# Wafer-Scale Engine 3: The Largest Chip Ever Built

The Wafer-Scale Engine (WSE-3), which powers the Cerebras CS-3 system, is the largest chip ever built. The WSE-3 is 57 times larger than the largest GPU, has 52 times more compute cores, and 880 times more high-performance on-chip memory. The only wafer scale processor ever produced, it contains 4 trillion transistors, 900,000 AI-optimized cores, and 44 gigabytes of high performance on-wafer memory all at accelerating your AI work.

**Cerebras Wafer-Scale Engine**

*Fabrication process*
**5nm**

*Silicon area*
**46,225mm²**

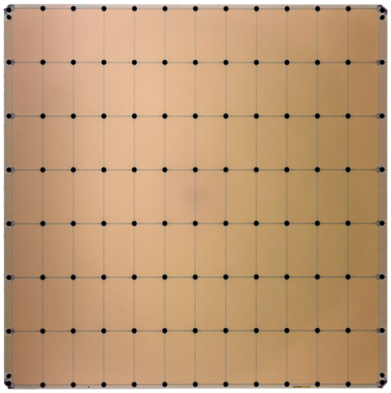
*Transistors*
**4 Trillion**

*AI-optimized cores*
**900,000**

*Memory (on-chip)*
**44GB**

*Memory bandwidth*
**21PB/s**

*Fabric bandwidth*
**214Pb/s**



Cerebras WSE-3
4 Trillion Transistors
46,225 mm² Silicon

Largest GPU
80 BillionTransistors
814 mm² Silicon

## Compute Designed for AI

Each core on the WSE-3 is independently programmable and optimized for the tensor-based, sparse linear algebra operations that underpin neural network training and inference for deep learning. The WSE-3 empowers teams to train and run AI models at unprecedented speed and scale, without the complex distributed programming techniques required to use a GPU cluster.

## Cluster-Scale in a Single Chip

Unlike traditional devices with tiny amounts of on-chip cache memory and limited communication bandwidth, the WSE-3 features 44GB of on-chip SRAM, spread evenly across the entire surface of the chip, providing every core with single-clock-cycle access to fast memory at an extremely high bandwidth of 21PB/s. This is 7,000x greater bandwidth than the leading GPU.

## High Bandwidth, Low Latency

The WSE-3 on-wafer interconnect eliminates the communication slowdown and inefficiencies of connecting hundreds of small devices via wires and cables. It delivers an astonishing 214 Pb/s interconnect bandwidth between cores. That's more than 3,715x the bandwidth delivered between the leading GPUs. The result is faster, more efficient execution for your deep learning work at a fraction of the power draw of traditional GPU clusters.

CEREBRAS SYSTEMS, INC.
1237 E. ARQUES AVE, SUNNYVALE, CA 94085 USA
CEREBRAS.NET