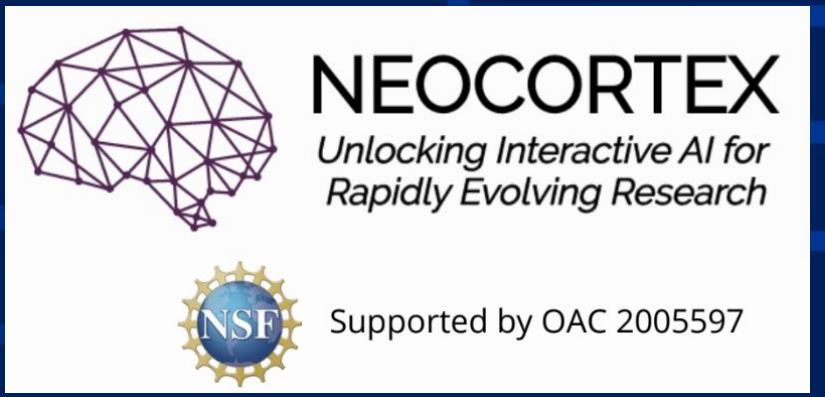
PSC

Carnegie Mellon University

University of Pittsburgh

Neocortex: Introduction to the Cerebras CS-3

Paola A. Buitrago
Neocortex, PI
Director, AI and Big Data | Pittsburgh Supercomputing Center



October 21st, 2025

Webinar Agenda

- Neocortex Program Overview
- Cerebras CS-3 technical overview
- Getting access to the CS-3s on the Cerebras Cloud via the Neocortex Program
- Managing your Allocation - Cloud portion
- How to get support and help?
- Questions and Answers
- Closing remarks

NEOCORTEX

Context – NSF Solicitation

NSF Solicitation – 19-587

Advanced Computing Systems and Services: Adapting to the Rapid Evolution of Science and Engineering Research

"The intent of this solicitation is to request proposals from organizations to serve as service providers … to provide advance cyberinfrastructure (CI) capabilities and/or services … to support the full range of computational- and data-intensive research across all science and engineering (S&E)."

Two categories:

- Category I, Capacity Systems: production computational resources.
- **Category II, Innovative Prototypes/Testbeds: innovative forward-looking capabilities deploying *novel technologies, architectures, usage modes*, etc., and exploring new target applications, methods, and paradigms for S&E discoveries.**

NEOCORTEX

The Neocortex Program

Acquisition and operation of ***Neocortex*** is made possible by the National Science Foundation:

NSF Award OAC-2005597:
Category II: Unlocking Interactive AI Development for Rapidly Evolving Research

Cerebras and HPE delivered *Neocortex*

NSF Solicitation and Award

NSF Solicitation - 19-587

Advanced Computing Systems and Services: Adapting to the Rapid Evolution of Science and Engineering Research

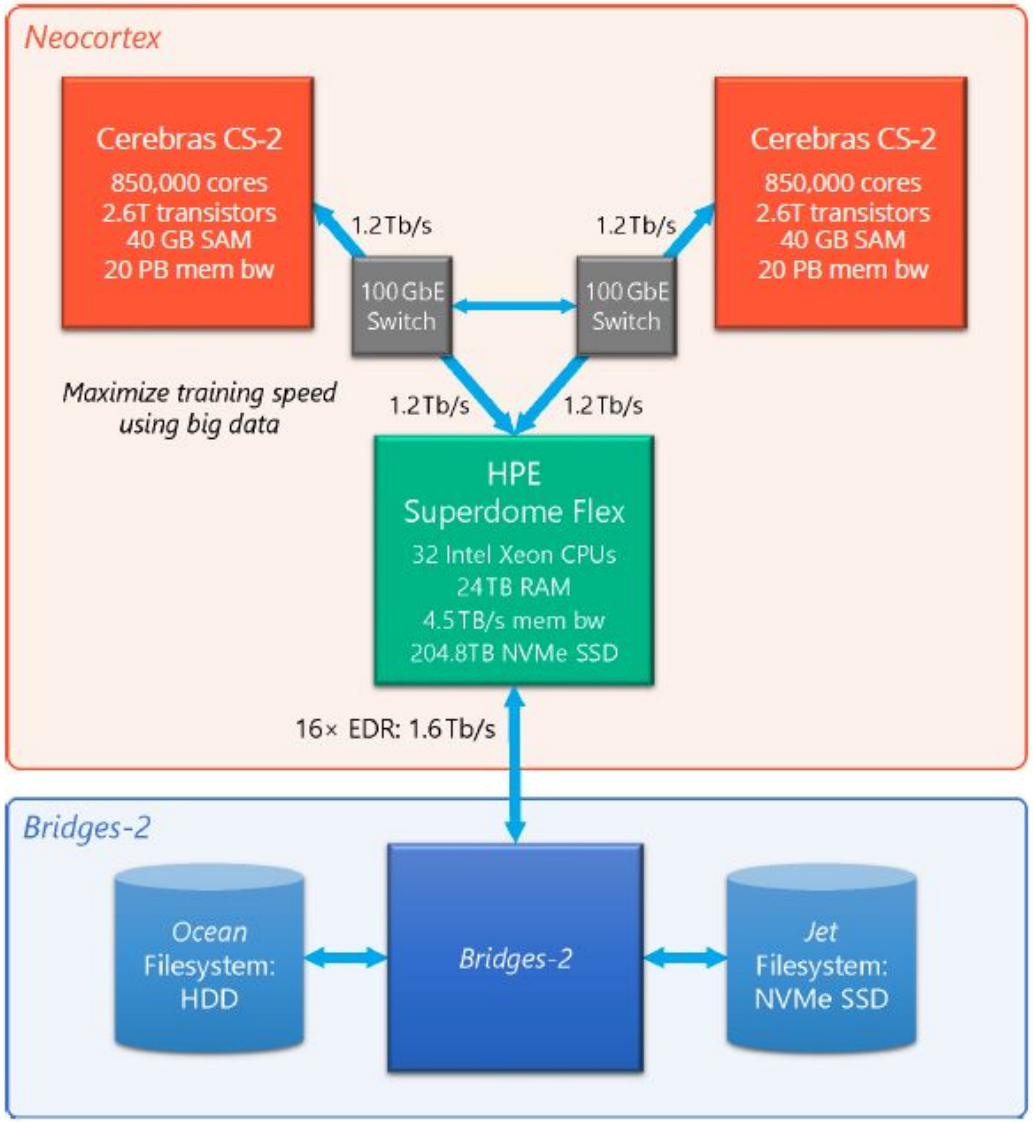
– **Category II, Innovative Prototypes/Testbeds:** innovative forward-looking capabilities deploying novel technologies, architectures, usage modes, etc., and exploring new target applications, methods, and paradigms for S&E discoveries.

*Acquisition and operation of Bridges, Bridges-AI, Bridges-2, and **Neocortex** are made possible by the National Science Foundation:*

NSF Award OAC-2005597 ($12.25M awarded to date):

Cerebras and HPE delivered *Neocortex*

Neocortex System Architecture Overview (Original System)



The main AI accelerators are the Cerebras Wafer Scale Engines (WSE-2)

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

NEOCORTEX

Neocortex System - Developer Cloud Access to CS-3

The main AI accelerators are two

Cerebras Wafer Scale Engines (WSE-3)

Specifications

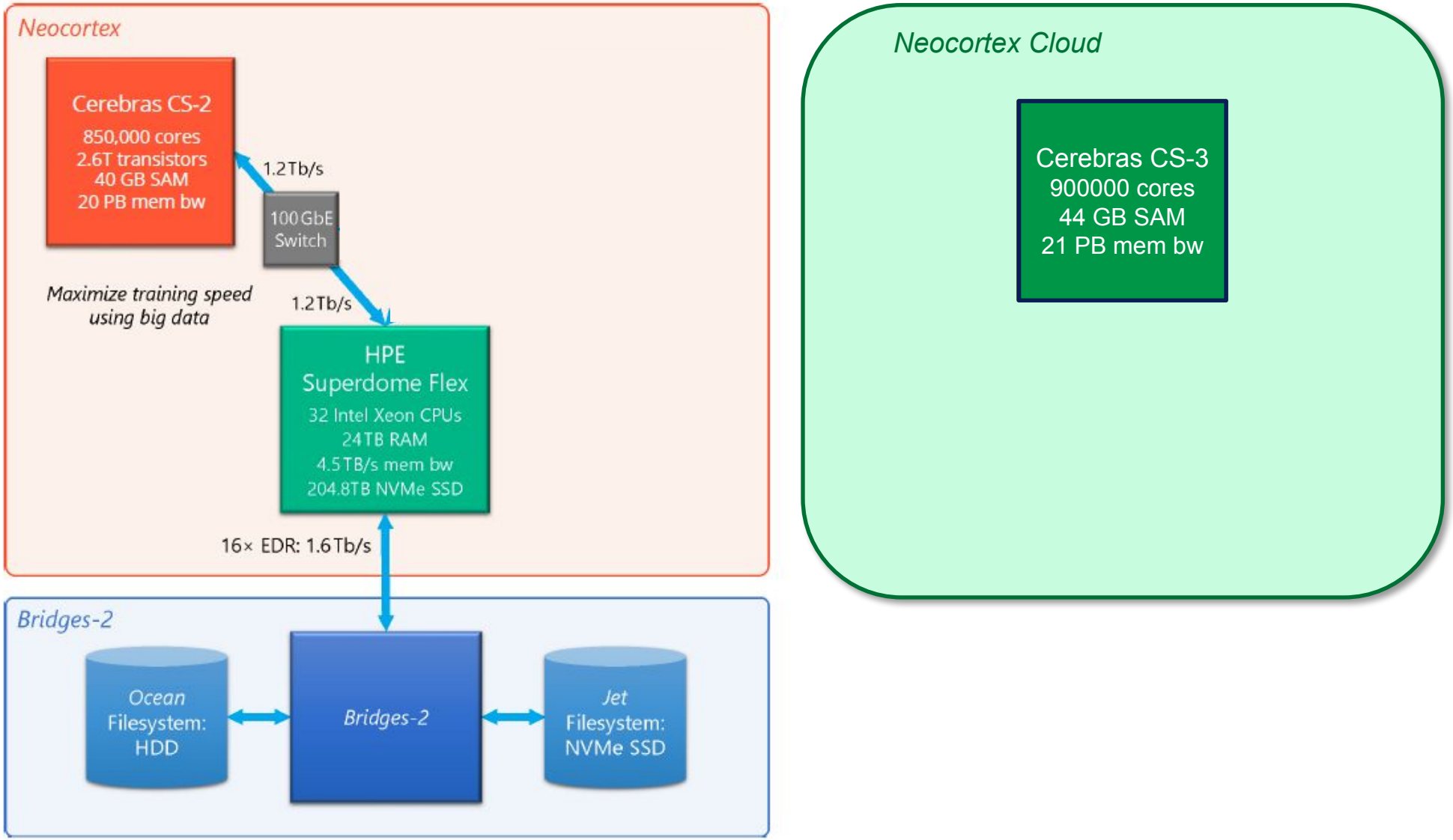
- 900,000 compute cores
- 125 PetaFLOPs of AI Performance
- 44 GB on-chip memory
- 12 to 1,200 TB of off-chip model memory
- 21 PB/sec memory bandwidth
- 214 PB/sec core-to-core bandwidth

Available for selected projects.

Reach out to neocortex@psc.edu if you are interested!

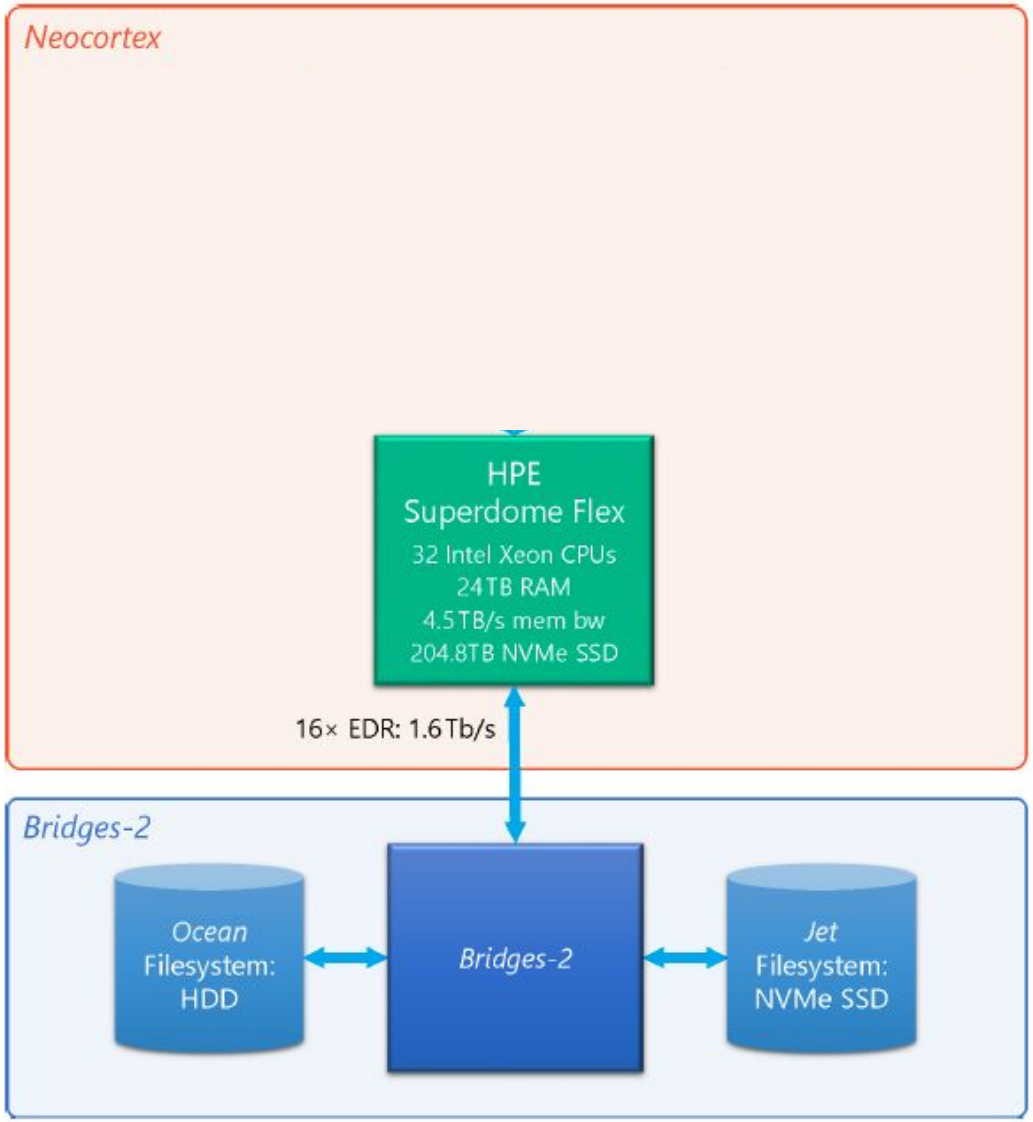
NEOCORTEX

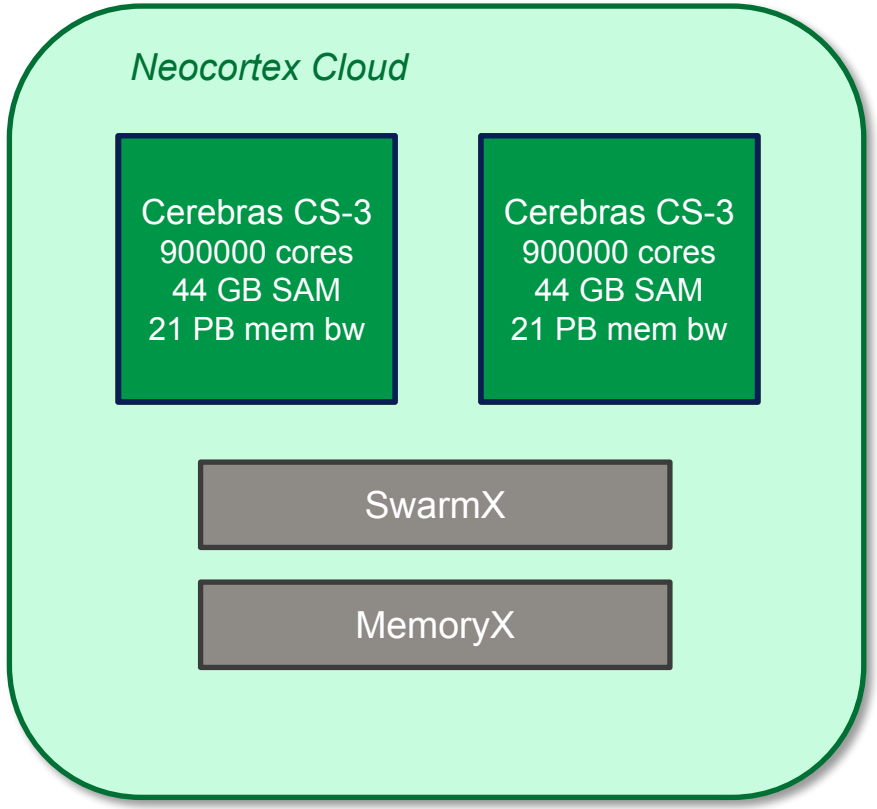
Neocortex System Architecture Overview (Hybrid Model)



NEOCORTEX

Neocortex System Architecture Overview (Cloud Model)





NEOCORTEX

Neocortex System – Getting Access

You can access the Neocortex System through the ACCESS and the NAIRR programs.

ACCESS

NAIRR

NEOCORTEX

To Learn More and Participate

Byteboost program	https://www.stonybrook.edu/ookami/ByteBoost.php
Access program	https://access-ci.org/
NAIRR pilot	https://nairrpilot.org/
Neocortex documentation	https://portal.neocortex.psc.edu/docs/index.html
Contact us with additional questions, inputs, requests > Office hours (Wednesdays, 2 pm ET)	Email: neocortex@psc.edu https://www.psc.edu/resources/neocortex/neocortex-office-hours/

NEOCORTEX

Webinar Agenda

- Neocortex Program Overview
- **Cerebras CS-3 technical overview**
- Getting access to the CS-3s on the Cerebras Cloud via the Neocortex Program
- Managing your Allocation - Cloud portion
- How to get support and help?
- Questions and Answers
- Closing remarks

Managing your Grant

- **No change** from current processes discussed in Allocations - Neocortex Documentation:
  - Accounting information from the CS3 Cloud is sent to the PSC database. You can view and manage it by logging into Neocortex, just like before the move to Cloud.
- If you have an ACCESS grant, you can manage Supplements, Extensions, and Renewals as discussed in https://allocations.access-ci.org/how-to
- If you have a NAIRR grant, you can manage Supplements, Extensions, and users as discussed in https://nairrpilot.org/help/faq

Help and Support

- **No change** from current processes:
  - Email: neocortex@psc.edu
  - Slack: Neocortex Slack - Neocortex Documentation
  - Office hours (Wednesdays, 2 pm ET except holidays): **Zoom**
  - Calendly - on demand - specialized 30 min session:
    - ML/AI Office hours
    - SDK Office hours
- **Cerebras CS3 documentation:**
  - http://training-docs.cerebras.ai/
    - Setup & Installation and the Modelzoo CLI to get up and running with basic commands.
    - Modelzoo Overview to build a mental model of the library and understand where they should begin based on their skill level.
    - The Converter Tool for leveraging pretrained models.
    - Writing a Custom Training Loop for learning how to integrate existing workflows via cstorch.
- **Stay tuned for your invitation to our upcoming Tutorial webinar!**

Next-Gen Wafer Scale Advantage

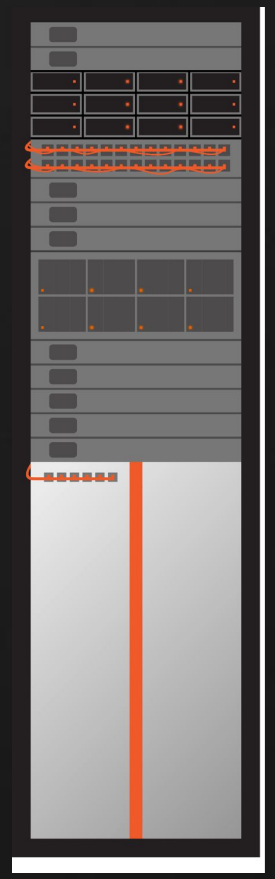
The CS-3 consistently delivers 1.5x+ AI acceleration with the same footprint

	WSE-3	WSE-2
Transistor Count	4 trillion	2.6 trillion
AI Compute Speed	1.5-2x	1x
Memory Bandwidth	24 Petabytes/s	20 Petabytes/s
Fabric Bandwidth	245 Petabits/s	220 Petabits/s

cerebras

Wafer-Scale Cluster

The world's most scalable AI supercomputer



MemoryX

Model weights stored in DDR5 & Flash with CPU running optimizer plus other operations

SwarmX

High-performance interconnect fabric broadcasts weights to CS-3 and returns gradients to MemoryX

CS-3

Up to 2048 CS-3 systems compute activations and gradients

1 terabyte → 1 petabyte

1 CS-3 → 2048 CS-3s

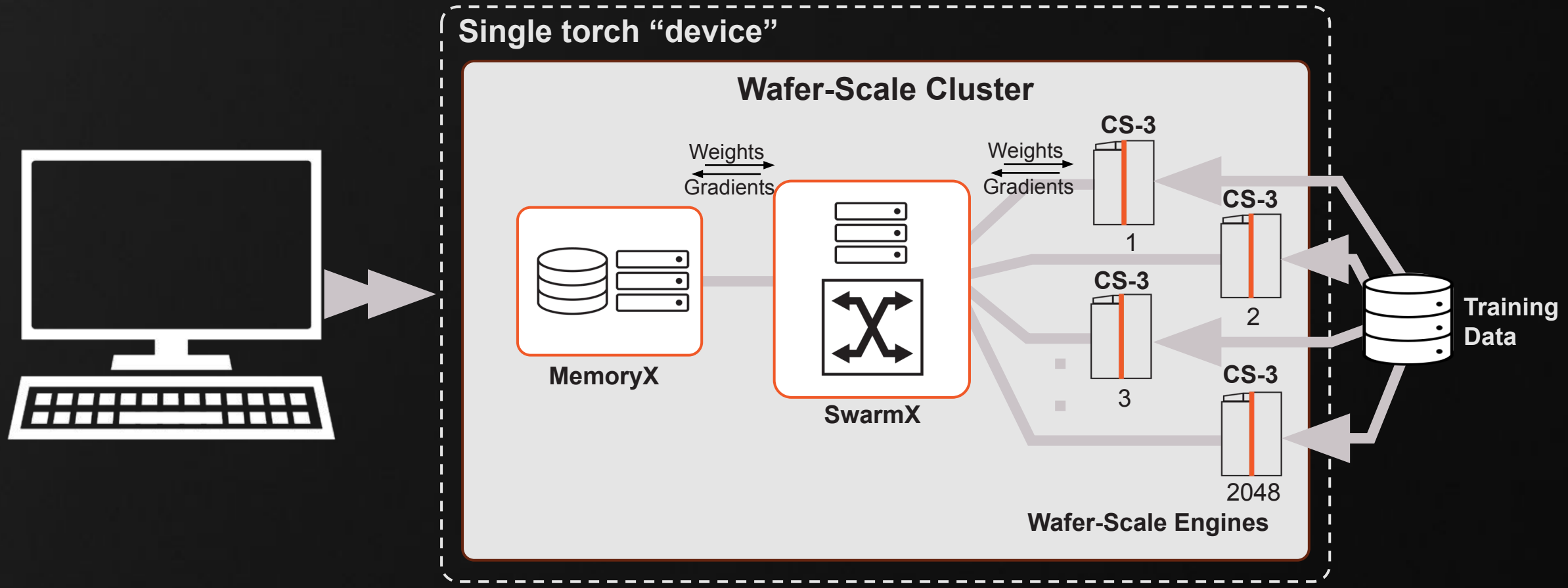
125 petaflops → 256 exaflops

1 billion parameters → 24 trillion parameters

cerebras

Wafer-Scale Cluster

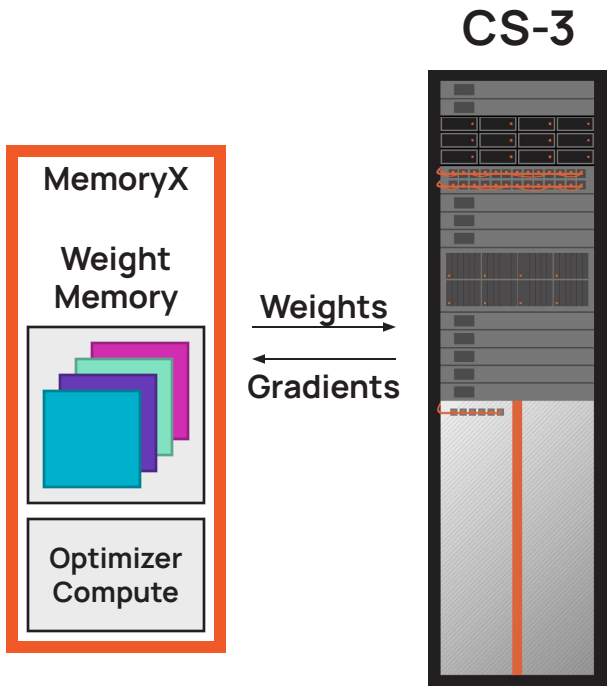
Cluster exposed as a single device in PyTorch with scaling complexity handled automatically



MemoryX External Memory
Virtually Unlimited Model Weight Capacity

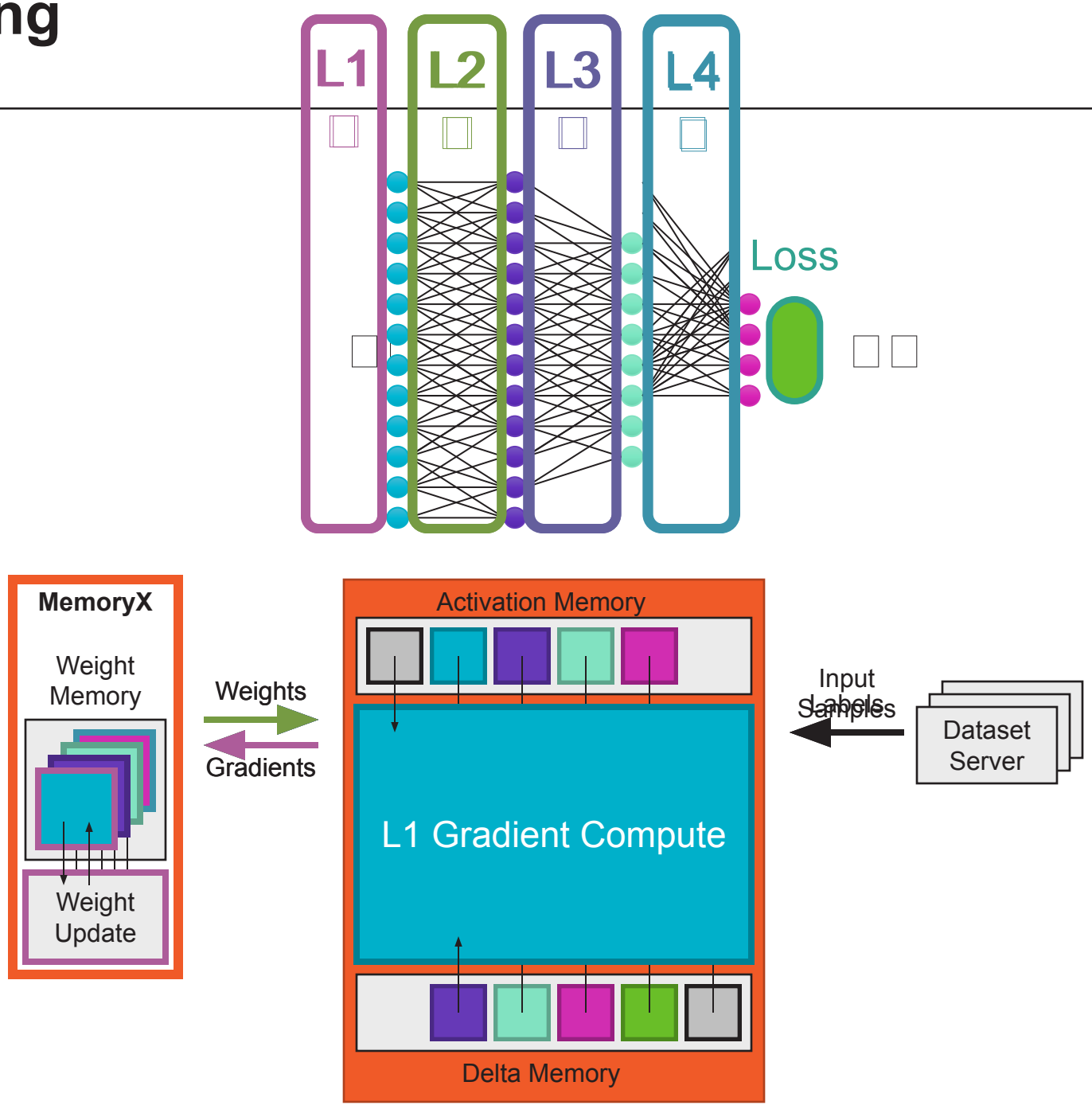
- Model capacity is not limited by wafer memory
- Weights are streamed onto the wafer
- Weights trigger compute using mesh dataflow
- Weights are never stored on-wafer

- Decouples weights and optimizer compute
- Gradients get streamed out of the wafer
- Weight update occurs inside MemoryX



Memory hierarchy capable of massive models on single device

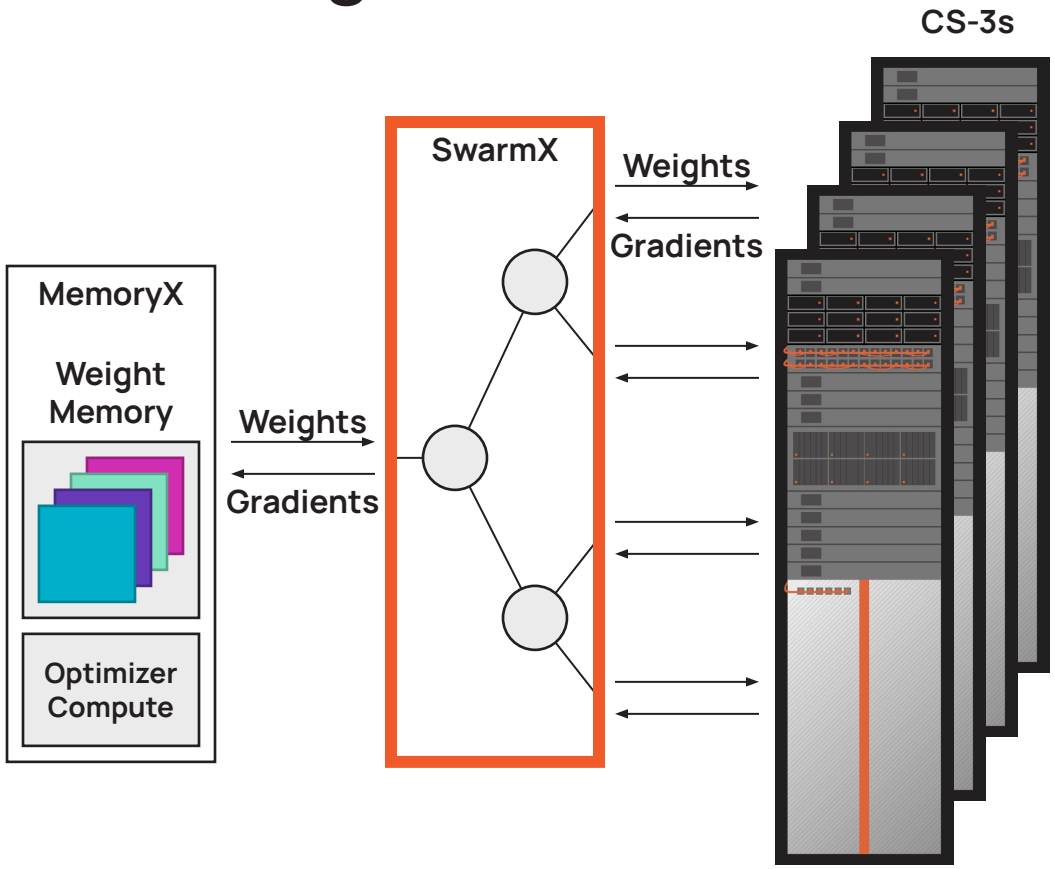
Weight Streaming In Action



SwarmX Fabric
Purpose Built Interconnect for Simple Scaling

- Data-parallel only training across CS-3s
- Weights are broadcast to all CS-3s
- Gradients are reduced on way back

- Multi-system scaling with the same execution model as single system
- Same system architecture
- Same network execution flow
- Same software user interface



Scaling to cluster compute while operating like a single device

Cerebras Modelzoo: Batteries Included

- Models
  - NLP
  - Vision
  - Multimodal
- Data Processors
  - NLP
  - Vision
  - Multimodal
- Tools
  - Checkpoint Converters
  - Config Converters
- Common
  - Class Registry
  - Checkpoint Utils
  - Run Utils
  - Eval Harnesses
- Trainer
  - Training logic
  - Callbacks
  - Loggers
  - Config Validation
- Layers & Losses
  - Transformer Blocks
  - Position Embeddings
  - Attention Variants
  - Losses for LM, Bert, DPO, CLIP...

cerebras

Streamlined Setup

- The modelzoo is now locally pip-installable!
- Specifying manual mount and python paths is a thing of the past.
- Simple, three-step setup:
  - `git clone https://github.com/Cerebras/modelzoo.git && cd modelzoo`
  - `pip install -r requirements.txt`
  - `pip install -e .`

The Modelzoo Command-Line Interface

- The modelzoo now comes with an accessible command-line interface (CLI)

- Supported commands include:
  - Model training and evaluation
  - Conversion of model configs and checkpoints
  - Summarizing supported models and breaking down a given model config
  - Data preprocessing and breaking down a given `modelzoo DataProcessor`

- The modelzoo also has an LLM assistant! It can:
  - Answer questions about Cerebras' hardware, software, and documentation
  - Process natural language instructions to execute modelzoo CLI commands (with permission)
  - All you need is an API key from inference.cerebras.ai

cerebras

Simplified Job Execution and Scale-Out

```
cszoo fit gpt3_1010b.yaml --num_csx=1
```

```
cszoo fit gpt3_1010b.yaml --num_csx=16
```

No fuss required.

cerebras

Modelzoo CLI: Converting to Huggingface

```
cszoo checkpoint convert \
--model llama \
--src-fmt cs-auto \
--tgt-fmt hf \
--config pretraining_tutorial/model_config.yaml \
--output-dir pretraining_tutorial/to_hf \
pretraining_tutorial/model/checkpoint_0.mdl
```


cerebras

Cerebras SDK on CS-3 Wafer-Scale Cluster

- Cerebras SDK programs which run on the simulator can run on a node of the Wafer-Scale Cluster.
- Wrapper scripts are provided to compile and run:

`cslc` → `SdkCompiler()` wrapper

`SdkRuntime` host script → `SdkLauncher()` wrapper

- See https://sdk.cerebras.net/appliance-mode for examples and instructions.
- See https://sdk.cerebras.net/csl/code-examples/ for examples of WSE-2 vs. WSE-3 microcode.

Accessing the Cerebras CS-3 Cloud Server

Mei-Yu Wang

Machine Learning Research Scientist,

Pittsburgh Supercomputing Center

October 21st, 2025

Getting Started: Request Access

To begin using the CS-3 cloud, you'll need to provide our team with the following information:

- Full name
- Email address
- Allocation ID
- SSH public key

Submit your request through email to **neocortex@psc.edu**.

How to Generate Your SSH public key

- If you do not already have an SSH key, generate one using the following command (replace with your actual email):

```
ssh-keygen -t rsa -C "your_email@example.com"
```

Do not forget your password if you choose to set one!

- For Windows 10/11, You can open **Open Command Prompt** or **PowerShell,** then run the command above. Alternatively, you can use **PuTTY** to generate the key.
- After generation, display the public key:

```
cat $HOME/.ssh/id_*.pub
```

Share only your **public key** with the Neocortex team (email neocortex@psc.edu) while keeping your private key secure.

For more information, please visit the SSH Project webpage.

- Example Output:

```
ssh-rsa
AAAAB3NzaC1yc2xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxkL2ZsBhZwalFt+gxX mwang7@psc.edu
```

VPN Setup: Download and Install

Once your access is set up, you will receive:

- **Cerebras Cloud Credentials**
- **VPN Configuration Instructions**

VPN Connection¶

1. Download the **GlobalProtect** VPN client from: https://access01.vpn.cerebras.net
2. Use your **Cerebras-provided VPN credentials** to log in.
3. Configure the VPN with **Portal Address**: access01.vpn.cerebras.net
4. Connect to the VPN by entering your credential and password.

Connecting to the Cerebras CS-3 Cloud

- After establishing a VPN connection, access the system via SSH:

  ```
  ssh -i $HOME/.ssh/id_KEY <cerebras_username>@cg3-us27.dfw1.cerebrascloud.com
  ```

  Replace `<cerebras_username>` with your assigned username and `id_KEY` with your private key.

- No password prompt will appear (authentication via SSH key)
- To verify VPN connectivity:

  ```
  ping cg3-us27.dfw1.cerebrascloud.com
  ```

Contact us with additional questions, inputs, requests:

Email: neocortex@psc.edu

You can find detailed instructions about accessing Cerebras CS-3 cloud here: