Huawei Atlas 800T A3

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

OVERVIEW

6.00	1.0 TB	25.6 TB/s	—
FP16 PFLOPS Dense	Aggregate Memory	Aggregate Bandwidth	Aggregate Power

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP16	6.0	750.0	–

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	—
TDP per GPU	600 W
FP16 Efficiency	—
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; scales to a 384-NPU Atlas 900 A3 SuperPoD

NOTES

The training member of Huawei's Atlas 800 A3 family: a 10U air-cooled compute node with eight Ascend 910 NPUs and four Kunpeng 920 CPUs, rated at 6.0 PFLOPS of FP16 with 1,024 GB of on-chip memory at 3.2 TB/s per NPU and 784 GB/s of bidirectional device-to-device bandwidth. Huawei sells three models on this chassis at 6.0, 5.0 and 4.48 PFLOPS FP16, which works out at 750, 625 and 560 TFLOPS per NPU; this is the top one, and the 4.48 model is the separately listed Atlas 800I A3. Multiple nodes combine into an Atlas 900 A3 SuperPoD of up to 384 cards. Huawei publishes only an FP16 figure for this machine, and no INT8: a previously held INT8 value of 12.0 PFLOPS was 8 x 1,500 TFLOPS, a per-NPU rate Huawei states nowhere, and has been removed. Power is also not attributable: Huawei's family page gives 16.2 kW as the maximum input power for the Atlas 800 A3 line, but that line covers three compute variants and the Atlas 800I A3's own page separately says 14.6 kW, so no figure is recorded here rather than guessing which model the 16.2 kW describes. Huawei labels nothing dense or sparse.