VOLUME 01
ISSUE 10
JULY 2025
www.gecnewswire.com

The AI Times

SUPPLEMENT OF ENTERPRISE CHANNELS MEA

INSIGHTS FOR A SMARTER WORLD

WORK REWIRED

A NEW ERA OF HUMAN-MACHINE COLLABORATION

INSIDE: TOP 10 SLMS

INTERVIEW

Redefining AI infrastructure

SambaNova's CEO and co-founder Rodrigo Liang speaks to us about how SambaManaged enables organisations to quickly launch profitable AI inference services leveraging existing power and network infrastructure.

RODRIGO LIANG

CEO AND CO-FOUNDER, SAMBANOVA'S

Could you walk us through the vision behind setting up the company?

When I started the company eight years ago, it was very clear that the AI market was going to become massive. We saw real opportunities to create more efficient infrastructure for deploying advanced AI technologies.

Fast forward to today—we've built silicon that is 10 times more efficient than what's available from companies like NVIDIA. Our technology enables organizations to deploy complex models much faster and with significantly higher performance.

That's really the core of our vision: even though existing AI infrastructure is good and has seen great success, we believed—and still believe—there's a better, more efficient way to scale. Especially in a world where power constraints are growing and the cost of compute is rising rapidly, we saw the need for alternative infrastructure.

So that's what we do: we build everything—from our own chips to custom AI models—designed to run in secure, private environments at scale.

Do you see infrastructure as the biggest bottleneck in AI right now?

I'd say there are three main challenges, but the biggest one is power.

Power has become the most significant constraint. In some cases, you may not even have access to a data center with sufficient power. Or if you do, the data center might not have access to reliable energy sources to continue scaling.

We're seeing data centers sitting idle, or countries that simply don't have additional grid capacity to expand. Think of certain regions in Southeast Asia or parts of Europe—power limitations are very real.

So yes, power is probably the number one constraint for AI infrastructure today.

SambaManaged promises deployment in just 90 days—a dramatic improvement over industry norms. What makes this speed possible?

So here's the thing—if you think about what it actually takes for enterprises to get into the AI game, there are several major hurdles.

First, you need to secure the infrastructure, but once you do that, you quickly realize that most AI models require a data center that can support 140 kilowatts with liquid cooling—and very few data centers in the world can handle that.

So, companies begin this long journey:

they secure infrastructure, then upgrade their data centers, which alone can take 18 to 24 months. After that, they need to build the expertise to deploy and manage the AI models. It's a long, complicated, and costly process.

What we've done with SambaManaged is streamline that entire path. Because our power consumption is so low—just 10 kilowatts per rack—we can deploy in most existing data centers without any need for infrastructure overhauls.

And since we offer a full-stack solution, we also eliminate two other barriers.

First, even after solving the power issue, many organizations struggle with deploying models and software efficiently. Because SambaNova delivers a fully managed stack, we handle that complexity.

Second, there's the issue of talent—most companies don't have teams with the expertise to run and scale these models. With SambaManaged, SambaNova's team manages everything, from infrastructure to model deployment.

So for companies that want to enter the AI space, SambaManaged offers the technology, the infrastructure, and the deployment expertise—all within 90 days. It's a game-changer.

How do you actually achieve such high performance per watt, and what impact is this having on total cost of ownership?

It's significant. It all starts with our custom-built chip. If you want to create AI infrastructure that's not dependent on NVIDIA GPUs, you need to design the foundation differently—and that's exactly what we did.

Our chip is inherently more efficient, which means it drastically reduces power requirements right at the core. Once you've collapsed the power consumption at the chip level, you're able to unlock a cascade of benefits.

For example, we can host multiple models within the same efficient rack, achieving much higher density per rack. This means you're running more compute power in less space, with lower energy consumption.

Additionally, our overall system performance is five times faster, so you get the same throughput with fewer racks. This directly contributes to a lower total cost of ownership (TCO)—fewer racks, less power, lower cooling requirements, and more performance.

In short, we give you higher efficiency, better scalability, and dramatically improved economics for AI infrastructure.

How do you handle cooling and thermal management with just 10 kilowatts of air-cooled power?

Exactly—that's the beauty of using just 10 kilowatts. Most data centers are already equipped to handle that level of air cooling, so we don't need to introduce the cost, effort, or complexity of retrofitting them for liquid cooling. And since the majority of data centers around the world are air-cooled, this makes our solution highly compatible and deployable with existing infrastructure.

How easy is it to scale SambaManaged from a pilot to full production?

It's incredibly straightforward. For example, you can run a full deep learning model with 670 billion parameters—which typically requires dozens of racks—on just one of our racks. That's because a single rack can handle a full 670B parameter model at full precision, which very few other platforms in the world can support.

As your needs grow, you simply add more racks based on user demand. It's a true modular scale-out architecture—start with one, scale to many, without the need to re-engineer the environment.

"As your needs grow, you simply add more racks based on user demand. It's a true modular scale-out architecture—start with one, scale to many, without the need to re-engineer the environment."

Can you walk us through the difference between the managed and self-service deployment models, and what types of customers typically choose each?

If you look at some of the hyperscale clouds and neo-cloud providers, they already have their own ecosystems. For example, Google has Vertex AI, and AWS offers Bedrock—these are sophisticated orchestration platforms that allow multiple users to access and operate different parts of the AI stack.

But for newer players—like telcos, neo-cloud providers, data center operators, or even on-premise enterprises that don't have mature AI capabilities—it can be quite complex.

Managing all that incoming traffic and deploying models efficiently requires a high level of expertise. Without that expertise, it becomes difficult to optimize performance or total cost of ownership (TCO).