SambaNova SYSTEMS

PRODUCT SHEET

SambaNova Suite

World record performance on the most efficient, accurate, and secure AI platform for enterprises and governments

SambaNova delivers the only complete AI solution offering the world's fastest inference performance with full accuracy, security and data privacy, model ownership, and the flexibility to power the generative AI workloads of today and the agentic AI systems of tomorrow.

The full stack suite

SambaNova Suite delivers a comprehensive platform for generative and agentic AI implementations. The full suite includes the SambaNova DataScale® system, the SambaStudio® software, and the innovative SambaNova Composition of Experts (CoE) Model architecture. These components combine into a powerful platform that delivers unparalleled performance, ease of use, accuracy, data privacy, and the ability to power every use case across the world's largest organizations.

SambaNova DataScale

The world's fastest platform for inference and the only one with the ability to train and run hundreds of models on a single, energy efficient node

SambaStudio

Intuitive software that enables organizations to quickly and easily deploy, fine tune, add, or remove models and control access, all from a single endpoint

Composition of Experts

Model architecture that combines multiple models to deliver greater efficiency, performance, accuracy, and capabilities than is possible from a single model

SambaNova Suite delivers the SambaStudio software to enable easy management of your entire AI ecosystem from a single endpoint, the SambaNova DataScale SN40L platform, and our unique Composition of Experts (CoE) model architecture to support every AI workload in the organization. All as part of a complete system with world record performance and accuracy, in a small footprint.

The most scalable, flexible platform to power every AI workload

SambaNova brings enterprises and governments the functionality and accuracy of large models, with the performance to meet the needs of every user and application, without the inherent data privacy risks or high costs associated with cloud-based model providers. All with a total cost of ownership (TCO) that is 10x better than the competition.

The superior GPU alternative - Built for AI

Powered by the SambaNova Cerulean SN40L™ Reconfigurable Dataflow Unit™ (RDU), the fourth generation of our chip, which is purpose-built for the most demanding AI workloads. Its revolutionary dataflow design and massive memory footprint make it the proven performance leader for training and inference.

Complete data privacy and security

Protect your private data with the SambaNova platform. With SambaNova you always own your models and control your data. Deployable in your private cloud, in your data center, or wherever you need, SambaNova gives you total privacy so you can get the most value from your data.

The highest performance and accuracy

SambaNova is the leader in performance and accuracy with world record throughput with full precision accuracy for the largest open source model, Llama 3.1 405B. We also delivered over 1000 tokens per second with full accuracy for Llama 3 8B. We deliver the highest performance and accuracy for models large and small.

The highest accuracy - Fine-tuning and RAG

Achieve the highest accuracy by securely fine-tuning models with your private data. Once models are fine tuned on your data, your organization then owns the model in perpetuity, eliminating vendor lock-in. The SN40L enables organizations to perform continuous training on their models so that they are always up to date. Retrieval augmented generation (RAG) can be used to keep the model up to date on the latest information, such as customer orders or shipping information.

The only complete, full stack solution

Take advantage of the first, full stack solution, from our fourth generation chip, the SN40L, through our complete software stack, and the latest open-source models, delivered as part of an ensemble of models, the Composition of Experts. No other platform delivers this broad range of capabilities, in a complete system, with world record performance and accuracy, in such a small footprint.

Scalability and deployment flexibility

The SambaNova platform enables organizations to start with the right compute resources and models for their applications. As the needs of the organization grow, additional models that provide new capabilities can be added. When new models become available they can be added easily.

SambaNova Suite can be deployed in any way to meet the needs of the organization. Customers can start with a cloud-based or on-premises deployment. Resources can be easily scaled up as needed, without the need for costly and complex configuration.

SambaNova®
SYSTEMS

SambaNova Suite™

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

Learn more at SambaNova.AI

in 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.