Cerebras CS-3 Inference Cluster (64-node)

64x WSE-3 | Wafer Scale Engine 3 | cluster | 2024

OVERVIEW

—	—	1344000.0 TB/s	1.9 MW
Peak Dense	Aggregate Memory	Aggregate Bandwidth	Aggregate Power

CONFIGURATION

Accelerator	WSE-3
Count	64
Memory per GPU	—
Bandwidth per GPU	—

POWER & EFFICIENCY

Aggregate Power	1.9 MW
TDP per GPU	—
Efficiency	—
Aggregate Bandwidth	1344000.0 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 sixty-four-node Cerebras inference cluster, provisioned as thirty-eight cabinets: thirty-two WSE cabinets holding two CS-3 systems each, plus six NET cabinets. Provisioned heat load is approximately 1.86 MW, about 1.75 MW on water and 0.11 MW on air, Cerebras's own provisioned total including networking. The cluster holds 2.816 TB 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.