SambaNova DataScale SN40L

The Hardware System for Running High Performance AI Workloads

Unlock the fastest system for AI model training and 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, SambaNova DataScale® is the system that powers the industry leading SambaNova Cloud platform and is the building block for the SambaNova Supercomputer.

Powered by the SambaNova SN40L Reconfigurable Dataflow Unit™ (RDU), the SambaNova DataScale SN40L delivers unprecedented performance across all model sizes to enable government agencies, research organizations, and enterprises to train and deploy the most demanding generative and agentic AI workloads and achieve world record performance with the largest and most challenging models.

The fastest platform for inference

Delivering world record performance and accuracy, across the latest large and small models with the highest accuracy

Reduced Power Consumption

Run dozens of models and switch between them in microseconds, on a single rack that only consumes 10kWs

Designed for Scale

Start with as little as one node and a few models and scale efficiently to meet the needs of any size organization

The DataScale 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 very 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.

DataScale SN40L-16 Specifications

Components

SN40L RDUs	16x Cerulean SN40L Reconfigurable Dataflow Unit (RDU) chips, with 520MB SRAM, 64GB HBM, 768GB DDR each
Processing Power	10.2 TFLOPs @BF16
Total RDU SRAM	8GB
Total RDU HBM	1TB
Total RDU attached DDR	12TB
Host processor	2x 64 core CPUs, 2TB DDR memory
Host boot and storage	4x 960GB (2x RAID 1)
Host storage	6x7.6TB NVMe disks
Networking	High performance 400/200GbE data switch
Management	• 1GbE 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 Training:14kW-17kW, Typical 16kW
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 (900m)
Rack Weight	1068lbs (485kg)

SambaNova® SYSTEMS

SambaNova DataScale®

To learn more about how SambaNova Systems 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

Customers turn to SambaNova to quickly deploy state-of-the-art AI capabilities to meet the demands of the AI-enabled world. Our purpose-built enterprise-scale AI platform is the technology backbone for the next generation of AI computing. We enable customers to unlock the valuable business insights trapped in their data. Our flagship offering, SambaNova Suite, overcomes the limitations of legacy technology to power the large complex foundation models that enable customers to discover new services and revenue streams, and boost operational efficiency. Headquartered in Palo Alto, California, SambaNova Systems was founded in 2017 by industry luminaries, and hardware and software design experts from Sun/Oracle and Stanford University. Investors include SoftBank Vision Fund 2, funds and accounts managed by BlackRock, Intel Capital, GV, Walden International, Temasek, GIC, Redline Capital, Atlantic Bridge Ventures, Celesta, and several others.