



Cerebras CS-3 Inference Cluster (8-node)

8x WSE-3 Wafer Scale Engine 3 cluster 2024

OVERVIEW

— Peak Dense	— Aggregate Memory	168000.0 TB/s Aggregate Bandwidth	240.0 kW Aggregate Power
-----------------	-----------------------	---	------------------------------------

CONFIGURATION

Accelerator	WSE-3
Count	8
Memory per GPU	—
Bandwidth per GPU	—

POWER & EFFICIENCY

Aggregate Power	240.0 kW
TDP per GPU	—
Efficiency	—
Aggregate Bandwidth	168000.0 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

An eight-node Cerebras inference cluster, where a node is one CS-3 system. Cerebras provisions it as five cabinets: four WSE cabinets holding two CS-3 systems each, plus one NET cabinet. Provisioned heat load is approximately 240 kW, of which about 220 kW is carried by water and 20 kW by air. Note that this is Cerebras's own provisioned figure and is higher than eight times the 27 kW system power, because it includes the networking cabinet. The cluster holds 352 GB of on-chip SRAM and no DRAM. Cerebras rates each engine at 125 petaFLOPS but footnotes the figure as sparse, and because the sparsity is unstructured no dense equivalent is derived.