Wafer-Scale Innovation for Next-Gen AI Acceleration

The Cerebras Revolution

May 22, 2025

Alexander.Mikoyan@cerebras.net

VP Europe

Innovation from the heart of the Silicon Valley

Established in 2015 to build a new class of system for the future of AI & HPC

A full acceleration solution: chip, system, software, compiler, ML, SDK, services

Offices

Silicon Valley | Toronto | Bangalore

Customers

North America | Asia | Europe

TIME 100 MOST INFLUENTIAL COMPANIES 2024

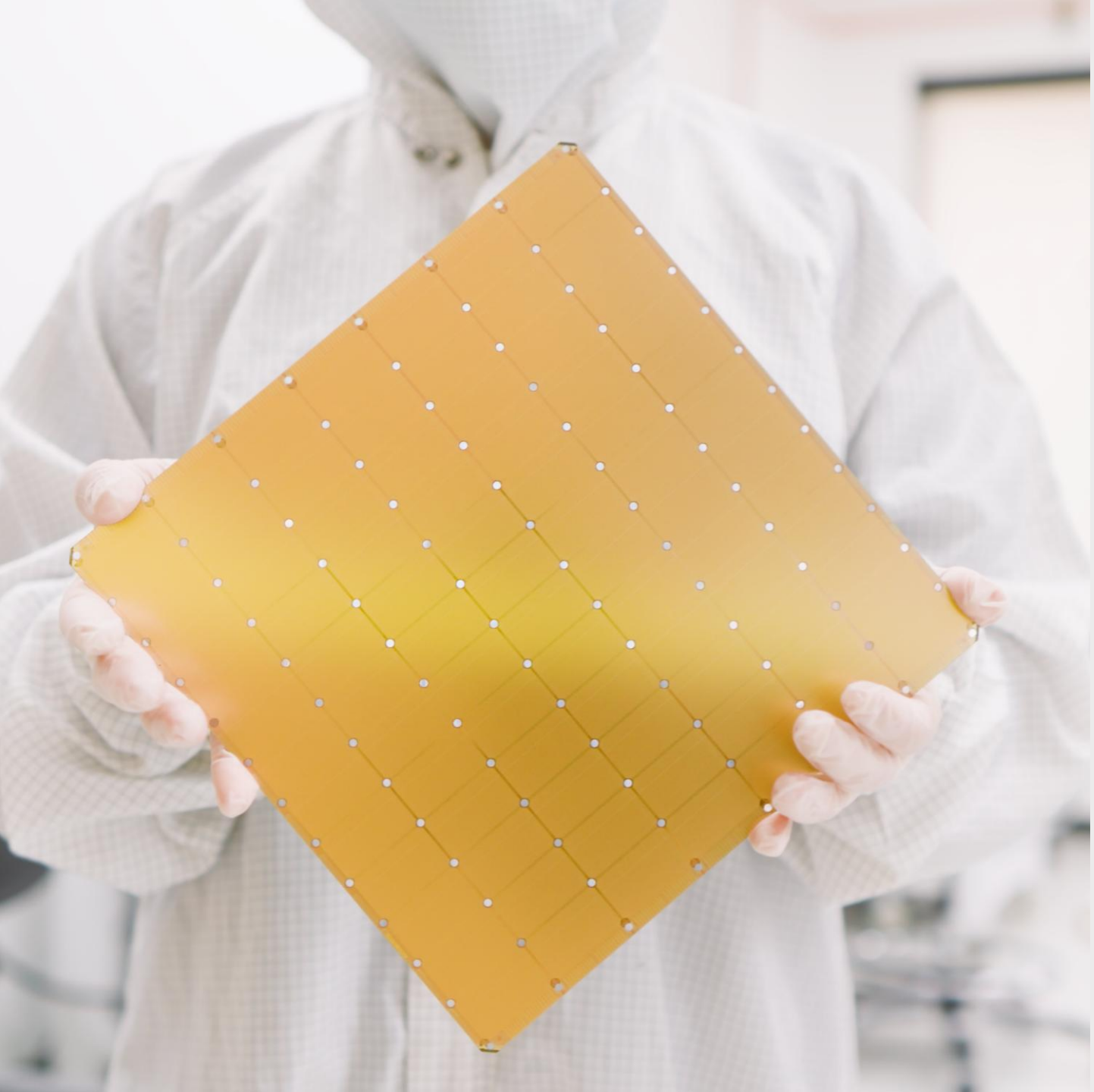
Time Magazine

Forbes AI 50

FORTUNE 50 AI INNOVATORS

Fortune 50 AI

cerebras



Wafer-Scale Engine

The foundation of our technology

4 trillion transistors

46,225 mm^2 silicon

900,000 cores optimized for sparse linear algebra

125 Petaflops of AI compute

44 Gigabytes of on-chip memory

25 PByte/s memory bandwidth

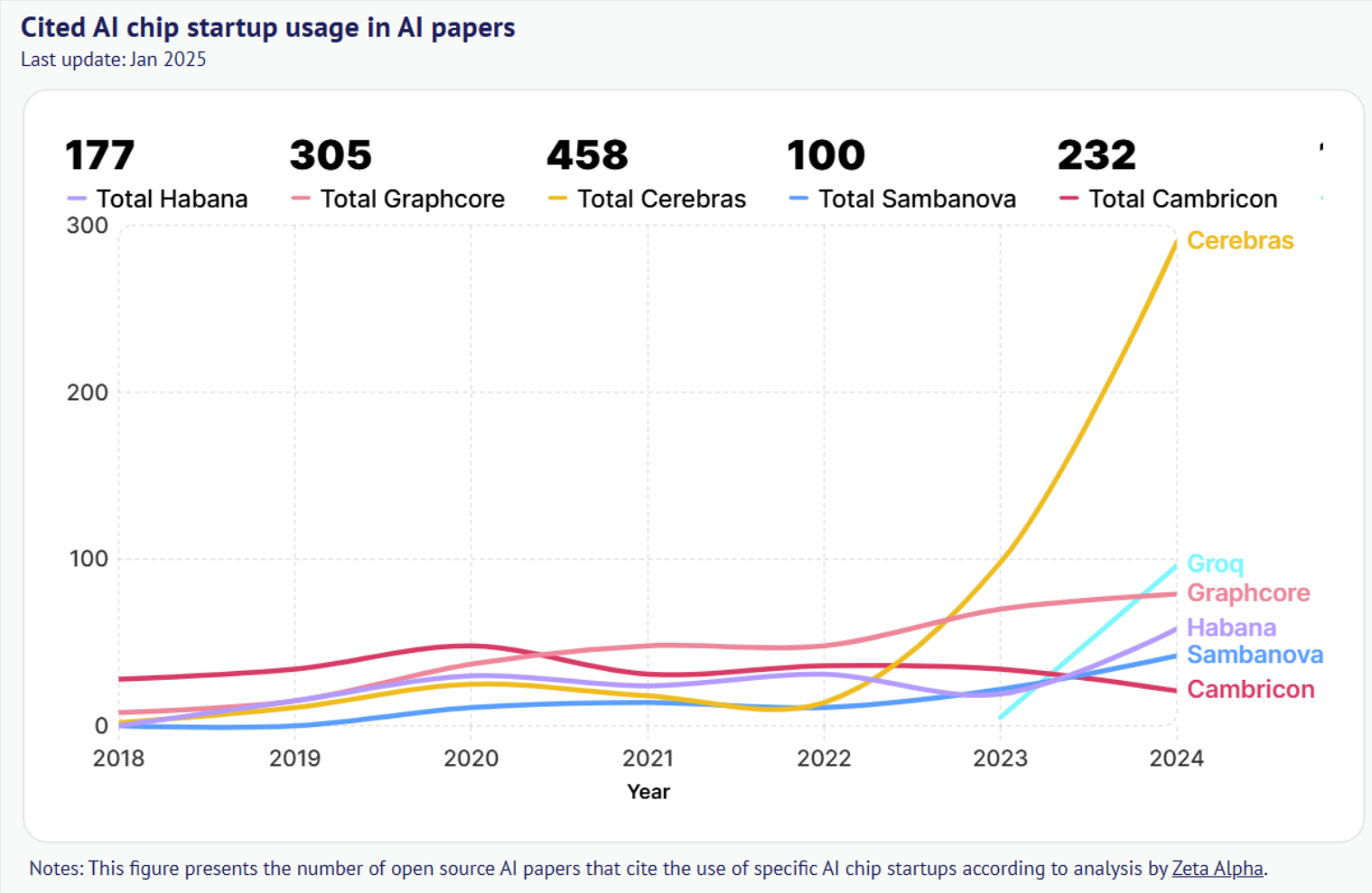
30 Pbyte/s fabric bandwidth

5nm TSMC process

3rd generation in production – WSE-3

WSE is the most cited AI processor outside GPU/CPU circle

Clear leadership across both training & inference

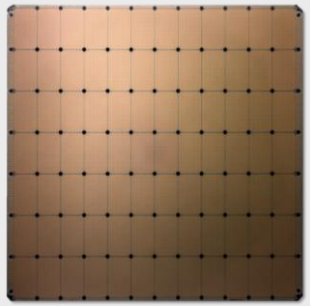


See State of AI Report Compute Index

WSE powers mature system architecture...

Simple clustering using lower-cost Ethernet. Comes fully integrated. Liquid cooled.

Wafer Scale Engine (WSE)



- ✓ 52x more cores
- ✓ 880x more fast memory on processor
- ✓ 8,000x more memory bandwidth
- ✓ 3,715x more communication bandwidth

Comparisons with the Nvidia H100

CS-3 Server

- ✓ Houses the WSE-3 in 16 RUs of a standard rack
- ✓ 125 PetaFLOPS AI compute
- ✓ Standard Ethernet 12x100Gbps I/O

Cerebras AI Supercomputer

- ✓ Up to 2,048 CS-3 servers (up to 256 ExaFLOPS AI compute)
- ✓ Training models with up to 24T parameters as one logical device or hosting model replicas for inference
- ✓ 97% less code complexity by using unified memory and data parallel only training

cerebras

… that delivers great value across AI & HPC domains

Faster, easier, more efficient

AI Training

Efficiently train & fine-tune the highest quality GenAI models for their domains and use cases

Key Propositions

1. **1/10th of efforts from idea-to-value**
   Faster iterations, higher precision and higher quality custom models
2. **Easy to program**
   Just use PyTorch. WSE auto-scales with no complexity
3. **One-third the power**
   Of leading commercially available GPU systems

AI Inference

Serve the most modern GenAI models at the fastest inference speeds in the market

Key Propositions

1. **20x-70x more rapid responses**
   Unlocking real-time interactive-ty, more thinking & new agentic use cases
2. **Easy to adopt**
   Switch from OpenAI API compatible LLMs in 30 seconds
3. **Unmatched economics**
   Best price/performance & energy/performance for low-latency use cases

High Performance Computing

Run some advanced simulations (physics/math) impossible or impractical on traditional gear

Key Propositions

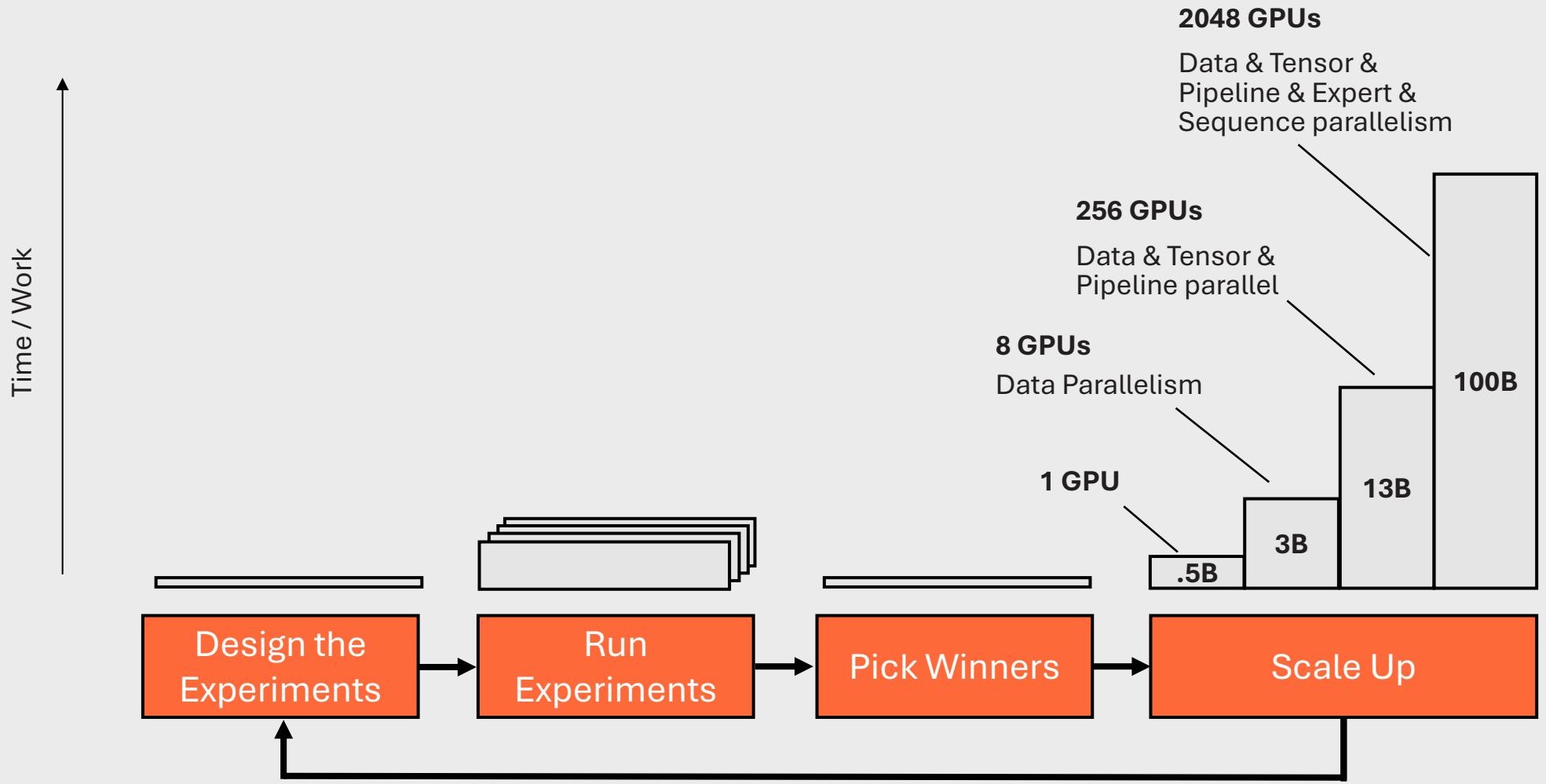
1. **>100x faster result**
   Deliver transformational research results in days rather than years
2. **Easy to use tools**
   SDK, Simulator, Compiler, code examples are available
3. **Much more efficient**
   Even a one system could overperform largest computers for some simulations

cerebras

Training

Getting good model quality at scale on GPUs is not easy…

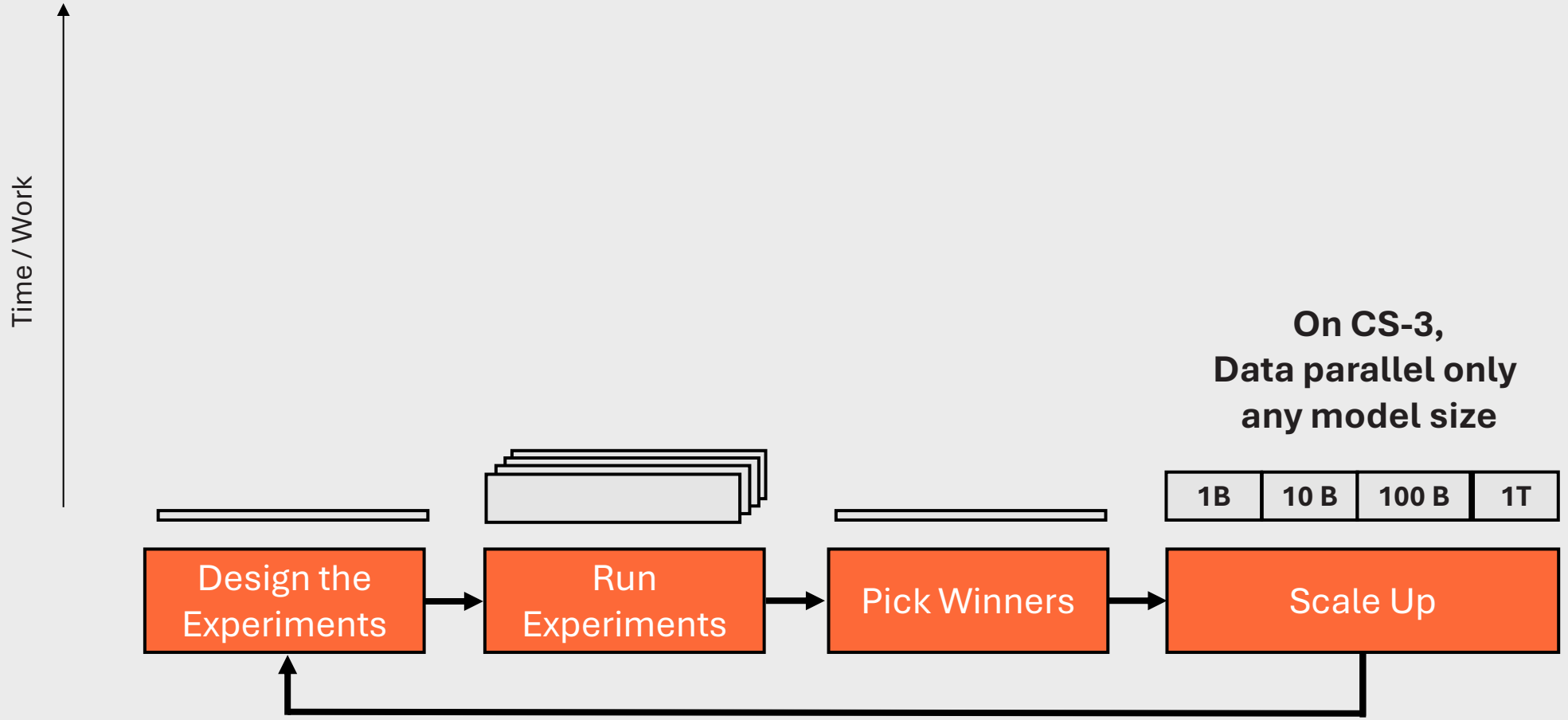
Engineering for multiple parallelisms is expensive – time and money



cerebras

… but scaling does not add any complexity on Cerebras

Cerebras gets practitioners to high-quality large models faster & more efficiently

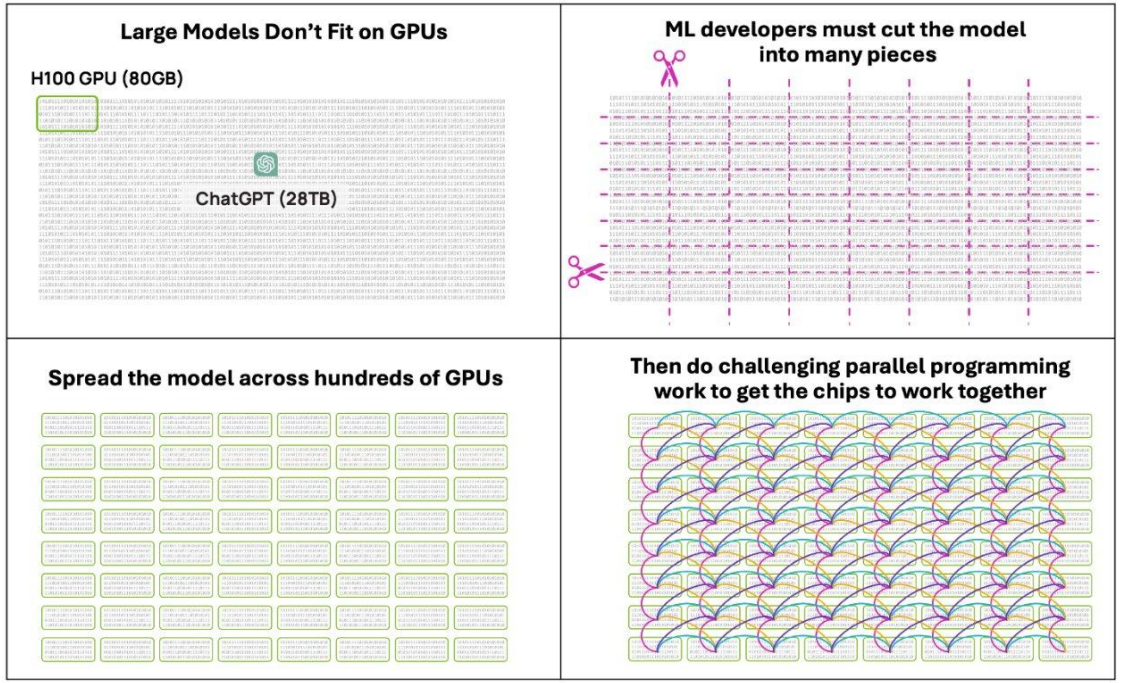


cerebras

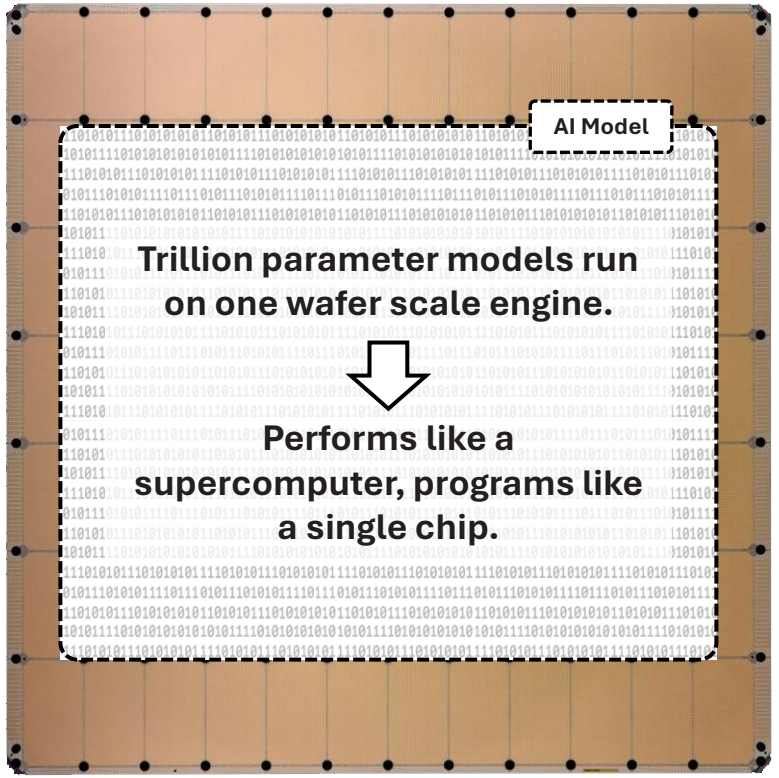
Train on CS-3 clusters with single chip simplicity

Handle largest models on Cerebras – no parallel programming, no CUDA!

GPUs



Cerebras WSE



Training large models on GPUs can be slow and resource-intensive. Each Cerebras WSE can train even the largest modern models, and is designed to allow simple independent scaling across thousands of CS-3s, with no complex distribution.

cerebras

CS-3 cluster: optimal architecture for data parallel training

Scaling cluster compute while operating like a single device

Weights are stored on MemoryX server

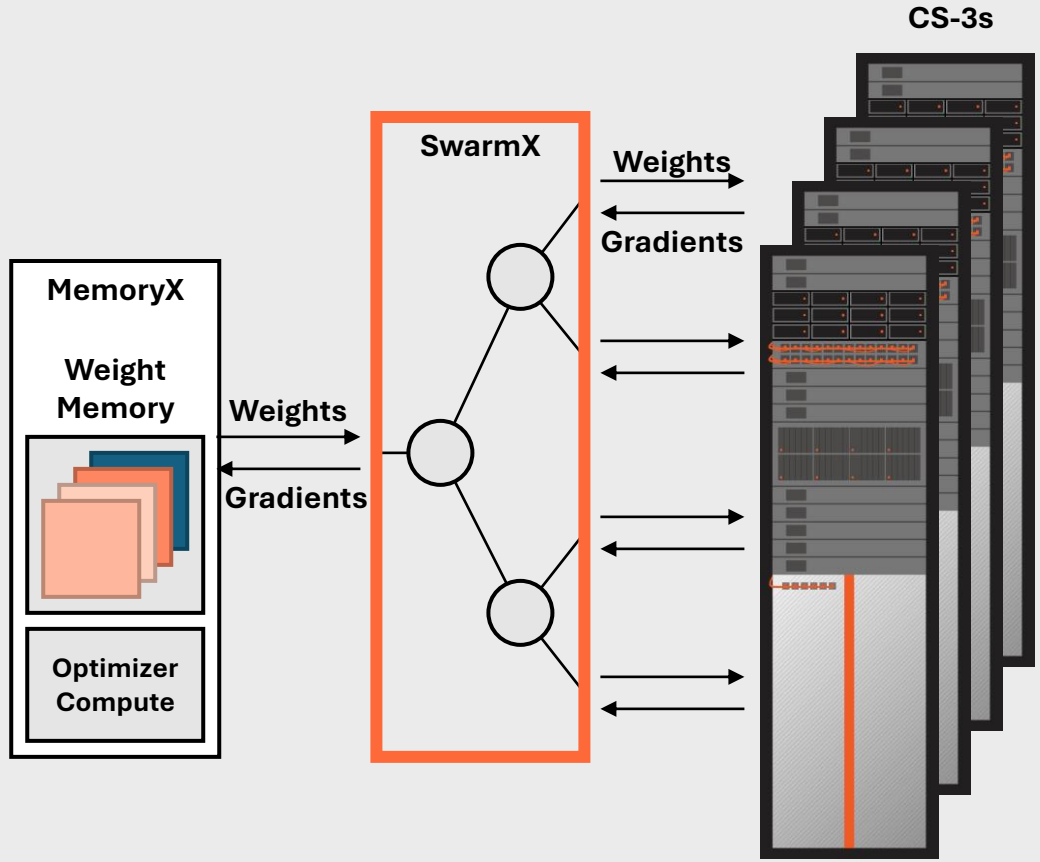
- Standard x86 server with DRAM housing up to trillions of parameters
- Executes weight updates
- Increase model memory independently of compute (CS-3s) for larger models

Data-parallel only training across CS-3 cluster

- Weights are broadcast via SwarmX (x86) to all CS-3s
- Gradients are reduced on the way back

Same execution model: single or multi-node cluster

- Same system architecture
- Same network execution flow
- Same software user interface



See documentation for more information: https://training-docs.cerebras.ai/rel-2.5.0/concepts/cerebras-wafer-scale-cluster

cerebras

Scaling from 1 CS-3 to 2048 CS-3s – just change one line

Impossible with GPUs

```
python run.py
--params params.yaml
--num_csx = 1
--model_dir = model_dir
--num_steps = 1000
--mode=train
```

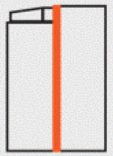
Where's your dataset?

How many nodes?

Where to store weights?

How many training steps?

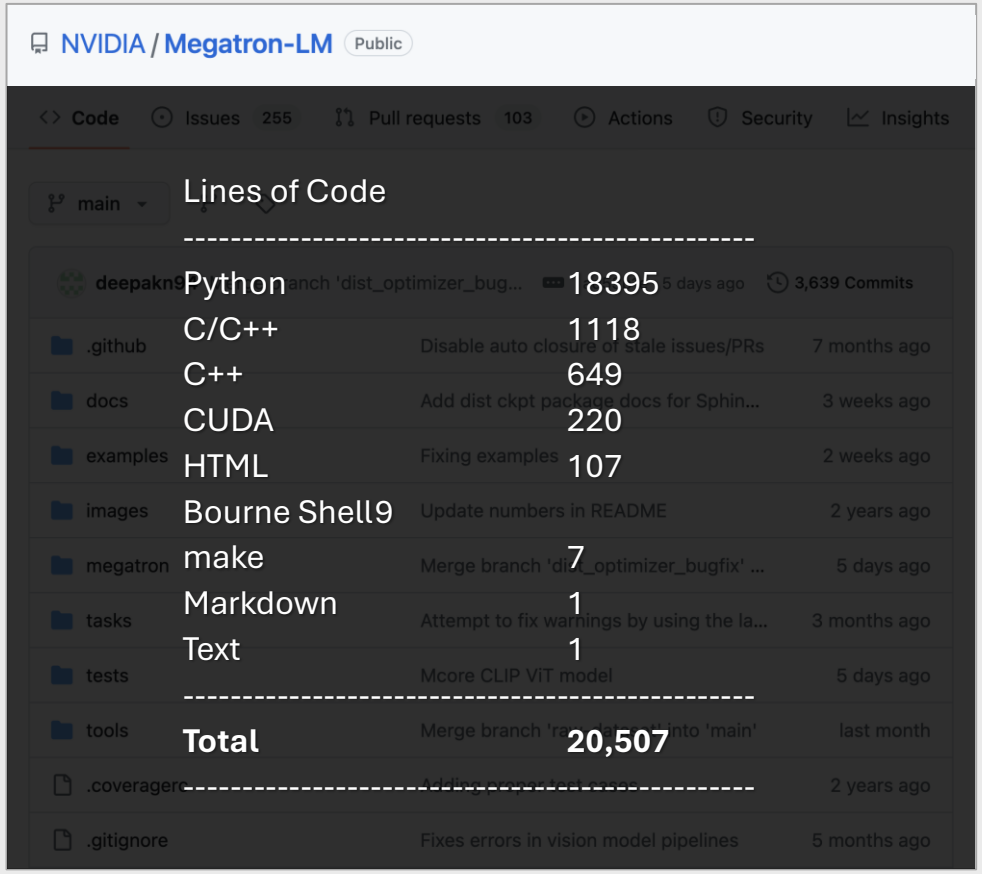
Train, evaluate or infer?



cerebras

Simpler building & training, reducing time/cost to solution

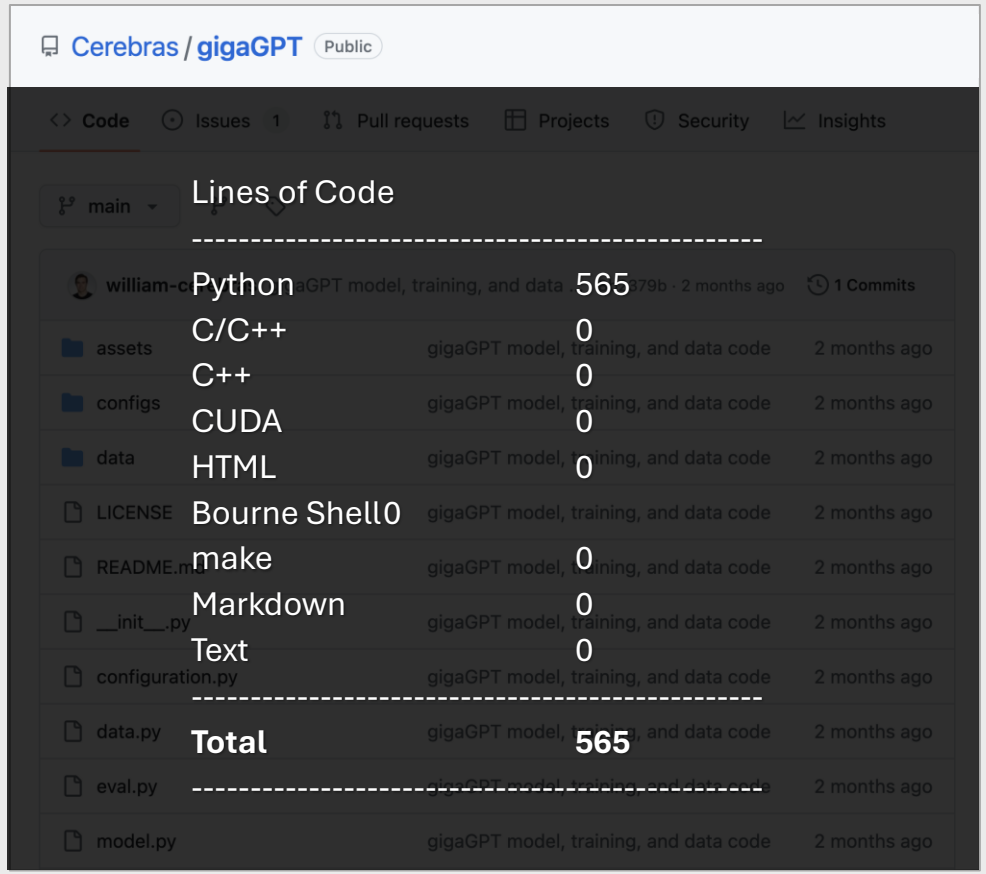
We meet ML practitioners at PyTorch level



Nvidia's GPT-175B Model

20,000 lines of code, weeks to implement

Teams of systems engineers



Cerebras' GPT-175B Model

565 lines of code, 1 Day to implement

1 ML Practitioner

See: Introducing gigaGPT: GPT-3 sized models in 565 lines of code - Cerebras

cerebras

Example: Mayo Clinic's Genomic Model trained on Cerebras

Rapid development accelerated by Cerebras AI platform. Weeks, not years.

1B

parameters

10x larger
than AlphaFold

1T

tokens

Mayo's in-house
patient data

Highest Accuracy

For clinical trials

Outperforms today's
best models

Unprecedented accuracy in disease prediction

87% for Rheumatoid Arthritis

96% for cancer predisposition

83% for cardiovascular conditions

"Our clinicians will be able to make more informed decisions based on genomic data, ***significantly reducing the time it takes to find the right treatment*** *and – more importantly – reducing the physical toll on patients."*

Matthew Callstrom. M.D., Ph.D.
Medical Director for Strategy, Chair – Dept. of Radiology

- Private US academic medical centre focused on integrated healthcare, education, and research.
- 7,300 physicians & scientists
- $660M/year on research, more than 3,000 full-time research staff.

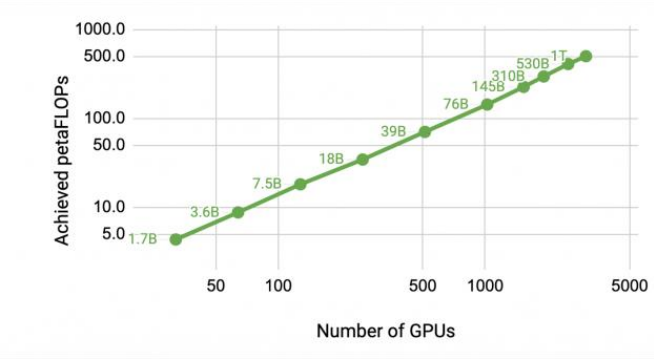
MAYO CLINIC

cerebras

Example: 1T LLM on one CS-3, scaling training linearly

1% power, 3% code complexity, linear scaling. Cerebras is the **only** AI hardware in the world that can do this

1 Per Nvidia, a trillion-parameter model requires over 3,000 GPUs to train or fine-tune*.



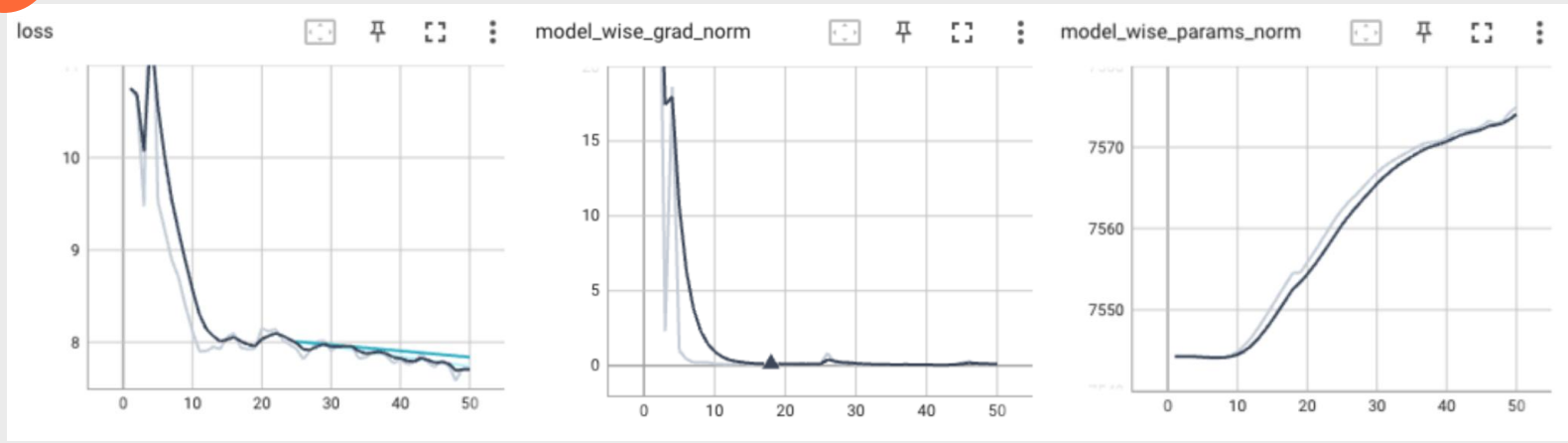
2 One CS-3 system with MemoryX device (pre-configured x86 server)

MemoryX holds the weights – 55 TB, equivalent to 287 Nvidia B200 GPUs in memory capacity.

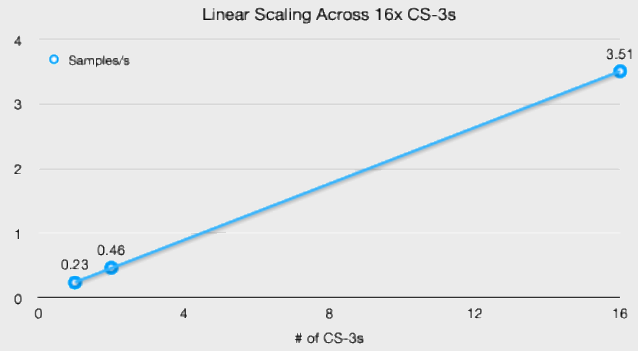
Only two racks – 1% the power & space of GPU infra.

* https://developer.nvidia.com/blog/scaling-language-model-training-to-a-trillion-parameters-using-megatron/

3 In this PoC 50 training steps performed, verifying loss & stable training dynamics



4 Scaled up to 16 systems with near linear performance scaling (15.3x)



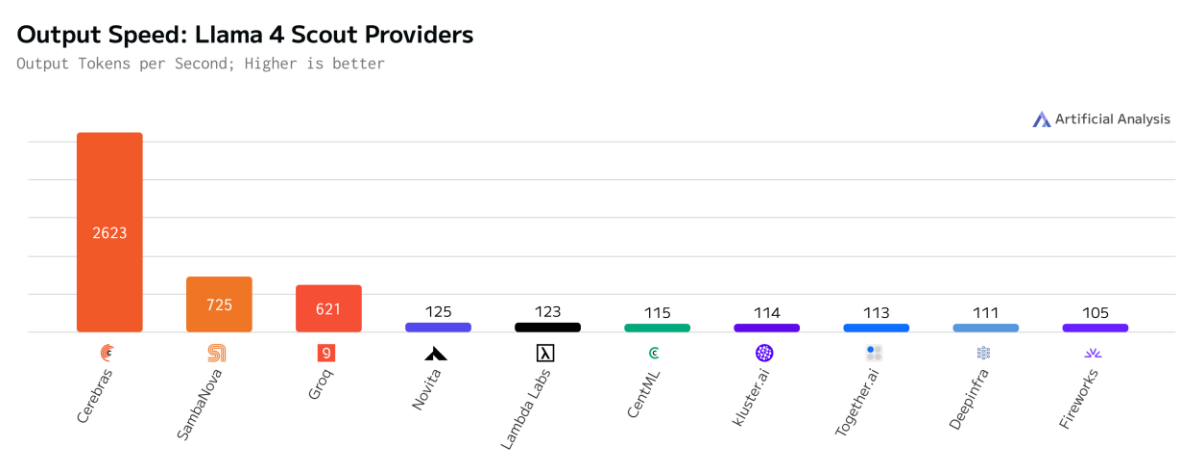
See Cerebras Demonstrates Trillion Parameter Model Training on a Single CS-3 System - Cerebras

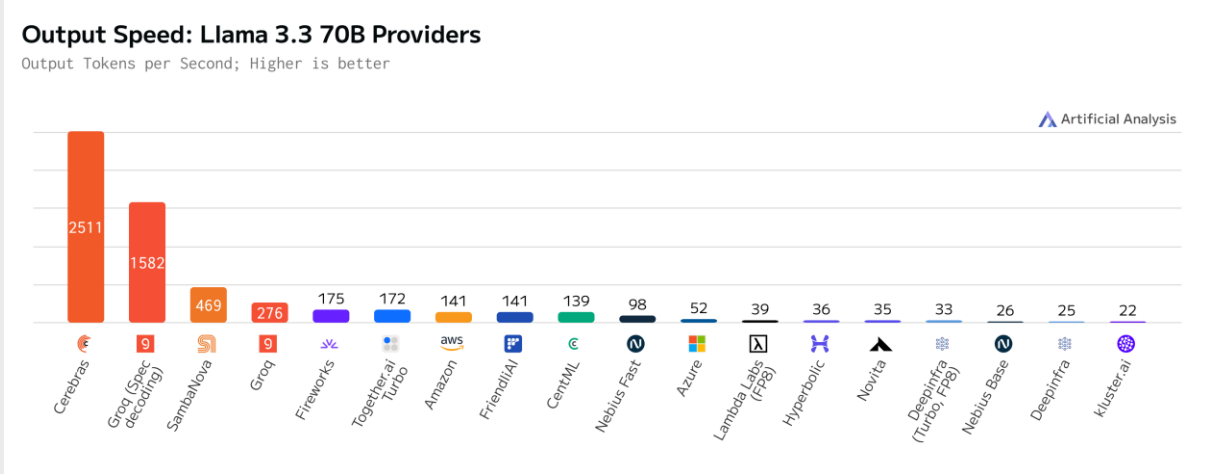
cerebras

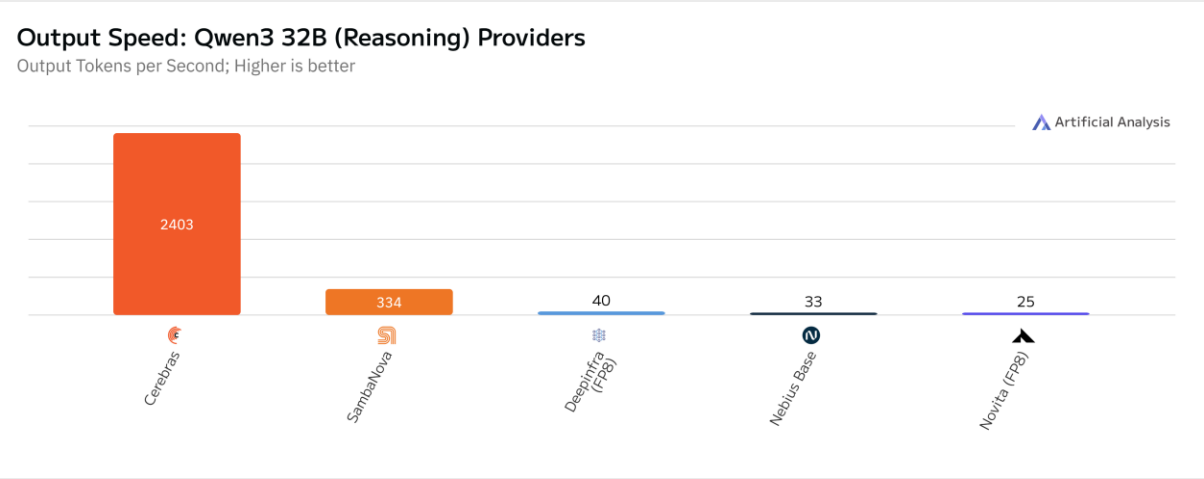
Inference

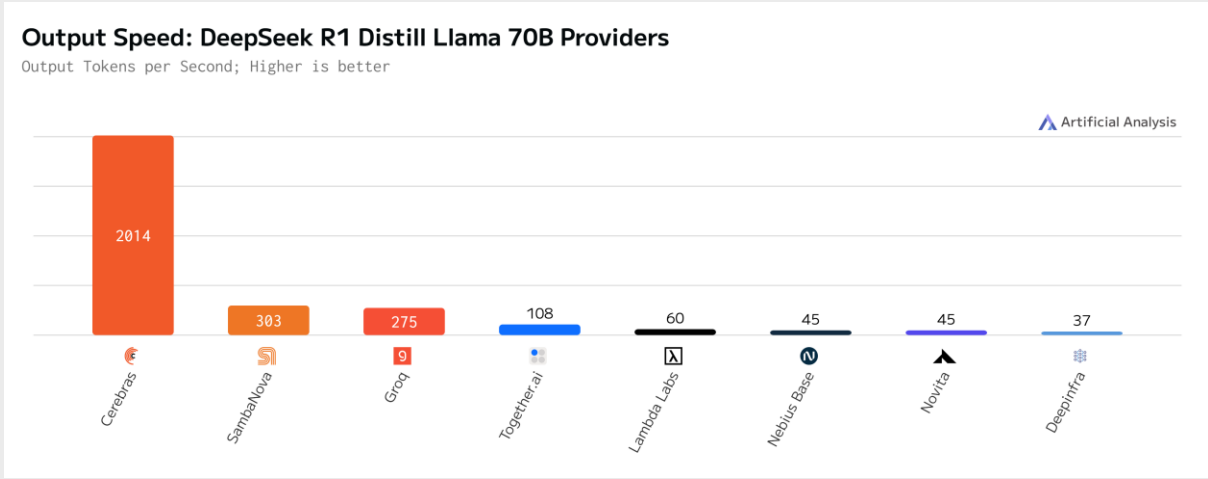
The fastest LLM inference is always on Cerebras

Whatever the model, Cerebras leads thanks to larger memory bandwidth; GPU infra is behind memory wall.





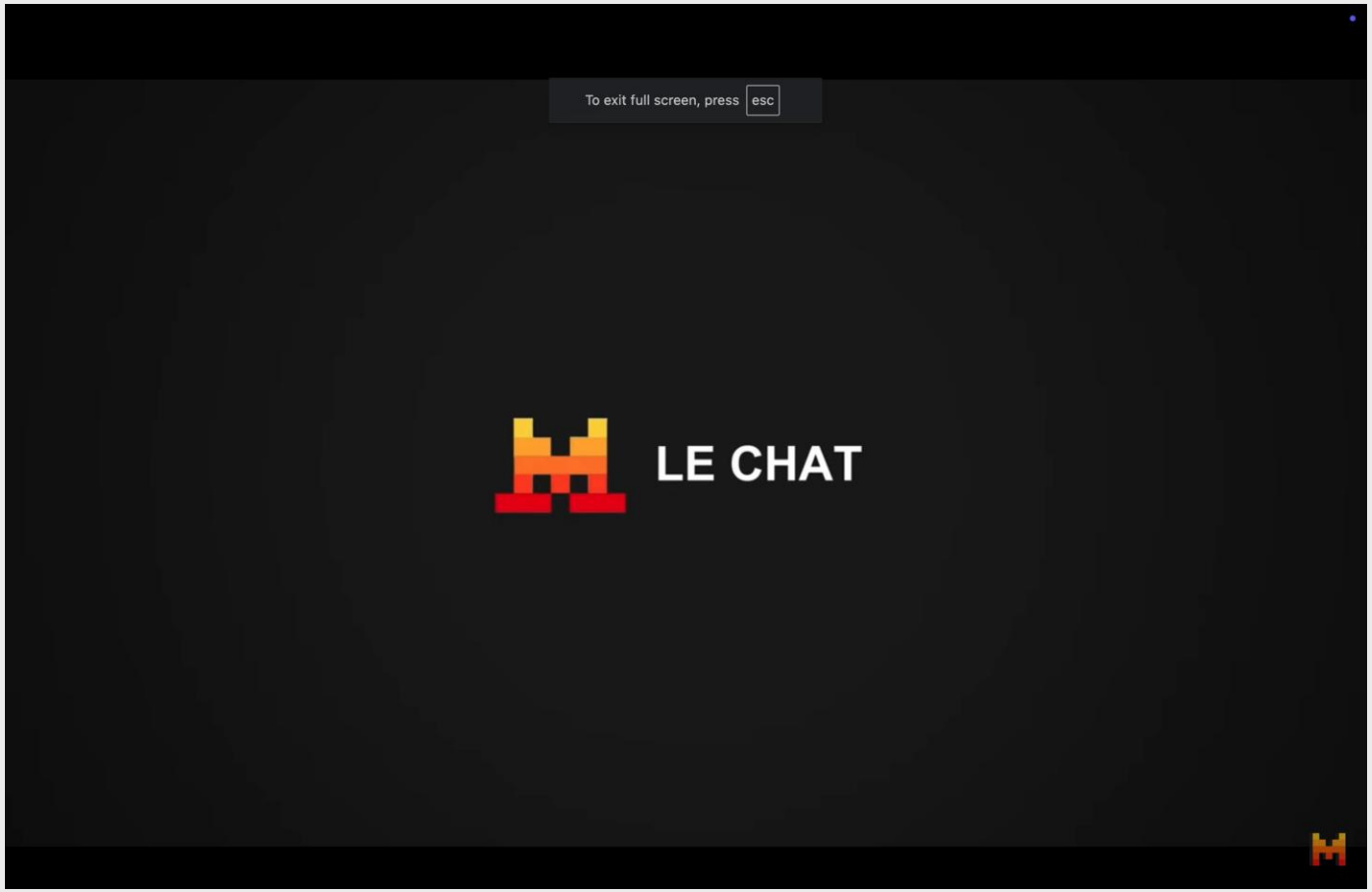




cerebras

Cerebras powers Le Chat by Mistral & others. Speed really matters.

Larger art-of-the-possible: better UX, reasoning, agents, real-time, match HPC



MISTRAL AI_

World's fastest **AI assistant**

deepseek Qwen

World's fastest **reasoning**

Meta AI

World's fastest **open models Llama 3.1, 3.3, 4, Llama API**

perplexity

World's fastest **search**

See The all new le Chat: Your AI assistant for life and work | Mistral AI

cerebras

Fast inference unlocks completely new applications

100 steps of reasoning at real time speed

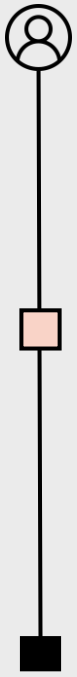
Inference Step

Emerging models reason for longer via additional steps

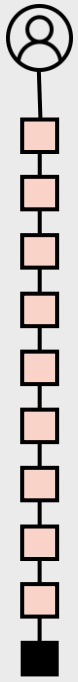
They plan, strategize, self-reflect, and refine their outputs

Each step adds compute & waiting time

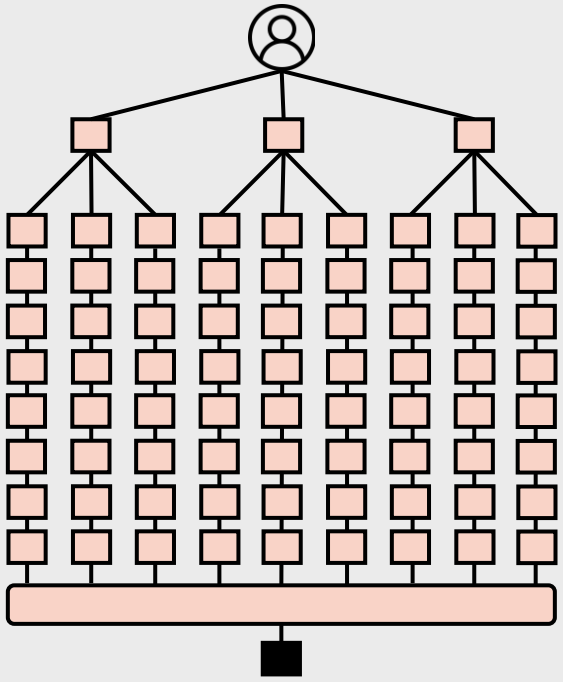
Simple Inference
Llama 3.3 70B



Reasoning
Qwen3-32B



Agentic + Reasoning



Steps per inference query	**1**	**~10**	**~100**
Compute required for inference	**1x**	**~10x**	**~100x**

cerebras

WSE is also a new scientific instrument

Delivering incredible performance for some modelling

Performance gains that make algorithm rewriting absolutely worth it

Benefit from strong scaling…

- Use WSE fabric that is **high bandwidth** and **low-latency**
  - Excellent parallel efficiency for non-linear and highly communicative codes
- Use **900k cores** and fit problems on an individual chip that traditionally take 10s to 100s of small compute nodes
  - Each core is individually programmable

Molecular dynamics, particle simulation, non-linear problems with iterative solvers

… or overcome data access constraints

- Utilize uniformly distributed **on-compute 44 GB of SRAM** that is 1 cycle away from a processing element
  - Speeds up memory access by orders of magnitude
- The CS-3 system is today capable of **1.2 Tb/s bandwidth onto the chip**
  - Stream data onto the chip as required

Stencil-based PDE solvers, signal processing, sparse tensor math

cerebras

All key components brought together on a single tile

Each core is individually programmable

Fabric

Vector streaming engine
Activation selection filter
Sparse stream support

Control

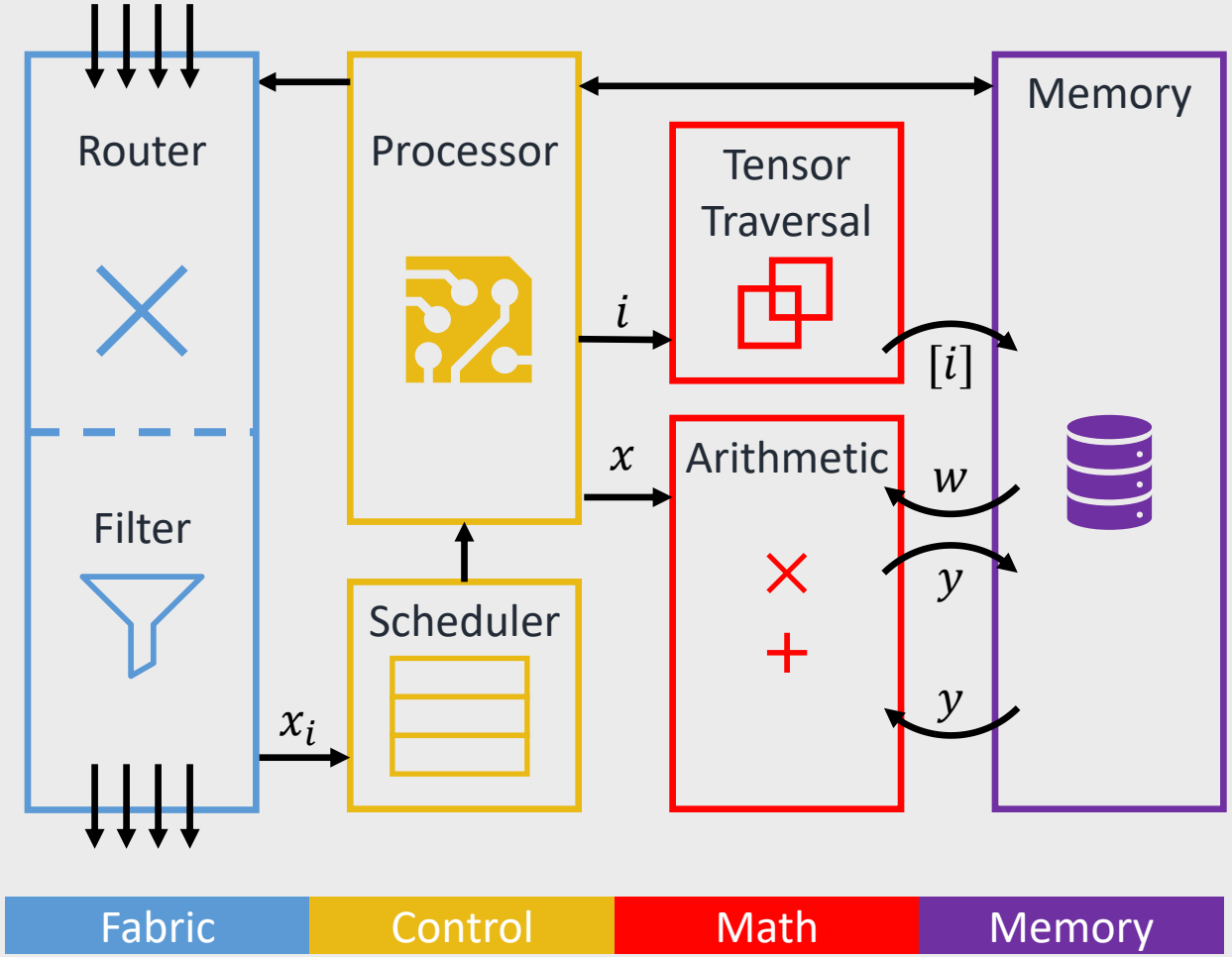
Dataflow scheduling
13 execution threads

Math

Floating point and integer
Tensor access patterns

Memory

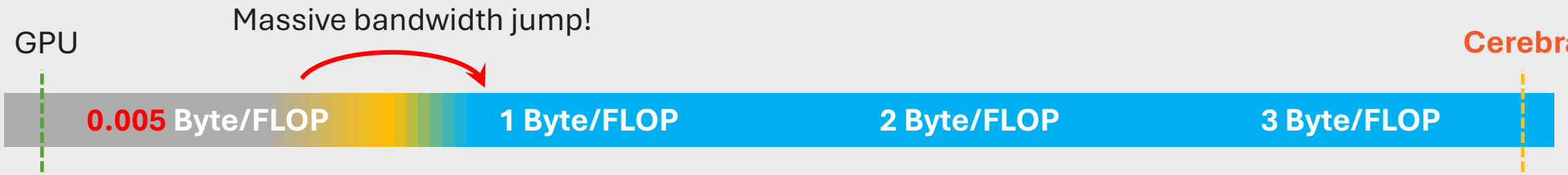
Single cycle access latency
Eight Reads + Four Writes



cerebras

Full memory performance at all BLAS Levels

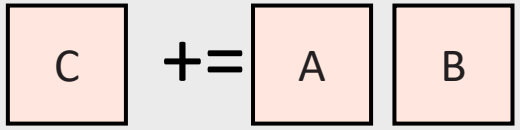
Thanks to massive on-chip memory compared to GPU's on-chip cache



Matrix-Matrix

GEMM

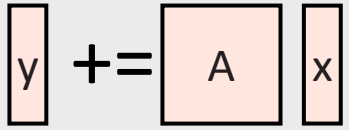
$C \leftarrow \alpha AB + \beta C$



Matrix-Vector

GEMV

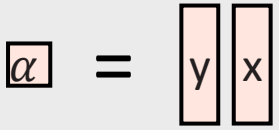
$y \leftarrow \alpha Ax + \beta y$



Vector-Vector

DOT

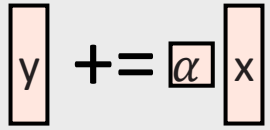
$\alpha \leftarrow (x, y)$



Vector-Scalar

AXPY

$y \leftarrow \alpha x + y$

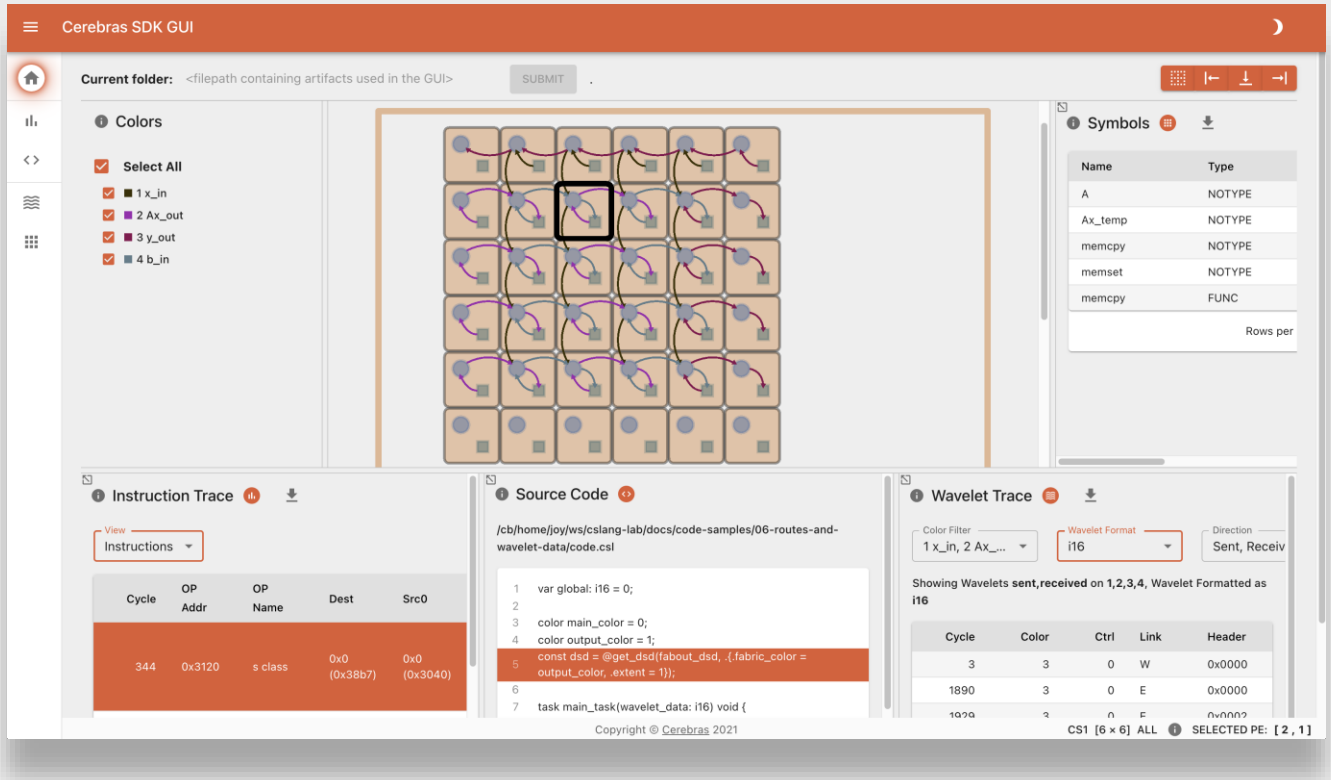


cerebras

Freely available Cerebras SDK and other tools

A parallel-computing platform & API enable custom programs (“kernels”)

Language	CSL: Cerebras Software Language
	Host APIs with Python
Libraries	Optimized primitives
Tools	Simulator
	Debugger
	Visualization



Discourse

SDK Access

CSL Code examples

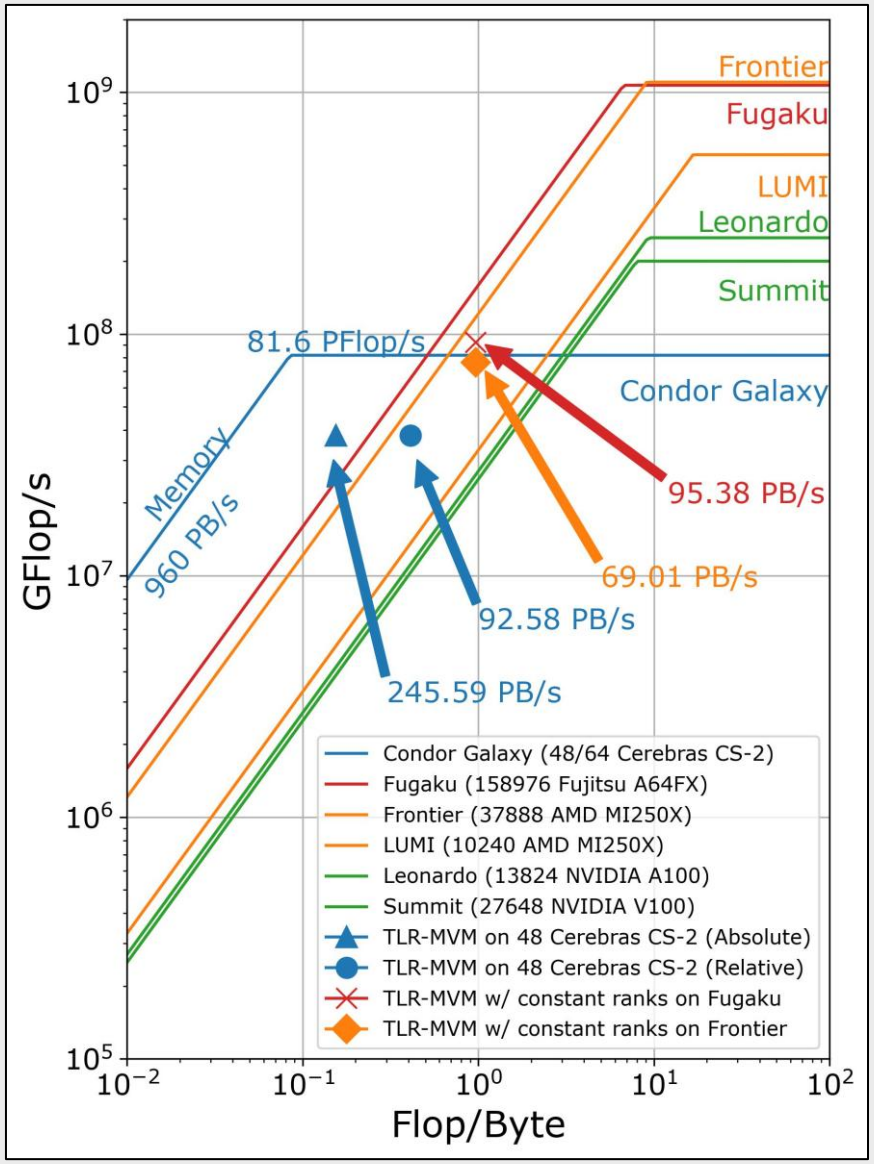
- Tutorials
- GEMV
- GEMM
- Cholesky
- Decomposition
- 1D, 2D, 3D FFT
- 7-Point Stencil SpMV
- Power Method
- Conjugate Gradient
- Preconditioned Conjugate Gradient
- Fin. Diff. Stencil
- Computations
- Mandelbrot Set Generator
- Shift-Add Multiplication
- Hypersparse SpMV
- Histogram Computation

Cerebras & KAUST break records on seismic processing

2023 Gordon Bell Prize finalist

- Seismic processing algorithms are memory-bound problems, limited by the memory access speeds of other architectures.
- Redesigned a Tile Low-Rank Matrix-Vector Multiplication (TLR-MVM) algorithm for Cerebras CS-2, taking advantage of the ultra high memory bandwidth
- Simulation on Cerebras Condor Galaxy-1 AI supercomputer
- Achieved sustained memory bandwidth of **92.38 PB/s** across 48 CS-2 systems – higher than Frontier (#1 TOP500), comparable to Fugaku (#4 TOP500)

Paper: https://dl.acm.org/doi/10.1145/3581784.3627042

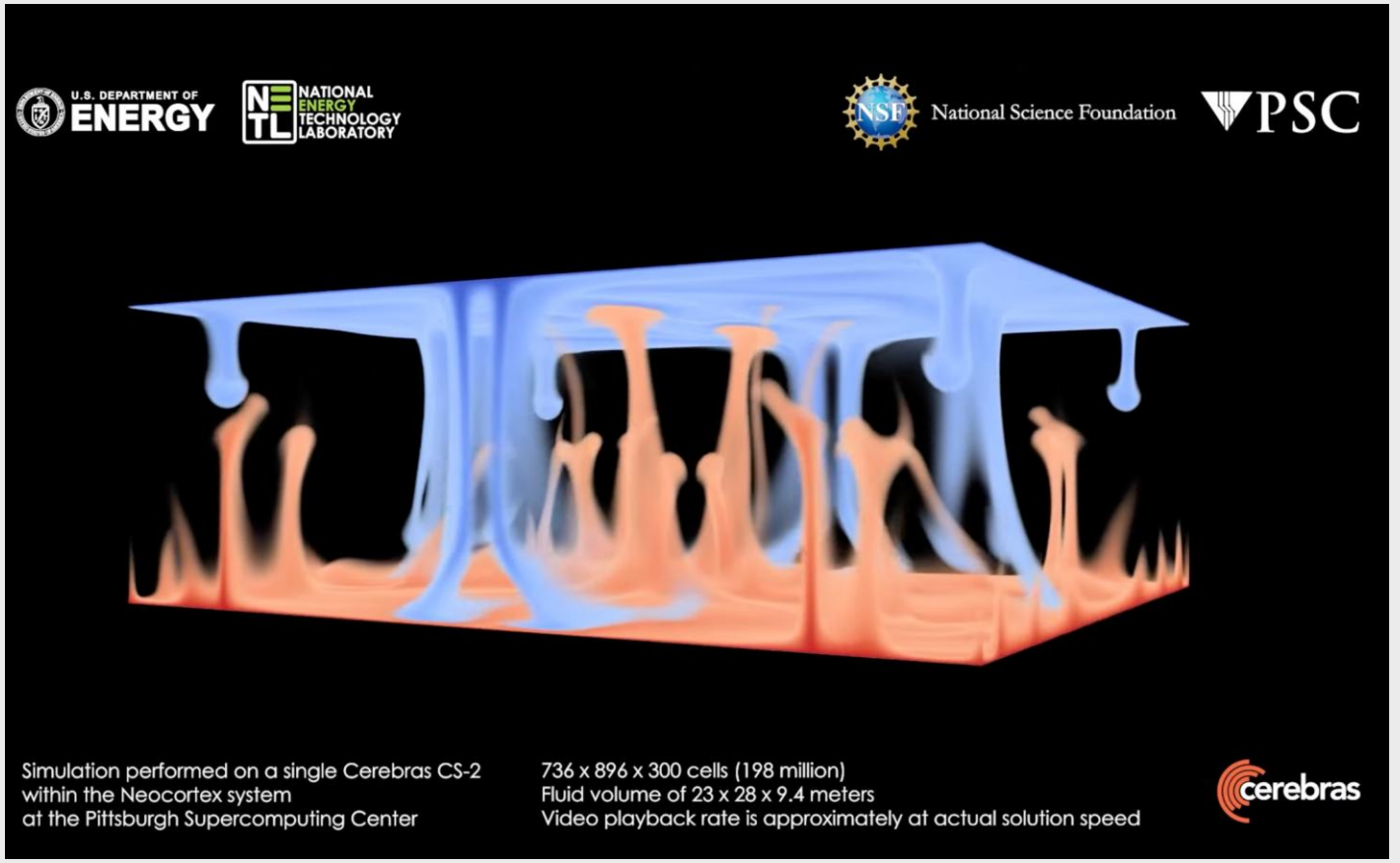


NETL achieves near-real time solution of CFD simulation

HPCWire 2023 Editor's Choice Award for Best Use of HPC in Industry

- CFD on traditional architectures is limited by memory bandwidth and communication
- CS-2 runs **470x faster** simulation of Rayleigh-Bénard convection vs. Joule 2.0 supercomputer
- CS-2 is **1000x more power efficient** vs. Joule 2.0 supercomputer

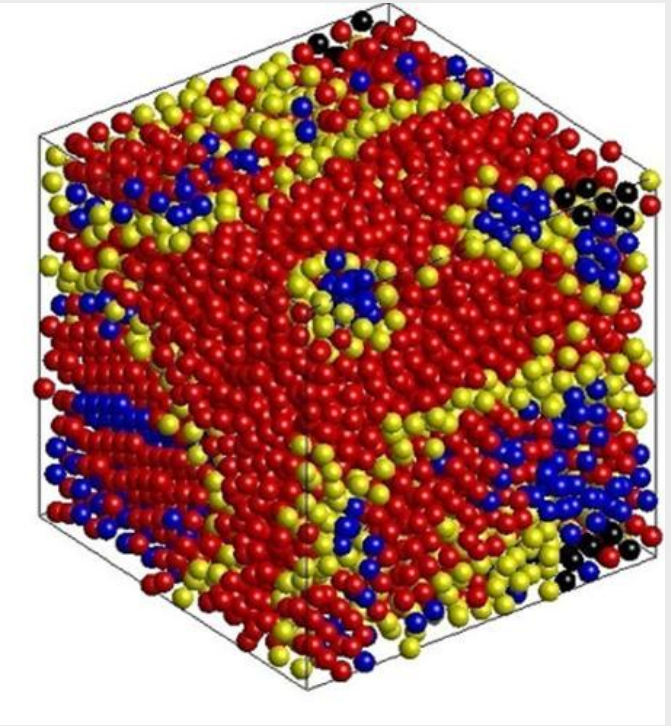
NETL Blog: netl.doe.gov/node/11762
Cerebras Blog: cerebras.net/blog/real-time-computational-physics-with-wafer-scale-processing

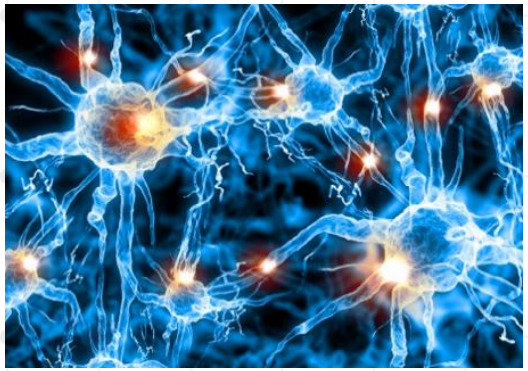
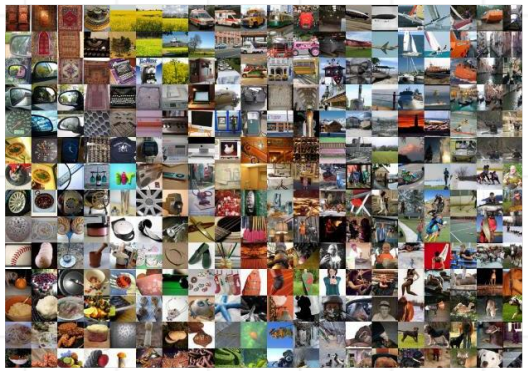
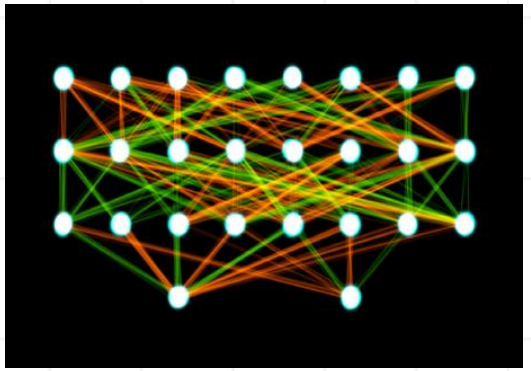
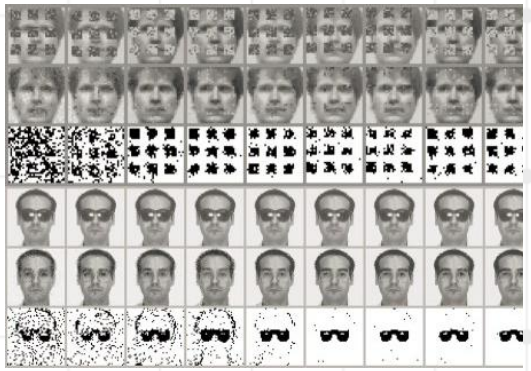
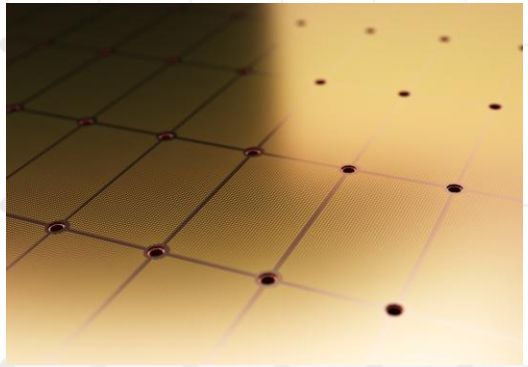


Molecular dynamics on CS-2 1 day = 2 years on Exascale

Gordon Bell Finalist 2024

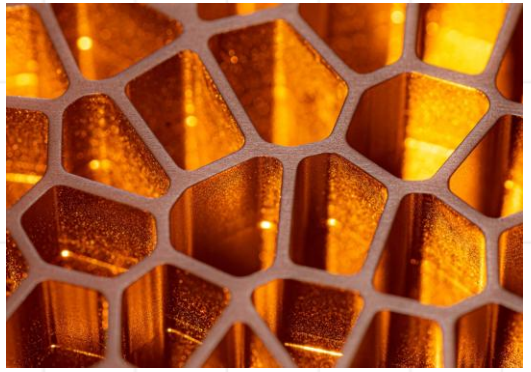
- Embedded Atom Model (EAM) is a molecular dynamics method with an interatomic potential suited for modelling metallic systems
- Strong scaling applies more than one core per simulated atom
- Simulation timestep **1,000x faster than today's SOTA**
- Investigate long time-scale system properties previously infeasible to compute
- Larger molecular systems can scale to cluster of Cerebras nodes with same timestep performance
- Extensions for biomolecules possible



www.cerebras.net




cerebras