

Cerebras CS-3 Rack

2x WSE-3

Wafer Scale Engine 3

rack

2024

OVERVIEW

—	—	42000.0 TB/s	54.0 kW
Peak Dense	Aggregate Memory	Aggregate Bandwidth	Aggregate Power

CONFIGURATION

Accelerator	WSE-3
Count	2
Memory per GPU	—
Bandwidth per GPU	—

POWER & EFFICIENCY

Aggregate Power	54.0 kW
TDP per GPU	—
Efficiency	—
Aggregate Bandwidth	42000.0 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

8x QSFP-DD transceiver, 4x100G; 214 Pb/s on-wafer fabric per engine

NOTES

The Cerebras CS-3 Rack packages two CS-3 systems into a single 48 RU cabinet measuring 600 mm by 1200 mm, drawing 54 kW in total: 50.4 kW carried by water and 3.6 kW by air. Networking is eight QSFP-DD transceivers at 4x100G, with a 1 GbE management port, a console port and a PoE flow controller interface. Power is 2+1 redundant at 240V/415Y, three phase, 60 A on IEC 60309 plugs, and water couples through 1.5 inch sanitary fittings with two pairs of supply and return. Between them the two engines hold 88 GB of on-chip SRAM and no DRAM. Cerebras publishes 125 petaFLOPS per engine but footnotes that figure as sparse, and the sparsity is unstructured, so no dense figure is derived.