Neocortex and *Bridges-2*: A High Performance AI+HPC Ecosystem for Science, Discovery, and Societal Good

Paola A. Buitrago and Nicholas A. Nystrom

Pittsburgh Supercomputing Center, Carnegie Mellon University,
Pittsburgh, PA 15213, United States
{paola}@psc.edu

Abstract. Artificial intelligence (AI) is transforming research through analysis of massive datasets and accelerating simulations by factors of up to a billion. Such acceleration eclipses the speedups that were made possible through improvements in CPU process and design and other kinds of algorithmic advances. It sets the stage for a new era of discovery in which previously intractable challenges will become surmountable, with applications in fields such as discovering the causes of cancer and rare diseases, developing effective, affordable drugs, improving food sustainability, developing detailed understanding of environmental factors to support protection of biodiversity, and developing alternative energy sources as a step toward reversing climate change. To succeed, the research community requires a high-performance computational ecosystem that seamlessly and efficiently brings together scalable AI, general-purpose computing, and large-scale data management. The authors, at the Pittsburgh Supercomputing Center (PSC), launched a second-generation computational ecosystem to enable AI-enabled research, bringing together carefully designed systems and groundbreaking technologies to provide at no cost a uniquely capable platform to the research community. It consists of two major systems: Neocortex and Bridges-2. Neocortex embodies a revolutionary processor architecture to vastly shorten the time required for deep learning training, foster greater integration of artificial deep learning with scientific workflows, and accelerate graph analytics. Bridges-2 integrates additional scalable AI, high-performance computing (HPC), and high-performance parallel file systems for simulation, data pre- and post-processing, visualization, and Big Data as a Service. Neocortex and Bridges-2 are integrated to form a tightly coupled and highly flexible ecosystem for AI- and data-driven research.

Keywords: Computer architecture · Artificial intelligence · AI for Good · Deep learning · Big Data · High-performance computing.

1 Introduction

Scalable artificial intelligence (AI) is of vital importance for enabling research, yet computational resources to support developing accurate models have largely

been based on processor technologies developed for other kinds of applications, and infrastructure to support scaling has been implemented mostly in software, limiting its effectiveness and ease of use. This paper describes a new, ambitious computer architecture for supporting AI-enabled research that balances the most powerful processors ever built with high-performance computing and data infrastructure. The two systems—*Neocortex*, which vastly shortens the time required for deep learning training, and *Bridges-2*, which provides great capacity for the many facets of rapidly evolving research—are integrated into a computational ecosystem to enable research in AI and its applications across all fields of study. They are being deployed at the Pittsburgh Supercomputing Center (PSC), a joint research center of Carnegie Mellon University and the University of Pittsburgh.

In 2012, the artificial neural network AlexNet [11] demonstrated the power of deep neural networks (DNNs) by dramatically decreasing the error rate in image classification and surpassing other machine learning (ML) approaches by 10.8% in the 2012 ImageNet competition. AlexNet achieved a top-5 error rate of 15.3%, with human-level accuracy being 5.1%. AlexNet consists of 8 network layers and 62,378,344 parameters, and it requires 7.25×10^8 flops. It took over five days to train on two NVIDIA GTX 580 GPUs.

The AlexNet result was significant because it convincingly demonstrated the ability of deep neural networks to automatically learn representations. AlexNet surpassed decades of traditional machine learning based on explicit feature engineering and other statistics. Inspired by AlexNet, researchers began developing more deeper, more sophisticated networks with progressively better results. Concurrently, domain scientists started applying the networks being created – and creating their own – to challenging problems in medical imaging, weather, cosmology, and many other fields.

In 2015, a new network, ResNet-152 [8], achieved top-5 error rate of only 4.49%, surpassing human-level accuracy. What changed were that ResNet-152 is an example of a residual network, and it is extremely deep: 152 layers. It has 60,192,872 parameters and requires 1.13×10^{10} flops, over 15 times that for AlexNet. This pattern is repeated across image classification and segmentation, time series analysis, natural language processing, and other fields to which deep learning is applied with great degrees of success: deeper, more complex networks better learn representations and result in higher accuracy. Neural networks for time series analysis and natural language processing (NLP) require recurrence and are much larger, for example, 330 million parameters for BERT [6] and 8.3 billion parameters for Megatron-LM [16]. In 2020, the GPT-3 language model presented another example of larger models yielding more accurate inferences. GPT-3 has 175 billion parameters and required 3.14×10^{23} flops (10 petaflop-years) to train [1]. *Training time is the primary bottleneck in applying AI to research, and the increasing complexity of deep learning models amplifies exacerbates the time required for training.*

Concurrently, researchers have begun to apply deep learning to a wide range of fields in science and engineering with remarkable results. For example, Kasim

et al. demonstrated speedups of 100,000 to 2,000,000,000 for a variety of applications including inertial confinement fusion (ICF), a global ocean biogeochemical model (MOPS), and a global aerosol-climate model (GCM) using Deep Emulator Network SEarch (DENSE) to develop and train neural network models [9]. The models are then used as emulators, i.e., as surrogates that replace computationally demanding calculations with much faster inferencing. Using a different approach, Smith et al. demonstrated billion-fold speedup in quantum chemistry with neural network potentials and transfer learning while approaching gold-standard accuracy of CCSD(T)/CBS calculations [17]. In large-scale data analytics, Khan et al. developed a neural network classifier for galaxies in the Dark Energy Survey (DES) that achieves state-of-the-art accuracy of 99.6% and also showed how it can be combined with unsupervised recursive training to prepare for extremely large sky surveys such as will be obtained from the Large Synoptic Survey Telescope (LSST) project [10].

The benefits of high-accuracy models are great. Such models can be applied to analyze and extract information from large datasets and to create surrogate models that substitute for expensive calculations in simulation codes to decrease time-to-solution by orders of magnitude without loss of accuracy. But first, the models must be trained.

Training deep neural networks often takes days, weeks, or even months. For some applications such as image segmentation in radiology, there already exist deep neural networks that are known to work reasonably well. For many other applications, developing a model first requires building and optimizing a neural network architecture. Different types of networks better suited to different types of applications, and the field is evolving rapidly, with new network types frequently emerging. Once a network architecture is selected, and also to choose between network architectures, hyperparameters must be optimized, requiring additional sets of runs. The time requirement can be prohibitive. *It is this challenge that Neocortex is designed to overcome.*

The following sections describe a unique, heterogeneous system architecture for scalable AI, data pre- and post-processing, and simulation. Section 2 summarizes related work. For context, section 3 provides an overview of the integrated system. Sections 4 and 5 then describe the Neocortex and Bridges-2 architectures, respectively. Section 6 concludes with a summary of the ecosystem's novel capabilities and expected opportunities.

2 Related Work

The heterogeneous architecture of Bridges-2 is an evolution of the *Bridges* system [13, 14], which pioneered the convergence of HPC, AI, and Big Data. Bridges, which was designed in early 2014 and entered production in April 2016, tightly integrated dual-socket CPU nodes, large-memory four- and sixteen-socket CPU nodes, GPU nodes, and a parallel, disk-based file system with an overarching interconnect fabric. Bridges enabled complex workflows running concurrently on different kinds of compute nodes for which individual components were best-

suited. Dedicated nodes containing solid-state disks (SSDs) for high IOPs and hard disk drives (HDDs) for large capacity supported persistent databases and web portals for different kinds of research ("science gateways"). Bridges was the world's first deployment of the Intel Omni-Path Architecture (OPA) fabric.

In November 2018, the authors developed and deployed Bridges-AI [4] as an expansion to Bridges. Bridges-AI consists of two types of AI-optimized nodes: an NVIDIA DGX-2 enterprise AI research system and nine Hewlett Packard Enterprise (HPE) Apollo 6000 Gen10 servers. The DGX-2 contains sixteen NVIDIA Tesla V100 GPUs with 32GB of HBM2 memory (aggregate 512 GB HBM2), interconnected by the NVSwitch at 2.4 TB/s bisection bandwidth, 30TB NVMe of SSD, two Intel Xeon Platinum 8168 CPUs, and 1.5 TB of CPU memory. Its 10,240 tensor cores deliver 2 Pf/s of performance. Until recently, the DGX-2 was the world's most powerful AI system. The nine Apollo 6000 servers each have eight V100 GPUs with 16 GB of HBM2 memory, 7.68 TB NVMe SSD, two Intel Xeon Gold 6148 CPUs, and 192 GB of CPU memory. They provide additional substantial capacity for deep learning training for models and data that don't require the DGX-2. When Bridges-AI entered production in January 2019, it expanded the aggregate AI capacity of the NSF XSEDE ecosystem by 300%.

The optimization of advanced cyberinfrastructure for AI research is highly complex due to the rapid advance of hardware and software technologies and the differences between models that are important for social networks and business versus models that address the very large images, volumes, time series, and multimodal data of research applications. The Open Compass [2] project aims to evaluate the potential of new AI technologies for research, going beyond standard benchmarks such as MLPerf to also evaluate representative research applications, and developing and sharing best practices.

As more is learned, there exists the potential to apply AI to improve the design of large-scale computer systems and specific workloads. Concurrently, AI can be applied to increase supercomputers' performance, reliability, and usability and to improve user experience. This is the subject of one of the authors' (Buitrago's) *Calima* project, and it is addressed in the report of the NSF Workshop on Smart Cyberinfrastructure [3].

3 Integrated Neocortex + Bridges-2 AI+HPC Ecosystem

Neocortex and Bridges-2, which are detailed in the following sections, are being integrated with each other, Bridges-AI [4], and wide-area networks to national and international cyberinfrastructure, instruments, campuses, and clouds. Figure 1 illustrates the computational and data components and bandwidths of the combined system.

From a hardware architecture perspective, the goals are *capability*, *performance*, and *efficiency*. Capability arises from processing nodes that are separately specialized for different components of research workflows and that have unified access to high-performance data storage. Performance arises from node architectures that are individually optimized for deep learning and other ma-

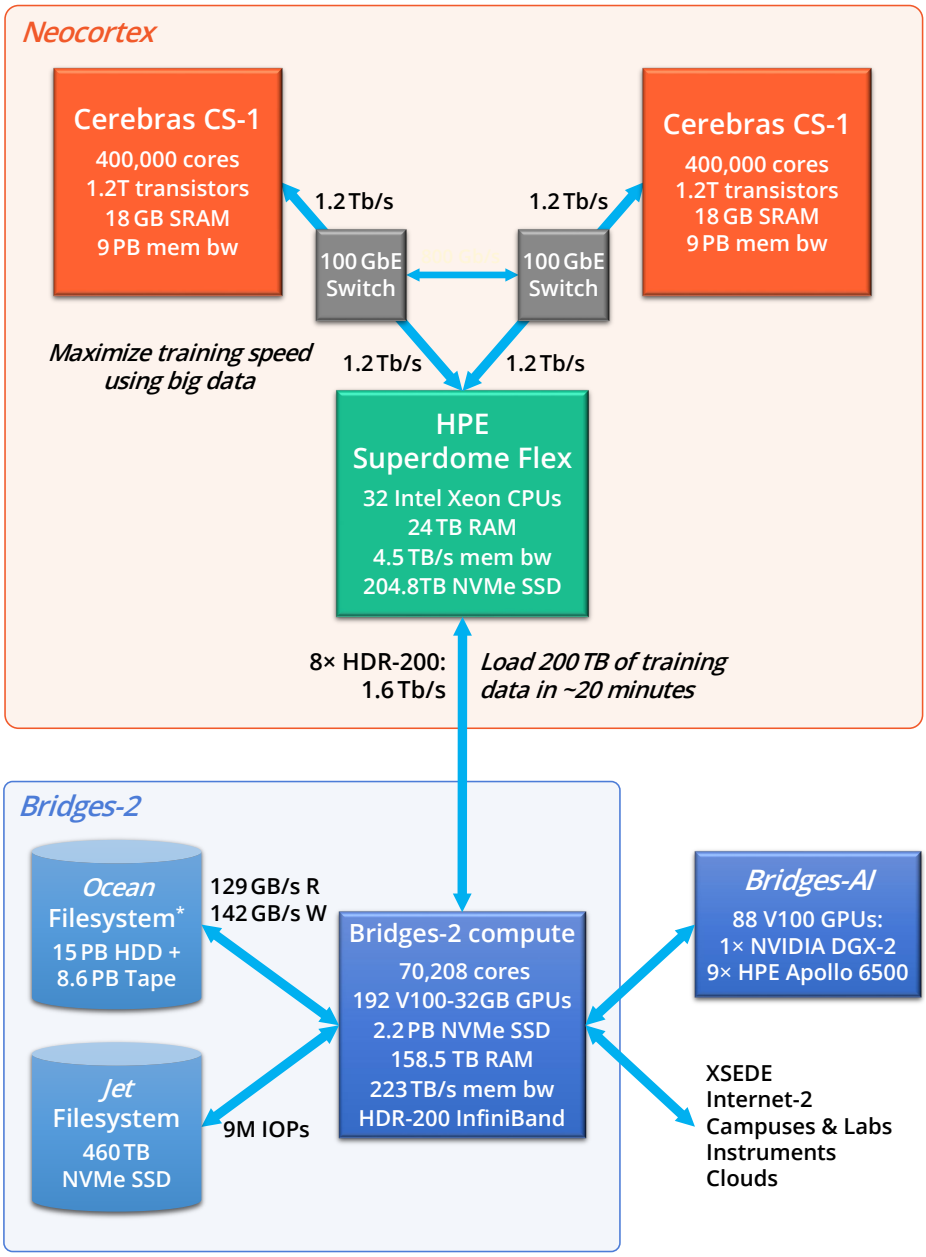

Fig. 1. High-level architecture of the Neocortex and Bridges-2 ecosystem for AI, HPC, and data. Bandwidths are balanced to enable efficient access to data and rapidly staging large-scale data from Bridges-2's Ocean file system to Neocortex's local NVMe flash file system. This facilitates training on Neocortex and doing pre- and post-processing on Bridges-2, as well as equitable access to Neocortex for a large number of users and research projects.

chine learning, high performance computing, and large-memory tasks. Efficiency arises from balanced bandwidth across the various data paths within the system.

A key metric for the combined system is efficiently transferring data from Bridges-2 to Neocortex, for which loading 200 TB of training data into Neocortex from Bridges-2 can be achieved in approximately 20 minutes, assuming that the data is well-distributed in Bridges-2's large (15 PB) disk-based Ocean file system, resident in its flash-based Jet file system, or resident in RAM.

4 Neocortex

In early summer 2020, an innovative and unprecedented AI supercomputer, Neocortex, was awarded by the National Science Foundation. Neocortex, which captures groundbreaking new hardware technologies, is designed to accelerate AI research in pursuit of science, discovery, and societal good.

Neocortex is a highly innovative resource designed to accelerate AI-powered scientific discovery by vastly shortening the time required for deep learning training, foster greater integration of artificial deep learning with scientific workflows, and provide revolutionary new hardware for the development of more efficient algorithms for artificial intelligence and graph analytics. Its scale democratizes access to game-changing compute power otherwise only available to tech giants, allowing students, postdocs, faculty, and other researchers who require faster turnaround on training to analyze data and integrate AI with simulations. A primary goal of Neocortex is to inspire the research community to tackle big ideas, no longer constrained by computational resources, and scale their AI-based research and integrate AI advances into their research workflows. Neocortex allows users to apply more accurate models and train on larger data. It also allows scaling model parallelism to unprecedented levels, avoiding the need for expensive and time-consuming hyperparameter optimization.

Neocortex is designed to enable three exciting areas of research. First, the WSE takes processor architecture to an unprecedented scale. Providing the research community with access to that unique and remarkable capability is vital to understand the potential of the WSE approach. Second, as powerful as the WSE is, there are models too large for one WSE. Neocortex uniquely couples two CS-1 systems using a large-memory "front end" to enable research into scaling across multiple WSEs. Third, Neocortex is designed to enable important research for societal good. Examples include discovering the fundamental causes of rare diseases and providing insights into treatments, revealing the low-level mechanisms of cancer to improve understanding of its causes and progression despite its complexity, and improving crops' resistance to climate change to alleviate world hunger.

4.1 Neocortex Overview

Neocortex couples two Cerebras CS-1 AI servers with a large shared memory HPE Superdome Flex HPC server to achieve unprecedented AI scalability with

excellent system balance. Each Cerebras CS-1 is powered by one Cerebras Wafer Scale Engine (WSE) processor, a revolutionary high-performance processor designed specifically to accelerate deep learning training and inferencing [12]. The Cerebras WSE is the largest chip ever built, containing 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 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 common machine learning frameworks such as TensorFlow and PyTorch, which for computational efficiency are mapped onto an optimized graph representation and a set of model-specific computation kernels. The CS-1 also supports native code development. Support for the most popular deep learning frameworks and automatic, transparent acceleration will researchers with ease of use. Table 1 summarizes the architectural characteristics of the subsystems of Neocortex.

Table 1. Neocortex architectural characteristics. Each of the two Cerebras CS-1 systems features a Cerebras Wafer Scale Engine (WSE) processor.

Cerebras CS-1

AI Processor	Cerebras Wafer Scale Engine (WSE) 400,000 Sparse Linear Algebra Compute (SLAC) cores 1.2 trillion transistors 46,225 mm^2 18 GB SRAM on-chip memory 9.6 PB/s memory bandwidth 100 Pb/s interconnect bandwidth
System I/O	1.2 Tb/s (12 × 100 GbE interfaces)

HPE Superdome Flex

CPUs	32 × Intel Xeon Platinum 8280
Memory	24 TiB RAM, aggregate bandwidth 4.5 TB/s
Data storage	32 × 6.4 TB NVMe SSDs 204.6 TB aggregate 150 GB/s read bandwidth
Network to CS-1 systems	24 × 100 GbE interfaces 1.2 Tb/s (150 GB/s) to each CS-1 2.4 Tb/s aggregate
Network to Bridges-2	16 × HDR-100 InfiniBand 1.6 Tb/s aggregate

The two Cerebras CS-1 systems and the HPE Superdome Flex are balanced to allow running the CS-1 systems concurrently on different models or together on a single model. This includes the bandwidth of the NVMe SSD file system in

Neocortex, the bandwidth to each CS-1, and the even higher RAM bandwidth of Superdome Flex.

4.2 Cerebras CS-1 and Wafer Scale Engine

The Cerebras CS-1 is first available system featuring the Cerebras Wafer Scale Engine (WSE) processor, which is the largest chip ever built. Fabricated using a whole silicon wafer, the Cerebras WSE measures 46,225 [2] and contains 400,000 AI-optimized cores and 1.2 trillion transistors. It includes an on-chip 100 Pb/s fabric as a fully configurable 2D mesh with no software overhead. 18 GB of SRAM provides memory latency of only one clock and memory bandwidth of 9.6 PB/s. The Cerebras CS-1 contains one WSE processor, twelve 100 GbE ports, twelve 3 kW power supplies, and self-contained water cooling in a 15U enclosure.

The matrix and vector values of deep neural networks are mostly zeros, which arises from operations such as ReLU (rectified linear unit; 90% natural sparsity) and dropout (30% natural sparsity). For example, Transformer has 50-98% zeros [7]. The inherent sparsity of deep neural networks is not aligned with GPUs and CPUs, the memory subsystems of which have been designed to maximize the efficiency of dense operations. For networks with high sparsity, there is little to no cache reuse. This mismatch manifests as low performance resulting from the high latency incurred when fetching non-sequential data from memory or other processors, potentially across a PCI Express bus. The latency for remote fetches, i.e., at least a microsecond, is at least three orders of magnitude greater than accessing data that is already in cache, only a few clocks away, i.e., on the order of a nanosecond.

The Cerebras WSE overcomes the latency barrier through mutually reinforcing architectural advances in on-chip memory, in-processor communications, optimized compute cores, and software. These synergistic advances overcome the latency barrier by making memory accesses local and explicitly addressing sparsity.

The 400,000 Sparse Linear Algebra (SLA) cores of the WSE are optimized for deep learning. They contain no caches or other unnecessary features that would introduce overhead. The SLA cores are fully programmable, supporting arithmetic, logical operations, load/store, and branching, and they implement optimized tensor operations specific to deep learning. The SLA cores are engineered to exploit sparsity, containing fine-grained dataflow scheduling through which compute is triggered by data. Multiples are performed only for non-zero operands. Both fine- and coarse-grained sparsity are supported to accommodate activations and weights being zero at both the individual and block levels [5].

The WSE includes 18 GB of on-chip SRAM (static RAM), yielding 9 PB/s of memory bandwidth and a latency of only one clock cycle. The distribution of SRAM across the wafer supports sparsity to run all SLA cores at full speed [12].

The Cerebras Swarm communication fabric interconnects the 400,000 cores on the WSE. It is a flexible, all-hardware, 2D mesh that delivers 100 Pb/s of bandwidth, hardware routing, and single-word active messages. Link latency and energy cost are extremely low. The Swarm fabric is fully reconfigurable,

allowing optimized communication paths to be implemented for each model, avoiding overheads and improving power efficiency [12].

The Cerebras software stack abstracts the WSE's sophisticated features to allow translation from models expressed in TensorFlow and PyTorch to highly efficient implementations on the WSE. The Cerebras Graph Compiler builds a dataflow representation from the user's model, mapping it onto an intermediate representation and optimized low-level kernels. A place-and-route step maps the model onto the WSE, creating a datapath that is optimized for locality and communications [18].

This hardware and software co-design enables great efficiency and new approaches to model parallelism. For example, by placing an entire network on the WSE at once, data can be streamed through a multi-stage pipeline, effectively running all layers simultaneously.

4.3 HPE Superdome Flex

The HPE Superdome Flex system is a high-end, modular, shared-memory server engineered for mission-critical AI and HPC workloads. For Neocortex, a large Superdome Flex was selected as the most powerful, user-friendly front-end for the two Cerebras CS-1 systems. The scalability of the Superdome Flex allows it to be robustly provisioned to drive the CS-1 systems independently or together. The Superdome Flex builds on experience with large shared-memory servers, which have been observed to support scaling with high ease of use (e.g., Blacklight [15]).

The Superdome Flex in Neocortex consists of 8 chassis connected by an internal interconnect to create a single-system image (SSI) spanning 32 high-end CPUs, 24 TB of hardware cache-coherent shared memory, 204.8 TB (raw) of high-bandwidth NVMe PCIe flash storage, 24 100 GbE ports, and 16 HDR-100 InfiniBand ports. The full 24 TB of RAM is cache-coherent across all 32 CPUs, supported by HPE Superdome Flex ASICs with coherency unit of one cache line (64 bytes). The internal Superdome crossbar interconnect, supported by two HPE Superdome Flex ASICs in each chassis, supports 850 GB/s of bisection bandwidth. The single-system image lets users quickly and conveniently train on their data without having manually to distribute it across a cluster of servers, saving them time and avoiding load imbalance to maximize efficiency.

The Superdome Flex is fully populated with 32 Intel Xeon Platinum 8280 CPUs, which have 28 cores, 56 hardware threads, base and maximum turbo frequencies of 2.70 GHz and 4.00 GHz, respectively, 38.5 MB of cache, and 3 UPI links. The 24 TB memory is comprised of of 192×128 GiB DDR4-2933 RDIMMs, with aggregate memory bandwidth of 4.5 TB/s.

Local storage consists of 32 NVMe 6.4 TB PCIe flash cards, for 204.6 TB raw capacity and 150 GB/s read bandwidth, matching the 150 GB/s network connection to a Cerebras CS-1. The local storage is managed by HPE Data Management Framework (DMF) for user-friendly, efficient data transfer from Bridges-2 over InfiniBand at up to 1.6 Tb/s.

Twenty-four 100 GbE network interface cards (NICs) provide 2.4 Tb/s of Ethernet connectivity, with 1.2 Tb/s (150 Gb/s) to each of the two Cerebras CS-1

systems in Neocortex. Sixteen HDR InfiniBand host channel adapters (HCAs), mounted on sixteen PCI Express Gen 3 ×16 ports, connect to Bridges-2's HDR InfiniBand fabric at 1.6 Tb/s.

The HPE Superdome Flex ASIC differentiates the Superdome Flex from other servers by providing cache-coherent shared memory spanning 32 CPUs. For Neocortex, the SD Flex's 24 TB of cache-coherent shared memory backed by over 200 TB of high-bandwidth NVMe flash storage ease training on very large datasets, avoiding the laborious task of splitting datasets across worker nodes and possibly generating load imbalances.

4.4 Neocortex Interconnect

Each Cerebras CS-1 is connected to the HPE Superdome Flex by twelve 100 Gb/s Ethernet ports, for aggregate 1.2 Tb/s (150 GB/s) from the Superdome Flex to each CS-1 and 2.4 Tb/s (300 GB/s) combined. Each of the Mellanox SN3700cM 32-port switches has eight ports remaining, which are interconnected between the switches to enable research involving communications directly between the two Cerebras CS-1 systems.

4.5 Neocortex Software

The Cerebras Software Stack [18] translates models from widely used frameworks such as TensorFlow and PyTorch to executables for the Cerebras CS-1, as summarized above. Neocortex's Superdome Flex runs the CentOS 8 operating system and is configured with containers, frameworks, libraries, and tools to support the Cerebras CS-1.

5 Bridges-2

Bridges-2 builds on, improves, and extends concepts proven in Bridges [13] to take the next step in pioneering converged, scalable HPC, AI, and data; prioritize researcher productivity and ease of use; and provide an extensible architecture for interoperation with complementary data-intensive projects, campus resources, and clouds. Funded by the National Science Foundation, Bridges-2 is a "capacity" resource, designed to enable rapidly evolving research and an extremely wide range of applications.

Bridges-2 contains 566 nodes, 70,208 CPU cores, and 192 GPUs. Its peak floating-point rates are 5.175 Pf/s fp64 and 24 Pf/s mixed-precision/tensor. It contains 158.5 TiB of memory with 223.4 TiB/s of memory bandwidth, 2.2 PB of node-local NVMe SSD, 15 PB (usable) disk in a high-performance Lustre file system, and 8.6 PB tape (estimated, assuming 20% compression). High bandwidth for efficient data movement was prioritized over raw flops. Figure 2 illustrates the high-level architecture of Bridges-2.

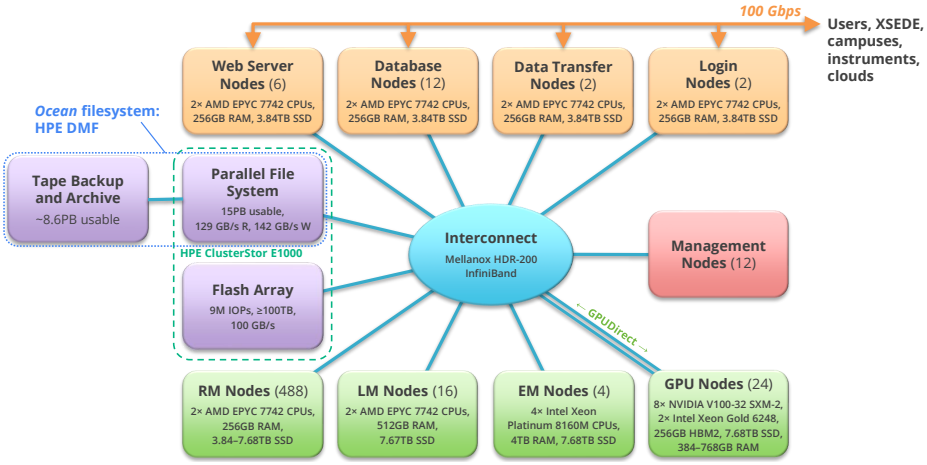

Fig. 2. Bridges-2 consists of four types of compute nodes—Regular Memory (RM), Large Memory (LM), Extreme Memory (EM), and Graphics Processing Unit (GPU)—interconnected with each other, file systems, and utility and management nodes by a high-performance fabric. Persistent data is maintained in the hierarchical Ocean file system. Data requiring high IOPs, such as for deep learning training, is cached to the Jet flash file system. Utility nodes serve persistent databases and distributed (web) services, data transfer (100 Gbps), and logins. Management nodes serve system configuration management, scheduling, logging, and other administrative functions.

5.1 Innovations

Bridges-2 introduces six important innovations beyond Bridges, in addition to greatly improving all aspects of system performance. These innovations, which reflect the evolution of research applications, are as follows:

- **An all-flash filesystem**, *Jet*, provides 9 IOPs (measured on 4 kB reads) of random-access I/O performance to support deep learning training on data that is much larger than node-local storage capacity. Jet has 460.8 TB of capacity (raw) and supports at least 100 GB/s of read/write bandwidth.
- **Enhanced GPU nodes** amplify scalable deep learning. GPU nodes each have eight NVIDIA Tesla V100-32GB SXM2 GPUs (aggregate 256 GB HBM2 memory per node), up to 768 GB of CPU memory, and dual-rail Mellanox HDR-200 InfiniBand (IB) between GPU nodes.
- **Full-system HDR-200 InfiniBand** doubles link bandwidth relative to Bridges and provides 200M messages/s injection rate, and $< 1\mu$s latency, and numerous advanced features for performance, flexibility, and to scale GPU applications, including GPUDirect RDMA communications between GPUs on different nodes.
- **AMD EPYC 7742 ("Rome") CPUs** support PCI Express Gen 4 (31.5 GB/s for 16 lanes), enabling full use of HDR-200 InfiniBand. They also yield excellent performance with 64 cores each.

- **Bridges-2 supports full-system AI.** Its 24 GPU nodes (192 NVIDIA Tesla V100-32GB SXM2 GPUs) provide high scalability and capacity for deep learning training, and its AMD EPYC 7742 CPUs have ample cores (64) for high-performance inferencing, including coupling of surrogate models with simulations. The unified architecture also allows for online training.
- **A hierarchical storage system** provides project storage (disk) and expandable archive and disaster recovery storage (tape), using HPE DMF to expose a single name space with rule-based replication and migration.

5.2 Compute and Utility Nodes

Bridges-2 contains four types of compute nodes:

- **488 Regular Memory (RM) nodes** each have 2 AMD EPYC 7742 ("Rome") CPUs, 256 GB of DDR4-3200 memory, 3.84–7.68 TB NVMe SSD local storage, and 1 HDR-200 IB adapter. RM nodes are HPE Apollo Gen10 plus chassis containing HPE ProLiant XL225n Gen10 plus Servers. RM nodes are used for HPC, data analytics and pre- and post-processing, and other general-purpose computing ranging from 1 core to 61k cores. HPC jobs can be run across all 62,464 (61k) cores of RM nodes.
- **16 Large Memory (LM) nodes** are similar to RM nodes, differing only in containing twice the memory (512 GB) and 7.68 TB NVME SSD. LM nodes are used for genomics and tasks similar to those for RM nodes but that need more memory. Large-memory HPC jobs can be run across all 2,048 (2k) cores of EM nodes, and especially demanding HPC jobs can be run across all 64,512 (63k) cores of combined RM and EM nodes.
- **4 Extreme Memory (EM)** nodes each have 2 Intel Xeon Platinum 8260M ("Cascade Lake") CPUs, 4 TB of DDR4-2933 memory, 7.68 TB NVMe SSD, and 1 HDR-200 IB adapter. EM nodes are HPE ProLiant DL560 Gen10 servers. EM nodes are used for genome sequence assembly and other tasks that require large shared memory.
- **24 Graphics Processing Unit (GPU) nodes** each have 8 NVIDIA Tesla V100-32GB SXM-2 GPUs (aggregate 256 GB HBM2 memory), 2 Intel Xeon Gold 6248 ("Cascade Lake") CPUs, 384–768 GB of DDR4-2933 memory, 7.68 TB NVMe SSD, and 2 HDR-200 IB adapters. GPU nodes are HPE Apollo 6500 Gen10 servers. GPU nodes are used for deep learning, other machine learning, visualization, and accelerated simulation. Preference is given to the 768 GB GPU nodes for deep learning training.

Bridges-2 utility nodes are identical to RM nodes but dedicated to specific purposes (i.e., not available for routine scheduling via Slurm). Of the 22 utility nodes, 6 are dedicated to serving web portals (for example, domain-specific "Science Gateways") that provide HPC, Big Data, and Software as a Service, 12 are dedicated to serving persistent databases to power workflows and web portals, 2 are Data Transfer Nodes for high-bandwidth transfers from and to wide-area networks, and 2 are login nodes. Services and databases running on web server

and database nodes are typically isolated in virtual machines and potentially also containerized. If additional web or database nodes come to be needed, RM nodes can be repurposed accordingly.

5.3 File Systems

Bridges-2 supports four file systems: *Ocean*, *Jet*, local, and memory.

The **Ocean** file system is hierarchical, providing user-friendly, seamless management of disk and tape subsystems in a single name space using the HPE Data Management Framework (DMF). The disk component of Ocean is an HPE ClusterStor E1000 storage system, with 15 PB of usable capacity (21 PB raw) and 129 GB/s and 142 GB/s read and write bandwidth, respectively. It runs Lustre, for which 10 data server pairs each serve 2.1 PB (raw) capacity. The tape component of Ocean is an HPE StoreEver MSL6480 Tape Library, initially populated with 5 modules (scalable to 7), where each module holds 80 LTO-8 Type M tape cartridges. Its raw capacity is 7.2 PB. Based on historical data, approximately 20% compression is expected, which occurs at line speed, increasing effective capacity to approximately 8.6 PB. Bandwidth is 50 TB/hour. The tape subsystem is expected to be used for archiving and disaster recovery (DR), and it is expandable, should the need and external support arise, to serve specific projects requiring great amounts of archive/DR capacity.

The **Jet** file system uses NVMe flash storage devices to provide 9M IOPs, at least 100 GB/s of read/write bandwidth, and 460.8 TB of raw capacity. The Jet file system is used to cache moderately large data for which high bandwidth is needed, for example, deep learning training.

Local and memory filesystems exploit NVMe SSD and RAM, respectively, on each compute node, which can substantially increase bandwidth for deep learning training, scratch files, and other ephemeral storage requirements.

5.4 Interconnect

A Mellanox HDR-200 InfiniBand fabric provides high communications performance both between compute nodes (for HPC jobs) and to and from Bridges-2's file systems. It is configured in a leaf-spine topology with 12 spine switches and 26 leaf switches, which cost-effectively supplies ample bandwidth for Bridges-2 diverse workload. The oversubscription is 2.3:1. Dual-rail HDR-200 (400 Gb/s) is used to interconnect Bridges-2's GPU nodes, doubling the inter-node bandwidth to more effectively scale deep learning training across nodes.

5.5 User Environment

The Bridges-2 user environment supports an extremely wide range of applications, libraries, and frameworks. Bridges-2 supports Singularity for containerized applications, including NVIDIA GPU Cloud containers. Conversion from Docker containers is typically straightforward. Both batch and interactive access are

supported. System resources are managed by Slurm, and a user-friendly *interact* command is implemented to obtain immediate access to resources ranging from a single core to multiple nodes. Interactivity has proven invaluable on Bridges for analytics, development, debugging, and visualization, and it has been possible to provision resources for interactive use with very low impact on overall utilization.

6 Summary

Neocortex and Bridges-2 form a unique computational ecosystem for scalable AI, data processing, analytics, and management, and high-performance simulation. Their design was strongly influenced by consideration of applications across diverse fields of research, especially for societal good. The innovations that differentiate this ecosystem are great innovation hardware architecture, fully integrated heterogeneous node types to optimally support components of research workflows, and a unified data management system consisting of in-processor memory, conventional memory, flash, disk, and tape layers. Specifically, Neocortex introduces the Cerebras Wafer Scale Engine, the largest processor ever built, to the open research community to accelerate deep learning training by orders of magnitude, potentially to interactive rates, and it couples two Cerebras CS-1 systems through a very large memory HPE Superdome Flex "front end" to explore scaling models to multiple CS-1 systems. Bridges-2 provides high capacity for data pre- and post-processing, other types of machine learning, simulation, and large-scale data management, and archiving through integration of multiple nodes types and hierarchical data storage using a high-performance 200 Gb/s fabric, with 400 Gb/s between its GPU-accelerated AI nodes, also to support scalable deep learning. Both systems are available at no cost for open research.

Acknowledgments

Thanks to Natalia Vassilieva for collaboration on the Cerebras CS-1. The Bridges system, including Bridges-AI, is supported by NSF award number 1445606. The Bridges-2 system is supported by NSF award number 1928147. The Neocortex system is supported by NSF award number 2005597. The Open Compass project is supported by NSF award number 1833317.

References

1. Brown, T.B., et al.: Language models are few-shot learners (2020)
2. Buitrago, P.A., Nystrom, N.A.: Open Compass: Accelerating the adoption of AI in open research. In: Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (Learning). PEARC '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3332186.3332253

3. Buitrago, P.A., Nystrom, N.A.: Strengthening the adoption of AI in research and cyberinfrastructure. In: Pascucci, V., et al. (eds.) Report from the NSF Workshop on Smart Cyberinfrastructure 2020. Alexandria, Virginia (February 2020)
4. Buitrago, P.A., Nystrom, N.A., Gupta, R., Saltz, J.: Delivering scalable deep learning to research with Bridges-AI. In: Crespo-Mariño, J.L., Meneses-Rojas, E. (eds.) 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, pp. 200–214. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-41005-6_14
5. Cerebras Systems: Cerebras wafer scale engine: An introduction (2019)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding (2018)
7. Gale, T., Elsen, E., Hooker, S.: The state of sparsity in deep neural networks (2019)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
9. Kasim, M.F., et al.: Up to two billion times acceleration of scientific simulations with deep neural architecture search (2020)
10. Khan, A., Huerta, E.A., Wang, S., Gruendl, R., Jennings, E., Zheng, H.: Deep learning at scale for the construction of galaxy catalogs in the Dark Energy Survey. Physics Letters B **795**, 248–258 (2019). https://doi.org/https://doi.org/10.1016/j.physletb.2019.06.009
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc. (2012)
12. Lie, S.: Wafer scale deep learning. In: Hot Chips 31 (2019)
13. Nystrom, N.A., Buitrago, P.A., Blood, P.D.: Bridges: Converging HPC, AI, and Big Data for enabling discovery. In: Vetter, J.S. (ed.) Contemporary High Performance Computing: From Petascale toward Exascale, Volume Three. Contemporary High Performance Computing, CRC Press, Boca Raton, FL (2019)
14. Nystrom, N.A., Levine, M.J., Roskies, R.Z., Scott, J.R.: Bridges: A uniquely flexible HPC resource for new communities and data analytics. In: Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure. XSEDE '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2792745.2792775
15. Nystrom, N.A., Welling, J., Blood, P.D., Goh, E.L.G.: Blacklight: Coherent shared memory for enabling science. In: Vetter, J.S. (ed.) Contemporary High Performance Computing: From Petascale toward Exascale. Contemporary High Performance Computing, Taylor & Francis Group, Boca Raton, FL (2013)
16. Shoeybi, M., et al.: Megatron-LM: Training multi-billion parameter language models using model parallelism (2019)
17. Smith, J.S., et al.: Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. Nature Communications **10**(1), 2903 (2019). https://doi.org/10.1038/s41467-019-10827-4
18. Vassilieva, N., Buitrago, P.A., Nystrom, N.A., Sanielevici, S.E.: Technical overview of the Cerebras CS-1, the AI compute engine for Neocortex (webinar) (2020), https://www.cmu.edu/psc/aibd/neocortex/technical-overview-webinar.html