Welcome!

Thank you for joining us today! As we wait for everyone to get settled, we'd like to bring a few things to your attention:

1. This webinar is being recorded. The recording will be available via the official YouTube channel and the Neocortex webpage this week.
2. There will be 45 minutes of presentation followed by Q&A. To maintain a quality experience for everyone, please mute your microphone during the presentations.
3. We hope you will participate in this interactive webinar by:
   - Asking questions to our team via the Q&A Zoom feature.

   These questions will seed the Q&A session in the final 15 minutes.
4. This webinar abides to the XSEDE code of conduct.

XSEDE Code of Conduct

XSEDE has an external code of conduct which represents our commitment to providing an inclusive and harassment-free environment in all interactions regardless of race, age, ethnicity, national origin, language, gender, gender identity, sexual orientation, disability, physical appearance, political views, military service, health status, or religion. The code of conduct extends to all XSEDE-sponsored events, services, and interactions.

Code of Conduct: https://www.xsede.org/codeofconduct

Contact:

- Event organizer: *PSC*
- XSEDE ombudspersons:
  - Linda Akli, Southeastern Universities Research Association (akli@sura.org)
  - Lizanne Destefano, Georgia Tech (lizanne.destefano@ceismc.gatech.edu)
  - Ken Hackworth, Pittsburgh Supercomputing Center (hackworth@psc.edu)
  - Bryan Snead, Texas Advanced Computing Center (jbsnead@tacc.utexas.edu)
- Anonymous reporting form available at https://www.xsede.org/codeofconduct.

Neocortex Overview and Call for Proposals

Paola A. Buitrago
Neocortex, Principal Investigator & Project Director
Director, AI and Big Data, Pittsburgh Supercomputing Center

October 4, 2021

Overview

- The Neocortex System: Context
- The Neocortex System: Motivation
- Hardware Description
- Early User Program and Exemplar Use Cases
- Call for Proposal

The Neocortex System

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

Context – NSF Award

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

NSF Award OAC-2005597 ($5M awarded to date):
Category II: Unlocking Interactive AI Development for Rapidly Evolving Research

Cerebras and HPE delivered *Neocortex*

Context – Project Goals

***Neocortex*, Unlocking Interactive AI Development for Rapidly Evolving Research**

A new NSF funded advanced computing project with the following goals:

- Deploy *Neocortex* and offer the national open science community revolutionary hardware technology to accelerate AI training at unprecedented levels.
- Explore, support and operate *Neocortex* for 5 years.
- Engage a wide audience and foster adoption of innovative technologies.

Context – Project Goals

***Neocortex*, Unlocking Interactive AI Development for Rapidly Evolving Research**

A new NSF funded advanced computing project with the following goals:

- ~~Deploy *Neocortex* and offer the national open science community revolutionary hardware technology to accelerate AI training at unprecedented levels.~~
- Explore, support and operate *Neocortex* for 5 years.
- Engage a wide audience and foster adoption of innovative technologies.

Neocortex Timeline

June 1, 2020	Award start date; preparatory activities begin – System and user environment, documentation, content, dissemination, etc. – Broadly invite researchers for the Early User Program
Fall 2020	Start of delivery, installation, initial testing
Feb 2021	System fully deployed and integrated
	Users gain early access
Summer 2021	Conclusion of Early User Program & Acceptance Testing
Aug 2021	Start of **Neocortex Testbed Operations**
Oct 2021	Call for Proposals

Why did we Propose Neocortex?

"Prior to 2012, AI results closely tracked Moore's Law, with compute doubling every two years. Post-2012, compute has been doubling every 3.4 months."

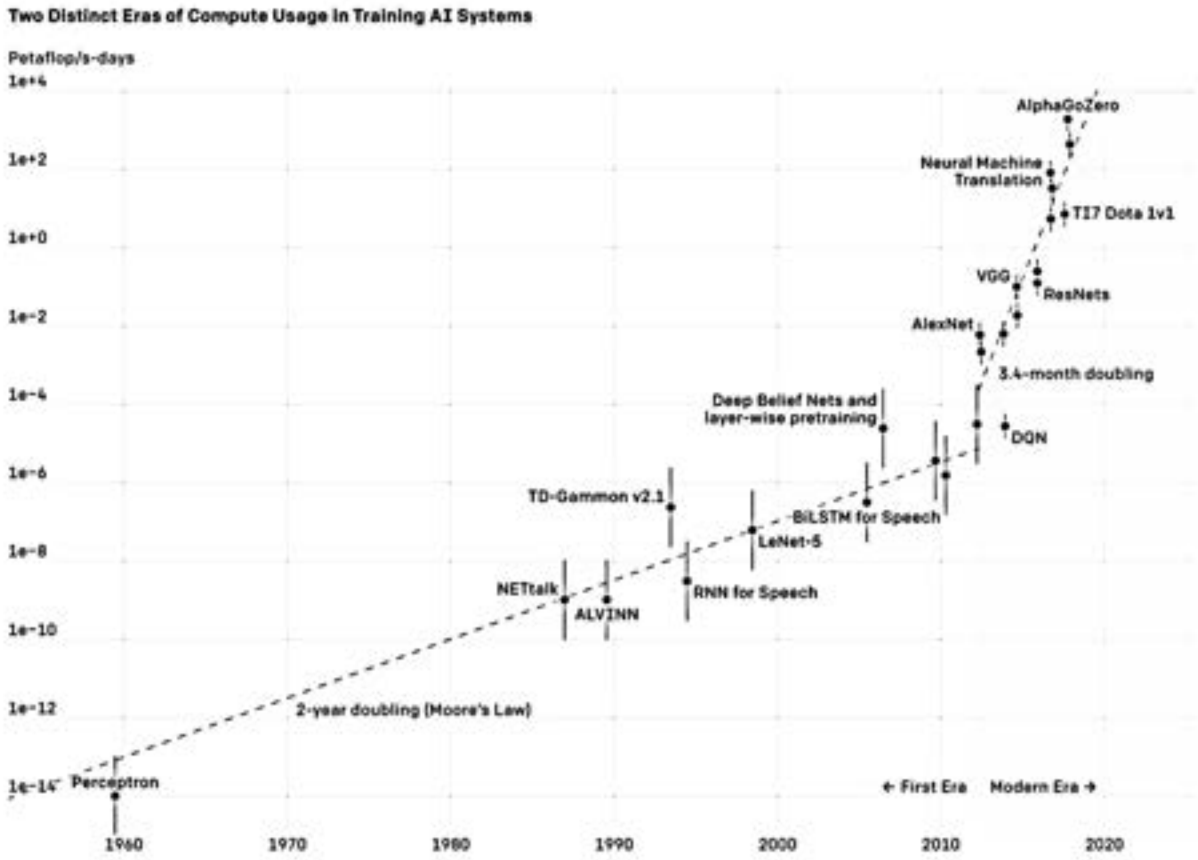


Figure from D. Amodei, D. Hernandez, G. SastryJack, C. Greg, and B. Sutskever. (2019, November 7). *AI and Compute*, OpenAI Blog. https://openai.com/blog/ai-and-compute.

NEOCORTEX

Why did we Propose Neocortex?

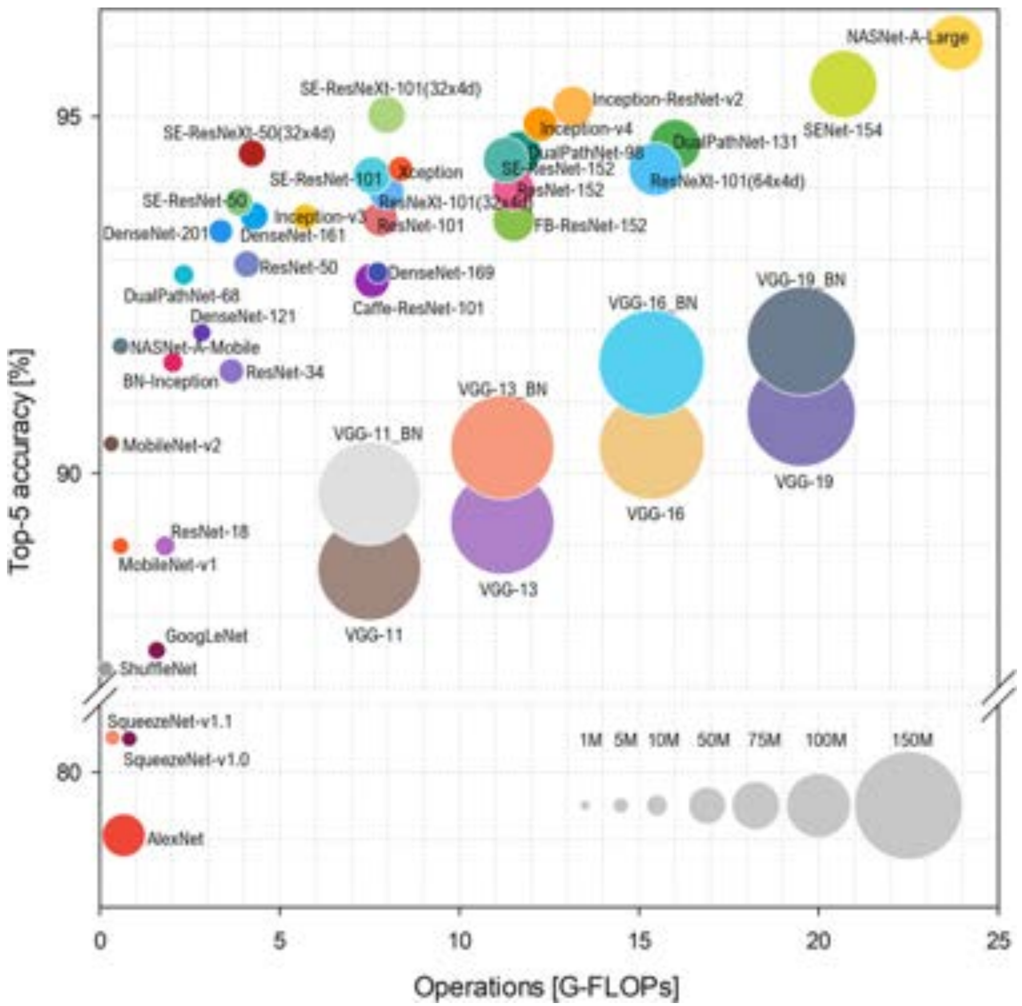


Figure from S. Bianco, R. Cadene, L. Celona, and P. Napoletano, *Benchmark Analysis of Representative Deep Neural Network Architectures*, IEEE Access, vol. 6, pp. 64270– 64277, 2018. arXiv:1810.00736v2.

Network	Published	Parameters
BERT Large	October 11, 2018	340M
PEGASUS Large	December 18, 2019	568M
GPT-2 (48 layers)	February 2019	1.5B
Megatron-LM	August 13, 2019	8.3B
GPT-3 (96 layers)	June 3, 2020	175 B
Switch Transformers	Jan 11, 2021	1.6 T

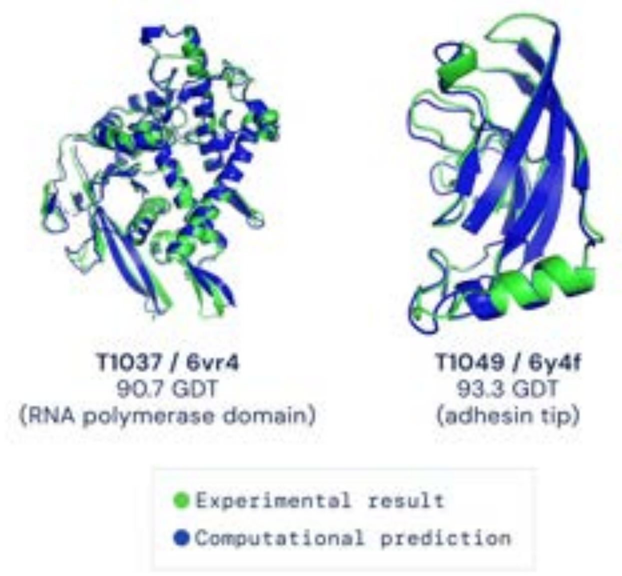
Sources of Additional Complexity

Generative Adversarial Networks (GANs)
Domain Adaptation
Reinforcement Learning (RL)

NEOCORTEX

Driving Use-Cases

- Transform and accelerate AI-enabled research
- Development of new and more efficient AI algorithms and graph analytics
- Foster greater integration of artificial deep learning with scientific workflows
- Democratize access to game changing compute power
- Explore the potential of a groundbreaking new hardware architecture
- Support research needing large-scale memory (genomics, brain imaging, simulation modeling)
- Augmenting traditional computational science with rapidly-evolving methodologies and technologies
- User-centric and interactive computing modalities



Animation from https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology. Retrieved on August 2021.

NEOCORTEX

Neocortex Hardware Description

Cerebras CS-1

Each CS-1 features a *Cerebras* WSE (Wafer Scale Engine), the largest chip ever built.

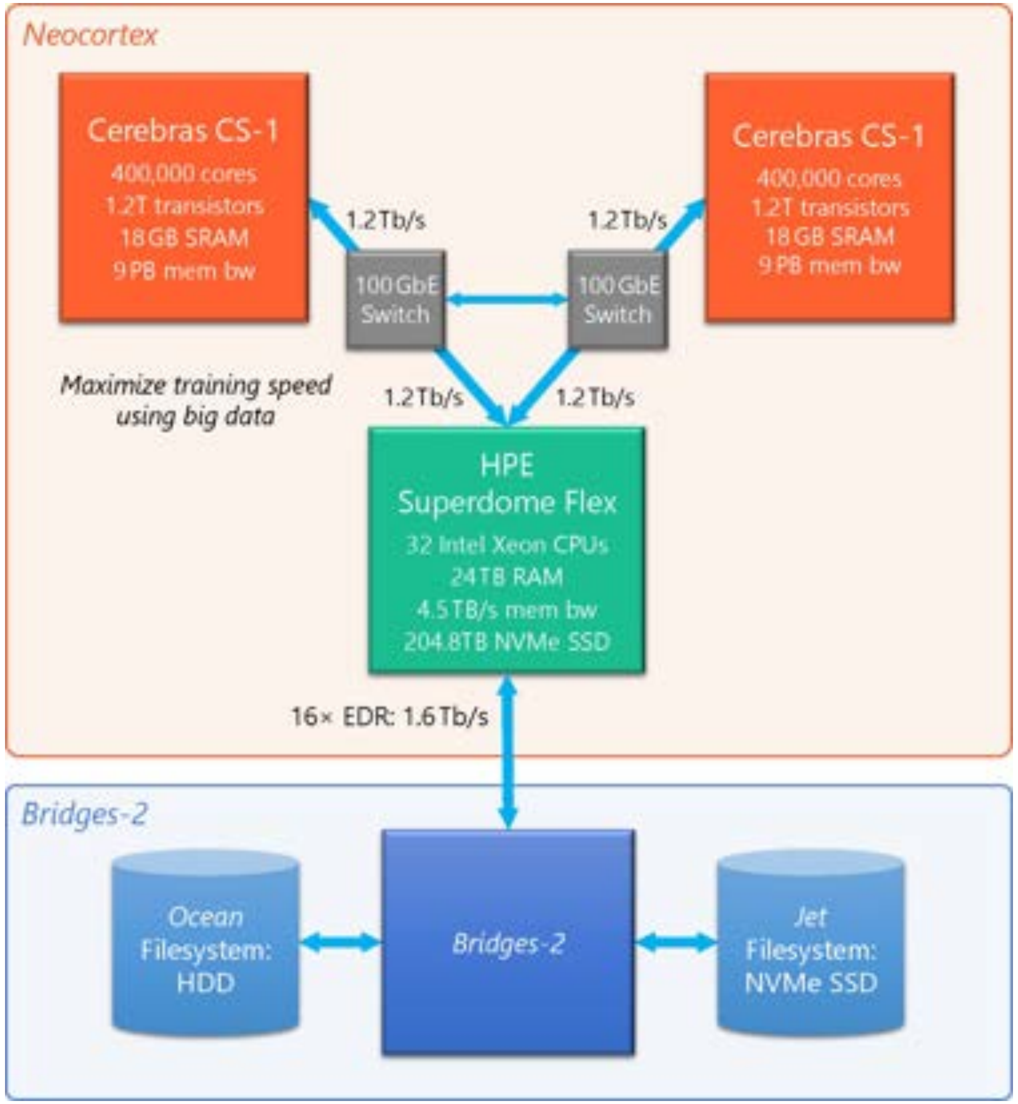
AI Processor	*Cerebras* Wafer Scale Engine (WSE) ◦ 400,000 Sparse Linear Algebra Compute (SLAC) Cores ◦ 1.2 trillion transistors ◦ 46,225 mm^2 ◦ 18 GB SRAM on-chip memory ◦ 9.6 PB/s memory bandwidth ◦ 100 Pb/s interconnect bandwidth
System I/O	1.2 Tb/s (12 × 100 GbE ports)

HPE Superdome Flex

Processors	32 x Intel Xeon Platinum 8280L, 28 cores, 56 threads each, 2.70-4.0 GHz, 38.5 MB cache (more info).
Memory	24 TiB RAM, aggregate memory bandwidth of 4.5 TB/s
Local Disk	32 x 6.4 TB NVMe SSDs ◦ 204.6 TB aggregate ◦ 150 GB/s read bandwidth
Network to CS-1 systems	24 x 100 GbE interfaces ◦ 1.2 Tb/s (150 GB/s) to each Cerebras CS-1 system ◦ 2.4 Tb/s aggregate
Interconnect to Bridges-2	16 Mellanox HDR-100 InfiniBand adapters ◦ 1.6 Tb/s aggregate
OS	Red Hat Enterprise Linux

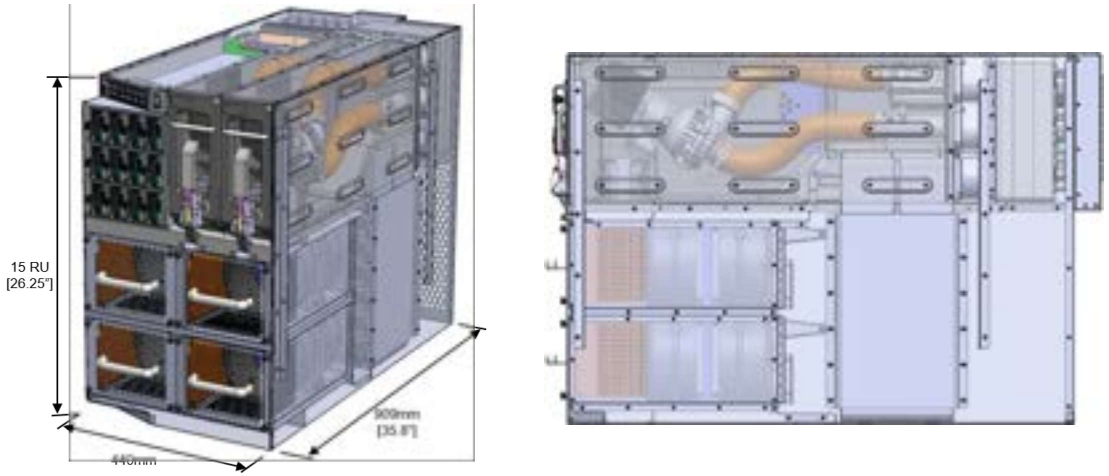
NEOCORTEX

Neocortex System Overview



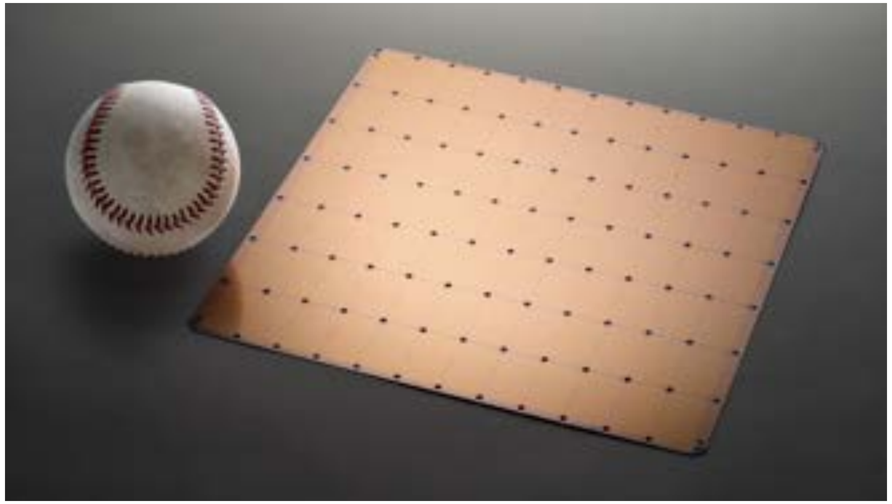
NEOCORTEX

The CS-1 Server



Interior view of the Cerebras CS-1

Wafer Scale Engine (WSE) Processor



Cerebras CS-1 – The WSE

- Powered by the Cerebras Wafer Scale Engine (WSE):
- Largest chip ever built: 46,225 mm2 silicon, 1.2 trillion transistors
- 400,000 AI optimized cores
- 18 GB on chip memory—all 1 clock cycle from the cores
- 9.6 PByte/s aggregate memory bandwidth
- 100 Pbit/s fabric bandwidth
- System IO: 12 x 100 GbE
- System power: 20 kW
- Ingests TensorFlow, PyTorch, etc.

Cerebras CS-1 server, 15 RU

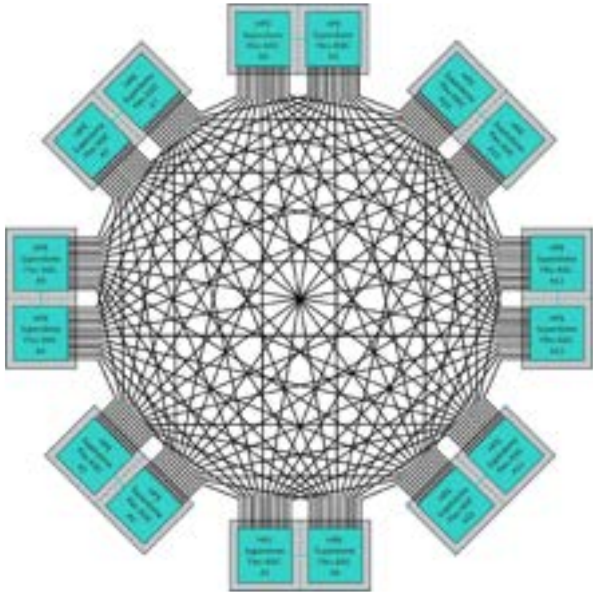
NEOCORTEX

The HPE Superdome Flex

HPE Superdome Flex HPC Server

The HPE Superdome Flex:

- Provides substantial capability for preprocessing and other complementary aspects of AI workflows.
- Enables training on very large datasets with exceptional ease.
- Supports both CS-1s independently and (will support them) together to explore scaling.



Superdome crossbar topology – 850 GB/s of bisection bandwidth

Early User Program (EUP)

- Reviewed 42 project proposals
- Applicants from 21 institutions
- Welcomed 17 for the EUP

Neocortex EUP applicants and their institutions.

NEOCORTEX

Dissemination Activities

- 5 Webinars and training opportunities delivered by the project.
- 14 speaking engagements in conference, university classes, and academic workshops.
- 2 scientific paper published.
- Project website and social media channels.


System Integration of Neocortex, a Unique, Scalable AI Platform


Neocortex and Bridges-2: A High Performance AI+HPC Ecosystem for Science, Discovery, and Societal Good

Paola A. Buitrago, Julian Uran, and Nicholas A. Nystrom. 2021. System Integration of Neocortex, a Unique, Scalable AI Platform. In PEARC21: Practice & Experience in Advanced Research Computing Conference Series, July 19–22,Virtual Conference.ACM, New York, NY, USA, 4 pages..

Paola A. Buitrago and Nicholas A. Nystrom. 2021. Neocortex and Bridges-2: A High Performance AI+HPC Ecosystem for Science, Discovery, and Societal Good. In High Performance Computing, Sergio Nesmachnow, Harold Castro, and Andrei Tchernykh (Eds.). Springer International Publishing, Cham, 205–219.

NEOCORTEX

Clinical Diagnosis and Prognosis in Acute Settings Using Deep Learning

Sardar Ansari, PhD Jonathan Motyka, University of Michigan

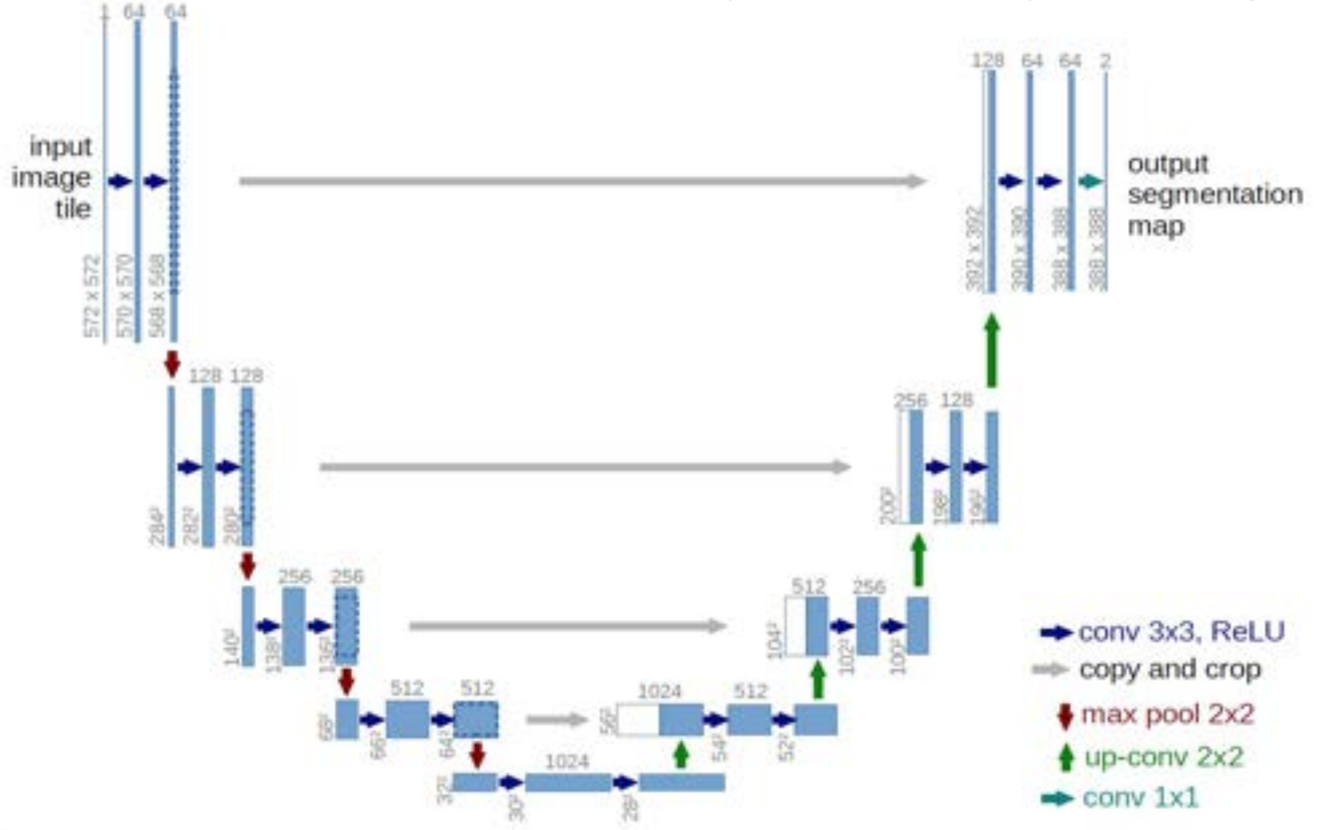


Figure 1. U-Net architecture as depicted in original publication [1].

1. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. LNCS, vol. 9351, pp. 234–241. Springer(2015)

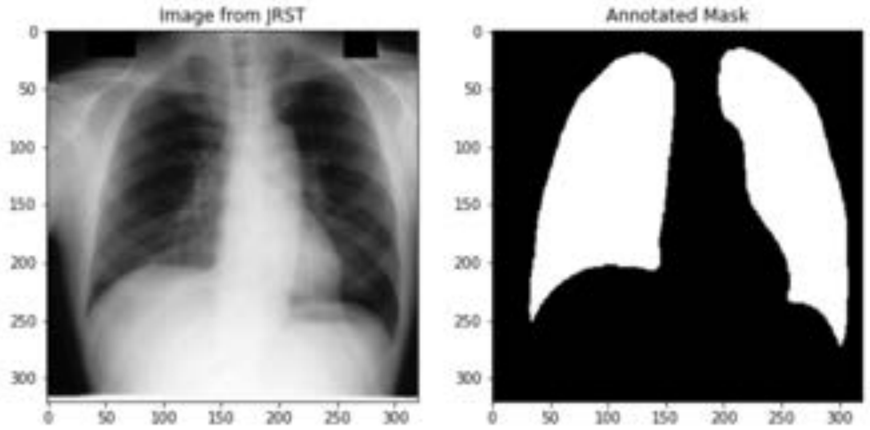


Figure 2 Image and annotated lung mask of a sample image from the JRST public dataset.

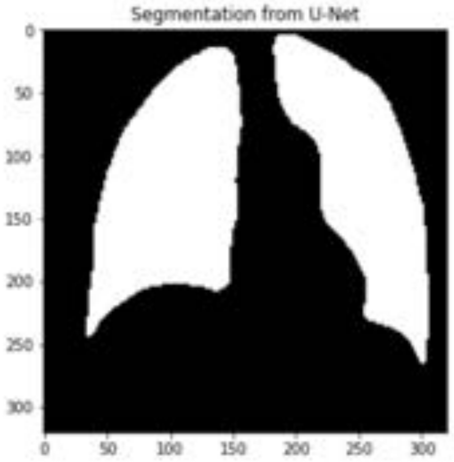


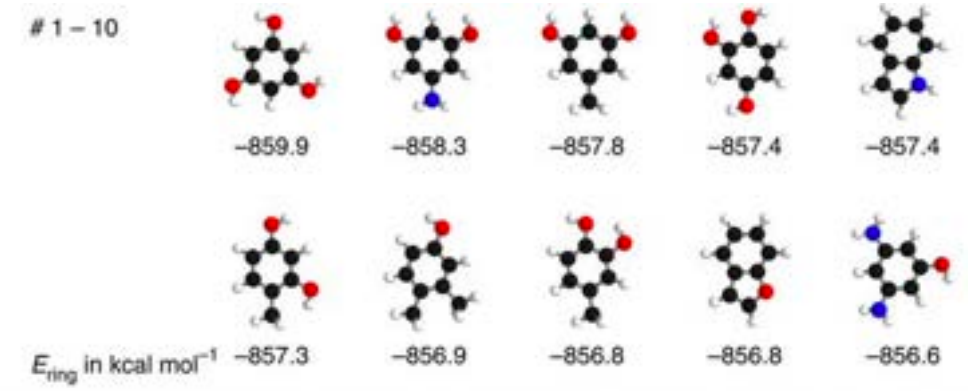
Figure 3 Lung segmentation as inferred by a U-Net trained with the JRST images.

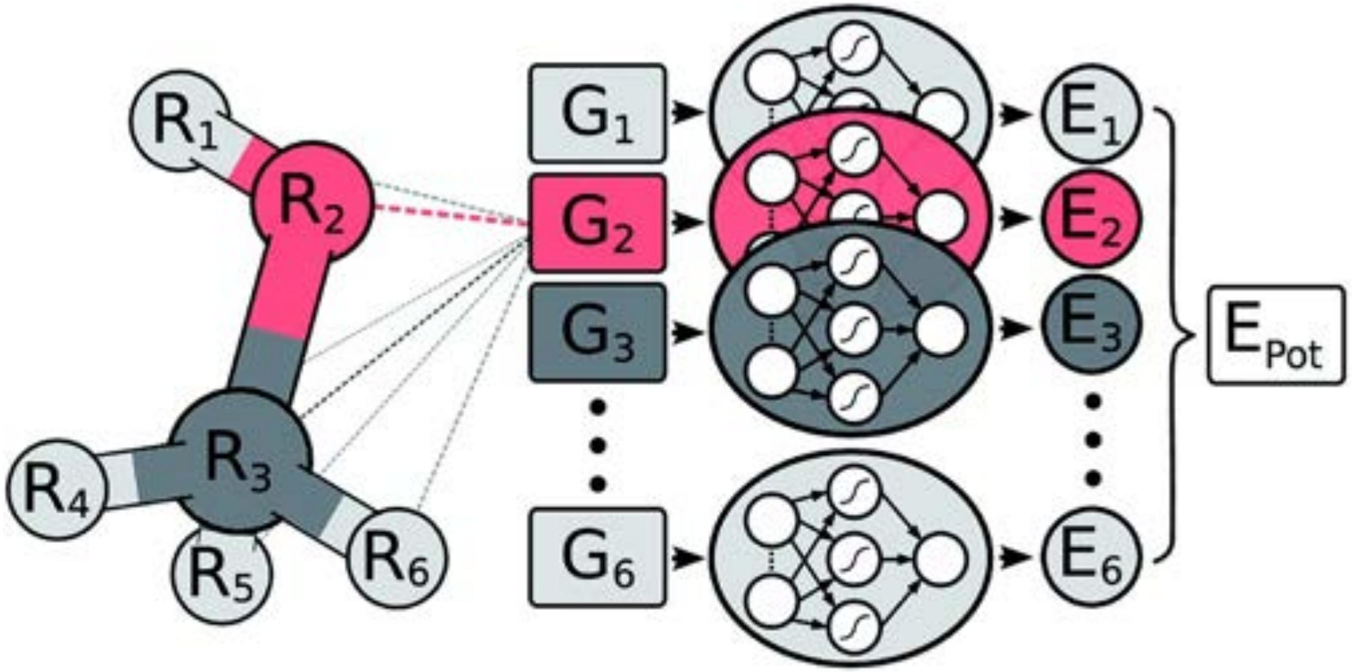
SDF CS-1

NEOCORTEX

Example Case: Behler-Parinello Neural Networks on Neocortex

Keith Phuthi (CMU), Matthew Guttenberg (CMU), Venkat Viswanathan (CMU)





Atomic Coordinates, forces and energies → Invariant Atomic Features → Atomic NNs with shared weights

Examples of isomers of $C_7O_2H_{10}$ and their molecular energies. Image from Schütt, K. T. *et al.* (2017) 'Quantum-chemical insights from deep tensor neural networks', *Nature Communications*, 8(1), p. 13890. doi: 10.1038/ncomms13890.
Gastegger, Michael, et al. "Machine Learning Molecular Dynamics for the Simulation of Infrared Spectra." Chemical Science, vol. 8, no. 10, 2017, pp. 6924–35. pubs.rsc.org, doi:10.1039/C7SC02267K.

Call for Proposals (CFP)

- All details to become fully available within a week in the Neocortex webpage. Stay tuned!
- Open to almost all U.S.-based university and non-profit researchers.
- Applications welcomed and processed through EasyChair.
- CFP open for a month.
- Applications will be evaluated as they come in. Apply as soon as convenient!
- Lightweight application via a short form.
- Follow-up meetings might be scheduled to confirm scope of the project and suitability.

Call for Proposals (CFP)

- Users expected to be onboarded by late November.
- Allocations to Neocortex resources and Bridges-2 will be initially granted for a year by default.
- Close collaboration and constant communication between domain projects, PSC, and vendors is expected.
- Feedback and user experiences are welcomed to further enrich the project.
- More technical details on the Cerebras servers, the ML frameworks, and applications supported, in the second part of the webinar to be presented by Dr. Natalia Vassilieva.

Thank you to all those contributing to *Neocortex*!

NSF

cerebras

Hewlett Packard Enterprise

NEOCORTEX

Unlocking Interactive AI for Rapidly Evolving Research

Neocortex Team

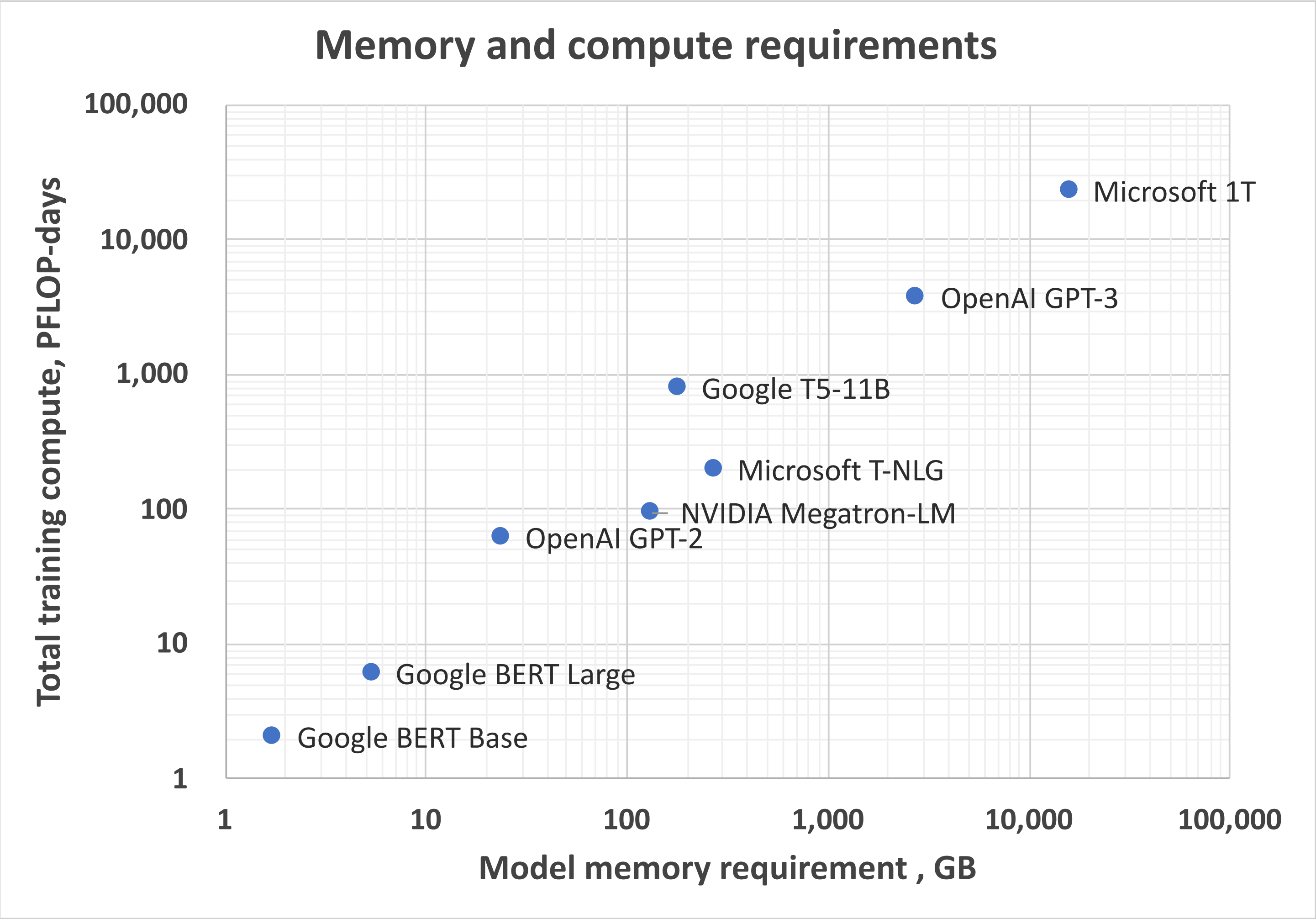
To Learn More and Participate

Watch the Neocortex website for updates!	https://www.cmu.edu/psc/aibd/neocortex/
Join the neocortex-updates list	https://www.cmu.edu/psc/aibd/neocortex/newsletter-sign-up.html
Apply to upcoming CFP	https://www.cmu.edu/psc/aibd/neocortex/
Stay tuned for an upcoming Cerebras technologies user group	https://www.cmu.edu/psc/aibd/neocortex/
Contact us with additional questions, input, or requests	neocortex@psc.edu

Cerebras CS-1: the AI Compute Engine for Neocortex

Technical Overview

Modern models need more compute than can fit on a single die

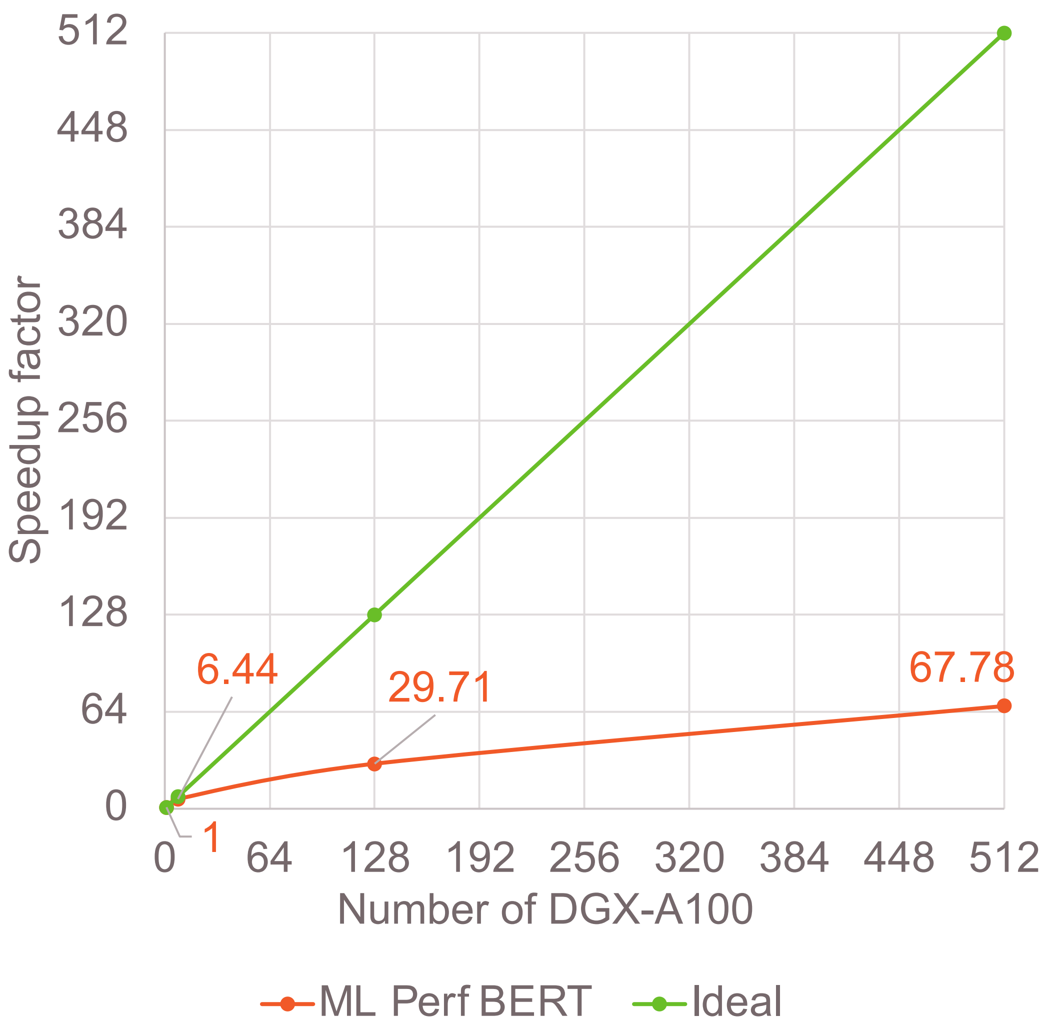


1 PFLOP-day is about
1 x DGX-2H or 1 x DGX-A100
busy for a day

Estimated time-to-train:

- OpenAI GPT-2:
  about 50 days
  on **1 DGX-A100** (8 A100)
- OpenAI GPT-3:
  about 20 years
  on **1 DGX-A100** (8 A100)

Distributed training is not the best option



MLPerf 1.0 results for BERT training

# DGX-A100	# A100 (80GB)	Batch per GPU	Total batch	Time to accuracy (min)	Speed-up
1	8	56	448	21.69	**1.0**
8	64	48	3072	3.37	6.4
128	1024	3	3072	0.73	29.7
512	**4096**	3	12288	0.32	**67.8**

We need **more compute per device,**
and ability to **rely less on data parallel training**

Cerebras Wafer Scale Engine

The Most Powerful Processor for AI

	WSE-1	**WSE-2**
AI-optimized cores	400,000	850,000
Memory on-chip	18 GB	40 GB
Memory bandwidth	9 PByte/s	20 PByte/s
Fabric bandwidth	100 Pbit/s	220 Pbit/s
Silicon area	46,225 mm^2	46,225 mm^2
Transistors	1.2 Trillion	2.6 Trillion
Fabrication process	16 nm	7 nm

Cluster-scale acceleration on a single chip

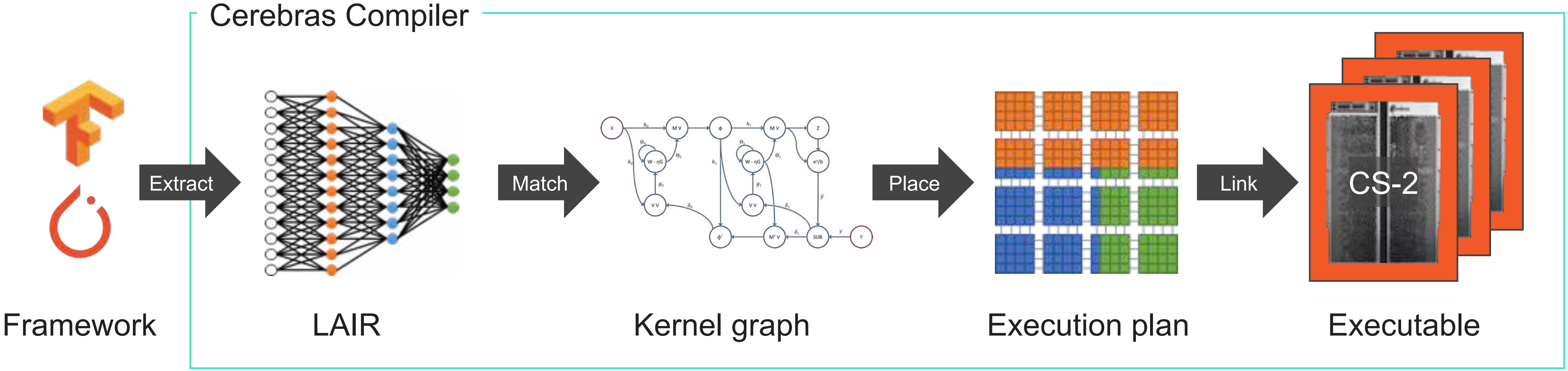
Cerebras CS-1 and CS-2: Cluster-scale Performance in a Single System

The world's most powerful AI computers

A **full solution** in a single system

- Powered by WSE
- Programmable via TF, other frameworks
- Install, deploy easily into a standard rack
- For datacenter or heavy edge deployment

The Cerebras Software Platform

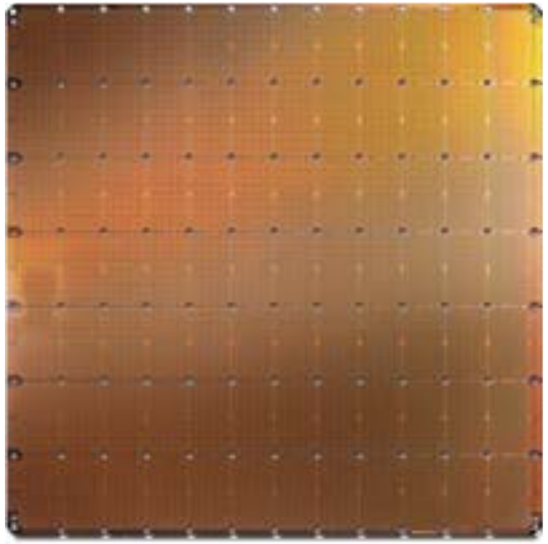


Cluster-scale performance with the programming ease of a single node

The Cerebras Solution

CS System

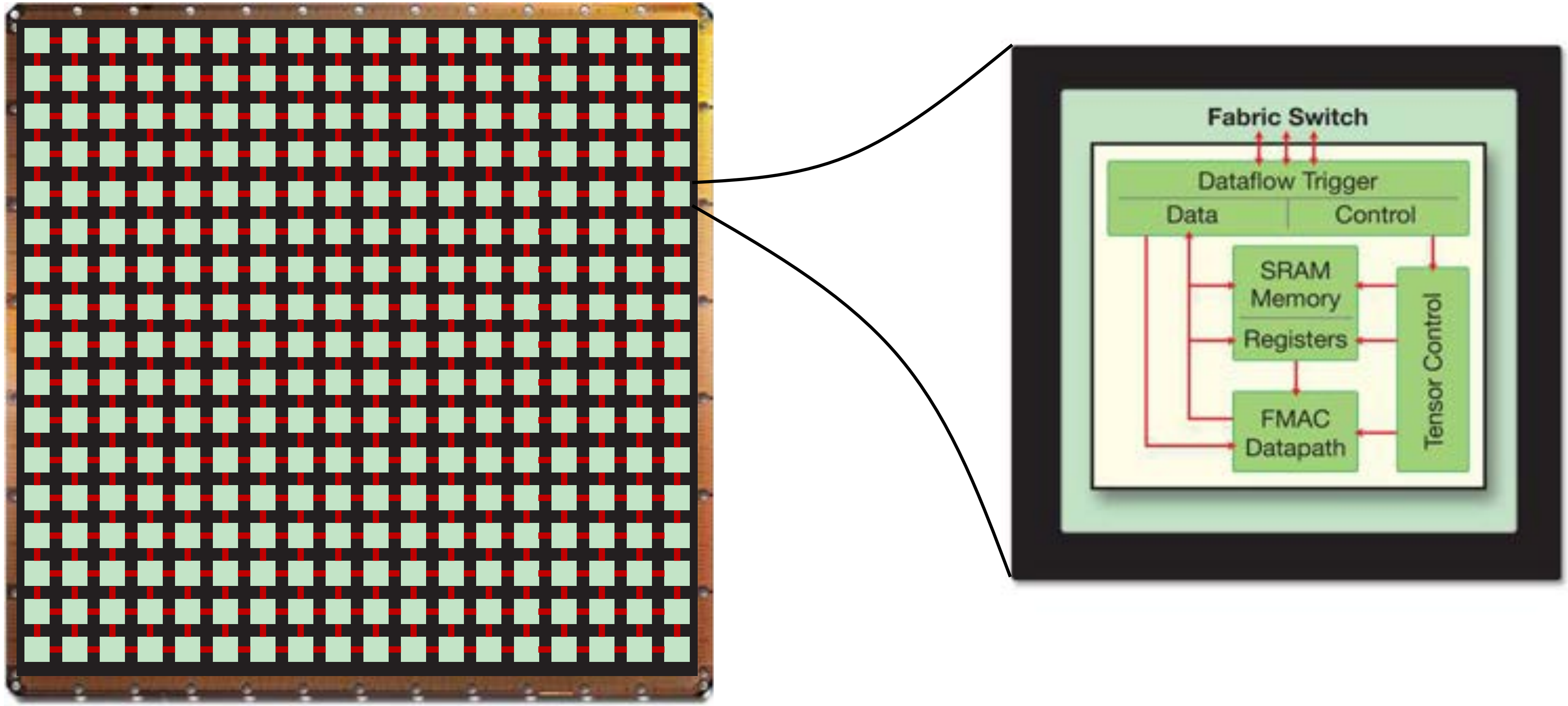
Wafer Scale Engine



Cerebras Software Platform

The Wafer-Scale Engine (WSE)

2D Mesh of 400,000 Fully Programmable Processing Elements



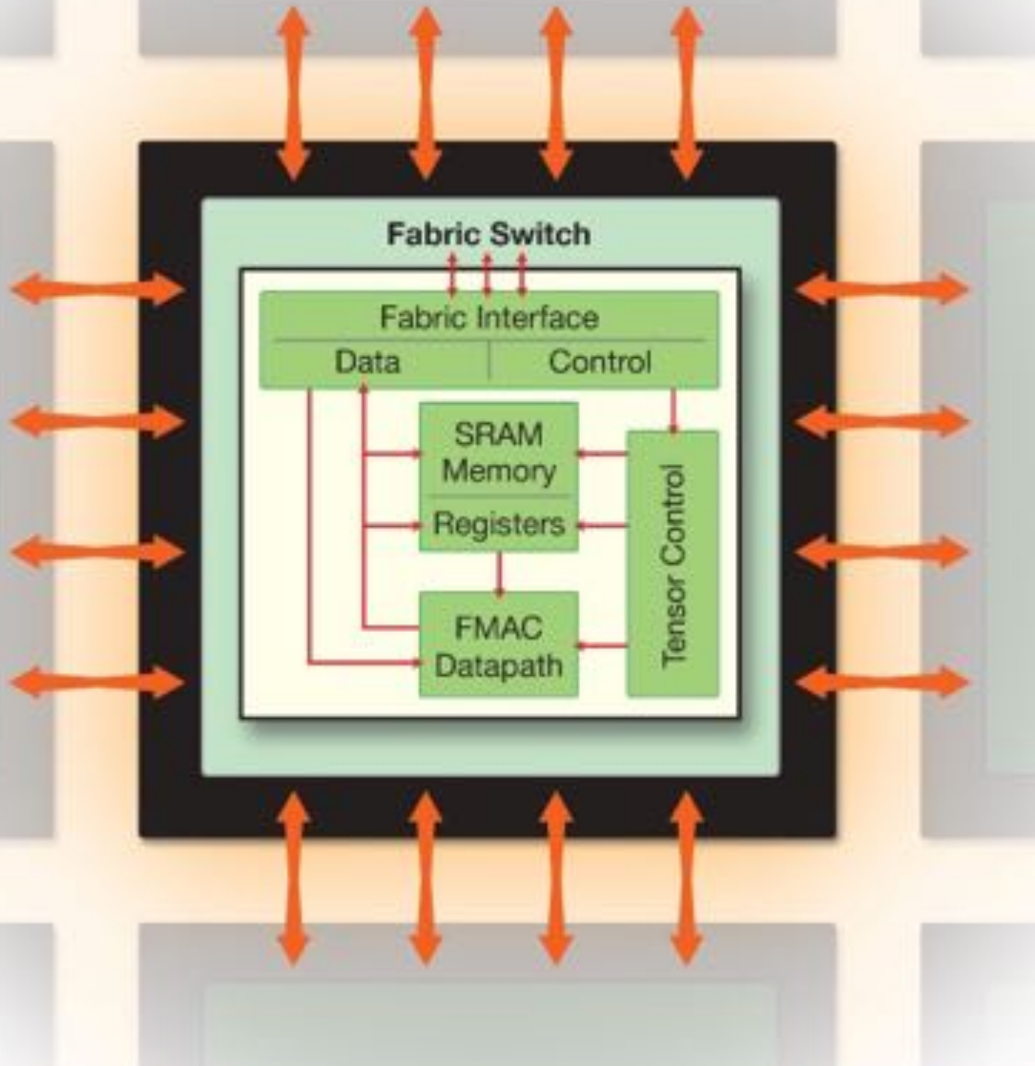
Designed for Deep Learning

Each component optimized for Deep Learning

Compute
- Fully-programmable core, ML-optimized extensions
- Dataflow architecture for sparse, dynamic workloads

Memory
- Distributed, high performance, on-chip memory

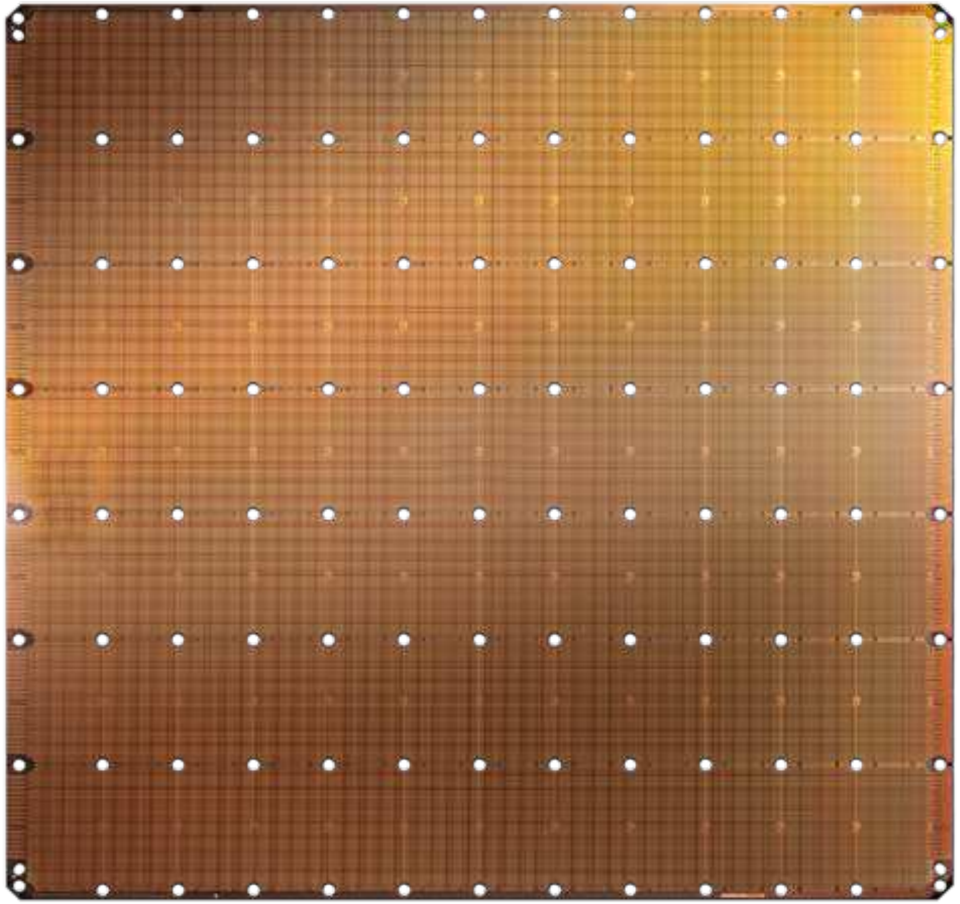
Communication
- High bandwidth, low latency fabric
- Cluster-scale networking on chip
- Fully-configurable to user-specified topology



Advantages of Wafer Scale

Wafer-scale enables:

- More AI optimized cores →
  Enormous compute on a single chip
- More high speed, on chip memory →
  No memory bottlenecks
- More fabric bandwidth at low latency →
  No communication bottlenecks



Cluster-scale acceleration on a single chip

Advantages of the WSE for DL and HPC

Full Performance on All BLAS Levels
Enabled by Massive Memory Bandwidth

Matrix-Matrix	Matrix-Vector	Vector-Vector	Vector-Vector
GEMM	**GEMV**	**DOT**	**AXPY**
$C \leftarrow \alpha AB + \beta C$	$y \leftarrow \alpha Ax + \beta y$	$\alpha \leftarrow (x, y)$	$y \leftarrow \alpha x + y$
C += A B	y += A x	α += y x	y += α x

Sparse GEMM is one AXPY per non-zero element

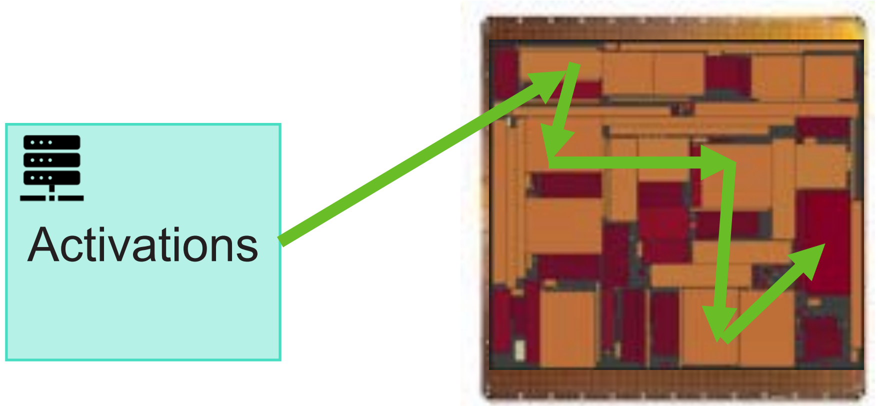
Software and Programming

How to program the CS-1

- Deep learning:
  - High-level programming via ML frameworks (TF, PyTorch), with Cerebras Graph Compiler
  - Ability to create custom kernels with Cerebras Kernel SDK
- Hybrid AI + HPC:
  - Today, a hybrid approach, an ML framework to define DL model, C++ interface to send inference requests to CS-1 directly from HPC codes
  - Tomorrow, ability to run hybrid workloads on the WSE
- HPC: C-level programming interface with Cerebras Kernel SDK

Two execution modes for Deep Learning

Pipelined

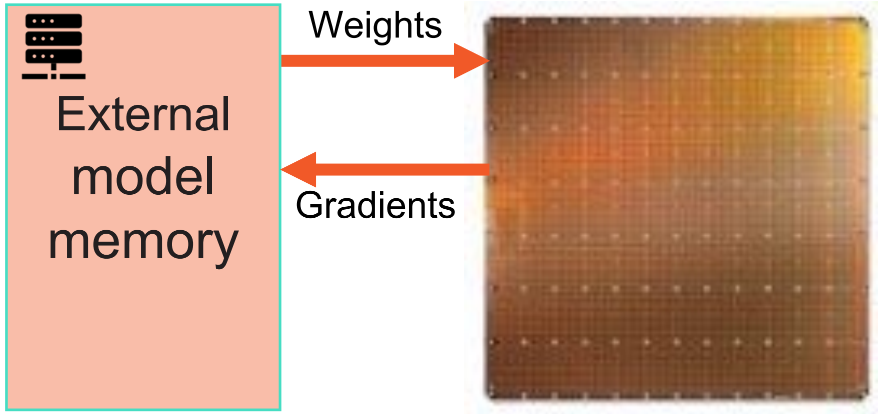


Weights on wafer

Arranged in **space**

Activations **stream**

Weight Streaming



Activations on wafer

Arranged in **time**

Weights **stream**

TensorFlow 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)
```

PyTorch example

```
import torch_xla.core.xla_model as xm
from cerebras.models.common.pytorch.PyTorchBaseModel import PyTorchBaseModel

import torch_xla.distributed.data_parallel as dp

class Model(PytorchBaseModel):
    def __init__(self):
        Pass
    def forward(self, x):
        Pass

device = xm.xla_device()

model = Model().to(device)
optimizer = optim.Adam (…)
```

Use XLA Device instead of GPU/CPU device

Cerebras SDK

A general-purpose parallel-computing platform and API allowing software developers to write custom programs for a Cerebras System. Consists of:

- Language
  - Device: Domain-specific language, based on C, uses familiar parallel programming concepts
  - Host: Python APIs
- Libraries
  - Communication primitives (scatter, gather, broadcast, allreduce, etc.,)
  - Neural network kernels (convolution, tanh, etc.,)
  - BLAS … and more to come
- Tools
  - Simulator
  - Debugger
  - Performance analysis
  - Visualization

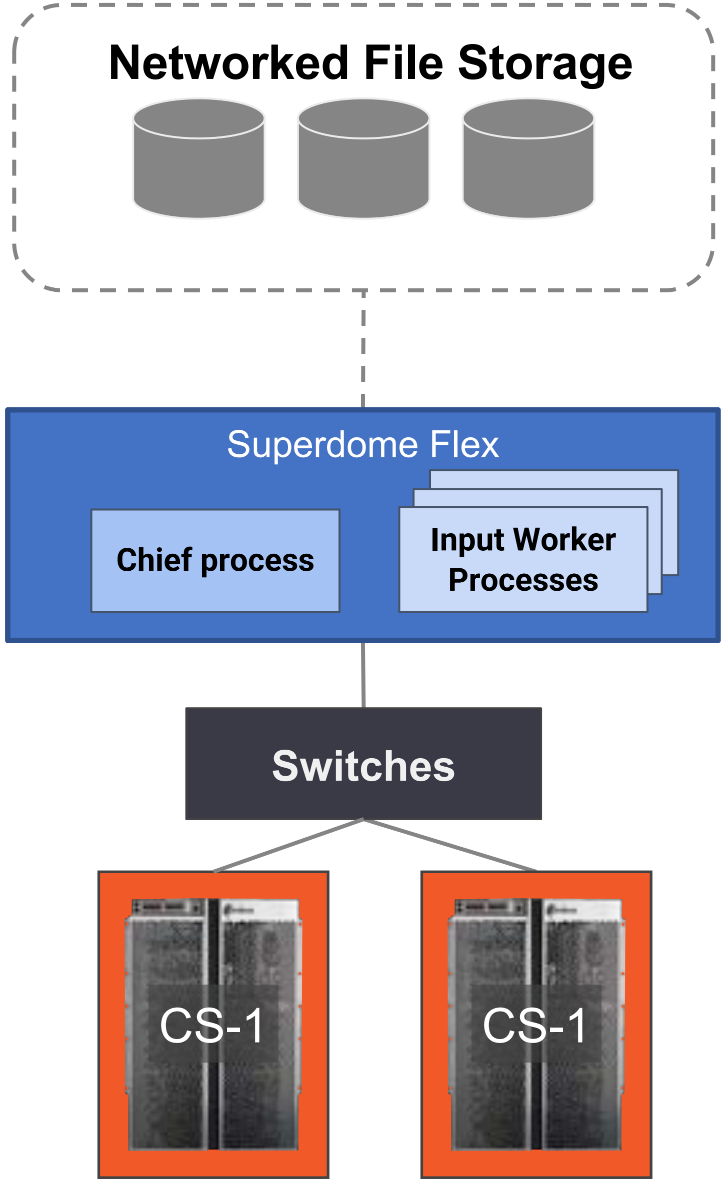
Beta Release in mid-October, 2021

Neocortex Execution mode

- CS-1 is a **network-attached** accelerator
- User launches job to orchestrator
- Spins up Cerebras SW container on standard CPUs
- Chief compiles network and manages CS-1
- Input workers pull input data from storage, run the input pipeline, and stream data to CS-1

No need to hyper-optimize input function

Just spin up more CPU input workers



Value to users

Training time reduced from weeks to hours, from days to seconds
→ 100s new hypotheses tested in the same time period

Enable orders of magnitude more data in a training sets
→ More data in less time improves results

High throughput inference at low latency
→ Employ larger models and datasets in production with higher throughput

Explore networks and methods not possible on GPUs
→ Larger deeper networks, extraordinarily sparse networks, very wide shallow networks, etc.

Focus areas for the upcoming CFP

Proposals in the following areas are encouraged:

- Domain-specific natural language processing via self-supervised pretraining of attention-based models
- Language modelling with wide and shallow LSTM models
- Sequence-to-sequence modelling with Transformers (e.g., machine translation)
- Self-supervised pre-training of protein embeddings with BERT-style models
- Self-supervised pre-training of attention-based DNA-language model
- Training and high-throughput inference with small multi-layer perceptrons (for e.g., virtual drug screening)

Additional materials

www.cerebras.net

Whitepapers, blog posts, customer stories