



Cerebras CS-3 Inference Cluster (88-node)

88x WSE-3 Wafer Scale Engine 3 cluster 2024

OVERVIEW

— Peak Dense	— Aggregate Memory	1848000.0 TB/s Aggregate Bandwidth	2.5 MW Aggregate Power
-----------------	-----------------------	--	----------------------------------

CONFIGURATION

Accelerator	WSE-3
Count	88
Memory per GPU	—
Bandwidth per GPU	—

POWER & EFFICIENCY

Aggregate Power	2.5 MW
TDP per GPU	—
Efficiency	—
Aggregate Bandwidth	1848000.0 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

The largest Cerebras inference cluster configuration published, at eighty-eight nodes provisioned as fifty-two cabinets: forty-four WSE cabinets holding two CS-3 systems each, plus eight NET cabinets. Provisioned heat load is approximately 2.54 MW, about 2.4 MW on water and 0.14 MW on air, Cerebras's own provisioned total including networking. Fully loaded a WSE cabinet weighs 843 kg and a NET cabinet 963 kg. The cluster holds 3.872 TB of on-chip SRAM and no DRAM at all. Each engine is rated at 125 petaFLOPS, footnoted by Cerebras as sparse; because that sparsity is unstructured rather than 2:4 structured, no dense equivalent can be derived and none is published.