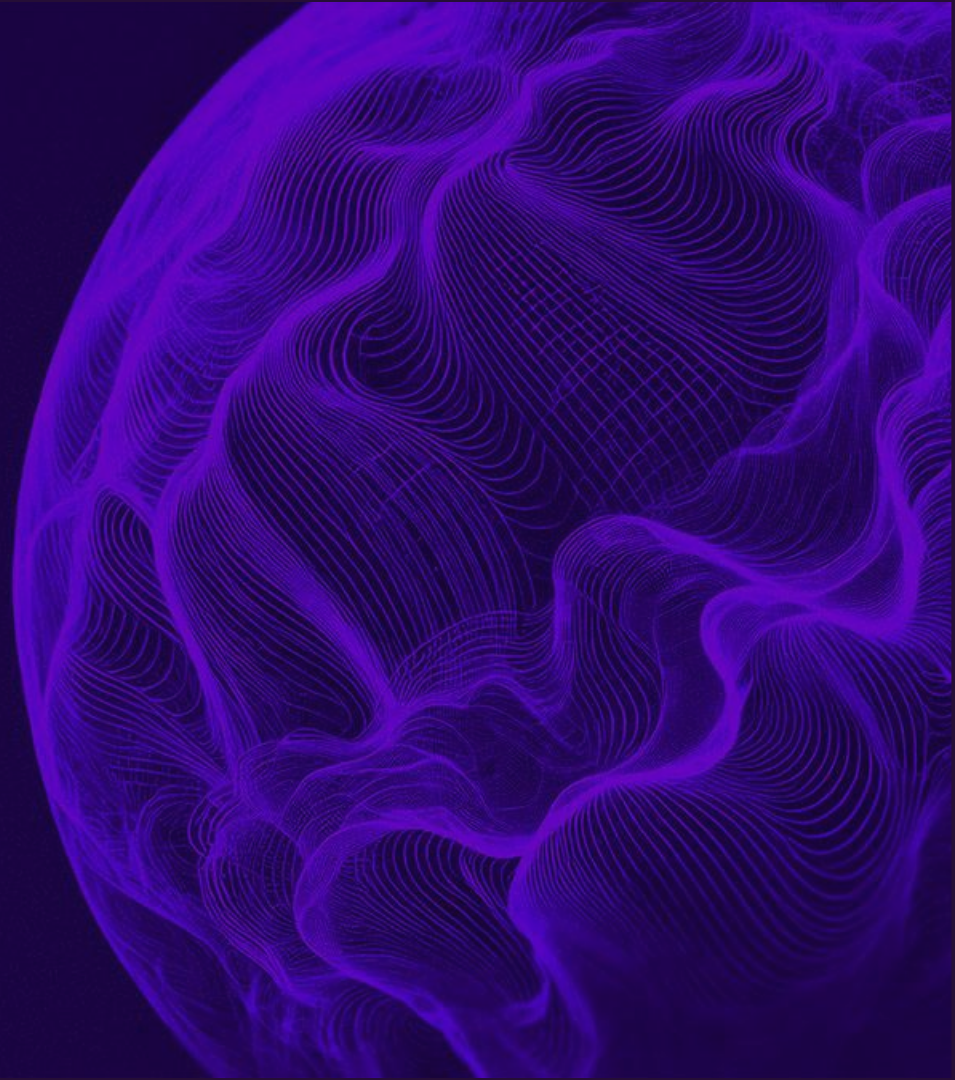


sambanova

Accelerating Agentic and Generative AI with SambaNova

Petro Junior Milan
Principal AI Engineer

Tutorial at Supercomputing 2025
16 November 2025, St. Louis, MO

sambanova

Agenda

1 Background

2 Hardware and Software Architecture Overview

3 On-Premises and Cloud Deployments

4 AI Use Cases

Market: Journey to Agentic AI Systems

Agentic AI increase inference calls per query by 100x

Key

Unique Model

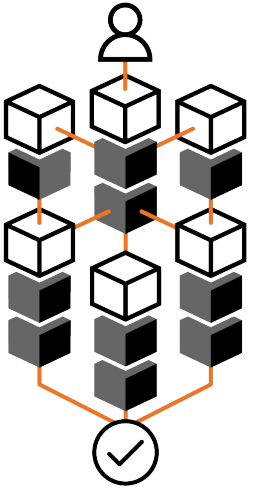
Unique Step

- **Agentic AI can increase inference calls per query by 100x**

Simple Chatbot

Chain of Thought

Agentic System



Tokens generated per query	1X	10X	Up to 100X
Models Used per query	1	1	~5-10
		GPUs consume too many resources	

SambaNova SYSTEMS

sambanova

Big Model Trends

Models are getting bigger

- Llama 2 70B (2023)
- Llama 3 405B (2024)
- DeepSeek R1 671B (2025)
- Llama 4 Behemoth?
- DeepSeek R2?

Context Lengths are getting bigger

- Models were 4K just a few years ago
- Then 128K became the standard
- Now context lengths are growing to over 1M

Models are getting sparser

- Dense models were the norm (Llama 2 & 3 / GPT3)
- Then coarse Mixture of Experts (Mixtral / GPT4 both around 16 experts)
- Then fine grained Mixture of Experts (DeepSeek / Llama 4 both around 128+ experts)

Reasoning is getting unlocked by test time compute

- All the best models today are reasoning models (DeepSeek R1, OpenAI o3)
- These models generate many more tokens, which means token generation speed is more important than ever

sambanova

Who We Are

Snapshot

- Founded in 2017 by industry luminaries and originated at Stanford University
- Fully integrated generative AI platform, from 4th generation hardware to pre-trained models
- $1B+ funding raised
- We are over approx. 400 employees with offices in Palo-Alto (HQ), Austin, London, Bangalore, Tokyo, Singapore, Australia

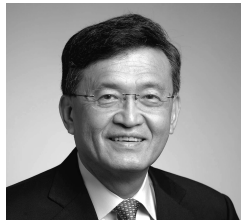
Founded by pioneers in AI

Clyde Hosein
President & COO

Rodrigo Liang
Co-founder & CEO

Kunle Olukotun
Co-founder & Chief Technologist & Stanford Professor

Christopher Ré
Co-founder & Stanford Professor



Lip-Bu Tan
Chairman

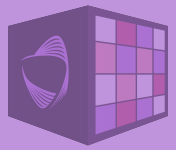
Sophisticated, long-term investors

BlackRock Capital Investment Corporation™ · SoftBank Investment Advisers · TEMASEK · GV · intel Capital · SAMSUNG CATALYST FUND · GIC · Micron · SK telecom · WALDEN INTERNATIONAL · Celesta Capital

Full Stack AI Inference Platform

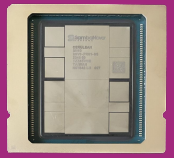
Products:
SambaCloud, SambaStack,
SambaManaged

Model Bundle



Fully Integrated Rack-Level System
SambaRack

SambaNova's AI Accelerator Chip
Reconfigurable Dataflow Unit (RDU)



Performance
Optimized for inference

Sustainability
Most efficient energy consumption

Efficiency
10x fewer racks/chips than the competition

Inference Optimized
Supports custom checkpoints

Scalability Optimized
The building block for AI clouds and supercomputers

Rack Optimized
Comprehensive rack-level solution

sambanova

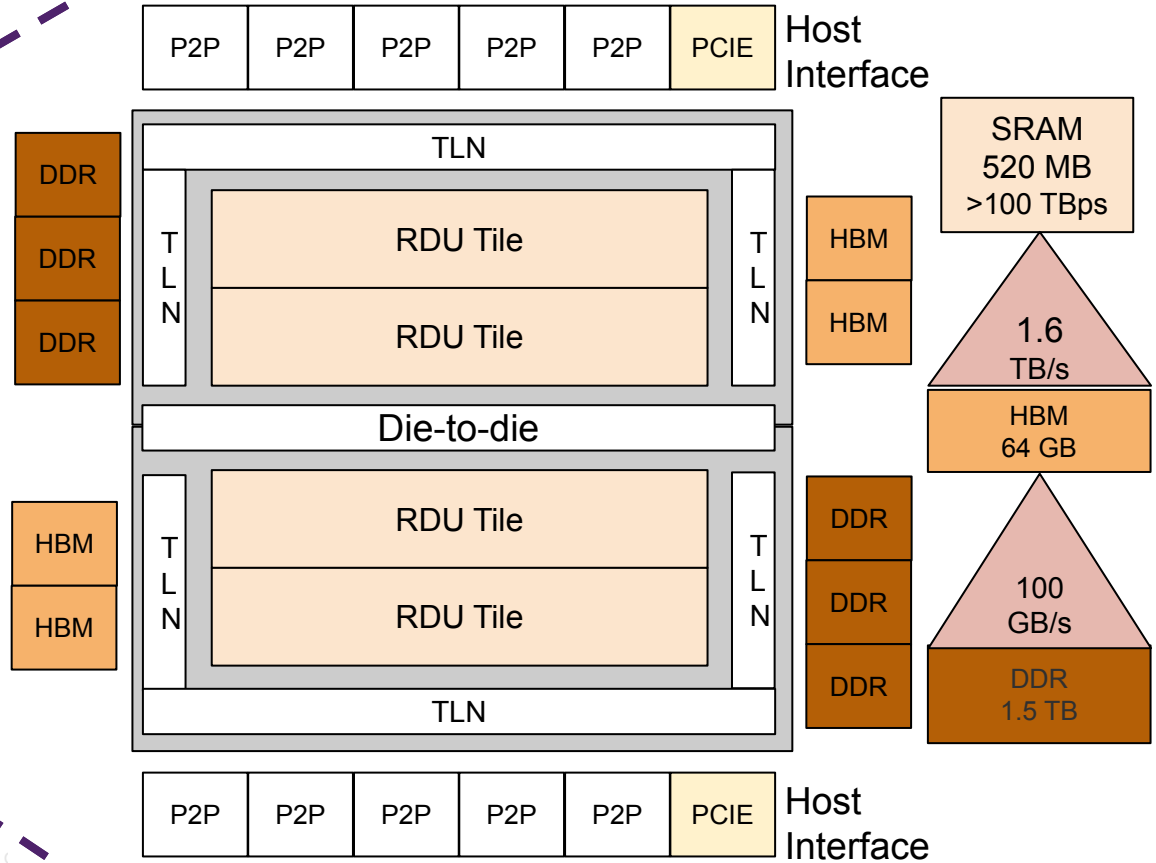
SN40L: SambaNova's 4th Gen AI Chip

Reconfigurable Dataflow Unit (RDU)
Native multi-tenancy support with fast model switching
Ideal for production inference, multi-tenancy, agentic workflows

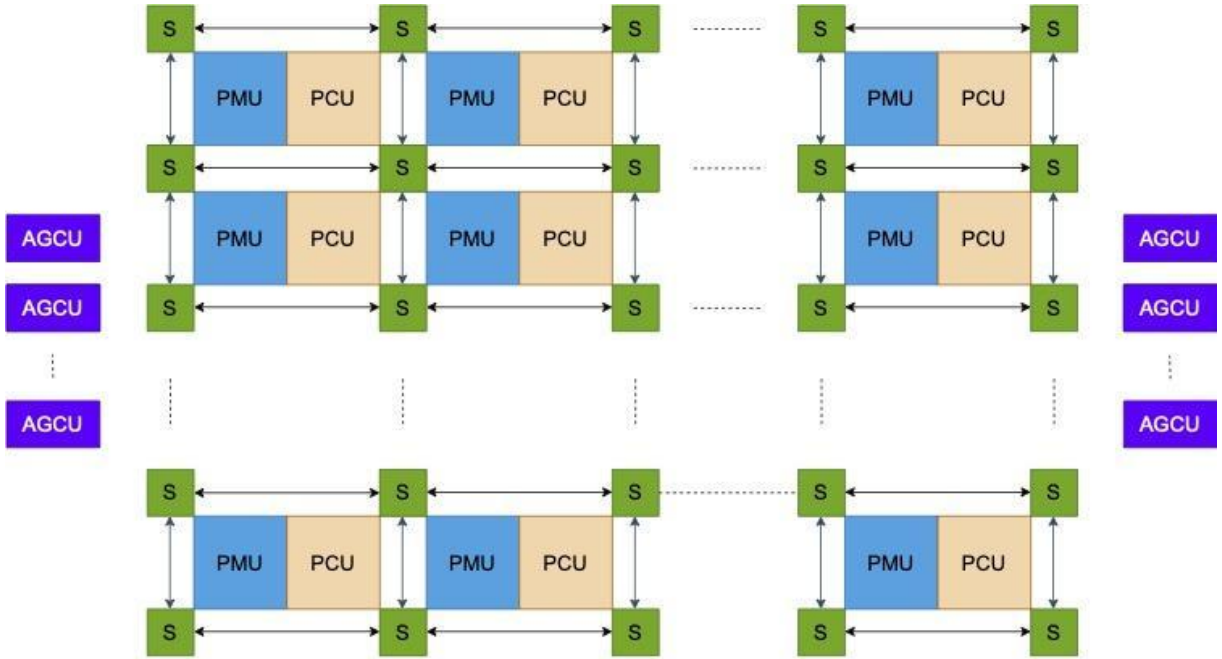
- 3-tier Dataflow Memory
- 520 MB On-Chip SRAM Memory → Very fast memory for high speed inference with caching
- 64 GB High Bandwidth Memory → Switch between models in as little as 2 milliseconds
- 1.5 TB High Capacity DDR Memory → Hold large number of models in memory

sambanova

SN40L Chip: Overview



SN40L: Tile Architecture



1040 PCUs and PMUs

PCU: Compute unit

PMU: Memory unit

S: Mesh switches

AGCU: Portal to off-chip memory and IO

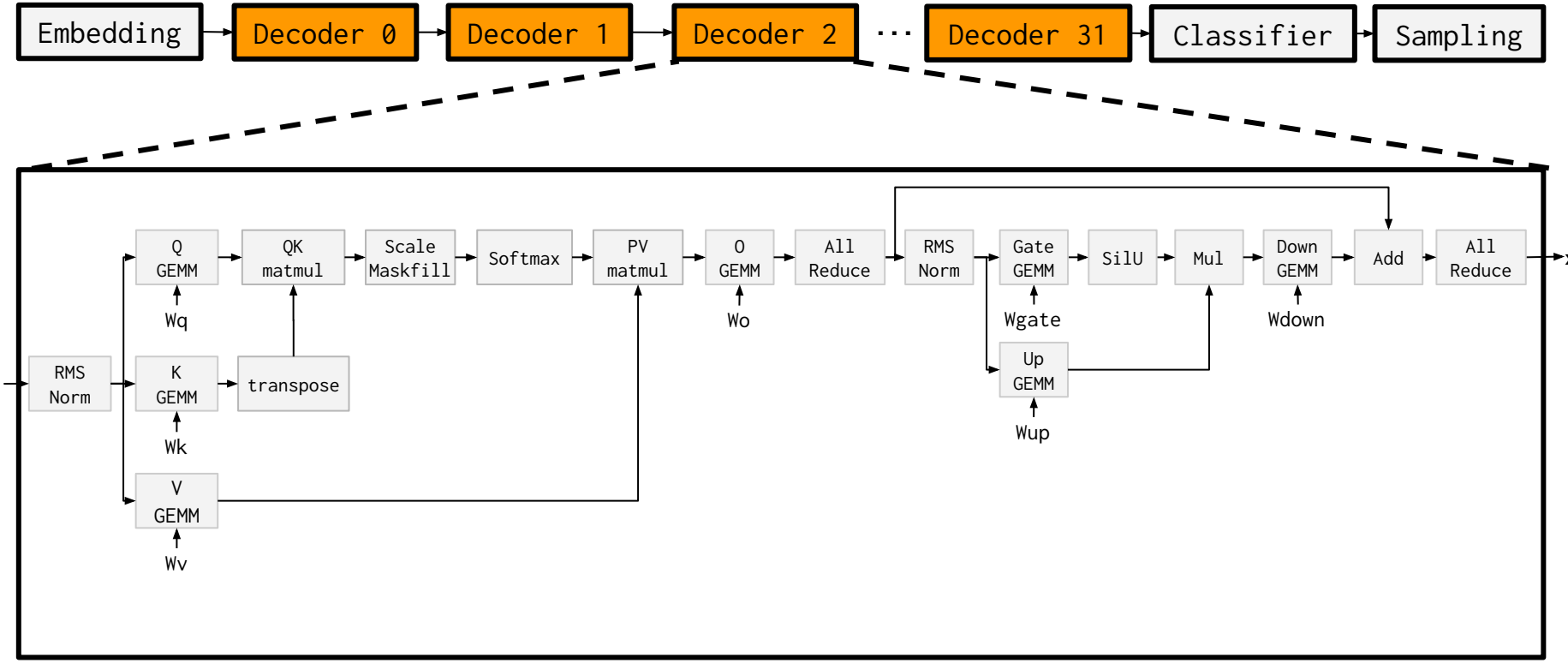
sambanova

Structure of a Transformer

Example: Llama3.1 8B

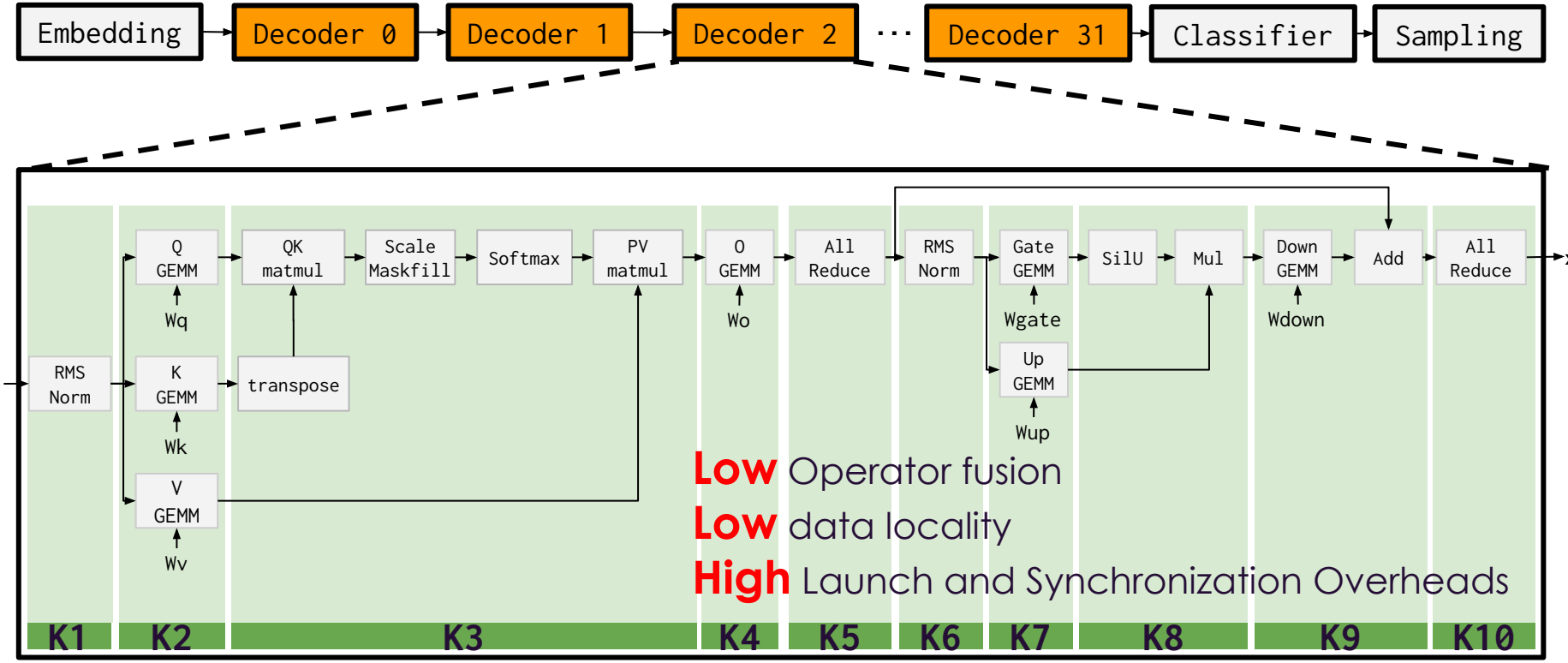
Structure of a Transformer

Example: Llama3.1 8B



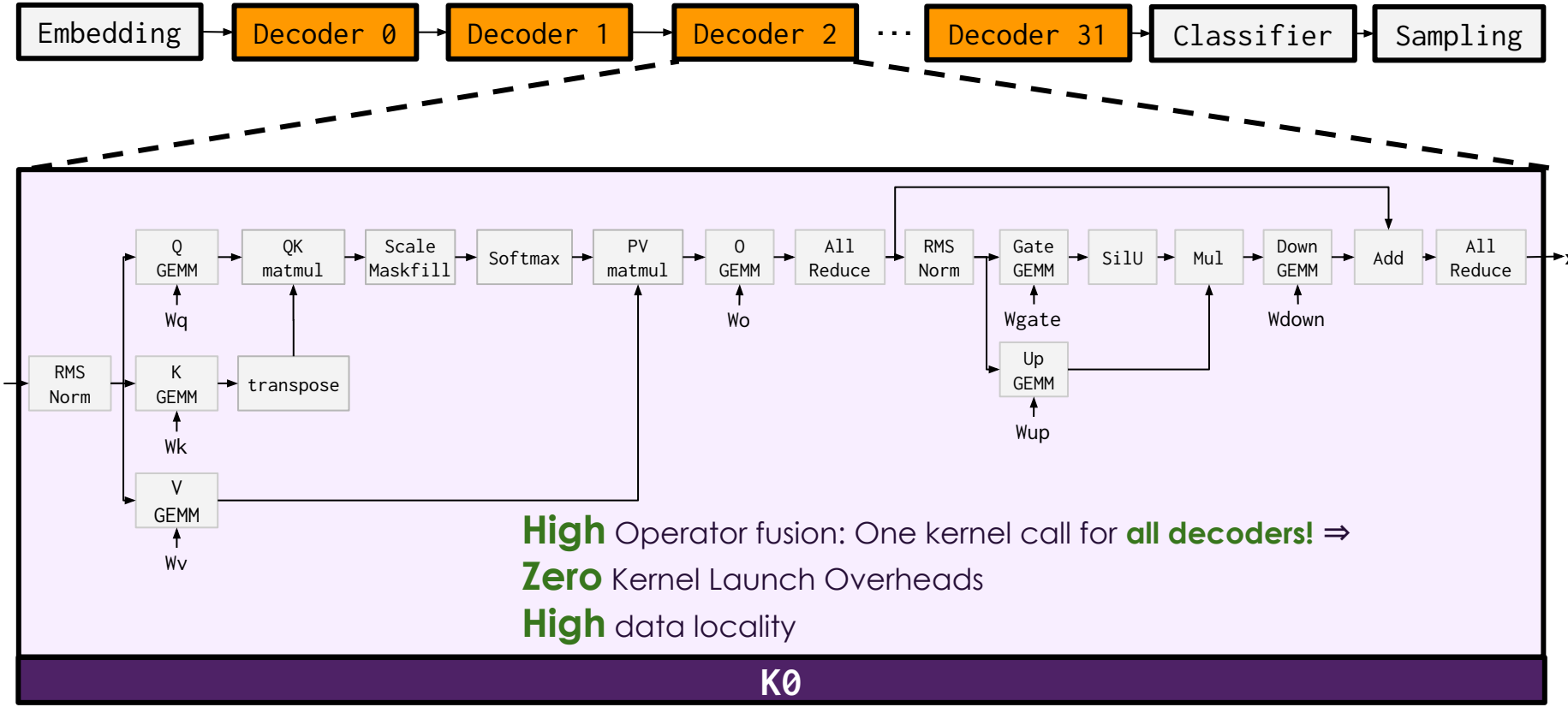
GPUs Incur Overhead with Low Operator Fusion

Example: Llama3.1 8B



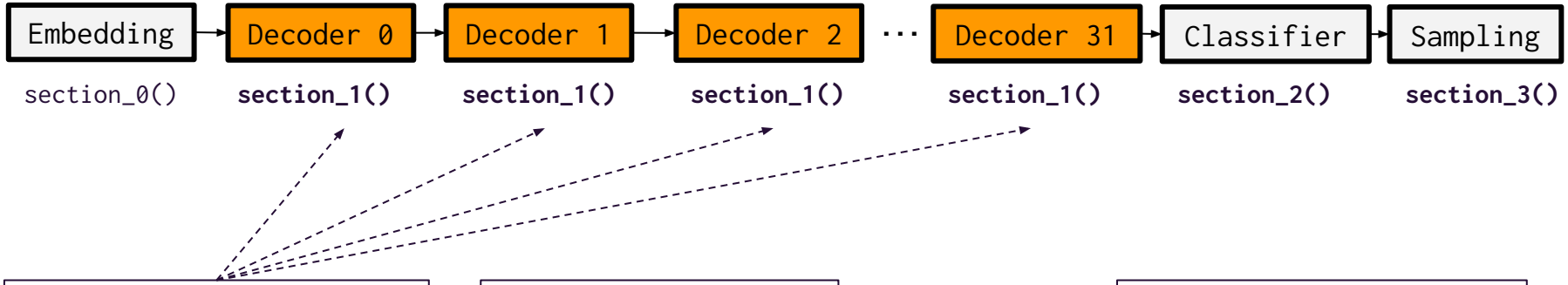
SN40L RDU Fuses the Entire Decoder into One Kernel!

Example: Llama3.1 8B



RDU Hardware Graph Orchestration: Temporal Fusion

Example: Llama3.1 8B



Single Section Decoder
+
Repeated Section Calls
=
New fusion opportunities!

```
section_1(arg):
  Foo;
  Bar;

// Repeat 32 times
section_1(arg0)
section_1(arg1)
…
section_1(arg31)
```

Kernel Looping

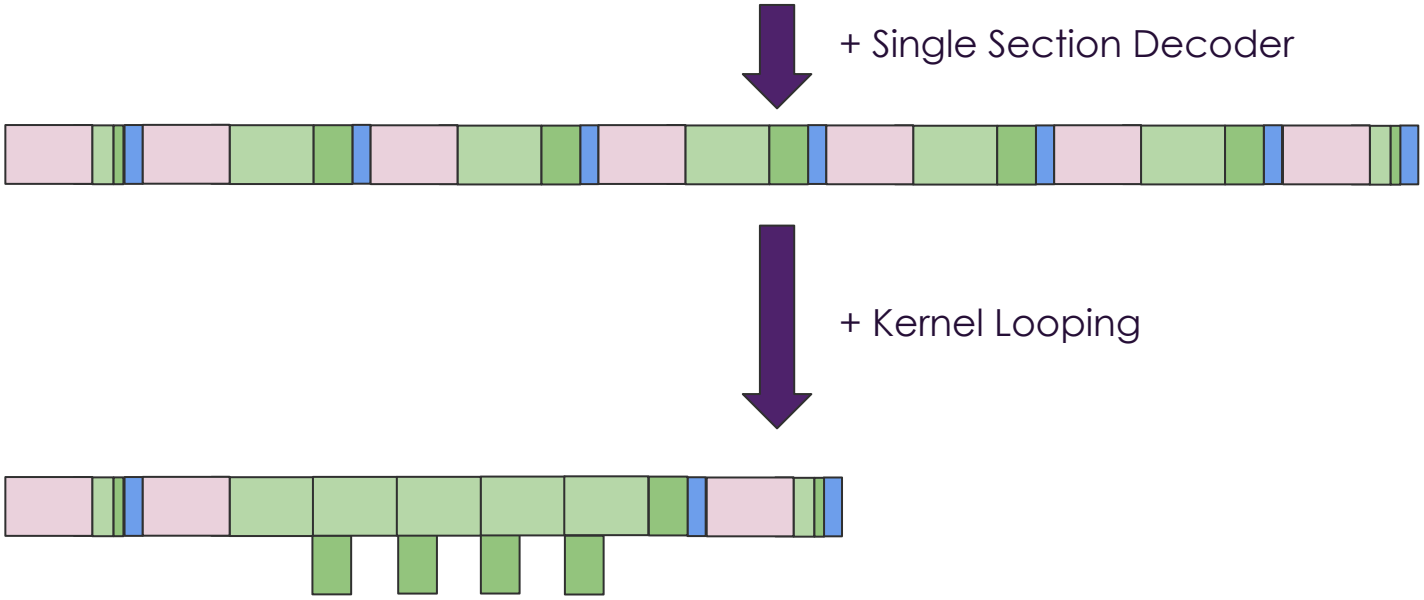
```
section_1(args[]):
  for i in range(32):
    Foo;
    Bar;

// 1 time
section_1(args)
```

1 kernel launch for all decoders!

Executing Transformer on RDU

Example: Llama3.1 8B

Baseline: **100** tokens/s

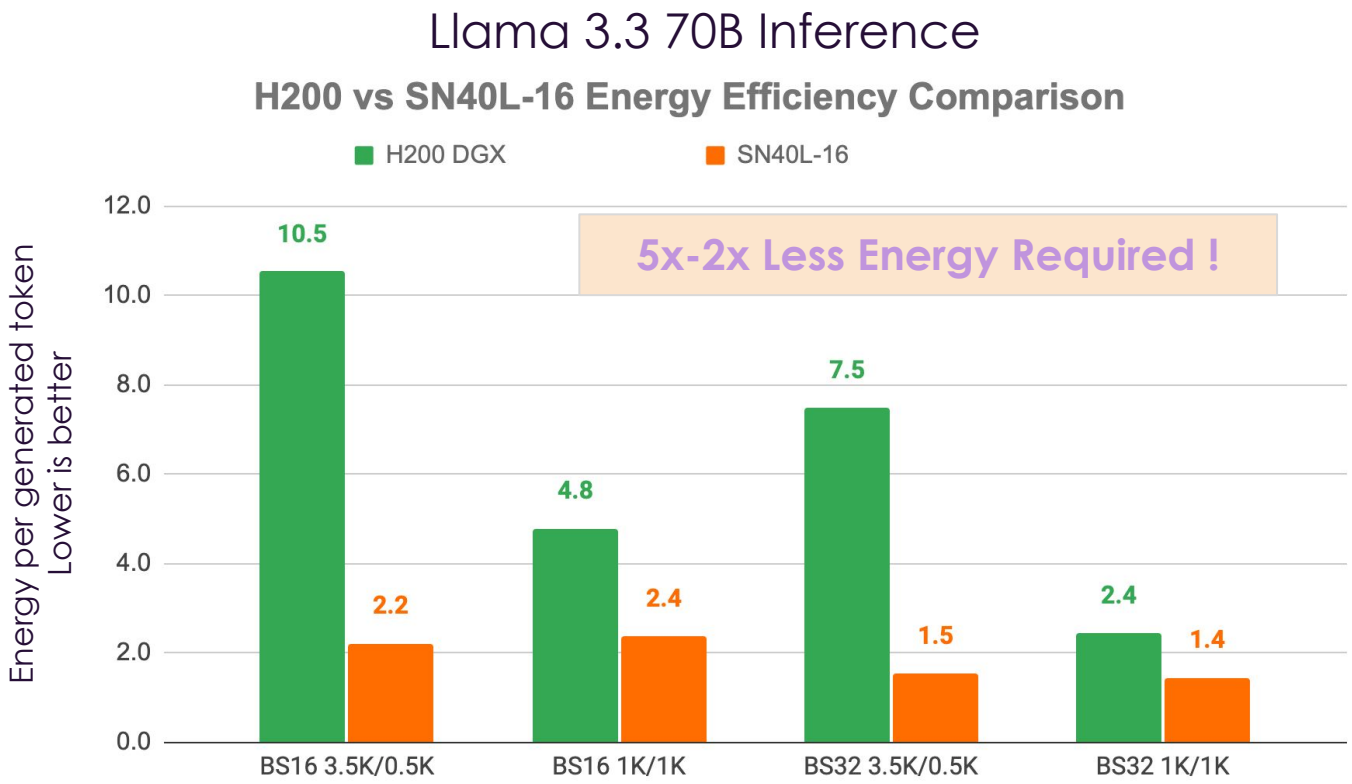


= **500** tokens/s
16 SN40L

= **1170** tokens/s
16 SN40L

sambanova

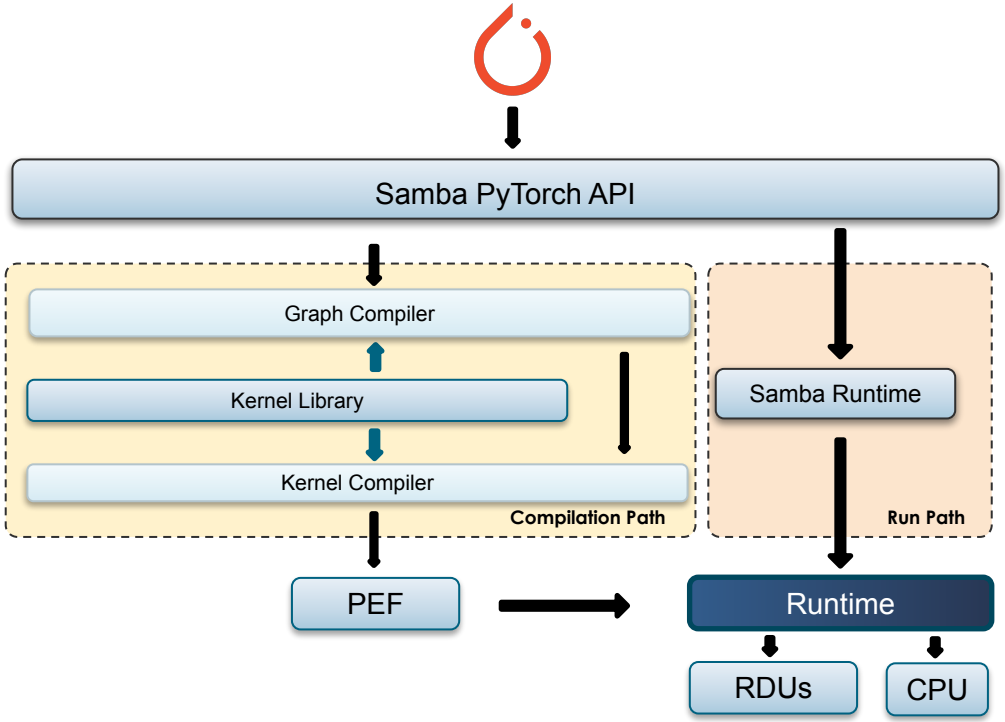
SN40L-16 vs H200 DGX Llama3.3 Inference Energy Efficiency



sambanova

Samba Compilation Flow

- **Samba**
  - SambaNova PyTorch compilation & run APIs
- **Graph compiler**
  - High-level ML graph transformation & optimizations
- **Kernel compiler**
  - Low-level RDU operator kernel transformation & optimizations
- **Kernel library**
  - RDU operator implementations



sambanova

SambaStack

Inference on SambaStack has a variety of features to serve your various use cases:

Model Bundling:

- Combine multiple models in a single OpenAI API compatible inference endpoint
- Run all models simultaneously, indicating the target model in each request

Speculative Decoding Pairs:

- Combine a main model with a draft model in a single endpoint to increase inference speeds

Bring Your Own Checkpoint (BYOC):

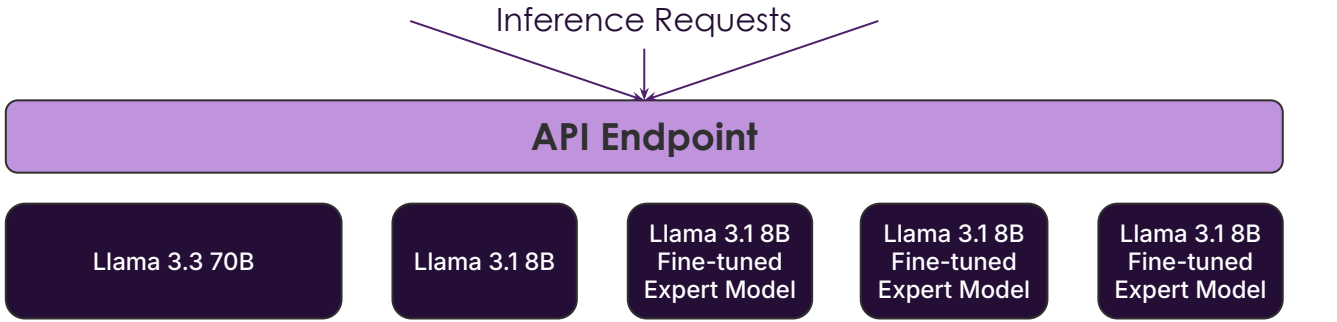
- Import external model checkpoints and run inference on them
- Imported model checkpoints can be externally fine-tuned or base models from Hugging Face

Self-Managed Through Standard Open-Source Tools:

- Kubernetes v 1.30+, Helm v 3.8.0+, and others

sambanova

More Models with Less Hardware



- Reduced startup costs
- Reduced serving costs - less HW more Models
- Simplify complex workflows
- Adding and maintaining new experts

sambanova

AI Computing at Scale in Multiple US National Laboratories

- Apply large language models to complex science problems
- Code generation
- Trustworthiness and security
- Drug discovery
- Climate science
- Brain mapping
- Physics simulations
- Cancer research

Refs: [1, 2, 3, 4, 5, 6, 7, 8]

sambanova

SambaCloud

sambanova

SambaCloud

Open source models available today

Meta	Qwen3	deepseek	OpenAI	MISTRAL AI_
Llama 4-Maverick Llama-3.3-70B-Instruct Llama-3.3-Swallow-70B-Instruct-v0.4 Llama-3.1-8B-Instruct	Qwen3-32B	DeepSeek-R1 0528 DeepSeek-V3.1 DeepSeek-V3.1-Terminus DeepSeek V3-0324 DeepSeek-R1-Distill-Llama-70B	GPT-OSS-120B Whisper large v3	Mistral E5 7B

SambaNova is delivering the largest and most capable open source models, with unparalleled performance to unlock new capabilities that have been impossible to achieve

sambanova

SambaCloud

Over 5-10x faster tokens/sec/request, Full accuracy

Model	SambaNova
Llama 3.1 8B	916
Llama 4 Maverick	704
OpenAI's gpt-oss 120B	543
DeepSeek V3-0324	250
DeepSeek R1-0528	173

Independently Benchmarked by Artificial Analysis

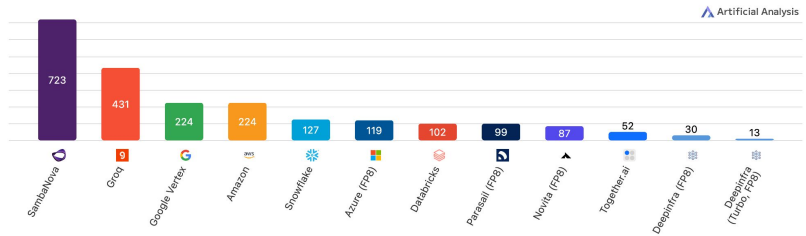
*Speeds benchmarked on Nvidia GPUs on Azure

Lightning Fast Inference for the Largest Models

Llama 4 Maverick

Output Speed: Llama 4 Maverick Providers

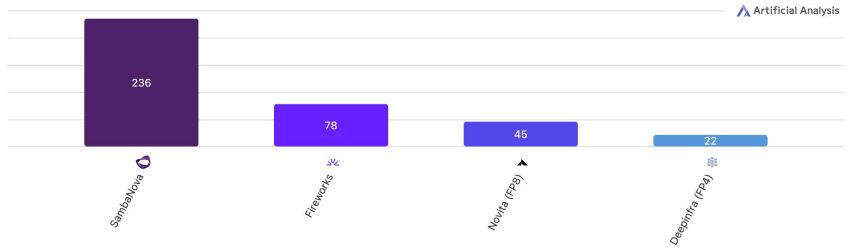
Output Tokens per Second; Higher is better; 1,000 Input Tokens



DeepSeek V3.1 Terminus

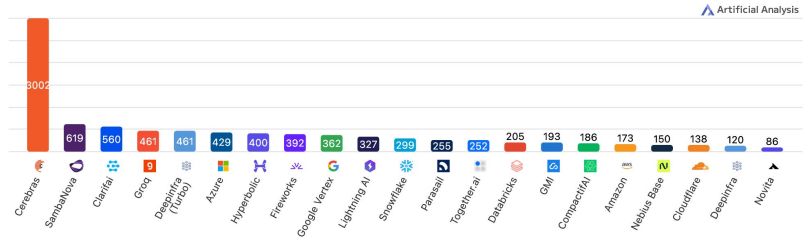
Output Speed: DeepSeek V3.1 Terminus Providers

Output Tokens per Second; Higher is better; 1,000 Input Tokens



Output Speed: gpt-oss-120B (high) Providers

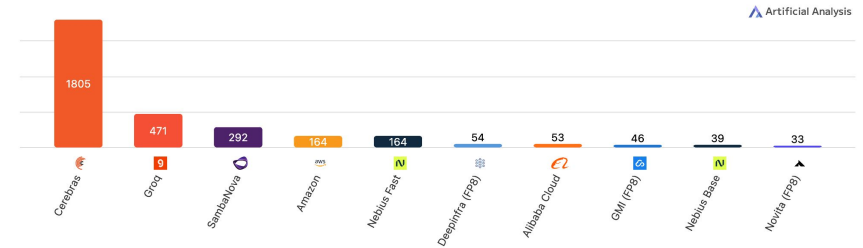
Output Tokens per Second; Higher is better; 1,000 Input Tokens



GPT OSS 120B

Output Speed: Qwen3 32B Providers

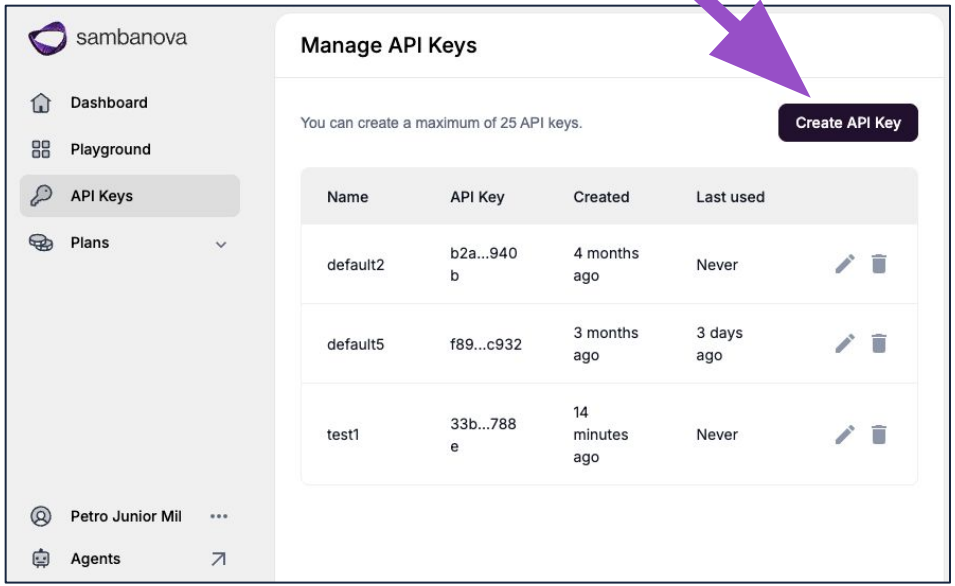
Output Tokens per Second; Higher is better; 1,000 Input Tokens

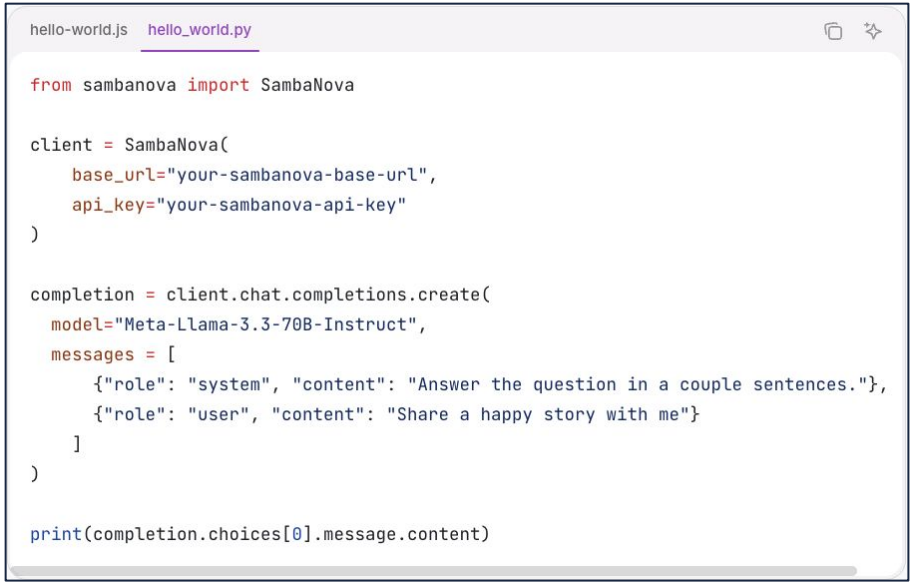


Qwen3 32B

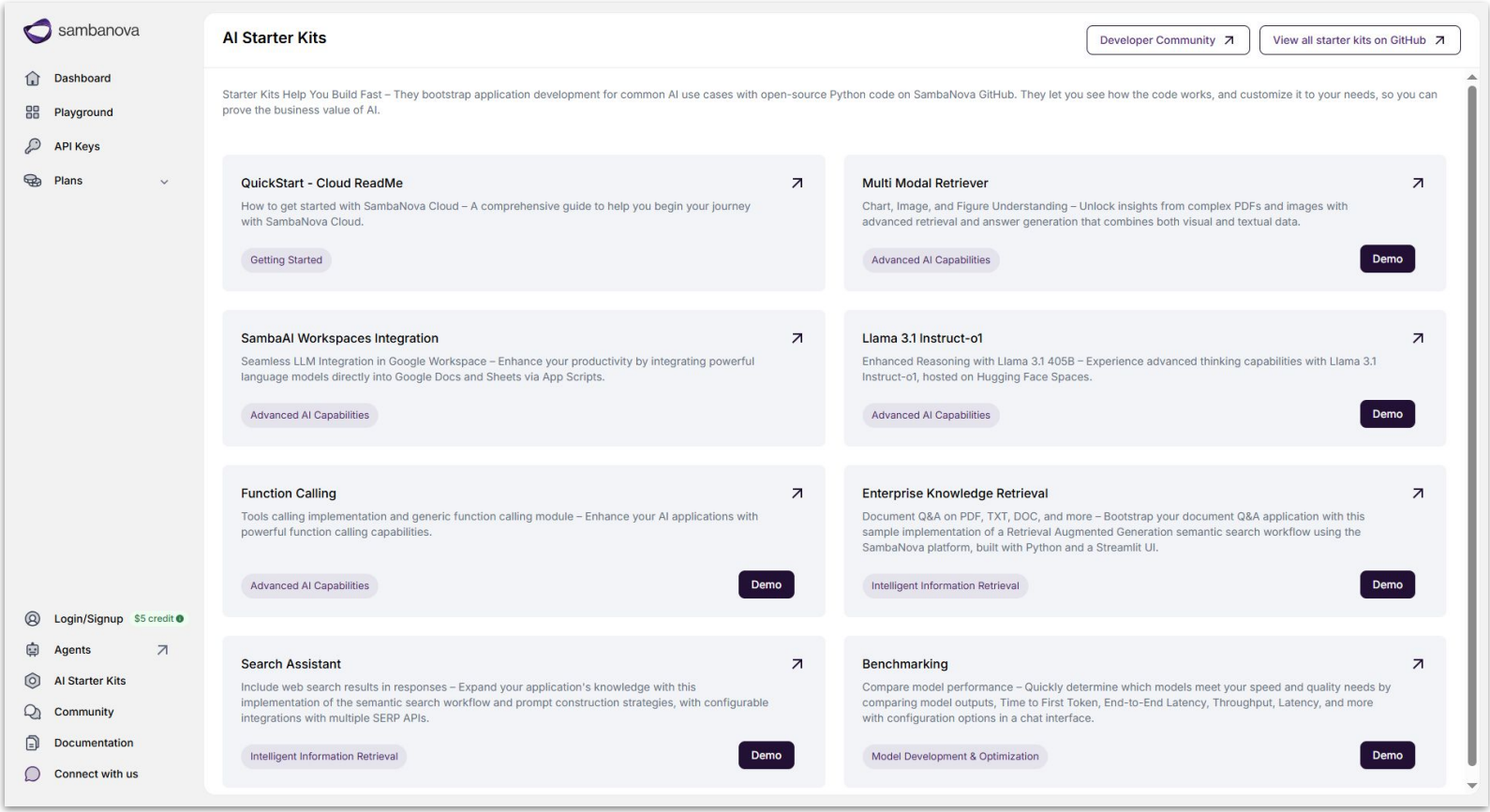
sambanova

Get Your API Key at cloud.sambanova.ai





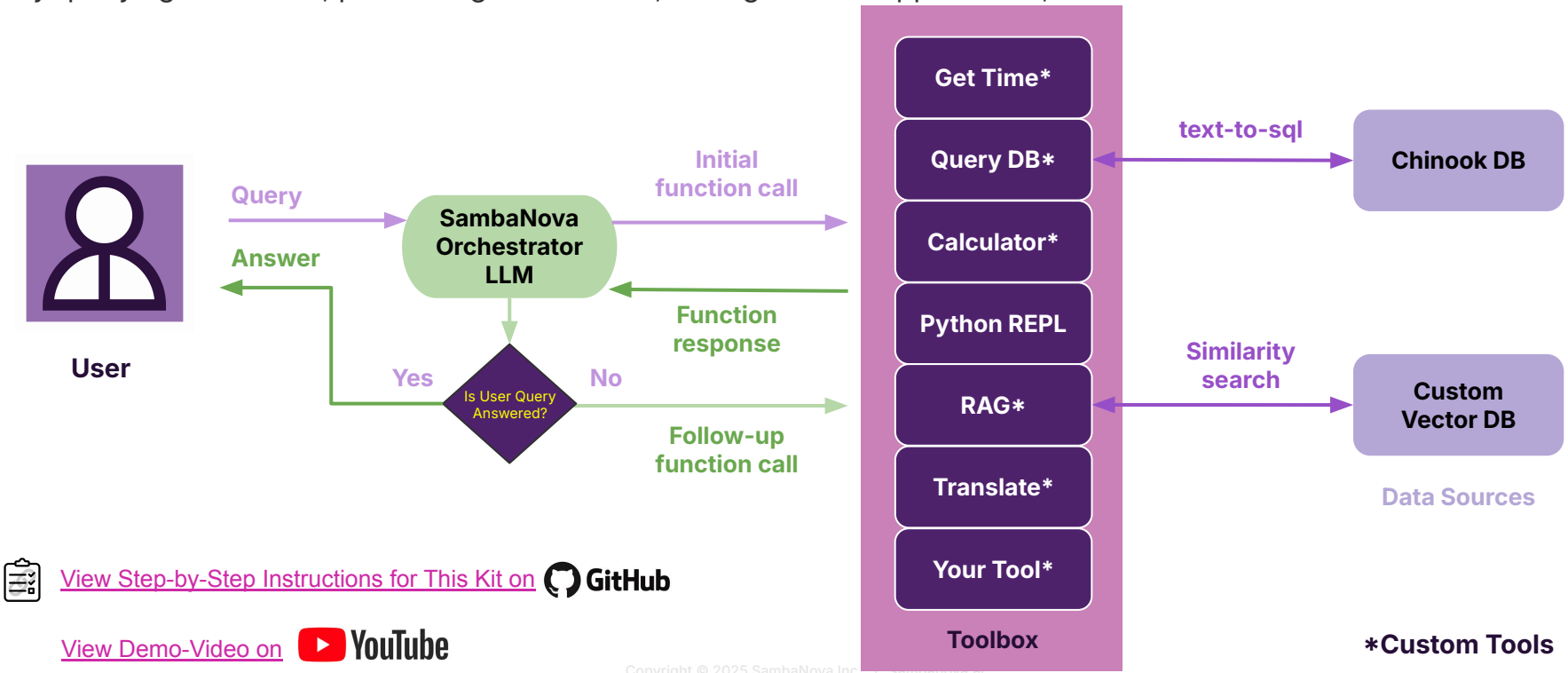
sambanova



sambanova

Tool/Function Calling

Easily enable business use cases by adding function calls into NLP applications that do more than understand text by querying databases, performing calculations, talking to other applications, and more.



View Step-by-Step Instructions for This Kit on GitHub

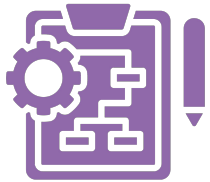
View Demo-Video on YouTube

sambanova

Agentic Applications

Input Objective

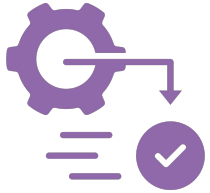
User provides high-level goals



Planning

Agent decomposes goals into tasks

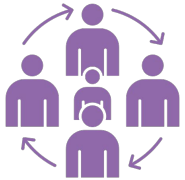
1 Large LLM



Execution

Agent performs or delegates tasks

Many Small LLMs



Feedback and Learning

System improves through iteration

✅**Reduces manual intervention**

✅**Increases efficiency and speed**

✅**Enables complex multi-step task automation**

Agents Demo (Try it yourself!)

sambanova

Integration with External Frameworks / Platforms

LLM Frameworks	Hugging Face, LangChain, LiteLLM, LlamaIndex
Agent Frameworks	Agno, LangGraph, crewai, Semantic Kernel
Coding Assistants	CLINE V3.4, CURSOR, CLAUDE CODE, BLACKBOX
Low-code platforms, vector DBs, and voice libraries	Langflow, milvus, hume, Eleven Labs

For complete list, visit our Integrations page

sambanova

THANK YOU

petro-junior.milan@sambanova.ai

sambanova

Try It for free!

https://cloud.sambanova.ai

Papers about SN40L

https://arxiv.org/abs/2405.07518

"SambaNova SN40L: Scaling the AI Memory Wall with Dataflow and Composition of Experts", MICRO '24

https://arxiv.org/abs/2410.23668

"Kernel Looping: Eliminating Synchronization Boundaries for Peak Inference Performance"