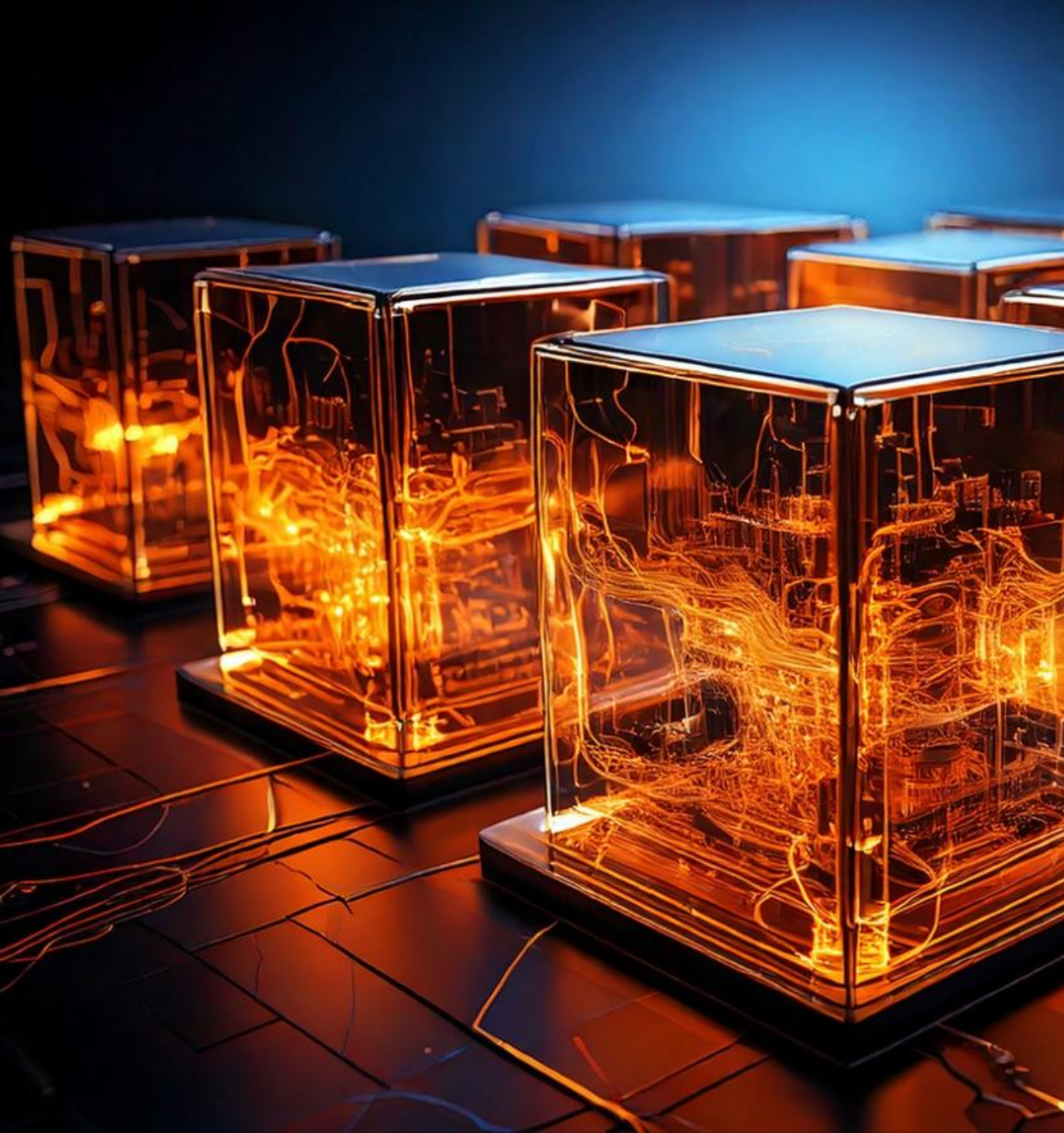
SambaNova SYSTEMS

Accelerating Agentic and GenAI with SambaNova Systems

Petro Junior Milan
Principal AI Engineer

Tutorial at ISC High Performance 2025
13 June 2025, Hamburg, Germany



Agenda

1. Background
2. Hardware and Software Architecture Overview
3. On-Premises and Cloud Deployments
4. AI Use Cases

SambaNova SYSTEMS

Market: Journey to Agentic AI Systems

Agentic AI increase inference calls per query by 100x

Key

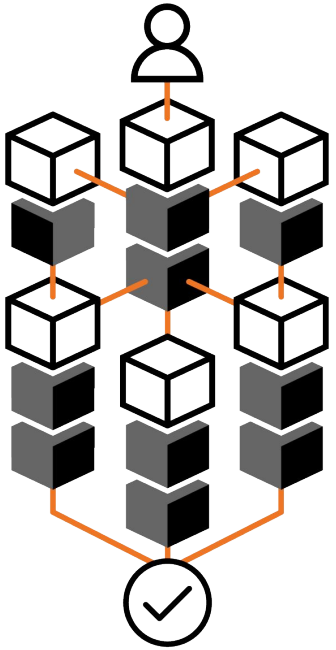
Unique Model

Unique Step

Simple Chatbot

Chain of Thought

Agentic System



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

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

SambaNova SYSTEMS

Big Model Trends

Models are getting bigger

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

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)

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

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 SYSTEMS

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

Founded by pioneers in AI

Lip-Bu Tan
Executive Chairman

Rodrigo Liang
Co-founder & CEO

Kunle Olukotun
Co-founder & Chief Technologist & Stanford Professor

Christopher Ré
Co-founder & Stanford Professor

Sophisticated, long-term investors

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

SambaNova SYSTEMS

How Do We Solve: Full Stack AI Platform

Model Bundling
Composition of Experts

Systems Management:
Inference Endpoint Deployment
SambaNova Suite, SambaNova Cloud

Fully Integrated Rack-Level System
DataScale

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 training and fine tuning

Scalability Optimized
The building block for AI clouds and supercomputers

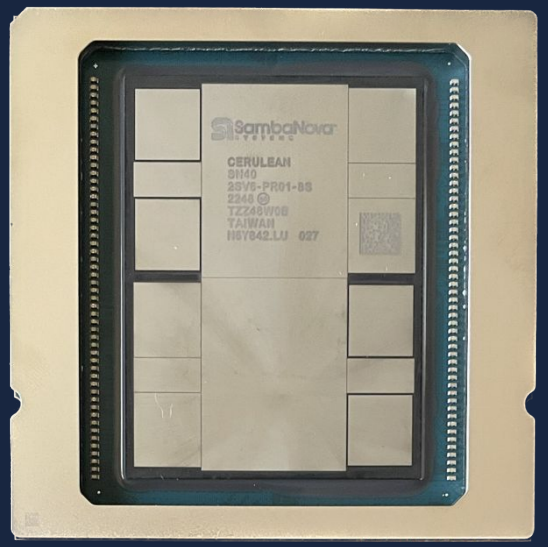
Rack Optimized
Comprehensive rack-level solution

SambaNova® SYSTEMS

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

SambaNova SYSTEMS
Cerulean SN40L RDU

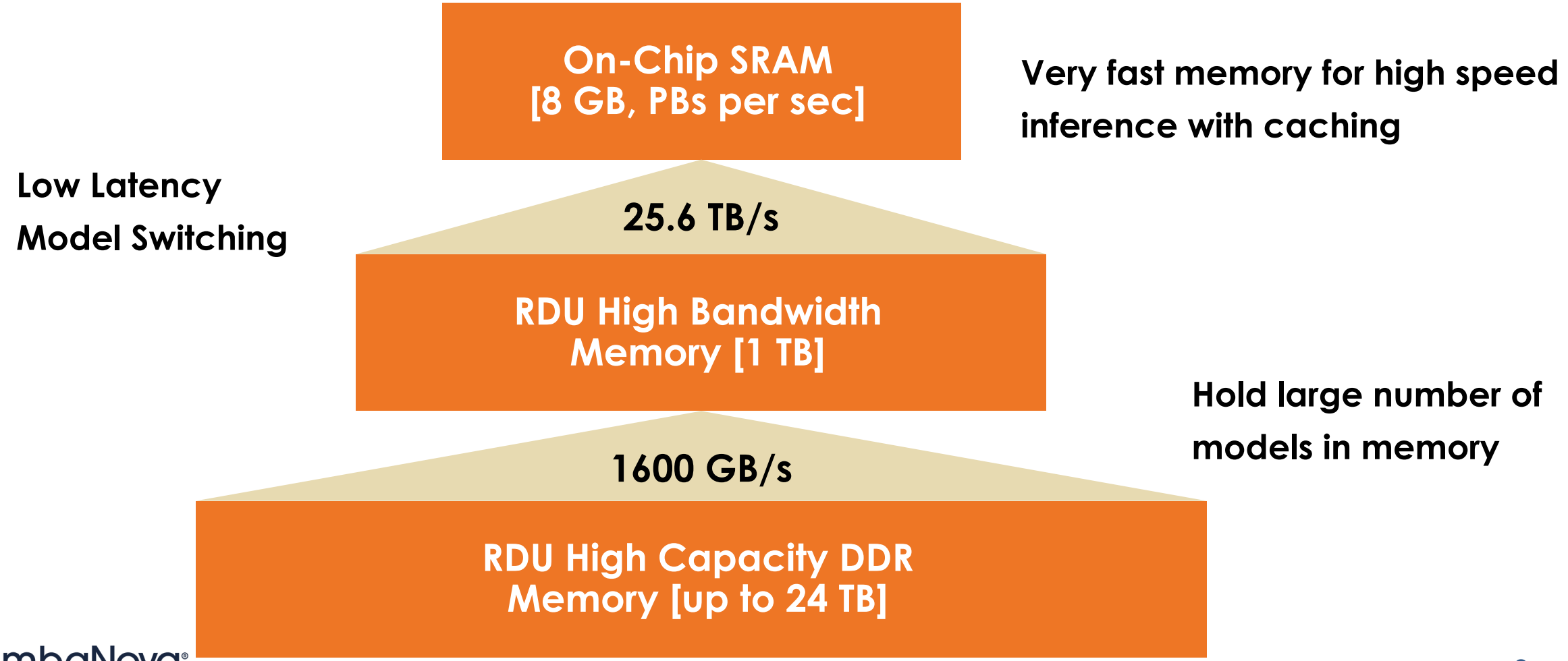


- **5nm TSMC**
- 102B Transistors
- 1,040 RDU Cores
- 638 TFLOPS (bf16)
- **3-tier Dataflow Memory**
- 520 MB On-Chip Memory
- 64 GB High Bandwidth Memory
- Up to 1.5 TB High Capacity Memory

SambaNova SYSTEMS

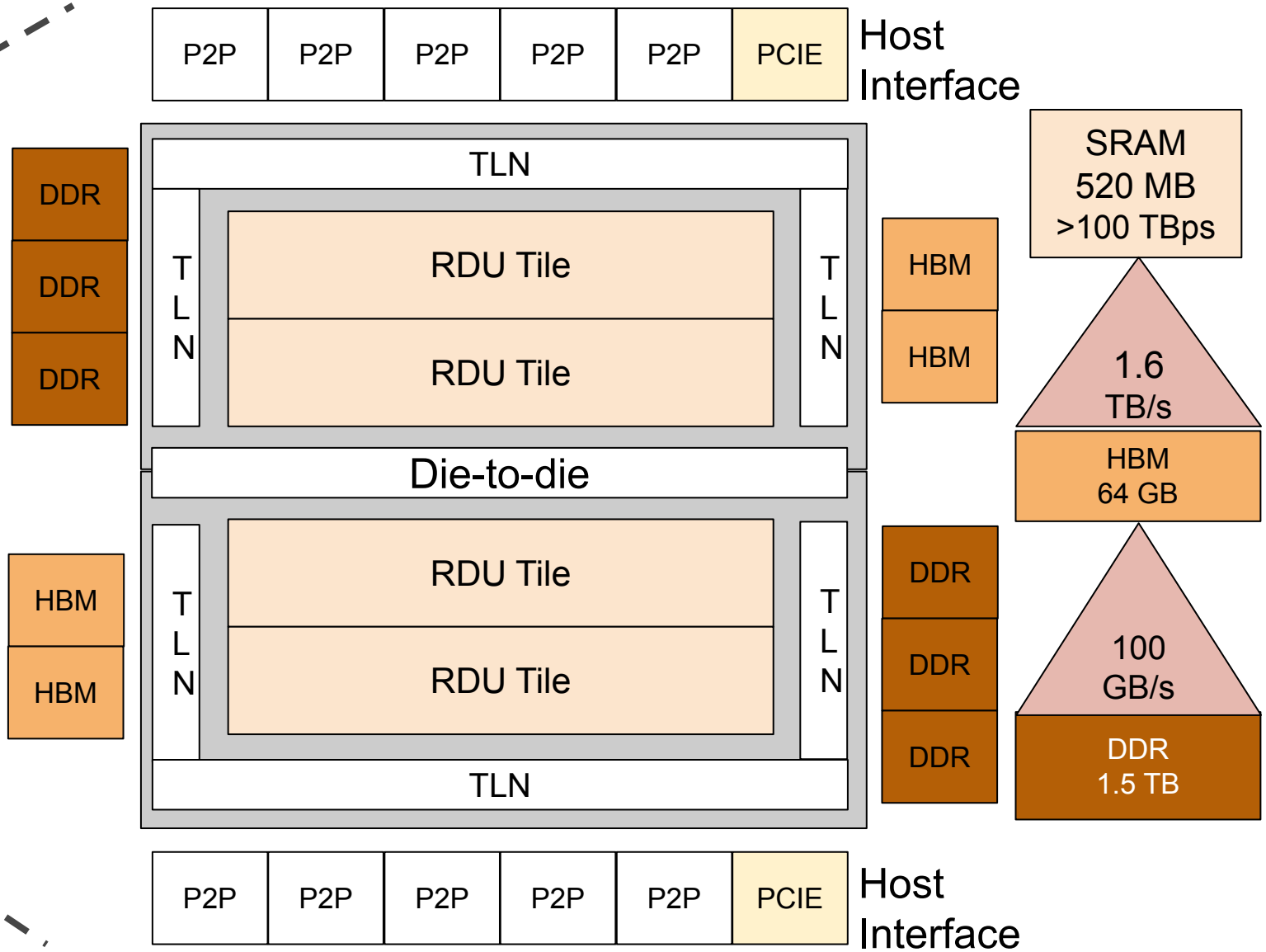
SN40L: Three Tier Memory Architecture

3-tier Memory System with SRAM, HBM, and DDR

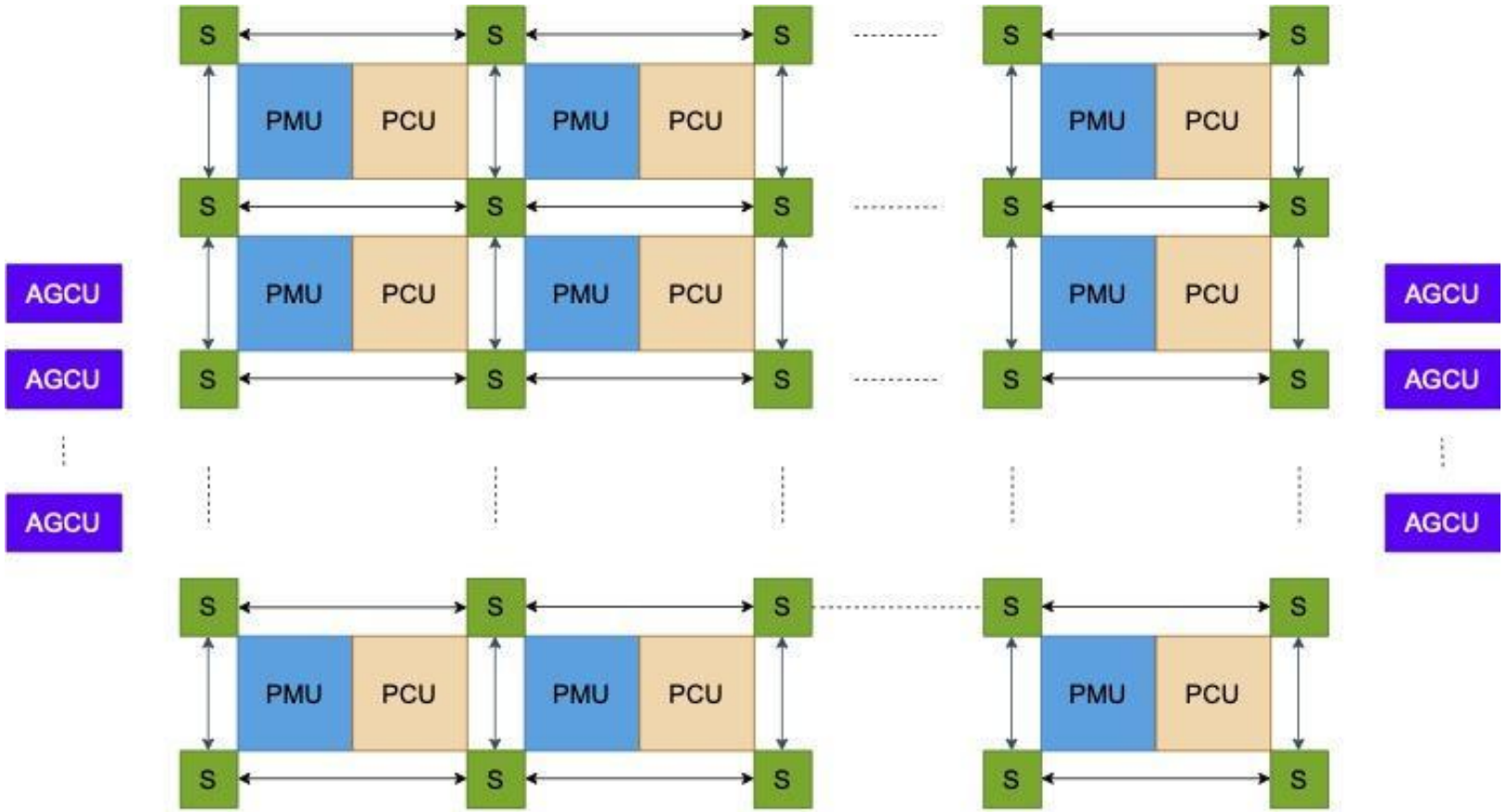


SambaNova SYSTEMS

SN40L Chip: Overview



SN40L: Tile Architecture



1040 PCUs and PMUs

PCU: Compute unit

PMU: Memory unit

S: Mesh switches

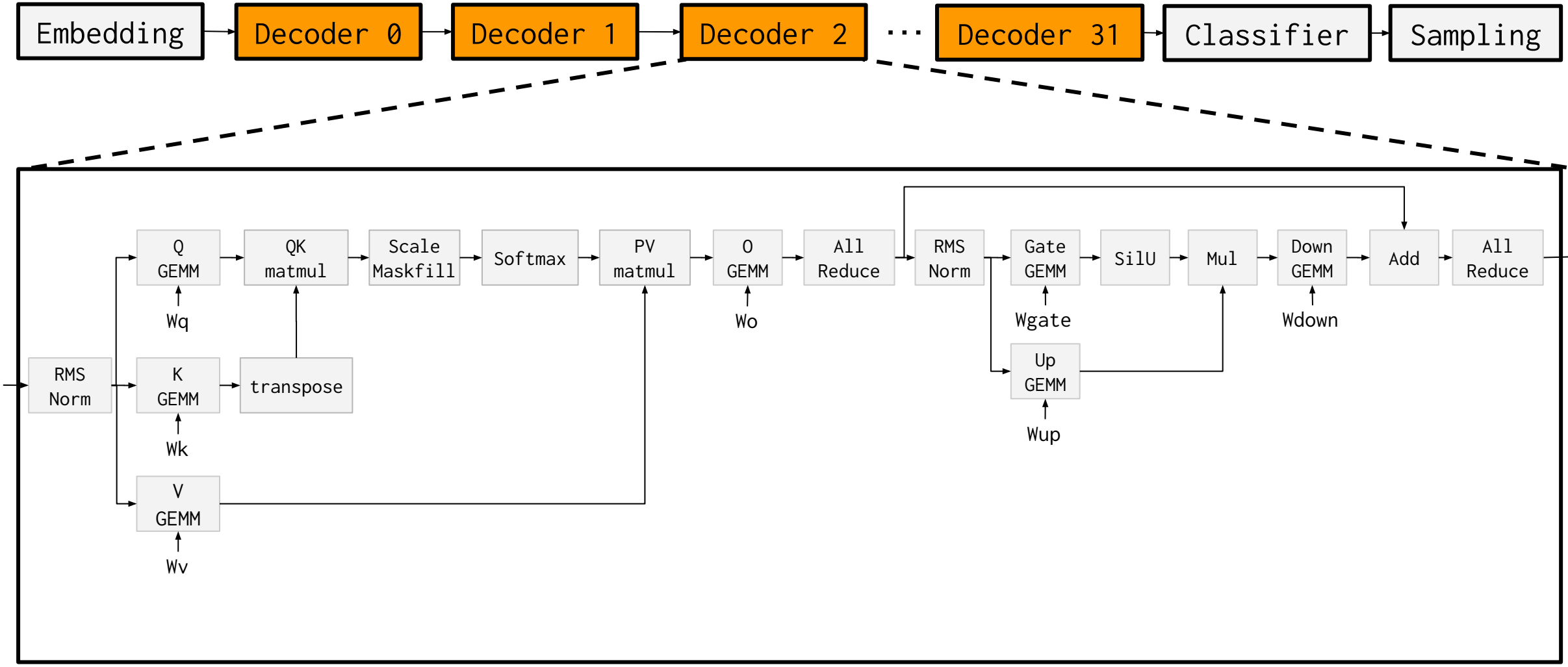
AGCU: Portal to off-chip memory and IO

Structure of a Transformer

Example: Llama3.1 8B

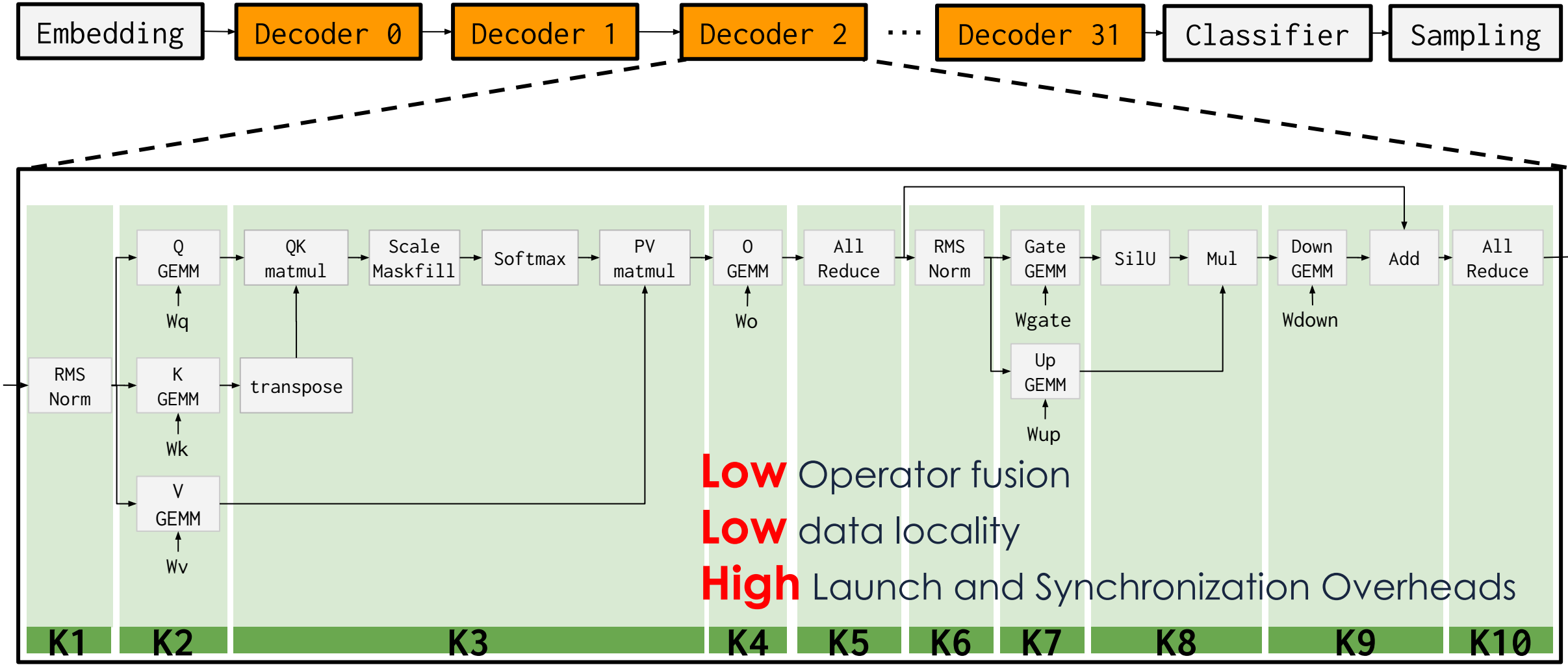
Structure of a Transformer

Example: Llama3.1 8B



GPUs Incur Overheads with Low Operator Fusion

Example: Llama3.1 8B



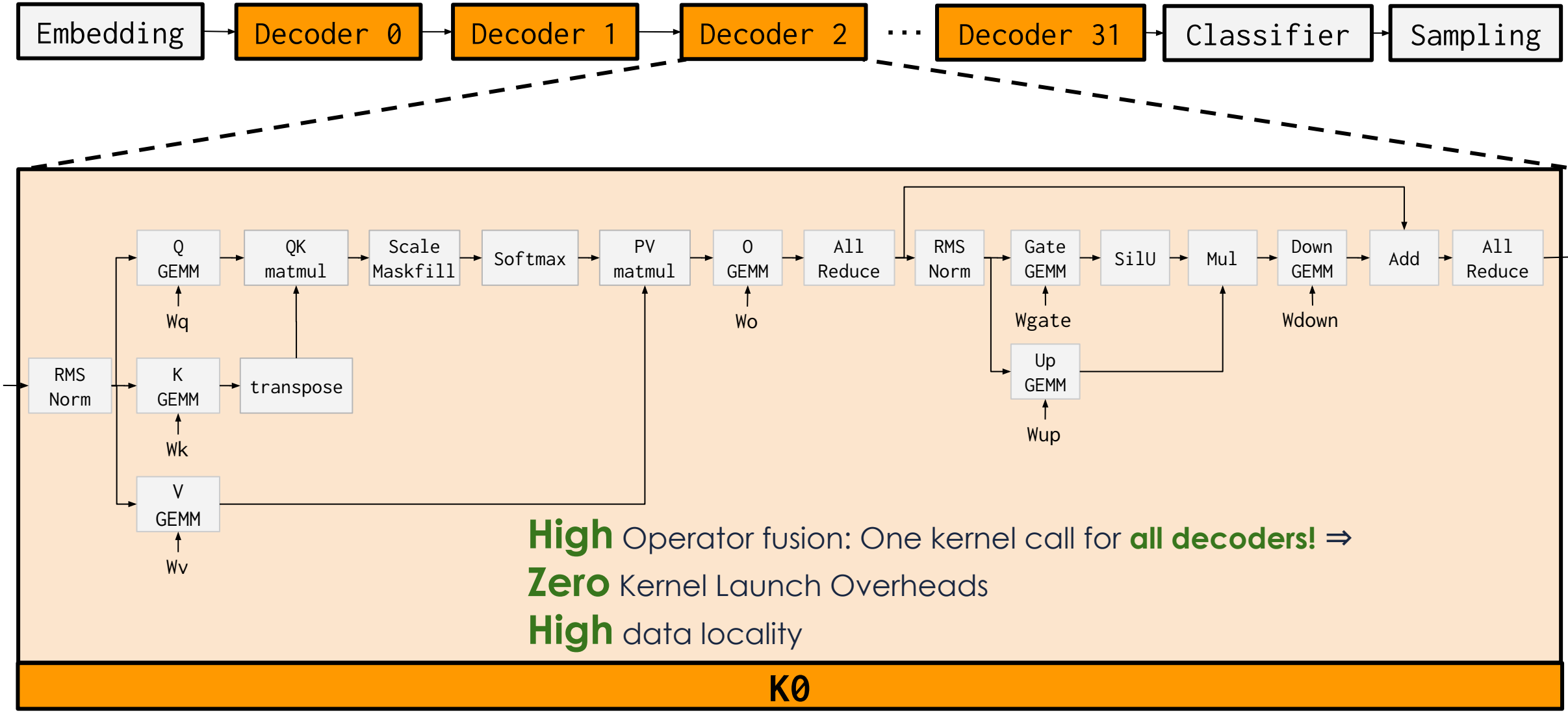
Low Operator fusion

Low data locality

High Launch and Synchronization Overheads

SN40L RDU Fuses the Entire Decoder into One Kernel!

Example: Llama3.1 8B



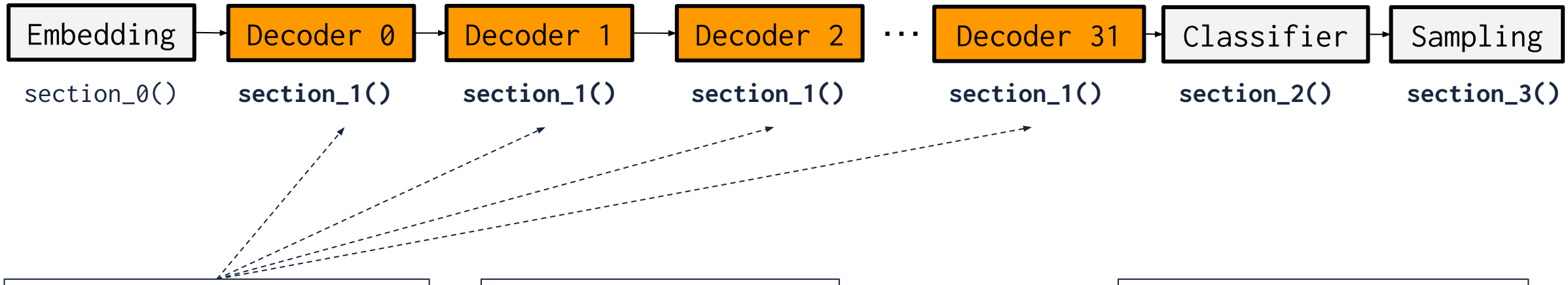
High Operator fusion: One kernel call for **all decoders!** ⇒

Zero Kernel Launch Overheads

High data locality

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

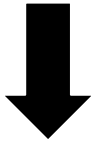
```
section_1(args[]):
  for i in range(32):
    Foo;
    Bar;

// 1 time
section_1(args)
```

1 kernel launch for all decoders!

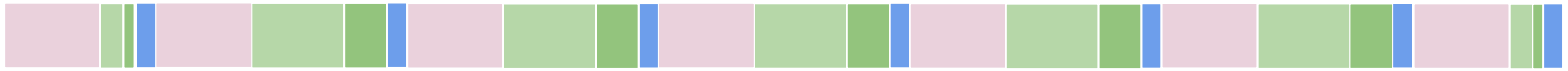
Executing Transformer on RDU

Example: Llama3.1 8B

Baseline: **100** tokens/s



+ Single Section Decoder



= **500** tokens/s
16 SN40L

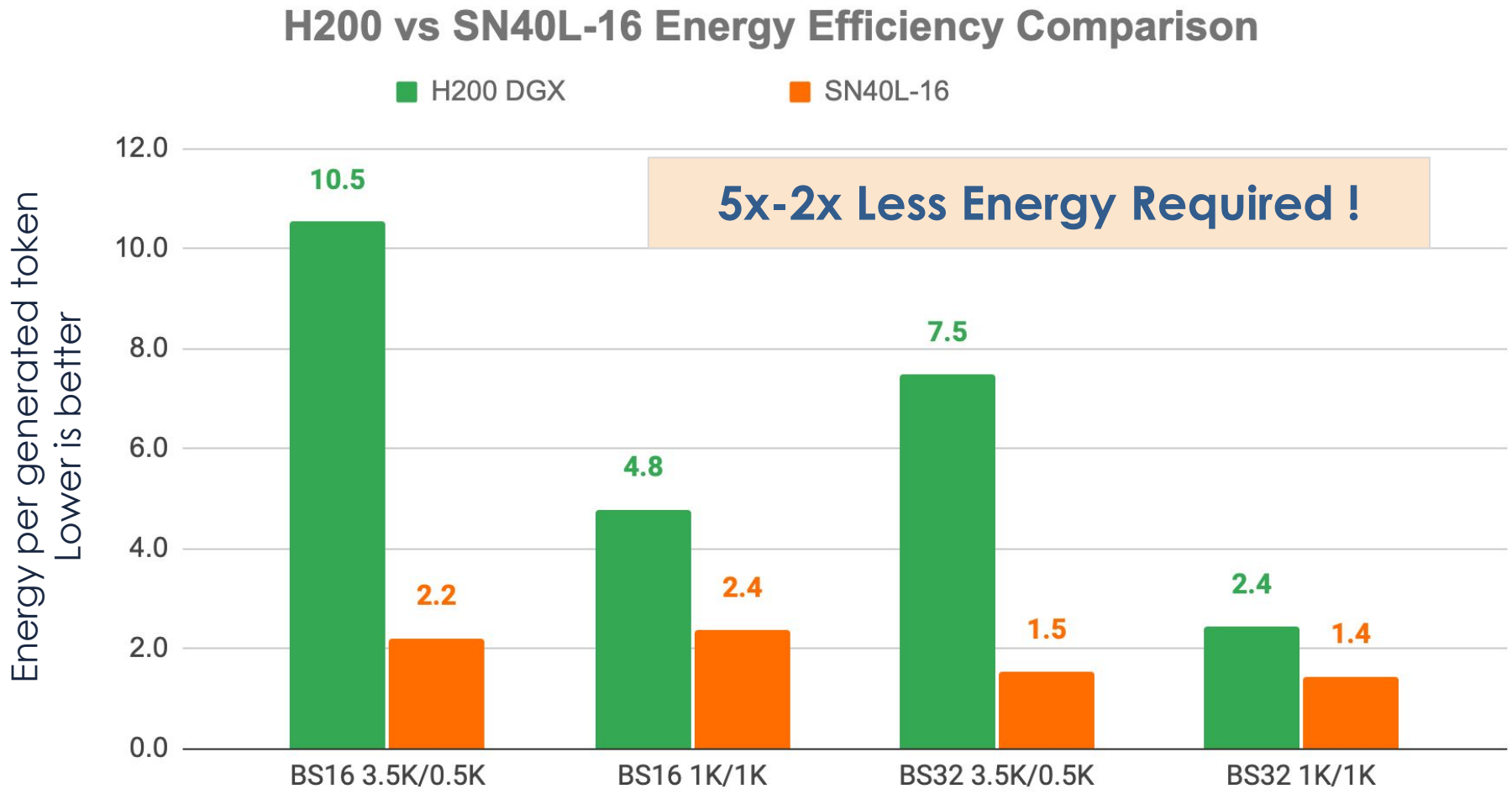
+ Kernel Looping

= **1170** tokens/s
16 SN40L

SambaNova SYSTEMS

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

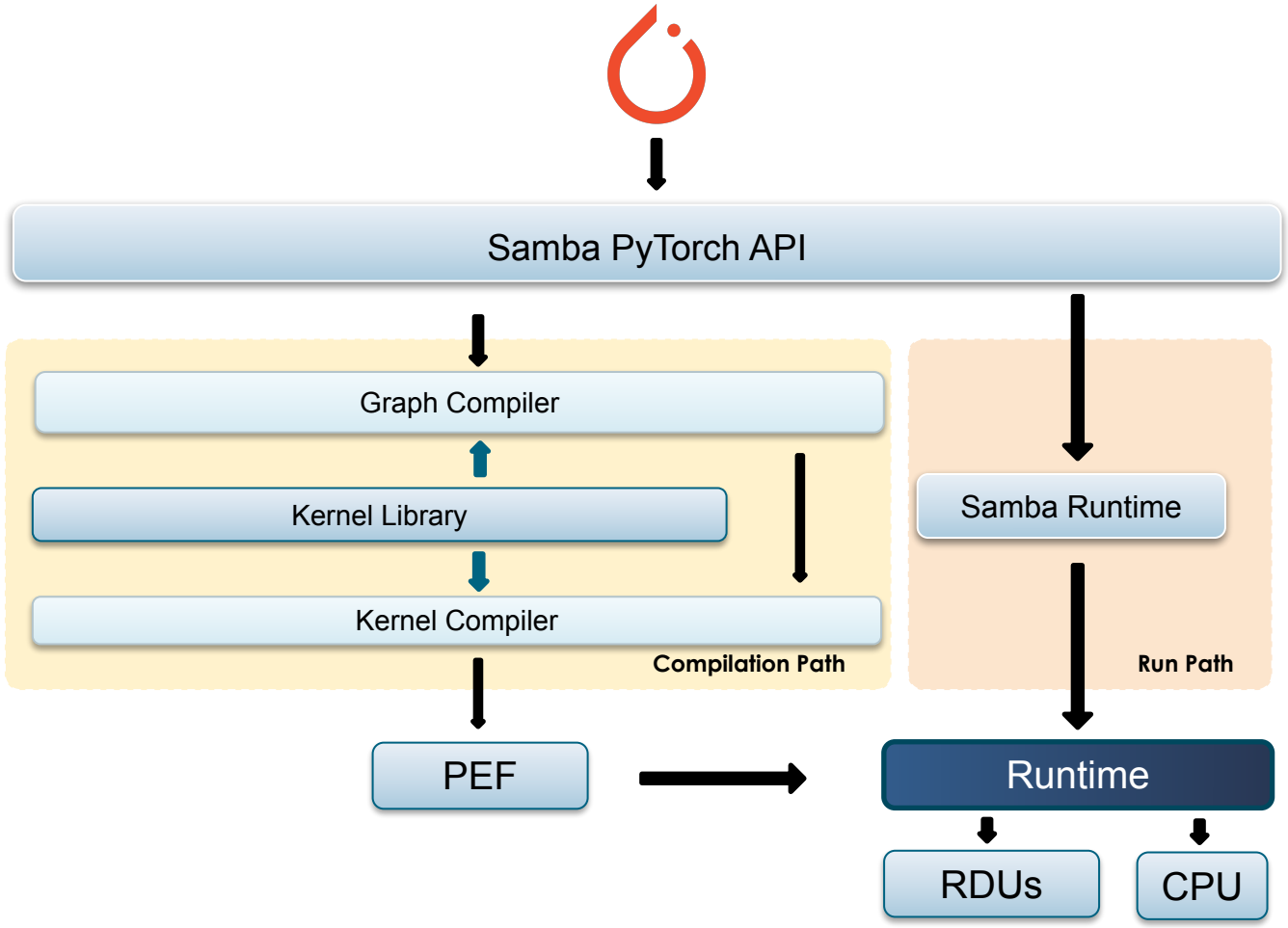
Llama 3.3 70B Inference



SambaNova SYSTEMS

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



PyTorch to SambaFlow: How Does It Work?

- Import your model from PyTorch
- Run **`samba.from_torch_model_(...)`** to convert model parameters to SambaTensors
  - Convert input Torch Tensors with **`samba.from_torch_tensor(...)`**
- Run **`samba.session.compile(...)`** to compile the model
  - Sometimes requires adaptations for compatibility
- Start running via **`samba.session.run(...)`** and **`samba.utils.trace_graph(...)`**

- Watch this video for more details!
  - Includes a ResFFNLogReg example

SambaNova SYSTEMS

SambaNova Suite

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

Model Bundling:

- Combine multiple models in a single composition of experts (CoE) 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

SDK/CLI

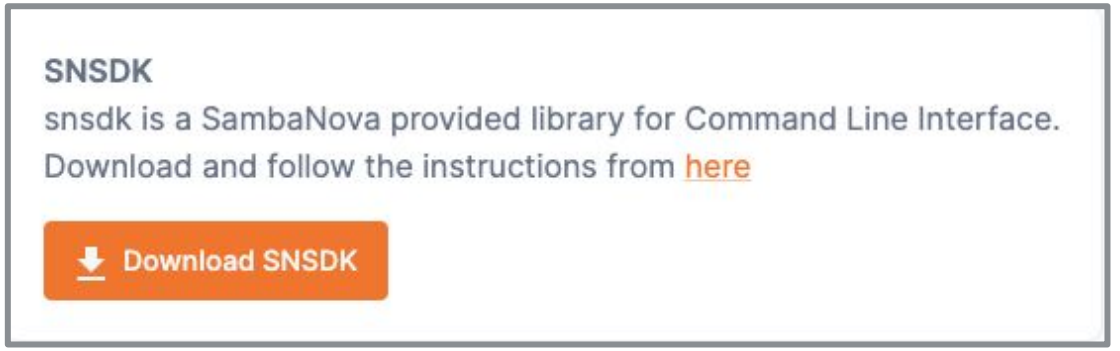
Our CLI and SDK tooling allows you to both **programmatically manage** SambaNova Suite and **build applications** on top of deployed endpoints.

SNSDK:

- Python SDK great for operationalizing endpoints deployed in your suite environment
- Supports OpenAI endpoint format

SNAPI:

- Programmatic interface for SambaNova Suite
- Can perform all actions available in the GUI
- Great for automating routine tasks



SambaNova SYSTEMS

AI Computing at Scale in Multiple US National Laboratories

- Fine-tuning and inferencing large language models to apply AI 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]

Example Models to Run at ALCF

DataScale - SN30					
NLP models	BERT, Llama 2, Llama 3, GenslM, Mistral, Deepseek Coder	**Vision models**	Unet2D, Unet 3D, DeepVIT, Vit, AutophaseNN	**Science models**	Uno

- Get more details on SambaNova Public Docs
  - SambaFlow developer documentation
  - SambaNova documentation at ALCF

SambaNova SYSTEMS

Example Models Available at ALCF

SambaNova Suite - SN40L					
Text and/or reasoning models	Llama 3.1 8B/70B/405B Llama 3.2 1B/3B Llama 3.3 70B Mistral-7B Mixtral-8x22B DeepSeek-R1/ V3 Qwen2.5-7B QwQ-32B	**Multimodal models**	Llama-3.2-11B/ 90B-Vision Qwen2-Audio-7B	**Embedding models**	E5 Large V2 E5-Mistral-7B

- Get more details on SambaNova Public Docs
  - SambaStudio

SambaNova SYSTEMS

SambaNova Cloud

SambaNova Cloud

Open-Source Models Available Today

Meta	Qwen3	deepseek	OpenAI	MISTRAL AI_
Meta Llama 4 Maverick Meta Llama 4 Scout Meta Llama 3.3 70B Meta Llama 3.2 1B/3B Meta Llama 3.1 8B/405B Meta Llama Guard 3-8B	Qwen3 32B QwQ 32B Qwen2-Audio-7B-Instruct	DeepSeek R1 671B DeepSeek V3 0324 DeepSeek-R1-Distill-Llama-70B	OpenAI 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 Cloud

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

Model	SambaNova
Llama 3.2 1B	2477
LLama 3.1 8B	1170
Llama 3.3 70B	458
Llama 3.1 405B	174
Llama 4 Maverick	785
DeepSeek R1 671B	263
DeepSeek V3 0324	263

SambaNova® SYSTEMS

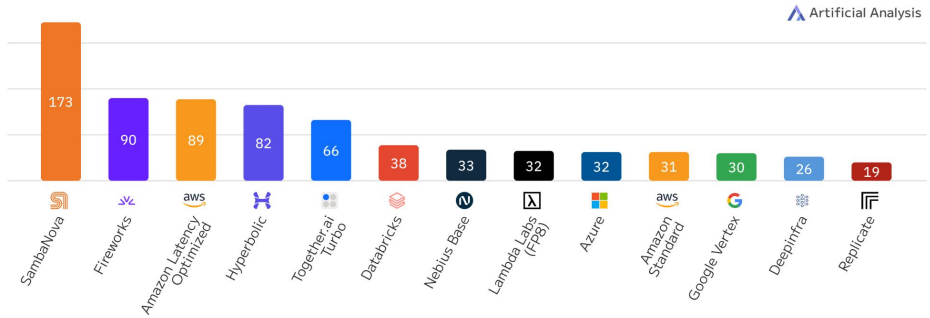
Independently Benchmarked by Artificial Analysis

Lightning Fast Inference for the Largest Models

Llama 3.1 405B

Output Speed: Llama 3.1 405B Providers

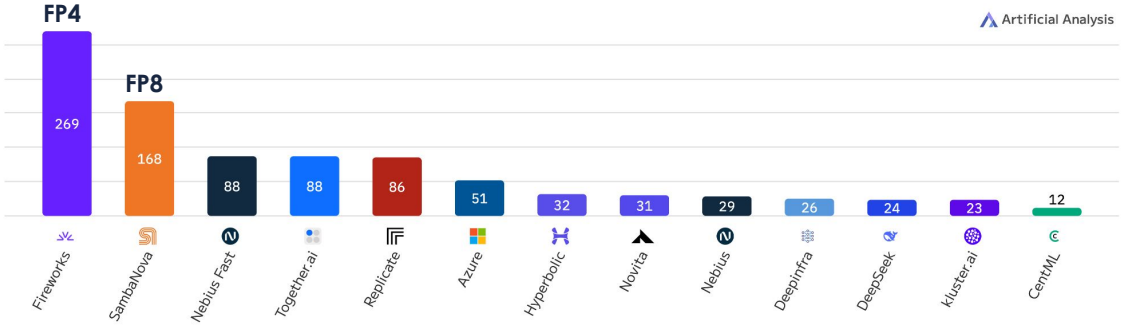
Output Tokens per Second; Higher is better



DeepSeek v3

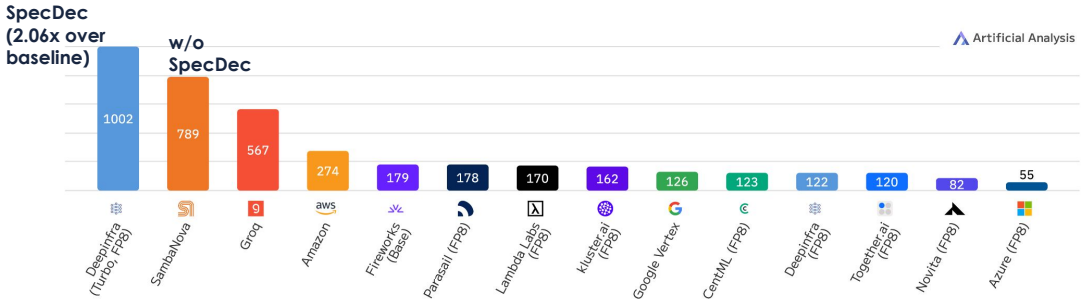
Output Speed: DeepSeek V3 0324 (Mar '25) Providers

Output Tokens per Second; Higher is better



Output Speed: Llama 4 Maverick Providers

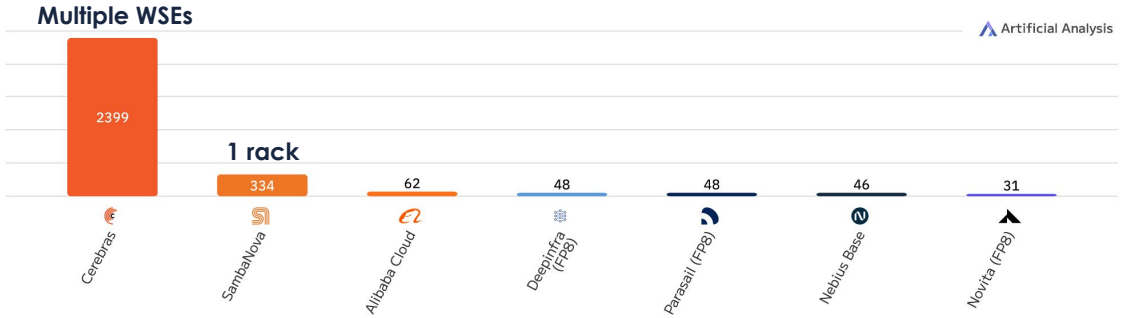
Output Tokens per Second; Higher is better



Llama 4 Maverick

Output Speed: Qwen3 32B (Reasoning) Providers

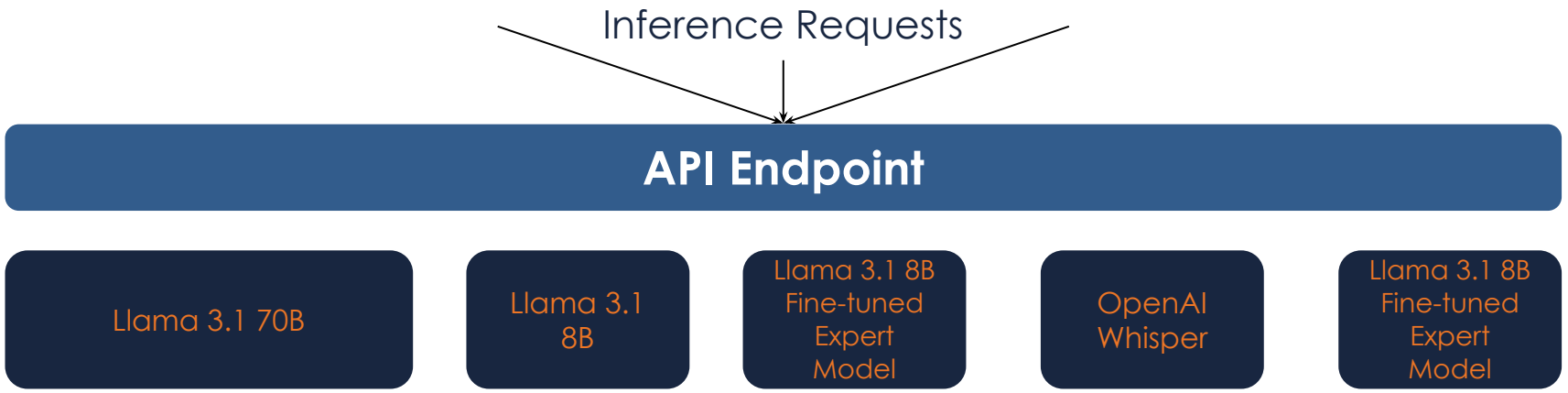
Output Tokens per Second; Higher is better



Qwen3 32B

SambaNova SYSTEMS

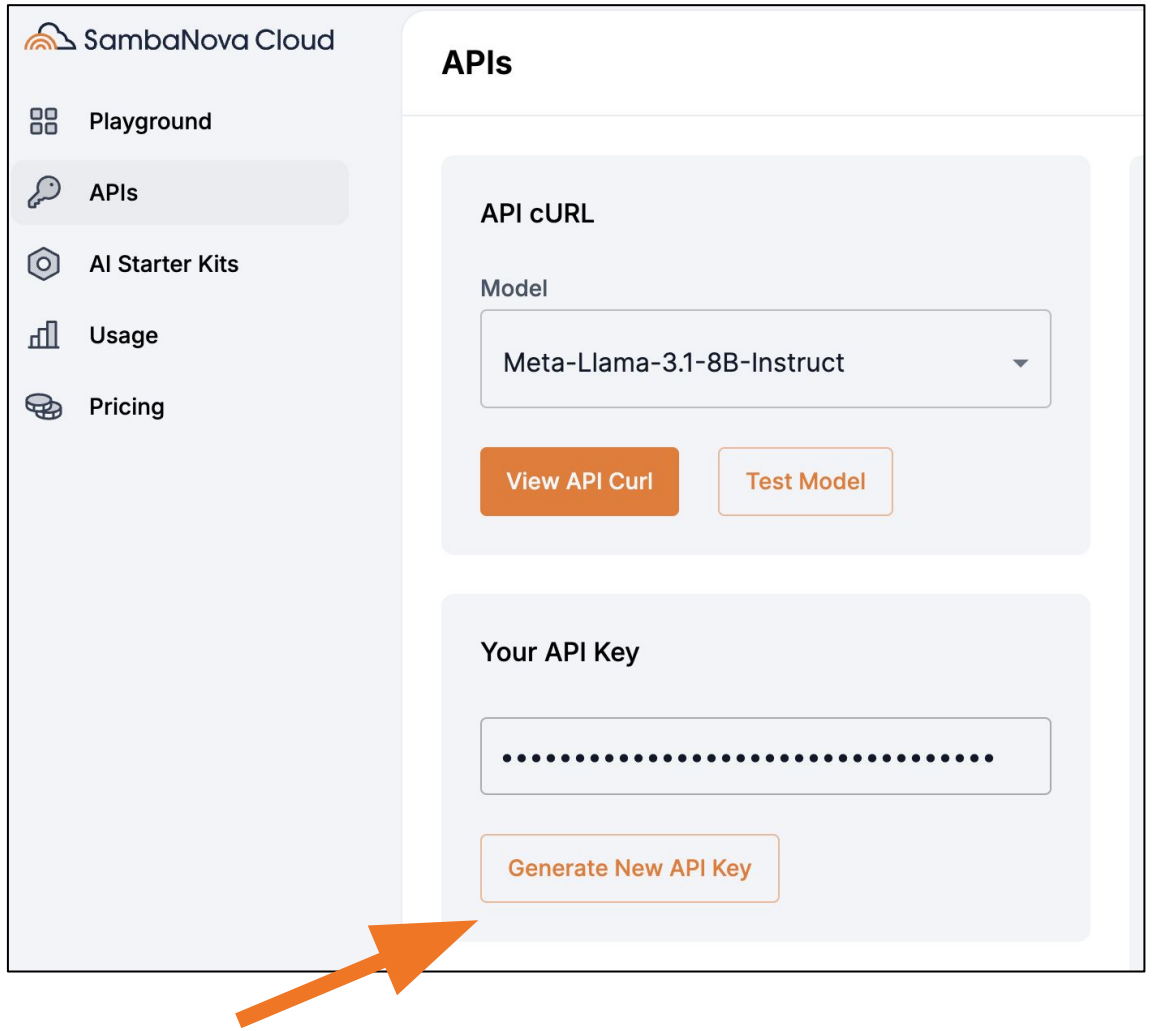
More Models with Less Hardware

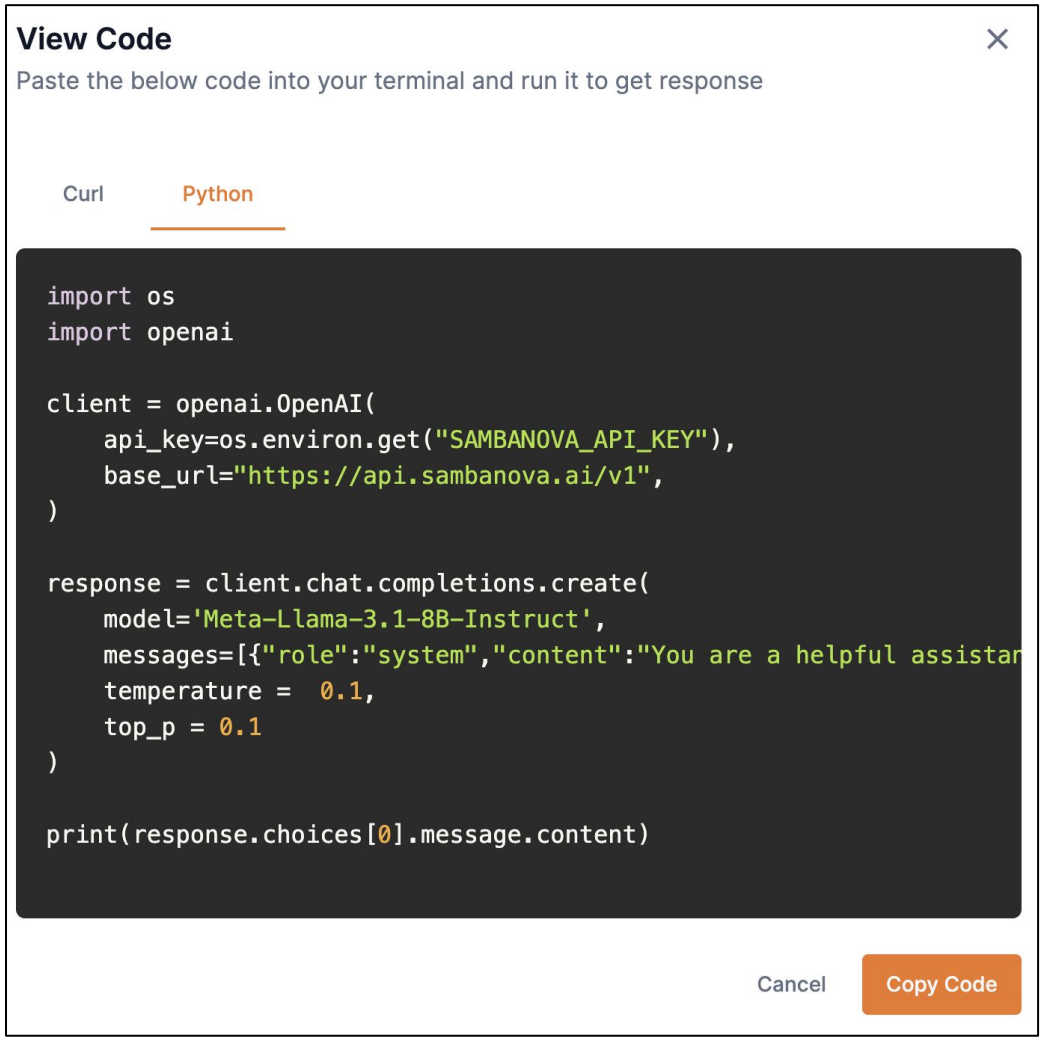


- Reduced startup costs
- Reduced serving costs - less HW more Models
- Simplify complex workflows
- Adding and maintaining new experts

SambaNova® SYSTEMS

Get Your API Key at cloud.sambanova.ai





SambaNova SYSTEMS

AI Starter Kits

SambaNova Cloud

- Playground
- APIs
- AI Starter Kits
- Usage
- Pricing

Login/Signup

Community

AI Starter Kits

Developer Community

View all starter kits on Github

Starter Kits Help You Build Fast – They bootstrap application development for common AI use cases with open-source Python code on SambaNova GitHub. They let you see how the code works, and customize it to your needs, so you can prove the business value of AI.

QuickStart - Cloud ReadMe

How to get started with SambaNova cloud – A comprehensive guide to help you begin your journey with SambaNova Cloud .

Getting Started

Function Calling

Tools calling implementation and generic function calling module – Enhance your AI applications with powerful function calling capabilities.

Advanced AI Capabilities

Demo

Financial Assistant

Enterprise-grade accuracy – Generate sophisticated, complex, and accurate responses by employing multiple agents in a chain to focus/decompose queries, generate multi-step answers, summarize them, and double-check accuracy.

Advanced AI Capabilities

Demo

Enterprise Knowledge Retrieval

Document Q&A on PDF, TXT, DOC, and more – Bootstrap your document Q&A application with this sample implementation of a Retrieval Augmented Generation semantic search workflow using the SambaNova platform, built with Python and a Streamlit UI.

Intelligent Information Retrieval

Demo

Search Assistant

Include web search results in responses – Expand your application's knowledge with this implementation of the semantic search workflow and prompt construction strategies, with configurable integrations with multiple SERP APIs.

Intelligent Information Retrieval

Demo

Benchmarking

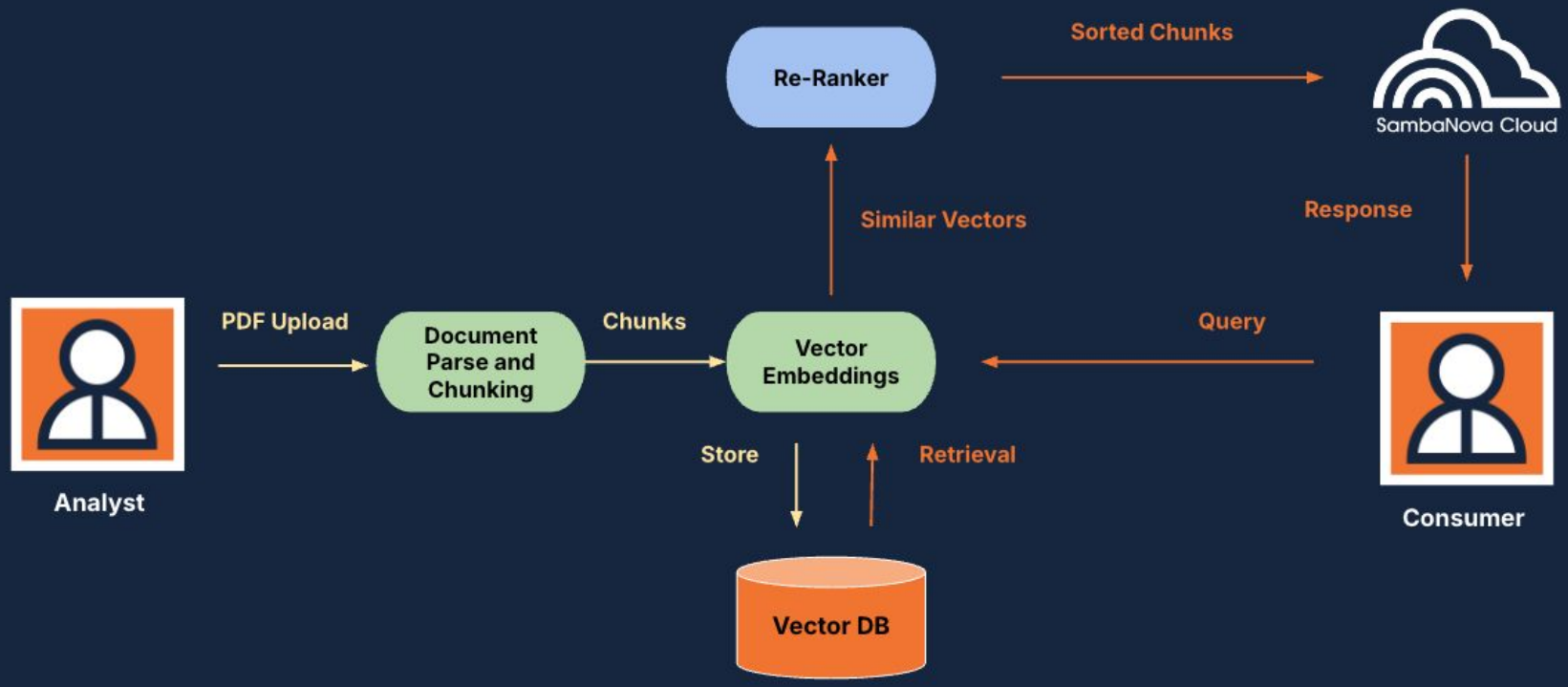
Compare model performance – Quickly determine which models meet your speed and quality needs by comparing model outputs, Time to First Token, End-to-End Latency, Throughput, Latency, and more with configuration options in a chat interface.

Model Development & Optimization

Demo

Retrieval Augmented Generation (RAG)

Bootstrap your document Q&A application with this sample implementation of a RAG semantic search workflow using the SambaNova platform, built with Python and a Streamlit UI.



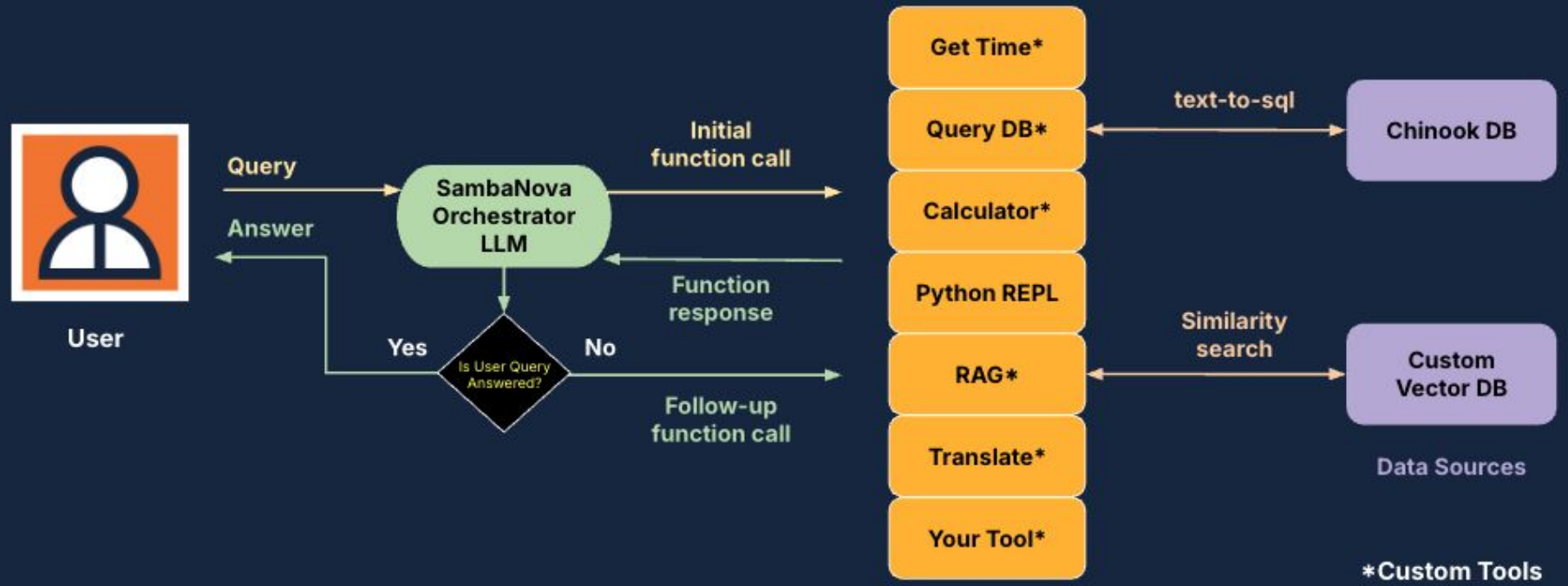
View Step-by-Step Instructions for This Kit on GitHub

View Demo-Video on YouTube

SambaNova SYSTEMS

Function Calling / Tool Use

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.



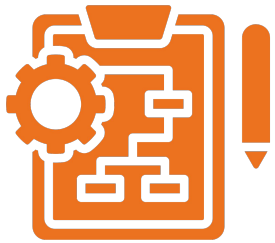
View Step-by-Step Instructions for This Kit on GitHub

View Demo-Video on YouTube

Agentic Applications

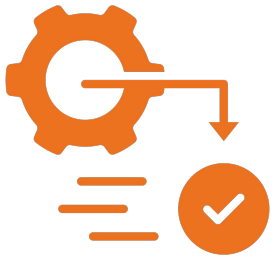
Input Objective

User provides high-level goals



Planning

Agent decomposes goals into tasks



Execution

Agent performs or delegates tasks

Feedback and Learning

System improves through iteration

✅ **Reduces manual intervention**

✅ **Increases efficiency and speed**

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

Deep Research Agents Demo (Try it yourself!)

SambaNova SYSTEMS

Integration with External Frameworks / Platforms

LLM Frameworks	Hugging Face, LangChain, LiteLLM, LlamaIndex
Agent Frameworks	Agno, AG, crewai, Semantic Kernel
Coding Assistants	CLINE V3.4, Continue, windsurf
Low-code platforms, vector DBs, and voice libraries	Langflow, Dify, milvus, Eleven Labs

For complete list, visit our Integrations page

SambaNova SYSTEMS

THANK YOU!

petro-junior.milan@sambanova.ai

SambaNova® SYSTEMS

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"