Accelerating AI and HPC for science at wafer-scale with Cerebras Systems

Argonne Training Program on Extreme-Scale Computing (ATPESC)

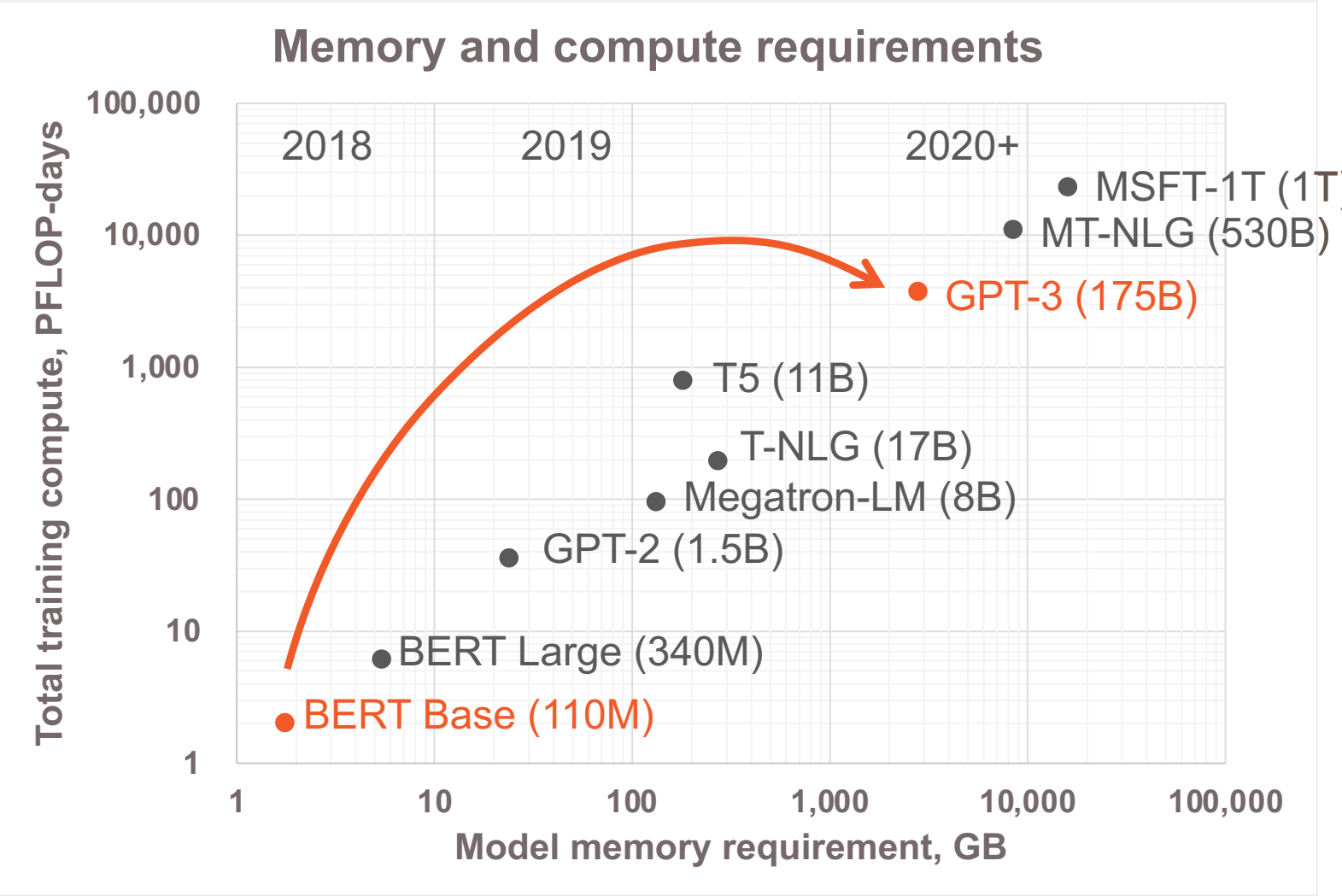
Dr Andy Hock*; Vice President, Product
Cerebras Systems

01 August 2022

* andy@cerebras.net

Introduction

The challenge and opportunity: growth in AI compute

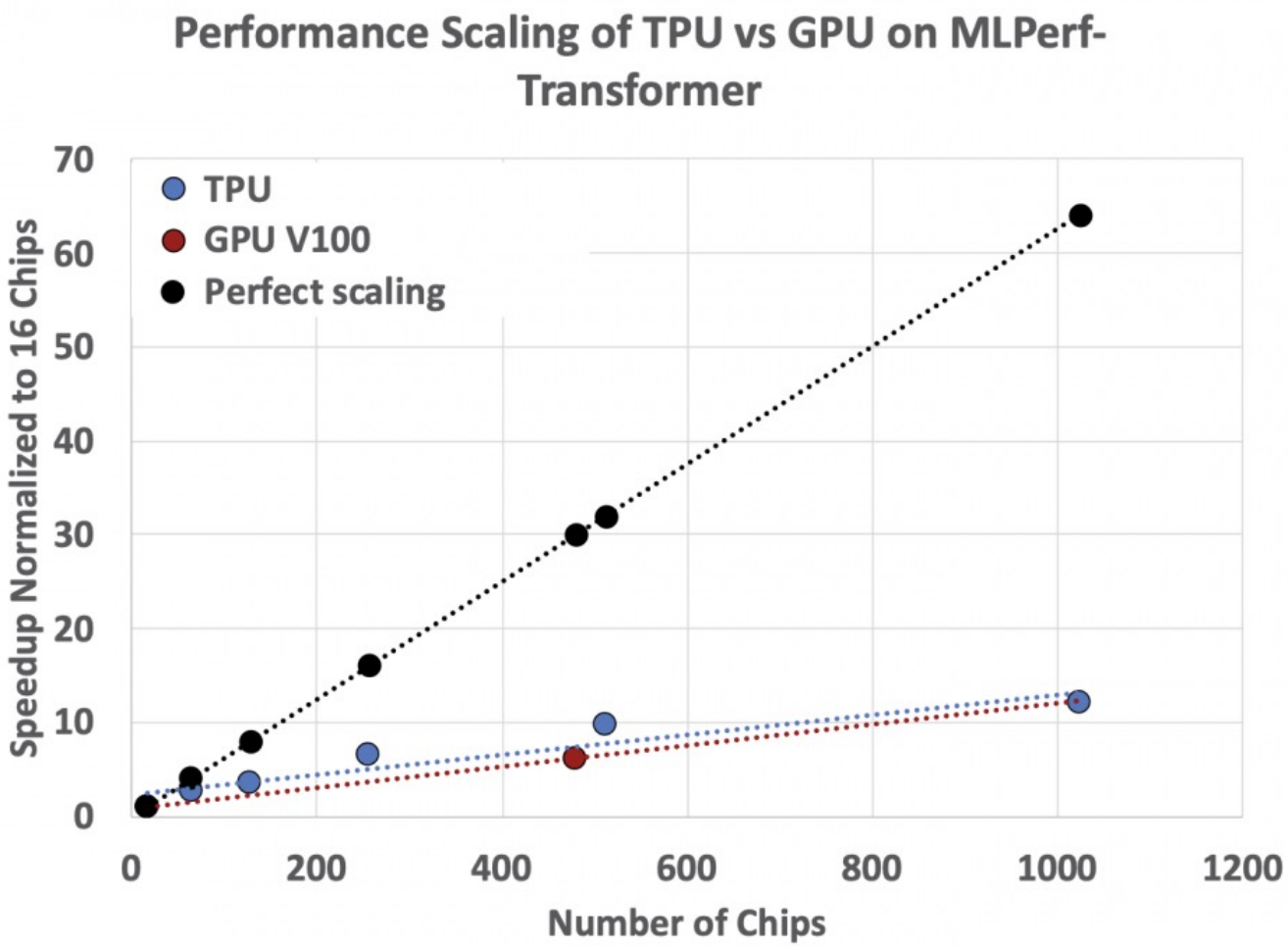


1800x more compute

In just **2 years**

Tomorrow, **multi-trillion** parameter models

Traditional cluster not the optimal path to scale



Time to solution scaling and efficiency falls as cluster size increases:

- Cost of communication grows
- Individual device utilization falls
- Total # epochs to train goes up

Here e.g. we see ~500 chips needed to achieve ~100-160x acceleration.

We **need a new compute solution for extreme-scale deep learning**

Figure. TPU and GPU performance on MLPerf-Transformer 0.6, from Rogers and Khary (2021). An Academic's Attempt to Clear the Fog of the Machine Learning Accelerator War, in *ACM Sigarch Computer Architecture Today.*

Limits of existing scale-out approaches

State-of-the art and emerging workloads need massive **memory**, massive **compute**, and massive **communication**.

On giant clusters of small devices, **all three become intertwined, distributed problems**.

Need to do inefficient, fine-grained partitioning and coordination of memory, compute, and communication across thousands of devices.

Distribution complexity scales dramatically with cluster size

Cerebras Systems

Design, build, and deploy a new class of computer system that delivers orders of magnitude more performance for AI and HPC

Founded in 2016

400+ engineers across HW, SW, ML

Offices
Silicon Valley | San Diego | Toronto | Bangalore

Customers
North America | Asia | Europe

Our solution

Cerebras Wafer-Scale Engine (WSE-2)

The Largest Chip in the World

850,000 cores optimized for sparse linear algebra

46,225 mm² silicon

2.6 trillion transistors

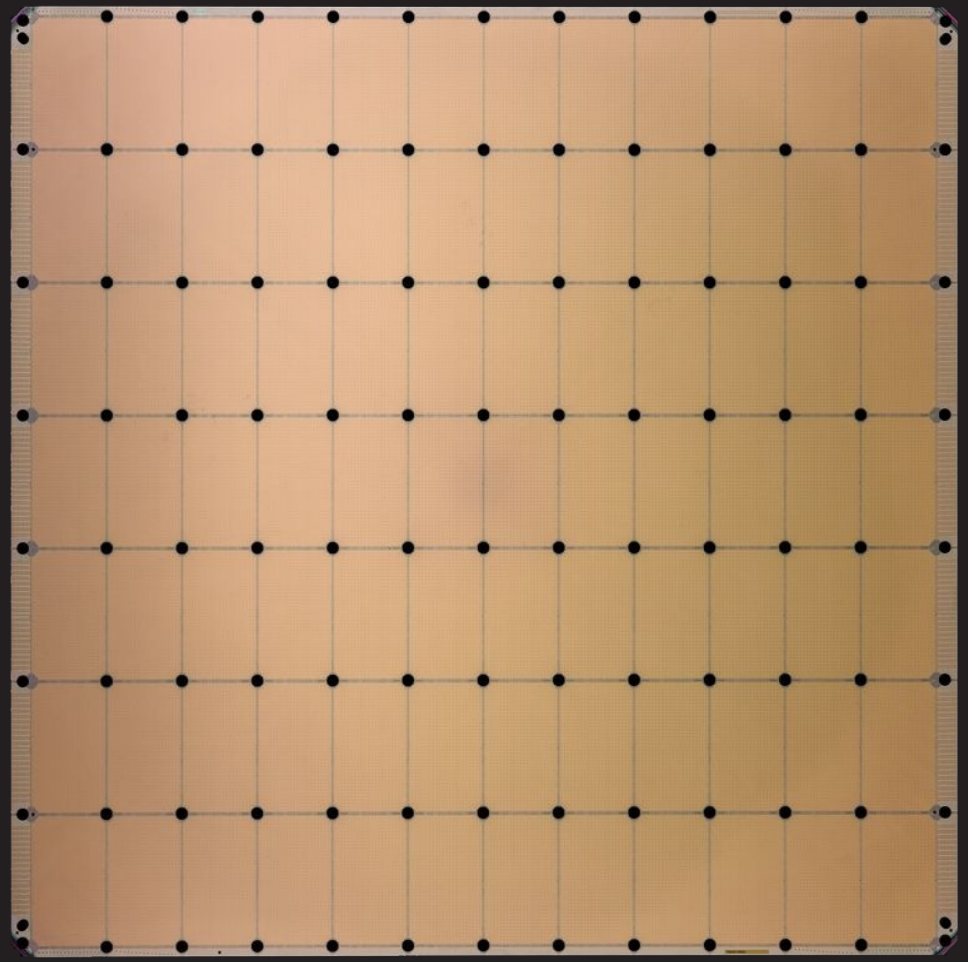
40 Gigabytes of on-chip memory

20 PByte/s memory bandwidth

220 Pbit/s fabric bandwidth

7nm process technology

Cluster-scale acceleration on a single chip



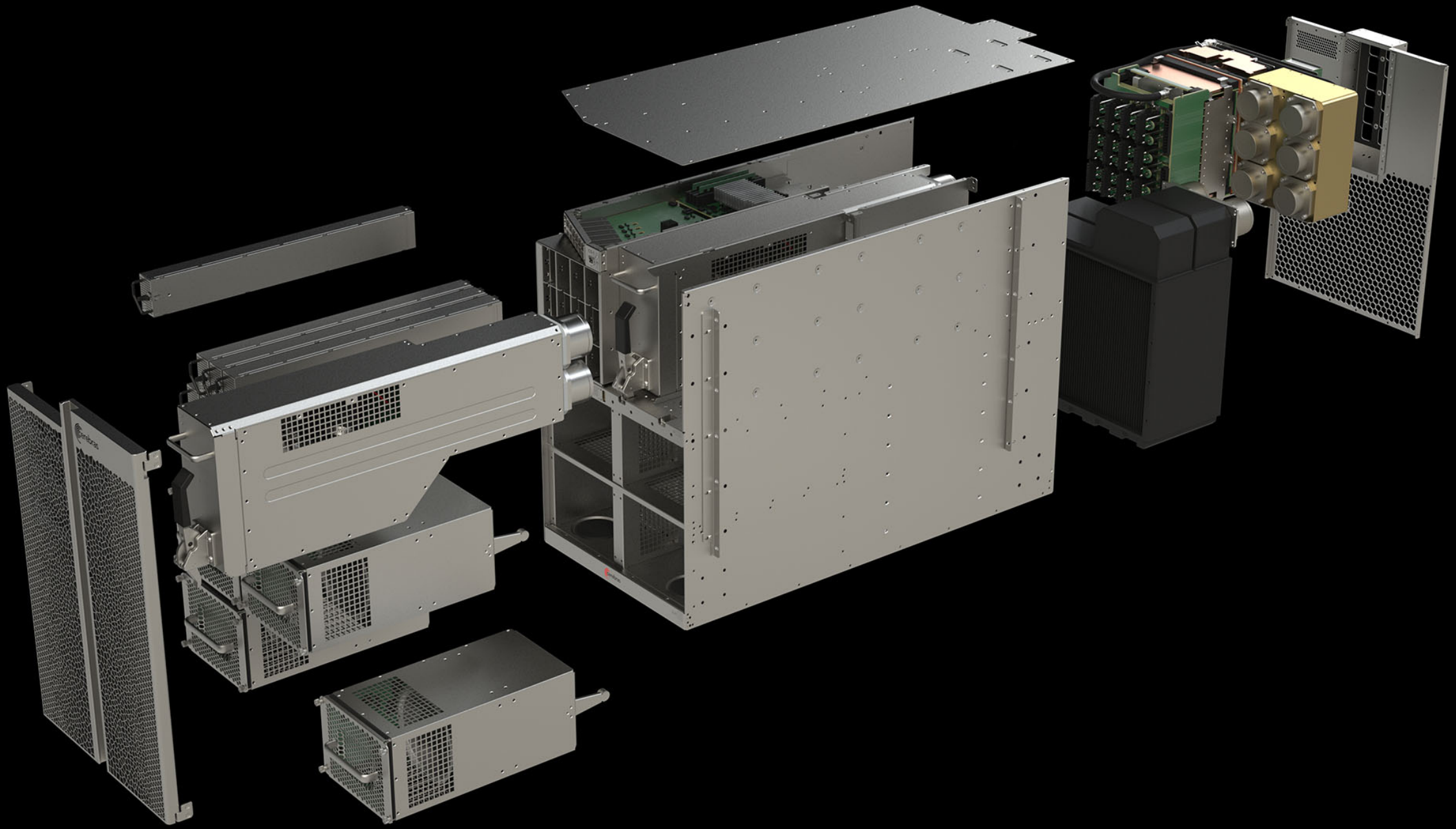
Cerebras WSE-2 7nm
2.6 Trillion Transistors
46,225 mm² Silicon

Largest GPU
54.2 Billion Transistors
826 mm² Silicon

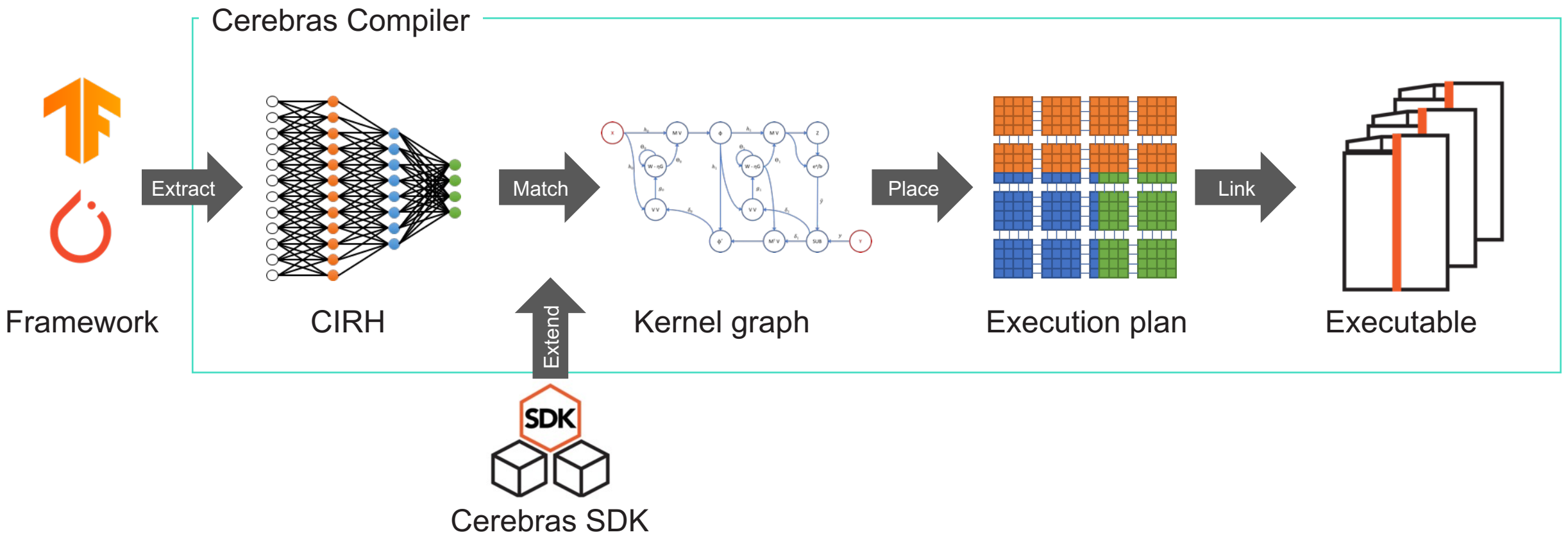
Cerebras CS-2 System

The world's most powerful AI computer

- ✓ Standard rack mount and integration
- ✓ Easy install, setup
- ✓ Available on-prem or remote / cloud



The Cerebras Software Platform



Program a cluster-scale resource with the ease of a single node

Easy to program with TensorFlow and PyTorch (TF example)

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

def model_fn(features, labels, mode, params);
  ...
  return spec

def input_fn(params):
  ...
  return dataset

est = Estimator(
    model_fn,
    config=CSRunConfig(cs_ip, params)
    params=params,
    model_dir='./out',
    )
est.train(input_fn, steps=100000)
```

```
$ cs_run python run.py --mode train --cs_ip $CS_IP
```

Import `CerebrasEstimator`

Import `CSRunConfig`

Define `model_fn` and `input_fn` as usual

Instantiate `Estimator`

Call `estimator.train()` instead of `model.fit()`

Launch run with orchestrator (like Slurm)

Value and use cases

What our customers are saying

"We have a cancer-drug response prediction model that's running many **hundreds of times faster on that chip (Cerebras) than it runs on a conventional GPU**"

Rick Stevens
Associate Director

"Training which historically took over **2 weeks to run on a large cluster of GPUs was accomplished in just over 2 days**"

Nick Brown
Head of AI

"On a Cerebras CS-1 system we **pre-trained our EBERT model…in ~2.5 days…which we estimate would have taken ~24 days of training on a GPU cluster with 16 nodes**."

Kim Branson
Senior VP AI

"We count on the CS-2 system to boost our multi-energy research and **give our research 'athletes' that extra competitive advantage**."

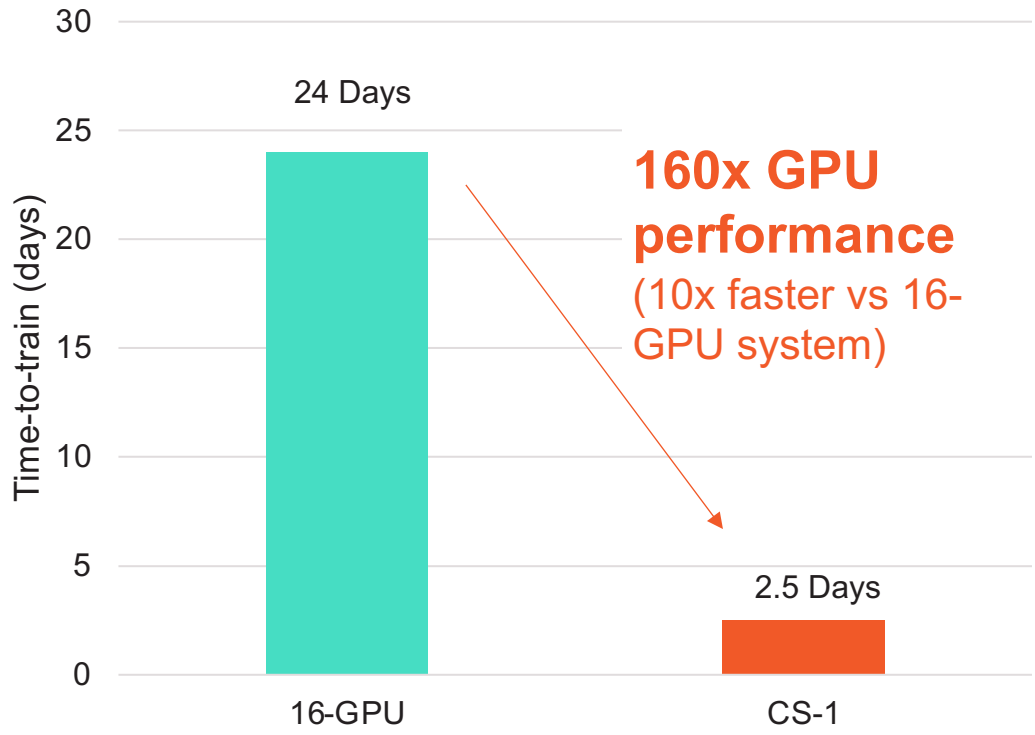
Vincent Saubestre,
CEO and President, TotalEnergies, USA

Large language models for science

Objective: Accelerate genetic validation of drug targets using novel technique that includes epigenomic data in NLP models, rather than genome-only models

Challenge: Training this complex model with massive datasets would take several weeks on a 16-GPU cluster, making rapid experimentation impractical

Outcome: ~10X training speedup over 16 GPUs empowered researchers to experiment with epigenomic data and demonstrate superior results to DNA-only datasets



"The training speedup afforded by the Cerebras system enabled us to explore architecture variations in a way that would have been prohibitively time and resource intensive on a typical GPU cluster"

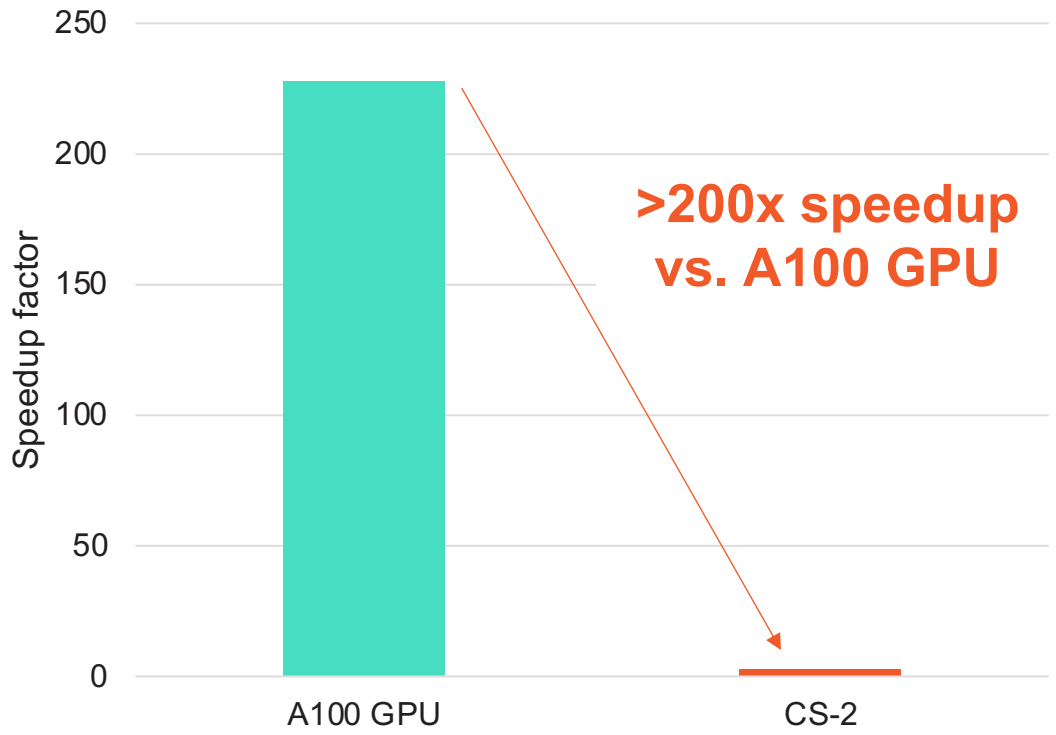
"Epigenomic Language Models Powered by Cerebras", Dec 2021. arxiv.org/abs/2112.07571

Large-scale HPC, AI-powered modeling & simulation

Objective: Enable order-of-magnitude speedups on a wide range of simulations: batteries, biofuels, wind flows, drillings, and CO2 storage

Challenge: Participate in Total study to evaluate hardware architectures, using finite difference seismic modelling code as a benchmark

Outcome: Cerebras CS-2 system outperformed a A100 AI GPU by >200X using code written in the Cerebras Software Language (CSL). System now installed and running at customer facility in Houston, TX



"We count on the CS-2 system to boost our multi-energy research and give our research 'athletes' that extra competitive advantage."

Dr. Vincent Saubestre, CEO and President, TotalEnergies Research & Technology USA

TotalEnergies

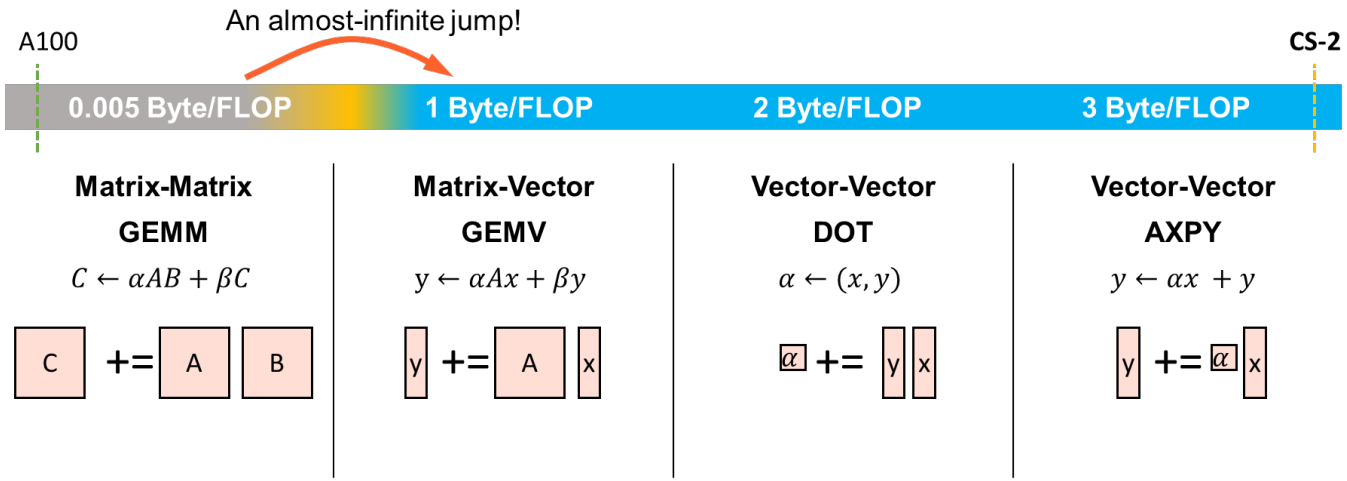
See Jaquelin et al 2022. *Massively scalable stencil algorithm*. https://arxiv.org/pdf/2204.03775.pdf

National Energy Technology Laboratory

Towards Real-Time CFD

Cerebras system solves sparse linear equations 200x faster than Joule 2.0 supercomputer*

Sparse GEMM performance enabled by massive memory bandwidth.

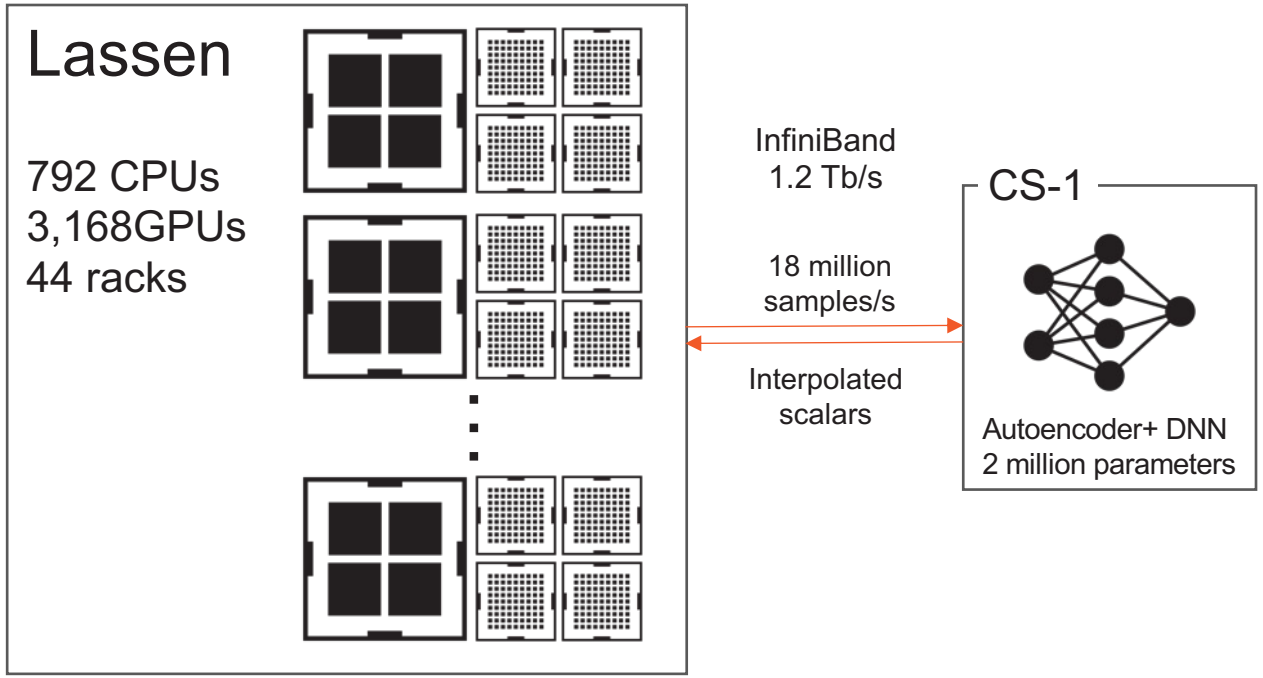


* See Rocki et al., “Fast Stencil-Code Computation on a Wafer-Scale Processor” SC20. arxiv.org/abs/2010.03660

Photo: Christian Kuhna, CC BY 3.0

AI surrogate models accelerating cognitive simulation AI+ HPC for physics at LLNL

Heterogeneous system-level optimization for converged AI + HPC workloads



AI-augmented MD for CoVID-19 research at ANL

Task:
Direct molecular dynamics simulations by learning behavior of previous runs

Challenge:
CVAE is quadratic in time and space complexity and can be prohibitive to train.

Outcome:
Impressive performance out of the box
Throughput comparable with 100 GPUs

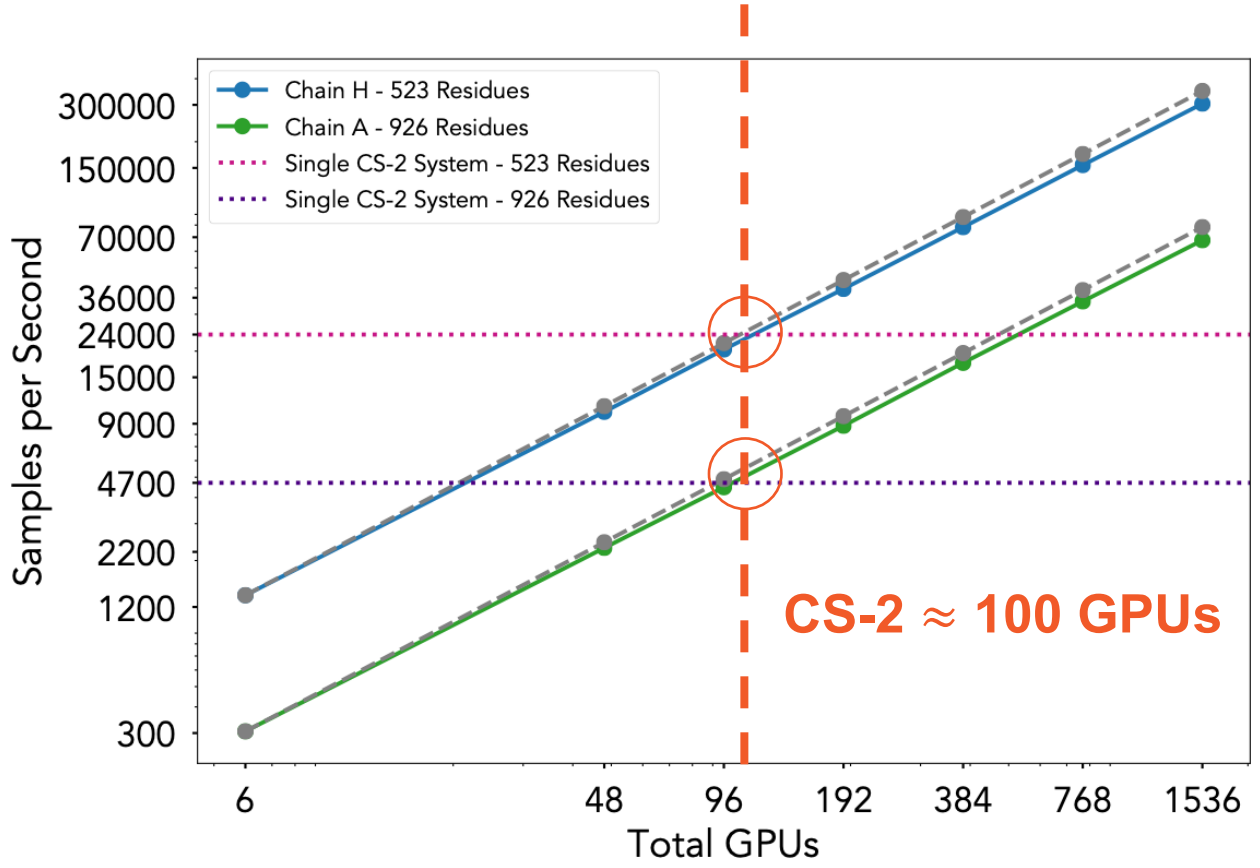


Figure taken from original paper

"Out of the box, we get about 100× improvement on the Wafer-Scale Engine over a single V100 GPU"

Venkatram Vishwanath—data science team lead at Argonne Leadership Computing Facility, ANL

Argonne NATIONAL LABORATORY

Wrapping up

Conclusions / wrapping up

- Cerebras. First **wafer-scale systems for AI + HPC**.
- Orders of magnitude **more performance, simple single-node programming**.
- Recent work **training 1-20B parameter models single systems and clusters**. Going bigger.

Conclusions / wrapping up

- Cerebras. First **wafer-scale systems for AI + HPC**.
- Orders of magnitude **more performance, simple single-node programming**.
- Recent work **training 1-20B parameter models single systems and clusters**. Going bigger.

- Thank you. Extreme-scale computing is **foundational for iterative science at scale**.
- Happy to be here with you as part of the **ATPESC community** – exciting program ahead!
- We encourage you to think big. **Come find us, work with us to go even bigger** 🤓

Conclusions / wrapping up

- Cerebras. First **wafer-scale systems for AI + HPC**.
- Orders of magnitude **more performance, simple single-node programming**.
- Recent work **training 1-20B parameter models single systems and clusters**. Going bigger.

- Thank you. Extreme-scale computing is **foundational for iterative science at scale**.
- Happy to be here with you as part of the **ATPESC community** – exciting program ahead!
- We encourage you to think big. **Come find us, work with us to go even bigger** 🤓

- Curious to **learn more**? www.cerebras.net for specs, docs, code examples.
- Want to **get access**? Reach out to us or our cloud partner Cirrascale.
- Systems for **scientific research** ANL, PSC, NCSA, EPCC, LRZ, more.