Introduction to Neocortex

ECSS Symposium, October 20, 2020

PI: Paola Buitrago
PSC

Co-PI: Sergiu Sanielevici
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
Context – NSF Award

Acquisition and operation of Neocortex is 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

PSC Partnership with Cerebras and HPE
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.
"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.”

Motivation

Some Recent Transformer-type Networks

<table>
<thead>
<tr>
<th>Network</th>
<th>Published</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT Large</td>
<td>October 11, 2018</td>
<td>340M</td>
</tr>
<tr>
<td>PEGASUS Large</td>
<td>December 18, 2019</td>
<td>568M</td>
</tr>
<tr>
<td>GPT-2 (48 layers)</td>
<td>February 2019</td>
<td>1.5B</td>
</tr>
<tr>
<td>Megatron-LM</td>
<td>August 13, 2019</td>
<td>8.3B</td>
</tr>
<tr>
<td>GPT-3 (96 layers)</td>
<td>June 3, 2020</td>
<td>175 B</td>
</tr>
</tbody>
</table>

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)
Properties of the right compute for DL

Compute is …

• Massive, more than can fit on a traditional single die
• Optimized for linear ops on sparse tensors, to execute most common ops fast, to exploit sparsity in models and data
• Flexible, to support evolving models
Properties of the right compute for DL

Compute is …

- **Massive**, more than can fit on a traditional single die
- Optimized for **linear ops on sparse tensors**, to execute most common ops fast, to exploit sparsity in models and data
- **Flexible**, to support evolving models

Memory is …

- **Large, high-bandwidth, tightly coupled with compute**, so utilization doesn’t depend on batch size
Properties of the right compute for DL

Compute is …

- **Massive**, more than can fit on a traditional single die
- Optimized for **linear ops on sparse tensors**, to execute most common ops fast, to exploit sparsity in models and data
- **Flexible**, to support evolving models

Memory is …

- **Large, high-bandwidth, tightly coupled with compute**, so utilization doesn’t depend on batch size

Fabric is …

- **High bandwidth, low-latency** for seamless model and data parallelism
- **Fully configurable** for each workload
Outline

• Introduction

• Neocortex for Research and Education
  – Target Applications
  – System Architecture

• Early User Program

• Summary

• Q&A
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

Mesocortex, allocortex
Basal mid- and forebrain, hypothalamus, thalamus
Pontine tegmentum
Medulla oblongata
Peripheral & enteric nervous systems
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, managed by DMF
Who is this designed for?

• People doing demanding deep learning training.

• Users with complementary projects on Bridges-2.

• Examples:
  - Large sets of medical images.
  - 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.
Neocortex: System Overview

PSC's HPC+AI+HPDA Ecosystem

Neocortex

Cerebras CS-1
400,000 cores
1.2T transistors
18GB SRAM

Cerebras CS-1
400,000 cores
1.2T transistors
18GB SRAM

100 GbE Switch
12×100GbE: 1.2 Tb/s
12×100GbE: 1.2 Tb/s
24×100GbE: 2.4 Tb/s

HPE Superdome Flex
32 Intel Xeon CPUs
24 TB RAM
204.8TB NVMe SSD

Maximize training speed using big data

8× HDR-200: 1.6 Tb/s
Load 200TB training data in ~20 minutes

Bridges-2
>70,000 cores
192 V100-32GB GPUs
15 PB Lustre (disk)
8 PB archive

Bridges-2 filesystem
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.

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

NOVEMBER 19, 2019

PRESS RELEASE | ARGONNE NATIONAL LABORATORY
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 PByte/s memory bandwidth
- 100 Pbit/s fabric bandwidth

- System IO: 12 x 100 GbE
- System power: 20 kW
- Ingests TensorFlow, PyTorch, etc.
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
- fmac [z] = [z], [w], a
  
  \[
  \begin{align*}
  \text{3D} & \quad \text{3D} & \quad \text{2D scalar}
  \end{align*}
  \]
- Data flow architecture, sparsity “harvesting”
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 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

<table>
<thead>
<tr>
<th>Specification</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processors</td>
<td>Intel Xeon, TBA</td>
</tr>
<tr>
<td>Memory</td>
<td>24 TiB RAM, aggregate memory bandwidth of 4.5 TB/s</td>
</tr>
<tr>
<td>Local Disk</td>
<td>32 x 6.4 TB NVMe SSDs</td>
</tr>
<tr>
<td></td>
<td>- 204.6 TB aggregate</td>
</tr>
<tr>
<td></td>
<td>- 150 GB/s read bandwidth</td>
</tr>
<tr>
<td>Network to CS-1 systems</td>
<td>24 x 100 GbE interfaces</td>
</tr>
<tr>
<td></td>
<td>- 1.2 Tb/s (150 GB/s) to each Cerebras CS-1 system</td>
</tr>
<tr>
<td></td>
<td>- 2.4 Tb/s aggregate</td>
</tr>
<tr>
<td>Interconnect to Bridges-2</td>
<td>16 Mellanox HDR-100 InfiniBand adapters</td>
</tr>
<tr>
<td></td>
<td>- 1.6 Tb/s aggregate</td>
</tr>
<tr>
<td>OS</td>
<td>Red Hat Enterprise Linux</td>
</tr>
</tbody>
</table>
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)

- Initiating the partnership between the Neocortex team and the research community.
- At least 30 days of actual system use, part of system acceptance.
- Participating research groups will receive one-on-one assistance from PSC and Cerebras experts to port, scale, and optimize their workflows.
- Areas of focus in the EUP are on generating scientific outcomes, achieving high performance, and maximizing ease of use.
- The EUP includes providing feedback to PSC and Cerebras to document the results of the EUP and inform the evolution of the system.
- There is no cost to participate in the Neocortex EUP or to use Neocortex for open research following the EUP.
EUP Preparation

• Proposals were invited via EasyChair, from Aug 31, 2020 to September 29, 2020
• Received 42 proposals in research areas including AI Theory, Bioinformatics, Neurophysiology, Materials Science, Electrical and Computer Engineering, Medical Imaging, Geophysics, Civil Engineering, IoT, Social Science, Drug Discovery, Fluid Dynamics, Ecology, and Chemistry.
• The *Neocortex* team has initiated follow-up discussions with the proposers to hone in on activities suitable for the EUP and plan work possible after the EUP.
• This will be followed by preparatory training prior to system access.
• Work with the user community leverages ECSS NIP
Neocortex Target Timeline

June 1, 2020  Award start date; preparatory activities begin
- System and user environment, documentation, content, dissemination, etc.
- Broadly invite participation in the Early User Program
- Logistics, machine room preparation, etc.

September 2020  Accept applications for the Early User Program

November 2020  System delivery, installation, initial testing

December 2020  Early User Program, Acceptance Testing
| Watch the Neocortex website for updates! | [https://www.cmu.edu/psc/aibd/neocortex/](https://www.cmu.edu/psc/aibd/neocortex/) |
| Contact us with additional questions, input, or requests | [necortex@psc.edu](mailto:necortex@psc.edu) |
Summary

- **Neocortex** is an upcoming NSF-funded innovative advanced computing system that will be made available in late 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 in Q4 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 focus strongly on outreach, training, and user support – leveraging ECSS NIP.

- Please, get in contact if you consider you have an application that you think could benefit from Neocortex and would like to explore this with our team: necortex@psc.edu
Thank you to all those contributing to Neocortex!

Andrew K. Adams  Rajanie Prabha*
Paola Buitrago*  Sergiu Sanielevici*
Ken Hackworth  Amanda Slimick
Ed Hanna  Julian Uran*
David Moses  John Urbanic*

* ECSS members
Outline

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