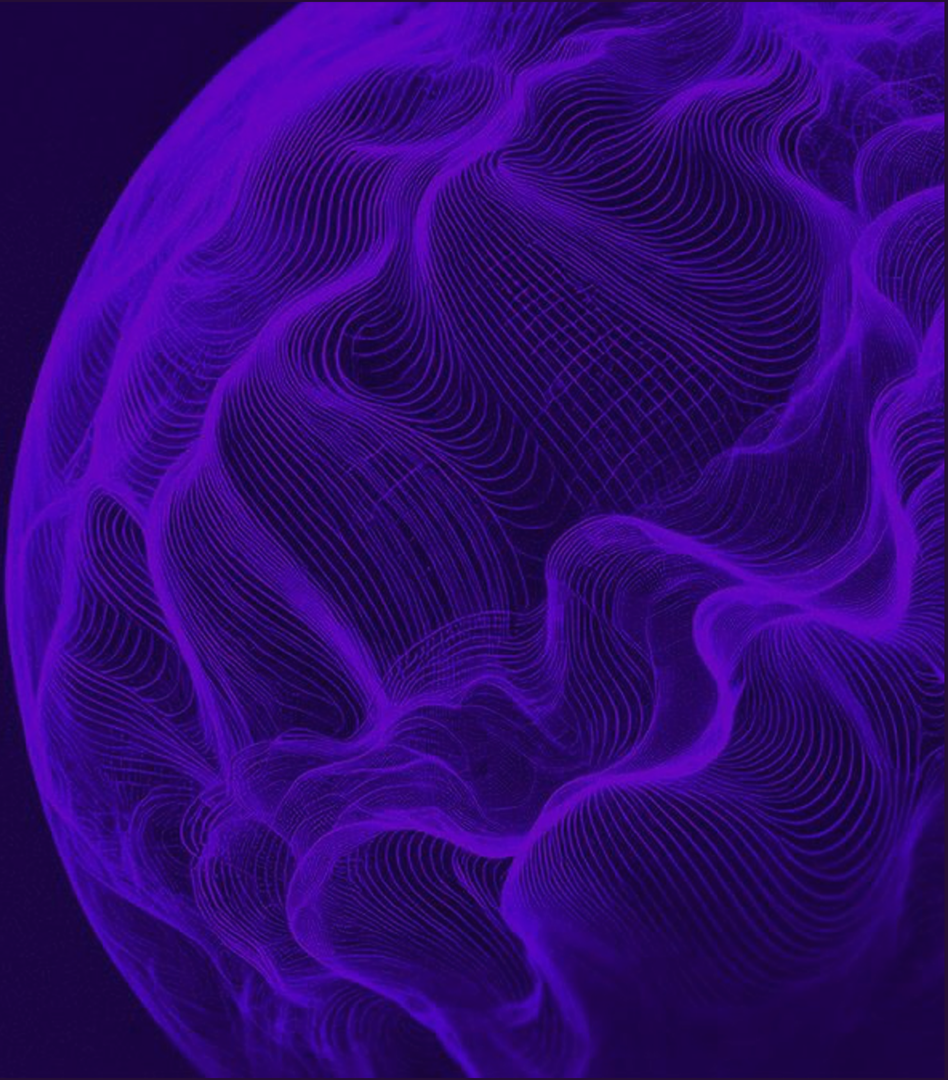
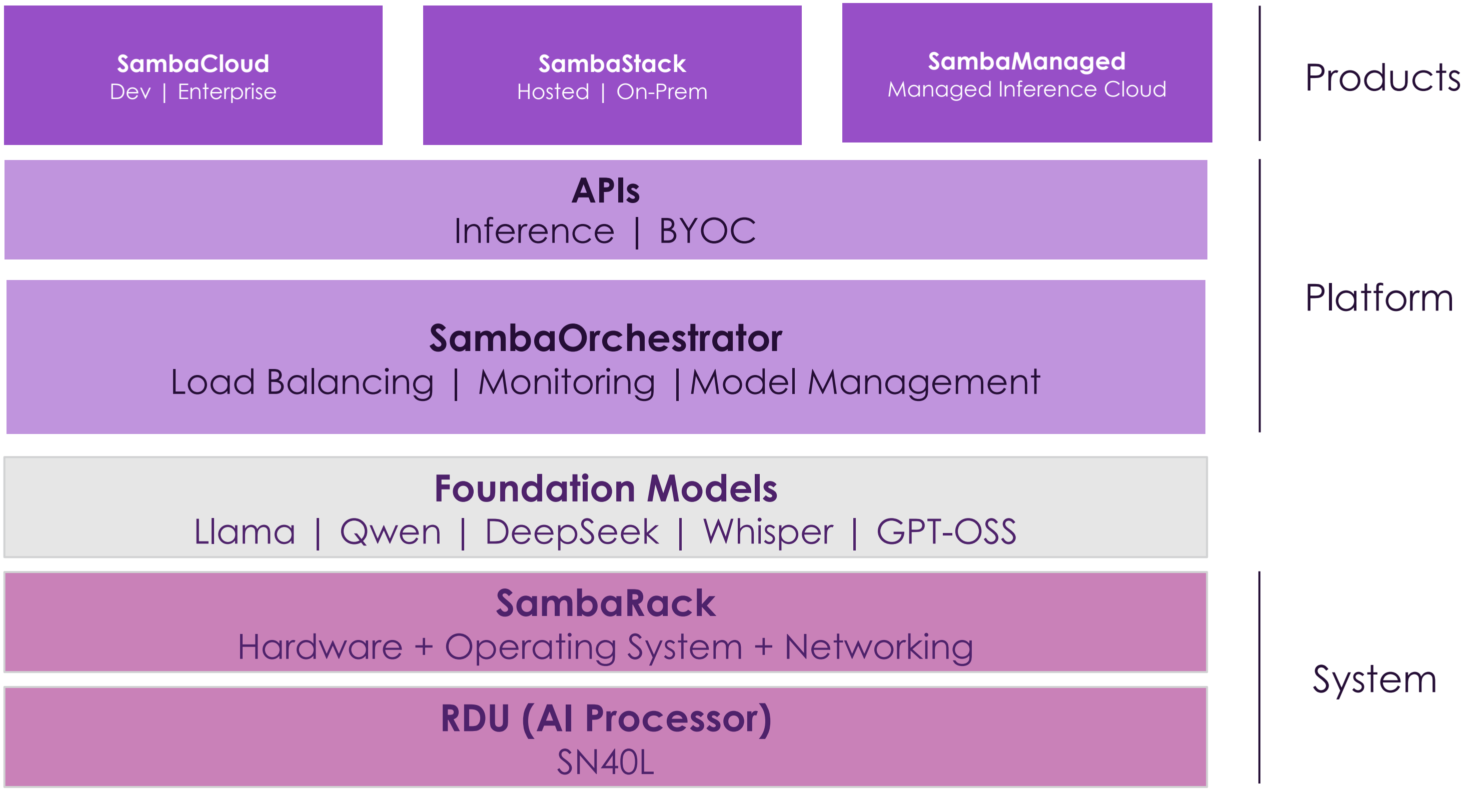
sambanova

LSSDA Workshop

Raghu Prabhakar
Senior Architect



From Chips to Models



SambaCloud
Dev | Enterprise

SambaStack
Hosted | On-Prem

SambaManaged
Managed Inference Cloud

Products

APIs
Inference | BYOC

SambaOrchestrator
Load Balancing | Monitoring | Model Management

Platform

Foundation Models
Llama | Qwen | DeepSeek | Whisper | GPT-OSS

SambaRack
Hardware + Operating System + Networking

RDU (AI Processor)
SN40L

System

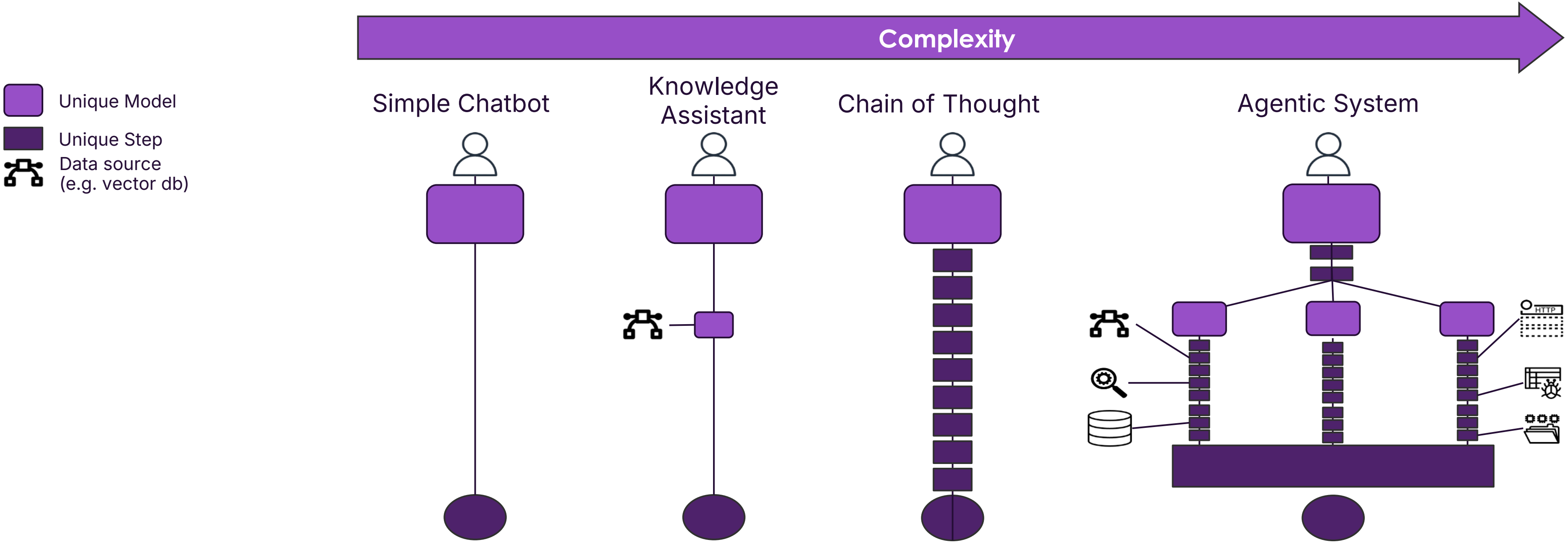
SambaCloud

cloud.sambanova.ai

Meta	Qwen3	deepseek	OpenAI	MISTRAL AI_
Meta Llama 4-Maverick-17B-128E-Instruct Meta Llama 3.3 70B Meta Llama 3.2 1B/3B/70B Meta Llama 3.1 8B Meta Llama Guard 3-8B	Qwen3-32B Qwen QwQ 32B	DeepSeek-R1 0528 DeepSeek V3-0324 DeepSeek-R1-Distill-Llama-70B DeepSeek-V3.1	OpenAI Whisper large v3 OpenAI GPT-OSS-120b	Mistral E5 7B

#1 - #2 in the world on performance on almost all models

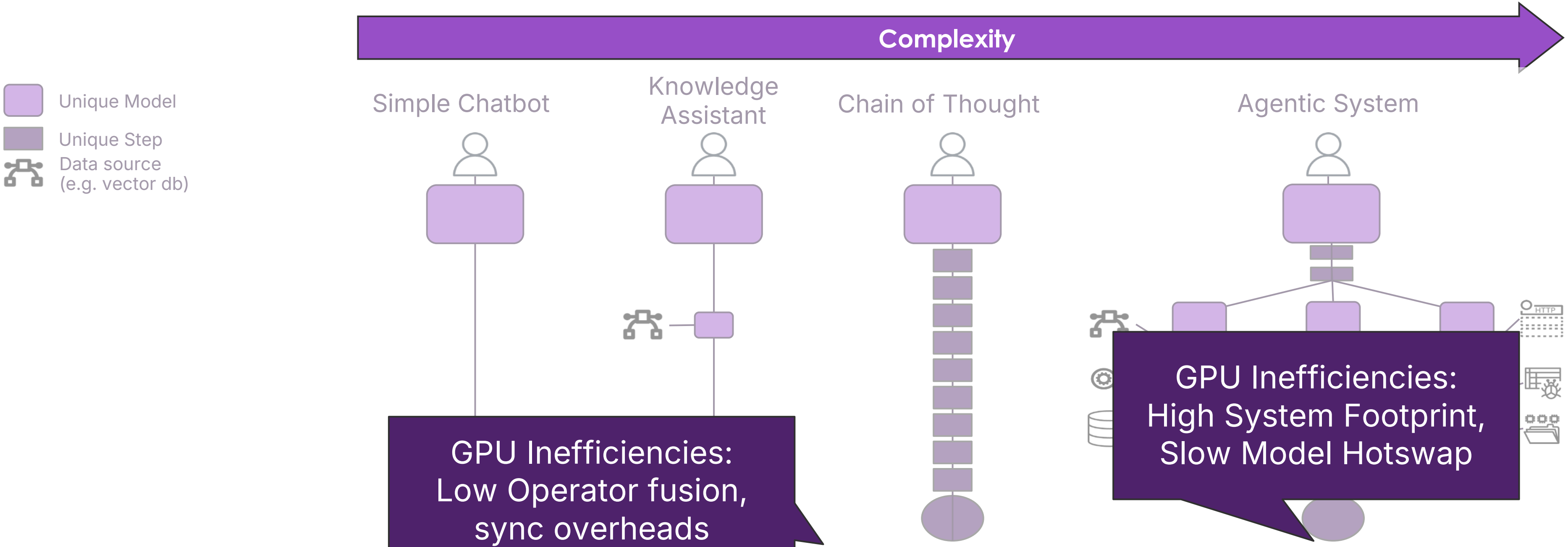
Journey to Agentic AI Systems



Tokens generated per query	1X	1X	10X	Up to 100X
Models Used per query	1	2	1	~5-10
			GPUs are challenged	GPUs fail

sambanova

Journey to Agentic AI Systems



GPU Inefficiencies:
Low Operator fusion,
sync overheads

GPU Inefficiencies:
High System Footprint,
Slow Model Hotswap

Tokens generated per query	1X	1X	10X	Up to 100X
Models Used per query	1	2	1	~5-10
			GPUs are challenged	GPUs fail

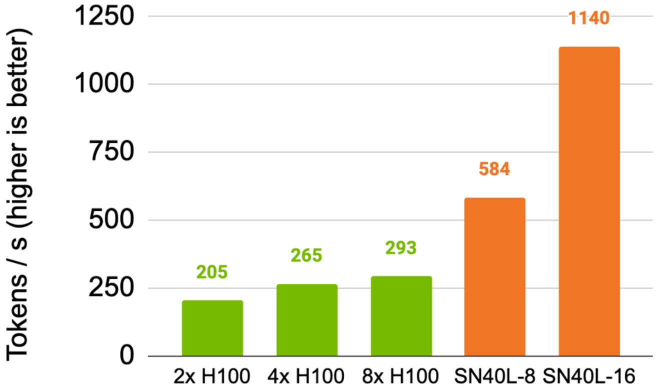
SN40L Reconfigurable Dataflow Unit

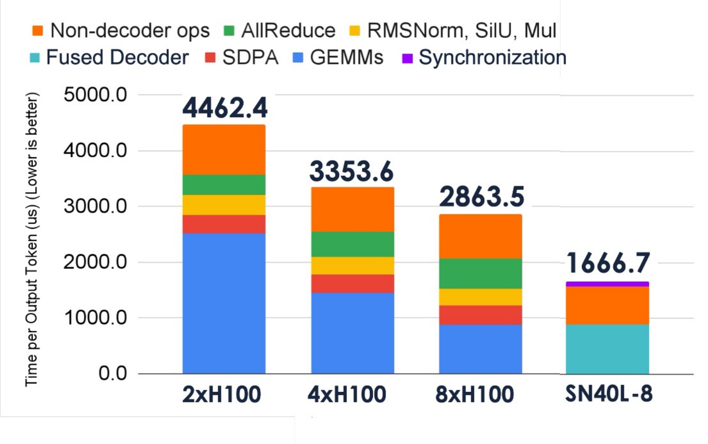
Native multi-tenancy support with fast model switching
Ideal for production inference, multi-tenancy, agentic workflows

- 3-tier Dataflow Memory
- 520 MB On-Chip SRAM Memory → Aggressive Pipelining and Operator Fusion with Dataflow
- 64 GB High Bandwidth Memory → Software-managed cache for low-latency and high-throughput inference
- 1.5 TB High Capacity DDR Memory → Hold and switch between large number of models and contexts in milliseconds

sambanova

Dataflow ⇒ High Performance





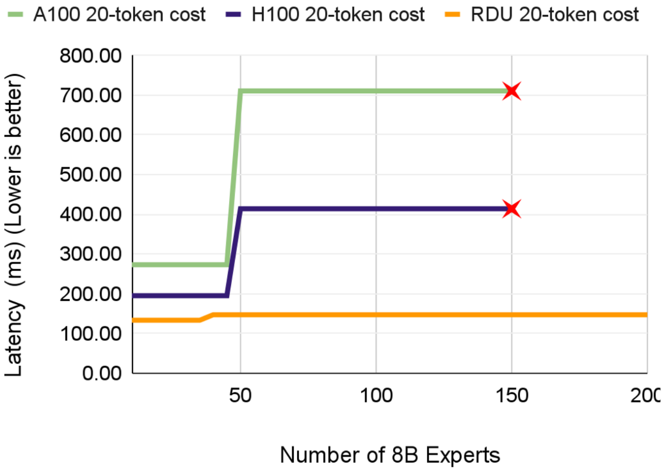
Overlap compute, memory access, chip-to-chip communication

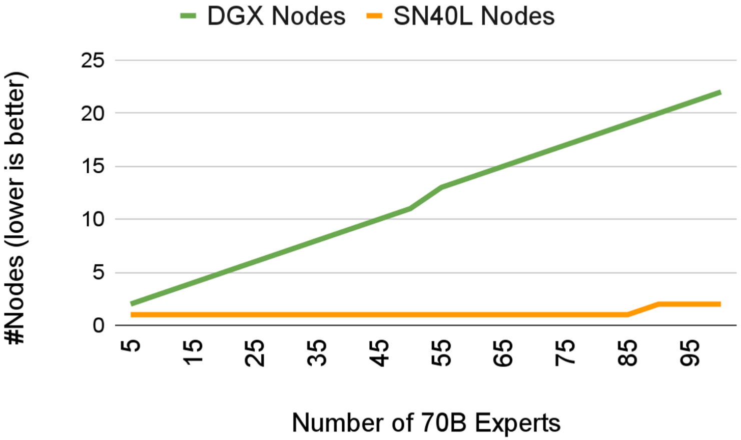
- Fully overlap allreduce with weight load and compute
- Allreduce does not consume HBM capacity or bandwidth

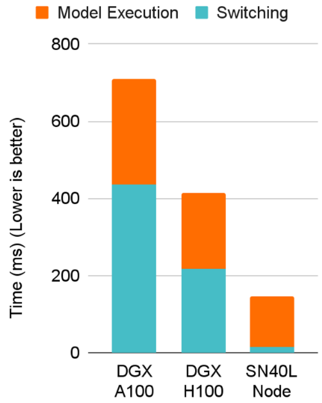
sambanova

3-Tier Memory ⇒ Faster Model Hotswap

Host and Serve Trillions of Parameters at a fraction of the cost of GPUs







3.7x

Faster

vs. DGX H100

19x Smaller

Machine Footprint

15x Faster

Model Switching