sambanova

# Sambanova SambaStack

## The purpose built, full stack platform for high speed AI inference

The future of AI will be powered by systems that serve the most accurate, high performance production scale AI inference in the most efficient way. The SambaStack platform delivers on the performance, efficiency, and management capabilities to power the most demanding AI workloads of today and tomorrow. Capable of powering ever larger models, including Mixture of Experts (MoE) models, which have higher efficiency and can be adapted to an organization's specific data to provide the most accurate answers. SambaStack has the performance and model management capabilities to power reasoning and chain-of-though models to be used in agentic configurations without impacting the user experience in terms of time of output latency. Finally, it has the ability to run these models efficiently in terms of space, power, and ultimately cost.

| SambaRack | Model Bundles | Bring Your Own Checkpoint |
|---|---|---|
| The fastest and most efficient platform for inference on the largest open source models | Power all your models on the platform that can run multiple models simulaneously | Pretrain models on your exisitng infrastructure and run them on sambanova |

**The largest models**
Run the largest open-source models, including the latest DeepSeek and Llama 4models and to run them with the largest context length. Power these large, complex models at speeds that allow for complex agentic configurations without impacting the user experience.

**Built for efficiency**
In addition to being faster than other systems, the SambaStack platform is also more efficient. Other systems either run one model per rack and some even require multiple racks to run even a single model. Even worse, many require specialized liquid cooling and consume massive amounts of power. SambaStack run the largest models on a single, air-cooled system that only consumes an average of 10kW of power.

**Powered by the SN40L RDU**
The SambaStack platform is powered by the fourth generation SambaNova SN40L Reconfigurable Dataflow Unit (RDU). This advanced AI accelerator takes advantage of a dataflow architecture to dramatically increase the efficiency of how inference is performed. Combined with a unique three-tiered memory design that enables the platform to both run the largest models and to run multiple models simultaneously, the SN40L delivers unmatched inference speed to power the most demanding AI applications.

**Ready for agentic**
The large memory of the SN40L enables the SambaStack platform to run multiple models on a single system. Since models are held in memory, switching between models is done in as little as 2 microseconds, making it the ideal solution for agentic workloads.

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.