SambaNova SYSTEMS

EN

BACK TO RESOURCES

PREVIOUS | NEXT

May 29, 2024

SambaNova has broken the 1000 t/s barrier: why it's a big deal for enterprise AI

SambaNova is the clear winner of the latest large language model (LLM) benchmark by Artificial Analysis. Topping the Leaderboad at over 1000 tokens per second (t/s), Samba-1 Turbo sets a new record for Llama 3 8B performance on a single SN40L node and with full precision.

With speeds like this, enterprises can expect to accelerate an array of use cases and will enable innovation around unblocking agentic workflow, copilot, and synthetic data, to name a few. This breakthrough in AI technology is possible because the purpose-built SambaNova SN40L Reconfigurable Dataflow Unit (RDU) can hold hundreds of models at the same time and can switch between them in microseconds.

Speed for today and tomorrow

For today’s workloads, this speed will result in immediate efficiencies in application chains used today to solve complex business problems without compromise on quality. Today, SINGLE models aren’t able to conclusively solve business problems with quality. Applications delivering real business value are making many model calls as part of an application. These model calls add up to unacceptably slow performance for high quality answers. SambaNova's 1000 t/s inference speed and Composition of Experts (CoE) architecture allows multiple passes to be completed by multiple model types, producing true business-quality answers in seconds instead of minutes or hours.

AI technology is already advancing beyond human prompts and chatbots. Soon, agentic AI - a system where AI models or other tools, not humans, consume the outputs in a chain sequence - will become a primary way businesses use AI. Instead of a single call and response, people will ask an AI to chain together commands to complete a complex set of tasks. For instance, a medical researcher may ask Samba-1 a single prompt: Place a chart into my presentation based on the UN Greenhouse Gas Inventory dataset. Behind the scenes, a series of small specialized AI models write code to pull the dataset, feed them into a spreadsheet, deliver a graph and embed it into a presentation. With each request, there is a handoff that takes time and behind that time, a system that requires memory, power, cooling, and storage. Speed is the first step in allowing this chain to execute efficiently - but not just tokens per second - this is where time to first token and total duration also matter. Beware of vendors touting singular speed metrics - they may be flashy, but success for the enterprise is not just about being fast.

At SambaNova, we care about speed in service to the enterprise. This means moving from t/s to ANSWERS per second; business-accurate, reliable, validated answers iteratively curated by agentic chains. On Samba-1, those chains are being accelerated at every layer of our stack, from chip speed (leading raw throughput) through chain speed in the API (leading in model-handoff optimizations) and personalized accuracy of running customized zero-lock-in open-source models fine tuned on customer data behind their firewalls, allowing an enterprise’s best data to be brought to bear on the problem, consistently outperforming generalized models.

WE WANT TO HEAR FROM YOU!

Join our developer community on X to share your feedback. Check out Samba-1 Turbo to try for yourself today.

Keith Parker

BACK TO RESOURCES

PREVIOUS | NEXT

PRODUCTS

RESOURCES

INDUSTRIES

ABOUT

CONTACT

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.

Our Headquarters:
2200 Geng Road, Unit 100
Palo Alto, CA 94303
(650) 263-1153

Terms & Conditions
Privacy Policy