SambaNova SYSTEMS

HPC User Forum at ANL

Jennifer Glore
Vice President, Customer Engineering

Where is the Industry Going?

- AI is experiencing rapid model innovation and a growing need for mission-aligned models, alongside a shift towards agentic AI.
- These trends are emphasizing the need for flexible system packing, faster speeds, rapid model switching, and greater interactivity.

Support Large Volume of Models with Less Hardware

Reduced Power: Lower Provisioning and Consumption

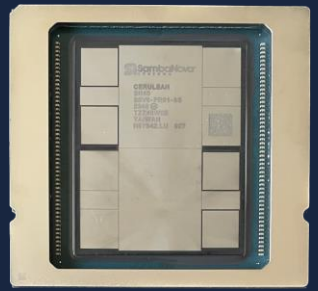
Meet the Performance Needs for the Future of AI

SambaNova SYSTEMS

SambaNova Key Innovations

Chip: Reconfigurable Dataflow Unit (RDU)

- Dataflow architecture
- Native multi-tenancy support with fast model switching
- Fast inferencing: Llama3 8B, full precision, 1000 tokens per second on only 16 chips (fast.snova.ai)
- .75 Terabytes of memory directly attached per chip, 6 TB per node
- Ideal for production inference, multi-tenancy, agentic workflows
- Supports fine-tuning and pre-training



Platform: Samba-1 Composition of Experts (CoE)

- Multi-tenant model platform enables serving of 100s of models/agents on single 8-socket system
- Millisecond switching of fine-tuned models for production inferencing
- Supports direct or auto-route to any of 100s of fine-tuned models
- Dynamic chaining of live agents/models enables dynamic workflow adaptation
- Secure access controls on a per model/agent, per prompt basis

SambaNova SYSTEMS