Huawei Atlas 800I A3

8x Ascend 910C | Da Vinci v2 | rack | 2025

OVERVIEW

4.48	1.0 TB	25.6 TB/s	14.6 kW
FP16 PFLOPS Dense	Aggregate Memory	Aggregate Bandwidth	Aggregate Power

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP16	4.5	560.0	–

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

CONFIGURATION

Accelerator	Ascend 910C
Count	8
Memory per GPU	128 GB
Bandwidth per GPU	3.2 TB/s

POWER & EFFICIENCY

Aggregate Power	14.6 kW
TDP per GPU	600 W
FP16 Efficiency	0.31 TFLOPS/W
Aggregate Bandwidth	25.6 TB/s

SYSTEM DETAILS

Vendor: Huawei
Family: Atlas
Architecture: Da Vinci v2
Form Factor: rack
Release Year: 2025
Process Node: SMIC 7nm

INTERCONNECT

UnifiedBus (UB): 784 GB/s bidirectional D2D per NPU; 8 x 400GE QSFP-DD direct from NPU over RoCE, plus 56 x 400GE QSFP-DD over the bus protocol

NOTES

The inference member of Huawei's Atlas 800 A3 family: a 10U air-cooled rack server with eight Ascend 910 NPUs and four Kunpeng 920 CPUs, rated at 4.48 PFLOPS of FP16 with 1,024 GB of on-chip memory across the eight NPUs at 3.2 TB/s each. That works out at 560 TFLOPS per NPU against the 750 of the training-oriented Atlas 800T A3 in the same chassis, which is a genuine inference bin rather than a transcription difference: Huawei's own family page lists three Atlas 800 A3 models at 6.0, 5.0 and 4.48 PFLOPS FP16. Memory is offered as 8 x 128 GB or 8 x 64 GB; the 128 GB configuration is recorded here. Networking is eight 400GE QSFP-DD ports direct from the NPUs over RoCE plus fifty-six more over Huawei's bus protocol, with 784 GB/s of bidirectional device-to-device bandwidth. Power is up to 14.6 kW from six hot-swap 3.0 kW supplies in 5+1 redundancy, air cooled. Huawei publishes no INT8 figure for this machine and labels nothing dense or sparse, so only the FP16 rate it states is recorded and nothing is derived.