



Cerebras CS-3 Inference Cluster (16-node)

16x WSE-3 Wafer Scale Engine 3 cluster 2024

OVERVIEW

— Peak Dense	— Aggregate Memory	336000.0 TB/s Aggregate Bandwidth	475.0 kW Aggregate Power
-----------------	-----------------------	---	------------------------------------

CONFIGURATION

Accelerator	WSE-3
Count	16
Memory per GPU	—
Bandwidth per GPU	—

POWER & EFFICIENCY

Aggregate Power	475.0 kW
TDP per GPU	—
Efficiency	—
Aggregate Bandwidth	336000.0 TB/s

SYSTEM DETAILS

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

INTERCONNECT

On-wafer fabric per engine, 214 Pb/s; cluster networking via dedicated NET cabinets

NOTES

A sixteen-node Cerebras inference cluster, provisioned as ten cabinets: eight WSE cabinets holding two CS-3 systems each, plus two NET cabinets. Provisioned heat load is approximately 475 kW, about 440 kW on water and 35 kW on air. That is Cerebras's own provisioned total including networking, not sixteen times the 27 kW system figure. The cluster holds 704 GB of on-chip SRAM and no DRAM. Each engine is rated at 125 petaFLOPS, footnoted by Cerebras as sparse; the sparsity is unstructured, so no dense figure is derived.