WEKA | cerebras

SOLUTION BRIEF

Cerebras CS-2 With WEKA Data Platform for AI

Purpose-Built Deep Learning Delivering Performance at Unprecedented Speeds and Scale Through a Systems Approach

The Deep Learning Problem

Deep learning has emerged as one of the most important computational workloads of our generation. Its applications are widespread and growing. But deep learning is profoundly computationally intensive. Between 2015 and 2020, the compute used to train the largest models increased by 300,000x. In other words, AI compute demand is doubling every 3.5 months. Because of this voracious demand, AI is constrained by the availability of compute; not by applications or ideas. Testing a single new hypothesis—training a new model—can take weeks or months and can cost hundreds of thousands of dollars in compute time. This is a significant drag on the pace of innovation, and many important ideas are ignored simply because they take too long to test.

AI Insights in Minutes, Not Months

NThe CS-2 is the industry's fastest AI accelerator. It reduces training times from months to minutes, and inference latencies from milliseconds to microseconds. And the CS-2 requires only a fraction of the space and power of graphics processing unit-based AI compute.

The CS-2 features 850,000 AI optimized compute cores, 40GB of on-chip SRAM, 20 PB/s memory bandwidth and 220Pb/s interconnect, all enabled by purpose-built packaging, cooling, and power delivery. It is fed by 1.2 terabits of I/O across 12 100Gb Ethernet links. Every design choice has been made to accelerate deep learning, reducing training times and inference latencies by orders of magnitude.

Powered by the 2nd Generation Wafer-Scale Processor

The CS-2 is powered by the largest processor ever built—the industry's only 2.6 trillion transistor silicon device. The WSE2 is 56 times larger than the largest graphics processing unit and contains 123x more compute cores and 1,000x more high performance on chip memory delivering more than any other deep learning processor in existence.

Seamless Software Integration

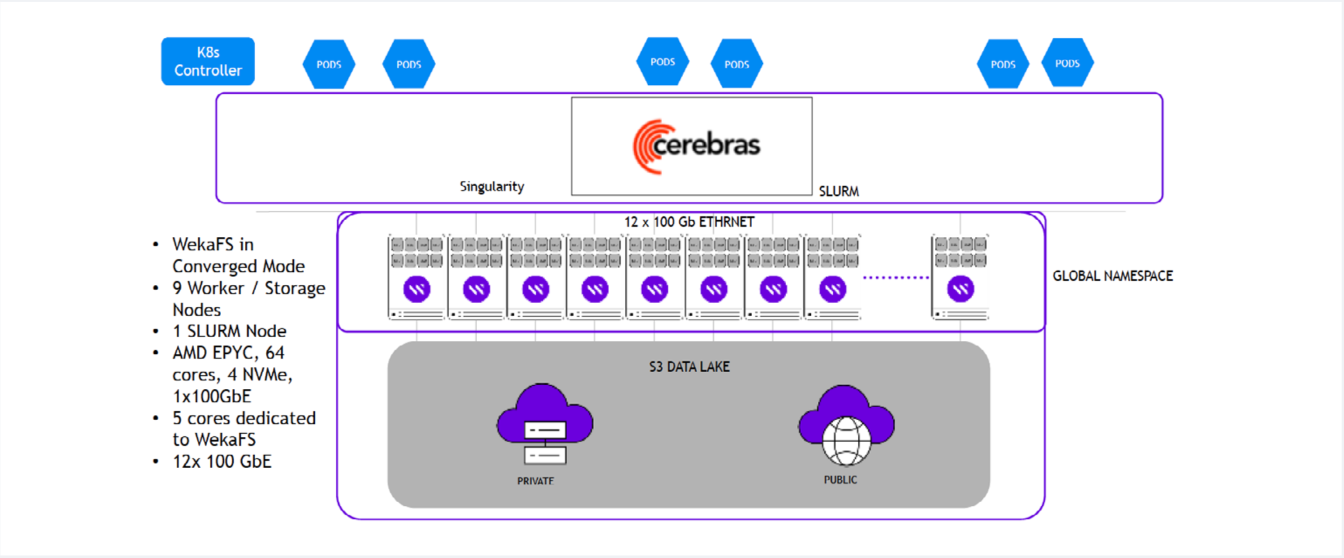
The Cerebras software platform integrates with popular machine learning frameworks like TensorFlow and PyTorch, so researchers can use familiar tools and rapidly bring their models to the CS-2. The platform is fully programmable and provides both an extensive library of primitives for standard deep learning computations, as well as a familiar C-like interface for developing custom kernels and applications.

Best of Breed Solution With Weka Data Platform for AI

Timely access to datasets has become the last mile problem for Deep Learning. WEKA's data platform for AI is built on the WEKAFS™ filesystem, to address the storage challenges posed by modern applications that leverage Cloud, AI accelerators and Deep Learning through innovations in how Flash is used. WEKA data platform offers the simplicity of NAS, the performance of SAN or DAS and the scale of object storage.

FIG. 1 CS-2 Wafer Scale Engine 2 Processor built by Cerebras.

AI accelerators can spend half their time waiting for data, that means you end up waiting for results. WEKA's data platform addresses the storage challenges posed by today's enterprise AI workloads running on-premises, in the cloud or bursting between platforms. With WEKA, you accelerate every stage of the data pipeline from data ingestion, to feature engineering to model training and inference, improving training times and reducing time to market.



- WekaFS in Converged Mode
- 9 Worker / Storage Nodes
- 1 SLURM Node
- AMD EPYC, 64 cores, 4 NVMe, 1x100GbE
- 5 cores dedicated to WekaFS
- 12x 100 GbE

FIG. 2 Example architecture implementation with Cerebras CS-2 and WEKA.

- Maximum performance using POSIX, S3, NFS, K8 CSI protocols with Singularity and SLURM
- Lead BERT, U-NET tfrrecords model training benchmarks over local NVMe
- Converged mode where ETL processing is run on same load / storage servers
  - Eliminates additional tiers, extend capacity by attaching S3 bucket to same namespace
- Best Economics
  - Better Storage efficiency than RAID 0 striping; configurable data protection to 16+4
  - NVMe Flash performance tier with S3/HDD bucket as capacity Tier
- Ease of Use at Scale
  - No need to manually shard data with local NVMe storage
  - Better data management with snap2object, scale to Exabytes with billion files in a directory and cloud bursting as needed

Conclusion

Cerebras Systems and WEKA Data Platform for AI have come together to build a new class of systems to accelerate artificial intelligence work. This thinking is pervasive in our ethos and manifests in our designs. The CS-2 achieves best-in-industry performance through technical innovation across software, chip design and system hardware. The performance and scale unlock entirely new classes of models, learning algorithms, and researcher opportunities. WEKA Data Platform for AI delivers high performance, with the best economics and software defined data management. All aspects of the solution work in concert to deliver unprecedented AI performance and ease-of-use.

Please reach out to info@cerebras.net or info@weka.io for more information.