

# 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 PSC's YouTube channel and the Neocortex webpage the week of July 20, 2020.
2. There will be 40 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 Zoom chat.
  - Completing the Zoom polls that will appear during the webinar.These questions will seed the Q&A session in the final 20 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](mailto:akli@sura.org))
  - Lizanne Destefano, Georgia Tech ([lizanne.destefano@ceismc.gatech.edu](mailto:lizanne.destefano@ceismc.gatech.edu))
  - Ken Hackworth, Pittsburgh Supercomputing Center ([hackworth@psc.edu](mailto:hackworth@psc.edu))
  - Bryan Snead, Texas Advanced Computing Center ([jbsnead@tacc.utexas.edu](mailto:jbsnead@tacc.utexas.edu))
- Anonymous reporting form available at <https://www.xsede.org/codeofconduct>.

# Introduction to *Neocortex*

*July 15, 2020*

Paola Buitrago  
PSC

Nick Nystrom  
PSC



**NEOCORTEX**  
*Unlocking Interactive AI for  
Rapidly Evolving Research*



NSF Award 2005597

# Outline

- Introduction
- Neocortex for Research and Education
  - Target Applications
  - System Architecture
- Early User Program
- Summary
- Q&A

# Outline

- Introduction
- Neocortex for Research and Education
  - Target Applications
  - System Architecture
- Early User Program
- Summary
- Q&A



## 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 in production operations 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.

# 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 are delivering *Neocortex*

All trademarks, service marks, trade names, trade dress, product names, and logos appearing herein are the property of their respective owners.

# 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* in 2020 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.



# Motivation

*“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.”*

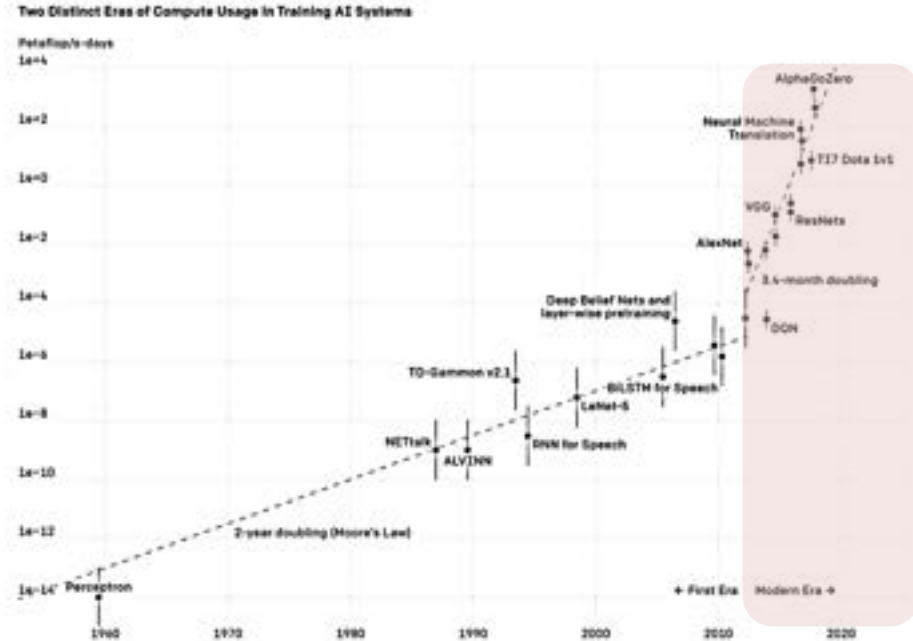


Figure from D. Amodei, D. Hernandez, G. Sastryjack, C. Greg, and B. Sutskever. (2019, November 7). *AI and Compute*, OpenAI Blog. <https://openai.com/blog/ai-and-compute>.

# Motivation

## Convolutional Neural Networks (CNNs)

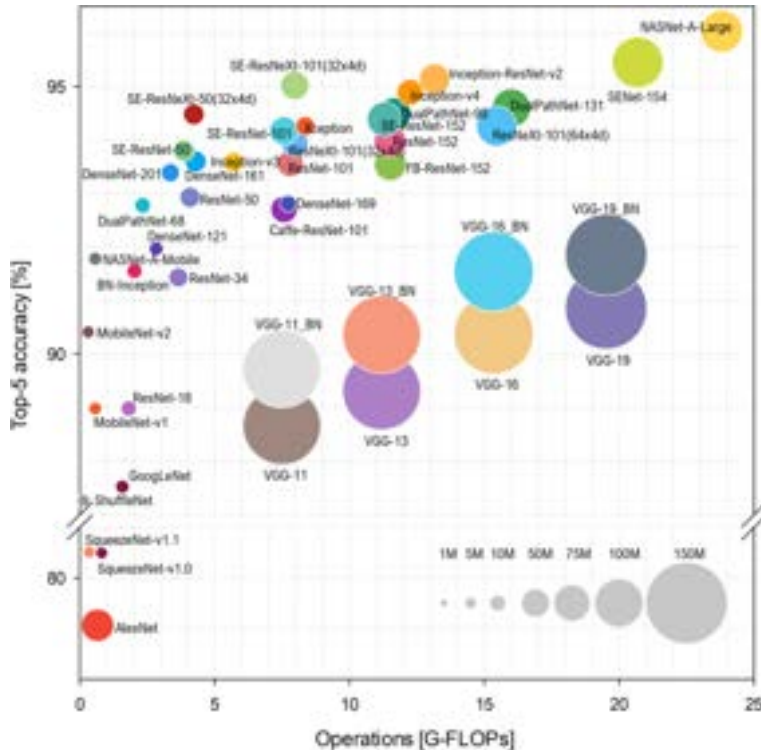


Figure from S. Bianco, R. Cadene, L. Celona, and P. Napolitano, *Benchmark Analysis of Representative Deep Neural Network Architectures*, IEEE Access, vol. 6, pp. 64270–64277, 2018. arXiv:1810.00736v2.

## Some Recent Transformer-type Networks

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

## Sources of Additional Complexity

Generative Adversarial Networks (GANs)

Domain Adaptation

Reinforcement Learning (RL)

# Motivation - ML Workflows

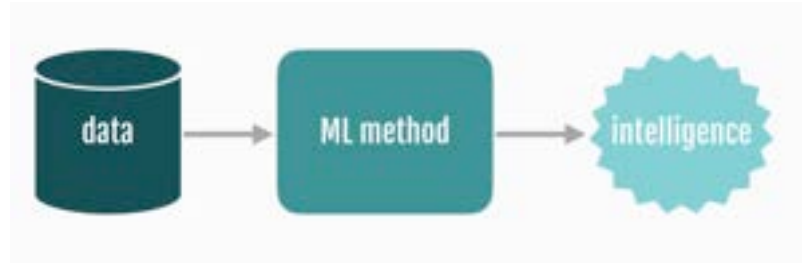


Figure from H. Miller, and V. Smith, Deep Learning, *Machine Learning with Large Datasets*, CMU, 2019. Retrieved on May 12, 2019 from <https://10605.github.io/>

# Motivation - ML Workflows

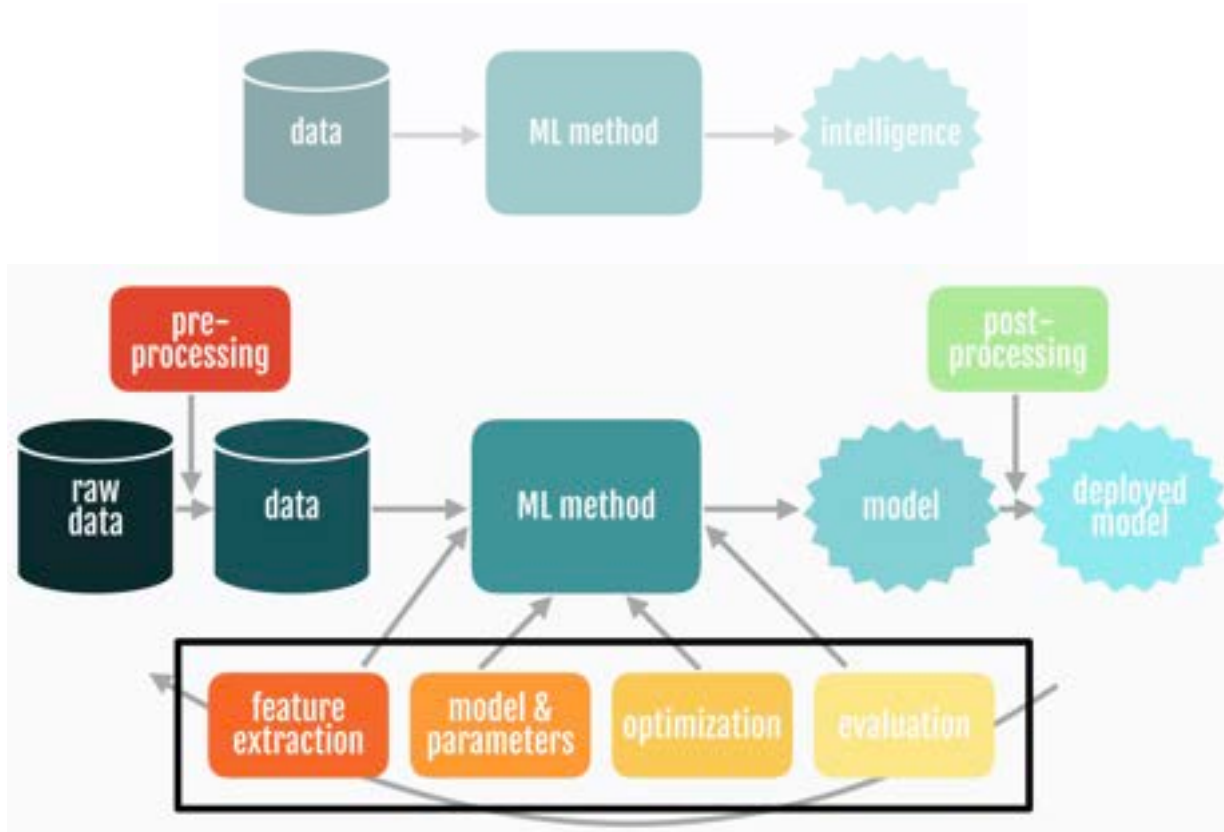


Figure from H. Miller, and V. Smith, Deep Learning, *Machine Learning with Large Datasets*, CMU, 2019. Retrieved on May 12, 2019 from <https://10605.github.io/>

***“Hardware capabilities and software tools both motivate and limit the type of ideas that AI researchers will imagine and will allow themselves to pursue. The tools at our disposal fashion our thoughts more than we care to admit.” (LeCun, 2019)***

*“Is DL-specific hardware really necessary? The answer is a resounding yes. One interesting property of DL systems is that the larger we make them, the better they seem to work. While this property is true for networks trained with supervised learning, the trend is to rely increasingly on unsupervised, self-supervised, weakly supervised or multi-task learning, for which **larger networks perform even better**. The demands on DL-specific hardware will undoubtedly increase.” (LeCun, 2019)*

LeCun, Y. (2019). 1.1 Deep Learning Hardware: Past, Present, and Future. 2019 IEEE International Solid-State Circuits Conference - (ISSCC), 12-19.

# Outline



- Introduction
- Neocortex for Research and Education
  - Target Applications
  - System Architecture
- Early User Program
- Summary
- Q&A

# Who is this designed for?



- People doing demanding deep learning training.
- Users with complementary projects on Bridges-2.
- Examples:
  - Large sets of medical images.
  - Challenging training for astrophysics, weather, genomics, and other sciences.
  - Simulation runs (on Bridges-2) + surrogate model training on Neocortex.
- Users exploring models that fit the following:
  - Models with separable convolutions.
  - Models with induced sparsity.
  - Graph neural networks.
  - Models with sparse attention.
  - Sequential models.
  - Model that would benefit from model parallelism.

# Introducing *Neocortex*



An innovative computing resource that will accelerate scientific discovery by:

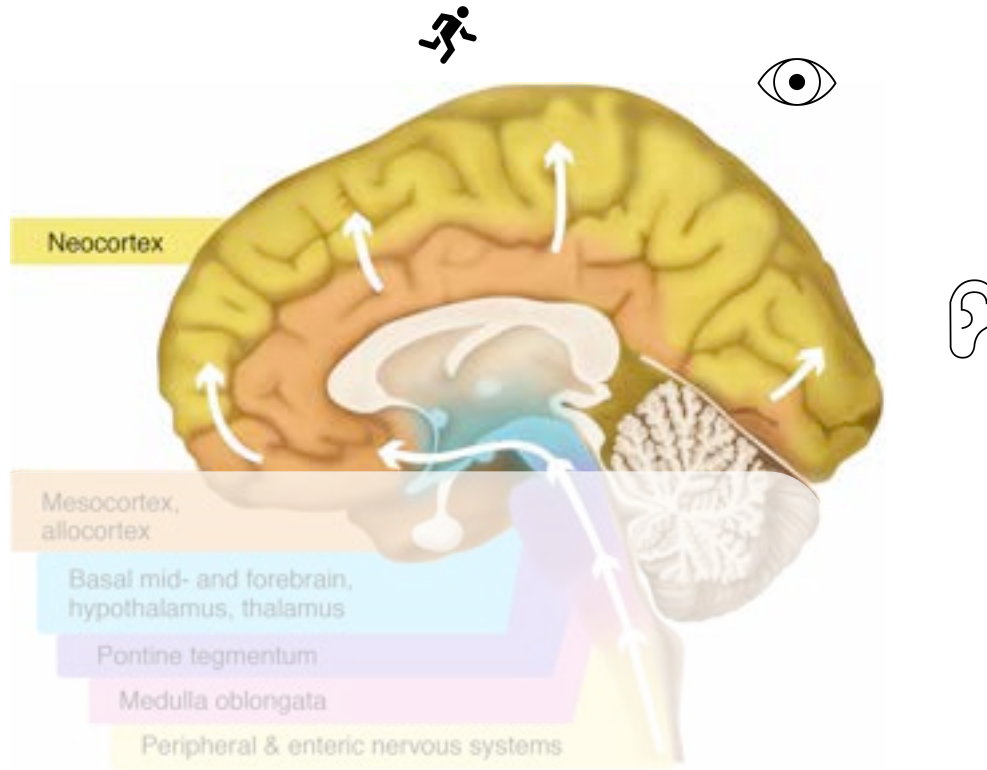
1. Vastly shortening the time required for DL training
2. Foster integration of artificial deep learning with scientific workflows
3. Provide revolutionary new hardware for the development of more efficient algorithms for artificial intelligence and graph analytics

Offered at **no cost** to the national open-science community.

Potential users that do not follow under this class can still get access under different terms (contact us for more info).



# Introducing *Neocortex*

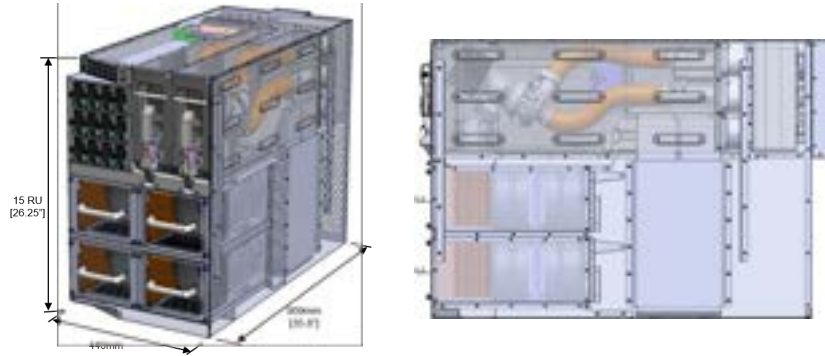


# Neocortex: Resource Specification



- Primary Compute System:
  - 2 Cerebras CS-1 servers
  - HPE Superdome Flex
  - Federated with Bridges-2
- Interconnect:
  - Superdome Flex to each CS-1: 12 100Gb/s ethernet links
  - Superdome Flex to Bridges-2: 8 HDR-200 links
- Storage
  - 205 TB of NVMe SSD
  - 24 TB RAM
  - Bridges-2 15PB Lustre filesystem and 8+PB tape archive, managed by DMF

# The CS1 server



**Interior view of the  
Cerebras CS-1**

## Wafer Scale Engine (WSE) Processor



# Early CS-1 Activities



Two of the CS-1 systems running in the Cerebras Systems lab in Los Gatos, California.

PRESS RELEASE | ARGONNE NATIONAL LABORATORY

## Argonne National Laboratory Deploys Cerebras CS-1, the World's Fastest Artificial Intelligence Computer

NOVEMBER 19, 2019

News

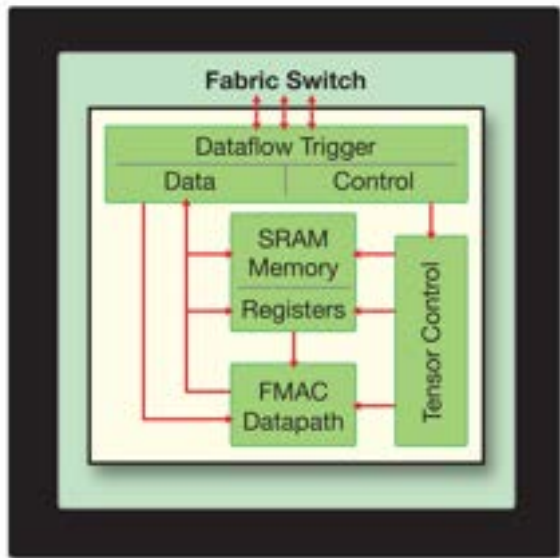
# Cerebras CS-1 – The WSE

Powered by the Cerebras Wafer Scale Engine (WSE):

- Largest chip ever built: 46,225 mm<sup>2</sup> silicon,  
1.2 trillion transistors
- 400,000 AI optimized cores
- 18 GB on chip memory—all 1 clock cycle from the cores
- 9 PByte/s 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**



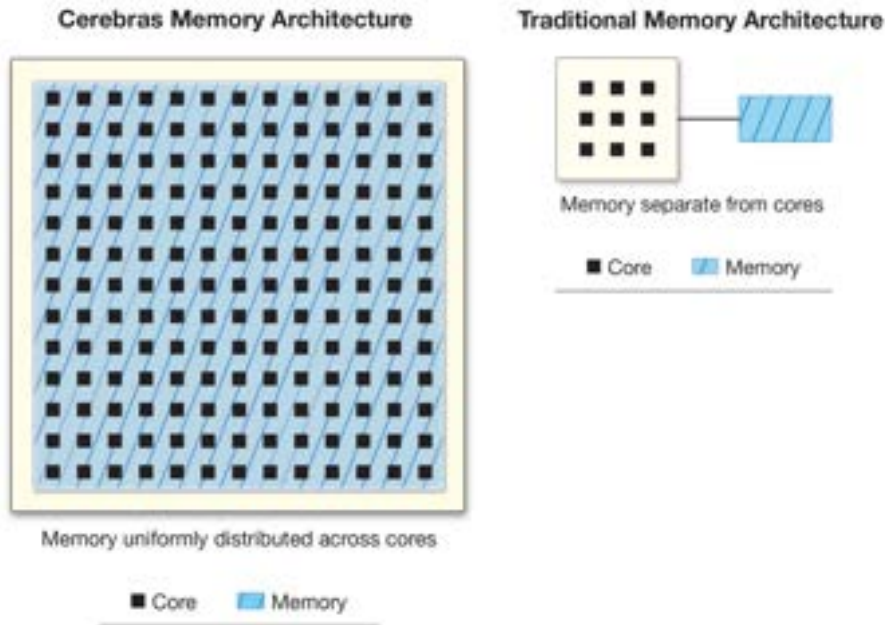
Fully programmable Sparse Linear Algebra cores optimized for tensor operations

- Full array of general instructions with ML extensions
- Flexible general ops for control processing (e.g. arithmetic, logical, load/store, branch)
- Optimized tensor ops for data processing, tensors as first-class operands
- $\text{fmac} [z] = [z], [w], a$ 
  - 3D 3D 2D scalar
- Data flow architecture, sparsity “harvesting”

# WSE - Memory

Distributed, high performance, on-chip memory

- All memory local - 1 clock from core
- Small batches including batch size 1 at full utilization
- Scale without big batches; without developing big batch learning rate schedules, etc.



# WSE - Fabric

To connect cluster-scale compute in a single system

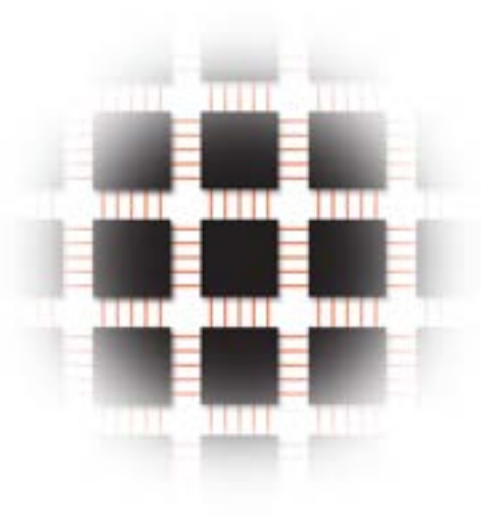
**System advantage:** no communication bottlenecks.

→ Model-parallel training is easy.

**Usability advantage:** no orchestration / sync headaches.

**ML advantage:** train with small batches at high utilization.

→ No need for tricky learning schedules and optimizers.

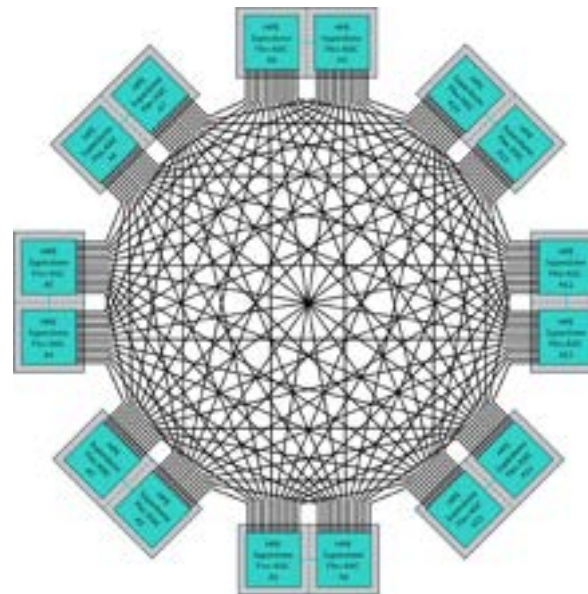




# The HPE Superdome Flex



HPE Superdome Flex HPC Server



Superdome crossbar topology – 850  
GB/s of bisection bandwidth

# The HPE Superdome Flex



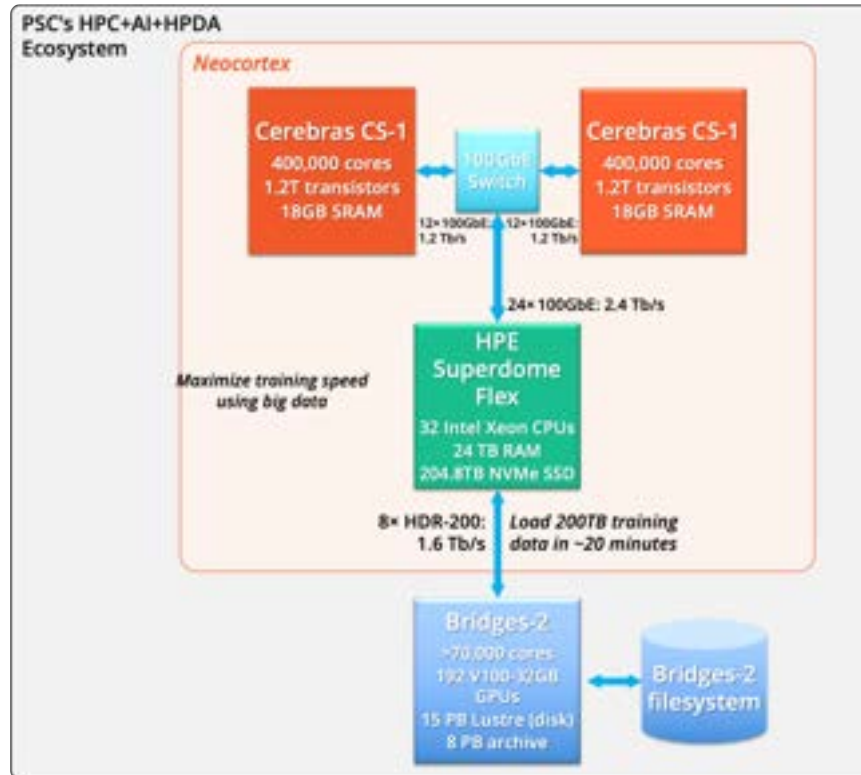
The HPE Superdome Flex will:

- Provide substantial capability for preprocessing and other complementary aspects of AI workflows.
- Enable training on very large datasets with exceptional ease.
- Support both CS-1s independently and together to explore scaling.

## Superdome Flex System Specifications

<b>Processors</b>	Intel Xeon, TBA
<b>Memory</b>	24 TiB RAM, aggregate memory bandwidth of 4.5 TB/s
<b>Local Disk</b>	32 x 6.4 TB NVMe SSDs <ul style="list-style-type: none"><li>◦ 204.6 TB aggregate</li><li>◦ 150 GB/s read bandwidth</li></ul>
<b>Network to CS-1 systems</b>	24 x 100 GbE interfaces <ul style="list-style-type: none"><li>◦ 1.2 Tb/s (150 GB/s) to each Cerebras CS-1 system</li><li>◦ 2.4 Tb/s aggregate</li></ul>
<b>Interconnect to Bridges-2</b>	16 Mellanox HDR-100 InfiniBand adapters <ul style="list-style-type: none"><li>◦ 1.6 Tb/s aggregate</li></ul>
<b>OS</b>	Red Hat Enterprise Linux

# Neocortex: System Overview



# Software Stack



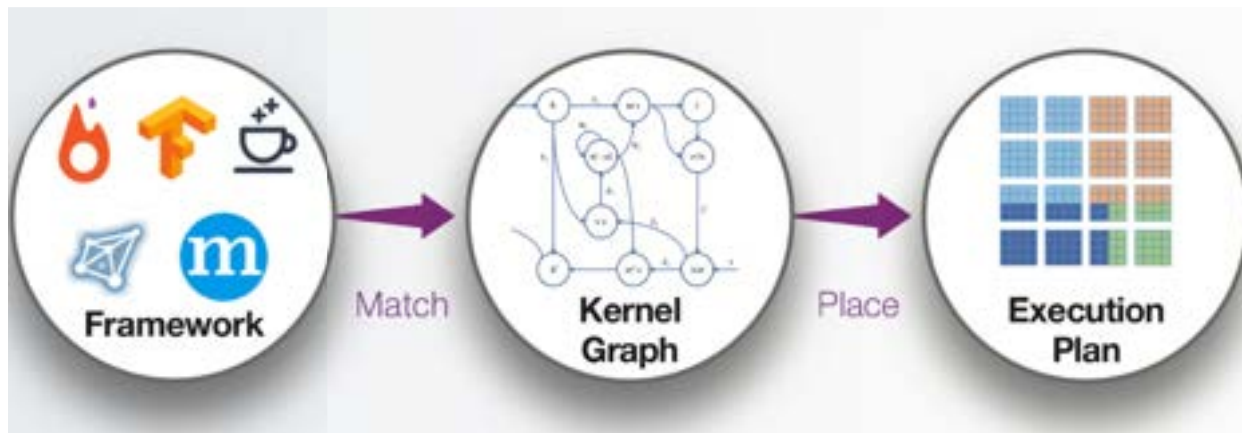
The user starts as usual by developing their ML model in existing opensource frameworks.

Cerebras integrates with popular ML Frameworks so researchers can write their models using familiar tools.

+ LAIR: A low-level programmable C++ interface.

# Cerebras Graph Compiler programs CS-1

- Model is specified through a high-level python framework
- Stretchable kernels are mathematical operators that run on any number of cores
- Kernels are placed and routed on the entire processor



# Outline

- Introduction
- Neocortex for Research and Education
  - Target Applications
  - System Architecture
- **Early User Program**
- Summary
- Q&A

# Early User Program (EUP)



- Achieve scientific progress early and at no charge
- Preliminary user guide, frontline support and advanced support from Day 0
- For news updates, including on the opportunity to apply for access, please review the project webpage:  
<https://www.cmu.edu/psc/aibd/neocortex/>
- EUP to take place in the fall.
- If you participate, you will be asked to provide feedback after 4 weeks of the EUP.
- *Ideally, EUP activities will result in scientific progress.*

# Early User Program (EUP)

- Some example that characterize the kind of applications we would like to include in the EUP
  1. Attention-based encoder-only or decoder-only models, such as BERT- or GPT-type deep neural networks;
  2. LSTM-based RNN models;
  3. Kipf-Welling-style Graph Convolution Neural (GCN) networks for graph-level predictions.

Please, get in contact if you consider you are interested in an application that would benefit from Neocortex and would like to join our EUP.



# Neocortex Target 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
- July 2020** Accept applications for Early User Program
- Summer 2020** Delivery, installation, initial testing
- Fall 2020** Early User Program, conclusion of Acceptance Testing

# To Learn More and Participate



Join our coming webinars

<https://www.cmu.edu/psc/aibd/neocortex/>

Catch our coming PEARC20 plenary (July 29)

<https://pearc.acm.org/pearc20/program/schedule/>

Join the Early User Program (more info coming)

<https://www.cmu.edu/psc/aibd/neocortex/>

Watch the Neocortex website for updates!

<https://www.cmu.edu/psc/aibd/neocortex/>

Contact us with additional questions, input, or requests

[neocortex@psc.edu](mailto:neocortex@psc.edu)

# Summary

- *Neocortex* is an upcoming NSF-funded innovative advance computing system that will be made available in the fall 2020 by PSC.
- *Neocortex* captures promising AI hardware technology (Cerebras CS-1) that is bound to transform AI-enabled research and development of new AI algorithms.
- *Neocortex* will be available, at no cost, to a group of early users starting on the fall of 2020.
- This new system will be integrated with upcoming Bridges-2 and will feature two Cerebras CS-1s and a 24 TB RAM HPE Superdome Flex.
- One of the main goals of *Neocortex* is to engage, inspire and enable a strong community around the new technologies. We will focus strongly on outreach, training, and user support.
- Please, get in contact if you consider you have an application that would benefit from *Neocortex* and would like to join our EUP.

# Thank you to all those contributing to Neocortex!



Andrew K. Adams  
Paola Buitrago  
Ken Hackworth  
Ed Hanna  
Dave Moses  
Nick Nystrom

Rajanie Prabha  
Sergiu Sanielevici  
Amanda Slimick  
Julian Uran  
John Urbanic  
Bryan Webb



**NEOCORTEX**  
*Unlocking Interactive AI for  
Rapidly Evolving Research*

# Outline

- Introduction
- Neocortex for Research and Education
- Early User Program
- Summary
- Q&A