

d-Matrix Corsair Inference Server

8x Corsair

Corsair

rack

OVERVIEW

19.2 INT8 PFLOPS Dense	2.0 TB Aggregate Memory	— Aggregate Bandwidth	— Aggregate Power
----------------------------------	-----------------------------------	--------------------------	----------------------

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
INT8	19.2	2400.0	—
INT4	76.8	9600.0	—

All figures are dense. Vendors commonly headline the sparse number, which is twice the dense one.

CONFIGURATION

Accelerator	Corsair
Count	8
Memory per GPU	256 GB
Bandwidth per GPU	—

POWER & EFFICIENCY

Aggregate Power	—
TDP per GPU	—
INT8 Efficiency	—
Aggregate Bandwidth	—

SYSTEM DETAILS

Vendor: d-Matrix	Family: Corsair	Architecture: Corsair
Form Factor: rack	Release Year: —	Process Node: —

INTERCONNECT

PCIe-based scale-up; 512 GB/s card-to-card DMX Bridge

NOTES

Eight Corsair cards in one server, rated by d-Matrix at 19.2 PFLOPS of MXINT8 and 76.8 PFLOPS of MXINT4, both labelled dense on its own page. Memory is two-tier and unusual: 16 GB of on-die performance memory running at 1,200 TB/s, backed by up to 2 TB of capacity memory. There is no HBM anywhere in the machine. d-Matrix quotes 60,000 tokens per second at 1 ms per token for Llama3 8B on a single server. It publishes no power figure for the card, the server or the rack, so total system power is left blank rather than estimated. Note that the 1,200 TB/s belongs to the on-die SRAM and is deliberately not recorded as memory bandwidth, which on this site means DRAM bandwidth; d-Matrix publishes no bandwidth for the capacity tier.