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.
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.

PSC

NEOCORTEX

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 Summer 2022 Call for Proposals

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

June 1, 2022

Overview

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

The Neocortex System

NEOCORTEX

Unlocking Interactive AI for Rapidly Evolving Research

Cerebras and HPE are delivering *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.**

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

Hewlett Packard
Enterprise

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 2021
Feb 2022	Neocortex CS-2 upgrade
June 2022	Summer 2022 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."

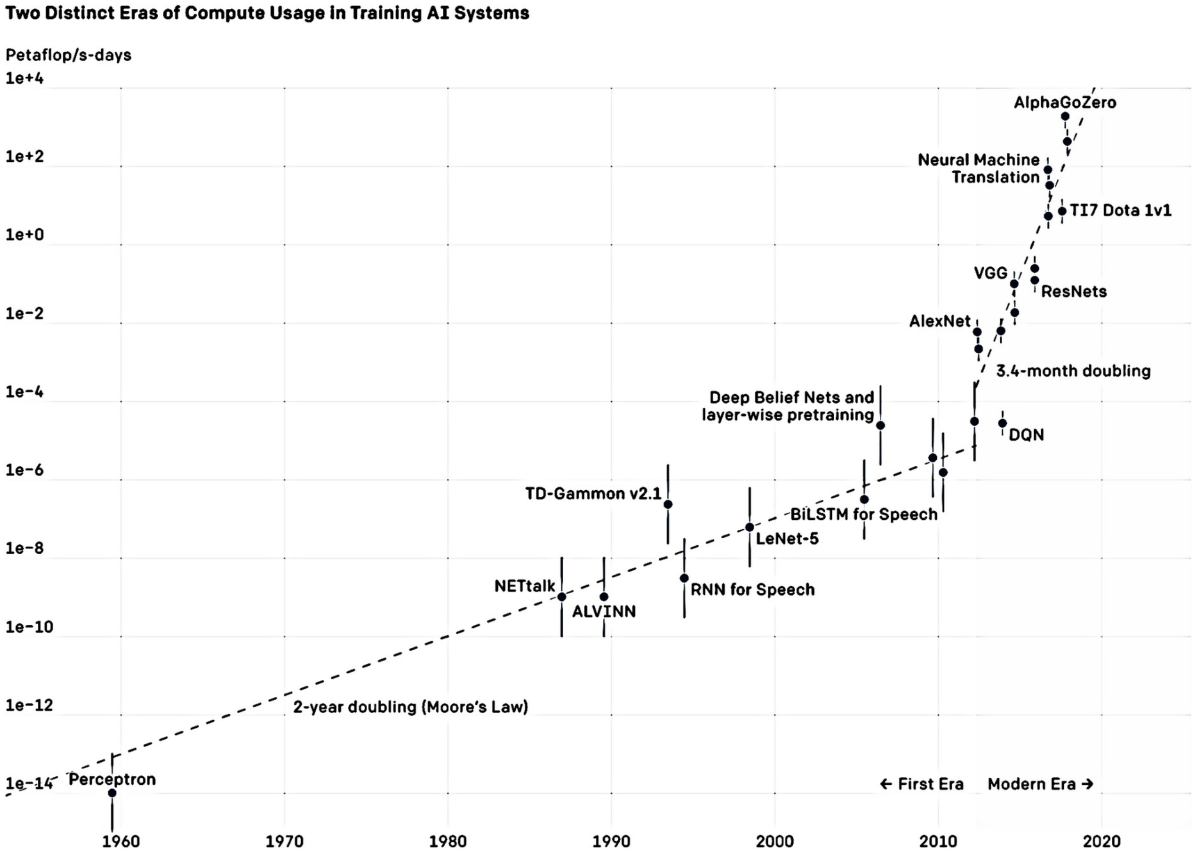
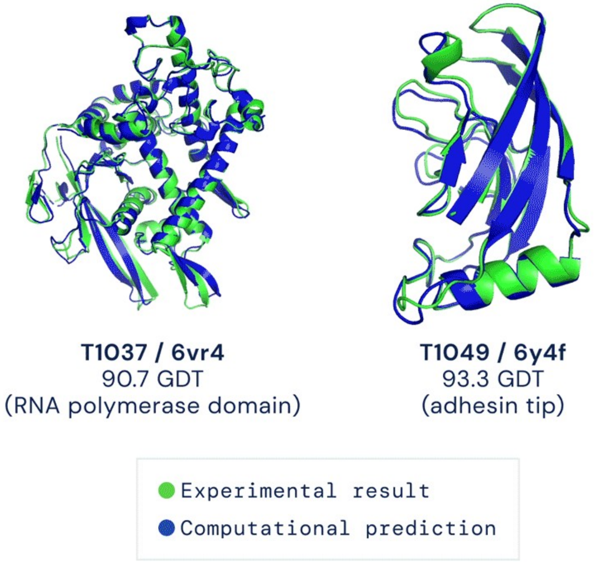


Figure from D. Amodei, D. Hernandez, G. SastryJack, C. Greg, and B. Sutskever. (2019, November 7). *AI and Compute*, OpenAI Blog. https://openai.com/blog/ai-and-compute.

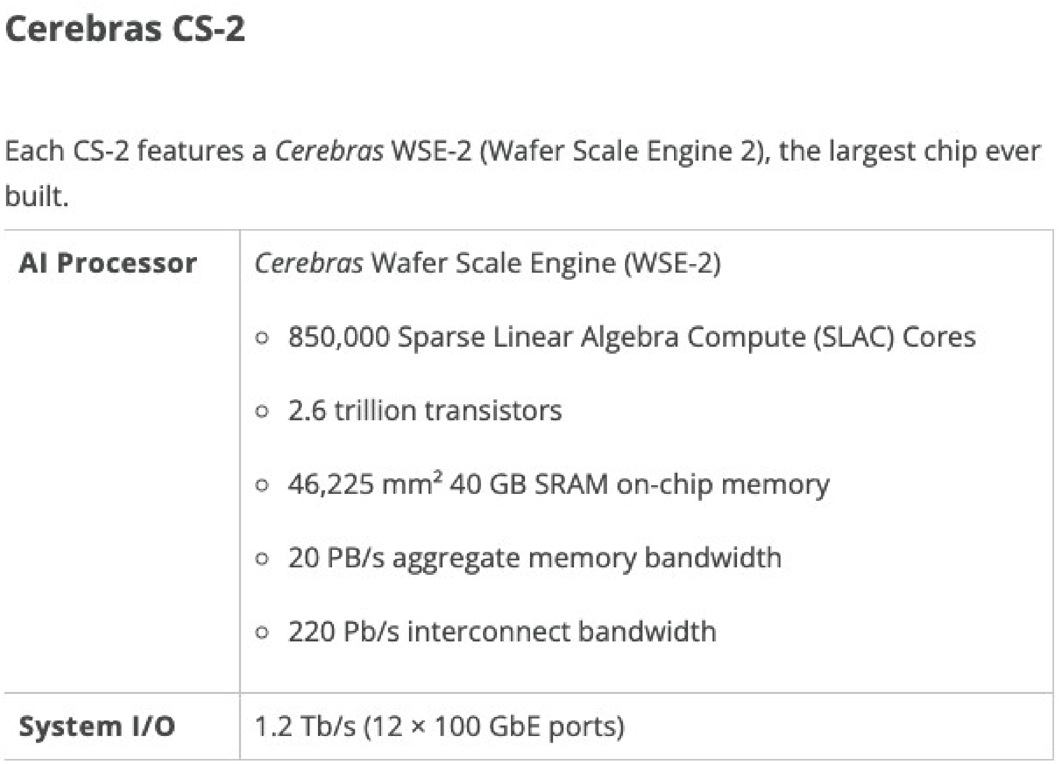
Driving Use-Cases

- 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.

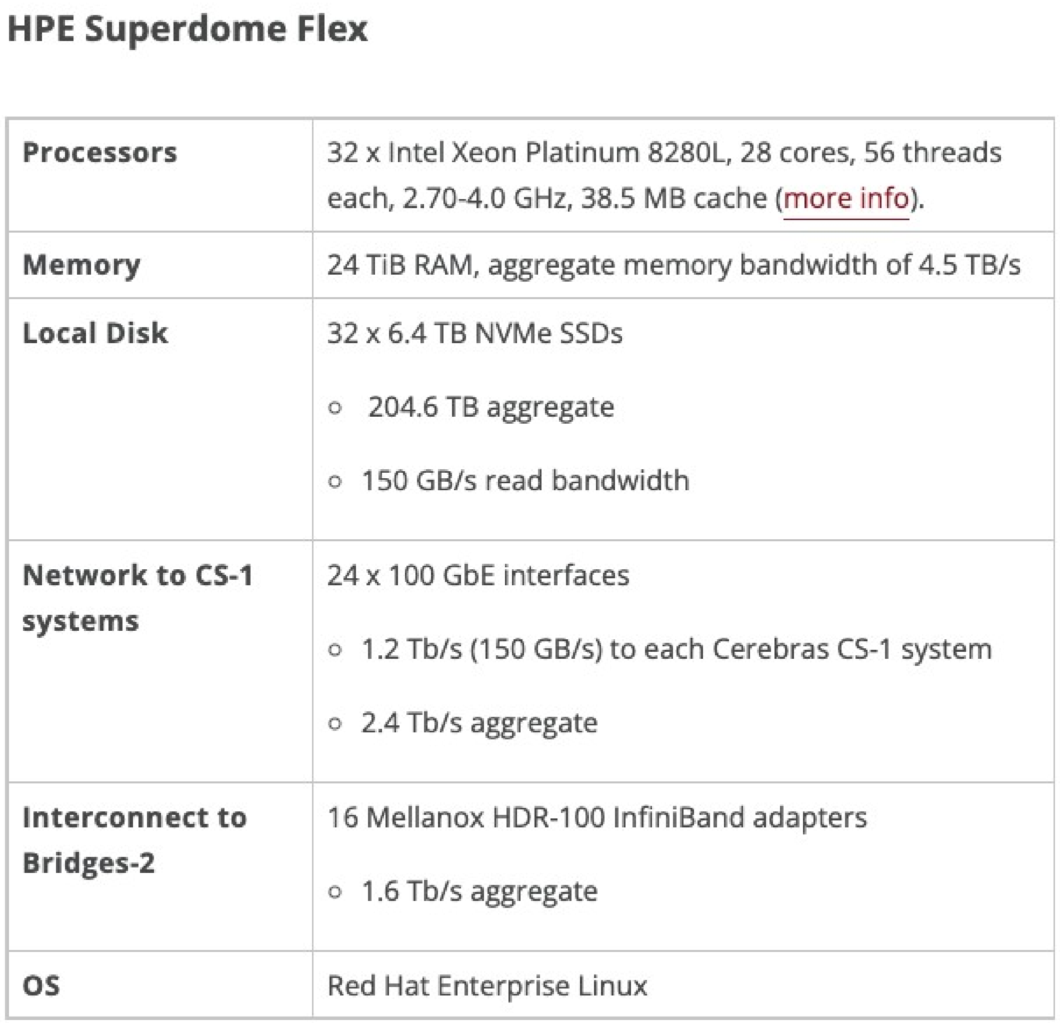
Neocortex Hardware Description



Cerebras CS-2

Each CS-2 features a *Cerebras* WSE-2 (Wafer Scale Engine 2), the largest chip ever built.

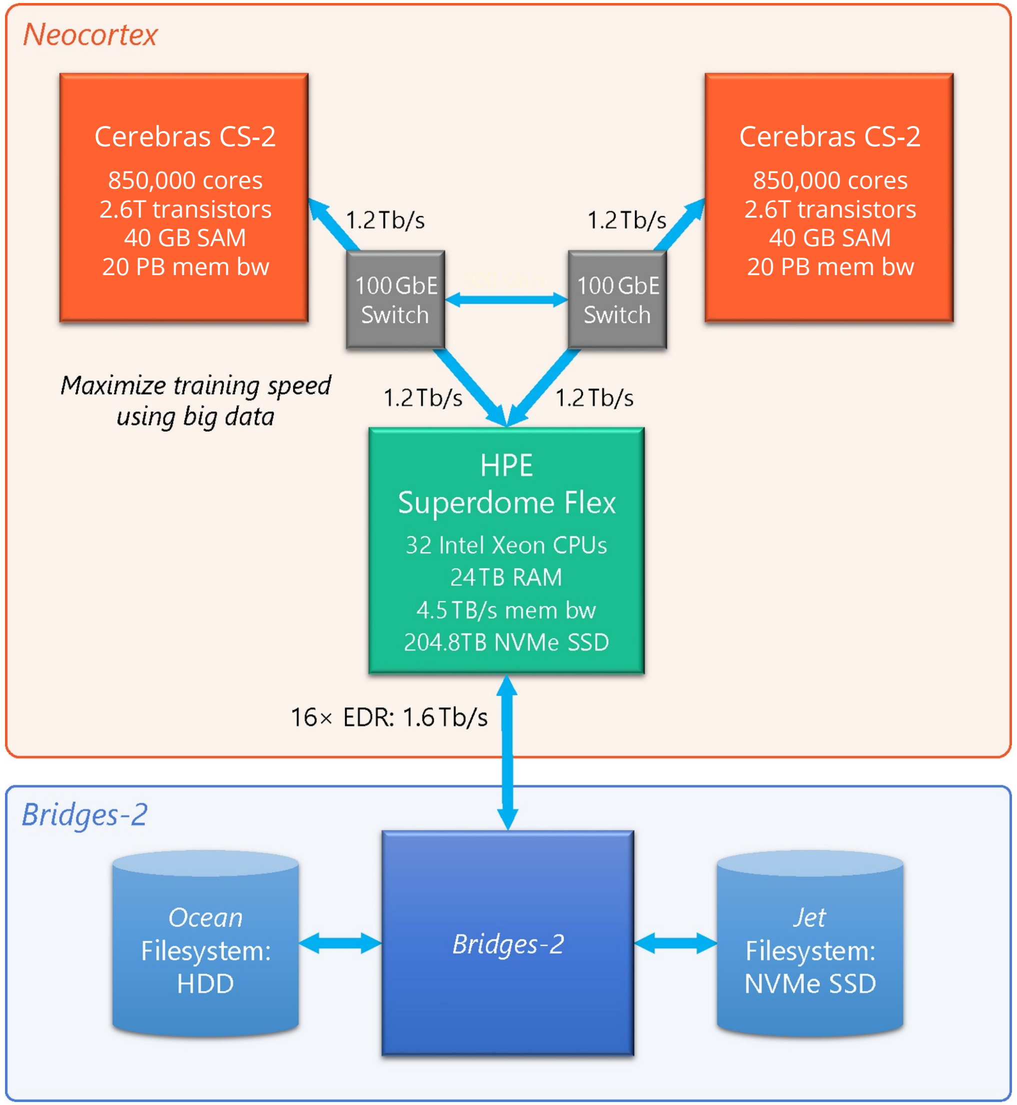
AI Processor	*Cerebras* Wafer Scale Engine (WSE-2) ○ 850,000 Sparse Linear Algebra Compute (SLAC) Cores ○ 2.6 trillion transistors ○ 46,225 mm^2 40 GB SRAM on-chip memory ○ 20 PB/s aggregate memory bandwidth ○ 220 Pb/s interconnect bandwidth
System I/O	1.2 Tb/s (12 × 100 GbE ports)



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 System Overview



Maximize training speed using big data

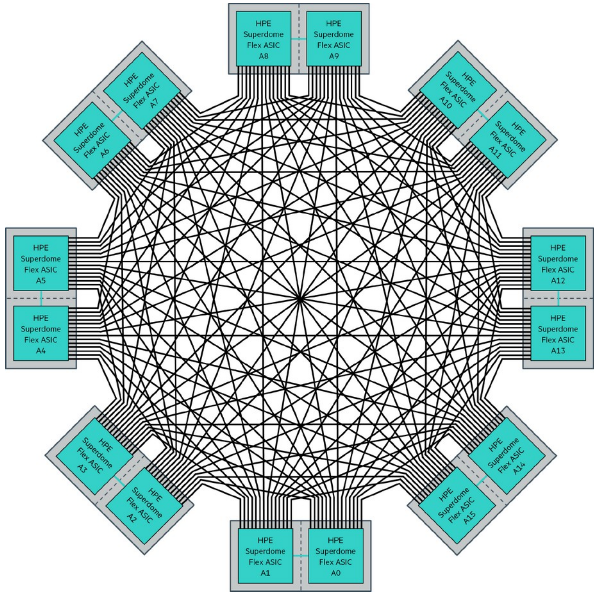
NEOCORTEX

The HPE Superdome Flex

HPE Superdome Flex HPC Server

The HPE Superdome Flex:

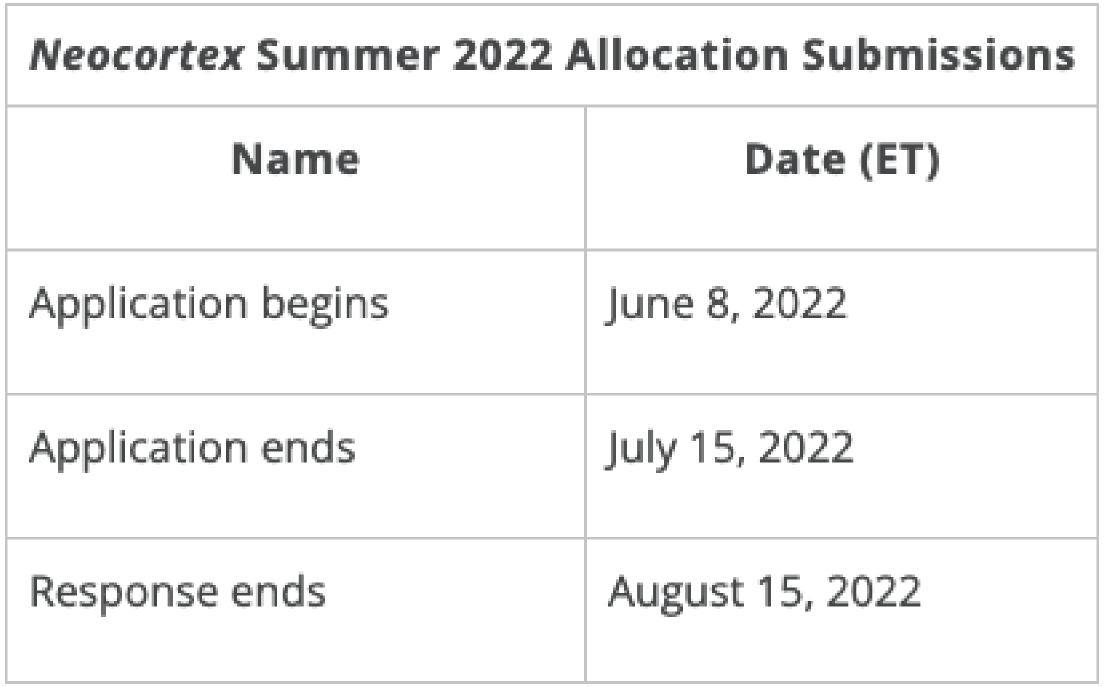
- Provides substantial capability for preprocessing and other complementary aspects of AI workflows.
- Enables training on very large datasets with exceptional ease.
- Supports both CS-1s independently and (will support them) together to explore scaling.


HPE Superdome Flex ASIC A8
HPE Superdome Flex ASIC A9
HPE Superdome Flex ASIC A7
HPE Superdome Flex ASIC A6
HPE Superdome Flex ASIC A10
HPE Superdome Flex ASIC A11
HPE Superdome Flex ASIC A5
HPE Superdome Flex ASIC A4
HPE Superdome Flex ASIC A12
HPE Superdome Flex ASIC A13
HPE Superdome Flex ASIC A3
HPE Superdome Flex ASIC A2
HPE Superdome Flex ASIC A14
HPE Superdome Flex ASIC A15
HPE Superdome Flex ASIC A1
HPE Superdome Flex ASIC A0

Superdome crossbar topology – 850 GB/s of bisection bandwidth

Summer 2022 Call for Proposals

- All details available in the official webpage: https://www.cmu.edu/psc/aibd/neocortex/2022-06-cfp-summer-2022.html



Neocortex Summer 2022 Allocation Submissions

Name	Date (ET)
Application begins	June 8, 2022
Application ends	July 15, 2022
Response ends	August 15, 2022

Summer 2022 Call for Proposals

- Open to almost all U.S.-based university and non-profit researchers.
- Applications welcomed and processed through EasyChair.
- Applications welcomed for a period of 5 weeks.
- 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 mid August.
- 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. Checkpoint sessions every 3 months or so.
- 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/2022-06-cfp-summer-2022.html
Contact us with additional questions, input, or requests	neocortex@psc.edu