Cerebras Wafer-Scale Cluster

Exa-scale performance, single device simplicity

Accelerate AI Model Training

The Cerebras Wafer-Scale Cluster (WSC) is a revolutionary technology suite that efficiently handles the enormous computational needs of AI model training. It centers around the CS-3 system, powered by the 3rd generation Wafer-Scale Engine (WSE-3)—the world's largest AI-optimized processor. The WSC integrates MemoryX for high-capacity, off-chip model weight storage, and SwarmX for effective weight broadcasting and gradient reduction across the cluster. This setup allows the WSC to adeptly train multi-trillion parameter models, achieving near-perfect linear-scaled performance and simplifying the complexity seen in traditional distributed computing.

Powered by the 3rd Generation Wafer-Scale Engine

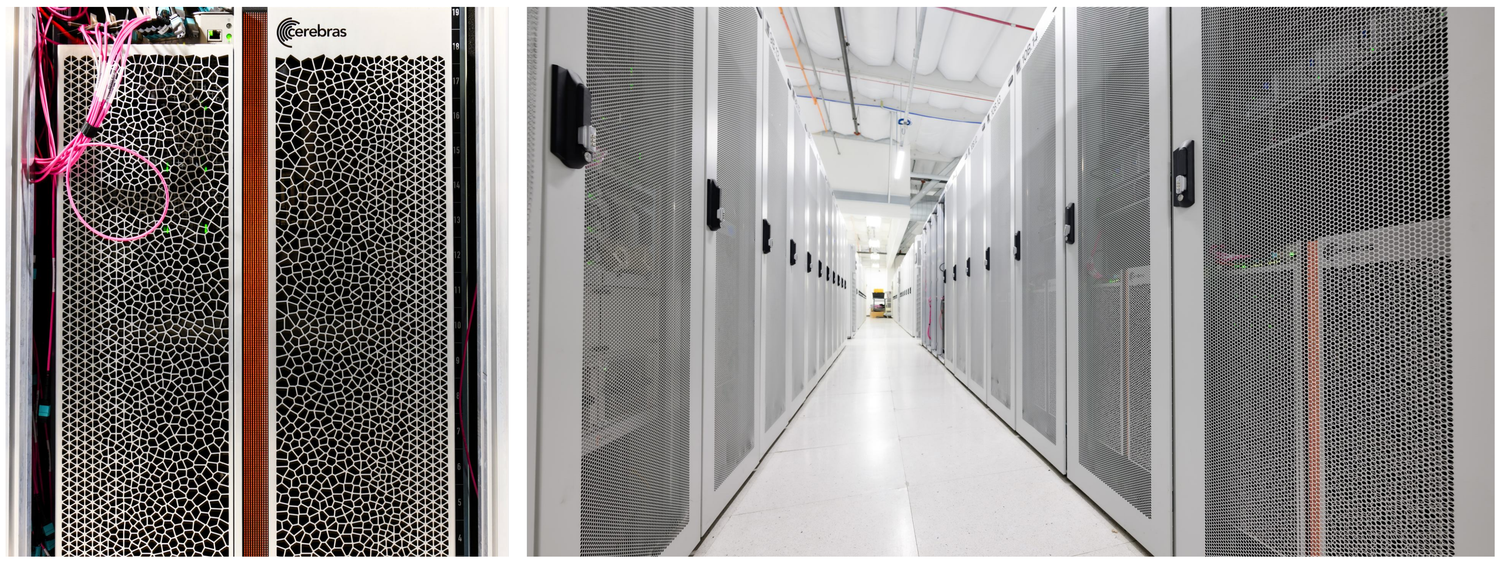
The Cerebras WSE-3 is 46,250 square millimeters of silicon, 4 trillion transistors, 900K cores, 44 GB on-chip memory, and delivers an unparalleled 125 petaFLOPS of AI compute. It surpasses all other processors in AI-optimized cores, memory speed, and on-chip fabric bandwidth.

Simplifying Large-Scale AI Computing

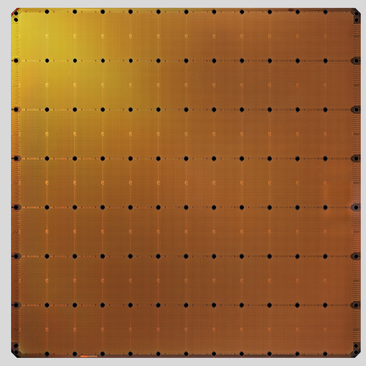
Conventional systems struggle to scale, hampered by the challenges of synchronizing vast arrays of processors across many nodes. The Cerebras WSC thrives, seamlessly integrating its components for large-scale, parallel computation and providing a straightforward, data-parallel programming interface.

Seamless Software Integration

The Cerebras Software Platform seamlessly integrates with ML frameworks like PyTorch, enabling researchers to use single or clustered CS-3s without altering their workflow.



Cerebras CS-3 (left) and Cerebras Wafer-Scale Clusters in our datacenter (right)



Cerebras Wafer-Scale Engine

Hardware

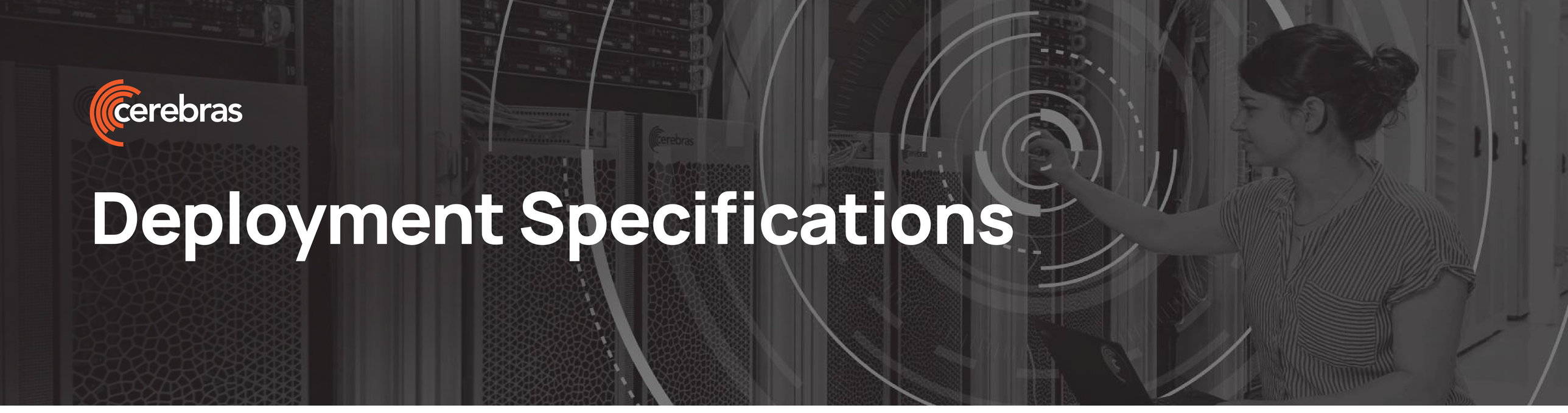
- Features the Cerebras Wafer-Scale Engine, with 125 petaFLOPS of AI compute
- High memory-compute bandwidth with on-chip memory on every compute core for faster generative AI training
- High bandwidth, low latency networking to interconnect multiple WSE's
- Optimized for handling trillions of data tokens with rapid data access and processing

Software

- Supports PyTorch 2.0
- Push-button scaling of compute resources and models
- Advanced cluster management for efficient job scheduling, model training, and data management

Services

- Data center design, installation, and post-installation testing and support
- White glove ML service from dataset preparation to custom ML applications built on your data



Deployment Specifications

Specifications

- 900,000 compute cores
- 125 PetaFLOPs of AI Performance
- 44 GB on-chip memory
- 12 to 1,200 TB of off-chip model memory
- 21 PB/sec memory bandwidth
- 214 PB/sec core-to-core bandwidth

Size & Weight (per Node)

- 42U x 1,200mm x 600mm EIA rack
- 800 kg (1,764 lbs)

Power (per Rack)

- 34kW provisioned power
- 4x 60A/208V drops
- Redundant and hot swappable in 6+6, 8+4, or 9+3 configurations
- Inlets: 12x IEC 60320 C20
- Inputs: 200-240 VAC, 16A, 50/60 Hz
  - Independent single-phase inputs
  - Protection: each inlet is individually protected with an external 16A (20A UL) circuit breaker

Network (per Node)

- Integrated optical multimode transceivers
- 12x 100GbE Data Ports (OM4 MPO/MTP-12)
  - 100GBase-SR4 link
  - Accepts MPO/MTP-12 fiber strand push-on cables
- Use Type-B cross-over OM4 MPO/MTP-12 50/125µm multi-mode fiber patch cable to plug into industry-standard 100GBase-SR4 optical module

Management (per Node)

- 1x 1GbE Management Port (RJ45)
- 1x Console Port (RJ45)
- 1x Power Management Port (RJ45)

Cooling (per Node)

- Air flow: 1,800CFM liquid-cooled, 2,800CFM air cooled
- Water temperature: 20 ± 2˚C
- Flow rate: 100 ± 10 L/min
- Can be deployed with an external liquid coolant loop or liquid-to-air cooling
- Internal coolant loop: 1+1 redundant hot-swappable pumps

Wafer Scale Cluster Hardware (per Node)

The Wafer-Scale Cluster streamlines large language model training with support for trillions of parameters and tokens through our proprietary MemoryX design. Weights are streamed in parallel to CS-3s for computation, with dedicated servers and switches for each cluster.

Each Wafer-Scale Cluster comes with a pre-determined design and specification for servers and switches based on the cluster size and customer needs.

Cerebras Wafer-Scale Clusters are customized to customer specifications. Here is an example of one construct:

- 1x Cerebras CS-3
- 1x Cerebras Wafer Scale Engine-3
- 8x Input Pre-Processing Servers
- 3x MemoryX enclosures
- 2x SwarmX Switches
- 1x Management Server
- 1x SwarmX enclosure
- 1x Management Switch
- 1x MemoryX Switch
- 1x Console Server