

Cerebras CS-3

1x WSE-3

Wafer Scale Engine 3

rack

2024

OVERVIEW

—	—	21000.0 TB/s	27.0 kW
Peak Dense	Aggregate Memory	Aggregate Bandwidth	Aggregate Power

CONFIGURATION

Accelerator	WSE-3
Count	1
Memory per GPU	—
Bandwidth per GPU	—

POWER & EFFICIENCY

Aggregate Power	27.0 kW
TDP per GPU	—
Efficiency	—
Aggregate Bandwidth	21000.0 TB/s

SYSTEM DETAILS

Vendor: Cerebras	Family: CS-3	Architecture: Wafer Scale Engine 3
Form Factor: rack	Release Year: 2024	Process Node: TSMC 5nm

INTERCONNECT

1.2 Tb/s system I/O; 214 Pb/s on-wafer fabric

NOTES

The Cerebras CS-3 is a single 16 RU liquid-cooled system built around one Wafer Scale Engine 3, drawing up to 27 kW. Its memory is 44 GB of on-chip SRAM running at 21 PB/s, with no DRAM of any kind, and system I/O is 1.2 Tb/s. Cerebras rates it at 125 petaFLOPS of FP16, but a footnote on its own datasheet states that FLOPS are shown as sparse, and the sparsity is unstructured rather than 2:4 structured, so no dense equivalent can be derived by halving and none is published. This entry therefore carries no precision figures rather than a sparse number that would be compared against the dense figures used everywhere else on this site. Cooling is water at 100 L/min per system, 20 +/- 2 C, with 50 um filtering and PG25 chemistry.