



SambaNova SambaRack SN40L-16

16x SN40L RDU rack

OVERVIEW

10.2 BF16 PFLOPS Dense	1.0 TB Aggregate Memory	28.8 TB/s Aggregate Bandwidth	10.0 kW Aggregate Power
---------------------------	----------------------------	----------------------------------	----------------------------

PERFORMANCE METRICS

Precision	Peak (PFLOPS)	Per GPU (TFLOPS)	TFLOPS/W
BF16	10.2	638.0	—

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

CONFIGURATION

Accelerator	SN40L
Count	16
Memory per GPU	64 GB
Bandwidth per GPU	1.8 TB/s

POWER & EFFICIENCY

Aggregate Power	10.0 kW
TDP per GPU	—
BF16 Efficiency	1.02 TFLOPS/W
Aggregate Bandwidth	28.8 TB/s

SYSTEM DETAILS

Vendor: SambaNova	Family: SambaRack	Architecture: RDU
Form Factor: rack	Release Year: —	Process Node: 5nm

INTERCONNECT

Peer-to-peer RDU interconnect with hardware collective offload; PCIe host attach per socket

NOTES

Sixteen SN40L Reconfigurable Dataflow Units in a single air-cooled rack, which SambaNova positions as able to run models the size of DeepSeek R1 671B and Llama 4 Maverick without liquid cooling. The rack draws an average of 10 kW. Compute is 10.2 PFLOPS of BF16 and memory is three-tier: 8.3 GB of on-chip SRAM, 1 TB of HBM at 28.8 TB/s, and 24 TB of high-capacity DDR. That third tier is the point of the architecture, letting a rack hold many models in DDR and move the active one into HBM in microseconds. Note that the 24 TB of DDR is deliberately not recorded as the system memory figure here, which carries the 1 TB of HBM instead, because the two tiers are different things and conflating them would suggest 24 TB of high-bandwidth memory. SambaNova publishes no rack-level compute figure; the 10.2 PFLOPS is sixteen times the per-socket figure from its own architecture paper, which is safe because its two published rack memory figures of 1 TB and 24 TB are likewise exactly sixteen times the socket. Neither source labels the compute dense or sparse. SambaNova also publishes per-user token rates for this configuration, shown separately below.

