Cerebras CS-4

3x WSE-3 | Wafer Scale Engine 3 Turbo | rack | 2026

OVERVIEW

—	—	129600.0 TB/s	—
FP16 TFLOPS Dense	Aggregate Memory	Aggregate Bandwidth	Aggregate Power

PERFORMANCE METRICS

Precision	Peak (TFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP16	–	–	–

All figures are dense. Vendors commonly headline the sparse number, which is twice the dense one.

CONFIGURATION

Accelerator	WSE-3
Count	3
Memory per GPU	—
Bandwidth per GPU	—

POWER & EFFICIENCY

Aggregate Power	—
TDP per GPU	—
FP16 Efficiency	—
Aggregate Bandwidth	129600.0 TB/s

SYSTEM DETAILS

Vendor: Cerebras
Family: CS-4
Architecture: Wafer Scale Engine 3 Turbo
Form Factor: rack
Release Year: 2026
Process Node: TSMC 5nm

INTERCONNECT

Direct Wafer Links, switch-free, plus RoCE v2 RDMA over Ethernet; 7.2 Tbit/s system I/O; 160.5 PB/s on-wafer fabric

NOTES

The Cerebras CS-4 is a rack built from three Wafer Scale Engine 3 Turbo processors, and the first system on the Cerebras Nexus platform. Each wafer is the same silicon as the WSE-3 in the CS-3, four trillion transistors across 46,225 mm2 on TSMC 5 nm with 900,000 AI cores and 44 GB of on-chip SRAM, run at roughly twice the throughput: 250 PFLOPS and 43.2 PB/s of memory bandwidth per wafer against 125 PFLOPS and 21 PB/s before it. Three of them give the rack 750 PFLOPS, 132 GB of SRAM and 129.6 PB/s of memory bandwidth, with no DRAM anywhere in the system. Cerebras gets the extra throughput from packaging rather than from new silicon: power conversion now sits 0.5 mm from the wafer instead of the roughly 50 mm of a conventional accelerator board, which nearly removes board-level loss and lets the rack push twice as much power into each wafer. Compute, power and I/O are separate modules, the wafer and its power conversion, liquid cooling, I/O and control electronics folded into a rear-mounted Wafer-Scale Backpack with half the component count of the previous generation. The new I/O module doubles bandwidth to 2.4 Tbit/s per wafer and runs two ways, standards-based RoCE v2 RDMA over Ethernet for mixed fleets and switch-free Direct

Wafer Links between wafers, which brings wafer-to-wafer latency down to about two microseconds. Cerebras rates the rack at 750 PFLOPS but footnotes that figure as sparse FP16, and its sparsity is unstructured rather than 2:4 structured, so no dense equivalent can be derived by halving and none is published.