Neocortex: An Innovative Resource for Accelerating AI and HPC Development for Rapidly Evolving Research

Mei-Yu Wang
ML Research Scientist, Pittsburgh Supercomputing Center

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

Carnegie Mellon University
University of Pittsburgh

March 21, 2023

NEOCORTEX
Unlocking Interactive AI for Rapidly Evolving Research

Supported by OAC 2005597

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.**

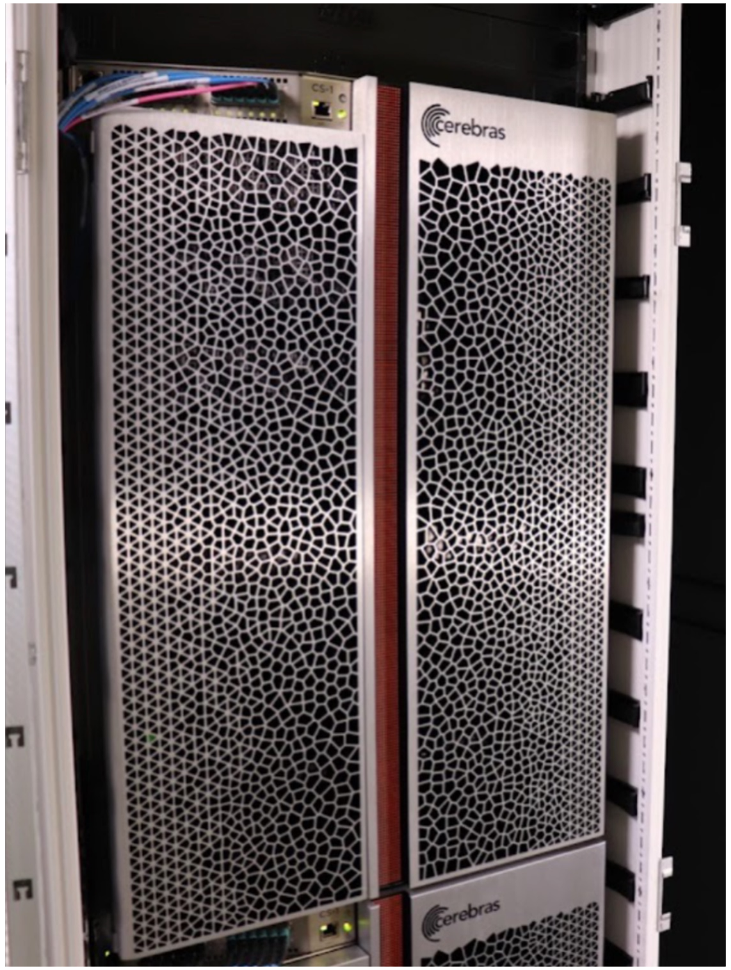
Context – NSF Award

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

NSF Award OAC-2005597 ($12.25M awarded to date):
Category II: Unlocking Interactive AI Development for Rapidly Evolving Research

Cerebras and HPE delivered *Neocortex*

The Neocortex System

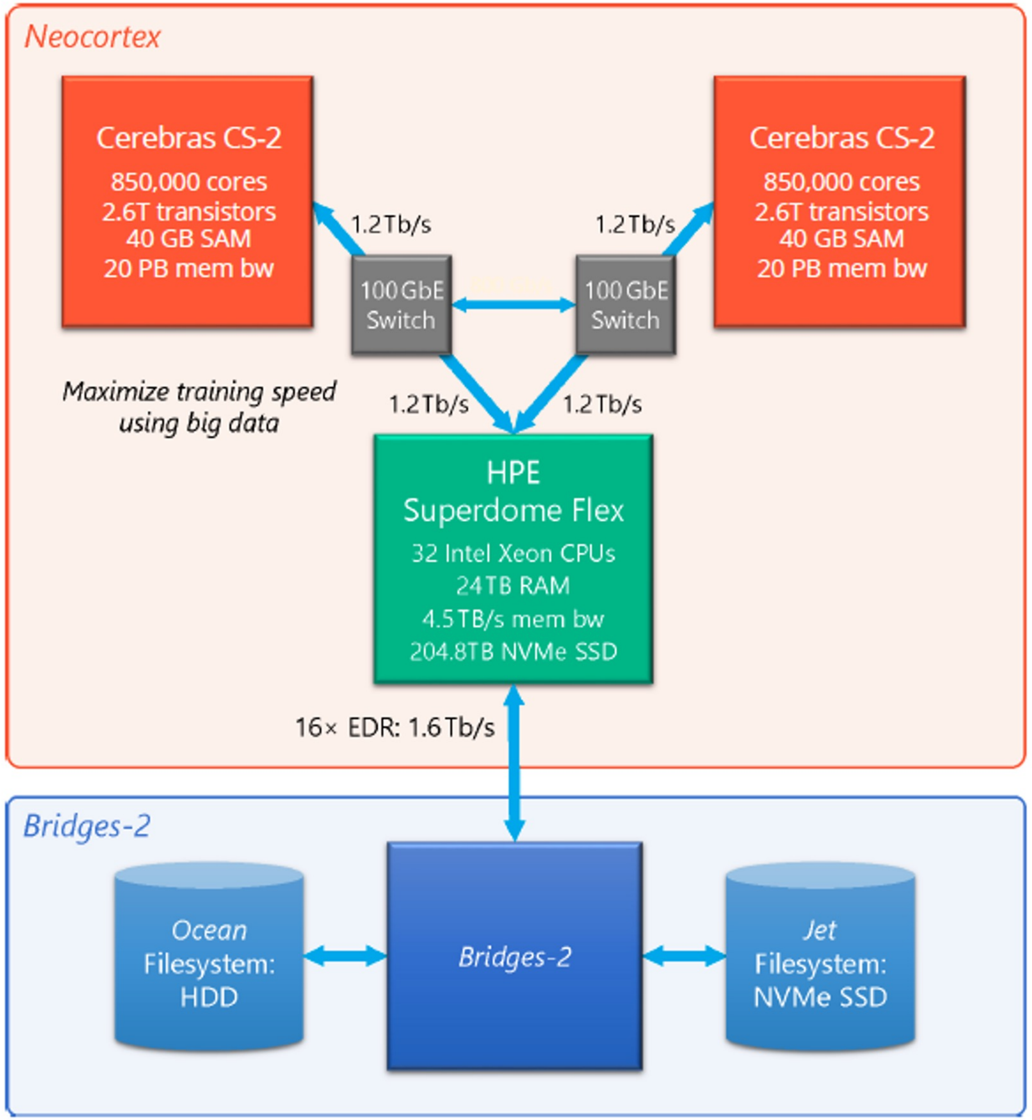


Left: one of the Cerebras CS servers

Middle: Two CS-2 servers (left) and Superdome Flex (right)

Right: Bridges-2

Neocortex System Overview

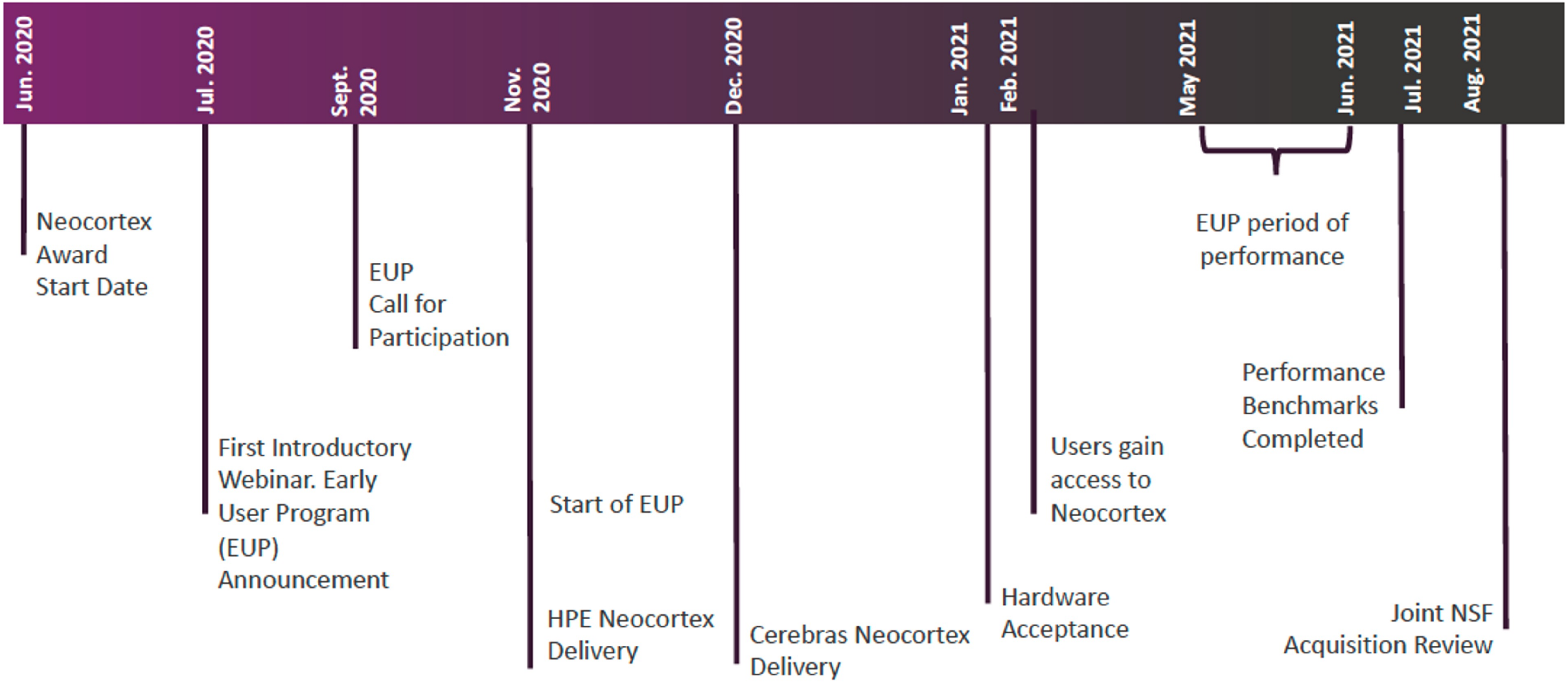


The main AI accelerators are the Cerebras Wafer Scale Engines (WSE-2)

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

Overall Project Timeline (1/2)



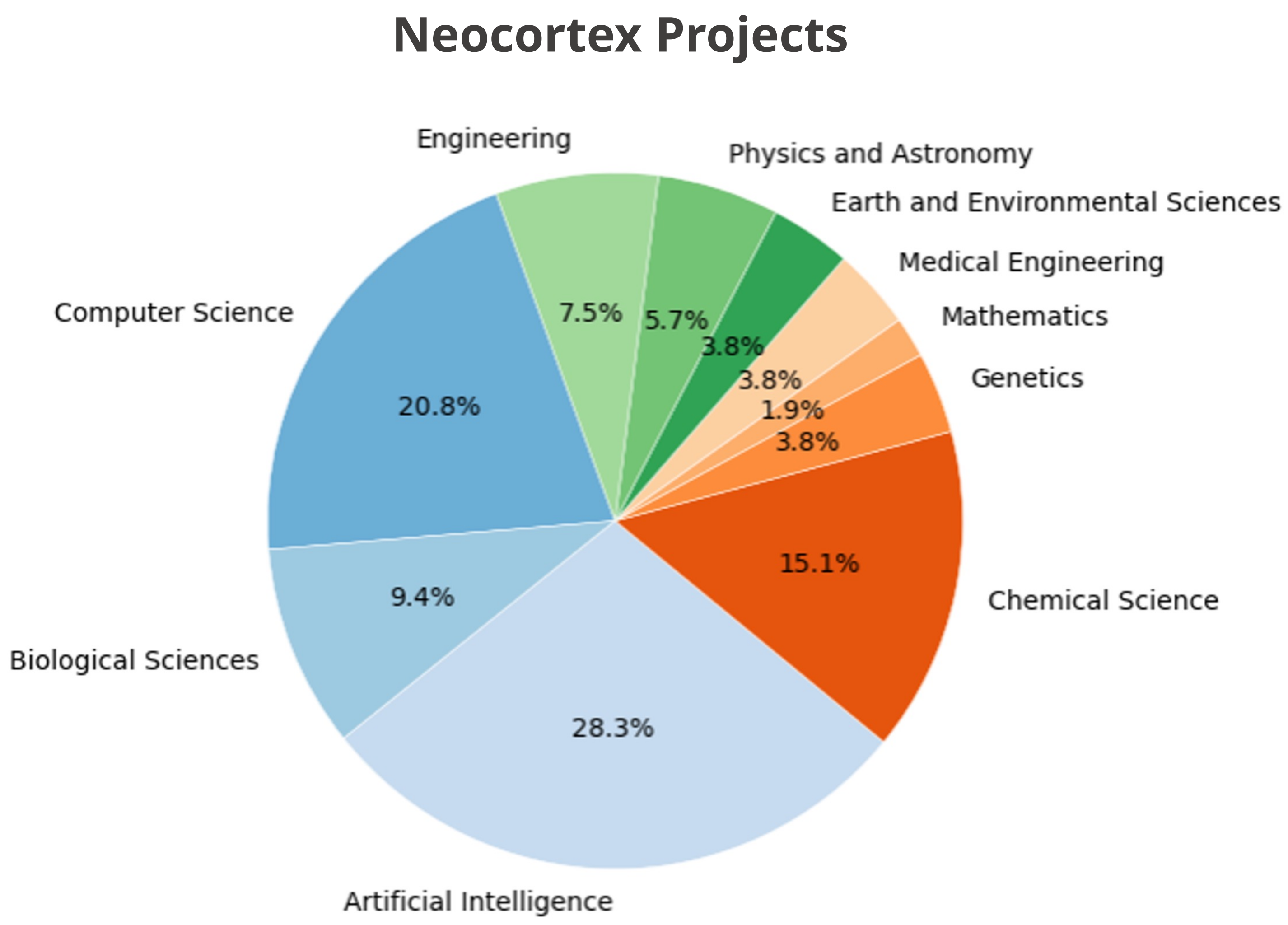
NEOCORTEX

Overall Project Timeline (2/2)

- **Oct. 2021** — System Overview & Fall 2021 CFP Announcement Webinar
- **Nov. 2021** — Fall 2021 CFP Submission Period
- **Jan. 2022** — CFP Fall 2021 User onboarding
- **Mar. 2022** — Neocortex CS-2 servers upgrade
- **Apr. 2022** — First Cerebras technologies Developer Community meeting
- **Jun. 2022** — System Overview & Summer 2022 CFP Announcement Webinar
- **Jul. 2021** — Supplemental request for funding – Accelerated adoption of Neocortex; Summer 2022 CFP Submission Period
- **Sep. 2022** — Joint NSF Testbed Ops. Y1 Review

NEOCORTEX

Projects Hosted by Neocortex



Active Project Details

Project Title: Physics informed distributed dynamic simulations of large-scale power grids for Optimization and Stability Analysis

Amar Ramapuram, Arizona State University

Project Abstract: The electric power system is the nation's critical infrastructure. It consists of millions of individual devices that are sparsely interconnected through transmission and distribution lines. As the current and power only flows along these lines, the properties of a device, such as voltage and current drawn, is only influenced directly by the behavior of the immediate neighboring devices and the properties of the interconnecting lines. Non-convex Optimization on power grid operations can be reformulated into a primal-dual dynamical system that has distributed dynamics and whose equilibrium is the optimal solution. Similarly, stability analysis in power grids is performed by simulating the non-linear dynamical equations for various disturbances and observing the evolution of voltages and currents over long time scales. We can also calculate stability metrics using the voltage evolution during the simulations to understand how close the system is to a collapse. Conventional approaches for simulating the power grid dynamics have taken advantage of the sparse nature of the power grid using techniques such as sparse solvers, etc. However, these approaches have not utilized the distributed nature of the power grid dynamics due to the lack of the right computing architecture that can leverage this property. Neocortex fills this void by having sufficient memory to hold the entire state of the system in memory while also having ultrafast communication between neighboring computing cores. We can recast the dynamic simulations into a form where the evolution of a state of a grid component (generator, motor, transmission line, etc.) is based on the various states of the neighboring components. These dynamics can be simulated in a near real-time fashion. We envision that there is likely to be a speedup of ~20x (based on the analysis of the NETL CFD solution using neocortex) compared to existing approaches for systems >10k elements.

More info:
https://www.cmu.edu/psc/aibd/neocortex/2022-03-neocortex-active-projects.html

NEOCORTEX

Advantages of Neocortex over Traditional Accelerators: DL

- Cluster-scale performance in a single chip
- High on-chip memory and high memory bandwidth
- Efficient computation for large deep learning models

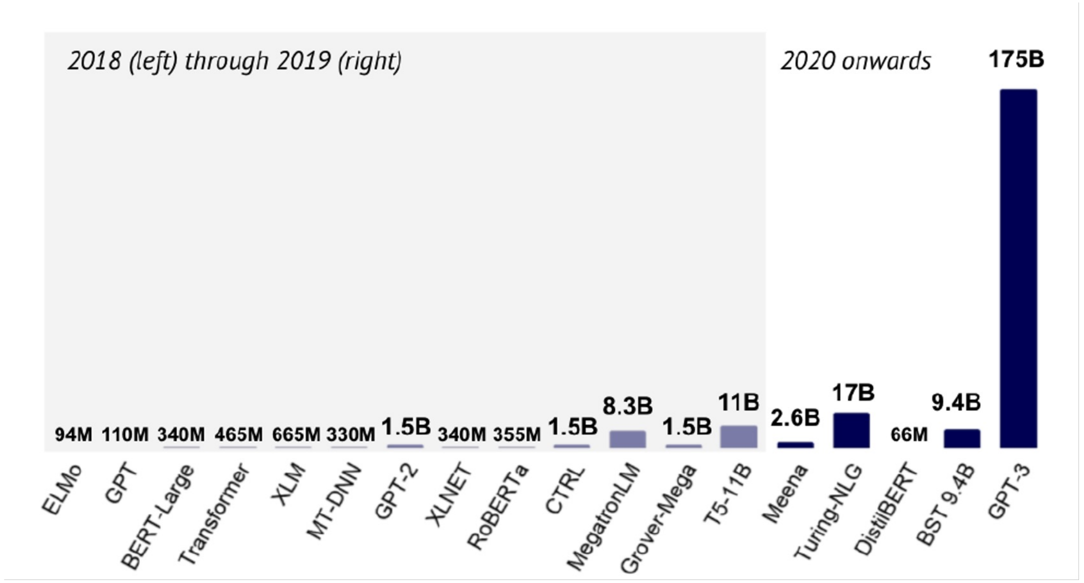
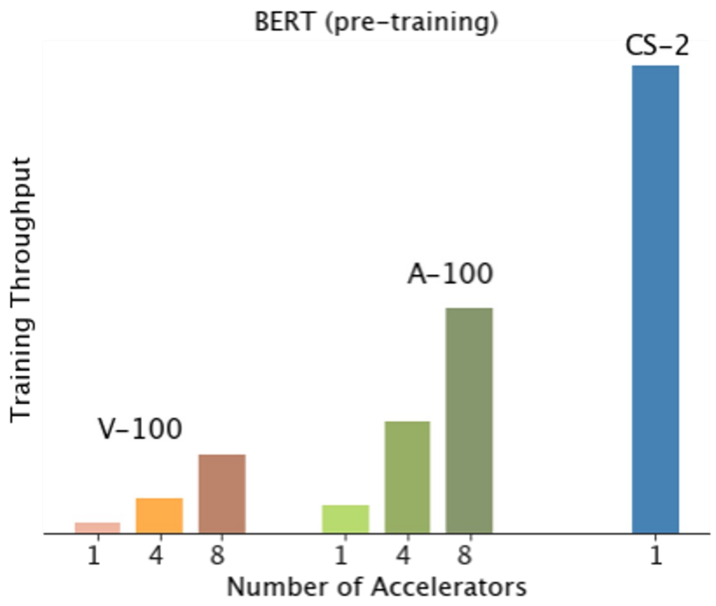


Figure from N. Benaich and P. Schwaller, State of AI Report, (2020) Available at https://www.stateof.ai/.

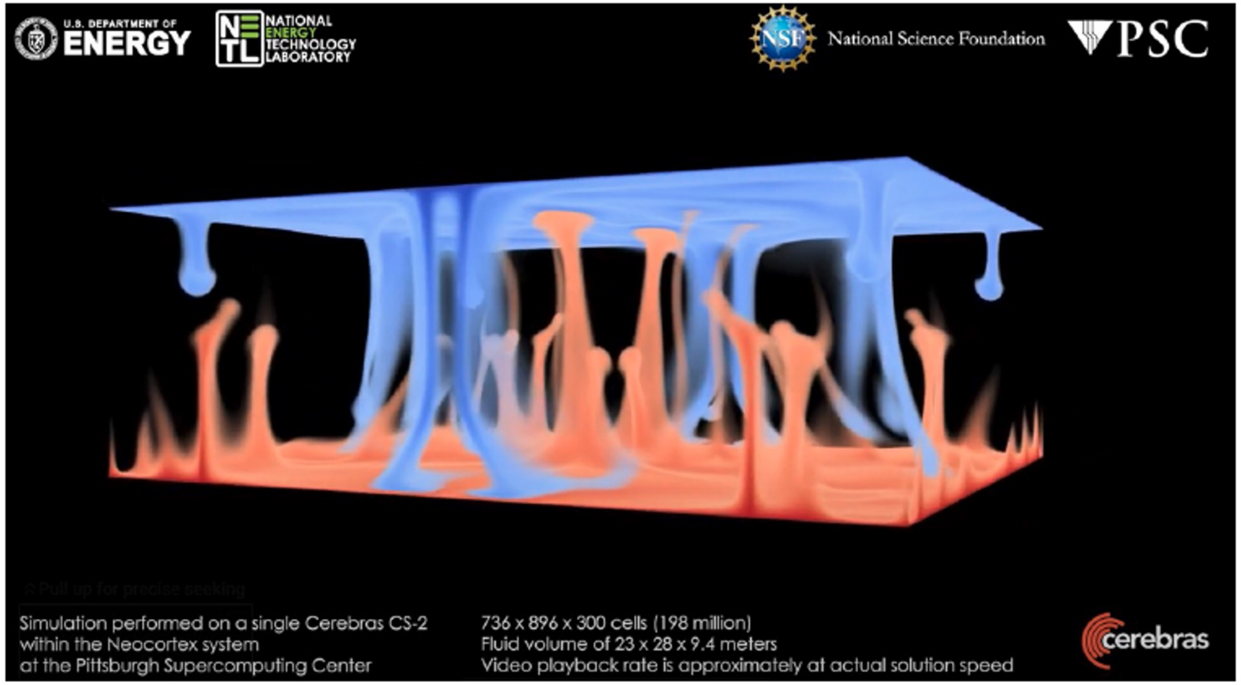


NEOCORTEX

Advantages of Neocortex over Traditional Accelerators: HPC

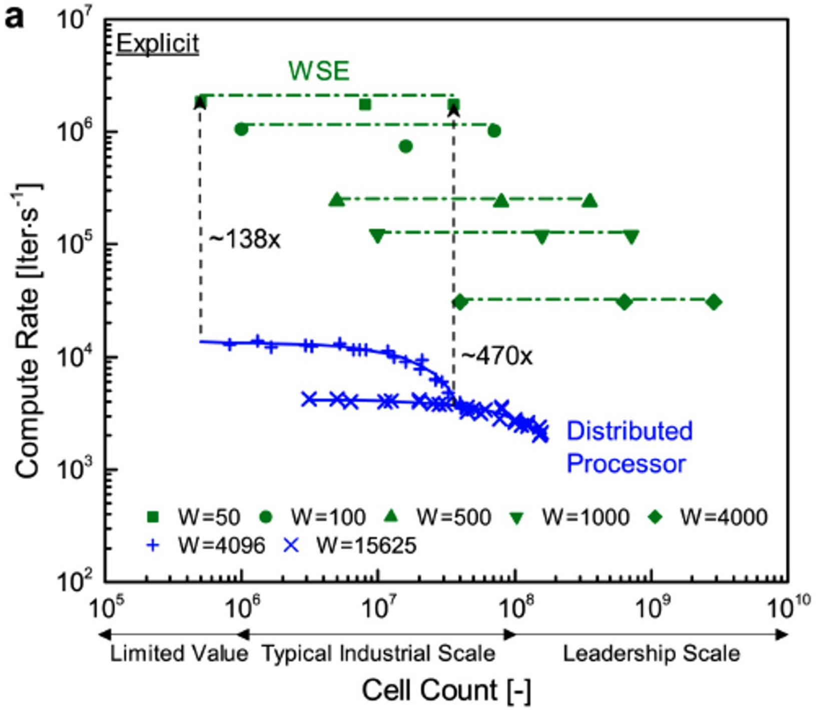
Credit: Dirk Van Essendelft, NETL

CFD Demonstration Completely on WSE



https://www.youtube.com/watch?v=5ad9f70ORvQ

Several Hundred Times Faster Than Distributed Computing



https://arxiv.org/abs/2209.13768

NEOCORTEX

Spring 2023 Call for Proposals

Details are available in the official website:
https://www.cmu.edu/psc/aibd/neocortex/2023-03-cfp-spring-2023.html

***Neocortex* Spring 2023 Allocation Submissions**	
Name	**Date (ET)**
Application begins	**March 15, 2023**
Application ends	April 12, 2023 (Anywhere on Earth time zone)
Response ends	May 10, 2023
Allocation starting date	User access to start mid-May 2023 (rough estimate)

- Open to all **U.S.-based university** and **non-profit researchers**. Offered **at no cost** for researchers advancing **open-science works**.
- Applications will be evaluated as they come in.
- Lightweight application via a short form.
- Allocations to Neocortex and the associated Bridges-2 resources will be initially granted for a year by default.
- Onboarding meetings will be scheduled to confirm the scope of the project and suitability.
- Close collaboration and constant communication between domain projects, PSC, and vendors is expected. Checkpoint sessions every 3 months or so.
- Feedbacks and user experience sharing are expected from users to further enrich the project.

Tracks of Supported Applications – as of March 2023

1 Cerebras modelzoo ML models

Transformers:

- BERT, GPT (GPT-2, GPT-3, GPT-J), Linformer, RoBERTa, T5, Transformer

MLP & 2D UNet (limited)

https://portal.neocortex.psc.edu/docs/models_supported.html

2 ML Models similar to the Cerebras modelzoo models

ML models that are a combination of layers/operations supported by Cerebras software stack.

https://docs.cerebras.net/en/latest/index.html

TensorFlow

Class:

CerebrasEstimator

- Based on TF Estimator, takes over executions after XLA compilation
- TensorFlow 2.2

PyTorch

Python Module:

cerebras.framework.torch

- Based on PyTorch XLA
- Wrappers for Dataloader, Module, Session
- PyTorch 1.11

3 General Purpose SDK

General purpose programming with the Cerebras SDK to write custom program ("kernels"), not integrated with Tensorflow nor PyTorch.

4 WFA (WSE Field-Equation API)

Domain specific programming on structured grids. It would involve a set of PDEs discretized to first/second order in implicit or explicit methods.

NEOCORTEX

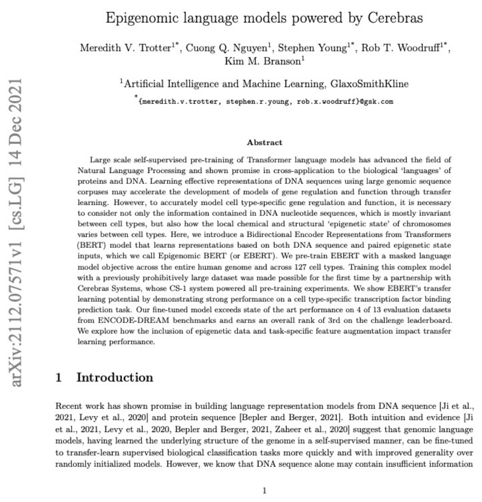
Topics of Interest for ML Applications (1/2) 1 2

- <u>Neocortex is best suited for running Transformer style models such as BERT, GPT, Transformer, T5, and ViT.</u>
- **Transformer style models** cover a wide range of tasks such as:
  - ***Sequence classification*** - sentiment analysis, molecule properties
  - ***Sequence annotation*** - extractive summarization, protein binding site identification
  - ***Sequence generation*** - abstractive summarization, candidate drug generation
  - ***Sequence to sequence mapping*** - Natural language translation, code translation
  - ***Representation learning for biological sequences*** (genome, epigenome, protein)

Topics of Interest for ML Applications (2/2)

1 2

- Example projects/models:
  - GSK: new sequence modeling for genetic medicine: "Epigenomic language models powered by Cerebras," Trotter and et al., 2021
  - PubMedBERT: "Domain-specific language model pretraining for biomedical NLP," Gu and et al., 2021
  - AntiBERTa: "Deciphering the language of antibodies using self-supervised learning," Leem and et al., 2021
  - TAPE: "Evaluating protein transfer learning with TAPE," Rao and et al., 2019
  - SMILES-BERT: "SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction," Wang and et al, 2019



Epigenomic language models powered by Cerebras

Meredith V. Trotter[1*], Cuong Q. Nguyen[1], Stephen Young[1*], Rob T. Woodruff[1*], Kim M. Branson[1]

[1]Artificial Intelligence and Machine Learning, GlaxoSmithKline

*{meredith.v.trotter, stephen.r.young, rob.x.woodruff}@gsk.com

arXiv:2112.07571v1 [cs.LG] 14 Dec 2021

Abstract

Large scale self-supervised pre-training of Transformer language models has advanced the field of Natural Language Processing and shown promise in cross-application to the biological 'languages' of proteins and DNA. Learning effective representations of DNA sequences using large genomic sequence corpuses may accelerate the development of models of gene regulation and function through transfer learning. However, to accurately model cell type-specific gene regulation and function, it is necessary to consider not only the information contained in DNA nucleotide sequences, which is mostly invariant between cell types, but also how the local chemical and structural 'epigenetic state' of chromosomes varies between cell types. Here, we introduce a Bidirectional Encoder Representations from Transformers (BERT) model that learns representations based on both DNA sequence and paired epigenetic state inputs, which we call Epigenomic BERT (or EBERT). We pre-train EBERT with a masked language model objective across the entire human genome and across 127 cell types. Training this complex model with a previously prohibitively large dataset was made possible for the first time by a partnership with Cerebras Systems, whose CS-1 system powered all pre-training experiments. We show EBERT's transfer learning potential by demonstrating strong performance on a cell type-specific transcription factor binding prediction task. Our fine-tuned model exceeds state of the art performance on 4 of 13 evaluation datasets from ENCODE-DREAM benchmarks and earns an overall rank of 3rd on the challenge leaderboard. We explore how the inclusion of epigenetic data and task-specific feature augmentation impact transfer learning performance.

1 Introduction

Recent work has shown promise in building language representation models from DNA sequence [Ji et al., 2021, Levy et al., 2020] and protein sequence [Bepler and Berger, 2021]. Both intuition and evidence [Ji et al., 2021, Levy et al., 2020, Bepler and Berger, 2021, Zaheer et al., 2020] suggest that genomic language models, having learned the underlying structure of the genome in a self-supervised manner, can be fine-tuned to transfer-learn supervised biological classification tasks more quickly and with improved generality over randomly initialized models. However, we know that DNA sequence alone may contain insufficient information

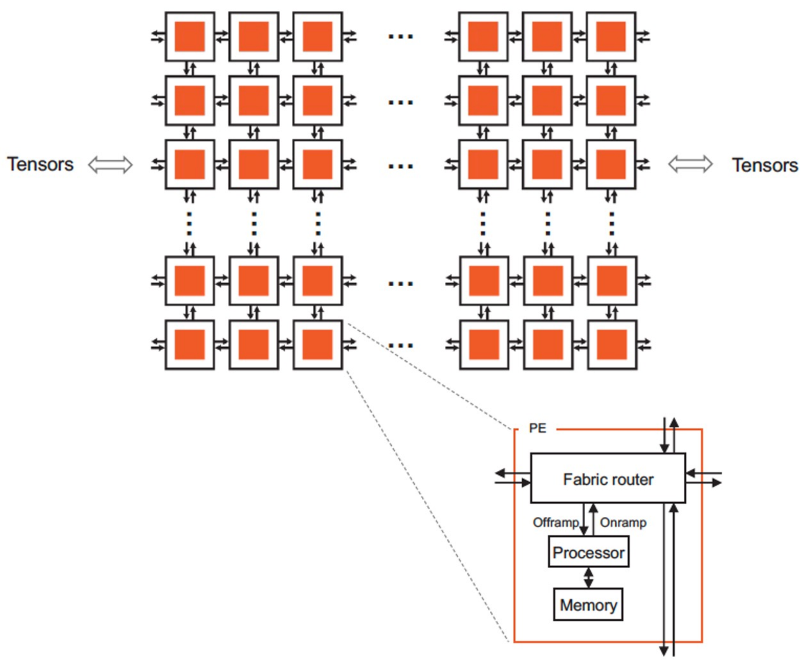
1

NEOCORTEX

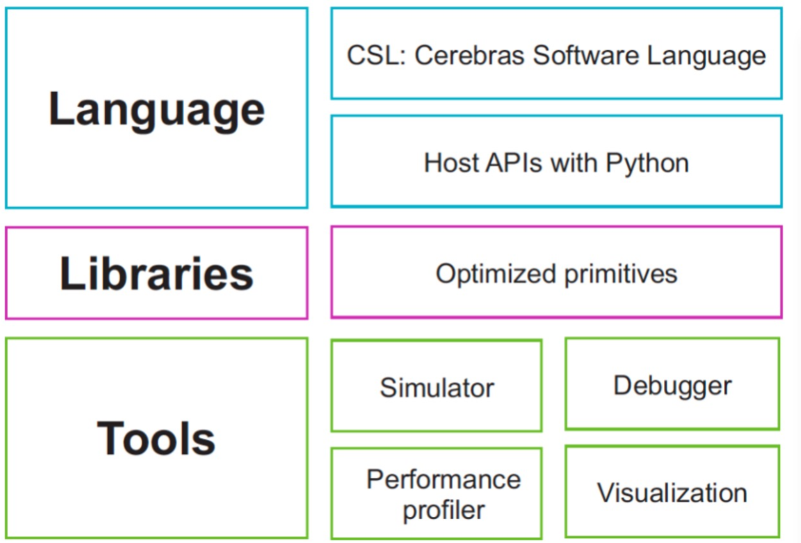
Cerebras SDK 3

CS-2 Dataflow Programming

To the programmer, the CS-2 appears as a logical 2D array of 850k individually programmable Processing Elements (PEs)



Credit: Leighton Wilson, Cerebras



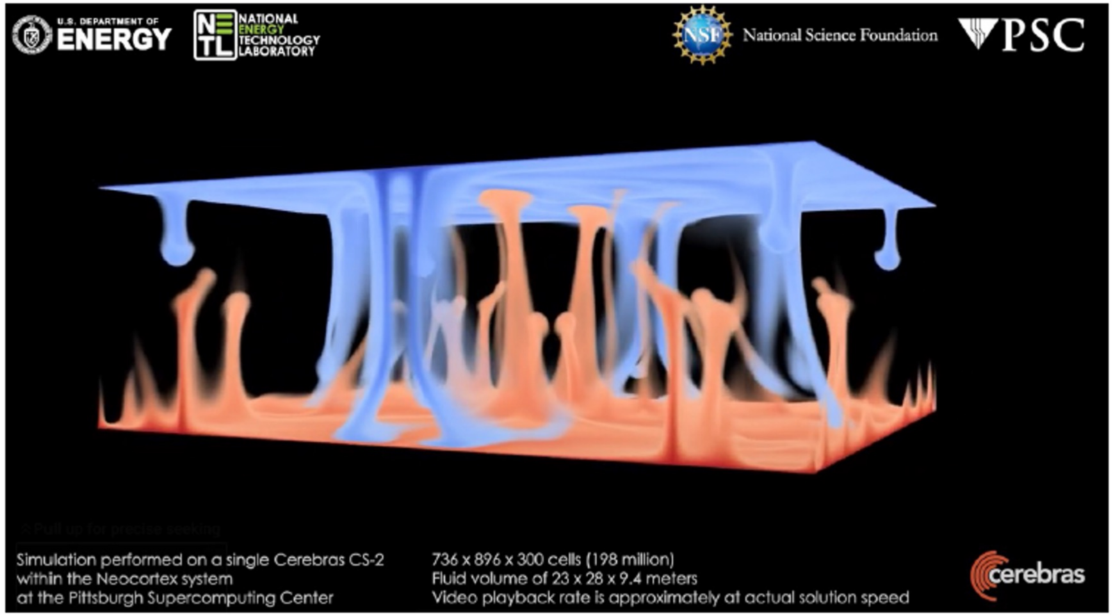
- **Benefit/Features:**
  - **High bandwidth and low-latency,** allowing for high parallel efficiency for non-linear and highly communicative code.
  - **40GB on-ship SRAM** uniformly across the chip that is **1 cycle** away from the PE. Capable of **1.2 Tb/s bandwidth** onto the chip.
  - **1 cycle** for PE-to-PE communication and read/write.

- **Topic of interest:**

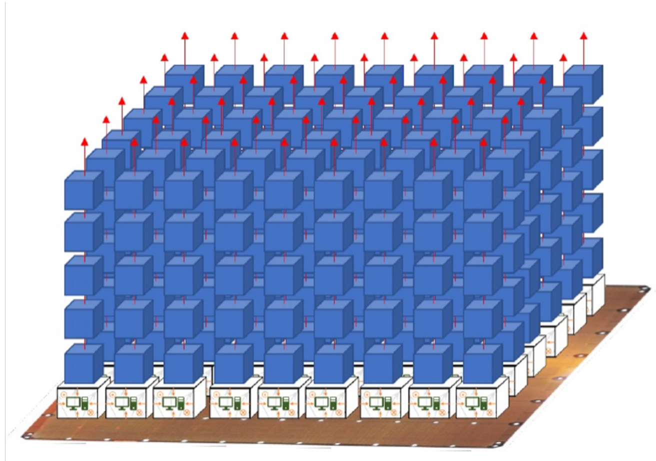
  Structured grid based PDE and ODE solvers, dense linear algebra, sparse linear algebra, particle methods with regular communication, Monte Carlo type problems that can fill the wafer, towards development of HPL, HPCG type benchmarks, custom ML kernels.

WFA (WSE Field-Equation API) (1/2) 4

Credit: Dirk Van Essendelft, NETL

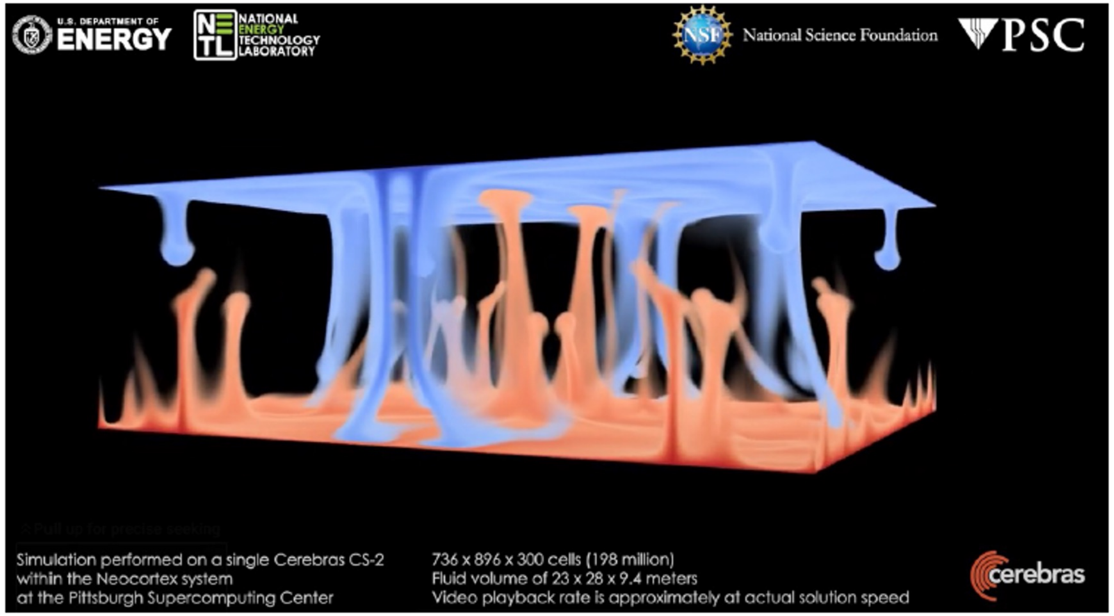


Demo video link: https://www.youtube.com/watch?v=5ad9f70ORvQ

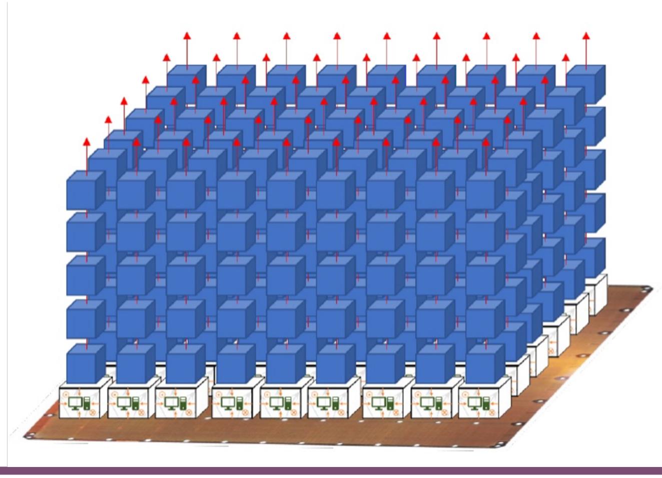


- Solving spatial-temporal problems on structured grids.
- Achieving **several hundred times faster** than distributed computing (see paper for details https://arxiv.org/abs/2209.13768)
- Simple Numpy-like Python front end (see documentation: https://dirk-netl.github.io/WSE_FE/) .

WFA (WSE Field-Equation API) (2/2) 4



Demo video link: https://www.youtube.com/watch?v=5ad9f70ORvQ



Credit: Dirk Van Essendelft, NETL

- **Project Guidelines**:
  - Problem Requirements
    - Must lay out on a Hex grid (3d or many 2d parallel)
    - Should involve Spatial Locality
    - Should be Data Intense
    - Single Precision, <40GB
  - Problem Examples
    - Computational Fluid Dynamics (FVM, FDM, FEM, LBM)
    - Structural Mechanics
    - Geomechanics
    - Weather/Climate
    - Materials – Ising Model, Density Functional Theory
    - CNN/RNN inference

NEOCORTEX

To Learn More and Participate

Apply to the Spring 2023 CFP (3/15/2023 – 4/12/2023)	**https://www.cmu.edu/psc/aibd/neocortex/2023-03-cfp-spring-2023.html**
Join the neocortex-updates email list	https://www.cmu.edu/psc/aibd/neocortex/newsletter-sign-up.html
Watch the Neocortex website for updates	https://www.cmu.edu/psc/aibd/neocortex/
Contact us with additional questions, inputs, requests	Email: neocortex@psc.edu

Thank you!