



Huawei Atlas 900 A3 SuperPoD

384x Ascend 910C Da Vinci v2 pod 2025

OVERVIEW

307 BF16 PFLOPS Dense	49.2 TB Aggregate Memory	1228.8 TB/s Aggregate Bandwidth	— Aggregate Power
--------------------------	-----------------------------	------------------------------------	----------------------

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
FP16	307.2	800.0	—
BF16	307.2	800.0	—
INT8	614.4	1600.0	—

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

CONFIGURATION

Accelerator	Ascend 910C
Count	384
Memory per GPU	128 GB
Bandwidth per GPU	3.2 TB/s

POWER & EFFICIENCY

Aggregate Power	—
TDP per GPU	600 W
BF16 Efficiency	—
Aggregate Bandwidth	1228.8 TB/s

SYSTEM DETAILS

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

INTERCONNECT

UnifiedBus (UB) all-to-all optical; 784 GB/s bidirectional D2D per NPU

NOTES

Huawei's 384-NPU supernode, built from twelve 47U liquid-cooled compute cabinets plus four air-cooled bus-equipment cabinets, each 2,250 by 600 by 1,150 mm. It carries up to 384 Ascend 910 NPUs and 192 Kunpeng 920 CPUs, with 384 x 128 GB of on-chip memory at up to 3.2 TB/s each and 784 GB/s of bidirectional device-to-device bandwidth. Huawei rates it at up to 307.2 PFLOPS of FP16, and also lists 288.7 and 240.3 PFLOPS variants; the top figure works out at exactly 800 TFLOPS per NPU, which is the Ascend 910C's own published rate, so the BF16 and INT8 figures here are that chip's published 800 and 1,600 TFLOPS multiplied by 384. Cooling is liquid for the compute cabinets and air for the bus cabinets, with an operating range of 5 to 40 C. Note that Huawei publishes no power figure for this machine: its supply row states only the input voltage, three-phase 380V AC on two feeds. A figure of 559 kW previously held here appeared in no Huawei source and has been removed rather than shipped. Huawei labels nothing dense or sparse. This machine is also

sold as CloudMatrix 384, which is a deployment name for the same 384-NPU configuration rather than a separate product.