SambaRack™ SN40L-16

The hardware system for running high performance AI Workloads

Unlock the fastest system for AI model inference with the capability to run multiple models, including the latest and largest open source models with highest performance. Deployable as an on-premises solution and in hosted data centers, SambaRack is the system that powers the industry leading SambaCloud platform and SambaStack.

The fastest platform for inference	Reduced Power Consumption	Designed for Scale
Delivering world record performance and accuracy, across the latest large and small models with the highest accuracy	Run dozens of models and switch between them in microseconds, on a single rack that only consumes 10kW	Start with as little as one node and a few models and scale efficiently to meet the needs of any size organization

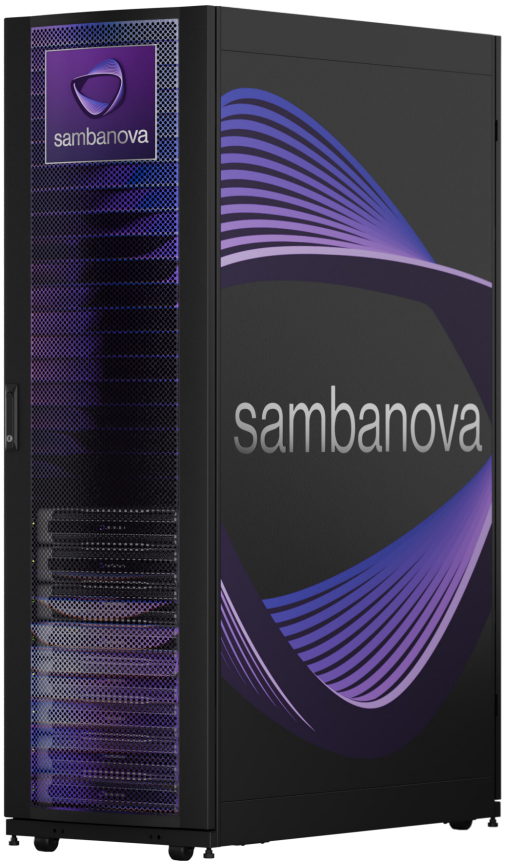
The SambaRack system takes advantage of the unique SambaNova SN40L Reconfigurable Dataflow Unit (RDU) to deliver exceptional performance in a small footprint. The SN40L is able to deliver this extreme performance thanks to its revolutionary dataflow architecture and large memory footprint.

Dataflow architecture

The SN40L is purpose-built for AI. Breaking free from the limitations of legacy technologies, the SN40L uses a dataflow architecture and revolutionary software stack that maps AI algorithms to the processor and dynamically reconfigures the processor for optimal performance. This eliminates the redundancy inherent to GPU architectures.

Three tiered memory architecture

Purpose-built to power the largest AI models, the SN40L has a three tiered memory architecture that includes very large memory, high bandwidth memory, and extremely fast memory. The result is that a single system node can support up to 5 trillion parameters consisting of up to hundreds of separate models. With terabytes of addressable memory, the SN40L is ideal for custom and chained models, and can switch between models in microseconds which is orders of magnitude faster than legacy GPUs.



SambaNova is the leading purpose-built AI system for generative and agentic AI implementations, from chips to models that gives enterprises full control over their model and private data. We take the best models, optimize them for fast tokens and higher batch sizes, the largest inputs and enable customizations to deliver value with simplicity.

© SambaNova, all rights reserved

Hardware Specifications

SN40L RDUs	16x Cerulean SN40L™ Reconfigurable Dataflow Unit (RDU), with 520MB SRAM, 64GB HBM, 768GB DDR each
Processing Power	10.2 PFLOPs @BF16
Total RDU SRAM	8GB
Total RDU HBM	1TB
Total RDU attached DDR	12TB
Host processor	2× 64 core CPUs, 2TB DDR4 memory
Host boot and storage	4× 960GB (2x RAID 1 + 2 hot-spares)
Host storage	6× 7.6TB NVMe disks RAID 10 (21TB usable)
Networking	High performance 400/200 GbE data switch
Management	• 1 GbE switch • Serial console server
Software	Red Hat Enterprise Linux OS

Environmental Specifications

System Dimensions	Height: 78.5″ (1994 mm) | Width: 24.0″ (610 mm) | Depth: 50″ (1270 mm)
Power consumption	Inference: 7kW-14.5kW, Typical 10kW
Operating Temperature	59° F to 86° F (15 C to 30° C)
Operating Humidity	20% to 80% (non-condensing)
Operating Altitude	Up to 9842ft (3000m); derated by @ 1.8F (1C) per 984ft (300m) above 2952ft
System Weight	1068lbs (485kg)

sambanova

Sambanova SambaRack™

To learn more about how sambanova can accelerate and transform your organization with generative AI, **schedule a meeting.**

Learn more at sambanova.ai

linkedin.com/company/sambanova

@SambaNovaAI

info@sambanova.ai

SambaNova is the leading purpose-built AI system for generative and agentic AI implementations, from chips to models that gives enterprises full control over their model and private data. We take the best models, optimize them for fast tokens and higher batch sizes, the largest inputs and enable customizations to deliver value with simplicity.

© Sambanova, All rights reserved