

d-Matrix Corsair SquadRack

64x Corsair

Corsair

rack

OVERVIEW

154 INT8 PFLOPS Dense	16.4 TB Aggregate Memory	— Aggregate Bandwidth	— Aggregate Power
---------------------------------	------------------------------------	--------------------------	----------------------

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
INT8	153.6	2400.0	—
INT4	614.4	9600.0	—

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

CONFIGURATION

Accelerator	Corsair
Count	64
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

Scale-out over PCIe or Ethernet; 512 GB/s card-to-card DMX Bridge within each server

NOTES

d-Matrix's rack-scale configuration, marketed as SquadRack and specified on its product page under the heading "Inference Rack": eight Corsair Inference Servers holding 64 cards in total. It carries 128 GB of on-die performance memory at 9.6 PB/s and up to 16.4 TB of capacity memory, with no HBM at all, and d-Matrix frames the two tiers as separate operating modes — up to 100B-parameter models held in performance mode, and 1T+ frontier models in capacity mode. Scale-out is over PCIe or Ethernet. d-Matrix quotes 30,000 tokens per second at 2 ms per token for Llama3 70B on a single rack. Note that d-Matrix publishes no compute figure for the rack: the 153.6 PFLOPS MXINT8 and 614.4 PFLOPS MXINT4 here are eight times its published server figures, which is safe because its own rack memory figure of 128 GB is likewise exactly eight times the server's 16 GB, and both reconcile against the 64 cards. No power figure is published for any part of the product line.

