QCT QuantaGrid D75H-10U

8x B300 | Blackwell Ultra | rack

OVERVIEW

—	—	—	—
Peak Dense	Aggregate Memory	Aggregate Bandwidth	Aggregate Power

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP8	–	–	–
FP4	–	–	–

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

CONFIGURATION

Accelerator	B300 SXM 288GB
Count	8
Memory per GPU	288 GB
Bandwidth per GPU	8.0 TB/s

POWER & EFFICIENCY

Aggregate Power	—
TDP per GPU	1400 W
Efficiency	—
Aggregate Bandwidth	—

SYSTEM DETAILS

Vendor: QCT
Family: QuantaGrid
Architecture: Blackwell Ultra
Form Factor: rack
Release Year: —
Process Node: 4nm

INTERCONNECT

One NVIDIA HGX B300 8-GPU baseboard; scale-out is eight ConnectX-8 OSFP ports carrying 800G Ethernet or InfiniBand at one port per GPU for GPU-Direct RDMA

NOTES

QCT's 10U Blackwell Ultra server, built on the NVIDIA HGX B300 platform with two Intel Xeon 6 processors. Networking is eight ConnectX-8 OSFP ports at 800G, one per accelerator, for one-to-one GPU-Direct RDMA, and there is room for four full-height full-length PCIe Gen 5 slots. QCT quotes 72 PFLOPS of FP8 and 144 PFLOPS of FP4 without saying whether either is dense or sparse. Both match NVIDIA's sparse figures for this baseboard exactly, and neither matches the dense figures of 36 and 108, so they are recorded as sparse. Worth noting QCT calls the FP8 number a training figure, which would normally imply dense; the number itself settles the basis, not the label. Power is twelve 3,200 W supplies in N+N redundancy, which is capacity rather than draw, so no system power is shown.