System Integration of *Neocortex*, a Unique, Scalable AI Platform

Paola A. Buitrago*
paola@psc.edu
Pittsburgh Supercomputing Center,
Carnegie Mellon University/
University of Pittsburgh
Pittsburgh, PA, USA

Julian Uran*
julian@psc.edu
Pittsburgh Supercomputing Center,
Carnegie Mellon University/
University of Pittsburgh
Pittsburgh, PA, USA

Nicholas A. Nystrom
nicholas.nystrom@peptilogics.com
Peptilogics, Inc.
Pittsburgh, PA, USA

ABSTRACT

To advance knowledge by enabling unprecedented AI speed and scalability, the Pittsburgh Supercomputing Center (PSC), a joint research center of Carnegie Mellon University and the University of Pittsburgh, in partnership with Cerebras Systems and Hewlett Packard Enterprise (HPE), has deployed *Neocortex*, an innovative computing platform that accelerates scientific discovery by vastly shortening the time required for deep learning training and inference, fosters greater integration of deep AI models with scientific workflows, and provides promising hardware for the development of more efficient algorithms for artificial intelligence and graph analytics. *Neocortex* advances knowledge by accelerating scientific research, enabling development of more accurate models and use of larger training data, scaling model parallelism to unprecedented levels, and focusing on human productivity by simplifying tuning and hyperparameter optimization to create a transformative hardware and software platform for the exploration of new frontiers. *Neocortex* has been integrated with PSC's complementary infrastructure. This papers shares experiences, decisions, and findings made in that process. The system is serving science and engineering users via an early user access program. Valuable artifacts developed during the integration phase have been made available via a public repository and have been consulted by other AI system deployments that have seen *Neocortex* as an inspiration.

CCS CONCEPTS

• **Computer systems organization** → **Neural networks**; **Neural networks**; • **Hardware** → **Emerging technologies**; **Analysis and design of emerging devices and systems**; • **Computing methodologies** → **Artificial intelligence**; **Artificial intelligence**.

KEYWORDS

artificial intelligence, AI, deep learning, high performance computing, HPC, neural networks, specialized processors, NSF

*Both authors contributed equally to this research.

PEARC '21, July 18–22, 2021, Boston, MA, USA

ACM ISBN 978-1-4503-8292-2/21/07.
https://doi.org/10.1145/3437359.3465604

ACM Reference Format:
Paola A. Buitrago, Julian Uran, and Nicholas A. Nystrom. 2021. System Integration of *Neocortex*, a Unique, Scalable AI Platform. In *Practice and Experience in Advanced Research Computing (PEARC '21), July 18–22, 2021, Boston, MA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3437359.3465604

1 INTRODUCTION

To advance knowledge by enabling unprecedented AI speed and scalability, the Pittsburgh Supercomputing Center (PSC), a joint research center of Carnegie Mellon University and the University of Pittsburgh, in partnership with Cerebras Systems and Hewlett Packard Enterprise (HPE), has deployed *Neocortex* [1], an innovative computing resource that is accelerating scientific discovery by vastly shortening the time required for deep learning training/inference, fostering greater integration of deep AI models with scientific workflows, and providing revolutionary innovative hardware for the development of more efficient algorithms for artificial intelligence and graph analytics. *Neocortex* advances knowledge by accelerating scientific research, enabling development of more accurate models and use of larger training data, scaling model parallelism to unprecedented levels, focusing on human productivity by simplifying tuning and hyperparameter optimization, and providing a transformative hardware and software platform for the exploration of new frontiers.

Neocortex is the first system of its kind. It is by design a testbed that aims to inform future system acquisitions within PSC and by the national advanced computing community.

Neocortex is the first architecture that couples Cerebras CS-1 servers with a large-memory front end, specifically, HPE Superdome Flex server, to enable scaling and increase ease of use. This integration advances the frontiers of scaling to multiple CS-1 systems, high corecount servers driving many PCI Express lanes, and internode bandwidth. This paper aims to describe integration challenges, design considerations, and the ways in which the *Neocortex* system was architected and configured to respond to them. Results of the integration in the form of software artifacts have been made available as a public repository for the enrichment of our wider HPC community. It includes detailed configuration information and scripts that can be easily reused for other systems with similar architectures.

2 SYSTEM COMPONENTS

The novel *Neocortex* architecture couples two powerful Cerebras CS-1 AI servers with a large shared memory HPE Superdome Flex HPC server to achieve unprecedented AI scalability with carefully designed system balance (Figure 1).

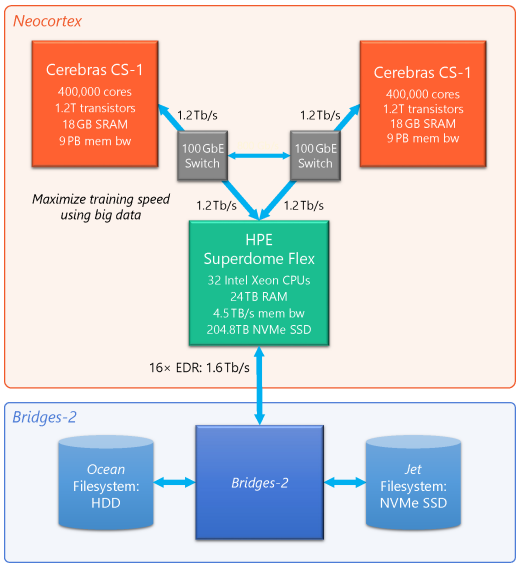

Figure 1: *Neocortex* System Architecture.

2.1 Cerebras CS-1: AI Engines

Each Cerebras CS-1 is powered by one Cerebras Wafer Scale Engine (WSE) processor, an unprecedented high-performance processor designed specifically to accelerate deep learning training and inferencing. The Cerebras WSE contains 400,000 AI-optimized cores implemented on a 46,225 mm^2 wafer with 1.2 trillion transistors. An on-chip fabric provides 100 Pb/s of bandwidth through a fully configurable 2D mesh with no software overhead. The Cerebras WSE includes 18 GB of SRAM accessible within a single clock cycle and at 9 PB/s bandwidth. The Cerebras WSE is uniquely engineered to enable efficient sparse computation, wasting neither time nor power multiplying the many zeroes that occur in deep networks. The Cerebras CS-1 software can be programmed with familiar AI frameworks such as TensorFlow and PyTorch, which for execution efficiency are mapped onto an optimized graph representation and model-specific computation kernels. The CS-1 also supports native code development. Support for the most popular deep learning frameworks and automatic, transparent acceleration provide researchers with exceptional ease of use.

2.2 HPE Superdome Flex: Front End

The HPE Superdome Flex ("SDF") HPC server of *Neocortex* is a compute and memory intensive, user-friendly front end for the Cerebras CS-1 servers. It enables flexible pre- and post-processing of data flowing in and out of the attached WSEs, preventing bottlenecks from taking full advantage of the WSE capability, and implementing advanced deep learning functions such as augmentation, hyper-parameter and model optimization, and ensemble learning. The Superdome Flex is robustly provisioned with 24 TB of RAM, 204.8 TB of high-performance NVMe flash storage, 32 Intel Xeon Platinum 8280 CPUs (896 cores in aggregate), and 24 100 GbE network interface cards to maximize flexibility for scaling applications across both CS-1 systems. Internally, the HPE Superdome Flex is interconnected by a custom memory fabric ASIC for cache-coherent hardware shared memory sustaining 850 GB/s. Its large, fast memory and high compute performance enable training on very large datasets with ease, avoiding the laborious task of splitting and load-balancing datasets across worker nodes.

2.3 Bridges-2: Data & General Computing

Neocortex is federated via 16 InfiniBand HDR-100 connections (aggregate 1.6 Tbps) with Bridges-2 [1], an NSF-supported capacity resource into which Bridges-AI [2] is also being integrated. This yields great benefits to the user community including access to the Bridges-2 filesystem to manage persistent data; general-purpose computing for data preprocessing and traditional machine learning; interoperation with data-intensive projects using Bridges-2; and high-bandwidth external network connectivity to other XSEDE Service Providers, campus, labs, and clouds.

Two additional servers were added as management nodes for logins, editing, and running Slurm services. These provide effective access to the eight individual SDF chassis that have been configured as two separate socket partitions for high availability.

3 INTEGRATION GUIDELINES

Users run jobs on the *Neocortex* system by applying the Cerebras container (implemented in Singularity) to data that is resident on Superdome Flex' internal NVMe SSD-based filesystem. Large data or data from simulations may originate in Bridges-2. Consequently, the integration goals for *Neocortex* were as follows:

- Enable running a single job across both CS-1 systems.
- Enable running multiple jobs concurrently.
- Utilize the full 1.2 Tbps bandwidth to each CS-1.
- Maximize the bandwidth from Bridges-2's *Jet* filesystem.
- Maximize ease of running jobs via the Slurm scheduler.

4 INTEGRATION APPROACH

Integration consisted of three aspects: *physical integration*, consisting of datacenter and cooling; *networking and kernel*, consisting of connectivity and kernel-level optimizations; and *systems software and applications*, consisting of scheduling and application optimizations. Integration test are also presented.

4.1 Physical Integration

All integration and hosting of the CS-1s, Superdome Flex, networking equipment, Coolant Distribution Unit (CDU), power-delivery hardware, and Bridges-2 takes places in PSC's datacenter. This single facility is critical for integrating the systems. Adaptation was required to support the power and cooling requirements of the Cerebras hardware.

4.1.1 Chilled Water Cooling. Each CS-1 consumes up to 20 kW of power in only 16 U (28 inches) of space, requiring specialized cooling to avoid overheating. The system involves two cooling loops. A primary loop mixes water with glycol coolant that traverse the CS-1s using pumps, then transfers heat onto a secondary loop (i.e., the datacenter's chilled water supply) via a heat exchanger. A Coolant Distribution Unit (CDU) between the datacenter's chilled

water and the Cerebras CS-1 units supplies water at the required temperature and pressure; in our case, lowering the pressure and raising the temperature. After significant analysis, we chose the Motivair MCDU-25 [3], which has 625 kW of thermal capacity.

4.2 Networking and Kernel Optimizations

4.2.1 Multi-homed system. To reach the target data transfer speeds, multiple network interfaces were added to the Superdome Flex. Each of the eight chassis has two HDR-100 InfiniBand network interfaces, and each socket partition has six 100 GbE network interfaces, for a total of 1.6 Tbps over InfiniBand to the *Bridges-2* filesystems and 1.2 Tbps over Ethernet to each CS-1. The Superdome Flex is a multi-homed system that must actively use all of its network interfaces optimally to attain the target transfer speeds.

Having multiple interfaces can create issues at the Operating System (OS) level. Two options to address these issues are: 1) have each interface use a different subnet, or 2) use the same subnet for all of the interfaces. The first approach requires more processing on the network layer (layer 3), requiring routing for all packets. For the second approach, each interface has a different IP address under the same network, and no routing is required for transferring data over Ethernet as both the Superdome Flex and the CS-1 network interfaces are connected to the same layer 2 (switching) device, allowing faster data transfer at the cost of additional configuration.

For implementing the single subnet approach, there is a known behavior in which the OS (Linux) binds IP addresses to the system itself and not to the network interfaces, requiring modification to several settings to ensure that communication takes place using individual network paths:

- Address Resolution Protocol (ARP) flux: Required for the OS to reply to ARP calls using only the interface that is being queried and not any of the other interfaces.
- Policy-based routing [5]: Required so the traffic is sent in parallel over all the SDF 100 GbE network interfaces to each of the CS-1's network interfaces. The default behavior is for the OS to use the first NIC based on the network distance to the target. This is done by defining multiple routing tables and routes and rules, and specifying that packets originating from each of the local 100 GbE IP addresses should be transferred only over that specific device. Interface bonding, an alternative, is not currently supported by the Cerebras network stack.

4.2.2 Superdome Flex Kernel Optimizations. Several kernel optimizations were performed to optimize data transfer:

- Total number of processes: The default (OS) ulimit value was increased from 1024 (which was too low) to a value appropriate to the jobs to be executed, for example, 4096.
- Network window size: Small window sizes perform poorly when transferring large volumes of data. Increase the default window by trying different combinations for identifying the performance sweet-spot.
- Security settings were optimized to allow clean starts of the Mellanox drivers and unimpeded flow of network traffic to and from the Superdome Flex.
- Additional processes: To ensure fast transfer rates, only critical service processes are run on the Superdome Flex. Anything that is deemed non-critical is stopped and disabled.
- Flash NVMe SSDs RAID configuration: To allow datasets of considerable size (~10 TB) to be read locally (without network transit) at high speed, local flash storage is aggregated into RAID arrays for capacity and performance. Each chassis has eight 3.2 TB NVMe SSDs, which were configured into a performant RAID0 configuration of 25 TB aggregate. The highest bandwidth can be achieved by the four CPUs connected in the same chassis as the local RAID0 array. To make it easy to use the NVMe array local to the CPU resources granted by Slurm, an epilog script sets an environment variable pointing to the disk array in the NUMA domain.

4.3 Systems Software and Applications

4.3.1 Slurm configuration. A base Slurm configuration is used, to which the CS-1 boxes were added as special resources with the accounting and configless slurm extensions added for tracking usage and only managing the configuration files from the central Slurm host, respectively.

Initially, the CS-1 supports only one training process at a time. This will be extended to multiple training processes in the future. Starting an additional training processes terminate any ongoing training job, so it was crucial to enforce that only one researcher can use a CS-1 box at a time.

This was achieved by setting the CS-1s as generic resources on the Slurm configuration [4], requiring any training job to explicitly request a CS-1 so it is granted exclusive access and making the CS-1s "special resources" that must be booked before being used.

Since both CS-1 boxes are in the same layer 2 (switching) network and there are no authorization controls, an additional mechanism to GRES was required for enforcing exclusive use of the CS-1s. This involved two steps:

- A script was added to the Slurm epilog, setting the CS-1 IP address (required for starting training jobs) and passing it on to the jobs programmatically as an environment variable to be used by the Cerebras container used for training.
- OS-level firewall policies for the network connectivity between the Superdome Flex and the CS-1s were set to be restricted by default and only allowed while Slurm has granted access to each special resource over the lifetime of a job.

4.3.2 Application-Level Optimizations for Improved Throughput. Transferring data using a single thread will be limited by the CPU process running the transfer. Multiple threads must be started across the available CPU resources to allow each of the CPUs to use all of the available network interfaces evenly. The following steps are required for fast transfers:

- Multiple threads or processes must be started for copying the data, in which the number of parallel instances should be determined based on the dataset characterization, such as the total number of files to use and their average size.
- The input data is expected to be readily accessible to the processes and threads running on the CPUs, whether it is being accessed via a very fast shared-filesystem or locally available on the NVMe high-speed storage.

- The processes for copying data are expected to be pinned by users to specific CPU cores and RAM mappings to guarantee that all of the resources available are used evenly. This can be done by specifying the NUMA nodes to be used when starting processes, for example, with numactl.
- Direct I/O: There are different ways to read and write information, and even if the training jobs that are started by the Cerebras container already have a set logic for streaming data into the CS-1, any other data transfers that have to be performed should use direct I/O for best performance.

4.4 Integration Tests

The following tests confirmed correct integration of *Neocortex*:

- Compilation and training using the Cerebras Singularity container: The latest container is available on the shared filesystem and/or the SDF local storage. A health test and a sample training job successfully run on the CS-1 boxes.
- Even CPU cores usage for jobs: The jobs span tasks (processes) evenly across all available CPU cores on the SDF to avoid overloading the network interface cards, NVMe drives, and memory, which have a set NUMA affinity defined by the hardware topology of the Superdome Flex internals.
- Connectivity: All the network interfaces (both Ethernet and InfiniBand) successfully ping the remote IP addresses (CS-1s, shared filesystem), and the target IP address field from the response is the same as the origin IP address from the network interface used for the test. Additionally, all of the Ethernet IP addresses successfully ping all of the CS-1 IP addresses, including the "control" IP address.
- Exclusive access to especial resource: Only one user is able to use a CS-1 at a time. Additional Slurm jobs wait until a CS-1 becomes available again to satisfy the request.
- Access to shared filesystem: The users have access to the shared filesystems from both the login node and the Superdome Flex socket partitions.
- Data transfer speeds: When copying data from the shared filesystem into the Superdome Flex and following the application-level optimizations suggested in section 4.2.2, the transfer rate is at least twice as much as when using a single interface, with the best case scenario being 8× faster transfer using all of the available interfaces.
- Data streaming for training: The data streaming rate from the Superdome Flex into the CS-1s makes use of all of the available Ethernet interfaces by the worker processes started, while a process manager takes care of coordinating each of the data transfer processes.
- Directory permissions: Users have no problems writing into the locations enabled for *Neocortex*. Those are: 1) the *Bridges-2* shared filesystem, 2) their local $HOME directories, and 3) the NVMe arrays in the Superdome Flex.

5 EARLY RESULTS

As a result of a successful project integration phase, the *Neocortex* system has been fully deployed, and science and engineering early users have gained access early 2021. Fifteen initial research projects across diverse domains including drug discovery, genomics, molecular dynamics, climate research, computational fluid dynamics, signal processing and medical imaging analysis are using the system to speed up their AI training. Science and performance outcomes from these early users will be published in the project website (https://www.cmu.edu/psc/aibd/neocortex/) and in subsequent publications. The experiences and specialized system configuration required for the successful integration of the advanced components of *Neocortex* have been captured, to the extent possible, in the form of scripts and configurations files which are publicly available in the Public *Neocortex* Configurations and Scripts Repository on GitHub, available by request at https://github.com/pscedu/neocortex-public/. The *Neocortex* team is also available to support the deployment of similar systems by sharing experiences and providing guidance.

6 CONCLUSIONS

The *Neocortex* system is an NSF-supported, highly specialized supercomputer designed to enable discovery in science and engineering research by vastly shortening the time required for deep learning training and inference, fostering greater integration of deep AI models with scientific workflows, and enabling the development of more efficient algorithms for artificial intelligence and graph analytics. *Neocortex* is the first system to couple two Cerebras CS-1 servers and an HPE Superdome Flex. The innovative nature of the project presented unique challenges that were successfully addressed through a research collaboration with Cerebras Systems and HPE Labs. The system integration phase of the project has been advanced successfully, and valuable artifacts have been made available for the wider community via a public repository. The *Neocortex* project has since supported the deployment of similar systems across the world and shared the resources and experience gained during the project integration phase. The *Neocortex* system is serving early users from various areas of science and engineering that involve ambitious AI training.

ACKNOWLEDGMENTS

Neocortex is supported by the National Science Foundation grant OAC-2005597. The authors would like to thank Cerebras Systems and HPE for their ongoing help and support. The authors would also like to thank the entire PSC team for all of their many contributions.

REFERENCES

[1] 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, Switzerland, 205–219.

[2] Paola A. Buitrago, Nicholas A. Nystrom, Rajarsi Gupta, and Joel Saltz. 2020. Delivering Scalable Deep Learning to Research with Bridges-AI. In *High Performance Computing: 6th Latin American Conference, CARLA 2019: Turrialba, Costa Rica, September 25–27, 2019: Revised Selected Papers (Communications in Computer and Information Science, Vol. 1087)*, Juan Luis Crespo-Mariño and Esteban Meneses-Rojas (Eds.). Springer International Publishing, Switzerland, 200–214. https://doi.org/10.1007/978-3-030-41005-6_14

[3] Motivair Corporation. 2019. *High Capacity Coolant Distribution Unit.* Technical Report. https://www.motivaircorp.com/uploads/files/brochures/CDU%20brochure_2019.pdf

[4] SchedMD LLC. 2020. *Generic Resource (GRES) Scheduling.* https://slurm.schedmd.com/gres.html

[5] Red Hat, Inc. 2021. *Red Hat Customer Portal.* https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/7/html/networking_guide/configuring-policy-based-routing-to-define-alternative-routes