# Cerebras CS-3 Inference Cluster (32-node)

32x WSE-3 | Wafer Scale Engine 3 | cluster | 2024

## OVERVIEW

| — | — | 672000.0 TB/s | 945.0 kW |
|---|---|---|---|
| Peak Dense | Aggregate Memory | Aggregate Bandwidth | Aggregate Power |

## CONFIGURATION

| | |
|---|---|
| Accelerator | WSE-3 |
| Count | 32 |
| Memory per GPU | — |
| Bandwidth per GPU | — |

## POWER & EFFICIENCY

| | |
|---|---|
| Aggregate Power | 945.0 kW |
| TDP per GPU | — |
| Efficiency | — |
| Aggregate Bandwidth | 672000.0 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 thirty-two-node Cerebras inference cluster, provisioned as twenty cabinets: sixteen WSE cabinets holding two CS-3 systems each, plus four NET cabinets. Provisioned heat load is approximately 945 kW, about 875 kW on water and 70 kW on air, Cerebras's own provisioned total including networking. The cluster holds 1.408 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.