for organizations.

---

9 https://arxiv.org/pdf/2101.00063.pdf

The original Cerebras CS-1 system was launched in November of 2019 and was the world's most powerful solution for AI computing with the largest chip ever made. In April of 2021, Cerebras Systems announced the CS-2 system, based on our second-generation wafer-scale engine, the WSE-2, which more than doubled the CS-1's performance. The CS-2 brings 850,000 AI-optimized processor cores, 40 GB of high-performance on-wafer memory, 20 Petabytes of memory bandwidth, and 220 Petabits of core-to-core communication bandwidth.

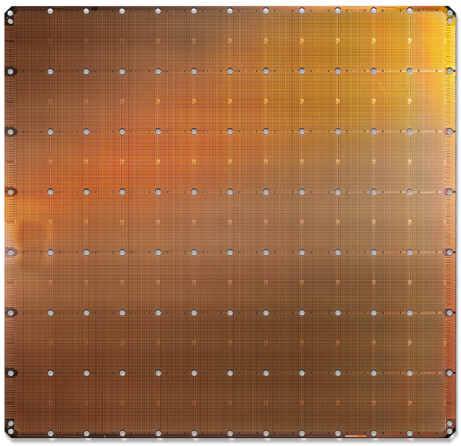
By way of comparison to the NVIDIA A100 GPU, the WSE-2 device is 56x times larger, has 123x more AI compute cores, 1,000x more on-chip memory, and 12,733x more memory bandwidth — all to execute your AI work far faster, at higher efficiency. See **Figure 1**.

	Cerebras WSE-2	NIVIDIA A100	Cerebras Advantage
Chip Size	46,225 mm^2	826 mm^2	56x
Cores	850,000	6912 + 432	123x
On-chip memory	40 Gigabytes	40 Megabytes	1,000x
Memory bandwidth	20 Petabytes/sec	1.6 Terabytes/sec	12,733x
Fabric Bandwidth	220 Petabytes/sec	600 Gigabytes/sec	45,833x

Figure 1. Specifications of the Cerebras WSE-2 for deep learning compared to the NVIDIA A100 graphics processing unit.

This collection of performance attributes allows the CS-2 system to execute deep learning compute on BERT models at far greater performance and efficiency than legacy, general-purpose processors such as GPUs. The CS-2 typically reduces time to solution for AI work by orders of magnitude; delivering wall-clock compute greater than an entire cluster (dozens to 100s) of GPUs, at a fraction of the space and power.

In addition to faster wall-clock compute times, the CS-2 offers a simpler programming model: this cluster-scale compute device is as easy to program as a single desktop machine. The CS-2 is programmable as a single node with standard ML frameworks like TensorFlow and PyTorch, enabling faster setup and iteration for deep learning work —more details in our Ease of Programming section below. This combination of high-performance and straightforward programming allows you to scale up performance quickly and radically reduce wall-clock training time and overall time to solution.



Scaling BERT Training and Time to Solution with Cerebras Systems

Faster solution times mean that machine learning researchers and engineers are able to reach learning faster and at a more efficient cost. However, GPU clusters are complex, time consuming and require extensive software modifications. In the end, these clusters still have poor performance training times. Cerebras solutions allow ML researchers to program easier and train faster to reach a solution faster.

Training Time

Faster training times expand the possibilities for deep learning. In the following section, we report on customer findings for training BERT and BERT-like models across industries.

In the first work with a web customer, the Cerebras CS-1 was benchmarked against a large BERT-style model running on an NVIDIA DGX-A100 (running 8x A100 GPUs). As the findings below show, a single CS-1 was 9.5x faster than the NVIDIA DGX-A100 reducing end-to-end pre-training from over 9 days on the DGX-A100 (218.5 hours) to less than 1 day on a single, first-generation CS-1 (23.1 hours; see Figure 2).

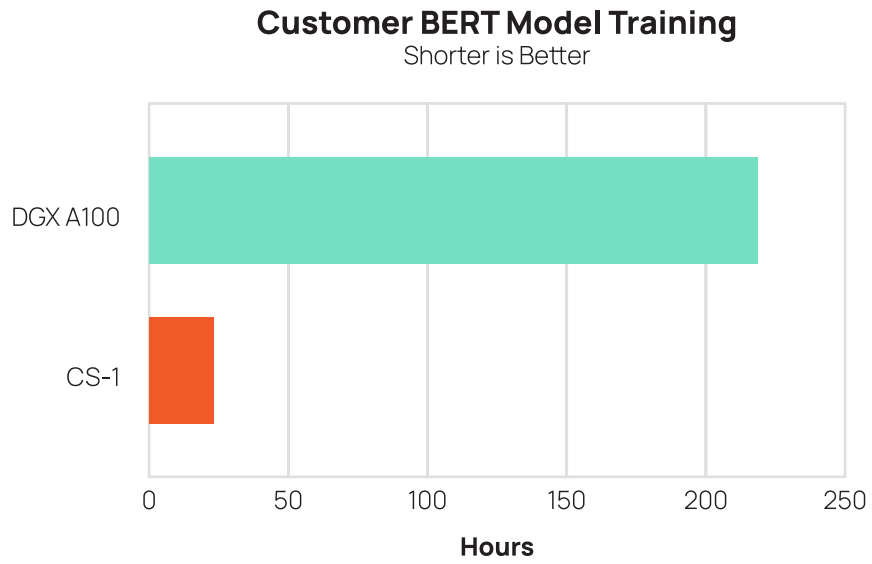


Figure 2. Wall-clock pre-training time CS-1 vs DGX-A100 for a customer BERT model, same TensorFlow model and data. Shorter wall-clock training time is better.

This is a major advantage in wall-clock training time, but an even greater advantage in compute. It would be tempting to then say, for example, that a single CS-1 delivered the compute performance of 76 NVIDIA A100's because it was 9.5x faster than an 8 A100 DGX system. But this would be inaccurate. GPUs do not scale linearly.

Adding a second set of 8 GPUs does not produce 2x wall-clock acceleration of the first 8 GPUs. In fact, data shows that it produces on the order of 1.6x. The programming challenges of scaling and the sublinear characteristics of compute scaling are well known and discussed extensively below. The result, however, is that a single CS-1 outperforms a cluster of more than 100 GPUs. And the CS-2 is twice as fast.

In addition to the known sublinear performance scaling of clusters of graphics processing units, building large clusters of small processors is complex, time consuming and requires extensive modifications to software — this includes the programming model for the machine learning model itself. The minibatch size needs to be changed, hyperparameters need to be changed and the learning rate needs to be changed. A different version of TensorFlow or PyTorch also needs to be used to enable a model to run on a cluster of graphics processing units.

However, **since Cerebras Systems designs single, high-performance machines, reaching cluster-scale acceleration requires no modifications to the model. There is no need to do extensive hyperparameter tuning and no need to increase batch sizes.** The BERT model written for a single GPU can be run on the CS-1 or CS-2 by typing only a few lines of code.

The combination of ease-of-use and blisteringly fast performance enables researchers working with large NLP models like domain-specific BERTs on Cerebras solutions to achieve higher accuracy results out of the box for their downstream tasks. For our customers, faster training time translates to faster and more cost-effective deep learning research on new model architectures and datasets. This also unlocks higher cadence re-training so that production models better match the evolving statistics of our customers' user data. As one of our customers working in cancer research notes,

> We want to be able to train these models fast enough so the scientist that's doing the training still remembers what the question was when they started.
>
> Rick Stevens, Argonne National Lab's Associate Lab Director for computing, environment, and life sciences (CELS)[10]

As mentioned above, 9.5x faster wall-clock training time represents an even greater compute advantage of the CS-1 over the DGX-A100 because wall-clock training time scales sub-linearly on GPU systems. In other words, if you want to train a model 2 or 3 times faster, you need more than 2 or 3 more GPU systems. We can expect even faster training time and greater advantage in models like this with the CS-2 system, as it is more than 2x the performance of the CS-1.

10 https://www.technologyreview.com/2019/11/20/75132/ai-chip-cerebras-argonne-cancer-drug-development/

We can see this empirically in **Figure 3**, which shows the actual BERT wall-clock training time scaling achieved by NVIDIA DGX-A100 systems based on their results reported in MLPerf Training v0.7[11]. Ideal linear scaling is shown in orange, reported actual wall-clock scaling factor is shown in blue.

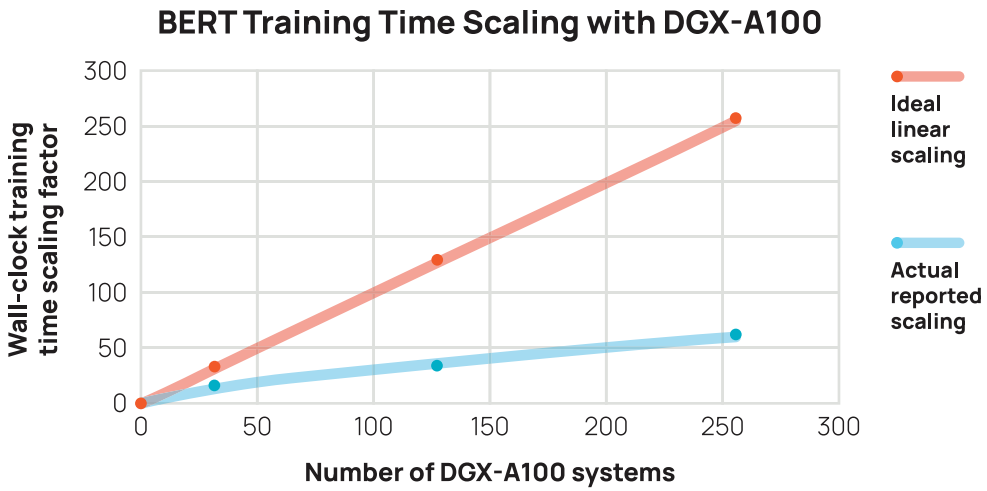


Figure 3. Actual wall-clock BERT training time scaling using DGX-A100 systems (blue) compared to ideal linear scaling (orange). Achieved multi-DGX system scaling is significantly sublinear and approximately follows a power law trend as more systems are added.

Using this data, we can show that for this type of workload, the CS-1 delivers 9.5x faster training which is the wall-clock compute equivalent of approximately 21-22 DGX-A100s. So in this case, our customer would need to buy 21-22 DGX-A100 systems to achieve the same training time they achieved with a single CS-1 machine — that is approximately 168-176 A100 GPUs.

For our customers, faster training time translates to faster and more cost-effective deep learning research on new model architectures and datasets. This unlocks higher cadence re-training so that production models better match the evolving statistics of our customers' user data. We see this type of performance gain in other industries as well, such as with pharmaceutical giant AstraZeneca. According to Nick Brown, Head of AI Engineering[12]:

> Cerebras opens the possibility to accelerate our AI efforts, ultimately helping us understand where to make strategic investments in AI. Training which historically took over 2 weeks to run on a large cluster of GPUs was accomplished in just over 2 days — 52hrs to be exact... This could allow us to iterate more frequently and get much more accurate answers, orders of magnitude faster.

11 MLPerf v0.7 Training NLP benchmark BERT training on Wikipedia data. MLPerf name and logo are trademarks. See www.mlperf.org for more information.

12 https://larslynnehansen.medium.com/accelerating-drug-discovery-research-with-new-ai-models-a-look-at-the-astrazeneca-cerebras-b72664d8783

Ease of Programming

The Cerebras software stack makes achieving higher levels of training performance on large NLP models easy. The effort required is as simple as a few lines of code, shown here.

The CerebrasEstimator is a wrapper developed by our team for TensorFlow.

Users simply import CerebrasEstimator, define their model (e.g. BERT-Large), input function, relevant parameters, and training script using standard TensorFlow semantics. The entire process is captured below:

```
from cerebras.tf.cs_estimator import CerebrasEstimator
from cerebras.tf.run_config import CSRunConfig

    est_config = CSRunConfig(
    cs_ip=params["cs_ip"],
    cs_config=cs_config,
  )
  est = CerebrasEstimator(
    model_fn=model_fn,
    model_dir=`./out`
    config=est_config,
    params=params,
      use_cs=True
  )
  est.train(input_fn, max_steps=100000, use_cs=True)
```

CerebrasEstimator is subclassed from the official TensorFlow Estimator, and using it is easy and familiar. The user simply provides an IP address for their Cerebras system on top of the standard Estimator specification and sets a flag use_cs=True to direct training and inference at the Cerebras device.

With a CS-2, users can quickly experiment with alternative model architectures, hyperparameters, and different batch sizes by changing only a few lines of code. There is no additional work to scale the network across multiple small devices or to deal with communication and synchronization issues.

End-to-end model development tasks such as model setup, hyperparameter optimization, scaling, and performance optimization can be done in days or weeks on a CS-1 system; compared to months on a traditional GPU cluster setup.

Ease of Programming + Faster Training = Faster Time to Solution

Combining hardware performance and software ease-of-use has a multiplicative effect on overall time-to-solution. In a typical GPU cluster setup, days or weeks can be spent choosing, validating, and optimizing hyperparameters to achieve acceptable device utilization, performance, and model convergence to target accuracy.

In partnership with one of our life sciences customers, we recently compared the time-to-solution for a domain-specific BERT NLP model development project from model concept to model in production using a GPU cluster versus our CS-1 (**Figure 4**).

We considered the same model and dataset and included steps for software setup: model definition, functional debugging, performance optimization, and initial model training and training experiments to develop a production-ready implementation.

This work showed that the CS-1 reduced end-to-end time to solution from research concept to production model from 18 weeks on a GPU cluster to 4 weeks on CS-1. Programming and compute time were reduced by more than 3 months, saving our customer engineering costs and allowing them to accelerate new AI innovation.

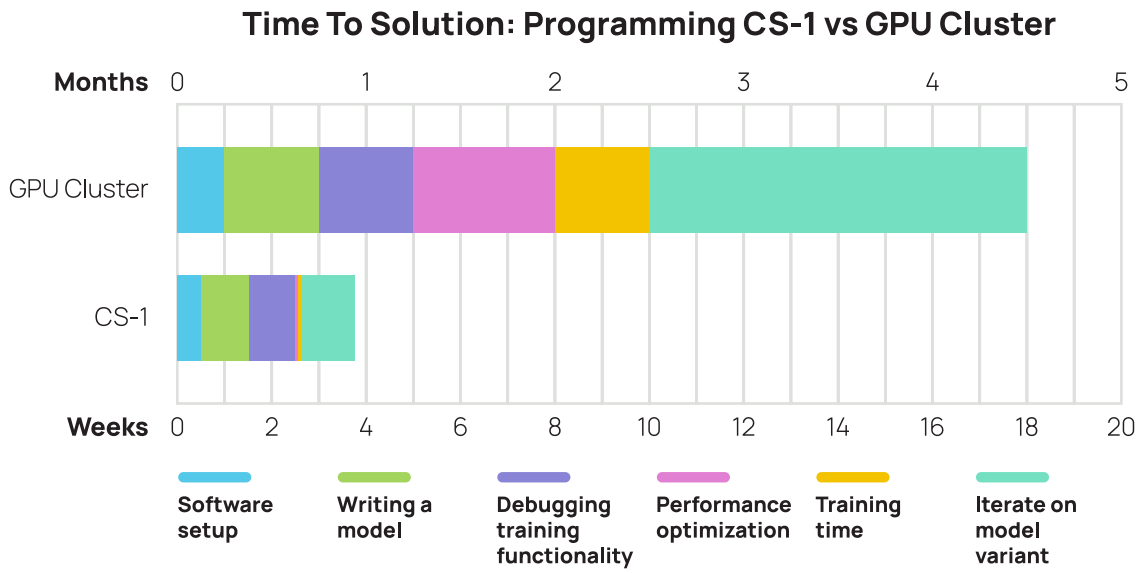
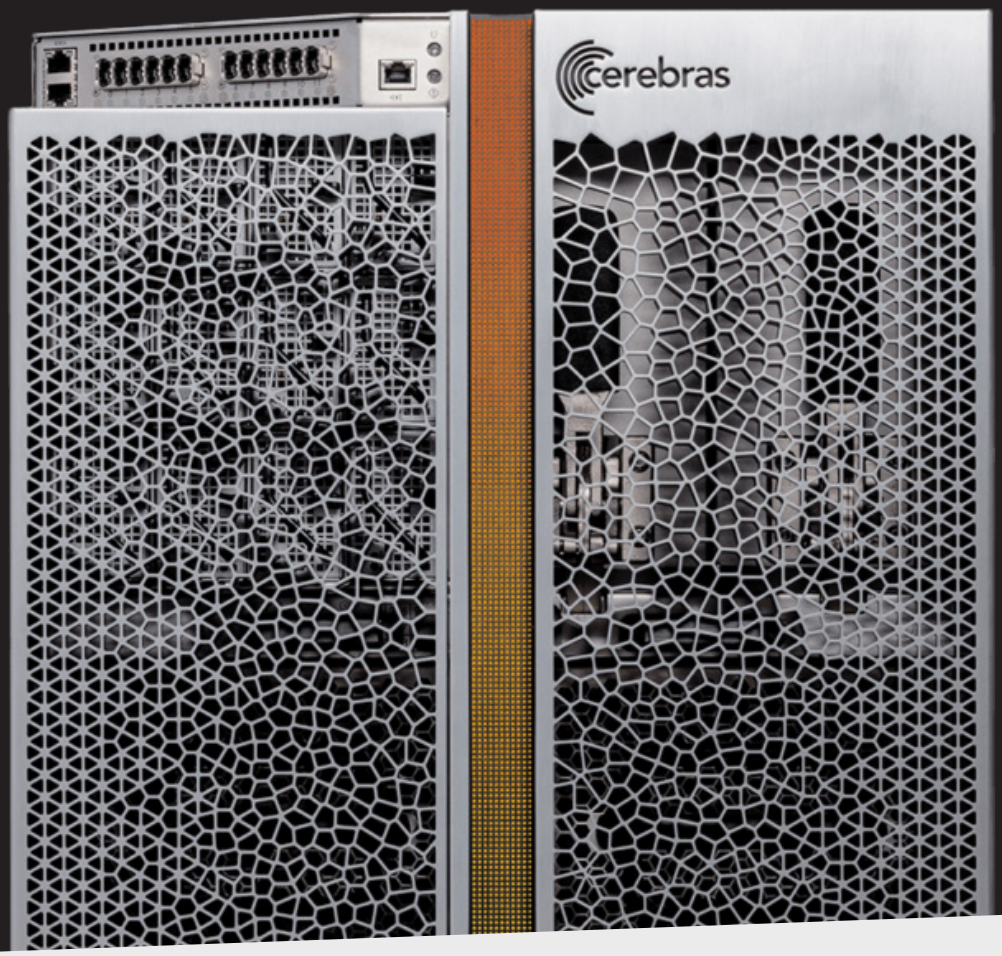


Figure 4. Overall time to solution from research idea to model in production — including programming and computation steps shown in the figure key — Cerebras CS-1 vs customer GPU cluster.

Conclusion

BERT and BERT-like models have a widespread impact in natural language processing and beyond. From natural language queries to protein sequence analysis BERT and BERT-like models are transforming the analysis of text and other sequential data. The Cerebras CS-1 and CS-2 systems bring the power of these networks to a widespread audience. By simplifying deployment, making them easy to use and dramatically reducing training times, Cerebras Systems solutions extend the reach and the impact of BERT and BERT Like models across industry and government customers alike.

To learn more or to see a demo, please contact us at cerebras.net/get-demo.



cerebras

Cerebras Systems is revolutionizing compute for Deep Learning with the CS-2 powered by the Wafer Scale Engine. The Wafer Scale Engine delivers more compute, more memory, and more communication bandwidth for artificial intelligence research at previously-impossible speeds and scale. Pioneering computer architects, computer scientists, and deep learning researchers have come together to build a new class of computer system that accelerates AI by orders of magnitude beyond the current state of the art.